Performance or Trust? Why Not Both. Deep AUC Maximization with Self-Supervised Learning for COVID-19 Chest X-ray Classifications

Siyuan He — National Research Council Canada
Pengcheng Xi — National Research Council Canada
Ashkan Ebadi — National Research Council Canada
Stéphane Tremblay — National Research Council Canada
Alexander Wong — University of Waterloo
Email: {*Firstname.Lastname*}@nrc-cnrc.gc.ca; a28wong@uwaterloo.ca

Abstract

Effective representation learning is the key in improving model performance for medical image analysis. In training deep learning models, a compromise often has to be made between performance and trust, both of which are essential for medical applications. Moreover, models optimized with cross-entropy loss tend to be over-confident in incorrect predictions and under-confident in correct predictions. In this work, we integrate a new surrogate loss with self-supervised learning for computer-aided screening of COVID-19 patients using radiography images. In addition, we adopt a new quantification score to measure a model's trustworthiness. Ablation study is conducted for both the performance and the trust on feature learning methods and loss functions. Comparisons show that leveraging the new surrogate loss on self-supervised models can produce label-efficient networks that are both high-performing and trustworthy.

1 Introduction

COVID-19 continues to affect our daily lives. In the fight against the pandemic, computer-aided screening of patients using radiography images has served as a complementary approach to standard polymerase chain reaction (PCR) test. Recent research has been conducted for developing deep learning models with improved performance, given limited data from COVID-19 patients [1, 2]. Nevertheless, how much trust we have in the deep learning models remains another challenge [3].

In improving model performance, effective representation learning is the key and it can be realized through un-supervised training when data labels are missing or expensive to collect. A series of self-supervised models have achieved comparable performance to the supervised ones on benchmark data sets [4, 5]. They learn image representations through minimizing an embedding distance between image pairs derived from the same image, while maximizing the distance between the pairs from different images.

Regarding model trust, a deep classification neural network optimized on cross-entropy loss tends to be over-confident in its incorrect predictions and under-confident in correct predictions. In this work, we investigate a new surrogate loss function named deep AUC Maximization [6] and integrate it with a self-supervised model named MoCo [4]. To our best knowledge, this is the first integration of the loss function with self-supervised learning. In addition, we validate the models through quantitative comparisons to gain insight into the models' trustworthiness. Our assumption is that, by adopting the new surrogate loss function to self-supervised models, we no longer need to sacrifice model trust for performance but can achieve both.

Our contributions are threefold:

- We proposed the use of a new surrogate loss on self-supervised models to improve representation learning and maximize metrics pertinent to the task of screening COVID-19 patients.
- We showed that the use of a new surrogate loss can produce models that are more trustworthy than those optimized with cross-entropy loss.
- We provided case-by-case ablation studies of varying representation learning and loss functions to demonstrate the advantages of the newly adopted loss function.

2 Literature Review

Self-supervised learning has gained momentum in learning visual representations. It can be categorized into generative and discriminative approaches. As a discriminative method, Momentum Contrast (MoCo) trains a visual representation encoder by matching an encoded query to a dictionary of encoded keys through a contrastive loss [4]. The query encoder are shared with the key encoder, which gets slow updates in order to achieve consistency in learning visual representations.

In medical AI, contrastive learning has led to improved representation learning. In [7], a model named MoCo-CXR proved that linear models trained on MoCo-CXR-pretrained representations outperform those without MoCo-CXR-pretrained representations. Due to the scacity of COVID-19 patient data, the MoCo model has been applied to predicting patient deterioration based on chest X-rays [8].

Area under the Receiver Operating Characteristic curve (AUC) is widely used in medical image analysis for evaluating the performance of a neural network. Recently, Yuan *et al.* [6] proposed a novel surrogate loss over the standard cross-entropy loss to directly optimize for the AUC metric. AUC maximization, as the authors claim, can lead to the largest increase in a network's performance. This new surrogate loss function was integrated with supervised deep learning models and it achieved the first place in the Stanford CheXpert competition [6].

3 Methodology

3.1 Model Architecture

Our approach leverages deep AUC maximization [6], a novel surrogate loss proposed for medical image classification, with self-supervised pre-training to maximize label efficiency, performance, and model trust. In our experiments, we compare the loss function against traditional cross-entropy (CE) optimization on both self-supervised and supervised models. The self-supervised model is built on the MoCo framework [4] and it is pre-trained on MIMIC-CXR dataset [9]. All models are then fine-tuned on COVIDx dataset [10] for validations. DenseNet-121 is chosen as the backbone architecture throughout our experiments [11].

3.2 Datasets

The MoCo model has been pre-trained on the MIMIC-CXR dataset for predicting patient deterioration [8]. The dataset is composed of 377,110 chest radiographs [9]. As the dataset was constructed before the COVID-19 pandemic, it does not contain any positive chest X-ray samples of COVID-19.

Table 1: Data split for COVIDx8B

Split	Negative	Positive	Total
Train	13,794	2,158	15,952
Test	200	200	400

We perform end-to-end fine tuning on the COVIDx dataset [10]. The latest version COVIDx8B consists of 15,952 chest radiographs for training and 400 for testing (Table 1). Each sample in the dataset is labelled as either COVID-19 positive or negative. Stratified 5-fold cross validation is conducted on the training split during the fine-tuning stage to evaluate model performance.

3.3 Experiment Setups

The DenseNet-121 model pre-trained on the MIMIC-CXR using the MoCo framework has a projection dimension of 128, whereas the supervised model pre-trained on ImageNet has a projection dimension of 1,000. For end-to-end fine tuning, the parameters of the last fully connected layer of both pre-trained models are replaced and randomly initialized with a single output neuron for binary classification. We apply a sigmoid layer over the raw logits of the model to obtain a probability distribution. All input images are resized to 224x224, center cropped and normalized. Only random horizontal flipping is used for data augmentation as further augmentations were noted to provide little improvement for classification [8].

AUC Maximization. We adpot a novel surrogate loss function introduced by [6] to maximize the area under the Receiver Operating Characteristic curve. For end-to-end fine tuning, we use a learning rate of 0.1 for all layers of the DenseNet model. Then, we optimize the network with the new surrogate loss function to maximize the AUC metric. Lastly, we train for 30 epochs while decaying the learning rate by a factor of 10 at the 15th epoch.

CE Optimization. For standard end-to-end fine tuning, we set the learning rate at $1e-3$ for all layers of the DenseNet model. Following similar procedures in [8], we use cosine annealing learning rate decay to reduce the learning rate. Finally, we use the SGD optimizer on cross-entropy loss with a momentum of 0.9 and weight decay of $1e-4$ to fine tune the model for 30 epochs.

During each validation fold, we first compute an optimal threshold by maximizing F1-score on the validation split. Then, we save the model corresponding to the best validation accuracy. Finally, we evaluate the saved models on the unseen test split.

4 Experimental Results

4.1 Supervised vs. Self-Supervised Pre-training

We first examine the performance difference between traditional supervised pre-training on ImageNet and self-supervised contrastive pre-training on MIMIC-CXR. Tables 2 and 3 show significant improvements in the precision metric of the negative class and the sensitivity metric of the positive class for both CE optimization and AUC maximization. In medical image analysis, this improvement is key as maximizing the positive sensitivity score is necessary to lower false-negatives.

However, this increase in performance comes at the cost of model trust. We examine the trustworthiness of each model by calculating a trust score of the positive class. As per the procedures introduced in [3], we compute a score that rewards well-placed confidence and penalizes undeserved overconfidence. In Table 4, we notice that in the case of CE optimization, supervised models are drastically more trustworthy than self-supervised models. Moreover, throughout our CE optimization experiments, we observed that self-supervised models are less confident in its correct predictions (overcautious) than its supervised counterparts.

4.2 CE Optimization vs. AUC Maximization

Our comparisons of CE Optimization against AUC Maximization show improvements across standard metrics as well as overall model trust-worthiness. Both Table 2 and Table 3 show improvements in the precision and sensitivity metrics regarding the supervised models. Moreover, Fig. 1 demonstrates an increase in the AUC scores of the supervised models. When examining self-supervised models, AUC maximization still achieves comparable performance to CE optimization.

Furthermore, we observe significant gains in model trust scores, especially in the context of self-supervised models. Table 4 shows a nearly 1% increase in supervised pre-training and a nearly 6% increase in self-supervised pre-training. Moreover, when using AUC maximization, we do not see the same disparity in model trust between supervised and self-supervised models. Therefore, unlike CE optimization, we can freely leverage AUC maximization with self-supervised pre-training to improve performance without sacrificing model trust. As shown in Tables 2, 3 and 4, AUC maximization allows us to achieve top metrics without trading off model trust for performance.

Table 2: Precision scores on the unseen COVIDx8B test split. The best metric out of each optimization method is bolded. The best metric across methods is denoted by *.

Pre-trained Model	Negative	Positive
Supervised (CE Opt)	0.8960 ± 1.6%	0.9956 ± 0.4%
Self-Supervised (CE Opt)	**0.9295*** ± 1.6%	**0.9978** ± 0.4%
Supervised (AUC Max)	0.9134 ± 1.4%	1.000
Self-Supervised (AUC Max)	**0.9251** ± 0.6%	**1.000***

Table 3: Sensitivity scores on the unseen COVIDx8B test split. The best metric out of each optimization method is bolded. The best metric across methods is denoted by *.

Pre-trained Model	Negative	Positive
Supervised (CE Opt)	0.9960 ± 0.3%	0.8840 ± 2.1%
Self-Supervised (CE Opt)	**0.9980** ± 0.4%	**0.9240*** ± 1.9%
Supervised (AUC Max)	1.000	0.9050 ± 1.7%
Self-Supervised (AUC Max)	**1.000***	**0.9190** ± 0.7%

Table 4: Trust scores calculated from each experiment on the positive class. The best score overall is bolded.

Cost Function	Supervised	Self-Supervised
CE Opt	0.929 ± 0.9 %	0.879 ± 1.4 %
AUC Max	**0.938** ± 1.0 %	0.937 ± 1.0 %

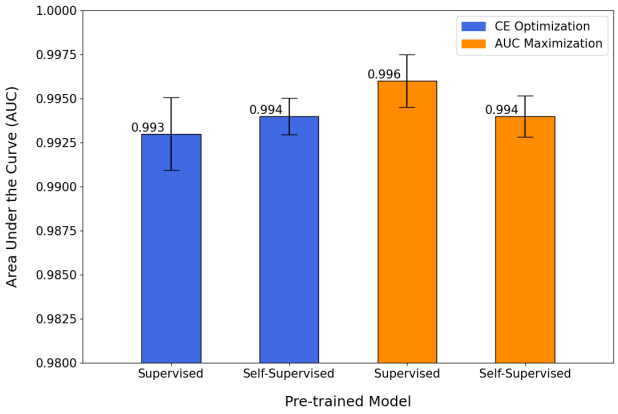

Fig. 1: Area under the Receiver Operating Characteristic curve. Error bars represent the standard deviation across cross-validation runs.

5 Conclusion

This work demonstrates that we no longer need to sacrifice model trust for performance. Integrating AUC maximization can produce more trustworthy and better performing models. By extending the AUC maximization paradigm [6] to self-supervised pre-training, we showed that we can significantly improve key metrics while also maintaining model trust.

We expect that our study on self-supervised learning with AUC maximization will contribute to the classification of both COVID-19 and future illnesses. More often than not, we cannot afford to collect large amount of labeled samples at the onset of a pandemic. Therefore, it is important that we exploit existing data, apply effective representation learning to maximizing model performance, and gain optimal model confidence.

References

[1] J. Zhang, P. Xi, A. Ebadi, H. Azimi, S. Tremblay, and A. Wong, "Covid-19 detection from chest x-ray images using imprinted weights approach," *ICLR 2021 Workshop: Machine Learning for Preventing and Combating Pandemics*, 2021.

[2] H. As' ad, H. Azmi, P. Xi, A. Ebadi, S. Tremblay, and A. Wong, "Covid-19 detection from chest x-ray images using deep convolutional neural networks with weights imprinting approach," *Journal of Computational Vision and Imaging Systems*, vol. 6, no. 1, pp. 1–3, 2020.

[3] A. Wong, X. Y. Wang, and A. Hryniowski, "How much can we really trust you? towards simple, interpretable trust quantification metrics for deep neural networks," 2021.

[4] 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 (CVPR)*, June 2020.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," 2020. [Online]. Available: https://arxiv.org/abs/2002.05709

[6] Z. Yuan, Y. Yan, M. Sonka, and T. Yang, "Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[7] H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar, "Moco pretraining improves representation and transferability of chest x-ray models," in *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*, ser. Proceedings of Machine Learning Research, M. Heinrich, Q. Dou, M. de Bruijne, J. Lellmann, A. Schläfer, and F. Ernst, Eds., vol. 143. PMLR, 07–09 Jul 2021, pp. 728–744. [Online]. Available: https://proceedings.mlr.press/v143/sowrirajan21a.html

[8] A. Sriram, M. Muckley, K. Sinha, F. Shamout, J. Pineau, K. Geras, L. Azour, Y. Aphinyanaphongs, N. Yakubova, and W. Moore, "Covid-19 prognosis via self-supervised representation learning and multi-image prediction," 2021.

[9] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, p. 317, Dec 2019. [Online]. Available: https://doi.org/10.1038/s41597-019-0322-0

[10] L. Wang, Z. Q. Lin, and A. Wong, "Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, p. 19549, Nov 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-76550-z

[11] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jul 2017, pp. 2261–2269. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.243