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 its 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 develope 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 difficult 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. Zhang et al. [12] 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 ith 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] takes 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.

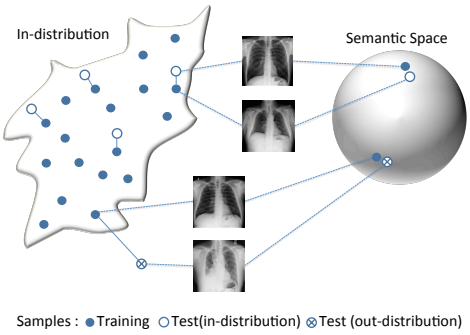

Fig. 1. Illustration of mapping examples from the in-distribution onto a unit-hypersphere. In this representation, feature vectors from the in-distribution are semantically similar if they approximately align and semantically diverse if they are separated by a large angle.

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. [13] shows 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 dominance of larger eigenvalues in euclidean distance metric as shown in [14] 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 [15] 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 [16]. 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.

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

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

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 [14]. 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 [17]	0.980	-	0.949
Pneumonai Detection using Radiomic Features[18]	0.923	-	-
ConVIRT [12]	-	-	0.908
Unsupervised methods trained on normal samples			
UAE[7]	0.89	0.78	0.83
Deep Anomaly Detection[19]	0.838	0.704	0.752
Generative Adversarial one-class classifier[5]	0.802	-	0.841
Ours	**0.940**	**0.828**	**0.866**

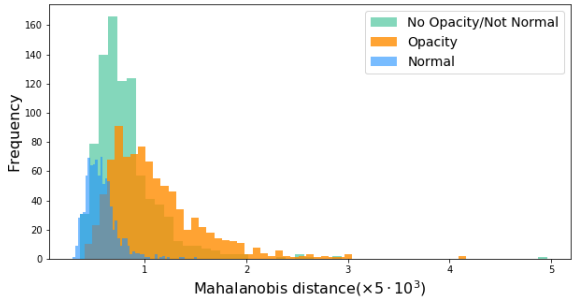

Fig. 2. Distributions over the anomaly detection score trained only on Normal samples and applied to the test sets of Normal as in-distribution, Opacity and "No Opacity/Not Normal" as out-distributions.

4.2. Pre-training and Label Efficiency of Multilabel Classification

In addition to self-supervised anomaly detection task, we evaluate the learned representaion by its performance in KNN accuracy and the impact it has on downstream task of multilable classification. For the pre-training task we use the same data split statistics as in [17] including 21152 training samples (14159 abnormal and 6993 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 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 [17]. To see the effect of self-supervised pre-training, we start with 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 AUROC score of 94.4% when fine-tuning with all labeled training data and AUROC score of 82.97% when using only 100 labeled samples which are selected randomly from training data.

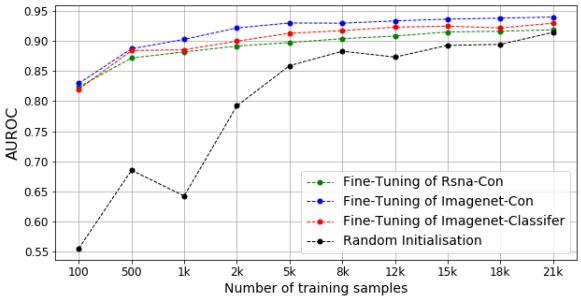

Fig. 3. Self-supervised pre-training increases the downstream classification task performance with small fraction of training samples. Rsna-Con and Imagenet-Con are finetunings of models with different model initialisation in self-supervised pre-training as follows: randomly initialised and initialised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

5. 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.

6. 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] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.