MASTER: Market-Guided Stock Transformer for Stock Price Prediction

Tong Li, Zhaoyang Liu, Yanyan Shen, Xue Wang, Haokun Chen, Sen Huang

Shanghai Jiao Tong University
Alibaba Group

The AAAI Conference on Artificial Intelligence

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Alibaba Group

Stock Price Forecasting

- Stock price forecasting uses the historical data of stocks to predict their future trends.
  - Profitable stock investment.
  - Close price of stock u at day t: $c_{u,t}$
  - Return ratio, the relative change of close price in d days:

$$\tilde{r}_u = \frac{c_{u,\tau+d} - c_{u,\tau+1}}{c_{u,\tau+1}}$$

- Stock price patterns are intricate.
  - Multiple factors: macroeconomic factors, capital flows, investor sentiments …
  - The mixing of factors interweaves the stock market as a correlated network.

Modeling Stock Correlation

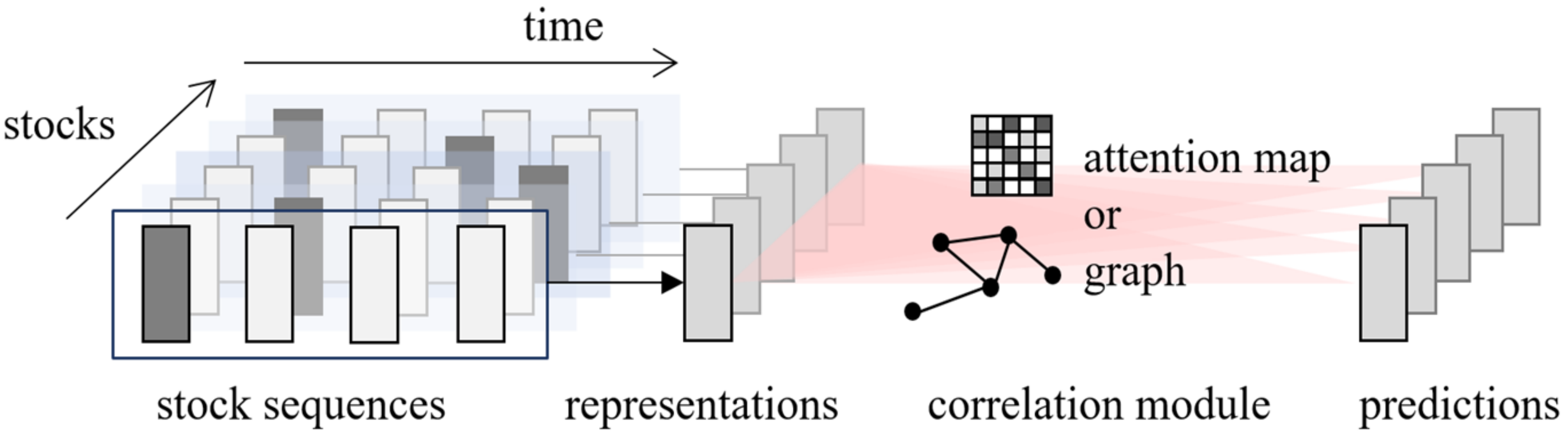
1. Static : Predefined concepts, relationships or rules.

- Example:
  - Industry graph - stocks in the same industry are connected to each other.
- relationship ≠ real-time correlation
- not generalizable when events such as company listing, delisting or change in main business happen.

2. Dynamic: Attention mechanism.

- Data-driven, more flexible, and applicable to the time-varying stock sets.

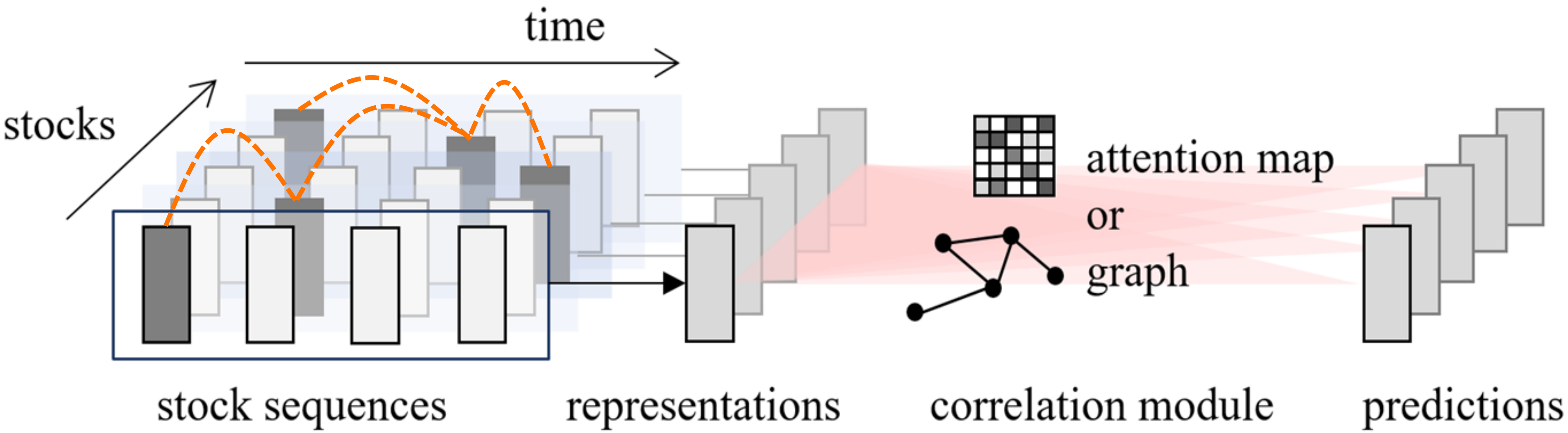
Framework of Existing Works



1. Use sequential encoder to summarize the historical sequence of stock features and obtain stock representation.
2. Establish overall stock correlation and aggregate information to refine each stock representation.

Limitation: They cannot model the realistic stock correlation.

Framework of Existing Works



1. Use sequential encoder to summarize the historical sequence of stock features and obtain stock representation.
2. Establish overall stock correlation and aggregate information to refine each stock representation.

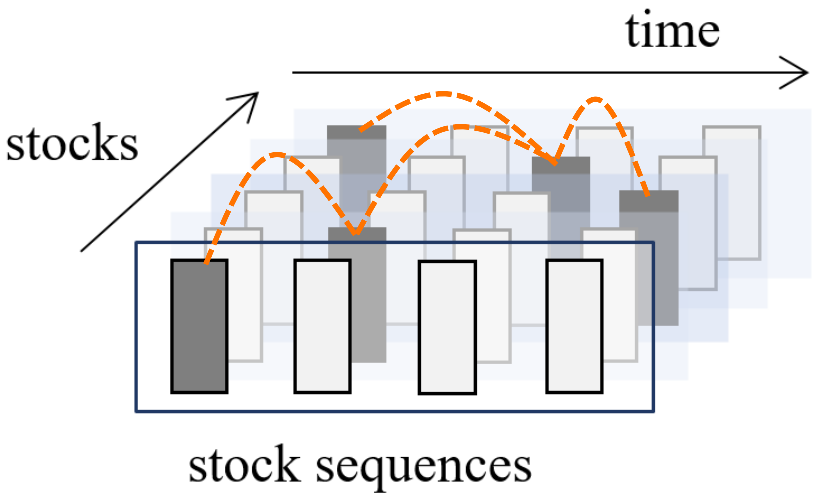
Limitation: They cannot model the realistic stock correlation.

Realistic Stock Correlation

- The dominating factors of stock prices constantly change.
- Different stocks may react to the same factor with different delays.

Instead of holding true through the whole look back window, realistic stock correlation:

1. **Momentary**: highly dynamic
2. **Cross-time**: residing in misaligned time steps.



Example:

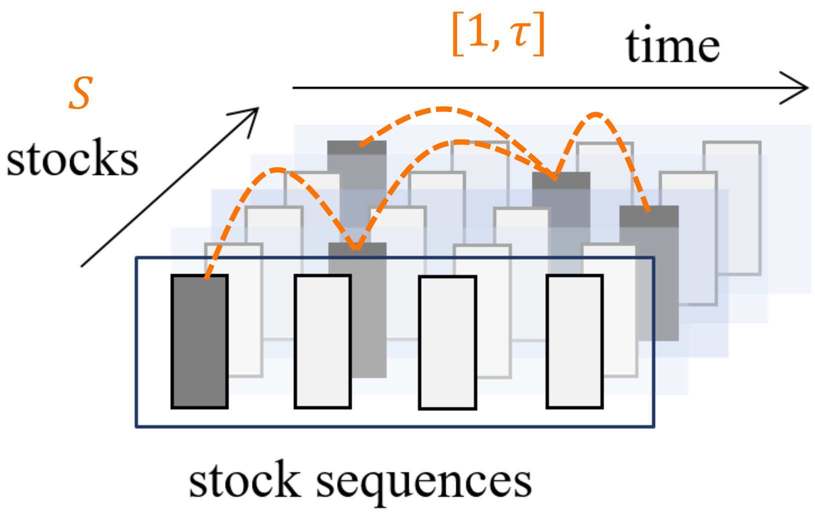
Upstream companies’ stock prices may react faster to a shortage of raw materials than those of downstream companies.

Difficulties: Complex Attention Field

To simulate the correlation, calculate pair-wise attention among all $\tau \times S$ feature vectors.

1. Large and complex attention field vs. stock data hunger

- Limited observation: around 250 trading days per year
- Clustering approaches are sensitive to initialization, unsuitable in stock domain.
- **Our solution**: aggregate information from different time steps and other stocks alternatively.



Difficulties: Market Variation

2. The stock correlation is different under varying market status.

- Example: in a bull market, the correlation are more significant due to investors' optimism.
- With market variation, the features come into effect and expire.
- Traditional investors repeatedly conduct statistical examination to select features.
- **Our solution**: incorporate the market information to perform automatic feature selection.

Preliminaries

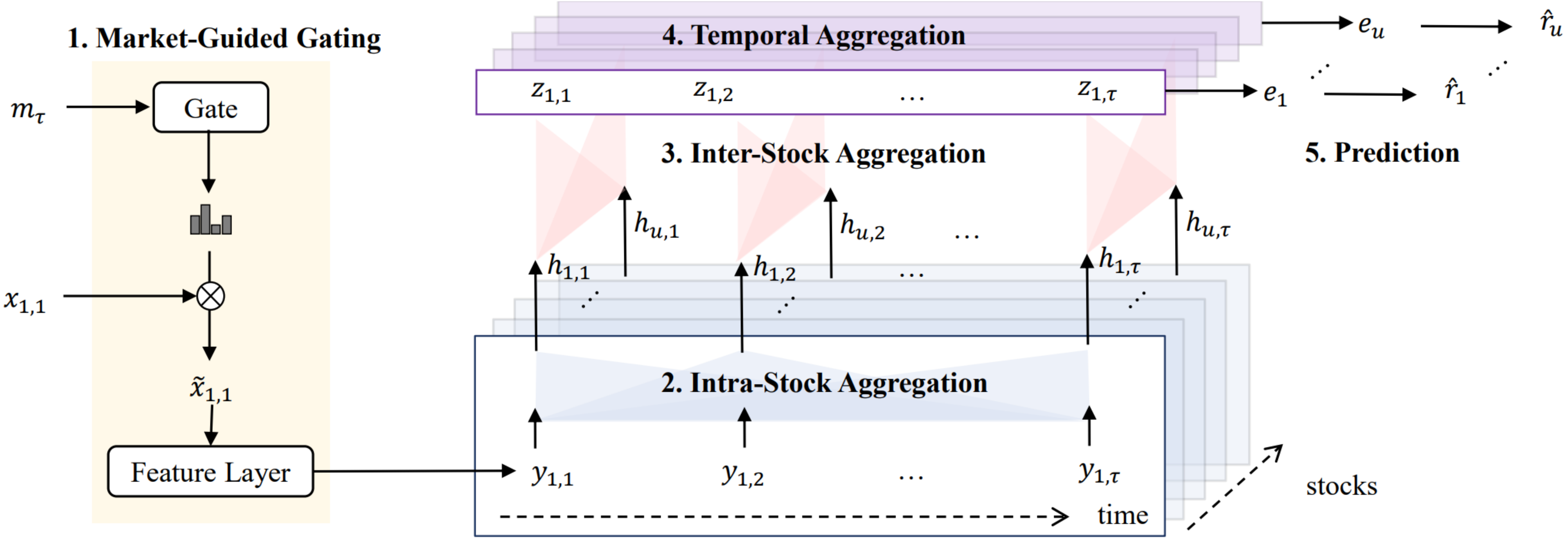
Input:

- Stock feature sequences $\{x_{u,t}\}_{u \in S, t \in [1,\tau]}$, where $x_{\{u,t\}} \in \boldsymbol{R}^{F}$
- Market status vector $m_\tau \in \boldsymbol{R}^{F'}$
    - Market index price (historical and current)
    - Market index trading volume (historical and current)

Output:

- Normalized Return Ratios $\{r_u\}_{u \in S}$, $r_u = \mathrm{Norm}_S(\tilde{r}_u)$
    - Encode the labels with ranking information.

MASTER: Overview



MASTER: Market-Guided Gating

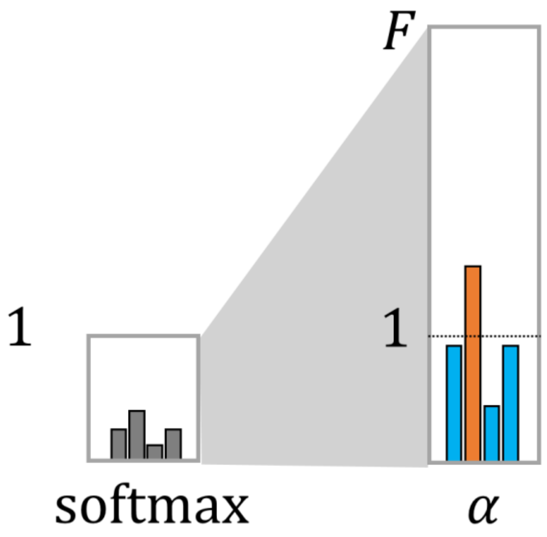
Input: m_τ

Output: $\alpha, |\alpha| = F$, one scaling coefficient for each feature.

$$\alpha(m_\tau) = F \cdot \mathrm{softmax}_\beta(W_\alpha m_\tau + b_\alpha)$$

- Softmax compels a competition among features to distinguish effective ones.
- β : temperature parameters.
- F : adjust the coefficient range to be $[0, F]$
  - the coefficient can either enlarge or shrink the magnitude.

$$\tilde{x}_{u,t} = \alpha(m_\tau) \circ x_{u,t}$$



MASTER: Intra-Stock Aggregation

We perform intra-stock aggregation first.

- Smaller attention field.
- The feature of a single stock is distributed simpler.

(1) For each stock u, we gather its feature sequence, and encode each feature with

$$Y_u = ||_{t\in[1,\tau]}\mathrm{LayerNorm}(f(\tilde{x}_{u,t}) + p_t). \qquad p\text{: positional codes.}$$

(2) Transform Y_u into Q_u^1, K_u^1, V_u^1 .

(3) Compute multi-head attention and send to feed forward layers.

$$H_u^1 = ||_{t\in[1,\tau]}h_{u,t} = \mathrm{FFN}^1(\mathrm{MHA}^1(Q_u^1, K_u^1, V_u^1) + Y_u)$$

MASTER: Inter-Stock Aggregation

(1) For each time step t, we gather the embedding of all stocks

$$H_t^2 = ||_{u \in S} h_{u,t}$$

(2) Transform H_t^2 into Q_t^2, K_t^2, V_t^2 .

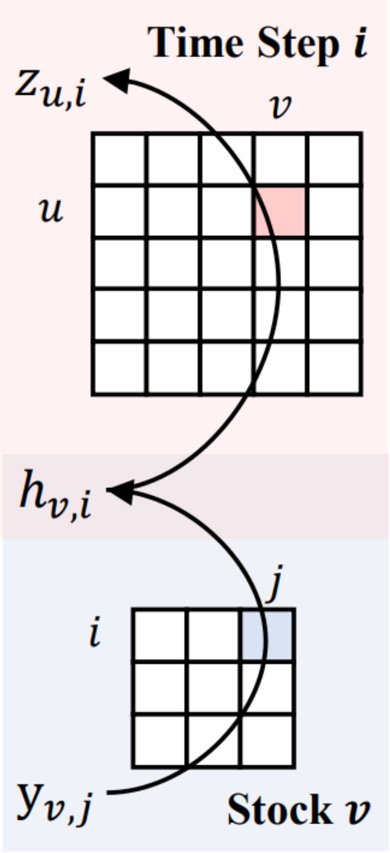
(3) Compute multi-head attention and send to feed forward layers.

$$Z_t = ||_{u \in S} z_{u,t} = \mathrm{FFN}^2(\mathrm{MHA}^2(Q_t^2, K_t^2, V_t^2) + H_t^2)$$

Correlation from (v, j) to (u, i):

1. The local details of $y_{v,j}$ is conveyed to $h_{v,i}$ by the intra-stock aggregation of stock v.
2. Transmit $h_{v,i}$ to $z_{u,i}$ by inter-stock aggregation at time step i.

cross-time correlation



MASTER: Temporal Aggregation & Prediction

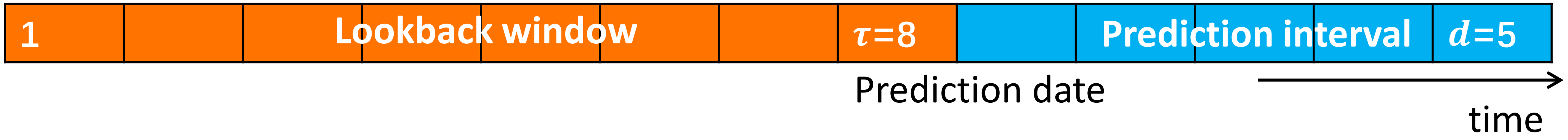
- For each stock u, MASTER produces a series of temporal embedding $z_{u,t}, t \in [1, \tau]$.
- We use the latest temporal embedding to query from others, and summarize them into the comprehensive stock embedding:

$$e_u = \sum_{t\in[1,\tau]} \lambda_{u,t} z_{u,t}\,, \qquad \lambda_{u,t} = \frac{\exp(z_{u,t}^T W_\lambda z_{u,\tau})}{\sum_{\mathrm{i}\in[1,\tau]} \exp(z_{u,i}^T W_\lambda z_{u,\tau})}$$

- Regression: $\hat{r}_u = g(e_u)$
- Optimization: $L = \sum_{u\in S} \mathrm{MSE}(r_u, \hat{r}_u)$

Experiments: Settings

- **Chinese market， Stock sets:** CSI300， CSI800
- **Dataset Split:**
  - **Training -** 2008 Q1~2020 Q1, **Validation** - 2020 Q2, **Test** – 2020 Q3, 2022 Q4
- **Prediction Setting**



- **Baselines:** XGBoost, LSTM,GRU, TCN, Transformer, GAT, DTML
- **Evaluation metrics:**

Ranking-based - IC, ICIR, RankIC, RankICIR, **Portfolio-based -** AR, IR.

Experiments: Overall Performance

Table 1: Overall performance comparison. The best results are in bold and the second-best results are underlined. And * denotes statistically significant improvement (measured by t-test with p-value < 0.01) over all baselines.

Dataset	Model	IC	ICIR	RankIC	RankICIR	AR	IR
CSI300	XGBoost	0.051 ± 0.001	0.37 ± 0.01	0.050 ± 0.001	0.36 ± 0.01	$\underline{0.23 \pm 0.03}$	1.9 ± 0.3
	LSTM	0.049 ± 0.001	$\underline{0.41 \pm 0.01}$	0.051 ± 0.002	0.41 ± 0.03	0.20 ± 0.04	$\underline{2.0 \pm 0.4}$
	GRU	0.052 ± 0.004	0.35 ± 0.04	0.052 ± 0.005	0.34 ± 0.04	0.19 ± 0.04	1.5 ± 0.3
	TCN	0.050 ± 0.002	0.33 ± 0.04	0.049 ± 0.002	0.31 ± 0.04	0.18 ± 0.05	1.4 ± 0.5
	Transformer	0.047 ± 0.007	0.39 ± 0.04	0.051 ± 0.002	$\underline{0.42 \pm 0.04}$	0.22 ± 0.06	2.0 ± 0.4
	GAT	$\underline{0.054 \pm 0.002}$	0.36 ± 0.02	0.041 ± 0.002	0.25 ± 0.02	0.19 ± 0.03	1.3 ± 0.3
	DTML	0.049 ± 0.006	0.33 ± 0.04	$\underline{0.052 \pm 0.005}$	0.33 ± 0.04	0.21 ± 0.03	1.7 ± 0.3
	MASTER	$\mathbf{0.064^* \pm 0.006}$	$\mathbf{0.42 \pm 0.04}$	$\mathbf{0.076^* \pm 0.005}$	$\mathbf{0.49 \pm 0.04}$	$\mathbf{0.27 \pm 0.05}$	$\mathbf{2.4 \pm 0.4}$
CSI800	XGBoost	0.040 ± 0.000	0.37 ± 0.01	0.047 ± 0.000	0.42 ± 0.01	0.08 ± 0.02	0.6 ± 0.2
	LSTM	0.028 ± 0.002	0.32 ± 0.02	0.039 ± 0.002	0.41 ± 0.03	0.09 ± 0.02	0.9 ± 0.2
	GRU	0.039 ± 0.002	0.36 ± 0.05	0.044 ± 0.003	0.39 ± 0.07	0.07 ± 0.04	0.6 ± 0.3
	TCN	0.038 ± 0.002	0.33 ± 0.04	0.045 ± 0.002	0.38 ± 0.05	0.05 ± 0.04	0.4 ± 0.3
	Transformer	0.040 ± 0.003	$\mathbf{0.43 \pm 0.03}$	0.048 ± 0.003	$\mathbf{0.51 \pm 0.05}$	0.13 ± 0.04	1.1 ± 0.3
	GAT	$\underline{0.043 \pm 0.002}$	0.39 ± 0.02	0.042 ± 0.002	0.35 ± 0.02	0.10 ± 0.04	0.7 ± 0.3
	DTML	0.039 ± 0.004	0.29 ± 0.03	$\underline{0.053 \pm 0.008}$	0.37 ± 0.06	$\underline{0.16 \pm 0.03}$	$\underline{1.3 \pm 0.2}$
	MASTER	$\mathbf{0.052^* \pm 0.006}$	$\underline{0.40 \pm 0.06}$	$\mathbf{0.066 \pm 0.007}$	$\underline{0.48 \pm 0.06}$	$\mathbf{0.28^* \pm 0.02}$	$\mathbf{2.3^* \pm 0.3}$

Experiments: Stock Transformer Architecture

Table 2: Experiments on CSI300 to validate the effectiveness of proposed stock transformer architecture. The best results are in bold and the second-best results are underlined.

Model	IC	ICIR	RankIC	RankICIR	AR	IR
(MA)STER	$\mathbf{0.064 \pm 0.003}$	$\mathbf{0.43 \pm 0.02}$	$\mathbf{0.074 \pm 0.004}$	$\mathbf{0.48 \pm 0.04}$	$\mathbf{0.25 \pm 0.03}$	$\mathbf{2.1 \pm 0.3}$
(MA)STER-Bi	$\underline{0.058 \pm 0.005}$	$\underline{0.38 \pm 0.04}$	$\underline{0.066 \pm 0.008}$	$\underline{0.41 \pm 0.05}$	$\underline{0.19 \pm 0.03}$	1.6 ± 0.2
Naive	0.041 ± 0.008	0.30 ± 0.05	0.046 ± 0.007	0.32 ± 0.04	0.18 ± 0.05	1.6 ± 0.6
Clustering	0.044 ± 0.003	0.36 ± 0.02	0.049 ± 0.005	0.39 ± 0.04	0.18 ± 0.04	$\underline{1.7 \pm 0.3}$

(MA)STER: An ablation of MASTER without the Market-Guided Gating.

(MA)STER-Bi: Substitute the transformer layer with Bi-LSTM.

Naïve: Directly compute pair-wise attention among $\tau \times S$ feature vectors.

Clustering: Apply Local Sensitive Hashing to break down the attention field.

Experiments: Market-Guided Gating

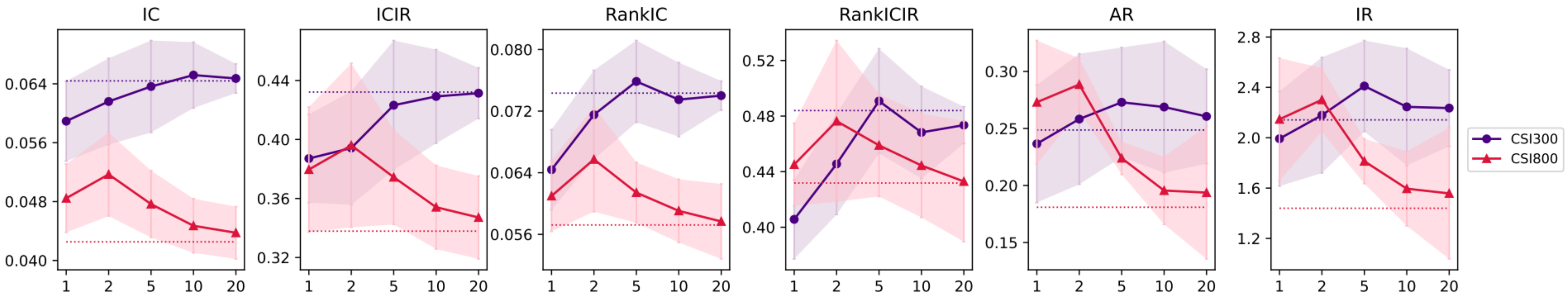


Figure 4: MASTER performance with varying β. The horizontal dash lines are performance without market-guided gating.

Gate temperature:

a smaller β forces a stronger feature selection while a larger β turns off the gating effect.

Experiments: Visualization of Attention Maps

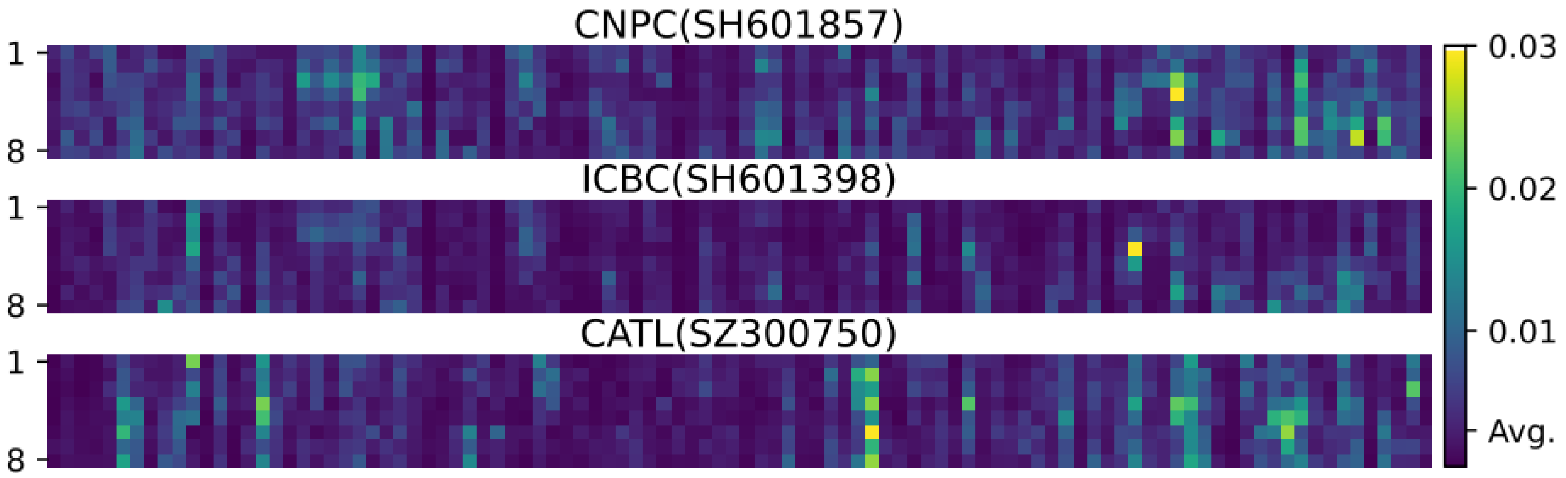


Figure 5: The correlation towards three target stocks on Aug 19th, 2022. The y-axis is time steps in the lookback window and the x-axis is source stocks. *Avg.* denotes the evenly distributed value.

Experiments: Visualization of Attention Maps

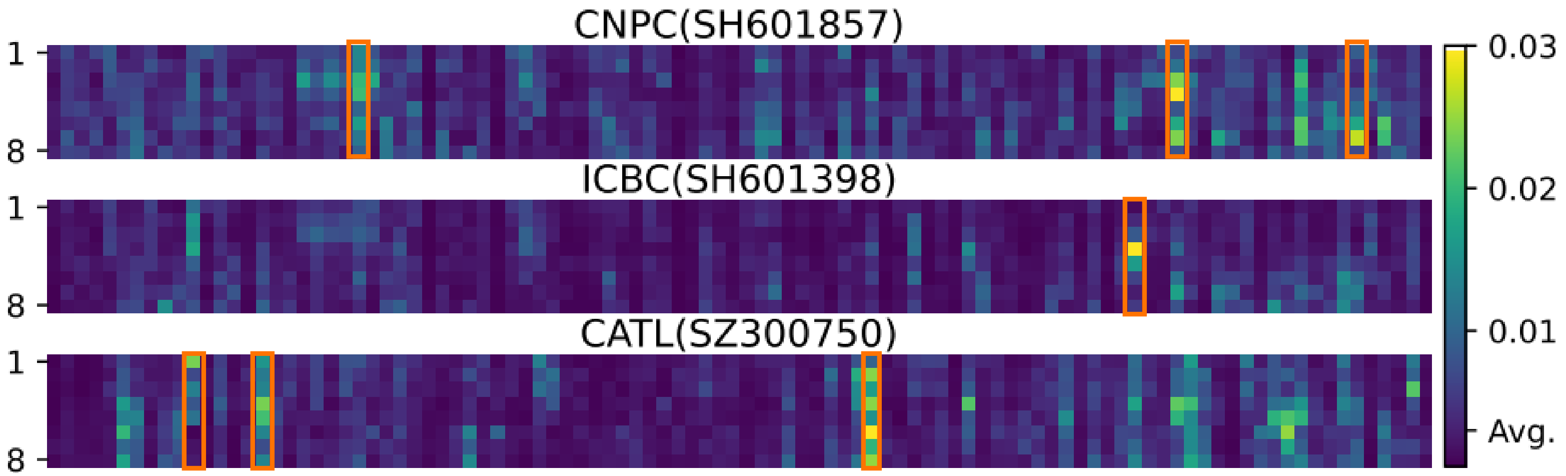


Figure 5: The correlation towards three target stocks on Aug 19th, 2022. The y-axis is time steps in the lookback window and the x-axis is source stocks. *Avg.* denotes the evenly distributed value.

Experiments: Visualization of Attention Maps

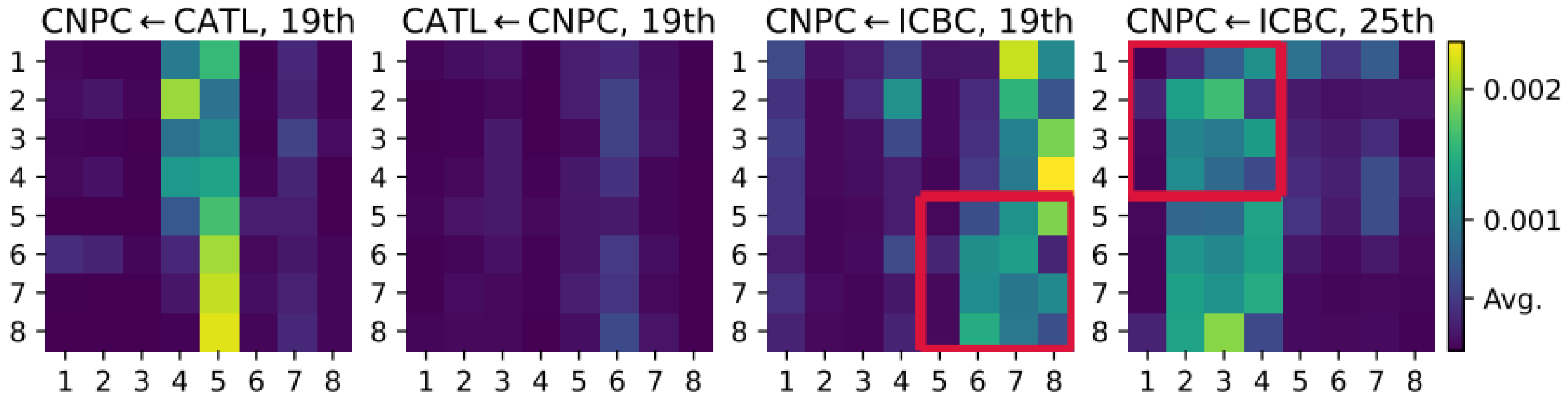


Figure 6: Cross-time correlation of stock pairs on Aug 19th and 25th, 2022. The x-axis is the source time steps and the y-axis is the target time steps.

Conclusion

- We introduce a novel method MASTER for stock price forecasting, which models the realistic stock correlation and guide feature selection with market information.
- Future work can explore to mine stock correlations of higher quality and study other uses of market information.
- Data & Code: github.com/SJTU-Quant/MASTER
- Email: 2017lt@sjtu.edu.cn

Thank you!