Improving Explainability of Image Classification in Scenarios with Class Overlap: Application to COVID-19 and Pneumonia

Edward Verenich‡ *, Alvaro Velasquez*, Nazar Khan†, Faraz Hussain‡
‡Clarkson University, Potsdam, NY
{verenie, fhussain}@clarkson.edu
†Punjab University College of Information Technology
nazarkhan@pucit.edu.pk
*Air Force Research Laboratory, Rome, NY
{edward.verenich.2, alvaro.velasquez.1}@us.af.mil

Abstract—Trust in predictions made by machine learning models is increased if the model generalizes well on previously unseen samples and when inference is accompanied by cogent explanations of the reasoning behind predictions. In the image classification domain, generalization can be assessed through accuracy, sensitivity, and specificity. Explainability can be assessed by how well the model localizes the object of interest within an image. However, both generalization and explainability through localization are degraded in scenarios with significant overlap between classes. We propose a method based on binary expert networks that enhances the explainability of image classifications through better localization by mitigating the model uncertainty induced by class overlap. Our technique performs discriminative localization on images that contain features with significant class overlap, without explicitly training for localization. Our method is particularly promising in real-world class overlap scenarios, such as COVID-19 and pneumonia, where expertly labeled data for localization is not readily available. This can be useful for early, rapid, and trustworthy screening for COVID-19.

Index Terms—explainability, trust in AI, class overlap, COVID-19, data-starved, deep learning, object localization.

I. INTRODUCTION

The use of deep neural networks for image classification and object detection in imagery is well established in the computer vision domain. As neural networks became increasingly used in real world applications, such as assisting medical diagnosis, the phenomenon of class overlap became more apparent [1]. Recent work on detecting COVID-19 using X-ray imagery has also shown that class overlap degrades classifier performance [2]. By training convolutional neural networks (CNNs) to account for classes with similar conditions, the model becomes less certain. This is in part due to overlap in class activations triggered by the same image. This paper presents a new technique to distinguish between COVID-19 and regular pneumonia in X-ray imagery in a more explainable fashion by using class activation maps.

The standard approach in deep learning to reduce uncertainty is to provide more training data to the model, which is not always possible and primarily addresses model uncertainty that is due to model parameters. Another method to reduce decision uncertainty is to localize target objects, hence increasing confidence in the prediction. Localization through supervised training with labeled bounding boxes to compute the reward is a widely used approach for reducing decision uncertainty in image classification [3]. However, the lack of labeled data and inherent noise present in novel situations results in additional predictive uncertainty. Consider, for example, X-ray imagery of confirmed COVID-19 patients, where X-ray images were taken to analyze pulmonary complications, yet expert localization of COVID-19 specific attributes was not performed by radiologists, i.e. no bounding boxes on X-ray images of COVID-19 relevant regions were annotated [4]–[8].

Ghoshal et al. [9] note that there are two distinct kinds of predictive uncertainty in deep learning. First, epistemic uncertainty, or uncertainty in model parameters, which decreases with more training data. Second, aleatoric uncertainty, that accounts for noise in observations due to class overlap, label noise, and varying error term size across values of an independent variable. Aleatoric uncertainty cannot be easily reduced by increasing the size of the training set.
Fig. 2: High-level view of our dual-network technique for handling class overlap. Given a class of interest $C_1$ and another possibly overlapping class $C_2$, our approach involves training two separate binary expert networks ($N_1$, $N_2$). Each input image ($I$) is fed to both expert networks to obtain class activation maps ($CAM_1$, $CAM_2$) which are then used by our directed kernel function $K$ to localize regions in $I$ where the expert network for the class of interest ($N_1$) is more confident.

At the heart of our approach is the use of class activation maps (CAMs) for improved localization of the regions responsible for the image being in a specific class (e.g. COVID-19) as opposed to some other overlapping class (pneumonia). As an example, Figure 1 shows activation maps for two separate models trained for classifying carwheels and cars, respectively. The figure depicts significant overlap in the regions responsible for categorizing the image in the two classes. Our goal is localizing regions in the image which are more responsible for its classification in a specific class of interest (carwheel), in particular where bounding boxes are not available during training.

In unexpected public health emergencies, such as the coronavirus pandemic, labeled datasets with bounding box annotations are unlikely to be available to the community at an early stage. Such scenarios preclude the possibility of training a model for localization. In such situations, is it possible to improve localization when there is no way to train for it?

This work proposes a method for improved localization in order to enhance explainability of image classifications in data regimes with significant class overlap, thus mitigating aleatoric uncertainty. Our results show that training per-class binary CNN models and applying our new kernel function on their class activation maps can extract and better localize objects from overlapping classes.

### II. Related Work

Image classification and object localization have been successfully utilized for diagnostic purposes in radiology for pneumonia detection using chest X-rays [10]. With the recent emergence of the COVID-19, a number of methods and models have been proposed, as surveyed by Shi et al. [11], to detect the disease using medical imaging. Wang and Wong [4] released early work on using convolutional neural networks for COVID-19 detection from X-ray images. Alqudah et al. [12] used convolutional neural networks to classify X-ray images as well as extract features and pass them to other classifiers. Ghoshal et al. [9] observed that most methods focused exclusively on increasing accuracy without accounting for uncertainty in the decision and proposed a method to estimate decision uncertainty. Zhou et al. [13] showed that discriminative localization is possible without explicitly training for object detection with labeled object bounding boxes. However, noise related to overlapping classes has not been considered in these works. This work builds on these ideas and extends the work of Zhou et al. [13] on discriminative localization to specifically address the problem of overlapping classes and explainability in image classification.

### III. Approach

Our goal is to localize image regions responsible for a class of interest (e.g. COVID-19), given possibly overlapping classes (e.g. pneumonia/COVID-19). Our method consists of training locally independent expert networks as binary classifiers for two different classes that are possibly overlapping, e.g. classifiers for COVID-19/No-COVID-19 and pneumonia/No-pneumonia. These binary expert networks are then leveraged as expert classifiers on each input image as part of a dual-network architecture as shown in Figure 2. CAMs obtained from the two networks are then passed to our novel kernel function ($K$) that localizes image regions responsible for the class of interest. Note that, in our approach, a binary expert network is a classifier that classifies its input as either in-class or out-of-class for a specific class e.g. a COVID-19/No-COVID-19 classifier is a binary expert network for COVID-19. Similarly, a Pneumonia/No-Pneumonia classifier is an expert network for Pneumonia. However, we do not consider a COVID-19/Pneumonia classifier as a binary expert network.
The activation values in the Activation layer contain 2048 when a forward pass is performed on a network for inference. A hook to the Activation layer to store its activation values before the Global Average Pooling layer. We add the Activation layer, in case of our ResNet based networks it is the final convolutional layer.

Requirements, utilizing a Global Average Pooling layer after to obtain a class score. A ResNet architecture meets these pooling layer (average or max), and a fully connected layer network architecture needs an activation layer, followed by a classification.

Class activation maps allow us to localize objects of a given class by mapping regions of an image to the most active values in the activation layer of a network. To obtain CAMs of each of the binary expert classifiers we follow the approach described in the activation layer of a network. To obtain CAMs of each of an expert network and locations of activation (f) and weight (w) tensors, respectively. Finally we compute a weighted sum using activation values and weights for a predicted class as in Equation 1 to produce a 7x7 tensor that represents a CAM for the predicted class c.

\[ CAM_c = w_1^c \cdot f_1 + w_2^c \cdot f_2 + ... + w_{2048}^c \cdot f_{2048} \]  

In order to identify regions in the image that are important to the predicted class, we superimpose the 7x7 CAM on the original image using bilinear sampling to scale the CAM to appropriate size, which in this case is 224x224.

**B. Amplified Directed Divergence Kernel**

In order to extract overlapping features between two classes, we need a directed divergence or difference measure. For tensors (x, x') we need a measure that will amplify only positive differences in activation values in (x – x'), because we are interested in recovering features in tensor x with higher values than tensor x', but not vice versa i.e. \( K(x, x') \neq K(x', x) \). We introduce a kernel method called Amplified Directed Divergence Kernel (ADDK) that accepts two tensors (x, x') of equal shape and returns another tensor of the same shape with amplified positive differences of (x – x') as shown in Equation 2. The kernel method ensures that a maximum value of a given tensor is not zero in the normalization step. Normalization with maximum tensor values has shown promising empirical results, but we plan to explore other normalization techniques in the future.
Fig. 4: Results of our approach applied to a natural image dataset with overlapping classes. The class of interest is carwheel and car is the overlapping class. Each row contains triples of images with class activation maps superimposed on the original image. The first two images in each set show activation maps from the expert binary models, and the third image shows the output of our directed kernel function that localizes features relevant to the class of interest. The amplification parameter $\alpha$ was 5.

$$K(x, x') = \exp(\alpha(x/\max(x) - x'/\max(x'))))$$  \hspace{1cm} (2)

The parameter $\alpha$ controls amplification of directed differences where higher amplification will concentrate the resulting heat map to a smaller region. To illustrate the kernel function operation, a simplified example with $\alpha = 15$ is shown in Equation 3.

$$K \begin{pmatrix} 1 & 1 & 5 \\ 0 & 6 & 4 \\ 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 8 & 0 & 7 \\ 1 & 4 & 3 \\ 1 & 2 & 1 \end{pmatrix} = \begin{pmatrix} .0 & 12.2 & .5 \\ .2 & 1808 & 79.8 \\ .2 & .3 & .2 \end{pmatrix}$$ \hspace{1cm} (3)

Tensors $x$ and $x'$ in $K(x, x')$ represent CAM outputs from respective binary expert models on the same image. Tensor sizes have been reduced and the result rounded for clarity.

IV. EXPERIMENTS

We apply our proposed dual-network technique to localize regions indicating COVID-19 in X-ray imagery. However, due to the absence of localized and labeled bounding boxes for COVID-19 in X-rays, the computed localizations cannot be easily validated. Therefore, we also tested our technique on a natural imagery dataset for which the localizations can be visually validated.

To train the expert binary models, we utilized transfer learning to mitigate the problem of training a robust image classifier with a small number of training samples from a novel class of interest. We used a pretrained ResNet-152 architecture [14] and replaced the final connected layer with appropriate classes and fine-tuned it with new data. Stochastic gradient descent was used with a learning rate of 0.001 and momentum of 0.9. Training was performed for 30 epochs and the best performing model based on validation accuracy was selected.

A. Natural imagery

We selected two categories of images with significant class overlap, viz. carwheel and car, the former being the class of interest. We fine-tuned two CNNs pretrained on ImageNet as binary experts for car and carwheel. The models were
trained solely for classification and not for object localization. Furthermore, the carwheel expert model was fine-tuned with only thirteen images in order to simulate an environment with a novel class of interest that is likely to be data-starved (e.g. COVID-19). The CAMs obtained from these two experts were passed through our novel kernel function to obtain a heat map that localizes regions in the image where the expert network for carwheel was more confident.

Figure 4 shows results consisting of sets of three images; the first two are the CAMs obtained from the carwheel and car expert networks respectively, and the third is the heat map computed using our kernel function that significantly improves the localization of the class of interest (carwheel). This enhances the explainability of the classification decision in this class overlap scenario.

B. Medical imagery

We utilized the COVID-19 chest X-ray dataset and extracted COVID-19 samples with the posteroanterior view of the X-ray. The dataset by Kermany et. al was used for the Pneumonia X-ray images, which are also available from Kaggle. Both data sets were processed into training, validation, and test splits using the 60/20/20 ratio. Table I shows the dataset sizes during training, validation, and testing.

Figure 5 shows triples of X-ray images with superimposed class activation maps for predictions obtained from expert binary models (images one and two) with the third image showing the heat maps computed using our kernel. The intended use of our method is to examine positive classifications from two possibly overlapping classes (i.e. COVID-19, Pneumonia) and extract discriminative features pertaining to the class of interest, i.e. COVID-19. Triples (a)-(f) show positive classifications of COVID-19 and Pneumonia by their respective binary expert

|                         | Train+Val | Test  |
|-------------------------|-----------|-------|
| COVID-19                | 79/140    | 27/47 |
| Pneumonia              | 2563/949  | 856/317 |
Our work improves the explainability of classification decisions in scenarios with overlapping classes. We do this by training more confident models on simpler binary classification problems. Our approach uses these simpler binary models for enhancing explainability by improved localization, without training for localization.

We believe that our technique can be extended towards other useful applications. An example is monitoring training progress of classification models on data without localization ground truths, while having subject matter experts assess whether the model is discriminating proper or expected regions based on class. This is useful to assess that the model is learning some causal relationship between data and class rather than some spurious correlation induced by data artifacts, i.e. certain classes have some artificial mark produced by the collection process. Finally, by pairing subsequent evolutions of the same model, i.e. continuous retraining on new data, our method can extract shifts in activation maps induced by retraining and hence detect model drift or covariate shift of the data.

We believe that our technique is promising in addressing uncertainty related to noisy data and further development will enable its use in numerous applications. One direction of future work is to investigate image upsampling methods in order to better map class activations to original imagery. Whether variations of our kernel function can help improve explainability by better localization can also be explored. More experiments are needed to observe the effects of transformations used during training of expert models. For example, some domains such as natural images benefit from random geometric variations during training, while others, i.e. posterioranterior X-ray imagery, do not, as subject positioning is relatively constant between data points. Finally, the effects of training expert models from scratch instead of utilizing pretrained models that are fine-tuned can also be explored.
REFERENCES

[1] R. Alejo, R. M. Valdovinos, V. García, and J. H. Pacheco-Sanchez, “A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios,” Pattern Recognition Letters, vol. 34, no. 4, pp. 380–388, 2013.

[2] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. Rajendra Acharya, “Automated detection of covid-19 cases using deep neural networks with x-ray images,” Computers in Biology and Medicine, vol. 121, p. 103792, 2020.

[3] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, “Accurate object localization in remote sensing images based on convolutional neural networks,” IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 5, pp. 2486–2498, 2017.

[4] L. Wang and A. Wong, “Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images,” arXiv preprint arXiv:2003.09871, 2020.

[5] J. P. Cohen, P. Morrison, and L. Dao, “Covid-19 image data collection,” arXiv 2003.11597, 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset

[6] https://github.com/aghung/figure1-COVID-chestxray-dataset figure 1 COVID-19 Chest X-ray Dataset Initiative.

[7] https://github.com/aghung/Actualmed-COVID-chestxray-dataset actualmed COVID-19 Chest X-ray Dataset Initiative.

[8] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. A. Emadi, and et al., “Can ai help in screening viral and covid-19 pneumonia?” IEEE Access, vol. 8, p. 132665132676, 2020.

[9] B. Ghoshal and A. Tucker, “Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection,” ArXiv, vol. abs/2003.10769, 2020.

[10] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Y. Ding, A. Bagul, C. Langlotz, K. S. Shpanskaya, M. P. Lungren, and A. Y. Ng, “Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning,” CoRR, vol. abs/1711.05225, 2017.

[11] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, and Y. Shi, “Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19,” IEEE Reviews in Biomedical Engineering, vol. PP, pp. 1–1, 04 2020.

[12] A. M. Alqudah, S. Qazan, and A. Alqudah, “Automated systems for detection of covid-19 using chest x-ray images and lightweight convolutional neural networks,” Research Square, 2020.

[13] B. Zhou, A. Kholas, A. Lapadruza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.

[14] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, pp. 770–778.

[15] D. Kermany, K. Zhang, and M. Goldbaum, “Labeled optical coherence tomography(oct) and chest x-ray images for classification,” Cell, 2018.