A high-performing explainable deep few-shot learning network for ultrasound COVID-19 detection

Jessy Song, Ashkan Ebadi, Adrian Florea, Pengcheng Xi, Alexander Wong

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Motivated by the lack of large number of well-annotated dataset during the onset of a novel disease, we present a **high-performing, interpretable few-shot learning network that detects positive COVID-19 cases with limited examples of ultrasound images**. Extensive experiments are conducted to evaluate model performance under different encoder architectures, number of training shots and classification problem complexity. When trained with only 5-shots, network classifies between positive and negative COVID-19 cases with 99.3% overall accuracy, 99.5% recall and 99.25% precision for positive cases. Network explainability is evaluated with two visual explanation tools (i.e., Grad-CAM [1] and GSInquire [2]) to ensure validity of network's decision-making process.

Our contribution is at least 2 folds:

- Network demonstrates ability to classify COVID-19 positive and negative cases with high performance when trained with only 5 shots
- Network accountability is assessed using an explainability module and validated by a practicing clinician.

Data and Methods

The COVIDx-US dataset v1.4. [3] is used for this study. The dataset contains 29,651 processed LUS images of 1) patients with COVID-19 infection, 2) non-COVID-19 infection, 3) other lung conditions, and 4) normal control cases. Dataset used for the study is prepared after filtering out of linear probe test, rescaling and augmentation by rotation. Figure 1 presents an overview of the analysis workflow.

The few-shot classification with a prototypical network involves three steps: 1) encoding of the images in a query set of M unlabelled images and a support set of N labelled images, 2) generating class prototypes from support sets, and 3) assigning labels to query samples based on the distance to the class prototypes. Figure 2 shows the network architecture.

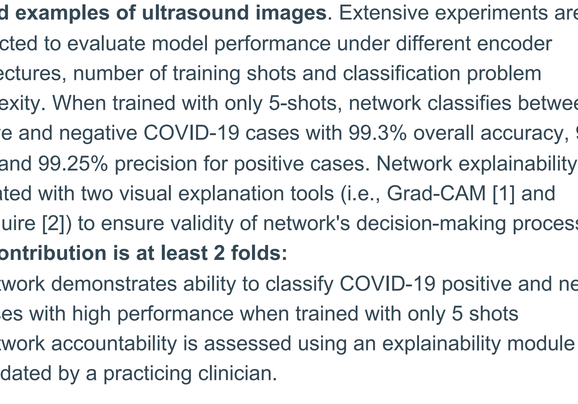

Figure 1: Study Analysis Overview

[1] Due to limit in space, not all encoder types or full results are presented.

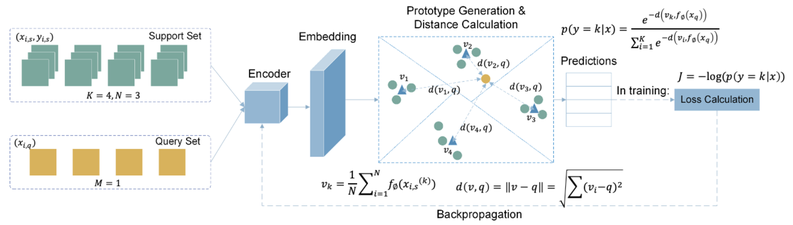

Figure 2: Network Architecture Overview

To comprehensively assess the performance of the proposed network in detecting COVID-19 positive cases from ultrasound images, we evaluated the impact of various training conditions such as (1) image encoders, (2) number of shots available for training, and (3) classification task types with increasing class numbers indicates increasing problem complexity.[1]

Experiment Parameters		Description
Models	1	Pre-trained ResNet 18 with ImageNet, with trainable parameters on the last 4 convolutional layers and final connected layer
	2	Pre-trained ResNet 50 with ImageNet, with trainable parameters on the last 3 convolutional layers and final connected layer
Training Shots		5, 10, 20, 30, 40, 50, 75, 100, 150, 200
Classification Formulation	2-way	Classify between negative/positive COVID-19 cases. Data from 'normal', 'non-COVID-19' and 'other' class, are combined as COVID-19 negative class.
	3-way	The 'other' class is excluded, as it contains data from multiple lung conditions which results in high in-class variations and may disrupt network's learning process due to the lack of uniformity in the data.
	4-way	Network is trained to classify 'COVID-19', 'normal', 'non-COVID-19' and 'other' class, same as the cases present in the original dataset.

Table 1: Experiment Description

Quantitative Results

		5-shot			100-shot		
Class #	Model	Accuracy	Precision	Recall	Accuracy	Precision	Recall
2-way	1	0.993	0.9925	0.995	1	1	1
	2	0.9965	0.9967	0.997	0.9999	0.9999	1
3-way	1	0.9987	0.9992	0.997	1	1	1
	2	0.9947	0.9942	0.994	1	1	1
4-way	1	0.9817	0.9975	0.997	0.9884	1	1
	2	0.985	0.9917	0.993	0.9902	1	1

Table 2: Experiment Results Summary

Across all classification types and models, performance metrics increases from 5-shot and plateaus after 75-shot. Both networks demonstrate the ability to classify COVID-19 with precision and recall above 98% consistently under both 5-shot and above 99% under 100-shot condition.

The increasing classes in 3-way and 4-way classification types reduces the performance of the network as problem complexity increases, but it is compensated as the number of shots increases, since more training examples improves class prototypes' representativeness.

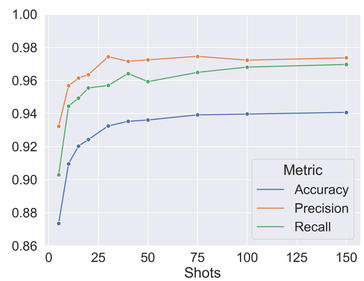

Figure 3: Model 1 binary classification performance under varied training shots

Qualitative Results

Our contributing clinician reviewed a randomly selected set of images to ensure network captures clinically relevant patterns. Figure 3 presents two Grad-CAM (identifies important areas with the gradients flowing into the last convolutional layer) [1] and GSInquire (identifies critical area with generative synthesis approach) [2] annotated COVID-19 ultrasound example images, enclosing two most common patterns deemed as critical for the final classification decision made by the network.

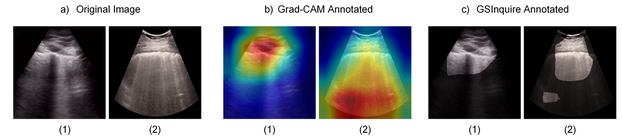

Figure 4: Two COVID-19 positive case examples correctly classified by the network with high confidence, a) original images, b) Grad-CAM annotated images and c) GSInquire annotated images.

The annotated images identified 1) the lung pleura region at the top (Figure 4(1)), and 2) the bottom region (Figure 4(2)) to have high importance. Clinically, the B-lines represented in Figure 4(2) more closely represent disease-related patterns within the ultrasound image. To improve model explainability, we experimented with removing the pleura region of the images by cropping so that network focuses on the disease-defining features at mostly the bottom of the images, and the strategy demonstrates initial improvements in annotations from both tools.

Limitations and Future Work

Several research directions can be explored to further improve the network.

1. Image augmentation and preparation techniques can be experimented to include linear probe data (excluded in this study).
2. In this work, we experimented with simple cropping to remove the pleura region of the images. A more procedural image segmentation step could be added to include only clinically relevant areas of the images for network construction to further improve network explainability.
3. We used a public dataset (i.e., COVIDx-US) that includes data of various sources and quality. Further filtering to include only high quality images with clinically-relevant features may improve network explainability.

References

[1] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.

[2] Z. Q. Lin, M. J. Shafiee, S. Bochkarev, M. S. Jules, X. Wang, and A. Wong, "Do explanations reflect decisions? A machine-centric strategy to quantify the performance of explainability algorithms," CoRR, vol. abs/1910.07387, 2019. [Online]. Available: http://arxiv.org/abs/1910.07387

[3] A. Ebadi, P. Xi, A. MacLean, A. Florea, S. Tremblay, S. Kohli, and A. Wong, "Covidx-us: An open-access benchmark dataset of ultrasound imaging data for ai-driven covid-19 analytics," Frontiers in Bioscience-Landmark, vol. 27, no. 7, 2022.