Deep Learning Classification on Chest X-ray Images with CNN, ResNet parallelized on GPU

Team 15: Zifeng Jiang

## Introduction

Pneumonia is an infection of one or both of the lungs caused by bacteria or viruses. It is a serious infection in which the air sacs fill with pus and other liquid. It is a very severe disease which may cause people die, so it is very important to discover it before it’s too late.



In order to reduce the burden on doctors and make the computers be able to automatically identify whether a human has pneumonia or the human is healthy by chest X-ray images, I will build this project with 2 deep learning models with parallel methods to help the chest X-ray images classification mission.

The goal of this project is to do model comparison between those 4 deep learning models and parallelization of data preprocessing, model selection and training. In the end, I will compare the execution time between series and parallel programming.

## Methodology

1. Data Preprocessing

First, I will read all the images from folder. Then, resize all the images’ size to 128x128(Every image is larger than 128x128). Next, loading all the images into an array with labels. Then, normalize image array pixel values between 0 and 1. Finally, re-split the training and validation dataset in order to do k-fold validation.

1. Deep Learning Models

In this project, I will build 4 deep learning models. The first model will be a regular CNN model. The second one will be a deeper CNN model with more filters and smaller kernel size. The third model will be a hand script ResNet model with bottleneck structure. The last model will be a pretrained ResNet101 model. All the models will be trained by augmented images.

1. Parallel Methods

In this project, I will use multiple GPUs with multiple parallel methods such as dask array, Joblib to embarrassingly parallel and multiprocessing. The NumPy array will be mainly used to load the pixel data from the images so that I can use Joblib or dask array to optimize it. And the k-fold cross validation and grid search can be parallelized in embarrassingly parallel with multiprocessing.

1. Time Evaluation

The result of the project will be mainly evaluated by the length of time costed when running this project. I will compare the parallel speedup performance on different number of GPUs such as 1, 2 and 4.

## Dataset introduction

The dataset is organized into 3 folders (train, test, validation) and contains subfolders for each image category (Pneumonia/Normal). There are 5,863 X-Ray images (JPEG) and 2 categories (Pneumonia/Normal). Chest X-ray images were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children’s Medical Center, Guangzhou. All chest X-ray imaging was performed as part of patients’ routine clinical care.

The size of the dataset is 1.15GB.

## Data source

This data source comes from the " Chest X-Ray Images (Pneumonia) " dataset on Kaggle. Its link is as follows:

<https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia>

The original version of this dataset can be found at this link:

<https://data.mendeley.com/datasets/rscbjbr9sj/2>