Machine Learning Engineer Nanodegree

Capstone Proposal

Ahmed Wael
February 27th, 2019

Domain Background

Pneumonia is an infection that can affect lungs' air sacs [1]. This infection can range in its seriousness from being mild to life-endangering. While pneumonia can affect people of any age, it's more common and serious in young children and elderly [2]. Pneumonia has different types such as bacterial pneumonia, viral pneumonia, aspiration pneumonia, fungal pneumonia, or even hospital-acquired pneumonia [2].

While its treatment is currently available, pneumonia is the world leading cause of death among children under 5 years old, killing over 2,400 children a day in 2015 and accounting for 16% of all children under 5 years old deaths [3]. Although it's more threatening for the young people, more than 1 million adults get hospitalized and more of 50,000 of them die in the US alone from this infection. The risk is even higher in developing countries such as Egypt, as about 90% of the deaths caused by pneumonia are in these countries [4].

Problem Statement

The identification of the infection is done by using X-ray chest image performed by a physician [5]. This can be converted into a classification problem which by classifying if patient has Pneumonia or not using the X-ray chest image. While the accuracy metric is important in any machine learning problem, in this scenario, the recall metric is more meaningful [6]. This is because the problem is more sensitive to false negatives -the sick patients that were diagnosed as healthy -, more than the false positives – the healthy people that were diagnosed as sick-

Datasets and Inputs

The dataset used is available on Kaggle with the name 'Chest X-Ray Images (Pneumonia) [7]. It consists of 5,863 labeled images (JPEG) split into 2 categories which are normal X-ray images and infected X-ray images. Moreover, the dataset is organized

into 3 folders -train, test, val-. The dataset was collected and preprocessed from Guangzhou Women and Children's Medical Center as part of patients' routine clinical care.

The dataset was screened for quality control by getting rid of unreadable or low-quality images. Then, the labeling was performed by two expert physicians, and got rechecked by a third expert.

Solution Statement

The dataset can be considered as an input for a machine learning algorithm such as Convolutional Neural Networks (CNN), which then output the class of the image either 1 (Pneumonia) or 0 (Normal). CNN has been used in the past few years to classify challenging datasets, especially in the medical field [8] [9]. Using back-propagation optimization techniques can solve the problem by forward propagation at first for calculating the weights of the neural network nodes and then back-propagation to minimize the error, update the weights and converge. The metric that can used for evaluating the model is the recall metric or more generally the F1 score which is a weighted average between the precision and recall. After converging the model and getting a sufficient score, the weights can be used on any X-ray image to classify it.

Benchmark Model

An ideal benchmark model for this problem is a shallow neural network with one hidden layer. The metrics for this model will be the same as the purposed solution, which is the recall or the F1 score. Moreover, they can be calculated using the same technique as the purposed model, using forward and backward propagation. A more advanced benchmark model can be a high-end neural network with its weights and use it for comparison with the purposed model.

Evaluation Metrics

The evaluation metrics that can be used to quantify the performance of both the benchmark model and the solution model are the recall, specificity, and the F-beta score [6]. While accuracy can be insightful in classification problems in general, it should not be used when the target variable classes are not balanced [6]. In the X-ray dataset, the infected category is 3 times bigger than the normal category, therefore, the accuracy metric should not be used.

Recall metric is a good fit for medical problems as it tries to capture all cases that have pneumonia with the answer as pneumonia. Specificity is the opposite of the recall, and it is needed to make sure that the recall score is not misleading and the model is not labeling all the cases as having pneumonia, as then the specificity would be 0%. The last metric that can be used is the F-beta score, which is a weighted harmonic mean between the precision and the recall, therefore having more freedom in having a combination between them.

Project Design

The model can be constructed by using a few convolutional layers that are used for basic features detection such as edges and shapes. Pooling layers can be used between each one or two convolutional layers in order to reduce the number of trainable parameters. A few fully connected layers can be used at the end in order to capture more complex features. Finally, the last layer should have two nodes corresponding the two different classes of the classification problem.

The activation function for all layers expect the final layer would be the ReLU activation function as it is a piecewise activation function so it is more prone to overfitting. The last layer would have a SoftMax activation function as it is used to calculate the probability of the input either being of the first or the second class, while being summed up to one [8].

The loss function that should be used is the log loss or the cross-entropy loss as minimizing the cross-entropy is the same as maximizing the likelihood [9]. Backpropagation and updating the weights will be done using Adam or RMSprop optimizers, while other optimizers will be experimented for better output.

The project design would make use of the transfer learning techniques, as training a convolutional neural network from scratch and designing it and going back and forth until it does not either underfit nor overfit, is very time consuming and unnecessary.

Therefore, by noticing that the dataset is large and different than the pre-trained models such as ResNet or AlexNet, the model can have ResNet or AlexNet pretrained weights as initial weights for the network, and changing only the last Fully Connected Layer (FCL) to have two nodes and initialize its weights randomly. If this does make the model converge faster, then random initialization in all the layers can be used.

Reference

1]"Pneumonia - Symptoms and causes", *Mayo Clinic*, 2019. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/pneumonia/symptoms-causes/syc-20354204. [Accessed: 26- Feb- 2019].

[2]"Pneumonia", *nhs.uk*, 2019. [Online]. Available: https://www.nhs.uk/conditions/pneumonia/. [Accessed: 26- Feb- 2019].

[3] *Thoracic.org*, 2019. [Online]. Available: https://www.thoracic.org/patients/patient-resources/resources/top-pneumonia-facts.pdf. [Accessed: 26- Feb- 2019].

[4] *Derpharmachemica.com*, 2019. [Online]. Available: https://www.derpharmachemica.com/pharma-chemica/possible-etiological-agents-of-pneumonia-in-egyptian-infants-and-children.pdf. [Accessed: 26- Feb- 2019].

[5] P. MM, "Detection of Pneumonia in chest X-ray images. - PubMed - NCBI", *Ncbi.nlm.nih.gov*, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25214377. [Accessed: 26- Feb- 2019].

[6]"Beyond Accuracy: Precision and Recall – Towards Data Science", *Towards Data Science*, 2019. [Online]. Available: https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c. [Accessed: 26- Feb- 2019].

[7]"Chest X-Ray Images (Pneumonia)", *Kaggle.com*, 2019. [Online]. Available: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia/home. [Accessed: 26- Feb- 2019].

[8]"How SoftMax activation function works?", *DataCamp Community*, 2019. [Online]. Available: https://www.datacamp.com/community/news/how-softmax-activation-function-works-6qya1z8uxxx. [Accessed: 26- Feb- 2019].

[9]"Understanding Entropy, Cross-Entropy and Cross-Entropy Loss", *Medium*, 2019. [Online]. Available: https://medium.com/@vijendra1125/understanding-entropy-cross-entropy-and-softmax-3b79d9b23c8a. [Accessed: 26- Feb- 2019].