DEEP FEATURE TRANSFER FROM DEEP LEARNING MODELS INTO MACHINE LEARNING ALGORITHMS TO CLASSIFY COVID-19 FROM CHEST X-RAY IMAGES

SUMMARY

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by SARS-CoV-2. It was first reported on December 2019 in Wuhan, China, and declared as a pandemic on 11 March, 2020. Even though the disease is a severe acute respiratory illness, it affects various organs and causes several symptoms such as fever, dry cough, tiredness, the loss of taste or smell, diarrhoea, headache, aches and pains, sore throat, and conjunctivitis. As of the beginning of July 2021, over 185 million people have been infected and more than 4 million people died because of COVID-19. For that reason, one of the most important issues is the diagnosis of COVID-19. Although the most basic method to diagnose COVID-19 is Polymerase Chain Reaction (PCR) test, different techniques have been being experimented and developed. Since COVID-19 has a huge effect on lungs, diagnosis methods based on lung characteristics and images are emphasized. However, there are various illnesses affecting lungs. Hence, it has been an important challenge and problem to find a procedure to classify COVID-19 with high success rate.

In this thesis, we suggested use of deep feature transformation from deep learning models to machine learning (ML) algorithms to classify patients COVID-19 infected via chest X-rays. In addition to image data, we also used the demographic information of patients during ML process to contribute to the information coming from deep features. Chapter 1 gives background information about our problem, the purpose of our study, the related literature survey and the structure of this thesis. The basic information about our image data, chest X-rays, are given in Chapter 2.

The problem we focused on is a binary classification between COVID-19 patients and other people. In order to solve this problem, we used data set containing 131 COVID-19 and 123 non-COVID-19 labeled data. Then, we divided the data set into train and test sets so that 80% of the total data is in the train set, and then augmented the train set with horizontal flip, vertical flip, 90 degrees of rotation, 180 degrees of rotation, and 270 degrees of rotation to increase the size of the train set. Thus, at the end, we yielded 630 COVID-19 and 588 non-COVID-19 labeled data in train set, and 26 COVID-19 and 25 non-COVID-19 labeled data in test set. The augmented data was used on CNN experiments only. At the beginning of Chapter 5, the data set and augmentation technique was detailed.

The deep learning models we used are Convolutional Neural Network (CNN) models such that AlexNet, ResNet-18, ResNet-34, ResNet-50, VGG16, and VGG19. We particularly experimented three different optimization methods for each CNN model such that SGD with momentum, Adam and Adam with decoupled weight decay. The objective loss function was to minimize cross-entropy loss function which was common for each model. Each image sample was resized to 227 x 227, center cropped, converted to gray-scale, and then normalized. Chapter 3 consists of the

introduction to deep learning, basic information about CNNs, and how to perform transfer learning. Two types of transfer learning were used in this study, which are transferring pre-trained model weights into CNN models and transferring deep features extracted from CNN models into ML algorithms. Pre-trained CNN models are the models that previously trained on ImageNet data set on the record, and we performed re-train after initializing the models with these recorded weights. Deep feature transfer learning is extracting the features of CNN model from the fully-connected block of model, and using it as feature matrix in another artificial intelligence technique such as ML algorithms.

The ML algorithms we used are supervised learning algorithms such that Support Vector Machines (SVM), Logistic Regression (LR), K-Nearest Neighbor (KNN) and Linear Discriminant Analysis (LDA). We experimented different regularization techniques, which are Lasso known as L1 norm and Ridge known as L2 norm, on stated ML algorithms. Chapter 4 consists of the introduction to ML, basic information about algorithms and regularizers, and cross-validation technique. We performed 10-Fold cross-validation on train set to obtain the generalized hyper-parameter choices besides hyper-parameters specific to our initially split test set. The algorithms and experiments were applied to the feature set of demographic information, the deep transferred feature set, and the combination of transferred features and demographic information separately. The demographic information feature matrix clearly consists of two feature columns such that age and sex information. The length of transferred deep features for each sample is thousand. Hence, the combined feature matrix contains thousand two columns.

All experiments for CNN and ML are detailed in Chapter 5, including data pre-processing and hyper-parameter tuning techniques for ML specifically. Grid search was used to find optimal parameters for each feature matrix and algorithm. The source code for experiments was mainly carried out in Python programming language, and a small part was done in R programming language. The CNN models were applied using PyTorch library in Python, and the ML algorithms were applied using Sklearn library in Python. Only regularized LDA algorithm was coded in R programming language using TULIP library. We performed our CNN experiments on GPU to have faster and parallel processes. Since we did not have an opportunity to reach a physical computer including GPU that we can use during our experiments, we performed the experiments on the Google Colaboratory platform. It is a partially-free platform for Gmail users to implement CUDA to use its provided GPU. After collecting CNN results and *.pth files containing the best model weights, the ML experiments were performed locally on CPU.

We explained the performance measurement techniques in Chapter 6 together with experiment results for CNN and ML processes. The best result was achieved by using ResNet-50 model with Adam optimizer. The metrics on this result are 92.16%, 0.9216, 0.9215, 0.9216, 0.9216, 0.9215 for the accuracy, sensitivity, specificity, precision, F1 score, and AUC score respectively. Since we experimented for obtaining optimal hyper-parameters for both generalized and specific to our test data, the results for both were reported too. For the feature matrix of demographic information, the best results for both generalized and chosen test set hyper-parameters are the same, and achieved with KNN algorithm. The metrics on this result are the accuracy of 56.86%, the sensitivity of 0.5686, the specificity of 0.5837 the precision of 0.6863, the F1

score of 0.4955, and the AUC score of 0.5745. For the deep feature matrix obtained from ResNet-50 model weights, SVM with Ridge penalty, LR, LR with penalty, and KNN algorithms had the same results according to generalized hyper-parameters. The metrics on this results are the accuracy, sensitivity, specificity, precision, F1 score and AUC score of 92.16%, 0.9216, 0.9230, 0.9243, 0.9215, 0.9223 respectively. Finally, for the combined feature matrix of demographic information and extracted deep features obtained from ResNet-50 model weights, SVM with Ridge penalty, LR, LR with Ridge penalty, and KNN algorithms had the same results as well according to generalized hyper-parameters. The metrics on this results are the accuracy, sensitivity, specificity, precision, F1 score and AUC score of 92.16%, 0.9216, 0.9230, 0.9243, 0.9215, 0.9223 respectively.

In conclusion, according to stated results, we yielded an improvement of using regularization with linear discriminant analysis and Lasso regularizer. We did not have an improvement by combining demographic information with deep features. However, we anticipate an improvement with this image and non-image data combining technique by using more data samples and more information about patients such as doctors report, tobacco product use, associated genetic diseases, respiratory test information, etc. Finally, even though we could not see an improvement from CNN testing results to ML testing results in terms of the accuracy, sensitivity and F1 score, the specificity and precision improved, as we discussed in Chapter 7, a data set with more samples and these samples inspected by subject matter experts, such as specialist radiologists for our X-rays, would allow the study to have better metric results and better comparison opportunities between experimental phases.

Keywords: COVID-19, Chest X-Ray, Data augmentation, Binary classification, Demographic information, Deep learning, Convolutional Neural Networks, Pre-trained CNN models, Transfer learning, Deep feature extraction, Deep feature transfer, Machine learning, Supervised learning, Regularization, Lasso, Ridge, Grid search.