



LSTM	#Layers=2, $D = 256$, $lr = 10^{-3}$
GRU	#Layers=2, $D = 256$, $lr = 10^{-4}$
TCN	#Layers=3, #channels=32, kernel size=3, $lr = 10^{-4}$
Transformer	#Layers=1, $D = 256$, $lr = 10^{-4}$
GAT	#Layers in backbone LSTM=2, $D = 512$, $lr = 10^{-4}$, #Layers in graph encoder=1
DTML	#Layers in backbone LSTM=2, $D = 256$, $\beta = 0.1$, $lr = 10^{-5}$.

Table 1: Baseline implementation details.

Baselines

Baseline implementation details.

The hyperparameters of baselines on CSI300 are in Table 1. On CSI800, the hyperparameters are the same except for DTML, $\beta = 1.0$.

Discussion on graph-based baselines.

We use a fully-connected graph in the correlation module of reported GAT baseline, as suggested by Qlib. Although most previous works [1,2,3] in the literature leverage industry graphs, where edges are connected between companies in the same industry, all experimented graph-based methods [1,3,5] report inferior results with industry graphs on the Chinese market, as in Figure (a). As argued in our paper, the predefined graphs mostly describes long-standing relationships rather than real-time proximity of stock prices. Alternatively, the fully-connected graph allows pairwise correlation calculation without strong human prior. With fully-connected graphs, MASTER still outperforms ESTIMATE, RSR, and GAT, showing the superiority of our transformer-based architecture.

Additional Experiments

Realistic Assessment

MASTER ranks stocks by profitability while the realistic profit is also affected by the trading strategy. Figure (b) reports AR on CSI300 with/without cost under widely-adopted *top-30*, *drop-N* strategy, with $N=5, 10$ to constrain the turnover rate on different levels.

Aggregation Order

In stock prediction, in order to capture correlations between any $(stock_1, time_1)$ and $(stock_2, time_2)$ pairs, we use intra-stock (temporal) aggregation followed by inter-stock aggregation to break down the large and complex attention field. Figure (c) reports the comparison with reversed aggregation order.

Prediction Interval

The prediction interval d determines the labels. Smaller d makes the labels more random and harder to learn, while larger d may miss out on immediate profits. Figure (d) report the AR when d varies on CSI300. We set $d = 5$ so that most models can gain their best returns, while it can also be set according to actual need.

Lookback Window Length

The lookback window length τ is a hyperparameter. Figure (e) report the AR when τ varies on CSI300. Interestingly, we found that longer lookback window not necessarily improve the model performance. We set $\tau=8$ so that most models can gain their best returns.

Reference

- [1] Huynh T. T., et al. Efficient integration of multi-order dynamics and internal dynamics in stock movement prediction. In WSDM 2023.
- [2] Sawhney, R., et al. Stock selection via spatiotemporal hypergraph attention network: A learning to rank approach. In AAAI 2021.
- [3] Feng, F., et al. Temporal relational ranking for stock prediction. In TOIS 2019.
- [4] Yang, X., et al. Qlib: An ai-oriented quantitative investment platform. arXiv preprint arXiv:2009.11189 (2020).
- [5] Veličković, P., et al. Graph attention networks. arXiv preprint arXiv:1710.10903 (2017).