A Trustworthy Framework for Medical Image Analysis with Deep Learning

Kai Ma[a] (k78ma@uwaterloo.ca), **Siyuan He**[a], **Pengcheng Xi**[a,b], **Ashkan Ebadi**[a,b], **Stephane Tremblay**[b], **Alexander Wong**[a]
[a]Faculty of Engineering, University of Waterloo, Canada [b]National Research Council Canada

Introduction

Background

Computer vision and machine learning are playing an increasingly important role in **computer-assisted diagnosis**. While promising, the application of **deep learning** to medical imaging still faces challenges. These include:

- **Limited data** — Datasets for novel diseases are often very small, requiring transfer learning, which is dependent on label quality.
- **Class imbalance** — More negative (benign) data samples than positive (malignant) ones, resulting in biased models.
- **Low trustworthiness** — Trust is important in medical settings. Deep NNs optimized with the cross-entropy loss function tend to be overly cautious for the minority class, while being overconfident for the majority class.

Goal

The goal(s) of this study were to:

1. Use a combination of **supervised and self-supervised contrastive pre-training** to make transfer learning more efficient and less reliant on labels.
2. Combat data imbalance and trustworthiness problems (quantified by a trust score [1]) by using **Deep AUC Maximization** with AUC min-max margin loss.

Methodology

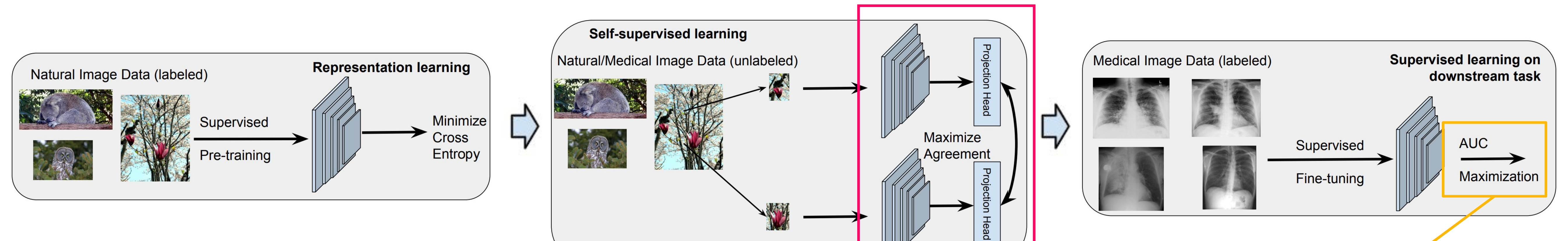


Figure 1. Overview of proposed framework

Dataset: COVIDx8B

Table 1: Data split for COVIDx 8B

Split	Negative	Positive	Total
Train	13,793	2,158	15,951
Test	200	200	400

- Chest radiograph (X-ray) dataset for COVID-19 binary classification
- Unseen test split
- Small and unbalanced

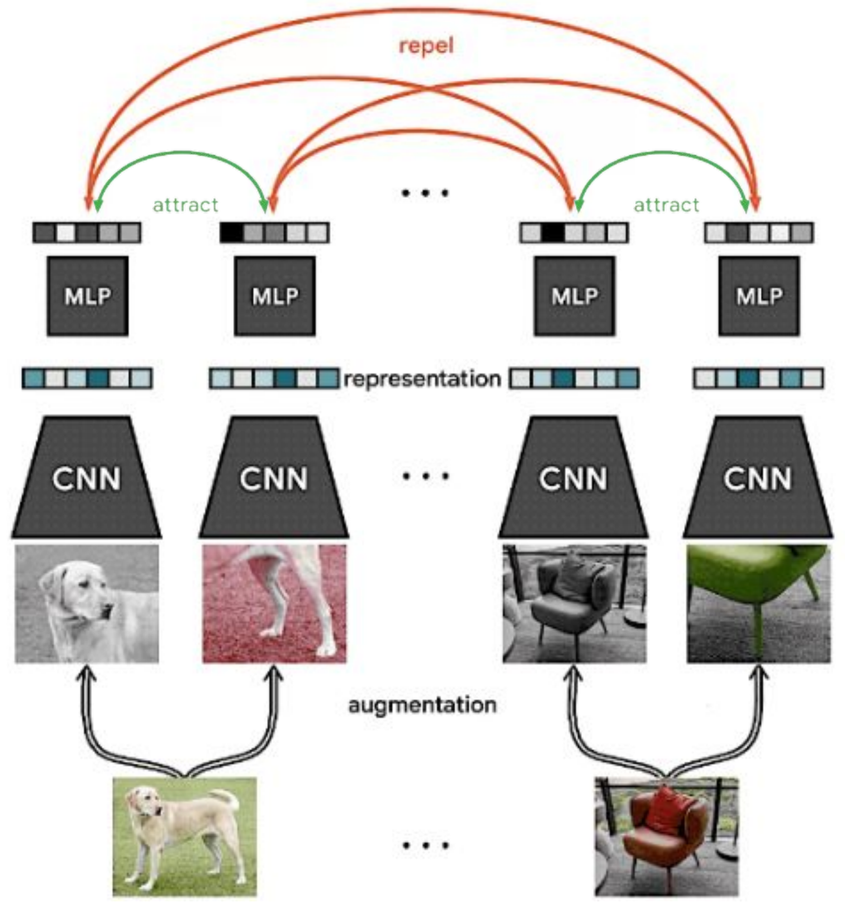
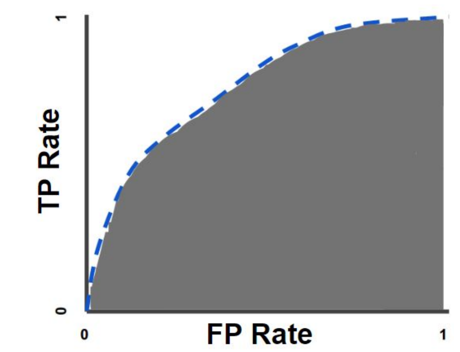


Figure 2. Contrastive SSL with SimCLR [2]



AUC score — Area under the ROC (receiver operating characteristic) curve

- We optimize with AUC instead of cross-entropy loss for better trust

- Naive pairwise loss based on definition of AUC score from scalability issues
- AUC margin loss is defined as:

Minimize variance of prediction scores on pos./neg. data

$$A_M(\mathbf{w}) = A_1(\mathbf{w}) + A_2(\mathbf{w}) + \max_{\alpha \geq 0} 2\alpha(m - a(\mathbf{w}) + b(\mathbf{w})) - \alpha^2$$

Tunable margin hyperparameter

Squared hinge function (widely used in margin-based SVM classifiers)

Figure 3. AUC maximization and AUC min-max margin loss[3]

Trust score computation: Given a question x, an answer y with respect to a model M, such that $y = M(x)$, and z representing the correct answer to x, we then use $R_{y=z \mid M}$ to denote the space of all questions where the answer $\mathbf{y}$ given by model M matches the correct answer z. Likewise, we use $R_{y \neq z \mid M}$ to denote the space of all questions where the answer $\mathbf{y}$ given by model does not match the correct answer. We also define the confidence of M in an answer y to question x as $C(y \mid x)$. Thus, *question-answer trust* of an answer y given by model M of a question x, with knowledge of the correct answer z, is defined as:

$$Q_z(x,y) = \begin{cases} C(y \mid x)^{\alpha}, & \text{if } x \in R_{y=z|M} \\ (1 - C(y \mid x))^{\beta}, & \text{if } x \in R_{y \neq z|M} \end{cases}$$

with α and β denoting reward and penalty relaxation coefficients.

Results

- Out-performs past SOTA models (custom architecture, designed for task)

Table 2: Model performance and trust scores for COVID-Net CXR-2, COVID-Net CXR-3 and COVID-CXR-SSL

Model	Precision		Sensitivity		Trust
	Pos.	Neg.	Pos.	Neg.	
COVID-Net CXR-2	0.970	0.955	0.956	0.970	-
COVID-Net CXR-3	0.990	0.975	0.975	0.990	-
COVID-CXR-SSL	**1.000**	**0.980**	**0.980**	**1.000**	**0.964**

- Ablation study shows improvement in performance & trust from each module

Table 3: Ablation study on model performance and trust scores for different model architectures

Architecture	Precision		Sensitivity		Trust
	Pos.	Neg.	Pos.	Neg.	
SL	1.000	0.885	0.870	1.000	0.918
SL+SSL	1.000	0.939	0.935	1.000	0.952
SL+SSL+AUC	**1.000**	**0.952**	**0.950**	**1.000**	**0.954**

- Generalization capabilities with other SSL model architectures and pre-training datasets

Table 4: Model performance and trust scores for different SSL plug-ins

SSL plugin	Precision		Sensitivity		Trust
	Pos.	Neg.	Pos.	Neg.	
MoCo (*MIMIC-CXR*)	0.995	0.896	0.884	0.995	0.909
MoCo (*ImageNet*)	0.998	0.934	0.930	0.998	0.937
SimCLR (*ImageNet*)	**1.000**	**0.952**	**0.950**	**1.000**	**0.954**

Conclusions

- We propose a **general deep learning framework** for medical image analysis which can be used to build high-performing, high-trust models;
- We show that fine-tuned models with **self-supervised pre-training** surpass supervised ones for COVID-19 classification, including state-of-the-art deep learning models designed specifically for the task;
- **AUC maximization** with margin loss leads to more effective feature learning and **higher trustworthiness**, effectively dealing with the problems of class imbalance and prediction under/over-confidence

Future Directions

- Validation with other datasets, including different diseases
- Exploring model explainability and interpretability
- Generative models to address data scarcity & imbalance issues
- Exploring different backbone architectures — Vision Transformers instead of CNNs

This work was supported by National Research Council Canada's Pandemic Response Challenge Program. It was also supported by the Natural Sciences and Engineering Research Council of Canada and Canada Research Chairs Program. The authors thank the COVID-Net team of University of Waterloo's VIP lab for their continued support.

[1] 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," CoRR, vol. abs/2009.05835, 2020. [Online]. Available: https://arxiv.org/abs/2009.05835
[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in International conference on machine learning 2020 Nov 21 (pp. 1597-1607). PMLR.
[3] 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 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021 pp. 3020-3029.