CS 294-082 Final Project

By Parth Baokar, Leonard Milea, Yu-Han Pang

Dataset resource: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

1. Introduction

1.1 Project Background:

With the success of convolutional neural networks (CNNs) in traditional image classification challenges (LeNet, AlexNet, etc.), there has been a great interest in utilizing these architectures in a medical setting to help streamline doctors' responsibilities. Powerful technology that can help replace some of the tasks of doctors can also be utilized in low-resource settings, which would prove to be an exciting development in standardizing quality of care across domestic and international communities. These networks have found a niche in aiding in the medical diagnoses through image, and here we focus on one specific diagnosis: pneumonia.

1.2 Project Motivation:

Pneumonia is a fairly prevalent respiratory disease, is among the leading causes of death within the United States, and disproportionately affects marginalized and poorer communities [1]. Early detection of pneumonia can drastically improve the survival rate, and is often performed through chest X-rays [2]. In this project, we will attempt to analyze and improve a deep learning model in its memory footprint and training speed which classifies medical x-ray images of lungs into normal and pneumonia classes. This deep learning model was first developed for a class at National Tsing Hua University in Taiwan by our team member Yu-Han Pang. We want to build a better learner to detect pneumonia from chest X-rays.

1.3 Project data set:

1. **What is the variable the machine learner is supposed to predict?**

The machine learner aims to predict whether a given chest X-ray image is that of normal lungs or that of lungs with pneumonia.

2. **How accurate is the labeling?**

The chest X-ray images are all graded by two expert physicians, and any low quality scans are removed [3]. The evaluation set was also checked by a third physician in case there were any clear grading errors. Patients who had X-ray examinations also had follow up tests to confirm a diagnosis of pneumonia, so it is very unlikely that there are any incorrect labels. There could be an erroneous label only if the confirmation of diagnosis also led to an incorrect result, which is a separate medical error that cannot be accounted for.

3. **What is the annotator agreement (measured)?**

The grades of the individual physicians are not provided and so we cannot measure an annotator agreement ourselves. It is also unclear whether or not all the images that were provided are only those that the two/three annotators agreed upon from the article.

2. Experimental Design of the Machine Learning Model

2.1 Project context:

1. **What is the required accuracy metric for success?**

We have a two-fold metric for success that require measuring accuracy on the provided test set. First, the model should perform at least better than the random guessing accuracy, which can be calculated by 1 / (number of classes) = ½. Moreover, we should also be better than just predicting the most common class (pneumonia) >62.5%.

2. **How much data do we have to train the prediction of the variable?**

There is 4976 images in our dataset where each image is 64*64. We have a validation set of 16 images and a test set of 624 images.

3. **Are the classes balanced?**

The classes are not very balanced in our training dataset. We have 1101 images belonging to class 0 (normal) and 3875 images belonging to class 1 (pneumonia). The classes are balanced in the validation set, and comparatively more balanced within the test set when compared to the training set with 234 normal X-rays and 390 pneumonia X-rays.

4. **How many modalities could be exploited in the data?**

We only have 1 modality of data since we only have images to work with. Additionally since we are working with X-ray images which are by nature grayscale, there is only one channel within each image.

5. **Is there temporal information?**

No, we only have static images so there is no temporal information.

2.2 Noises and Bias:

1. **How much noise are we expecting?**

Although labelings are determined by professional doctors, wrong labeling may still exist. For example, a Chest X-ray image may seem normal but it may contain abnormal parts in the image, which is too small to be discovered by human beings. These kind of wrong labeling is the noise of our training data.

2. Do we expect bias?

We can expect imbalance classes (more normal images) in our training data. Moreover, since there were exclusion criteria based on age, our image all came from children, there may be natural biases inside the dataset. Furthermore, although Low quality scans are filtered out, but could have high intensity images, lower resolution, etc.

3. Memory Equivalent Capacity

3.1 What is the memory equivalent capacity as a dictionary?

The MEC of the data as a dictionary is 135191 bits according to Brainome.

3.2 What is the expected Memory Equivalent Capacity for a neural network?

When training the neural network, the expected number of bits is 24581, according to Brainome.

4. Generalization

4.1 What is the expected generalization in bits/bit and as a consequence the average resilience in dB?

The Brainome-generated neural network achieves a generalization of 0.07 bits/bit. Resilience is therefore then −23.098 dB (20 $\log_{10}$ Generalization).

4.2 Is that resilience enough for the task?

−23.098 dB of resilience can be converted to bits/bit of resilience by multiplying by 20. This means that the model's resilience is -461.96 bits/bit. By calculating the average variance of each column, we get average variance = 0.023. Since 0.023 is larger than -30.457, we do not have the resilience for the task.

4.3 How bad can adversarial examples be?

Adversarial examples are special inputs created with the purpose of confusing a neural network, resulting in the misclassification of a given input. These adversarial examples are often due to overfitting and strongly related to the resilience of the network. We have 2 convolution layers, 2 pooling layers, 1 flatten layer, 2 dense layers and 2 dropout layers in our model. Since there are only very few convolutional layers in this model, we might expect that the model would be moderately affected by adversarial examples. Convolutional layers often act as a denoising operation,

4.4 Do we expect data drift?

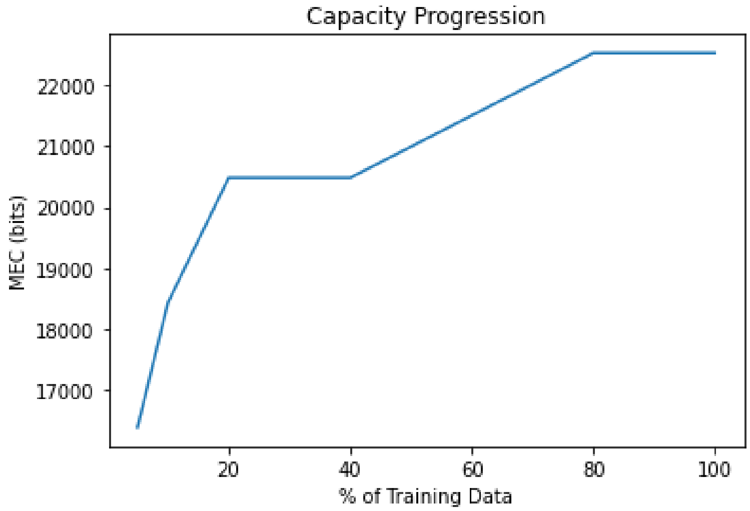
In the long term, we can reasonably expect data drift. The training data is from children of ages 1-5, and as people age, chest X-rays can change to mirror the changing bone structures. There is also the possibility of improvement in X-ray technology that leads to higher resolution images or changes in medical practice which can change the framing of chest X-rays as well. Often, sex of the patient can change the ending result of X-rays as well. The models built on these images could potentially be susceptible to these changes.

5. Capacity Progression

5.1 Is there enough data?

According to Brainome's output we may have enough data to generalize, with a yellow warning from Brainome.

5.2 How does the capacity progression look like?



The capacity progression provided by Brainome is as follows:

Capacity Progression:	at [5%, 10%, 20%, 40%, 80%, 100%]
Ideal Machine Learner:	8, 9, 10, 10, 11, 11

6. Train the Machine Learner for Memorization

6.1 Train your machine learner for accuracy at memory equivalent capacity.

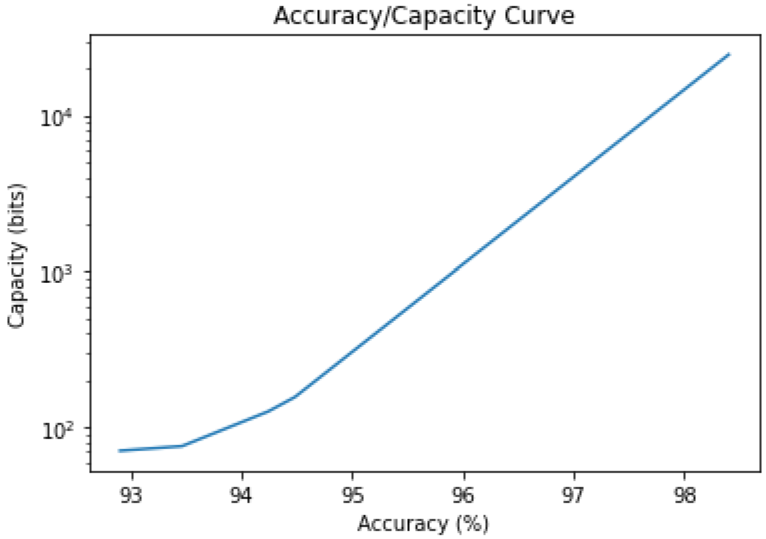
We trained our model for accuracy at MEC. In the end we got 99.80% on the training accuracy and 97.26% for the combined model accuracy, with an MEC of 24581 bits.

6.2 Can you reach near 100% memorization? If not, why (diagnose)?

We were able to get 100% memorization.

7. Train the Machine Learner for Generalization

7.1 Train your machine learner for generalization: Plot the accuracy/ capacity curve.



7.2 What is the expected accuracy and generalization ratio at the point you decided to stop?

We chose a Neural Network as our model and decided to stop at a combined model accuracy of 94.24% and generalization ratio of 24.16 bits/bit.

7.3 Do you need to try a different machine learner?

No, we don't need to try a different machine learner.

7.4 How well did your generalization prediction hold on the independent test data?

Our test data got a 92.3% accuracy so our generalization prediction held very well.

7.5 Explain results. How confident are you in the results?

Since the model reaches 94.35% accuracy on the training set and 93.61% on the validation set, the combined model accuracy of 94.24% is greater than 77.8 % and the MEC is 127 bits, which is significantly less than the MEC for memorization. Thus, we can be confident in our results.

8. Possible Quality Assurance Measures

8.1 Comment on any other quality assurance measures possible to take/the authors should have taken. Are there application-specific ones?

While the images used in the test set were not of the same patients in the test set, it would be better for images to be gathered on non-pediatric patients. As mentioned earlier, X-ray images are subject to change based on age and sex so it would better test the resilience of the model and its applicability in various settings.

Since our data has a high information density specifically around the lungs in an X-ray, using an effort of 5 (-e 5 on Brainome) helps to improve results through extended training time .

8.2 If time is present: How did you deal with it?

There is no temporal information in the classification of chest X-ray images in this project.

9. Repeatability and Reproducibility

9.1 How does your experimental design ensure repeatability and reproducibility?

Repeatability: We have provided the code for memorization and generalization that can run and saved as the ipynb file. We also documented all results of all commands we have tried.

Reproducibility: The result of each model we have tried has been documented and detailed explanation is also added as a README for reference.

The link to our github repo can be found here:

https://github.com/YuHan0215/CS294-082Final_Project.git

Contributions

Yu-Han worked on the written questions Q1-Q5, and helped Leo with training for memorization and generalization and Q9.

Leo worked on training for memorization and generalization and Q9.

Parth worked with Yu-Han on the written questions Q1-Q5, and Q8.

References

[1] Jain V, Vashisht R, Yilmaz G, et al. Pneumonia Pathology. [Updated 2021 Aug 4]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK526116/

[2] Htun TP, Sun Y, Chua HL, Pang J. Clinical features for diagnosis of pneumonia among adults in primary care setting: A systematic and meta-review. Sci Rep. 2019;9(1):7600. Published 2019 May 20. doi:10.1038/s41598-019-44145-y

[3] Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., ... & Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell, 172(5), 1122-1131.