

Financial Time Series Forecasting using CNN and Transformer

Zhen Zeng, Rachneet Kaur, Suchetha Siddagangappa,
Saba Rahimi, Tucker Balch, Manuela Veloso

J. P. Morgan AI Research, New York, NY, USA

Abstract

Time series forecasting is important across various domains for decision-making. In particular, financial time series such as stock prices can be hard to predict as it is difficult to model short-term and long-term temporal dependencies between data points. Convolutional Neural Networks (CNN) are good at capturing local patterns for modeling short-term dependencies. However, CNNs cannot learn long-term dependencies due to the limited receptive field. Transformers on the other hand are capable of learning global context and long-term dependencies. **In this paper, we propose to harness the power of CNNs and Transformers to model both short-term and long-term dependencies within a time series, and forecast if the price would go up, down or remain the same (flat) in the future.** In our experiments, we demonstrated the success of the proposed method in comparison to commonly adopted statistical and deep learning methods on forecasting intraday stock price change of S&P 500 constituents.

Introduction

Time series forecasting is challenging, especially in the financial industry (Pedersen 2019). It involves statistically understanding complex linear and non-linear interactions within historical data to predict the future. In the financial industry, common applications for forecasting include predicting buy/sell or positive/negative price changes for company stocks traded on the market. Traditional statistical approaches commonly adapt linear regression, exponential smoothing (Holt 2004; Winters 1960; Gardner Jr and McKenzie 1985) and autoregression models (Makridakis, Spiliotis, and Assimakopoulos 2020). With the advances in deep learning, recent works are heavily invested in ensemble models and sequence-to-sequence modeling such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997). In Computer Vision domain, Convolutional Neural Networks (CNN) (Ren et al. 2015; Ronneberger, Fischer, and Brox 2015) have shown prominence in learning local patterns which are suitable for modeling short-term dependencies, although not suitable for modeling long-term dependencies due to limited receptive field. Most recently, Transformers (Vaswani et al. 2017), have shown great success in Nat-

ural Language Processing (NLP) (Devlin et al. 2018; Brown et al. 2020; Smith et al. 2022) domain, achieving superior performance on long-term dependencies modeling compared to LSTM.

Our contributions is the following: we leverage the advantages of CNNs and Transformers to model short-term and long-term dependencies in financial time series, as shown in Figure 1. In our experiments, we show the advantage of the proposed approach on intraday stock price prediction of S&P 500 constituents, outperforming statistical methods including Autoregressive Integrated Moving Average (ARIMA) and Exponential Moving Average (EMA) and a state-of-the-art deep learning-based autoregressive model DeepAR (Salinas et al. 2020).

Related Works

Time series forecasting

Typical forecasting techniques in the literature utilize statistical tools, such as, exponential smoothing (ETS) (Holt 2004; Winters 1960; Gardner Jr and McKenzie 1985) and autoregressive integrated moving average (ARIMA) (Makridakis, Spiliotis, and Assimakopoulos 2020), on numerical time series data for making one-step-ahead predictions. These predictions are then recursively fed into the future inputs to obtain multi-step forecasts. Multi-horizon forecasting methods such as (Taieb, Sorjamaa, and Bon-tempi 2010; Marcellino, Stock, and Watson 2006) directly generate simultaneous predictions for multiple pre-defined future time steps.

Machine learning and deep learning based approaches

Machine learning (ML) approaches have shown to improve performance by addressing high-dimensional and non-linear feature interactions in a model-free way. These methods include tree-based algorithms, ensemble methods, neural network, autoregression and recurrent neural networks (Hastie, Tibshirani, and Friedman 2001). More recent works have applied Deep learning (DL) methods on numeric time series data (Bao, Yue, and Rao 2017; Gensler et al. 2016; Romeu et al. 2015; Sagheer and Kotb 2019; Sutskever, Vinyals, and Le 2014). DL automates the process of feature extraction and eliminates the need for domain expertise.

Since the introduction of transformers (Vaswani et al. 2017), they have become the state of the art model to im-

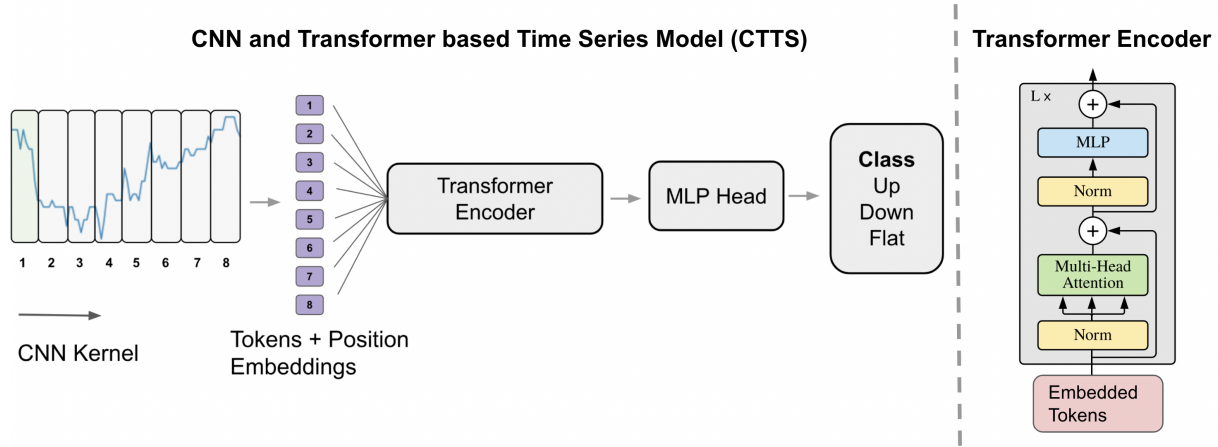


Figure 1: Overview of the proposed approach. Quick peak of the transformer encoder architecture on the right (Dosovitskiy et al. 2020)

prove the performance of NLP applications. The commonly used approach is to pre-train on a large dataset and then fine-tune on a smaller task-specific dataset (Devlin et al. 2018). Transformers leverage from multi-headed self-attention and replace the recurrent layers most commonly used in encoder-decoder architectures. In contrast to RNNs and LSTMs where the input data is sequentially processed, transformers bypass the recursion to ingest all inputs at once; thus, transformers allow for parallel computations to reduce training time and do not suffer from long-term memory dependency issues.

The remainder of the paper is organized as follows. We discuss our proposed methodology for time series modeling. Further, we discuss our benchmark baseline models along with the performance evaluation metrics and report the experimental results. Finally, we highlight some concluding remarks and future directions for this study.

Method

Our proposed method is called **CNN and Transformer based time series modeling (CTTS)** as shown in overview Figure 1.

Preprocessing

We used standard min-max scaling to standardize each input time series within $[0, 1]$. Given a raw stock prices time series \mathbf{x} , the standardized time series were calculated as

$$\mathbf{x}_{standardized} = \frac{\mathbf{x} - \min(\mathbf{x})}{\max(\mathbf{x}) - \min(\mathbf{x})}$$

This $\mathbf{x}_{standardized}$ is then passed into our model to learn the sign of the change in the very next time step, which is defined as our prediction target.

Forecasting

As shown in Figure 1, we use 1D CNN kernels to convolute through a time series, projecting each local window into

an embedding vector that we call a token. Each token carries the short-term patterns of the time series. we then add positional embedding (Vaswani et al. 2017) to the token and pass that through a Transformer model to learn the long-term dependencies between these tokens. The transformer model outputs a latent embedding vector of the time series, which is then passed through a Multilayer Perceptron (MLP) with softmax activation in the end to generate sign classification outputs. The output is in the form of probability for each class (up, down, flat), where the probability for all 3 classes sum up to 1.

Experiments

In our experiments, we benchmarked CTTS - our proposed method against 4 methods. 1) DeepAR - a state-of-the-art autoregressive recurrent neural networks-based time series forecasting method, 2) AutoRegressive Integrated Moving Average (ARIMA), 3) Exponential Moving Average (EMA) and 4) naive constant class predictions. We benchmarked the performance using sign prediction accuracy and thresholded version of the sign prediction accuracy (discussed later) using our model prediction probabilities. We ran our experiments on a Linux machine with 8 16GB NVIDIA T4 Tensor Core GPUs, and using PyTorch v1.0.0 DL platform in Python 3.6. In all models, we set a fixed random seed for reproducible results.

Experimental setup

Data We used the intraday stock prices of S&P 500 constituent stocks obtained from licensed Bloomberg data service (blo 2022). The data was sampled at 1 minute interval for the year 2019 (52 weeks, each week has 5 trading days). For every stock, we sampled 7 time series for each week. Data from the first three quarters (weeks 1 to 39) were used for training and validation. Data was randomly split and 80% was used for training and the remaining 20% for validation.

Method	2-class \uparrow	2-class* \uparrow	3-class \uparrow	3-class* \uparrow
EMA	53.2%	59.9%	39.5%	41.7%
ARIMA	50.9%	51.8%	37.5%	38.4%
DeepAR	51.1%	53.6%	37.4%	38.7%
CTTS	56.7%	66.8%	44.1%	55.2%

Table 1: Summary of sign accuracy over the last quarter of 2019. Our proposed method CTTS outperforms the baselines DeepAR, ARIMA and EMA. 2/3-class* refers to the thresholded version of the sign accuracy. The distribution of signs in the ground truth test set are as follows: price goes up | down | remains flat: 37.1% | 36.5% | 26.4%, respectively. If naively predicting the majority class (up) for all time series, the sign accuracy will only be 37.1%.

We had around 507K training and 117K validation samples. Data from the last quarter (weeks 40 to 52) was used for testing, totaling 209K time series. **For each time series, the first 80 time steps (input) were used to forecast the sign of price change at the 81st time step (target).** The overall test set performance is denoted by the aggregate for the evaluation metrics over the test set.

CTTS In our experiments, we used cross entropy loss, and AdamW (Loshchilov and Hutter 2017) optimizer with batch size of 64 and max epoch of 100. we used CNN kernel size of 16 and stride 8. Transformer with depth 4 and 4 self-attention heads, with an embedding dimension of 128, and a drop rate of 0.3 to prevent overfitting.

Baselines

DeepAR DeepAR forecasting algorithm is a supervised learning algorithm for forecasting 1D time series using autoregressive recurrent neural networks. DeepAR benefits from training a single model jointly over all of the time series. We used a batch size of 128, Adam optimizer, and the normal distribution loss. The model was trained for a maximum of 300 epochs, with early stopping mechanism set to a patience of 15. The base learning rate was $1e-3$, adjusted by learning rate scheduler with decay factor 0.1 and a patience of 5. We also used dropout with probability 0.1. DeepAR generated multiple (200) samples of the prediction target for each time series and we defined the prediction probability over the three classes as the proportion of samples predicted per class.

ARIMA Autoregressive Integrated Moving Average (ARIMA) models capture autocorrelations in the data using a combination approach of autoregressive model, moving average model, and differencing (Wilks 2011). We compared the continuous valued ARIMA forecasts with the last known price in the input data to define the predicted sign, and the corresponding prediction probability was defined as the percentage of the absolute delta between the predicted and the last known price with respect to the standard deviation of the past 80 data points, capped by 1.

EMA An exponential moving average (EMA) is a type of moving average that places a greater weight and significance

on the most recent data points. We used the "estimated" initialization method, which treats the initial values like parameters, and chooses them to minimize the sum of squared errors. Similar to ARIMA, we delta between the predicted and the last known price to define the predicted sign, and the corresponding prediction probability.

Constant Class Prediction We tested the three naive baselines where we always predict a constant sign i.e., either the predicted price always goes up, down, or remains flat. The prediction probabilities for these were set to 1 for the winning class, and 0 for the remaining two classes.

Metric

We evaluated all our models on 3-class [class 1-price goes up, class 2-price goes down, class 3-price remains flat] and 2-class [class 1-price goes up or remains flat, class 2-price goes down]. We used the averaged sign prediction accuracy over all test samples to evaluate our models. Higher accuracy implies more accurate forecasts.

Further, we evaluate a thresholded version of the sign accuracy, where we defined our threshold as the 75th percentile of all predicted probabilities of dominating classes. We retain only the samples that exceed the threshold and compute the sign prediction accuracy over these retained samples.

Results & Discussion

We summarize the quantitative benchmark results from the sign accuracy in Table 1. We show the sign accuracy for both 2-class and 3-class tasks as explained above. Note that random guess in 2-class task leads to 50% of accuracy, and random guess in 3-class task leads to 33% of accuracy. As shown in the 2-class and 3-class columns of Table 1, CTTS outperformed all baseline methods in both cases. This demonstrates the benefit of combining CNNs and transformers for time series forecasting.

Further, we evaluate a thresholded version of the sign accuracy, where we defined our threshold as the 75th percentile of all predicted probabilities of dominating classes. We retain only the samples that exceed the threshold and compute the sign prediction accuracy over these retained samples. As shown in 2-class* and 3-class* columns of Table 1, the accuracy after thresholding over the prediction probability has increased for all methods. In addition, the gain of accuracy increase is the most for our proposed method CTTS. This shows that the class probabilities that CTTS outputs are reliable. Specifically, high-confidence predictions from CTTS are often correct, thus thresholding filters out erroneous low-confidence predictions, leading to a boost in the sign accuracy.

Another highlight is the significant gap that CTTS has achieved for thresholded 3-class* accuracy compared to other baselines. This can be harnessed in the future for trading decision-making. For example, a straightforward trading decision can be buy/sell/hold stocks when the predicted class is up/down/flat, respectively. Given that CTTS's predicted probabilities are reliable as discussed earlier, the amount of stock shares to buy/sell/hold can depend on the

predicted probability, the higher the probability, the more shares to consider.

Conclusion

In this paper, we tackle the challenging problem of time series forecasting of stock prices in the financial domain. In this paper, we demonstrated the combined power of CNN and Transformer to model both short-term and long-term dependencies within a time series. In our experiments over intraday stock price of S&P 500 constituents in year 2019, we demonstrated the success of the proposed method CTTS in comparison to ARIMA, EMA, and DeepAR, as well as the potential for using this method for downstream trading decisions in the future.

References

2022. Bloomberg Market Data. <https://www.bloomberg.com/professional/product/market-data/>. Accessed: 2022-08-16.
- Bao, W.; Yue, J.; and Rao, Y. 2017. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PloS one*, 12(7): e0180944.
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Gardner Jr, E. S.; and McKenzie, E. 1985. Forecasting trends in time series. *Management Science*, 31(10): 1237–1246.
- Gensler, A.; Henze, J.; Sick, B.; and Raabe, N. 2016. Deep Learning for solar power forecasting—An approach using AutoEncoder and LSTM Neural Networks. In *2016 IEEE international conference on systems, man, and cybernetics (SMC)*, 002858–002865. IEEE.
- Hastie, T.; Tibshirani, R.; and Friedman, J. 2001. The elements of statistical learning. Springer series in statistics. New York, NY, USA.
- Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.
- Holt, C. C. 2004. Forecasting seasonals and trends by exponentially weighted moving averages. *International journal of forecasting*, 20(1): 5–10.
- Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Makridakis, S.; Spiliotis, E.; and Assimakopoulos, V. 2020. The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1): 54–74.
- Marcellino, M.; Stock, J. H.; and Watson, M. W. 2006. A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of econometrics*, 135(1-2): 499–526.
- Pedersen, L. H. 2019. *Efficiently inefficient: how smart money invests and market prices are determined*. Princeton University Press.
- Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.
- Romeu, P.; Zamora-Martínez, F.; Botella-Rocamora, P.; and Pardo, J. 2015. Stacked denoising auto-encoders for short-term time series forecasting. In *Artificial Neural Networks*, 463–486. Springer.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.
- Sagheer, A.; and Kotb, M. 2019. Time series forecasting of petroleum production using deep LSTM recurrent networks. *Neurocomputing*, 323: 203–213.
- Salinas, D.; Flunkert, V.; Gasthaus, J.; and Januschowski, T. 2020. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191.
- Smith, S.; Patwary, M.; Norick, B.; LeGresley, P.; Rajbhandari, S.; Casper, J.; Liu, Z.; Prabhumoye, S.; Zerveas, G.; Korthikanti, V.; et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.
- Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.
- Taieb, S. B.; Sorjamaa, A.; and Bontempi, G. 2010. Multiple-output modeling for multi-step-ahead time series forecasting. *Neurocomputing*, 73(10-12): 1950–1957.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wilks, D. S. 2011. *Statistical methods in the atmospheric sciences*, volume 100. Academic press.
- Winters, P. R. 1960. Forecasting sales by exponentially weighted moving averages. *Management science*, 6(3): 324–342.