COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Samson Qian (University of California, San Diego), August 18, 2021

Title: *Generating Explanations for Chest Medical Scan Pneumonia Predictions*

YouTube Recording with Slides

August 2021 CIC Webinar Information

Transcript Editor: Macy Moujabber

Transcript

Lauren Close:

I'd like to introduce our final speaker today, Samson Qian. Samson is one of our three winners of the inaugural CIC undergraduate student paper challenge which was conducted earlier this spring. So we welcome Samson and we're very excited to share his research as an emerging scholar. So Samson, please go ahead and take it away.

Samson Qian:

Slide 1

Thank you Lauren. So today I'll be doing a presentation which is also slightly different from the previous ones about generating explanations for machine learning predictions on viral and bacterial pneumonia.

Slide 2

And so the overarching goal of this research is to take various chest x-ray patients who have bacteria and viral pneumonia as well as healthy patients and build a machine learning model to classify between these patients using these images. And so is it possible to build a machine learning model that can accurately identify these different classes, but also interpret this machine learning model in order to understand where the model is looking at when it's generating predictions? And so this introduces the idea of using explainability algorithms, also known as Explainable AI, to analyze a model and the data it's predicting on in order to understand exactly where the model is looking at. And the ultimate goal is to see if this type of framework can help radiologists in their work in diagnosing different patients.

Slide 3

So a quick overview of Explainable AI. There are various types of Explainable AI algorithms, like, more like a family branch of different types of methods. For this research, particularly, we'll be focusing on Post-Hoc algorithms meaning algorithms to analyze complicated models after they're trained on the data instead of during the process or before. And the benefit of this is you're able to use more complex models, such as deeper convolutional neural networks in order to learn the features in the x-ray. And there's a trade-off between interpretability and accuracy, sometimes. And so deeper and more complex models are harder to interpret than simpler ones, so these algorithms provide a way to understand these more complex models. So the first type of algorithm is known as LRP, Layer-wise Relevance Propagation, and this is a method that I'll be talking more about a bit later, but this essentially looks into the model's layers and weights and understands what pixels in the image contribute most to activating a model's internal structure.

The second one is something known as LIME and this method is model-agnostic, meaning it doesn't depend on what type of model you used. This method differs from LRP in that it starts from the data first, rather than the model and it looks at individual subsections of the data to find what regions in the image contribute most to a model's prediction.

And then the next one is called Grad-CAM which is similar to LRP sort of but instead of looking at each layers and each weights, it takes a look at the convolution layers in the model and then it takes on the gradients that when you fit an image, the gradients of the last convolutional layer, and then it produces sort of, like, a activation map of the gradients.

And the last type of algorithm is sort of more of a novel algorithm known as Contrastive LRP which is a modification of the regular LRP, but it takes the relevances and then it differs between different classes such as viral bacterial pneumonia, so you can visualize the difference between the classes more.

Slide 4

So building a convolutional neural network to identify the features in these type of patients requires a complex model structure such as VGG16 and ResNet50 which are two state-of-the-art models that do a very good job in image classification, and as you can see from the model structures, it's a very deep convolutional network which is able to pick up on many features in the images. And so these two models give some of the best results in classifying between the three different classes of patients.

Slide 5

And so the x-ray images were all collected from Mendeley data and a bunch of image pre-processing is required on the images to fit it into the model. So making sure that there's a similar amount of image examples from all different types of classes and then doing some pre-processing on the images to fit into the model.

Slide 6

And here's like an image of just an overview of the model performance during training. So on the left you have the accuracy plot for the training set in validation set of images. And as you train the model more and more each iteration, each epoch, the accuracy grows more and then the loss which is what we're trying to minimize on the right side is steadily decreasing. So this implies- this shows that our model is picking up on the features in the data- the images that we're feeding it and it's doing a good job in classifying between the chest x-rays.

Slide 7

So here's just more analysis on the model built, and it's very important to have an accurate model in order to identify or examine what the model is looking at in each image. So you can see a confusion matrix on the left which shows what the model is confusing and in general is doing a very good job between the three classes besides distinguishing between viral and bacterial pneumonia which is doing- it's getting confused on a few examples. And so it's important to run these explainability algorithms to understand what the model is looking at.

Slide 8

So here's a brief overview of LRP, Layer-wise Relevance Propagation, again, and essentially what this does is you have your neural network with each layer- a hidden layer inside. And it propagates your image from the output layer backwards. So instead of forward propagating, it propagates on the image backwards and then you compute relevant scores at each neuron, which is these circles right here. And these relevant scores represent how important each pixel in the image is to a prediction that the model makes.

Slide 9

And so here you can take a look- on the right is an example of LRP. And as you can see, it produces sort of like a heat map based on the individual pixels in the image and that shows how important each pixel in the image is contributing to the model's prediction on the class. And then on the left you can see a comparison of LIME which is more of a region-based method that takes regions in the image, and then finds which regions are most important to making a model prediction.

Slide 10

And then here's more examples of LIME for the three different classes. What LIME does- it doesn't specifically go into each layer of the model to examine the weights, but instead it takes sub-regions chosen specifically on the image and runs on the model to determine which of the sub-regions is most important to the model's prediction and highlights on those specific regions.

Slide 11

And then another type of algorithm that I discussed was called Grad-CAM and Grad-CAM essentially is an algorithm sort of like LRP, but instead it examines the gradients of the convolutional layers in your model. And so this method also takes an image and then feeds it into the model and then produces an activation map which is a heat map that represents all the gradients in a convolutional layer and that picks up where exactly the model is looking at.

Slide 12

So here are some examples of Grad-CAM instead. It's sort of like a combination between LRP and LIME, kind of. It produces a heat map of a specific region on the image, and you can see what the model is focusing on and what regions it's not focusing on.

Slide 13

And last variation which is most contrasted LRP is a slight modification to the original LRP method, but instead, it takes the relevance scores and applies a modification that distinguishes the relevance between various types of classes. And so you can identify or distinguish between viral and bacterial pneumonia a lot more clearly using these different types of relevant scores. And so it's very important overall to look and compare these various types of explainability methods in order to understand what your model is specifically looking at. And so there are many types of algorithms that you can use, but in order to get a thorough understanding of what your model has learned- what your model is looking at when it's making predictions- it's important to compare these types of methods and while you're building an accurate model to help patients.

Slide 14

And just some ongoing research right now- Explainable AI is a continuously growing field. More algorithms are being developed every day and more data is being collected to build more accurate models and different types of data such as CT scans can also help with radiologists work in diagnosing patients. So running these algorithms on different types of models built to examine which model has learned the data more accurately than others.

Slide 15

And yeah, thank you so much for listening to my presentation and I would like to acknowledge Dr. Michael Pazzani, who's also on the call right now for all of his help in conducting his research, collecting data, and implementing these algorithms to identify pneumonia. And thank you. This is my email if you want to contact me.

Slide 16

Thank you so much Samson. It's really exciting to see your research evolve and we'll be, you know, watching your career and your scientific research with interest. We really appreciate you sharing your work with us, and congratulations again on winning our student paper challenge.