Building a Time Series Vision Transformer with Small, High-resolution Datasets

Keywords: ViT, time series, transformer, high-resolution, machine learning

Extended Abstract

Motivation. Market crashes have negative consequences for economies around the world. However, there is a lack of focus on applying high-temporal knowledge of markets to predict market crashes in existing early-warning systems, resulting in potential incompleteness in the information processing of market data during predictions. In this talk, we will focus on demonstrating our progress in building a transformer model to forecast changes in stock market daily return correlations over multiple temporal periods.

Approach and Methodology. We obtained cross-correlations of the daily returns of 457 stocks of S&P 500 stocks over 1575 trading days from 3 January 2017 to 28 April 2023 using the quasi-differentiation method [1]. Then, we performed topological data analysis (TDA) to obtain distance matrices of the cross-correlations at various thresholds. The resulting multiscale datasets are divided into 9 walk-forward time-series cross-validation folds as shown in Figure 1a. Inspired by advancements in vision transformer developments [2], we constructed a decoder-only transformer model to autoregressively generate next-period distance matrices. Due to the small number of datasets, we incorporate shifted-patch tokenization and locality self-attention to account for the limited data [3]. Finally, we performed interpretability checks to ensure that the transformer model is learning correctly.

Results. We obtained very encouraging results. First, as shown in Fig 1b, the validation results suggest that later folds, covering later periods, yield better training results than the others. We suspect that this is partly influenced by the removal of old delisted stocks from our datasets, which causes information to be lost in early training periods. Secondly, we observed that the diagonal elements consistently received less attention than their off-diagonal counterparts, pre- and post-patching, which is expected since the cross-correlations between identical stocks are always 1. Furthermore, we studied the heatmap of the learned attention temperature and observed that the attentions manage to capture structures within the distance matrices, such as fewer attention on the diagonal elements compared to the off-diagonal ones, as shown in Fig 1d.

Conclusions and Outlook. We have successfully developed a transformer model to provide insight into the cross-correlation changes in the subsequent 6 trading months. We believe that our model may be further improved via hyperparameter tuning such as increasing the depth of the decoder layers and the incorporation of multiple length scales of cross-correlation threshold datasets. We will perform more interpretability checks to increase the robustness of our transformer models against unseen data structures.

References

[1] Siew Ann Cheong, Zheng Tien Kang, and Peter Tsung-Wen Yen. Quasi-differentiation and its applications to noisy time series data from complex systems. *Scientific Reports*, 15(1):39080, Nov 2025.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[3] Seunghoon Lee, Seunghyun Lee, and Byung Cheol Song. Improving vision transformers to learn small-size dataset from scratch. *IEEE Access*, 10:123212–123224, 2022.

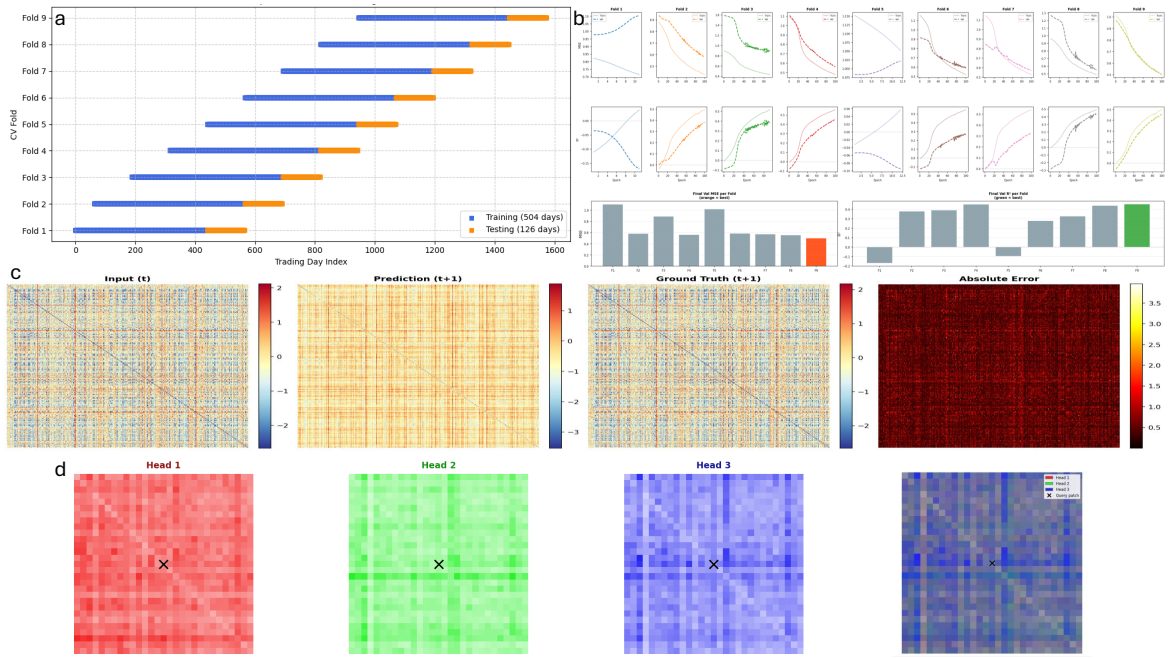

Figure 1: **Data splitting and Training results.** a) The cross-correlation threshold datasets comprising 1575 trading days from 3 Jan 2017 to 28 April 2023 are split into 9 folds via *sklearn.model_selection.TimeSeriesSplit* method. Each fold covers 2 trading years of training data (504 trading days) and 6 trading months of validation data (126 trading days). b) Multi-fold CV results for w = 35 days. Fold 9 has the lowest MSE and highest R^2 value.
c) Snapshot of a prediction instance using the first day of the last validation set (fold 9).
d) Individual head attention maps with color contrasts (w = 35 days) from the last decoder block (6). Central patch is marked ”X”. RGB colors are chosen to allow better magnitude contrast within heads. The rightmost plot combines all three attention heads to form a composite attention map.)