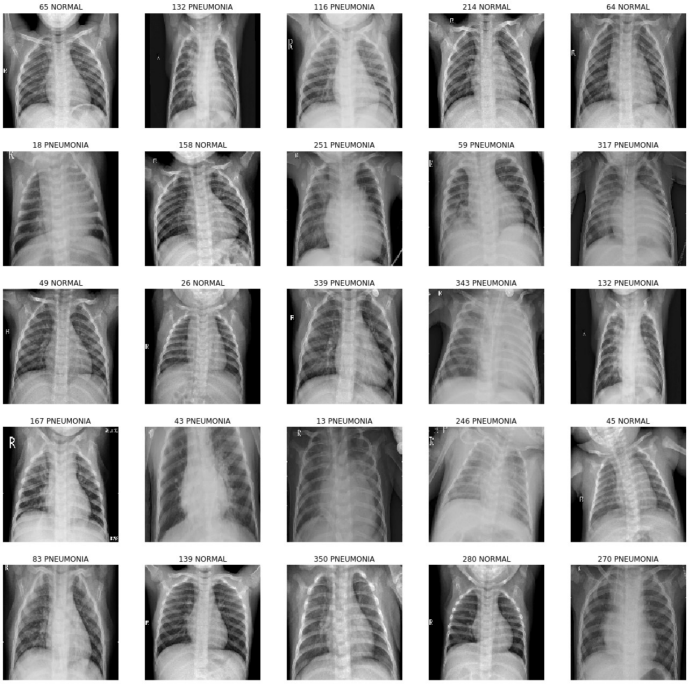

COMP47590

Advanced Machine Learning

Assignment 2: Pneumonia Detection

Introduction

Medical applications are amongst the fastest growing areas of application of deep learning today. Clinical screening offers a particularly exciting field of use as automated, or semi-automated, systems could massively increase the numbers of people who can be screened. In this assignment we will build machine learning models to detect pneumonia from chest x-ray images. Figure 1 shows an example of a normal image and two examples with pneumonia.

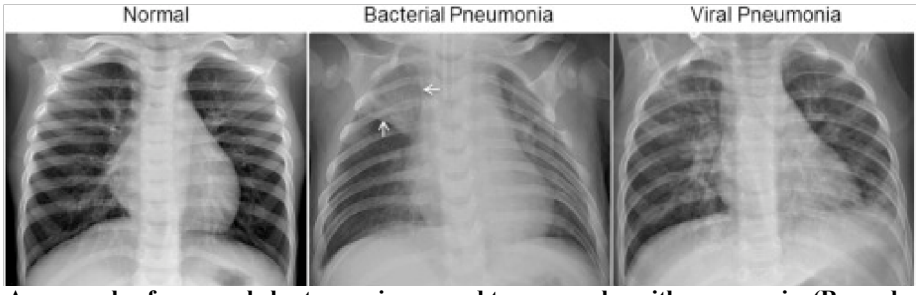

Figure 1: An example of a normal chest x-ray image and two examples with pneumonia. (Reproduced from (Kermany et al, 2018)).

In this assignment you will build convolutional neural network based detectors of pneumonia from chest x-ray images. The chest x-ray set contains a training set of 5,235 images and a test set of 627 images.

Tasks

Perform the following tasks:

1. Load the chest x-ray dataset and become familiar with its contents. **Note:** Print some sample images.
2. As a simple baseline train a **logistic regression** model to perform the classification task.
   Note: You will need to convert images into a vector representation to allow this. We suggest converting the images to greyscale and resizing them (perhaps to 162*128).
3. Train a **convolutional neural network** model to recognise pneumonia in these images.
   Note: Consider the following:
   - Try the classic LeNet-5 model architecture.
   - You should consider converting the images to greyscale.
   - You should consider shrinking the images (perhaps to 162*128).
   - Perform a suitable evaluation experiment to determine how effective the model trained is.
   - You could consider using approaches to handle the class imbalance in the dataset.
4. Use **data augmentation** techniques (for example image rotations or slight colour changes) to increase the size of the training dataset. Use the augmented

dataset to train a model with the same model architecture that you used in Part 3.
Note: Think carefully about the types of transformations that are appropriate to use in this context (for example flipping images horizontally or vertically is not appropriate as x-ray images have a fixed frame of reference).

5. Change your model so that it uses a pre-trained VGG16 model (trained on the ImageNet dataset) that is fine tuned to the pneumonia detection problem.
Note: Consider unfreezing one of the convoluito0onal layers in the network as well as the dense layers.
6. Compare the performance of the 4 models built in parts 2 to 5 using the supplied test set. Write a reflection on the results of this comparison. You could describe the most accurate model, why you think a certain model performed well, and the computational overhead of different techniques and anything else you think is interesting. Use no more than 300 words.

Notes

The following notes may be useful:

- **Can I Use Scikit-Learn, Keras And Other Python Packages?** Yes, and you absolutely should!
- **It's Taking Forever!** The dataset is reasonably large. If you find things are taking too long feel free to down-sample the dataset (make sure to use the same down sampling in all experiments). Submissions will not be penalised for this. Google Colaboratory (https://colab.research.google.com/) might be a useful resource for accessing computation.
- **Can I Work In A Team?** Teams of up to two people are allowed. All team members will receive the same mark. There is no penalty for submitting as a team, and no reward for submitting as an individual.
- **What No Templates?** No there are no templates, but there are plenty of template notebooks from the tasks performed during the course that you can use as a starting point.

Submission

The key submission details for the assignment are as follows:

- **Submission date:** Friday 1st May 2020 before 23:59.
- **Submission method:** Submissions should be made through the module Moodle site.
- **Submission format:** Submissions should compose a zip file containing the following: ○ One or more Jupyter notebooks addressing each of the tasks required ○ .html exports of each Jupyter notebook after execution that contains all output
    - ○ Any other files required to execute your code (e.g. saved model files – but do not upload the datasets)

Marking

Marking of tasks will be based on the following weighting.

- Task 1 5% Data loading and exploration
- Task 2 15% Logistic regression model
- Task 3 35% Simple CNN
- Task 4 15% Data augmentation
- Task 5 20% Pre-trained model
- Task 6 10% Reflection

References

Kermany, Daniel S., Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, et al. "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning." *Cell* 172, no. 5 (February 22, 2018): 1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010. https://www.cell.com/cell/fulltext/S0092-8674(18)30154-5