Heinrich Heine
Universität Düsseldorf

Self-Supervised Representation learning for Anomaly Detection

Inaugural-Dissertation

submitted to

the Faculty of Mathematics and Natural Sciences at the Heinrich Heine University

for the attainment of the degree of the Doctor of Philosophy

Author

Nima Rafiee

Mathematical Modeling of Biological Systems Institute

Thesis Director

Prof. Dr. Markus Kollmann Heinrich Heine University, Düsseldorf, Germany

Thesis Co-Director

Prof. Dr. Timo Dickscheid Heinrich Heine University, Düsseldorf, Germany

June 2022

To my mother and father
To my wife Pegah,
To my sister and brother, Laleh and Ramin.

Statement of authorship

I hereby declare that this dissertation is the result of my own work. No other person's work has been used without due acknowledgment. This dissertation has not been submitted in the same or similar form to other institutions. I have not previously failed a doctoral examination procedure.

Düsseldorf, Germany, June 2022
Nima Rafiee

Acknowledgements

I have spent 4 fruitful years at the Institute of Mathematical Modeling of Biological Systems. I had the chance to work and be directed by an excellent scientist, Prof. Dr. Kollmann as my supervisor. The discussions we had around understanding the problems and different solutions were an endless enjoyment and an opportunity for me to learn.

and Chris, thanks for your helps from the very first day I came to Düsseldorf. I will miss Chrisian jokes at Mensa.

and Linlin, (the Dadash), thanks for all your support in the hard times I had.

and Rahil, thanks for being such a great colleague. For sure the work I have done was not possible without your help.

Finally, I wish to present my sincere thanks to Prof. Dr. Timo Dickscheid for accepting to read the thesis. I am sure that his remarks and suggestions will be so precious to this work.

Düsseldorf, Germany, October 2022
Nima Rafiee

Abstract

Machine learning in general and deep learning, in particular, has been recognized with various predictive and descriptive applications. Object detection, data clustering, and classifying samples into predefined categories are only some tasks in which promising results have been achieved using machine learning approaches. Another important application is Out-of-distribution (OOD) or anomaly detection. In General, anomaly detection refers to the problem of detecting whether or not a sample belongs to the distribution of an already seen training dataset. Anomaly detection has gained a lot of applications in real-world problems such as detecting anomalous items in manufacturing production lines using image processing, medical diagnosis in medical imaging, detecting abnormalities in internet traffic, and so on.

Similar to other areas, machine learning approaches, including supervised and unsupervised methods, have been extensively leveraged for the task of anomaly detection. However, significant challenges and limitations with these two methods have remained untackled. Recently self-supervised methods have been introduced aiming at learning a representation from data that benefits diverse downstream tasks, including anomaly detection. The main focus of this thesis is to introduce and study the use of these methods for the task of anomaly detection in natural object-centric and medical images. In particular, some limitations of supervised methods, such as the lack of annotated data and the likelihood score issue of the unsupervised method, have been tackled.

Zusammenfassung

Das maschinelle Lernen im Allgemeinen und das Deep Learning im Besonderen haben sich bei verschiedenen prädiktiven und beschreibenden Anwendungen bewährt. Die Erkennung von Objekten, das Clustering von Daten und die Klassifizierung von Proben in vordefinierte Kategorien sind nur einige der Aufgaben, bei denen mit Ansätzen des maschinellen Lernens vielversprechende Ergebnisse erzielt wurden. Eine weitere wichtige Anwendung ist die Out-of-distribution (OOD) oder Anomalie-Erkennung. Im Allgemeinen bezieht sich die Erkennung von Anomalien auf das Problem, zu erkennen, ob eine Probe zur Verteilung eines bereits gesehenen Trainingsdatensatzes gehört oder nicht. Die Erkennung von Anomalien hat in der Praxis viele Anwendungen gefunden, z. B. die Erkennung von anomalen Objekten in Fertigungsstraßen mit Hilfe der Bildverarbeitung, die medizinische Diagnose in der medizinischen Bildgebung, die Erkennung von Anomalien im Internetverkehr usw.

Ähnlich wie in anderen Bereichen wurden Ansätze des maschinellen Lernens, einschließlich überwachter und unüberwachter Methoden, in großem Umfang für die Erkennung von Anomalien eingesetzt. Allerdings gibt es bei diesen beiden Methoden erhebliche Herausforderungen und Einschränkungen, die bisher nicht angegangen wurden. Kürzlich wurden selbstüberwachte Methoden eingeführt, die darauf abzielen, eine Repräsentation von Daten zu erlernen, die verschiedenen nachgelagerten Aufgaben, einschließlich der Erkennung von Anomalien, zugute kommt. Das Hauptaugenmerk dieser Arbeit liegt auf der Einführung und Untersuchung des Einsatzes dieser Methoden für die Erkennung von Anomalien in natürlichen, objektzentrierten und medizinischen Bildern. Insbesondere wurden einige Einschränkungen der überwachten Methoden, wie das Fehlen kommentierter Daten und das Problem der Likelihood-Scores der unüberwachten Methode, in Angriff genommen.

Contents

List of Figures

Abbreviation

Artificial Intelligence	AI
Machine Learning	ML
Deep Learning	DL
Neural Network	NN
Self-Supervised Learning	SSL
Deep Neural Network	DNN
Feedforward Neural Network	FFDNN
Out-of-Distribution	OOD
Empirical Risk Minimization	ERM
Maximum A Posterior	MAP
Mutual Information	MI
Convolutional Neural Network	CNN
In-distribution	In-dist
One-Class Classifiers	OCC
Kullback-Leibler divergence	KL divergence
Multilayer Perceptron	MLP
Rectified Linear Unit	ReLU
Independent Identically Distributed	iid
Normalized Temperature-scaled Cross-entropy	NT-Xent
Contrastive Predictive Coding	CPC
Area Under the Receiver Operating Characteristics	AUROC

Chapter 1

Introduction

In this thesis, we study the problem of anomaly detection using a recently emerged approach in machine learning (ML) known as self-supervised learning (SSL). The general focus of this study is on anomaly detection in the image domain and, in particular natural and medical images. Following, an introduction to anomaly detection and machine learning is provided. Different categories of ML are briefly introduced. In future chapters, ML materials related to this study will be explained in more detail.

1.1 Anomaly Detection

Anomaly detection is the process of detecting whether or not a sample is driven by input data distribution or, in other words, if a given sample follows the same patterns as the train data. The concept of anomaly detection is used by several different names, such as out of distribution detection (OOD) and novelty detection. Different naming is due to different areas in which the problem is studied or the application for which an anomaly detection solution is designed. However, there is a slight difference for approaches titled novelty detection. The goal of novelty detection is to find the unseen patterns in training data and then incorporate them as the attribute of normal (in-distribution) data. However, the goal for most of the anomaly detection problems is to detect abnormal and faulty samples during inference time, such as detecting damaged items in the production line using image processing. Figure 1.1 illustrates examples of anomaly samples. Normal or, in other words, in-distribution samples marked with the plus sign occupy three regions in 2-dimensional space. In-distribution samples in each area are presented with different colours where each colour shows a different class. OOD samples, marked with red circles, are located at a different distance from normal samples. For example, o_3 is located close to in-distribution samples of class 2, which makes the detection models prone to recognise it as one of the class 2 in-distribution samples.

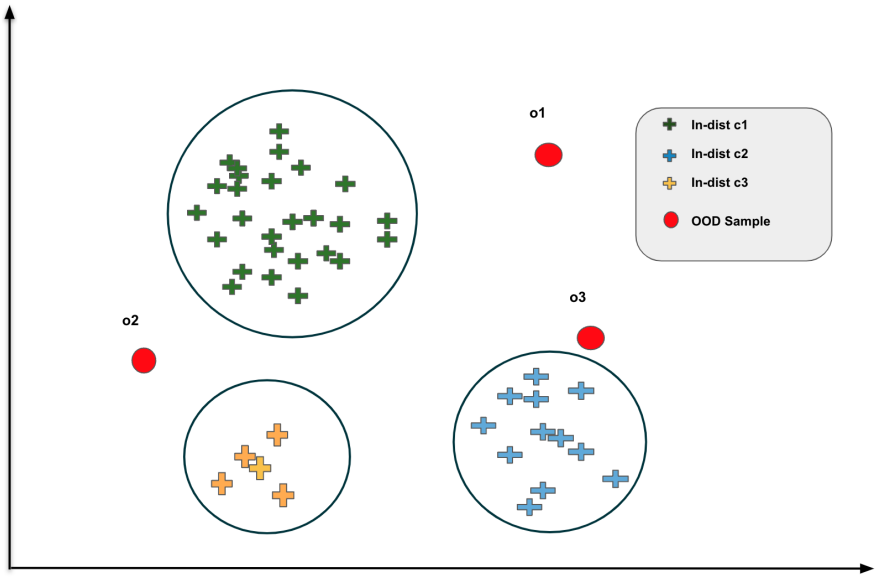


Figure 1.1: In-distribution samples are marked with the plus sign where different colour shows different classes. OOD samples are shown with red circles. OOD samples can be located at different distances from the region of normal samples.

1.2 Introduction to Machine Learning

Machine learning is a branch of artificial intelligence and computer science. In general, ML mainly focuses on an automated way of data analysis through statistical inference. Particularly, ML can be defined as a set of techniques used to extract unknown statistical patterns from data and then use these patterns to perform predictions for unseen data or to be used in decision making problems [12]. ML algorithms can be divided into two groups of traditional techniques and recently emerged deep models. Traditional ML approaches include diverse methods such as support vector machines(SVM) [13] and decision trees [14]. A common aspect of traditional ML algorithms is that input data features upon which the model performs the training and prediction have to be considerably hand-engineered using human knowledge. On the other hand, deep models apply an automated way of feature extraction with less human knowledge intervention. This is especially important for the anomaly detection task where the input data are complex and high-dimensional where hand-engineered feature extraction is not feasible, such as image data. However, even with deep models, some human knowledge can be leveraged to force special learning bias, such as using convolutional neural networks [15] for image data where there exists spacial information redundancy. On the other hand, traditional ML approaches benefit more from interpretability. For example, decision trees have a more clear decision making process compared to deep models. Additionally, deep models in practice are designed with a large number of stacked parameters followed by linear and non-linear functions. Thus have a high learning capacity or high variance compared to the traditional ML techniques, which makes

them data-hungry in order to avoid overfitting problem [16].

Based on the used ground truth target, machine learning algorithms can be divided into supervised-, unsupervised- and self-supervised- learning categories

1.2.1 Supervised Learning

Supervised learning is mainly about leveraging human-generated labels into the training procedure. Supervised learning is commonly used with classification methods where the task is to assign input data to some categories predefined by human knowledge. Classification can also be considered as a mapping of input data into a lower-dimensional and discrete space. Email spam filtering, predicting the different types of flowers from their image, and category of news from its content are some examples of classification tasks [12]. Regression is another type of supervised learning, and it differs from classification as the labels are real-valued numbers and are located in a continuous space. Predicting future stock markets or future weather temperatures are some examples of regression learning. One of the main limitation of supervised learning is the existence of annotated data which is not always available. This becomes more challenging when deep models are used as they need an enormous number of samples for training [17].

1.2.2 Unsupervised Learning

The main difference between unsupervised training and supervised one is that in the latter, we have the pairs of input and target (x, y) whereas, in the former, we only use input data x, or in other words, input data is also the target. Supervised learning can be understood as simply mapping from x to y or predicting the distribution $p(y|x)$. However, unsupervised learning approaches mainly focus on understanding the input data by fitting $p(x)$ and learning a representation from data not based and limited to human-made labels. Autoencoders, Generative adversarial networks (GANs) [18], and density estimators such as VAEs [19] and PixelRNN Autoregressive models [20] are all examples of unsupervised learning.

1.2.3 Self-Supervised Learning

Self-supervised learning (SSL) is a relatively new approach compared to the two other mentioned techniques. SSL methods aim at learning useful features (representation) from input data through the solving of a supervised proxy task. The learned features can be used in several downstream tasks such as classification, object- and anomaly- detection. The proxy objectives are usually designed independently of the final downstream task, and the labels are generated automatically using the input content.

1.3 Challenges of Anomaly Detection

The problem of anomaly detection might seem straightforward in the first place, as one can argue that simply by defining the normal behavior, we can decide a decision boundary, and what locates outside of this decision boundary is considered as anomalous. However, there are plenty of challenges that have been addressed in existing research. Following, some of these challenges will be discussed.

1.3.1 What is considered as anomalous

The definition of anomaly samples is not the same in all applications. For example, analyzing time-series data related to the health and medical domain has a high sensitivity, and a slight deviation is considered an anomaly, whereas, in the business domain, the same level of deviation can be tolerated as normal behavior.

1.3.2 ML models with a lower level of abstraction

A more complicated scenario arises with abnormality diagnosis using medical images. It can be that a model detects an image without any abnormality as an anomaly only because the image is taken with a different device than the one used for the training data. In this situation, according to the high-level semantics, the image is a normal one taken from a healthy person but coming from a completely different distribution, which may be considered an anomaly by the trained ML model. The problem here is that the notion of anomaly defined by a human considering high-level semantics and easy for us to generalize it to other normal samples independent of the device used to take the image. However, even current SOTA ML models still can not capture the information as high-level as humans, and at their best, they can generalize to the test data driven from the same distribution as training data known as in-distribution test. There is a reverse problem known as an adversarial attack where an anomaly sample is manipulated so that high-level semantics seem to be similar from a human perspective, while the sample is not part of the in-distribution data.

1.3.3 Existence of outlier samples

Usually, in-distribution data contain noise or outliers. Outliers are in-distribution samples which assigned a lower likelihood by a trained model compared to the average of train data. Having these samples inside the training dataset, a model is encouraged to draw a broader boundary for normal samples, which increases the chance of anomaly samples being located inside the boundary. This problem becomes more severe for the datasets containing high dimensional samples and the models that perform the prediction on high dimensional space of input data but not on a lower-dimension representation. The issue with high dimensional data is

that for a fixed number of samples when the number of features (dimension of the space) increases, the result will be a sparse space where the noise and outliers have a stronger effect on the decided boundary by the ML model.

1.3.4 Lack of annotated data

Another issue with anomaly detection is the lack of labels when a supervised model is used to detect anomalies. In particular, for diagnosis using medical images, it can be that there are few to no samples for a specific disease, thus making the detection a challenging problem. To address the limitations of supervised methods, many different unsupervised methods such as VAEs [19] and PixelCNNs [21] have been leveraged for the task of anomaly detection [22–25]. The idea behind using these density estimators as anomaly detectors is that they can provide a higher likelihood for in-distribution data compared to OOD samples. However, in a recent study, it is shown that given high dimensional and complex data such as natural images, these methods can fail to detect OODs by assigning a higher likelihood to the samples from a different dataset than training one [26]. Table 1.1 shows the results reported in [26] for two training datasets of FashionMNIST and CIFAR-10 when GLOW structure [27] is used as a density estimator.

Table 1.1: Comparing the log-likelihood of Glow architecture for training datasets FashionMNIST and CIFAR-10. Log-likelihood is calculated in bits per dimension (BPD), and the lower the value, the higher the likelihood. **Left:** When training on FashionMNIST, test data from MNIST get higher likelihood. **Right:** SVHN test data get higher likelihood when the training data is CIFAR-10.

GLOW Trained on FashonMNIST	
Evaluate Set	Avg. BPD
FashionMNIST-Train	2.902
FashionMNIST-Test	2.958
MNIST-Test	1.833

GLOW Trained on CIFAR10	
Evaluate Set	Avg. BPD
CIFAR10-Train	3.386
CIFAR10-Test	3.464
SVHN-Test	2.389

Motivated by the explained challenges and limitations, the use of SSL methods for the task of OOD detection is studied. It is shown how representations learned by SSL methods can achieve competing results compared to supervised methods. Additionally, results show that SSL methods are not prone to the failures explained for unsupervised density estimators.

1.4 Thesis Outline

Chapter 2 provides an introduction to the problem of anomaly detection and machine learning based algorithms used to approach this problem. Supervised and Unsupervised solutions and their limitations are explained to better understand the motivation for the use of SSL methods.

Chapter 3 explains fundamental concepts of ML in more details.

Chapter 4 is generally about introducing different self-supervised methods. In particular, some examples of early methods are introduced, but the main focus is on more recent approaches since they can be considered as an improvement to their priors in terms of the performance of learned representation for the downstream tasks.

Chapter 5 addresses the problem of pneumonia detection in chest X-ray images. A contrastive based self-supervised method is combined with a Mahalanobis distance score function. This model is able to improve on previous detectors that use only healthy images during training.

Chapter 6, tackles the task of anomaly detection by proposing a self-supervised self-distillation method that leverages negative samples. Different way of creating negative samples and their impact on model performance for OOD detection in the domain of natural object-centric images is studied.

Chapter 7 investigates the use of negative sampling in abnormality detection of medical images. In particular, this chapter studies the applicability of methods used to create negative samples in the domain of natural object-centric images for the task of abnormality detection in medical imaging.

Chapter 8 provides a conclusion to this thesis and the possible topics for the future studies

Chapter 2

Anomaly Detection

2.1 Introduction

As explained in the previous chapter, anomaly or out of distribution (OOD) detection is the problem of recognising if a given sample has the same distribution as training data or if it is drawn from a different one. In this chapter, supervised and unsupervised approaches to the problem of OOD detection are explained, and their limitations are introduced.

2.2 Supervised Approach to Anomaly Detection

In this section, supervised algorithms for the task of anomaly detection will be explained. Two main approaches that use supervised training are to detect anomaly samples based on softmax response or calculate a distance such as Mahalanobis distance.

2.2.1 Anomaly detection based on softmax response

The main idea behind this approach is that the trained model to classify the labels for in-distribution data has lower confidence when it is fed with an out of distribution sample [28]. The confidence is measured by the amount of softmax response that the model provides for each category of in-distribution data. Subject to this, if the softmax response is below a given threshold for all the existing classes, then the sample is recognized as an anomaly. Figure 2.1 shows the distribution of different confidence scores for in-distribution and OOD samples assigned by the trained model.

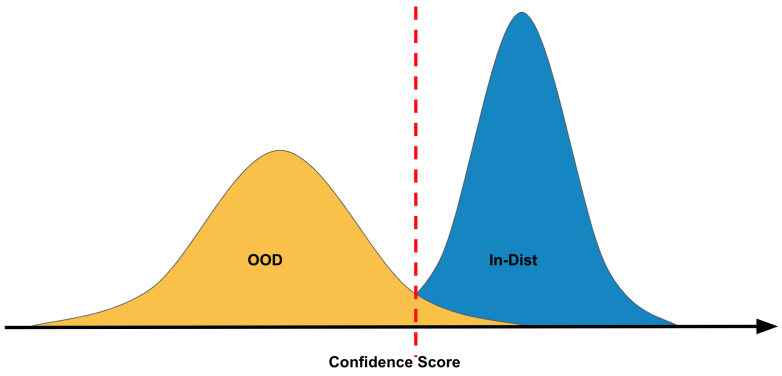


Figure 2.1: Methods trained in a supervised manner should predict a class for in-distribution data with higher confidence compared to an OOD sample.

2.2.2 Anomaly detection based on Mahalanobis distance

The main idea behind this method is that OOD samples have a larger Mahalanobis distance [29] to the training data compared to in-distribution samples [30]. Similar to the previous method, in this approach, the model is trained to predict the labels of the in-distribution data. After the training, for all training samples and a given test sample, an embedding v_i is generated by extracting the output of the last layer in the model. Note that the last layer refers to the final layer before the decision-making layer, which is a softmax in this case. Having the embedding vectors of the training samples extracted, a Gaussian distribution $\mathcal{N}(\mu_k, \sum)$ is fitted for each of the existing in-distribution classes k based on the mean of that class μ_k and a shared covariance matrix. A given test sample is recognized as an anomaly if the minimum Mahalanobis distance $MD_k = (v_{test} - \mu_k)^T \sum^{-1} (v_{test} - \mu_k)$ is larger than a threshold. Figure 2.2 shows an example of an OOD sample for which the Mahalanobis distance is calculated from training data with two different classes.

Despite their success, supervised learning approaches face specific challenges and limitations that make use of these methods not always feasible and practical. Following, some of these challenges are introduced.

2.2.3 Lack of labeled and annotated data

One of the main issues with supervised approaches is the lack of labels for many of the real-world datasets, such as rare labeled data in medical images. A recently noticed problem is detecting pneumonia from X-ray images where plenty of X-ray images belong to healthy people, but very few images from people with a particular disease [31]. Figure 2.3 shows two X-ray images from both a healthy person and a person diagnosed with pneumonia.

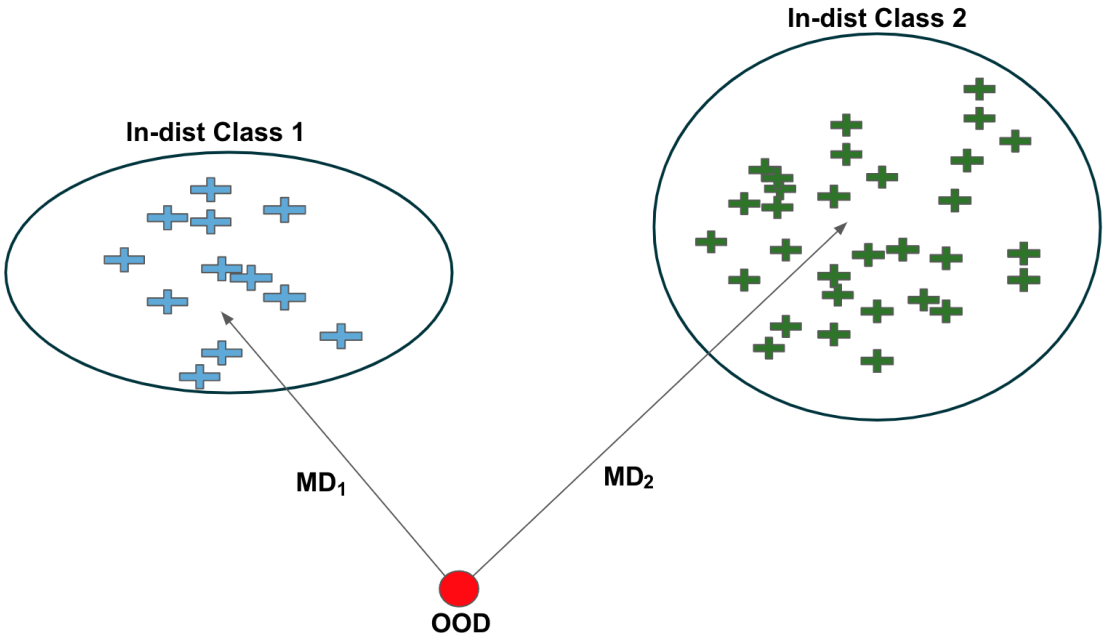


Figure 2.2: For the OOD sample, two Mahalanobis distances of MD_1 and MD_2 from each class of the training data are calculated. For a given threshold ϵ, the sample is recognised as OOD if $min(MD_1, MD_2) > \epsilon$.

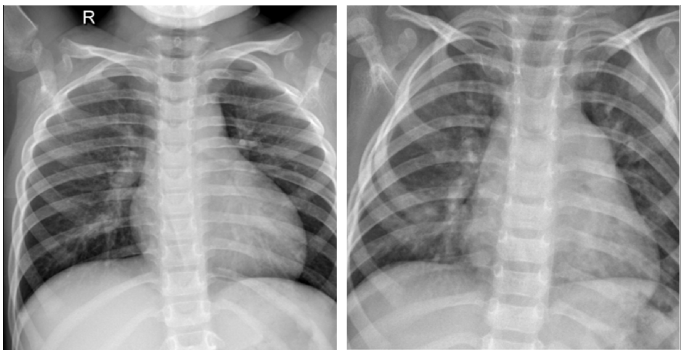


Figure 2.3: **Left:**X-ray image of a healthy person. **Right:** X-ray image of a person with pneumonia. There are few medical images related to a particular disease compared to medical images of healthy people. Image from [1].

2.2.4 Hard discriminative problem

Another issue with supervised classifiers is solving a complex discriminative task where the model has low confidence over the existing classes, even for in-distribution samples. In figure 2.4, we can see an illustration of softmax responses for easy and complex discriminative problems, whereas, for the easy one, the model provides a high softmax response for the selected class given in-distribution samples. On the other hand, for a complex discriminative problem, the model provides softmax responses with less confidence; therefore, given an anomaly sample, it is challenging

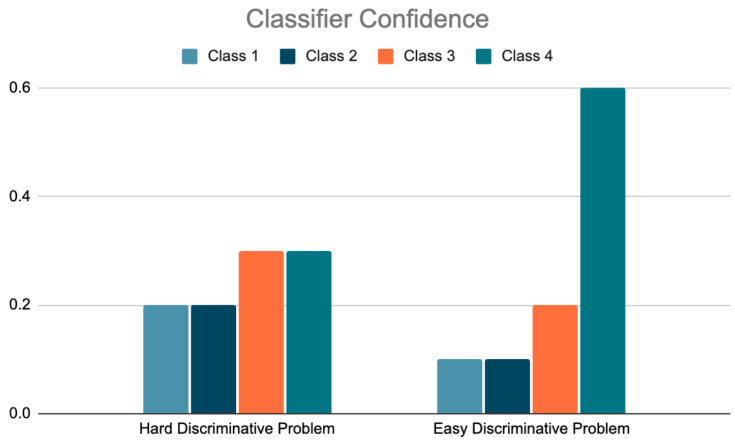


Figure 2.4: **Right:** Easy discriminative problem where it is easier to detect OOD samples based on softmax response.**Left:** Hard discriminative problem where the model confidence is low even for assigning a class to in-distribution data that makes it challenging to recognize an OOD sample based on softmax response.

to recognize it as an OOD sample or an in-distribution one.

2.2.5 Assigning high confidence to anomaly samples

Ideally, we need models with the decision boundary that have high confidence (low uncertainty) for the in-distribution data and low confidence (high uncertainty) for OOD samples. However, for deep neural networks, it is shown that the model can have high confidence even for OOD samples and be uncertain only on boundaries of decision boundary [2, 32, 33]. Figure 2.5 shows the ideal decision boundary and the decision boundary of a deep neural network trained to classify between two categories [2].

This situation can especially happen when new classes are added during the time, such as identifying a bacteria given its genomics sequence. This problem is referenced in [3]. As depicted in figure 2.6, the number of discovered bacteria types increases yearly. This means that the trained classifier is required to detect the newly discovered bacteria as an unknown type. However, the model is prone to assign the new bacteria to one of the existing classes with high confidence.

This issue with supervised approaches can have a severe and dangerous effect in real-world applications such as disease diagnosis in the medical field. Recognising a pathogenic sample as healthy is a tremendously expensive decision at the cost of human life.

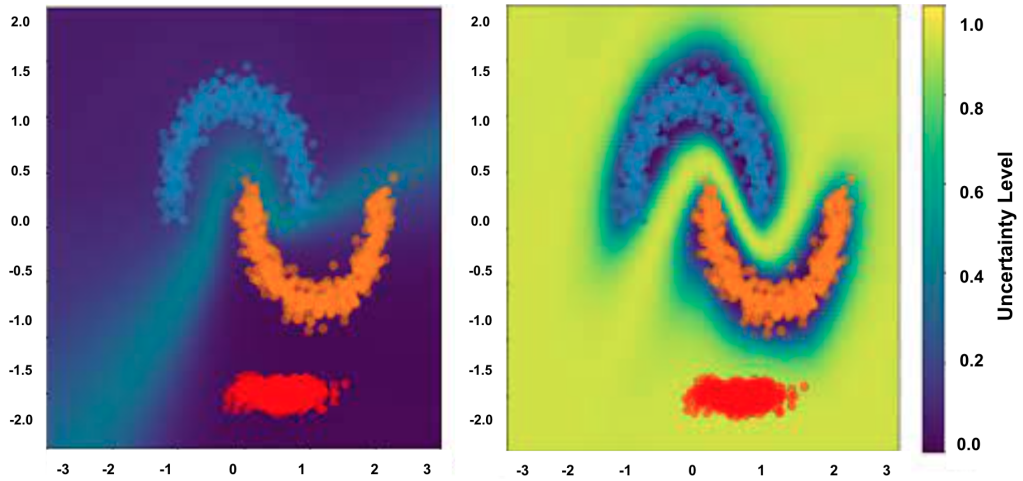


Figure 2.5: **Left:** Uncertainty of decision boundaries in deep neural networks. **Right:** Ideal decision boundary. Blue and orange dots are in-distribution data with different classes, and red dots represent OOD samples. Image from [2].

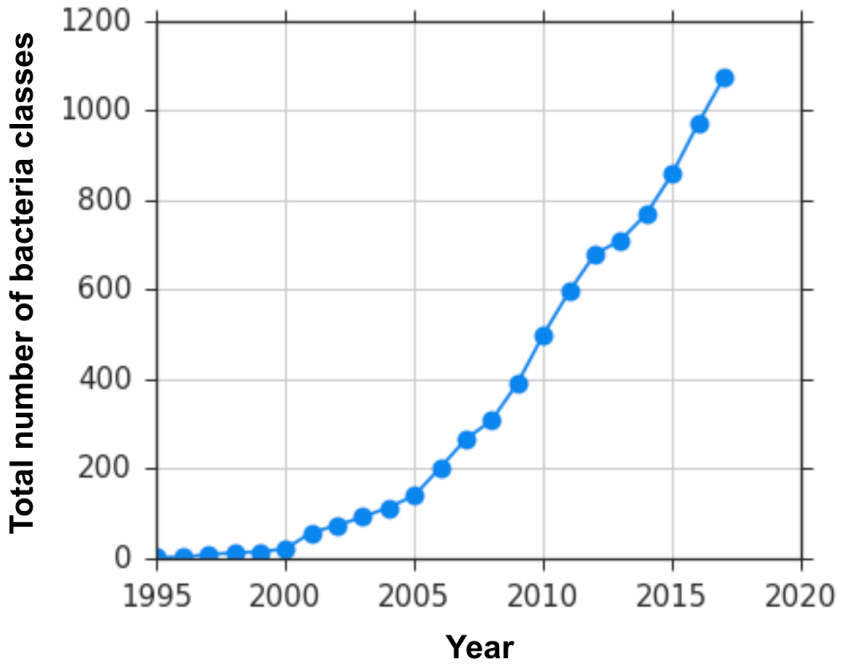


Figure 2.6: New bacteria categories are discovered over the years [3].

2.2.6 Supervised methods learn superficial features

In a study, it is shown that the features learned by a supervised method are superficial compared to features learned using a self-supervised method, where superficial means having no knowledge about high-level semantics such as object information. The details of the mentioned study will be explained, and then the near-OOD problem is introduced, where it is crucial for the model to be able to capture high-level

semantics to perform well in detecting OOD samples.

To compare the captured features by two supervised and self-supervised models, the authors in [4] train two models, one using a supervised method and the other leveraging a self-supervised approach on ImageNet dataset [34]. ImageNet is a large-scale dataset including images built based on the structure from WordNet [35] and includes 80000 of WordNet synsets, each with 500 to 1000 images. Next, a new set of images are added to the previous dataset by applying a scrambling-like transformation on ImageNet samples that destroys the object semantic but preserves the local statistics as depicted in figure 2.7.

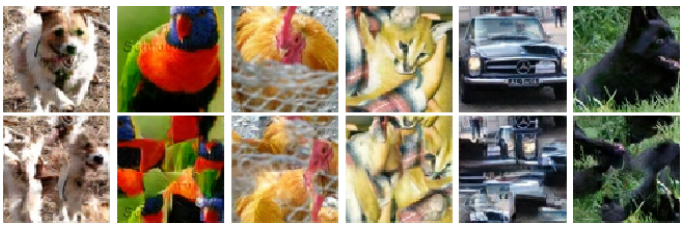


Figure 2.7: Top row: original images. Bottom row: Augmented images. The augmentation should keep the local statistics intact while changing the global semantics. A good image representation should be able to discriminate the augmented class from the original images [4].

The two trained models then are frozen therefor; their parameters are not updated in the next steps. The objective of the next step is to discriminate between these two types of samples by adding a linear layer to the two pretrained models. As it is depicted in figure 2.8, the model pretrained in a self-supervised manner provides higher accuracy. As the objective is to discriminate the samples with correct object semantics from samples with destroyed one but with similar low-level statistics, we can argue that features captured self-supervised pretrained model includes high-level semantics such as the knowledge of the object inside the image.

Near-OOD and Far-OOD

Out of distribution (OOD) detection can be categorised into near-OOD and far-OOD. With far-OOD problems, it is rather easy to detect OOD samples even by capturing shallow statistics such as color statistics. An example of a far-OOD problem could be detecting samples from street view house numbers (SVHN) dataset [36] against the CIFAR-10 dataset that includes natural images of different animals and vehicles up to 10 different classes [37]. Some of the randomly sampled images from SVHN and CIFAR-10 datasets are depicted in figure 2.9

In contrast to far-OOD, for the models to rely only on shallow statistics may not be efficient for the problem of near-OOD detection. Instead, machine learning models need to capture higher level semantics in order to be able to detect near-

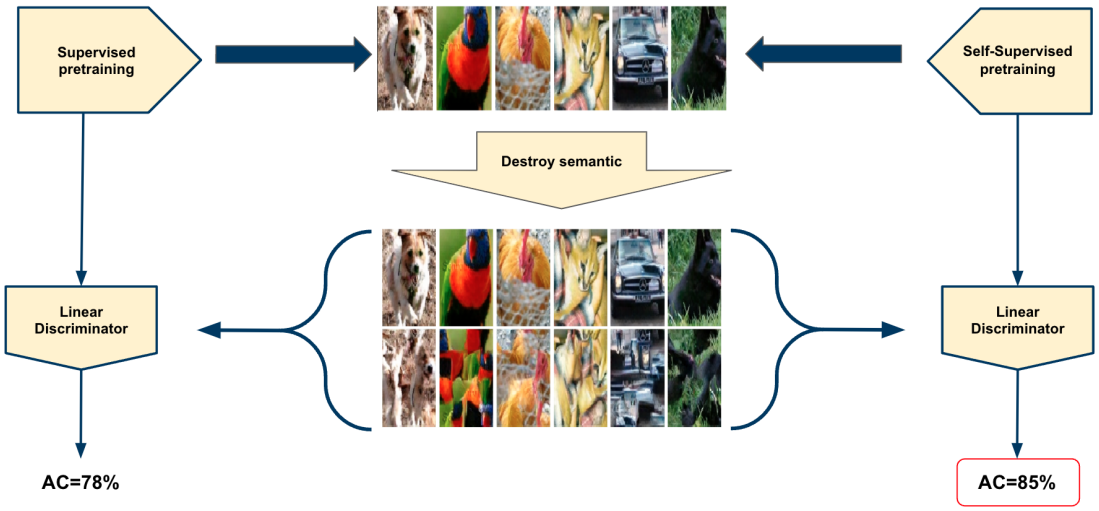


Figure 2.8: Two models are trained on the ImageNet dataset. One uses a supervised method, and the other is trained in a self-supervised manner. Extra samples are added to the dataset by applying an augmentation that destroys the object semantic but keeps low-level statistics such as color histogram. By adding a linear discriminator layer to each model, the goal is to evaluate which of these two models discriminate samples of correct semantics from the samples with object semantics destroyed. The self-supervised trained model achieves higher accuracy.

Figure 2.9: **Right:** Sample images from CIFAR-10 dataset. **Left:** Sample images from SVHN dataset.

OOD samples. An example of such a problem is to detect samples of CIFAR-10 from CIFAR-100. Images in dataset CIFAR-100 can take 100 different classes, such as different types of flowers, animals, and vehicles, where each class contains 600 images. There is no exact label overlap between images of CIFAR-10 and CIFAR-100; however, some categories are similar. For example, there is a possibility to have two different types of animals but with the same color skin or the same background, such as having a grassland background. Figure 2.10 shows some randomly sampled images from CIFAR-100 and CIFAR-10, where some of the images from these two datasets have similar color patterns.

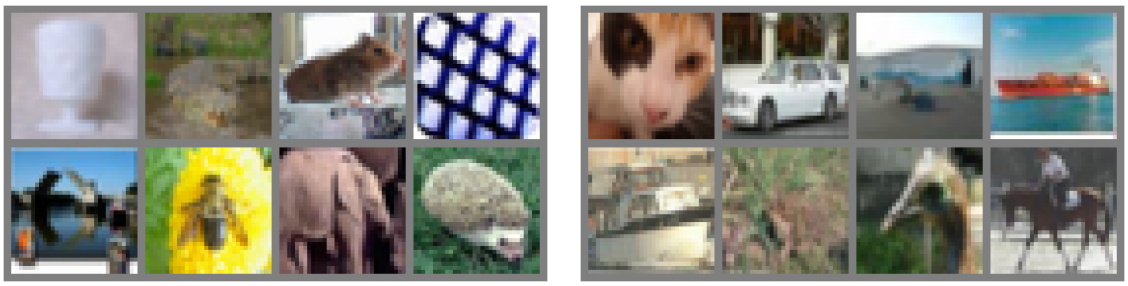


Figure 2.10: **Right:** Sample images from CIFAR-10 dataset. **Left:** Sample images from CIFAR-100 dataset. Some samples in these two datasets are similar and share low-level statistics. However, the object semantic is different.

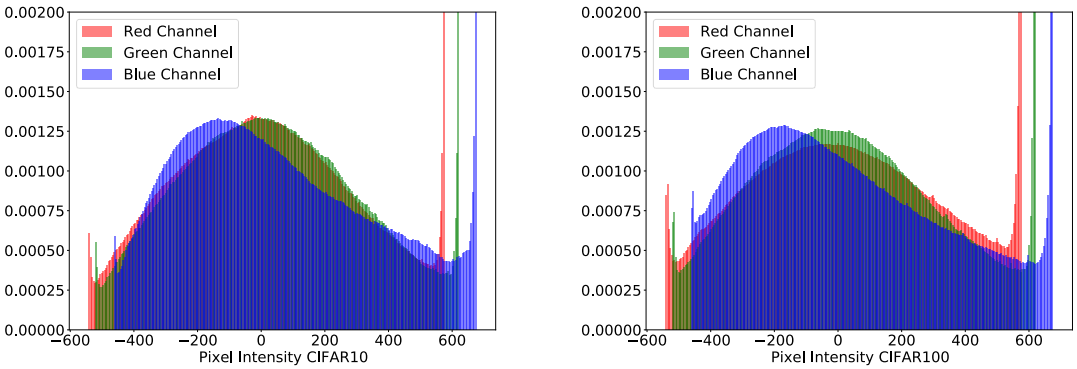


Figure 2.11: **Right:** Color distribution for samples of CIFAR-100 dataset. **Left:** Color distribution for samples of CIFAR-10 dataset. As a low-level statistic, the color distributions of these two datasets are approximately similar.

In addition to depicted samples from both CIFAR-10 and CIFAR-100 datasets, as it is shown in figure 2.11, we can see that the color histogram of these two datasets follows an approximately similar distribution. This implies that for successfully detecting images from one of these datasets against the other, a model is required to capture high-level semantics such as object semantics.

2.3 Unsupervised Approach to Anomaly Detection

According to limitations of supervised methods, such as lack of annotated data, unsupervised approaches, such as generative models, seem to be a natural choice. Some of the probabilistic generative models such as VAE [19], auto-regressive [20], and normalising flows [38] are studied for the problem of anomaly detection in different areas [39–42]. The idea here is to try to approximate the training data

distribution by learning a likelihood score. Ideally, after training, these likelihood based models should assign a higher likelihood score to in-distribution data than the dataset of OOD samples. However, different studies show that these models can either fail to discriminate between in-distribution and OOD samples [3] or, on average, assign a higher likelihood to OOD samples compared to in-distribution data [26, 43].

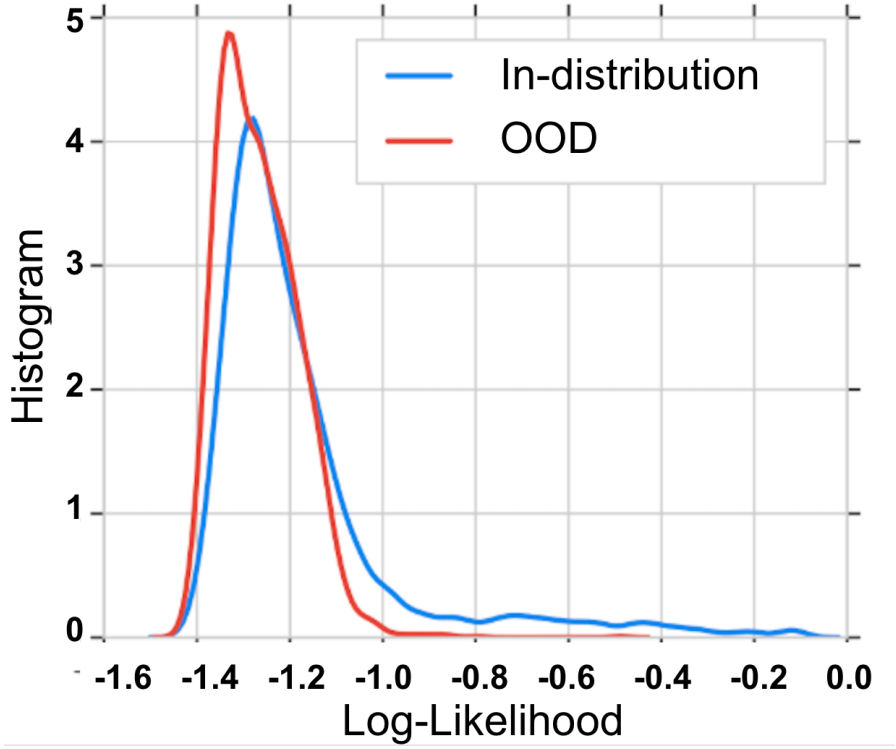


Figure 2.12: Log-likelihood distributions of the trained deep generative model for in-distribution and out of distribution genome samples significantly overlap. Image from [3].

In [3], two datasets of bacterial genomics sequences are considered. Based on the date of discovery and if that class of bacteria is discovered before a given date, the sample is considered as in-distribution and out of distribution otherwise. Then a deep generative model is trained on the in-distribution dataset with the aim that the likelihood score of the trained model is smaller for OOD samples. However, as depicted in figure 2.12, the two distributions that present the log-likelihood scores significantly overlap.

The authors in [3] study a similar experiment in which a deep generative model is trained on Fashion-MNIST dataset [44], and samples from MNIST [45] are considered as OOD. MNIST dataset contains 60000 training and 10000 test examples of handwritten digits between 0 to 9 that are centered and size-normalised. The Fashion-MNIST dataset contains the same number of train and test samples as

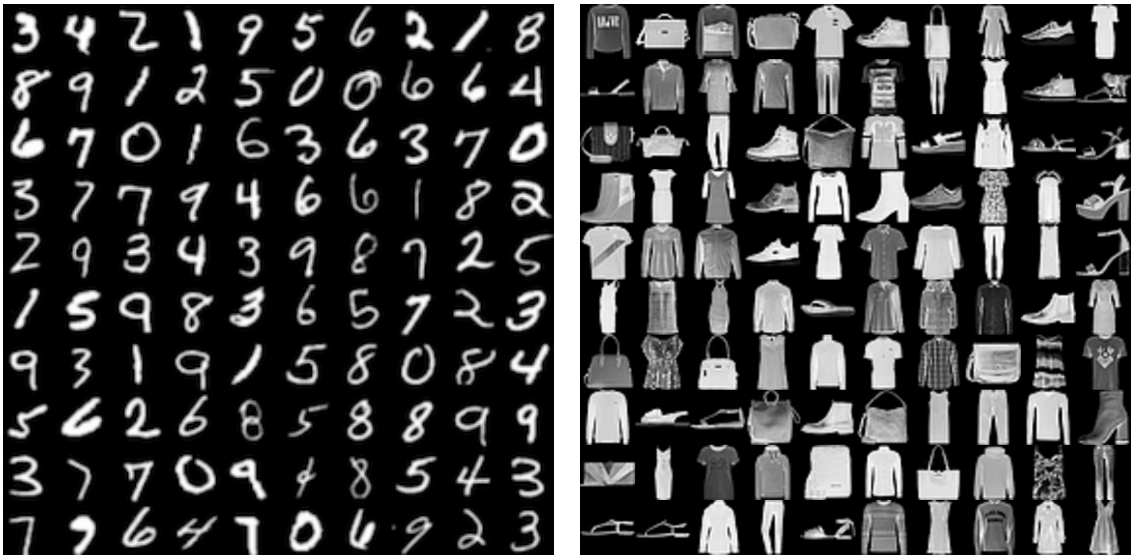


Figure 2.13: **Right:** Samples from Fashion-MNIST dataset. images belong to fashion articles. **Left:** Handwritten image samples from MNIST dataset.

MNIST. However, the samples belong to images from fashion articles. Figure 2.13 shows examples of these two datasets.

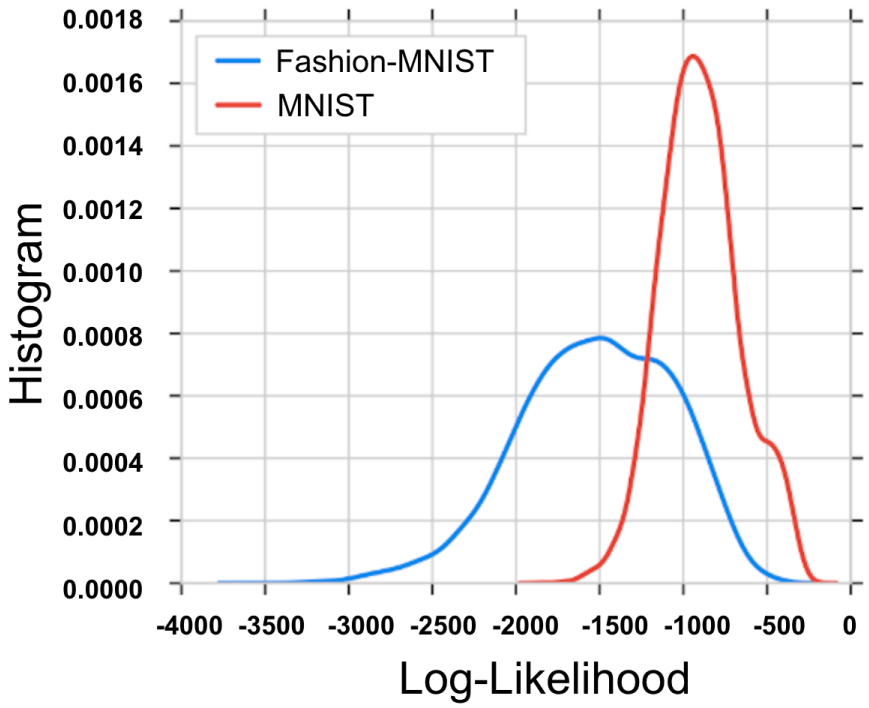


Figure 2.14: Trained deep generative models, on average, assigns a higher likelihood score to OOD samples of MNIST when it is trained on the Fashion-MNIST dataset. Image from [3].

Surprisingly, as it is depicted in figure 2.14, the trained deep generative model

assigns a higher likelihood score to the OOD samples of MNIST. As it is discussed by the authors, the reason for this failure in detection is that the likelihood can be affected by background statistics instead of the object semantic, including GC (guanine-cytosine) content and pixels with zero value in the genomics and Fashion-MNIST datasets, respectively. To train the deep generative models, the likelihood is computed on the input space, and the model needs to consider every single pixel in the image where the background pixels constitute the major part. The same detection failure is also introduced in [26] for detecting SVHN against CIFAR and using PixelCNN [21] and Glow [27] architectures.

To overcome this issue, a likelihood ratio is introduced in [3]. The idea here is to train two models with different focuses. First, the model is optimised to capture only background statistics, and the second model is trained as before, which captures both background statistics and high-level semantics. The likelihood ratio can be defined as in equation 2.1.

$$LLR(\mathbf{x}) = log\frac{p_\theta(\mathbf{x})}{p_{\theta_0}(\mathbf{x})} \tag{2.1}$$

where p_θ is the likelihood score generated by the model trained on in-distribution data and p_{θ_0} is the likelihood score generated by the model optimised on perturbed training data. The perturbation should destroy the semantic information enforcing the model to rely only on background statistics. The likelihood ratio approach is relatively successful in overcoming the mentioned problem; however, the performance lacks behind the recently proposed self-supervised methods [46–48]. Additionally, it is not always clear how to distinguish the background content.

Chapter 3

Machine Learning Basics

3.1 Introduction

In this chapter, some of the fundamental concepts in machine learning (ML) will be explained. These concepts are necessary to understand the content of the following chapters.

3.2 Deterministic vs Probabilistic Models

Models used in ML can be categorised in two different categories. First, models as a deterministic parametric function, and second, probabilistic models [49]. Each of these approaches has its own benefits and challenges. The main task in deterministic parametric models is to find the best value for the model parameters through an optimisation task. However, in probabilistic models, the goal is to learn a distribution over the model parameters instead of a single best value. The focus of this chapter, in particular, and this thesis in general, is the first approach; However, the probabilistic models will be shortly introduced after some of the essential concepts in deterministic parametric models are explained. Unless otherwise mentioned, by ML models, we mean deterministic parametric models used in machine learning algorithms.

3.2.1 Deterministic Parametric Models

ML models, as a deterministic function, map a particular input to an output where input and output can have different dimensions. For example, equation 3.1 shows an ML model which maps input samples of dimension D from real value space to

an output of dimension 1 of real value space.

$$f : \mathbb{R}^D \rightarrow \mathbb{R} \tag{3.1}$$

ML models are usually associated with a set of parameters θ with initial values which can be optimized based on a defined objective function. The optimization process can be done by various means, such as stochastic gradient descent, which is a numerical optimization technique [50]. ML optimization methods aim to find values for the model parameters that satisfy the objective function for training data and the same objective for unseen data. The ability of the model to have a good performance on unseen data is known as generalization.

3.2.2 Independent and Identically Distributed Samples

Considering ML problems, samples are assumed to be independent and identically distributed (i.i.d.). By independent, we mean statistical independence, and by identical, we mean that all the samples follow the same distribution with shared parameters.

3.2.3 Empirical Risk Minimization

To quantify the performance of the model during the training optimization, we need to define a loss function $\ell(.)$ based on model prediction and the true values. Having N training samples $(\mathbf{x_1}, y_1), ..., (\mathbf{x_n}, y_n)$, The training goal is to find values for θ that minimize the average loss for all N training data. Minimizing the average loss function on training data is called empirical risk minimization (ERM) [49]. Note that as samples are i.i.d., the empirical average is a good estimation for the mean of the population. Assuming $\hat{y}_i$ as model output, we can formally define ERM as in equation 3.2.

$$\mathbf{R}_{emp}(f, \mathbf{X}, \mathbf{y}) = \frac{1}{N} \sum_{n=1}^{N} \ell(y_n, \hat{y}_n) \tag{3.2}$$

where $\hat{y}_n = f_\theta(\mathbf{x})$.

3.2.4 Maximum Likelihood Estimation

As explained in the previous section, to minimize the empirical risk, we need to define and minimize a loss function. Maximum likelihood estimation (MLE) $P(\mathbf{x}|\theta)$ is a method that enables us to find a model that best fits the training data [51]. The goal is to maximize $P(\mathbf{x}|\theta)$ or, in practice, minimize the negative log of it given by equation 3.3.

$$\mathcal{L}_{train}(\theta) = -\log p(\mathbf{x}|\theta). \tag{3.3}$$

Intuitively, having fixed observed training data $\mathbf{x}$, by minimizing $\mathcal{L}_{train}(\theta)$ we find the most likely parameter θ. Unless otherwise mentioned, we use $\mathcal{L}$ as $\mathcal{L}_{train}$.

3.2.5 Maximum A Posteriori Estimation

Optimizing the maximum likelihood, we assume no prior knowledge about the model parameters θ. Maximum A Posteriori (MAP) estimation enables us to include more specific knowledge regarding the model parameters. MAP estimation can be understood by considering a prior distribution for model parameters θ. Multiplying this prior to the likelihood term and using the Bayes theorem allows us to compute a posterior distribution $p(\theta|\mathbf{x})$ on model parameters as formulated in equation 3.4.

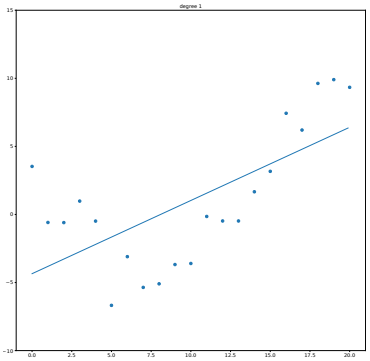
$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}. \tag{3.4}$$

3.2.6 Overfitting

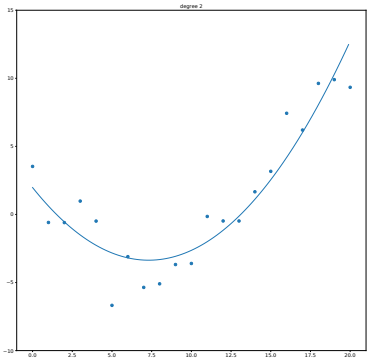
One crucial issue that must be considered when fitting a model to data is to avoid overfitting. This can happen when we choose a complex model with a large number of parameters that can model every detailed variation in the data, including available noise samples. Figure 3.1 shows four different linear models with four polynomial degrees. The higher the polynomial degree, the higher the capacity of the model to fit every single data point.

A model with high flexibility can fit every sample in training data instead of learning the existing patterns, failing to generalize its predictive performance on unseen data. To detect if overfitting is happening, we can compare the gap for model performance on training data $\mathcal{L}_{train}$ and for the entire population. This gap is known as the generalization gap [5]. However, we do not have access to the entire population; thus, we use an approximation of this gap using a subset of the whole data known as the test set $\mathcal{L}_{test}$. Figure 3.2 shows the difference between $\mathcal{L}_{train}$ and $\mathcal{L}_{test}$ when polynomial models with different degrees of freedom are used for training.

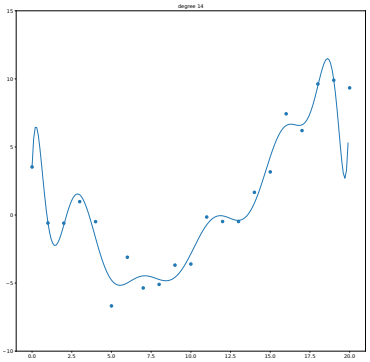
Several techniques have been studied to avoid overfitting, such as early stopping, adding more training samples, and regularization. Early stopping is stopping the training iterations before the model can fit all the training samples. Similarly, adding more training data prevents the model from using all of its capacity to memorize each sample. Regularization can be performed in different ways but by imposing limitations on the values that model parameters can take. MAP estimations can be understood as applying a regularization since we are considering a prior distribution over the model parameters. Having data distribution $p(\mathbf{x})$ fixed, the model parameters θ can be estimated by minimizing the negative log of $p(\theta|\mathbf{x})$



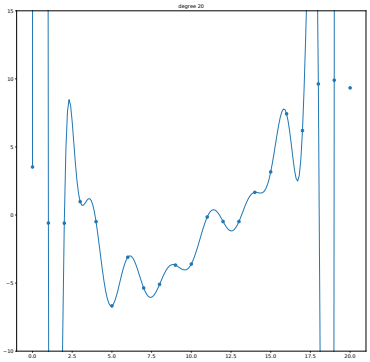
(a) Degree 1



(b) Degree 2



(c) Degree 14



(d) Degree 20

Figure 3.1: A linear model fit with four different polynomial degrees. **a**: Polynomial of degree 1. **b**: Polynomial of degree 2. **c**: Polynomial of degree 14. **d**: Polynomial of degree 20. Diagrams are reproduced from [5].

as mentioned in equation 3.5.

$$\hat{\theta} = \arg\min_{\theta} -\log p(\theta|\mathbf{x}) = \arg\min_{\theta} -\log p(\mathbf{x}|\theta) - \lambda \log p(\theta) \tag{3.5}$$

where λ controls the strength of regularization. Assuming a normal distribution $\mathcal{N}(0, 1)$ for θ, equation 3.5 can be rewritten as equation 3.6

$$\hat{\theta} = \arg\min_{\theta} -\log p(\mathbf{x}|\theta) + \lambda \|\theta\|_2 \tag{3.6}$$

which is known as Ridge regularization.

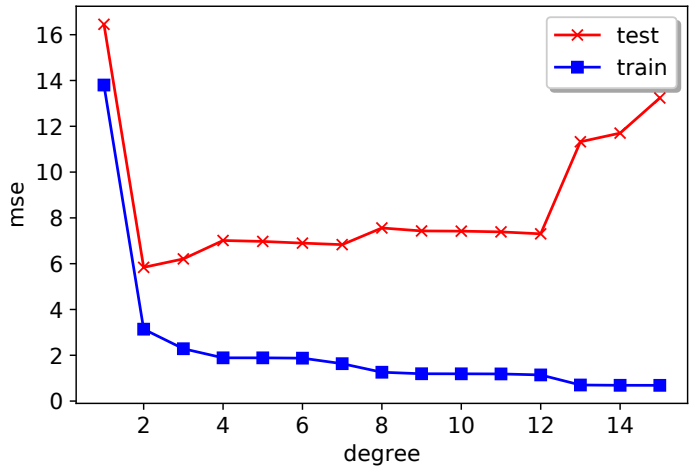


Figure 3.2: Increasing the degree of freedom for the polynomial model results in overfitting. Reproduced from [5].

3.2.7 Probabilistic Models

The key goal of MLE and MAP methods is to estimate a single value for θ through an optimization task. Finding the best value θ^*, the prediction task can be performed for any given sample to estimate $p(\mathbf{x}|\theta^*)$. However, sometimes having the parameter value which maximizes the posterior is not enough, and it is required to have full information about the posterior distribution for robust decision making. To achieve this, probabilistic modeling leverage the Bayesian theorem.

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}, \quad p(\mathbf{x}) = \int p(\mathbf{x}|\theta)p(\theta)d\theta \tag{3.7}$$

As it is evident in equation 3.7, the key problem in probabilistic models is to solve the integration, whereas, in models as function, the critical problem is to perform a point estimation for θ^*.

3.3 Deep Neural Networks

Deep neural networks (DNNs) have their root in feedforward neural networks (FFDNN) or multi-layer perceptrons (MLP). An MLP consists of a stack of layers wherein each layer, a linear combination of outputs from the previous layer is followed by an activation function (see figure 3.3). The activation function can be both linear and non-linear. However, using a linear one reduces the MLP into a linear model. In contrast, the idea behind MLP is to create a non-linear relation between input and output, which is not the case with models such as linear and logistic regression.

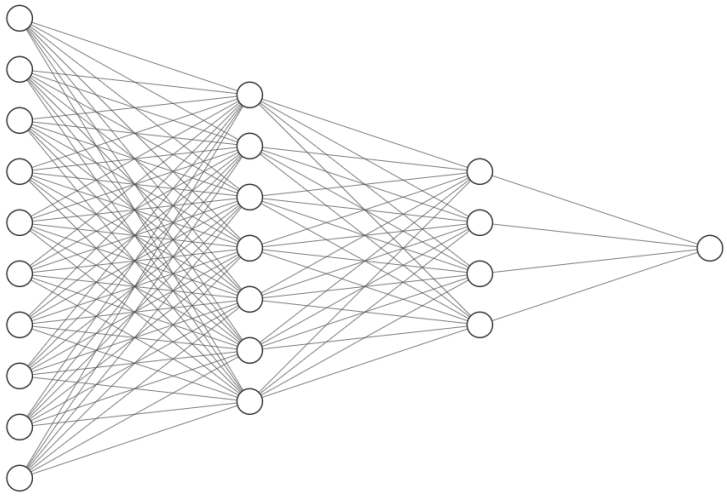


Figure 3.3: FFDNN consists of a stack of layers when each node in each layer is connected to all the other nodes in the previous and next layers.

Formally for each layer, l, output z_l can be calculated using equation 3.8.

$$\mathbf{z}_l = f_l(\mathbf{z}_{l-1}) = \phi_l(\mathbf{b}_l + \theta_l \mathbf{z}_{l-1}) \tag{3.8}$$

where ϕ is an activation function, $\mathbf{b}_l$ is the bias term, and θ_l is the parameters of the layer l, which can be optimized to fit the training data best. In general, we can opt from both sets of non-differentiable and differentiable functions for activation functions. However, the former comes with the difficulty that we need to hand engineer the model parameters θ. In contrast, with the latter, we benefit from techniques such as stochastic gradient descent (SGD)and loss backpropagation to update the parameters.

3.3.1 Backpropagation

To minimize the loss function -the error from model prediction and the true target- we can calculate the derivative of the loss function with respect to the given parameters. The idea of calculating the loss derivative for parameters of each layer using the chain rule starting from the last layer and propagating to the first layer is referred to as backpropagation [12].

3.3.2 Stochastic Gradient Descent

Considering our loss function $\mathcal{L}$ as a differentiable function of our model prediction, we can calculate the minimum of this function using gradient descent (GD) algorithms depicted in equation 3.9 when a closed-form solution can not be found analytically.

$$\theta^t = \theta^{t-1} + \eta \nabla \mathcal{L}(\theta^t) \tag{3.9}$$

where η is the learning rate and defines the size of steps. Intuitively, using GD, each parameter in the model is updated with small steps in the direction of the negative gradient, which intuitively means toward the direction where the decrease in the error function happens with the highest rate.

As explained earlier in the previous section, having N training examples, instead of maximizing the likelihood over the joint probability of all examples, the negative log-likelihood of the sum over the training data is minimized so that

$$\mathcal{L}(\theta) = -\sum_{n=1}^{N} \log p(\mathbf{x}_n|\theta) \tag{3.10}$$

where $\mathbf{x}_n$ is the training sample and θ is the model parameter. Subject to this, the $\nabla\mathcal{L}(\theta)$ can be calculated as in equation 3.11.

$$\nabla\mathcal{L}(\theta) = -\sum_{n=1}^{N} \nabla \log p(\mathbf{x}_n|\theta). \tag{3.11}$$

Calculating the gradient descent for complete training examples-known as batch gradient descent- can be time consuming, especially when we have datasets with an enormous number of samples. Instead of using all the samples in each update, we can randomly select a subset of them and apply the gradient descent. This method is referred to as stochastic gradient descent (SGD). Taking enough small learning rate, SGD can effectively converge to a local minimum [52].

3.3.3 Convolutional Neural Network

The connection between two layers of an FFDNN, as depicted in figure 3.3, looks like a fully connected graph where every node in the current layer is connected to all the nodes in the previous and next layer. This gives the model high flexibility and capacity, which can cause the model to overfit. Another issue with FFDNNs is when they are fed with structured data such as image samples. Image data usually come with a spatial structure, such as the reoccurrence of objects or patterns in different positions. FFDNNs have no bias regarding these spatial structures, as the weights are not shared. As depicted in figure 3.4 change in object position results in a completely different response. Convolutional neural networks (CNNs) can solve these problems by replacing the matrix multiplication $\theta\mathbf{x}$ with a convolution operation. Intuitively, convolution works by sliding parameterized filters over the input. These filters can have different sizes, e.g., the common sizes used for image data are 3×3, 5×5, and 7×7. These filters are small size (compared to input image size) weight matrices with shared parameters, thus considerably reducing the number of model parameters, making them less prone to overfitting [5]. Recently many different convolutional neural network models such as ResNet [53], VGGNet

[54], and DensNet [55] with promising results in image classification and object detection have been introduced.

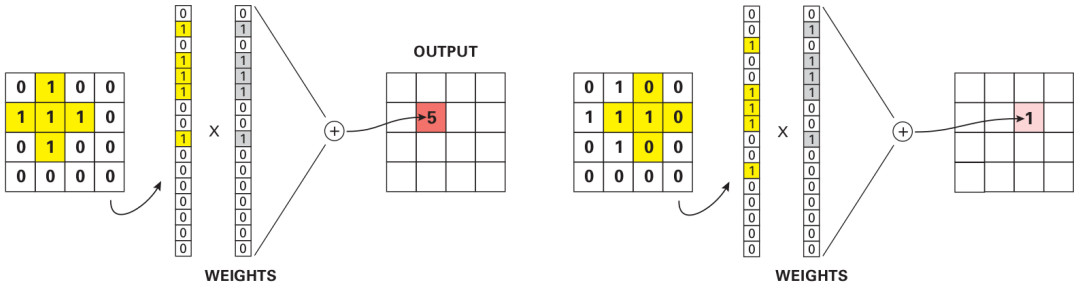


Figure 3.4: MLPs are not translation invariant; thus, changing the position of the same pattern inside the image results in a different response from the model. Taken from [5].

3.3.4 Vision Transformers

Transformer models were first introduced by [56] and used for the text data. The main idea behind the transformer is using a set of quadratic operations named multi-head attention, which is a set of inner products between embeddings of all input text tokens. Intuitively these inner products calculate the pairwise interaction between all words of a given text to help the model capture existing long-range correlations among the given words. Vision transformers (ViTs) use the same approach for image data. Calculating the pairwise interaction between all pixels of an image is computationally not feasible thus, the authors in [57] divide each image into several patches and use the embedding of each patch as the input for attention calculation. One issue with ViTs is that these models are prone to overfitting and usually have good performance if enough samples and data augmentations are available [58, 59].

Chapter 4

Self-supervised Representation Learning

4.1 Introduction

Self-supervised learning (SSL) has recently emerged and is used in deep learning research, particularly in the visual and text domain. One main aspect of SSL is removing the need for human-generated labels. Recently researchers have approached this goal either using some automated way of generating labels for a pretext task independent of the final downstream task [7–9, 60] or by considering each sample(instance) as one class and defining pretext task based on different augmentations(views) of each instance[6, 61, 62].

One of the early motivations for SSL was the scores reported in [63] for supervised learning of object recognition. As reported, there is a large margin between the top-5 and the top-1 classification error. Moreover, the images for the highest and the second highest response of softmax output are more likely to be correlated or even be from the same category. Figure 4.1 shows that given an input image of a leopard, correlated samples such as images of a jaguar and cheetah have higher softmax activation compared to images of a bookcase and lifeboat. These findings reveal the fact that even with typical discriminative learning, it is possible to learn features that capture similarity among semantically similar categories.

Similar to findings from supervised learning, it has been shown that the features learned by SSL based methods also can be used to improve on various down-stream tasks such as classification, object detection, and anomaly detection, although the pretext task is defined independently of the final task[8, 9, 46, 47].

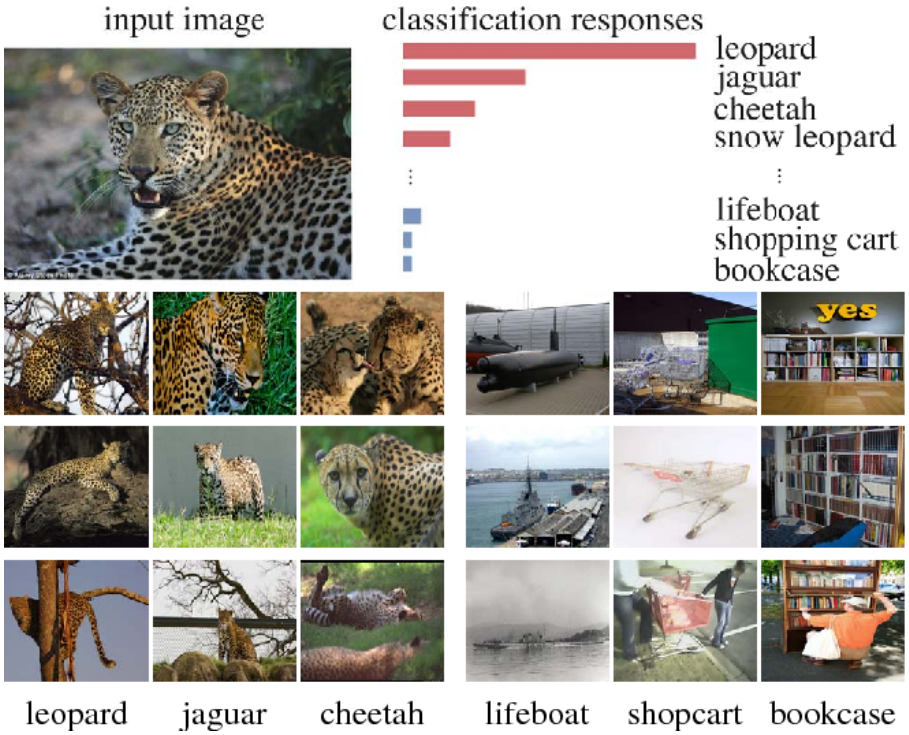


Figure 4.1: A model trained to classify labeled data generates a higher response for images of similar categories for a given class. In the figure, for the class of leopard, images of similar classes, such as jaguar and cheetah, receive higher softmax probability compared to images from a completely different class, such as the lifeboat, shop cart, and bookcase. The image is taken from [6].

4.2 Challenges in Supervised Methods

As already mentioned in previous chapters, supervised learning always has the challenge of lacking annotated data. However, a recent study shows that even with the existence of annotated data, supervised methods do not always guarantee learning a meaningful representation[64]. A study conducted by [64] shows that the signals learned by a supervised trained model can bias the network to focus on superficial patterns instead of recognizing the object inside the image. In an experiment, they show that while a color-preserving style transfer is applied from an image of a leopard to an image of a car, a pretrained classifier assigns a higher likelihood to the class leopard rather than the class car.

In a different study by [4], the representation learned by a supervised method and an SSL method is compared in terms of having a better sense of objects positioned inside the image. To do this, authors in [4] train a binary linear classifier on features extracted from a model pretrained on ImageNet labels and from a model pretrained using the SSL method. The aim of the binary linear classifier is then to discriminate between the original image and a second class made by scrambling the original images in a way that local statistics are preserved. The result shows that

when features from SSL pretraining are used, the binary linear classifier achieves an accuracy of 85% whereas, for the features extracted from supervised pretraining, the accuracy is 78%, which shows that the SSL pretraining provides a representation with a sounder understanding of the objects inside the image.

SSL methods discussed in this section can be divided into two groups. The first group uses an autoencoding approach, and the pretext task is defined on input space, whereas the second group defines the pretext task on latent space, which is a mapping of input data into a usually lower-dimensional space. Early methods for the latter approach use different heuristics to automatically generate pseudo labels based on the image content, such as solving a jigsaw puzzle [9] or predicting the degree of applied rotation [60]. On the other hand, subject to shortcomings of these methods, some recently introduced methods explore the techniques where each sample(instance) is considered as one class, and the task is to maximize the agreement between different views of each instance. Following, some of these methods will be explained in more depth.

4.3 Pretext Tasks on Input Space

The loss function used in these pretext tasks is defined on input space similar to the objective functions of autoencoder models.

4.3.1 Feature Learning by Inpainting

Inpainting is one of the early methods used to define a pretext task. In [7], one part of the input image is masked, and a convolutional neural network autoencoder is trained to reconstruct the entire image, including the missing part. Figure 4.2 shows an overall view of inpainting autoencoding. Despite their diversity, usually, image data are highly structured in terms of pixel patterns. This helps the model to consider the context of pixels surrounding the missing part and therefore encouraging learn high-level semantic features. Autoencoding in inpainting is partly similar to denoising autoencoders [65] since both avoid the encoder to learn features, which are a simple compression of input data like in the standard autoencoding method. Inpainting autoencoder differs from denoising one as in the former, a large part of the image is missing, and contextual information is required for reconstruction. However, only low-level corruption is applied to the input image in the latter. Intuitively in inpainting autoencoders, the encoder should capture higher-level semantics to be able to fill enough large size missing parts; however, in denoising one, the encoder that captures more of the low-level features can solve the denoising task. Similar approaches in the text domain have been used to learn a high-level semantic feature representation of text data[66]. Formally, given input data x, the reconstruction loss for the inpainting pretext task can be formulated as

in equation 4.1

$$\mathcal{L}_{rec}(x) = \|M \odot (x - F((1 - M)) \odot x)\|_2^2 \tag{4.1}$$

where M is a binary mask. Corresponding dropped regions of the image have the value 1 in M and 0 for the rest of the pixels.

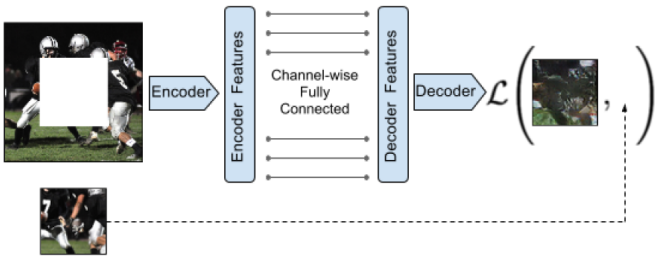


Figure 4.2: An overall view of feature learning by inpainting approach. Part of the image is masked and fed into the encoder. The generated encoder output is passed to the decoder through a channel-wise fully connected layer. The use of a channel-wise fully connected layer is to help each unit in the decoder to have information about the entire image. The image is taken from [7].

Despite the promising results reported in [7], this method has some limitations. One of the problems with this method is that it is not clear what is the optimal way to define the region for dropping the pixels. Additionally, this method can be prone to shortcut learning as features processed through convolutional layers can latch onto the features of boundary pixels, which helps the model decrease the loss without learning the context of the image.

4.4 Pretext Tasks on Latent Space

In the previous methods, the pretext task is defined on the decoder output and in input space which means that the decoder needs to generate pixel-wise correct predictions and encourages the encoder to capture low-level statistics. This can be problematic when we seek to capture high-level semantics, especially in the vision domain, e.g., where we are interested in using the trained encoder to do object detection tasks with transfer learning. A solution to this problem is to define pretext tasks on latent space instead of input space. In [8], an input image x is divided into non-overlapping patches, and a convolutional neural network is optimized to predict the relative position of two randomly sampled patches of the given image. In particular, as depicted in figure 4.3, for a given image, the first patch is randomly sampled, and the second patch is selected from each of the possible eight neighboring locations. The model is then trained to predict the location.

Splitting an image into patches and predicting the relative position of two randomly sampled patches comes with some issues regarding shortcut learning. As

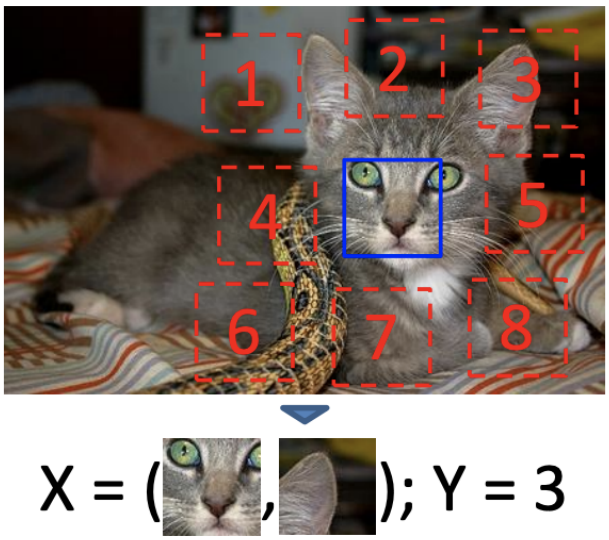


Figure 4.3: The model is trained to predict the relative location of two randomly sampled patches. The image is taken from [8].

stated in [8], low-level signals such as pixel patterns around the patch boundary or common textures in some of the patches could be a source of information leak. Exposed to this information leak, the model can optimize the loss function without the need to capture high-level semantics. To address the mentioned problem, a gap is included between patches, and each patch location is jittered by up to 7 pixels. Surprisingly, a chromatic aberration -resulting from the lens differently focusing different wavelengths- can also be a source of information leak. To alleviate this problem, authors in [8] propose to randomly replace 2 of 3 color channels with Gaussian noise.

4.4.1 Feature Learning by Solving Jigsaw Puzzles

One difficulty with the approach used in [8] is that the model always sees two randomly selected patches separately from other patches and is never fed with the complete image information. This makes it difficult for the model to find each patch's association as part of a specific object within an image, especially when patches have similar textures and patterns. Subject to this, in [9], a new pretext task is defined based on the arrangement of all patches(parts). In this method, at first, a subset of all possible permutations is selected. Then a given image is divided into non-overlapping patches, and patches are reordered based on a randomly chosen arrangement from a defined subset and fed to the model. The model is then optimized to predict the index of applied permutation from the selected subset. Figure 4.4 shows an overall view of the context prediction model. Formally, the model prediction for each arrangement S can be modeled as a conditional proba-

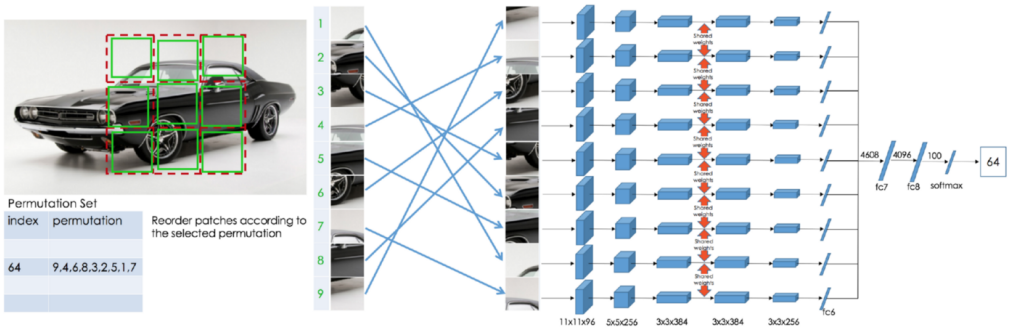


Figure 4.4: An overall view of the method used in [9]. A part of the image is randomly selected and cropped from the original image (shown by the red dashed box). The selected part is divided into a 3×3 grid, and each cell is randomly cropped. These cells are then randomly reordered based on a chosen permutation, and the model is optimized to predict the index of this permutation. Image is taken from [9].

bility density function (pdf) as stated in equation 4.2

$$p(S|A_1, A_2, ..., A_9) = p(S|F_1, F_2, ..., F_9) \prod_{i=1}^{9} p(F_i|A_i) \tag{4.2}$$

where A_i is the i-th part of the input and $\{F_i\}_{i=1,...,9}$ is the represented feature by the model. When generating only one random jigsaw puzzle per image is like considering S as a unique and fixed arrangement. This can encourage the model to learn a shortcut which is the absolute position of each image part, while the objective was to associate each F_i with the semantic attributes of A_i to identify the relative positions of image parts. To solve this issue, authors in [9] considered S as a list of tile positions $S = (L_1, ..., L_9)$ and by generating several random jigsaw puzzles per image, the conditional pdf can be rewritten by a new factorization of independent terms as in equation 4.3

$$p(L_1, L_2, ..., L_9|F_1, F_2, ..., F_9) = \prod_{i=1}^{9} p(L_i|F_i) \tag{4.3}$$

where each L_i is now completely determined by F_i. It is necessary to make sure that within the selected subset, among all the possible permutations, each part has an almost equal chance for each of the locations. This stops the model from learning the positions instead of patterns. To do this, the chosen configurations within the selected subset should have a large average Hamming distance. Additionally, similar to [8], a random gap is applied between the parts to avoid shortcut learning as a consequence of edge continuity.

4.4.2 Rotation Prediction

The approach [8, 9], where the pretext task is defined on model output for patches of an image, requires considerable human intervention in order to avoid shortcut learning. Additionally, even with human intervention, it is still not clear what is the best way to create patches from a given image. Moreover, some careful consideration is required for model architecture design and training like the siamese-ennead convolutional network used in [9] to avoid shortcut learning (see 4.4). In [60], a different pretext task is defined without the need to divide each image into patches. The authors in [60] propose the geometric transformation prediction as a pretext task to learn high-level semantic features. To achieve this, a set of K discrete geometric transformations $G = \{g(.|y)\}_{y=1}^{K}$ is defined where $g(.|y)$ is transformation with label y which is applied to the input image x. The transformed image x^y is then defined as $x^y = g(x|y)$. Finally, the pretext task is defined by optimizing the loss function $\mathcal{L}$ for N training images as stated in equation 4.4

$$\mathcal{L} = -\frac{1}{N * K} \sum_{i=1}^{N} \sum_{y=1}^{K} log(F^y(g(x_i|y)|\theta)) \tag{4.4}$$

where $F^y(g(x_i|y)|\theta)$ is the prediction for transformation with label y from a convolutional network model F parameterized by θ.

4.5 Instance Based Discrimination

A common issue with the methods that have been introduced is that they hugely rely on the human-level heuristics used to define the pretext task, limiting the generality of learned features for diverse downstream tasks. Additionally, these heuristics usually come with many challenges to avoid shortcut learning. Moreover, it is not clear what is the optimal heuristic. Following, some of the studies which address this problem will be introduced. The main focus will be on contrastive based methods and self-distillation using student and teacher approaches.

4.5.1 Contrastive Methods

In an attempt by [6, 61, 62], each instance in the training data is considered as one class. The pretext task is then defined as maximizing the distance between the representation of what is known as similar samples while minimizing it between the representation of different samples. Similar representations could be representations from an image generated in different time steps as in [6] or from different augmentation of an image as in [61]. Subject to this, having a training dataset of N samples $x_1, ..., x_N$, we have N distinguished classes.

In [6], all representations of train samples are stored in a memory bank. Then,

as stated in equation 4.5, a non-parametric softmax is proposed to model the probability of recognizing $\mathbf{v}_i$ in the memory bank as the representation corresponded to sample x_i

$$p(\mathbf{v}_i|\mathbf{v}) = \frac{exp(\mathbf{v}^T\mathbf{v}_i/\tau)}{\sum_{j=1}^{N} exp(\mathbf{v}_j^T\mathbf{v}/\tau)} \tag{4.5}$$

where $\mathbf{v}_i = f_\theta(\mathbf{x}_i)$ and f_θ is a convolutional neural network with parameters θ. Note that every $\mathbf{v}$ is stored in a memory bank after it is normalised with L_2 norm; thus $\|\mathbf{v}\| = 1$. Figure 4.5 shows a general overview of the algorithm used in [6].

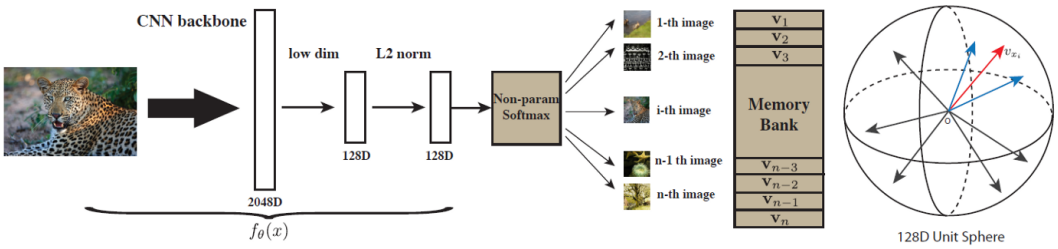


Figure 4.5: A general overview of approach used in [6]. An image is passed through a CNN and mapped to a lower-dimensional space. The loss is then calculated on the generated representation of the image and the representation stored in the memory. The image is taken from [6].

Calculating the softmax is costly when an enormous number of instances(classes) exist in the training dataset. To address this problem, authors in [6] adapt the noise-contrastive estimation (NCE) [67]approach where a multi-class classification is cast into a set of binary classification tasks discriminating between data instances and noise data. Considering a dataset of size N, a uniform distribution $P_n = \frac{1}{N-1}$ is assumed for noise samples which in [6] includes all the other $M = N - 1$ samples except the x_i. Suppose that the binary classifier outputs label $D = 1$ if $\mathbf{v}_i$ corresponds to $\mathbf{v}$ then, the probability can be figured out as in equation 4.6.

$$h(\mathbf{v}_i, \mathbf{v}) := p(D = 1|\mathbf{v}_i, \mathbf{v}) = \frac{p(\mathbf{v}_i|\mathbf{v})}{p(\mathbf{v}_i|\mathbf{v}) + Mp_n(\mathbf{v}_j)} \tag{4.6}$$

Calculation of $p(D = 1|\mathbf{v}_i, \mathbf{v})$ and $p(D = 0|\mathbf{v}_i, \mathbf{v})$ can be realised by at first considering the joint probability $P(D = 1, \mathbf{v}|\mathbf{v}_i)$

$$p(D, \mathbf{v}_i|\mathbf{v}) = \begin{cases} \frac{1}{M+1}p(\mathbf{v}_i|\mathbf{v}) & if \quad D = 1 \\ \frac{M}{M+1}p_n(\mathbf{v}_j) & if \quad D = 0 \end{cases} \tag{4.7}$$

then each of these probabilities can be calculated using the Bayes theorem.

$$
\begin{aligned}
p(D=1|\mathbf{v}_i,\mathbf{v}) &= \frac{p(D=1,\mathbf{v}_i|\mathbf{v})}{p(D=1,\mathbf{v}_i|\mathbf{v})+p(D=0,\mathbf{v}_i|\mathbf{v})} = \frac{1}{1+\frac{Mp_n(\mathbf{v}_j)}{p(\mathbf{v}_i|\mathbf{v})}} \\
p(D=0|\mathbf{v}_i,\mathbf{v}) &= \frac{p(D=0,\mathbf{v}_i|\mathbf{v})}{p(D=1,\mathbf{v}_i|\mathbf{v})+p(D=0,\mathbf{v}_i|\mathbf{v})} = \frac{1}{1+\frac{p(\mathbf{v}_i|\mathbf{v})}{Mp_n(\mathbf{v}_j)}}
\end{aligned}
\tag{4.8}
$$

The objective function can be written as minimising $J_{NCE}(\theta)$ as in 4.9.

$$
J_{NCE}(\theta) = -E_{p_d}[\log h(\mathbf{v}_i,\mathbf{v})] - ME_{p_n}[log(1-h(\mathbf{v}_j,\mathbf{v}))] \tag{4.9}
$$

Where $\mathbf{v}_j$ is a randomly sampled image according to noise distribution p_n and p_d is the actual data distribution. The denominator in $p(\mathbf{v}_i|\mathbf{v})$ is assumed to be a constant and is approximated using the Monte Carlo method using the samples extracted from initial batches. Hyperparameter M, as mentioned, is the number of drawn samples according to noise distribution, usually referred to as negative samples. The larger the number of these samples is, the more accurate the NCE approximation will be. However, as reported in [6], to reach a reasonably good performance, it is not essential to use all the remaining training samples, excluding the x_i. For example, in [6], it is shown that for CIFAR10 with $N = 50,000$ examples setting $M = 4,096$ has almost the same performance as setting $M = 49,999$.

The proposed method in [6] is further improved by the approach SimCLR introduced in [61]. In this method, the objective is to maximize agreement between two representations $\mathbf{v}_i$ and $\hat{\mathbf{v}}_i$ generated by different augmentation of the same image while distracting from other images using a contrastive loss formulated in equation 4.10.

$$
\mathcal{L}_{cst} = \frac{-1}{2N}\sum_{i=1}^{2N}\log\frac{e^{(s(\mathbf{v}_i,\hat{\mathbf{v}}_i)/\tau)}}{\sum_{k=1}^{2N}\mathbb{1}_{[k\neq i]}e^{(s(\mathbf{v}_i,\mathbf{v}_k)/\tau)}} \tag{4.10}
$$

where $\mathbf{v}_i = g_{\theta 2}(f_{\theta 1}(\mathbf{x}_i))$. $f_{\theta 1}$ is a ResNet structure [53] which extract the features $\mathbf{h}_i = f(\mathbf{x}_i)$. In general different structures can be used for $f(.)$, such as transformer architecture [68]. $g(.)$ is the projection head that maps the represented feature $\mathbf{h}_i$ into the space $\mathbf{v}_i = g(\mathbf{h}_i)$ on which the contrastive loss is calculated. s is a metric that shows a similarity between two representation vectors, also known as the critic or similarity measure. In [61], Cosine similarity is used to calculate $s(.,.)$. Figure 4.6 shows a general view of the SimCLR method.

The objective function in equation 4.10 differs from what has been introduced in equation 4.9 as it is calculating a categorical cross-entropy loss similar to InfoNCE loss introduced in [69] but different from it as uses a non-parametric similarity measurement. Moreover, unlike the [6] SimCLR does not use any memory bank to calculate the denominator in equation 4.10 but benefits from negative examples

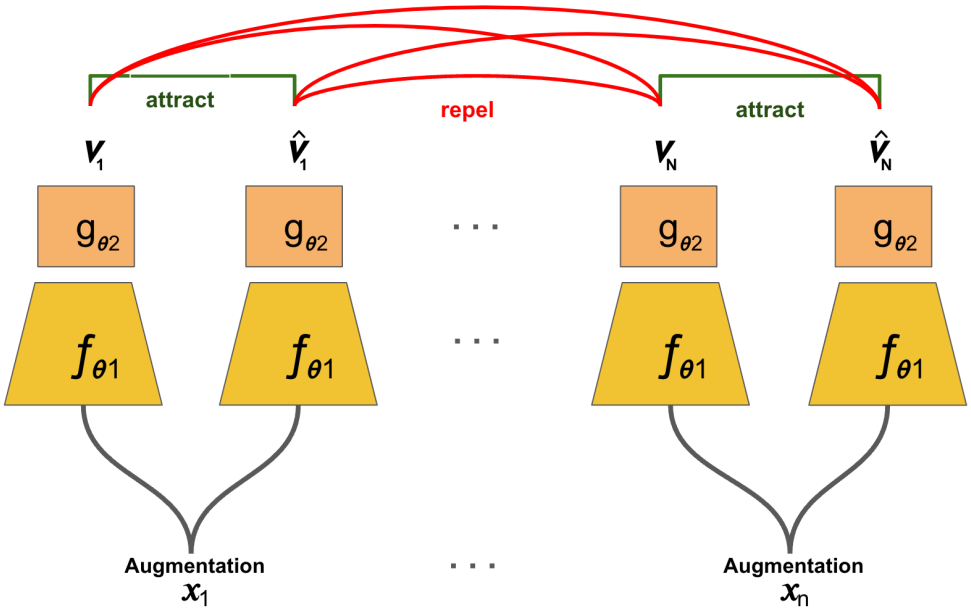


Figure 4.6: A general schematic of SimCLR model. Representation of different augmentations of the same image are attracted to each other while the representations of different images are repelled.

coming from a large batch size. Commonly, pairs of two different augmentations of the same sample are referred to as positive, and pairs of two different instances as negative.

It can be proved that minimizing the loss function in equation 4.10 is effectively maximising a lower bound on mutual information between $\mathbf{v}_i$ and $\mathbf{v}_j$ [69].

$$MI(\mathbf{v}_i, \hat{\mathbf{v}}_i) \geq \log(2N-1) - \mathcal{L}_{cst} \tag{4.11}$$

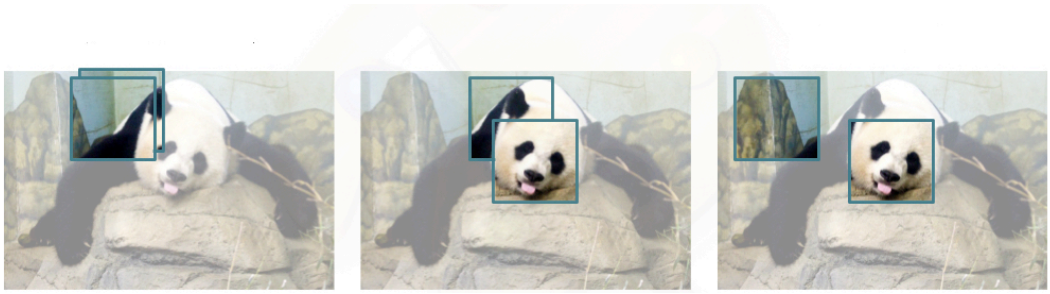


Figure 4.7: **Left**: Too much noise. **Center**: Reasonable amount of shared information. **Right**: Miss information. image is take from [10].

One important aspect of the SimCLR method is the choice of effective image augmentations. In a study conducted by [10], it is shown that the optimal augmentation is downstream task dependent. However, some general directions may

be realized to better understand the effect of data augmentation. As mentioned before, by minimising the $\mathcal{L}_{cst}$ we are indeed tightening a lower bound on $MI(\mathbf{v}_i, \hat{\mathbf{v}}_i)$. Subject to this and intuitively, it can be figured out that too weak augmentation leaves much pixel information in the augmented views and consequently into the generated representations, which provide the model with an easy task to solve, similar to the previously mentioned one mentioned shortcut learning. On the other hand, useful information such as object semantics will be lost with too strong augmentation. To further study the impact of mutual information between two views on downstream tasks, authors in [10] did an experiment utilizing patch distance. In this experiment, two different patches from the same image are extracted as two different views. The amount of shared information between patches is controlled by calculating a Euclidean distance between them, and mutual information is estimated using InfoNCE [69]. Figure 4.7 shows three different distances between patches and the amount of information they shared. After SSL pretraining with different patch distances, generated representations are fed into a linear model for the downstream classification task. As it is depicted in figure 4.8, the relation between the amount of mutual information and performance on downstream tasks looks like a reversed U shape which means that both high and low amounts of shared information between views can downgrade the model performance on the final downstream task.

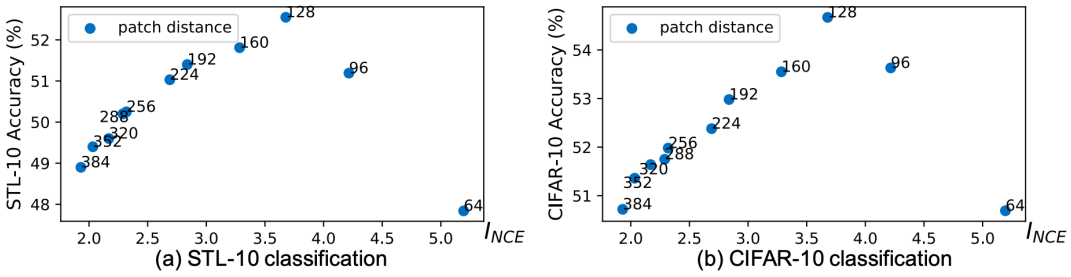


Figure 4.8: Too high and too low shared information between the generated patches from an image results in poor performance for both CIFAR10 and STL datasets. The reverse U shape shows that the right amount of shared information is necessary for the best performance. diagram is taken from [10].

4.5.2 Student Teacher Models

Contrastive methods have achieved competing results compared to supervised methods. However, they are facing some difficulties in reaching their best performance. First, they require a massive number of negative samples against which to contrast. Second, Even though, given a batch of examples, the contrastive loss functions are leveraged with automatic hard negative mining, it is not clear how to choose this batch among the entire dataset. Third, it is impossible to detect whether a sample is negative since some of the examples within the negative set may have similar semantics as the positive one.

The student-teacher model offers an alternative approach to learning high-level features. Generally, in these approaches, the model learns to predict the same target feature regardless of perturbation applied to inputs while dose not requiring negative examples. In particular, these models were inspired by a technique named knowledge distillation introduced by [70]. The main idea behind the proposed method by [70] is to distill the knowledge from a bigger model pretrained with the supervision name as the teacher to a smaller model named as the student by optimizing the student to predict the teacher target for the same input. The teacher generated targets can be of diverse types, such as classification logits and intermediate representations extracted from hidden middle layers of the teacher model.

Inspired by [70] but with some modifications, a plethora of SSL methods such BYOL [71], SWAV [72], SimSiam [73] and DINO [62] have been introduced. Unlike [70], where the teacher has a larger size than the student, these methods use the same network size for both of them. In [70], the student prediction should match the teacher's output for the same image. However, in SSL methods, the student and teacher are fed with different perturbations of an image, similar to making different views in contrastive based methods. Moreover, in [70], it is assumed that we have access to a teacher with enough good performance on prediction tasks, while in SSL methods, both student and teacher start with random initialization. In the above mentioned self-distillation based SSL method, the student model is updated directly through the gradient backpropagation. On the other hand, the teacher model is updated iteratively using an exponential moving average (EMA) method.

$$\theta_T^{t+1} = \alpha\theta_T^t + (1-\alpha)\theta_S^{t+1} \tag{4.12}$$

where θ_T and θ_S are the teacher and student parameters respectively. α controls the weight between using previous teacher or current student parameters. Subject to equation 4.12, the teacher is usually referred to as mean or momentum teacher. The momentum teacher, during the iterations, acts like a temporal ensemble of the student by averaging over its parameters.

One important issue with these SSL methods is to avoid mode collapse. The most common mode collapse among self-distillation SSL methods is that teacher and student map all the different samples to a fixed point. Each of these approaches has a different focus to avoid mode collapse. For example, BYOL [71] and SimSiam [73] use an extra prediction head for the student. Additionally, the report that stopping gradient backpropagation for the teacher model is crucial to avoid mode collapse. SimSiam is similar to BYOL, but the teacher and student have shared parameters. On the other hand, DINO [62] uses a teacher output post-processing approach to avoid mode collapse, including centering and sharpening. Centering is applied by subtracting the mean feature. Sharpening is performed by decreasing

the softmax temperature.

Chapter 5

Anomaly Detection in chest X-ray Images

5.1 Summary

5.1.1 Motivation

Medical diagnoses have been widely associated with medical imaging. Among different techniques of imaging for medical purposes, X-ray images have been widely used. These images, which are a form of electromagnetic radiation, have a less negative effect on the body due to their noninvasiveness. Bone fractures, certain tumors, pneumonia, and dental problems are examples of medical diseases that can be detected using X-ray images. In recent years, the use of machine learning approaches to process and extract knowledge from X-ray images has vastly gained attention [24, 74–79].

A significant problem in medical imaging is detecting cardiothoracic and pulmonary abnormalities that are the causes of mortality every year. A widely studied approach is to use supervised methods to discriminate normal or healthy images from abnormal ones, such as images taken from people with cardiothoracic and pulmonary abnormalities [75, 78].

Despite their promising results, supervised methods are limited to the existence of annotated data [22]. It is usually the case that for many of the abnormalities, there is not enough training data available compared to the enormous available normal samples, thus making using the supervised method challenging for such highly unbalanced datasets. To overcome the limitations of supervised methods, reconstruction based unsupervised methods are used [74, 80, 81]. The main idea behind

these methods is that a trained reconstruction based model on normal samples fails to reconstruct the abnormal samples, thus introducing a way to detect them. Nevertheless, this approach in practice can also reconstruct the abnormal samples even when they are only trained on normal samples [22, 24].

5.1.2 Training procedure

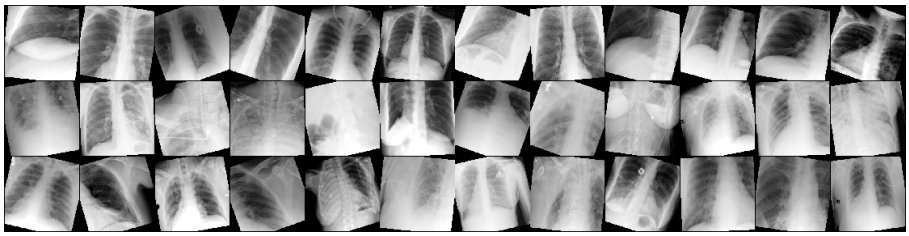


Figure 5.1: Examples of augmented images from RSNA dataset [11].

Table 5.1: Data augmentation used for fine-tuning

Transformation	PyTorch snippet
Resize	transforms.Resize(256)
Cropping	transforms.CenterCrop(224)
Horizontal Flip	transforms.RandomHorizontalFlip($p = 0.5$)
Color Jitter	transforms.ColorJitter$(0.3, 0.3, 0, 0)$
Random Affine	transforms.RandomAffine(15, translate=(0.1, 0.1), scale=(0.9, 1.1))
Normalization	transforms.Normalize()

During the training, the model has access only to normal images. The training procedure starts by creating positive pairs (x_i, x_j) by applying image augmentation on the same input image from the RSNA dataset. Figure 5.1 shows examples of augmented images sampled from the RSNA dataset, and table 5.1 shows the augmentation pipeline developed using PyTorch framework [82]. The objective is to pull together the mapping of positive pairs (z_i, z_j) in the latent space and push them away from the mapping of other samples in a randomly selected batch from the whole dataset. To do this, similar to [61], a Normalized Temperature-scaled Cross-entropy (NT-Xent) loss function is utilised as described in equation 5.1.

$$\mathcal{L}_{i,j} = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp(sim(z_i, z_k)/\tau)} \tag{5.1}$$

where $\mathbb{1}_{k \neq i}$ is an indicator function evaluating to 1 if $k \neq i$.

Adopted from [83], to generate the mapping of input images z_i, a convolution neural network encoder named as query encoder is used where the backbone is similar to the ResNet50 structure [53]. The same structure is used for another encoder that generates the mapping z_j of the positive pair plus the mapping of other images in the selected batch. This encoder is called the momentum encoder.

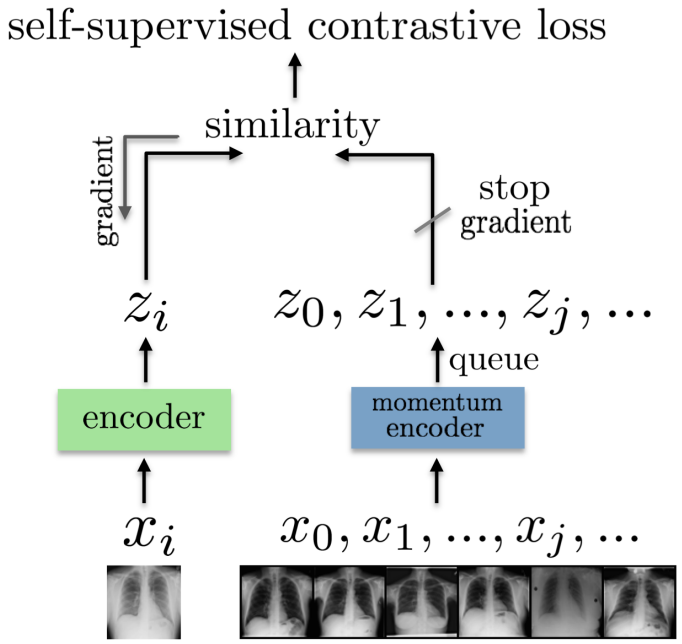


Figure 5.2: The query encoder is updated through the gradient backpropagation from the NT-Xent loss, whereas the momentum encoder is updated using the momentum-based moving average of the query encoder.

The query encoder is updated through the gradient backpropagation from the NT-Xent loss, whereas the momentum encoder is updated using the momentum-based moving average of the query encoder. Figure 5.2 shows the general schematic of the proposed method.

5.1.3 Evaluation procedure

In General, after the training steps, the detection performance for a given test sample x is evaluated by applying the Mahalanobis distance [29] on the learned features $h(x)$ of the query encoder. $h(x)$ can be accessed by dropping the last fully connected layers and normalising the output of the penultimate layer before the fully connected layer. Note that the momentum encoder is not used during the evaluation process. To calculate the Mahalanobis distance, a K-means clustering with $K = 1$ is applied to all the training data. Then, the distance is compared to cluster mean μ_m. Equation 5.2 shows Mahalanobis distance $s(x)$ for the given test sample of x.

$$s(x) := (h(x) - \mu_m)^T \Sigma_m^{-1} (h(x) - \mu_m) \tag{5.2}$$

The lower the $s(x)$, the higher the chance that the test sample x belongs to the distribution of training samples of healthy people. The performance of different methods can be compared using the area under the receiver operating characteristic

(AUROC) score.

5.1.4 Experimental results

Table 5.2: OOD detection performance (AUROC).

Methods	Opacity	No Opacity	All
Methods making use of label information			
Automated Abnormality Classification [75]	0.980	-	0.949
Pneumonia Detection using Radiomic Features[84]	0.923	-	-
ConVIRT [85]	-	-	0.908
Unsupervised methods trained on normal samples			
UAE[81]	0.89	0.78	0.83
Deep Anomaly Detection[74]	0.838	0.704	0.752
Generative Adversarial one-class classifier[24]	0.802	-	0.841
Ours	**0.940**	**0.828**	**0.866**

RSNA dataset includes samples of three different categories as follows:

- "Normal": samples from healthy people.
- "Opacity": samples of people with opacity suspicious for pneumonia.
- "No Opacity/Not Normal": samples of people that may have lung opacity but not the opacity suspicious for pneumonia.

The learned representation is evaluated on three different tasks, including the detection of "Normal" vs. "Opacity", "Normal" vs. "No Opacity/Not Normal", and "Normal" vs. all of "Opacity" and "No Opacity/Not Normal". As it is depicted in table 5.2, the proposed method performs better compared to the previous state of the art results.

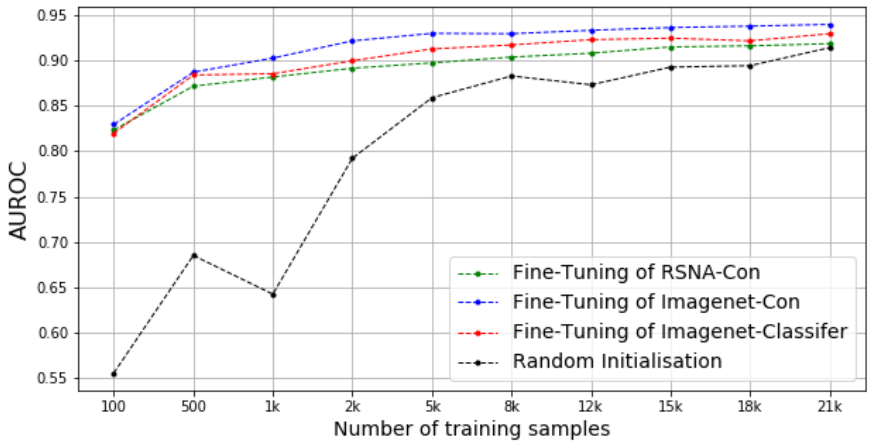


Figure 5.3: RSNA-Con and Imagenet-Con are models trained in a self-supervised manner with two datasets of RSNA and ImageNet. Imagenet-Classifier is the fine-tuning of a classifier model already trained on ImageNet in a supervised manner. Random initialisation is performing classification with random weight initialisation.

In addition to the evaluation explained above, the effect of using a self-supervised approach as a pretraining step is studied. The self-supervised pretraining can help the model to capture useful signals from a particular dataset. This is especially helpful for cases where, compared to unlabeled data, few annotated samples are available. Pretraining on unlabeled data can help the model to reach a reasonable performance by only applying fine tuning on the few existing annotated samples. In figure 5.3 result of different pretraining experiments is depicted. Different differ from the aspect of the dataset used for pretraining. As it is illustrated, pretraining achieves better performance than when the classification is performed with random initialisation of model weights.

5.2 Pneumonia Detection with Semantic Similarity Scores

R. Gholamipoor, N. Rafiee, M. Kollmann. Pneumonia Detection with Semantic Similarity Scores. *ISBI*, 2022.

Status: Published.

Contributions: The research and preparation of this manuscript were done jointly by R. Gholamipoor and N. Rafiee under the supervision of Prof. M. Kollmann.

PNEUMONIA DETECTION WITH SEMANTIC SIMILARITY SCORES

Rahil Gholamipoor[*1]*, Nima Rafiee*[*1]*, Markus Kollmann*[1,2]

Department of Computer Science[1], Department of Biology[2]
Heinrich Heine University, Düsseldorf, Germany
{rahil.gholamipoorfard, nima.rafiee, markus.kollmann}@hhu.de

X-ray images have been widely used for medical diagnoses of cardiothoracic and pulmonary abnormalities due to their noninvasiveness. Advancement in computer-aided diagnostic technologies, such as deep supervised methods, can help radiologists with a reliable early treatment and reduce diagnosis time. Nevertheless, these methods are prone to the small number of labeled samples and are limited to a specific abnormality. In this paper, we combined a self-supervised contrastive method with a Mahalanobis distance score to develop an abnormality detection method that uses only healthy images during the training procedure. We were able to outperform previous unsupervised methods for the task of Pneumonia detection. We show that representation learned by the self-supervised method improves the supervised tasks for Pneumonia detection.

1. INTRODUCTION

Chest X-ray has been used for medical screening in order for the detection of cardiothoracic and pulmonary abnormalities, which are one of the causes of mortality worldwide. Radiologists widely use chest X-ray images to diagnose lung-related diseases such as pneumonia. A computer-aided diagnostic approach would be very helpful to allow radiologists to detect potential abnormalities in chest X-ray images for early care and treatment. Recently supervised deep learning approaches have achieved promising results in abnormality detection for these images. Hendrycks et al. [1] proposed the maximum value of posterior distribution from the classifier as a baseline method to detect anomalies and Liang et al. [2] improved performance using temperature scaling and input pre-processing. However, these approaches [3] require large, annotated datasets for training which is not always feasible. Additionally, it is in general challenging to acquire enough supervised data for rare pathologies. To address these problems, many approaches have exploited unsupervised or semi-supervised frameworks to use unlabeled data for extracting generalizable features in medical images [4, 5]. Among unsupervised approaches, reconstruction-based methods assume that anomalies cannot be represented and reconstructed accurately by a model trained only on normal data. However, in practice, these models can also reconstruct abnormal samples fairly well and thus fail to detect them [5, 6]. To overcome this problem, Mao et al. [7] trained an autoencoder model to not only reconstruct the corresponding normal version of any input but also estimate the uncertainty of reconstruction at each pixel to enhance the performance of anomaly detection. In [8], an autoencoder is trained while a constraint is additionally imposed on the lower-dimensional representation of the data in which features of the same X-ray images under random data augmentations are invariant, while the features of different images are scattered.

Recently the effectiveness of self-supervised contrastive learning has been proven in different domains, e.g. the visual domain [9, 10], which enables learning of robust representations through unlabeled data. Azizi et al. [11] investigated the effect of self-supervised pre-training on the classification downstream task on the CheXpert dataset [12]. Zhang et al. [13] improved on supervised-based pneumonia detection using a contrastive-based pre-training and leveraging image description as an extra modality. In this paper, we utilize a self-supervised contrastive method to construct an anomaly detection score based on Mahalanobis distance for anomaly detection. To the best of our knowledge, we achieved state-of-the-art results for anomaly detection among all methods that can be applied to unlabeled data.

2. METHOD

2.1. Contrastive Learning

Given unlabeled training data, self-supervised contrastive representation learning aims to train a feature extractor, g_θ, to discriminate similar samples from dissimilar ones. Using image transformations that keep the semantics, each image is augmented twice, referred to as positives. The function g_θ is optimized to pull semantically similar samples together while pushing away from other images, referred to as negatives. Assuming that (x_i, x_j) is a positive pair for the i^{th} image from a batch of N images, τ is a scalar temperature parameter and $sim(u, v) = \frac{u^T v}{\|u\|\|v\|}$ denotes the dot product between $l2$ normalized u and v (i.e. cosine similarity). Contrastive learning

*Equal contribution

minimizes the following loss for a positive pair of examples (i, j), referred to as Normalized Temperature-scaled Cross-entropy (NT-Xent):

$$\mathcal{L}_{i,j} = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp(sim(z_i, z_k)/\tau)} \quad (1)$$

where $\mathbb{1}_{k \neq i}$ is an indicator function evaluating to 1 iff $k \neq i$. z_i denotes the output feature of the contrastive layer. Intuitively, this loss is the log loss of a (2N)-way softmax-based classifier that tries to classify x_j as a positive sample for x_i. One can define the contrastive feature $z(x)$ directly from the encoder g_θ, i.e., $z(x) = g_\theta(x)$ [10], or apply an additional projection layer f_ϕ, i.e., $z(x) = f_\phi(g_\theta(x))$ [9]. The contrastive loss (Eq.1) can be minimized by different mechanisms that differ in how the negative samples are maintained. Chen et al. [9] take negatives from the same batch but it requires a large batch size to provide a large set of negative pairs. Alternatively, Eq.1 can be minimized with sufficient number of negative pairs without using large batch sizes by maintaining negatives in a queue [10]. The encoded representations of the current mini-batch are enqueued while the oldest are dequeued. Unlike [9] in which only one encoder is used, following [10] we use two encoders, a query encoder and a slowly progressing key encoder, implemented as a momentum-based moving average of the query encoder.

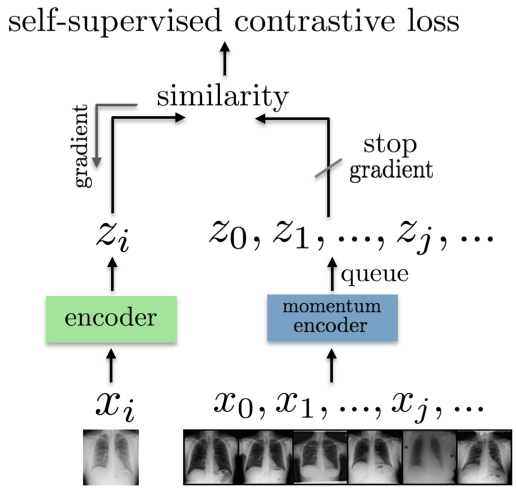


Fig. 1. The query encoder is updated end-to-end by back-propagation while the key encoder maintains a queue and is updated with momentum-based moving average. We got our best results when the model is pre-trained on ImageNet dataset.

2.2. Score Function for Anomaly Detection

Mahalanobis distance-based confidence score We use Mahalanobis distance on feature space $h(x)$ of the trained contrastive encoder as a score function for anomaly detection. Mahalanobis distance achieved promising results for supervised anomaly detection. Lee et al. [14] show that with a well-trained softmax classifier, applying Mahalanobis distance on feature space using the class means and the feature covariance matrix can reach the state of the art results on supervised anomaly detection. To measure the Mahalanobis distance for a given test sample x first, we apply K-means clustering with $K = 1$ on the feature space $h(x)$ of training data. This clustering helps to reduce computation time as we only compare the distance with the cluster mean. The anomaly score $s(x)$ for a test sample x is given by the Mahalanobis distance

$$s(x) := (h(x) - \mu_m)^T \Sigma_m^{-1} (h(x) - \mu_m) \quad (2)$$

where μ_m and Σ_m are the mean and covariance of the feature vectors from the training data. The reason to use the Mahalanobis distance is to remove the dominance of larger eigenvalues in euclidean distance metric as shown in [15] eigenvalues have an approximately inverse correlation with anomaly detection performance.

3. EXPERIMENTAL SETUP

3.1. Dataset

RSNA [1]. The Radiological Society of North America (RSNA) Pneumonia Detection Challenge dataset [16] is a publicly available dataset of frontal view chest radiographs. Each image was labeled as "Normal", "No Opacity/Not Normal" or "Opacity". The Opacity group consists of images with opacities suspicious for pneumonia, and images labeled "No Opacity/Not Normal" may have lung opacity but no opacity suspicious for pneumonia. The RSNA dataset is a subset of the National Institutes of Health (NIH) Chest X-Ray dataset [17]. It contains $26,684$ X-rays with $8,851$ normal, $11,821$ no lung opacity/not normal and $6,012$ lung opacity.

3.2. Self-supervised Contrastive Training

Experiments were carried out using ResNet50 neural network architecture. Following [9], two fully connected layers are used to map the output of ResNet to a 128-dimensional embedding space where the contrastive loss is applied. We perform training on RSNA with initialization from ImageNet self-supervised pre-trained weights. We train at batch size 128 for 100 epochs using SGD optimiser. The temperature τ in Eq.(1) is set as 0.07. At training time, we apply the following augmentations: (1) a 224×224-pixel crop is taken from a randomly resized image (2) random rotation by an angle sampled from the uniform distribution $U(-20, 20)$ (3) random horizontal flip with probability 0.5 (4) brightness and contrast adjustments.

[1] https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data

3.3. Evaluation Methodology

We evaluate the results using Area Under the Receiver Operating Characteristic curve (AUROC), which has the advantage to be scale-invariant "it measures how well predictions are ranked, rather than their absolute values" and classification-threshold-invariant "it measures how well anomaly samples are separated from the normal samples".

4. EXPERIMENTAL RESULTS

4.1. Self-Supervised Anomaly Detection

For Mahalanobis distance, the highest performance achieved from the last layer, the output after the average pooling layer, before the MLP head [15]. On RSNA dataset, to detect anomalies, we consider three different cases: "Normal" vs. "Opacity"; "Normal" vs. "No Opacity/Not Normal" and "Normal" vs. all "Opacity and No Opacity". In Table 1, we compare our method with both supervised methods and unsupervised methods trained on only healthy images. We averaged AUROC values over 5 different train/test splits.

Table 1. OOD detection performance (AUROC).

Methods	Opacity	No Opacity	All
Methods making use of label information			
Automated Abnormality Classification [18]	0.980	-	0.949
Pneumonai Detection using Radiomic Features[19]	0.923	-	-
ConVIRT [13]	-	-	0.908
Unsupervised methods trained on normal samples			
UAE[7]	0.89	0.78	0.83
Deep Anomaly Detection[20]	0.838	0.704	0.752
Generative Adversarial one-class classifier[5]	0.802	-	0.841
Ours	**0.940**	**0.828**	**0.866**

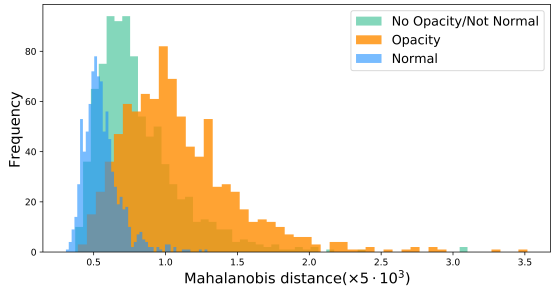


Fig. 2. Distributions over the anomaly detection score trained only on Normal samples and applied to the test sets of Normal as in-distribution, "Opacity" and "No Opacity/Not Normal" as out-distributions.

4.2. Pre-training and Label Efficiency of Multilabel Classification

In addition to the self-supervised anomaly detection task, we evaluate the learned representation by its performance in KNN accuracy and the impact it has on multilable classification downstream task. For the pre-training task, we use the same data split statistics as in [18] including $21,152$ training samples ($14,159$ abnormal and $6,993$ normal samples). We use the same optimization config as for the anomaly detection task. The self-supervised pre-trained model achieved 1-NN accuracy of 79.01%. For the classification task, we replace the projection head of the contrastive encoder with a classification head, projecting the data into a one-dimensional scalar value and fine-tune the whole model with binary cross-entropy loss and same optimization config as in [18]. To see the effect of self-supervised pre-training, we start with a small fraction of training data and compare model AUROC performance on test data for different case studies. Figure 3 shows that self-supervised pre-training can significantly help with label efficiency and causes a considerable performance improvement when we have a small fraction of labeled samples for the downstream task. We achieve an AUROC score of 94.4% when fine-tuning with all labeled training data and an AUROC score of 82.97% when using only 100 labeled samples which are selected randomly from training data.

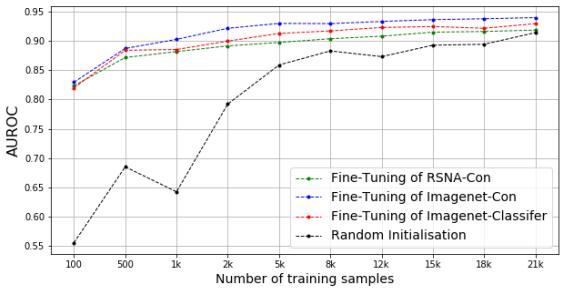


Fig. 3. Self-supervised pre-training increases the downstream classification task performance with small fraction of training samples. RSNA-Con and Imagenet-Con are fine-tunings of models with different model initialisation in self-supervised pre-training as follows: randomly initialised and initialised with Imagenet. Imagenet-Classifier stands for fine-tuning an already trained imagenet classifier and Random Initialisastion is performing classification with random weight initialisation.

5. ABLATION STUDIES

5.1. Data Augmentation Details

In our setting, to train the self-supervised contrastive encoder, we utilize random crop (resize to 224×224), random rotation (image rotation by angle θ from range $(-20, 20)$), random horizontal flip, brightness and contrast adjustments as the data augmentations. Brightness and contrast adjustments are composed by color jittering. The details of these augmentations are provided in Table 2.

Table 2. Data augmentation used for contrastive training

Transformation	PyTorch snippet
Cropping	transforms.RandomResizedCrop(224, scale=$(0.08, 1.0)$)
Rotation	transforms.RandomRotation(20)
Horizontal Flip	transforms.RandomHorizontalFlip($p = 0.5$)
Color Jitter	transforms.ColorJitter$(0.4, 0.4, 0, 0)$
Normalization	transforms.Normalize()

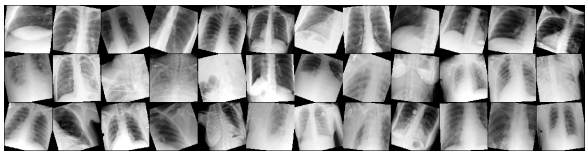


Fig. 4. Examples of augmented images from RSNA dataset

5.2. Ablation on Batch Size

Training with small batches. Table 3 confirms that large batches are not necessary for a good performance in our anomaly detection problem. We scale the learning rate linearly with the batch size [21]. We averaged AUROC values for "Normal" vs. "Opacity" over 5 different train/test splits for each batch size.

Table 3. Effect of batch size

Batch size	AUROC
256	0.926
128	0.940
64	0.916

5.3. Ablation on Data Augmentation

Because of the less diverse nature of X-ray images, in our experiments, we used strong data augmentations in order to prevent over-fitting and improve anomaly detection performance. In Table 4, we change the strength of each augmentation individually while keeping the rest unchanged. We averaged AUROC values for "Normal" vs. "Opacity" over 5 different train/test splits.

Table 4. Effect of data augmentation

Transformation	AUROC
transforms.RandomResizedCrop(224, scale=($\mathbf{0.4}$, 1.0))	0.934
transforms.RandomRotation(**10**)	0.928
transforms.ColorJitter($\mathbf{0.25}$, $\mathbf{0.25}$, 0, 0)	0.926

5.4. Fine-tuning Implementation Details

To do the fine-tuning, we use the same data augmentation as used in [18]. Table 5 shows the augmentation details together with related PyTorch code. We use a batch size of 128 for all experiments where the training samples are more than 1000 images and 64 where we have 100 and 500 training samples. Other optimisation hyper-parameters are the same for all experiments. Table 6 summarises the optimisation hyper-parameters used for fine-tuning.

Table 5. Data augmentation used for fine-tuning

Transformation	PyTorch snippet
Resize	transforms.Resize(256)
Cropping	transforms.CenterCrop(224)
Horizontal Flip	transforms.RandomHorizontalFlip($p = 0.5$)
Color Jitter	transforms.ColorJitter$(0.3, 0.3, 0, 0)$
Random Affine	transforms.RandomAffine(15, translate=(0.1, 0.1), scale=(0.9, 1.1))
Normalization	transforms.Normalize()

Table 6. Hyper-parameters for fine-tuning training

Hyper-parameter	Default value
Number of epochs	50
Learning rate	10^{-2}
Weight decay	10^{-4}
Optimizer	SGD
Momentum of SGD	0.9

6. CONCLUSION

In this work, we proposed a self-supervised contrastive learning framework for X-ray anomaly detection trained only with the normal images to make our method future-ready for yet unknown anomalies. The self-supervised representations are highly effective for the task of anomaly detection in our framework. We define an anomaly detection score based on Mahalanobis distance applicable for detecting anomalies. We find that our approach outperforms all previous unsupervised methods on the RSNA pneumonia detection challenge dataset. This work may allow for improving radiology workflow and clinical decision-making.

7. REFERENCES

[1] Dan Hendrycks and Kevin Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," 2018.

[2] Shiyu Liang, Yixuan Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," 2020.

[3] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," 2018.

[4] Diana Davletshina, Valentyn Melnychuk, Viet Tran, Hitansh Singla, Max Berrendorf, Evgeniy Faerman, Michael Fromm, and Matthias Schubert, "Unsupervised anomaly detection for x-ray images," 2020.

[5] Yuxing Tang, Youbao Tang, Mei Han, Jing Xiao, and Ronald M. Summers, "Abnormal chest x-ray identification with generative adversarial one-class classifier," 2019.

[6] Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G van Leeuwen, and Keelin Murphy, "Deep learning for chest x-ray analysis: A survey," *Medical Image Analysis*, p. 102125, 2021.

[7] Yifan Mao, Feifei Xue, Ruixuan Wang, Jianguo Zhang, Wei-Shi Zheng, and Hongmei Liu, "Abnormality detection in chest x-ray images using uncertainty prediction autoencoders," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part VI*. 2020, vol. 12266 of *Lecture Notes in Computer Science*, pp. 529–538, Springer.

[8] Behzad Bozorgtabar, Dwarikanath Mahapatra, Guillaume Vray, and Jean-Philippe Thiran, "SALAD: self-supervised aggregation learning for anomaly detection on x-rays," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part I*. 2020, vol. 12261 of *Lecture Notes in Computer Science*, pp. 468–478, Springer.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," 2020.

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, "Momentum contrast for unsupervised visual representation learning," 2020.

[11] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi, "Big self-supervised models advance medical image classification," 2021.

[12] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," 2019.

[13] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," 2020.

[14] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," 2018.

[15] Vikash Sehwag, Mung Chiang, and Prateek Mittal, "SSD: A unified framework for self-supervised outlier detection," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. 2021, OpenReview.net.

[16] George Shih, C. C. Wu, S. Halabi, M. Kohli, L. Prevedello, T. Cook, Arjun Sharma, J. Amorosa, V. Arteaga, M. Galperin-Aizenberg, R. Gill, M. Godoy, Stephen Hobbs, J. Jeudy, A. Laroia, P. Shah, D. Vummidi, K. Yaddanapudi, and Anouk Stein, "Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia.," *Radiology. Artificial intelligence*, vol. 1 1, pp. e180041, 2019.

[17] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3462–3471.

[18] Yu-Xing Tang, You-Bao Tang, Yifan Peng, Ke Yan, Mohammadhadi Bagheri, Bernadette A Redd, Catherine J Brandon, Zhiyong Lu, Mei Han, Jing Xiao, and Ronald M Summers, "Automated abnormality classification of chest radiographs using deep convolutional neural networks," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–8, 2020.

[19] Yan Han, Chongyan Chen, Ahmed Tewfik, Ying Ding, and Yifan Peng, "Pneumonia detection on chest x-ray using radiomic features and contrastive learning," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 247–251.

[20] Takahiro Nakao, Shouhei Hanaoka, Yukihiro Nomura, Masaki Murata, Tomomi Takenaga, Soichiro Miki, Takeyuki Watadani, Takeharu Yoshikawa, Naoto Hayashi, and Osamu Abe, "Unsupervised deep anomaly detection in chest radiographs," *Journal of Digital Imaging*, pp. 1–10, 2021.

[21] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," 2018.

5.3 Conclusion

In this study, a self-supervised contrastive based method is proposed. This method overcomes the limitation of supervised methods as it is not limited to performing a discrimination task for a particular abnormality disease. On the other hand, it learns the concept and semantics of normal samples; thus, any other sample from a certain disease can be recognised as an abnormality. Additionally, it does not sever from the problem of reconstruction based unsupervised methods as the samples are mapped into a lower-dimensional latent space.

Having the learned features from the trained model, Mahalanobis distance is used as an anomaly score for a given test sample. As depicted in the experimental result section, the proposed method achieves better performance compared to the previous state of the art results for the task of pneumonia detection by accessing only normal samples. Additionally, the efficacy of self-supervised learning as a pretraining procedure is studied when the model is pretrained using different datasets.

Chapter 6

Self-Supervised Anomaly Detection by Self-Distillation and Negative Sampling

6.1 Summary

6.1.1 Motivation

As it is explained in section 2.1, anomaly detection or out of distribution (OOD) detection refers to a type of problem where the goal is to detect whether a given sample is derived from a distribution same as the training data or from a different one.

Many supervised approaches have been introduced to tackle this problem, and promising results [28, 86]. However, the need for annotated data has always been a limitation of the supervised approach. This is especially the case for real-world applications of OOD detection, such as abnormality detection in medical images or where the type of OOD data can be changed over time. Additionally, as it is explained in 2.1, representation learned using supervised methods can capture superficial features compared to self-supervised methods.

Motivated by this, a plethora of unsupervised density estimation methods such as PixelCNNs [21] and reconstruction based methods such as VAEs [19] have been used in different areas such as abnormality detection in medical imaging [22]. The idea behind using the density estimator methods is that in-distribution data should get a higher likelihood score compared to OOD samples. Similarly, the reason behind using reconstruction based method is that these methods should fail to

reconstruct the OOD samples. However, despite their achievements, in [26], it is shown that sometimes they can fail when performing in the regime of high-dimensional and complex data such as natural images in a way that they can assign a higher likelihood to OOD samples or can reconstruct the anomaly samples fairly well [22, 24].

To address the above mentioned problems, recent studies investigate representation learned using self-supervised methods for the task of anomaly detection [46, 47, 87]. One advantage of SSL methods compared to supervised ones is that no OOD samples are required in the training phase. These methods learn to gain some level of understanding of normal samples and to draw a decision boundary to discriminate normal samples from anomalous ones. Moreover, it is shown that the used self-supervised methods are not prone to the problem of unsupervised density estimators and reconstruction based methods [46, 47].

In [88], a representation is learned using the prediction of the degree of rotation applied on input training images. The representation is used to detect anomaly samples using cosine similarity. The learned representation is further improved by using a contrastive based method in [47]. The authors in [46] use the contrastive method plus negative samples that are generated by applying rotation on input data. An extra head is also used to predict the rotation degree applied to training images. In our study, we use a self-supervised representation learning approach based on self-distillation and the use of negative samples. Self-distillation approach helps to avoid the limitations of contrastive methods explained in section 4.5.2. Despite the promising result of using negative samples in [46], the effect of different negative samples made by applying different shifting transformations on different original images has not been thoroughly studied. In this study, different experiments have been conducted to compare the distinct effects of different negative samples. Moreover, a sensitivity score is introduced based on which we can gain some intuition regarding the model performance and the effect of hyperparameters without accessing the OOD samples.

6.1.2 General overview of training procedure

Figure 6.1 depicts the general workflow of the proposed method. The network structure is generally adapted from DINO [62]. Similar to DINO, our framework creates positive pairs by applying augmentations extracted from set $\sim \mathcal{T}$. One of the samples in the positive pair is passed to the student network and the other one to the teacher. For the teacher network, a centering is applied to the output to avoid the mode collapse and then a softmax with temperature τ_t to sharpen the probability scores. A softmax with temperature τ_s is also applied to the student output. These two outputs are then passed to a cross-entropy loss where the objective is to assign the same class to the positive pair.

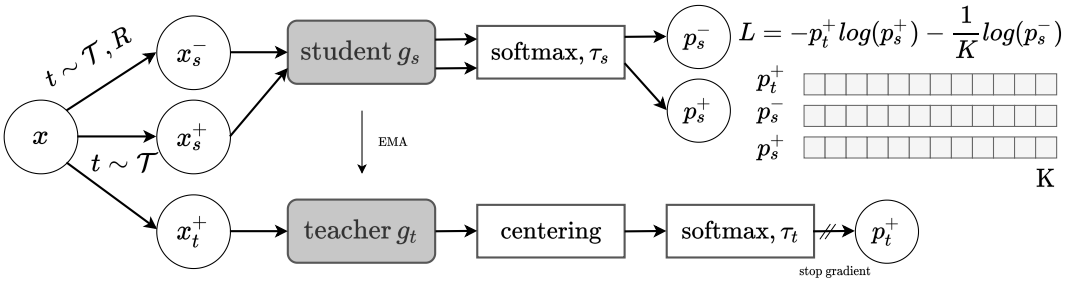


Figure 6.1: An overview of the proposed contrastive self-distillation framework, consisting of student and teacher networks, g_s and g_t, that map two random transformations of the same image, $x_s^+ \sim \mathcal{T}(x)$ and $x_t^+ \sim \mathcal{T}(x)$ to the same class. Negative views, x^-, arise from first applying a shifting transformation R, such as random rotation, followed by $\mathcal{T}$ to either an in-distribution image x or an auxiliary image x_{aux}.

In the proposed method, an extra set $\mathcal{R}$ is introduced from which shifting transformation can be extracted. To create the negative sample, extracted shifting transformation is applied together with augmentations from set $\mathcal{T}$. The negative sample is then passed to the student network, and the output is fed into a cross entropy loss where the objective is to encourage a uniform spread of negative samples over all the K existing classes.

6.1.3 Negative samples

To create negative samples, different shifting transformations can be applied to training samples of in-distribution data, samples of an auxiliary dataset, or a combination of both. For the auxiliary dataset, we used ImageNet for most of the experiments. Additionally, samples of debiased tiny images (DTI) [89] are also used to compare the results with ImageNet as the auxiliary dataset. For DTI, samples whose labels are the same as samples of CIFAR10 and CIFAR100 datasets tried to be excluded. For shifting transformations, the following augmentations are used:

- rotating by $r \sim R = \mathcal{U}(\{90°, 180°, 270°\})$, where $\mathcal{U}$ is the uniform distribution.
- random permutation of each part of the evenly partitioned image in N patches (Perm-N) with $N = 4$ and $N = 16$
- random permutation of all pixels of an image referred to as Pix.Perm.

Figure 6.2 shows a schematic of creating negative samples.

On the model side, the vision transformer ViT-small [90] is used as the backbone of the network. Vision transformers can capture the long-range correlation between input features and show better performance compared to convolutional

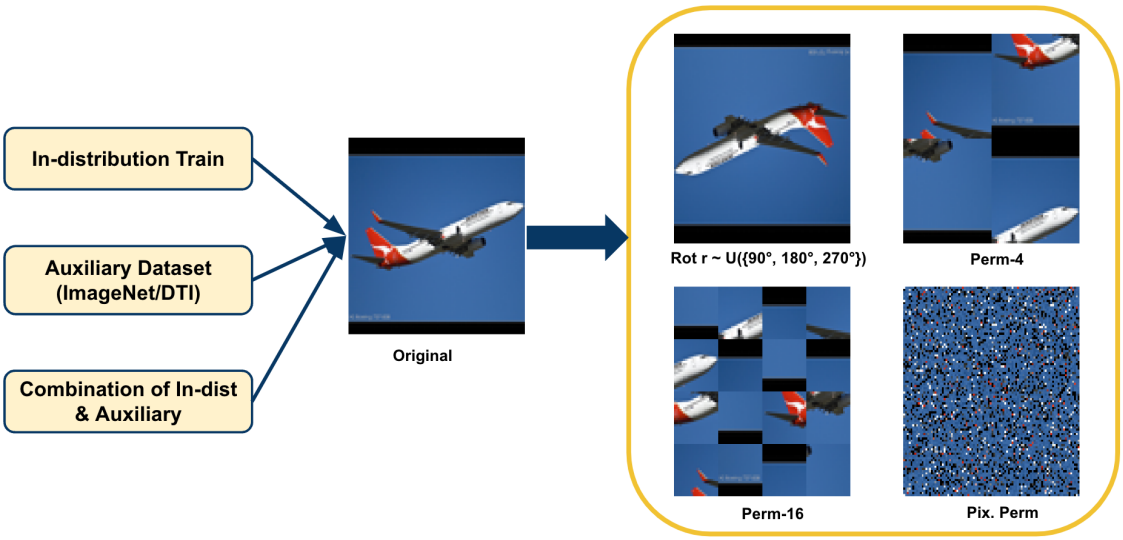


Figure 6.2: Negative samples are created by applying shifting transformations on images from the in-distribution train, auxiliary dataset (ImageNet/DTI), or a combination of both. Shifting transformation set includes **Rot**: rotating by $r \sim R = \mathcal{U}(\{90°, 180°, 270°\})$, where $\mathcal{U}$ is the uniform distribution. **Perm**4 **and** **Perm**16: Patch permutation where each image is divided into 4 and 16 patch divisions. **Pix.Perm**: pixel permutation of an input image.

neural networks [91]. Two fully connected layers then map the backbone output to a K dimensional space.

6.1.4 Evaluation procedure

To evaluate the performance of the model, fully connected layers are dropped, and the backbone output of the teacher network is used. For a given test sample x_{test} and training data sample x_m the backbone network outputs f_{test} and f_m, respectively. An average of temperatured cosine similarity between f_{test} and all the training samples are considered as anomaly score. Equation 6.1 shows the anomaly score for a given test sample x.

$$\mathcal{S}(x) = -\frac{1}{M}\sum_{m=1}^{M} \exp\left(\frac{1}{\tau}\cdot\frac{f_{test}^T f_m}{\|f_{test}\|\|f_m\|}\right), \tag{6.1}$$

where M is the number of the training samples.

6.1.5 Experimental results

The proposed method is compared to previous self-supervised methods with state of the art results on the task of anomaly detection. 6.1 shows the comparison results. As it is depicted in the table, for the CIFAR-10 dataset as in-distribution,

Table 6.1: The Area Under the Receiver Operating Characteristic Curve (AUROC) scores for OOD detection without label supervision. Last two columns shows the results for the proposed method. Rot.ImgN and Combined stand for rotation on ImageNet and rotation on combination of ImageNet and in-distribution data respectively.

						Ours	
$\mathcal{D}_{train}^{in}$	$\mathcal{D}_{test}^{out}$	Geometric*[88]	SSD[47]	CSI[46]	MTL†[92]	Rot.ImgN	Combined
CIFAR10	CIFAR100	91.91	90.63	89.20	93.24	92.51	**94.20**
	SVHN	97.96	99.62	99.80	**99.92**	99.69	**99.92**
	ImageNet30	–	90.20	87.92	–	**94.16**	93.40
	TinyImageNet	92.06	92.25	92.44	92.99	**96.28**	95.02
	LSUN	93.57	96.51	91.60	95.03	**98.08**	97.52
	STL10	–	70.28	64.25	–	**77.29**	74.34
	Places365	92.57	95.21	90.18	93.72	**97.14**	96.01
	Texture	96.25	97.61	98.96	–	**99.16**	98.69
CIFAR100	CIFAR10	74.73	69.60	58.87	**79.25**	69.96	67.63
	SVHN	83.62	94.90	96.44	87.11	96.00	**97.17**
	ImageNet30	–	75.53	71.82	–	**84.82**	75.36
	TinyImagenet	77.56	79.52	79.28	80.66	**81.41**	79.75
	LSUN	71.86	79.50	61.83	74.32	**85.03**	74.55
	STL10	–	72.76	64.26	–	**79.96**	71.70
	Places365	74.57	79.60	65.48	77.87	**81.67**	72.79
	Texture	82.39	82.90	**87.47**	–	80.65	77.33

* Requires labels for the supervised training loss. Results reported from [92].
† Requires labels to select the optimal transformations.

the proposed method achieves Superior performance compared to other methods for two types of negative samples of Rot.ImgN and Combined, which stand for rotation on ImageNet and rotation on the combination of ImageNet and in-distribution data, respectively.

Extensive experiments have been conducted to better understand the result of different negative samples. Table 6.2 shows the result of these experiments. As it is illustrated, applying rotation results in the best performance. The result can be interpreted as evidence that an effective shifting transformation should change the high-level semantics but keep the low-level statistics intact. Intuitively, the model should capture high-level semantics to be able to distinguish between such negative examples from their in-distribution counterparts. Consequently, the model that captures higher level semantics can perform better for the near-OOD problem explained in Chapter 2.

One challenging task in the OOD detection problem is to evaluate the performance of the model without accessing the OOD samples. This is important because OOD samples are not always available during the training. Moreover, these samples might be drawn from different distributions. To get some intuition regarding how the model performs, an analysis is conducted to examine the relation between the model performance on OOD sets and the model performance for 10-NN accuracy

Table 6.2: AUROC scores for OOD Detection with CIFAR10 as $\mathcal{D}^{in}_{train}$ and different $\mathcal{D}_{neg}$. ImgN denotes ImageNet samples.

Negative Sampling:	None	Auxiliary								In-Dist
$\mathcal{D}^{in}_{train}$	DINO $\lambda = 0$	ImgN	Rot. ImgN	Rot. 360 ImgN	DTI	Perm-16 ImgN	Perm-4 ImgN	Rot. DTI	Pix. Perm.	Rot. In-Dist.
CIFAR100	90.29	90.46	92.51	88.62	93.77	88.32	89.57	93.77	87.67	**93.96**
SVHN	99.38	99.50	99.69	99.42	99.86	99.59	99.13	99.86	99.62	**99.92**
ImageNet30	88.81	89.96	94.16	88.95	93.39	89.17	84.71	**96.04**	87.46	91.69
TinyImageNet	91.07	94.14	**96.28**	91.60	94.53	89.72	91.27	95.64	89.39	94.27
LSUN	92.20	93.41	98.08	93.24	98.56	94.58	89.32	**99.12**	93.33	94.93
STL10	66.50	77.65	77.29	72.41	72.02	69.22	68.81	**81.49**	68.55	69.11
Places365	91.28	93.12	97.14	92.58	97.03	92.77	87.63	**98.12**	91.89	93.53
Texture	96.21	95.01	**99.16**	93.93	97.55	93.38	89.86	95.11	93.08	98.29
Average	89.47	91.66	94.29	90.09	93.34	89.59	87.54	**94.89**	88.87	91.96

on in-distribution test data. Moreover, a sensitivity score is introduced. The sensitivity score is calculated as the AUROC score of part of the training data excluded as the new in-distribution test and the original in-distribution test. In other words, the sensitivity score shows how strongly the model rejects the in-distribution test as anomaly samples. The result shows that for the range close to AUROC=50%, the higher the sensitivity score is, the better the model performs on OOD samples. Figure 6.3 top diagrams show the correlation between 10-NN and AUROC scores for CIFAR-10 as in-distribution and CIFAR-100 and Texture [93] as OOD datasets. Figure 6.3 bottom diagrams show the correlation between the sensitivity score and AUROC score for the same dataset setup as the top diagrams.

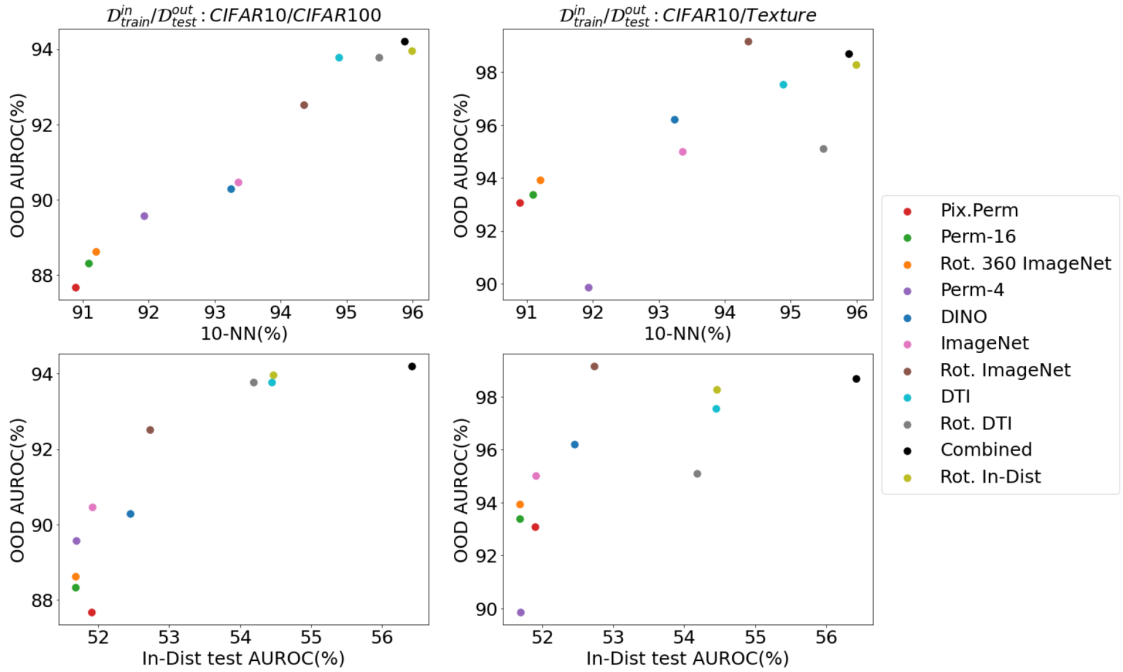


Figure 6.3: Different models trained on CIFAR10 for two OOD datasets, CIFAR100 (left column) and Texture (right column). Points in each plot indicate different negative sampling strategies (colors are shared). **Top row**: correlation between OOD detection AUROC and 10-NN accuracy on in-distribution test. **Bottom row**: correlation between OOD detection AUROC and sensitivity score. Models with higher sensitivity close to a range of 50% have higher OOD detection performance.

6.1.6 Conclusion

In this study, a new self-supervised framework based on self-distillation and negative sampling for the task of OOD detection is introduced. The study has three main areas of focus. First, modifying the self-distillation method introduced in [62] to account for both positive and negative examples by adding an auxiliary term to the original objective. The main idea is to pull the positive pairs with the same high-level semantics close to each other while pushing the negative sample with different high-level semantics but the same low-level statistics away. Second, experimenting with the effect of different negative samples when different shifting transformations are applied to in-distribution training data, images of an auxiliary dataset, and a combination of both. Finally, study the correlation of AUROC performance on OOD detection by the performance of the model considering the metrics calculated on in-distribution test data such as 10-NN and moreover, introduce sensitivity score and study the correlation between this score and AUROC performance on OOD datasets.

6.2 Self-Supervised Anomaly Detection by Self-Distillation and Negative Sampling

Nima Rafiee, Rahil Gholamipoorfard, Nikolas Adaloglou, Simon Jaxy, Julius Ramakers, Markus Kollmann, 2022.

Status: Accepted for *ICANN*.

Contributions: The author contributed with training the models, evaluations, visualization, and writing under the supervision of Prof. Markus Kollmann.

Self-Supervised Anomaly Detection by Self-Distillation and Negative Sampling

Nima Rafiee[1][0000−0002−3193−9534], Rahil Gholamipoor[1][0000−0001−8207−7295], Nikolas Adaloglou[1][0000−0003−4938−6322], Simon Jaxy[1][0000−0002−7076−4108], Julius Ramakers[1][0000−0002−2925−152X], and Markus Kollmann[1,2][0000−0002−5317−5408]

[1] Department of Computer Science, Heinrich Heine University, Düsseldorf, Germany
[2] Department of Biology, Heinrich Heine University, Düsseldorf, Germany
{rafiee,rahil.gholamipoorfard,nikolaos.adaloglou,simon.jaxy,ramakers,kollmann}@hhu.de

Abstract. Detecting whether examples belong to a given in-distribution or are out-of-distribution (OOD) requires identifying features that are specific to the in-distribution. In the absence of labels, these features can be learned by self-supervised representation learning techniques under the generic assumption that the most abstract features are those which are statistically most over-represented in comparison to other distributions from the same domain. This work shows that self-distillation of the in-distribution training set together with contrasting against negative examples derived from shifting transformation of auxiliary data strongly improves OOD detection. We find that this improvement depends on how the negative samples are generated, with the general observation that negative samples that keep the statistics of lower level features but change the global semantics result in higher detection accuracy on average. For the first time, we introduce a sensitivity score using which we can optimise negative sampling in a systematic way in an unsupervised setting. We demonstrate the efficiency of our approach across a diverse range of OOD detection problems, setting new benchmarks for unsupervised OOD detection in the visual domain.

Keywords: Anomaly Detection · Self-Supervised Learning · Self-Distillation · Negative Sampling.

1 Introduction

OOD detection or anomaly detection is the problem of deciding whether a given test sample is drawn from the same in-distribution as a given training set or belongs to an alternative distribution. Many real-world applications require highly accurate OOD detection for secure deployment, such as in medical diagnosis. Despite the advances in deep learning, neural network estimators can generate systematic errors for test examples that are far from the training set [25]. For example, it has been shown that Deep Neural Networks (DNNs) with ReLU activation functions can make false predictions for OOD samples with arbitrarily high confidence [12].

A major challenge in OOD detection is the case where the features of outlier examples are statistically close to the features of in-distribution examples, which is frequently the case for natural images. In particular, it has been shown that deep density estimators like Variational Autoencoders (VAEs) [16], PixelCNNs [33], and

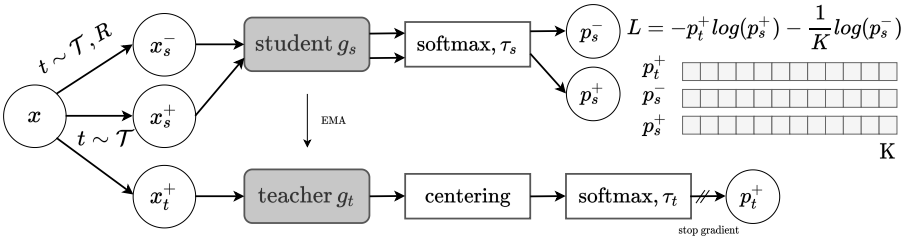


Fig. 1. An overview of the proposed contrastive self-distillation framework, consisting of student and teacher networks, g_s and g_t, that map two random transformations of the same image, $x_s^+ \sim \mathcal{T}(x)$ and $x_t^+ \sim \mathcal{T}(x)$ to the same class. Negative views, x^-, arise from first applying a shifting transformation R, such as random rotation, followed by $\mathcal{T}$ to either an in-distribution image x or an auxiliary image x_{aux}.

normalising flow models [28] can on average assign higher likelihood to OOD examples than to examples from the in-distribution [22]. This surprising finding can be partially attributed to an inductive bias from upweighting local pixel correlations as a consequence of using convolutional neural networks.

A challenging scenario of anomaly detection is near OOD detection [35], where the OOD distribution samples are statistically very similar to the in-distribution. A particular challenging OOD detection task is given by CIFAR10 [18] as in-distribution and CIFAR100 [18] as OOD, where the larger number of classes in CIFAR100 makes it harder to identify features that are specific to the in-distribution. For the cases where the in and out distributions are not closely related, we refer to as far OOD.

State-Of-The-Art (SOTA) performance has been obtained for the CIFAR10/CIFAR 100 near OOD detection task, using pretrained classification models on ImageNet-21K [8]. However, as CIFAR100 and CIFAR10 share many of their classes with ImageNet but the classes among themselves are mutually exclusive, the pretrained model effectively solves the OOD detection problem for this special case. The advantage of using pretrained models as OOD detectors drops if there is no class overlap with the OOD test set, such as for SVHN [8]. To overcome these limitations, a plethora of self-supervised pretext tasks have been proposed that provide a richer learning signal that enables abstract feature learning [4, 2, 11]. These advancements in self-supervised learning have shown remarkable results on unsupervised anomaly detection [31, 30, 35] by solely relying on the in-distribution data.

More recently, it has been suggested to include dataset-specific augmentations that shift the in-distribution – so-called negative samples. The core idea behind using shifting transformations is to concentrate the learned representation in feature space. This can result in a more conservative decision boundary for the in-distribution [14]. However, in-distribution shifting requires dataset-specific prior knowledge [21]. Therefore, a bad choice of augmentations may result in rejecting the in-distribution test samples, which reduces the OOD detection performance.

In this paper, we propose an improved version of the DINO [3] framework together with a sensitivity score for the problem of OOD detection. The main contributions of this work are summarized as follows:

- We propose a self-supervised self-distillation method that leverages unlabelled data for OOD detection which aims at drawing a tight, not necessary simply connected, decision boundary between the in-distribution data and an auxiliary negative distribution.
- We introduce an auxiliary loss that encourages unlabeled negative samples to be uniformly assigned to the existing in-distribution soft-classes.
- To the best of our knowledge, for the first time we introduce a sensitivity score defined by the AUROC value between the in-distribution training set and the in-distribution test set. Using sensitivity score, we can intuitively compare the effect of different negative auxiliary sets and to find optimal values by grid search for training hyperparameters without the access to OOD validation set.
- Finally, we show that the proposed framework does not only improve OOD detection performance but also improves representation learning for the in-distribution, as measured by the K-Nearest Neighbour (K-NN) accuracy.

2 Related Works

Supervised OOD detection methods. In-distribution classification accuracy is highly correlated with OOD performance [8]. This has motivated supervised OOD detection approaches to learn representations from classification networks [13, 19]. This can be achieved by directly training a classifier on the in-distribution or by pre-training on a larger dataset .

Fine-tuning pretrained transformers [34] has shown promising OOD scores. In computer vision, Koner et al. [17] leveraged the contextualization capabilities of pretrained Vision Transformer (ViT) [7] by exploiting the global image context. Such models heavily rely on the classes of the pretrained dataset, which often include classes from both the in and out distribution. Although, supervised pretraining can form a good boundary for OOD detection, it has two limitations, firstly the pretraining dataset should share labels with both in and out distributions, and secondly impeded OOD performance is observed when the distributions have overlapping classes.

Mohseni et al. [21] recently presented a 2-step method that initially learns how to weight the in-distribution transformations based on a supervised objective. Then, the selected shifting transformations are applied in a self-supervised setup for OOD detection. Human-level supervision is still required to learn the best shifting transformations for each training dataset. In *Geometric* [15], Hendrycks et al. defined a self-supervised task to predict geometric transformations to improve the robustness and uncertainty of deep learning models. They further improve their self-supervised technique with supervision through outlier exposure.

Unsupervised OOD detection methods. Existing label-free OOD detection approaches can be separated in density-based [27, 23], reconstruction-based [26, 38], and self-supervised learning [9, 15] methods. Density-based methods aim to fit a probability distribution (e.g. Gaussian) to the training data and then use it for OOD detection. Reconstruction-based methods assume the network would generalize less for unseen

OOD samples. Meanwhile, recent studies [22] revealed that probabilistic generative models can fail to distinguish between training data and OOD inputs.

Self-supervised methods have recently shown that adopting pretext tasks results in learning general data representations [1] for OOD detection. Choi et al. [5] used blurred data as adversarial examples to discriminate the training data from their blurred versions. In *CSI* [31], Tack et al. leverage shifting data transformations in contrastive learning for OOD detection, combined with an auxiliary task that predicts which shifting transformation was applied to a given input. In *SSD* [30], the authors further improved contrastive self-supervised training by developing a cluster-conditioned OOD detection method in the feature space.

Outlier Exposure (OE). OE leverages auxiliary data that are utterly disjoint from the ODD data [14]. Furthermore, OE assumes that the provided auxiliary samples are always OOD. To guarantee this, one needs human supervision to remove the overlap between auxiliary and in-distribution. OE has been successfully applied to training classifiers, by enforcing the auxiliary samples to be equally distributed among the in-distribution classes. In contrast to [14], we attempt to teach the network better representations for OOD detection by incorporating auxiliary data into a self-distillation soft-labeling framework.

Finally, since the proposed method does not require labels, there is no information whether the in-distribution data are meaningfully similar to the auxiliary ones. In this aspect, this work is different from OE, as it only requires the in-distribution to be sufficiently statistically underrepresented. To ensure the latter an additional transformation is applied on the auxiliary data (e.g. rotation).

3 Proposed Method

3.1 The vanilla DINO framework

The DINO framework uses two identical networks $g_s = g(x|\theta_s)$ and $g_t = g(x|\theta_t)$ called student and teacher, which differ by their sets of parameters θ_s and θ_t, respectively. For each transformed input image x, both networks produce K-dimensional output vectors, where K is the number of soft-classes. Both outputs enter a temperature-scaled softmax functions $p_t = softmax(g_t, \tau_t)$ and $p_s = softmax(g_s, \tau_s)$ defined by:

$$p^i(x) = \frac{\exp\left(g^i(x)/\tau\right)}{\sum_{k=1}^{K} \exp\left(g^k(x)/\tau\right)}, \tag{1}$$

where $p^i(x)$ is the probability of x falling in soft-class i and τ_s, τ_t are the student and teacher temperatures. In contrast to knowledge distillation methods, the teacher is built from previous training iterations of the student network. To do so, the gradients are back-propagated only through the student network and the teacher parameters are updated with the Exponential Moving Average (EMA) of the student parameters

$$\theta_t \leftarrow m\theta_t + (1-m)\theta_s, \tag{2}$$

where $0 \leq m \leq 1$ is a momentum parameter. For $\tau_t < \tau_s$, the training objective is given by the cross entropy loss for two non-identical transformations x'', x' of an image x drawn from the in-distribution training set $\mathcal{D}^{in}_{train}$

Table 1. AUROC scores for OOD detection without label supervision.

$\mathcal{D}^{in}_{train}$	$\mathcal{D}^{out}_{test}$	Geometric*[15]	SSD[30]	CSI[31]	MTL†[21]	**Ours** Rot.ImgN	**Ours** Combined
CIFAR10	CIFAR100	91.91	90.63	89.20	93.24	92.51	**94.20**
	SVHN	97.96	99.62	99.80	**99.92**	99.69	**99.92**
	ImageNet30	–	90.20	87.92	–	**94.16**	93.40
	TinyImageNet	92.06	92.25	92.44	92.99	**96.28**	95.02
	LSUN	93.57	96.51	91.60	95.03	**98.08**	97.52
	STL10	–	70.28	64.25	–	**77.29**	74.34
	Places365	92.57	95.21	90.18	93.72	**97.14**	96.01
	Texture	96.25	97.61	98.96	–	**99.16**	98.69
CIFAR100	CIFAR10	74.73	69.60	58.87	**79.25**	69.96	67.63
	SVHN	83.62	94.90	96.44	87.11	96.00	**97.17**
	ImageNet30	–	75.53	71.82	–	**84.82**	75.36
	TinyImagenet	77.56	79.52	79.28	80.66	**81.41**	79.75
	LSUN	71.86	79.50	61.83	74.32	**85.03**	74.55
	STL10	–	72.76	64.26	–	**79.96**	71.70
	Places365	74.57	79.60	65.48	77.87	**81.67**	72.79
	Texture	82.39	82.90	**87.47**	–	80.65	77.33

* Requires labels for the supervised training loss. Results reported from [21].
† Requires labels to select the optimal transformations.

$$\mathcal{L}_{pos} = -\sum_{x'' \in G} \sum_{\substack{x' \in V \\ x' \neq x''}} p_t(x'') \log(p_s(x')). \tag{3}$$

Additionally, DINO uses the multi-crop strategy [2], wherein M global views $G = \{x^g_1, ..., x^g_M\}$ and N local views, $L = \{x^l_1, \cdots, x^l_N\}$, are generated based on a set of transformations $\mathcal{T}$, e.g. crop and resize, horizontal flip, Gaussian blur, and color jitter. Global views are crops that occupy a larger region of the image (e.g. $\geq 40\%$) while local views cover small parts of the image (e.g. $\leq 40\%$). All $V = G \cup L$ views are passed through the student network, while the teacher has only access to the global views such that local-to-global correspondences are enforced. The trained teacher network is used for evaluation.

3.2 Negative samples

The learning objective (Eq. 3) assigns two transformed views of an image to the same soft-class. The applied transformations $\mathcal{T}$ are chosen to be sufficiently strong and diverse, such that the generated images generalise well over the training set but keep the semantics of the image they were derived from. The transformations are designed to learn higher-level features such as labels that represent semantic information and avoid learning lower-level features, such as edges or the color statistics over pixels [4]. The quality of the learned representation can be quantified by evaluating the K-NN accuracy for an in-distribution test set $\mathcal{D}^{in}_{test}$, using as higher-level feature vector an activity map of the network near the last layer. For OOD detection, the feature vector

representation should be enriched by in-distribution-specific features and depleted by features that frequently appear in other distributions from the same domain. This can be achieved by designing a negative distribution D_{neg} that keeps most of the low-level features of the in-distribution but changes the high-level semantics.

For example, a negative distribution for natural images can be realised by additionally rotating in-distribution images or images from a related auxiliary distribution by $r \sim R = \mathcal{U}(\{90°, 180°, 270°\})$, where $\mathcal{U}$ is the uniform distribution. It has been shown that using rotation as an additional positive transformation degrades the performance in the contrastive learning setup, where the objective is to maximize the mutual information between positive examples [4]. Motivated by this, authors in [31] report a performance gain for OOD detection by using rotation to generate negative examples.

3.3 Auxiliary objective

In addition to the self-distillation objective Eq. 3 we define an auxiliary task to encourage the student to have a uniform softmax response for negative examples. This task can be realised by a similar objective as Eq. 3 but with changed temperature $\tau_t \to \infty$ and transformations $\mathcal{T}$ applied to examples x from the negative set $\mathcal{D}_{neg}$, defined as:

$$\mathcal{L}_{neg} = -\frac{1}{K} \sum_{x' \in V} \log p_s(x'). \tag{4}$$

The total loss of our proposed method is defined by a linear combination of the two objectives

$$\mathcal{L}_{total} = \mathcal{L}_{pos} + \lambda \mathcal{L}_{neg}, \tag{5}$$

where $\lambda > 0$ is a balancing hyperparameter.

3.4 Sensitivity Score

Intuitively the sensitivity score is the degree of rejection of in-distribution examples which gives us a measure about the sensitivity of the OOD score to examples that have very similar features statistics to $\mathcal{D}^{in}_{train}$. To calculate the sensitivity score we randomly extract B samples from $\mathcal{D}^{in}_{train}$ without replacement as $\mathcal{D}^{sens}_{train}$ and denote the remaining train samples as $\mathcal{D}^{ref}_{train}$. We define the sensitivity score as the AUROC value between $\mathcal{D}^{sens}_{train}$ and $\mathcal{D}^{in}_{test}$, where $\mathcal{D}^{ref}_{train}$ is used as new train data during the evaluation.

4 Experiments

The proposed method is based on the vanilla DINO [3] implementation[3]. Unless otherwise specified, we use ViT-Small (ViT-S) with a patch size of 16. We use $N = 8$ local views for both positives and negatives, but two global positive views and one global negative view. Global views are resized to 256×256 while local views to 128×128. The temperatures are set to $\tau_t = 0.01$ and $\tau_s = 0.1$. In each epoch, we

[3] https://github.com/facebookresearch/dino

linearly decrease τ_t starting from 0.055 for CIFAR10 and from 0.050 for CIFAR100 to 0.01 during training. Sensitivity score is used to find optimal $\lambda = 1$. we set $K = 4096$ for all experiments. We use the Adamw optimizer [20] with an effective batch size of 256. The learning rate lr follows the linear scaling rule of $lr = \mathrm{lr}_{\mathrm{base}} \times \mathrm{batchsize}/256$, where $\mathrm{lr}_{\mathrm{base}} = 0.004$. All models are trained for 500 epochs. Experiments were conducted using 4 NVIDIA-A100 GPUs with 40GB of memory. The image augmentation pipeline $\mathcal{T}$ is based on [10, 3]. Finally, weight decay and learning rate are scaled with a cosine scheduler.

Table 2. AUROC scores for OOD Detection with CIFAR10 as $\mathcal{D}^{in}_{train}$ and different $\mathcal{D}_{neg}$. ImgN denotes ImageNet samples.

Negative Sampling:	None	Auxiliary								In-Dist
$\mathcal{D}^{out}_{test}$	DINO $\lambda = 0$	ImgN	Rot. ImgN	Rot. 360 ImgN	DTI	Perm-16 ImgN	Perm-4 ImgN	Rot. DTI	Pix. Perm.	Rot. In-Dist.
CIFAR100	90.29	90.46	92.51	88.62	93.77	88.32	89.57	93.77	87.67	**93.96**
SVHN	99.38	99.50	99.69	99.42	99.86	99.59	99.13	99.86	99.62	**99.92**
ImageNet30	88.81	89.96	94.16	88.95	93.39	89.17	84.71	**96.04**	87.46	91.69
TinyImageNet	91.07	94.14	**96.28**	91.60	94.53	89.72	91.27	95.64	89.39	94.27
LSUN	92.20	93.41	98.08	93.24	98.56	94.58	89.32	**99.12**	93.33	94.93
STL10	66.50	77.65	77.29	72.41	72.01	69.22	68.81	**81.49**	68.55	69.11
Places365	91.28	93.12	97.14	92.58	97.03	92.77	87.63	**98.12**	91.89	93.53
Texture	96.21	95.01	**99.16**	93.93	97.55	93.38	89.86	95.11	93.08	98.29
Average	89.47	91.66	94.29	90.09	93.34	89.59	87.54	**94.89**	88.87	91.96

4.1 Datasets and negative sample variants

We evaluate our method on CIFAR10 and CIFAR100 as in-distribution data. For auxiliary datasets, we use ImageNet [29] and Debiased 300K Tiny Images (DTI) [14]. The latter is a subset with 300K images from [32], where images belong to CIFAR10, CIFAR100, Places365 [37], and LSUN [36] classes are removed. To avoid shortcut learning (due to different image resolutions), we resize the auxiliary data to the size of the in-distribution data before applying any augmentation. For OOD detection, we consider common benchmark datasets, such as SVHN [24], Places365, Texture [6] and STL10. The following cases are considered for generating negative samples:

- DINO: no negatives are included ($\lambda = 0$).
- ImgN: samples from ImageNet.
- DTI: samples from Debiased Tiny Images.
- Rot.: samples are randomly rotated by $r \sim R = \mathcal{U}(\{90°, 180°, 270°\})$.
- Rot.360: samples are rotated by an angle randomly sampled from range $(0°, 360°)$.
- Perm-N: randomly permutes each part of the evenly partitioned image in N patches.
- Pix. Perm: randomly shuffles all the pixels in the image.
- Rot. In-Dist: a random rotation $r \sim R$ is applied to the in-distribution data.
- Combined: sample from both Rot. In-Dist and Rot. ImageNet are used.

4.2 Evaluation protocol for OOD detection

The DINO network structure $g(x)$ used in this work consists of a ViT-S as backbone, which maps the input x to a d-dimensional feature vector $f \in \mathbb{R}^d$, and two fully connected layers as head, which converts the features vector f to a K-dimensional output vector that enters the softmax layer. We define an anomaly detection score, $\mathcal{S}$, for the OOD test data $\mathcal{D}_{test}^{out}$ by computing the cosine similarity between the feature vector for a test image f_{test} and all features vectors f_m of the in-distribution training set. Instead of taking the maximum cosine similarity as a OOD score, we opt for a temperature weighted non-linear score,

$$\mathcal{S}(x) = -\frac{1}{M}\sum_{m=1}^{M} \exp\left(\frac{1}{\tau}\cdot\frac{f_{test}^T f_m}{\|f_{test}\|\|f_m\|}\right), \tag{6}$$

with $\tau = 0.04$ which is found by maximizing the sensitivity score. The value $\tau = 0.04$ is the average over optimal values for different datasets that typically lie in the range $[0.02, 0.08]$. The score is used to evaluate OOD performance by reporting the Area Under the Receiver Operating characteristic Curve (AUROC) between a given OOD test set and the in-distribution test set.

4.3 Experimental results

In Table 1, quantitative results are reported for CIFAR10 and CIFAR100 as in-distribution. We report results with Rot. ImgN as well as combining them with in-distribution rotated samples (Combined). When using CIFAR10 as $\mathcal{D}_{train}^{in}$, the proposed method shows superior performance in 6 out of 8 (75%) OOD datasets compared to current SOTA self-supervised methods. Surprisingly, we surpass hybrid methods, where self-supervised training is combined with human-labelled images. By further leveraging in-distribution negatives, we are able to surpass all other methods by 3.57% and 0.96% against self-supervised and supervised methods, respectively. Our results are roughly consistent for CIFAR100 as $\mathcal{D}_{train}^{in}$. We report superior performance in 6 out of 8 (75%) OOD datasets. Far OOD datasets have a substantial benefit, such as LSUN where we report a 5.53% gain against the best self-supervised method. Our results on near OOD, CIFAR10, are on par with self-supervised methods [31], while lacking behind supervised methods.

In Table 2, we investigate several ways to generate negative samples, as detailed in Section 4.1. It can be observed that by rotating both ImageNet and DTI with R, both distributions demonstrate an average performance gain of 2.63% and 1.55% respectively compared to no additional transformation.

It is worth noting that we abstain from reporting the performance of DTI in Table 1, since labels were used to form this subset of 300K images. Finally, we report an inferior (or on par) average AUROC score when employing Pix. Perm, Perm-4, and Perm-16 against the vanilla DINO method using ImageNet as the auxiliary dataset.

5 Discussion

Do negative samples lead to more condensed in-distribution representations? To understand the impact of the introduced negative sampling methods, we

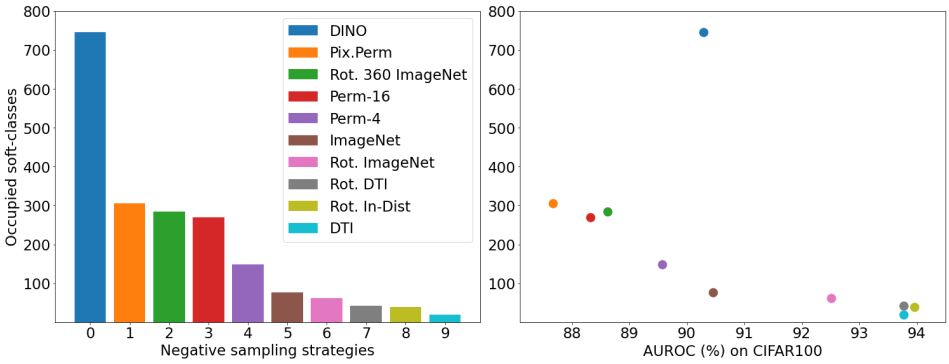


Fig. 2. We define a soft-class as "occupied" if the probability assigned to that soft-class is greater than the average probability of all K soft-classes. Colors indicate multiple $\mathcal{D}_{neg}$ and are shared within the two plots. The teacher network g_t is used to generate p_t from $\mathcal{D}^{in}_{test}$. Training is performed on CIFAR10. **Left**: $\mathcal{D}^{in}_{test}$ occupy less soft-classes with negative sampling compared to the DINO baseline. **Right**: relationship of occupied soft-classes with respect to AUROC score in CIFAR100.

investigate how many of the $K = 4096$ soft-classes are "occupied" by the $\mathcal{D}^{in}_{test}$ after training on CIFAR10. A soft-class is considered occupied if the probability assigned to the intended soft-class computed from test data is greater than the average soft-class probability. As depicted in Fig. 2 (left), negative sampling reduces the occupied classes compared to the DINO baseline. This observation is independent of how $\mathcal{D}_{neg}$ is created. More specifically, Rot. ImageNet, Rot. DTI, and Rot. In-Dist use roughly the same number of soft-classes and achieve SOTA AUROC scores on CIFAR100. By combining the aforementioned qualitative evaluations with Table 2, we claim that by contrasting $\mathcal{D}^{in}_{train}$ against $\mathcal{D}_{neg}$ a more condensed representation can be learnt.

Is OOD detection related to in-distribution classification? To answer this question, we investigate if there is a relationship between the OOD detection performance and the K-NN accuracy, determined from human-generated labels. To do so, we use CIFAR10 as $\mathcal{D}^{in}_{train}$ and CIFAR100 and Texture as $\mathcal{D}^{out}_{test}$, as representative cases of near OOD and far OOD, respectively. We find that the OOD AUROC score is positively correlated with K-NN accuracy for both near and far OOD detection (Fig. 3, top row).

Is the performance gain from use of transformers or auxiliary loss function? The performance gain stems from a more compact representation of high-level features for the in-distribution. This can be seen from the high K-NN values, that can be partially attributed to the DINO self-distillation framework (CIFAR10 K-NN accuracy of 93.2% for vanilla DINO vs 87.1% for CSI) and in part due to the negative loss (4.82% AUROC improvement with Rot.ImgN compared to vanilla DINO on CIFAR10). We highlight that K-NN correlates positively with OOD AUROC values (Fig. 3, top row).

Can an arbitrary auxiliary dataset be detrimental? Auxiliary negative datasets can be detrimental if they are semantically too close to the in-distribution, which

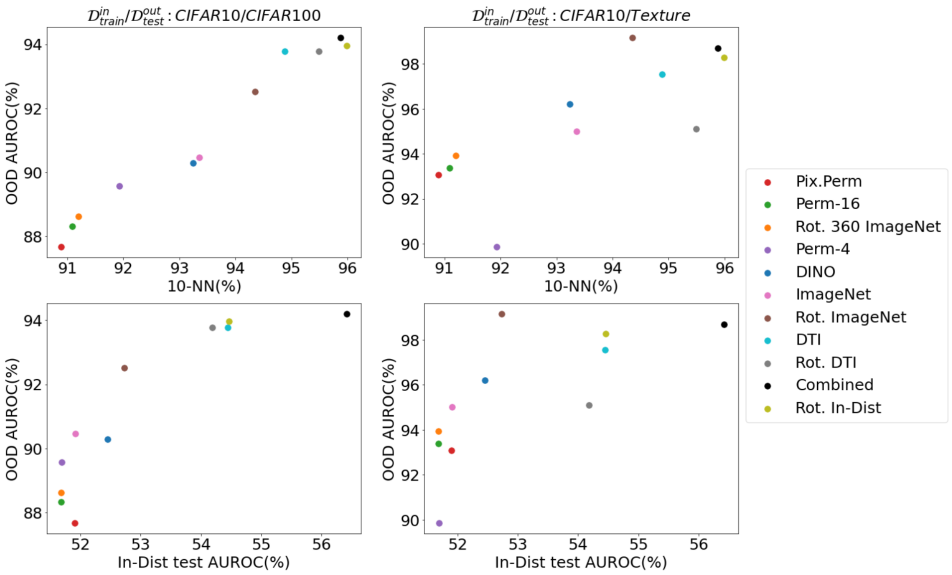


Fig. 3. We evaluate different models trained on CIFAR10 for two OOD datasets, CIFAR100 (left column) and Texture (right column). In each plot, points indicate different negative sampling strategies (colors are shared). **Top row**: correlation between OOD detection AUROC and K-NN accuracy on $\mathcal{D}^{in}_{test}$. **Bottom row**: correlation between OOD detection AUROC and AUROC score of $\mathcal{D}^{in}_{train}$ vs. $\mathcal{D}^{in}_{test}$. We observe models with higher sensitivity to detect $\mathcal{D}^{in}_{test}$ as outliers have higher OOD detection performance.

explains why non-rotated ImgN gives worse AUROC than Rot. ImgN for CIFAR10, despite the former being closer to the in-distribution. However, this effect can be detected by our sensitivity score, which is higher for Rot. ImgN (Fig. 3, bottom row). **How to choose good negative examples?** We use the sensitivity score to select $\mathcal{D}_{neg}$ (dataset + augmentation). Sensitivity values significantly higher than 0.5 indicate that negative examples are close enough to induce a difference between $\mathcal{D}^{in}_{train}$ and $\mathcal{D}^{in}_{test}$, but far enough to avoid a significant overlap of $\mathcal{D}^{in}_{train}$ with $\mathcal{D}_{neg}$ (see sensitivities of ImgN vs. Rot. ImgN, Fig. 3, bottom row).

6 Conclusion

In this work, we presented a new general method for self-supervised OOD detection. We demonstrated how self-distillation can be extended to account for positive and negative examples by introducing an auxiliary objective. The proposed objective introduces a form of contrastive learning, which pushes negative samples to be uniformly distributed among the existing in-distribution soft-classes. Additionally, we introduced a sensitivity analysis technique with which we can compare negative datasets and find the optimal values for the negative loss weight and the evaluation temperature without accessing the OOD validation set. The proposed method outperforms current SOTA for self-supervised OOD detection methods in the majority

of OOD benchmark datasets for both CIFAR10 and CIFAR100 as $\mathcal{D}^{in}_{train}$. We hope that the provided insights of our analysis will shed light on how to choose negative samples in more challenging vision domains.

References

1. Alexey, D., Fischer, P., Tobias, J., Springenberg, M.R., Brox, T.: Discriminative, unsupervised feature learning with exemplar convolutional, neural networks. IEEE TPAMI **38**(9), 1734–1747 (2016)
2. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882 (2020)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. arXiv preprint arXiv:2104.14294 (2021)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
5. Choi, S., Chung, S.Y.: Novelty detection via blurring. arXiv preprint arXiv:1911.11943 (2019)
6. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3606–3613 (2014)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Fort, S., Ren, J., Lakshminarayanan, B.: Exploring the limits of out-of-distribution detection. arXiv preprint arXiv:2106.03004 (2021)
9. Golan, I., El-Yaniv, R.: Deep anomaly detection using geometric transformations. Advances in neural information processing systems **31** (2018)
10. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems **33**, 21271–21284 (2020)
11. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
12. Hein, M., Andriushchenko, M., Bitterwolf, J.: Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem (2019)
13. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)
14. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. arXiv preprint arXiv:1812.04606 (2018)
15. Hendrycks, D., Mazeika, M., Kadavath, S., Song, D.: Using self-supervised learning can improve model robustness and uncertainty. Advances in Neural Information Processing Systems **32** (2019)
16. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
17. Koner, R., Sinhamahapatra, P., Roscher, K., Günnemann, S., Tresp, V.: Oodformer: Out-of-distribution detection transformer. arXiv preprint arXiv:2107.08976 (2021)

18. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
19. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690 (2017)
20. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam (2018)
21. Mohseni, S., Vahdat, A., Yadawa, J.: Shifting transformation learning for out-of-distribution detection (2021)
22. Nalisnick, E., Matsukawa, A., Teh, Y.W., Gorur, D., Lakshminarayanan, B.: Do deep generative models know what they don't know? arXiv preprint arXiv:1810.09136 (2018)
23. Nalisnick, E.T., Matsukawa, A., Teh, Y.W., Lakshminarayanan, B.: Detecting out-of-distribution inputs to deep generative models using a test for typicality. (2019)
24. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.: Reading digits in natural images with unsupervised feature learning (2011)
25. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 427–436 (2015)
26. Pidhorskyi, S., Almohsen, R., Doretto, G.: Generative probabilistic novelty detection with adversarial autoencoders. Advances in neural information processing systems **31** (2018)
27. Ren, J., Liu, P.J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., Lakshminarayanan, B.: Likelihood ratios for out-of-distribution detection. Advances in Neural Information Processing Systems **32** (2019)
28. Rezende, D.J., Mohamed, S.: Variational inference with normalizing flows (2016)
29. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
30. Sehwag, V., Chiang, M., Mittal, P.: Ssd: A unified framework for self-supervised outlier detection. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=v5gjXpmR8J
31. Tack, J., Mo, S., Jeong, J., Shin, J.: Csi: Novelty detection via contrastive learning on distributionally shifted instances. Advances in neural information processing systems **33**, 11839–11852 (2020)
32. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE transactions on pattern analysis and machine intelligence **30**(11), 1958–1970 (2008)
33. Van Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: International conference on machine learning. pp. 1747–1756. PMLR (2016)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
35. Winkens, J., Bunel, R., Roy, A.G., Stanforth, R., Natarajan, V., Ledsam, J.R., MacWilliams, P., Kohli, P., Karthikesalingam, A., Kohl, S., et al.: Contrastive training for improved out-of-distribution detection. arXiv preprint arXiv:2007.05566 (2020)
36. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015)
37. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017)
38. Zong, B., Song, Q., Min, M.R., Cheng, W., Lumezanu, C., Cho, D., Chen, H.: Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: International conference on learning representations (2018)

Chapter 7

Abnormality Detection for Medical Images Using Self-Supervision and Negative Samples

7.1 Summary

7.1.1 Introduction

In recent years advances in computer-aided technologies have made significant impacts on medical diagnosis, especially in the field of medical imaging. Despite the success of machine learning algorithms in medical imaging, many of the recent advanced approaches in data driven models have mostly focused on the natural images domain. In particular, recent progresses in the self-supervised representation learning of natural images has not been studied thoroughly in the medical imaging domain. This is important as one of the main challenges in medical diagnosis is the lack of annotated images for different abnormalities compared to images taken from healthy people. Therefore, we can benefit from self-supervised methods that can be trained using only normal images and then used for abnormality detection. Motivated by this, the application of negative sampling in self-supervised representation learning based on self-distillation is investigated for the task of abnormality detection. To examine the general applicability of this method, three different types of medical image datasets are used, including X-ray, colonoscopy, and ophthalmology images for pneumonia, polyp, and glaucoma detection, respectively.

For the X-ray images of the Radiological Society of North America (RSNA)

dataset is used [11]. This dataset is publicly available and includes a frontal view of chest radiographs with $8,851$ samples labeled as "Normal", $11,821$ as "No Opacity/Not Normal", and $6,012$ labeled as "Opacity". For glaucoma detection, the LAG dataset is used [94]. LAG includes the total number of $4,854$ samples with $1,711$ positive glaucoma (abnormal) and $3,143$ images of negative glaucoma (normal). For polyp detection, Hyper-Kvasir is used [95]. Hyper-Kvasir is one of the largest publicly available gastrointestinal images containing $2,100$ normal and 1000 images with a polyp.

7.1.2 Training procedure

To learn the representation, we benefit from the general applicability of the method introduced in the previous chapter. Positive pairs are created by applying random augmentation extract from set $\mathcal{T}$. Negative examples are created by applying shifting transformation on in-distribution training or samples of an auxiliary dataset. In this study, two types of auxiliary datasets are examined. First, the ImageNet dataset from the natural image domain, and Second, an auxiliary dataset from the related domain of the in-distribution dataset if available. Figure 7.1 shows a general overview of the model.

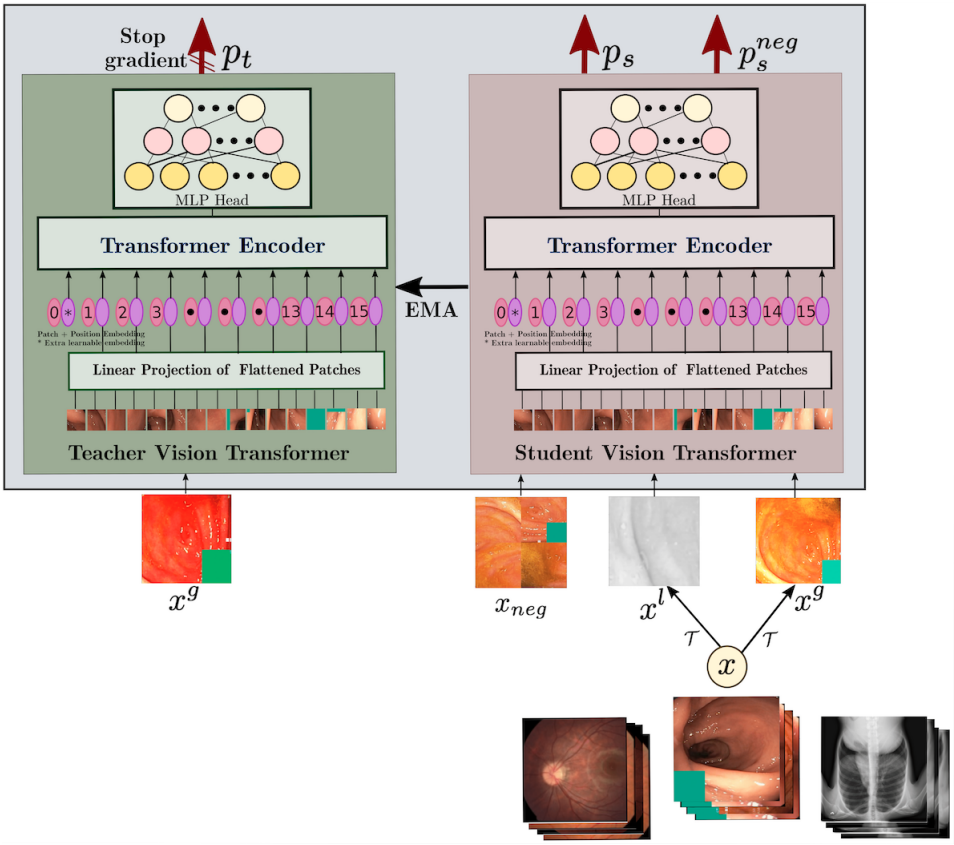


Figure 7.1: Overview of the proposed self-supervised framework, comprising student network (right) and teacher network (left). Student and teacher map two randomly augmented views of the same image to the same class. x^g and x^l are global and local views of image x where $x^g \sim \mathcal{T}(x)$ and $x^l \sim \mathcal{T}(x)$. A negative sample, x_{neg}, is generated by applying first a shifting transformation, such as random rotation, followed by $\mathcal{T}$ to either an in-dist image x or an auxiliary image x_{aux}.

7.1.3 Evaluation metrics

To evaluate the performance of the models trained on the three different datasets, for a given test sample x, cosine similarity score $\mathcal{S}_{cs}(x)$ and Mahalanobis distance $\mathcal{S}_{md}(x)$ are calculated as in equation 7.1 and equation 7.2 respectively.

$$\mathcal{S}_{md}(x) := (f_{test} - \mu_m)^T \Sigma_m^{-1} (f_{test} - \mu_m) \tag{7.1}$$

where μ_m and Σ_m are the mean and covariance of the all feature vectors f_m from the training data.

$$\mathcal{S}_{cs}(x) := -\max_m \exp\left(\frac{f_{test}^T f_m}{\|f_{test}\|\|f_m\|}\right) \tag{7.2}$$

7.1.4 Experimental results

As depicted in tables 7.1 and 7.2, the proposed method can achieve better performance compared to the previous studies when the negative samples are created based on the combination of in-distribution train and domain related auxiliary dataset.

Table 7.1: AUROC of OOD detection method trained on **RSNA** dataset.

Method	$\mathcal{D}_{ood}$: Opacity		$\mathcal{D}_{ood}$: No Opacity	
Unsupervised methods trained on normal samples				
UAE [81]	0.89		0.78	
Deep AD [74]	0.838		0.704	
[96]	**0.940**		0.828	
Score	$\mathcal{S}_{md}$	$\mathcal{S}_{cs}$	$\mathcal{S}_{md}$	$\mathcal{S}_{cs}$
Ours	**0.941**	0.764	**0.841**	0.714

Table 7.2: AUROC results on **Hyper-Kvasir** and **LAG** datasets.

Method	$\mathcal{D}_{train}$: Hyper-Kvasir, $\mathcal{D}_{ood}$: Polyp		$\mathcal{D}_{train}$: LAG, $\mathcal{D}_{ood}$: Glaucoma	
Unsupervised methods trained on normal samples				
CAVGA-R_u [97]	0.928		0.819	
IGD [98]	0.937		0.857	
CCD [99]	0.972		0.874	
Score	$\mathcal{S}_{md}$	$\mathcal{S}_{cs}$	$\mathcal{S}_{md}$	$\mathcal{S}_{cs}$
Ours	**0.996**	0.994	0.849	**0.879**

Figure 7.2 shows the result of comparing different methods of creating negative samples. The left plot shows that using a combination of in-distribution data and auxiliary datasets always performs better for these three medical datasets. The right plot shows that domain related auxiliary datasets achieve superior results compared to out of domain auxiliary datasets such as ImageNet.

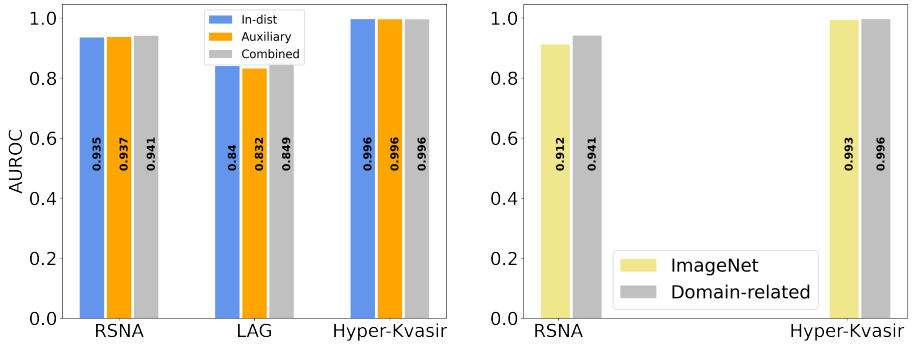


Figure 7.2: **Left**. AUROC results based on $\mathcal{S}_{md}$ for different negative sets were generated from in-dist train data, an auxiliary dataset, or a combination of both. **Right**. AUROC results across different auxiliary datasets where we take images from an in-domain medical dataset or out-domain.

The effect of different shifting transformations is studied and the results are compared for all three datasets. As it is illustrated in table 7.3, rotation (Rot) results in a more effective negative sample for both RSNA and LAG datasets. However, for the Hyper-Kvasir dataset, patch permutation with 4 division (Perm-4) gives better results.

Table 7.3: The impact of different shifting transformations on AUROC results. Reported scores are for $\mathcal{S}_{md}$ ($\mathcal{S}_{cs}$).

In-dist Dataset	NoNeg	Shifting transformations				
		Rot	Rot-360	Perm-4	Perm-16	Pixel-Shuffle
RSNA	0.925(0.888)	**0.941**(**0.764**)	0.933(0.766)	0.924(0.634)	0.908(0.887)	0.925(0.733)
LAG	0.799(0.862)	**0.849**(**0.879**)	0.831(0.866)	0.807(0.873)	0.814(0.881)	0.797(0.860)
Hyper-Kvasir	0.974(0.875)	0.989(0.915)	–	**0.996**(**0.994**)	0.985(0.960)	0.994(0.985)

7.1.5 Conclusion

In this study, the applicability of the proposed method in Chapter 6 on abnormality detection in medical images is investigated. Three different types of medical images are used to study the impact of negative samples created using in-distribution train data, domain related auxiliary data, and ImageNet dataset. According to the experimental results combination of in-distribution train data and domain related auxiliary datasets create more effective negative samples. Additionally, rotation as a shifting transformation results in better performance for most of the datasets.

7.2 Abnormality detection for medical images using self-supervision and negative samples

Nima Rafiee, Rahil Gholamipoorfard, Markus Kollmann, 2022.

Status: Submitted to *MICCAI*.

Contributions: The research and preparation of this manuscript were done jointly by Nima Rafiee and Rahil Gholamipoorfard under the supervision of Prof. Markus Kollmann.

Abnormality Detection for Medical Images Using Self-Supervision and Negative Samples

Nima Rafiee[1*][0000−0002−3193−9534], Rahil Gholamipoor[1*][0000−0001−8207−7295], and Markus Kollmann[1,2][0000−0002−5317−5408]

[1] Department of Computer Science, Heinrich Heine University, Düsseldorf, Germany
[2] Department of Biology, Heinrich Heine University, Düsseldorf, Germany
{nima.rafiee,rahil.gholamipoorfard,markus.kollmann}@hhu.de

Abstract. Recent progress in computer-aided technologies has considerable impact on helping experts with a reliable and fast diagnosis of abnormal samples. In particular, self-supervised and self-distillation techniques have advanced automated out-of-distribution (OOD) detection in the image domain. Further improvements for OOD detection have been observed by including negative samples derived from shifting transformations of real images. In this work, we study different ways of creating negative samples for medical images and how effective they are when leveraging them in a self-supervised self-distillation framework. We investigate the impact of various types of negative examples by applying different shifting transformations on samples when they are derived from in-distribution training data, from an auxiliary dataset, or a combination of both. For the case of the auxiliary dataset, we compare the OOD detection performance when auxiliary samples are extracted from an in-domain or an out-domain. Our approach uses only data belonging to healthy people during the training procedure and does not require any additional information from labels. We demonstrate the efficiency of our technique by comparing abnormality detection performance on diverse medical datasets, setting new benchmarks for pneumonia, polyp, and glaucoma detection from X-ray, colonoscopy, and ophthalmology images.

Keywords: Abnormality detection · Self-supervised learning · Medical imaging.

1 Introduction

In recent years, computer-aided diagnosis in medical image screening has gained increased attention. In particular, detecting whether a sample includes some abnormality can help medical experts with faster and more reliable decision making. Diagnosis problems can be frequently assigned to the problem of out-of-distribution (OOD) detection in machine learning and statistical inference. OOD detection or anomaly detection refers to the problem of detecting if a test

[*] Equal contribution

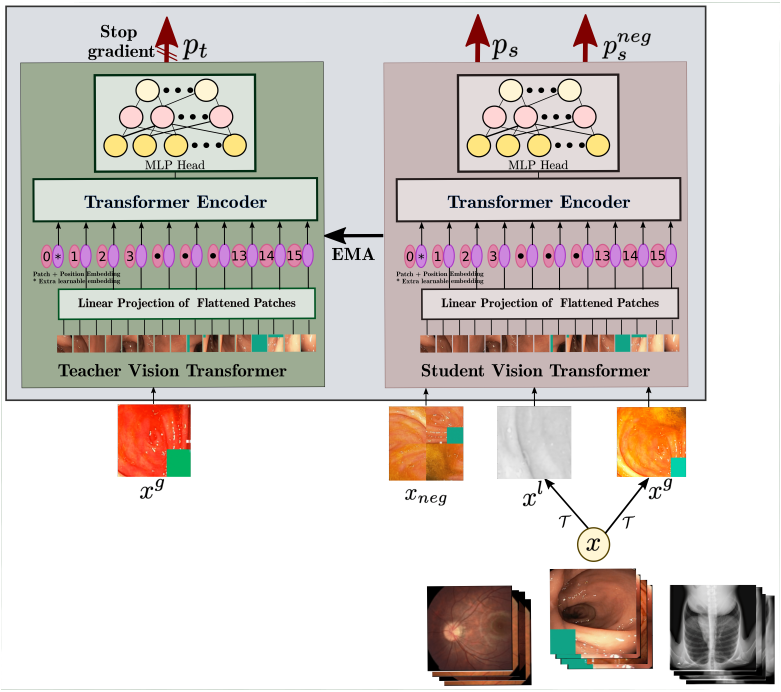


Fig. 1: Overview of the proposed self-supervised framework, comprising student network (right) and teacher networks (left). Student and teacher map two randomly augmented views of the same image to the same class. x^g and x^l are global and local views of image x where $x^g \sim \mathcal{T}(x)$ and $x^l \sim \mathcal{T}(x)$. Negative sample, x_{neg}, is generated by applying first a shifting transformation, such as random rotation, followed by $\mathcal{T}$ to either an in-dist image x or an auxiliary image x_{aux}.

sample has the same distribution as training data or is drawn from a different distribution. Diverse techniques developed for computer vision problems have been successfully applied to abnormality detection in the medical field. In [25, 29] deep supervised methods are used to classify X-ray and colonoscopy images. Despite the promising results, these approaches rely on annotated samples for abnormalities that are not available or only available to very limited number. Typically, the number of healthy samples outnumbers abnormal ones, which results in a challenging unbalanced classification problem. To overcome these problems, many studies have investigated the use of unsupervised or semi-supervised methods [7, 26, 19]. These methods aim at detecting abnormalities by learning the distribution of healthy/normal data. A well-studied category of unsupervised methods named Variational Autoencoders (VAEs) [13] uses reconstruction error. The assumption is that abnormal samples can not be reconstructed equally accurately as training data (lower likelihood) where the model only uses normal images during the training. However, it has been shown that in practice, these models can

be prone to reconstruct the abnormal samples fairly well, which lowers the detection performance [25, 2]. Furthermore, it is shown that these density estimation based methods can assign a higher likelihood to OOD samples compared to in-distribution (in-dist) test data [18]. Recently the effectiveness of self-supervised learning has received considerable attention in different domains, such as the visual domain [5], which enables learning robust representations from unlabeled data. Due to their efficiency, self-supervised pretext tasks such as predicting geometric transformations [10] or contrastive learning [27, 30, 9, 22, 24] have been designed for OOD detection in both natural and medical images. In [24] negative samples, drawn from shifting transformation of train data, are incorporated into a contrastive method to further tighten the decision boundary around normal samples resulting in an improved OOD detection score. This approach is also supported by Ref. [11] where supervised and density estimator models are exposed to some auxiliary datasets and negative samples. In [20] a self-distillation approach similar to DINO [4] is used with negative samples in order to compensate for limitations of contrastive based methods. Despite the numerous studies of leveraging negative samples in natural images, we believe it has remained untapped in the field of medical image processing. In this work, we study different ways of creating negative examples by applying shifting transformations on in-dist train data, samples from an auxiliary dataset or a combination of both. We show how these different negative samples can affect the performance of abnormality detection when leveraging them into a self-distillation self-supervised method. With the general assumption that effective transformations are the ones that change the high-level semantic while keeping the low-level statistics, we can achieve state-of-the-art (SOTA) results on abnormality detection for three different medical datasets including detecting pneumonia, polyp, and glaucoma from chest X-ray, colonoscopy, and from ophthalmology images with only access to normal samples. Additionally, we compare two evaluation metrics, cosine similarity and Mahalanobis distance, for OOD detection.

2 Method

In this section, we describe our proposed approach, Fig. 1. Similar to [4], our framework use teacher and student networks that have the same architecture, Vision Transformer [8] (ViT), and use distillation during training. Student and teacher are parametrized by two identical networks $g_s = g(x|\theta_s)$ and $g_t = g(x|\theta_t)$ which have different set of parameters. For an augmented input image x, both student and teacher output K-dimensional vectors including soft-classes. The probability of x falling in soft-class k is computed using temperature-scaled softmax function defined as

$$p_s^k(x) = \frac{\exp\left(g_s^k(x)/\tau_s\right)}{\sum_{i=1}^{K} \exp\left(g_s^i(x)/\tau_s\right)}, \tag{1}$$

where $\tau_s > 0$ is student temperature. The same formula, Eq. 1, holds for teacher with temperature τ_t. The student parameters are updated by back-propagating

the gradients through the student network while the teacher parameters are updated with the Exponential Moving Average (EMA) of the student parameters

$$\theta_t \leftarrow m\theta_t + (1-m)\theta_s, \tag{2}$$

where $0 \leq m \leq 1$ is a momentum parameter. For $\tau_t < \tau_s$, the training objective is given by the cross entropy (CE) loss for two non-identical transformations x'', x' of an image x drawn from the training set, $\mathcal{D}_{train}$

$$\mathcal{L} = -\sum_{x'' \in G} \sum_{\substack{x' \in G \cup L \\ x' \neq x''}} p_t(x'') \log p_s(x'). \tag{3}$$

We additionally use the multi-crop strategy [3], wherein M global views $G = \{x_1^g, ..., x_M^g\}$ and N local views, $L = \{x_1^l, \cdots, x_N^l\}$, are generated based on a set of transformations $\mathcal{T}$. Global views usually cover a larger region of the original image while local views cover smaller as they are results of a stronger cropping. All global and local views are passed through the student network, while the teacher has only access to the global views encouraging local-to-global correspondence. The CE loss, Eq. 3, is minimized such that two transformed views of an input image are assigned to the same soft-class. The applied transformations $\mathcal{T}$ are chosen to be strong and diverse enough such that the generated images generalise well over the training data. The transformations are designed in order to learn higher-level features, semantic information, and avoid learning lower-level features.

2.1 Auxiliary objective for OOD detection

For OOD detection, the representations should be enriched by in-dist specific features and deprived of features that frequently appear in other distributions from the same domain. This can be achieved by designing a negative distribution $\mathcal{D}_{neg}$ that keeps most of the low-level features of the in-dist data but changes the high-level semantics.
In addition to the self-distillation objective, Eq. 3, we define an auxiliary task to encourage the student to have a uniform softmax response for negative examples. This can be done by a similar objective as Eq. 3 when temperature $\tau_t \to \infty$

$$\mathcal{L}_{neg} = -\frac{1}{K} \sum_{x_{neg} \in \mathcal{D}_{neg}} \log p_s(x_{neg}). \tag{4}$$

The total loss of our proposed method is defined by a linear combination of the two objectives

$$\mathcal{L}_{total} = \mathcal{L} + \lambda \mathcal{L}_{neg}, \tag{5}$$

where $\lambda > 0$ is a balancing hyperparameter.

2.2 Negative samples

A negative distribution $\mathcal{D}_{neg}$ can be realised by additionally applying shifting transformations to samples from $\mathcal{D}_{train}$ or from an auxiliary set augmented by $\mathcal{T}$. We consider the following shifting transformations to shape $\mathcal{D}_{neg}$.

- NoNeg: no negative samples are included ($\lambda = 0$).
- Rot: samples are randomly rotated by $r \sim \mathcal{U}(\{90°, 180°, 270°\})$.
- Rot-360: rotation by an angle randomly sampled from range (0°, 360°).
- Perm-n: random permutation of image patches where the image is sliced in n square patches.
- Pixel-Shuffle: randomly shuffles all pixels in the image.

2.3 Evaluation protocol for OOD detection

Different studies have shown the advantage of using Mahalanobis distance and cosine similarity as two metrics for OOD detection [24, 22, 9]. We compare the effectiveness of these two metrics for different medical datasets in section 4. To calculate scores, we drop the fully connected head and use normalised ViT backbone output as feature vector f for calculating evaluation scores. For each given test sample x, we calculate Mahalanobis distance based anomaly score, $\mathcal{S}_{md}(x)$, as

$$\mathcal{S}_{md}(x) := (f_{test} - \mu_m)^T \Sigma_m^{-1} (f_{test} - \mu_m) \tag{6}$$

where μ_m and Σ_m are the mean and covariance of the all feature vectors f_m from the training data, $\mathcal{D}_{train}$. We calculate the cosine similarity based anomaly score $\mathcal{S}_{cs}(x)$ for test sample x

$$\mathcal{S}_{cs}(x) := -\max_m \exp\left(\frac{f_{test}^T f_m}{\|f_{test}\|\|f_m\|}\right) \tag{7}$$

Detection is assessed with Area Under the Receiver Operating Characteristic curve (AUROC).

3 Experimental Setup

3.1 Dataset

We assess our model performance on three different health screening medical imaging benchmarks, chest X-ray images, colonoscopy images and fundus images for glaucoma detection.

RSNA. The Radiological Society of North America (RSNA) Pneumonia Detection Challenge dataset [23] is a publicly available dataset of frontal view chest radiographs. Each image was labeled as "Normal", "No Opacity/Not Normal" or "Opacity". The Opacity group consists of images with opacities suspicious for pneumonia, and images labeled "No Opacity/Not Normal" may have lung opacity but no opacity suspicious for pneumonia. The RSNA dataset contains 26,684

X-rays with 8,851 normal, 11,821 no lung opacity/not normal and 6,012 lung opacity.

Hyper-Kvasir. The Hyper-Kvasir dataset is the largest public gastrointestinal dataset [1]. The data were collected during real examinations and partially labeled by experienced endoscopists. The dataset contains 110,079 images from patients, with 10,662 labelled images. Following [27] we take 2,100 images from "cecum", "ileum" and "bbps-2–3" cases as normal and 1000 abnormal images from "polyp" as abnormal. We take 1,600 images for training set and 500 images for test set.

LAG. The LAG dataset is a large scale image dataset for glaucoma detection [14], containing 4,854 images with 1,711 positive glaucoma (abnormal) and 3,143 negative glaucoma (normal) scans. For consistent comparison, following [27], we take 2,343 normal images for training and 800 normal images and 1,711 abnormal images for testing.

Table 1: AUROC of OOD detection method trained on **RSNA** dataset

Method	$\mathcal{D}_{ood}$: Opacity		$\mathcal{D}_{ood}$: No Opacity	
Unsupervised methods trained on normal samples				
UAE[16]	0.89		0.78	
Deep AD[17]	0.838		0.704	
[9]	**0.940**		0.828	
Score	$\mathcal{S}_{md}$	$\mathcal{S}_{cs}$	$\mathcal{S}_{md}$	$\mathcal{S}_{cs}$
Ours	**0.941**	0.764	**0.841**	0.714

3.2 Auxiliary Dataset

For auxiliary dataset, we compare use of samples from ImageNet or from a in-domain one if any available. For RSNA dataset of X-ray images we use CheXpert [12] and for Hyper-Kvasir dataset of colonoscopy images all unlabled Hyper-Kvasir images are taken as in-domain. For LAG dataset, we only use ImageNet due to unavailability of any in-domain dataset. We highlight that we do not use any label information to shape negative samples.

3.3 Training

Our proposed method has the same structure as DINO implementation. we use ViT-Small (ViT-S) backbone for all different training data. A patch size of 16 and $N=8$ local views for both positives and negatives, but two global positive views and one global negative view are used. All global views are resized to 256×256 while local views to 96×96. The temperatures for teacher and student network are set to $\tau_t = 0.01$ and $\tau_s = 0.1$. During training, τ_t is linearly decreased from 0.04 to 0.01 in each epoch. λ and K are set to 1 and 4096 respectively for all our experiments. We use the Adamw optimizer [15] with an effective batch size of 256. For the base learning rate $\mathrm{lr}_{\mathrm{base}}$, we use 0.001 for Hyper-Kvasir and LAG datasets and 0.002 for RSNA. For each dataset, we trained the model for 700

epochs. We conducted our experiments using 4 NVIDIA-A100 GPUs with 40 GB of memory. The image augmentation pipeline $\mathcal{T}$ is based on DINO except that for Hyper-Kvasir dataset we rotate all positive views with same randomly chosen angle to avoid information leak from position of existing green boxes in images. Finally, weight decay and learning rate are scaled with a cosine scheduler.

Table 2: AUROC results on **Hyper-Kvasir** and **LAG** datasets

Method	$\mathcal{D}_{train}$: Hyper-Kvasir, $\mathcal{D}_{ood}$: Polyp		$\mathcal{D}_{train}$: LAG, $\mathcal{D}_{ood}$: Glaucoma	
Unsupervised methods trained on normal samples				
CAVGA-R_u [28]	0.928		0.819	
IGD [6]	0.937		0.857	
CCD [27]	0.972		0.874	
Score	$\mathcal{S}_{md}$	$\mathcal{S}_{cs}$	$\mathcal{S}_{md}$	$\mathcal{S}_{cs}$
Ours	**0.996**	0.994	0.849	**0.879**

4 Experimental Results

We compare the proposed method with unsupervised methods trained on only healthy images. We report our results for both $\mathcal{S}_{md}$ and $\mathcal{S}_{cs}$ scores. In Table 1, on RSNA dataset, our method outperforms the contrastive self-supervised SOTA method [9] when taking $\mathcal{S}_{md}$ as the anomaly score. In Table 2, we inspect the anomaly detection performance on the Hyper-Kvasir dataset for polyp detection and on the LAG dataset for glaucoma detection. Our method can surpass the recently proposed self-supervised anomaly detection method, CCD [27] on both polyp and glaucoma detection where we take $\mathcal{S}_{md}$ and $\mathcal{S}_{cs}$ respectively.
In Table 3, the impact of different shifting transformations, as explained in section 2.2 is explored. We found out that transformations such as Rot have better performance than excluding negative samples or using non-effective transformations such as Pixel-Shuffle. This result supports our general assumption about a good negative transformation that changes high-level semantics and keeps low-level statistics. Note that for Hyper-Kvasir, positive views are rotated by the same angle randomly selected from $\mathcal{U}(\{90°, 180°, 270°\})$ thus, we skip applying Rot-360 as a shifting transformation. In Fig. 2 [Left], we examine the effect of creating negative samples by applying shifting transformation on samples from each in-dist training, auxiliary dataset, or a combination of both. For RSNA and LAG dataset, as it is shown, the AUROC score increases where a combination of both is used, while for Hyper-Kvasir, we see no difference. Moreover, the use of only auxiliary datasets shows slightly better performance for RSNA compared to only taking in-dist negative samples on the other hand for LAG in-dist negative samples have higher score. The reason can be that for RSNA the in-domain auxiliary datasets are from a broader distribution compared to in-dist train data with a higher chance of resembling OOD samples but for LAG even though the ImageNet dataset has a broader distribution, in-dist negatives are harder negative samples which can be more advantageous [21]. The evaluation on taking in-domain or out-domain auxiliary datasets is shown in Fig. 2

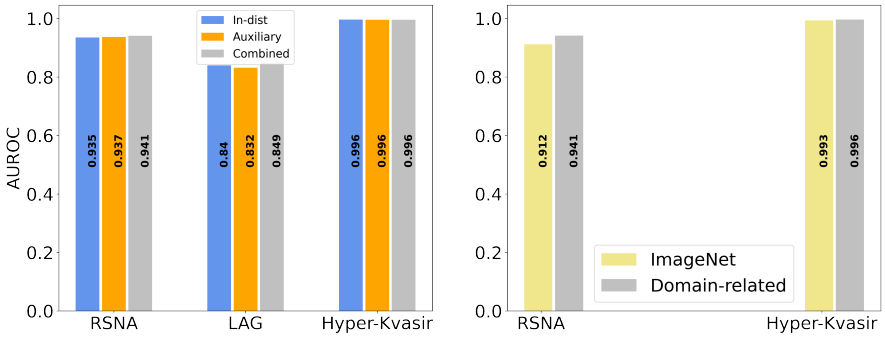


Fig. 2: **Left**. AUROC results based on $\mathcal{S}_{md}$ for different negative sets where generated from in-dist train data, an auxiliary dataset or a combination of both. **Right**. AUROC results across different auxiliary datasets where we take images from an in-domain medical dataset or out-domain.

[Right]. For RSNA X-ray images, OOD detection performance is improved by a large margin when negative samples are from an in-domain auxiliary set. However, for Hyper-Kvasir, the out-domain auxiliary has approximately the same performance as the in-domain.

Table 3: The impact of different shifting transformations on AUROC results. Reported scores are for $\mathcal{S}_{md}$ ($\mathcal{S}_{cs}$).

In-dist Dataset	NoNeg	Shifting transformations				
		Rot	Rot-360	Perm-4	Perm-16	Pixel-Shuffle
RSNA	0.925(0.888)	**0.941**(**0.764**)	0.933(0.766)	0.924(0.634)	0.908(0.887)	0.925(0.733)
LAG	0.799(0.862)	**0.849**(**0.879**)	0.831(0.866)	0.807(0.873)	0.814(0.881)	0.797(0.860)
Hyper-Kvasir	0.974(0.875)	0.989(0.915)	–	**0.996**(**0.994**)	0.985(0.960)	0.994(0.985)

5 Conclusion

In this study, we present a self-supervised method which leverages self-distillation and negative samples for the task of abnormality detection without accessing label information. We study different ways of creating negative samples by applying shifting transformations on in-dist training data, an auxiliary dataset, or a combination of both. Additionally, we compare the impact of having auxiliary samples from domain-related distribution or from a different domain such as ImageNet. Moreover, we compare the abnormality detection performance using two different evaluation metrics including cosine similarity and Mahalanobis distance. A major motivation behind this work is that we take only normal samples during training which makes our method suitable for yet unknown abnormalities. In anomaly detection, our method outperforms SOTA methods on the RSNA, Hyper-Kavsir and LAG datasets.

References

1. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific data **7**(1), 1–14 (2020)
2. Çallı, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K.G., Murphy, K.: Deep learning for chest x-ray analysis: A survey. Medical Image Analysis p. 102125 (2021)
3. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882 (2020)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. arXiv preprint arXiv:2104.14294 (2021)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
6. Chen, Y., Tian, Y., Pang, G., Carneiro, G.: Unsupervised anomaly detection and localisation with multi-scale interpolated gaussian descriptors. arXiv e-prints pp. arXiv–2101 (2021)
7. Davletshina, D., Melnychuk, V., Tran, V., Singla, H., Berrendorf, M., Faerman, E., Fromm, M., Schubert, M.: Unsupervised anomaly detection for x-ray images (2020)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
9. Gholamipoor, R., Rafiee, N., Kollmann, M.: Pneumonia detection with semantic similarity scores (2021)
10. Golan, I., El-Yaniv, R.: Deep anomaly detection using geometric transformations. Advances in neural information processing systems **31** (2018)
11. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. arXiv preprint arXiv:1812.04606 (2018)
12. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison (2019)
13. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
14. Li, L., Xu, M., Wang, X., Jiang, L., Liu, H.: Attention based glaucoma detection: a large-scale database and cnn model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10571–10580 (2019)
15. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam (2018)
16. Mao, Y., Xue, F.F., Wang, R., Zhang, J., Zheng, W.S., Liu, H.: Abnormality detection in chest x-ray images using uncertainty prediction autoencoders. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 529–538. Springer (2020)
17. Nakao, T., Hanaoka, S., Nomura, Y., Murata, M., Takenaga, T., Miki, S., Watadani, T., Yoshikawa, T., Hayashi, N., Abe, O.: Unsupervised deep anomaly detection in chest radiographs. Journal of Digital Imaging pp. 1–10 (2021)

18. Nalisnick, E., Matsukawa, A., Teh, Y.W., Gorur, D., Lakshminarayanan, B.: Do deep generative models know what they don't know? (2019)
19. Perera, P., Nallapati, R., Xiang, B.: Ocgan: One-class novelty detection using gans with constrained latent representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2898–2906 (2019)
20. Rafiee, N., Gholamipoorfard, R., Adaloglou, N., Jaxy, S., Ramakers, J., Kollmann, M.: Self-supervised anomaly detection by self-distillation and negative sampling. arXiv preprint arXiv:2201.06378 (2022)
21. Robinson, J., Chuang, C.Y., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. arXiv preprint arXiv:2010.04592 (2020)
22. Sehwag, V., Chiang, M., Mittal, P.: Ssd: A unified framework for self-supervised outlier detection. arXiv preprint arXiv:2103.12051 (2021)
23. Shih, G., Wu, C.C., Halabi, S., Kohli, M., Prevedello, L., Cook, T., Sharma, A., Amorosa, J., Arteaga, V., Galperin-Aizenberg, M., Gill, R., Godoy, M., Hobbs, S., Jeudy, J., Laroia, A., Shah, P., Vummidi, D., Yaddanapudi, K., Stein, A.: Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiology. Artificial intelligence **1 1**, e180041 (2019)
24. Tack, J., Mo, S., Jeong, J., Shin, J.: Csi: Novelty detection via contrastive learning on distributionally shifted instances. In: 34th Conference on Neural Information Processing Systems (NeurIPS) 2020. vol. 33, pp. 11839–11852 (2020)
25. Tang, Y.X., Tang, Y.B., Peng, Y., Yan, K., Bagheri, M., Redd, B.A., Brandon, C.J., Lu, Z., Han, M., Xiao, J., Summers, R.M.: Automated abnormality classification of chest radiographs using deep convolutional neural networks. npj Digital Medicine **3**(1), 1–8 (2020)
26. Tang, Y., Tang, Y., Han, M., Xiao, J., Summers, R.M.: Abnormal chest x-ray identification with generative adversarial one-class classifier (2019)
27. Tian, Y., Pang, G., Liu, F., Chen, Y., Shin, S.H., Verjans, J.W., Singh, R., Carneiro, G.: Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 128–140. Springer (2021)
28. Venkataramanan, S., Peng, K.C., Singh, R.V., Mahalanobis, A.: Attention guided anomaly localization in images. In: European Conference on Computer Vision. pp. 485–503. Springer (2020)
29. Wang, Y., Feng, Z., Song, L., Liu, X., Liu, S.: Multiclassification of endoscopic colonoscopy images based on deep transfer learning. Computational and Mathematical Methods in Medicine **2021**
30. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text (2020)

Chapter 8

Conclusion and Future Works

Out of distribution detection is the problem of detecting samples that are drawn from a different distribution rather than the already seen training dataset. Anomaly detection has gained many applications in the vision domain, such as fault detection in manufacturing production lines and medical image diagnosis using supervised and unsupervised machine learning approaches. However, there are substantial challenges and limitations attributed to these approaches. Motivated by recently introduced self-supervised learning (SSL) methods, in this work, we proposed and studied SSL frameworks to learn representations that benefit the downstream task of anomaly detection for natural and medical images.

In Chapter 2, supervised and unsupervised approaches to anomaly detection are explained in detail. Some of the limitations of both approaches are discussed. For the supervised approach, the lack of labeled data, learning superficial features, assigning high confidence to anomaly samples, and the hard discriminative problem are explained. For unsupervised density estimators, the issue of assigning higher likelihood to OOD samples is investigated.

In chapters 3 and 4, the required machine learning background and a literature review of self-supervised methods are explained, respectively.

In Chapter 5, a contrastive SSL method is used to learn representation from medical X-ray images aiming at detecting pneumonia. Mahalanobis distance is then applied to the learned features to calculate a score function. This score function is used to calculate a similarity measure between a given test sample and the training data. A common problem in medical image diagnosis is the lack of annotated data. Moreover, many of the existing X-ray image samples belong to healthy people, and the number of samples that include abnormalities is much less than the normal ones. Thus, the use of the supervised method is challenging as the training datasets are highly imbalanced. The proposed method in Chapter 4 benefits from using only normal samples and removes the need for the existence of abnormalities. It is

also illustrated that self-supervised pretraining can considerably improve sample efficiency when there exist an enormous number of unlabeled data and only a few annotated ones.

In Chapter 6, motivated by the limitation of the contrastive SSL method used in Chapter 5, such as the requirement for large batch size and issues regarding the definition of negative examples in contrastive objective explained in Chapter 4, a new SSL framework for anomaly detection is introduced. The proposed frameworks leverage self-distillation using a teacher-student structure and negative sampling. A systematic way is proposed to create a negative set by applying shifting transformation on either in-distribution training data or an auxiliary dataset. A sensitivity score which is the AUROC value between the in-distribution training set and the in-distribution test set is introduced. Without access to the OOD validation set, the sensitivity score is used to intuitively compare the effect of different negative sets and find optimal values by grid search for training hyperparameters. To calculate the score function, a cosine metric is used to measure the similarity between a given test sample and the training data.

In Chapter 7, the principles of the proposed method in Chapter 6 are used for the task of abnormality detection in medical images. Recent progress in anomaly detection of natural images has not been studied thoroughly in abnormality detection using medical images. Motivated by this, in Chapter 7, the general applicability of self-supervised anomaly detection using self-distillation and negative sampling is studied in the field of medical images. The method is applied for abnormality detection on three different types of medical datasets, including pneumonia, polyp, and glaucoma detection from X-ray, colonoscopy, and ophthalmology images. Note that similar to the approach used in Chapter 5, we only use normal samples of training data, and the model is not exposed to any abnormality during the training. For the score function, we compared both cosine similarity and Mahalanobis distance.

It is observed that the cosine similarity and Mahalanobis metric have different performances for different types of datasets. A more systematic evaluation and analysis of this difference can be a topic for future works. Moreover, it is observed that type and strength of applied transformation used for both generalization to in-distribution test data and to create the negative samples play an important role. The current transformations which are made by human knowledge and their strength are adjusted intuitively can limit the performance of learned representation. Thus the use of any automatic and self-performance feedback-oriented method can further improve the robustness of the features learned using SSL methods.

Chapter 9

Publications

2019

1. Linlin Zhao, Gereon Poschmann, Daniel Waldera-Lupa, Nima Rafiee, Markus Kollmann, Kai Stühler. OutCyte: a novel tool for predicting unconventional protein secretion. *Scientific reports*, 2019.
Contributions: Nima Rafiee contribute to this research by helping in model development and putting the model under production.
Status: Published in Scientific Reports volume 9, Article number: 19448, 19 December 2019, DOI:10.1038/s41598-019-55351-z

2022

2. Nima Rafiee, Rahil Gholamipoorfard, Nikolas Adaloglou, Simon Jaxy, Julius Ramakers, Markus Kollmann. Self-Supervised Anomaly Detection by Self-Distillation and Negative Sampling, 2022.
Contributions: Nima Rafiee contributed with research, training, evaluation, visualization and writing under the supervision of Prof. M. Kollmann.
Status: Published in ICANN 2022: 31st International Conference on Artificial Neural Networks, Bristol, UK, September 6–9, 2022, Proceedings; Part IV., DOI: 10.1007/978-3-031-15937-4_39

3. R. Gholamipoor, N. Rafiee, M. Kollmann. Pneumonia Detection with Semantic Similarity Scores. *ISBI*, 2022.
Contributions: The research and preparation of this manuscript were done jointly and contributed equally by R. Gholamipoor and N. Rafiee under the supervision of Prof. M. Kollmann.
Status: Published in 2022 IEEE 19th International Symposium on Biomedical

Imaging, 28-31 March 2022, (ISBI), DOI:10.1109/ISBI52829.2022.9761494

4. Nima Rafiee, Rahil Gholamipoor, Markus Kollmann. Abnormality detection for medical images using self-supervision and negative samples, 2022.
Contributions: The research and preparation of this manuscript were done jointly by Nima Rafiee and Rahil Gholamipoorfard under the supervision of Prof. M. Kollmann. **Status**: Submitted to *ISBI 2023*.

Bibliography

[1] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Prasadha, J. Pei, M. Y. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V. A. Huu, C. Wen, E. D. Zhang, C. L. Zhang, O. Li, X. Wang, M. A. Singer, X. Sun, J. Xu, A. Tafreshi, M. A. Lewis, H. Xia, and K. Zhang, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092867418301545

[2] J. Z. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax-Weiss, and B. Lakshminarayanan, "Simple and principled uncertainty estimation with deterministic deep learning via distance awareness," ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[3] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. A. DePristo, J. V. Dillon, and B. Lakshminarayanan, *Likelihood Ratios for Out-of-Distribution Detection*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[4] S. Jenni, H. Jin, and P. Favaro, "Steering self-supervised feature learning beyond local pixel statistics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6408–6417.

[5] K. P. Murphy, *Probabilistic Machine Learning An Introduction*. The MIT Press, 2022.

[6] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance-level discrimination," *CoRR*, vol. abs/1805.01978, 2018.

[7] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros, "Context encoders: Feature learning by inpainting," 2016.

[8] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1422–1430.

[9] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European conference on computer vision*. Springer, 2016, pp. 69–84.

[10] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6827–6839.

[11] G. Shih, C. C. Wu, S. Halabi, M. Kohli, L. Prevedello, T. Cook, A. Sharma, J. Amorosa, V. Arteaga, M. Galperin-Aizenberg, R. Gill, M. Godoy, S. Hobbs,

J. Jeudy, A. Laroia, P. Shah, D. Vummidi, K. Yaddanapudi, and A. Stein, "Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia." *Radiology. Artificial intelligence*, vol. 1 1, p. e180041, 2019.

[12] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[13] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[14] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip *et al.*, "Top 10 algorithms in data mining," *Knowledge and information systems*, vol. 14, no. 1, pp. 1–37, 2008.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[17] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." in *International Conference on Learning Representations*, 2019.

[18] 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*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[19] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014.

[20] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1747–1756.

[21] A. van den Oord, N. Kalchbrenner, L. Espeholt, k. kavukcuoglu, O. Vinyals, and A. Graves, "Conditional image generation with pixelcnn decoders," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper/2016/file/b1301141feffabac455e1f90a7de2054-Paper.pdf

[22] E. Çallı, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, "Deep learning for chest x-ray analysis: A survey," *Medical Image Analysis*, p. 102125, 2021.

[23] Z. Niu, K. Yu, and X. Wu, "Lstm-based vae-gan for time-series anomaly detection," *Sensors*, vol. 20, no. 13, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/13/3738

[24] Y. Tang, Y. Tang, M. Han, J. Xiao, and R. M. Summers, "Abnormal chest x-ray identification with generative adversarial one-class classifier," 2019.

[25] Y. Zhou, X. Liang, W. Zhang, L. Zhang, and X. Song, "Vae-based deep svdd for anomaly detection," *Neurocomputing*, vol. 453, pp. 131–140, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231221006470

[26] E. T. Nalisnick, A. Matsukawa, Y. W. Teh, D. Görür, and B. Lakshminarayanan, "Do deep generative models know what they don't know?" in *International Conference on Learning Representations*, 2019.

[27] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/d139db6a236200b21cc7f752979132d0-Paper.pdf

[28] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *Proceedings of International Conference on Learning Representations*, 2017.

[29] G. Mclachlan, "Mahalanobis distance," *Resonance*, vol. 4, pp. 20–26, 06 1999.

[30] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Curran Associates Inc., 2018, p. 7167–7177.

[31] R. Gholamipoor, N. Rafiee, and M. Kollmann, "Pneumonia detection with semantic similarity scores," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 2022, pp. 1–5.

[32] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *CoRR*, vol. abs/1412.6572, 2015.

[33] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 427–436.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[35] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[36] N. Yuval, W. Tao, C. Adam, B. Alessandro, W. Bo, and N. Andrew, "Svhn: Reading digits in natural images with unsupervised feature learning," 2011.

[37] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[38] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 1530–1538. [Online]. Available: https://proceedings.mlr.press/v37/rezende15.html

[39] R. Yao, C. Liu, L. Zhang, and P. Peng, "Unsupervised anomaly detection using variational auto-encoder based feature extraction," in *2019 IEEE International Conference on Prognostics and Health Management (ICPHM)*, 2019, pp. 1–7.

[40] D. Kim, H. Yang, M. Chung, S. Cho, H. Kim, M. Kim, K. Kim, and E. Kim, "Squeezed convolutional variational autoencoder for unsupervised anomaly detection in edge device industrial internet of things," in *2018 International Conference on Information and Computer Technologies (ICICT)*, 2018, pp. 67–71.

[41] A. Nanduri and L. Sherry, "Anomaly detection in aircraft data using recurrent neural networks (rnn)," *2016 Integrated Communications Navigation and Surveillance (ICNS)*, pp. 5C2–1–5C2–8, 2016.

[42] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan, "Hybrid models with deep and invertible features," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 4723–4732. [Online]. Available: https://proceedings.mlr.press/v97/nalisnick19b.html

[43] H.-J. Choi and E. Jang, "Generative ensembles for robust anomaly detection," *ArXiv*, vol. abs/1810.01392, 2018.

[44] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," 2017, cite arxiv:1708.07747Comment: Dataset is freely available at https://github.com/zalandoresearch/fashion-mnist Benchmark is available at http://fashion-mnist.s3-website.eu-central-1.amazonaws.com/. [Online]. Available: http://arxiv.org/abs/1708.07747

[45] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[46] J. Tack, S. Mo, J. Jeong, and J. Shin, "Csi: Novelty detection via contrastive learning on distributionally shifted instances," *Advances in neural information processing systems*, vol. 33, pp. 11 839–11 852, 2020.

[47] V. Sehwag, M. Chiang, and P. Mittal, "Ssd: A unified framework for self-supervised outlier detection," *arXiv preprint arXiv:2103.12051*, 2021.

[48] I. Golan and R. El-Yaniv, "Deep anomaly detection using geometric transformations," *Advances in neural information processing systems*, vol. 31, 2018.

[49] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for machine learning*. Cambridge University Press, 2020.

[50] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[51] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[52] L. Bottou, "On-line learning and stochastic approximations," in *In On-line Learning in Neural Networks*. Cambridge University Press, 1998, pp. 9–42.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[54] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015, pp. 730–734.

[55] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[57] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale." OpenReview.net, 2021.

[58] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers and ; distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 10 347–10 357. [Online]. Available: https://proceedings.mlr.press/v139/touvron21a.html

[59] A. P. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your vit? data, augmentation, and regularization in vision transformers," *Transactions on Machine Learning Research*, 2022. [Online]. Available: https://openreview.net/forum?id=4nPswr1KcP

[60] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[61] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[62] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," *arXiv preprint arXiv:2104.14294*, 2021.

[63] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[64] L. A. .Gatys, A. S. Ecker, and M. Bethge, "Texture and art with deep neural networks," *Current Opinion in Neurobiology*, vol. 46, pp. 178–186, 2017, computational Neuroscience.

[65] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," ser. ICML '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 1096–1103. [Online]. Available: https://doi.org/10.1145/1390156.1390294

[66] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018, cite arxiv:1810.04805Comment: 13 pages. [Online]. Available: http://arxiv.org/abs/1810.04805

[67] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[68] S. A. A. Ahmed, M. Awais, and J. Kittler, "Sit: Self-supervised vision transformer," *CoRR*, vol. abs/2104.03602, 2021. [Online]. Available: https://arxiv.org/abs/2104.03602

[69] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[70] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015.

[71] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, k. kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent - a new approach to self-supervised learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 21 271–21 284.

[72] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," vol. 33, pp. 9912–9924, 2020.

[73] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 15 750–15 758.

[74] T. Nakao, S. Hanaoka, Y. Nomura, M. Murata, T. Takenaga, S. Miki, T. Watadani, T. Yoshikawa, N. Hayashi, and O. Abe, "Unsupervised deep anomaly detection in chest radiographs," *Journal of Digital Imaging*, vol. 34, pp. 418–427, 2021.

[75] Y.-X. Tang, Y.-B. Tang, Y. Peng, K. Yan, M. Bagheri, B. A. Redd, C. J. Brandon, Z. Lu, M. Han, J. Xiao, and R. M. Summers, "Automated abnormality classification of chest radiographs using deep convolutional neural networks," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–8, 2020.

[76] Z. Li, C. Wang, M. Han, E. Xue, W. Wei, J. Li, and F.-F. Li, "Thoracic disease identification and localization with limited supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[77] S. Gündel, S. Grbic, B. Georgescu, S. K. Zhou, L. Ritschl, A. Meier, and D. Comaniciu, "Learning to recognize abnormalities in chest x-rays with location-aware dense networks," in *CIARP*, 2018.

[78] E. J. Hwang, S. Park, K.-N. Jin, J. Kim, S. Choi, J. Lee, J. M. Goo, J. Aum, J.-J. Yim, and C. M. Park, "Development and validation of a deep learning-based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs," *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, vol. 69, 11 2018.

[79] E. J. Hwang, S. Park, K.-N. Jin, J. I. Kim, S. Y. Choi, J. H. Lee, J. M. Goo, J. Aum, J.-J. Yim, J. G. Cohen, G. R. Ferretti, C. M. Park, for the DLAD Development, and E. Group, "Development and Validation of a Deep Learning–Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs," *JAMA Network*

Open, vol. 2, no. 3, pp. e191 095–e191 095, 03 2019. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2019.1095

[80] B. Bozorgtabar, D. Mahapatra, G. Vray, and J. Thiran, "SALAD: self-supervised aggregation learning for anomaly detection on x-rays," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part I*, ser. Lecture Notes in Computer Science, vol. 12261. Springer, 2020, pp. 468–478. [Online]. Available: https://doi.org/10.1007/978-3-030-59710-8_46

[81] Y. Mao, F. Xue, R. Wang, J. Zhang, W. Zheng, and H. Liu, "Abnormality detection in chest x-ray images using uncertainty prediction autoencoders," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part VI*, ser. Lecture Notes in Computer Science, vol. 12266. Springer, 2020, pp. 529–538. [Online]. Available: https://doi.org/10.1007/978-3-030-59725-2_51

[82] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[83] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[84] Y. Han, C. Chen, A. Tewfik, Y. Ding, and Y. Peng, "Pneumonia detection on chest x-ray using radiomic features and contrastive learning," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 247–251.

[85] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," 2020.

[86] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," 2020.

[87] J. Winkens, R. Bunel, A. G. Roy, R. Stanforth, V. Natarajan, J. R. Ledsam, P. MacWilliams, P. Kohli, A. Karthikesalingam, S. Kohl, A. T. Cemgil, S. M. A. Eslami, and O. Ronneberger, "Contrastive training for improved out-of-distribution detection," *CoRR*, vol. abs/2007.05566, 2020. [Online]. Available: https://arxiv.org/abs/2007.05566

[88] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[89] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," *Proceedings of the International Conference on Learning Representations*, 2019.

[90] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and

N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[91] X. Chen, C.-J. Hsieh, and B. Gong, "When vision transformers outperform resnets without pretraining or strong data augmentations," *arXiv preprint arXiv:2106.01548*, 2021.

[92] S. Mohseni, A. Vahdat, and J. Yadawa, "Multi-task transformation learning for robust out-of-distribution detection," *CoRR*, vol. abs/2106.03899, 2021. [Online]. Available: https://arxiv.org/abs/2106.03899

[93] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[94] L. Li, M. Xu, X. Wang, L. Jiang, and H. Liu, "Attention based glaucoma detection: a large-scale database and cnn model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 571–10 580.

[95] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen *et al.*, "Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *Scientific data*, vol. 7, no. 1, pp. 1–14, 2020.

[96] R. Gholamipoor, N. Rafiee, and M. Kollmann, "Pneumonia detection with semantic similarity scores," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 2022, pp. 1–5.

[97] S. Venkataramanan, K.-C. Peng, R. V. Singh, and A. Mahalanobis, "Attention guided anomaly localization in images," in *European Conference on Computer Vision*. Springer, 2020, pp. 485–503.

[98] Y. Chen, Y. Tian, G. Pang, and G. Carneiro, "Unsupervised anomaly detection and localisation with multi-scale interpolated gaussian descriptors," *arXiv e-prints*, pp. arXiv–2101, 2021.

[99] Y. Tian, G. Pang, F. Liu, Y. Chen, S. H. Shin, J. W. Verjans, R. Singh, and G. Carneiro, "Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 128–140.