Capstone Proposal:

Deep Learning Chest X – Rays to Detect Pneumonia

Domain Background

Worldwide, pneumonia is an illness that can be fatal for older adults and children under five It is estimated that pneumonia is the #1 reason why children have been hospitalized and approximately 1 million adults are hospitalized every year with 50,000 dying from the disease (America Thoracic Society, 2018). Furthermore, given that pneumonia can occur as a complication in the recently ongoing COVID-19 pandemic leading to a rise in cases worldwide, tools to help quickly diagnose this complication is worth exploring (World Health Organization, 2020). Pneumonia affects the lungs and makes breathing difficult, which can be fatal (America Thoracic Society, 2018). Diagnosis of pneumonia requires expertise, time, and resources which could be scarce depending on the facility, country, and situation.

Problem Statement

A common way pneumonia is diagnosed is through checking x – rays images. Traditionally, qualified professionals, such as doctors, would need to view the images and make the diagnosis. The aforementioned resources can be scarce in locations and facilities lacking qualified personnel and during times of high demand. Machine learning, specifically deep learning, can be leveraged to assess chest x – ray images and classify whether the patient the x -ray image belongs to has pneumonia or not.

Datasets and Inputs

The input data to be used for this project is the “Chest X-Ray Images (Pneumonia)” dataset from *Kaggle* (Mooney, 2018). Each data point is a chest x -ray image in jpeg format from children between one to five years old from a hospital in Guangzhou, China. There are 5,863 x – ray images and two categories, Normal / Pneumonia (Mooney, 2018). The images have been obtained as part of the patient’s health check-up. These images have been sorted into folders marked NORMAL / PNEUMONIA for use in training, testing, and validating any classification algorithms. The image classifications were determined by physicians. These classified images would be used to train, test, and validate a machine learning model.

Solution Statement

One solution to reducing the resources needed to diagnose pneumonia is to automate the diagnosis of the chest x – ray images using a deep learning algorithm, specifically a neural network algorithm. The training set of the data would be used to train the model. The testing set of the data would be used to determine how well the trained model does on unseen data. The final validation data would be used as the final dataset to see if the trained model is consistent with the results seen from training and testing the data. The Kaggle dataset has been organized to allow for the training and testing iterations of creating the final deep learning model. Deep learning python frameworks would be used to create the final training solution.

Benchmark Model

Image classification methods have been assessed through the MNIST dataset and the most effective type of deep learning for classifying the MNIST dataset has been the convolutional net (LeCun , “The MNIST database of handwritten digits”, n.d.). The chest x-rays will be resized to to a length and width of 224 by 224. As part of a Udacity introduction to convolutional networks, the settings for the multi-layer perception was used in the benchmark model for this dataset as part of “2. Additional Materials: Convolutional Neural Networks” / “Lesson 1: Convolutional Neural Networks” / “Notebook: MLP Classification, MNIST” [Udacity, “Machine Learning Engineer Nanodegree”, n.d.]. The configuration of the benchmark multi-layer model using the pytorch api is as follows (Udacity, n.d.):

```
MLPNet(
  (fc1): Linear(in_features=150528, out_features=512, bias=True)
  (fc2): Linear(in_features=512, out_features=512, bias=True)
  (fc3): Linear(in_features=512, out_features=2, bias=True)
  (dropout): Dropout(p=0.2)
)

The criterion for loss is CrossEntropyLoss and the optimizer is stochaistic
gradient descent.
```

Evaluation Metrics

Metrics used to evaluate models would be the following classification metrics: accuracy, precision, and recall. Accuracy will be the metric that shows overall how many cases have been determined correctly compared to all cases or ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives) [Koehrson, 2018]. Recall measures “the ability of a model to find all relevant cases in the dataset” or “true positives / (true positives + false positives)” [Koehrson,

2018]. Precision measures the ability of a “classification model to identify only the relevant data points” or “true positives / (true positives + false negatives)” [Koehrson, 2018].

Overall Project Design

The workflow to create a model to automate diagnosis of the chest x -rays are the following:

1. Sampling / ingestion of the jpeg data of chest x -rays from Kaggle.
2. Potential transformation of the image data for analysis (Udacity, n.d.).
3. Application of differing deep learning architectures. This would include multi-layer perceptions and convolutional networks (Lecun, n.d.; Udacity, n.d.)
4. Training data: Applying the training data on each of the deep learning architectures.
5. Testing data: Gathering the evaluation metrics on each of the testing data from the trained models. The model that has the maximum accuracy, precision, and recall will be selected as the best model (Koehrson, 2018).
6. Final validation dataset: Based on the trained model with the best evaluation metrics from the test data, the model will be assessed on this final validation dataset to confirm the best model’s performance. The model’s performance on this validation data should not significantly differ from the selected model’s performance in the aforementioned step.

References

American Thoracic Society. (2018). Top 20 Pneumonia Facts—2018. Retrieved from https://www.thoracic.org/patients/patient-resources/resources/top-pneumonia-facts.pdf

Koehrsen, W. (2018, March 10). Beyond Accuracy: Precision and Recall. Retrieved from https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c

LeCun. (n.d.). The MNIST database of handwritten digits. Retrieved from http://yann.lecun.com/exdb/mnist/

Mooney, P. (2018, March 24). Chest X-Ray Images (Pneumonia). Retrieved from https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

Udacity. (n.d.). Machine Learning Engineer Nanodegree - 2. Additional Materials: Convolutional Neural Networks.

World Health Organization. (2020). Coronavirus. Retrieved from https://www.who.int/health-topics/coronavirus