Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

MASTERTHESIS

On the Potential and Limits of Zero-Shot Out-of-Distribution Detection

submitted by

Fabian Meyer

MIN-Faculty

Department of Informatics

Course of studies: Master Informatics

Matrikelnummer: 6816480

Submission date: 16.04.2023

First examiner: Prof. Dr. Chris Biemann

Second examiner: Dr. Florian Wilhelm

Supervisors: Sven Müller, Florian Schneider, Xintong Wang

Fabian Meyer:
On the Potential and Limits of
Zero-Shot Out-of-Distribution Detection
Master Thesis, Informatics
Universität Hamburg

Abstract

The growing prevalence of artificial intelligent (AI) systems in nearly all aspects of everyday life has also led to their integration into critical domains, e.g. in nuclear power plants, autonomous vehicles, and the detection of fatal diseases. Given that these systems are initially designed and tested within controlled, closed-world environments, they may face unanticipated inputs when deployed in real-world scenarios, leading to uncertainty in their interpretation and response. To mitigate the risk of incorrect decision-making, Out-of-Distribution Detection (OOD detection) techniques ensure that AI systems make decisions only for data that originates from familiar distributions. Zero-Shot Out-of-Distribution Detection (Z-OOD detection) is a special case recently introduced, which builds on the zero-shot classification paradigm.

In this thesis, we explore the potential and limitations of Z-OOD detection for image classification by leveraging the capabilities of recent multi-modal architectures, such as the Clip model. To test the generalizability of the approach, we conduct large-scale benchmarks on 12 datasets with strong semantic shifts in the data using the two published Z-OOD detection methods, Maximum Concept Matching (MCM) and Zero-Shot Out-of-Distribution Detection based on Clip (ZOC), followed by a challenging comparison with a smaller semantic shift. The robustness of the methods is tested under different conditions, such as image corruption, and attempts are made to determine the lower bound of task difficulty of the methods. We investigate correlations with difficulty metrics from OOD detection and assess their predictive power.

The thesis also aims to understand whether advancements in domain adaptation methods can be transferred to OOD detection. To accomplish this, we test the methodology in a few-shot setup and compare it against benchmark results. Our findings indicate that Z-OOD detection is generally effective, especially in far-OOD scenarios. However, challenges arise in near-OOD cases, where the underlying Clip model faces difficulties in classification. We propose the Universal Clip-based Confusion Log Probability (UC-CLP) as a universal indicator of the difficulty of selected In-Distribution/OOD splits, improving comparability within the field.

Finally, we propose T-MCM and T-ZOC as domain-adapted few-shot OOD detection methodologies, with T-MCM demonstrating a lightweight, fast-adapting approach. The performance of these methods depends on the success of domain adaptation, showing potential for improvement.

Acknowledgements

I certainly would not have been able to complete this research project on my own, so I would like to take this opportunity to thank people who helped me to achieve this. First, I would like to thank Dr. Florian Wilhelm for the opportunity to write my thesis with inovex GmbH. Then, I would like to thank Prof. Dr. Chris Biemann for allowing me to write the master thesis in his research group.

Special thanks go to Florian Schneider, Sven Müller, Lars Engellandt, and Xintong Wang, who accompanied me throughout the entire process from finding the topic to the final revision. They provided me with fruitful discussions, advice, had an open ear for any amount of confused thoughts, and offered any support I needed. It was a pleasure to work with you. I would also like to thank all friends and my colleagues in Hamburg, where I spent most of my time writing this thesis. Special thanks go to Christian Gill, who also provided me with great feedback.

To my beloved family and especially my parents, thank you for your unconditional support. I could not have achieved this without you. Lastly, I would like to thank Pia Čuk for taking a coffee break whenever I needed one. Hvala za vse.

Contents

1. Introduction **1**
1.1. Motivation . . . 2
1.2. Approach . . . 3
1.3. Research Questions . . . 4
1.4. Structure of this Work . . . 4

2. Background **7**
2.1. Learning Theory . . . 7
2.1.1. Supervised Learning . . . 8
2.1.2. Binary Classification . . . 9
2.2. Deep Learning Architectures . . . 11
2.2.1. Classification . . . 11
2.2.2. Natural Language Processing . . . 12
2.2.3. Computer Vision . . . 16
2.2.4. Clip . . . 18
2.2.5. Adapters . . . 20
2.3. Out-of-Distribution Detection . . . 22
2.3.1. Differentiation from Related Topics . . . 24
2.3.2. Zero-Shot Out-of-Distribution Detection . . . 26

3. Related Work **31**
3.1. Out-of-Distribution Detection . . . 31
3.1.1. Evaluation Protocols . . . 31
3.1.2. Out-of-Distribution Detection without Outlier Exposure . . . 32
3.1.3. Few-Shot Out-of-Distribution Detection . . . 33
3.1.4. Zero-Shot Out-of-Distribution Detection . . . 34
3.1.5. OOD Detection Difficulty . . . 34
3.2. Zero-shot Transfer . . . 36
3.2.1. Vision Language Models . . . 36
3.2.2. Task and Domain Adaption . . . 37

4. Methods and Datasets **39**
4.1. Methodology . . . 39
4.2. Datasets . . . 42
4.2.1. Selected Datasets . . . 42

4.2.2. Corruptions . . . 47
4.2.3. Metrics . . . 48
4.3. Model Details . . . 51
4.3.1. Clip . . . 51
4.3.2. Baseline . . . 51
4.3.3. MCM . . . 52
4.3.4. ZOC . . . 52
4.3.5. Adapter . . . 53
4.3.6. Prompt Engineering . . . 54
4.4. Optimization of Zero-Shot Out-of-Distribution Detection Methods . . . 54
4.4.1. Domain Adaption for Image Classification . . . 54
4.4.2. Domain Adaption for Clip-based Out-of-Distribution Detection . . . 55

5. Experimental Analysis of Zero-Shot Out-of-Distribution Detection 57
5.1. Exploratory Search for Far-OOD . . . 58
5.2. Exploratory Search for Near-OOD . . . 61
5.3. Exploratory Search on Robustness of Zero-Shot OOD detection . . . 65
5.4. Discussion . . . 67
5.4.1. Findings . . . 67
5.4.2. Potential Issues and Shortcomings . . . 68
5.4.3. Future Work . . . 69

6. Zero-Shot Methods as Basis for Few-Shot OOD Detection 71
6.1. Few-Shot Domain Adaption . . . 72
6.1.1. Results . . . 73
6.1.2. Interpretation . . . 75
6.2. Discussion . . . 76
6.2.1. Findings . . . 76
6.2.2. Potential Issues and Shortcomings . . . 77
6.2.3. Future Work . . . 77

7. Conclusion 79
7.1. Contributions . . . 79
7.2. Future Work . . . 81

Bibliography 83

A. Supplementary Material 93
A.1. Near-OOD with different Linear Probes . . . 93
A.2. Training Details Text Decoder . . . 94
A.3. Temperature & Prompt Ablation . . . 95
A.4. Zero-Shot Methods as Basis for Few-Shot OOD Detection . . . 96

List of Figures

2.1. Graphical representation of Attention 13
2.2. Transformer model architecture 15
2.3. Graphical representation of Clip 18
2.4. Graphical representation of TIP-Adapter 21
2.5. Taxonomy of the generalized Out-of-Distribution Detection framework . . 25
2.6. Graphical representation ZOC inference 27

4.1. Two image samples from the Caltech101 dataset 43
Samples from the DTD Dataset 44
4.2. Two image samples from the DTD dataset 44
4.3. Graphical overview of the applied image corruptions 48

5.1. Comparison of MCM Strategies for far-OOD 59
5.2. Comparison of different MCM strategies for near-OOD 62
5.3. Comparison of Out-of-Distribution Detection strategies for near-OOD . . 63
5.4. Correlation of difficulty metrics and AUROC scores with corruptions . . . 64
5.5. Influence of the corruptions on the AUROC 66
5.6. The influence of corruption on the difficulty metrics 67

6.1. Bar chart comparing MCM and adapted methods 75
6.2. Bar chart comparing adapted methods and benchmark 76

A.1. Logistic Regression and fully connected linear layer on top of Clip features for each dataset 94
A.2. Mean batch loss in each epoch in the Decoder Training 94
A.3. ZOC temperature scaling ablation for a subset of the datasets. 95
A.4. MCM temperature scaling and prompt ablation 96
A.5. Full AUROCS for all methods in near-OOD setup 97
A.6. K-shot ablation on datasets with 2 to 64 samples per class 98

List of Tables

4.1. Overview of the datasets utilized in this work 43
4.2. Corruption values for each severity level of the Snow corruption 47
4.3. Clip-ViT hyperparameters 51

5.1. ID/OOD Combinations with AUROC $>$ 99% 60
5.2. Pearson correlation matrix for far-OOD 60
5.3. Numerical near-OOD results 63

6.1. Accuracy comparison of domain adaption methods 73
6.2. Mean AUROC scores for different MCM based OOD detection methods . 74
6.3. Mean AUROC scores for different ZOC based OOD detection methods . . 74

A.1. Accuracies for each dataset using the named strategy 93
A.2. Full OOD Detection results with domain adaption and linear benchmark. Bold indicates best results 96
A.3. Full classification results with all methods used in this thesis. Log is short for logistic regression 97

Acronyms

AUROC Area Under Receiver Operating Characteristic. v, vii, 29, 39, 49, 57, 58, 62–67, 72, 74, 75

BPE Byte Pair Encoding. 12, 14

Clip Contrastive Language-Image Pre-Training. i, v, vii, 2–4, 10, 11, 15, 17–22, 26–29, 31, 33, 34, 36, 37, 39–42, 49–56, 58, 59, 68, 69, 71, 72, 76, 77, 79–81, 93, 94

CLP Confusion Log Probability. 32, 34, 35, 40, 41, 49, 58, 60, 63–69, 81

CNN Convolutional Neural Network. 16–18

CV Computer Vision. 11, 16

DTD Describable Textures Dataset. 44

FPR False Positive Rate. 29

ID In-Distribution. i, 3, 22–29, 31–35, 39, 40, 49, 50, 55–61, 64, 67–70, 72, 75, 81

MCM Maximum Concept Matching. i, v, vii, 29, 34, 39, 40, 52, 55, 56, 58, 61–66, 68, 71–76, 79, 80, 95

MLS Maximum Logit Score. 59, 63, 65, 95

MMD Maximum Mean Discrepancy. 34, 35, 40, 41, 50, 58, 60, 63–67, 69

MSP Maximum Softmax Probability. 29, 32–34, 40, 42, 58, 59, 63–65

NLP Natural Language Processing. 11–15, 17, 19, 20

OOD Out-of-Distribution. i, v, vii, 1, 3, 22–24, 28, 29, 31–35, 39, 40, 49, 50, 54, 55, 57–64, 67–70, 79–81

OOD detection Out-of-Distribution Detection. i, v, vii, 1–5, 19, 22–35, 39–42, 46, 49–52, 54, 56–58, 61, 63, 64, 68, 71, 72, 74–77, 79–81, 93, 95

RNN Recurrent Neural Network. 13, 14, 17

SOTA State-of-the-Art. 19, 41, 71, 77, 80

TPR True Positive Rate. 29

ViT Vision Transformer. vii, 17, 18, 51–53, 55, 58, 72

Z-OOD detection Zero-Shot Out-of-Distribution Detection. i, 1–5, 7, 22, 26, 27, 31, 32, 34, 36, 39–42, 49, 50, 54–58, 61, 63, 66–69, 71, 77, 79–81

ZOC Zero-Shot Out-of-Distribution Detection based on Clip. i, vii, 15, 27, 28, 34, 40, 41, 52, 55, 56, 58, 61–67, 71–77, 79, 80, 94, 95

ZSA Zero-Shot Accuracy. 58, 60, 63–69

1. Introduction

There are numerous real-world applications where detecting anomalies is crucial, as they can range from low-risk scenarios, such as bird species classification in an informational app, to high-risk situations like autonomous vehicle navigation or medical diagnostics. Artificial intelligence becomes increasingly integrated into various aspects of human life, so it is imperative to ensure that these models operate only within their intended domains to maintain safety and efficacy. Out-of-Distribution Detection (OOD detection) is a critical area of research within the realm of machine learning, which focuses on identifying input data that falls outside the scope of a closed-world classification model. In these models, only a limited set of classes are recognized, and any input data not belonging to these classes is considered Out-of-Distribution (OOD).

Despite rapid advancements in machine learning benchmarks and state-of-the-art systems, the issue of robustness against incorrect or malicious input is frequently overlooked. This is where OOD detection plays a vital role: it aims to identify inputs that are out of distribution while simultaneously classifying them. One common approach is to assess a classifier's confidence and reject input if the prediction confidence does not exceed a specific threshold. However, research has shown that cross-entropy-based classifiers tend to produce overconfident predictions for incorrect classifications [2, 29], which highlights the need for more reliable OOD detection methods.

The advent of large-scale pre-trained vision-language models has given rise to a new image classification paradigm: zero-shot classification. This method, which relies solely on an image and a set of textual labels, can classify images with remarkable accuracy without any task-specific training. Zero-shot classifiers measure the similarity between image and label representations in a shared hyperdimensional embedding space, which offers several advantages, such as saving computational resources and time, and alleviating the need for extensive data collection. Recently, zero-shot classification has been applied to OOD detection, demonstrating competitive performance compared to more complex detection methods.

This thesis will explore the potential and limitations of Zero-Shot Out-of-Distribution Detection (Z-OOD detection) through three distinct approaches. Firstly, we will analyze additional datasets beyond the existing research to include a variety of scenarios and data shifts. This analysis will assist in evaluating the generalizability of zero-shot methods, which is not explicitly mentioned but is inherent due to the limited options for adaptation in these methods. Second, we will evaluate the robustness of these techniques by subjecting them to challenging detection problems and image corruptions. Finally, we

will employ domain adaptation strategies for the multi-modal pre-trained backbone of these methods, seeking to enhance their performance in Z-OOD detection to leverage the methods towards few-shot OOD detection. Throughout this investigation, we will also examine difficulty metrics from the closely related field of OOD detection and assess their relevance in Z-OOD detection, ultimately aiming to predict the limitations and potential of these techniques.

1.1. Motivation

The primary inspiration for this thesis stems from the remarkable achievements of foundation models, particularly the multi-modal models such as OpenAI's Clip [74], which is known to be one of the first to achieve competitive results in classification tasks and related challenges without the need for task-specific training. Although Clip itself is a large model with substantial resource consumption during training, widespread deployment of such models could potentially reduce the resource demands of artificial intelligence systems. This is because, for downstream applications, there is no need for vast amounts of training data, an extensive human effort for task adaptation, or high computational power. Overall, the potential to conserve resources is inherent in these methodologies. Therefore, this thesis investigates the novel application of the zero-shot paradigm in the research area of OOD detection. The transfer of the paradigm to OOD detection prompted similar challenges, which inspired this work and are briefly discussed below, along with the general issue in OOD detection of the difficult comparison between different methods.

Clip Zero-Shot Classification and its Limits

All Z-OOD detection methods are based on Clip and its zero-shot classification mechanism. However, it is known that this approach has limitations. Firstly, there are specific image domains, such as OCR, where the classifier performs significantly worse than even the simplest computer vision techniques, though still better than random guessing. Other, more specialized datasets, such as satellite images [34] or cancer detection [97], do not yield reliable results. The question arises whether OOD detection works with these methods as long as classification is functional, or if there is a knowledge gap.

Transfer of Domain Adaptation Methods to OOD Detection

The aforementioned gap can often be closed using domain adaptation techniques, which can significantly improve or even enable the performance of Clip zero-shot classification. Novel methods manage to utilize additional layers with relatively few parameters and require comparatively little data. It remains to be seen whether these methods can be transferred to OOD detection.

Comparability of Methods

OOD detection research faces a challenge: artificial environments are created by combining in-distribution and out-of-distribution data, and the goal is to identify the OOD instances. The combination possibilities are virtually infinite, and only a small subset of these possibilities can be chosen to cover a specific range of requirements. These possibilities are often qualitatively delineated from one another, making concrete comparisons beyond this range, such as for practical applications, difficult. Quantifying the chosen setup could provide valuable insights in this regard.

1.2. Approach

The primary question guiding this thesis, and thus its title, is: What are the potential and limits of Z-OOD detection? First, we will address the question of potential: How reliable and functional are the methodologies based on Clip's zero-shot classifier? We will investigate whether a wide range of common datasets can be combined to form In-Distribution (ID) and OOD tasks. This includes seemingly simple combinations as well as more challenging ones. Additionally, the two currently published methods for Z-OOD detection are closely related, and we will explore whether one of them is superior or in which areas each method excels.

Initially, the experiments will focus on the realm of image data where Clip zero-shot classification is feasible. As the methods are based on the classifier, it is highly likely that their potential is initially limited by its capabilities, and the area where it is possible provides ample room for research. Within these boundaries, we will examine which shifts in the ID and OOD distributions can be solved and how difficult a problem can become before the methodologies fail.

Furthermore, we will investigate whether these methods have the fundamental potential to compete with more complex, fine-tuned approaches; that is, whether they can achieve this in the zero-shot setup or if task improvements in the Clip backbone can transfer and enhance these methods. To this end, we will utilize current domain adaptation strategies, which have already shown significant improvements in classification tasks with very little data, thus aligning with the resource-saving setup.

1.3. Research Questions

The research questions driving this thesis are listed here. They result from the above discussion and aim to demonstrate the potential and limitations of Z-OOD detection.

RQ. 1 **Is the performance of current Zero-Shot Out-of-Distribution Detection methodologies generalizable, i.e., transferable to datasets within the realm of Clip's zero-shot capabilities?**

RQ. 2 **Where are the boundaries of Zero-Shot Out-of-Distribution Detection methods with respect to different difficulty metrics for Out-of-Distribution Detection?**

RQ. 3 **How does the performance of Out-of-Distribution Detection methods using Clip's zero-shot classifier compare to traditional state-of-the-art Out-of-Distribution Detection methods?** This question focuses on applying domain adaptation techniques to the backbone model in order to enhance Out-of-Distribution Detection performance without altering the detection method itself.

1.4. Structure of this Work

To tackle the research questions, the remainder of this work is structured as follows: the next chapter introduces Z-OOD detection as a method to detect outliers in image classification tasks and provides the necessary theoretical background to understand this study, followed by related research in this area. The next two chapters present the experiments, focusing on the generalization and robustness of Z-OOD detection, and then exploring improvements through domain adaptation. Finally, we summarize the findings, discuss potential limitations, and suggest future work. The chapters are described in more detail below.

Chapter 2 provides the theoretical background to the methods used in this work. More specifically, it will describe the utilized machine learning and deep neural network architectures as well as a theoretical approach to Out-of-Distribution Detection.

Chapter 3 contextualizes this thesis by showing related work in Z-OOD detection and the overarching topic, OOD detection. Furthermore, we provide related work regarding Clip and adapter-based fine-tuning strategies for Clip-like models.

Chapter 4 provides detailed information about the experiments conducted in this work. This includes an overview of the datasets, architectures, training and fine-tuning strategies and experimental setup.

Chapter 5 investigates the applicability of Z-OOD detection as a novel approach to OOD detection in image classification across a wide array of domains, which are characterized by multiple datasets with distinct properties. More specifically, the methodology is evaluated using twelve different image classification benchmarks and numerous image corruptions to assess the robustness of these approaches.

Chapter 6 expands the methodology of current Z-OOD detection to few-shot OOD detection by domain-adaption of the utilized models.

Chapter 7 concludes this thesis with a summary of findings, together with an outlook on possible future work.

Following the list of abbreviations and the bibliography, additional material is provided in Appendix A.

2. Background

This chapter provides an overview of the essential concepts and methodologies that form the foundation of our research. Understanding these fundamental ideas is crucial for gaining insights into the potential and limitations of Zero-Shot Out-of-Distribution Detection (Z-OOD detection). The chapter is structured as follows: First, we present an introduction to learning theory, where we discuss the key principles that guide the design and evaluation of machine learning algorithms. Next, we delve into deep learning architectures, highlighting the critical components and techniques that have driven the success of modern artificial intelligence systems. Lastly, we explore the field of Z-OOD detection, discussing its importance in ensuring the robustness and reliability of artificial intelligent models when confronted with unforeseen inputs. This comprehensive background will equip readers with the necessary knowledge to comprehend the challenges and opportunities that Z-OOD detection presents in the context of the rapidly evolving landscape of artificial intelligence.

2.1. Learning Theory

This section introduces learning theory as the foundation of machine learning and deep learning. This introduction is followed by a more specific description of the two cases of learning used in this thesis, supervised learning [27, 80] and contrastive learning [44].
Machine learning aims to develop models that can learn from data, identify patterns, and make predictions. Formally, these predictions are made with a differentiable function f with parameters θ that makes the predictions $\hat{y}$ based on inputs x:

$$f(x, \theta) = \hat{y}.$$

Central to the process of learning is the definition of a learning objective, which outlines the goal that the model seeks to achieve. The learning objective typically involves minimizing an error function, also known as a loss function or objective function, that quantifies the discrepancy between the model's predictions and the ground truth. By optimizing the error function, the model adjusts its internal parameters to capture the underlying data distribution and improve its performance on the given task. Formally, the goal of the learning process is to find the best set of parameters θ_* that minimize a

loss function L with respect to the predictions $\hat{y}_n$:

$$\theta_* = \arg\min_\theta L(\hat{y}_i, ...)$$

To minimize the error, machine learning models iteratively adjust their parameters in response to the available data. This adjustment process, known as learning or training, relies on optimization algorithms, such as gradient descent [78] or its variants. Gradient descent-based models are very common in the machine learning community and are characterized by leveraging the gradient of the error function with respect to the model parameters to guide the updates during the training process, ultimately converging to a set of parameters that minimize the error function.
This thesis uses methods that rely on two different learning approaches, supervised learning and contrastive learning, which training processes are described in detail in the following subsections. The ultimate goal of supervised learning is to learn to predict from the training data so that the model can make accurate predictions when presented with new, unseen data. Contrastive learning is an approach that focuses on learning useful representations in a shared latent space by comparing similar and dissimilar data points. It is often used in contexts where crafted labelled data is scarce or unavailable, even though it does need a large amount of text-image pairs: These are usually crawled from the Internet, so the captions are used as labels for the images. [74].

Finally, after training on training data, the model is tested using test samples of the same data distribution, that are not part of the training. This is crucial for assessing the model's ability to generalize to new examples, which is a key objective in machine learning. The testing set typically contains a smaller portion of the available data (e.g., 20-30%) and is kept separate from the training set to ensure that the evaluation is unbiased and reflects the model's true generalization performance. This is called the *train-test paradigm*, which helps prevent overfitting. This occurs when a model performs exceptionally well on the training data but poorly on new, unseen data. Overfitting typically arises when the model learns to memorize the training data rather than capturing the underlying data distribution, resulting in poor generalization to the testing set. By evaluating the model's performance on the testing set, practitioners can gain insights into its generalization capabilities and make informed decisions about model selection, hyperparameter tuning, and other aspects of the machine learning process.

2.1.1. Supervised Learning

In this paradigm, the learning process is guided by a ground truth that pairs input examples with their corresponding target outputs. Supervised learning algorithms aim to establish a mapping between inputs and outputs by minimizing a predefined loss function, which quantifies the discrepancy between the model's predictions and the actual ground truth values. Common supervised learning tasks include classification, where

the goal is to assign discrete class labels to input instances, and regression, where the objective is to predict continuous values. Supervised learning has been successfully applied across various domains, such as image recognition [49, 33], natural language processing [95, 17], and financial forecasting [14, 18], demonstrating its versatility and effectiveness in solving diverse real-world problems.

2.1.2. Binary Classification

Binary classification, or single-label classification, is a fundamental task in machine learning, where the objective is to categorize input samples into one of two distinct classes. Given a set of input features, the binary classification model generates a prediction, which corresponds to either the positive or negative class. The model's performance is often evaluated using metrics such as accuracy, precision, recall, and the F1 score. Using accuracy is, especially for highly imbalanced data, oftentimes misleading, as predicting only the majority class will provide good results. Thus, the other mentioned metrics are oftentimes preferred In essence, binary classification serves as the foundation for understanding more complex classification problems, enabling the development of advanced algorithms for diverse applications [7].

Multi-Class Classification

Multi-class classification represents a specific case where the objective is to predict one of multiple discrete class labels k for a given input. This task differs from binary classification, which involves predicting between only two classes. Multi-class supervised learning requires the development of models that can discern and differentiate between the distinct classes present in the dataset.

To accommodate the complexity of multi-class problems, various algorithms and techniques have been developed. Some prominent approaches include:

One-vs-All or One-vs-Rest: This strategy involves training multiple binary classifiers, one for each class, to distinguish between instances of that class and instances of all other classes. During the prediction phase, the class with the highest confidence score or probability from the individual binary classifiers is assigned to the input.

One-vs-One: This approach trains a binary classifier for each pair of classes, resulting in a total of $k(k-1)/2$ classifiers. During the prediction phase, the input is passed through each classifier, and a majority voting scheme determines the final class assignment.

Neural networks: Deep learning models, such as convolutional neural networks [53] for image classification and recurrent neural networks [92] or transformers [95] for natural language processing, can inherently handle multi-class problems by adjusting the output layer to match the number of classes and employing an appropriate activation function. This approach is further described in Section 2.2.

Contrastive Learning

Contrastive learning is a self-supervised machine learning approach that aims to enhance the discriminative capabilities of deep learning models by leveraging the inherent structure of data. This method involves training models to distinguish between positive (similar) and negative (dissimilar) pairs of data samples, which in turn allows the model to learn meaningful representations. The focus is on the identification of similarities and differences between data points, so contrastive learning effectively mitigates the need for large amounts of labelled data, the reliance on costly manual annotations is reduced. Recent advancements in contrastive learning have led to significant improvements in various tasks, including computer vision and natural language processing, contributing to the development of more robust and efficient AI systems. In contrastive learning, the commonly used loss function is the contrastive loss, also known as the triplet loss or the InfoNCE loss [69], depending on the specific implementation. The objective of the contrastive loss function is to minimize the distance between positive (similar) pairs while maximizing the distance between negative (dissimilar) pairs in the latent feature space. The distance is usually measured as cosine similarity, which is defined for two n-dimensional vectors a and b as

$$\text{sim}_{\cos} = \cos(\Theta) = \frac{a \cdot b}{||a|| \cdot ||b||} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}. \tag{2.1}$$

As loss, there are as described multiple options. The InfoNCE [69] is often used for multi-modal text-image similarity learning, which is the foundation of models used in this thesis. The loss is adapted to fit text-image pairs [75]. The InfoNCE loss for feature vector x_i, a positive sample x_i^+ (the anchor) and a set of negative examples X^- is defined as:

$$L_{\text{InfoNCE}}(x_i, x_i^+, X^-) = -\log \frac{\exp(\text{sim}(x_i, x^+))}{\exp(\text{sim}(x, x^+)) + \sum_{i=1}^{N} \exp(\text{sim}(x, x_i^-))} \tag{2.2}$$

The numerator represents the similarity between the anchor and positive pair, while the denominator represents the sum of similarities between the anchor and all pairs, including the positive and negative samples. This loss basically classifies the positive pair correctly among the set of positive and negative pairs conditioned on the anchor sample. This way, the similarity to the positive pair increases and the similarity for all negative pairs decrease. The InfoNCE loss shown in equation 2.2 is not introduced for multi-modal learning, but can easily be adapted by replacing the x_i with features of an image, x_i^+ with features of a matching text and X^- with a batch of non-fitting text features, e.g. randomly sampled. This is the approach of Clip [74], which is used in this thesis.

2.2. Deep Learning Architectures

In recent years, deep learning has revolutionized the field of artificial intelligence, enabling groundbreaking advancements in various domains, including Natural Language Processing (NLP) and Computer Vision (CV). This section will provide a comprehensive overview of the key deep learning architectures that have contributed to these advancements, focusing on both Natural Language Processing (NLP) and Computer Vision (CV) models. Finally, multi-modal models will be explored with a specific focus on OpenAI's Clip model [74]. Clip bridges the gap between NLP and CV. By combining the advancements of both fields, Clip demonstrates the potential for a more unified and versatile AI landscape. It will end with domain adaption methods, especially Adapters [12, 73, 111, 26], that build up on foundation models [8] to adapt these models to special domains. This section aims to provide a solid foundation for understanding the key deep-learning architectures used in this thesis and their respective contribution.

2.2.1. Classification

Deep learning architectures employed for classification tasks are typically designed with a number of output neurons corresponding to the number of classes. Each neuron is associated with one output label, and the outputs represent the activations of these neurons [27]. The softmax function,

$$softmax(z_i) = \frac{e^{z_i}}{\sum j = 1^K e^{z_j}}$$

is frequently used to convert the outputs into class probabilities. A common loss function in this context is the Cross-Entropy Loss, which quantifies the dissimilarity between the predicted probabilities and the true class labels. The objective is to minimize this difference during the model training process. Cross-entropy loss is particularly well-suited for models that generate probability distributions [27].

In such models, the targets are one-hot encoded, meaning that a target t_i corresponding to input x_i is a vector in $\mathbb{R}^{|K|}$, consisting of $|K| - 1$ zeros and a single 1 at the position that corresponds to the label. For example, given the classes "cat", "dog", and "bird", a one-hot encoded target for a sample of a dog would be $t_{\text{dog}} = [0, 1, 0]$, with the second position in the vector representing "dog". The Cross-Entropy Loss penalizes activations at the first and last positions, thereby encouraging the output of the second position to approximate 1.

The equation for cross-entropy loss is given by:

$$L_{CE} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c})$$

where L_{CE}, with the number of classes C, the number of samples N, $y_{i,c}$ as the true label of

the i-th sample for class c, and $p_{i,c}$ as the predicted probability of the i-th sample belonging to class c, By minimizing the cross-entropy loss, a model learns to generate accurate probability estimates for each class, thereby enhancing classification performance [27, 7].

2.2.2. Natural Language Processing

NLP is a subfield of artificial intelligence that focuses on the interaction between computers and humans through natural language. This includes tasks such as text classification, named entity recognition, sentiment analysis and machine translation. Before the advent of deep learning architectures, oftentimes statistical and/or frequency-based methods such as the Bag-of-Words or tf–idf [83] were used. Even though these "traditional" methods can perform a variety of tasks, they lack in-depth representation and the perception of the subtleties of language. This includes multiple meanings, neo-composites and long-term dependencies. Computational advances and architectures improved these methods substantially in the last years, such that current state-of-the-art in almost every subdomain is achieved by deep learning architectures, which will be described in the following. As foundation, the next paragraph describes the process of tokenization, an important preprocessing technique to further process texts in neural networks.

Tokenization

Tokenization refers to the process of converting textual information into a format suitable for machine learning algorithms and deep neural networks. In this context, a token can represent a character, a subword, or a word. One of the major challenges when working with text is the near-infinite number of possible tokens, rendering it infeasible to store them all in a vocabulary. A vocabulary, in this context, denotes a set of known words for an algorithm. When a vocabulary does not encompass all possible words, unknown words (out-of-vocabulary words) are disregarded, which can render input sentences or entire texts unintelligible. Utilizing characters as vocabulary is feasible and circumvents this problem, as every word can be constructed from a language's characters. However, a word conveys more than just the sum of its parts; thus, working at the character level omits crucial aspects of language [7]. To strike a balance between capturing the distinct meaning of a word and the impracticality of storing all possible words, Byte Pair Encoding (BPE) [25, 85] is widely employed. BPE is a data compression algorithm repurposed for NLP as a subword tokenization technique. It addresses the issue of out-of-vocabulary words by decomposing text into subword units, allowing for a more efficient representation of rare or unseen words. BPE functions by iteratively merging the most frequent pairs of characters or character sequences in the training data, constructing a vocabulary of subword and word tokens. During tokenization, text is divided into the longest possible subwords present in the created vocabulary. By utilizing BPE, models can accommodate a wide range of linguistic variations while maintaining a manageable vocabulary size.

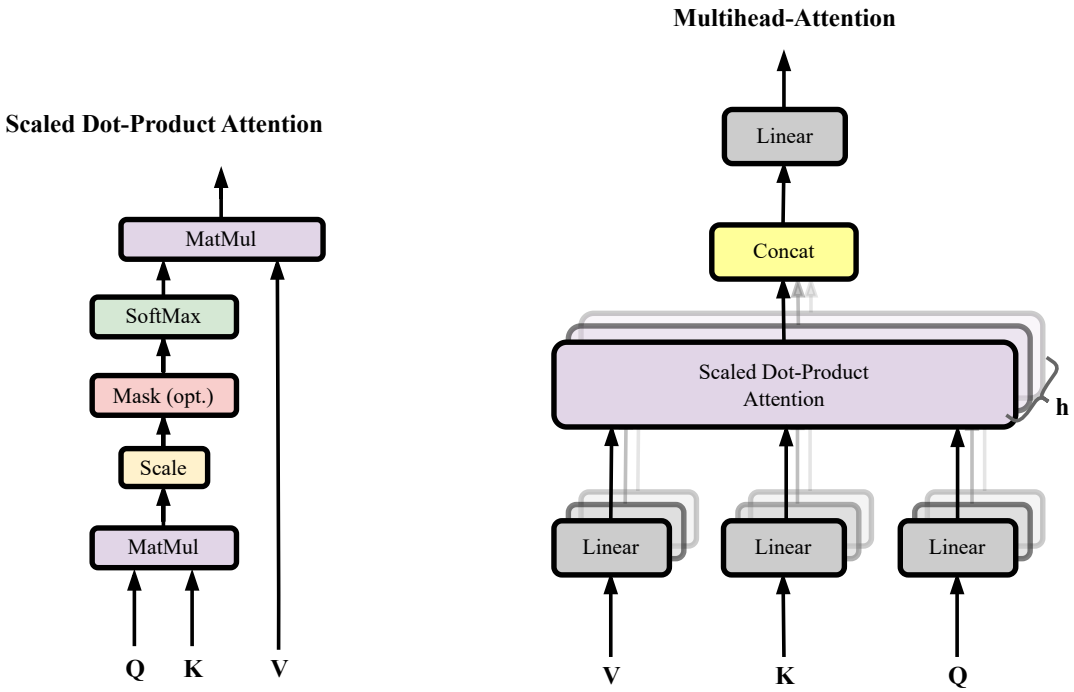


Figure 2.1.: From [95]: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

Recurrent Neural Networks

Recurrent Neural Networks (RNNs) [92] are a class of neural networks designed to handle sequential data by maintaining an internal hidden state that can capture information from previous time steps, thus addressing the issue of capturing long-term dependencies. RNNs are particularly well-suited for NLP tasks due to their ability to process variable-length input sequences, such as sentences or paragraphs. However, due to the design of RNNs, all previous information for an output at one timestep is stored in a single state, which creates a bottleneck in the architecture. Other variants of the RNN, e.g. the Long Short-Term Memory (LSTM) [40], and Gated Recurrent Unit (GRU) [13] tried to mitigate this problem.
RNNs have certain other limitations, such as difficulty in capturing long-range dependencies due to the vanishing gradient problem, which hampers their ability to learn complex language patterns. Moreover, RNNs are inherently sequential, making it difficult to parallelize computations and take advantage of modern hardware accelerators.

Attention Mechanism

The different attention mechanisms were introduced to address the limitations of RNNs, particularly their inability to efficiently capture long-range dependencies. The attention mechanism enables the model to assign different weights to parts of the input sequence according to their relevance to the current context, allowing the model to concentrate

on the most crucial information. Attention can be used in combination with RNNs, as demonstrated in the Seq2Seq model with attention, which has been extensively applied to tasks such as machine translation and summarization. Unlike RNN, the input is not processed sequentially but entirely, up to a model-dependent maximum sequence length. Textual input is encoded using BPE to account for the position, which is automatically handled in RNNs due to their sequential processing. A positional encoding is added to represent different positions. Consequently, the same word at the first position has a distinct representation compared to the same word at the last position. This approach allows the model to learn the sequential aspects of language effectively.

Scaled Dot-Product Attention

Figure 2.1 illustrates the two different forms of attention used in most cases. The inputs are called queries (Q), keys (K) and values (V), which are learnable linear projections from the output of the last layer or the input encodings combined with positional encodings, if it is the first attention block.

On the left, the Scaled Dot-Product Attention, which computes attention for queries Q and key-value pairs K and V with dimensions d_x as :

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V.$$

Q, K, V are all derived using three separate linear layers from the input of the same input, either the linear projected output from the previous block. Both are vectors of the same size, called d_{model}

Multihead-Attention Modern Transformer use Multihead-Attention, which basically splits the calculation of using Scaled Dot-Product Attention with d_{model} dimensional keys to h (the *Heads* of a Transformer model) different, learnable linear projections to d_k, d_k and d_v dimensions. Each of those projections can now attend to the input in parallel. In the end, the h outputs are concatenated and fed into the next attention layer via a linear projection. Formally, the Multihead-Attention is

$$\begin{aligned}\text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \\ &\text{with head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)\end{aligned} \tag{2.3}$$

where $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$ are the parameter matrices with learnable weights.

Transformers

The Transformer model [95] represents a paradigm shift in Natural Language Processing (NLP), as it entirely replaced recurrent structures with attention mechanisms. Transformers offer the advantage of being highly parallelizable, enabling efficient training on

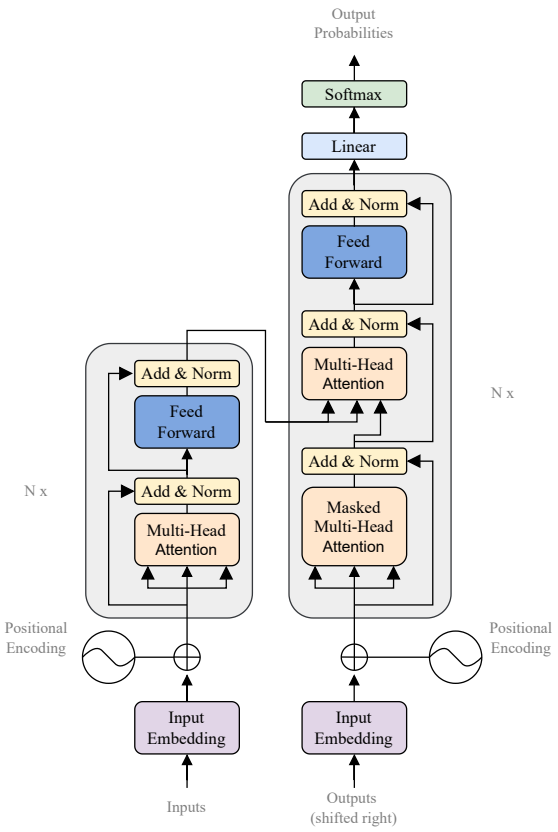


Figure 2.2.: From [95]: The Transformer - model architecture

large-scale datasets and significantly improving performance across a wide range of NLP tasks. The advent of the Transformer architecture led to the development of pre-trained language models, such as BERT [17], GPT models like the recently published GPT-4 [70], Llama [94], and many others. These models serve as foundation models [8] and have task-adapted, specialized successors and variations that cater to nearly every aspect of NLP.

The original Transformer proposed by Vaswani et al. [95] consists of an encoder-decoder architecture, with each component composed of multiple layers of multi-head self-attention and feed-forward neural networks. The encoder-decoder structure is visualized in Figure 2.2. However, in practice, both components are also used independently. This thesis relies on architectures that use only the encoder (e.g., Clip's vision and text encoder) or the decoder (e.g., ZOC's caption generator).

Using an encoder-only Transformer model is straightforward: after the last encoder block, the output is not fed into the decoder structure but to a task-specific final layer, such as for text classification. Task-specific loss is applied and the model weights are updated according to the specific optimizing process. For decoder-only models like GPT [75], the second Multi-Head Attention block, visible on the right side in Figure 2.2, which receives key-value pairs from the encoder output, is removed. The model takes the start input, called the prompt, and generates one token. This token is appended to

the prompt and fed back into the decoder to generate new output, a process known as auto-regression [30].

For other applications, it is possible to still use the Multi-Head Attention block without input from an encoder, but with other features, such as image features. In this way, a decoder can learn to generate textual output based on image features. Instead of using a prompt as the starting point for auto-regressive generation, only the image and the start token are fed into the decoder. Subsequently, at each step t, the output from step $t-1$ is fed back into the model.

2.2.3. Computer Vision

CV is a field of artificial intelligence that focuses on enabling machines to interpret and understand visual information. Traditional methods focus on line- and edge detection algorithms such as the Canny-Edge detector [11] and feature detection algorithms such as SIFT [62]. With the emergence of deep learning, the landscape of CV has undergone a significant transformation, leading to remarkable progress in various tasks, such as image classification, object detection, and semantic segmentation. Deep learning architectures for CV have evolved over time, starting with CNNs [53], followed by the development of Residual Networks (ResNets) [33], and more recently, the introduction of Vision Transformers [20].

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a class of neural networks designed to process grid-like data, such as images, by exploiting the inherent spatial structure of the input. A CNN consists of multiple layers, including convolutional layers, pooling layers, and fully connected layers. The convolutional layers are responsible for extracting local features from the input, while the pooling layers reduce spatial dimensions, and the fully connected layers perform high-level reasoning.
A convolutional layer consists of multiple learnable filters. Usually, these filters are of squared size (e.g. 3×3) which slide over the input image from the top left to the bottom right and multiply the pixel values at each position with the filter kernel values. The results are the sum of all multiplication results and process the whole image by sliding over the input. This way, an activation matrix for the filter is created. By adding multiple layers with different filter sizes, oftentimes increasing with depth, features of different sizes and shapes can be captured. One of the pioneering CNN architectures is the LeNet-5 [54] for digit recognition. Subsequently, the AlexNet [49], which is one of the first deep convolutional networks, achieved a breakthrough performance in the ImageNet Large Scale Visual Recognition Challenge [79] and sparked renewed interest in deep learning for computer vision.

Residual Networks

Despite the success of CNNs, increasing the depth of these networks led to issues such as vanishing gradients and degradation of performance, similar to the earlier described RNNs. ResNets [33] addressed these problems by incorporating residual connections or skip connections. In the standard formulation, these connections add the unchanged input l_i of layer f to the output of the layer. The formulation of a residual connection is simply

$$\text{res}(x) = f(x) + x.$$

These connections were already known from RNN architectures like the LSTM and allow the gradients to bypass certain layers, enabling the training of much deeper networks without suffering from the vanishing gradient problem. The function $f(\dot{})$ of ResNets usually consists of three convolutional layers, batch normalization layer and activation functions and the residual connection. These layers can be stacked up to at least 200 layers [5].
The ResNet50 (for 50 layers) became also one of the most used benchmarks for feature extraction and image classification in Computer Vision and is up to date used as a baseline in numerous works. Even though there are now superior architectures [20], ResNets are again the focus of research and still show very good performances in various tasks [59].

Vision Transformer

The Vision Transformer (ViT) [20] represents a shift in computer vision by adapting the Transformer architecture, originally propesed for NLP, to image data, also, with great impact. A major challenge for the usage of Transformer architectures was the size of input tokens: As sentences, or short text usually consist of a few up to hundreds of tokens (the words), a 224×224 RGB image, which is quite small, consists of about 150.000 token. Transformer are not suitable for such big inputs due to the quadratic runtime complexity of the self-attention mechanism. Therefore, the input image is processed in so-called patches, e.g. 32×32, which is usually indicated by a number after the vision transformer in research (e.g. ViT-B/32 for a Clip model with patch size 32). That means, a patch of 32×32 pixels is treated as one token for the self-attention mechanism, which reduces the tokens to $N = HW/P^2 = 49$ non-overlapping patches for height $H = 224$ and width $W = 224$. These are linearly embedded into a flat vector. These vectors, along with position embeddings, are then processed by the Transformer model.
Vision Transformers have shown competitive performance compared to CNNs on various benchmarks, such as ImageNet, suggesting that the self-attention mechanism can be effectively applied to visual data as well.

In conclusion, deep learning architectures for computer vision have evolved from CNNs to ResNets, culminating in the recent emergence of Vision Transformers. Each stage of de-

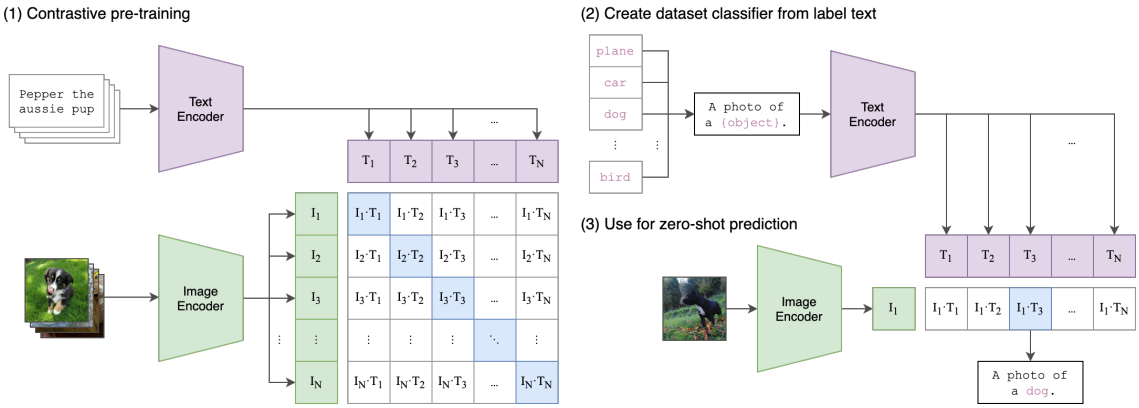


Figure 2.3.: Figure from [74]: Summary of Clips approach. It jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

velopment has brought advancements in modelling capabilities and performance, paving the way for increasingly sophisticated computer vision applications.

2.2.4. Clip

Clip [74] is a neural network architecture that jointly learns visual and textual representations by leveraging the power of Transformers [95] and contrastive learning [69]. Clip is designed to understand images and their semantic context within natural language descriptions. The model is pre-trained on a large dataset of 400 million images text pairs from the web, allowing it to perform well on various tasks, such as zero-shot image classification, object detection, and image captioning, without requiring task-specific fine-tuning. Clip consists of two primary components: a vision encoder and a text encoder. The architecture is also illustrated in Figure 2.3.

The vision encoder is a CNN or a ViT that processes input images and generates visual feature representations. The text encoder is a Transformer model that processes textual input and generates contextualized word embeddings. During training, the model forces both encoders to maximize the similarity between matching text image pairs while reducing the similarity to other pairs. This is achieved using the contrastive learning approach with InfoNCE loss [69]. As similarity metric, Clip uses cosine similarity. Notably, even though the model could use pre-trained weights, due to the large training set Clip is able to train all weights of both encoders from random initialization. The model is trained to correctly match images and their textual descriptions, learning meaningful joint representations that can be transferred to a wide range of downstream tasks, such as zero-shot classification or linear probing for downstream tasks.

Zero-Shot Transfer

Zero-shot transfer in machine learning refers to the ability of a model to perform well on tasks or classes it has not seen during training, by leveraging its knowledge learned from related tasks or classes. This ability is essential for achieving generalization in real-world scenarios, where data for all possible tasks or classes may not be available during the training phase [74].

An early approach to achieve zero-shot transfer is attribute-based zero-shot learning, where a semantic relationship between seen and unseen classes is established using attributes, allowing the model to make predictions for unseen classes based on their attributes [51]. A significant advancement was the introduction of word embeddings, particularly Word2Vec [63] and GloVe [72]. Socher et al. [88] demonstrated the possibility of using these embeddings for zero-shot transfer. Following the introduction of the Transformer [95] and the accompanying large-scale pre-trained language models such as BERT [17] and GPT [75], zero-shot transfer significantly improved in Natural Language Processing. The text features these models produced were so powerful that simply adding (and fine-tuning) a task-specific head was sufficient to approach State-of-the-Art (SOTA) in many different tasks [17]. Recently, with the advent of Vision Language Models such as Clip [74], language and vision understanding have been combined, enabling zero-shot transfer across a wide range of tasks. These models learn a joint representation space for images and text, allowing them to perform zero-shot transfer for tasks like image classification, object detection, and more. Furthermore, the joint feature space enables zero-shot classification, a true zero-shot methodology that eradicates the need to train a task-specific head. This functionality has also been transferred to other domains, such as OOD detection.

One of the key advantages of Clip is its ability to perform zero-shot transfer given only an image and a set of candidate textual labels. Clip can perform classification by ranking the labels based on their cosine similarity to the image's visual representation. This is also illustrated on the right side in Figure 2.3. The classification effort is thus reduced to encoding the image with the encoder and a one encoding a batch of textual labels with the text encoder and a subsequent matrix multiplication of the normalized features. Note that for normalized embeddings the cosine similarity is equal to the dot product of vectors.

This zero-shot capability stems from the model's pre-training on a diverse dataset, which enables it to learn rich visual and textual representations that are applicable across various tasks and domains [74]. This demonstrates the potential for more versatile and robust AI systems and is therefor called a foundation model [8].

2.2.5. Adapters

Adapters [41] are a parameter-efficient technique for fine-tuning deep learning models, preserving the original model's weights while learning task-specific information through the addition of a small number of trainable parameters. Adapters have been successfully applied to Transformer models in natural language processing, such as BERT [17], to enable efficient transfer learning. Integrating adapters into the Clip model can further improve its performance on downstream tasks while maintaining the benefits of parameter efficiency and task-specific adaptation.

There are two types of adapters for Clip. The first is identical to that used in NLP Transformer architectures [41], and recently also for vision transformers [12]. In the context of Clip, these adapters can be integrated into both the vision and text encoders individually. The adapter module is typically inserted after each layer of the original architecture, consisting of a small feed-forward neural network with linear layers. The adapter module also includes layer normalization and skip connections to maintain the original information flow. To fine-tune the full architecture to a task, all weights of the original Clip model are frozen and only the weights of the adapters are trained. The second type of adapters are not plugged into the transformer architecture of the encoder but are external small feed-forward structures which interact with the encoder outputs of the model [111, 26]. These adapters are specifically designed to serve as few-shot domain adaption modules. TIP-Adapter [111] is a recent strategy with state-of-the-art results in few-shot domain adaption and used in this theses for such tasks.

TIP-Adapter

Figure 2.4 shows a detailed illustration of TIP-Adapters [111]. The model uses a so-called cache model of a few-shot training set as domain knowledge, which is ultimately added up to Clip's "general" knowledge using only matrix multiplication and addition. Given a pre-trained Clip model and a dataset of images I_K with K images for each of the N classes and corresponding labels L_N. The adapter is a key-value cache created from this K-shot training set containing knowledge of every of the N classes. For each image, the knowledge is the $L2$-normalized C dimensional feature obtained by Clips vision encoder together with a N-dimensional One-Hot encoded class label. For the NK training samples, the visual features are called $F_{\text{train}} \in \mathbb{R}^{NK \times C}$, the keys of the cache model and the corresponding labels $L_{\text{train}} \in \mathbb{R}^{NK \times N}$, the values. Formally noted,

$$F_{\text{train}} = \text{VisualEncoder}(I_K),$$

$$L_{\text{train}} = \text{OneHot}(L_N),$$

Both together form the key-value cache, where F_{train} are treated a keys and L_{train} as

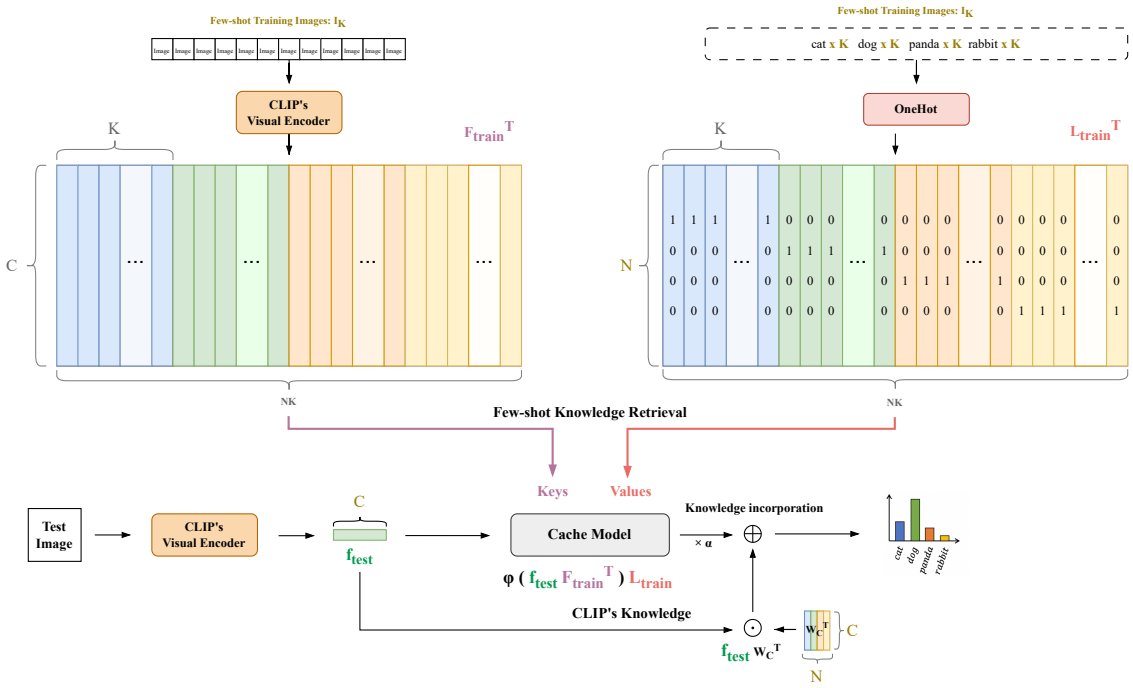


Figure 2.4.: From [111]: Given a K-shot N-class training set, we construct a cache model to adapt Clip on downstream tasks. It contains few-shot visual features F_{train}^T encoded by Clip and their ground-truth labels L_{train}^T under one-hot encodings. After retrieval from the cache model, the few-shot knowledge is incorporated with Clip's pre-trained knowledge, achieving the training-free adaption

values. To further improve the classification abilities, the cache model can be treated as initialization point for a trainable linear layer that can be fine-tuned via SGD [78] to surpass the frozen cache model. During fine-tuning, the Tip-Adapter's predictions are supervised with few-shot training data and cross-entropy loss, updating the weights in the cache model. The key weights are unfrozen, while the value weights and the two encoders in the Clip model remain fixed. This approach allows adaptive affinity estimation and boosts distance calculation between training and testing images in the embedding space. The Tip-Adapter requires only a small number of epochs for fine-tuning and achieves strong performance with fast convergence and limited resources.

For accessing the knowledge during inference with a test image, first, the Clip features $f_{\text{test}} \in \mathbb{R}^{1\times N}$ are obtained by querying the vision encoder. Now, the features are used as query to calculate the affinity A between the test image and the few-shot cache model. The affinity is

$$A = exp(-\beta(1 - f_{\text{test}} F_{\text{train}}^T)),$$

where β is a *sharpness* hyperparameter, which controls the influence of low-similarity training samples on the output. The prediction of the cache model is obtained via linear combination of cached values weighted by A as $AL_{\text{train}} \in \mathbb{R}^{1\times N}$.
Given the Clip zero-shot prediction, which is the matrix multiplication of the normalized

class features W_c^T and f_{test}, so $\text{pred}_{\text{Clip}} = f_{\text{test}} W_C^T$ and AL_{train}, the predictions (logits) of the whole model, and α, the parameter controlling the residual ratio, are given by

$$pred_{\text{TIP}} = \alpha AL_{\text{train}} + \text{pred}_{\text{Clip}} \tag{2.4}$$

The key aspect, why TIP-Adapter where chosen in favor of other few-shot fine-tuning methods, is that in Equation 2.4 the Clip prediction, especially those of the visual encoder, are unchanged and are thus suitable to be used in Zero-Shot Out-of-Distribution Detection methods (see Section 4.3 for details) without changing any of the pre-trained models.

2.3. Out-of-Distribution Detection

This section introduces Out-of-Distribution Detection (OOD detection) and introduces relevant terms and concepts. It should be noted in particular that the definitions of OOD detection are not uniform and can vary slightly. In this work, the common definition and delimitation used in current surveys on the topic [108, 82] will be used. This section is structured as follows: first, key terms are defined and the background to this work is explained. Finally, OOD detection is classified in more detail and distinguished from related topics. We refer to In-Distribution (ID) as each image associated with a known label and Out-of-Distribution (OOD) as each image not associated to a known label.

Closed-World Assumption

The closed-world assumption [22, 108], also known as closed-set assumption [108, 65], is a principle in machine learning that assumes all possible information about a system is known and available to the learning algorithm. This means that the algorithm does not consider any information that is not contained in the training data. In other words, the algorithm assumes that all relevant information is already present in the data, and does not attempt to reason about information that is missing or unknown. While this assumption can simplify the learning process and make it more efficient, it can also limit the generalizability of the model to new situations where relevant information may be missing or unknown. Closely related is the term **closed-world classifier**, which describes the classifier trained on the closed set of labels.

Outlier Exposure

Outlier exposure [38] is a concept in machine learning that refers to the intentional exposure of a model to outlier data points during training. This approach is often used to improve the model's robustness to outliers, which are data points that deviate significantly from the majority of the data. By exposing the model to outliers during training, it can learn to identify and handle them better during inference, leading to more accurate

predictions on real-world data. However, the effectiveness of outlier exposure depends on the type and distribution of outliers in the data, as well as the specific modelling techniques used.

Task Difficulty

Evaluating and defining the difficulty of an Out-of-Distribution Detection (OOD detection) task in machine learning involves assessing various factors that influence the complexity. Aspects to consider when evaluating the difficulty of an OOD detection task include:

- Distribution overlap: If the ID and OOD class samples have significant overlap in their feature space, it becomes challenging to distinguish between them. The more distinct the distributions, the easier a task becomes [36].
- Dimensionality: High-dimensional data spaces make it difficult to detect ID samples for multiple reasons. Firstly, the curse of dimensionality makes it difficult to detect outliers by distance [7]. Secondly, as the dimension of the embedding space increases, the features become more uniformly distributed [98]. Consequently, the overlap between ID and OOD samples increases.
- Data complexity: The complexity of the data, such as variations in texture, colour, shapes, and sizes of the objects, can affect the difficulty. More complex data may require more advanced techniques or larger models to distinguish between ID and OOD samples [38].

Aside from these factors, there may be additional considerations. In the following section, an overview of methods that attempt to assess difficulty is presented.

The most widely used metric is the Openness Score [84], which quantifies the degree to which a recognition system is exposed to unknown classes, providing a measure of how well the system can handle new or unseen classes during testing. Given the seen training classes C_{train}, the seen testing classes C_{test}, and the target classes C_{target}, which is the total number of classes to be identified (i.e., all seen and unseen classes), the openness is defined as:

$$\text{openness} = 1 - \sqrt{\frac{2 \times |C_{\text{train}}|}{|C_{\text{test}}| + |C_{\text{target}}|}}.$$

As the definition indicates, defining openness requires some degree of outlier exposure. Therefore, the metric is not suitable for predicting difficulty or performance in real-world scenarios. Although openness and performance are strongly correlated [84], the score does not consider visual similarity that can exist between classes. Consequently, different datasets with identical openness may yield different performances. Nonetheless, openness is still in use and helps define difficulty to some extent, as it is easier to detect

IDsamples from one outlier class (low openness) and thus distribution, or from many other distributions (high openness).

A commonly used distinction made in OOD detection is between near and far outliers [104]. Near-OOD describes outliers, that are *close* (semantic, statistically, etc) to the ID data while far-OOD describes the opposite. Both have distinct features and challenges, but in general, near-OOD is perceived as the harder challenge [104, 82]. Nevertheless, both are active research fields, because while most near-OOD methods also do work for far-OOD (vice versa is not necessarily the case), the latter can be detected with methods that need less resources or have other benefits.

Out-of-distribution Detection

Out-of-Distribution Detection (OOD detection) is a critical problem in machine learning that refers to the task of identifying samples that are significantly different from the training data distribution. This problem arises in many real-world scenarios, such as detecting unseen data or identifying anomalous samples that can cause errors or unexpected behaviour in the system.

One popular approach for OOD detection is based on the outlier exposure principle, which suggests training the model on both in-distribution (ID) and out-of-distribution (OOD) samples. This approach exposes the model to a wide range of data, including samples that are significantly different from the training data distribution, thereby improving the model's ability to identify OOD samples.

However, this approach assumes a closed-world assumption, which means that the OOD samples are drawn from a known distribution that is distinct from the ID distribution. In practice, this assumption may not hold, as the OOD samples may come from an unknown distribution that is similar to the ID data, making them difficult to detect using traditional OOD detection methods.

Several novel techniques for OOD detection have been proposed that do not rely on the closed-world assumption. These methods include density-based approaches, such as using a density ratio to distinguish between ID and OOD samples, e.g. [1, 81] (157-161), and deep generative models, which can learn to generate samples from the ID distribution and identify OOD samples by measuring their distance to the learned distribution, e.g. [16, 46].

Overall, OOD detection is a challenging problem in machine learning, and developing effective methods for detecting OOD samples is critical for ensuring the reliability and safety of machine learning systems in real-world applications.

2.3.1. Differentiation from Related Topics

Figure 2.5 shows an overview of related topics and classifies them by four bases: The shift to detect (covariate, semantic), the ID data type (single class, multiple classes), whether the ID data needs to be classified and transductive vs inductive learning. Transductive

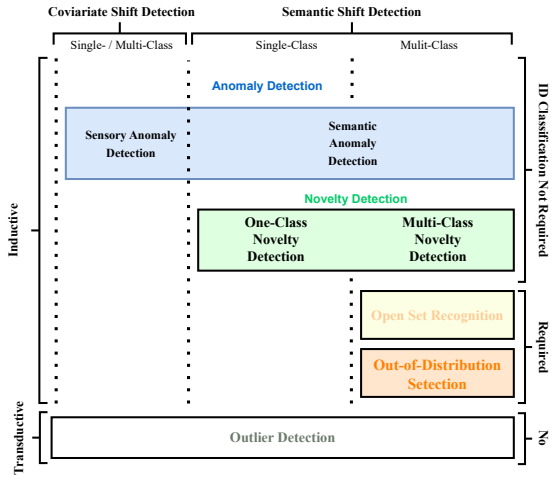


Figure 2.5.: Taxonomy of the generalized OOD detection framework by [108]: Illustrated by classification tasks. Four bases are used for the task taxonomy: 1) Distribution shift to detect: the task focuses on detecting covariate shift or semantic shift; 2) ID data type: the ID data contains one single class or multiple classes; 3) Whether the task requires ID classification; 4) Transductive learning task requires all observations; inductive tasks follow the train-test scheme. Note that ND is often interchangeable with AD, but ND is more concerned with semantic anomalies. OOD detection is generally interchangeable with OSR for classification tasks

means, that all samples are available at training time, thus this includes methods with outlier exposure. In the following, the two most important differentiations are briefly explained, the distributional shifts and the classification objective.

Covariate Shift and Semantic Shift

Semantic shift in context of OOD detection refers to a change of the labels and the related concepts of the instances of these labels. Consider the ID data to be classes of different birds. A semantic shift would mean the occurrence of any other class that is not a bird. Also, the occurrence of any bird class, which is not in the training dataset, is a semantic shift.

Covariate shift, on the other hand, refers to a change in the distribution of input variables that can affect the performance of a machine learning model. For example, if a model is trained on data from one geographical region, but is applied to data from a different geographical region, there may be differences in the distribution of input variables (smaller or bigger birds, colour differences) that can lead to reduced accuracy or precision in outlier detection. Covariate shifts can also arise when there are changes in the data collection process, such as changes in sensor calibration or sampling rates. Covariate shift is commonly more used to evaluate model generalization than robustness [108].

Out-of-Distribution Detection addresses semantic shift.

Classification objective

The classification objective separates anomaly detection and novelty detection from OOD detection (see Figure 2.5). In anomaly detection, the classification of the ID data is not part of the objective. It is a technique used to identify data points that are significantly different from the majority of data points. It is typically used in unsupervised learning, where the algorithm learns to identify patterns and relationships in the data without being explicitly trained on what constitutes an anomaly. Anomaly detection is used to identify rare or unusual events, such as fraud detection, network intrusion detection, or equipment failure prediction.
OOD detection and also Open Set Recognition [84] are methods, that should recognise if a given input is outside the distribution of the training data. It is commonly used in supervised learning, where the algorithm has been trained on a specific set of inputs and outputs. Both are used to identify when the model is given inputs that are unlike anything it has seen during training. This can help prevent the model from making incorrect predictions or providing unreliable results.

Out-of-Distribution Detection and Open Set Recognition

As stated in Figure 2.5, the terms are often used interchangeably or described with identical properties [104, 21, 108, 6]. In Yang et al. [108] the difference and similarity is described as "[OOD-Detection] canonically aims to detect test samples with semantic shift without losing the ID classification accuracy. However, OOD detection encompasses a broader spectrum of learning tasks and solution space." This can be interpreted as models that do detect outliers by using the classifier's output confidence [36, 56].

2.3.2. Zero-Shot Out-of-Distribution Detection

Zero-Shot Out-of-Distribution Detection (Z-OOD detection) describes models with the ability to deliver robust performance and generality without relying on training using ID samples [65, 21]. Unlike traditional OOD detection methods, which often necessitate training from scratch or fine-tuning on a specific ID dataset, Z-OOD detection capitalizes on pre-trained models to accomplish two goals: 1) accurately classify test samples from seen classes and 2) detect samples not belonging to any of the seen classes. This is achieved using only the names of the seen classes, without any training data or the need for building a closed-world classifier. As of now, in the research area of image classification, there are two published methodologies, which both rely on the large-scale pre-trained multi-modal model Clip, which enables zero-shot classification [74]. Thus, the methods rely not only on images but also on textual labels. All Z-OOD detection methods do rely on this functionality or on the functionality on related multi-modal models.

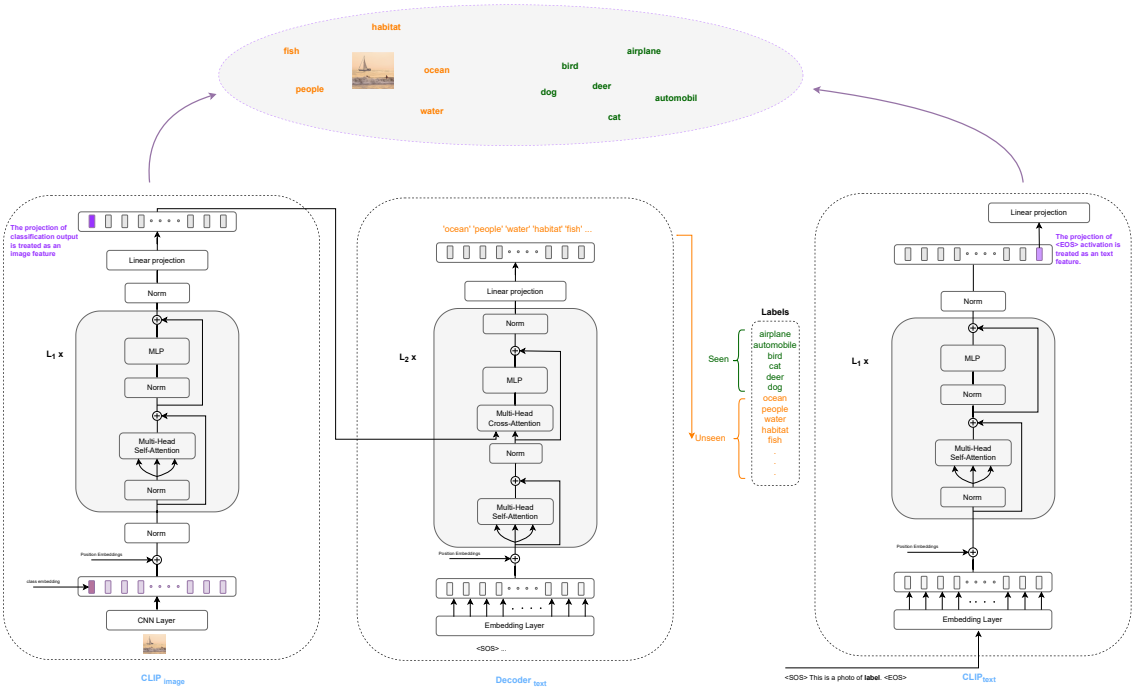


Figure 2.6.: From [21]: The diagram illustrates the inference steps of ZOC for a sample from an unseen class 'boat'. The available seen class labels (shown in green) are $Y_s = \{$'airplane', 'automobile', 'bird', 'cat', 'deer', 'dog'$\}$. In the first step, the image is encoded through $\text{Clip}_{\text{image}}$ and then image description is generated in the output of $\text{Decoder}_{\text{text}}$. The description is in fact a set of candidate unseen labels Y_u. In the second step, $Y_s \cup Y_u$ are encoded through $\text{Clip}_{\text{text}}$ on the right. The purple ellipsoid shows Clip's feature space where the relevant labels are aligned with the image. Clip quantifies the alignment by calculating the cosine similarity of each encoded label to the encoded image. Then $S(x)$ is obtained according to Eq. 2.5. The score is high for this image as it is more similar to the set of Y_u than Y_s. The inference relies on Clip pre-trained encoders as well as Y_u generated by $\text{Decoder}_{\text{text}}$ (best viewed in color)

ZOC

The Zero-Shot Out-of-Distribution Detection based on Clip (ZOC) [21] is a recent OOD detection that distinguishes itself among numerous methods (see Chapter 3) as the first published technique to detect outliers using solely general pre-trained models, without task-specific training. This approach is referred to as Zero-Shot Out-of-Distribution Detection (Z-OOD detection). The method employs Clip's zero-shot classifier to measure the similarity of a given input image to known labels and a set of image-specific labels. These unseen labels, denoted as Y_u, are generated for each image using a description generator based on the image's Clip features. The known labels Y_s along with unseen labels Y_u, are used to calculate class probabilities using Clip 's zero-shot classifier and the softmax function.

Decisions about whether a given image is an outlier or an ID image are made based on these probabilities. The model can generate meaningful labels for test images without

any outlier exposure, effectively utilizing zero-shot classification for outlier detection. Figure 2.6 illustrates this process in detail. The underlying intuition is that if an image is ID, there exists a semantic label describing a discernible object within the image (e.g., a dog for the label 'dog'). Clip 's zero-shot classifier generates features for the label and the image with high similarity. The description generator creates labels based on the image, but if the label 'dog' is already present in the seen labels Y_s, it will be removed. As a result, no unseen label in Y_u should be as similar to the image as the existing label 'dog'. Conversely, if the input is an OOD image (e.g., a pig) without a matching label in Y_u, the most similar label will be among the generated labels.

Architecture: The model comprises a Clip [74] model, which includes a trained vision encoder $\text{Clip}_{\text{vision}}$, text encoder $\text{Clip}_{\text{text}}$, and a description generator Decodertext. This description generator is a Transformer-Decoder that generates image descriptions based on image features from Clipvision. Image descriptions can be generated by extracting image features from the vision encoder and feeding them through the caption generator. Consequently, a description I_{TEXT} for an image I_{test} is generated as follows:

$$I_{\text{TEXT}} = \text{Decoder}_{\text{text}}(\text{Clip}_{\text{vision}}(I_{\text{test}})).$$

The description is split into individual words, which serve as generated unseen labels Y_u. The complete set of labels for an image is given by $Y = Y_s \cup Y_u$. By passing each label through $\text{Clip}_{\text{text}}$, the features are generated as $W_{\text{test}} = \text{Clip}_{\text{text}}(Y)$. Further details on the caption generator's training and the model are provided in Chapter 4.

Next, the cosine similarity between image features and all labels $Y_{\text{full}} = Y_s \cup Y_u$ is calculated, and a softmax score is computed over the similarities with C_{zoc}. Instead of employing the maximum softmax score $\max(\hat{y}_1, \ldots, \hat{y}_k)$, which is standard for OOD detection [36], ZOC utilizes the summed probability of the generated labels. Formally, the ZOC score for an input image I is defined as follows:

$$\text{score}(I) = \sum_{c \in C_{\text{gen}}} p(y_c|I). \tag{2.5}$$

In summary, ZOC leverages the generalization capabilities of recent transformer architectures, combined with the vision-language gap bridging provided by Clip, to create a form of outlier exposure with generated labels that do not contradict real-world settings, where outliers cannot be provided[1]. This is unlike other work with outlier exposure, as the unseen labels here are not derived from predefined OOD content but are image-specific. This approach is theoretically applicable to numerous scenarios where acquiring training data is challenging, making it a versatile and innovative solution for OOD detection.

[1]Outlier exposure implicitly assumes knowledge, what incorrect input will be given to a model

MCM

Maximum Concept Matching (MCM) [65] has been recently introduced as a zero-shot OOD detection technique, building upon the foundation of an earlier used baseline [21]. This method, also referred to as "MSP + Clip" there, integrates the Maximum Softmax Probability [36] with Clip's zero-shot classifier [74]. During inference, the image similarity to all known labels is computed, forming a closed-world classifier within the zero-shot classification paradigm.

The classifier's softmax output, $\hat{y}_k$, represents the probability that image I belongs to class c_k. A probability threshold t is determined such that if $\max(\hat{y}_1, \ldots, \hat{y}_k) < t$, the image is classified as OOD, otherwise as ID. In essence, if the classifier's confidence in the image belonging to a known class falls below a specific level, the image is deemed not part of the set of known classes..

A notable characteristic of this method lies in its end-to-end application speed. Assuming the model is deployed and zero-shot labels are pre-calculated, inference requires only a single forward pass through the image encoder and one matrix multiplication to produce output logits, subsequently utilized for softmax score calculation. Consequently, the method exhibits $O(n)$ complexity for inference and virtually no domain adaptation overhead for new data distributions. Furthermore, this approach significantly reduces the demand for training data, ultimately eliminating the need altogether.

Metrics

- AUROC: The Receiver Operating Characteristic [9] is a well-known criterion for OOD detection and the de-facto standard to measure performance [84, 82].
  The ROC is a graphical plot illustrating the relationship between True Positive Rate (TPR) and False Positive Rate (FPR) for different threshold values of the model. Formally, TPR and FPR are defined as follows:

$$\text{TPR} = \frac{\text{true positive}}{\text{true positive + false negative}}$$
$$\text{FPR} = \frac{\text{false positive}}{\text{false positive + true negative}}$$

  The Area Under Receiver Operating Characteristic (AUROC) is threshold-independent and defined between 1 and 0. 1 indicates a perfect detection, while .5 is the worst outcome for this scenario, as it is the result an uninformed guesser would reach, so no useful information is processed. Values $<$ .5 indicate that information is processed wrong, as the result is worse than guessing. In practice, that means the predictions can simply be flipped, thus the absolute difference from .5 can be considered relevant, however, that practice ignores that the model does something wrong with the given information, which should be understood and fixed instead of being

ignored. Thus, in this work, this practice will not be applied.

- FPR@TPR: The False Positive Rate (FPR) at the True Positive Rate (TPR) is a performance metric commonly used in the evaluation of binary classification models, thus also useful for OOD detection. The FPR measures the fraction of false positive predictions made by the model, while the TPR measures the fraction of true positive predictions. The FPR@TPR metric provides a way to visualize the trade-off between the two quantities and is often plotted on a ROC curve, which shows the relationship between the FPR and TPR as the classification threshold is varied.

- F1-Score: The F1 score is the harmonic mean of precision and recall. Precision is the fraction of relevant instances (True Positives, TP) among the retrieved ones, which are the TP and false negatives (FN). The Recall is the fraction of relevant instances that are retrieved by the model:

$$\begin{aligned} \text{F1} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\ \text{with Precision} &= \frac{TP}{TP + FP} \\ \text{and Recall} &= \frac{TP}{TP + FN} \end{aligned}$$

  The F1 score is a good overall measure of the model's performance, taking into account both false positive and false negative predictions. A high F1 score indicates that the model has a good balance between precision and recall, while a low F1 score indicates that the model has either a high false positive rate or a high false negative rate.

3. Related Work

This chapter discuss and present related approaches to those in this thesis with the goal to contextualize this work. First, it will provide an overview to Out-of-Distribution Detection (OOD detection) with a special focus on methods that use foundation models [8] and build up on those. The second part will focus on multi-modal models such as Clip [74] and adapters [41, 26] as domain adaption methods, which are of relevance to the third research question.

3.1. Out-of-Distribution Detection

The section starts with an overview of evaluation protocols for OOD detection, which enables the reader to understand further use of the terms performance, score and improvements regarding OOD detection, before presenting approaches to relevenat aspects of OOD detection. Existing OOD detection approaches differ in their use of outlier exposure, which refers to whether the model has access to any OOD samples during the training process, such as in [56, 37], or not, like in [21, 43]. This work primarily focuses on OOD detection without outlier exposure. From this perspective, approaches are presented that can handle scarce data (Few-shot OOD detection) and no data at all, called Zero-Shot Out-of-Distribution Detection (Z-OOD detection). Furthermore, methods for assessing the difficulty of OOD detection tasks are introduced, which are useful for improving the comparability of different techniques.
For a comprehensive overview of OOD detection strategies and methods, readers are referred to current surveys such as [82, 108].

3.1.1. Evaluation Protocols

Evaluating Procedure

In the field of OOD detection and related areas, such as Anomaly Detection and Open Set Recognition, there is a lack of standardization in the evaluation procedures. The CIFAR10 and CIFAR100 datasets [48] are commonly used, but they are not yet a standard dataset like ImageNet [49] is for image classification. The selection of ID and OOD classes is also not standardized. Some researchers split one dataset into ID and OOD classes, such that these classes are related [21, 96], while others use OOD classes that are self-defined as "close" to the ID data to make the task harder [104, 65]. Self-defined means, that there is no clear foundation used in the publication, even though Winkens et al. [104] provided the

CLP to define near- and far-ood. Some others create ID/OOD splits by choosing different datasets from a group of "semantic shift datasets" [96], which includes ImageNet as a large-scale evaluation for category shift and multiple fine-grained classification datasets, two of which (Caltech CUB and Stanford Cars) are also used in this research.
The general issue in the research on OOD detection is the infinite combinations of ID/OOD and each combination yields different properties that influence the performance. All combinations can occur in reality, so the selection of some combinations is a necessary design decision, but the significance and generalizability of the results to the whole research topic is challenging to estimate. It is open to question whether an improvement in a specific niche translates to the entire area. Nevertheless, a benchmark or more standardized testing procedure would be beneficial for comparing results. Summarizing, it can be said that the most published work either uses combinations of well-known datasets [65, 104] or explicitly decide to use a harder near-OOD setting [21, 24], where either close related classes from different datasets are used or the dataset is split into ID and OOD.

3.1.2. Out-of-Distribution Detection without Outlier Exposure

There are currently very few proposed methods for Zero-Shot Out-of-Distribution Detection (Z-OOD detection) without any data exposure, so with no trained closed-world classifier as well as no access to OOD data. Thus, a broader look at the topic is presented, and recent methods for OOD detection are included with ID exposure and trained closed-world classifiers.
A solid baseline for OOD detection is the Maximum Softmax Probability (MSP) [36], which follows the assumption that a trained closed-world classifier should be more confident predicting a class for ID data than predicting an erroneous class, as it is always the case for OOD data. Even though neural networks are sometimes overconfident when predicting erroneous classes, the general trend is that more confident predictions tend to be more accurate [36]. The experiments showed, that the confidence based approach combined with recent feature extractors, can outperform more sophisticated approaches.

Generalized ODIN (G-ODIN) [43] is based on ODIN [56]. However, it eliminates the need for OOD data in the training process. The authors argue that this step makes their model more generalizable, hence the name Generalized-ODIN. The original ODIN trains a vanilla classifier, f_θ, on D_{IN}, the In-Distribution data for a given task and uses a scoring function $S(X; f_\theta)$. This scoring function has parameters that are learned with OOD data. The scoring function applies temperature scaling [32] for a more realistic confidence estimation of the classifier's output. Additionally, ODIN utilizes specific input preprocessing to further enhance performance, which also relies on OOD data. G-ODIN enhances the temperature scaling methodology and the input preprocessing so that the learned confidence output of the network is not dependent on any OOD data but outperforms ODIN on every reported test [56].

Class Anchor Clustering [64] utilizes anchored class centers in the logit space, which encourages the formation of dense clusters around the seen classes (i.e. the ID data). The OOD samples are detected simply based on the distance to the class centers. This approach can also be easily improved by using more effective feature extractors, such as Clip, which shows better semantic feature extraction than the utilized feature extractor in the original paper.

In light of remarkable advancements in sophisticated methods for OOD detection, the findings of Vaze et al. [96] present an intriguing and potentially paradigm-shifting perspective. By employing state-of-the-art training strategies and scaling the confidence of the classifier, the researchers demonstrated that OOD detection scores based on the classifier's softmax output could match or even surpass several current state-of-the-art methods across various experimental setups. This evidence suggests that a well-performing closed-world classifier is not merely a baseline but, in fact, competes with the top-performing methods in the OOD detection domain.

In summary, several methods have been developed for OOD detection with varying degrees of reliance on OOD data exposure. While the MSP serves as a simple and effective baseline, more advanced techniques such as G-ODIN and Class Anchor Clustering aim to improve performance by refining confidence estimation or leveraging class center distances in the logit space, respectively. These methods can be further enhanced by employing more powerful feature extractors or adapting them for specific application scenarios. As research in this area continues to evolve, new techniques and approaches will emerge to address the challenges of OOD detection in different contexts, ultimately improving the robustness and reliability of classification models.

3.1.3. Few-Shot Out-of-Distribution Detection

Some noteworthy techniques in the domain of few-shot OOD detection include few-shot learning algorithms, such as prototypical networks [87], and meta-learning methods, such as Model-Agnostic Meta-Learning (MAML) [23]. These methods capitalize on the limited OOD data to adapt and generalize the model to new, unseen OOD samples. Furthermore, the incorporation of transfer learning and pre-trained models, like Clip [74], has shown potential for enhancing few-shot OOD detection performance. Despite the progress in this area, few-shot OOD detection remains an ongoing research topic, with the goal of developing more effective and robust techniques to handle the challenges posed by limited OOD data.

3.1.4. Zero-Shot Out-of-Distribution Detection

The development of zero-shot classifiers has significantly advanced with the introduction of the Clip model, which has enabled new techniques for OOD detection. Fort, Ren, and Lakshminarayanan [24] initially explored the potential of Clip -based zero-shot classifiers, demonstrating promising results in various experiments. Their research primarily focused on improving OOD detection performance through the incorporation of outlier exposure. Very similar work was done by Liznerski et al. [60], which main purpose was also to further dive into outlier exposure but also showed promising results with Z-OOD detection.

Following these initial efforts, Zero-Shot Out-of-Distribution Detection based on Clip (ZOC) [21] was proposed as the first true zero-shot method for OOD detection. Notably Ming et al. [65] later published a method called Maximum Concept Matching (MCM), that built upon the same MSP baseline and introduced temperature scaling, adopting the ideas from ODIN [104] to enhance OOD detection performance.

Currently, ZOC and MCM represent the two primary true zero-shot OOD detection methods in the field. Both approaches leverage the capabilities of Clip -based zero-shot classifiers while incorporating additional techniques to optimize their performance. As research in this area continues, further advancements in zero-shot OOD detection methods are expected, ultimately contributing to the development of more robust and accurate classification models.

3.1.5. OOD Detection Difficulty

Liang, Li, and Srikant [56] adapted the commonly used metric in statistics, the Maximum Mean Discrepancy (MMD), with a Gaussian RBF kernel [89, 31, 91]. MMD defines the distance between the ID and OOD distributions, thus assessing the difficulty of the task based on the visual content of the distributions. A high similarity (a low MMD) makes it more difficult for a machine learning system to discriminate between ID and OOD samples, and vice versa. Consequently, the MMD tends to be negatively correlated with OOD detection performance [56]. Although this approach demonstrates the correlation and is both intuitive and comprehensible, it has two limitations: First, applying the L^2 RBF kernel directly in the image space only identifies nearly identical images [104]. Second, when applying it to high-dimensional feature spaces, it will suffer from the curse of dimensionality, and the MMD will lose its significance. Especially in models trained with Contrastive Learning [69], such as Clip, the embeddings of all images tend to be very similar, further reducing the significance of the MMD.

To address the limitations of the MMD, Winkens et al. [104] propose the Confusion Log Probability (CLP) to measure the difficulty of OOD detection tasks. The CLP is based on the probability with which a classifier confuses ID and OOD samples. As a result, the

classifier also needs access to ID samples. In the original publication, an ensemble of five ResNet-34 [33] classifiers was trained individually on the union of all datasets used in their work. In their specific case, this led to training a 486-way classifier. An ensemble is chosen because of their well-calibrated predictions compared to single classifiers [50]. As the OOD detection score is often directly derived from a classifier's output, the metric is close to the problem, interpretable, and, like MMD, negatively correlated with the OOD detection score. Moreover, it does not suffer from the curse of dimensionality. However, as the example from [104] shows, it is computationally expensive to calculate, and there are design decisions that can directly influence the CLP, such as the choice of classification model, the number of models in an ensemble, learning rates, etc. The CLP is also an intra-task metric, meaning that it can only be used to compare task difficulty between different combinations in one setting, as the classifier needs to be trained on all classes. Therefore, adding and removing classes directly influences the CLP. With this method, a general CLP that quantifies difficulty comparably across different combinations and use cases is not possible.

Derived from CLP, the terms *near*-OOD and *far*-OOD were introduced [104]. A low CLP indicates near-ID, and a high CLP indicates far-OOD; near-OOD is perceived as more challenging. The terms are widely accepted and used, e.g., in [21, 76, 82], but often without the direct link to the CLP. Instead, they rely on the semantic definition Winkens et al. [104] provided: "[they] distinguish between near OOD regimes where inlier and outlier distributions are meaningfully similar, and far OOD regimes where the two are unrelated." The problem with this rather vague definition is that both near and far can be defined arbitrarily. Usually, splitting one dataset into an ID and OOD part is considered near-OOD. However, in many research datasets, such as CIFAR-10/100 [48], the classes are arbitrary ('airplane', 'dog', 'car', etc.). In contrast, some datasets have all classes sharing semantics, e.g., birds. Splitting different bird classes into ID and OOD results in very closely related distributions. Splitting the CIFAR datasets means splitting only loosely connected classes and is closer to a far-OOD setting ('dog' and 'airplane' are as close as 'dog' and 'rocket', for example). Thus, merely defining near-/far-OOD does not capture all the necessary information, and using the CLP would improve interpretability.

3.2. Zero-shot Transfer

The published Z-OOD detection methods all rely on zero-shot transfer from the multi-modal model Clip [74]. Even though Clip gained lots of attention in research, there are several other approaches and methods, which also allow zero-shot image classification and therefor might be suitable for Z-OOD detection. This section provides an overview of these methods and also shows alternative task and domain adaption methods, which enable the models to adapt to domains not yet covered by foundational knowledge.

3.2.1. Vision Language Models

In addition to the previously discussed Clip, there are several other vision-language models capable of performing zero-shot downstream tasks, and therefore, they are potentially suitable for Zero-Shot Out-of-Distribution Detection methods. However, to the best of knowledge, no method or adaptation has been published for these models in this context. Here, a brief overview of competing approaches in this area is provided. For a more comprehensive examination of current methods and models, a recent blog post [19] provides an comprehensive overview.
Locked-image Tuning (LIT) [110] is a method that employs contrastive training [44] to align image and text models while preserving the benefits of their pre-training. The authors find that the optimal configuration involves locked pre-trained image models combined with unlocked text models. LIT teaches a text model to extract useful representations from a pre-trained image model for new tasks. The proposed method is versatile and works reliably with multiple pre-training methods (supervised and unsupervised) as well as diverse architectures (ResNet [33], Vision Transformers [20]).

A Foundational Language and Vision Alignment Model (Flava) [86] learns separate unimodal vision and language representations and combines them with a multi-modal encoder, which is also a Transformer [95]. The model utilizes masked image modelling [4] and masked language modelling [93] for the encoders, while a novel contrastive, masked multi-modal modelling (MMM) loss and image-text matching (ITM) loss are employed over paired image-text data.
BridgeTower [107] addresses the limitations of existing visual-language representation learning approaches, which typically involve unimodal encoders that learn to extract, align, and fuse both modalities simultaneously in a deep cross-modal encoder, or use the last-layer unimodal representations from deep pre-trained unimodal encoders as input for the top cross-modal encoder. The authors propose BridgeTower, a method that introduces multiple bridge layers to connect the top layers of unimodal encoders with layers of the cross-modal encoder. This approach enables effective bottom-up cross-modal alignment and fusion between visual and textual representations of different semantic levels from pre-trained unimodal encoders within the cross-modal encoder.

3.2.2. Task and Domain Adaption

Numerous strategies exist to improve the classification accuracy for models like Clip. In addition to non-task-related factors, such as selecting appropriate hyperparameters and defining well-tuned loss functions, it is beneficial to choose strategies that adapt the model to the task domain. Foundation models, although covering many aspects of common data, may not learn all relevant features and information from a target domain during pretraining. Domain adaptation is the most widespread method to align the model with the target domain, which can be achieved in multiple ways: A traditional approach is fine-tuning the entire model on a given dataset [42] or only selected weights, e.g., the last layers of a model. In theory, this leads to improvements, but it has many pitfalls in practice, such as the risk of overfitting the transformers from Clip on the training data and losing generalization abilities. WiSE-FT [105] addresses this issue by introducing a method that enables domain adaptation without sacrificing generalization ability.
Aside from these traditional approaches, recent research has explored domain adaptation methods without fine-tuning, leveraging the power of pre-trained models like Clip. Prompting, as discussed in a comprehensive survey by [58], represents one such method. CoOp [112] is an example of a prompting technique that learns prefix prompts for Clip's encoder to enable domain adaptation. Building upon this, Clip-Adapter [26] introduces a two-layer Multi-Layer Perceptron with a residual connection at the end, serving as a bottleneck adapter for improved performance. TIP-Adapter [111] further extends this concept by incorporating a cache model, which not only allows efficient adaptation but also supports few-shot learning scenarios, as detailed in the background section. Finally, another line of research explores the use of adapters [41] that can be integrated into the transformer layers of the model. These adapters can either be pre-trained or fine-tuned for task-specific applications, providing a flexible and efficient approach to domain adaptation.

4. Methods and Datasets

The methodology chapter of this scientific work provides a comprehensive overview of the methods and techniques used to address the research questions. It begins with a general introduction to the field of Out-of-Distribution Detection (OOD detection) and the benchmark methods utilized in this study. This is followed by a comprehensive description of the Zero-Shot Out-of-Distribution Detection (Z-OOD detection) methods used in the experiments.

The chapter also includes a detailed description of the datasets used in the experiments, along with the corruptions applied to test the robustness of the Z-OOD detection methods. This section provides a qualitative and quantitative description of the datasets, their properties, and the metrics used to evaluate the results.
The methodology chapter also provides a comprehensive description of the model architectures and training procedures used in the experiments. This includes a detailed explanation of the adapter layer injection technique used for few-shot domain adaptation in Clip. The chapter concludes with a summary of the methods used in this study, ensuring the reproducibility of results and a clear understanding of the methodology used to address the research questions.

4.1. Methodology

This section outlines the methodology used to address each of the research questions in this thesis. The task of OOD detection is a binary detection task in all experiments, where In-Distribution (ID) data is labelled as 0 and Out-of-Distribution (OOD) data is labelled as 1. Unless otherwise stated, the detection score or performance of a method is measured using the Area Under Receiver Operating Characteristic (AUROC).

Research Question 1: Is the performance of current zero-shot OOD detection methodologies generalizable, i.e., transferable to datasets within the realm of Clip's zero-shot capabilities?

To address this question, we first define the term *generalizable* since it is not possible to make this statement without exhaustive testing over all possible data. We approach it by first selecting multiple datasets (12) and running all far-OOD combinations with Maximum Concept Matching (MCM), which has a low inference time. The main part of the

comparison is between the different zero-shot OOD detection methods in near-OOD settings. This approach was chosen because experiments have shown that both MCM and Zero-Shot Out-of-Distribution Detection based on Clip (ZOC) work well in far-OOD scenarios, but there is little to no comparison in hard experiments. In total, we run 144 (132 far-OOD, 12 near-OOD) different combinations with zero-shot OOD detection methods to determine their limits and potential.

The focus of this work is the classification of real-world pictures of physical objects, as this task is perceived as the most widespread one in image classification. The selected datasets differ in high-level features and statistics, such as image size, backgrounds, and shapes of the objects, number of classes, and number of samples. The OOD detection specific metrics, Confusion Log Probability (CLP) and Maximum Mean Discrepancy (MMD) classify all tasks of the near-OOD part as hard [104, 56, 21]. That means, the inter-class differences, compared to the intra-class differences, are very small in these tasks.

Regarding the difficulty of OOD detection and the selection of ID/OOD splits, this work focuses on difficulty metrics rather than on high-level semantic features.

To ensure high similarity in the near-OOD experiments between ID and OOD samples, the classes of each dataset are randomly divided into 40% ID classes and 60% OOD classes. All experiments are run ten times with different random ID/OOD splits to mitigate the high variance this setup entails. The influence of openness on the OOD detection performance is well known and not further investigated here, as it is kept high enough to ensure the tasks are not trivial to solve. All experiments are conducted with $openness \approx 24\%$. The decision to use classes from the same dataset as OOD samples is considered difficult in related work, with CLP $= [-2.486, -0.955]$ and MMD $= [0.001, 0.035]$ between ID and OOD samples [56, 104, 21].

The methodology selected in this thesis aims to enhance the comparability with relevant literature in the research area, which currently has limited comparability (as discussed in Chapter 3). From a practical perspective, it is also reasonable to include more samples that are similar to in-distribution (ID) data, as image classification tasks may not encompass all aspects of a particular domain, such as flowers. Meant is, that if a flower classification system is deployed, it is likely that it does not know every flower that exists, but will likely be queried with some of the unknown flowers, rather than be queried with objects that are very dissimilar, e.g. with images of a dog.

This thesis compares and evaluates two published Z-OOD detection methods: ZOC [21] and MCM [65], which is also used as MSP [36] modified with Clip's zero-shot classifier earlier in [21] as a baseline. If not otherwise mentioned, we use the same zero-shot classification template (see Section 4.3) for all datasets and experiments.

As the baseline, we have added the MSP methodology with Clip as the pre-trained feature extractor and a logistic regression classifier on top, referred to as Clip-L [74], trained on the training split of the dataset. The training set here refers to the selected ID classes per run, and the baseline has no exposure to any out-of-distribution (OOD) samples.

All methods utilize the small vision transformer backbone, "ViT-B/32", from Clip, which serves as the foundation for ZOC and is still computable by consumer GPUs. This is a crucial factor in terms of reproducibility. Ideally, all possible vision backbones should be utilized, but this is not feasible in this context as it would require the computationally intensive retraining of the description generator in ZOC (as discussed in Section 4.3). The experiments and methodologies do not rely on Clip exclusively, and other image-language foundation models could also be used (see Chapter 3). Clip is selected as the backbone of all experiments due to its widespread usage as the most prominent vision-language model for zero-shot classification and its use in related literature [21, 65].

Research Question 2: Where are the boundaries of zero-shot OOD detection methods with respect to different difficulty metrics for Out-of-Distribution detection?

To answer this question, the baselines for CLP, MMD, and zero-shot accuracy are established with the data sets and data splits shown above. The methodology is then expanded by altering the images with common corruptions (see Section 4.2.2) to increase the difficulty of the task until the Z-OOD detection methods are no longer applicable, which is defined in this work as not better than randomly guessing.

Five datasets are selected based on their zero-shot closed-world classification accuracy, as it has been shown to be strongly correlated to the OOD detection performance [96]. These datasets are manually corrupted with increasing severity, following a methodology described in prior work [35], which is used there to test the robustness of the classifier. Overlap in the accuracy of the dataset-corruption combinations is expected and desired, as it may provide further insight into the reliability of closed-world accuracy as a predictor for OOD detection scores and the influence of other measurable factors.

The experiments are performed by corrupting all images in a dataset with the same corruption. The closed-world accuracy, CLP, and MMD for the corrupted image dataset are reported, and the OOD detection is performed as described earlier. The experiments are performed using three different corruptions (Gaussian blur, snow, and brightness) at three different severity levels. The corruptions were selected based on their general and random nature (Gaussian blur), their occurrence in real-world phenomena (snow), and their impact on images taken outside of laboratory conditions (brightness).

Research Question. 3:How does the performance of out-of-distribution detection methods using Clip's zero-shot classifier compare to traditional State-of-the-Art Out-of-Distribution Detection methods?

The objective of Research Question 3 is to evaluate the performance of Z-OOD detection methods against traditional state-of-the-art OOD detection methods and if they can benefit from Clip domain-adaption methods, similar to regular image classification.

To compare Z-OOD detection methods with traditional OOD detection methods, a baseline is established using the Maximum Softmax Probability (MSP) method with a fine-tuned classifier using Clip vision transformer backbone as feature extractor. This method is selected as it has shown consistent and competitive results in previous studies [96, 65]. Instead of the vision backbone used in the other experiments, we here use a larger Clip model (see Section 4.3) which provides improved closed-world accuracy. Still, there is no outlier exposure in the training process. The baseline method has access to all training samples of each dataset. To obtain softmax scores for the images, a closed-world classifier is trained on the training split of each dataset and hyperparameters are tuned on a validation set. For further details on the models and training, see Section 4.3.

Additionally, the goal is to improve the Z-OOD detection methods. Temperature scaling is used on the output scores before obtaining the OOD detection scores [65]. the findings of previous studies that show an improvement in closed-world accuracy results in an improvement in the OOD detection score are considered [96]. We chose to improve classification accuracy with adapter-based domain adaption. Hence, to further improve the Z-OOD detection methods, a few-shot domain adaptation method, TIP-Adapters [111], is used. This method utilizes very few training samples and has the advantage of being computationally inexpensive and fast to use in real-world scenarios.

It is important to note that all methods that involve the use of training samples from the data distribution are no longer zero-shot methodologies. The main advantage of zero-shot methodologies is their independence from training samples, which are often expensive to obtain. However, if sufficient data is available, methods that involve the use of training samples are generally superior to few-shot domain adaptation methods in image classification tasks.

4.2. Datasets

4.2.1. Selected Datasets

Experiments regarding Zero-Shot Out-of-Distribution Detection (Z-OOD detection) are currently limited to very few datasets [21], thus we try to include many well-known image datasets, covering a broad range of attributes, such as the number of classes, the number of images per class, but also different difficulties and metrics, also described in this section. The datasets are chosen that there is negligible inter-dataset overlap between classes, images and even high-level features. The majority of datasets thus share no labels or even concepts e.g kinds of flowers in the Flowers102 dataset [68] and types of cars in the Stanford Cars dataset [47]. We provide a brief semantic description of each dataset in this section. A quantitative comparison of the datasets is provided by Table 4.1. To ensure reproducibility, we selected datasets that are publicly available. All datasets consist of at least ten different labels with multiple images per class.

Dataset Name	Content	Semantic Labels	Image Size	Classes	Samples
CALTECH101	Pictures of objects	Describes object	200-300px	101	8,623
Caltech CUB	Pictures of birds species	Bird species	120-500px	200	11,988
CIFAR10	Pictures of objects from common objects	Object	32 x 32 px	10	60,000
CIFAR100	labelled subsets of 80 million tiny images	Object	32 x 32 px	100	60,000
DTD	collection of textural images	Textures	300 - 640 px	47	5,640
Fashion MNIST	An MNIST-like fashion product database	Article of clothing	28x28 px	10	70,000
Flowers102	Images of flowers	Flower species	500-1.200px	120	8,189
GTSRB	Images of german road signs	Description of roadsigns	25-244px	43	39,270
TinyImageNet	Subset of ImageNet	Object	64x64 px	200	110,000
MNIST	Handwritten digits	Digits from 0-9	28x28 px	10	70,000
Stanford Cars	Images of cars	Make, Model, Year	41-5,616 px	196	16,185
SVHN	House numbers	Digits from 0-9	32x32 px	10	73,257

Table 4.1.: Overview of the datasets utilized in this work

Caltech101

The Caltech101 dataset [55] consists of 8,623 images belonging to 101 distinct categories. There are between 40 and 800 images per category, while most categories consist of 50 images. The image sizes are roughly 300 $\times$ 200 pixels. The image contents vary a lot, since many are photos of everyday objects in natural settings, while others are cropped images with no background at all. Figure 4.1 shows an example for each type of image from the dataset. The images were taken from the Google Search Engine image search by entering the respective class name. See Figure 4.1 for samples.

(a) A starfish in its natural habitat. Labelled as 'starfish'

(b) A cropped image of a camera. Labelled as 'camera'

Figure 4.1.: Two image samples from the Caltech101 dataset. The samples depict the background differences in the dataset

Caltech CUB

The Caltech-UCSD Birds-200-2011 dataset [99], also known as Caltech CUB, is described as a challenging dataset of 200 bird species, the classes. The high-level features in every image are very similar to each other, e.g. each species has a beach, wings, etc. The dataset consists of roughly 12,000 images, on average 60 per species. Each image has an edge length between 120 and 500 pixels and contains a single animal.

CIFAR10 & CIFAR100

The CIFAR and CIFAR100 datasets [48] each consist of 60,000 colour images, categorized into 10 (CIFAR10) and 100 (CIFAR100) classes respectively. The images are uniformly distributed over the classes. Each image is a 32×32 pixel colour image. Compared to other datasets in this work, the classes of both CIFAR datasets are disjunct and share few semantic features. Classes are e.g. 'airplane', 'cat', and 'ship' in CIFAR10. In CIFAR100 the classes are grouped into 20 superclasses, which are not further utilized in this work but show, that there is a higher overlap in the classes than in CIFAR10. Classes are e.g. 'dolphin', 'whale', 'poppies', 'sunflowers', 'lawn-mower', and 'woman'.

DTD

The Describable Textures Dataset (dtd) [15] consists of 5460 images showing 47 different textures. The images are uniformly distributed over all classes, hence 120 images are present for each category. It is designed to gain insights into how textural information is processed by an intelligent system. The DTD differs from the majority of visual datasets in research and this thesis, which focus on the classification of objects present in the images, not the textures of the objects in the images. The image size ranges between 300×300 and 640×640 pixels with at least 90% of the surface representing the texture. All images were taken from Google Image search and Flickr[1] by entering the respective texture phrase. See Figure 4.2 for samples.

(a) Texture labelled 'fibrous' (b) Texture labelled 'lined'

Figure 4.2.: Two image samples from the DTD dataset

[1] https://www.flickr.com/

Fashion-MNIST

Fashion MNIST dataset [106] consists of 70,000 images of fashion products. the images are split into a training set, consisting of 60,000 images and a testing set, consisting of 10,000 images by the publishers. Each image is a 28×28 greyscale picture and is associated with a label from the ten available classes. The images are uniformly distributed with 6,000 images per class in the train split and 1,000 images per class in the test split. The dataset is proposed as a more challenging classification task than the well-known MNIST dataset. [54]. All images were taken from an online shopping platform, sharpened, resized and converted to 8-bit greyscale images.

Flowers102

The Flowers102 dataset [68] consists of 8189 images from 102 different flower classes. The images are split into training, validation, and test set by the publishers. The training and validation sets each consist of ten images per flower class, so 1020 images per split. The test set consists of 6149 images, with between 40 to 250 images per class. The images are collected from the web, some were taken by the researchers themselves. Each image is rescaled so that the smallest dimension is 500 pixels and the original aspect ratio is preserved. The different classes in this dataset have a higher feature similarity between classes than other datasets presented here, such as the CIFAR datasets [48] as the flowers share many high-level features, such as flower, stem, and leaves. Originally, the labels are one-hot-encoded. For this work, we use the semantic labels provided by Radford et al. [74], which are manually crafted from descriptions in the original publication [68].

GTSRB

The German Traffic Sign Recognition Benchmark (GTSRB) [90] consists of 39270 photos of 43 different German traffic signs. The dataset is split into a training set, consisting of 26640 images and a test set, consisting of 12630 images. Each class consists of 210 to 2250 images with a rough average of 900 images per class. The dataset was created from ten hours of video recorded while driving on different types of German streets during the daytime. Each physical sign exists only once in the dataset, meaning, every image of a class is taken with different illumination, background, etc. All images have a resolution of 1360 ×1024 pixels. The traffic signs in the image itself are between 15×15 and 222×193 pixels. This dataset has also a higher inter-class similarity than object classification datasets such as the CIFAR datasets [48], as traffic signs share many high-level features and are therefore considered more difficult [74]. Originally, the labels were one-hot-encoded. To allow for zero-shot classification, we utilize semantic labels provided by Radford et al. [74].

MNIST

The MNIST dataset [54] consists of 70,000 samples of images of handwritten digits from zero to nine. The dataset is split into 60,000 train and 10,000 test samples. The images are uniformly distributed over all classes. The images were sampled from the NIST dataset, normalized and centred to 28×28 8-bit greyscale images. The image classification task is considered solved, with models achieving more than 99% accuracy [100, 3, 10, 39], hence considered easy. The dataset is still utilized as a benchmark for different algorithms and image-related tasks, such as OOD detection.

Tiny ImageNet

The Tiny ImageNet dataset [52] consists of 110,000 images of 200 classes of different objects and is a subset of the Imagenet large scale visual recognition challenge (ILSVRC) [79]. The dataset is split into 100,000 training images (500 per class), and 10,000 test images (50 per class). Each image is downsampled to 64×64 pixels.

Stanford Cars

The Stanford Cars dataset [47] consists of 16185 images of 196 different distinct classes. The classes are fine-grained car fabricates with the oldest model dating from in 1990. Each semantic label consists of the manufacturer, the model and the model year (e.g. "Audi RS 4 Convertible 2008"). The dataset is split into a training set (8144 images) and a testing (8041 images) by randomly dividing each class in half. On average, there are 41 images per class per split. The classes were sampled from a crawled list of all types of cars made since 1990. The images themselves were taken from the web with these classes and manually labelled by trained annotators [47]. The images share a high inter-class similarity for high-level features, such as colours and shapes. Thus, image classification task is considered very difficult, even for humans [47]. The images have different edge lengths, ranging from 41 to 7800 pixels.

SVHN

The Street View Housing Numbers (SVHN) dataset [67] consists of 99289 images with ten different classes. The dataset is split into training (73257 images) and testing (26032 images), nearly uniformly distributed per class. Each image is a 32×32 colour image, cropped from Google street view images of housing numbers to represent a single digit. For example, house number 648 is split into three images with labels 6, 4, and 8 similar to the MNIST dataset [54] regarding image size and classes, but significantly harder to recognize since the images show vast intra-class variations and photometric distortions [67].

4.2.2. Corruptions

In this part, image corruptions [35] are introduced. We briefly explain the reasoning behind using corruptions as well as present the selected corruptions for the methodology of this work.

Compared to the human visual system, computer vision systems are not robust to small changes in images querying these systems. While the human system can easily cope with such small changes and even more abstract changes [35], many computer vision systems are easily fooled by such changes. But in practice, these changes do occur very often, due to technical errors or limitations (resolution, motion blur), environmental influences (rain, snow, illumination) and more. Thus Hendrycks and Dietterich [35] introduced ImageNet-C, a dataset consisting of 15 different, algorithmically generated corruptions of five increasing severity levels. The corruptions are selected so that they resemble many of the naturally occurring image corruptions.

The three corruptions utilized in the methodology of this work and the severity levels are:

1. *Gaussian Noise* can appear in low-lighting conditions. It adds a random value, drawn from a Gaussian distribution with mean $= 0$, to each pixel of the original RGB image. The three severity levels define the standard deviation of the distribution. The severity values are $0.08, 0.18, 0.38$.

2. *Brightness* of an image is varying with daylight and artificial illumination intensity. The corruption transforms RGB images to HSV images and adds a positive value to the V-channel. The severity values are $0.1, 0.3, 0.5$

3. *Snow* is a visually obstructive form of precipitation. It adds multiple white stains with motion blur to the original images and puts a grey veil over the image. The severity values are presented in Table 4.2

The algorithms and severity level values are identical with the corruptions used in Imagenet-C [35], only minor adjustments for different image shapes and formats are applied. Greyscale images, such as samples from MNIST [54] are transformed into RGB images. Figure 4.3 shows sample images for the used corruptions.

Severity	Gaussian - mean	Gaussian - std	zoom factor	min threshhold	motion - radius	motion - std	coluor intensity
1	0.1	0.3	3.0	0.5	10.0	4.0	0.8
2	0.55	0.3	4.0	0.9	12.0	8.0	0.65
3	0.55	0.3	2.5	0.85	12.0	12.0	0.55

Table 4.2.: Corruption values for each severity level of the Snow corruption

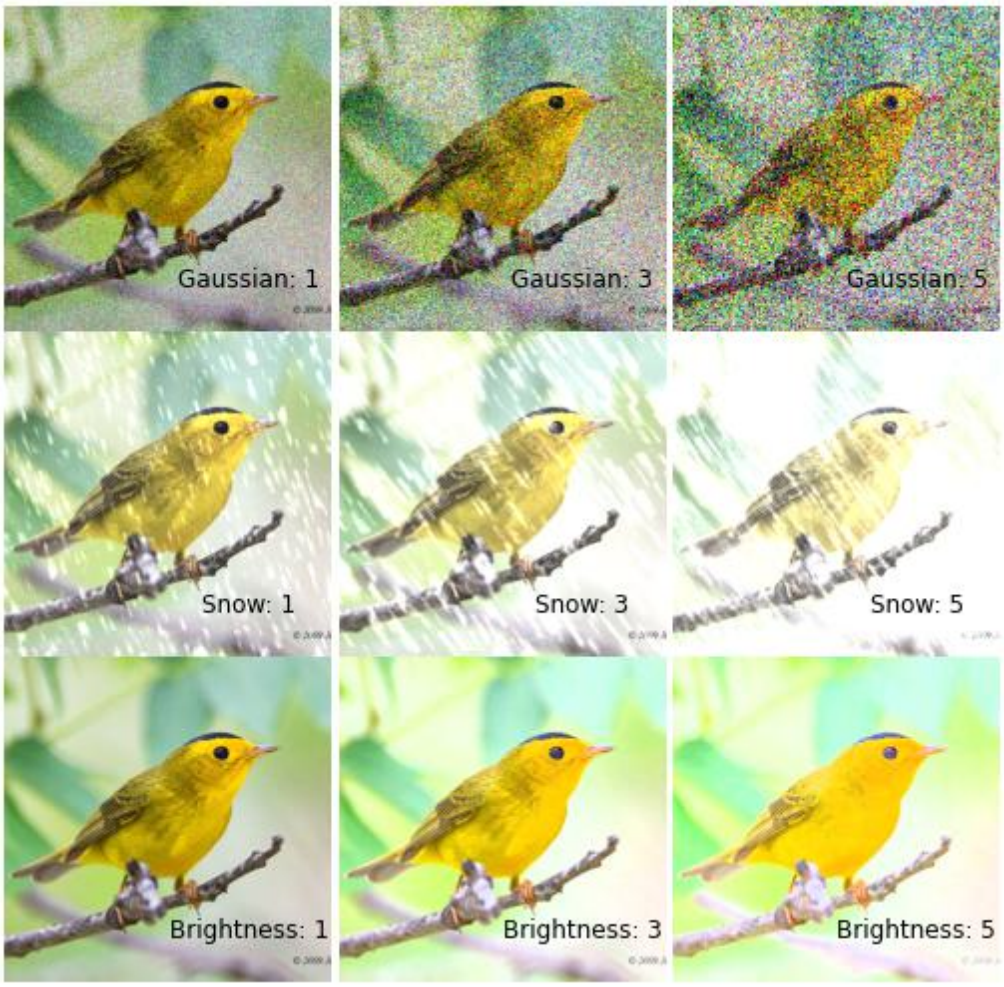


Figure 4.3.: An image of a Wilson's warbler from the CUB dataset [99] corrupted with three increasing severities from left to right. Each row presents on corruption. The title on each image names the corruption and the severity and is here added for clarity

4.2.3. Metrics

In this part, the most important metrics for this thesis are briefly described, as well as a brief explanatory statement about why they were chosen.

Accuracy

The accuracy is a measure of observational error and is commonly utilized for classification tasks. The accuracy shows, how close a given set of measurements, here the output of a classifier, are to their true values, here the ground truth or labels. For multiclass classification, the accuracy is defined as follows:

$$\text{Accuracy} = \frac{\text{correct classifications}}{\text{all classifications}}$$

The simplicity makes this metric easily comparable and interpretable, even though the accuracy does not account for heavily imbalanced datasets. As all datasets utilized in this work are nearly or fully balanced, the accuracy is a reliable indicator of the classification performance of a model.

AUROC

AUROC is a commonly used metric for evaluating the performance of a binary classification model. As we define the task of Out-of-Distribution Detection as binary classification with the classes In-Distribution (label 0) and OOD (label 1), the AUROC score is the main metric to compare different models. It is also the standard in related work. The AUROC provides a single scalar value to evaluate the performance of a OOD detection method.

Confusion Log Probability

In this study, only one Clip zero-shot classifier [74] is used instead of an ensemble of classifiers, which are traditionally used due to their well-calibrated predictions [50]. An ensemble of the same classifier is not necessary, as the predictions of the classifier are identical. The choice of the Clip zero-shot classifier was made because it serves as the backbone for all Z-OOD detection methodologies, and it is expected that the Clip based Confusion Log Probability (CLP) will be more predictive than other classifiers. Additionally, using a single classifier may improve comparability, as this setting is available to all researchers and can be applied to any image dataset task with consistent results. Even though an ensemble of classifiers reduces the risk, fine-tuned classifiers still have the chance of different outcomes, as they rely on many learnable parameters and hyperparameters.

Formally, the CLP for two labelled datasets D_{in} and D_{out} with the corresponding sets of classes C_{in} and C_{out}, the classification is performed on the joint dataset $D = D_{\text{in}} \cup D_{\text{out}}$ and the extended label set $C = C_{\text{in}} \cup C_{\text{out}}$. The expected probability of a test sample x to be predicted as class k is given by:

$$c_k(x) = \hat{p}^j(\hat{y} = k|X).$$

The confusion of of OOD samples D_{test} with inlier classes C_{in}, the CLP is then

$$\text{CLP}_{c_{\text{in}}}(D_{\text{test}}) = \log\left(\frac{1}{|D_{\text{test}}|}\sum_{x\in D_{\text{test}}}\sum_{k\in C_{\text{in}}} c_k(x).\right)$$

A high CLP score indicates that test samples are near-OOD, so hard to separate from the original data for the classifier, a low CLP score indicates far-OOD samples, or an *easier* task.

MMD

The Maximum Mean Discrepancy (MMD) is a measure of statistical distance between two datasets, here between the ID and the OOD datasets. The MMD is negatively correlated with the detection performance, and hence can be used to determine the difficulty of a task in the OODD domain, where there are few standardized benchmarks.
The MMD is calculated using a kernel function on the embeddings of the images, with a high MMD indicating low intra-class distances (first two sums of Equation 4.1) and high inter-class distances between in-distribution (ID) and OOD samples. In this study, the MMD is calculated using the exponential of the squared Euclidean distance between image embeddings, and the embeddings are obtained using the Clip model.

Formally, given two image sets, $V = \{v_1, \ldots, v_m\}$ and $W = \{w_1, \ldots, w_m\}$ and a kernel function $k(.,.)$, the MMD between V and W is defined as:

$$\text{MMD}(V, W) = \frac{1}{\binom{m}{2}} \sum_{i \neq j} k(v_i, v_j) + \frac{1}{\binom{m}{2}} \sum_{i \neq j} k(w_i, w_j) - \frac{1}{\binom{m}{2}} \sum_{i \neq j} k(v_i, w_j) \tag{4.1}$$

with $k(x, x') = \exp\left(-\frac{||x-x'||_2^2}{2\sigma^2}\right)$, as also used in related work [56, 91].
The scores are calculated based on the Clip embeddings of each image. A high MMD indicates that intra-class distances (first two summands of Equation 4.1) are low compared to the inter-class distance (last summand of Equation 4.1), which means ID and OOD are harder to distinguish and vice versa in OOD detection.

Openness

The openness measures the difficulty of an OOD detection task by comparing the number of seen classes in ratio to the unseen classes [84]. In the original formulation, there is a separation between the number of ID train classes N_{train} and the number of ID target classes N_{target} when testing. N_{test} is the total number of classes. Formally, openness is defined as follows:

$$\text{openness} = (1 - \sqrt{\frac{2 \times N_{\text{train}}}{N_{\text{test}} + N_{\text{target}}}}) \times 100$$

As there is no training, so no training classes, in Z-OOD detection and we do not experiment with different openness settings, N_{train} is always equal to N_{target}. The openness is negatively correlated to the OOD detection performance. A higher openness means a higher proportion of unseen classes tested against the ID classes, which increases the likelihood that an OOD sample is similar to an ID class.
The openness does not account for the number of classes in the OOD detection and not for any other property of a classification task, such as semantic similarity of classes or similarity of ID and OOD. Also, it is a strictly artificial measure, as in real-world scenarios there cannot be a number of target classes. Thus, it is not reliable as an indicator of the difficulty of an OOD detection task [104] and kept constant.

4.3. Model Details

This section describes the model architectures for OOD detection and provides the training details for selected models and the domain adaption methods used to increase the performance of these models.

4.3.1. Clip

The Clip model used in this work is a combination of a vision encoder and a text encoder. The vision encoder is based on the "ViT-B/32" model, which is a base variant of the vision transformer that can be run on consumer GPUs with 12 GB RAM. The patch size for tokenizing the images is 32, which balances the trade-off between preserving details and the number of tokens to process. The images are resized and centre cropped to match the input resolution specified in Table 4.3. This specific vision encoder is chosen for its compatibility with hardware restrictions and for comparability to related work.

The text encoder is a Transformer that operates on a vocabulary size of 49,152, using lower-cased byte pair encoding [85] and a maximum sequence length of 76. The text sequences are bracketed with a start token "[SOS]" and an end token "[EOS]". The representation of the text is obtained from the activation of the highest layer of the Transformer at the [EOS] token, as described in the original publication [74].

Model	Embedding Dimension	Input Resolution	V - Layers	Patch Size	Tokens	T- Layers	T - Heads
ViT-B/32	512	224	12	32	49	12	8
ViT-L/16@336	768	336	24	16	196	12	12

Table 4.3.: Clip-ViT hyperparameters. "V" stands for Vision Encoder [20] model, "T" for the Text Encoder of Clip [74].

4.3.2. Baseline

As a baseline method, an adapted version of the acrshortmsp [36] is adopted. This approach involves training a logistic regression classifier on the in-distribution (ID) data and using the maximum class probability as the prediction. The logistic regression is implemented using the L-BFGS solver from the scikit-learn library [71], with a maximum of 1,000 iterations. The L2 regularization strength is determined by evaluating 96 logarithmically spaced values in the range from 10^{-6} to 10^{6}.

We also compared the logistic regression with a fully connected linear classification head trained with Adam [45] optimizer and multiple hyperparameter settings for 1000 epochs but we found the performance of the logistic regression is on average better than the linear classification head, even with many hyperparameter settings. This holds for both, the classification accuracy on the OOD detection performance. See Appendix A.1 for full results. This also holds for the benchmark used to compare to the fine-tuned OOD detection models, where we train the classifier on the features of the ViT-L/14@336px vision encoder.

4.3.3. MCM

The methodology for MCM [65] involves the use of a pre-trained Clip model, as described in Section 4.3.1. The base temperature value τ is set to 1.0, as recommended in the original publication [65] for the main experiments. An deeper look on the large-scale experiments on the influence of τ will be provided.

4.3.4. ZOC

For ZOC [21], we also use the pre-trained Clip model described in Section 4.3.1. The text decoder used is from BERT large [17] with 24 Transformer layers and a hidden size of 1024. During inference, candidate labels are selected from the top 35 levels of annotations, with a maximum of 77 iterations. The seen labels are filtered from the candidate labels, but stopwords are not removed. In order to ensure comparability with the original publication [21], the recommended training method for the decoder is followed, although alternative methods exist. In a recently published work [65], the method is used with GPT-2 [75] as initial weights, which is not further investigated here. The temperature parameter for zero-shot classification is set to 0.01, as suggested in the original publication [21]. Additional insights into the influence of τ on ZOC will also be provided.
From all seen labels Y_s and generated labels Y_u we generate zero-shot labels with the recommended template "This is a photo of a {TOKEN}", which is very similar to the zero-shot template from Clip [74]. We define the OOD detection confidence score as

$$S(i) = 1 - \sum_{y \in Y_s} P(y|i),$$

the accumulative sum of the probability of Y_u (see Chapter 2 for further details).

Training Details

The only part of the model that needs to be trained is the $\text{Decoder}_{\text{TEXT}}$. Theoretically, this model can be considered a foundation model [8], as it is not specifically trained for this task or specific datasets (hence ZOC claimes to be a zero-shot method). Such a model is, in contrast to Clip not easily online available with training similar to the original publication [21]. Thus, to ensure comparability, we decide to train it following the described method.
The text encoder is pre-trained on the BBC extreme dataset [66] and is available on the Hugging Face platform[2] [77].

The fine-tuning process is performed using the MS-COCO 2017 dataset [57], which is commonly used for training image captions. The ViT-B/32 model is used as the image feature generator. The training is performed using the teacher-forcing method [103] and the Adam optimizer [45] with a constant learning rate of 10^{-5} for 25 epochs. The training

[2] www.huggingface.org

and validation sets are based on the officially released data splits for the MS-COCO 2017 release [57]. The model after the iteration with the lowest loss on the validation split is used as final model the experiments.

4.3.5. Adapter

This subsection explaines the model and training details of the utilized TIP-Adapters [111].

Model Details

The adapters used in this study, TIP-Adapter, are identical to those used in the publication by Zhang et al. [111] The adapter stores knowledge from a few-shot training set in a cache model. The experiments are conducted with two types of cache models: either a cache model with no trainable parameters constructed from the training data (further referred to as TIP) or a trainable cache model with weights initialized from the training data, called TIP-f (TIP-fine-tuned). The fine-tuned version is a fully connected layer with the size of the cache model and uses the generated features as initial weights for faster domain adaption [111]. For a more detailed explanation of the cache model (see Chapter 2. The size of the cache model is defined by the number of classes (N), the number of images per dataset (K), and the dimension of the vision encoder (C) as $\dim_{\text{cache}} = NK \times C$. For $K = 16$ and $C = 512$, the output size of the ViT-B/32, the number of trainable parameters per dataset ranges from $16 \times 512 \times 10 = 81,920$ (for 10 classes, the minimum number of classes in all datasets) to $16 \times 512 \times 200 = 1,638,400$ (for TinyImagenet's 200 classes), which is relatively small compared to the 87 million parameters in the ViT-B/32 model. The adapter is trained on a maximum of $200 \times 16 = 3,200$ images. $K = 16$ is selected as it is considered few-shot learning [112, 26, 111] and thus still close to the original zero-shot setting.

Training Details

The cache model is built using the few-shot training data and 10 augmentation epochs. In each of these epochs, a random part from each image is cropped and resized to the input shape of the Clip vision encoder. The cropped parts are $50 - 100\%$ of the original image. Afterwards, each image is randomly flipped.
For the TIP-Adapter-F, AdamW [61] is chosen as the optimizer with a learning rate of 10^{-3} and the stabilizing parameter $\epsilon = 10^{-4}$. The training runs for 20 epochs, and the model with the best accuracy on a held-out validation set is selected as the best model. A hyperparameter search for the α and β is conducted on the same held-out validation set. All combinations of ten evenly spaced numbers over the interval $[0.1, 10]$ are selected, resulting in 100 different combinations. Initially, both values are set to 1.0.

4.3.6. Prompt Engineering

Prompt engineering and prompt tuning play a significant role in closed-world accuracy, as reported in previous studies [112, 74]. Despite evidence that different prompts can have a significant impact on closed-world accuracy, the influence of prompts on zero-shot out-of-distribution detection performance is not as pronounced [65]. This can be attributed to the possibility that using the same prompt for every image shifts the images in the same direction, which can alter the prediction of zero-shot classifiers that rely on distance metrics such as cosine similarity. However, a threshold-based metric, such as AUROC, is less affected by identical shifts. Further experiments are needed to examine this hypothesis but are not within the scope of this study. Therefore, the prompt "This is a photo of a TOKEN" was selected for all experiments.

4.4. Optimization of Zero-Shot Out-of-Distribution Detection Methods

In this section, we delineate the methodology employed for tailoring the Z-OOD detection techniques to various domains. Initially, we expound on the utilization of adapters in Clip image categorization, followed by the integration of the approach into OOD detection mechanisms. To the best of our knowledge, there is no adaptation of the proposed Z-OOD detection techniques for a few-shot configuration. The premise of this concept is grounded in the substantiated correlation between enhanced closed-world accuracy and the OOD detection score [96], as well as the research on few-shot domain adaptation for Clip [74] image classification [111].

4.4.1. Domain Adaption for Image Classification

For comprehensive information on the construction of the cache model, which serves as a means to query the TIP-logits, refer to Chapters 2 and 4.3. Moreover, we provide the algorithm for implementing the TIP-Adapter in Figure 2 in the form of pseudocode. Querying the pre-constructed cache model with a novel, unobserved image essentially generates a similarity matrix between the query image and the k-shot training set, exhibiting the same dimensions as the Clip zero-shot predictions. Subsequently, upon acquiring these cache values and the affinity (hereinafter referred to as *TIP knowledge*), this knowledge is combined with the predictions from the pre-trained Clip model (*Clip knowledge*) through a residual connection [33]. This approach facilitates the concurrent exploitation of knowledge derived from the few-shot cache model and the pre-trained Clip model [111].

Algorithm 2 Algorithm used to obtain a class prediction using TIP-Adapter [111]

Require: k-shot cache model (tip_cache_model)

```
# query clip & tip knowledge
clip_zeroshot_logits = clip_zeroshot(image, labels)
cache_values, affinity = tip_cache_model(image)
alpha, beta = 1.0, 1.0
# create the tip logits for classification
cache_logits = ((-1) * (beta - beta * affinity)).exp() @ cache_values
tip_logits = clip_zeroshot_logits + cache_logits * alpha
softmax_scores = softmax(tip_logits)
score, class_prediction = get_top_one_prediction(softmax_scores)
```

4.4.2. Domain Adaption for Clip-based Out-of-Distribution Detection

The described Z-OOD detection methodologies share a common trait: the determination of whether an image is ID or OOD is contingent upon the classifier's confidence. Theoretically, this confidence, in the context of MCM [65], is the proximity[3] of the representation of the nearest known class to the image in comparison to all other similarities (i.e., the maximum softmax score). For ZOC, the ratio of the cumulative probabilities of all generated labels is compared to the known classes, with this accumulation also interpretable as an additional label.

The unscaled softmax scores procured by Clip's zero-shot classifier exhibit high similarity and are nearly uniformly distributed. This similarity arises from two factors: the known outcome of Clip's Contrastive Training method [102, 101] and the high-dimensional feature space of the embeddings (512 for ViT-B/32), as uniformly distributed points in a high-dimensional sphere, tend to be equidistant [98]. In the literature and officially released code snippets[4], temperature scaling with $\tau = 0.01$ is applied, which enhances the classifier's confidence, as $\tau \to 0$ converges the probability to a point mass. It should be noted that, firstly, this scaling does not alter classification performance but significantly impacts Z-OOD detection [65], and secondly, the direct interpretation of softmax scores as confidence is met with scepticism [36, 32].

The incorporation of adapters amplifies the logit score for similar images from the constructed cache model, thereby increasing the softmax score of the label corresponding to these akin training images. No spike in similarity is anticipated if all images from the training set are approximately equidistant, which is more probable for an OOD sample. Intuitively, this implies that identifying a threshold for the ID/OOD separation becomes more feasible, as the closed-world classifier's confidence is elevated [111].

This effect mirrors the temperature scaling in MCM [65] for far-OOD experiments (i.e., reducing the uniformity of zero-shot predictions). However, the distinction lies in the fact that utilizing adapters not only escalates the softmax scores but also potentially alters the

[3] Proximity here signifies the cosine similarity between image features and class label features in Clip's embedding space

[4] e.g., `www.github.com/openai/CLIP`, `www.github.com/sesmae/ZOC`

model's predictions. For OOD detection, this is inconsequential, as all known labels are ID; thus, we are interested in high confidence scores for known classes when an input image belongs to one of these classes (i.e., the image is an ID sample). By merging a well-calibrated classifier with domain knowledge from the cache model, we anticipate enhanced OOD detection performance. In addition to this rationale, empirical evidence [96] has demonstrated that increased closed-world accuracy translates to an improved OOD detection score. We deduce that domain adaptation with TIP-Adapter holds a high likelihood of augmenting the OOD detection performance in comparison to zero-shot MSP / MCM.

The explication of this chosen method also accommodates other approaches for few-shot domain adaptation. However, as demonstrated in Chapter 2 and the training details in Section 4.3, the caption generator for ZOC is designed to predict captions based on the output of the original Clip vision encoder. The residual connection in TIP-Adapter (Line 7 in Algorithm 2) enables the utilization of the original caption generator, as the output remains unaltered. Employing other domain adaptation techniques necessitates retraining the caption generator or using a second Clip model to produce unchanged image features. While both options are viable, they invariably increase the effort required to implement them, potentially diminishing one of the advantages of Z-OOD detection. Consequently, further investigation into these alternatives was not pursued.

Incorporating a trained TIP-Adapter in MCM mirrors its use in closed-world classification (see Part 4.4.1). However, integrating TIP-Adapter into ZOC, which we designate as T-ZOC or with "-f" appended to indicate the fine-tuned version of TIP-Adapters, utilizing the same residual connection is not feasible, as there are no cache values for the generated labels. As a result, we pad the adapter weights ("tip logits" in Algorithm 2) with the neutral element in the softmax function. We recognize that appending the non-negative TIP similarities increases the mass for the ID probability; nevertheless, this alteration is unlikely to negatively affect the threshold-based AUROC.

5. Experimental Analysis of Zero-Shot Out-of-Distribution Detection

This chapter delves into the generalization ability and robustness of Zero-Shot Out-of-Distribution Detection (Z-OOD detection) methods in detecting images not originating from the initial image distribution, addressing the first and second research questions.

Given the infinite possibilities of images and classes, proving the general applicability of OOD detection for images in an open-ended task is impossible. To address this challenge, we select diverse datasets and create ID / OOD setups to test the applicability of these methods broadly and at the highest level of difficulty. Image corruptions are also introduced to further increase the difficulty, a method not yet tested on Z-OOD detection.

The experiments to answer research questions 1 and 2 are divided into three parts:

5.1 **Exploratory search for far-OOD**: Assessing the performance of Z-OOD detection methods on a large set of 132 far-OOD combinations for OOD detection.

5.2 **Exploratory search for near-OOD**: Evaluating the performance of Z-OOD detection methods on more difficult near-OOD detection setups, comparing the performance to a fine-tuned method across 12 different ID/OOD combinations with varying methods and hyperparameters.

5.3 **Exploratory search for the lower bound of Zero-Shot Out-of-Distribution Detection**: Determining the point at which Z-OOD detection methods no longer work[1] on 15 different setups by progressively corrupting the image with increasing severity.

These experiments aim to provide a comprehensive understanding of Z-OOD detection methods, which is left open since previous publications [21, 65] conduct specific tests on a smaller scale and focused more on a comparison to other OOD detection methods. This gap shall be closed. Also, this chapter provides benchmarks for further experiments in Chapter 6.

The results are compared to a traditional method and metrics used to measure difficulty in standard OOD detection are assessed for their suitability in determining the difficulty of Z-OOD detection. Additionally, the experiments will determine if one Z-OOD detection method is superior to the other. It will also become clear whether one of the two

[1] No longer works means that it has an AUROC of 0.5, which is achieved by an uninformed guesser

Z-OOD detection is superior to the other, or whether, as the previous publications [21, 65] suggests, no method is always superior to the other.

Precisely, we test all possible combinations of ID / OOD of the datasets among each other, which corresponds to a rather easy task (far-OOD), as well as a hard OOD detection with each dataset, namely by splitting the classes into 40% ID and 60% OOD. Thus for twelve datasets, we test $12 \times 11 = 132$ far-OOD combinations, and 12 for near-OOD. The large number of combinations can be tested due to MCM's minimal training and adjustments required, along with its short inference time. However, ZOC's higher inference time limits the scale of experiments. The focus is on the challenging near-OOD experiments, where both ZOC and MCM will be tested. For accessing the robustness of the methods, three corruptions with three severities on 5 selected datasets, so 45 different combinations will be investigated.

MMD and the CLP are measured for all combinations. In addition, we provide the Zero-Shot Accuracy (ZSA) of the ID data set. The AUROC score is primarily used as a performance measure. For MCM, we test the temperature values $\tau \in \{0.01, 1.0, 100.0\}$ where $\tau = 1.0$ corresponds to the recommendation from the paper [65] and $\tau = 0.01$ corresponds to the MCM baseline used in [21], where it is called *MSP+Clip*. $\tau = 100$ is used to test a less confident classifier.

Finally, we compare the above methods with a fine-tuned classifier that uses all training data of the respective dataset to adapt to the distribution. We compare the methods to a fine-tuned logistic regression classifier using maximum class probability as a confidence value for the AUROC score, serving as a proxy for supervised methods.

Once the general applicability of the methodology on all combinations has been investigated, we will look for the lower limit of Z-OOD detection with a smaller number of data sets. A maximum of five data sets and three corruptions, Gaussian Blur, Snow and Brightness, are selected and applied to images with increasing severity. The Out-of-Distribution Detection itself will again be in the near-OOD setting. It will be measured how hard the task becomes in ZSA, CLP and MMD. MCM and ZOC will be compared in this setup.

5.1. Exploratory Search for Far-OOD

Setup

In this study, the ViT-B/32 Clip model is employed for the current and subsequent experiments. We examine the MCM methodology, wherein features are generated for each image using the vision encoder of the Clip model. The last layer's output, which produces a 512-dimensional feature vector, is utilized for this purpose. Each known label (i.e., ID label) is inserted into the chosen template "This is a photo of a TOKEN." For example, the label "dog" transforms into "This is a photo of a dog" and serves as the label

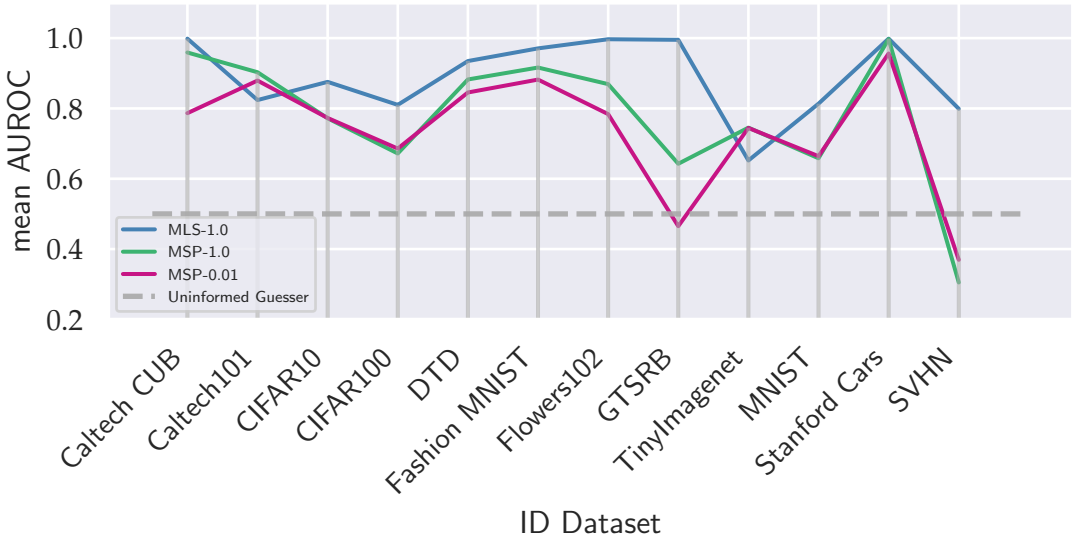


Figure 5.1.: Comparison of different MCM strategies for far-OOD. The y-axis shows mean AUROC scores over the 11 runs with the ID dataset shown on the x-axis. MSP with two temperatures is compared to MLS. The dotted line shows the worst possible outcome, an uninformed guesser. The lines do not indicate a dependency between plots, but are used to improve the comparability between the methods

representation. The text encoder of the selected Clip model then converts this sentence into a 512-dimensional vector. The token at the [EOS] position of the last layer is employed as the feature representation (see Section 4.3.1 for more details of the sentence representations).
Subsequently, the cosine similarity between an image and each label is calculated. The output (*logits*) serves as the basis for determining whether an image is ID or OOD, i.e. the similarity is over a task-specific threshold or below. This output is utilized directly to ascertain the maximum AUROC score (Maximum Logit Score (MLS)) and is converted into probabilities using the softmax function, which is then employed as the prediction (Maximum Softmax Probability). The label for OOD samples is 1, while that for ID samples is 0. Temperature scaling is applied to both methodologies, with a scale of 1.0 corresponding to the original score, 0.01 to a classifier that sharpens the softmax distribution (i.e., assigns higher probabilities to the most confident prediction), and 100.0, which produces less certain predictions. All results are reported from experiments conducted on the test split of the respective datasets.

Results

Figure 5.1 shows the mean AUROCs for different methods and temperatures in MCM [65]. Each point represents the ID dataset. The AUROC is the average of the scores of this dataset as ID and each other dataset as OOD. This means that it is not directly observable how well one dataset works with one another in this graphic. It is used as an indicator

ID Dataset	AUROC >99%
Caltech CUB	11 (100%)
Stanford Cars	11 (100%)
Flowers102	10 (90.9%)
GTSRB	10 (90.9%)
Fashion MNIST	3 (27.3%)
CIFAR10	1 (9.1%)
DTD	1 (9.1%)

(a) Grouped by ID datset

OOD Dataset	AUROC >99%
Flowers102	5 (45.5%)
Stanford Cars	5 (45.5%)
Caltech CUB	4 (36.4%)
CIFAR100	4 (36.4%)
DTD	4 (36.4%)
Fashion MNIST	4 (36.4%)
SVHN	4 (36.4%)
Caltech101	3 (27.3%)
CIFAR10	3 (27.3%)
GTSRB	3 (27.3%)

(b) Grouped by OOD Dataset

Table 5.1.: The number of ID / OOD combinations, where the AUROC is >99% which are in sum 47 of 132 (35%) combinations. In brackets, the share of the total number of experiments is displayed. MCM is used with MLS and temperature $\tau = 1.0$

to see which of the temperatures and methods performs best. It turns out that MLS has a higher score on average by over 10 points (0.89 vs < 0.78) than the best MSP setting with $\tau = 1.0$. Higher temperatures in MCM ($\tau = 100$) do not improve or even change the results, thus no higher temperatures are included.
Furthermore, there is no data set where the methodology does not work (i.e. is not better than randomly guessing) on average. A more detailed analysis shows that only the combination TinyImagenet - Fashion MNIST is very close to random guessing with AUROC $= 0.482$. There are also five combinations that are worse than 0.5. There is no trend for a dataset, the combinations include CIFAR10, Caltech CUB, GTSRB, SVHN, MNIST and TinyImagenet.

The following statements from now on refer to MCM with MLS. Table 5.1 shows that 35% of the combinations achieve over 99% AUROC score. Especially the ID data sets Caltech CUB and Stanford Cars always score above 99%. Flowers102 and GTSRB still score over 90% of the time (Table 5.1a). Table 5.1b highlights the results from the perspective of the OOD dataset. Flowers102 and Stanford Cars are perfectly detected as OOD five times. No other dataset is more often perfectly detected. Over 50% of the combinations score better than 95% (67/132). Only one combination (SVHN - Stanford Cars) was perfectly solved. The measured metrics for the experiments are in the following ranges: MMD $\in [0.015; 0.680]$, CLP $\in [-3.073; -0.049]$, ZSA $\in [0.250; 0.898]$. Table 5.2 displays the correlations to the AUROC score. Across all experiments, there is no higher absolute correlation than $r = 0.383$.

Metric	**r** $\uparrow$
MMD	0.383
CLP	-0.046
ID ZSA	-0.086
OOD ZSA	0.114

Table 5.2.: Pearson correlation matrix for far-OOD

Interpretation

In this investigation of 12 datasets encompassing 132 different far-OOD combinations, the methodology demonstrated its ability to perform significantly better than random guessing in the majority of cases. In fact, it achieved near-perfect results in 35% of the instances. Based on these experiments, it can be concluded that Z-OOD detection is effective in most combinations and has minimal limitations (0.008% with no informed guess). The recommendations from the MCM publication [65] to utilize MLS instead of MSP and apply $\tau = 1$ appear to be highly beneficial, as they yield the best average results. This is especially evident for the GTSRB dataset, where the difference exceeds 40% on average. AUROCs below 0.50 are challenging to interpret, indicating that the classifier misinterprets the data. In practice, these predictions are often simply reversed, so the absolute difference to 0.5 is assessed. As there are only four instances, no further analysis is conducted, and the predictions are reported as they occur.

A detailed examination of the results reveals that the semantic shift datasets [96] are relatively easy to solve in all combinations. In other words, if all classes in the ID share high-level features (e.g., only birds, cars, flowers, or street signs), perfect recognition is practically achieved with this method, provided that the OOD dataset exhibits a semantic shift.

In conclusion, Z-OOD detection consistently outperforms an uninformed guesser in far-OOD tasks and virtually solves the tasks when a strong semantic shift is present. The selected metrics, however, are not suitable for measuring far-OOD difficulty. In this scenario, a (human) analysis of the semantic classes or contents is likely more effective, as semantic shifts can be easily detected. For far-OOD detection it can be said that these methods do generalize well, but not always.

5.2. Exploratory Search for Near-OOD

Setup

The foundational setup remains identical to the one described in Section 5.1. Furthermore, ZOC [21] is incorporated and compared to MCM [65] and a fine-tuned baseline. ZOC is trained according to the prescribed guidelines and employed for inference as delineated in Section 4.3. The generated logits are converted into softmax probabilities using the temperature $\tau = 0.01$, which is the recommended value for ZOC. The sum of the probabilities of all generated labels will be used as the prediction score.
For training the fine-tuned OOD detection, the training set of the dataset is utilized. The validation split is employed to search for the optimal hyperparameter settings.

In the near-OOD setup, twelve different datasets are employed: Each dataset is divided into ID and OOD, establishing a near-OOD scenario. 40% of the classes of a dataset are

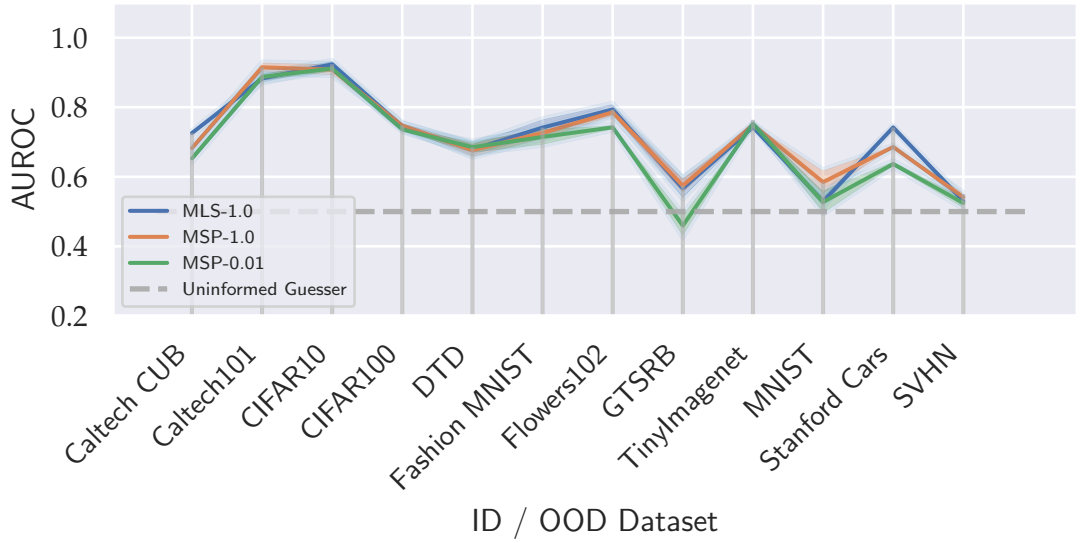


Figure 5.2.: Comparison of different MCM strategies for near-OOD. AUROCs are mean AUROC score averaged over 10 runs. For each run, the dataset classes are randomly split into 40% ID and 60% OOD classes. The shadows indicate the standard deviation. The dotted line shows the worst possible outcome, an uninformed guesser. The lines do not indicate a dependency between plots but are used to improve the comparability between the methods

designated as ID, while the remaining 60% are considered OOD. This process is repeated 10 times, and the average values, accompanied by the standard deviation, are reported. All reported results are derived from experiments conducted with the test split of each dataset.

Results

First, the comparison of the MCM strategies was carried out. Figure 5.2 displays the results. The differences between the individual strategies are 0.03 points ($0.69 - 0.72$) and have a Pearson correlation coefficient of $r > 0.95$. The average standard deviation is nearly identical at 0.037.

The comparison of the methodologies is shown in Figure 5.3. It shows that ZOC performs better on average than MCM with one point. These results also correlate with $r = 0.98$. The largest difference is 0.05 points (Fashion MNIST). The methodology has AUROCs $<= 0.55$ on the GTSRB, MNIST and SVHN datasets. The CIFAR10 dataset achieves the best score for the zero-shot methods with AUROC > 0.9 for both methodologies.

ZOC has the highest standard deviation of all methodologies with 0.048, the fine-tuned linear regression 0.033 and MCM 0.032. This is particularly evident in the GTSRB (std $= 0.110$) and MNIST (std $= 0.133$). Thus, there are several runs with those datasets,

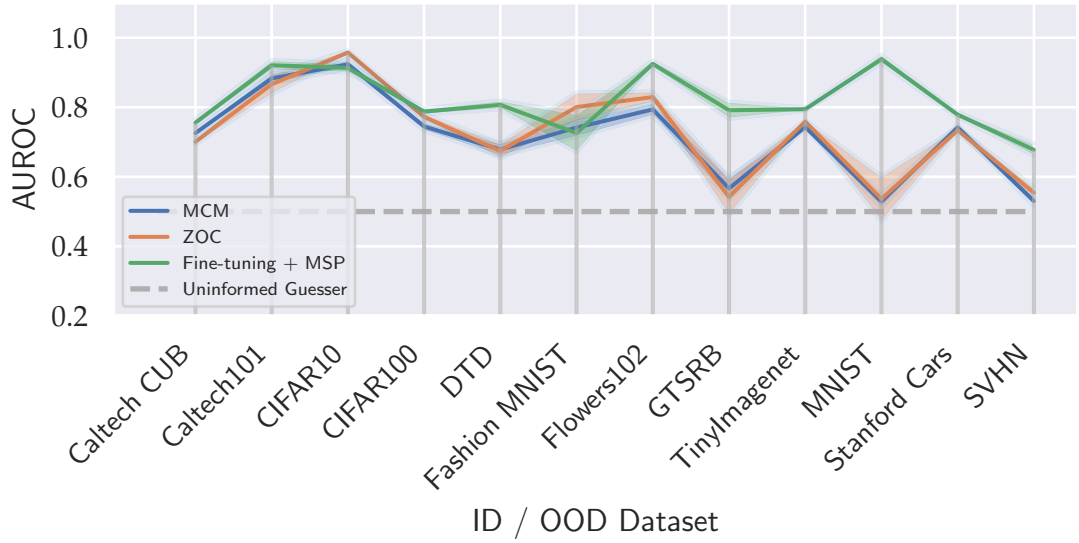


Figure 5.3.: Comparison of OOD detection strategies for near-OOD. AUROCs are mean AUROC scores averaged over 10 runs. For each run, the dataset classes are randomly split into 40% ID and 60% OOD classes. The shadows indicate the standard deviation. The dotted line shows the worst possible outcome, an uninformed guesser. The lines do not indicate a dependency between plots but are used to improve the comparability between the methods

where the AUROC is not better than the random guesser. The fine-tuned baseline has an average AUROC of 0.818. The two Z-OOD detection score 0.727 (ZOC) and 0.717 (MCM). Per dataset, the AUROC of the baseline is higher or on par on every dataset except Fashion MNIST. The highest difference is on the MNIST dataset, with a difference of 0.402. Looking only at the results on which ZOC performed significantly better than 0.5, the fine-tuned classifier has a higher AUROC by 0.350.

	AUROC ↑
MCM	**0.717** $\pm$ 0.032
ZOC	**0.727** $\pm$ 0.048
Fine-Tuned	**0.818** $\pm$ 0.033

Table 5.3.: Near-OOD results in AUROC score with the standard deviation

The correlation of the selected metrics to the AUROC is higher than in Section 5.1. CLP has a Pearson correlation coefficient $r = -0.94$, ZSA $r = 0.93$ and MMD $r = 0.71$. The correlations of CLP and ZSA are visualized in Figure 5.4.

Interpretation

In the near-OOD setup, the large disparities observed with temperature scaling in the far-OOD setup are not present. Instead, all values exhibit nearly identical performance. Moreover, there is only a minor difference between MLS and MSP. Utilizing the classifier with $\tau = 1.0$ continues to yield the best average results, corroborating the conclusions from MCM.

In the near-OOD setting, the two Z-OOD detection methods demonstrate highly sim-

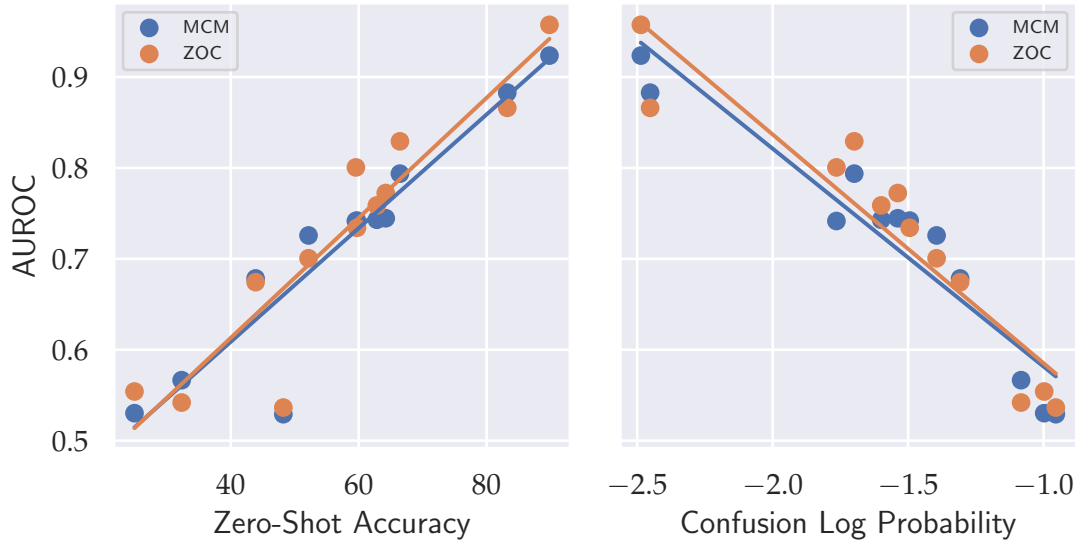


Figure 5.4.: Correlation of ZSA (left) and CLP (right) to the AUROC for zoc and MCM. The lines are fitted using least-mean-square regression and only for orientation. All points on a line would indicate perfect linear correlation

ilar performance. Although ZOC performs marginally better on average, it exhibits a notably higher standard deviation in the results. The findings from MCM, where MSP outperformed ZOC in specific experiments, are not substantiated in the near-OOD setup.

Barring a few exceptions (Fashion MNIST, CIFAR10), a fine-tuned methodology proves to be significantly superior and more reliable, as it detects all outliers considerably better than a random guesser could. The method has also the lowest standard deviation, i.e. it performs more similar independent of the ID/OOD split. The high standard deviation of the other methods indicates, that there are some split combinations, where the outlier detection is harder for them. It is worth noting that up to 110,000 (TinyImagenet) training samples were used in these cases to achieve these results. Excluding the exceptionally poor results of ZOC on the three datasets where it fails to reliably detect outliers, the difference is minimal. The selected metrics exhibit strong (MMD) to very strong (CLP, ZSA) correlations for the zero-shot methodologies, suggesting their utility for predicting the difficulty of an OOD detection task.

Additional experiments investigating the influence of temperature are presented in Appendix A.3. The results indicate that while temperature has a significant impact on performance in far-OOD experiments [65], its impact is considerably smaller in near-OOD experiments. For very low-temperature values ($\tau < 0.01$), the performance decreases by approximately 2 points but stabilizes thereafter. For ZOC, the default setting of $\tau = 0.01$ works best. The AUROC declines to 0.5 as the temperature increases.
Besides the temperature scaling insides, the Appendix also shows insights into the usage of prompts in Figure A.4: Using the default prompt, as stated in Chapter 4 in combination with MLS achieves the best results for all combinations.

5.3. Exploratory Search on Robustness of Zero-Shot OOD detection

Setup

The setup for all models and runs remains consistent with the two previous sections. ZOC and MCM are evaluated on a selected set of five datasets (Caltech101, Caltech CUB, Flowers102, GRSRB, Stanford Cars) and three corruptions each (Gaussian Blur, Brightness, Snow). These corruptions are applied at three increasing severity levels (1, 3, 5), utilizing the severity levels from the original publication [35]. Due to computational constraints, ZOC is executed for only five different splits in each setting. In total, 45 distinct setups are examined and compared to the 12 non-corrupted ZOC baselines from Section 5.2.

Results

Figure 5.6 shows the influence of the corruptions on the difficulty metrics with increasing severity. All corruptions do increase the difficulty, as evidenced by the metrics moving in the direction that is considered more challenging in the literature [56, 104, 96].
The ZSA is reduced by 67% on average ($0.570 \rightarrow 0.184$) from the normal image to the highest corruption. The MMD decreases by 55% ($0.011 \rightarrow 0.005$), and CLP decreases by 32% ($-1.564 \rightarrow -1.237$). The AUROCS of the two MCM strategies are very similar (MLS: AUROC $= 0.670$, MSP AUROC $= 0.669$ and a Pearson correlation $r = 0.98$). This correlation remains at the same level when comparing both strategies grouping by dataset, by corruption and by severity. Therefore, only MLS is used for further comparisons. Figure 5.5 demonstrates the decline in AUROC scores as the severity of corruptions increases, reaching the lowest point at severity level 5 for all combinations. The bottom graph reveals that the corruptions cause a consistent deterioration of the average AUROC scores across all datasets, with Gaussian Blur having the lowest AUROC value of 0.574. The top graph of Figure 5.5 displays, when grouped by datasets, that the datasets with poorer performance initially experience smaller declines. In contrast, Caltech101, which exhibits the highest score without corruption, maintains the highest score at every level while increasing its distance from other combinations (0.037 to 0.161). GTSRB, the poorest-performing dataset, does not exhibit significant improvement beyond the random guesser (AUROC $= 0.5$) after the application of any corruption except for Brightness. The AUROC score ranges between 0.515 and 0.567, with a standard deviation of 0.06.

MCM and ZOC are correlated with $r = 0.975$. ZOC has the lowest scores on the same 9 combinations as MLS. The AUROC scores differ by 0.04 (ZOC: 0.685, MLS: 0.681) and both have a standard deviation of 0.04.

All metrics continue to show a correlation with the AUROC score. ZSA has the lowest

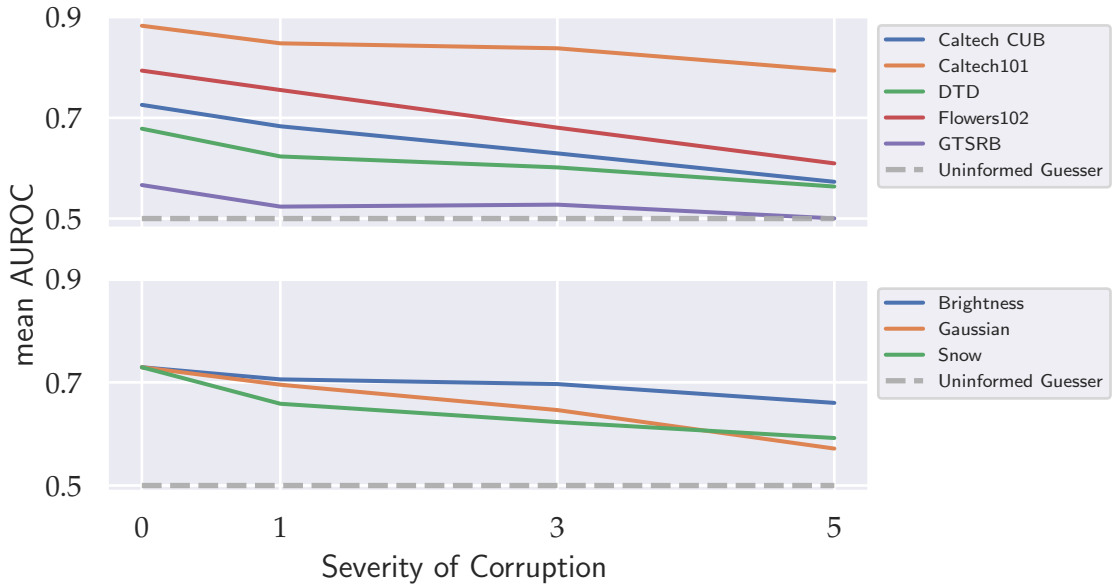


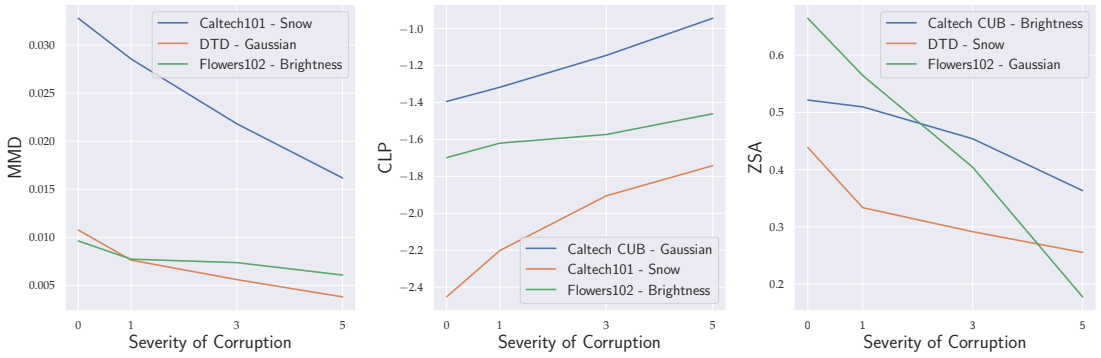
Figure 5.5.: The influence of the corruptions on the AUROC scores on the datasets grouped by dataset (top) and corruption (bottom) (MCM). Both graphs show the linear decline on the AUROC score with increasing corruption severity

correlation at $r = 0.41$, MMD has a correlation of $r = 0.72$, and CLP has the highest absolute correlation at $r = -0.92$. When considering only the 9 combinations that do not work, the ZSA correlation increases to $r = 0.57$, while the correlations of MMD and CLP decrease to $r = 0.03$. There is no clear lower limit at which Z-OOD detection no longer works. All correlations are almost identical for ZOC (ZSA: $r = 0.32$, MMD: $r = 0.71$, CLP: $r = -0.92$).

Interpretation

The incorporation of corruptions has heightened the difficulty of the tasks, which is evident in all assessed metrics. Thus, corruptions were useful to access the robustness of the methods, as intended to answer the second research question.
On one hand, this reveals that the methodologies, provided they function properly (excluding the datasets that failed in Section 5.2), exhibit robustness to mild and moderate corruptions. Gaussian Blur had the most substantial impact, while Brightness had the least. Upon examining examples of corruptions, these findings align with human perception. Particularly for small images, Gaussian Blur nearly distorts them to the point of being unrecognizable. In contrast, Brightness corruption generally retains high-level features, mainly reducing colour and contrast.

With the exception of GTSRB, the majority of the combinations are solvable by the tested methodologies. This suggests that there is robustness against corruptions. Additionally, the ability to handle increased difficulties, as measured by CLP and MMD, demonstrates robustness in this aspect as well. The results do not provide a clear in-



(a) MMD ↓ (b) CLP ↑ (c) ZSA ↓

Figure 5.6.: The influence of corruption on the difficulty metrics. Each graphic is captioned with the metric displayed. The arrow indicates, in which direction the difficulty of a task theoretically increases

dication of which method is more robust or better suited, as the marginally superior performance of ZOC is counterbalanced by a higher standard deviation.

The gathered metrics maintain a strong correlation, making them appropriate for estimating task difficulty. However, an exact lower threshold at which the Z-OOD detection methods cease to function could not be determined.

5.4. Discussion

The discussion examines the main findings and critically examines the methodology used, highlighting potential shortcomings. This section concludes with ideas and suggestions for future work.

5.4.1. Findings

On average, the Z-OOD detection is proficient at reliably detecting far-OOD outliers. However, there are critical systems where even better detection scores may be required, thus more sophisticated solutions can be necessary, e.g. for self-driving cars. Particularly when there is a strong semantic shift, outliers can be detected almost perfectly. This occurs when the ID images are highly similar, as with the Flowers102 dataset. The results for TinyImagenet are a notable exception, as it consistently achieves an AUROC of just 0.8 in all experiments, indicating a clear potential for improvement. This dataset consists of 200 partly disparate classes, which increases the probability of semantic overlaps with other datasets, e.g it is likely that the dataset contains images of cars, which are also a class in the Stanford Cars dataset or a bird class from the Caltech CUB class. This can also influence the results, as these cases can be wrong labelled inliers.

Consequently, many far-OOD methods can be addressed with an untrained method. It is recommended to perform benchmarks with the TinyImagenet dataset, as there is significant room for improvement. Nonetheless, the dataset's diverse classes raise the

question of whether this is a realistic use-case scenario.

In the chosen near-OOD setting, the results are considerably worse, with no combination achieving scores above 0.99. The comparison of the two Z-OOD detection methods reveals minimal differences, prompting the question of whether the overhead of ZOC (larger architecture, training of the caption generator) is worth it in practice. Training a caption generator, in particular, is highly resource-intensive.
However, the Z-OOD detection methods can also solve many challenging OOD detection cases much better than random guessing, but three out of twelve datasets cannot be solved: MNIST, SVHN, and GTSRB. The first two are highly similar to an OCR task and also pose significant problems in Clip classification, so it was not unlikely that this would transfer to OOD detection. For GTSRB, there are no direct indications without further analysis as to why it does not work, but the classification is already challenging here. The methods appear to be highly dependent on the Clip classifier, which is also reflected in the correlation with the ZSA.

The linear baseline is more reliable, as it performs significantly better than random on every dataset, albeit with training data. It must be acknowledged that Z-OOD detection is not yet at that level. However, in cases where Z-OOD detection works well, the results are quite similar.

In the near-OOD setting, the metrics could also be effectively used to determine the difficulty of the task. Thus, future Z-OOD detection research could utilize these to provide an estimate of the chosen setting and make the results more interpretable. This aspect is often overlooked. It has been demonstrated that a carefully calibrated MCM works almost identically to ZOC, which was not apparent from the two publications, where the methods are proposed [21, 65].

The limits of Z-OOD detection can be illustrated using corruptions. The correlation of the metrics with the boundaries once again demonstrates their usefulness in assessing a task. Unfortunately, it remains unclear how other OOD detection methods perform in this setting. Nevertheless, both Z-OOD detection methods can still handle many slight corruptions in images. A strict boundary, when the methods no longer work, was not found for any of the metrics. CLP appears to be promising, as it has the highest correlation to tasks. The accuracy of the closed-world classifier is also a good indicator, but, as accuracy is overall more dependent on the number of classes, it is less useful.

5.4.2. Potential Issues and Shortcomings

The chosen near-OOD setting is highly susceptible to chance, as evidenced by the high standard deviations. This means that the split in which a dataset is divided into ID and OOD can significantly impact the results, sometimes by over 10%. This reduces the over-

all significance. Particularly with datasets were outliers are poorly recognized (i.e. close to an AUROC score 0f 0.5), this means that on some runs, the outliers are not detected better than by a random guesser. In other words, it cannot be ruled out that datasets close to the boundary may or may not work after all. This methodology is also very resource-intensive. For most methods, the entire training must be repeated for each split to avoid data leakage. For example, training GANs [28], which traditionally takes a long time, may not be feasible. This was also the reason why there are no comparative values of fine-tuned methodologies for the far-OOD setting.

Furthermore, both CLP and MMD must be critically examined: First, a direct comparison in related work is only possible if the same Clip model is used. Second, this is only useful in research, as both MMD and CLP require access to outliers. This is precisely where the Z-OOD detection methodologies are intended to help. The only metric that remains is the ZSA, which at least only needs access to training data from the ID distribution.

Regarding the corruptions, it can be argued that the scenario is not directly transferable to a real world application, as all images were consistently altered with the same corruptions of the same intensity. A mixture of corruptions would better reflect the reality. The applied temperature scaling can also be viewed as a task-specific adjustment where data leakage occurs, i.e. instances of the dataset were used to improve the scores, despite claiming to be a zero-shot method. We have considered them as separate methodologies and carried them out accordingly. However, a comparison in this context is no longer in the true zero-shot setting.

The last point pertains more to the methodologies themselves and not to this work: for ZOC, the term "zero-shot" is interpreted in a way that no specific domain adaptation is made, which is true. Nevertheless, the caption generator had to be trained with significant resource expenditure. It cannot be ruled out that there is data leakage, and the domain ends up in the training. This is a general problem with foundation models and will not be discussed further here. However, to compare both Zero-Shot Out-of-Distribution Detection methods, we had to conduct training for one.

5.4.3. Future Work

In future work, the methodologies should be tested with images that were not included in the pre-training of any of the models. Although Foundation Models generalize very well, there are use cases where, for instance, the classification does not perform effectively. This was evident in this work, such as with the SVHN dataset. Further research could clarify whether generalization also applies in these cases, shedding light on the question of data leakage in the methods.

To better assess robustness, a comparison with other methodologies should be conducted in future work. Additionally, mixing corruptions and employing stronger ones could potentially reveal a limit. This limit could then serve as a benchmark against which improvements can be measured. The large number of experiments and the comparison with the original publications highlight that the selection of datasets for ID and OOD can influence which methodology is considered superior. A set of standard datasets and setups, as suggested by [96], would address this issue.

6. Zero-Shot Methods as Basis for Few-Shot OOD Detection

In this chapter, after examining the generalization capabilities and limits of Zero-Shot Out-of-Distribution Detection (Z-OOD detection) methodologies, we explore whether the core idea of these methodologies can be improved to achieve the performance of methods that have full access to training data. Thus, the third research question is addressed. The goal is to leverage the Z-OOD detection methods by applying State-of-the-Art (SOTA) domain adaptation methods to the Clip backbone and investigate if the adaption to a specific domain also transfers to Out-of-Distribution Detection (OOD detection) performance. This would make these methods more versatile for real-world usage.

The unique selling point of zero-shot methods is that they do not require any data or domain knowledge in the form of training data or otherwise, apart from semantically meaningful class names. Therefore, they can achieve impressive results and compete with fine-tuned baselines in many combinations, even in difficult ones. Since the release of large vision-text models like Clip, there has been significant interest in few-shot domain adaptation, aiming to improve the performance of a model with as few data examples as possible, usually for image classification [26, 111, 109, 12]. The TIP-Adapter architectures [111] stand out, as they have few or no trainable parameters, and thus bring significant improvements with minimal training effort and very few samples (<64). The TIP-Adapter methodology is particularly noteworthy because domain knowledge is added to the foundation knowledge of Clip via a residual connection. This means that the pre-trained Clip model, as used in Section 5, remains unchanged. As a result, both Z-OOD detection methodologies, MCM and ZOC, can be used with it without additional overhead besides the adapter. This would also have been the case for other adapter architectures for MCM, while the adjustments for ZOC would have been greater (see Chapter 4.3).

For the adaptations, we follow these steps in this chapter: First, we use TIP-Adapter [111] to adapt the Clip model to the respective domain (represented by a dataset). We set the boundary at 16 data points per class of a dataset, which is a number used for the main results in related works [111, 26]. The adaptability is measured by comparing the accuracy of the closed-world classifier with and without adaptation. Next, the domain adaptation is incorporated into each of the methodologies and tested in a series of experiments to determine whether this adaptation also transfers to Out-of-Distribution

Detection. The goal is to improve the results from Section 5.2 using this methodology and to compare how close it comes to high-performing fine-tuned OODD benchmarks.

In summary, in this chapter, we test domain adaptation with the two adapter models from TIP-Adapter, TIP and TIP-f. Then, we integrate these adapters into MCM (T-MCM) and ZOC (T-ZOC) and test their performance on the 12 near-OOD settings from Section 5.2. Finally, we compare them with a fine-tuned benchmark. For this purpose, the methodology of the baseline from Section 5.2 is used, but with the Clip ViT-L/16@336, the largest and best-performing Clip model [75].

6.1. Few-Shot Domain Adaption

In the following section, few-shot domain adaption for the Clip-based OOD detection method is investigated. First, domain adaption using TIP-Adapters [111] is applied and then it will be investigated, whether the adaption transfers and improves the OOD detection.

Setup

As the foundation for all experiments, the ViT-B/32 vision encoder serves as the base model. Both MCM and ZOC employ the same models and hyperparameters as described in Chapter 5.

Regarding the TIP and TIP-f approaches, adapters are trained utilizing up to 16 instances from the respective ID training dataset splits, maintaining equal numbers per class, up to a maximum of 16 instances. In exceptional cases (e.g., Caltech101), fewer than 16 instances may be used, when there is a class with less than 16 images. For each class then, the minimum number of instances per class is used. No outlier exposure is implemented. To generate TIP cache models, 10 iterations of random augmentations are applied. The TIP-f method trains the cache model weights for 20 epochs, using the AdamW optimizer [61] with a learning rate of 0.001 and an epsilon value of 10^{-4}. Optimal alpha and beta values are determined through a hyperparameter search ranging from 0.1 to 5. The Out-of-Distribution Detection (OOD detection) results are reported as AUROC scores. For TIP-adapted ZOC (T-ZOC), besides the summation of OOD label probabilities, the maximum softmax probability is also examined as a prediction. Each experimental setup is executed ten times with random ID/OOD splits, and the reported values represent the averages of these iterations. All 12 datasets are utilized in these experiments.

6.1.1. Results

Table 6.1.: Accuracy comparison of different classifying methods, averaged over all datasets. The Accuracy is displayed in percent. Accuracy is shown in percent. The best method result is highlighted.

	Accuracy ↑	Diff to baseline
Zero-Shot	57.4	-
TIP	**66.9**	9.57
TIP-f	66.7	9.38

Table 6.1 presents the classification outcomes for the 12 datasets utilizing pure zero-shot learning and two domain adaptation techniques. The highest accuracy is achieved by TIP at 66.9%, whereas TIP-f demonstrates an accuracy of 66.7%, surpassing the zero-shot baseline by more than 9 percentage points. The average values for the two adapter models are quite similar; however, disparities emerge when examining individual results. TIP without fine-tuning exhibits 11% higher accuracy on Flowers102 and 5.5% on MNIST. Conversely, the fine-tuned model displays an increase of 6.3 points on GTRSB and 7.9 on SVHN. The most significant differences are observed between the baseline and TIP on the MNIST dataset, with a 21.4-point increase (+45%). The largest discrepancy for TIP-F is evident on GTRSB, with an 18.9-point improvement (+58.4%). The smallest differences for the adapters are found on CIFAR10 (TIP-F+0.89 points, 0.9%) and SVHN (TIP +0.5 points, 0.2%). Comprehensive results are supplied in Appendix A.4.

Applying domain adaption to MCM

Table 6.2 displays the results averaged over datasets. Part 6.2a shows an improvement of 2.1 points compared to the baseline MCM using T-MCM and 1.2 using T-ZOC-f. In Table 6.2b, only the results of the datasets with an accuracy improvement of more than 10% due to the adapters are shown. This threshold is selected because it represents a significant improvement attributed to domain adaptation, which indicates effective domain adaptation. The aim is to measure the transfer of successful domain adaptation. Therefore, minor and negligible improvements are excluded, as no transfer can be concluded from these results. Here, the difference between the baseline and T-MCM is 6.2 points. The finetuned adapter version, T-MCM-f is +2.3 compared to the baseline.

Table 6.2.: Mean AUROC scores for different MCM based OOD detection methods. The standard deviation notes the mean over the standard deviation from the ten runs per dataset

	AUROC ↑
MCM	0.69 ± 0.0415
T-MCM	$\mathbf{0.722} \pm 0.035$
T-MCM-f	0.713 ± 0.035

(a) All 12 datasets

	AUROC ↑
MCM	0.649 ± 0.060
T-MCM	$\mathbf{0.711} \pm 0.049$
T-MCM-f	0.672 ± 0.060

(b) The five datasets with +10% accuracy with domain adaption

Table 6.3.: Mean AUROC scores for different ZOC based OOD detection methods. The standard deviation notes the mean over the standard deviations from the ten runs per dataset

	AUROC ↑
ZOC	0.727 ± 0.044
T-ZOC	$\mathbf{0.737} \pm 0.031$
T-ZOC-f	0.733 ± 0.042

(a) All 12 datasets

	AUROC ↑
ZOC	0.675 ± 0.071
T-ZOC	$\mathbf{0.719} \pm 0.038$
T-ZOC-f	0.681 ± 0.069

(b) The five datasets with +10% accuracy with domain adaption

Applying domain adaption to ZOC

T-ZOC and T-ZOC-f both achieve a higher AUROC score than the standard ZOC architecture. Table 6.3a displays these values. T-ZOC (+1.0) has the highest AUROC. The finetuned adapter strategies show less improvement (+0.6) but are still higher than the baseline ZOC. Looking only at the datasets with good domain adaption, the difference between ZOC and both T-ZOC (+4.4) and T-ZOC-f (+1.6) is higher.

Comparison of different strategies

Figure 6.1 juxtaposes various approaches on datasets that demonstrate notable enhancements (+9.9%) in accuracy attributable to adapters. The comparison reveals that no single method consistently outperforms the others. In only one of the five cases (MNIST), the MCM baseline surpasses one of the adapted variants. For the GTSRB dataset, both techniques yield superior OOD detection performance compared to the uninformed guesser, with a 10% improvement for T-ZOC. In Figure 6.2, we present a comparative analysis of

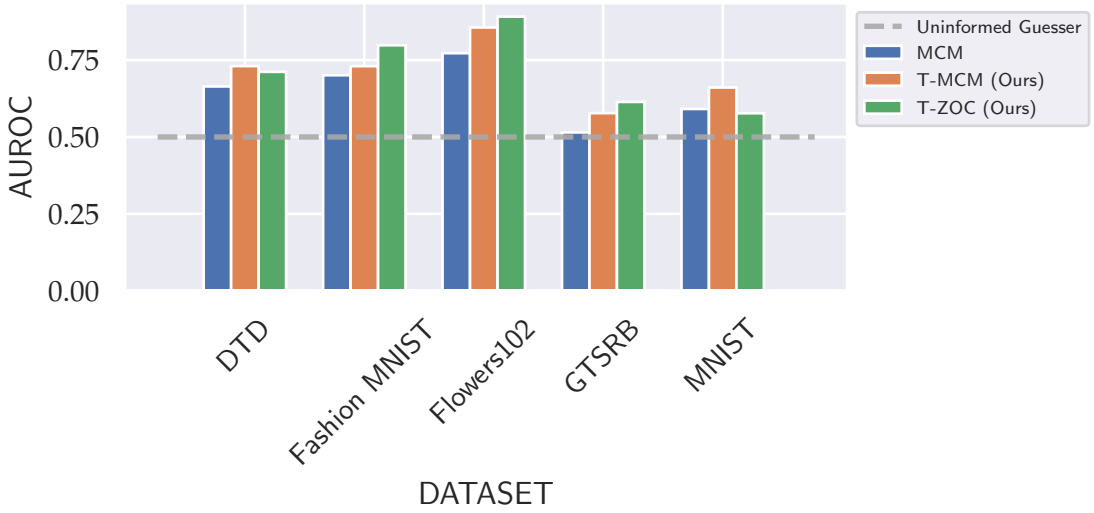


Figure 6.1.: Bar chart showing the baseline (MCM) to both adapted strategies for datasets with good domain adaption

the methods with the highest AUROC values and the fine-tuning benchmark. The results reveal that in the context of Fashion MNIST dataset, one of the evaluated methods, namely T-ZOC, demonstrates better performance with a $+5.7\%$ improvement over the fine-tuned model. However, for all other datasets, the fine-tuned model outperforms all the other evaluated methods, with the maximum improvement of $+29.1$ observed for the MNIST dataset. Full results are in Appendix A.2.
The few-shot domain adaption also significantly outperforms a fine-tuned linear baseline. The linear probe method is not able to detect outlier until at least 16 images per class are available, while both adapted strategies perfrom at least on par with the zero-shot baseline and improve with more available training data. See Appendix A.6 for the results on selected datasets.

6.1.2. Interpretation

On average, domain adaptation significantly outperforms the zero-shot baseline in terms of accuracy. However, the extent of improvement varies across datasets, with SVHN and CIFAR10 exhibiting almost no improvement while Flowers102 demonstrates strong adaptation. The performance of OOD detection also varies depending on the quality of domain adaptation. When domain adaptation is effective, with a noticeable improvement of at least 10%, the resulting performance gains tend to transfer to OOD detection as well. Specifically, on average, the improvements in OOD detection accuracy can increase by up to 4.4% with only 16 training images per ID class. Nevertheless, compared to a benchmark, there remains a significant difference for most datasets, with only Fashion MNIST surpassing the baseline. A linear probe method trained on 16 instances is worse on each dataset compared to both, the zero-shot baseline and the adapted methods.

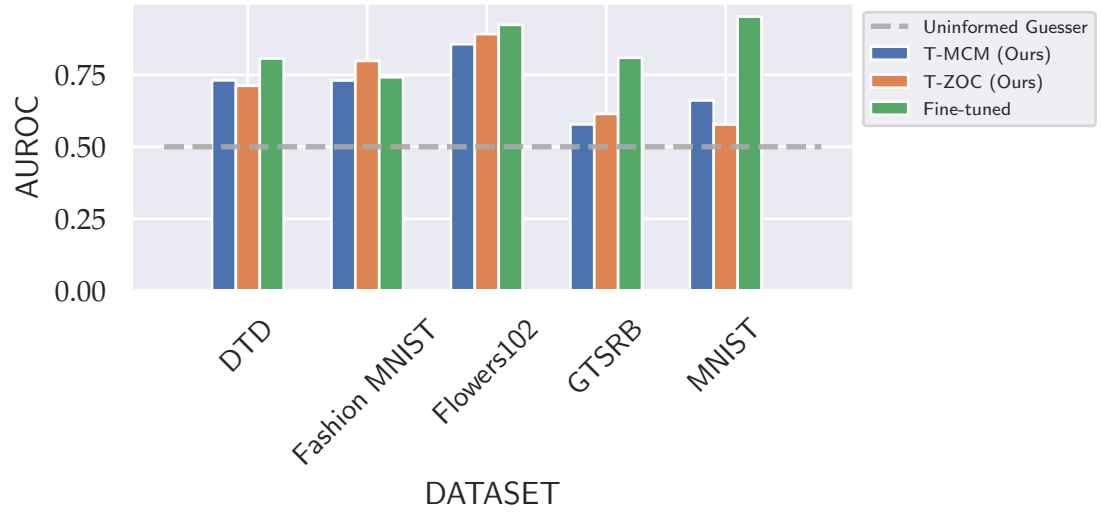


Figure 6.2.: Bar chart showing the finetuned benchmark OOD detection (MCM) to both adapted strategies for datasets with working domain adaption

In other cases, few-shot adaptations show significantly higher performance and thus emerge as superior alternatives to the zero-shot baseline.

6.2. Discussion

This section discusses the results from the conducted experiments on few-shot domain adaption.

6.2.1. Findings

The findings of this study suggest that Clip domain adaptation can enhance the performance of zero-shot methodologies, but only when the adaptation is highly effective and significantly improves accuracy. The effectiveness and reasons, when and why this is the case, are not further investigated. When this is the case, the benefits also extend to OOD detection methodologies. Notably, both ZOC and MCM approaches are sensitive to the quality of Clip embeddings. If the embeddings are distinguishable and accurately classified, this improvement also transfers to OOD detection methodologies.

However, the study also highlights that in most cases, this method falls short of the benchmark performance, with the exception of Fashion MNIST and, to a lesser extent, CIFAR10. Nevertheless, when only a few data points per class are available, these methods can still be improved using Clip domain adaptation. These results are in line with recent research showing a close relationship between zero-shot accuracy and OOD detection performance [96].

6.2.2. Potential Issues and Shortcomings

This study covers several research areas, including Z-OOD detection, OOD detection, few-shot learning, and Clip domain adaptation. As such, there may be some thematic overlaps and challenges in appropriately comparing the results. The extension of Z-OOD detection methodologies is a new research area, making it challenging to compare with related work. In this study, the comparison was made with fully supervised methods in the OOD detection research area, with a focus on fast and reliable methods that deliver good results without high complexity. However, there may be other benchmarks or research areas that could be explored further, such as few-shot learning.

One potential issue with few-shot domain adaptation is that the results of TIP adapters depend on hyperparameters and an optimization process, which assumes additional data on which the methodology is tested. This may not be the case in practice and can affect the few-shot nature of the approach.

Another point to consider is that all of the methodologies mentioned in this study rely on Clip and are highly dependent on the extracted features. Therefore, there may be a limited perspective on the OOD detection research area if further methodologies without Clip are not included. To address this, a different Clip backbone was used in the benchmark comparison in this study, which has significantly different classification results on the datasets. These results can be found in Appendix A.5, where the closed-world classification results of all methods used in this thesis are presented.

6.2.3. Future Work

Future work should address the issues and shortcomings identified in this study. Specifically, a comparison with other few-shot OOD detection methods should be conducted to evaluate the performance of the proposed approaches in comparison to other SOTA methodologies. Additionally, the effectiveness of fine-tuning in few-shot learning should be explored further to understand whether it is a suitable approach for this task.

Further methods beyond TIP adapters should be explored to improve domain adaptation. This study highlights that the proposed methodologies are suitable for enhancing domain adaptation, and exploring additional methods could lead to further improvements.

To enhance the ZOC approach, improvements could be made to the caption generator, which has not been further adapted for the task. Optimizing the caption generator for the few-shot setting could potentially improve its performance on this task. Additionally, since the caption generator is based on transformers, adapter types could also be used to improve its performance, as demonstrated in recent work [73, 65].

7. Conclusion

This thesis delved into the potential and limits of Zero-Shot Out-of-Distribution Detection (Z-OOD detection) for image classification, using novel approaches that leverage the zero-shot classification capabilities of recent multi-modal architectures. The work starts with large-scale far-OOD experiments using MCM, followed by a challenging near-OOD comparison of ZOC and MCM. Throughout the investigation, various hyperparameters were examined and assessed for all setups. In Section 5.3, the robustness of the proposed method was tested under different conditions, such as image corruption, and attempts were made to determine the lower bound of the method. Additionally, this study explored correlations with difficulty metrics from Out-of-Distribution Detection (OOD detection) and assessed their predictive power.

Furthermore, the thesis aimed to understand whether advancements in domain adaptation methods could be transferred to OOD detection. The methodology was tested in a few-shot setup and compared against benchmark results to accomplish this. The following sections detail the key findings by addressing the research questions and elaborating on additional contributions made by this work. In the concluding section, recommendations for future research are provided, drawing on the insights gained from the experimental chapters in this thesis.

7.1. Contributions

Research Questions Answered

A brief discussion of the research questions will be presented, followed by their respective answers in the context of this thesis.

RQ. 1 **Is the performance of current Z-OOD detection methodologies generalizable, i.e., transferable to other datasets and use-cases?**
The results demonstrate that Z-OOD detection is generally effective, particularly in far-OOD scenarios. In near-OOD cases, certain challenges arise where the underlying Clip model has difficulties in classification, leading to inherited problems in the methods. Apart from written number datasets, the methods were successful in all other dataset combinations.

RQ. 2 **Where are the boundaries of Z-OOD detection methods with respect to different difficulty metrics for good detection?**
The methods were found to be quite robust, handling mild to medium corruptions and consistently performing significantly better than random guessing. However, severe corruptions led to method failures. No other meaningful metric-based threshold was identified beyond which the methods ceased to work.

RQ. 3 **How do the Z-OOD detection methods compare against traditional State-of-the-Art OOD detection methods?**
A fully fine-tuned linear probe on top of a pre-trained Clip model on full datasets proved superior, primarily due to the absence of blind spots associated with Z-OOD detection when dealing with images of digits. When Z-OOD detection performed well, it was comparable to traditional methods. Additionally, by incorporating domain adaptation for few-shot OOD scenarios, the methods demonstrated potential for improvement with successful domain adaptation.

Further Contributions

Alongside the research questions, this thesis made other contributions:

- As the results showed, that far-OOD is a task which is solved for many combinations without any fine-tuning, we showed that the focus of research should be on hard near-OOD and robustness, as zero-shot methods can already effectively detect far-OOD instances.
- A large-scale comparison of both Z-OOD detection methods revealed that ZOC performed better on average but with higher variance. MCM with Maximum Logit Score was found to be superior, which is consistent with existing literature.
- The influence of temperature scaling was tested, revealing that no scaling (scaling with $\tau = 1.0$) was optimal for MCM and a scaling of $\tau = 0.01$ was best for ZOC across all scenarios.
- This thesis proposed T-MCM and T-ZOC as domain-adapted few-shot OOD detection methodologies. T-MCM, in particular, is a lightweight method with fast adapting speed. If domain adaptation to a specific domain is successful, the adaption showed to transfer to OOD detection.

Lastly, this thesis identified a strong correlation between the CLP with Clip's zero-shot classifier and all methods, including the baseline linear probe classifiers. The model and classifier are universally available, easily implementable, and only require semantic labels. These labels are accessible for the majority of research datasets or can be crafted as demonstrated by Radford et al. [74]. Consequently, we propose the Universal Clip-based Confusion Log Probability (UC-CLP) as a universal indicator of the difficulty of selected

ID / OOD splits. The introduction of UC-CLP is expected to serve as a valuable indicator of the difficulty of various ID/OOD settings, thereby enhancing comparability within the field, which has hitherto remained limited.

7.2. Future Work

In conclusion, a thorough comparison of the robustness of the proposed approach with other contemporary methods would enhance the evaluability of the results, providing deeper insights into the validity of the metrics, particularly the CLP as universal difficulty estimators. Although we have examined the approach for different methods, including Z-OOD detection and linear probe classifiers, it is important to note that all these methods are based on Clip features, creating a strong association between the UC-CLP and these techniques.

Future work should also focus on comparing the proposed few-shot OOD detection methods with other few-shot OOD detection approaches to ensure seamless integration into the broader context. While the demonstrated efficacy of the proposed approach is promising, surpassing Z-OOD detection and performing better than a fine-tuned method, it is important to acknowledge that the chosen method also relies on Clip features, thereby sharing the same limitation addressed by domain adaptation. Consequently, future research should incorporate comparisons with non-Clip-based methods.

Furthermore, domain adaptation to more distinct distributions needs to be explored. Although the observed performance improvement is a valuable and intriguing finding, the transition from tasks that were previously unachievable to those that are now feasible is of greater interest and represents the true advancement. By pursuing this line of inquiry, the limitations of Clip-based OOD detection can be surmounted.

Bibliography

[1] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara. "Latent space autoregression for novelty detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, USA, 2019, pp. 481–490.

[2] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. "Concrete problems in ai safety". In: *arXiv preprint arXiv:1606.06565* (2016).

[3] S. An, M. Lee, S. Park, H. Yang, and J. So. "An ensemble of simple convolutional neural network models for mnist digit recognition". In: *arXiv preprint arXiv:2008.10400* (2020).

[4] H. Bao, L. Dong, S. Piao, and F. Wei. "Beit: Bert pre-training of image transformers". In: *arXiv preprint arXiv:2106.08254* (2021).

[5] I. Bello, W. Fedus, X. Du, E. Cubuk, A. Srinivas, T.-Y. Lin, J. Shlens, and B. Zoph. "Revisiting resnets: Improved training and scaling strategies". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 22614–22627.

[6] Abhijit Bendale and Terrance E. Boult. "Towards open set deep networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, BC, Canada, 2016, pp. 1563–1572.

[7] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[8] R. Bommasani et al. "On the opportunities and risks of foundation models". In: *CoRR* abs/2108.07258 (2021).

[9] A. P. Bradley. "The use of the area under the roc curve in the evaluation of machine learning algorithms". In: *Pattern Recognition* 30.7 (1997), pp. 1145–1159.

[10] A. Byerly, T. Kalganova, and I. Dear. "No routing needed between capsules". In: *Neurocomputing* 463 (2021), pp. 545–553.

[11] J. Canny. "A computational approach to edge detection". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* Pami-8 (1986), pp. 679–698.

[12] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo. "Adaptformer: Adapting vision transformers for scalable visual recognition". In: *arXiv preprint arXiv:2205.13535* (2022).

[13] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. "Learning phrase representations using rnn encoder–decoder for statistical machine translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1724–1734.

[14] E. Chong, C. Han, and F. C. Park. "Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies". In: *Expert Systems with Applications* 83 (2017), pp. 187–205.

[15] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. "Describing textures in the wild". In: *IEEE Conference in Computer Vision and Pattern Recognition*. Columbus, OH, USA, 2014, pp. 3606–3613.

[16] L. Deecke, R. Vandermeulen, L. Ruff, S. Mandt, and M. Kloft. "Image anomaly detection with generative adversarial networks". In: *Machine Learning and Knowledge Discovery in Databases*. Dublin, Ireland, Springer International Publishing, 2019, pp. 3–17.

[17] J. Devlin, M.-W. Chang, L. Lee, and K. Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, MN, USA, 2019, pp. 4171–4186.

[18] X. Ding, Y. Zhang, T. Liu, and J. Duan. "Deep learning for event-driven stock prediction". In: *Proceedings of the 24th International Conference on Artificial Intelligence*. Buenos Aires, Argentina: AAAI Press, 2015, pp. 2327–2333.

[19] A. Dirik and S. Paul. *A dive into vision-language models*. Accessed on 27.03.2023. URL: `https://huggingface.co/blog/vision_language_pretraining`.

[20] A. Dosovitskiy, A. Kolesnikov, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, Mostafa D., N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai. "An image is worth 16x16 words: transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[21] S. Esmaeilpour, B. Liu, E. Robertson, and L. Shu. "Zero-shot out-of-distribution detection based on the pretrained model clip". In: *Proceedings of the AAAI conference on artificial intelligence*. Online, 2022, pp. 6568–6576.

[22] G. Fei and B. Liu. "Breaking the closed world assumption in text classification". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, CA, USA, 2016, pp. 506–514.

[23] C. Finn, P. Abbeel, and S. Levine. "Model-agnostic meta-learning for fast adaptation of deep networks". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. Sydney, Australia: JMLR.org, 2017, pp. 1126–1135.

[24] S. Fort, J. Ren, and B. Lakshminarayanan. "Exploring the limits of out of distribution detection". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 7068–7081.

[25] P. Gage. "A new algorithm for data compression". In: *C Users Journal* 12.2 (1994), pp. 23–38.

[26] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao. "Clip-adapter: Better vision-language models with feature adapters". In: *arXiv preprint arXiv:2110.04544* (2021).

[27] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. 2016.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial nets". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger. Vol. 27. Montreal, QC, Canada: Curran Associates, Inc., 2014.

[29] I. J. Goodfellow, J. Shlens, and C. Szegedy. "Explaining and harnessing adversarial examples". In: *arXiv preprint arXiv:1412.6572* (2014).

[30] A. Graves. "Generating sequences with recurrent neural networks". In: *arXiv preprint arXiv: 1308.0850* (2013).

[31] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. "A kernel two-sample test". In: *Journal of Machine Learning Research* 13.25 (2012), pp. 723–773.

[32] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. "On calibration of modern neural networks". In: *International conference on machine learning*. Sydney, Australia, 2017, pp. 1321–1330.

[33] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA, 2016, pp. 770–778.

[34] P. Helber, B. Bischke, A. Dengel, and D. Borth. "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12.7 (2019), pp. 2217–2226.

[35] D. Hendrycks and T. Dietterich. "Benchmarking neural network robustness to common corruptions and perturbations". In: *arXiv preprint arXiv:1903.12261* (2019).

[36] D. Hendrycks and K. Gimpel. "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks". In: *Proceedings of the 5th International Conference on Learning Representations*. Toulon, France, 2017.

[37] D. Hendrycks, X. Liu, E. Wallace, A. Dziedzic, R. Krishnan, and D. Song. "Pretrained transformers improve out-of-distribution robustness". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online, 2020, pp. 2744–2751.

[38] D. Hendrycks, M. Mazeika, and T. Dietterich. "Deep anomaly detection with outlier exposure". In: *arXiv preprint arXiv:1812.04606* (2018).

[39] D. Hirata and N. Takahashi. "Ensemble learning in cnn augmented with fully connected subnetworks". In: *arXiv preprint arXiv:2003.08562* (2020).

[40] S. Hochreiter and J. Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[41] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. "Parameter-efficient transfer learning for nlp". In: *International Conference on Machine Learning*. Long Beach, USA, 2019, pp. 2790–2799.

[42] J. Howard and S. Ruder. "Universal language model fine-tuning for text classification". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia, 2018, pp. 328–339.

[43] Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira. "Generalized odin: Detecting out of distribution image without learning from out of distribution data". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Online, 2020, pp. 10948–10957.

[44] P. H. Le-Khac, G. Healy, and A. F. Smeaton. "Contrastive representation learning: A framework and review". In: *Ieee Access* 8 (2020), pp. 193907–193934.

[45] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[46] I. Kobyzev, S. D. Prince, and M. A. Brubaker. "Normalizing flows: An introduction and review of current methods". In: *IEEE transactions on pattern analysis and machine intelligence* 43.11 (2021), pp. 3964–3979.

[47] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. "3D object representations for fine-grained categorization". In: *4th International IEEE Workshop on 3D Representation and Recognition*. Sydney, Australia, 2013.

[48] A. Krizhevsky and G. Hinton. *Learning multiple layers of features from tiny images*. Tech. rep. Toronto, ON, Canada, 2009.

[49] A. Krizhevsky, I. Sutskever, and G. E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. Lake Tahoe, NV, USA: Curran Associates, Inc., 2012.

[50] B. Lakshminarayanan, A. Pritzel, and C. Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in neural information processing systems* 30 (2017).

[51] C. H. Lampert, H. Nickisch, and S. Harmeling. "Attribute-based classification for zero-shot visual object categorization". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.3 (2014), pp. 453–465.

[52] Y. Le and X. Yang. "Tiny imagenet visual recognition challenge". In: *Cs 231n* 7.7 (2015), p. 3.

[53] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. "Backpropagation applied to handwritten zip code recognition". In: *Neural computation* 1.4 (1989), pp. 541–551.

[54] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[55] F.-F. Li, M. Andreeto, M. Ranzato, and P. Perona. *Caltech 101*. 2022.

[56] S. Liang, Y. Li, and R. Srikant. "Enhancing the reliability of out-of-distribution image detection in neural networks". In: *6th International Conference on Learning Representations*. Vancouver, BC, Canada, 2018.

[57] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. "Microsoft coco: Common objects in context". In: *Proceedings of the European Conference on Computer Vision*. Zurich, Switzerland, 2014, pp. 740–755.

[58] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing". In: *ACM Computing Surveys* 55.9 (2023), pp. 1–35.

[59] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. "A convnet for the 2020s". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, LA, USA, 2022, pp. 11976–11986.

[60] P. Liznerski, L. Ruff, R. A. Vandermeulen, B. J. Franks, K.-R. Müller, and M. Kloft. "Exposing outlier exposure: What can be learned from few, one, and zero outlier images". In: *arXiv preprint arXiv:2205.11474* (2022).

[61] I. Loshchilov and F. Hutter. "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101* (2017).

[62] D.G. Lowe. "Object recognition from local scale-invariant features". In: *Proceedings of the Seventh IEEE International Conference on Computer Vision*. Vol. 2. Kohala Coast, HI, USA, 1999, 1150–1157 vol.2.

[63] T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[64] D. Miller, N. Sunderhauf, M. Milford, and F. Dayoub. "Class anchor clustering: A loss for distance-based open set recognition". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. Online, 2021, pp. 3570–3578.

[65] Y. Ming, Z. Cai, J. Gu, Y: Sun, W. Li, and Y. Li. "Delving into out-of-distribution detection with vision-language representations". In: *arXiv preprint arXiv:2211.13445* (2022).

[66] S. Narayan, S. B. Cohen, and M. Lapata. "Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018.

[67] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. "Reading digits in natural images with unsupervised feature learning". In: (2011).

[68] M.-E. Nilsback and A. Zisserman. "Automated flower classification over a large number of classes". In: *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. Bhubaneswar, India, 2008, pp. 722–729.

[69] A. van den Oord, Y. Li, and O. Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018).

[70] OpenAI. *Gpt-4 technical report*. 2023. arXiv: `2303.08774 [cs.CL]`.

[71] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine learning in python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[72] J. Pennington, R. Socher, and C. D. Manning. "Glove: Global vectors for word representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*. Baltimore, MD, USA, 2014, pp. 1532–1543.

[73] J. Pfeiffer, A. Rücklé, C. Poth, A. Kamath, I. Vulić, S. Ruder, K. Cho, and I. Gurevych. "Adapterhub: A framework for adapting transformers". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Systems Demonstrations*. Online: Association for Computational Linguistics, 2020, pp. 46–54.

[74] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I Sutskever. "Learning transferable visual models from natural language supervision". In: *International Conference on Machine Learning*. Online, 2021, pp. 8748–8763.

[75] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9.

[76] J. Re, S. Fort, J. Liu, A. G. Roy, S. Padhy, and B. Lakshminarayanan. "A simple fix to mahalanobis distance for improving near-ood detection". In: *arXiv preprint arXiv:2106.09022* (2021).

[77] S. Rothe, S. Narayan, and A. Severyn. "Leveraging pre-trained checkpoints for sequence generation tasks". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 264–280.

[78] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.

[79] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115 (2015), pp. 211–252.

[80] S. J Russell. *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010.

[81] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli. "Adversarially learned one-class classifier for novelty detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA, 2018, pp. 3379–3388.

[82] M. Salehi, H. Mirzaei, D. Hendrycks, Y. Li, M. H. Rohban, and M. Sabokrou. "A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges". In: *arXiv preprint arXiv:2110.14051* (2021).

[83] G. Salton and M. J. McGill. *Introduction to modern information retrieval*. McGraw-Hill, Inc., 1986.

[84] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. "Toward open set recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 35.7 (2012), pp. 1757–1772.

[85] R. Sennrich, B. Haddow, and A. Birch. "Neural machine translation of rare words with subword units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 1715–1725.

[86] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela. "Flava: A foundational language and vision alignment model". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, LA, USA, 2022, pp. 15638–15650.

[87] J. Snell, K. Swersky, and R. Zemel. "Prototypical networks for few-shot learning". In: *Advances in neural information processing systems* 30 (2017).

[88] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. "Zero-shot learning through cross-modal transfer". In: *Advances in neural information processing systems* 26 (2013).

[89] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R.G. Lanckriet. "Hilbert space embeddings and metrics on probability measures". In: *Journal of Machine Learning Research* 11 (2010), pp. 1517–1561.

[90] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. "The german traffic sign recognition benchmark: A multi-class classification competition". In: *The 2011 international joint conference on neural networks*. San Jose, CA, USA, 2011, pp. 1453–1460.

[91] D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. "Generative models and model criticism via optimized maximum mean discrepancy". In: *arXiv preprint arXiv:1611.04488* (2016).

[92] Ilya Sutskever. "Training recurrent neural networks". PhD thesis. 2013.

[93] W. L. Taylor. "Cloze procedure: A new tool for measuring readability". In: *Journalism quarterly* 30.4 (1953), pp. 415–433.

[94] T. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971* (2023).

[95] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[96] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman. "Open-set recognition: A good closed-set classifier is all you need". In: *arXiv preprint arXiv:2110.06207* (2021).

[97] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling. "Rotation equivariant cnns for digital pathology". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference*. Granada, Spain, 2018, pp. 210–218.

[98] R. Vershyninm. *High-dimensional probability: An introduction with applications in Data Science*. Cambridge University Press, 2018.

[99] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. *Caltech cub*. Tech. rep. California Institute of Technology, 2011.

[100] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, and R. Fergus. "Regularization of neural networks using dropconnect". In: *International conference on machine learning*. Atlanta, GA, USA, 2013, pp. 1058–1066.

[101] F. Wang and H. Liu. "Understanding the behaviour of contrastive loss". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos, CA, USA: IEEE Computer Society, 2021, pp. 2495–2504.

[102] T. Wang and P. Isola. "Understanding contrastive representation learning through alignment and uniformity on the hypersphere". In: *International Conference on Machine Learning*. online, 2020, pp. 9929–9939.

[103] R. J. Williams and D. Zipser. "A learning algorithm for continually running fully recurrent neural networks". In: *Neural Computation* 1.2 (1989), pp. 270–280.

[104] J. Winkens, R. Bunel, A. Guha Roy, R. Stanforth, V. Natarajan, J. Ledsam, P. Kohli, P. MacWilliams, A. Karthikesalingam, S. Kohl, A. Cemgil, S. Eslami, and O. Ronneberger. "Contrastive training for improved out-of-distribution detection". In: *arXiv preprint arXiv:2007.05566* (2020).

[105] M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. G. Lopes, H. Hajishirzi, A. Farhadi, H. Namkoong, and L. Schmidt. "Robust fine-tuning of zero-shot models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, LA, USA, 2022, pp. 7959–7971.

[106] H. Xiao, K. Rasul, and R. Vollgraf. *Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms*. 2017. eprint: `cs.LG/1708.07747`.

[107] X. Xu, C. Wu, S. Rosenman, V. Lal, W. Che, and N. Duan. "Bridgetower: Building bridges between encoders in vision-language representation learning". In: *arXiv preprint arXiv:2206.08657* (2022).

[108] J. Yang, K. Zhou, Y. Li, and Z. Liu. "Generalized out-of-distribution detection: A survey". In: *arXiv preprint arXiv:2110.11334* (2021).

[109] Q. Ye and X. Ren. "Learning to generate task-specific adapters from task description". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online, 2021, pp. 646–653.

[110] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer. "Lit: Zero-shot transfer with locked-image text tuning". In: *Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, LA, USA, 2022, pp. 18123–18133.

[111] R. Zhang, W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li. "Tip-adapter: Training-Free adaption of clip for few-shot classification". In: *Proceedings of the European Conference on Computer Vision*. Tel Aviv, Israel: Springer Nature Switzerland, 2022, pp. 493–510.

[112] K. Zhou, J. Yang, C. Change Loy, and Z. Liu. "Learning to prompt for vision-language models". In: *International Journal of Computer Vision* 130.9 (2022), pp. 2337–2348.

Appendix A.

Supplementary Material

A.1. Near-OOD with different Linear Probes

In the following, we present a classification comparison for selecting the baseline in the experiments for Chapter 5. Table A.1 shows the results. The logistic regression performs slightly better on average. Figure A.1 shows that the logistic regression classifier performs clearly better on Out-of-Distribution Detection (OOD detection) and is therefor used as baseline in the experiments.

Table A.1.: Accuracies for each dataset using the named strategy on top of image features from Clip's ViT-B/32 image encoder. Values are in %

	Linear Acc ↑	Logistic Acc ↑
Caltech101	95.5	95.07
Caltech CUB	72.17	71.14
CIFAR10	94.7	94.74
CIFAR100	78.7	79.09
DTD	72.8	72.87
Fashion MNIST	88.85	90.41
Flowers102	94.49	94.1
GTSRB	85.84	86.44
TinyImagenet	73.99	74.82
MNIST	97.71	98.6
Stanford Cars	78.48	77.88
SVHN	59.44	65.14

Table A.1 shows the accuracies for each dataset for different linear probe settings. The means are nearly identical (Linear: 82.72, Logistic: 83.36).

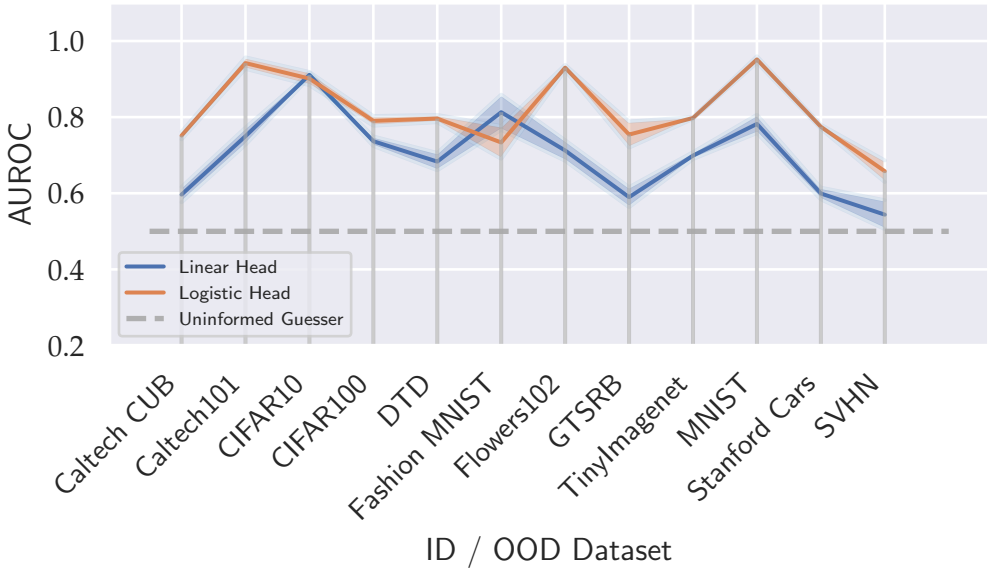


Figure A.1.: Logistic Regression and fully connected linear layer on top of Clip features for each dataset. AUROCs are the mean averaged over 10 runs. The shadows indicate the standard deviation. All experiments are conducted with the ViT-B/32 vision encoder

A.2. Training Details Text Decoder

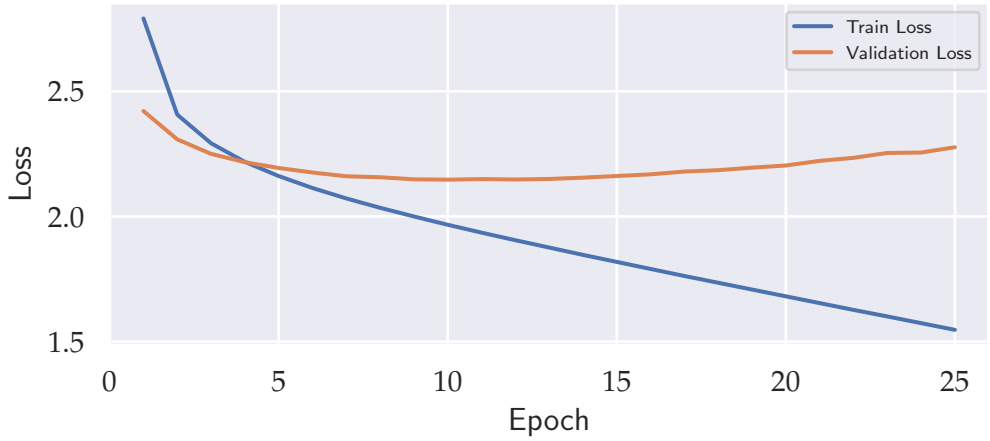


Figure A.2.: Mean batch loss in each epoch in the Decoder Training

This section provides the training details for the image decoder used in the ZOC approach for all experiments. The progression of the loss over the training epochs is shown in Figure A.2. The train loss decreases consistently, while the validation loss increases after epoch 14. This could be due to overtraining, however, no further experiments were conducted to confirm this hypothesis. The training was conducted using the teacher-forcing method [103] and the Adam optimizer [45] with a constant learning rate of 10^{-5} for 25 epochs. The training and validation sets were based on the officially released data

splits for the MS-COCO 2017 release [57], which are consistent with the methodology used in [21]. The model after 14 iterations is selected, as it had the lowest validation loss, which suggests it may have the best generalization to unseen data.

A.3. Temperature & Prompt Ablation

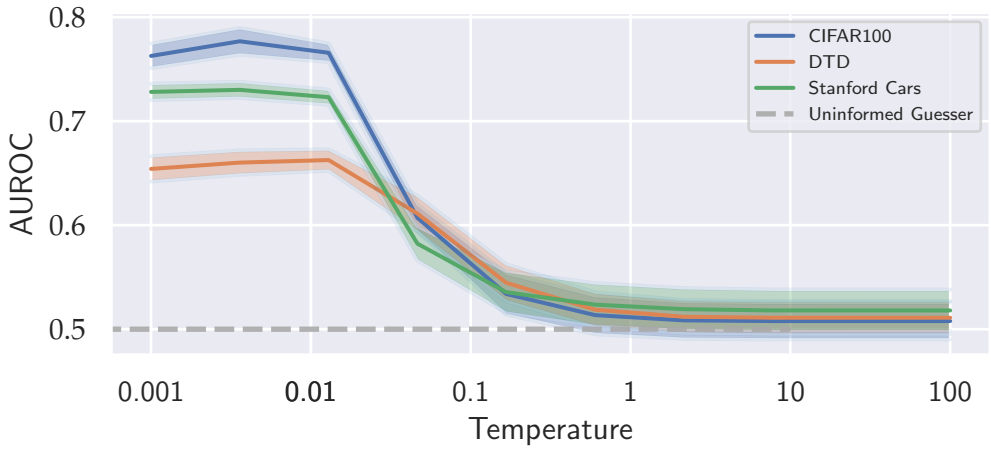


Figure A.3.: ZOC temperature scaling ablation for a subset of the datasets. The shadow indicates standard deviation

This section shows the influence of temperature scaling in the softmax scores on the OOD detection performance. Figure A.3 shows, that ZOC performs best with very small temperatures ($\tau < 1.0$) is at $\tau = 1.0$ already close to the performance of an uninformed guesser. For MCM Figure A.4 shows that it is exactly the other way round: The best worst performance is for temperaturs $\tau < 1.0$ and stabilizes afterwards.
Figure A.4 also shows results for different prompts: The best and most stable combination is using MLS with the default prompt "A photo of TOKEN".

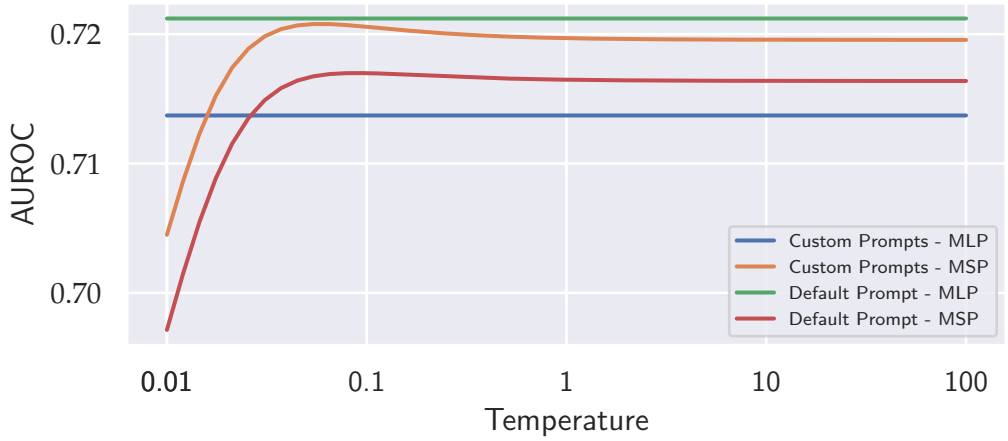


Figure A.4.: MCM temperature scaling and prompt ablation. The values are means over all datasets. Custom prompts origin from the respective publication, or are manually crafted by Radford et al. [74] , default prompt as stated in Chapter 4.1

A.4. Zero-Shot Methods as Basis for Few-Shot OOD Detection

1 This section provides further insights into domain adaption with TIP-Adapter [111]. Table A.3 shows the classification results, which are the indicator of good domain adaption. The fully fine-tuned logistic regression, especially with the ViT-L model has the highest accuracy overall, which is also visualized in Figure A.5.

	MCM	T-MCM	T-MCM-f	ZOC	T-ZOC	T-ZOC-f	Log ViT-L/16@336px
MNIST	0.591	0.661	0.622	0.603	0.577	0.604	**0.952**
CIFAR100	0.735	0.717	0.77	0.76	0.718	0.774	**0.791**
SVHN	0.533	0.547	0.555	0.533	0.514	0.53	**0.668**
GTSRB	0.515	0.578	0.549	0.525	0.614	0.548	**0.809**
Caltech CUB	0.639	0.665	0.643	0.724	0.719	0.725	**0.753**
Fashion MNIST	0.701	0.73	0.706	0.785	**0.798**	0.77	0.742
Stanford Cars	0.645	0.678	0.65	0.739	0.749	0.737	**0.776**
TinyImagenet	0.75	0.725	0.774	0.759	0.727	0.773	**0.791**
Flowers102	0.773	0.856	0.775	0.816	0.891	0.807	**0.924**
DTD	0.664	0.731	0.71	0.648	0.712	0.674	**0.807**
Caltech101	0.87	0.897	0.897	0.879	0.906	0.901	**0.94**
CIFAR10	0.9	0.881	0.907	**0.956**	0.912	0.954	0.913

Table A.2.: Full OOD Detection results with domain adaption and linear benchmark. Bold indicates best results

The Full AUROCS are displayed in Table A.2, the accuracies of all classification approaches in Table A.3.

Figure A.6 shows, that both, MCM and T-MCM are better than the linear probe for all setups with less than 65 samples per class.

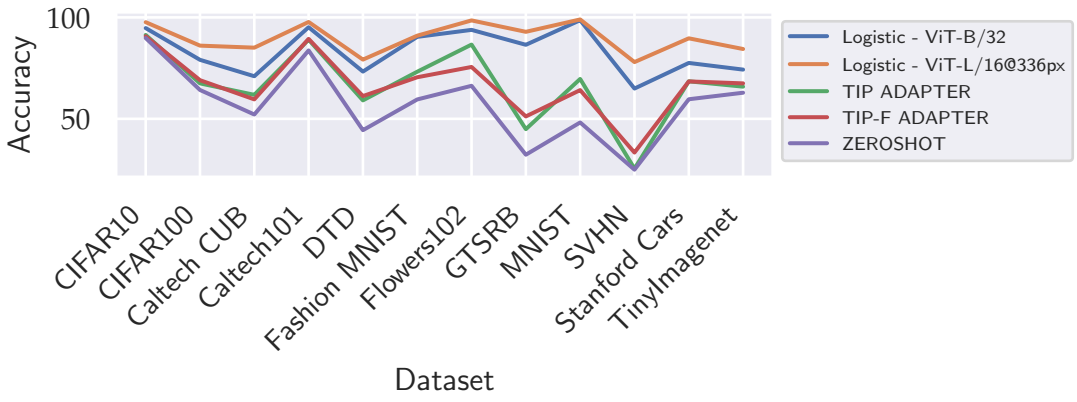


Figure A.5.: Full AUROCS for all methods in near-OOD setup. All adapted methods are trained according to Chapter 4. The best results are in bold.

	ZEROSHOT	TIP-F	TIP	Log ViT-B/32	Log ViT-L/16@336px
Fashion MNIST	59.56	70.502	73.229	90.41	91.0
DTD	44.415	61.218	59.106	73.298	79.202
SVHN	24.946	33.382	25.455	64.932	77.977
TinyImagenet	62.88	67.481	65.868	74.27	84.42
Stanford Cars	59.657	68.45	68.504	77.553	89.69
CIFAR10	89.83	90.72	91.367	94.74	97.65
Caltech CUB	52.14	59.598	61.836	71.004	85.088
GTSRB	32.32	51.199	44.921	86.5	92.898
CIFAR100	64.23	69.061	67.483	79.09	86.09
MNIST	48.22	64.17	69.657	98.62	98.99
Flowers102	66.287	75.575	86.609	93.853	98.52
Caltech101	83.722	89.404	89.038	95.16	97.753

Table A.3.: Full classification results with all methods used in this thesis. Log is short for logistic regression

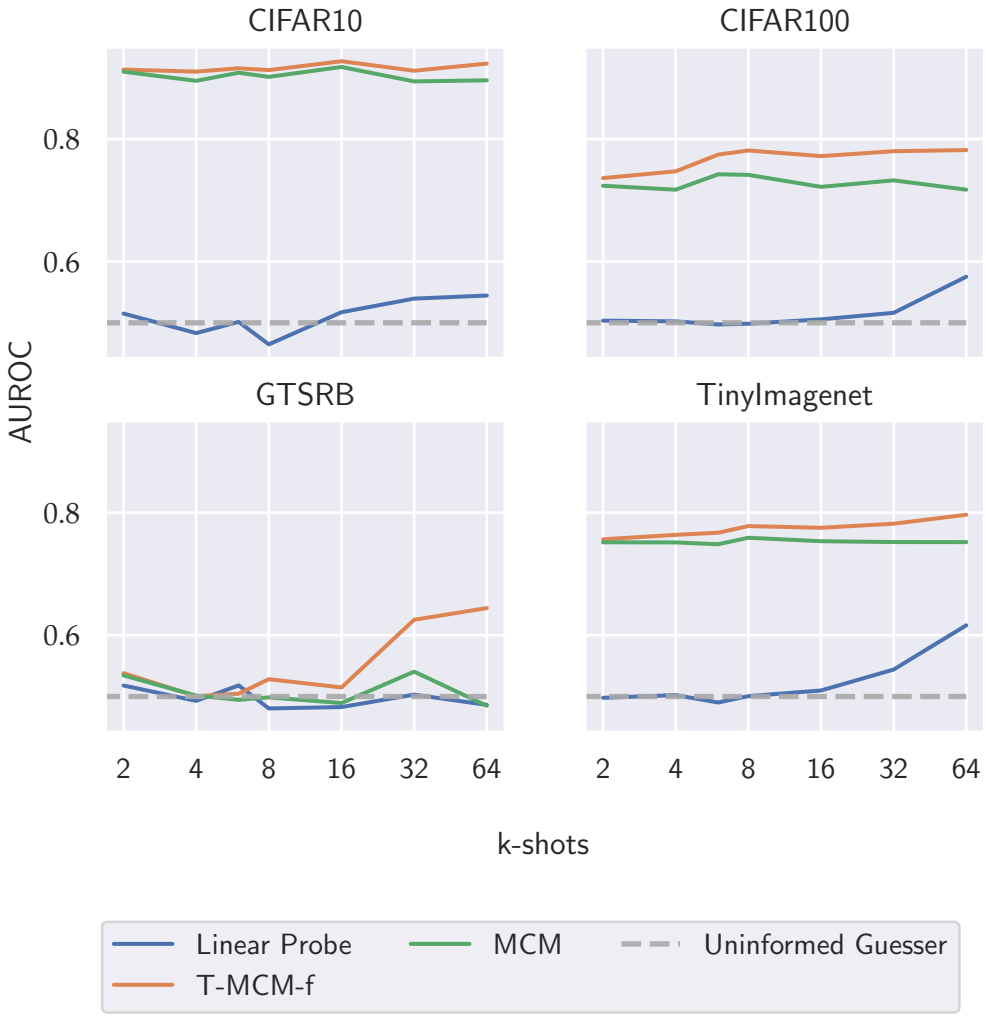


Figure A.6.: K-shot ablation on datasets with 2 to 64 samples per class. The linear probe is only able to detect outliers better than an uninformed guesser with at least 32 samples per class, while all other methods also perform with 2 samples.

Erklärung der Urheberschaft

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit im Masterstudiengang Informatik selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht.

Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der elektronischen Abgabe entspricht.

Hamburg, den 16.04.2023 Fabian Meyer

Veröffentlichung in der Informatik-Bibliothek

Ich stimme der Einstellung der Arbeit in die Bibliothek des Fachbereichs Informatik zu.

Hamburg, den 16.04.2023 Fabian Meyer