Poster

Evaluation of Transfer Learning for Protein Function Prediction

Student: Harry Thomas Chirayil; Supervisor: Dr. Claus Horn
Institute: IAS

Abstract

The theoretical basis of the Transfer Learning methodology and python codes to compute the enzyme activity of protein sequences are presented in the thesis. The concept of Transfer Learning is of high importance in neural network. It is considered as one of the Machine Learning success drivers in implementations based on deep learning models in today's world. The focus of this thesis is to study and understand the Transfer Learning concept and use this to generate a new deep learning model that has been developed from a model learnt with bigger datasets that can be utilized for learning over the unseen new smaller datasets .

Introduction

Proteins are made up of basically twenty different types of amino acids which can evolve over time, and it's been said that there are more than millions of different types of combinations of these in reality. This can be a matter of analysis to determine how fit or efficient they are related to the numerous evolutions and patterns of folding in proteins. Hence determining the physical enzyme activity of these proteins is very critical as it can be measured and used for the achievement of higher level of fitness over the changes in structure.
The protein sequences and corresponding factor values which basically is our input to the model has been obtained from the dataset hub of the university. Here we are using Transfer learning method, which is a machine learning method where a model also called pretrained model in this scope, developed for a task is reused as the starting point for a building another model with unseen but related data thereby the learning or the knowledge gained on the initial process is utilized.
The focus of this thesis is to study the Transfer Learning in detail, develop and implement deep learning model to predict the enzymes activities of protein sequences based on the Transfer Learning approach and incorporate the fine-tuning method to fine tune the model and explore its importance in transfer learning methodology.

Design

Logistic Regression Model – Baseline Model

Logistic regression is considered one of the fundamental classification techniques in machine learning. It is fast and relatively uncomplicated and convenient for the people to interpret the results. In this thesis, a logistic regression model is to be developed to compare the performance of the neural network models subjected to Transfer Learning approach.

Generally, the process of creating a logistic regression model is to find a regression function such that the predicted response of the model is as close to the actual responses for each observation. In case of a binary classification problem , the actual response can be only 0 or 1. The regression function can be used then to predict the outputs for new and

Convolutional Neural Network Model

CNNs are generally feedforward artificial neural networks with alternating convolutional and subsampling layers. In CNN itself , the one-dimensional layers has gained popularity especially when the training data is small or application specific. They have immediately achieved good performance levels in several fields such as personalized biomedical data classification and early diagnosis, structural health monitoring, anomaly detection,etc. The convolutional layer learns local patterns of data in convolutional neural networks. It helps to extract the features of input data to provide the output.

In addition to these characteristics, they can also combine both feature extraction and classification tasks into a single body unlike traditional Artificial Neural Networks (ANNs). The conventional Machine Learning (ML) methods usually perform certain pre-processing steps and then use fixed and hand-crafted features which are not only sub-optimal but may usually require a high computational complexity whereas the CNN-based methods can extract the "learned" features directly from the raw data of the problem at hand to maximize the classification accuracy. This is indeed the key characteristic for improving the classification performance significantly which made CNNs attractive to complicated engineering applications.

As the model is developed from a pretrained model, it is very crucial to modify the input shape of the layer of the model. This was taken care of when the CNN model was built.

SL.No	Model	Accuracy
1.	Logistic Regression	0.84
2.	CNN with freezing	0.88
3.	Fine-tuned CNN	0.93

Fig 1: Validation accuracy of the developed models.

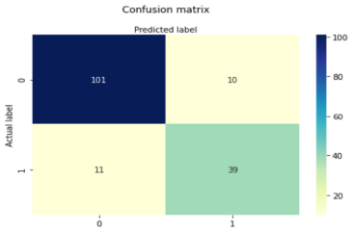

Fig 2: Confusion matrix of Regression Model

Fig 3: Plot of loss function of CNN Model

Fine-Tuned Convolutional Neural Network Model

The development of the fine-tuned CNN model comes as the model as described in the part above has been formulated and integrated along with the pretrained model.

According to the TL methodology, the fine-tuning the model can be performed in many ways according to the need and specific results that need to be obtained from the process. Once the initial top layers are configured , then the layers below them can be fine-tuned or unfreezed depending upon the feature extraction. At last the model is fitted and evaluated to see how well the fine-tuned model performs to the available dataset in hand.

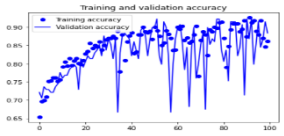

Fig 4: Accuracy plot of CNN Model

Fig 5: Accuracy plot of Fine-tuned CNN Model

Results

The implementation of the three model designs has been tested with the different protein sequence files obtained from the ZHAW datahub.

The overall accuracy of the logistic regression model is 84% which is considered as good in classification problems and the performance is further evaluated using confusion matrix and Receiver Operating Characteristic (ROC) curve.

CNN model developed had an accuracy of 88% while keeping the sequence layers of the pretrained model freezed. The model performance over the epochs was visually plotted for analysis using the built-in functionality called history callbacks.

Fine-tuned CNN model got an accuracy of 93% when the sequence layers of the pretrained model were unfreezed and integrated for implementing Transfer Learning methodology.

Conclusions

- Transfer Learning methodology is studied and implemented using deep learning architecture.
- Fine-tuning concept can be used to increase accuracy of machine learning models.

References

[1] Francois Chollet. DEEP LEARNING with Python.(2018).

[2] Ste Nicholas C Wu, Lei Dai, C Anders Olson , James O Lloyd-Smith, Ren Sun. Adaptation in protein fitness landscapes is facilitated by indirect paths.(2016).