5470 Final Project: (Re-)Imag(in)ing Price Trends

Zhanmiao Huang, Xuanyu Shen

May 5, 2024

1 Introduction

Predicting future returns using past price data is a key area of research, often approached through traditional hypothesis testing methods. However, recent work explores machine learning techniques to predict future returns of stocks using past price data, leveraging convolutional neural networks (CNNs) to analyze two-dimensional images encoding past price dynamics. This project replicates and extends experimental results from the paper [1]. Extensions include providing Grad-CAM visualizations for model interpretation. Our results show improved perception regarding accuracy and extraction capabilities of CNN-based stock predictors.

2 Dataset

We investigate the 20-day OHLC in images from 1993 to 2019 in 64*60 grey scale that depicting price (daily open, close, high, low and a moving average price) and volume bar (daily trading volume). The labels are the returns of the next time horizon with 1 indicating positive (up) return and 0 otherwise (down). There are in total approximately 50% up and 50% down labels.

3 Methodology

CNN Workflow The CNNs are trained to do the binary classification. The workflow follows from basic procedure of training, model tuning, and finally prediction. To reduce the effect of time, we choose data in years 1993, 1998, 2003, 2008, 2013, 2018 as training dataset, years 1994, 2000, 2004, 2009, 2014, 2019 as testing dataset, years 1999 and 2010 as validation dataset. The reason for not using entire dataset from 1993 to 2019 is the limitation of time. The activation function in use is Leaky ReLU. The loss during training is measured by cross-entropy between labels and optimized by AdamW optimizer with a learning rate 1e-5. The right figure shows an example of input image and the structure of baseline model (Jiang et al., 2021).

output: probability of positive return in 20 days, softmax output $\hat{y}$

$$loss = -y \log(\hat{y}) - (1-y)\log(1-\hat{y}) \tag{1}$$

Note that: trained with 10000 data randomly selected each epoch to save cost

4 Predicting Results

The loss on the training data steadily decreases over 20 epochs, while the loss on the validation data remains relatively constant. We then achieve prominently higher accuracies on the total testing set (2001-2019), around 52% before training and 55% after training, an increase of 3%. In particular, the accuracy can be as high as 57% on some randomly selected subsets of the testing data.

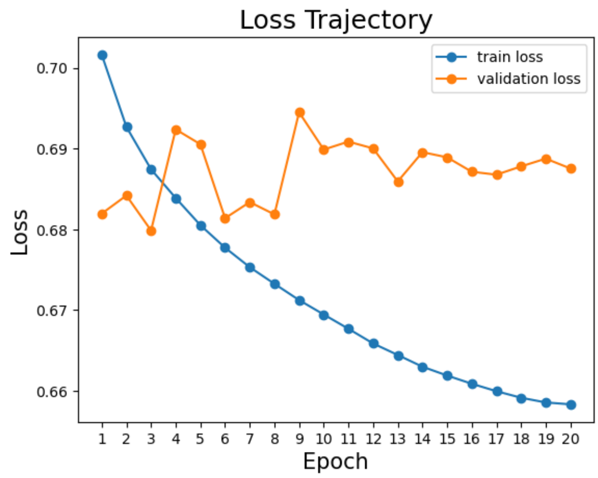


Figure 1: loss

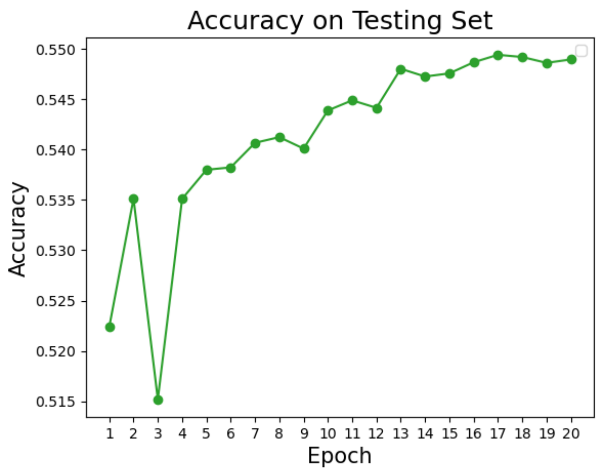
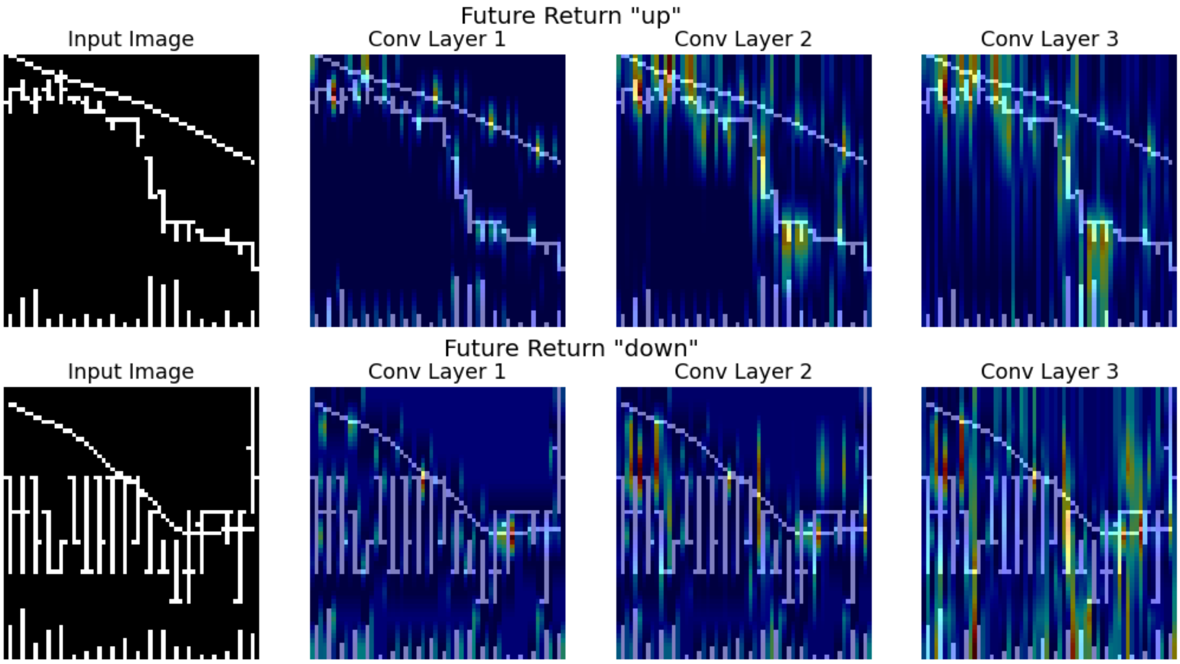


Figure 2: **for all test dataset, the accuracy of predicting future return. accuracy obviously increase with epochs**, 20 epochs-3% rise in accuracy

5 Visualization and Interpretation of CNN

Grad-CAM To visualize and interprete the CNN model, we employ Grad-CAM (Gradient-weighted Class Activation Map, [2]), which can produce images processed by each layer of the CNN and illustrate the most important features in triggering up or down return predictions. In the figures below, the images in each row are represented in the order of CNN layers. Input image is shown as greyscale as processed in CNN, while the images in convolutional layers are are shown in RBG format. The brighter regions correspond to higher activation, such as the special transition in price trend and high volume bar in striking colors.



6 Conclusion

The general conclusion from the simulation is that the CNN is able to detects part of complicated technical patterns by increasing the price predicting accuracy.

References

[1] J. Jiang, B. Kelly, and D. Xiu. (re-) imag (in) ing price trends. *The Journal of Finance*, 78(6):3193–3249, 2023.

[2] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.