Classification of Chest X-Rays Using Learning to Defer via Analysis of GradCAM Maps

Brad Levin BALEVIN@MIT.EDU
Massachusetts Institute of Technology

Paul Ruh PAULRUH@MIT.EDU
Massachusetts Institute of Technology

Claire Yin YINC@MIT.EDU
Massachusetts Institute of Technology

Touchpoint Summary

After our preliminary project proposal, we gathered more information about the full scope of our previous project and reached many obstacles regarding computational complexity and project entailment. Specifically, the computational needs did not align with our expectations, and given the complex nature of the network architecture of Liu et al, we anticipated many obstacles in reimplementing this model. With the time constraint of less than two weeks till the project poster session, we did not think this initial project would give us enough time to explore and implement our own ideas. Thus, we as a group decided it would be best to switch gears.

We have rerouted our project to tackle a new problem on the same dataset and staying in the same realm of healthcare. Now, we are implementing a system that involves two main components: a first classifier that predicts diseases from chest radiographs, and a second classifier that takes in the GradCAM and prediction of the first classifier to learn to defer. By leveraging GradCAM mapping information, we hope to provide a system that will provide optimal accuracy for chest disease classification by learning when classifiers fail to provide correct diagnoses. Given the extreme consequences of medical misdiagnosis, this is crucial to having a reliable diagnostic system. The big picture idea is that such a system can then be more effectively implemented in the clinical diagnosis process.

For the first model, the image classifier, we will use a pretrained DenseNet. For the GradCAM error model, we will use a deep neural network. During the training process, we will use CheXpert labels as our ground truth values. Our evaluation process will compare the two models' performances with other baseline models, computing AUC and accuracy of the image classifier both without deferral and with deferral.

1. Introduction

This project will focus on the classification of the presence of Pleural Effusion, Pneumothorax, and Edema in chest X-rays. These are three conditions that are commonly identified in chest X-rays. Our goal will be to improve the overall process of diagnosing these patients through a learning to defer model. This is important because while classification of X-rays is an important task, it is one that can be made more efficient by relying on doctors only when necessary. Our aim is for our classifications to achieve a similar accuracy to that of a radiologist, and recognizing the instances in which it is recommended to trust the clinician's diagnoses to the model's. This process poses a significant challenge of a model adaptively working with the decision maker. We argue that the process of analyzing chest X-ray radiographs is one where the model and clinician operate simultaneously such that the model only predicts if its predictions are accurate to an acceptable degree.

Our thought is that a model's ability to correctly classify an image can be traced back to how it decides on the class to predict. Through this line of thinking we came up with our project of using GRAD-CAM data (which uses gradient information in the final layer of a CNN to determine relative importance of different regions in the image) to determine whether to use our prediction or defer to a clinician. We believe that a CNN is more likely to produce the correct class when the classification of an image is affected by a concentrated region as opposed to being very spread out throughout the image.

Technical Significance The steps for our work are broken down into three sections. The first step uses convolutional neural networks (CNN) to determine which of the possible diseases, if any is present in the image. Then we will use Gradient Class Activation Maps (GradCAMS) to inspect the network outputs and determine which regions of the convolutional stack at the final convolutional layer were the most informative in the prediction. The final section takes the GradCAM output for an image and returns the probability that that image was classified correctly. The decision of whether or not to defer will be made by comparing the likelihood of an incorrect prediction and the cost of deferring. The cost of deferring will be determined with help from our clinical mentors.

Clinical Relevance This work focuses on generating an accurate prediction of the presence of a disease from a chest X-ray image. This task has been explored before, but previous works have mostly focused on generating legible or accurate reports without consideration of how the tool can be used in tandem with a radiologist Rubin et al. (2018). Radiologists analyze a large volume of chest x-rays so automation would decrease their workload and time spent on radiology reports to allow them to focus on other parts of their job. Ideally this predictive model will output an accurate prediction in the areas where radiologists struggle and defer to their expertise in situations where the clinician is most likely to be correct.

We have reached out to our clinical mentors to become aware of the situations in which our model can be most useful. We will continue communication with them to assure that learning to defer allows the model to work with the radiologists to make the best predictions which is extremely important when stakes are the life or death of a patient.

2. Datasets

Widely used in machine learning applications to healthcare, including NLP and image understanding, the MIMIC-CXR dataset contains over 300,000 chest X-rays paired with free-text radiology reports. Radiographs in this dataset contain chests with chronic cardiopulmonary conditions and treatment devices (pacemakers, tubes, etc.) that are correctly positioned. This is the largest dataset available that has reports and images paired. We are using a modified version of this dataset which contains structured labels made from running Chexpert Irvin et al. (2019) to determine which conditions are present in the image Johnson et al. (2019).

3. Methods

Our goal is to learn to identify the most clinically accurate labels as possible, and defer to a clinicians expertise when necessary. To achieve this we build and train two separate models, an image classification model, and an expected loss model.

Image Classifier

We will use a Convolutional Neural Network to determine which of our three diseases are present. To train this model we will start with PyTorch's pretrained Densenet. We will use transfer learning on this model to fine tune this model to our chest x-rays and labels, an idea implemented for detecting COVID-19 in chest x-rays in Punn and Agarwal (2020). This was chosen because this architecture is known to perform well for image classification and this reduces the amount of training that has to be done Iandola et al. (2014). Furthermore, it performs just as well as ResNet, obtaining the same accuracy while requiring fewer parameters Iandola et al. (2014); He et al. (2016). We will modify this architecture so the last layer has 3 sigmoid outputs to produce a probability for our three conditions (Pleural Effusion, Pneumothorax, Edema). We do this instead of a softmax output because multiple conditions can be present in one image.

In the training process, we will implement batch normalization and dropout, techniques available through DenseNet-121's implementation, to prevent overfitting. Our loss function will be a cross-entropy loss.

Once this prediction is made we will produce a saliency map via an implementation of GRAD-CAM Selvaraju et al. (2017). This output, along with our prediction for this image will be stored for later use.

GRAD-CAM Error Model

Once our image classifier is trained we will start a second model. This model will be a deep neural network, taking in the GRAD-CAM output for one image and outputting a probability of classifying the conditions in that image correctly. The input for training this model will be created from GRAD-CAM outputs from a separate set of images that our image classifier was not trained on. With the labels for these data being whether our generated predictions matched the ground truth labels taken from Chexpert.

Through working with Dr. Nhi Vo we will develop a cost for letting a doctor handle the diagnosis (a number $\in (0, 0.5)$). This will be based on time spent on the diagnoses as well as the probability of a clinician misdiagnosing the patient (we have reached out and are waiting for a response). The full pipeline will work as follows:

1. An image will be passed through our image classifier which will decide on a prediction of what conditions are present (not to be shared with the user yet), and produce GRAD-CAM data for this prediction.

2. This GRAD-CAM output will be passed into our second model to determine a probability of our initial prediction being correct.

3. If the probability we are incorrect is greater than our defined cost of deferring, then we will not make a prediction and allow a clinician to look at the image. Otherwise we will output our predicted labels.

4. Evaluation Approach

We will split our MIMIC-CXR data into two train, test, and validation sets with 7/1.5/1.5 ratio splits, maintaining each patient's images within one set. This will ensure no overlap between the three sets. We will also have separate train sets for each of the models to prevent the second model from learning from images that the first model was trained on, which the first classifier is likely to perform better on and label more images correctly.

As aforementioned, we will be using CheXpert labels as ground truths when evaluating our models. The metrics we will use are accuracy and AUC. We will compare our first model's performance with baseline models, such as a dual CNN and a simple deep learning model, PCANet Chan et al. (2015); Rubin et al. (2018).

In order to evaluate our second model and our overall L2D system, we will compute the accuracy and AUC on the subset of data that has been classified as accurate by the second model. When using these metrics, we need to take into account the fact that false negatives can be detrimental as missing a diagnosis may be life threatening.

5. Discussion and Related Work

The idea of GRAD-CAM or a related result is something that has been very sought after ever since CNNs first became popular. When observing a trained model's performance it seems like a black box blindly spitting out predictions. It would be great to see where that prediction is actually coming from, in essence what part of the image is the CNN looking at. These ideas and an implementation are discussed in Simonya et. al Simonyan et al. (2013). We decided to focus on a version of this idea found in Selvaraju et. al Selvaraju et al. (2017).

The learning to defer model is one that has been explored thoroughly. It is a clever idea with very helpful implications as if executed well can greatly reduce the burden on a doctor. Learning to defer offers the utility of human classification and decision-making to the process of machine learning and solely automated classification. Madras discusses the ideas and virtues of this topic in detail Madras et al. (2018). Learning to defer is also shown to as a method to decide whether a second opinion from a doctor is needed. This is shown in Raghu et al. (2018) where this problem is explored through the task of diagnosing patients with Diabetic Retinopathy from retinal fundus images. Their goal is to make the predictions from CNNs ‘more transparent’.

6. Project Details

6.1. Timeline

Our proposed timeline is described below. We hope to accomplish certain tasks within the dates provided. More detailed tasks will be included as we continue developing ideas and exploring more nuanced directions.

4/29 Complete the new project proposal. Meet with Matthew and reach out to Dr. Nhi Vo to discuss project direction, obstacles, and other nuances.

4/29 - 5/4 Complete preprocessing of MIMIC-CXR, exploration of CheXpert, development of loss functions (through discussion with Dr. Vo), and development of first classification model.

5/4 - 5/7 Build baseline models and secondary model, incorporating evaluation methods as decided earlier.

5/7 - 5/10 Complete paper writing, perform evaluations and tune hyperparameters.

5/10 - 5/12 Create and complete poster.

6.2. Division of Labor

Though we will all contribute to each aspect of the project, we have delegated each member to manage certain aspects of the project. We will meet biweekly to delegate new tasks, discuss progress, and share obstacles and new ideas. This will keep us accountable for our contributions to the project, and keep us on track to following our proposed timeline.

6.2.1. Brad

Brad will lead the cloud management and data exploration segment of the project, including the MIMIC-CXR exploration/analysis and use of CheXpert.

6.2.2. Paul

Paul will lead part of the model development and evaluation, including the baseline models and first classification model.

6.2.3. Claire

Claire will lead the other part of the model development which is the second L2D model, and the evaluation process of this model.

7. Results to date

As we recently shifted projects, we have not yet produced results. We are beginning the implementation process and will have an accelerated workflow for the next couple weeks. So far we have only worked with acquiring the data and setting up a platform to work in. The data is rather large (500GB) but we are able to access it through a Google Cloud Platform instance and mount the data from a cloud storage bucket. From there we have narrowed down the images we are going to use to ones that contain either a positive or negative label from one of the conditions we are looking at. This takes out images that only have uncertain scores for all of our conditions, because we would not have a ground truth label for these.

8. Questions

1. When deciding on the implementation of our image classifier we were unsure of whether it would be best to use a pretrained CNN or simply train our own. There was some literature, though not a lot about using transfer learning specifically for chest X-rays, but it did not achieve great performance. We figure this would save us a lot of time and resources training, but are not sure if this would be better than simply training a model all the way through.

2. We currently do not balance the training data for the second model between correctly and incorrectly predicted images. Should the expected prediction capacity of our first model be represented in our second model's loss function, or should we introduce balancing its train set? Alongside, should the second model be performing with prior knowledge of the first model's expected performance? Or, should it choose to defer independently of the first model's performance?

References

Tsung-Han Chan, Kui Jia, Shenghua Gao, Jiwen Lu, Zinan Zeng, and Yi Ma. Pcanet: A simple deep learning baseline for image classification? *IEEE transactions on image processing*, 24(12):5017–5032, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

Alistair EW Johnson, Tom J Pollard, Seth Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 1(2), 2019.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, pages 6147–6157, 2018.

Narinder Singh Punn and Sonali Agarwal. Automated diagnosis of covid-19 with limited posteroanterior chest x-ray images using fine-tuned deep neural networks. *arXiv preprint arXiv:2004.11676*, 2020.

Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Robert Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. Direct uncertainty prediction for medical second opinions. *arXiv preprint arXiv:1807.01771*, 2018.

Jonathan Rubin, Deepan Sanghavi, Claire Zhao, Kathy Lee, Ashequl Qadir, and Minnan Xu-Wilson. Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks. *arXiv preprint arXiv:1804.07839*, 2018.

Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.