Interim Report

Course Info: COMP 4801 Final Year Project

Project Title: Financial Market Prediction by Deep Learning Neural Network

Students: 3035142596 LIU Jiayao

3035183552 ZHANG Yuqing

Date of Submission: 21/01/2018

1. Project Background

1.1 Deep Learning

Deep learning is a new field of machine learning, the motivation of which is to establish and simulate the neural network of a human’ s brain to study and analyze information such as images, sounds and texts etc. Deep learning combines low-level features to form more abstract high-level attributes, categories, or features to discover distributed representations of data.

Deep learning has been applied to multiple fields including computer vision, speech recognition, natural language processing etc. For example, combining deep learning with the acoustic model in speech recognition can reduce speech recognition error rate by 30% (Taghi & Hoseinzade, 2013).

1.2 Convolutional Neural Network

Convolutional Neural Network (CNN), biologically-inspired variants of MLPs, is one of the artificial neural networks.

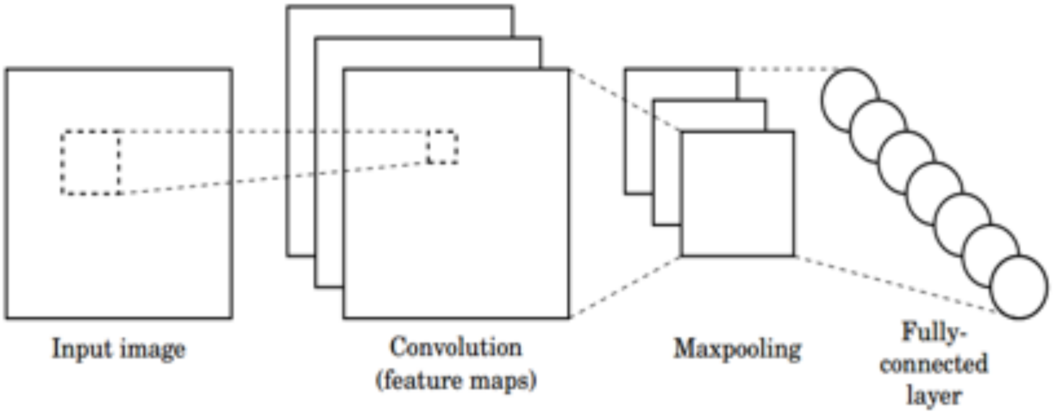

Figure 1. A CNN with three types of layers

As shown in Figure 2, there are three types of layers in convolutional neural networks, which are convolutional layer, pooling layer and fully-connected layer. The convolutional and pooling layer first filter and abstract the features and then pass the features to the fully connected layer where each neuron is connected to all the neurons of the next layer to produce the output. The structure of fully connected layer is similar to the layer in a feed forward neural network. These three types of layers are stacked together to form a CNN architecture.

Compared to feed forward neural network, CNN decreases the number of parameters required in the model because of parameter sharing. Parameter sharing assumes that if one feature is useful at one particular position, then the feature should also be useful at another position.

1.3 Dow Jones Industrial Average

Launched in May of 1896, Dow Jones Industrial Average (DJIA) is a price-weighted stock market index. This index is used as a measure of the development of industrial compositions in the US stock market. Dow Jones Industrial Average includes 30 blue chip companies of different industries in the United States, making it a good indicator of the US stock market.

1.4 Related Work

Deep learning is used widely in the field of financial prediction in recent decades. One related study applies an autoencoder which is composed of stacked restricted Boltzmann machines to extract features of stock prices (Takeuchi & Lee, 2013).

According to the preliminary research, CNN has significant achievements in image recognition. There is a particular study (Krizhevsky, Sutskever, & Hinton, 2012) which applies CNN to classify the 1.2 million high-resolution images into 1000 classes. ReLU nonlinearity is utilized to extract the non-linearity in this practice because CNN with ReLUs is trained faster than that with tanh units. The model is trained on multiple GPUs because of the large size of the training data. The pooling operation of the CNN is max pooling, which is possible to be used in our model. After the pooling operation, the features are abstracted to 4096 neurons in the fully-connected layers in total.

2. Project Objective

The objective of this project is to develop a CNN model to predict the trend of DJIA. We intend to predict the trend of DJIA of the 12th week, which covers five business days, based on the trading data of previous 11 weeks, which covers 55 business days.

The target accuracy of our model is greater than 0.5 which is better than a random guess.

3. Project Methodology

The input data include 10 variables consisting of the 6 stock variables for each of the component company of DJIA as well as 4 macro variables. The output is the predicted trend of DJIA. In essence, our stock trend prediction problem is a supervised binary classification problem because the data will be labelled and the model will generate a discrete binary result (0 or 1) to represent the trend.

3.1 Data Collection

We have collected 11 variables with the span of 30 years (Oct. 1, 1998 - Sept. 31, 2017) as follows:

Table 1. Breakdown of the eleven variables in the input dataset

Variable Name	No. of Variables	Variable compositions	Variable Frequency	Data Source
Dow Jones Industrial Average	1	Value	Daily	Yahoo Finance, Bloomberg
Stock data (each company)	6	Open	Daily	
		Close	Daily	
		High	Daily	
		Low	Daily	
		Volume	Daily	
		Adjusted Close	Daily	
Macro data	4	GDP	Quarterly	Federal Reserve Economic Data (FRED)
		Interest Rate	Monthly	
		Unemployment Rate	Monthly	
		Inflation Rate (CPI)	Monthly	

3.2 Data Preprocessing

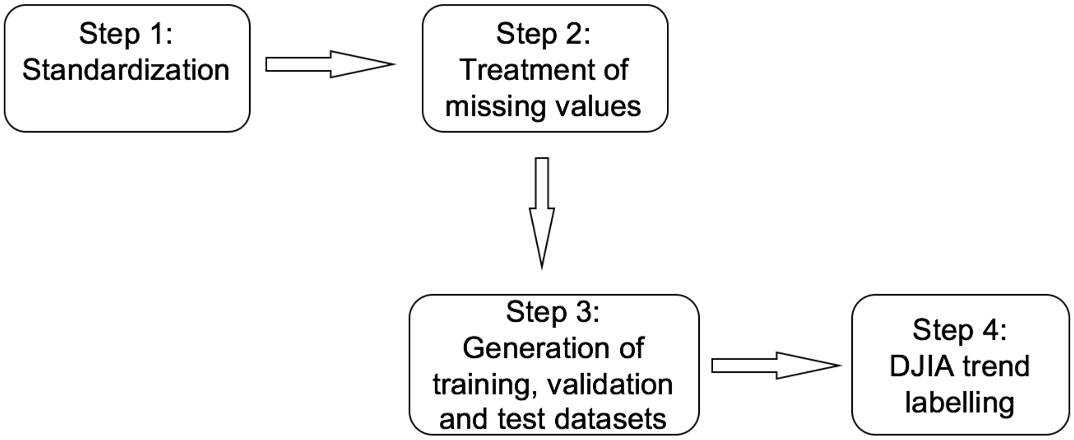
Step 1: Standardization
Step 2: Treatment of missing values
Step 3: Generation of training, validation and test datasets
Step 4: DJIA trend labelling

Figure 2. Four technical procedures of Data Preprocessing

Step 1 and Step 2 in Figure 2 are common data preprocessing techniques. Step 1 centers the data and transfers the variables into a uniform scale. which eliminates the effect caused by different scales of different variables. In Step 2, we propose to take the closing price of the last valid trading day as the missing day’ s price caused by trading suspension.

After the first two steps, the data will be cleaned into samples with the span of 12 weeks. The samples will be shuffled and split into training, validation and test datasets according to a commonly used ratio 8:1:1. In Step 4, the trend of DJIA of the 12th week in each sample is labelled by comparing the value of the first business day to that of the last business day. If DJIA of the last business day of this week increases, then it is labelled as 1, otherwise labelled as 0.

3.3 Assessment of the Model

By fitting the model to the three datasets, the trained model will produce a predicted $\hat{Y}$ for each sample. The accuracy of a dataset can be acquired by comparing the predicted values to the original labelled values in the specified dataset. The definition of the accuracy of a dataset is as follows:

$$Accuracy = \frac{the\ number\ of\ correctly\ classified\ samples\ in\ the\ dataset}{total\ number\ of\ samples\ in\ the\ dataset}$$

Apart from accuracy, processing time is also a significant factor regarding efficiency in the assessment. The model should take a reasonable time to generate a result.

3.4 Determining Hyperparameters and Parameters

The architecture of the constructed CNN model is contingent upon hyperparameters and parameters while the latter are also dependent on the former.

The following hyperparameters and parameters are to be determined to build the CNN model:

Hyperparameters
- Number of layers
- Types of layers
- Convolutional layer: no. of filter, size of the filter(s), stride, zero-padding, activation function
- Pooling layer: pooling function
- Fully connected layer: no. of hidden units, activation function

Parameters
- Weights and bias of the convolutional layer(s) and pooling layer(s)

There are some successful CNN architectures including LeNet, AlexNet and VGGNet etc. on image recognition. In addition, some common practices such as max pooling and ReLU can be adopted in our project. We can take reference from these structures and tune the hyperparameters based on the data of our project. The parameters will be learnt through backpropagation.

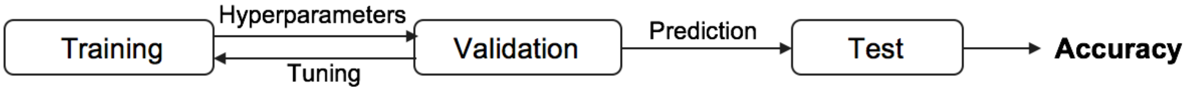

Figure 3. Demonstration regarding fitting model on the datasets

In Figure 3, with a specified combination of hyperparameters which we take reference from the existing CNN architectures, the parameters including weights and bias terms of each layer are learned from the training dataset through backpropagation. Then the hyperparmeters will be tuned according to the performance of validation dataset. After the set of hyperparameters and parameters is settled, fitting the model on test dataset produces the accuracy of the model.

3.5 Technology Support

The model will be implemented in python. In 3.1 Data Collection, Yahoo Finance API in python is employed to extract the stock data from Yahoo Finance. In addition, we make use of the package pandas in python to clean the raw data into a well-structured data frame in data preprocessing. There exist a number of frameworks for deep learning. Tensorflow, a framework that uses computational graph to execute neural network, will be applied in our model training process. It not only contains useful functions to calculate back propagation but is also able to run on a GPU. In contrast to CPU, GPU possesses more computational units and provides the possibility of parallel computing. Considering the large quantity of data and the complexity of the model, a GPU may be required to implement the model to reduce the processing time.

4. Difficulties in Data Collection

We have encountered the problem of data loss because of the change of components of DJIA in the specified 30 years. Once a company went bankruptcy, merged with or was acquired by another company, the database of Wall Street Journal Yahoo Finance will not keep the record any more. It is found that the stock data of the component companies are intact after 2005. Therefore, one solution is to only use the data after 2005. If we allow overlapping by trading day in the data, we will still get adequate number of samples. We also propose to find the intact stock data of the component companies that still exist or once got replaced in the DJIA composition. All these eligible companies (more than 30) will be placed in the input data.

5. Project Progress

The schedule of the project is as follows:

Procedures	**Expected Time**	**Progress**
Preliminary research	Before Oct. 15	Completed
Data collection	Before Jan. 21	In progress
Data preprocessing	Before Feb. 15	Not completed
Model Training	Before Apr. 15	Not completed

5.1 What Has Been Accomplished

Preliminary research has been completed and the study objective has been decided. We are currently in the stage of data collection and have identified the problem for which a possible solution is proposed.

5.2 What Will Be Done

The data collection job will be completed before Jan.21. After determining the initial CNN model, which will also shed light on the dimension of the input data, the data will be cleaned into samples with appropriate dimensions and we will also be able to generate the training, validation and test datasets in the coming month.

Reference

Taghi S. & Hoseinzade S. (2013). Forecasting S&P 500 index using artificial neural networks and design of experiments. Journal of industrial Engineering International.

Takeuchi, L., & Lee, Y. Y. A. (2013). Applying deep learning to enhance momentum trading strategies in stocks. In Technical Report. Stanford University.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).