CS482/682 Final Project Report Group 16

COVID Defense Force: ML Division

Alexander Lu(alu27); Daijie Bao(dbao1); Jihoon Kim(jkim620); Xiaotong Hu(xhu66)

1 Introduction

Background Medical image data typically requires highly trained specialists for image interpretation and diagnosis. Clinical translation of deep learning frameworks for medical image classification and regression tasks that directly influence patient care may be limited by the inability of end users to understand the predictions that are made.

The STOIC 2021 challenge involves the prediction of COVID test positivity and disease severity at 1 month given an input chest CT scan. Explainable and interpretable neural networks can not only improve end user comfort with these traditionally black box systems, but also guide novel understanding of which features may be predictive of disease severity to improve clinician understanding of COVID-19 disease pathogenesis. Here, we explored the performance and potential for use of ProtoPNet for binary image classification of medical imaging data.

ProtoPNet and Related Work ProtoPNet has been shown to process images with comparable accuracy to SOTA models in computer vision while preserving model interpretability through a novel module that memorizes prototypical regions of input data, which are used to compute similarity scores against features extracted from a CNN before pooling and projecting with a linear layer to produce the classification output[1].

ProtoPNet has also been used for interpretable detection of COVID-19 from chest x-ray images[2]. Other work has also demonstrated additional extensions on ProtoPNet and applications to other medical imaging tasks [3, 4].

2 Methods

Dataset 2000 CT training volumes were obtained from the STOIC 2021 grand challenge. Due to prohibitive computational overhead, we first randomly sampled 400 volumes, maintaining the original class frequency. We then sampled axial slices from each volume, with each slice centered in the thorax. These 400 slices were then divided into training, validation, and hold-out testing sets (325, 25, 50 slices respectively).

Setup, Training and Evaluation We propose a binary classification task, predicting COVID positivity status given an axial slice. We compared an MLP model (baseline 1), a fine-tuned VGG-19 (baseline 2), and a ProtoPNet with VGG-19 backbone configured to learn 1000 prototypes per class. The MLP is a simple 3 layer network using one hidden state. The fine-tuned VGG-19 model features a custom classifier and is initialized identically to the backbone of ProtoPNet for baseline understanding of feature extraction performance. Hyperparameters were selected on the validation set, and evaluated on the hold-out test set using accuracy and F1 score. ProtoPNet was trained for 100 epochs on a Quadro RTX 6000. VGG model is obtained from Torchvision. ProtoPNet starter code obtained from public repository, which we extended and modified [1]. We qualitatively assessed interpretability using saliency maps for VGG-19 and and prototypes learned by the ProtoPNet.

3 Results

Baselines Upon 300 epochs of training, the MLP model achieves an accuracy of 49.6% on the test set,

and the fine-tuned VGG-19 Model reports 60% accuracy on the test set after 100 epochs of training. Saliency maps were generated on the test set (example in Figure 1).

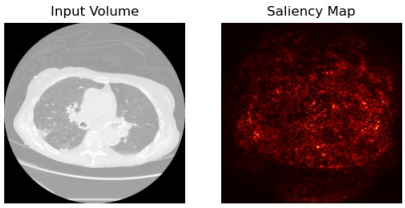

Figure 1: Example Saliency Map of the fine-tuned VGG-19 Model

ProtoPNet Training the ProtoPNet achieves only 60% accuracy on the test set, despite rounds of hyperparameter tuning and adjustments to the dataset. Examples of prototypes can be seen in Figure 2.

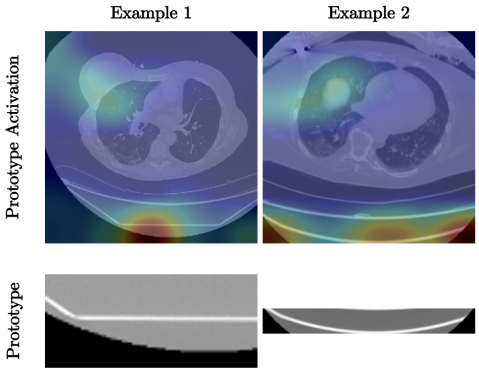

Figure 2: Examples of Prototype activations and Prototypes of ProtoPNet

4 Discussion

Comparing ProtoPNet against Baselines ProtoPNet achieved lower performance relative to the baselines on the testing set, despite numerous rounds of hyperparameter tuning. Given that our baseline model (Figure 1) shows saliency maps resembling features relating to patient anatomy, we believe difficulty fitting ProtoPNet arises from struggles in optimizing prototypes. Together with reports on the ProtoPNet repository's issues forum describing difficulty replicating the reported performance of ProtoPNet, we believe the complex multi-stage optimization required to tune the feature extraction, the prototype update, and the similarity score projection yield a poorly conditioned problem that is subject to numerous spurious minima. The result is difficulty fitting the ProtoPNet, resulting in overall mediocre performance.

Interpretability of ProtoPNet The prototypes learned by ProtoPNet should enable interpretation of network predictions. While the prototypes are indeed easily interpretable (attention maps with respect to the original input image + prototypical input image regions), we qualitatively observe redundancy between the 1000 learned prototypes, with some bearing no relevance to the task (e.g. prototypes frequently latched on air around the patient, the table under the patient, etc.). While this could arise due to many factors, the consistency with which the model optimization arrived at these spurious prototypes suggests poor conditioning in the optimization process and difficulty extending ProtoPNet to new tasks. Although we suspect that learning prototypes that align with prior expectations of task-critical image content should lead to strong task performance, we were unable to tune the network to learn such prototypes.

Further, single axial slices may contain insufficient information to distinguish COVID from non-COVID patients, evidenced by our low baseline performance. Though this is certainly an issue, ProtoPNet performance inferior to the baseline demonstrates that the ProtoPNet model is also part of the problem.

Future Directions Though ProtoPNet demonstrates strong potential for application to medical images, we could not reproduce its ability to identify meaningful prototypes. In the future, it would be interesting to explore mechanisms through which explicit prior knowledge could be directly introduced, either as initialization of the prototypes or conditioning the prototypes on a prior distribution of domain knowledge.

5 Reference

[1] Chen C, Li O, Tao D, et al. This looks like that: deep learning for interpretable image recognition. Advances in neural information processing systems, 2019, 32.

[2] Singh, G. and Yow, K. C. (2021). An Interpretable Deep Learning Model for Covid-19 Detection With Chest X-Ray Images. IEEE access : practical innovations, open solutions, 9, 85198–85208. https://doi.org/10.1109/ACCESS.2021.3087583

[3] Donnelly, Barnett, and Chan. Deformable ProtoPNet: An Interpretable Image Classifier Using Deformable Prototypes. https://arxiv.org/abs/2111.15000

[4] Mohammadjafari et al. Using ProtoPNet for Interpretable Alzheimer's Disease Classification. https://assets.pubpub.org/gzl17h3e/21624570427985.pdf