ADVANCING CONVOLUTIONAL NEURAL NETWORKS: NOVEL TECHNIQUES AND EVALUATIONS FOR ENHANCED ROBUSTNESS AND GENERALIZATION

ARTEM PILZAK

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the
Doctor of Philosophy in Experimental Psychology

School of Psychology
Faculty of Social Sciences
University of Ottawa

Table of Contents

List of Figures

List of Tables

Table of Acronyms

(Sorted alphabetically)

ANN	Artificial Neural Network
CNN	Convolutional Neural Network
CWA	Channel-Wise Attention
DAL	Dynamic Attention Layer
DFM	Dominant Feature Masking
DG	Domain Generalization
DORA	Discovery of Relations by Analogy
FCN	Fully Convolutional Networks
FGSM	Fast Gradient Sign Method
GAN	Generative Adversarial Network
Grad-CAM	Gradient-weighted Class Activation Mapping
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
IRM	Invariant Risk Minimization
LISA	Learning and Inference with Schemas and Analogy
MNIST	Modified National Institute of Standards and Technology
MSP	Maximum Softmax Probability
NNF	Noisy/Noise-Free
OOD	Out-of-Distribution
PGD	Projected Gradient Descent
ReLU	Rectified Linear Unit
RICAP	Random Image Cropping and Patching
RSC	Representation Self-Challenging
SENet	Squeeze-and-Excitation Networks
SIFT	Scale-Invariant Feature Transform
SVM	Support Vector Machines

VEB	Versatile Evaluation Benchmark
YOLO	You Only Look Once

Acknowledgements

This thesis is the culmination of five years of research and represents one of the most challenging yet rewarding experiences of my life. I am immensely proud of this work and grateful to everyone who has supported me throughout this journey.

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Jean-Philippe Thivierge. The past seven years working under his guidance have been incredibly rewarding. I began as a volunteer in his lab, completed my honours thesis under his supervision, and continued through to my doctoral studies. Dr. Thivierge's unwavering support, motivation, and insightful feedback have been instrumental in the completion of this thesis. His mentorship has not only shaped my research skills but also fostered my growth as a scientist. Without his guidance and encouragement, this work would not have been possible.

I would also like to extend my sincere thanks to my thesis committee: Dr. Sylvain Chartier, Dr. Denis Cousineau, and Dr. Charles Collin. Their valuable input and constructive criticism have greatly improved the quality of this research.

I am grateful to my colleagues in the ComBiNe lab for their collaboration and support. Special thanks to Camille Godin and Megan Boucher-Routier for their contributions and stimulating discussions that have enriched my research experience. I also extend my appreciation to other lab members whose support has been invaluable.

I am deeply thankful to my parents, Victoria Vladimirova and Oleg Pilzak, for their unwavering support, motivation, and belief in me. From the very beginning, they have been my biggest advocates, providing me with the strength and confidence to pursue my dreams. Their constant encouragement has been the backbone of my academic journey, offering me the emotional and practical support necessary to overcome the many challenges I encountered. They have always been there to celebrate my successes and to provide comfort during difficult times. Without their love, guidance, and sacrifices, I would not

have come this far. Their faith in my abilities has been a driving force, pushing me to strive for excellence and never give up. I owe my accomplishments to their enduring support and dedication.

I also want to thank my eldest brother, George Pilzak, for being a source of inspiration and support throughout my studies. His guidance and encouragement have been invaluable. His insights and advice have often provided clarity during complex situations, and his unwavering belief in my potential has been a constant source of motivation. George's support has significantly contributed to my growth both personally and academically.

Most importantly, I want to express my deepest gratitude to my wife, Marie Anne Campagna, for her unwavering support, unconditional love, and for always being my number one cheerleader. Throughout my academic journey, she celebrated my successes and stood by me during the most challenging times. Whenever my research faced obstacles, she was there to provide the support and motivation I needed to push forward. Her belief in me has been a driving force, helping me to overcome difficulties and achieve my goals. Marie Anne has been my steadfast partner for over 10 years, sharing both the joyful moments and the struggles. Her patience, understanding, and encouragement have been invaluable. I am forever grateful for her presence in my life, and I am more than happy to share the next chapters with her, looking forward to many more years of love, growth, and partnership.

Preface

Published Work

This thesis document is based on the following published work:

Pilzak, A., & Thivierge, J. P. (2022). Generating Robust Convolutional Networks by Injecting Partial Noise in the Training Data. In 2022 *IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)* (pp. 1-5). IEEE.

***Contributions*: A.P.** conceptualized the project, curated and investigated the data, developed the methodology and software, and performed the visualizations and validations. **A.P.** also wrote the original draft. J.P.T. supervised the project, secured funding, validated the methodology and results, and contributed to reviewing and editing the manuscript. Both authors discussed the results, revised the manuscript, and approved the final version.

As well as the following submitted work:

Pilzak, A., & Thivierge, J.P. (2024). Enhancing Out-of-Distribution Learning in Computer Vision through Dominant Feature Masking. [Manuscript submitted for peer review]

***Contributions*: A.P.** was responsible for the project's conceptualization, data curation, and investigation, and developed the methodology and software. Additionally, **A.P.** performed the visualizations and validations and wrote the initial draft. J.P.T. provided project supervision, secured funding, validated the methodology and results, and contributed to reviewing and editing the manuscript. The authors collaboratively discussed the results, revised the manuscript, and gave final approval.

Pilzak, A., & Thivierge, J.P. (2024, December). Improving Generalization in Convolutional Neural Networks with a Dynamic Attention Layer. In 2024 International Conference on Sustainable Technology and

Engineering (i-COSTE) (TBA). IEEE.

***Contributions*: A.P.** took the lead in conceptualizing the project, managing data curation and investigation, and developing the methodology and software. **A.P.** also carried out the visualizations and validations, and authored the original draft. J.P.T. supervised the project, obtained funding, ensured the validation of the methodology and results, and contributed to the manuscript's review and editing. Both authors engaged in discussions of the results, revised the manuscript, and approved the final version.

Abstract

This thesis advances the field of computer vision by addressing the critical challenges of out-of-distribution (OOD) generalization and robustness. Computer vision systems, particularly those based on convolutional neural networks (CNNs), have shown remarkable performance in controlled settings with in-distribution data. However, their ability to generalize to new, unseen data that differs significantly from the training set remains a substantial obstacle. This issue is especially pertinent in real-world applications where models frequently encounter data with varying conditions, such as changes in lighting, occlusions, or entirely new objects and scenarios. Ensuring that these systems can maintain high performance and reliability under such diverse conditions is crucial for their deployment in critical areas like autonomous driving, healthcare, and surveillance. Through a comprehensive study, this research explores innovative methodologies to enhance the robustness and generalization capabilities of CNNs, providing practical solutions to these challenges. To address this issue, we introduce four novel contributions to the field. First, the "noisy/noise-free" (NNF) training method improves the generalization capabilities of CNNs by incorporating a balanced mix of noisy and noise-free images in the training set, significantly enhancing overall generalization and classification accuracy on noisy datasets. Second, we present Dominant Feature Masking (DFM), a data augmentation approach that strategically conceals dominant features within images to encourage the network to learn both primary and secondary attributes, thus improving OOD prediction performance. Third, we develop the Versatile Evaluation Benchmark (VEB), a comprehensive evaluation framework that rigorously tests the domain generalization and OOD performance of CNNs across various challenging scenarios. Fourth, the thesis introduces the Dynamic Attention Layer (DAL), a novel layer that dynamically adjusts attention weights based on feature relevance during training, allowing CNNs to focus on both dominant and secondary features, thereby enhancing their robustness and adaptability. Our findings demonstrate significant improvements in the generalization performance of CNNs, particularly in handling diverse and unpredictable scenarios. The research offers valuable insights and practical solutions for developing more robust and reliable computer vision systems

Chapter 1: General Introduction

History of Computer Vision

Computer vision, a subfield of artificial intelligence, aims to enable machines to interpret and understand the visual world (Szeliski, 2010; Goodfellow, Bengio, & Courville, 2016). The history of computer vision is rich and spans several decades, evolving from early theoretical explorations to practical applications that impact numerous industries today.

Early Beginnings (1950s-1960s)

The origins of computer vision can be traced back to the 1950s and 1960s when researchers began to explore the possibility of automating image recognition and interpretation. The initial focus was on understanding how the human visual system works and attempting to replicate these processes in machines. During this period, researchers primarily worked on simple image processing tasks such as edge detection, pattern recognition, and the segmentation of images into meaningful parts (Roberts, 1965).

One of the pioneering works in this era was by Lawrence Roberts in 1963, who developed the block world model (Roberts, 1965). This model was an early attempt to interpret three-dimensional structures from two-dimensional images. The block world model represented objects in a scene as combinations of simple geometric shapes, such as cubes and rectangular prisms, which could be easily processed by early computer algorithms. Roberts' work helped establish fundamental principles that guided subsequent research in computer vision. His Ph.D. thesis is often considered one of the first comprehensive studies in the field, focusing on deriving three-dimensional information from two-dimensional photographs. This approach laid the groundwork for future developments in 3D computer vision and object recognition.

Formative Years (1970s-1980s)

The 1970s and 1980s marked significant progress in computer vision, driven by advancements in

computing power and the development of new algorithms. Researchers began to explore more complex tasks, such as motion analysis, stereo vision, and object recognition (Ballard & Brown, 1982). Notable milestones during this period include David Marr's theory of vision, which proposed a computational approach to understanding visual perception. Marr's work introduced three levels of analysis for understanding vision: the computational level, the algorithmic level, and the implementational level, emphasizing the importance of both the algorithmic and implementational aspects in developing computer vision systems (Marr, 2010).

Marr's theory of vision provided a structured framework for analyzing visual information processing. The computational level addresses the 'what' and 'why' questions: what is the goal of the visual process, and why is it necessary? At this level, the focus is on understanding the purpose of visual processing and the nature of the tasks it performs, such as recognizing objects or navigating through an environment (Marr, 2010).

The algorithmic level deals with the 'how' question: how can the visual tasks be performed? This involves specifying the algorithms and representations used to solve the problems identified at the computational level (Marr, 2010). It includes defining the processes for edge detection, segmentation, and pattern recognition, which transform raw visual inputs into meaningful structures and objects.

The implementational level concerns the 'where' question: where in the physical world do these processes occur? This level focuses on the physical realization of the algorithms, whether in biological systems, like the human brain, or in artificial systems, like computer hardware (Marr, 2010). It addresses the specifics of how visual processing is physically carried out, including the neural mechanisms in the brain or the design of computer circuits and processors.

By distinguishing these three levels, Marr's theory provided a comprehensive framework that guided research in computer vision, bridging the gap between understanding visual perception and developing practical algorithms and systems. This approach underscored the need to consider all levels of

analysis to create effective and efficient computer vision models.

Another critical development was the introduction of the "feature extraction" technique, which involves identifying and isolating significant parts of an image, such as edges, corners, and textures. The Scale-Invariant Feature Transform (SIFT) algorithm (Lowe, 1999), developed by David Lowe in 1999, became one of the most widely used feature extraction methods, enabling robust object recognition even under varying scales and orientations.

During this period, Kunihiko Fukushima developed the "neocognitron" in 1980, a hierarchical, multilayered artificial neural network inspired by the visual cortex of the brain (Fukushima, 1980; Fukushima, 1988). The neocognitron was designed to mimic the structure and function of the human visual system, incorporating alternating layers of "S-cells" for local feature extraction and "C-cells" for spatial pooling. This architecture allowed the neocognitron to detect simple patterns such as edges and corners in the initial layers and progressively build more complex and invariant representations in the deeper layers. By simulating the hierarchical processing of visual information, the neocognitron laid the foundational concepts for modern CNNs, significantly influencing subsequent advancements in computer vision.

The Rise of Machine Learning (1990s-2000s)

The 1990s and 2000s witnessed a paradigm shift in computer vision with the advent of machine learning techniques. Instead of manually designing features and rules for image interpretation, researchers began to train models on large datasets to learn patterns and representations. The development of support vector machines (SVMs) and other machine learning algorithms provided powerful tools for classification and regression tasks in computer vision (Cortes & Vapnik, 1995; Freund & Schapire, 1997).

During this era, there was also a significant focus on developing methods for face recognition, a challenging problem with numerous practical applications. The introduction of the Viola-Jones object

detection framework in 2001 (Viola & Jones, 2001), which utilized Haar-like features and a cascade of classifiers, revolutionized real-time face detection and laid the groundwork for many subsequent advancements. Haar-like features, which are digital image features used to capture texture differences, enabled rapid and efficient computation using an integral image (Lienhart & Maydt, 2002). This framework's ability to perform real-time detection on standard hardware significantly advanced the field of computer vision and inspired numerous subsequent innovations in object recognition.

Deep Learning Revolution (2010s-Present)

The 2010s marked the beginning of the deep learning revolution, which transformed computer vision by leveraging the power of deep neural networks (LeCun, Bengio, & Hinton, 2015; Schmidhuber, 2015). Deep learning is a subset of machine learning that involves training artificial neural networks with many layers (hence "deep") to learn representations of data with multiple levels of abstraction. These neural networks can automatically learn features from raw data, significantly improving the performance of models on complex tasks (LeCun, Bengio, & Hinton, 2015; Goodfellow, Bengio, & Courville, 2016).

The breakthrough moment for deep learning in computer vision came in 2012 when Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) with their deep convolutional neural network (CNN) model, AlexNet (Krizhevsky, Sutskever, & Hinton, 2012). AlexNet significantly outperformed traditional methods and demonstrated the potential of deep learning for complex visual tasks. The success of AlexNet showcased the effectiveness of deep learning in handling large-scale image datasets and paved the way for rapid advancements in the field. Its deep architecture, combined with innovative techniques like data augmentation (enhancing training data diversity) and efficient training algorithms, significantly improved image classification accuracy. This demonstrated the potential of deep neural networks for complex visual tasks, setting a new standard in computer vision research.

It is also important to acknowledge the contributions of Yann LeCun, one of the pioneers of deep

learning and CNNs. In the late 1980s and early 1990s, LeCun developed the concept of convolutional neural networks, a type of artificial neural network designed to process structured grid data like image, and applied them to handwritten digit recognition, a precursor to modern deep learning applications (LeCun et al., 1989). His work on the LeNet architecture (LeCun et al., 1998) laid the foundation for later developments in deep learning and computer vision, including the breakthroughs achieved by AlexNet.

Since then, deep learning has become the dominant approach in computer vision, leading to rapid advancements and new applications. CNNs have been particularly successful in tasks such as image classification, object detection, and semantic segmentation (He et al., 2016; Szegedy et al., 2015). The development of architectures like VGGNet, GoogLeNet, ResNet, and more recently, transformer-based models like Vision Transformers, has pushed the boundaries of what is possible in computer vision (Simonyan & Zisserman, 2014; Szegedy et al., 2015; He et al., 2016; Dosovitskiy et al., 2021). Transformer-based models, originally designed for natural language processing, use self-attention mechanisms, which allow the model to dynamically focus on different parts of the input data by calculating attention scores that indicate the relevance of each part to others. This enables more effective feature extraction and improved performance in vision tasks (Vaswani et al., 2017). Deep learning has enabled significant progress in areas such as medical imaging, autonomous driving, and augmented reality (Litjens et al., 2017; Grigorescu et al., 2020; Azuma, 2016). Techniques like generative adversarial networks (GANs), which consist of two neural networks competing with each other to generate realistic images, have also opened up new possibilities for image synthesis and manipulation (Goodfellow et al., 2014). By enabling the creation of high-quality synthetic images, GANs have revolutionized fields like data augmentation and artistic creation.

Current Trends in Computer Vision

Today, computer vision continues to evolve rapidly, driven by advances in hardware, algorithms, and the availability of large-scale datasets. Research is increasingly focusing on areas such as:

- 3D Vision: Techniques for reconstructing and understanding three-dimensional scenes from images and videos.
- Self-supervised Learning: Methods for training models with minimal labeled data, leveraging large amounts of unlabeled data.
- Real-time Processing: Efficient algorithms and hardware for real-time image and video analysis, crucial for applications like robotics and augmented reality.
- Explainability: Developing methods to make computer vision models more interpretable and transparent, addressing concerns about their black-box nature.

As computer vision technology continues to advance, it is poised to revolutionize various fields, including healthcare, transportation, retail, and entertainment. The integration of computer vision with other AI technologies, such as natural language processing and robotics, holds great promise for creating intelligent systems capable of interacting with the world in more sophisticated and human-like ways.

Artificial Neural Networks and Human Selective Attention

The relationship between neural networks and human cognition has been a subject of interest in both neuroscience and artificial intelligence research. One area where this connection is particularly evident is in the concept of selective attention, which is the cognitive process of focusing on certain stimuli while ignoring others (Posner & Petersen, 1990). This concept has inspired various mechanisms in neural network models, enhancing their ability to process and prioritize information efficiently.

Selective attention allows humans to manage the vast amount of sensory information received from the environment by filtering out irrelevant details and focusing on what is important (Posner & Petersen, 1990). This process is crucial for tasks such as recognizing objects in a cluttered scene, following a conversation in a noisy room, and performing complex activities that require sustained focus. The brain's

ability to selectively attend to certain stimuli is mediated by networks of neurons that dynamically adjust their activity based on the relevance and significance of the information being processed (Posner & Petersen, 1990).

Research has shown that selective attention operates through multiple mechanisms, including bottom-up and top-down processes (Desimone & Duncan, 1995). Bottom-up attention is driven by the properties of the stimuli, such as brightness, color, or motion, which automatically capture our attention. In contrast, top-down attention is guided by our goals, expectations, and prior knowledge, allowing us to focus on specific aspects of a scene while ignoring others. These mechanisms work together to enable efficient and adaptive perception in dynamic environments (Desimone & Duncan, 1995).

Incorporation of Selective Attention in Neural Networks

Inspired by the human cognitive process of selective attention, researchers have developed attention mechanisms in neural networks to improve their performance on various tasks. These mechanisms allow neural networks to dynamically weight the importance of different parts of the input data, enhancing their ability to focus on relevant features and ignore irrelevant ones.

One of the most influential attention mechanisms is the self-attention mechanism, introduced in the context of natural language processing but also applicable to computer vision (Vaswani et al., 2017). Self-attention enables the network to weigh the contribution of each input element relative to all other elements, allowing it to focus on the most pertinent parts of the data. This mechanism has been integral to the development of transformer models, which have achieved state-of-the-art performance in various tasks (Parmar et al., 2018; Tay et al., 2021).

In computer vision, attention mechanisms can be applied to tasks such as image classification, object detection, and segmentation. For example, in image classification, an attention mechanism can help the network focus on the most distinctive features of the object, improving accuracy (Wang et al., 2017). In object detection, attention can aid in identifying and localizing objects within complex scenes by

prioritizing regions of the image that are more likely to contain objects of interest (Zhou et al., 2019).

Benefits of Attention Mechanisms in Neural Networks

The integration of attention mechanisms into neural networks offers several advantages:

- Improved Accuracy: By focusing on relevant features, attention mechanisms enhance the accuracy of predictions and classifications (Hu et al., 2018). This is particularly useful in tasks where the signal-to-noise ratio is low, such as in medical imaging or autonomous driving.

- Efficiency: Attention mechanisms can reduce computational load by allowing the network to ignore irrelevant parts of the data, leading to faster processing times (Xu et al., 2015). This efficiency is crucial for real-time applications where processing speed is critical.

- Robustness: Networks with attention mechanisms are better equipped to handle noisy or cluttered inputs, improving their robustness to real-world variations (Jetley et al., 2018). This robustness ensures that the models maintain high performance even in challenging environments.

- Interpretability: Attention mechanisms can provide insights into the decision-making process of the network, making it easier to understand and interpret the model's behavior (Xu et al., 2015). This transparency is valuable for applications in healthcare and finance, where understanding the model's decisions is essential for trust and accountability.

Convolutional Neural Networks (CNNs)

CNNs are a class of deep learning models that have proven exceptionally effective in processing and analyzing visual data (Krizhevsky, Sutskever, & Hinton, 2012; LeCun, Bengio, & Hinton, 2015). Unlike traditional neural networks, CNNs are specifically designed to recognize patterns and structures within images, making them the backbone of modern computer vision tasks such as image classification, object

detection, and segmentation.

CNN's Architecture Overview

A typical CNN architecture consists of several key layers as demonstrated in Figure 1. The input layer receives the raw pixel values of an image. For example, a colored image with a resolution of 224x224 pixels would have an input layer consisting of 224x224x3 (height x width x color channels) values.

The convolutional layer is the core building block of a CNN. It performs the convolution operation, which involves sliding a small matrix, known as a filter or kernel, over the input image to produce a feature map. This process helps in detecting local patterns such as edges, textures, and shapes. Each filter is trained to recognize different features, and multiple filters can be used in a single convolutional layer. Filters are small, trainable matrices (e.g., 3x3 or 5x5) that are applied across the image to detect specific features. The stride refers to the step size by which the filter moves across the image, and padding is sometimes added around the image border to control the spatial size of the output feature map.

After the convolution operation, an activation function is applied to introduce non-linearity into the model. The most commonly used activation function in CNNs is the Rectified Linear Unit (ReLU) (Nair & Hinton, 2010), which sets all negative values to zero and leaves positive values unchanged. This is important in deep neural networks as it helps address the problem of vanishing gradients when backpropagating across many layers, ensuring effective training and faster convergence (Nair & Hinton, 2010).

The pooling layer performs down-sampling (or subsampling) of the feature maps to reduce their spatial dimensions and the number of parameters, which helps in controlling overfitting and improving computational efficiency (LeCun et al., 1989; Scherer, Müller, & Behnke, 2010). The most common type is max pooling (Scherer, Müller, & Behnke, 2010), which selects the maximum value from a patch of the feature map. Max pooling takes the maximum value from each patch of the feature map, typically using a 2x2 filter with a stride of 2, while average pooling takes the average value from each patch of the feature

map. A stride refers to the number of pixels by which the filter moves over the feature map, with a stride of 2 meaning that the filter shifts two pixels at a time.

After several convolutional and pooling layers, the final processing in the neural network is done via fully connected layers. These layers are similar to those in traditional neural networks, where each neuron is connected to every neuron in the previous layer. The output from the convolutional layers is flattened into a one-dimensional vector before being fed into fully connected layers. The output layer produces the final prediction. For classification tasks, this layer typically uses the SoftMax activation function, which converts the raw output scores into a probability distribution over the classes by exponentiating each score and then normalizing by the sum of all exponentiated scores (Bridle, 1990).

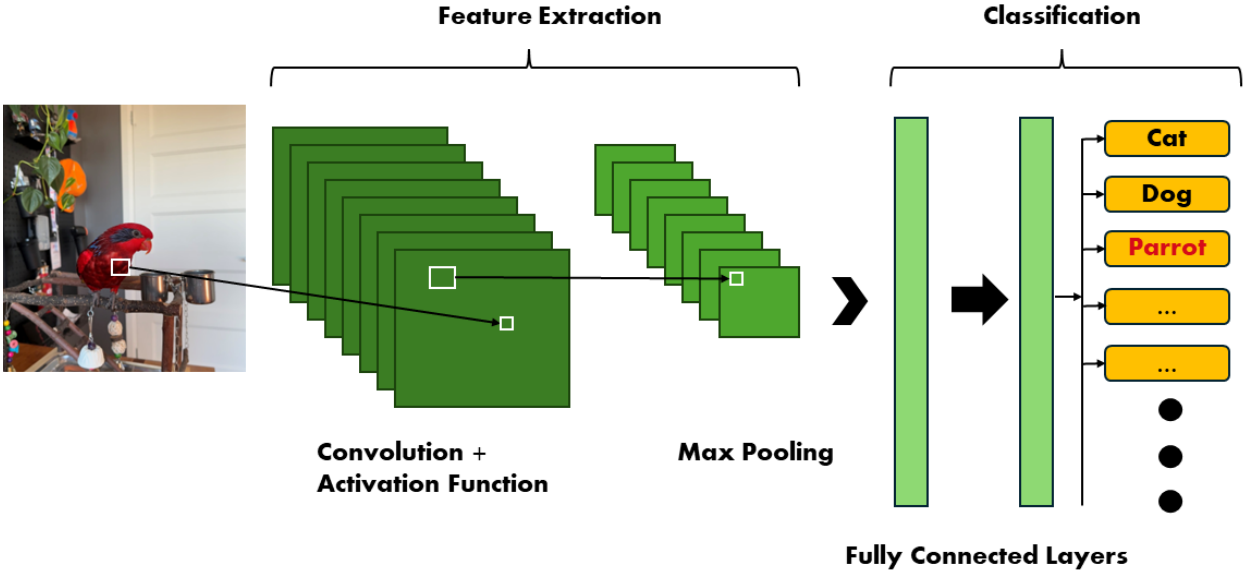

Figure 1.1 An example of CNN's architecture.

CNN's Advantages

CNNs offer several advantages that make them particularly well-suited for image processing tasks. One of the primary benefits is parameter sharing (LeCun et al., 1989). In convolutional layers, the same filter (or set of filters) is used across different parts of the input image, which significantly reduces the number of parameters compared to fully connected networks where each neuron connects to every

neuron in the previous layer (LeCun, Bottou, Bengio, & Haffner, 1998). This reduction in parameters not only makes CNNs more computationally efficient but also helps mitigate the risk of overfitting, which occurs when the network captures too many detailed features of the training data at the detriment of generalization, especially when dealing with limited training data.

Another key advantage is local connectivity (LeCun et al., 1998). Unlike fully connected networks where each neuron is connected to every input feature, neurons in a convolutional layer are only connected to a small, localized region of the input image. This local connectivity enables the network to learn spatial hierarchies of features, starting from low-level features such as edges and textures in the initial layers to high-level features like object parts and full objects in deeper layers. By focusing on local patterns, CNNs can effectively capture the spatial dependencies and structures inherent in visual data (LeCun et al., 1998).

Additionally, CNNs benefit from their hierarchical structure (LeCun, Bengio, & Hinton, 2015). Early layers capture low-level visual features, while deeper layers build on these to recognize more complex patterns and objects. This hierarchical feature learning is particularly effective for image recognition tasks where understanding both fine details and broader context is important (Krizhevsky et al., 2012).

CNNs also excel in feature extraction (LeCun et al., 1998). By automatically learning relevant features from raw pixel data, CNNs eliminate the need for manual feature engineering, which refers to the pre-selection of features by a programmer before training the network—a process that is both time-consuming and prone to human bias. The learned features are often more informative and generalizable, leading to better performance on a wide range of visual tasks (Krizhevsky et al., 2012).

Another significant advantage of CNNs is their ability to handle high-dimensional data efficiently (Krizhevsky et al., 2012). Images are naturally high-dimensional, and fully connected networks struggle to process such data without an explosion in the number of parameters. CNNs, with their convolutional and pooling operations, manage high-dimensional data more effectively by preserving spatial hierarchies and

reducing dimensionality progressively.

Finally, CNNs are highly scalable and can be adapted for various complex tasks beyond simple image classification (LeCun, Bengio, & Hinton, 2015). For instance, architectures like Faster R-CNN and YOLO (You Only Look Once) extend the basic CNN framework to perform object detection, where the network not only classifies objects but also predicts their locations within the image (Ren et al., 2015; Redmon et al., 2016). Similarly, fully convolutional networks (FCNs) and U-Net are designed for semantic segmentation, where each pixel in the image is classified into a specific category (Long et al., 2015; Ronneberger et al., 2015).

Importance of Secondary Features

While CNNs excel at identifying dominant features that are prevalent in the training data, there is a risk of becoming overly fixated on these features (Storcheus et al., 2015; Tang et al., 2022; Ye et al., 2021). This phenomenon occurs when the network disproportionately focuses on the most common features at the expense of other potentially important but less frequent features. For instance, in a model trained to recognize dogs, the network might focus predominantly on the facial features if these are the most common aspects in the training images. In human visual perception, findings indicate that the primary visual cortex (V1) and secondary visual cortex (V2) process different types of visual information. Research shows that the brain integrates color, form, and motion through distinct yet interconnected pathways, ensuring a comprehensive understanding of visual stimuli (Sincich & Horton, 2005). Similarly, parallel processing pathways help the brain segregate form, color, movement, and depth, emphasizing the importance of balanced feature processing (Livingstone & Hubel, 1988). Additionally, research has shown that both high-level and low-level features play crucial roles in visual recognition. High-level features, also known as dominant features, are processed rapidly and can often lead to quick, albeit sometimes superficial, recognition (Grill-Spector & Weiner, 2014). Low-level features, also known as secondary features, although processed more slowly, contribute to a deeper and more nuanced

understanding of a visual stimulus (Reddy & Kanwisher, 2006). This layered approach in human perception allows for a robust and flexible understanding of visual scenes, ensuring that both prominent and subtle aspects are considered. This highlights the necessity for CNNs to balance the learning of both dominant and secondary features to improve their generalization and robustness, especially if we want to bridge computer vision with human perception capabilities.This fixation on dominant features can impair the model's ability to generalize to new, unseen data or to scenarios where the dominant features are absent or obscured (Geirhos et al., 2018). To counter this, it is crucial to ensure that CNNs also learn and utilize secondary features. Secondary features are subtler and less prominent but provide valuable additional information that can improve the model's robustness and adaptability (Zhou, 2000; Hassaballah & Awad, 2016; Selvaraj, Veloso, & Rosenthal, 2018). Secondary features might include subtle patterns, textures, and contextual details that are not immediately obvious but contribute to the overall understanding of the input data (Hassaballah & Awad, 2016). For example, in the case of the dog recognition model, secondary features might include the shape of the body, the texture of the fur, or the presence of specific background elements as demonstrated in Figure 2. By incorporating these secondary features into the learning process, the model can develop a more holistic understanding of the data, making it better equipped to handle variations, occlusions and novel instances (Chen et al., 2021).

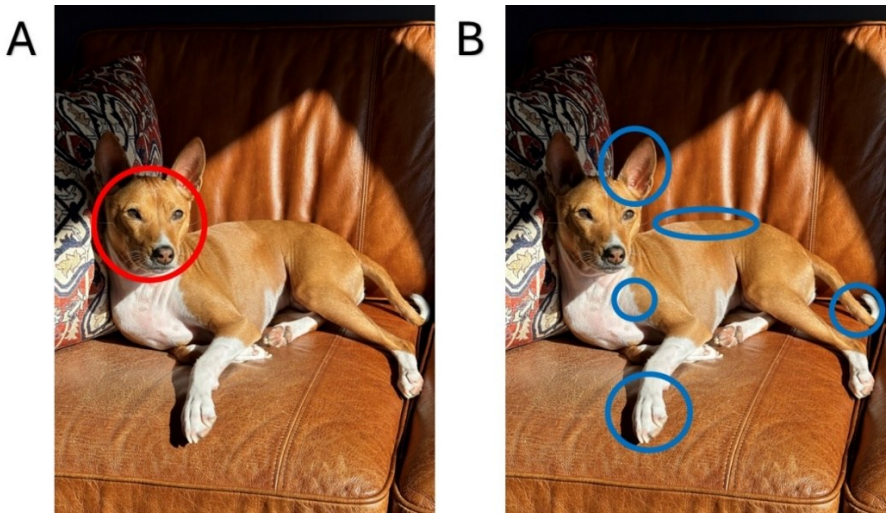

Figure 1.2 An example of dominant (A) and secondary features (B).

Incorporating secondary features into the learning process involves several strategies. One approach is data augmentation, which artificially increases the diversity of the training data by applying transformations such as rotation, scaling, and color adjustments (Shorten & Khoshgoftaar, 2019). This helps the model to learn a broader range of features and become less reliant on any single characteristic.

Another strategy is to use techniques like dropout and regularization during training (Srivastava et al., 2014). These methods prevent the model from becoming too dependent on any one feature by randomly omitting certain features during training or by penalizing large weights. This encourages the network to spread its attention across a wider array of features, including secondary ones.

Finally, multi-task learning can be employed, where the model is trained on multiple related tasks simultaneously (Caruana, 1997). This encourages the network to learn a more comprehensive set of features that are useful across different tasks, thereby improving its robustness and generalization capabilities.

Nevertheless, there exists a substantial gap in our understanding of how to enhance CNNs' ability to effectively learn and utilize secondary features. This thesis aims to bridge this gap by investigating innovative techniques specifically designed to augment secondary feature recognition using advanced data augmentation methods and the strategic incorporation of attention mechanisms. Secondary features, which are often subtle and less prominent compared to primary features, can play a crucial role in the overall accuracy and robustness of CNNs (Zhou, 2000; Hassaballah & Awad, 2016; Selvaraj, Veloso, & Rosenthal, 2018). These features might include fine textures, minor shape variations, and contextual details that are not immediately apparent but are essential for a comprehensive understanding of the input data (Hassaballah & Awad, 2016).

To address this, the thesis will explore how novel data augmentation techniques can be employed to emphasize secondary features during the training process. Traditional data augmentation methods, such as rotation, scaling, and color adjustments, have been widely used to improve model robustness by

artificially increasing the diversity of the training data. However, these methods often do not explicitly target the enhancement of secondary features. By developing new augmentation strategies that focus on subtle variations and less prominent characteristics, this research aims to train CNNs to recognize and prioritize these secondary features more effectively. These techniques will be tailored to mimic the complex and varied conditions that models are likely to encounter in real-world scenarios, thereby improving their generalization capabilities.

In addition to data augmentation, attention mechanisms will be refined and integrated into the CNN architecture to dynamically focus on both primary and secondary features. Attention mechanisms, which have shown great promise in tasks such as natural language processing and computer vision, allow neural networks to weigh the importance of different parts of the input data (Vaswani et al., 2017). By adjusting these weights based on the relevance of the features, attention mechanisms can help ensure that secondary features are not overlooked. This approach aims to create a more balanced and holistic representation of the input data without directly manipulating the training data, enabling the model to utilize a wider range of features and improve its performance across diverse applications.

Through this comprehensive approach, the thesis will address the current knowledge gap and contribute to the development of CNNs that are more robust, versatile, and capable of generalizing beyond their training data. The findings from this research will have broad implications, potentially improving the performance of CNNs in various fields, including autonomous driving, healthcare, and surveillance, where recognizing subtle yet critical details is essential for success.

Out-of-Distribution generalization

Out-of-Distribution (OOD) prediction is essential for computer vision models to manage data that significantly deviates from their training sets (Liang, Li, & Srikant, 2017; DeVries & Taylor, 2018; Hendrycks et al., 2021; Fort, Ren, & Lakshminarayanan, 2021). OOD refers to instances where the model encounters inputs that are substantially different from the training data, often leading to a decrease in model

performance (see Figure 3). This capability is vital for the real-world application of computer vision systems, where encountering diverse and unpredictable data is common. For example, an autonomous vehicle's vision system trained in one region might encounter entirely different road conditions and signage in another, necessitating strong OOD prediction abilities to maintain safety and reliability.

The nature of OOD data is varied and difficult to categorize, ranging from minor changes like lighting and angle differences to significant shifts such as new objects or scenarios (Liang, Li, & Srikant, 2017; DeVries & Taylor, 2018; Hendrycks et al., 2021; Fort, Ren, & Lakshminarayanan, 2021). In the healthcare field, an OOD instance could involve rare diseases or medical imaging artifacts not present in the training data (Gao et al., 2018). This variability requires computer vision models to not only detect these anomalies but also respond appropriately, whether through accurate classification or flagging for further human analysis.

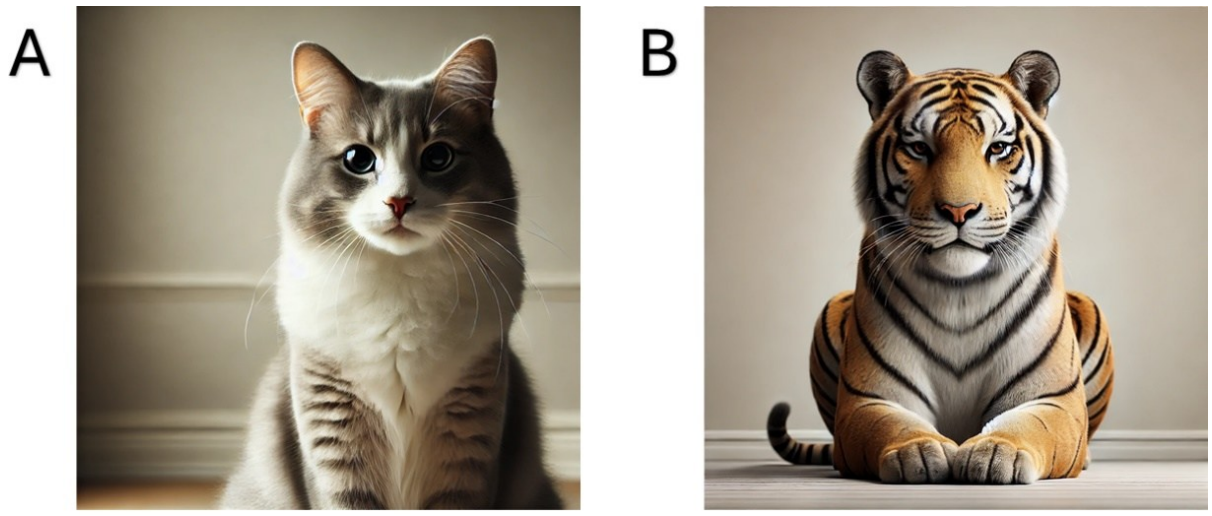

Figure 1.3 An example of a training sample (A) and OOD sample (B) with new attributes not previously encountered

Importance of OOD Prediction

The ability to accurately identify and handle OOD data is crucial for several reasons. In safety-critical applications like autonomous driving, failure to recognize and appropriately react to OOD data can lead to catastrophic outcomes (Amodei et al., 2016). For instance, an autonomous vehicle encountering

unexpected weather conditions or unfamiliar road signs must be able to recognize that these inputs differ from its training data and adjust its behavior accordingly. Similarly, in security applications, surveillance systems must identify unusual activities or objects that were not part of their training datasets to alert security personnel to potential threats (Hodge & Austin, 2004).

In the context of healthcare, OOD prediction is equally critical. Medical imaging systems trained on common diseases might encounter rare conditions in new patients. These systems must flag such cases for specialist review to ensure accurate diagnosis and treatment. For example, a system trained to detect lung abnormalities in chest X-rays must recognize when it encounters an atypical pattern that it has not seen before and alert the radiologist for a more detailed examination.

Research has shown that humans possess a remarkable ability to recognize and adapt to OOD samples, often outperforming deep neural networks in various scenarios involving visual distortions and novel stimuli. Geirhos et al. (2018) demonstrated that humans exhibit superior generalization capabilities compared to deep neural networks when faced with changes in texture, style, or context of visual stimuli, highlighting the robustness of human perception and its adaptability to unexpected variations in visual input. Similarly, Dodge and Karam (2017) conducted a comprehensive comparison of human and deep learning recognition performance under various visual distortions, revealing that while deep neural networks can achieve high accuracy under controlled conditions, they struggle significantly when presented with distorted or altered images. In contrast, human recognition performance remains relatively stable across a wide range of visual perturbations, showcasing the inherent flexibility and robustness of human visual perception. Wichmann et al. (2017) further explored the differences between human and machine vision by comparing their performance in a variety of recognition tasks, emphasizing that humans can leverage contextual information and prior knowledge to maintain high levels of performance even when confronted with unfamiliar or distorted visual inputs. Overall, these studies underscore the superiority of human visual perception in OOD tasks, pointing to the need for developing

more robust and adaptable machine learning models that can emulate human-like generalization abilities.

Challenges of OOD Prediction

OOD prediction presents several challenges. One major challenge is the inherent diversity and unpredictability of OOD data (Hendrycks & Gimpel, 2017). Since OOD data can range from subtle variations in familiar patterns to completely new and unseen objects, designing models to handle this variability is complex (Arjovsky, Bottou, Gulrajani, & Lopez-Paz, 2020). Traditional supervised learning methods, which rely heavily on the assumption that training and test data are drawn from the same distribution, often struggle with OOD data (Szegedy et al., 2014).

Another challenge is the lack of labeled OOD data for training (Bulusu et al., 2020). In many cases, it is impractical to anticipate all possible OOD scenarios and collect corresponding labeled data. This limitation necessitates the development of unsupervised or semi-supervised methods capable of identifying OOD instances without extensive labeled datasets (Hendrycks et al., 2019).

Techniques for OOD Detection

Several techniques have been proposed to improve OOD detection in computer vision models. These include:

- Confidence Scoring: One approach involves using the model's prediction confidence to identify OOD data. Typically, models are less confident when presented with OOD data. Techniques such as the Maximum Softmax Probability (MSP) can be used, where lower confidence scores indicate potential OOD instances (Hendrycks & Gimpel, 2017).

- Outlier Exposure: This method involves training the model on known outliers or artificially generated OOD examples. By exposing the model to a diverse set of non-training data, it can learn to distinguish between in-distribution and OOD samples (Hendrycks, Mazeika, & Dietterich, 2019).

- Auxiliary Models: Separate models or networks can be trained to specifically detect OOD

instances (Ren et al., 2019). These models can operate alongside the primary vision model to flag anomalies. For example, autoencoders can be used to reconstruct input data and detect anomalies based on reconstruction errors (Zhai et al., 2016).

- Ensemble Methods: Combining predictions from multiple models can improve OOD detection (Lakshminarayanan, Pritzel, & Blundell, 2017). Ensemble methods leverage the diversity of different models to provide more robust predictions and identify OOD data by analyzing the variance among the models' outputs.

- Bayesian Neural Networks: These networks incorporate uncertainty estimation into their predictions, making them more adept at identifying OOD data (Gal & Ghahramani, 2016). By modeling uncertainty, Bayesian neural networks can provide more informative predictions regarding the likelihood of an input being OOD.

Secondary Features and OOD learning

There exists a large gap in the literature regarding the role of secondary features in enhancing the performance of models in OOD learning. While considerable effort has been devoted to developing highly complex deep learning models to improve robustness (Schmidhuber, 2015; LeCun, Bengio, & Hinton, 2015), the potential of leveraging secondary features for OOD detection remains underexplored. This thesis aims to fill this gap by examining how the recognition and utilization of secondary features can enhance OOD detection capabilities in both advanced CNN models and simpler neural network architectures.

The research will focus on innovative strategies to emphasize secondary features during the training process, exploring their influence on the model's ability to identify and adapt to OOD instances. By implementing advanced data augmentation methods and refining attention mechanisms, this study will evaluate how enhancing secondary feature recognition can improve the model's detection of

anomalies and overall generalization performance. The objective is to create models that not only excel in identifying dominant features but also effectively utilize secondary features to make more accurate and reliable predictions across a wide range of real-world conditions. This investigation aims to provide new insights into the integration of secondary features in OOD learning, potentially leading to the development of more robust and adaptable computer vision systems.

Domain Generalization

Domain Generalization is a critical area in computer vision aimed at training models that can generalize to new, unseen domains without requiring additional fine-tuning (Blanchard et al., 2011). Unlike traditional approaches that assume the training and test data come from the same distribution, domain generalization seeks to create robust models capable of performing well across various domains. This involves developing techniques that enable models to capture features that are invariant across different domains, thus enhancing their adaptability and utility in real-world applications (Muandet et al., 2013). By learning to generalize from multiple source domains, these models can better handle the variability and unpredictability encountered in practical scenarios. For instance, Figure 4 illustrates the concept of domain generalization through the classification of mammals and reptiles. It shows how a model trained on specific instances (e.g., dog, cat, snake, chameleon) must accurately classify unseen instances (e.g., gorilla, bear, iguana, gecko), demonstrating the model's ability to generalize knowledge across different but related categories.

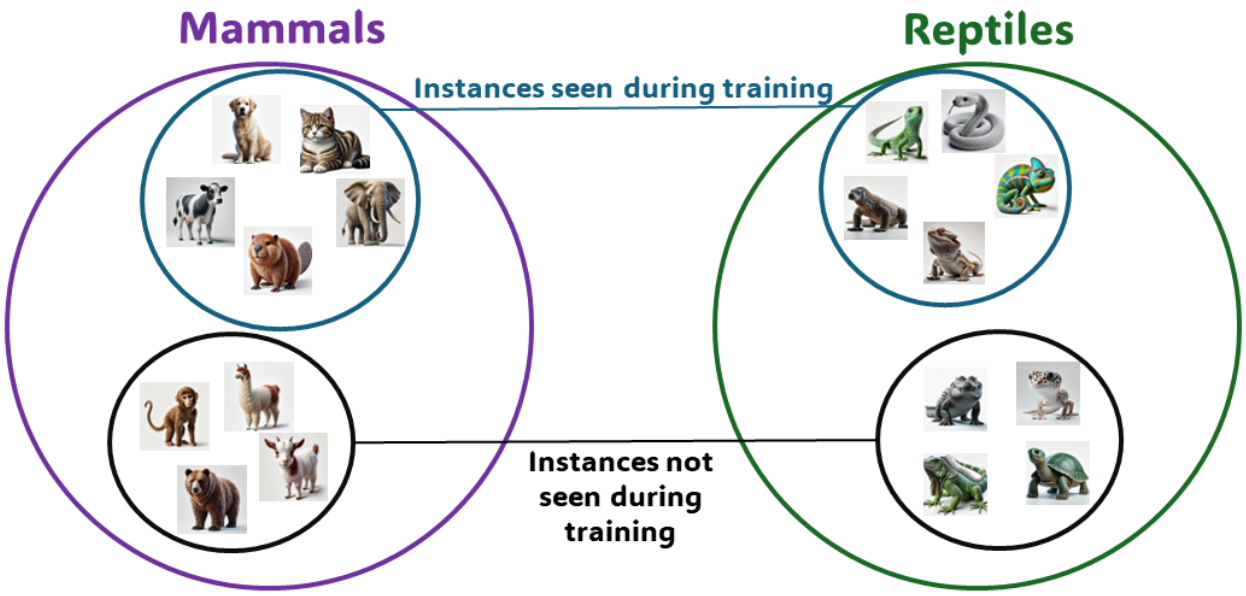

Figure 1.4 Models trained on specific instances of mammals and reptiles must accurately classify new, unseen examples within the same domains, demonstrating domain generalization.

Importance of Domain Generalization

In real-world scenarios, computer vision models often encounter data distributions that differ significantly from their training sets (Torralba & Efros, 2011). For instance, a model trained on images from urban environments in one city might need to be deployed in another city with different architectural styles, lighting conditions, and weather patterns. Without robust domain generalization, the model's performance can degrade significantly when faced with these new conditions (Gulrajani & Lopez-Paz, 2020).

The importance of domain generalization extends to various applications:

- Healthcare: Medical imaging models trained on data from one hospital may need to be deployed in other hospitals with different equipment, patient demographics, and imaging protocols. Robust domain generalization ensures consistent diagnostic performance across different healthcare settings (Chen et al., 2021).

- Agriculture: Models used for crop monitoring must generalize across different regions, soil types,

and weather conditions to provide accurate insights and recommendations to farmers (Kamilaris & Prenafeta-Boldú, 2018).

- Retail and E-commerce: Visual recognition systems in retail must recognize products across different stores with varying layouts, lighting, and presentation styles.

In human studies, domain generalization is a well-documented phenomenon where humans apply principles learned in one domain to new, superficially different domains (Doumas, Puebla, Martin & Hummel, 2022). This ability to generalize across domains is a fundamental aspect of human cognition, enabling robust and flexible understanding. For instance, research by Doumas et al. (2022) introduced a computational model integrating Learning and Inference with Schemas and Analogy (LISA) and Discovery of Relations by Analogy (DORA) frameworks to explain how humans learn and represent relations. The model shows that humans can learn structured representations from simple visual inputs and perform zero-shot learning through analogical inference, reflecting human cognitive development. Zero-shot learning refers to the ability to recognize and categorize objects or concepts without having seen them before, based on understanding relationships and features learned from other contexts. Studies by Geirhos et al. (2018), Dodge and Karam (2017), and Wichmann et al. (2017) further highlight how humans outperform AI systems in recognizing and generalizing to OOD data. These findings underscore the significance of relational representations and structured learning in achieving human-like generalization, suggesting that enhancing AI systems with similar capabilities could bridge the gap between human and machine intelligence.

Techniques for Achieving Domain Generalization

Several techniques have been developed to improve domain generalization in computer vision models. These approaches typically aim to enhance the model's robustness to domain shifts by exposing it to a diverse set of training conditions or by encouraging it to learn domain-invariant features.

Data augmentation. The simplest yet most effective techniques for enhancing domain generalization in computer vision models is data augmentation (Shorten & Khoshgoftaar, 2019). At its core, data augmentation involves artificially increasing the diversity of the training dataset by applying a series of transformations to the original images. These transformations can include a wide array of operations, such as rotation, scaling, cropping, flipping, color jittering, and adding noise, among others (Perez & Wang, 2017). Each of these transformations serves to simulate different conditions and variations that a model might encounter in real-world scenarios.

Building on this foundation, previous studies explored adding noise to training data as a means to improve model robustness, particularly in noisy environments where real-world data often deviates from ideal conditions (Grandvalet, Canu, & Boucheron, 1997; Zur, Jiang, Pesce, & Drukker, 2009). These efforts aimed to mitigate the model's sensitivity to noise and enhance its performance on challenging datasets. Expanding on this idea, this thesis systematically examines the optimal balance between noisy and noise-free images in training data, alongside retraining strategies that enable better generalization across various noise types. By adjusting the proportion of noise injected into the dataset, the study identifies levels that maximize generalization without sacrificing accuracy on clean data. This detailed analysis not only sheds light on the interplay between noise injection and model learning but also provides practical guidelines for effectively incorporating noise in diverse applications.

One of the primary benefits of data augmentation is that it helps to mitigate the problem of overfitting, which occurs when a model learns to perform exceptionally well on training data but fails to generalize to unseen data. By exposing the model to a broader range of variations during the training phase, data augmentation effectively increases the dataset's richness without the need for additional labeled data. This process is particularly valuable in situations where collecting and labeling large amounts of data is impractical or cost prohibitive (Shorten & Khoshgoftaar, 2019).

By simulating a variety of conditions during training, data augmentation helps the model to learn

a more generalized set of features, making it more resilient to variations in the input data. This means that when the model encounters data that deviates from the training distribution, it is better equipped to handle these discrepancies and still make accurate predictions (Shorten & Khoshgoftaar, 2019).

Data augmentation is not just about making slight modifications to the data; it's about creating a comprehensive training environment that anticipates and prepares for the diversity of the real world (Shorten & Khoshgoftaar, 2019). This approach is particularly valuable in domains like autonomous driving, where vehicles must operate under a wide range of conditions, from different weather patterns to varying traffic scenarios. For instance, augmenting the training data with images simulating rain, fog, or night-time conditions can significantly enhance the robustness of the model (Shorten & Khoshgoftaar, 2019).

Moreover, data augmentation can be combined with other techniques to further improve model performance. For example, mixing augmentation with transfer learning, where a model pre-trained on a large dataset is fine-tuned on a specific task, can leverage the benefits of both approaches, leading to even better generalization (Perez & Wang, 2017).

In addition to traditional methods, advanced data augmentation techniques have been developed to further enhance domain generalization. These include:

- Generative Adversarial Networks (GANs): GANs can generate new, synthetic images that are highly realistic and can be used to augment training datasets (Antoniou, Storkey, & Edwards, 2018).

- Adversarial Training: This involves creating adversarial examples that are intentionally designed to be challenging for the model, helping it to learn more robust features (Bai, et al., 2021).

- AutoAugment: An automated data augmentation technique that uses reinforcement learning to find the optimal set of augmentation policies for a given dataset (Cubuk et al., 2019).

Meta-Learning. "Learning to learn" involves training models on multiple tasks or domains to improve their adaptability to new tasks or domains (Finn, Abbeel, & Levine, 2017). In the context of domain generalization, meta-learning techniques train models on a variety of domains and evaluate their performance on held-out domains, optimizing the model's ability to generalize to new, unseen domains (Li et al., 2018).

Invariant Risk Minimization (IRM). It aims to learn invariant predictors that remain effective across different environments or domains (Arjovsky et al., 2020). The model is trained to minimize the risk (loss) while ensuring that the learned representations are invariant to changes in the domain (Arjovsky et al., 2020).

Self-Supervised Learning. It leverages unlabeled data to learn useful representations that can transfer across domains (Jing & Tian, 2020). By training models on tasks where the supervision signal is derived from the data itself (e.g., predicting the rotation angle of an image), self-supervised learning can produce robust features that generalize well to new domains (He et al., 2020).

Even with the advancements provided by techniques like Data Augmentation, Domain-Adversarial Training, Meta-Learning, IRM, and Self-Supervised Learning, domain generalization in computer vision still faces significant challenges (Hendrycks & Dietterich, 2019). These methods, while effective in many scenarios, often fall short when models encounter extreme domain shifts or highly varied real-world environments (Hendrycks & Dietterich, 2019). The limitations are particularly evident when the variations in test data are substantially different from the training data, which can result in a dramatic decrease in model performance (Gulrajani & Lopez-Paz, 2020).

One key issue is that many current techniques assume that the diversity in training data is representative of all possible variations in the real world, which is rarely the case (Gulrajani & Lopez-Paz, 2020). For example, while data augmentation can simulate different conditions, it may not capture all the nuances of domain shifts encountered in practical applications (Gulrajani & Lopez-Paz, 2020). Similarly,

domain-adversarial training and meta-learning can improve generalization but often require extensive computational resources and sophisticated model architectures, making them difficult to implement and scale (Gulrajani & Lopez-Paz, 2020). This thesis aims to address these challenges by investigating how the recognition and incorporation of secondary features can enhance domain generalization. By focusing on the subtle yet significant aspects of data that are often overlooked, this research will develop innovative methods to improve model robustness and adaptability. Through advanced data augmentation techniques and the integration of attention mechanisms, the study will enhance the ability of CNN models to generalize across diverse and unpredictable real-world environments, ultimately bridging the gap between current model performance and the demands of practical applications.

Benchmarks and Challenges in Domain Generalization

In the realm of domain generalization, several benchmark datasets, including PACS, Office-Home, and DomainNet, have been widely utilized to assess the robustness and generalization ability of CNNs (Li et al., 2017; Venkateswara et al., 2017; Peng et al., 2019). These datasets encompass images from multiple domains, mimicking the real-world variations that models might encounter. However, they fall short in fully addressing the specific challenges faced in the everyday deployment of CNNs (Gulrajani & Lopez-Paz, 2020).

Many domain generalization datasets lack the breadth and diversity necessary to represent real-world domains comprehensively. For instance, PACS mainly features images of objects in different artistic styles, such as photos, sketches, and paintings, which do not adequately capture the complexity of real-world variations. Similarly, Office-Home includes images of objects from various office and home environments, but the variations are limited to changes in backgrounds and perspectives, failing to reflect the full scope of real-world diversity. DomainNet, with its mix of clipart, infographics, and real-world photos, still does not provide a thorough range of variations encountered in practical applications (Niu et al., 2022).

To effectively demonstrate domain generalization, it is crucial to show a model's ability to generalize to new instances within a class that were not seen during training (Arpit et al. 2017). For example, showing different cars from various angles and backgrounds does not adequately capture the challenge of generalizing to new instances of a class. What is required are datasets that include entirely new instances within the same class, exhibiting attributes and features absent in the training set. This approach will test the network's ability to recognize fundamental characteristics of a class and adapt to unseen variations, providing a more rigorous evaluation of generalization capabilities in real-world scenarios.

This thesis will address this challenge by implementing a new benchmark to evaluate domain generalization and OOD predictions. This benchmark is designed to provide a more rigorous assessment of a model's generalization capabilities by incorporating tasks that test resilience to data transformations, intra-class variability, and adversarial conditions. By leveraging this benchmark, this research aims to offer a comprehensive evaluation framework that ensures CNNs are not only accurate on familiar data but also robust and adaptable to real-world variability.

Contribution of Thesis

The contributions of this thesis can be summarized as follows:

1) Introducing a novel training method termed "noisy/noise-free" (NNF) training, which enhances the robustness and generalization of CNNs by incorporating a specific proportion of noisy and noise-free images in the training set. This method significantly improves the classification accuracy of CNNs on noisy datasets by training them with a small but optimal amount of noisy data, as demonstrated using the MNIST dataset.

2) Introducing Dominant Feature Masking (DFM) as a novel data augmentation approach that strategically conceals the most prominent features within images, allowing neural networks to capture both dominant and non-dominant attributes. This method enhances the model's ability to generalize to OOD data, significantly improving OOD prediction performance without compromising in-distribution accuracy.

3) Introducing the Versatile Evaluation Benchmark (VEB) to rigorously test domain generalization and OOD performance of CNNs. VEB comprises three distinct datasets: augmented MNIST for resilience against common transformations, a novel dataset with unseen image classes for intra-class variability, and a DALL-E generated dataset to assess class differentiation under adversarial conditions. This benchmark enables comprehensive evaluation of models, ensuring robustness and adaptability in real-world scenarios

4) Developing the Dynamic Attention Layer (DAL). This novel layer dynamically adjusts attention weights based on selected percentiles during training. DAL enhances CNNs by allowing them to focus on both dominant and secondary features of the data, thereby improving their

generalization capabilities in OOD scenarios. The effectiveness of DAL was demonstrated across three different datasets, showing significant improvements in accuracy and robustness compared to traditional attention mechanisms and data augmentation techniques.

Overview of Thesis

The remainder of this thesis is structured as follows. First, we will investigate the novel training method involving partial noise injection in the training data. More specifically, we will discuss the architecture of the custom-built CNN, the implementation of the NNF training technique, and the experimental results demonstrating the efficiency of this method in different tasks. We will also compare the performance of our model trained on noisy datasets to those trained on noise-free datasets to assess CNN's generalization.

Next, we will examine the DFM technique aimed at improving model generalization. We will describe the methodology behind DFM, the new VEB used for evaluation, and the results that show enhanced accuracy and robustness, particularly in handling OOD scenarios. We will also discuss the comparative performance of DFM against other augmentation techniques and its impact on model generalization.

Finally, we will explore the use of DAL in improving generalization in CNNs. We will detail the methodology behind DAL, the datasets used for evaluation, and the results that show enhanced accuracy and reduced loss in both in-distribution and OOD scenarios. We will also discuss the optimal configuration of parameters for DAL and its impact on model performance.

Following the detailed discussion of the methodologies and results from these studies, we will address the broader implications for understanding and improving the robustness and generalization of computer vision models. To end, we will offer avenues for future research directions.

Chapter 2

Generating Robust Convolutional Networks by Injecting Partial Noise in the Training Data

Artem Pilzak[1], Jean-Philippe Thivierge[1,2]

[1]School of Psychology, University of Ottawa, K1N 6N5, Ottawa, Canada
[2]Brain and Mind Research Institute, University of Ottawa, K1N 6N5, Ottawa, Canada

Abstract

Convolutional Neural Networks (CNN) have emerged as a highly efficient deep learning algorithm for processing, classifying, and analyzing images. Currently, there is a high demand for networks that are robust to noise in the training and testing data. However, the presence of noise in images has been an ongoing problem when it comes to deploying CNN for real-world application. In this chapter, we propose a solution to this problem by incorporating noisy and noise-free images in the training set of a CNN network. Using the MNIST dataset, we demonstrate that marked improvements to classification accuracy only require a small proportion of noisy images in the training set. This result generalized across test sets with different amounts of noise and types of noise. In sum, our novel training method increases the generalization of a CNN model when processing noisy data.

Keywords—convolutional neural networks, computer vision, noisy images, image classification, MNIST.

Introduction

Convolutional Neural Networks (CNN) are artificial neural networks (ANN) that specialize in discriminating, classifying, and predicting images (Albawi, Mohammed, & Al-Zawi, 2017). They are known to be the state-of-the art when it comes to working with high-resolution and complex image datasets. From facial recognition to cancer detection (Charan, Khan, & Khurshid, 2018), these neural networks are capable of high generalization accuracy by not overfitting the training data. However, distortions to the input image inducted by noise cause testing performance to drop due to the lack of robustness to noisy images (Dodge & Karam, 2016).

Several strategies have been proposed to increase the robustness to noise of CNN. One of these strategies is training the artificial neural network with noisy input which allows the model to make better predictions on noisy data (Grandvalet, Canu, & Boucheron, 1997; Zur, Jiang, Pesce, & Drukker, 2009). However, this method has not characterized the minimal required amount of noisy inputs for an efficient increase of accuracy while minimizing the computational cost. We hypothesized that only a small proportion of noisy inputs is necessary for an efficient noise learning. In other words, a surplus of noisy inputs in the training set will not lead to a further increase in accuracy when compared to a smaller set. To the best of our knowledge, no study has been conducted to investigate the relation between the proportion of noisy examples and testing accuracy of a CNN.

In this chapter, we propose a new training method to increase the accuracy of CNNs on noisy datasets. This method, termed “noisy/noise-free” (NNF) training, consists of configuring a training dataset with a specific proportion of noise-free images. We reason that presenting noise-free images followed by noisy images to a CNN will allow it to learn the associated features better by providing it with noise-free exemplars and using its acquired knowledge as a point of comparison when learning noisy image features.

This chapter is structured as follows: the methods section describes our custom-built CNN followed by an explanation and application of NNF training. The results section demonstrates the

efficiency of NNF training. Additionally, we show that retraining with only a limited number of noisy images increases the testing accuracy of our CNN. Finally, we tested our model on images with correlated noise to show the robustness of the CNN to different data distributions.

Methods

Custom Convolutional Neural Network

In order to test the proposed NNF training method, we designed a model built via the Tensorflow Keras framework and consisting of twelve layers (Figure 2.1). All convolutional layers had a RELU activation function with He Uniform initializer. The first two layers were 2D convolutional layers with a shape of 28x28x32 and were responsible for the initial feature extraction of images. The third layer was a max-pool layer with a shape of 14x14x32. This layer was followed by two convolutional layers with a shape of 14x14x128. The sixth layer was a max-pool layer with a shape of 7x7x64 connected to the seventh convolutional layer with a shape of 7x7x128. This was followed by a convolutional layer with the same shape as the previous layer. The ninth layer was a max-pool layer with a shape of 3x3x128. To flatten the three-dimension input into a simple vector, a flat layer was introduced as the tenth layer with a shape of 1152. This flat layer was connected to a dense layer with 128 units and RELU activation function. Finally, the last layer was a dense layer with 10 outputs and the SoftMax activation function.

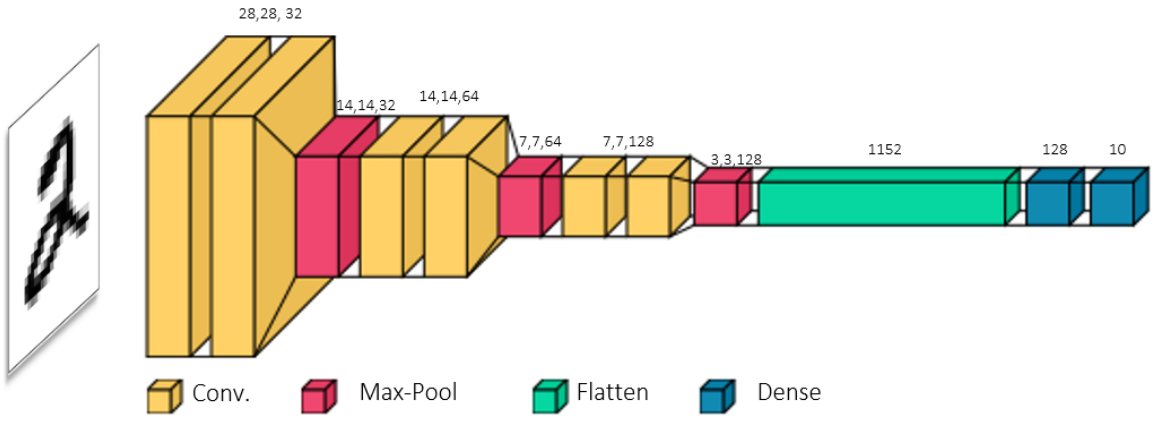

Figure 2.1 Architecture of the custom CNN with shapes and layers characteristics.

Data

We used the MNIST (Modified National Institute of Standards and Technology database) database (Figure 2.2) to train and test our novel training technique. The MNIST database is a large multi-purpose database of handwritten digits, from zero to nine, that is commonly used for training various image processing systems (Deng, 2012). The database contains 60,000 training images and 10,000 testing images in greyscale with dimensions 28x28x1.

Noise/Noise-Free Training

NNF training consists of training the model simultaneously with noisy and noise-free data. The optimal ratio of noisy to noise-free images in the training dataset is described below. During the training phase, the validation dataset consisted of noise-free and noisy data to increase the probability of achieving a global minimum in the error landscape and increasing the model's generalization. The testing phase consisted solely of noisy images.

Results

This section examines three different experiments involving noisy images as part of the training dataset. In the first experiment, we wanted to find the optimal proportion of noisy images in the training dataset while reaching the maximum testing accuracy. In the second experiment, we trained our model exclusive with noise-free images followed by retraining with a subset of noisy images to increase testing accuracy on the noisy dataset. Lastly, we trained our model on uncorrelated noise images and tested on correlated noise to examine if the CNN becomes robust to various types of noise.

Noise Training

The model was trained on Gaussian noise and tested on five different intensities of noise ranging from 0 to 1 (Figure 2.2). Here, noise intensity is defined as i.i.d. Gaussian noise,

$$\boldsymbol{N} = \boldsymbol{P} + \beta \boldsymbol{G}, \tag{1}$$

where $\boldsymbol{P} \in \mathbb{R}^{m \times n}$ represents a noise-free image with 28x28 dimensions, $\boldsymbol{G} \in \mathbb{R}^{m \times n}$ represents a matrix with entries drawn from a normal Gaussian distribution with same dimensions as matrix $\boldsymbol{P}$, β is a noise intensity constant β ranging from 0 to 1, and $\boldsymbol{N} \in \mathbb{R}^{m \times n}$ is the resulting noisy image.

Figure 2.2 shows five different noise intensities used to test the accuracy of our model. The model was constantly trained on β =0.75 noise intensity (Figure 2.2 (d)).

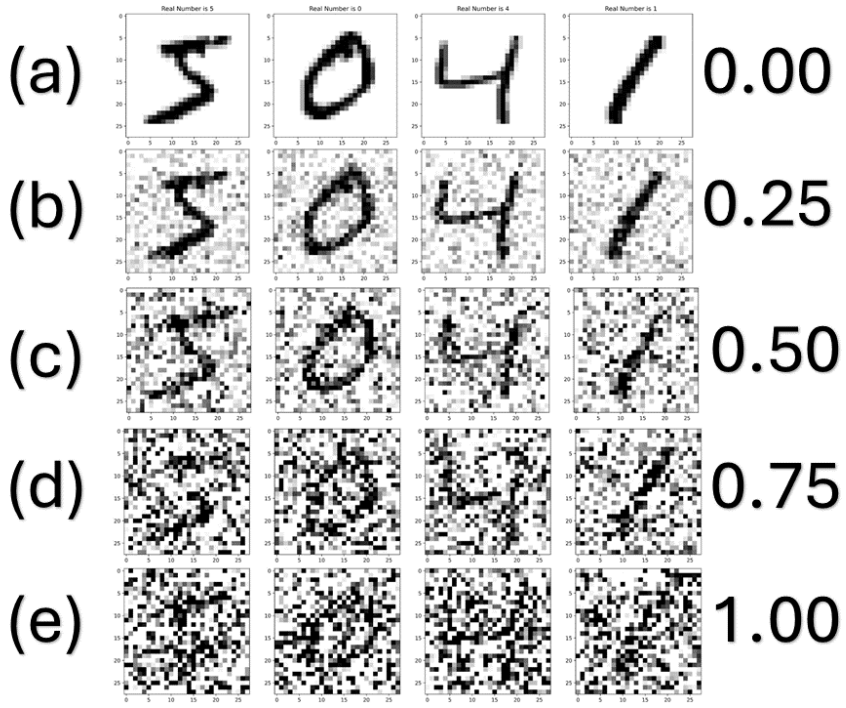

Figure 2.2 MNIST dataset with different intensities of Gaussian noise shown on the right side of the images.

NNF training with noisy and noise-free examples was split up in 15 different ratios where each proportion represented the number of noisy images in the dataset. For example, a 1% ratio corresponds to 600 noisy images and 59,400 noise-free images in the training dataset. The testing set consisted entirely of noisy images.

The model was trained and tested 10 times for each ratio and Gaussian noise with a total of 20 epochs for each trial. Thus, the network was trained and tested 1,500 times across different conditions. The results are shown in Figure 2.3. This figure shows the testing accuracy on four noise intensity factors across different ratios of noise versus noise-free images in the training set. The graph shows that the model does not require a large quantity of noisy images in the training set to reach a plateau in testing accuracy. Additionally, testing images with a greater noise intensity factor requires more noisy training exemplars to obtain the maximum accuracy. However, only 30% of noisy images in the training set are necessary to reach the testing accuracy limit for a noise intensity of 1.

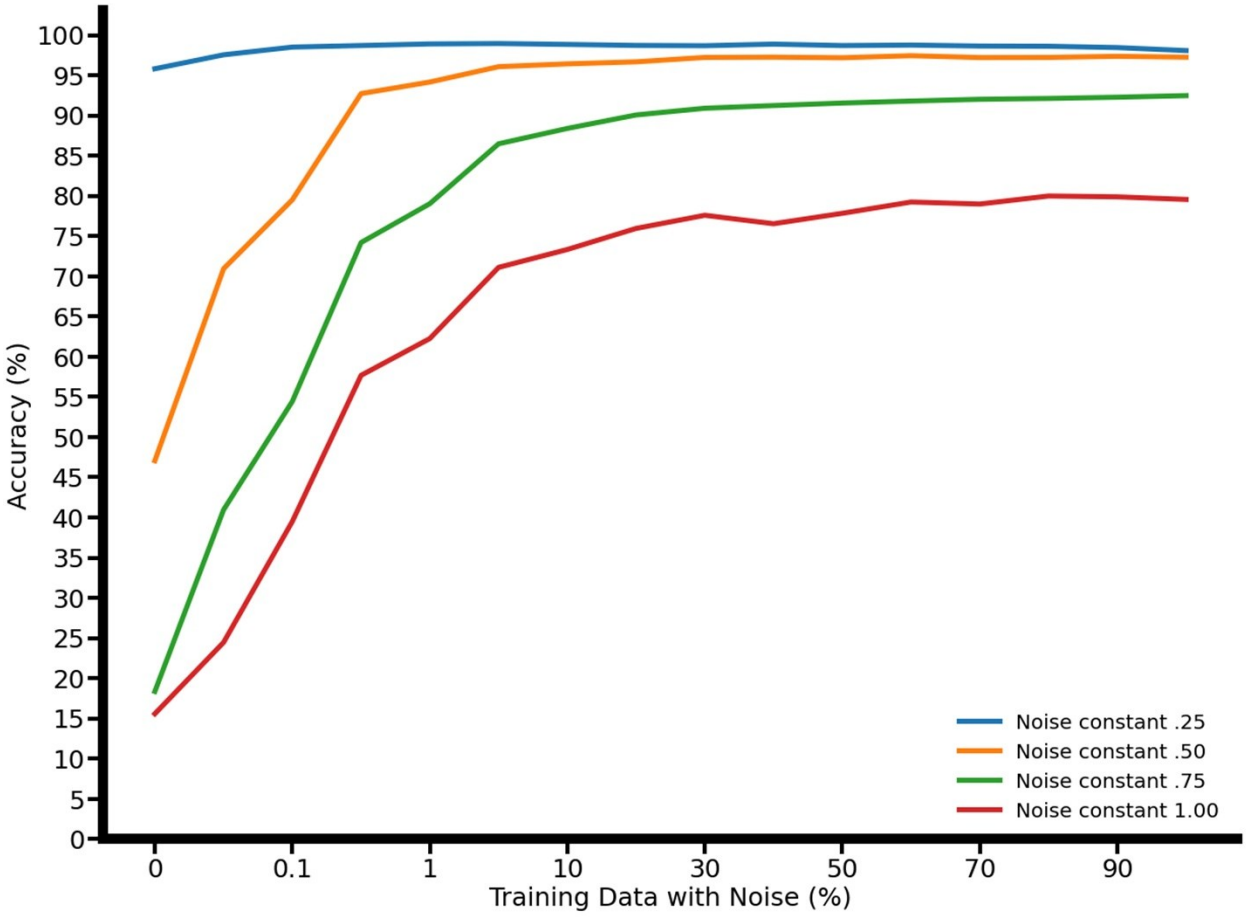

Figure 2.3 Testing the CNN model on different noise intensities across different sizes of training data with noise. Each line represents a different testing noise threshold.

Minimal Noise Injection

Next, our model was trained on noise-free images followed by retraining with only a small subset of noisy images to observe the difference in the testing accuracy (Figure 2.4). Once we have a trained network, it is beneficial to estimate how much retraining may be required to make it robust to noise.

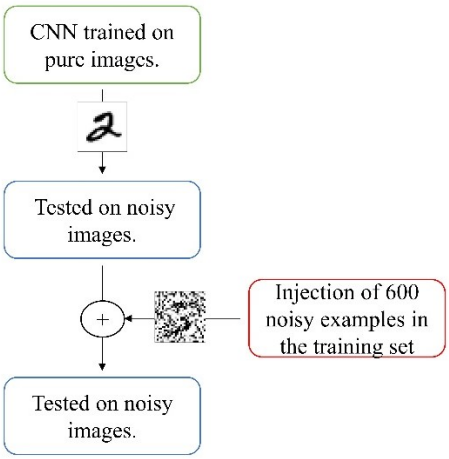

Figure 2.4 Diagram representing the training and retraining of the model.

The initial training of 10 epochs consisted of 60,000 noise-free images and was tested on 10,000 noisy images of a .75 noise intensity (Figure 2.2 (d)). The testing attained an average of 23% (Figure 2.5) without further improvement over time. Subsequently, we retrained the network on 10 epochs with only 600 noisy examples. Our model's accuracy increased to an average of 92% (Figure 2.5). Thus, the model's testing accuracy on noisy images drastically increases after retraining with only a few noisy images.

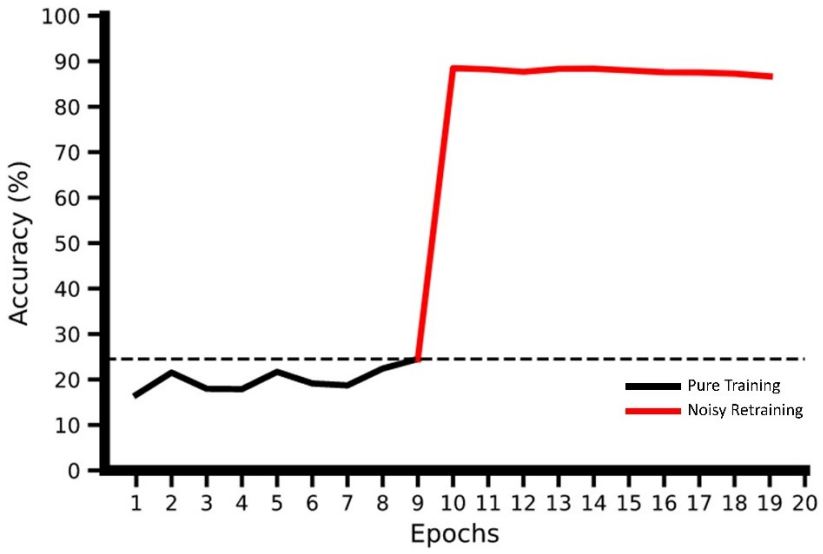

Figure 2.5 Accuracy of training the model with noise-free images, identified as the black solid line, followed by retraining, identified as red solid line, with 600 noisy images across time.

Correlated Noise Testing

Next, since our model was trained exclusively on uncorrelated noise, we aimed to test it on correlated noise to examine the robustness to noise of the CNN. To generate correlated noise images, we added evenly spaced horizontal bars in the background of the MNIST digits (Figure 2.6).

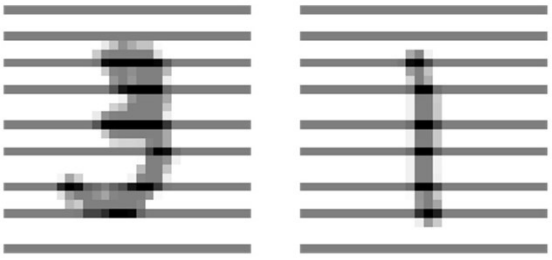

Figure 2.6 MNIST digits with horizontal bars in the background defined as correlated noise images.

We compared the testing accuracy of two models with identical architecture (Figure 2.1). The first model was trained solely on noise-free images and the second model implemented the NNF training method where only 20% of the training dataset consisted of Gaussian noise images. The results are shown

in Figure 2.7. The model with noise-free training had an average of 36% testing accuracy and the NNF training model had an average of 99% testing accuracy. The NNF training technique markedly improved the overall noise robustness of the model.

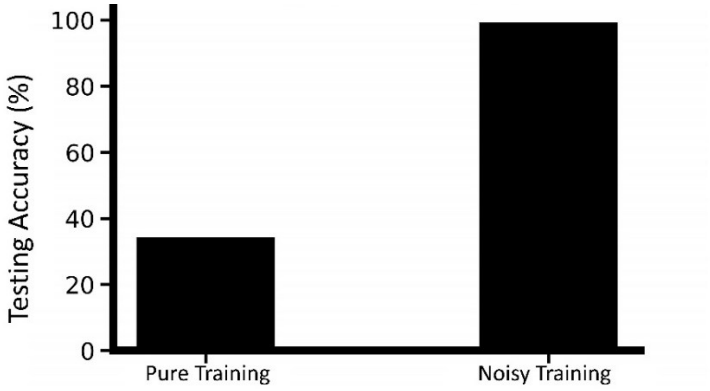

Figure 2.7 Testing correlated noise on two identical models but with different training techniques.

To demonstrate the impact of only few uncorrelated noise images on the correlated noise testing accuracy, we first trained a model on noise-free images, then retrained it with 600 noisy images. As demonstrated in figure 2.8, retraining the model with only a few noisy images increases the accuracy from 34% to 94%. To summarize, a model initially trained on noise-free images can become noise robust to correlated noise after being retrained with a subset of images with uncorrelated noise.

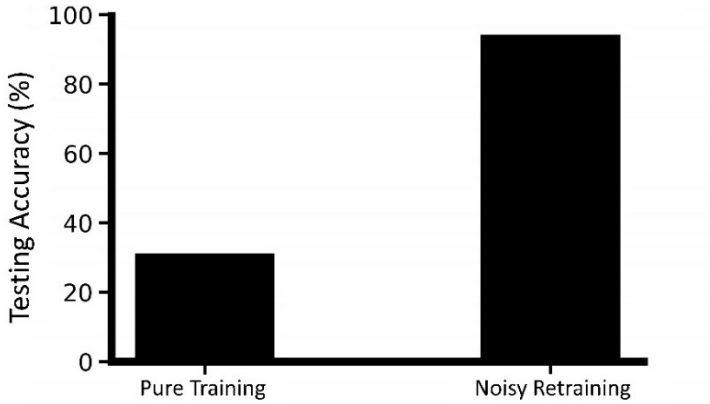

Figure 2.8 Testing correlated noise on a model that was initially trained on noise-free images followed by retraining with 600 noisy exemplars.

Discussion

The sensitivity of CNNs to noise is a largely unresolved problem; here, the contribution of our work is to address this problem by a novel method of NNF training where we systematically control the intensity and proportion of noisy images given as input to a CNN. We explored the link between the ratio of noisy-noise-free images and the testing accuracy on noisy images.

In the first set of experiment, we began by running simulations with different ratios of noisy and noise-free images while testing on different Gaussian noise intensities. Our custom-built CNN was trained on the MNIST dataset with 0.75 noise intensity factor (Figure 2.2(d)). Our results demonstrated that only a small percentage of noisy images is necessary in the training set for an accurate testing performance. As shown in Figure 2.3, only 10-15% of noisy images in the training set were required to achieve maximum accuracy. Nevertheless, if a test image contains a larger noise intensity factor, it will require more noisy examples in the training set. For the highest accuracy on a test image with a noise intensity factor of 1.00 (Figure 2.2 (e)), a training set consisting of 20-30% of noisy images is required. However, an image with such intensity is an extreme case. The opposite is also true: if images in the test set contain minimal noise, the network will require fewer noisy examples in the training set. For instance, an image with a noise intensity factor of 0.25 (Figure 2.2 (b)), will require 0.1% of noisy examples in the training.

All testing conditions revealed that surpassing the required number of noisy examples did not improve the accuracy of the model. As demonstrated in Figure 2.3, each condition of noise intensity attains a plateau in accuracy after reaching a certain percentage of noisy images in the training dataset. This allows more flexibility with the training of the model and selection of dataset. For example, extra exemplars in the training dataset could include additional noisy training conditions such as impulse, salt, pepper, and others for producing a more noise robust network. Further research is necessary to verify the minimal number of noisy examples for each type of noise during the training phase. It has been shown that a model trained on different types of noise has an advantage when dealing with noise in future data

when compared to a model strictly trained on noiseless data (Nazaré, Costa, Contato, & Ponti, 2017). Noise robust networks are necessary for successful applications of deep learning to real-life situation (Shafiee, Jeddi, Nazemi, Fieguth, & Wong, 2020). Additionally, the NNF training method allows to reduce the computational cost without sacrificing testing accuracy by eliminating unnecessary examples in the training set.

The reasoning behind successful training with just a small number of noisy examples is not fully understood. It is possible to consider that only a certain number of noisy images are necessary to achieve an optimal global minimum. A stream of noisy inputs with the same intensity and size can also cause the model to stop learning after a certain number of examples, since the network only needs a limited number of images to learn patterns and features (Figueroa, Zeng-Treitler, Kandula, & Ngo, 2012).

In the second set of experiments, we explored the performance of a CNN model that was retrained with a subset of noisy examples. We started by training the CNN with only noise-free images followed by testing on noisy images with an intensity factor of 0.75 (Figure 2.2 (d)). As shown in Figure 2.5, the testing accuracy averaged around 20%. Then, we retrained our network with only 600 noisy examples with the same intensity as the testing images and achieved an accuracy of 90%. This result demonstrates that CNNs can be efficiently retrained with only a few noisy examples with minimal computational cost since the whole process took 15 seconds (when trained with NVIDIA GeForce RTX 2060 SUPER GPU).

In the third set of experiments, we evaluated the robustness of our model when trained on uncorrelated noise and tested on correlated noise. In addition, to establish a concrete comparison, we trained another identical model with only noise-free images and tested on correlated images. Our results (Figure 2.7) revealed that the model trained with noisy uncorrelated images was efficient at predicting noisy correlated images with an accuracy of 99%. The model trained only on noise-free images achieved a testing accuracy of 34% when tested on correlated noise images. Additionally, we retrained the noise-

free model with a subset of 600 noisy exemplars and observed an increase in the testing accuracy on correlated noise images from 34% to 94% (Figure 2.8). Our model's high prediction accuracy on correlated noise demonstrates its capability of generalization (Figure 2.9) to unfamiliar stimuli (Kawaguchi, Kaelbling, & Bengio, 2017).

In Figure 2.9, we demonstrated the generalization of a model trained only on noise-free images and a model trained with NNF technique. As previously revealed, the model trained with NNF method has a higher generalization across different data features when compared to a network trained solely on noise-free images. The NNF network performs better on unfamiliar images, such as correlated noise exemplars, even without being exposed to this condition during training phase. Simultaneously, the network trained only with noise-free images has a poor accuracy across several different data features and exerts a poor robustness to noise.

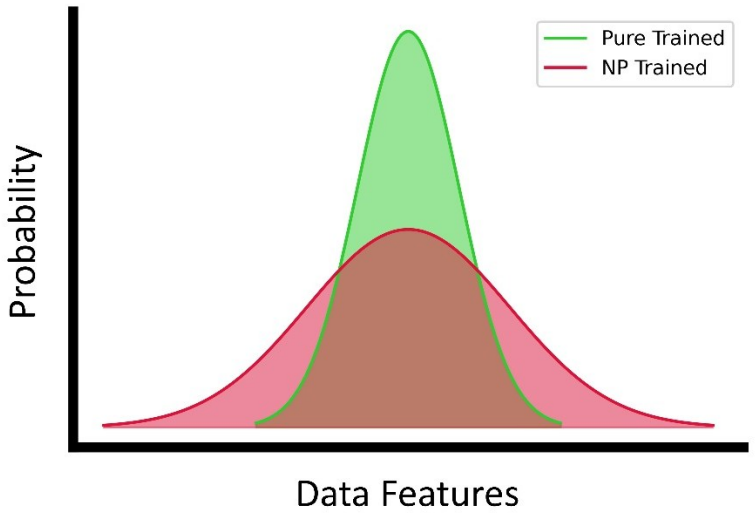

Figure 2.9 Width of generalization to different data features. Green distribution represents a model solely trained on noise-free images and red distribution represents a model trained with NNF technique.

Our NNF training technique leaves an open question of whether the training method is more important than the architectural complexity of the model. The significance of properly training any

artificial neural network has been overlooked by the deep learning community. The current literature has focused primarily on the number of layers contained in ANNs and their configuration rather than the contents of the training data. In our research, we demonstrated that a simple model with only twelve layers can achieve high testing accuracy on noisy images if properly trained. The NNF training method can also help to achieve broader generalization to new stimulus by allowing extra room within the training dataset for different types of noisy conditions. Future research should explore the possibility of reducing the computational cost associated with excessively deep neural networks by incorporating a smaller network with an efficient training technique. Moreover, it is important to balance out the complexity of the model and the training data to obtain the most efficient model.

Future research should focus on exploring the NNF training method with more complex dataset and different CNN networks. Our model was solely trained and tested on MNIST which is a low-quality greyscale dataset.

Conclusion

In this chapter, we proposed a novel training technique by only using a small set of noisy examples for achieving maximum testing accuracy with minimal computational cost. To best of our knowledge, this is the first work that explored the link between the ratio of noisy and noise-free images and testing accuracy in noisy conditions. We additionally demonstrated that a CNN model can be successfully retrained with only a small number of noisy examples to achieve a high testing accuracy when exposed to a new stimulus.

Acknowledgement

This work was supported by a Discovery grant to J.P.T. from the Natural Science and Engineering Council of Canada (NSERC Grant No. 210977).

References

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)* (pp. 1-6). IEEE.

Charan, S., Khan, M. J., & Khurshid, K. (2018, March). Breast cancer detection in mammograms using convolutional neural network. In *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)* (pp. 1-5). IEEE.

Deng, L. (2012). The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine, 29*(6), 141-142.

Dodge, S., & Karam, L. (2016, June). Understanding how image quality affects deep neural networks. In *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)* (pp. 1-6). IEEE.

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making, 12*(1), 8.

Grandvalet, Y., Canu, S., & Boucheron, S. (1997). Noise injection: Theoretical prospects. *Neural Computation, 9*(5), 1093-1108.

Kawaguchi, K., Kaelbling, L. P., & Bengio, Y. (2017). Generalization in deep learning. *arXiv preprint arXiv:1710.05468*.

Nazaré, T. S., Costa, G. B., Contato, W. A., & Ponti, M. (2017, November). Deep convolutional neural networks and noisy images. In *Iberoamerican Congress on Pattern Recognition* (pp. 416-424). Springer, Cham.

Shafiee, M. J., Jeddi, A., Nazemi, A., Fieguth, P., & Wong, A. (2020). Deep Neural Network Perception Models and Robust Autonomous Driving Systems: Practical Solutions for Mitigation and Improvement. *IEEE Signal Processing Magazine, 38*(1), 22-30.

Zur, R. M., Jiang, Y., Pesce, L. L., & Drukker, K. (2009). Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. *Medical Physics, 36*(10), 4810-4818.

Chapter 3

Enhancing Out-of-Distribution Learning in Computer Vision through Dominant Feature Masking

Artem Pilzak[1], Jean-Philippe Thivierge[1,2]

[1] School of Psychology, University of Ottawa, K1N 6N5, Ottawa, Canada
[2] Brain and Mind Research Institute, University of Ottawa, K1N 6N5, Ottawa, Canada

Abstract

Out-of-distribution (OOD) learning presents a major challenge in machine learning as models must effectively generalize to previously unseen data. This challenge is prevalent in deep learning models, which tend to focus on the most dominant features in images. This narrow focus impedes OOD learning, where critical features are concealed or absent during testing, leading to reduced prediction accuracy. To address this issue, we introduce a novel data augmentation approach termed Dominant Feature Masking (DFM), inspired by human visual holistic processing. DFM strategically conceals and reveals the most prominent features within images, allowing neural networks to simultaneously capture both dominant and non-dominant attributes, thereby enhancing adaptability to OOD data. We evaluated DFM using a novel set of learning challenges termed Versatile Evaluation Benchmark (VEB), which assesses model performance on three distinct tasks: (*i*) augmented MNIST images to test resilience against diverse transformations; (*ii*) a novel dataset of unseen image classes to examine performance on new instances within familiar categories; and (*iii*) a dataset created by DALL-E to challenge class differentiation with artificially mixed features. Our results demonstrate that DFM significantly improves OOD generalization compared to traditional augmentation techniques, achieving marked enhancements across various conditions without compromising in-distribution testing accuracy. These findings underscore the potential of DFM to improve the performance of computer vision systems in various real-world scenarios, making them more robust and adaptable to unexpected data variations. By leveraging VEB, researchers will gain a deeper understanding of their models' generalization performance, ensuring that CNNs are well-equipped to handle the complexities of real-world applications. The source code and VEB datasets are available at https://github.com/DeepVisionary/DFM .

Keywords*:* Out-of-Distribution Learning; Convolutional Neural Networks; Data Augmentation; Feature Acquisition; Domain Generalization.

Introduction

In recent years, computer vision has achieved multiple milestones, enabling machines to identify objects, discern patterns, and generate insights from images and videos (Feng et al., 2019; Khan et al., 2021; Voulodimos et al., 2018). Applications span from autonomous vehicles navigating complex roadways (Janai et al., 2020) to improved medical diagnoses facilitated through image analysis (Gao et al., 2018), impacting various industries and opening new frontiers of application.

However, a challenge within the field of computer vision is the problem of out-of-distribution learning (Hendrycks et al., (2021), where a network must generalize to entirely novel and previously unseen testing samples. This becomes increasingly pertinent as the scope of computer vision expands, leading to scenarios where novel data significantly diverges from the original training set (DeVries & Taylor, 2018; Fort et al., 2021). This divergence is compounded by the fact that it is unfeasible to acquire data that can represent all possible exemplars, especially when tackling unique tasks where data is limited. Such disparities between training and novel data can expose conventional models to errors. Addressing this challenge necessitates the development of innovative algorithms and strategies that empower computer vision systems to detect anomalies, adapt to unfamiliar environments, and maintain a high level of precision and reliability, even when confronted with unfamiliar inputs. Achieving robust out-of-distribution (OOD) learning is pivotal in ensuring the practical viability and effectiveness of computer vision solutions across diverse domains, while striving to rival the perceptual abilities of the human visual system.

In the realm of machine learning, an issue emerges when a neural network becomes overly fixated on learning the most prevalent feature within the training dataset. This phenomenon, described herein as 'fixated learning', occurs when the network's training data is biased toward a specific feature or characteristic (Storcheus et al., 2015; Tang et al., 2022; Ye et al., 2021). For instance, consider a neural network designed to recognize cats. Due to the nature of the training data, the network may

predominantly encounter images of cats' faces. As a result, the network's learning process may become disproportionately focused on this dominant feature. As the network trains, it increasingly anchors its decision-making on the detection of facial features.

The functionality of fixated learning is rooted in the inherent design of neural networks, which aim to minimize error during training. As these networks iterate through the data, they naturally gravitate towards features that recur most frequently, as recognizing and leveraging these features can expediently reduce the overall error across samples. Over successive training epochs, this inclination can become deeply ingrained, leading the network to become heavily reliant or 'fixated' on these dominant features. This can lead to a biased weighing towards these dominant features when making predictions, to the impairment of the overall performance.

The network's adaptability to variations or OOD samples diminishes because of this fixation. It becomes less capable of effectively recognizing exemplars that deviate from the training data, such as those with obscured or partially visible features, or those with novel attributes. While this might enhance accuracy on familiar (i.e., in-distribution) data, it can compromise the network's ability to generalize to OOD samples or scenarios where dominant features are absent or obscured.

Here, we consider the idea that for a network to proficiently infer OOD samples, it must transition from an over-reliance on dominant features of the training data (Figure 3.1 (a)), to embracing a broad spectrum of features present in the data (Figure 3.1 (b)). We introduce Dominant Feature Masking (DFM), an approach inspired by the concept of holistic processing observed in human perception (Richler et al., 2012). DFM guides networks to perceive images more comprehensively by selectively masking dominant features, thereby compelling the model to consider the entire visual content. Such an approach not only broadens the feature spectrum from which the network learns but also mirrors the human ability to understand complex visuals as an integrated whole rather than disjointed parts.

Holistic Processing and Dominant Features

Inspired by the techniques used by medical experts, who employ holistic processing to interpret complex medical imagery, there is a recognized need for neural networks to adopt a similar comprehensive view when analyzing visual data. According to Sheridan and Reingold (2017), medical experts often seamlessly integrate diverse visual information to make rapid and accurate diagnostic decisions. This global approach to visual analysis allows them to quickly detect critical abnormalities, often before these elements are directly in their line of sight, as demonstrated through eye-tracking studies. These findings underscore the capability of human experts to perceive and process medical images holistically, utilizing their entire field of view rather than focusing solely on the most obvious or prominent features.

In contrast, novice medical practitioners typically exhibit a more constrained approach to image interpretation. Novices tend to focus on specific, salient features within an image. This approach often results in slower diagnostic times and increased likelihood of overlooking critical subtleties in the images. For instance, while an expert might quickly identify a subtle anomaly in a peripheral area of a radiograph by integrating contextual visual cues from the entire image, a novice might miss the same anomaly due to a focus on the more central or prominent features. This difference in approach highlights the advantages of holistic processing, not only in human experts but also as a desirable capability for neural networks across various applications.

The disparity between the holistic approach of experienced practitioners and the more feature-focused strategy of novices emphasizes the importance of training and experience in developing effective diagnostic skills. Similarly, this contrast serves as a foundation for the principles behind DFM. By mimicking the holistic processing strategies of seasoned experts, DFM aims to train neural networks to overlook initial biases towards dominant features and instead engage in a broader, more integrative view of the visual data. This strategy ensures that networks do not merely recognize the most apparent features but

are also sensitive to subtler, yet diagnostically critical features that might otherwise be ignored. Thus, DFM not only addresses key challenges in machine learning but also aligns neural network training closer to the perceptual and cognitive strategies employed by human experts.

Additionally, DFM is user-friendly, requiring minimal manual tuning as it can be effectively implemented by simply adjusting a small number of parameters as described below.

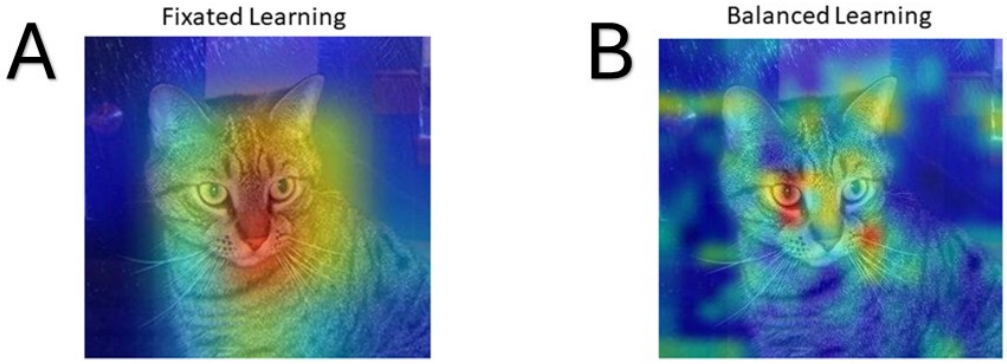

Figure 3.1 Comparing feature importance in fixated and balanced learning. a) Illustrates fixated learning with a predominant feature driving most decisions. b) Depicts balanced learning where several less-dominant features collectively influence the prediction. In (a) and (b), the heatmap depicts the importance attributed to features within the image, from blue (low importance) to red (high importance)

In DFM, the most dominant features are first identified using the network's Gradient-weighted Class Activation Mapping (Grad-CAM), which is a technique that uses the gradients of the target output flowing into the last convolutional layer to produce a heatmap highlighting important regions in the image for making predictions (Selvaraju et al., 2016; Selvaraju et al., 2017; Panati et al., 2022). Then, these features are systematically masked by applying a specific overlay to the identified regions, effectively obscuring them to vary the input data and improve the robustness of the model. The objective is to cultivate a broader feature acquisition process, thereby enhancing the network's resilience and adaptability when faced with OOD data (Bengio et al., 2013).

Domain Generalization Benchmark

To ensure the reliability of convolutional neural networks (CNNs), they must possess several key capabilities. Firstly, CNNs need to generalize effectively to new instances of the same class, handling variations in features or attributes that were not present during training (Arpit et al. 2017). This ensures that the model can maintain high accuracy and performance in real-world scenarios where such variations are common. Additionally, CNNs should be able to recognize and accurately predict images that have undergone transformations, such as cropping, occlusion, rotation, and the introduction of noise (Azulay, & Weiss, 2019). This ability to handle transformed images is crucial for maintaining robustness in diverse environments. Lastly, CNNs must be resilient against adversarial attacks, which involve inputs specifically designed to mislead the model (Goodfellow, Shlens, & Szegedy, 2014; Kurakin, Goodfellow, & Bengio, 2018). By being robust against such attacks, CNNs can ensure the security and reliability of their predictions in practical applications.

In the field of domain generalization (DG), several benchmark datasets, such as PACS, Office-Home, and DomainNet, have been widely adopted to evaluate the robustness and generalization capabilities of CNNs. These datasets incorporate images from multiple domains, simulating real-world variations that models might encounter. However, these datasets do not fully capture the specific challenges faced in everyday deployment of CNNs (Zhou et al., 2016). Many DG datasets do not sufficiently encompass the breadth and diversity of real-world domains. For example, PACS primarily includes images of objects in different artistic styles, such as photos, sketches, and paintings, which do not adequately reflect the complexity of real-world variations. Similarly, Office-Home comprises images of objects from different office and home environments, but the variations are limited to background changes and different perspectives, lacking the scope of real-world diversity. DomainNet, which includes images from multiple domains like clipart, infographics, and real-world photos, still falls short in presenting a comprehensive range of variations encountered in practical applications (Niu et al., 2022).

To convincingly demonstrate domain generalization, it is essential to show the ability to generalize to novel instances of a class, specifically those not seen during training. Presenting the same digits in different colors or with different backgrounds does not address the core problem of DG. Similarly, showing different cars from various angles and backgrounds does not fully capture the challenge of generalizing to new instances of a class. What is needed are datasets that include entirely new instances within the same class, exhibiting attributes and features not present in the training set. This approach will test the network's capacity to recognize fundamental characteristics of a class and adapt to unseen variations, thereby providing a more rigorous evaluation of generalization capabilities in real-world scenarios (Wang et al., 2021; Wang et al., 2022). For example, consider a Fruits and Veggies dataset as demonstrated in Figure 3.2. The training set might include common fruits like bananas, apples, and oranges, and vegetables like carrots, lettuce, and broccoli. However, the testing set would include new instances not seen during training, such as pineapples, mangoes, and pears for fruits, and eggplants, zucchinis, and bell peppers for vegetables. This intra-class variation ensures the model is tested on its ability to generalize beyond the specific instances it was trained on.

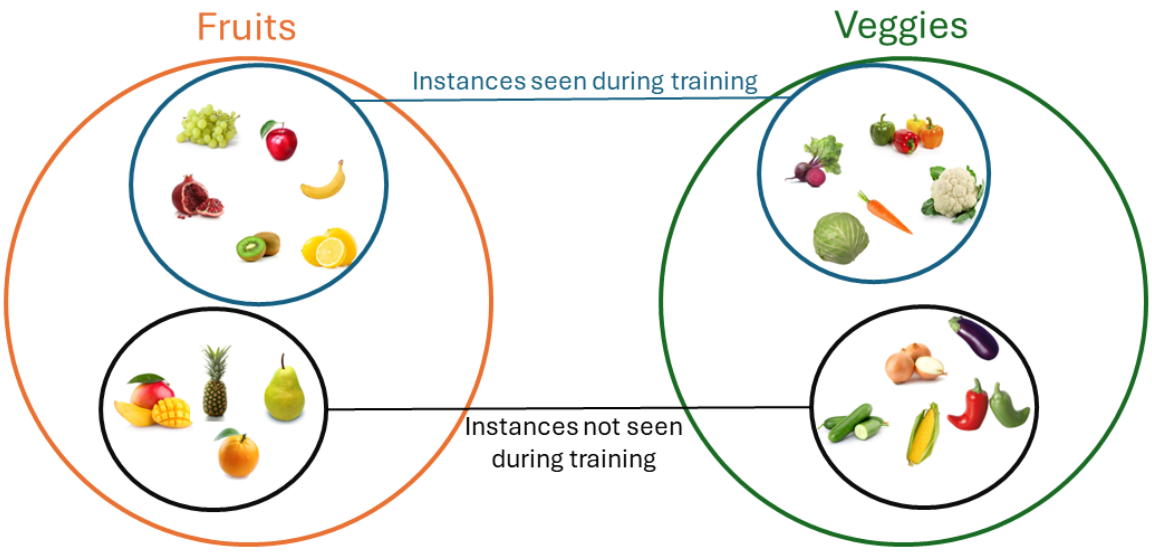

Figure 3.2 The concept of intra-class variation within the Fruits and Veggies dataset, highlighting the instances seen during training and those introduced during testing.

Furthermore, these datasets typically do not include adversarial samples, which are inputs intentionally designed to deceive models. This omission makes models trained on these datasets vulnerable to adversarial attacks, potentially compromising their robustness and security in practical applications (Chen et al., 2024; Zheng, Huai, & Zhang, 2024). Additionally, DG datasets lack corrupted images—images exposed to domain-shift feature variations like cropping, occlusion, rotation, and noise (Yoon et al., 2023). These types of image corruptions are common in real-world scenarios where images can be partially obscured, captured at different angles, or subjected to environmental noise. The absence of these corrupted images in DG datasets means that models are not adequately evaluated to handle such variations, leading to a significant gap in their robustness and generalization capabilities.

To address the limitations of current DG datasets, we propose a new benchmark designed to test DG, called the "Versatile Evaluation Benchmark" (VEB), consisting of three datasets: MNIST, Fruits and Veggies, and Cats and Dogs. Although state-of-the-art CNNs achieve extremely high precision on the in-distribution versions of these datasets, the proposed challenges will reveal how well these networks can truly generalize to more complex and varied scenarios. By introducing modified versions of these datasets, VEB provides a rigorous assessment of a model's ability to handle OOD conditions, ensuring that the networks are not only accurate on familiar data but also robust and adaptable in the face of real-world variability.

The first dataset included in VEB is the MNIST dataset, an improvement of the original Rotated MNIST (Ghifary, et al., 2015), which requires the network to learn the digits in their standard form, but where the testing set includes transformed images of the same digits with previously unseen modifications. These transformations include cropping, rotation, Gaussian noise, and vertical and horizontal pattern noise. By introducing these corrupted images, the challenge aims to test the network's ability to maintain high accuracy and performance when faced with variations that are common in real-

world scenarios. This aspect of the benchmark is crucial for ensuring the robustness of CNNs against everyday image distortions and transformations.

The second dataset is the Fruits and Veggies challenge, the network will be trained to recognize specific fruits and vegetables, but the testing set will feature new instances of these classes that the network has never seen before (Figure 3.2). This challenge is designed to evaluate how well the network generalizes to new instances within the same class, testing its ability to handle intra-class variability. The objective is to assess the model's performance in recognizing the fundamental features of fruits and vegetables despite differences in appearance, shape, and texture that naturally occur. This challenge addresses a critical aspect of generalization that has not been directly tested using previous benchmarks. This will demonstrate the model’s capacity to generalize from the training data to new, unseen examples, which is essential for practical applications where data can vary significantly.

Finally, VEB includes the Cats and Dogs dataset, where the network will be trained on standard images of cats and dogs but tested on DALL-E generated images of cats wearing dog masks and dogs wearing cat masks. This challenge is intended to assess the network's performance under adversarial conditions, where key features for recognition are intentionally masked. By introducing such adversarial examples, the challenge aims to test the resilience of the network against attempts to deceive it. This is critical for applications requiring high security and reliability, where adversarial attacks could compromise the system's integrity.

These specific challenges aim to evaluate and enhance the robustness, generalization, and adversarial resilience of CNNs. By addressing these aspects, the VEB will help to ensure that CNNs are not only high-performing on standard datasets but also capable of handling the complexities and variations encountered in real-world applications.

Contributions

The contributions of this study can be summarized as follows:

(a) Introduction of a novel data augmentation approach termed DFM, which involves applying masks to the predominant features extracted from a class activation map derived from the last convolutional layer of a trained model. The extent of masking can be adjusted using a single parameter (α), ranging from zero to one.

(b) Introduction of VEB, a comprehensive framework for evaluating the generalization capabilities of models to OOD samples. VEB consists of three unique datasets designed to test different aspects of model robustness: (i) augmented MNIST images to evaluate resilience against common data transformations; (ii) a collection of previously unseen images to gauge the ability to generalize to new examples within known categories; and (iii) a dataset generated by DALL-E, designed to challenge class differentiation by artificially mimicking features of opposing classes. This benchmark enables a thorough evaluation of DFM and other data augmentation strategies, ensuring that models are assessed under diverse and challenging conditions.

The remainder of this work is structured as follows. First, the Related Work section examines different data augmentation techniques for increasing generalization and robustness of neural networks. Second, the Methods section presents the two neural networks employed for this research and the training data, followed by a description of DFM. In the Results section, we detail the performance of models utilizing DFM alongside various data augmentation techniques during both in-distribution and OOD testing. The Discussion section contrasts DFM with existing data augmentation techniques, evaluating their effectiveness across three distinct datasets. Additionally, we delve into the limitations of DFM and propose directions for further exploration.

Related Work

Data Augmentation

In computer vision, data augmentation serves as a foundational technique aimed at enhancing a neural network's capacity to handle OOD data (Liang et al., 2017; Liu et al., 2022). Data augmentation

achieves this by subjecting the original training set to a variety of controlled transformations, introducing variability into the learning process (Shorten & Khoshgoftaar, 2019). This variability empowers the network to extract more generalized and resilient features from the training data.

The adoption of data augmentation has become widespread, particularly for its ability to boost neural network performance, especially when confronted with limited training data (Shorten & Khoshgoftaar, 2019). Traditional data augmentation methods involve the application of random transformations, such as rotations, translations, and flips to the original images (Maharana et al., 2022). This technique expands the size of the training dataset without requiring the acquisition of additional labeled samples. As a result, neural networks become more adept at handling the variations encountered in real-world scenarios. This learning experience reinforces their innate ability to adapt and extrapolate beyond the boundaries of the training distribution.

Nonetheless, it is essential to acknowledge a critical aspect of many data augmentation techniques: these transformations are executed randomly on the input images, without control over which features are masked (Mumuni & Mumuni, 2022). While this randomness contributes to diversifying the training data and enhancing the network's robustness, it also raises questions about its potential impact on the ability of the network to extract specific features from the data. Given the lack of control over what features are masked or concealed, some crucial features may be absent from the learning process. This concern becomes pronounced when attempting to mask specific regions of interest within images via data augmentation. A delicate balance between enhancing robustness and preserving critical features in the learning process highlights the challenge posed by data augmentation in the development of high-performance models (Taylor & Nitschke, 2018)

As an alternative, data augmentation may focus on other, non-random methods, including color adjustments, geometric transformations, or domain-specific operations (Ratner & al., 2017; Taylor & Nitschke, 2018). Nevertheless, implementing these modifications can be labor-intensive and may

necessitate manual adjustments to individual image. For instance, color adjustments may need fine-tuning to maintain visual quality, geometric transformations may involve precise resizing or reshaping, and domain-specific operations may demand tailored modifications to fit a dataset's unique characteristics. Balancing these advantages against the manual labor involved is essential when selecting the appropriate data augmentation strategy for a given machine learning task.

Mixed Sample Data Augmentation

Mixed sample data augmentation is an approach designed to enhance the diversity of training datasets in machine learning, particularly within the realm of deep learning and computer vision (Taylor & Nitschke, 2018). Unlike traditional augmentation techniques that apply simple transformations like rotation, scaling, or cropping to existing images, mixed sample data augmentation combines features or segments from multiple images to create novel composite samples (Zhang et al.. 2017). This methodology not only enriches the training data but also encourages models to learn more generalized representations by interpolating between different classes. Techniques such as MixUp, CutMix, GridMix, and RICAP exemplify this strategy (Baek et al., 2021; Takahashi et al., 2019; Yun et al., 2021; Zhang et al., 2017) each introducing unique variations in how images are blended and how their corresponding labels are combined, leading to improved model robustness and performance on a variety of tasks.

MixUp is one of the simplest yet most effective mixed sample data augmentation techniques (Zhang et al.. 2017). It creates new images by linearly interpolating between pairs of images and their labels. By blending images and their class labels in varying proportions, MixUp encourages models to generate predictions near class boundaries, thereby enhancing generalization and reducing overfitting. The advantage of this technique lies in its simplicity and the intuitive premise that linear combinations of features should lead to linear combinations of labels.

CutMix takes a slightly different approach by adopting a region-level mixing strategy (Yun et al., 2021). Instead of blending entire images, CutMix randomly cuts patches from one image and pastes them

onto another, mixing their labels in proportion to the area of the patch. This method forces the model to generate predictions based on incomplete information, enhancing its ability to focus on local features and improving robustness to occlusion and spatial variations. CutMix has shown effectiveness in challenging classification tasks where context and localized features play a crucial role (Yun et al., 2021).

GridMix extends the concept of image mixing by dividing images into a grid of cells and then randomly filling these cells with patches from different images (Baek et al., 2021). This grid-based approach generates a complex mosaic of features from multiple sources, compelling models to learn from a richer variety of patterns and contexts within a single training sample. GridMix can be particularly useful for tasks requiring fine-grained feature discrimination, as it diversifies the feature space available for learning in a structured yet randomized manner.

RICAP (Random Image Cropping and Patching) further diversifies the augmentation space by randomly cropping four different images and stitching them together to form a new composite image (Takahashi et al., 2019). The labels are mixed based on the contribution of each cropped region to the final image, encouraging models to recognize objects from partial views and enhancing their ability to generalize from limited or occluded visual information. RICAP challenges models to integrate disparate visual cues into coherent predictions, pushing the boundaries of what can be achieved with data augmentation.

While mixed sample data augmentation techniques have advanced the field by reducing overfitting and enhancing model generalization, they sometimes overlook the contextual coherence of images (Yan et al., 2024). That is, these methods might inadvertently crop and combine sections of images that lack meaningful content or that disrupt the natural structure of the objects being represented. For example, they could isolate a patch of fur without the context of the animal it belongs to, or slice through a critical feature, such as a wheel of a car, rendering it unrecognizable. This approach risks introducing ambiguity instead of clarity, teaching models to recognize patterns that are not indicative of physical

objects. The transition to techniques that respect the saliency of images promises a solution to this issue. By prioritizing the most informative parts of an image, saliency-focused augmentation ensures that the resulting training samples are not only diverse but also retain the essential characteristics necessary for models to learn realistic and meaningful representations.

Saliency in Neural Networks

Saliency in the context of neural networks refers to the identification and prioritization of the most influential parts of input data—typically, the sections of an image that significantly impact the network's output (Simonyan et al., 2013). In computational models, saliency detection helps highlight how neural networks perceive and process information, revealing the features deemed most important for making decisions or classifications. By analyzing saliency maps, which visually represent these critical areas, researchers can gain insights into a model's behavior, including its focus areas, potential biases, and areas for improvement. This understanding is crucial for refining performance, increasing interpretability, and ensuring fairness and transparency in automated decision-making processes.

Building upon the concept of saliency, SaliencyMix introduces an innovative approach to data augmentation that leverages the salient regions of images to create blended samples (Uddin et al., 2020). Unlike traditional augmentation methods that apply uniform transformations across the entire image, SaliencyMix selectively combines the most informative parts of different images, thus preserving the semantic significance of the mixed samples. It ensures that the resulting augmented images maintain a high degree of relevance and utility for training purposes. This technique not only enriches the training dataset with more diverse and challenging examples but also encourages models to learn from the most meaningful aspects of the data (Uddin et al., 2020). Consequently, SaliencyMix has the potential to enhance model robustness, particularly in complex classification tasks where discerning subtle differences between classes is essential. It represents a strategic melding of data augmentation with the nuanced

understanding of visual importance, paving the way for more sophisticated and effective training methodologies in machine learning.

Representation Self-Challenging

Representation Self-Challenging (RSC) (Huang et al., 2020) is a training strategy designed to enhance the generalization capability of CNNs by encouraging the network to leverage a broader set of features within the training data. RSC dynamically identifies and suppresses the most dominant features during training, forcing the network to utilize less prominent but potentially more generalizable features. This is achieved by monitoring the gradients of the feature representations across various layers. During each training iteration, RSC sets the gradients of the most predictive features to zero, effectively masking them and compelling the model to learn from other available features. This iterative process ensures that the network does not overly rely on a limited set of features, promoting a more diverse and robust representation of the input data.

Key differences can be highlighted between RSC and our proposed DFM approach. While RSC operates by dynamically masking gradients throughout the training process to encourage the use of non-dominant features, DFM explicitly identifies and masks dominant features using Grad-CAM, a visualization technique that highlights important regions in the input images (Selvaraju et al., 2016; Selvaraju et al., 2017; Panati et al., 2022). By applying these masks to the training data, DFM ensures that the model learns to recognize patterns from both dominant and non-dominant features, aiming to balance the use of these features. Additionally, DFM introduces a controllable parameter, α, that adjusts the intensity of masking, providing flexibility to tailor the level of feature suppression according to specific dataset characteristics. One significant advantage of DFM over RSC is its ability to visualize the decision-making characteristics of the network. By making it easier for users to understand which features the network relies on for its decisions, DFM enhances interpretability and transparency, making the model's behavior more accessible and comprehensible to users. Moreover, by focusing on the last convolutional layer, DFM targets high-

level, abstract features critical for final predictions, improving computational efficiency. RSC's gradient-based suppression across multiple layers, although thorough, can be more computationally intensive and demanding in terms of processing power, as it involves continuously updating gradients across the entire network. This makes DFM more efficient by concentrating its computational efforts on the most critical layer.

Methods

Custom neural network and ResNet-50

To assess the effectiveness of the proposed DFM method and benchmark it against various competing data augmentation strategies, we employed two distinct neural network models, namely a custom-designed model and ResNet-50. This dual-approach strategy was devised to analyze the impact of DFM across different model complexities, allowing us to observe its influence on both a relatively simple neural network architecture and a more sophisticated, cutting-edge model. This comparison provided a comprehensive understanding of how DFM performed in enhancing model robustness and generalization across the spectrum of neural network designs.

For the custom CNN, we designed an architecture within the TensorFlow Keras framework that comprises seventeen layers in total represented in Figure 3.3. Rectified Linear Unit (ReLU) activation functions were utilized across all layers, and weights were initialized using the He Uniform initializer (Lu et al., 2019). The initial layers, comprising three 2D convolutional layers with 3x3 filters, performed the initial feature extraction from input images. The first convolutional layer featured 32 filters, maintaining the same spatial dimensions as the input (160x160x3), followed by batch normalization and a max-pooling layer with a 3x3 pool size and stride of 3. This structure was then replicated with 64 filters in the second convolutional layer, followed by another batch normalization and max-pooling layer. Subsequently, a third convolutional layer with 128 filters, again followed by batch normalization and max-pooling, contributed to feature refinement. To prepare the data for fully connected layers, a flattening layer was introduced,

resulting in a vector of shape 1152. This vector was then connected to a dense layer with 200 units, applying the ReLU activation function and incorporating L2 regularization. The subsequent dense layers consisted of 100 units, 10 units, and a final output layer with a single unit, each employing ReLU activation, batch normalization, and L2 regularization. The final layer of our model is tailored to the classification type: a single-unit layer with a sigmoid activation function is used for binary classification, while a 10-unit layer with a SoftMax activation function is used for multi-class classification.

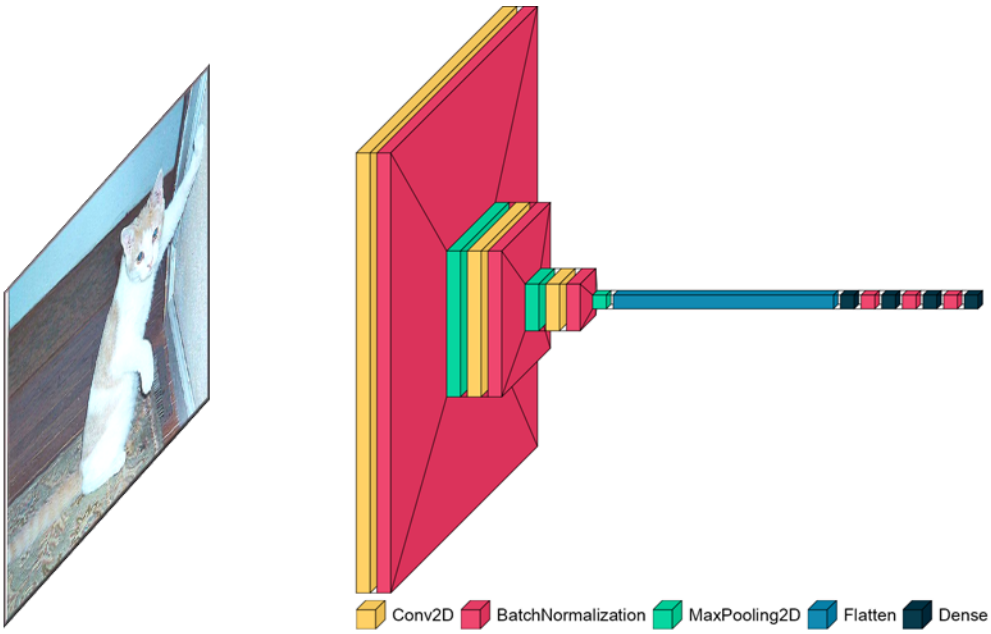

Figure 3.3 Architecture of the custom CNN

For the ResNet-50 network (Koonce & Koonce, 2021), the architecture of the model was enhanced for our specific needs by disabling the original top layer to integrate custom layers designed for our classification tasks. After flattening ResNet-50's output, we added two dense layers with ReLU activation, with the first having 1000 units and the second 500 units. The final layer of our model depends on the type of classification: for binary classification tasks, we used a single-unit layer with a sigmoid activation function, and for multi-class classification tasks, a 10-unit layer with SoftMax activation was employed.

For the training of both the custom CNN and the ResNet-50 models, we employed the Adam optimizer (Kingma & Ba, 2014) to facilitate backpropagation updates. Each model underwent a training regimen spanning 20 epochs, accommodating the varied complexities and requirements of every dataset involved. The training sessions were standardized with a batch size of 32 to ensure consistent gradient estimation across updates. Depending on the nature of the classification task—whether binary or multi-class—we tailored the loss functions accordingly, utilizing binary cross-entropy for binary classification tasks and categorical cross-entropy for multi-class scenarios. This approach allowed us to adjust the learning processes to the specific demands and characteristics of the input data, aiming to optimize performance and accuracy in distinguishing among the different classes presented in the training sets.

VEB

For our initial challenge and dataset evaluation, the MNIST dataset was employed (Deng, 2012). This dataset, comprising 60,000 training and 10,000 testing images of handwritten digits in a 28x28 single-channel format, was employed to test the custom model. The model underwent training on the original, unmodified images from MNIST. For testing, it encountered a variety of transformed images (see Figure 3.4), including random cropping, rotation, and the addition of Gaussian and patterned noise, aiming to assess its robustness and ability to generalize across altered data conditions reflective of real-world scenarios.

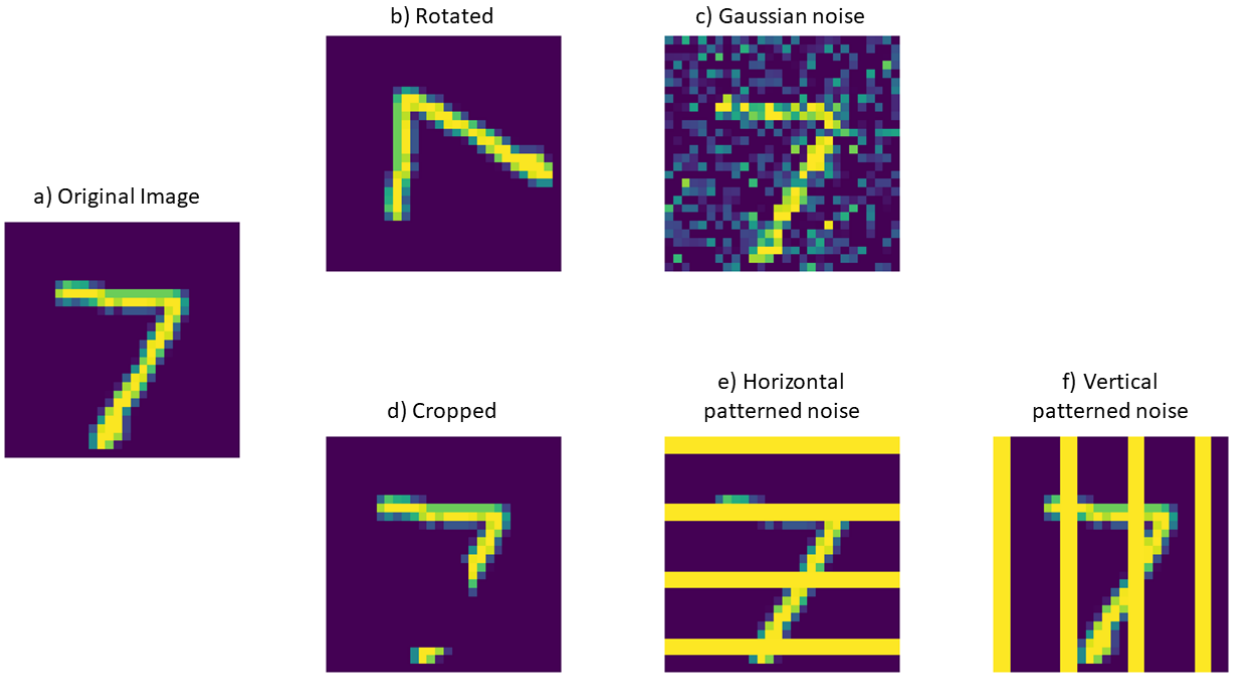

Figure 3.4 An example of a MNIST training sample (a) and transformed OOD sample (b-f) with novel attributes

The second challenge and dataset involved a classification task of fruits and vegetables, sourced from Kaggle (Seth, 2020). The training was conducted on a set of in-distribution images, including common fruits like bananas and apples, and vegetables such as cucumbers and carrots. This training set included 489 images of fruits and 518 of vegetables with each image resized to 150x150x3 pixel with RGB channels as demonstrated in Figure 3.5. To assess the generalization capabilities of our model, we tested it on OOD images, including fruits like watermelons and vegetables like eggplants, that were not seen during training. The test set comprised 348 fruit images and 790 vegetable images. This challenge aims to test the network's ability to generalize to new instances within the same class, evaluating its capacity to manage intra-class variability.

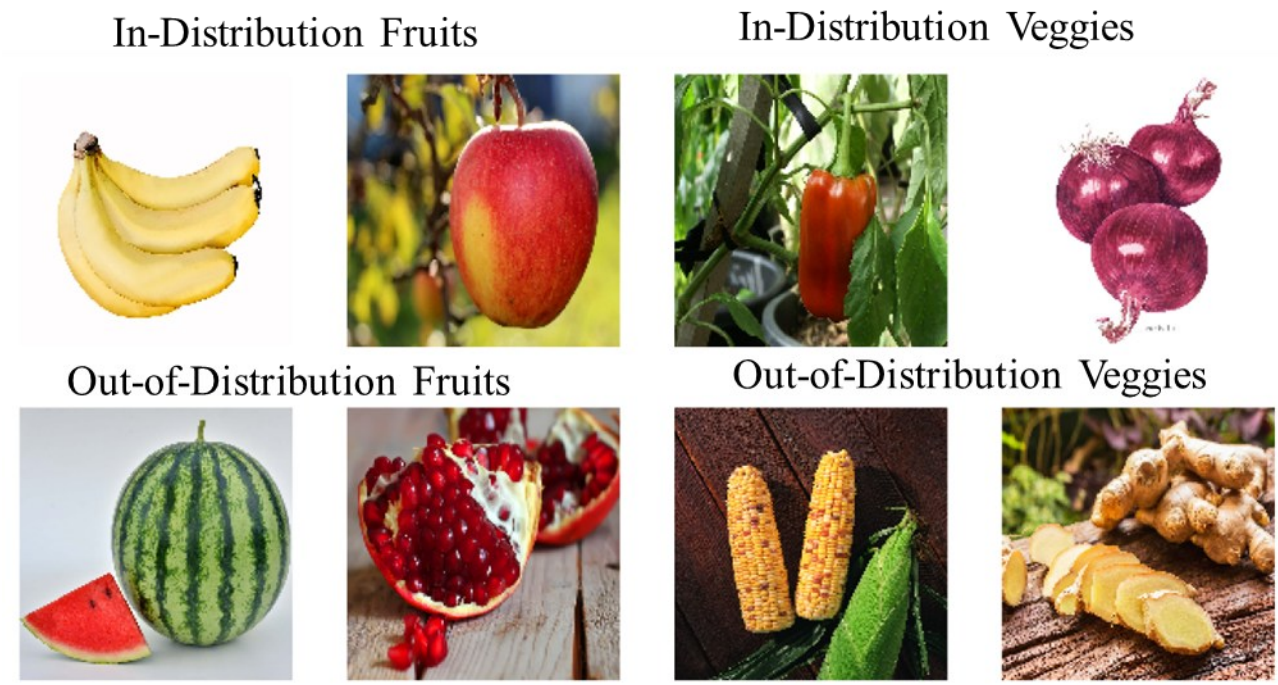

Figure 3.5 Training images of in-distribution fruits and veggies contrasted with OOD examples evaluated in generalization tests

For the third challenge, sourced from Kaggle (https://www.kaggle.com/datasets/tongpython/cat-and-dog), the dataset comprised two distinct categories: dogs and cats. It encompassed a total of 10,028 images distributed between 5,011 cat images and 5,017 dog images. We partitioned the dataset into 8,005 samples for training and 2,023 samples for in-distribution testing. All images are represented in RGB color format and possess dimensions of 160x160x3 (see Figure 3.6). For OOD testing, we crafted an evenly split set of 50 images, using DALL-E, that blend characteristics of both cats and dogs, such as cats with dog masks and dogs with cat masks. his approach was specifically designed to test the model's ability to distinguish between classes when presented with a mix of dominant features. This challenge aims to evaluate the network's ability to perform under adversarial conditions, where key recognition features are deliberately obscured.

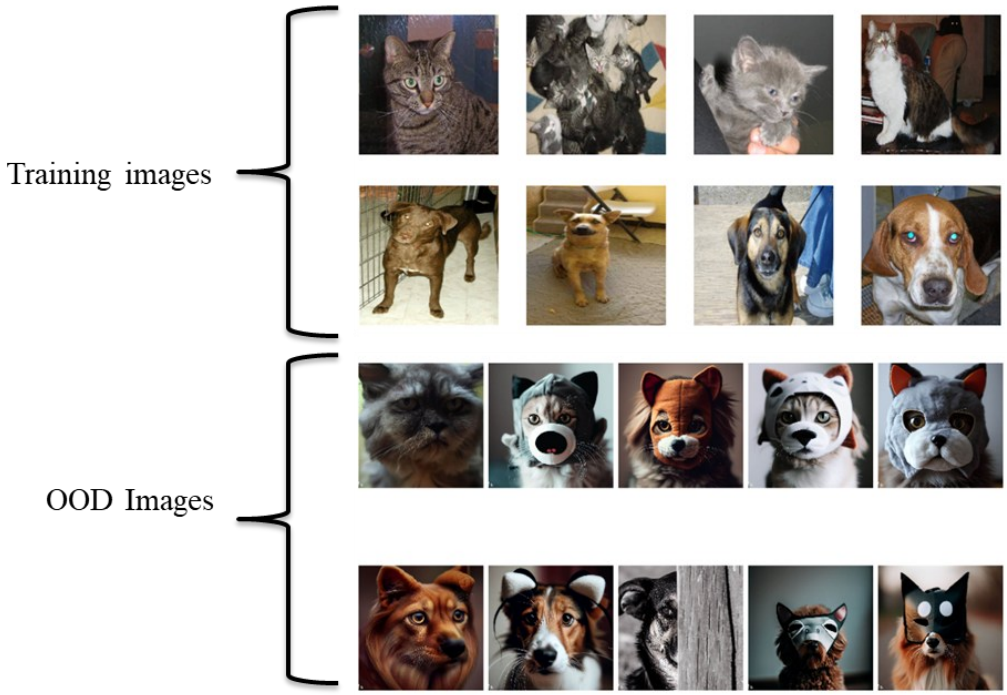

Figure 3.6 Training and OOD images for both classes

Grad-CAM

Grad-CAM was utilized to identify the essential features recognized by our custom deep learning model in image classification tasks, a technique previously employed to pinpoint significant elements within an input image (Selvaraju et al., 2017). Grad-CAM serves as an interpretive tool by highlighting areas within images that are pivotal to the model's output decisions predictions (Selvaraju et al., 2016; Selvaraju et al., 2017; Panati et al., 2022). We selected the last convolutional layer within this network, situated near the output layer, to implement Grad-CAM. The justification for this choice is that the final convolutional layer has the most comprehensive combination of features extracted from the input image, capturing high-level spatial information (Selvaraju et al., 2017). Through a forward pass of our dataset, we computed the gradients of the class scores relative to the feature maps of our chosen layer. These gradients then guided the creation of relevance heatmaps, which illustrated the influential features

driving the model's predictions. Adopting this method offered insight into the distinctive features in our dataset and enhanced the interpretability of the model's classification.

More formally, the Grad-CAM can be obtained via the following procedure. Given an input image $\mathbf{X}$ with dimensions $(w \times h)$, the feature maps $\boldsymbol{A}^k$ can be represented as:

$$\boldsymbol{A}^k(i,j) = f(\mathbf{X}), \tag{1}$$

where $f(\cdot)$ is the function representing the operations up to last convolutional layer in the CNN, k indexes the feature maps and (i,j) index the spatial dimensions of the feature map.

The gradient of the score for class $(c)(y^c)$ with respect to the feature maps A^k is calculated as:

$$\boldsymbol{G_k^c}(i,j) = \frac{\partial y^c}{\partial \boldsymbol{A}^k(i,j)}, \tag{2}$$

where the gradient $\boldsymbol{G_k^c}(i,j)$, which has 4 dimensions, represents how much each part of the feature maps $\boldsymbol{A}^k$ contributes to the class score y^c.

To derive the partial derivative $\frac{\partial y^c}{\partial A^k(i,j)}$, we first consider the gradient of the class score y^c with respect to the output of the fully connected layer o^{fc}:

$$\frac{\partial y^c}{\partial o^{fc}} = \mathbf{w} \cdot \frac{\partial g}{\partial o^{fc}}, \tag{3}$$

where $\mathbf{w}$ represents the weights from o^{fc} to y^c, and g denotes the activation function applied at the fully connected layer.

Next, we look at how the output of the fully connected layer changes with respect to the Global Average Pooling GAP output. GAP is a technique that averages the outputs across the spatial dimensions of feature maps, effectively reducing each map to a single scalar (Al-Sabaawi et al., 2020). This reduction

helps minimize overfitting by decreasing the number of trainable parameters and enables the model to handle various input sizes.:

$$\frac{\partial o^{fc}}{\partial \boldsymbol{A}^k(i,j)} = \frac{1}{N}. \tag{4}$$

In this equation, N is the total number of elements in the feature map $\boldsymbol{A}^k$ that contribute to the GAP output.

Finally, the gradient of y^c with respect to $\boldsymbol{A}^k(i,j)$ is obtained by combining these steps through the chain rule, yielding the relationship between the class score and the activations in the last convolutional layer:

$$\frac{\partial y^c}{\partial \boldsymbol{A}^k(i,j)} = \frac{\partial y^c}{\partial o^{fc}} \cdot \frac{\partial o^{fc}}{\partial \boldsymbol{A}^k(i,j)}. \tag{5}$$

Equation (5) summarizes the process of backpropagating the influence of the class score through the network to the activation of the last convolutional layer, illustrating the principle behind Grad-CAM and similar visualization techniques.

After obtaining the gradient at the last convolutional layer, we calculate the neuron importance weights β_k^c for class c by globally averaging the gradients G_k^c over the two spatial dimensions of the feature map,

$$\beta_k^c = \frac{1}{N}\sum_i \sum_j \boldsymbol{G_k^c}(i,j), \tag{6}$$

The weighted combination of feature maps is then computed as:

$$\mathbf{H}^c(i,j) = \text{ReLU}\left(\sum_k \beta_k^c \cdot \boldsymbol{A}^k(i,j)\right), \tag{7}$$

where $\mathbf{H}^c$ is the heatmap for class c, and the ReLU function ensures that only features with a positive influence on class c are considered.

Dominant Feature Masking (DFM)

After extracting the heatmap of a specific image in the dataset (Eq.7), we performed DFM to address the issue of fixated learning. Given the heatmap $\mathbf{H}^c$ with dimensions $(m \times n)$, where m represents the number of rows and n represents the number of columns, the objective is to resize it to $(w \times h)$ to match the dimensions of the original input. Using bilinear interpolation, the value at a new position (i', j') in $\mathbf{H}^c{}_{\text{resized}}$ is

$$\mathbf{H}^c{}_{resized}(i',j') = \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{H}^c(i,j) \cdot r(i',j',i,j), \tag{8}$$

where $r(i', j', i, j)$ is the interpolation kernel, calculating the original pixel $\mathbf{H}^c(i,j)$ contribution at (i', j'), defined as:

$$r(i',j',i,j) = (1-|i'-i|)(1-|j'-j|). \tag{9}$$

Here, (i', j') maps back to the original heatmap's coordinates, factoring the resize scale. This approach blends the four nearest neighbors in $\mathbf{H}^c$ to determine $\mathbf{H}^c{}_{resized}$, ensuring a smooth transition between pixel values.

In the thresholded heatmap, $\mathbf{H}^c{}_{resized}$ highlights the input image's regions that most significantly impact the CNN's predictions and directly correlates with its significance to decision-making. Higher

intensity values mark the most crucial regions influencing the model's predictions, pinpointing the critical features within the image.

Following the resizing of the heatmap (Eq.9), we applied a binary threshold governed by a free parameter α. This parameter delineates a threshold level, to mask features based on their significance in the decision-making process (Figure 3.7). The α parameter acts as a criterion for differentiating between features: a lower α threshold admits a broader range of features, encompassing both more and less dominant ones. Conversely, as α approaches 1, the selection criterion become stricter, isolating only the most significant features deemed important by the model,

$$\mathbf{H}^{c}{}_{thresholded}(i',j') = \begin{cases} 1\ if\ \mathbf{H}^{c}{}_{resized}(i',j') \geq\ \alpha \\ 0\ Otherwise. \end{cases} \quad (10)$$

After applying the above binary threshold, we obtained $\mathbf{H}^{c}{}_{thresholded}$, where a 0 indicates untouched areas of the original image and a 1 identifies dominant features targeted for masking. In processing the original image $\mathbf{X}$, each pixel corresponding to a 1 in $\mathbf{H}^{c}{}_{thresholded}$ at the same location is masked with a randomly generated color represented as a tuple of RGB values (r, g, b) selected from a predefined range, typically [0,255] for each color channel to cover the full spectrum of visible colors. This random color application is crucial to prevent the network from associating a particular class with a specific color, thereby encouraging the network to focus on structural features rather than color, with the aim of enhancing its generalization capability.

This process can be formally described as follows:

$$\mathbf{X}_{masked}(x,y) = \begin{cases} (r,g,b)\ , \mathbf{H}^{c}{}_{thresholded}(i',j') = 1 \\ \mathbf{X}(x,y), \mathbf{H}^{c}{}_{thresholded}(i',j') = 0. \end{cases} \quad (11)$$

where $\mathbf{X}_{masked}$ is the final image with masked dominant features. The whole DFM process is illustrated in the Figure 3.8.

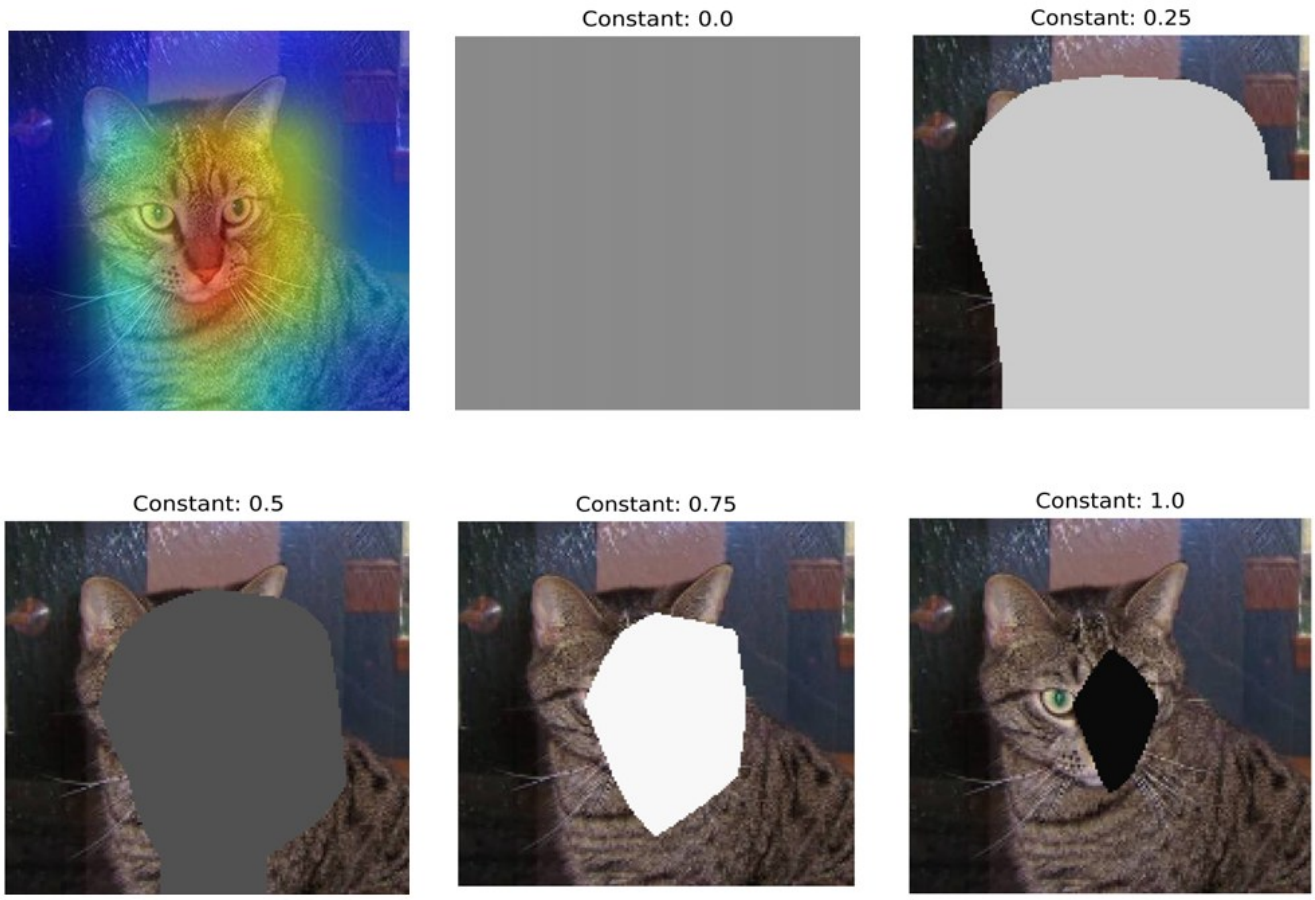

Figure 3.7 Impact of parameter α on DFM

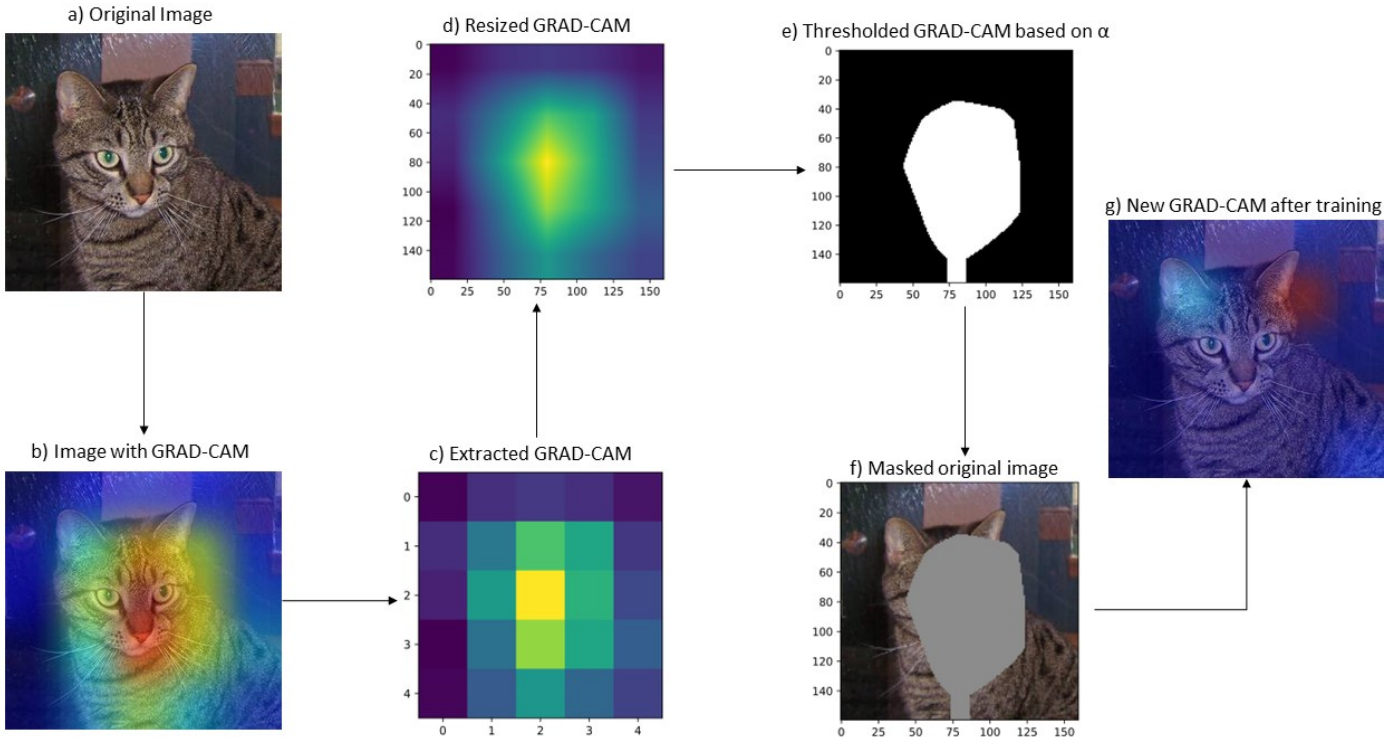

Figure 3.8 DFM procedure. a) Depicts the original image sourced from the dataset. b) Shows the application of Grad-CAM on the original image to extract the heatmap. c) Presents the heatmap derived from Grad-CAM. d) Displays the heatmap resized to match the dimensions of the original image. e) Highlights the binary thresholding with parameter α applied to the resized heatmap. f) Illustrates the process of masking the image using contour coordinates obtained during contour detection on the thresholded heatmap. g) Demonstrates Grad-CAM results after training the model with the masked image

Results

This section details a sequence of experiments conducted to assess the OOD generalization capabilities of our models using DFM relative to other data augmentation techniques. First, we evaluated DFM on the MNIST dataset using a custom model, analyzing its ability to boost model resilience against common data transformations typically not encountered during training. We then implemented the ResNet-50 model for a binary classification task involving Fruits and Vegetables dataset, aiming to determine how both in-distribution and OOD generalization are influenced by various augmentation methods. Following this, the Cats and Dogs dataset provided a platform to further analyze the

comparative effectiveness of DFM, particularly focusing on the model's accuracy in distinguishing between these common household pets under challenging conditions. Lastly, we examined how adjusting the α parameter within the DFM affects the recognition and visibility of features in the Cats and Dogs dataset, which mimics characteristics of opposing classes. Our goal is to determine an optimal α threshold that ensures effective masking; as illustrated in Figure 3.7, overly low α values result in excessive masking of the image, while excessively high values lead to insufficient masking, undermining the efficacy of our method and its generalization in testing scenarios.

Due to the internal nature of RSC, where dominant feature gradients are zeroed out during training, visual illustrations of RSC were not included for the following experiments. The impact of RSC is instead reflected in the performance metrics presented.

MNIST

In the analysis of the MNIST dataset, our experiments focused on the evaluation of DFM, with comparative assessments against previously described augmentation methods and the newly added RSC method as shown in Figure 3.9. Our custom convolutional neural network, when trained with DFM, exhibited notable improvements in resilience, particularly against image transformations that were not encountered during the training phase. According to the results presented in Table 3.1, the DFM-trained model surpassed other augmentation methods in recognizing digits within images subjected to Gaussian noise, horizontal patterned noise, and vertical patterned noise, achieving accuracies of 69.5%, 39.18%, and 35.16%, respectively. When comparing DFM to RSC, the DFM-trained model showed superior performance in these noise conditions. However, RSC also demonstrated significant resilience, with accuracies of 64.54%, 33.42%, and 16.5% for Gaussian noise, horizontal patterned noise, and vertical patterned noise, respectively. Moreover, DFM was outperformed only by RICAP in the cropped evaluation, attaining an accuracy of 85% against RICAP's 87%, and ranked second to SaliencyMix with an accuracy of 22.19% versus SaliencyMix's 24% in the rotated evaluation. RSC achieved an accuracy of

83.46% in the cropped evaluation and 18.34% in the rotated evaluation. Across in-distribution testing, all methods showed comparable performance. Hence, data augmentation did not interfere with in-distribution generalization.

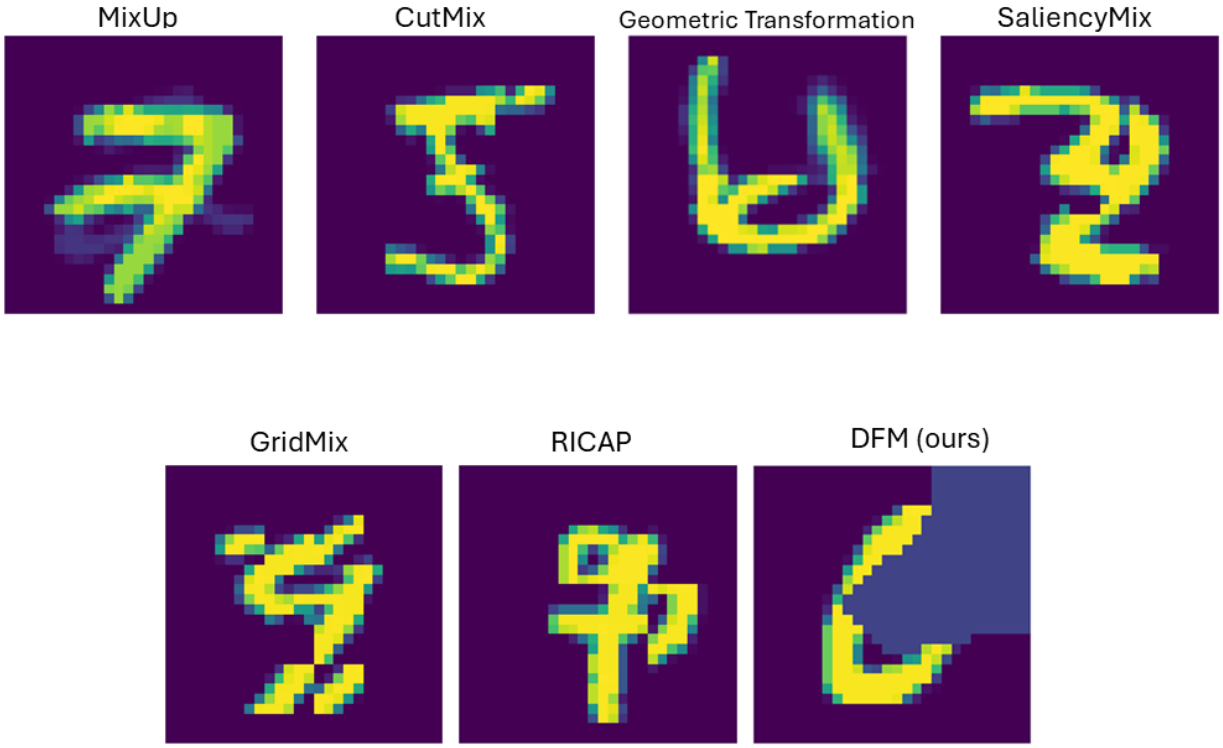

Figure 3.9 A set of MNIST digits processed through different data augmentation techniques, including MixUp, CutMix, Geometric Transformation, SaliencyMix, GridMix, RICAP, and DFM

Table 3.1 Evaluation of various data augmentation strategies like Pure, DFM, SaliencyMix, MixUp, CutMix, GridMix, RICAP, and Geometric Transformation on the MNIST dataset, showcasing their impact on model accuracy and loss amidst different image distortions, including Cropped, Rotated, Gaussian Noise, and Horizontal and Vertical Patterned Noise, along with in-distribution testing performance.

Method	Metric	Cropped	Rotated	Gaussian Noise	Horizontal Patterned Noise	Vertical Patterned Noise	In-Distribution Testing
Pure	Accuracy (%)	73.45	14.74	43.12	14.51	23.47	98.14
	Loss	1.1619	6.7618	2.1767	6.068	4.1794	0.1044
DFM	Accuracy (%)	85.13	22.19	**69.5**	**39.18**	**35.16**	99.45
	Loss	0.4512	4.3217	**1.1335**	**2.5341**	**2.1871**	0.0138
RSC	Accuracy (%)	83.46	18.34	64.54	33.42	16.5	99.42
	Loss	2.0654	14.56	2.452	11.7985	11.7985	0.094
SaliencyMix	Accuracy (%)	77.5	**24**	62.12	34.14	30.5	99.28
	Loss	0.9464	**3.2952**	1.7147	2.6705	2.6309	0.0139
MixUp	Accuracy (%)	80.49	13.24	46.49	12.38	25.85	99.01
	Loss	0.7419	7.8147	2.7438	5.1424	4.7813	0.0274
CutMix	Accuracy (%)	77.71	18.83	44.27	14.84	7.98	98.94
	Loss	0.8437	5.2457	2.8914	4.3736	11.4579	0.0564
GridMix	Accuracy (%)	80.14	20.63	56.15	17.96	17.46	98.96
	Loss	0.7147	4.8741	1.8473	5.5873	7.6265	0.0709
RICAP	Accuracy (%)	**87.01**	18.47	40.93	11.67	16.28	**99.54**
	Loss	**0.4378**	5.7319	3.0014	6.7509	6.4723	**0.0101**
Geometric Transform	Accuracy (%)	75.61	19.79	34.17	18.17	33.47	98.35
	Loss	1.0147	4.9871	4.1579	4.079	2.309	0.0614

Best results are highlighted in bold.

Fruits and Veggies

In the assessment of the Fruits and Veggies dataset, we employed the ResNet-50 model, enhanced with DFM, RSC, and other augmentation methods as depicted in Figure 3.10. Table 3.2 reveals that DFM allowed the ResNet-50 model to achieve an accuracy of 74.45% on previously unseen images of fruits and vegetables. SaliencyMix followed closely, attaining an accuracy of 72.14%, while RICAP ranked third with an accuracy of 70.04%. RSC achieved an accuracy of 73.34%, placing it second in this evaluation.

This evaluation highlights the proficiency of DFM in boosting the generalization capabilities of a sophisticated model like ResNet-50 when it encounters new and diverse data. It is worth noting that for in-distribution testing, the performance remained consistently high and comparable across all augmentation techniques.

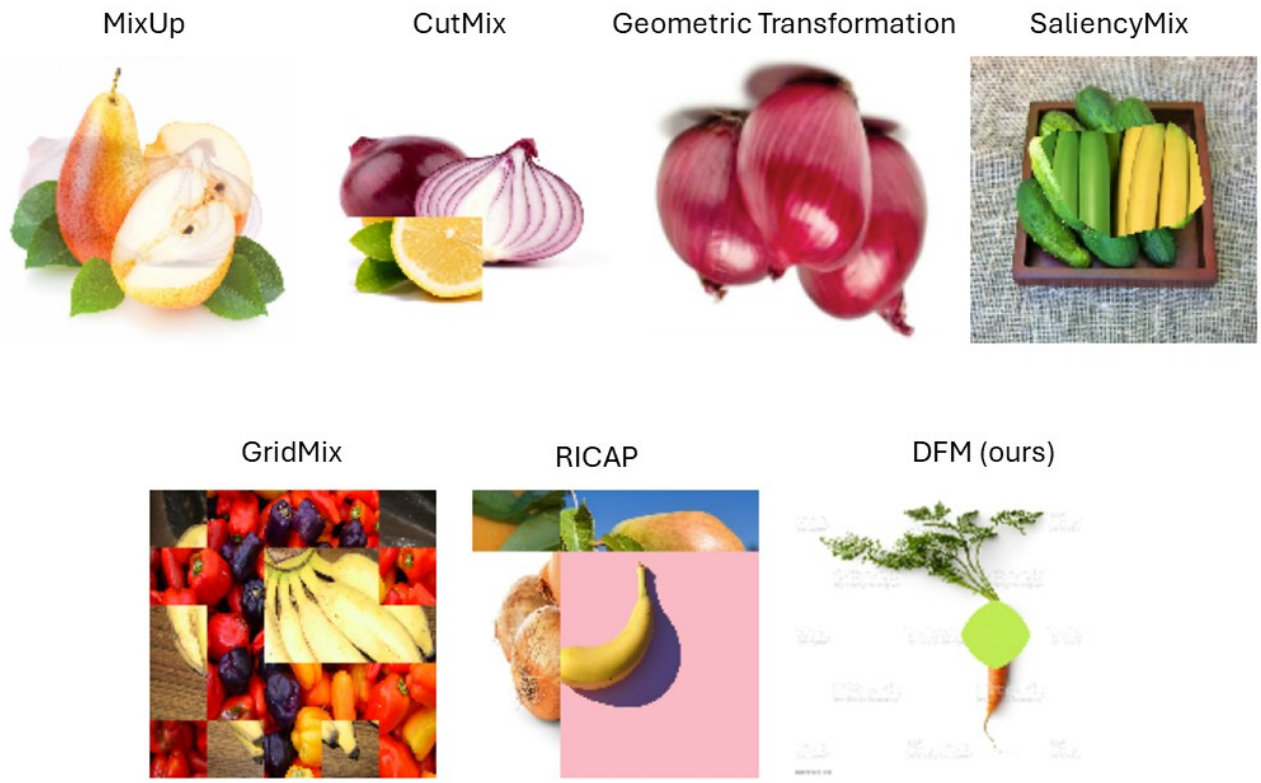

Figure 3.10 The application of various data augmentation techniques (MixUp, CutMix, Geometric Transformation, SaliencyMix, GridMix, RICAP, and DFM) on the Fruits and Veggies dataset

Table 3.2 The comparison of data augmentation techniques tested on the Fruits and Veggies dataset, reporting accuracy and loss for in-distribution and OOD data. Methods compared include Pure, DFM, SaliencyMix, MixUp, CutMix, GridMix, RICAP, and Geometric Transformation.

Method	Metric	In-Distribution	Out of Distribution
Pure	Accuracy (%)	97.48	63.68
	Loss	0.3959	1.8943
DFM	Accuracy (%)	98.76	**74.45**
	Loss	0.2014	**0.8819**
RSC	Accuracy (%)	98.14	73.34
	Loss	0.2149	0.887
SaliencyMix	Accuracy (%)	**98.87**	72.14
	Loss	**0.1975**	1.2467
MixUp	Accuracy (%)	97.57	68.19
	Loss	0.4512	3.1687
CutMix	Accuracy (%)	97.61	69.86
	Loss	0.2179	1.677
GridMix	Accuracy (%)	97.72	66.96
	Loss	0.8621	1.3573
RICAP	Accuracy (%)	97.86	70.04
	Loss	0.502	1.4216
Geometric Transform	Accuracy (%)	97.54	65.47
	Loss	0.3179	1.7841

Best results are highlighted in bold.

Cats and Dogs

Next, the ResNet-50 model was applied to the Cats and Dogs dataset. ResNet-50 was enhanced with various augmentation techniques, including DFM, as shown in Figure 3.11. This OOD task presented significant challenges: it involved discriminating between cats wearing dog masks and dogs wearing cat masks. For this task, DFM obtained an accuracy of 54% with a loss of 1.96. RSC achieved an accuracy of 50% with a loss of 2.457, closely followed by SaliencyMix, which also attained an accuracy of 50% but with a loss of 2.104, as detailed in Table 3.3. RICAP achieved an accuracy of 44% with a loss of 2.4328. In contrast, ResNet-50 trained without any augmentation technique managed an accuracy of just 38% and suffered a substantially higher loss of 12.72, highlighting the complexity of this OOD task and the effectiveness of augmentation methods in confronting such challenges. Yet, it is important to note that for in-distribution testing samples, performance remained consistently high across all augmentation conditions.

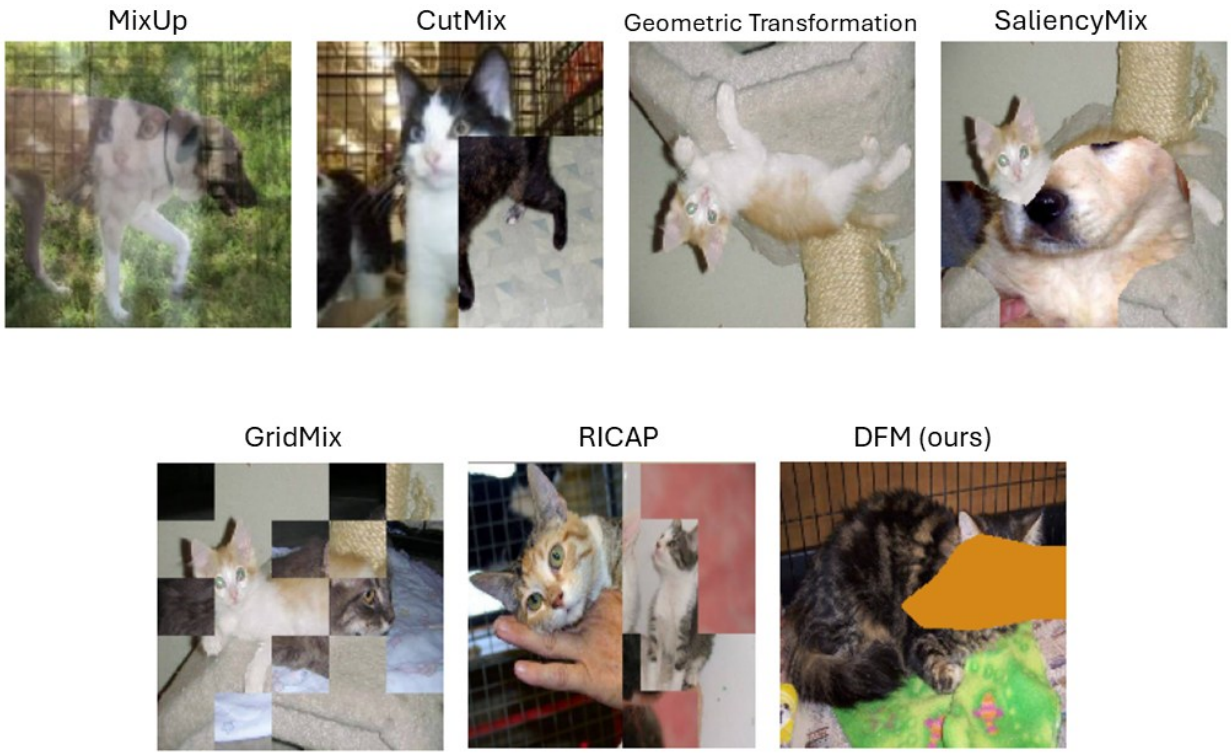

Figure 3.11 Various data augmentation techniques applied to the Cats and Dogs dataset, with visual examples from MixUp, CutMix, Geometric Transformations, SaliencyMix, GridMix, RICAP, and DFM methods

Table 3.3 The accuracy and loss metrics for several data augmentation methods on the Cats and Dogs dataset, showing performance for both in-distribution and OOD samples. The methods assessed include Pure, DFM, SaliencyMix, MixUp, CutMix, GridMix, RICAP, and Geometric Transformation.

Method	Metric	In-Distribution	Out of Distribution
Pure	Accuracy (%)	96.94	38
	Loss	0.2542	12.7237
DFM	Accuracy (%)	**97.13**	**54**
	Loss	**0.2244**	**1.9635**
RSC	Accuracy (%)	97.05	50
	Loss	0.2147	2.457
SaliencyMix	Accuracy (%)	97.07	50
	Loss	0.2246	2.104
MixUp	Accuracy (%)	95.23	36
	Loss	0.5886	4.453
CutMix	Accuracy (%)	96.74	38
	Loss	0.3201	3.2698
GridMix	Accuracy (%)	96.64	44
	Loss	0.2585	2.5634
RICAP	Accuracy (%)	97.01	44
	Loss	0.2297	2.4328
Geometric Transform	Accuracy (%)	96.48	38.61
	Loss	0.3542	5.7621

Best results are highlighted in bold.

Exploring the α parameter

In examining the impact of the α parameter on the efficacy of the DFM method, the ResNet-50 model was trained across different intensities of α values and tested on the OOD task from the Cats and Dogs dataset. The results, presented in Figure 3.12, indicate that an α intensity in the range of 0.55 to 0.75 yielded the highest generalization accuracy for OOD samples, with accuracies approximating 54-56% and a loss between 1.95-2.30. Conversely, excessively high or low α intensities were detrimental to the model's performance. Specifically, an α intensity of 1, which equates to minimal spatial extension of masking, mirrored the outcomes of training the model without augmentations, resulting in a mere 38% accuracy and a loss around 9.80. Conversely, an excessively low α intensity compromised not only OOD generalization—plummeting to 24% accuracy with a loss of 10.25—but also in-distribution testing performance, leading to an accuracy of 80% and loss of 2.24. An α intensity of 0.0, as showcased in Figure 3.7, would effectively mask the entire image, essentially omitting data from the dataset, thereby excluding the opportunity for model learning.

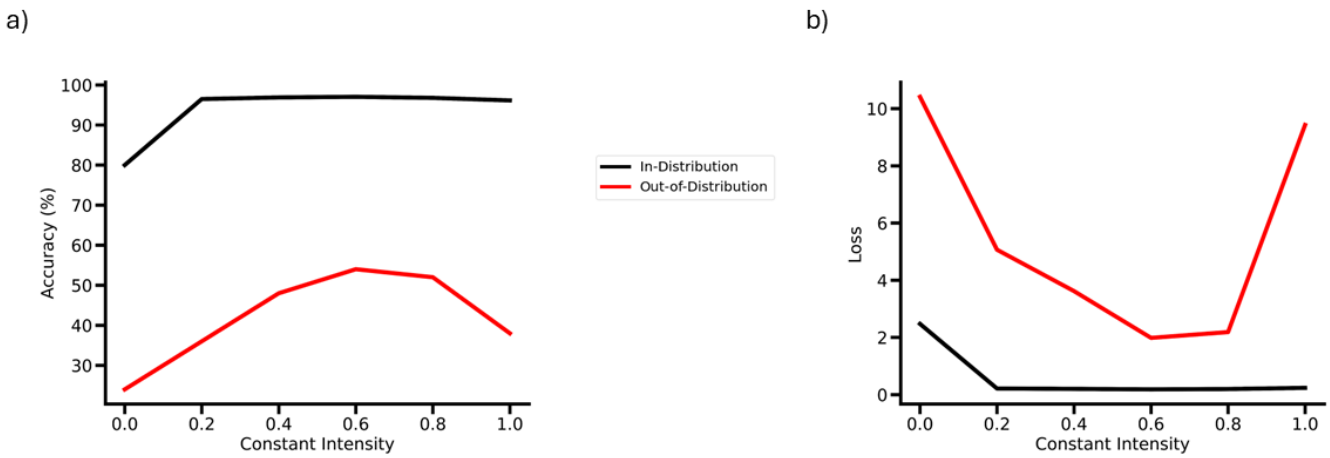

Figure 3.12 a) The accuracy graph showcases the performance of a model across varying intensities of α, both for in-distribution and OOD data. b) The loss graph contrasts the model's loss for in-distribution and OOD tests at different intensities of α

Discussion

The sensitivity of CNNs to OOD exemplars remains an unresolved problem (Geirhos et al., 2020). Our work addresses this issue by introducing a novel method of DFM training, where we control the intensity of masked images given as input to a CNN. Our findings offer insights into the efficacy of the DFM method for enhancing the OOD generalization capabilities of CNNs across three diverse datasets. Specifically, our results revealed that the DFM method not only allows for the customization of feature learning within the network but also achieves this without negatively impacting the training and testing accuracy on the in-distribution dataset, while outperforming traditional augmentation methods in challenging conditions. By fine-tuning the α parameter used for masking, we uncovered a range of optimal values that significantly improved the network's accuracy on OOD testing.

Additionally, we introduced new benchmarks for data generalization, which provide a comprehensive framework for evaluating the robustness of CNNs to various OOD conditions. These benchmarks offer a more rigorous assessment of model performance, ensuring that the improvements brought by DFM and other augmentation methods are thoroughly validated.

MNIST

The results from the MNIST dataset experiment offered solid evidence of the DFM method's efficacy, particularly in enhancing model resilience against common yet challenging image transformations. When compared with traditional augmentation techniques, DFM's performance in recognizing MNIST digits amid Gaussian noise and various types of patterned noise not only demonstrates its robustness but also highlights its ability to maintain high accuracy under adversarial conditions. This is indicative of DFM’s unique approach to encouraging a neural network to generalize beyond the most salient features commonly relied upon during training.

The ability of DFM to outperform other methods under these testing conditions suggests that it encouraged the model to utilize a broader range of features in the decision-making process. This is

particularly valuable in practical applications where models must perform reliably across a range of input variations that were not explicitly covered during training. The marked improvement in model accuracy with DFM under conditions of Gaussian and patterned noises points to its potential in applications where data can often be imperfect or noisy, such as in medical imaging or real-time video processing (Ghazal et al., 2007; Kumar et Nachamai, 2017).

Furthermore, the somewhat reduced performance of DFM compared to methods like RICAP and SaliencyMix in some tests highlighted the significance of selecting the right augmentation techniques for specific contexts. While DFM exceled in enhancing noise resilience, its comparative underperformance in certain scenarios suggests a need for further tuning of the α parameter or possibly combining DFM with other techniques to optimize performance across all types of distortions and noises.

Fruits and Veggies

The evaluation of the DFM technique on the Fruits and Veggies dataset reinforced its effectiveness in generalizing to new, unseen categories of data. DFM established a notable performance, achieving the highest accuracy among the augmentation techniques tested. This success demonstrated that DFM not only enhanced model performance across different types of image content but also showed its capability to handle complex and naturally varied subjects, such as the organic shapes found in fruits and vegetables. Unlike alphanumeric characters that tend to have defined and consistent forms, organic shapes vary widely in their appearance, making them more challenging to classify. DFM's ability to improve model accuracy in this context highlights its adaptability to a broad range of visual data beyond simple, structured elements.

The close competition between DFM and RSC, with the former slightly outperforming the latter, illustrated the benefits of DFM in handling datasets that contain intricate and subtle variations between classes. The ability of DFM to achieve 74.45% accuracy on this dataset—a significant improvement over traditional methods—underscored its potential in enhancing feature recognition in more complex and

less structured visual contexts. While RSC and DFM both aim to improve generalization by masking dominant features, DFM offers the added flexibility of controlling the masking intensity through the α parameter. This flexibility allows for fine-tuning that can push generalization slightly further than RSC. This is particularly crucial for tasks where distinguishing between categories depends heavily on subtle variations, which are common in natural images.

However, the narrow margin by which DFM leads SaliencyMix and RICAP also calls attention to the challenges of optimizing augmentation techniques for diverse types of data. Each method brings specific strengths that may be more or less suited to particular datasets or classification challenges. The results thus advocate for a strategic selection of augmentation methods based on the specific characteristics of the data and the desired outcome of the model. Future studies could explore a hybrid approach that combines the strengths of DFM with those of other methods to maximize performance across a broader array of tasks and conditions. This could be especially effective in domains such as agricultural technology or dietary monitoring, where precise recognition of diverse natural products is critical (Hassannejad et al., 2017; Patrício & Rieder, 2018).

Cats and Dogs

The analysis of the Cats and Dogs dataset with the DFM technique and other augmentation methods highlighted the considerable challenges posed by OOD scenarios, especially when the model is required to distinguish between visually similar categories under complex conditions. The OOD task—discriminating between cats wearing dog masks and dogs wearing cat masks—proved extremely difficult, yet DFM emerged as a strong performer, achieving the highest accuracy among the tested methods.

DFM's success in this challenging task, with an accuracy of 54% and a loss significantly lower than those achieved without any augmentation, emphasized its effectiveness in enhancing model robustness against highly confusing inputs. This performance was particularly notable given the complexity of the

task, where conventional features used for class discrimination are deliberately obscured, forcing the model to rely on subtler cues that might otherwise be overlooked.

However, the performance across all methods was not as high as seen in other datasets, highlighting the inherent difficulty of the task and the limitations of current augmentation techniques in extremely nuanced OOD conditions. The fact that DFM still outperformed other methods suggests its potential utility in applications where models must operate in unpredictable environments and make decisions based on incomplete or misleading information.

Given these results, future enhancements to DFM could focus on refining the balance between masking intensity and feature preservation to further improve its efficacy. Exploring combinations of DFM with techniques like SaliencyMix or RICAP might also yield better outcomes, leveraging the strengths of multiple approaches to handle complex visual categorization tasks more effectively. Such advancements could be particularly beneficial in fields like security and surveillance, where distinguishing between objects or entities that are deliberately disguised is crucial (Idrees et al., 2018; Sage & Young, 1998).

The α parameter

In our exploration of the α parameter within the DFM framework on the Cats and Dogs dataset, we found that fine-tuning the intensity of this parameter was crucial for achieving optimal model performance. The investigation, as illustrated in Figure 3.12, revealed that an α intensity between 0.55 and 0.75 yielded the best generalization accuracy for OOD samples, with accuracies ranging from 54% to 56% and loss rates between 1.95 and 2.30. This range represented a balanced approach where sufficient feature visibility is maintained while still enhancing the model's ability to generalize from obscured or altered input data.

However, our study also highlighted the sensitivity of model performance to variations in α intensity. When the α value was set too high, nearing 1, the masking effect was minimal, rendering the augmentation almost ineffective, akin to training without any augmentation, which resulted in

significantly lower accuracy and higher loss. Conversely, setting α too low resulted not only in reduced OOD generalization, evidenced by a mere 24% accuracy and a loss of 10.25, but also adversely affected the in-distribution testing, with the accuracy dropping to 80% and an increased loss of 2.24. This underscored the importance of carefully calibrating the α parameter to avoid over-masking, which can eliminate crucial data needed for effective learning.

A comparison between DFM and RSC in our experiments underscores the importance of such parameter tuning. While RSC also masks dominant features to enhance generalization, it lacks the fine-tuning capability provided by the α parameter in DFM. This flexibility allows DFM to achieve a more optimal balance in masking intensity, which can push generalization performance further than RSC. Both methods share the goal of improving the model's ability to handle OOD samples by forcing it to rely on a broader set of features, but the flexible nature of DFM provides a distinct advantage in fine-tuning the model's response to different types of data variations.

This parameter adjustment highlights the delicate balance required in applying DFM effectively: too much masking can obscure necessary features for classification, while too little does not sufficiently challenge the model to learn secondary features. Similarly, adjusting hyperparameters is a common practice in other data augmentation techniques to tailor their effects appropriately (Baek et al., 2021; Takahashi et al., 2019; Yun et al., 2021; Zhang et al., 2017). Therefore, future work should consider adaptive mechanisms that could dynamically adjust α based on the model's learning stage or feedback from its performance, potentially leading to even more refined and effective use of the DFM technique in complex image classification tasks.

Effect of DFM on different types of CNNs

Our evaluation of DFM across various neural network architectures demonstrated its adaptability and effectiveness, from simpler custom models to complex systems like ResNet-50. The application of DFM consistently enhanced model performance, ensuring robust generalization across OOD scenarios.

This broad applicability signified that DFM's approach to selectively masking features to promote non-dominant feature learning is effective regardless of the architectural complexity. Thus, DFM is a versatile tool that can be integrated into diverse neural network designs, enhancing their ability to handle real-world variability in data.

VEB

Another interesting takeaway from this study is that relying solely on training and testing accuracy as performance metrics provided a poor estimate of a model's ability to generalize to OOD samples. This is a crucial consideration, suggesting that the current evaluation metrics might not be adequate for assessing a CNN's adaptability (Mishra et al., 2020), especially since training and testing datasets have similar in-distribution features. Considering this, we introduced the VEB to provide a more comprehensive evaluation of a model's generalization capabilities.

The VEB helps to understand generalization performance by exposing models to a range of OOD scenarios, thereby highlighting their adaptability and robustness. As observed in our experiments, the accuracy on in-distribution testing is consistently close to 99%, demonstrating very high performance. However, not a single challenge was able to pass 85% accuracy except for the cropping task in the MNIST dataset, where RICAP achieved high accuracy. This result can be attributed to the nature of the RICAP augmentation technique, which closely resembles the cropping task.

The VEB evaluates model performance on three distinct tasks that are crucial for real-life deployment of CNNs: resilience to transformations, ability to handle intra-class variability, and resilience of the network against adversarial attacks. These tasks encompass a wide range of potential real-world variations, ensuring that the evaluation framework thoroughly tests the model's ability to handle unforeseen scenarios. The high accuracy on in-distribution datasets provides an overestimation of the network's performance, leading researchers to overestimate the extent to which their models may be robust and universally applicable.

Our work highlights the need for diverse benchmarks that evaluate performance on familiar data while also challenging models with unforeseen scenarios. Incorporating OOD samples into regular evaluation procedures is a proactive step in fostering models that are more robust and less prone to failure in the presence of unexpected data. The VEB framework serves as a standardized benchmark for this purpose, promoting the development of more reliable and adaptable models. Furthermore, our study encourages the creation of more diverse and challenging OOD datasets to further advance the field of machine learning.

By leveraging the VEB, researchers can gain a deeper understanding of their models' generalization performance, ensuring that CNNs are well-equipped to handle the complexities of real-world applications. This comprehensive approach to evaluation is essential for developing robust and effective machine learning models that can be confidently deployed in diverse and dynamic environments.

DFM and human holistic processing

Our research, inspired by the holistic processing techniques used by medical experts as discussed by Sheridan and Reingold (2017), demonstrates how such approaches allow for rapid and accurate integration of diverse visual information, enabling the detection of abnormalities even when they are not the primary focus of observation. Our implementation of DFM encourages neural networks to extend their focus beyond the most immediately noticeable features, enhancing the network's ability to identify crucial, yet less dominant features swiftly, thus improving accuracy in OOD scenarios. DFM aligns with principles of human holistic processing by allowing models to utilize global visual information effectively, mirroring human capacity to integrate broad visual data (Richler et al., 2012), thus enhancing the depth and breadth of network’s analytical capabilities. Like medical experts who adeptly interpret complex images, DFM enhances the generalization abilities of computer vision systems by training networks to recognize non-dominant features, thereby improving the model’s functionality in varied environments.

Limitations of DFM

While DFM has shown promise in enhancing the generalization and robustness of machine learning models, several limitations need to be addressed. The effectiveness of DFM depends on the precise calibration of the α parameter. If α is set too high, the model may not mask enough, failing to compel it to learn from non-dominant features. Conversely, if α is set too low, it can lead to excessive masking, obscuring vital information necessary for accurate classification. It is reasonable to maintain α within the range of 0.4 to 0.8 to avoid these extremes, as adjustments outside this range tend to produce these negative effects, as demonstrated in Figure 3.12. This sensitivity to α requires accurate tuning and could limit the rapid deployment or scalability of DFM in dynamic environments.

Moreover, DFM's current reliance on Grad-CAM for identifying salient regions assumes accurate recognition based on distinct features of each class. However, there's a potential risk that the model might recognize a class due to the absence of the features of the other class, rather than through the presence of its own defining characteristics. This issue highlights the need for exploring more advanced or alternative saliency detection methods that ensure the most influential features are being masked, which could further improve the effectiveness of DFM (Sundararajan et al., 2017; Zhao et al., 2015; Zhou et al., 2016).

Additionally, integrating DFM into existing architectures and training workflows can be complex and may require modifications to accommodate the specific demands of feature detection and masking. The computational overhead associated with these processes can also escalate resource requirements, making DFM less practical for large-scale or real-time applications.

To overcome these challenges, future research should focus on developing adaptive and computationally efficient masking strategies, possibly incorporating hybrid approaches that blend DFM with other augmentation or regularization techniques. This would enhance both the practicality and performance of DFM, making it more versatile and broadly applicable across a wider range of tasks and

deployment scenarios. Such advancements are crucial for realizing the full potential of DFM in making neural networks more robust and reliable in their decision-making processes.

Conclusion

In conclusion, the series of experiments conducted using DFM across various datasets and scenarios have provided valuable insights into its capability to enhance model generalization, particularly in OOD conditions. DFM has proven effective across a diverse set of challenges, from digit recognition in the MNIST dataset to complex image classifications involving fruits, vegetables, and animals in disguise. These findings emphasized DFM's potential as a robust augmentation technique that not only preserved but enhanced the model's ability to interpret and classify complex visual data under varied conditions.

However, the dependence on precise parameter tuning, particularly the α parameter, highlights the need for careful implementation and possibly automated adaptive mechanisms to maximize the efficacy of DFM. Future research should aim to refine these aspects, potentially integrating DFM with other computational techniques to create more robust and adaptable models. By continuing to explore and enhance DFM and similar methodologies, we can push the boundaries of what's possible in machine learning, making systems that are not only more accurate but also more capable of handling the unexpected variability of real-world data.

The introduction of the VEB provided a comprehensive framework to assess model performance on OOD tasks, highlighting the importance of evaluating resilience to transformations, intra-class variability, and adversarial attacks. The VEB framework enabled a thorough evaluation of DFM and other augmentation techniques, ensuring that the improvements in generalization capabilities were robust and reliable. By leveraging the VEB, researchers can gain a deeper understanding of their models' generalization performance, ensuring that CNNs are well-equipped to handle the complexities of real-world applications. This comprehensive approach to evaluation is essential for developing robust and

effective machine learning models that can be confidently deployed in diverse and dynamic environments.

Statements and Declarations

Funding

This work was supported by a Discovery grant to J.P.T. from the Natural Sciences and Engineering Council of Canada (NSERC Grant No. 210977).

Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this chapter.

Acknowledgements

We thank members of the ComBiNe Laboratory at the University of Ottawa for useful discussions about this work.

References

Al-Sabaawi, A., Ibrahim, H. M., Arkah, Z. M., Al-Amidie, M., & Alzubaidi, L. (2020, December). Amended convolutional neural network with global average pooling for image classification. In *International Conference on Intelligent Systems Design and Applications* (pp. 171-180). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-71187-0_16

Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., ... & Lacoste-Julien, S. (2017, July). A closer look at memorization in deep networks. In *Proceedings of the International Conference on Machine Learning* (pp. 233-242). PMLR.

Azulay, A., & Weiss, Y. (2019). Why do deep convolutional networks generalize so poorly to small image transformations?. *Journal of Machine Learning Research, 20*(184), 1-25.

Baek, K., Bang, D., & Shim, H. (2021). GridMix: Strong regularization through local context mapping. *Pattern Recognition*, *109*, 107594. https://doi.org/10.1016/j.patcog.2020.107594

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798-1828. doi: 10.1109/TPAMI.2013.50.

Chen, J., Ding, L., Yang, Y., Di, Z., & Xiang, Y. (2024). Domain adversarial active learning for domain generalization classification. *arXiv preprint* arXiv:2403.06174.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, *29*(6), 141-142. doi: 10.1109/MSP.2012.2211477.

DeVries, T., & Taylor, G. W. (2018). Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*. https://doi.org/10.48550/arXiv.1802.04865

Feng, X., Jiang, Y., Yang, X., Du, M., & Li, X. (2019). Computer vision algorithms and hardware implementations: A survey. *Integration*, *69*, 309-320. https://doi.org/10.1016/j.vlsi.2019.07.005

Fort, S., Ren, J., & Lakshminarayanan, B. (2021). Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, *34*, 7068-7081.

Gao, J., Yang, Y., Lin, P., & Park, D. S. (2018). Computer vision in healthcare applications. *Journal of healthcare engineering*, *2018*. https://doi.org/10.1155/2018/5157020

Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665-673. https://doi.org/10.1038/s42256-020-00257-z

Ghazal, M., Amer, A., & Ghrayeb, A. (2007). A real-time technique for spatio–temporal video noise estimation. *IEEE transactions on Circuits and Systems for Video Technology*, *17*(12), 1690-1699. doi: 10.1109/TCSVT.2007.903805.

Ghifary, M., Kleijn, W., Zhang, M., & Balduzzi, D. (2015). Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 2551-2559). IEEE Computer Society. https://doi.org/10.1109/ICCV.2015.293

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint* arXiv:1412.6572.

Hassannejad, H., Matrella, G., Ciampolini, P., De Munari, I., Mordonini, M., & Cagnoni, S. (2017). Automatic diet monitoring: a review of computer vision and wearable sensor-based methods. *International journal of food sciences and nutrition*, *68*(6), 656-670. https://doi.org/10.1080/09637486.2017.1283683

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., ... & Gilmer, J. (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8340-8349).

Huang, Z., Wang, H., Xing, E. P., & Huang, D. (2020). Self-challenging improves cross-domain generalization. In *Proceedings of the European Conference on Computer Vision (ECCV) 2020* (Vol. 12460, pp. 124-140). Springer. https://doi.org/10.1007/978-3-030-58592-1_8

Idrees, H., Shah, M., & Surette, R. (2018). Enhancing camera surveillance using computer vision: a research note. *Policing: An International Journal*, *41*(2), 292-307.

Janai, J., Güney, F., Behl, A., & Geiger, A. (2020). Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, *12*(1–3), 1-308. http://dx.doi.org/10.1561/0600000079

Khan, A. A., Laghari, A. A., & Awan, S. A. (2021). Machine learning in computer vision: a review. *EAI Endorsed Transactions on Scalable Information Systems*, *8*(32), e4-e4. https://doi.org/10.4108/eai.21-4-2021.169418

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koonce, B., & Koonce, B. (2021). ResNet 50. *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, 63-72.

Kumar, N., & Nachamai, M. (2017). Noise removal and filtering techniques used in medical images. *Orient. J. Comput. Sci. Technol*, *10*(1), 103-113. http://dx.doi.org/10.13005/ojcst/10.01.14

Kurakin, A., Goodfellow, I. J., & Bengio, S. (2018). Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security* (pp. 99-112). Chapman and Hall/CRC.

Liang, S., Li, Y., & Srikant, R. (2017). Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*. https://doi.org/10.48550/arXiv.1706.02690

Liu, Z., Xu, Y., Xu, Y., Qian, Q., Li, H., Jin, R., ... & Chan, A. B. (2022). An empirical study on distribution shift robustness from the perspective of pre-training and data augmentation. *arXiv preprint arXiv:2205.12753*. https://doi.org/10.48550/arXiv.2205.12753

Lu, L., Shin, Y., Su, Y., & Karniadakis, G. E. (2019). Dying relu and initialization: Theory and numerical examples. *arXiv preprint arXiv:1903.06733*. https://doi.org/10.4208/cicp.OA-2020-0165

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, *3*(1), 91-99. https://doi.org/10.1016/j.gltp.2022.04.020

Mishra, S., Arunkumar, A., Bryan, C., & Baral, C. (2020). Our evaluation metric needs an update to encourage generalization. *arXiv preprint arXiv:2007.06898*. https://doi.org/10.48550/arXiv.2007.06898

Mumuni, A., & Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array*, *16*, 100258. https://doi.org/10.1016/j.array.2022.100258

Niu, H., Li, H., Zhao, F., & Li, B. (2022). Domain-unified prompt representations for source-free domain generalization. *arXiv preprint* arXiv:2209.14926.

Panati, C., Wagner, S., & Brüggenwirth, S. (2022, September). Feature relevance evaluation using grad-CAM, LIME and SHAP for deep learning SAR data classification. In *2022 23rd International Radar Symposium (IRS)* (pp. 457-462). IEEE. doi: 10.23919/IRS54158.2022.9904989.

Patrício, D. I., & Rieder, R. (2018). Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. *Computers and electronics in agriculture*, *153*, 69-81. https://doi.org/10.1016/j.compag.2018.08.001

Ratner, A. J., Ehrenberg, H., Hussain, Z., Dunnmon, J., & Ré, C. (2017). Learning to compose domain-specific transformations for data augmentation. *Advances in neural information processing systems*, *30*.

Richler, J. J., Palmeri, T. J., & Gauthier, I. (2012). Meanings, mechanisms, and measures of holistic processing. *Frontiers in psychology*, *3*, 553. https://doi.org/10.3389/fpsyg.2012.00553

Sage, K., & Young, S. (1998, October). Computer vision for security applications. In *Proceedings IEEE 32nd Annual 1998 International Carnahan Conference on Security Technology (Cat. No. 98CH36209)* (pp. 210-215). IEEE. doi: 10.1109/CCST.1998.723792.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Batra, D. (2016). Grad-CAM: Why did you say that?. *arXiv preprint arXiv:1611.07450*. https://doi.org/10.48550/arXiv.1611.07450

Seth, K. (2020). Fruits and vegetables image recognition dataset [Data set]. Kaggle**.** https://www.kaggle.com/kritikseth/fruit-and-vegetable-image-recognition

Sheridan, H., & Reingold, E. M. (2017). The holistic processing account of visual expertise in medical image perception: A review. *Frontiers in psychology*, *8*, 252697. https://doi.org/10.3389/fpsyg.2017.01620

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, *6*(1), 1-48. https://doi.org/10.1186/s40537-019-0197-0

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*. https://doi.org/10.48550/arXiv.1312.6034

Storcheus, D., Rostamizadeh, A., & Kumar, S. (2015, December). A survey of modern questions and challenges in feature extraction. In *Feature Extraction: Modern Questions and Challenges* (pp. 1-18). PMLR.

Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In *International conference on machine learning* (pp. 3319-3328). PMLR.

Takahashi, R., Matsubara, T., & Uehara, K. (2019). Data augmentation using random image cropping and patching for deep CNNs. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(9), 2917-2931. doi: 10.1109/TCSVT.2019.2935128

Tang, K., Tao, M., Qi, J., Liu, Z., & Zhang, H. (2022, October). Invariant feature learning for generalized long-tailed classification. In *European Conference on Computer Vision* (pp. 709-726). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-20053-3_41

Taylor, L., & Nitschke, G. (2018, November). Improving deep learning with generic data augmentation. In *2018 IEEE symposium series on computational intelligence (SSCI)* (pp. 1542-1547). IEEE. doi: 10.1109/SSCI.2018.8628742.

Uddin, A. F. M., Monira, M., Shin, W., Chung, T., & Bae, S. H. (2020). Saliencymix: A saliency guided data augmentation strategy for better regularization. *arXiv preprint arXiv:2006.01791*. https://doi.org/10.48550/arXiv.2006.01791

Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, *2018*. https://doi.org/10.1155/2018/7068349

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., ... & Philip, S. Y. (2022). Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering, 35*(8), 8052-8072.

Wang, Z., Luo, Y., Qiu, R., Huang, Z., & Baktashmotlagh, M. (2021). Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 834-843).

Yan, L., Ye, Y., Wang, C., & Sun, Y. (2024). LocMix: local saliency-based data augmentation for image classification. *Signal, Image and Video Processing*, *18*(2), 1383-1392. https://doi.org/10.1007/s11760-023-02852-0

Ye, Z., Qin, S., Chen, S., & Huang, X. (2021). Dominant patterns: Critical features hidden in deep neural networks. *arXiv preprint arXiv:2105.15057*. https://doi.org/10.48550/arXiv.2105.15057

Yoon, J. S., Oh, K., Shin, Y., Mazurowski, M. A., & Suk, H. I. (2023). Domain generalization for medical image analysis: A survey. *arXiv preprint* arXiv:2310.08598.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6023-6032).

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*. https://doi.org/10.48550/arXiv.1710.09412

Zhao, R., Ouyang, W., Li, H., & Wang, X. (2015). Saliency detection by multi-context deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1265-1274).

Zheng, G., Huai, M., & Zhang, A. (2024, March). AdvST: Revisiting data augmentations for single domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 19, pp. 21832-21840).

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921-2929).

Zhou, K., Yang, Y., Qiao, Y., & Xiang, T. (2021). Domain generalization with mixstyle. arXiv preprint arXiv:2104.02008.

Chapter 4

Improving Generalization in Convolutional Neural Networks with a Dynamic Attention Layer

Artem Pilzak[1], Jean-Philippe Thivierge[1,2]

[1]School of Psychology, University of Ottawa, K1N 6N5, Ottawa, Canada
[2]Brain and Mind Research Institute, University of Ottawa, K1N 6N5, Ottawa, Canada

Abstract

This chapter addresses the challenge of out-of-distribution (OOD) learning in computer vision by introducing a novel Dynamic Attention Layer (DAL), designed to enhance Convolutional Neural Networks. DAL dynamically interchanges attention weights, based on selected percentiles during training, improving the network's ability to capture both dominant and secondary features of the data. We assessed DAL in three scenarios: (1) augmented MNIST images to evaluate resilience to transformations; (2)a novel dataset of unseen image classes to test performance on new instances; and (3) a DALL-E generated dataset to challenge class differentiation with mixed features. Our results show that DAL enhances accuracy and reduces loss in both in-distribution and OOD scenarios. The optimal configuration of parameters for DAL was found to maximize model performance, demonstrating its potential to address OOD learning challenges and contribute to more robust computer vision systems.

Keywords—attention layer, convolutional neural networks, out-of-distribution learning, feature acquisition, deep learning

Introduction

In recent years, computer vision has reached several significant milestones, empowering machines to recognize objects, detect patterns, and derive insights from images and videos (Feng, Jiang, Yang, Du, & Li, 2019; Khan, Laghari, & Awan, 2021; Voulodimos, Doulamis, Doulamis, & Protopapadakis, 2018). This technology's applications are vast, ranging from autonomous vehicles that can navigate complex road environments (Janai, Güney, Behl, & Geiger, 2020) to enhanced medical diagnoses through advanced image analysis (Gao, Yang, Lin, & Park, 2018). These advancements are revolutionizing various industries and unlocking new possibilities for applications.

Nevertheless, a significant challenge in computer vision is out-of-distribution (OOD) learning, where a network must generalize to entirely new and unseen testing samples (Hendrycks et al., 2021). As the scope of computer vision expands, the likelihood of encountering novel data that significantly deviates from the original training set increases (DeVries & Taylor, 2018; Fort, Ren, & Lakshminarayanan, 2021). This issue is exacerbated by the difficulty of obtaining data that encompasses all possible variations, particularly in unique tasks with limited data availability. Such disparities between training and novel data can cause conventional models to produce errors (Geirhos et al., 2020). Addressing this challenge requires the development of innovative algorithms and strategies that enable computer vision systems to detect anomalies, adapt to new environments, and maintain high precision and reliability even with unfamiliar inputs. Achieving robust OOD learning is crucial for the practical viability and effectiveness of computer vision solutions across diverse domains, aiming to match the perceptual and cognitive capabilities of the human visual system.

One effective approach to addressing OOD learning in computer vision is through data augmentation techniques (Liang, Li, & Srikant, 2017; Liu et al., 2022, Shorten & Khoshgoftaar, 2019). By applying transformations such as rotation, scaling, cropping, and color adjustments to existing data, data augmentation artificially increases the diversity of the training set. This helps models generalize better to

novel and unseen samples, improving their robustness and accuracy (Miao & Luo, 2022). Consequently, data augmentation enhances the performance of computer vision systems in OOD scenarios, ensuring high reliability even with unfamiliar inputs. However, there are downsides to this approach. Data augmentation can be time-consuming and may significantly increase the size of the dataset, necessitating more processing power and storage (Goceri, 2023). Moreover, it is not a guarantee that the augmented data will sufficiently cover all potential variations, meaning that the network may still struggle to learn and generalize effectively (Yan, Ye, Wang, & Sun, 2024). Therefore, it is important to explore adjustments to the learning dynamics within the neural network itself.

Attention Mechanisms

Attention mechanisms (Vaswani et al., 2017) allow models to focus on the most relevant parts of the input data, dynamically adjusting their focus based on the context of the task. This enables the network to better capture complex dependencies and patterns within the data, significantly improving performance and generalization. Among various attention mechanisms, channel-wise attention (CWA) (Hu, Shen & Sun, 2018; Wang, Wu, Zhu, Li, Zuo & Hu, 2020) specifically targets the importance of each channel in the feature maps produced by convolutional layers. By reweighting the importance of each channel based on their relevance, CWA helps the network prioritize more informative features while suppressing less relevant ones. This approach not only enhances the model's ability to generalize to new and unseen data but also improves its overall performance and robustness (Meng et al., 2022; Zhang, Ma, Zhao, Zhang, Sun, & Ma, 2022). Integrating CWA into the network architecture allows for a more nuanced understanding of the data, leading to more accurate and reliable predictions.

While attention mechanisms primarily focus on the most important features for decision making, secondary features also play a crucial role in achieving efficient generalization (Hassaballah & Awad, 2016; Selvaraj, Veloso, & Rosenthal, 2018; Zhou, 2000). Secondary features, though less prominent, can include subtle patterns, textures, or contextual details that provide valuable additional information. These

features may not dominate the model's attention during the initial training phase but are essential for capturing the full complexity of the data. By considering these secondary features, models can develop a more comprehensive understanding of the input, which is particularly important when encountering novel or diverse data (Chen, Li, Bai, Yang, Jiang, & Miao, 2021). Enhancing the detection and utilization of secondary features ensures that the model is not overly reliant on dominant characteristics alone, thereby increasing its robustness and adaptability in various scenarios (Koo & Cha, 2017; Shih, 2010). This balanced approach to feature importance can significantly improve the model's ability to generalize, making it more effective in real-world applications where data variability is common.

In this chapter, we propose a novel CWA layer to enhance the accuracy of Convolutional Neural Networks (CNNs) on both in-distribution and OOD datasets. In-distribution generalization refers to testing accuracy on data withheld from the training set that shares common features with it, while OOD datasets consist of entirely new data with unseen features, posing a greater challenge. This new layer, which we call the Dynamic Attention Layer (DAL), operates by dynamically interchanging attention weights based on selected percentiles during the training phase. While attention mechanisms like self-attention in transformers rely on pairwise interactions using queries, keys, and values to capture global dependencies (Vaswani et al., 2017), the DAL is based on CWA mechanisms. DAL focuses on reweighting the importance of individual feature map channels independently, making it distinct from transformer-based approaches. We hypothesize that by redistributing the attention weights, the network can better capture and utilize secondary features that are often overlooked by traditional models (Alzubaidi et al., 2021). This redistribution allows the model to balance the emphasis on both dominant and subtle features, thereby improving its overall generalization capabilities.

The remainder of this work is structured as follows. First, the Related Works section discusses existing research on CWA mechanisms and their applications. The Proposed Methods section then introduces a custom CNN incorporating the DAL and describes the training data used for this research,

followed by a detailed explanation of the DAL itself. In the Results section, we evaluate the performance of the custom CNN with DAL, comparing it to a traditional CWA layer and simple data augmentation techniques like rotation, cropping, and scaling during both in-distribution and OOD testing. Additionally, the Results section includes an examination of the optimal parameter choice for the percentiles swap. The Discussion section contrasts the effectiveness of DAL with the traditional CWA layer and simple data augmentation across three distinct datasets. Finally, we discuss the limitations of DAL and propose potential directions for future research.

Contribution

The contributions of this study can be summarized as follows:

(a) Introduction of a novel modified CWA layer termed the DAL, which dynamically interchanges attention weights based on selected percentiles during the training phase. The selection of percentiles can be done manually, where θ represents the higher bound percentile and τ represents the lower bound percentile.

(b) Evaluation of the effectiveness of DAL and the calculation of a score for classifying in-distribution and OOD samples. We benchmarked DAL against the traditional CWA layer and simple data augmentation using three datasets with unique OOD challenges: (i) augmented MNIST images to assess resilience to common data transformations; (ii) a collection of novel images to test the model's ability to generalize to new examples within familiar categories; and (iii) a custom dataset created using DALL-E, intended to test class differentiation by simulating features of different classes.

Related Works

CWA mechanisms have gained significant attention in recent research due to their potential to enhance the generalization capabilities of neural networks (Meng et al., 2022; Zhang, Ma, Zhao, Zhang,

Sun, & Ma, 2022). These mechanisms focus on adjusting the importance of each channel in the feature maps, enabling the network to prioritize more informative features. Various studies have explored different aspects of CWA and its applications.

One notable approach is the PLACE Dropout technique (Guo, Qi, Shi, & Gao, 2023), which introduces a novel layer-wise and channel-wise dropout mechanism. This technique improves domain generalization by selectively dropping out entire channels during training, thereby forcing the network to learn more robust representations that are less sensitive to the specific channels present during training. The PLACE Dropout has been shown to enhance the model's ability to handle OOD tasks effectively, making it a valuable tool in scenarios where the training data may not comprehensively represent all possible variations in the testing data.

Another significant contribution is the integration of spatial and channel-wise attention mechanisms to improve image captioning performance (Chen et al., 2017; Liu et al., 2022). The spatial attention mechanism focuses on relevant regions of the image, while the channel-wise attention mechanism enhances the network's ability to emphasize important feature channels. This combined approach not only improves the accuracy of image captions but also enhances the generalization capabilities of the network by providing a more detailed and context-aware feature representation. The results underscore the potential of combining different attention mechanisms to create more robust and versatile models.

Despite these advancements, there is a notable gap in research regarding the direct manipulation of attention weights to detect secondary features. Current studies have largely focused on global and primary features (Alzubaidi et al., 2021), leaving the exploration of secondary feature enhancement through direct attention weight manipulation as an open area for further investigation. By directly manipulating the attention weights, it becomes possible to enhance the network’s sensitivity to secondary features, leading to a more balanced and comprehensive feature representation. This approach can

improve the model's ability to generalize to novel and unseen data, ensuring that it is not overly reliant on dominant characteristics alone, which may be corrupted or absent in novel data presented during testing.

Overall, the field of CWA and its applications in enhancing neural network performance is rich with ongoing research. The proposed DAL seeks to build upon these foundations by introducing a novel approach to dynamically interchange attention weights based on selected percentiles. This method aims to fill the gap in current research by focusing on the detection and enhancement of secondary features, providing a new avenue for improving the generalization capabilities of CNNs.

Methods

Custom Convolutional Neural Network

For our custom CNN, we created an architecture using the TensorFlow Keras framework consisting of twenty-one layers (Figure 4.1). All layers utilized Rectified Linear Unit (ReLU) activation functions, with weights initialized using the He Uniform initializer (Lu, Shin, Su, & Karniadakis, 2019). The initial structure included three 2D convolutional layers with 3x3 filters for extracting features from input images. The first convolutional layer had 32 filters and maintained the input's spatial dimensions, followed by DAL, batch normalization, and a max-pooling layer with a 3x3 pool size and a stride of 3. This structure was replicated in the second convolutional layer with 64 filters, followed by another DAL, batch normalization, and max-pooling layer. A third convolutional layer with 128 filters, followed by DAL, batch normalization, and max-pooling, further refined the features. After each ReLU activation function in these convolutional layers, we added the DAL layer to dynamically adjust the importance of each channel. Attention layers were placed before max-pooling and right after the convolutional layers to ensure the network focused on the most informative features before spatial dimensions were reduced, thereby enhancing the effectiveness of feature recalibration (Wang, Wu, Zhu, Li, Zuo & Hu, 2020; Woo, Park, Lee, & Kweon, 2018).

To prepare the data for the fully connected layers, a flattening layer was used, resulting in a vector of shape 1152. This vector was then fed into a dense layer with 200 units, utilizing the ReLU activation function and incorporating L2 regularization. The following dense layers consisted of 100 units, 10 units, and a final output layer with a single unit, each using ReLU activation, batch normalization, and L2 regularization. The final layer of our model was tailored to the classification type: a single-unit layer with a sigmoid activation function was used for binary classification, while a 10-unit layer with a SoftMax activation function was used for multi-class classification.

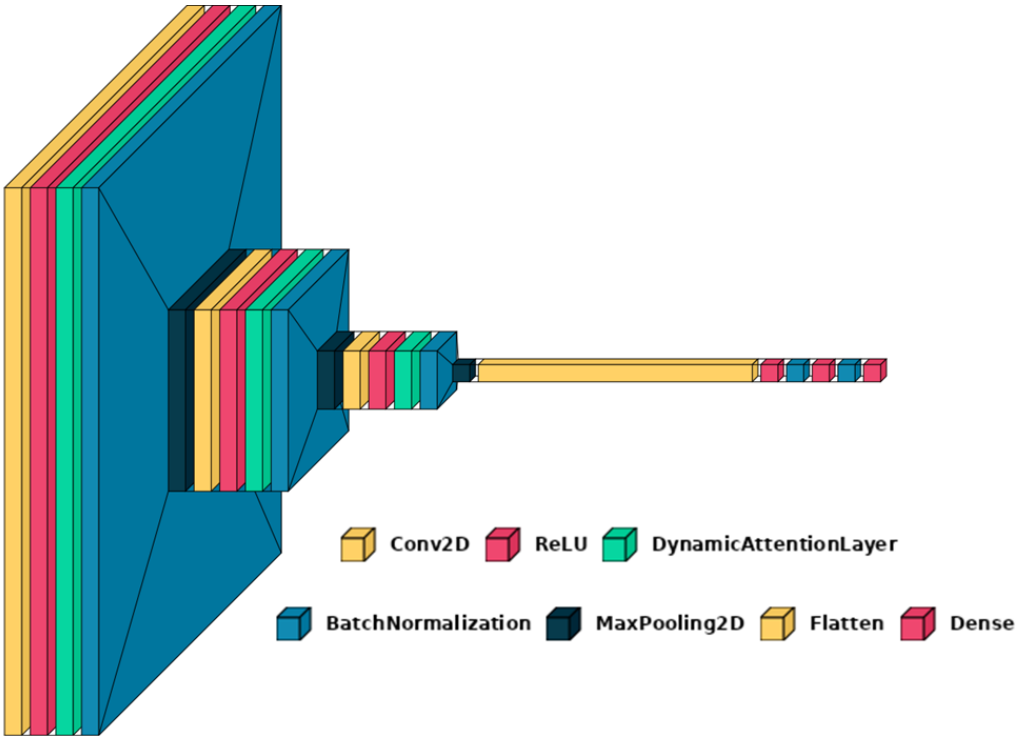

Figure 4.1 Architecture of the custom CNN.

For training the custom CNN model, we employed the Adam optimizer (Kingma & Ba, 2014) to perform backpropagation updates. The model was trained over 20 epochs to cater to the varying complexities of each dataset. We maintained a consistent batch size of 32 during training to ensure reliable gradient estimation. The loss functions were chosen based on the nature of the classification task: binary cross-entropy for binary classification tasks and categorical cross-entropy for multi-class scenarios.

This allowed us to tailor the learning process to the specific characteristics and demands of the input data, aiming to maximize performance and accuracy in discriminating between the different classes in the training sets.

Datasets

For our initial challenge and dataset evaluation, we utilized the MNIST dataset (Deng, 2012). This dataset consists of 60,000 training images and 10,000 testing images of handwritten digits in a 28x28 single-channel format. The custom model was trained on the original, unmodified MNIST images. For testing, the model was exposed to a variety of transformed images (Figure 4.2), including random cropping, rotation, and the addition of Gaussian and patterned noise. These transformations aimed to assess the model's robustness and ability to generalize across altered data conditions reflecting real-world scenarios.

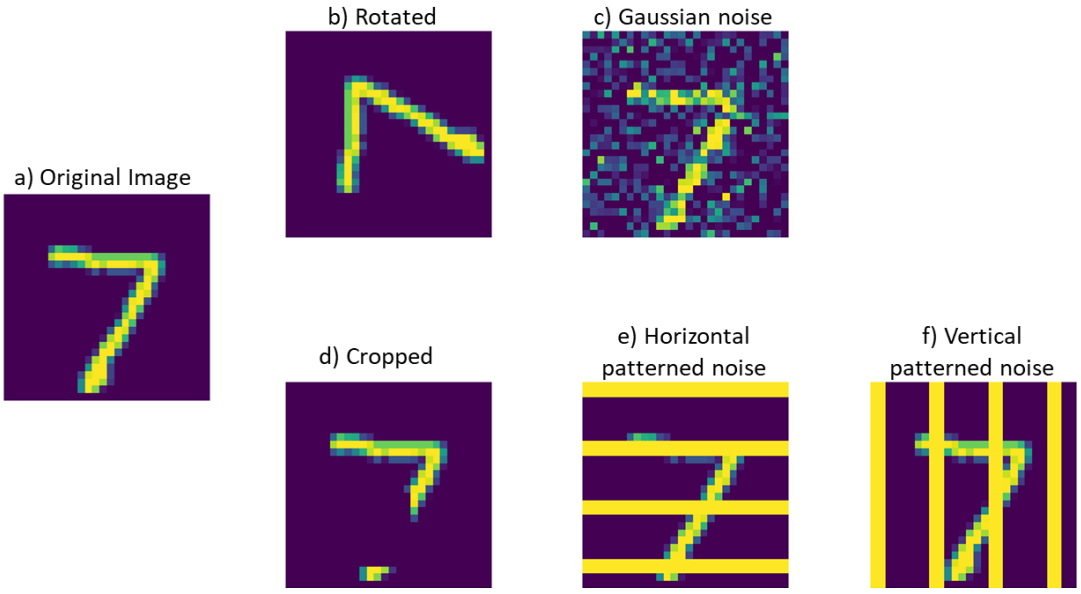

Figure 4.2 An example of an MNIST training sample (a) and transformed OOD samples (b-f) featuring novel attributes.

The second dataset and challenge involved a classification task with fruits and vegetables, sourced from Kaggle (Seth, 2020). The training set included images of familiar fruits like bananas and apples, as

well as vegetables such as cucumbers and carrots. This training set comprised 489 fruit images and 518 vegetable images, all resized to 150x150x3 pixels in RGB format. To test the model's generalization capabilities, we used an OOD test set featuring unseen fruits like watermelons and vegetables like eggplants as shown in Figure 4.3. The test set contained 348 fruit images and 790 vegetable images. This setup allowed us to assess the model's ability to handle data variations and its effectiveness in accurately classifying previously unseen images.

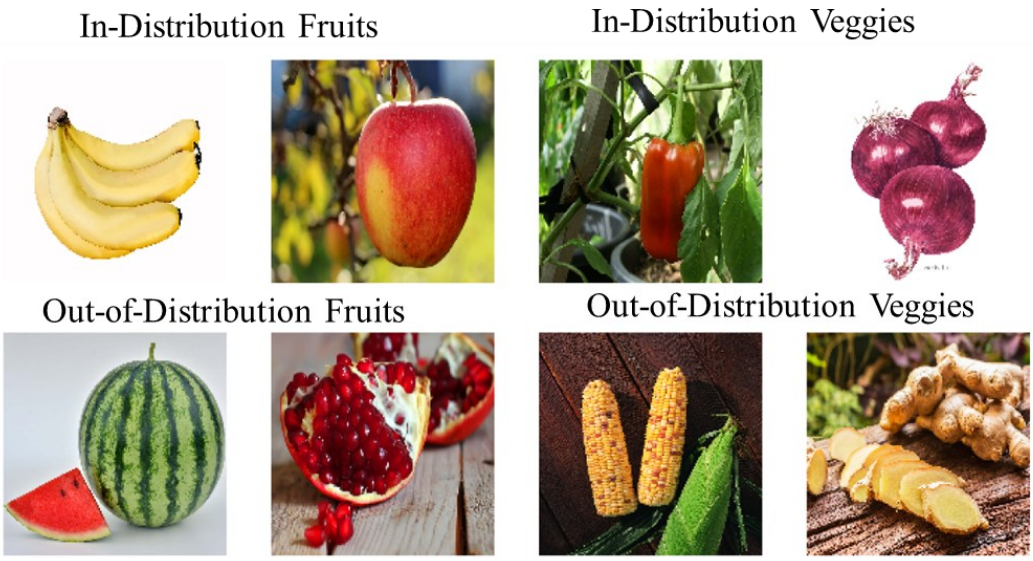

Figure 4.3 Training images of in-distribution fruits and vegetables compared with OOD examples assessed in generalization tests

For the third and final challenge, we used a dataset from Kaggle (https://www.kaggle.com/datasets/tongpython/cat-and-dog) featuring two categories: cats and dogs. The dataset included a total of 10,028 images, with 5,011 images of cats and 5,017 images of dogs. We split the dataset into 8,005 images for training and 2,023 images for in-distribution testing. All images were in RGB format with dimensions of 160x160x3 pixels as demonstrated in Figure 4.4. For OOD testing, we created a set of 50 images using DALL-E, evenly divided between those that blended features of both cats and dogs, such as cats with dog masks and dogs with cat masks. This setup was designed to evaluate the model's ability to differentiate between classes when presented with a mixture of key features.

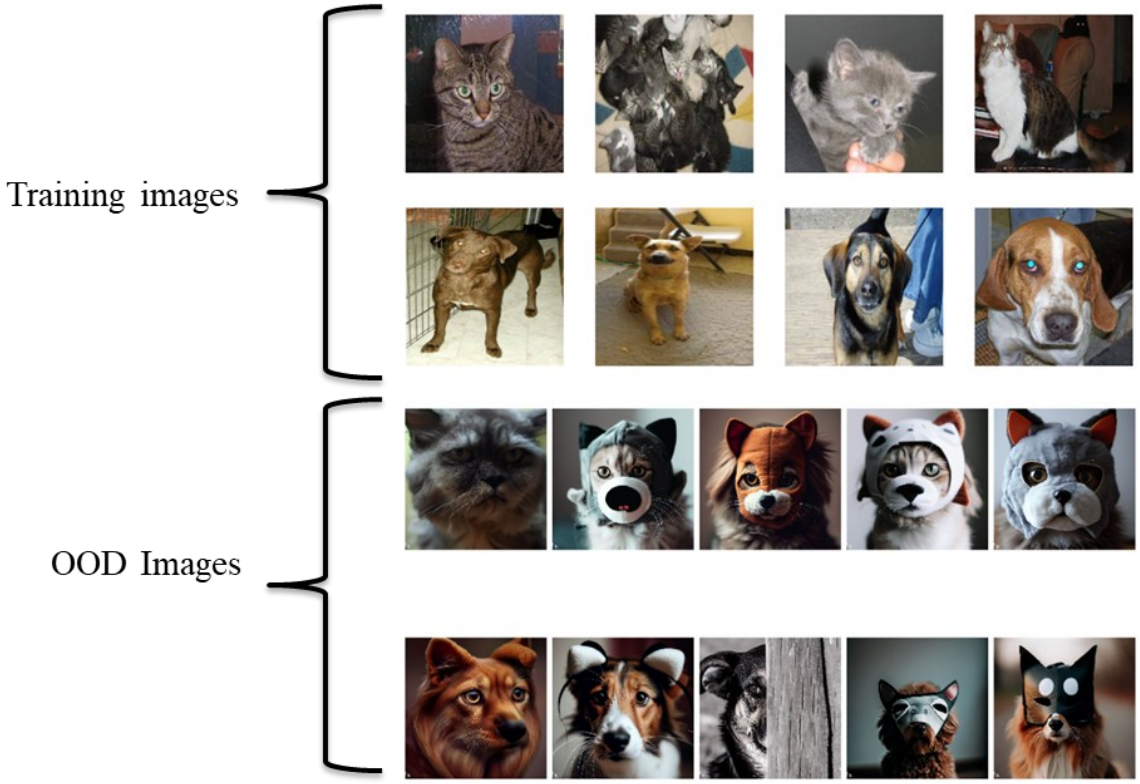

Figure 4.4 Training and OOD samples for both categories

Dynamic Attention Layer (DAL)

DAL can be obtained via the following procedure. Given an input image $X \in R^{H\times W\times C}$, where H is the height, W is the width, and C is the number of channels, the image is processed by a convolutional layer.

The convolutional layer applies a set of filters $W \in R^{k\times k\times C\times F}$, where $k \times k$ is the filter size, C is the number of input channels, and F is the number of filters. The output feature map $\boldsymbol{F}^{cn} \in R^{H'\times W'\times F}$, where *H'* and *W'* represent the height and width of the output feature map, is computed as:

$$\boldsymbol{F}^{\boldsymbol{cn}}_{i,j,f} = \sum_{c=1}^{C}\sum_{u=1}^{k}\sum_{v=1}^{k} X_{i+u-1,j+v-1,c} \cdot W_{u,v,c,f} \quad (1)$$

where (i,j) are the spatial positions in the output feature map, (u,v) are the spatial indices of the convolution filter, c is the channel index of the input image and f is the filter index.

The output of the convolutional layer is passed through a ReLU activation function, which is defined as:

$$A_{i,j,f} = \max\left(0, \boldsymbol{F}_{i,j,f}^{\boldsymbol{cn}}\right) \quad (2)$$

where A is the activation map.

After obtaining A, we move on to the DAL. The weights of DAL are defined as $V \in R^F$, where V_f is the attention weight for the f-th channel. The procedure for swapping the attention scores and calling the DAL can be summarized in the following pseudocode:

Algorithm 1 Swap Attention Scores

1: **function** SWAP_ATTENTION_SCORES(attention_scores, top_percentile=θ, low_percentile=τ)
2: Flatten the attention scores into *flat_attention_scores*
3: Calculate *num_scores* as the size of *flat_attention_scores*
4: Sort *flat_attention_scores* into *sorted_scores*
5: Calculate *top_index* as $\lfloor num\_scores \times \frac{top\_percentile}{100} \rfloor$
6: Calculate *low_index* as $\lfloor num\_scores \times \frac{low\_percentile}{100} \rfloor$
7: Find *top_threshold* as *sorted_scores*[top_index - 1]
8: Find *low_threshold* as *sorted_scores*[low_index - 1]
9: Find *top_indices* where *flat_attention_scores* $\geq$ *top_threshold*
10: Find *low_indices* where *low_threshold* $\leq$ *flat_attention_scores* $<$ *top_threshold*
11: Calculate *min_len* as min(*len(top_indices)*, *len(low_indices)*)
12: Truncate *top_indices* and *low_indices* to *min_len*
13: Gather *top_scores* from *flat_attention_scores* using *top_indices*
14: Gather *low_scores* from *flat_attention_scores* using *low_indices*
15: Update *flat_attention_scores* by swapping *top_scores* and *low_scores*
16: Reshape *updated_attention_scores* back to the original shape of *attention_scores*
17: **return** *updated_attention_scores*
18: **end function**

Algorithm 1 swaps attention scores based on upper bound percentile θ and lower bound percentile τ. It starts by flattening the attention scores for easier manipulation, sorting them in ascending order, and calculating indices for the upper bound and lower bound percentiles. Using these indices, it determines the upper bound and lower bound threshold values. Indices for scores within the selected percentiles are then extracted (lines 9 and 10). To ensure equal size for swapping, the smaller set of indices is truncated. The upper bound and lower bound scores are then swapped within the flattened attention

scores (as shown in Figure 4.5). Finally, the scores are reshaped back to their original size and returned at the end of the function.

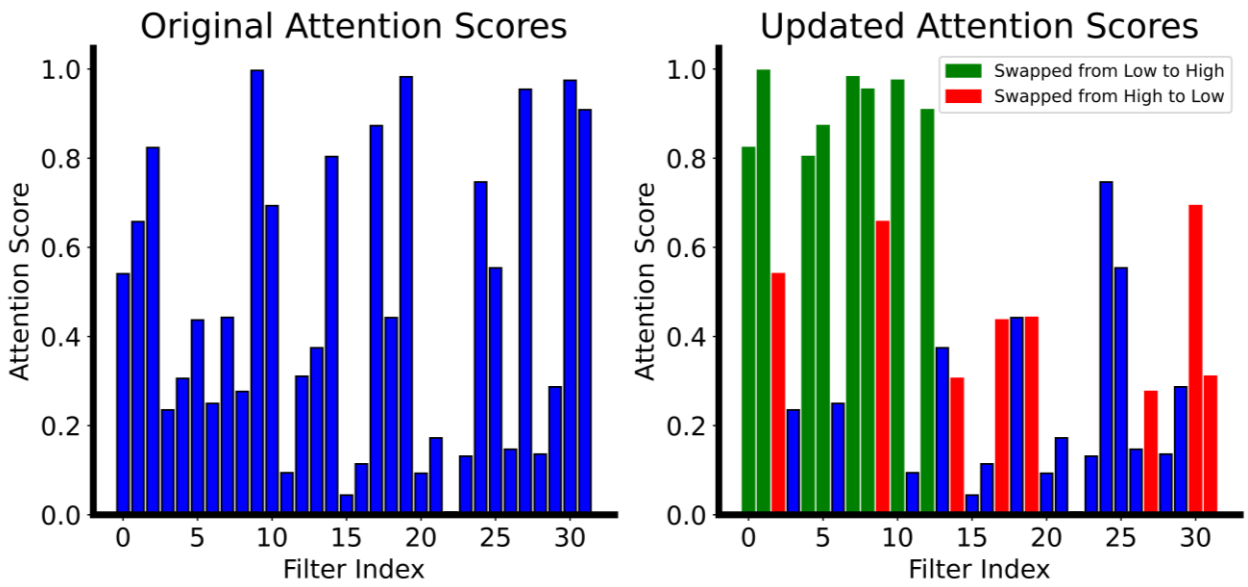

Figure 4.5 Demonstration of swapping attention scores based on percentile selection where θ=80 and τ=40.

Algorithm 2 is the layer call function that processes the layer within the CNN and splits the DAL sets of weights based on the batch size of the input during the training phase, not the inference phase. For example, if an input A has the following dimensions (32x16x16x8), the first dimension represents the batch size of 32. The weights of DAL, V, will have dimensions (32x8) where 32 is the number of sets of weights and 8 is the number of f in the input. These weights are split into V_f^{DAL} and V_f^{CWA} with dimensions (16x8). V_f^{DAL} is processed through the swapping function, while V_f^{CWA} remains unchanged. After the swapping, V_f^{DAL} and V_f^{CWA} are concatenated back into V with the original dimensions. The rationale behind splitting the batches into two sets is to enable the network to concurrently learn primary and secondary features, thereby enhancing its overall feature learning capability (Chen, Li, Bai, Yang, Jiang, & Miao, 2021; Hassaballah & Awad, 2016; Selvaraj, Veloso, & Rosenthal, 2018; Zhou, 2000). By swapping weights, our method does not discard dominant features or set them in opposition to secondary features;

instead, it promotes a balanced utilization of both, resulting in more accurate predictions. The final step is to pass V through the SoftMax function to convert the raw scores into a normalized probability distribution:

$$\alpha_f = \frac{exp(V_f)}{\sum_{f'} \exp(V_{f'})} \tag{3}$$

where f' is the summation index that runs over all possible channels in the set of scores V, and α_f is the normalized attention weight for the f-th channel.

Algorithm 2 Call Function

```
function CALL(inputs, Training=False)
    if Training = True then
        batch_size = shape(inputs)[0]
        half_batch = batch_size // 2
        attention_scores = channel_weights
        Expand attention_scores to match the dimensions of inputs
        Split attention_scores_expanded in half into attention_scores_to_modify by using half_batch as index
        Apply swap_attention_scores function to attention_scores_to_modify
        Concatenate modified_attention_scores and the rest of attention_scores_expanded
        Apply softmax to combined_attention_scores
        Apply combined_attention_scores to inputs to get modified_inputs
    else
        attention_scores = softmax(channel_weights)
        Expand attention_scores to match the dimensions of inputs
        Apply attention_scores_expanded to inputs to get modified_inputs
    end if
    return modified_inputs
end function
```

Unlike transformer-based self-attention (Vaswani et al., 2017), which computes pairwise interactions across all elements using queries, keys, and values, DAL operates independently on feature map channels. This design aligns with channel-wise attention mechanisms and focuses on dynamically redistributing channel importance rather than modeling global dependencies.

Results

MNIST Task

The first series of experiments focused on evaluating DAL and comparing it against traditional CWA and simple data augmentation techniques using the MNIST dataset. For simple data augmentation methods, we only used transformations not present in the OOD task, such as scaling, color adjustment, and stretching, to ensure the model was only exposed to transformations it had never seen before. Our custom CNN, incorporating DAL as illustrated in Figure 4.1, demonstrated significant improvements in robustness, especially against image transformations not encountered during the training phase. The results in Table 4.1 show that the model with DAL outperformed both CWA and simple data augmentation methods in recognizing digits across various challenging tasks.

Table 4.1 Evaluation of a custom CNN with DAL, CWA, and data augmentation on MNIST, assessing accuracy and loss under various distortions and in-distribution performance.

Method	Metric	Cropped	Rotated	Gaussian Noise	Horizontal Patterned Noise	Vertical Patterned Noise	In-Distribution Testing
Baseline	Accuracy (%)	73.45	14.74	43.12	14.51	23.47	98.14
	Loss	1.1619	6.7618	2.1767	6.068	4.1794	0.1044
DAL	Accuracy (%)	**83.5**	**21**	**72**	**28.5**	**28**	**98.69**
	Loss	**1.1366**	**4.3217**	**1.6738**	**4.7022**	**3.25**	**0.0616**
CWA	Accuracy (%)	75	19.5	47.5	23	24.5	98.45
	Loss	1.3777	6.7108	2.0475	5.8742	11.3151	0.0685
Data Augmentation	Accuracy (%)	74.61	17.79	44.17	18.17	23.98	98.35
	Loss	1.0147	4.9871	4.1579	4.079	5.309	0.0614

Best results are highlighted in bold.

Fruits and Veggies

In the assessment of the Fruits and Veggies dataset, we tested our custom CNN with integrated DAL and compared it to the CNN with original CWA and simple data augmentation transformations. Table 4.2 reveals that the CNN with DAL achieved an accuracy of 59% on previously unseen images of fruits and vegetables. The original CWA followed closely with an accuracy of 53%, while simple data augmentation techniques ranked third with an accuracy of 44%, and the baseline model achieved the lowest accuracy of 40%. This evaluation highlights the proficiency of DAL in enhancing the generalization capabilities of a CNN model when encountering new and diverse data. It is worth noting that for in-distribution testing, DAL also achieved the highest precision of 92% compared to the other methods.

Table 4.2 Assessment of a custom CNN with DAL, CWA, and data augmentation on Fruits and Veggies, evaluating accuracy and loss in OOD scenarios and in-distribution performance.

Method	Metric	In-Distribution	OOD
Baseline	Accuracy (%)	87	40
	Loss	0.7644	2.145
DAL	Accuracy (%)	**92**	**59**
	Loss	**0.6466**	**1.4374**
CWA	Accuracy (%)	89	53
	Loss	0.906	1.9847
Data Augmentation	Accuracy (%)	88	44
	Loss	0.5127	1.023

Best results are highlighted in bold.

Cats and Dogs

Next, we applied our custom CNN model with DAL to the Cats and Dogs dataset. This OOD task was particularly challenging as it involved distinguishing between cats wearing dog masks and dogs wearing cat masks. In this task, CNN with DAL achieved an accuracy of 46% and a loss of 2.8657. CNN with

original CWA closely followed, achieving an accuracy of 42% with a loss of 3.2721, as shown in Table 4.3. The simple data augmentation technique resulted in an accuracy of 38% and a loss of 5.7621. The baseline CNN, which lacked any attention layers or augmentation techniques, managed an accuracy of only 36% and had a significantly higher loss of 4.2354. This underscores the difficulty of this OOD task and demonstrates the advantage of incorporating attention layers. Additionally, for in-distribution testing, CNN with DAL achieved the highest accuracy of 84% with a loss of 0.4916.

Table 4.3 Evaluation of a custom CNN with DAL, CWA, and data augmentation on the Cats and Dogs dataset, analyzing accuracy and loss in OOD scenarios and in-distribution performance.

Method	Metric	In-Distribution	OOD
Baseline	Accuracy (%)	81	36
	Loss	0.5191	4.2354
DAL	Accuracy (%)	**84**	**46**
	Loss	**0.4916**	**2.8657**
CWA	Accuracy (%)	83	42
	Loss	0.5703	3.2721
Data Augmentation	Accuracy (%)	82	38
	Loss	0.5147	5.7621

Best results are highlighted in bold.

Exploring Optimal Percentile Selection

Finally, we performed an in-depth analysis using the Fruits and Veggies OOD dataset to determine the optimal values for θ (upper bound percentile) and τ (lower bound percentile) in our attention mechanism. This exploration aimed to identify the best combination of θ and τ to maximize model accuracy and minimize loss.

As illustrated in Figure 4.6, our results show that the highest accuracy and lowest loss are achieved when θ is set to the 80th percentile and τ is set to the 60th percentile, where we achieved an accuracy of

60% and a loss of 1.43. This combination indicates that swapping attention scores in these specific percentile ranges enhances the model's performance significantly. In previous experiments, we had θ set at 80 and τ at 40.

It is also important to note that when θ and τ are equal, the swapping mechanism has no effect. This is because the exchange will occur within the same set of scores, essentially nullifying the intended benefit and reverting the process to a simple channel-wise attention layer without the dynamic benefit of differentiated attention as introduced here.

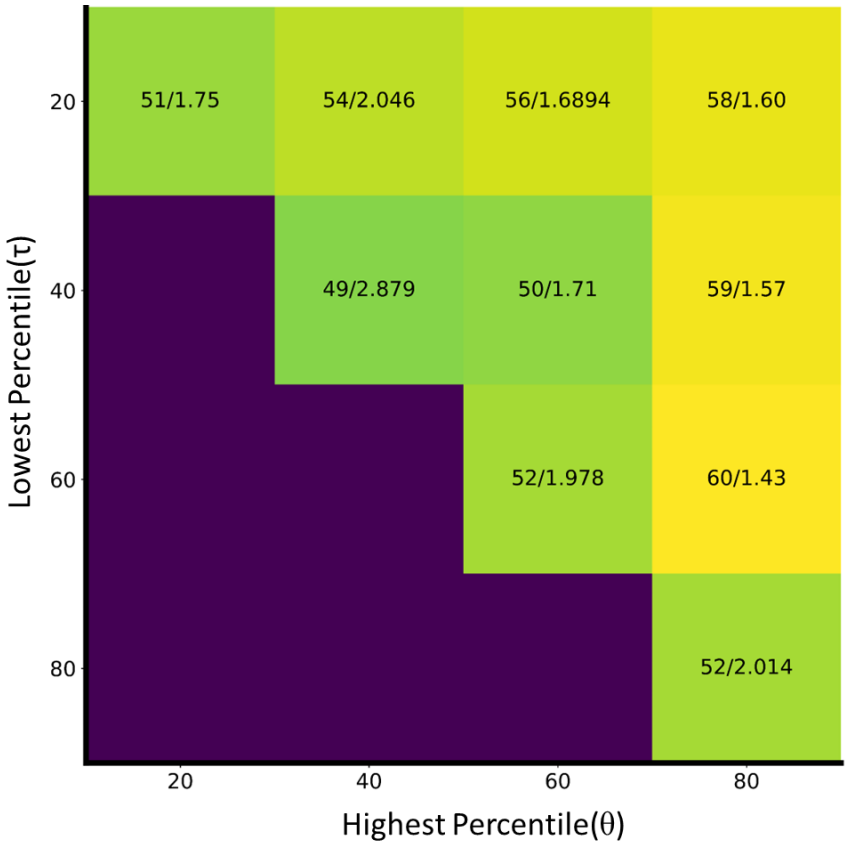

Figure 4.6 DAL's performance across different combinations of θ and τ. The grid shows accuracy (%) and loss values for various percentiles.

Discussion

The sensitivity of CNNs to OOD exemplars was an unresolved problem (Geirhos et al., 2020). The contribution of our work was to address this issue by introducing a novel attention layer, DAL, where

attention weights were dynamically interchanged based on selected percentiles during training. Our findings offered insights into the efficacy of the DAL for enhancing the OOD generalization capabilities of CNNs across three diverse datasets and challenges. Specifically, our results revealed that DAL not only increased accuracy on OOD distribution but also improved in-distribution accuracy, outperforming traditional CWA and simple data augmentation methods in challenging conditions. Additionally, by fine-tuning the θ and τ parameters used for selecting percentiles, we uncovered a range of optimal values that significantly improved the network's accuracy on OOD testing.

MNIST

In the analysis of the MNIST dataset, our experiments focused on evaluating the DAL and comparing it against the traditional CWA and simple data augmentation techniques. MNIST, being a benchmark dataset for handwritten digit recognition, provided a controlled environment to test the robustness of our proposed layer. Our custom CNN with DAL demonstrated significant improvements in robustness, especially against image transformations not encountered during the training phase.

The results in Table 4.1 showed that the model with DAL outperformed both CWA and data augmentation methods in recognizing digits across various challenging tasks. This enhancement could be attributed to DAL's ability to dynamically adjust attention weights, thereby improving the model's focus on both dominant and more subtle features. The findings suggested that DAL enhanced the model's generalization capabilities, making it more resilient to common distortions such as cropping, rotation, Gaussian noise, and patterned noise. This robustness is crucial for practical applications where the model may encounter unforeseen variations, such as noise in the input data like in medical imaging (Kumar & Nachamai, 2017) or real-time video processing (Ghazal, Amer & Ghrayeb, 2007). Furthermore, the ability of DAL to improve in-distribution accuracy suggested that it did not only perform well under novel conditions but also retained its effectiveness in standard settings.

Fruits and Veggies

The assessment of the Fruits and Veggies dataset involved testing our custom CNN with integrated DAL and comparing it to the CNN with original CWA and simple data augmentation techniques. This diverse dataset from Kaggle includes a mix of both in-distribution and OOD examples. The evaluation highlighted DAL's proficiency in enhancing the generalization capabilities of a CNN when encountering new data.

The CNN with DAL achieved an accuracy of 59% on previously unseen images, outperforming both CWA and data augmentation techniques, which achieved 53% and 44%, respectively. DAL's dynamic adjustment of attention weights allowed the model to better capture secondary features often overlooked by traditional models. For in-distribution testing, DAL achieved the highest precision of 92%, further demonstrating its superiority. By dynamically redistributing attention weights based on percentile selection, DAL ensured a balanced approach to feature importance, contributing significantly to the model's robustness and accuracy across both familiar and novel data variations. The integration of DAL could be particularly beneficial in fields like agricultural technology (Patrício & Rieder, 2018) or dietary monitoring (Hassannejad et al., 2017), where accurate identification of various natural products is essential.

Cats and Dogs

In the Cats and Dogs dataset, our custom CNN model with DAL was tested alongside CNNs enhanced with original CWA and simple data augmentation techniques. This dataset presented a particularly challenging OOD task, as it involved distinguishing between cats wearing dog masks and dogs wearing cat masks. Such a complex scenario tested the limits of a model's ability to generalize beyond its training data.

Despite this challenging task, the CNN with DAL achieved an accuracy of 46% and a loss of 2.8657, outperforming both CWA and data augmentation techniques. The baseline CNN, which lacked any

attention layers or augmentation techniques, managed an accuracy of only 36%, highlighting the added value of DAL in improving OOD performance. While 46% accuracy attained by

DAL is still relatively low, it underscores the inherent difficulty of this task for computer vision models. Despite all models performing below chance level (50%) in a binary classification task, DAL has both higher accuracy and lower loss, making it the relatively better performing model of them all. The DAL's performance is notable because it demonstrated the layer's capacity to discern subtle differences and secondary features that are crucial in such mixed-feature scenarios.

Additionally, for in-distribution testing, the CNN with DAL achieved the highest accuracy of 84% with a loss of 0.4916, compared to the baseline and other enhanced models. This high in-distribution accuracy indicated that DAL did not compromise the model's ability to perform well on familiar data while enhancing its OOD capabilities.

These results indicated that DAL enhanced the model's ability to differentiate between classes even when presented with a mixture of key features, showcasing its potential to handle complex real-world scenarios where traditional methods may falter. The dynamic nature of DAL ensured that the model could adapt its attention focus based on the relevance of features, providing a more robust and adaptable approach to feature selection and utilization. This capability is critical for applications requiring high precision in diverse and unpredictable environments, such as security (Idrees, Shah, & Surette, 2018) and surveillance (Sage & Young, 1998).

Optimal Percentile Selection

Our in-depth analysis using the Fruits and Veggies OOD dataset aimed to determine the optimal values for θ and τ in our attention mechanism. Additionally, we sought to understand the relationship between exchanging weights at different percentiles and their impact on model performance. This exploration aimed to identify the best combination of θ and τ that maximizes model accuracy and minimizes loss.

The optimal score was found when τ was set to the 60th percentile and θ was set to the 80th percentile, representing secondary and primary features, respectively. Interchanging these during training allowed the network to learn secondary features, enhancing its ability to generalize, which aligned with current literature on the importance of secondary features (Chen, Li, Bai, Yang, Jiang, & Miao, 2021; Hassaballah & Awad, 2016; Meng et al., 2022; Selvaraj, Veloso, & Rosenthal, 2018; Zhang, Ma, Zhao, Zhang, Sun, & Ma, 2022; Zhou, 2000). When looking at Figure 4.6, it became clear that the most critical selection in DAL was θ since a high θ produced optimal results. If θ was set to a lower bound percentile, τ had to be set lower as well, as τ could not be greater than θ. Exchanging weights at these lower percentiles did not create significant differences. Therefore, it was essential to keep θ at a higher bound percentile to introduce variability in the primary features. This approach ensured that the model captured a broader range of features, including subtle yet important secondary features, leading to better generalization and robustness. This strategy supported our hypothesis that attention mechanisms should focus on a balance between primary and secondary features to achieve optimal performance in diverse and unfamiliar data scenarios.

Computational Resources, Memory Usage, and Scalability of DAL

The network was trained on an NVIDIA RTX 3080 GPU. As shown in figure 4.7, training time decreased from 85.52 seconds at a batch size of 32 to 17.25 seconds at 256 with the DAL model, demonstrating efficient GPU utilization. Interestingly, the baseline CNN, despite having higher memory allocation across all batch sizes, trained faster than both CWA and DAL models. Memory usage for the baseline model remained consistently high, peaking at around 986 KB, while the DAL model peaked at about 944 KB, showing more efficient memory use.

The difference in memory allocation stems from the attention mechanisms in CWA and DAL, which dynamically focus on relevant features, reducing memory usage by avoiding unnecessary data

storage. In contrast, the baseline CNN allocates memory more statically across the network, leading to higher memory usage since all features are processed uniformly, regardless of their importance. While training times improved with larger batch sizes across all models, the gains diminished at higher levels, reflecting the natural limits of hardware parallel processing. Overall, figure 4.7 illustrates that DAL and CWA effectively balance computational efficiency with memory management, whereas the baseline model favors faster training times at the expense of higher memory consumption.

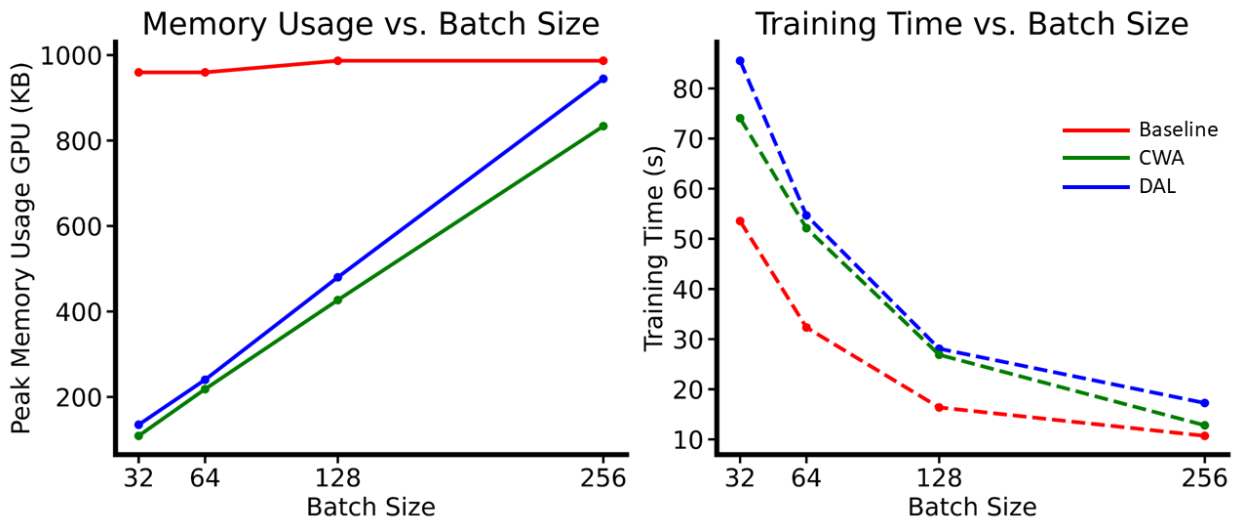

Figure 4.7 Scalability of Various Models: Memory Usage and Speed Across Batch Sizes during Training.

Limitations and Future Research

While our study demonstrated the efficacy of the DAL in improving the generalization capabilities of CNNs across various datasets, several limitations should be considered. First, the selection of optimal percentiles (θ and τ) was determined through empirical testing on specific datasets. This approach may not generalize well to all types of data, and further research is needed to develop a more systematic method for selecting these parameters.

While our experiments primarily involved binary classification tasks, this choice was intentional to highlight the limitations of simpler datasets in exposing the challenges faced by CNNs. The observed performance degradation in these tasks suggests that generalization issues would likely worsen in more complex, multi-class scenarios, emphasizing the applicability of DAL to a broad range of datasets and models. Expanding our current work to include a wider range of datasets will provide a more comprehensive understanding of DAL's effectiveness in multi-class datasets as well as more complex networks.

Future research could also explore the integration of DAL with other advanced neural network architectures and attention mechanisms. Combining DAL with techniques like self-attention (Vaswani et al., 2017) or transformer models (Islam, Elmekki, Elsebai, Bentahar, Drawel, Rjoub, & Pedrycz, 2023) could yield further improvements in model performance. Additionally, investigating the role of DAL in transfer learning scenarios, where pre-trained models are fine-tuned on new tasks, could provide insights into its versatility and utility in various domains.

A notable limitation of our study is that DAL was tested primarily on a custom CNN. To fully understand its potential and effectiveness, future studies should evaluate DAL on current state-of-the-art models. Testing DAL with modern architectures like ResNet, EfficientNet, and VGG models would provide a more comprehensive understanding of its benefits and potential limitations. These future explorations could solidify DAL's role in advancing the robustness and adaptability of neural networks in diverse and challenging scenarios.

Lastly, exploring the theoretical foundations of why certain percentiles (θ and τ) work better than others could lead to more informed and effective designs of attention mechanisms. Understanding the underlying principles that govern the effectiveness of percentile-based attention weight adjustments will help refine and enhance the DAL approach.

Conclusion

In this chapter, we introduced the DAL to enhance the generalization capabilities of CNNs for both in-distribution and OOD data. Our experiments on the MNIST, Fruits and Veggies, and Cats and Dogs datasets demonstrated that DAL significantly improved model accuracy and reduced loss compared to traditional CWA and simple data augmentation techniques. By dynamically interchanging attention weights based on selected percentiles, specifically θ=80% and τ=60%, DAL effectively captured both primary and secondary features, leading to better generalization. Despite the need for empirical parameter selection, DAL showed promise in addressing the challenges of OOD learning and enhancing the robustness of CNNs. Future research should focus on exploring its applicability to a wider range of datasets and integrating it with other advanced neural network architectures.

Acknowledgements

We thank members of the ComBiNe Laboratory at the University of Ottawa for useful discussions about this work.

References

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., ... & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data, 8*, 1-74.

Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., & Miao, Y. (2021). Review of image classification algorithms based on convolutional neural networks. *Remote Sensing, 13*(22), 4712.

Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., & Chua, T. S. (2017). Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5659-5667).

Deng, L. (2012). The MNIST database of handwritten digit images for machine learning research [Best of the Web]. *IEEE Signal Processing Magazine, 29*(6), 141-142.

DeVries, T., & Taylor, G. W. (2018). Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*. https://doi.org/10.48550/arXiv.1802.04865

Feng, X., Jiang, Y., Yang, X., Du, M., & Li, X. (2019). Computer vision algorithms and hardware implementations: A survey. *Integration, 69*, 309-320. https://doi.org/10.1016/j.vlsi.2019.07.005

Fort, S., Ren, J., & Lakshminarayanan, B. (2021). Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems, 34*, 7068-7081.

Gao, J., Yang, Y., Lin, P., & Park, D. S. (2018). Computer vision in healthcare applications. *Journal of Healthcare Engineering, 2018*. https://doi.org/10.1155/2018/5157020

Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence, 2*(11), 665-673. https://doi.org/10.1038/s42256-020-00257-z

Ghazal, M., Amer, A., & Ghrayeb, A. (2007). A real-time technique for spatio–temporal video noise estimation. *IEEE Transactions on Circuits and Systems for Video Technology, 17*(12), 1690-1699. https://doi.org/10.1109/TCSVT.2007.903805

Goceri, E. (2023). Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence Review, 56*(11), 12561-12605.

Guo, J., Qi, L., Shi, Y., & Gao, Y. (2023). PLACE Dropout: A progressive layer-wise and channel-wise dropout for domain generalization. *ACM Transactions on Multimedia Computing, Communications, and Applications, 20*(3), 1-23.

Hassaballah, M., & Awad, A. I. (2016). Detection and description of image features: An introduction. In *Image Feature Detectors and Descriptors: Foundations and Applications* (pp. 1-8).

Hassannejad, H., Matrella, G., Ciampolini, P., De Munari, I., Mordonini, M., & Cagnoni, S. (2017). Automatic diet monitoring: A review of computer vision and wearable sensor-based methods. *International Journal of Food Sciences and Nutrition, 68*(6), 656-670. [Online]. Available: https://doi.org/10.1080/09637486.2017.1283683

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., ... & Gilmer, J. (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 8340-8349).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7132-7141).

Idrees, H., Shah, M., & Surette, R. (2018). Enhancing camera surveillance using computer vision: A research note. *Policing: An International Journal, 41*(2), 292-307.

Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., & Pedrycz, W. (2023). A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications, 122666*.

Janai, J., Güney, F., Behl, A., & Geiger, A. (2020). Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision, 12*(1–3), 1-308. http://dx.doi.org/10.1561/0600000079

Khan, A. A., Laghari, A. A., & Awan, S. A. (2021). Machine learning in computer vision: a review. *EAI Endorsed Transactions on Scalable Information Systems, 8*(32), e4-e4. https://doi.org/10.4108/eai.21-4-2021.169418

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koo, K. M., & Cha, E. Y. (2017). Image recognition performance enhancements using image normalization. *Human-centric Computing and Information Sciences, 7*, 1-11.

Kumar, N., & Nachamai, M. (2017). Noise removal and filtering techniques used in medical images. *Oriental Journal of Computer Science and Technology, 10*(1), 103-113. [Online]. Available: http://dx.doi.org/10.13005/ojcst/10.01.14

Liang, S., Li, Y., & Srikant, R. (2017). Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*. https://doi.org/10.48550/arXiv.1706.02690

Liu, T., Luo, R., Xu, L., Feng, D., Cao, L., Liu, S., & Guo, J. (2022). Spatial channel attention for deep convolutional neural networks. *Mathematics, 10*(10), 1750.

Liu, Z., Xu, Y., Xu, Y., Qian, Q., Li, H., Jin, R., ... & Chan, A. B. (2022). An empirical study on distribution shift robustness from the perspective of pre-training and data augmentation. *arXiv preprint arXiv:2205.12753*. https://doi.org/10.48550/arXiv.2205.12753

Lu, L., Shin, Y., Su, Y., & Karniadakis, G. E. (2019). Dying ReLU and initialization: Theory and numerical examples. *arXiv preprint arXiv:1903.06733*. [Online]. Available: https://doi.org/10.4208/cicp.OA-2020-0165

Meng, R., Li, X., Chen, W., Yang, S., Song, J., Wang, X., ... & Pu, S. (2022, October). Attention diversification for domain generalization. In *European Conference on Computer Vision* (pp. 322-340). Cham: Springer Nature Switzerland.

Miao, Y., & Luo, W. (2022, January). Improve Generalization Ability of CNN by Data Augmentation and SE Block in Landmark Classification. In *2022 14th International Conference on Computer Research and Development (ICCRD)* (pp. 250-255).

Patrício, D. I., & Rieder, R. (2018). Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. *Computers and Electronics in Agriculture, 153*, 69-81. [Online]. Available: https://doi.org/10.1016/j.compag.2018.08.001

Sage, K., & Young, S. (1998, October). Computer vision for security applications. In *Proceedings IEEE 32nd Annual 1998 International Carnahan Conference on Security Technology (Cat. No. 98CH36209)* (pp. 210-215). doi: 10.1109/CCST.1998.723792

Selvaraj, S. P., Veloso, M., & Rosenthal, S. (2018, September). Classifier-based evaluation of image feature importance. In *GCAI* (pp. 162-175).

Seth, K. (2020). Fruits and vegetables image recognition dataset [Data set]. Kaggle. [Online]. Available: https://www.kaggle.com/kritikseth/fruit-and-vegetable-image-recognition

Shih, F. Y. (2010). *Image processing and pattern recognition: Fundamentals and techniques*. Hoboken, NJ: John Wiley & Sons.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*(1), 1-48. https://doi.org/10.1186/s40537-019-0197-0

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience, 2018*. https://doi.org/10.1155/2018/7068349

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11534-11542).

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 3-19).

Yan, L., Ye, Y., Wang, C., & Sun, Y. (2024). LocMix: local saliency-based data augmentation for image classification. *Signal, Image and Video Processing, 18*(2), 1383-1392. https://doi.org/10.1007/s11760-023-02852-0

Zhang, H., Ma, R., Zhao, Y., Zhang, Q., Sun, Q., & Ma, Y. (2022). Optimized Convolutional Neural Network Recognition for Athletes' Pneumonia Image Based on Attention Mechanism. *Entropy, 24*(10), 1434.

Zhou, X. S. (2000, June). Image retrieval: Feature primitives, feature representation, and relevance feedback. In *2000 Proceedings Workshop on Content-based Access of Image and Video Libraries* (pp. 10-14). IEEE.

Chapter 5: General Discussion

The general discussion of this thesis will encompass a comprehensive review of our research findings and their implications for the field of computer vision. We will begin by summarizing the key results from each study, including NNF, DFM, VEB, and DAL. Following this, we will delve into the specific contributions these studies make to advancing our understanding and improving the robustness and generalization capabilities of CNNs. Additionally, we will address the limitations inherent in each study, providing a critical analysis of their scope and potential weaknesses. Finally, we will outline future research directions that can build upon the insights gained from this work, aiming to further enhance the performance and applicability of CNNs in diverse real-world scenarios.

Summary of Main Findings

Enhancing CNN Robustness through NNF

This section delves into the key findings from the research on improving the robustness of CNNs by injecting partial noise into the training data. The study primarily aimed to address the ongoing challenge of CNN sensitivity to noisy data, which often hampers their real-world application.

The study introduced a novel training technique called NNF training. This method involves configuring the training dataset with a specific proportion of noisy and noise-free images. The rationale is that presenting noise-free images followed by noisy images allows the CNN to learn the associated features more effectively by providing it with clear exemplars first and then using this acquired knowledge to handle noisy image features. This sequential exposure to clean and noisy data enables the network to develop a more robust feature representation, making it better equipped to manage noise in real-world scenarios where data quality can vary significantly.

To test the proposed NNF training method, a custom CNN was built using the TensorFlow Keras framework. The architecture consisted of twelve layers, including convolutional layers for initial feature extraction, max-pooling layers for down-sampling, and dense layers for high-level reasoning. The network

was trained on the MNIST dataset, which is a large database of handwritten digits commonly used for training image processing systems. The choice of the MNIST dataset was strategic, as it provides a standardized benchmark for evaluating the performance of new algorithms and techniques. By employing a well-established dataset, the study ensured that the improvements observed could be directly attributed to the NNF training method rather than dataset-specific anomalies.

The NNF training method demonstrated marked improvements in classification accuracy with only a small proportion of noisy images in the training set. The research showed that even minimal noise injection significantly enhanced the generalization ability of the CNN when processing noisy data. Specifically, only 10-15% of noisy images in the training set were necessary to achieve maximum accuracy, with the exact proportion depending on the noise intensity of the test images. This approach allows for a reduction in computational cost by eliminating unnecessary noisy examples in the training set while maintaining high accuracy. This finding is particularly important for practical applications, as it suggests that robust models can be developed without the need for extensive and computationally expensive training on large sets of noisy data.

Further experiments revealed that retraining the network with a small subset of noisy images dramatically increased the testing accuracy on noisy data. For instance, retraining with just 600 noisy examples improved the accuracy from 23% to 92%. This finding underscores the efficiency of the NNF training method in enhancing noise robustness with minimal computational resources. The dramatic improvement in accuracy highlights the potential for quick and efficient model updates in dynamic environments where new types of noise might be encountered. By simply incorporating a small number of representative noisy samples, the model can rapidly adapt to new conditions, maintaining high performance without requiring extensive retraining.

The study also tested the robustness and generalization of the CNN trained on uncorrelated noise by evaluating its performance on correlated noise. Results showed that the model trained with NNF

training had significantly higher accuracy on correlated noise images compared to a model trained solely on noise-free images. This demonstrates the model's capability to generalize to unfamiliar stimuli, highlighting the effectiveness of the NNF training technique in producing noise-robust networks. The ability to generalize from uncorrelated to correlated noise indicates that the NNF method helps the network learn more abstract and transferable features, which are essential for robust performance across a wide range of real-world scenarios.

In summary, the NNF training method effectively improves the robustness of CNNs to noisy data by incorporating a small proportion of noisy images in the training set. This approach not only enhances generalization but also reduces computational cost, making it a valuable technique for developing robust convolutional networks for real-world applications. The findings of this study suggest that the NNF method could be widely adopted in various domains where data quality is a concern, offering a practical solution to the challenge of developing noise-resistant neural networks.

Dominant Feature Masking and Versatile Evaluation Benchmark

This section discusses DFM, a novel data augmentation technique aimed at improving the OOD generalization capabilities of CNNs. Traditional data augmentation methods, while effective in enhancing model performance on in-distribution data, often fall short in scenarios where the test data significantly deviates from the training data. DFM addresses this gap by masking the most salient features of training images, compelling the network to learn and utilize secondary, less dominant features. This approach is grounded in the concept of human holistic visual processing, where the focus extends beyond the most prominent elements of a scene to include a wider array of features.

This study also introduces the VEB, a comprehensive framework designed to rigorously test model robustness and generalization capabilities through a series of challenges across three distinct datasets. The first dataset, MNIST, consists of handwritten digits where the model is trained on clean images but tested on transformed images featuring cropping, rotation, and various types of noise to assess the CNN’s

performance on corrupted images. The second dataset, Fruits and Veggies, involves training the model on specific types of fruits and vegetables and testing it on new instances of the same classes not seen during training, evaluating the CNN's robustness to intra-class variation, which is crucial due to common domain shifts in real-world applications. The third dataset, Cats and Dogs, tests the network's ability to correctly classify adversarial images by training on normal images of cats and dogs but testing on adversarial images, such as cats wearing dog masks and dogs wearing cat masks. Overall, VEB assesses three important qualities of the CNN: resilience to transformation, recognition of intra-class variation, and robustness against adversarial attack.

In the MNIST dataset, the custom CNN trained with DFM exhibited remarkable resilience against various noise patterns and transformations, such as Gaussian noise, patterned noise, and occlusion. The model outperformed those trained with conventional augmentation techniques like SaliencyMix, MixUp, CutMix, and GridMix. This indicates that DFM helps the network learn more abstract and transferable features, essential for maintaining high performance when encountering noisy or transformed images. The ability to generalize from clean to noisy images is crucial for applications where data imperfections are common.

The Fruits and Veggies dataset presented a more complex challenge due to the organic shapes and natural variations inherent in the images. The ResNet-50 model trained with DFM achieved the highest accuracy on previously unseen fruits and vegetables, significantly outperforming other augmentation methods. This success illustrates DFM's potential to handle intricate and naturally varied subjects, making it a valuable tool for applications requiring precise recognition of diverse objects. The method's adaptability to different types of visual data underscores its effectiveness in enhancing model performance in more complex and less structured visual contexts.

In the Cats and Dogs dataset, DFM again proved to be the most effective augmentation technique. The model trained with DFM achieved the highest accuracy among the tested methods, demonstrating

its utility in distinguishing between visually similar categories under highly challenging conditions. This highlights DFM's ability to enhance model robustness against confusing inputs by forcing the network to rely on subtler cues, such as textures and secondary features, which are less likely to be manipulated in adversarial scenarios.

The study also explored the impact of the α parameter, which controls the intensity of feature masking, on the efficacy of DFM. Fine-tuning this parameter was found to be crucial for achieving optimal model performance. An α intensity range between 0.55 and 0.75 yielded the best generalization accuracy for OOD samples, providing a balance where sufficient feature visibility was maintained while still enhancing the model's ability to generalize from obscured or altered input data. This highlights the importance of parameter tuning in maximizing the benefits of DFM.

Additionally, the study highlighted the importance of considering secondary features in the learning process. Traditional augmentation methods often fail to account for these less prominent yet significant features, leading to models that are overly reliant on dominant features. By masking these dominant features, DFM encourages the network to learn from a broader spectrum of features, promoting a more balanced and holistic understanding of the input data. This approach aligns the training process more closely with human visual processing strategies, where a wide array of features is considered for accurate recognition and understanding (Desimone & Duncan, 1995).

The findings from this study suggest that DFM can be a powerful tool for improving the robustness and adaptability of computer vision models in diverse and dynamic environments. By encouraging networks to learn from a wider array of features, DFM not only enhances generalization but also improves the model's resilience against various types of noise and adversarial attacks. The method's adaptability and effectiveness across different datasets and tasks make it a valuable addition to the repertoire of data augmentation techniques in computer vision.

In sum, the study's findings underscore the potential of DFM to significantly enhance the OOD

generalization capabilities of CNNs. The method's ability to promote a more balanced and holistic feature learning process, coupled with its adaptability and robustness, makes it a valuable tool for improving the performance of computer vision models in real-world applications. The introduction of the VEB further ensures that improvements brought by DFM and other augmentation methods are thoroughly validated, providing a rigorous framework for assessing model robustness and generalization capabilities.

Dynamic Attention Layer

This section will discuss the implications of DAL, a novel approach developed in our work to address the challenges of OOD learning in computer vision. The primary objective of the DAL is to dynamically interchange attention weights based on selected percentiles (θ and τ) during the training phase. This technique aims to improve the network's ability to capture both dominant and secondary features, which are often overlooked by traditional attention mechanisms. By redistributing attention weights, the DAL ensures a balanced focus on various features, thereby enhancing the model's overall robustness and adaptability to new and unseen data.

The experiments conducted in this study utilized the VEB which consists of three distinct datasets to evaluate the effectiveness of DAL: augmented MNIST images, a novel dataset of unseen image classes, and a DALL-E generated dataset. The augmented MNIST images were used to assess the resilience of the model to transformations such as cropping, rotation, and various types of noise. The novel dataset of unseen image classes tested the model's performance on new instances within familiar categories, while the DALL-E generated dataset presented a unique challenge of class differentiation with mixed features.

The results demonstrated that the CNN with DAL significantly outperformed traditional Channel-Wise Attention (CWA) mechanisms and simple data augmentation techniques across all datasets. The DAL-enhanced model achieved higher accuracy and lower loss in both in-distribution and OOD scenarios, highlighting its superior generalization capabilities. Specifically, the DAL's ability to dynamically adjust attention weights allowed the network to better capture secondary features, which are crucial for

handling diverse and unpredictable data.

For the MNIST dataset, the DAL-enhanced CNN showed remarkable improvements in robustness against image transformations not encountered during the training phase. The model's performance on cropped, rotated, and noisy images was significantly better compared to the traditional CWA and data augmentation methods. This indicates that DAL enhances the model's ability to generalize across altered data conditions, reflecting real-world scenarios.

In the evaluation of the Fruits and Veggies dataset, the CNN with DAL achieved the highest accuracy on previously unseen images, demonstrating its proficiency in handling data variations. The dynamic adjustment of attention weights allowed the model to better capture secondary features, leading to improved generalization capabilities. This was particularly evident in the model's performance on OOD examples, where DAL outperformed both CWA and data augmentation techniques.

The Cats and Dogs dataset presented a particularly challenging OOD task involving distinguishing between cats wearing dog masks and dogs wearing cat masks. Despite the complexity of this task, the CNN with DAL achieved the highest accuracy and the lowest loss, underscoring the layer's capacity to discern subtle differences and secondary features. This showcases DAL's potential to handle complex real-world scenarios where traditional methods may falter.

Additionally, the study explored the optimal selection of percentiles (θ and τ) for the DAL, discovering that the best performance was achieved when θ was set to the 80th percentile and τ to the 60th percentile. This combination of percentiles maximized model accuracy and minimized loss, emphasizing the critical role of selecting appropriate percentiles for effective feature learning. By examining the Fruits and Veggies OOD dataset, the analysis aimed to understand the relationship between weight exchanges at different percentiles and their impact on model performance. The optimal values for θ and τ allowed the network to effectively learn secondary features, enhancing its ability to generalize. The study revealed that the most critical selection in DAL was θ, as a high θ produced optimal results.

Setting θ to a lower percentile required τ to be set lower as well, as τ could not exceed θ. Exchanging weights at these lower percentiles did not result in significant improvements, underscoring the importance of maintaining θ at a higher percentile to introduce variability in the primary features. This strategy ensured that the model captured a broader range of features, including subtle yet important secondary features, leading to improved generalization and robustness. The findings supported the initital hypothesis that attention mechanisms should balance primary and secondary features to achieve optimal performance in diverse and unfamiliar data scenarios.

In conclusion, the study presents DAL as a powerful tool for improving the generalization capabilities of CNNs in both in-distribution and OOD scenarios. The DAL's dynamic adjustment of attention weights based on selected percentiles allows the model to capture a broader range of features, leading to better performance across diverse datasets.

The Contribution of The Thesis to Advancing Knowledge

The primary contributions of this thesis lie in the development and evaluation of novel techniques that enhance the robustness and generalization capabilities of CNNs in computer vision. The advancements made through this research address critical gaps in the current understanding and application of domain generalization and OOD detection.

Introduction of Noise-Noise Free Injection

The research on injecting partial noise into the training data of CNNs delves into enhancing the robustness and generalization capabilities of these models. Previous studies have established the importance of incorporating noise into training data to improve model performance on noisy datasets (Grandvalet, Canu, & Boucheron, 1997; Zur, Jiang, Pesce, & Drukker, 2009). However, this thesis advances the existing knowledge by investigating the optimal ratio of noisy to noise-free images, retraining strategies, and the model's ability to generalize to different type of noise.

One of the key contributions of this thesis is the systematic examination of different noise ratios. By

varying the proportion of noise injected into the training data, the study identifies optimal levels that enhance the model's ability to generalize without compromising its accuracy on clean data. This detailed analysis provides a clearer understanding of how noise injection affects the learning process, offering guidelines for its effective application in different scenarios.

Additionally, the study investigates the impact of retraining models with noise-injected data. This aspect explores how retraining can further solidify the model's robustness, enabling it to better handle unforeseen variations in real-world data. The findings suggest that strategic retraining with noise can significantly boost a model's resilience, making it more reliable in practical applications where data imperfections are common.

The robustness and generalization capabilities of the CNN were further tested by evaluating its performance on correlated noise after being trained on uncorrelated noise. The results were positive, demonstrating that the model trained with NNF had a significantly higher accuracy on correlated noise images compared to a model trained only on noise-free images. This indicates that the NNF training method helps the network generalize better to unfamiliar types of noise, a critical aspect for deploying CNNs in dynamic environments.

In summary, this study contributes to the field by providing a deeper understanding of the balance between noisy and noise-free data in training CNNs. It offers a practical solution to improve the robustness of models in noisy conditions without requiring extensive computational resources. The insights gained from this research pave the way for developing more efficient training methods that can be applied to various real-world scenarios, ensuring that CNNs maintain high performance despite the presence of noise.

Introduction of VEB

In this work, we introduced VEB, a set of learning challenges that offers a comprehensive evaluation across multiple dimensions, significantly contributing to the robustness and generalization

capabilities of CNNs. One of the critical aspects of VEB is its assessment of resilience to transformation. This involves evaluating how well models perform when faced with various transformed versions of the same data, such as through cropping, rotation, and noise addition. By using the MNIST dataset for these tests, VEB ensures that the models are not just memorizing the data but genuinely learning to recognize and adapt to changes in the input, thereby demonstrating their ability to handle corrupted or altered images effectively.

Previous benchmarks like CIFAR-10 and ImageNet have been instrumental in advancing computer vision but often focus on standard accuracy metrics without adequately addressing the robustness to transformations and noise (Krizhevsky, Hinton, & Sutskever, 2012; Deng et al., 2009). CIFAR-10, which contains 10 classes with 60,000 images, primarily evaluates classification accuracy, while ImageNet, with over 14 million images across 1,000 classes, has been a cornerstone for large-scale image recognition. VEB, by emphasizing resilience to transformation, provides a more holistic evaluation that is crucial for real-world applications where data corruption is common (LeCun et al., 1998; Cireșan et al., 2011). Another dimension of VEB's comprehensive evaluation is the recognition of intra-class variation. This aspect is crucial for understanding a model's ability to generalize within the same class but across different instances that it has not seen before. Using the Fruits and Veggies dataset, VEB tests the model's performance on unseen variations within the same class. This challenge is particularly important as it mirrors real-world scenarios where a model might encounter variations of objects it was trained to recognize. Ensuring that a model can handle such intra-class variations is essential for its robustness and reliability in diverse applications.

Traditional benchmarks like CIFAR-100 and COCO capture some degree of intra-class variation but do not specifically target this aspect in a structured manner (Krizhevsky, Hinton, & Sutskever, 2012; Lin et al., 2014). CIFAR-100 contains 100 classes with 600 images each, emphasizing diversity across classes, while COCO (Common Objects in Context) provides a large-scale dataset with object segmentation and

recognition challenges in varied contexts. VEB's focus on intra-class variation represents a significant improvement, ensuring that models are evaluated on their ability to generalize within classes, which is critical for practical applications (Russakovsky et al., 2015). Moreover, VEB evaluates the robustness against adversarial attacks, which is a critical aspect of modern computer vision systems. The benchmark includes an evaluation using adversarial examples, such as cats wearing dog masks and vice versa, to test the model's ability to maintain accuracy against deliberate perturbations. This challenge highlights the model's capability to resist and correctly classify manipulated inputs designed to confuse it, which is vital for applications in security and autonomous systems where adversarial attacks can have serious consequences.

Benchmarks like Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) focus on adversarial robustness but often in isolated scenarios without integrating other robustness aspects (Goodfellow, Shlens, & Szegedy, 2014; Madry et al., 2018). FGSM generates adversarial examples by adjusting input data along the gradient of the loss function to maximize perturbations, while PGD is an iterative method that refines these perturbations over multiple steps to find more effective adversarial examples. VEB's inclusion of adversarial robustness within a comprehensive framework offers a more integrated approach to evaluating model performance under adversarial conditions (Carlini & Wagner, 2017). The use of diverse and representative datasets is another significant strength of VEB. By incorporating datasets that capture different types of challenges—MNIST for transformation resilience, Fruits and Veggies for intra-class variation, and Cats and Dogs for adversarial robustness—VEB ensures a comprehensive assessment of model performance. Notably, there are no other benchmarks available that test all these three crucial components simultaneously. This diverse selection of datasets helps to cover a broad spectrum of domain shifts, providing a thorough evaluation of how well models can generalize across various types of data. Such a wide-ranging approach is crucial for developing robust models capable of performing reliably in different environments.

Benchmarks like DAWNBench and MLPerf offer comprehensive evaluations but primarily focus on performance metrics without the same emphasis on varied domain shifts (Coleman et al., 2017; Mattson et al., 2019). DAWNBench evaluates deep learning training and inference costs, highlighting efficiency and speed, while MLPerf provides a broad suite of benchmarks to assess machine learning performance across diverse hardware and software configurations. However, they lack VEB's emphasis on testing models against varied domain shifts, which is crucial for promoting robustness and generalization (Li et al., 2017; Peng et al., 2019). VEB is structured to address real-world applications directly. The design of VEB aligns closely with practical scenarios where models often encounter varied and unpredictable data. By simulating these real-world conditions within the benchmark, VEB provides a realistic assessment of a model's readiness for deployment in practical applications. This alignment with real-world challenges ensures that the models evaluated by VEB are not only theoretically sound but also practically effective.

In summary, VEB's structured challenges and evaluation criteria provide a systematic approach to benchmarking, making it a reliable tool for comparing different models and techniques on a level playing field. This comprehensive and realistic evaluation framework helps advance the field of domain generalization by ensuring that CNNs are robust, adaptable, and ready for diverse and unpredictable real-world applications.

Development of DFM

The research on DFM has significantly advanced the field of domain generalization for CNNs. This study tackles the limitations of current data augmentation techniques by introducing DFM, a method that enhances a model's ability to generalize to OOD data. By selectively masking dominant features during the training process, DFM forces the network to learn and leverage secondary features, thereby improving its robustness and adaptability.

DFM has been rigorously tested using the VEB, which includes datasets such as MNIST, Fruits and

Veggies, and Cats and Dogs. These datasets represent a wide range of tasks and conditions, providing a comprehensive evaluation of DFM's effectiveness. The results consistently showed that models trained with DFM outperformed those using traditional augmentation techniques, particularly in handling OOD scenarios. In comparative studies, DFM demonstrated superior performance against methods like Mixup (Zhang et al., 2018), Cutout (DeVries & Taylor, 2017), and CutMix (Yun et al., 2019). For instance, while Mixup and CutMix aim to enhance generalization by blending images and labels, they do not explicitly address the masking of dominant features as DFM does, which has shown to be particularly effective in OOD settings.

One of the key contributions of DFM is the deeper understanding it provides into the decision-making processes of neural networks. By focusing on secondary features, DFM reduces the reliance on dominant features and encourages a more comprehensive feature learning approach. This shift not only improves the model's generalization capabilities but also helps demystify the 'black box' nature of neural networks. By understanding how secondary features influence decision-making, DFM brings neural network operations closer to human visual perception, where subtle details often play a crucial role in recognition and understanding (Geirhos et al., 2018; Dodge & Karam, 2017).

Furthermore, the study delves into the optimal levels of feature masking, providing insights into how much of the dominant features need to be masked to achieve the best results. This aspect of the research is crucial as it highlights the balance required between masking enough dominant features to encourage secondary feature learning without overly compromising the model's initial learning process. For example, while the RSC technique introduced by Huang et al. (2020) also focuses on challenging dominant feature representations, DFM’s adjustable masking parameter offers more flexibility and control, leading to better performance in diverse scenarios. RSC works by suppressing the dominant features during training to force the network to learn more generalized features, which helps in improving its performance on unseen data. However, the static nature of the suppression in RSC can sometimes lead

to suboptimal feature learning, especially in cases where the dominant features are critical for certain classifications. In contrast, DFM provides a dynamic and adjustable approach to masking dominant features, allowing for a more tailored feature suppression strategy that can be fine-tuned based on the specific requirements of the dataset and task at hand. This flexibility ensures that while the network is discouraged from overly relying on dominant features, it can still leverage them when necessary, thus maintaining a balance between generalization and specificity. This balance is particularly important in scenarios where both dominant and secondary features play crucial roles in accurate classification (Zhou, 2000; Hassaballah & Awad, 2016; Selvaraj, Veloso, & Rosenthal, 2018).

In comparative evaluations, DFM consistently outperformed other data augmentation techniques. The models trained with DFM demonstrated enhanced robustness and adaptability, showing significant improvements in accuracy when exposed to OOD data. This superiority underscores the effectiveness of DFM in preparing models to handle the complexities and variations encountered in real-world applications.

Overall, the contributions of this study to the field of domain generalization are substantial. The introduction of DFM provides a novel and effective method for improving the robustness of CNNs. It advances the theoretical understanding of feature learning in CNNs, offering practical tools and techniques that can be applied across various applications. By bridging the gap between machine learning models and human visual perception, DFM ensures that models can perform reliably even in the face of significant domain shifts.

Development of DAL

The study introducing DAL makes several notable contributions to the field of computer vision, particularly in enhancing OOD and domain generalization for CNNs. The motive behind developing DAL was to explore secondary feature enhancement without relying on external tools such as data augmentation. One of the primary contributions is the introduction of a novel dynamic attention

mechanism. The DAL adjusts attention weights based on selected percentiles during the training phase. This innovative approach allows the model to capture both dominant and secondary features effectively, addressing the common issue where traditional models overly focus on dominant features and overlook subtler, yet crucial, secondary features. By dynamically interchanging these weights, DAL provides a balanced feature representation, which is crucial for improving the robustness and generalization capabilities of CNNs.

The study demonstrates DAL's effectiveness through its ability to improve model accuracy and reduce loss in both in-distribution and OOD scenarios. The research evaluates DAL across various datasets, including augmented MNIST images for transformation resilience, a dataset of unseen image classes for generalization performance, and a custom DALL-E generated dataset for testing class differentiation with mixed features. DAL consistently outperforms traditional CWA and simple data augmentation techniques, showcasing its superior ability to generalize across different domains and conditions. For instance, while CWA methods, such as those used in Squeeze-and-Excitation Networks (SENet) (Hu et al., 2018), have shown improvements in focusing on important channels, they lack the flexibility and dynamic adjustment that DAL offers, resulting in less effective handling of OOD data. SENet enhances the representational power of a network by explicitly modeling the interdependencies between channels through a squeeze-and-excitation mechanism, which involves compressing (squeezing) the feature maps to capture global information and then scaling (exciting) them to highlight important features, but does not dynamically adjust to different data distributions, thereby limiting its effectiveness in more varied scenarios. Similarly, while data augmentation methods like Mixup (Zhang et al., 2017) and Cutout (DeVries & Taylor, 2017) help in improving robustness, they do not specifically address the learning of secondary features, making DAL a more comprehensive solution.

Another key contribution is the identification of optimal parameter configurations for DAL. The research finds that setting the upper bound percentile (θ) to 80% and the lower bound percentile (τ) to

60% yields the best performance. This configuration maximizes model accuracy and minimizes loss, highlighting the importance of appropriately selecting these parameters to enhance feature learning. The ability to determine optimal configurations for attention mechanisms is a significant step towards more efficient and effective model training. This is a marked improvement over static parameter settings commonly found in other attention mechanisms, such as in Convolutional Block Attention Module (Woo et al., 2018), which applies sequential channel and spatial attention mechanisms but lacks dynamic adjustment, potentially limiting performance gains.

The study employs a robust evaluation framework that includes diverse and challenging datasets. This thorough assessment ensures that DAL not only performs well in controlled settings but also excels in practical, real-world applications where data variability is common. This comprehensive evaluation reinforces the potential of DAL to be a valuable addition to various computer vision systems.

Furthermore, DAL's focus on both primary and secondary features offers a significant advancement in understanding feature importance within neural networks. Unlike traditional methods that may obscure the decision-making process, DAL sheds light on how different features contribute to the network's decisions. This enhanced transparency helps to clarify the role of subtle and dominant features in decision-making processes, providing valuable insights into the internal workings of neural networks and contributing to the development of more robust and generalizable models. This contrasts with methods like Dropout (Srivastava et al., 2014), which improve generalization by randomly setting a fraction of the input units to zero during training, thus preventing overfitting. However, Dropout does not provide insights into feature importance, thereby maintaining the 'black box' nature of deep learning models.Overall, the study on DAL introduces an innovative and efficient approach to enhance the robustness and generalization capabilities of CNNs without relying on external tools, providing practical solutions and techniques for real-world applications. This study not only enhances theoretical understanding but also provides tangible improvements in model performance, making significant strides

towards more reliable and adaptable computer vision systems.

Limitations and Future Directions

While this thesis has made significant contributions to the field of computer vision, it is also important to acknowledge its limitations. As Isaac Newton wisely stated, "What we know is a drop, what we don't know is an ocean." This highlights the continuous nature of scientific exploration and the vast amount of knowledge yet to be discovered. Despite the advancements presented in this work, there remain areas that require further investigation and refinement. Each method introduced, whether it be the NNF, VEB, DFM, or DAL, has its own set of constraints and areas for improvement. Recognizing these limitations is crucial for guiding future research directions and refining these techniques for broader and more effective applications. In the following sections, we will assess the limitations of each technique separately, providing a comprehensive understanding of the current challenges and potential avenues for future work.

Partial Noise-Injection

While the study on injecting partial noise into training data for CNNs has yielded significant insights, several limitations must be acknowledged. First, the research was conducted exclusively on the MNIST dataset, which is relatively simplistic and does not fully represent the complexity of real-world image datasets. Consequently, the generalizability of the findings to more complex and varied datasets remains uncertain. Future research should extend the evaluation of the NNF training method to datasets with higher complexity and more diverse image characteristics to confirm its efficacy across a broader range of scenarios.

Another limitation is the focus on Gaussian noise and correlated noise types. While these are common forms of noise encountered in practical applications, there are numerous other noise types, such as salt-and-pepper noise, speckle noise, and motion blur, which were not considered in this study. Future work should explore the impact of these different noise types on the robustness of CNNs when trained

using the NNF method. This would provide a more comprehensive understanding of how various noise conditions affect model performance and the potential for NNF training to generalize across them.

Additionally, the optimal ratio of noisy to noise-free images identified in this study is specific to the conditions tested. While the findings suggest that only a small proportion of noisy images is necessary for effective training, the exact ratio may vary depending on the dataset, noise type, and specific model architecture. Further research is needed to establish more generalizable guidelines for the proportion of noisy images required across different contexts.

The study also revealed that while NNF training improves robustness to noise, the underlying mechanisms behind this improvement are not fully understood. It is hypothesized that a certain number of noisy images are sufficient to achieve an optimal global minimum during training, but the precise dynamics of how CNNs learn from noisy data require deeper investigation. Future research should aim to elucidate these mechanisms to better understand how noise influences learning and to refine training techniques accordingly.

Moreover, the computational efficiency of the NNF method, while advantageous, was not compared with other noise robustness techniques in a detailed manner. Future studies should conduct a comparative analysis of the computational costs and benefits of NNF training relative to other methods, such as dropout, data augmentation, and adversarial training, to provide a clearer picture of its relative efficiency and practical viability (Cubuk et al., 2019; Goodfellow et al., 2014; Srivastava et al., 2014).

Finally, while the study demonstrated the potential of retraining models with a small subset of noisy images to significantly enhance robustness, the long-term stability and scalability of this approach remain uncertain. Future work should investigate the implications of retraining on larger and more varied datasets and assess the potential for integrating this method into continuous learning systems where models are periodically updated with new data (Lopez-Paz & Ranzato, 2017; Parisi et al., 2019).

In summary, while the partial noise injection study presents promising results, future research should

focus on validating these findings across more complex and varied datasets, exploring the impact of different noise types, establishing generalizable training guidelines, understanding the underlying learning mechanisms, comparing computational efficiency with other techniques, and evaluating the long-term stability and scalability of the retraining approach. These steps will be crucial for translating the theoretical benefits observed into practical, real-world applications.

Limitations of VEB

The introduction of the VEB represents a significant advancement in the assessment of CNN robustness and generalization capabilities. However, several limitations need to be addressed to enhance its utility and applicability further.

One primary limitation of VEB is its reliance on three specific datasets: MNIST, Fruits and Veggies, and Cats and Dogs. While these datasets are diverse and represent various challenges, they do not encompass the full spectrum of real-world variations that CNNs might encounter. The MNIST dataset, for example, is relatively simplistic and primarily evaluates resilience to transformations like noise and rotation. To fully assess the robustness of CNNs, future iterations of VEB should include more complex and varied datasets, such as those involving higher-resolution images, and different types of noise (Deng et al., 2009; Geirhos et al., 2018).

Additionally, VEB's current structure primarily evaluates the robustness of models against predefined transformations and adversarial examples. However, real-world applications often involve more dynamic and unforeseen challenges. Future research should focus on incorporating dynamic evaluation scenarios into VEB, where models are tested in real-time against a continuously evolving set of challenges (Lopez-Paz & Ranzato, 2017; Parisi et al., 2019). This would better mimic the unpredictable nature of real-world data and provide a more rigorous assessment of model robustness.

The VEB framework also currently lacks a detailed analysis of the computational efficiency of the models being evaluated. While robustness and generalization are critical, the practical deployment of CNNs also

requires considerations of computational resources and efficiency. Future enhancements to VEB should include metrics for evaluating the computational cost of achieving robustness, such as training and inference times, memory usage, and energy consumption (Canziani et al., 2016). This would provide a more holistic view of model performance, balancing robustness with practical efficiency.

Another area for improvement lies in the interpretability of the results obtained through VEB. While the benchmark provides a comprehensive evaluation of model performance across different challenges, the underlying reasons for a model's success or failure in specific tasks are not always clear. Future versions of VEB should incorporate tools for interpretability, such as visualization of feature importance and decision pathways, to help researchers understand why certain models perform better than others (Ribeiro et al., 2016; Selvaraju et al., 2017). This could lead to more targeted improvements in model design and training techniques.

Lastly, the current implementation of VEB does not account for the impact of continuous learning and model updates. In real-world applications, models are often updated with new data over time. Future enhancements to VEB should include mechanisms for evaluating how well models maintain their robustness and generalization capabilities after being incrementally updated with new data (Parisi et al., 2019; Shin et al., 2017; Zhou et al., 2020). This would provide insights into the long-term stability and adaptability of CNNs in dynamic environments.

In conclusion, while VEB offers a robust framework for evaluating CNN performance, future improvements should focus on incorporating more complex datasets, dynamic evaluation scenarios, computational efficiency metrics, interpretability tools, and mechanisms for continuous learning assessment. These enhancements will help ensure that VEB remains a comprehensive and practical tool for advancing the field of computer vision.

Limitations of DFM

While DFM has demonstrated significant advancements in enhancing the robustness and

generalization capabilities of CNNs, several limitations should be addressed to further improve its efficacy and applicability.

One of the primary limitations of DFM is its dependence on selective masking of dominant features during training. While this approach encourages the learning of secondary features and enhances generalization, the method for identifying and masking dominant features can be complex and computationally intensive. The current implementation requires careful tuning of parameters to balance the masking process, which might not be straightforward for all types of data or models. Future research could focus on developing more efficient and automated methods for identifying and masking dominant features, potentially leveraging advancements in unsupervised learning or self-supervised learning techniques (He et al., 2020; Jing & Tian, 2020; Zhai et al., 2019).

Additionally, the evaluation of DFM has primarily been conducted on a specific set of datasets included in the VEB. While these datasets provide a comprehensive assessment of model robustness, they do not cover the full range of possible real-world scenarios. Future studies should test DFM on a broader spectrum of datasets, including those with more complex and diverse features, to ensure its effectiveness across different domains and applications. Expanding the evaluation to include more varied datasets will help validate the generalizability of DFM and its applicability to a wider range of tasks.

The current implementation of DFM also does not fully address the scalability of the approach to very large datasets or high-resolution images. As the complexity and size of datasets increase, the computational demands of DFM can become prohibitive. Future research should explore optimization techniques to improve the scalability of DFM, such as parallel processing, distributed computing, or more efficient algorithms for feature masking (Abadi et al., 2016; Dean & Ghemawat, 2008). This will help make DFM more practical for use in large-scale applications.

Moreover, while DFM has shown promise in improving the generalization of CNNs, it is still reliant on the quality and diversity of the training data. In scenarios where the training data is inherently biased or lacks

sufficient diversity, the benefits of DFM may be limited. Future work could investigate methods to enhance the diversity and representativeness of training data, potentially through advanced data augmentation techniques or synthetic data generation. Additionally, combining DFM with other robust training methods could further enhance its effectiveness (Goodfellow, Shlens, & Szegedy, 2014; Perez & Wang, 2017; Shorten & Khoshgoftaar, 2019).

Lastly, the long-term impact of DFM on continuous learning and model updates has not been extensively studied. In practical applications, models often need to be updated with new data over time. Future research should examine how DFM affects the stability and adaptability of CNNs in continuous learning scenarios, ensuring that the benefits of feature masking are maintained as the model evolves. This could involve developing strategies for incremental learning that incorporate DFM principles, helping to maintain robustness and generalization in dynamic environments (Lopez-Paz & Ranzato, 2017; Parisi et al., 2019).

In summary, while DFM has made significant contributions to improving the robustness and generalization of CNNs, addressing its limitations in feature identification, interpretability, scalability, data diversity, and continuous learning will be crucial for its future development. By focusing on these areas, future research can enhance the practicality and effectiveness of DFM, making it a more versatile and widely applicable technique in the field of computer vision.

Limitations of DAL

The DAL introduces innovative approaches to improving OOD generalization and feature learning in CNNs. However, several limitations need to be addressed to maximize its potential and applicability.

One notable limitation of DAL is the complexity of its implementation. The dynamic adjustment of attention weights based on selected percentiles during training requires careful calibration to ensure optimal performance. Determining the best percentiles for different datasets and tasks can be computationally intensive and may require extensive experimentation. Future research could focus on

developing more automated methods for selecting these percentiles, potentially through adaptive learning algorithms or meta-learning techniques that can dynamically adjust based on the training data (Finn, Abbeel, & Levine, 2017; Li, Yang, Song, & Hospedales, 2018; Vilalta & Drissi, 2002).

Another challenge is the computational overhead associated with dynamically interchanging weights during training. This process can significantly increase the training time and resource requirements, particularly for large datasets or high-resolution images. Optimizing the efficiency of DAL to reduce computational demands without compromising performance is a crucial area for future research. Techniques such as parallel processing, hardware acceleration, or more efficient attention mechanisms could be explored to address this issue (Dean et al., 2012; Hennessy & Patterson, 2017; Vaswani et al., 2017).

While DAL enhances the model's ability to focus on both dominant and secondary features, it may still struggle with extremely subtle or rare features that are critical in certain applications. The current approach might not fully capture these nuances, leading to potential gaps in model performance. Future improvements could involve integrating DAL with other feature enhancement techniques, such as fine-grained feature extraction or multi-scale analysis, to ensure a more comprehensive feature representation (Hariharan et al., 2017; Lin et al., 2017).

The evaluation of DAL has primarily focused on specific datasets and controlled experiments, which may not fully capture the variability and complexity of real-world scenarios. To validate the generalizability of DAL, future studies should test it on a broader range of datasets, including those with diverse and unpredictable variations. This will help ensure that DAL is robust and effective across different domains and applications. Expanding the evaluation to include real-world data will provide a more accurate assessment of its practical utility.

Another limitation is the potential impact on model interpretability. While DAL aims to balance the importance of primary and secondary features, the dynamic adjustment of attention weights can

complicate the understanding of how specific features influence model predictions. Developing interpretability tools tailored to DAL, such as visualization techniques or explainable AI methods, could help researchers and practitioners gain deeper insights into the model's decision-making processes (Ribeiro et al., 2016; Selvaraju et al., 2017). This would make DAL more transparent and easier to trust in critical applications.

Additionally, the current implementation of DAL does not fully address the challenges of continuous learning and model updates. In real-world scenarios, models often need to be updated with new data over time. Ensuring that DAL can maintain its benefits in such dynamic environments is an important area for future research. Investigating methods for incremental learning and model adaptation that incorporate DAL principles will be essential for maintaining robustness and generalization over time (Lopez-Paz & Ranzato, 2017; Parisi et al., 2019).

Furthermore, while DAL has shown promise in enhancing OOD generalization, it is still reliant on the quality and diversity of the training data. In scenarios where the training data is biased or lacks sufficient diversity, the effectiveness of DAL may be limited. Future work could explore combining DAL with advanced data augmentation techniques or synthetic data generation to enhance training data diversity (Perez & Wang, 2017; Shorten & Khoshgoftaar, 2019). Additionally, integrating DAL with other robust training methods could further improve its performance (Goodfellow, Shlens, & Szegedy, 2014).

In summary, while DAL has made significant contributions to improving feature learning and OOD generalization in CNNs, addressing its limitations in implementation complexity, computational efficiency, feature capturing, evaluation breadth, interpretability, and continuous learning will be crucial for its future development. By focusing on these areas, future research can enhance the practicality and effectiveness of DAL, making it a more versatile and applicable technique in the field of computer vision.

Conclusion

In conclusion, this thesis has explored several innovative approaches to enhance the robustness and generalization capabilities of CNNs in the field of computer vision. Through the introduction of NNF training, DFM, VEB, and DAL, we have addressed some of the critical challenges that current models face in handling OOD data and domain shifts.

The NNF training demonstrated that incorporating noise during the training phase significantly improves the resilience of CNNs to various types of noise and corruption, showcasing the importance of robust training techniques in developing adaptable models. DFM further contributed to this field by emphasizing the role of secondary features in model decision-making, providing a more holistic understanding of feature importance and reducing the reliance on dominant features. This approach not only improved model robustness but also aligned more closely with human visual perception strategies.

VEB was introduced as a comprehensive framework for evaluating model performance across multiple dimensions, including transformation resilience, intra-class variation recognition, and robustness against adversarial attacks. By utilizing diverse and representative datasets, VEB offers a thorough assessment of model capabilities, highlighting areas for improvement and guiding future research.

The development of DAL added another layer of sophistication to our understanding of feature importance in neural networks. By dynamically adjusting attention weights based on selected percentiles during training, DAL enabled models to capture both dominant and secondary features effectively, enhancing their generalization capabilities without relying on external augmentation tools.

While these contributions have significantly advanced the field of computer vision, it is crucial to acknowledge the limitations of each method. The complexity of implementing DAL, the reliance of DFM on the quality and diversity of training data, and the computational overhead associated with these techniques present areas for future research and improvement. Addressing these limitations through automated methods, optimization techniques, and the integration of complementary approaches will

further enhance the robustness and practical viability of CNNs.

Overall, this thesis has provided valuable insights and practical tools for improving the performance and reliability of CNNs in diverse and unpredictable real-world scenarios. By bridging the gap between theoretical advancements and practical applications, these contributions pave the way for the development of more robust, adaptable, and interpretable models in the rapidly evolving field of computer vision.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265-283.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mane, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2020). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., ... & Lacoste-Julien, S. (2017, July). A closer look at memorization in deep networks. In *Proceedings of the International Conference on Machine Learning* (pp. 233-242). PMLR.

Azuma, R. T. (2016). A survey of augmented reality. *Presence: Teleoperators & Virtual Environments, 6*(4), 355-385.

Bai, T., Luo, J., Zhao, J., Wen, B., & Wang, Q. (2021). Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*.

Blanchard, G., Lee, G., & Scott, C. (2011). Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems* (pp. 2178-2186).

Bridle, J. S. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing* (pp. 227-236). Springer.

Bulusu, S., Kailkhura, B., Li, B., Varshney, P. K., & Song, D. (2020). Anomalous example detection in deep learning: A survey. *IEEE Access, 8*, 132330-132347.

Canziani, A., Paszke, A., & Culurciello, E. (2016). An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*. Retrieved from https://arxiv.org/abs/1605.07678

Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 39-57). IEEE.

Caruana, R. (1997). Multitask learning. *Machine Learning, 28*(1), 41-75.

Chen, C., Dou, Q., Chen, H., Qin, J., & Heng, P. A. (2019, July). Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 865-872).

Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., & Miao, Y. (2021). Review of image classification algorithms based on convolutional neural networks. *Remote Sensing, 13*(22), 4712.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Doctor AI: Predicting clinical events via recurrent neural networks. *arXiv preprint arXiv:1511.05942*.

Cireșan, D. C., Meier, U., Masci, J., Gambardella, L. M., & Schmidhuber, J. (2011). Flexible, high performance convolutional neural networks for image classification. In Twenty-Second International Joint Conference on Artificial Intelligence.

Coleman, C., Narayanan, D., Kang, D., Zhang, T., Nardi, L., Bailis, P., ... & Zaharia, M. (2017). Dawnbench: An end-to-end deep learning benchmark and competition. Training, 100(101), 102.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2019). AutoAugment: Learning augmentation policies from data. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

Recognition (pp. 113-123).

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, 51(1), 107-113. https://doi.org/10.1145/1327452.1327492

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Le, Q. V., ... & Ng, A. Y. (2012). Large scale distributed deep networks. *In Advances in Neural Information Processing Systems* (pp. 1223-1231).

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248-255). IEEE. https://doi.org/10.1109/CVPR.2009.5206848

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience, 18(1)*, 193-222.

DeVries, T., & Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552.

DeVries, T., & Taylor, G. W. (2018). Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*.

Dodge, S., & Karam, L. (2017). A study and comparison of human and deep learning recognition performance under visual distortions. In *26th International Conference on Computer Communication and Networks* (pp. 1-7). IEEE.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.

Doumas, L. A., Puebla, G., Martin, A. E., & Hummel, J. E. (2022). A theory of relation learning and cross-domain generalization. *Psychological review*, *129*(5), 999.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1126-1135).

Fort, S., Ren, J., & Lakshminarayanan, B. (2021). Exploring the limits of out-of-distribution detection. In *Advances in Neural Information Processing Systems, 34*.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119-139.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*(4), 193-202.

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks, 1*(2), 119-130.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning* (pp. 1050-1059).

Gao, J., Yang, Y., Lin, P., & Park, D. S. (2018). Computer vision in healthcare applications. *Journal of Healthcare Engineering, 2018*. [Online].

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems*.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems, 27*.

Grandvalet, Y., Canu, S., & Boucheron, S. (1997). Noise injection: Theoretical prospects. *Neural Computation*, *9*(5), 1093-1108.

Grigorescu, S., Trasnea, B., Cocias, T., & Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics, 37*(3), 362-386.

Gulrajani, I., & Lopez-Paz, D. (2020). In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*.

Hariharan, B., Arbelaez, P., Girshick, R., & Malik, J. (2017). Hypercolumns for object segmentation and fine-grained localization. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 447-456).

Hassaballah, M., & Awad, A. I. (2016). Detection and description of image features: an introduction. In *Image Feature Detectors and Descriptors: Foundations and Applications* (pp. 1-8).

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual

representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9729-9738.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770-778.

Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., ... & Song, D. (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*.

Hendrycks, D., Mazeika, M., & Dietterich, T. G. (2019). Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*.

Hendrycks, D., Mazeika, M., Kadavath, S., & Song, D. (2019). Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems* (pp. 15663-15674).

Hennessy, J. L., & Patterson, D. A. (2011). *Computer architecture: a quantitative approach*. Elsevier.

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review, 22*(2), 85-126.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE conference on*

computer vision and pattern recognition, 7132-7141.

Huang, L., Wang, C., Peng, H., & Liu, C. (2020). Self-Challenging Improves Cross-Domain Generalization. In European Conference on Computer Vision (pp. 124-140). Springer, Cham.

Jetley, S., Lord, N. A., Lee, N., & Torr, P. H. S. (2018). Learn to pay attention. *arXiv preprint arXiv:1804.02391*.

Jing, L., & Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(11), 4037-4058.

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture, 147*, 70-90.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems* (pp. 6402-6413).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation, 1*(4), 541-551.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324.

Li, D., Yang, Y., Song, Y. Z., & Hospedales, T. M. (2017). Deeper, broader and artier domain generalization.

In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 5542-5550).

Li, D., Yang, Y., Song, Y. Z., & Hospedales, T. M. (2018). Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).

Liang, S., Li, Y., & Srikant, R. (2017). Enhancing the reliability of out-of-distribution image detection in neural networks. In *Advances in Neural Information Processing Systems* (pp. 136-145).

Lienhart, R., & Maydt, J. (2002). An extended set of Haar-like features for rapid object detection. In *Proceedings of the International Conference on Image Processing* (Vol. 1, pp. I-I). IEEE.

Lin, T. Y., Dollar, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2117-2125).

Lin, T. Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., ... & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In European Conference on Computer Vision (pp. 740-755). Springer, Cham.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*, 60-88.

Livingstone, M. S., & Hubel, D. H. (1988). Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. *Journal of Neuroscience, 7*(11), 3416-3468.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3431-3440).

Lopez-Paz, D., & Ranzato, M. (2017). Gradient episodic memory for continual learning. In *Advances in*

Neural Information Processing Systems (pp. 6467-6476).

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision* (Vol. 2, pp. 1150-1157).

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

Mattson, P., Cheng, C., Diamos, G., Coleman, C., Micikevicius, P., Patterson, D., ... & Tang, H. (2019). MLPerf: An industry standard benchmark suite for machine learning performance. IEEE Micro, 40(2), 8-16.

Muandet, K., Balduzzi, D., & Schölkopf, B. (2013). Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning* (pp. 10-18).

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 807-814).

Niu, H., Li, H., Zhao, F., & Li, B. (2022). Domain-unified prompt representations for source-free domain generalization. *arXiv preprint* arXiv:2209.14926.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks, 113*, 54-71. https://doi.org/10.1016/j.neunet.2019.01.012

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., & Tran, D. (2018, July). Image

transformer. In *International conference on machine learning* (pp. 4055-4064). PMLR.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., & Wang, B. (2019). Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1406-1415).

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual review of neuroscience, 13(1)*, 25-42.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 779-788).

Ren, J., Liu, P., Fertig, E., Snoek, J., Poplin, R., DePristo, M., ... & Lakshminarayanan, B. (2019). Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems* (pp. 14707-14718).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems* (pp. 91-99).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144. https://doi.org/10.1145/2939672.2939778

Roberts, L. G. (1965). Machine perception of three-dimensional solids. In *Optical and Electro-Optical Information Processing* (pp. 159-197). MIT Press.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234-241).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115(3), 211-252.

Scherer, D., Müller, A., & Behnke, S. (2010). Evaluation of pooling operations in convolutional architectures for object recognition. In *Proceedings of the International Conference on Artificial Neural Networks* (pp. 92-101).

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, 85-117.

Selvaraj, S. P., Veloso, M., & Rosenthal, S. (2018). Classifier-Based Evaluation of Image Feature Importance. In *GCAI* (pp. 162-175).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618-626. https://doi.org/10.1109/ICCV.2017.74

Shin, H. C., Lee, H., Kim, J., Park, H., & Byun, H. (2017). Continual learning with deep generative replay. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2994-3004.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*(1), 1-48.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sincich, L. C., & Horton, J. C. (2005). The Circuitry of V1 and V2: Integration of Color, Form, and Motion. *Annual Review of Neuroscience, 28*, 303-326.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*(1), 1929-1958.

Storcheus, D., Rostamizadeh, A., & Kumar, S. (2015, December). A survey of modern questions and challenges in feature extraction. In *Feature Extraction: Modern Questions and Challenges* (pp. 1-18). PMLR.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1-9.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Szeliski, R. (2010). Computer vision: algorithms and applications. Springer Science & Business Media.

Tang, K., Tao, M., Qi, J., Liu, Z., & Zhang, H. (2022, October). Invariant feature learning for generalized long-tailed classification. In *European Conference on Computer Vision* (pp. 709-726). Cham: Springer Nature Switzerland.

Tay, Y., Bahri, D., Metzler, D., Juan, D. C., Zhao, Z., & Zheng, C. (2021, July). Synthesizer: Rethinking self-attention for transformer models. In *International conference on machine learning* (pp. 10183-10192). PMLR.

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition (pp. 1521-1528).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.

Venkateswara, H., Eusebio, J., Chakraborty, S., & Panchanathan, S. (2017). Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5018-5027).

Vilalta, R., & Drissi, Y. (2002). A perspective view and survey of meta-learning. *Artificial Intelligence Review, 18*(2), 77-95.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 1, pp. I-511). IEEE.

Wang, X., Girshick, R., Gupta, A., & He, K. (2017). Non-local neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794-7803.

Wichmann, F. A., Janssen, D. H. J., Geirhos, R., Aguilar, G., Schütt, H. H., Maertens, M., & Bethge, M. (2017). Methods and measurements to compare men against machines. *Electronic Imaging, Human Vision and Electronic Imaging*, 2017(14), 36-45.

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV) (pp. 3-19).

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., ... & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *International Conference on Machine Learning*, 2048-2057.

Ye, Z., Qin, S., Chen, S., & Huang, X. (2021). Dominant patterns: Critical features hidden in deep neural networks. *arXiv preprint arXiv:2105.15057*.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6023-6032).

Zhai, S., Cheng, Y., Lu, W., & Zhang, Z. (2016). Deep structured energy based models for anomaly detection. In *International Conference on Machine Learning* (pp. 1100-1109).

Zhai, X., Oliver, A., Kolesnikov, A., & Beyer, L. (2019). S4L: Self-supervised semi-supervised learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1476-1485.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.

Zhou, K., Yang, Y., Qiao, Y., & Loy, C. C. (2020). Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 29, 2230-2244. https://doi.org/10.1109/TIP.2019.2945766

Zhou, X. S. (2000). Image retrieval: feature primitives, feature representation, and relevance feedback. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*.

Zhou, X., Wang, D., & Krähenbühl, P. (2019). Objects as points. *arXiv preprint arXiv:1904.07850*.

Zur, R. M., Jiang, Y., Pesce, L. L., & Drukker, K. (2009). Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. *Medical physics*, *36*(10), 4810-4818.