PREDICTION OF EARNING SURPRISE USING DEEP LEARNING TECHNIQUE

Quan Liu*
Bloomberg Enterprise Quants
{qliu268, entquant}@bloomberg.net

Liwen Ouyang
Bloomberg Enterprise Quants
{louyang11, entquant}@bloomberg.net

Gilbert Xu
Bloomberg APAC Broker Disclosed and Research
yxu202@bloomberg.net

Friday 24th June, 2022

ABSTRACT

Earnings expectations and earnings forecasts are central focus of the financial industry. Many investors rely on earnings performance to make investment decisions. Stock price adjusts according to a company's ability to increase earnings as well as to meet or beat analysts' consensus estimates. Betting on earning reports is a popular strategy which focuses on the earning surprise: the percent difference between the actual and estimated earnings per share. In this paper, we first backtest the historical performance of the earning surprise betting, assuming the earning surprise could have been known, and deem it a profitable strategy. Based on this conclusion, we investigate several machine learning model architectures to predict the earnings surprise using historical fundamental data along with Bloomberg Estimate EPS. We backtest the strategy again using the earnings surprise projected by different models and find that the attention-based model, a cutting-edge technology widely used in sequence modeling, yields the best performance.

* Corresponding author

1 Introduction

Earnings per share (EPS) is a critical measure of a company's profitability. It is widely used by investors to forecast the company's stock return in the short and long term. Historically, there has been financial research focusing on the relation between earnings and stock market returns (Ball and Brown, 1968; Beaver, 1968; Albrecht et al., 1977). Bernard and Thomas (1989) showed that extreme changes in quarterly earnings are more persistent than investors expect. This result inspired research on predictions for unexpected earnings.

Recently, more and more research focuses on using machine learning to predict earnings. Ou and Penman (1989) used a step-wise logit regression to predict the sign of future earnings changes and formed a profitable hedge portfolio which longed in firms predicted to have an increase in earnings and shorted in firms predicted to have a decrease in earnings. Elend et al. (2020) compared long short-term memory (Section 3.1) networks to temporal convolution network (Lea et al., 2016) in the prediction of future EPS.

Despite the increasing interests in combining machine learning with earnings prediction, to the best of our knowledge, previous research has mainly focused on developed markets. In our opinion, the emerging markets are less efficient, and thus the earning prediction should yield more promising results. Another interesting observation is that most researches focus on the EPS prediction itself, without any benchmark, making it difficult to gauge the EPS level across different companies.

To address the issues above, in this paper, we focus on earnings predictions for the China A-share market. Instead of looking at the EPS level itself, we incorporate the Bloomberg Estimate EPS (BEST EPS) as a benchmark and try to predict the earning surprise, which is defined as percentage difference between Actual EPS and BEST EPS. We backtest a long-only EPS surprise betting strategy on China A-share market and deem it profitable. Based on that, we build long-only strategies using different deep learning architectures and compare their performance against the benchmark strategy (Section 2.4). Results show that the strategy based on attention model (Section 3.2) consistently beats the benchmark strategy and gives good performance in the out-of-time testing period.

2 Study of Earning Surprise

2.1 Motivation

As discussed in Section 1, we choose to model the EPS surprise instead of EPS level. The definition of EPS surprise is:

$$\text{EPS Surprise} = \frac{\text{Actual EPS} - \text{Expected EPS}}{|\text{Expected EPS}|} \tag{2.1}$$

It reflects the company's profitability compared to some benchmark target. If there is a significant divergence between the actual and expected EPS, the stock price will adjust quickly in the short term. It is often observed that the stock price of company which beats/misses expectation rallies/plunges after earning announcement. If one can reasonably predict the EPS surprise, one could benefit from a strategy based on the prediction. In other words, it is the EPS surprise that drives the stock return in the short term rather than the EPS level. The EPS betting strategy we will explore in this paper is rooted from this assumption: *there is a strong correlation between EPS surprise and stock return after announcement*. In Section 2.3 and Section 2.4, we will test this assumption and try to build trading strategies based on it.

2.2 Benchmark EPS

An important input for the EPS surprise is the Expected EPS in Eq 2.1. It should be a widely used and well recognized benchmark on the market. In this paper, we use Bloomberg Estimate (BEST) EPS, which reflects the consensus estimate for adjusted earnings per share, from **Asian Bloomberg Estimates Data**. Bloomberg's Estimates dataset delivers three types of key forward-looking measures: consensus estimates, analyst recommendations, and company guidance. Consensus estimates fields include all key Income Statement, Balance Sheet and Cash Flow measures from Sales, EPS to Net Debt and Free Cash Flow. Both Generally Accepted Accounting Principles (GAAP) and Adjusted fields are available for Net Income and EPS. For selected fields, measures of high, low, medium, four-week change, and number of contributions are also provided for additional context on the data distribution. For the purpose of this paper, we focus on the consensus estimate of quarterly EPS. To make a fair comparison to the benchmark, we use the estimate comparable EPS adjusted (IS900) as the actual EPS. This field provides a Bloomberg calculated or company reported EPS that aligns with BEST EPS.

2.3 Wizard Strategy

To test our assumption in Section 2.1, we first calculated the correlation between EPS surprise as defined in Eq 2.1 and the post announcement stock returns in China A-share market. Since most companies announce the earning reports post market, we define the post announcement return as:

$$r_w^t = \frac{P_{\text{close}}^{t+1} - P_{\text{close}}^t}{P_{\text{close}}^t}, \tag{2.2}$$

where t represents the announcement date and t+1 is the day after announcement.

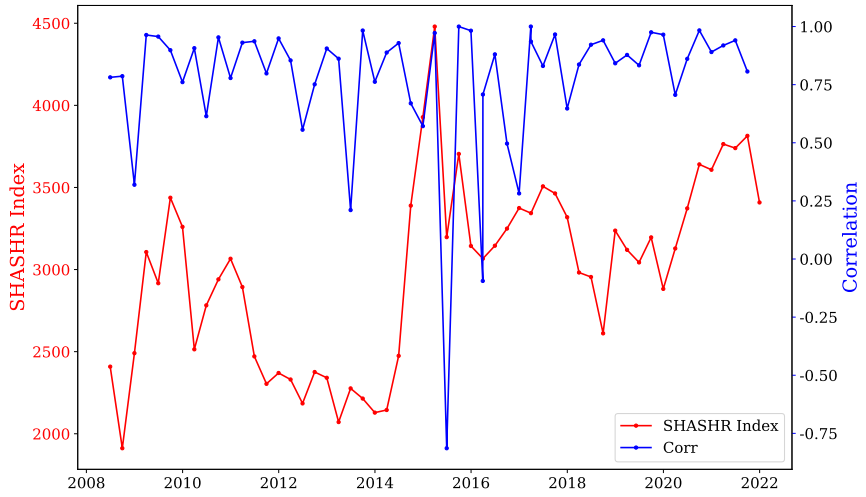

Figure 1: Correlation between EPS Surprise and Return

Fig 1 shows the correlation between EPS surprise and post announcement return for China A-share market along with the Shanghai Stock Exchange A-Share Index (SHASHR Index) level from 2008 Q2 to 2021 Q3. The average correlation across all periods is 0.77. This result validates our assumption in Section 2.1. An interesting observation is that the correlation is quite negative (-0.81) as of 2015 Q2. This period corresponds to the A-share market crash that happened in Aug 2015. It indicates that the assumption is generally true but may not hold under extreme market conditions.

Based on this finding, we construct a long-only strategy assuming we can perfectly predict the EPS surprise before the market close as of earning announcement date, as our wizard strategy:

- From 2014 Q1 to 2018 Q4, rank China A-share stocks based on the actual earning surprise within each quarter.
- Calculate the average threshold for the top one sixth: Calculate the top one sixth earning surprise for each of the 20 quarters and then calculate the mean of the 20. This quantile is chosen by maximizing the Sharpe ratio of the strategy.
- From 2019 Q1 to 2021 Q4, calculate the earning surprise for the universe of China A-share stocks, build a portfolio before market close as of the announcement date by (assuming we know the earning surprise in advance):
  - On the announcement day, buy the stock if the actual earning surprise is greater than the average historical top one sixth threshold.
  - If there are multiple stocks meet the condition, invest equally.
- Exit the positions before market close the day after earning announcement.

We backtest the performance of this strategy assuming:

- 3-month risk-free rate = 3%.
- Round-trip commission = 5 bps.
- No transaction cost (slippage, market impact etc).

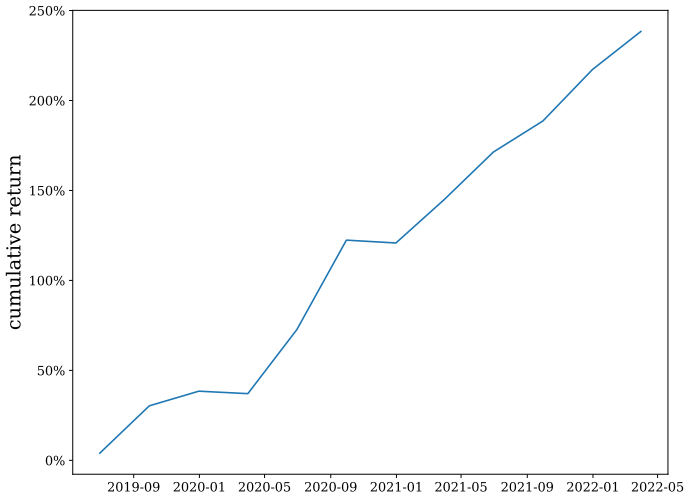

Figure 2: Backtesting for Wizard Strategy

The quarterly return is calculated as:

- Start a portfolio with unit 1.
- Re-invest all the PnL after each trade within the quarter.
- The quarterly return is the end of quarter portfolio value - 1.
- Re-base the portfolio value to 1 for the calculation of next quarter.

The cumulative return (cumulative sum of quarterly returns) is shown in Fig 2. The total return is 247% with a Sharpe ratio of 2.57. This wizard model only serves as an proxy for the upside potential of the strategy as it assumes we can perfectly predict actual earning surprises before the market close of the earning announcement days, which is ideal but not practical.

2.4 Benchmark Strategy

A more realistic strategy we can exploit is to modify the definition of the post announcement return as:

$$r_b^{t+1} = \frac{P_{\text{open}}^{t+2} - P_{\text{open}}^{t+1}}{P_{\text{open}}^{t+1}}, \tag{2.3}$$

where t represents the announcement date. Since we won't know the earning surprise in advance, the best we can do without prediction is to long the winners based on their EPS surprise at the open of the next day after earning announcement. This is essentially a momentum strategy which assumes the price will follow the earning momentum the day after earning announcement. We define this as our benchmark strategy. It can be described as:

- From 2014 Q1 to 2018 Q4, rank China A-share stocks based on the actual earning surprise within each quarter.
- Calculate the average thresholds for the top one sixth the same way as described for wizard strategy in Section 2.3.
- From 2019 Q1 to 2021 Q4, calculate the earning surprise for the universe of China A-share stocks, build a portfolio at the market open as of the day after announcement:
  - Buy the stock if the actual earning surprise is greater than the average historical top one sixth threshold.
  - If there are multiple stocks meet the condition, invest equally.
- Exit the positions at open of the next day. (The holding period for this strategy is from one day after announcement to two days after announcement).

The assumptions for backtesting (risk free rate, commission, and transaction cost) and quarterly return calculation methodology are the same as those of the wizard strategy.

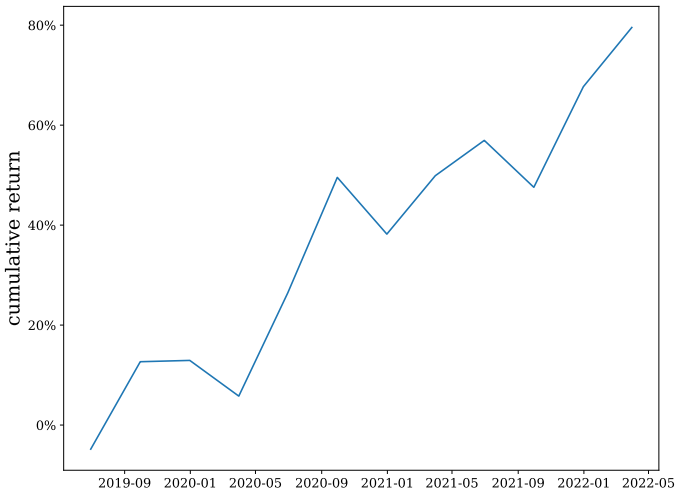

Figure 3: Correlation and backtesting for Benchmark Strategy

The cumulative return is shown in Fig 3. The total return drops to 89% with a Sharpe ratio of 1.05. This is as expected since the jump of stock prices is more likely to happen immediately after the earning announcement and be reflected in the gap between close on announcement day and open on next day. This strategy will serve as the benchmark to evaluate our strategies based on different models' prediction in Section 4.

3 Machine Learning Techniques Review

In this Section, we will give a brief review of some machine learning techniques that we use to predict the EPS surprise.

3.1 Long Short-term Memory

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture (Hochreiter and Schmidhuber, 1997) used in the field of deep learning. Unlike standard feed-forward neural networks, LSTM has feedback connections. A common LSTM unit is composed of a cell, with an input gate, an output gate and a forget gate (Fig 4). The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs.

3.2 Attention Model

The attention mechanism was originally proposed by Bahdanau et al. (2015). Instead of using the recurrent structure, it first calculates attention scores based on similarities of all pairs in the sequences and then forms a weighted sum of representations of sequence members as hidden state representations where the weights are the attention scores. Compared to LSTM, the attention based models can learn much longer sequence and allows for significantly more parallelism.

The tensors used to calculate similarity are called query (Q) or key (K) depending on whether they are the query itself or the reference. A tensor that is used as the representation is called value (V). Vaswani et al. (2017) generalize the concept and write it in a matrix form with a scaling factor $\sqrt{d_k}$ where d_k is the dimension of keys,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.1}$$

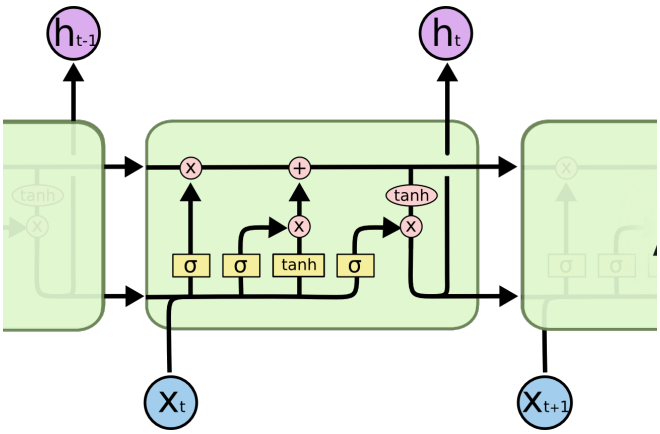

Figure 4: Architecture of LSTM (Olah, 2015)

In addition, the authors propose a multi-head attention mechanism which essentially projects keys, queries and values into lower dimensions, concatenates, and projects once again after performing the attention mechanism individually in parallel. In this paper, we will use the decoder architecture of transformer. Fig 5 is a

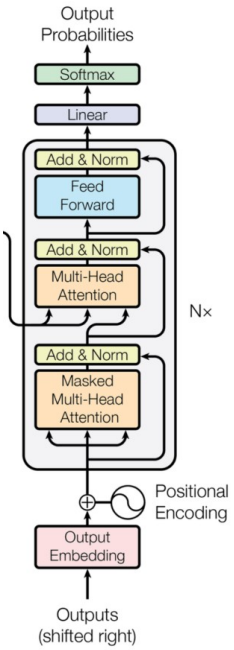

Figure 5: Architecture of the Decoder in Transformer Model

snapshot from Vaswani et al. (2017) to show the transformer decoder. The goal of the decoder is to generate text sequences. It begins with a start token, takes in a list of previous outputs as inputs as well as the encoder outputs that contain the attention information from the input and tries to generate next token. Since the decoder is auto-regressive and generates the sequence one by one, a look-ahead mask is used to prevent it from accessing future tokens.

4 Earning Surprise Prediction

4.1 Data Preparation

4.1.1 Study Universe and Input variables

The study universe we use for China A-share market include 2,392 public companies that have at least one quarterly Bloomberg Estimate EPS (BEST EPS) from 2014 Q1 to 2021 Q4. The numbers of companies with earning surprise (both BEST_EPS and actual EPS are available) by quarter are shown in Appendix A. The target for the model is the EPS surprise defined in Eq 2.1. The raw inputs for the model include 18 fundamental factors (Appendix B), 10 BEST factors (Appendix C), and industry classification to characterize the companies. For industry classification, we use Bloomberg Industry Classification System (BICS), a hierarchical classification of industries; all industry sectors are broken down to at least four levels, with some going as deep as eight levels. For this study, we use the first four levels. These are features commonly used in fundamental analysis and deemed to have strong predictive power for company's profitability. We also incorporate the 60-day SHASHR index return one day before announcement to gauge the overall market condition. Due to the limited number of features, feature selection is not performed for this study.

4.1.2 Imputation and Transformation

We use zero imputation and introduce two auxiliary indicators for each raw input, except the industry classification indicator. I_1 is used to indicate if the value is missing as of time t:

$$I_1^t(x) = \begin{cases} 0, & \text{if } x \text{ is null.} \\ 1, & \text{otherwise.} \end{cases} \tag{4.1}$$

and I_2 is a cumulative indicator used to show how many non-missing values does the time series accumulate up to time t:

$$I_2^t(x) = \sum_{i=0}^{t} I_1^i(x) \tag{4.2}$$

We also apply a hyperbolic tangent (tanh) function on the target EPS surprise to normalize all the EPS surprises between [-1,1]. This transformation can preserve the ranking order of the EPS surprise (since we only care about the ranking order of surprise for this trading strategy, the absolute value of EPS is less important in this case) and make the model easier to converge.

4.2 Model Architecture

We compare four model architectures in the paper for EPS surprise prediction: Linear model, Multilayer Perceptron (MLP), LSTM, and Attention-based model. Linear model and MLP (a few sequentially fully connected layers) are served as benchmark models while LSTM and Attention models (Section 3) are two candidates for our champion model. We use the data from 2014 Q1 to 2018 Q2 (18 quarters) as the training data and the data from 2018 Q3 to 2018 Q4 (2 quarters) as the development data. Development data is used for hyper-parameters tuning. After the optimal hyper-parameters are obtained, we re-train the model for every quarter from 2019 Q1 to 2021 Q4 (OOT data). For 2019 Q1, the model is re-trained using data from 2014 Q3 to 2018 Q4 (18 quarters) and for 2019 Q2, the model is re-trained using data from 2014 Q4 to 2019 Q1 (18 quarters) and so on and so forth. We use ensemble to reduce the variance of the model. For each period, 5 models with different random initializations are trained and we build the final portfolio by equally investing in the stocks generated by the five models.

The target we want to predict is the transformed EPS surprise as described in Section 4.1.2. We use the Mean Square Error (MSE) as the loss function. We use grid search to minimize the MSE on the development data for hyper-parameter tuning. The search spaces for LSTM and Transformer models are given in Appendix D. The architectures for LSTM and attention model are shown in Fig 6 and Fig 7 respectively.

4.3 Backtesting Result

We check the models' performances by backtesting on the out-of-time data. We simulate the calculation ten times to account for the randomness in models. The model-based strategies are constructed similarly as the wizard strategy described in Section 2.3:

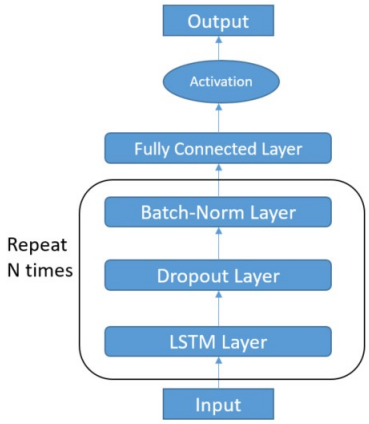

Figure 6: Architecture for our LSTM Model

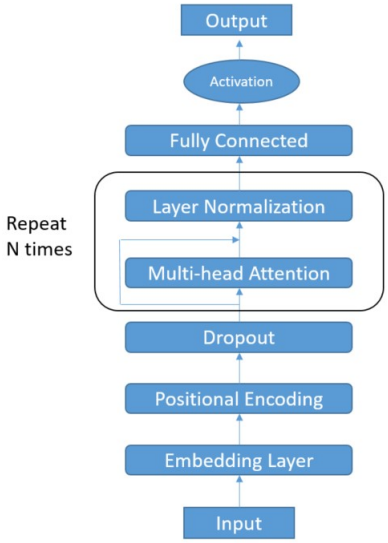

Figure 7: Architecture for our Attention-based Model

- For each quarter, we determine a threshold for *predicted EPS surprise* according to the *predicted EPS surprise last quarter*. Take 2019 Q1 as an example. The threshold is determined as the *top one sixth predicted EPS surprise as of 2018 Q4*.
- For each earning announcement day, we long the stocks whose *predicted EPS surprise* is greater than the threshold. If there are multiple stocks meet the condition, invest equally.
- We enter the positions before market close as of the earning announcement day and exit the positions before market close as of the day after announcement.

To evaluate the strategy's performance, we use the same assumptions as for the wizard strategy in Section 2.3 and focus on both the Sharpe ratio and the cumulative return during the testing period. Table 1 summarizes the results for ten simulations, along with the wizard and benchmark models. It shows that the attention model significantly outperforms other models in terms of both Sharpe ratio and cumulative return in the testing period.

Table 1: OOT backtesting results: 2019 Q1 to 2021 Q4. The numbers before and after $\pm$ are means and standard errors from ten simulations for model-based strategies.

Strategy	Sharpe Ratio	Cum. Return %
Wizard	2.57	247
Benchmark	1.05	89
Linear	0.57 ± 0.09	46 ± 5
MLP	0.68 ± 0.09	77 ± 10
LSTM	0.84 ± 0.01	73 ± 1
Attention	1.44 ± 0.09	110 ± 7

Fig 8 compares the cumulative returns for benchmark strategy (Section 2.4), average cumulative return of attention strategy, with shaded areas as the standard error bands, along with the cumulative return of investing in the SHASHR index. The β's for benchmark and model based strategies are 0.12 and -0.10 respectively. Both benchmark and attention based strategy significantly outperform the cumulative return of the index.

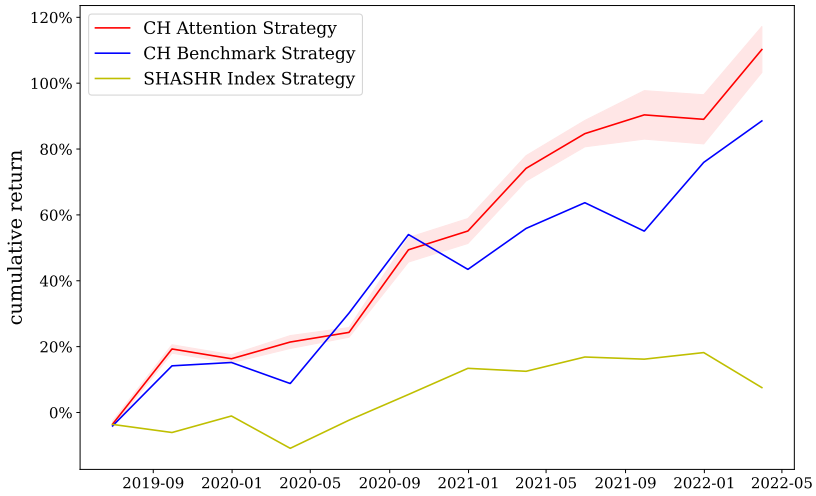

Figure 8: Model based strategies v.s. benchmark strategy. The solid lines for Attention is constructed by averaging the cumulative returns for ten simulation paths. Shaded area is the upper/lower bound corresponding to cumulative return $\pm$ 1 standard error.

In summary, among all the model based strategies, the attention-model-based strategy performs best and consistently outperforms the benchmark strategy and SHASHR index from 2019 Q1 to 2021 Q4.

5 Sensitivity Analysis for Transaction Cost

In previous sections, we show the backtesting results for benchmark and model based strategies without considering transaction cost (slippage, market impact, etc...) To better understand the strategies' performance, we perform a sensitivity analysis for cumulative returns with respect to transaction cost. Fig 9 shows that the cumulative returns decrease linearly when transaction cost increases. The cumulative returns will decrease by about 9% when transaction cost increases by 5 bps for both strategies. The break-even transaction cost for benchmark and model based strategies are 45 bps and 56 bps respectively.

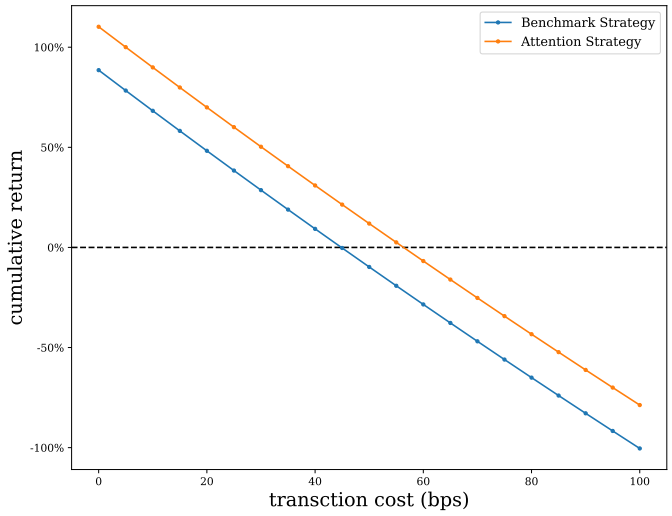

Figure 9: Sensitivity analysis with respect to transaction cost

6 Conclusion

In this paper, we first study the relation between EPS surprise and stock return one day after earning announcement for China A-share market. Based on the result, we develop a profitable trading strategy using the EPS surprise. Next, we investigate into different deep learning architectures to predict the EPS surprise and backtest the trading strategy based on the predicted EPS surprise. It shows that the attention-based model can yield a strategy that consistently beats the benchmark strategy we build in Section 2.4. Last but not least, we perform a sensitivity test with respect to transaction cost for both benchmark- and model-based strategies. It shows that the cumulative returns for both benchmark strategy and attention-model-based strategy decrease linearly when transaction cost increases.

Acknowledgments

We would like to thank Yang Wu, Young Li, Ruslan Tepelyan, Achintya Gopal, and Bo Pang for their fruitful insights and useful comments.

References

Albrecht, W. S., L. L. Lookabill, and J. C. McKeown (1977). The time-series properties of annual earnings. *Journal of Accounting Research*, 226–244.

Bahdanau, D., K. Cho, and Y. Bengio (2015, January). Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

Ball, R. and P. Brown (1968). An empirical evaluation of accounting income numbers. *Journal of accounting research*, 159–178.

Beaver, W. H. (1968). The information content of annual earnings announcements. *Journal of accounting research*, 67–92.

Bernard, V. L. and J. K. Thomas (1989). Post-earnings-announcement drift: delayed price response or risk premium? *Journal of Accounting research 27*, 1–36.

Elend, L., S. A. Tideman, K. Lopatta, and O. Kramer (2020). Earnings prediction with deep leaning. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, pp. 267–274. Springer.

Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural computation 9*(8), 1735–1780.

Lea, C., R. Vidal, A. Reiter, and G. D. Hager (2016). Temporal convolutional networks: A unified approach to action segmentation. In *European Conference on Computer Vision*, pp. 47–54. Springer.

Olah, C. (2015). Understanding lstm networks. `https://colah.github.io/posts/2015-08-Understanding-LSTMs` Access date: 2022-03-17.

Ou, J. A. and S. H. Penman (1989). Financial statement analysis and the prediction of stock returns. *Journal of accounting and economics 11*(4), 295–329.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 30, pp. 5998–6008. Curran Associates, Inc.

A Number of Companies with EPS Surprise

Table 2: Number of companies with EPS surprise by quarter

Period	Number of Companies
2014Q1	446
2014Q2	530
2014Q3	542
2014Q4	532
2015Q1	234
2015Q2	401
2015Q3	434
2015Q4	666
2016Q1	446
2016Q2	607
2016Q3	501
2016Q4	489
2017Q1	540
2017Q2	622
2017Q3	654
2017Q4	569
2018Q1	448
2018Q2	561
2018Q3	524
2018Q4	322
2019Q1	321
2019Q2	456
2019Q3	768
2019Q4	615
2020Q1	454
2020Q2	571
2020Q3	657
2020Q4	612
2021Q1	479
2021Q2	604
2021Q3	713
2021Q4	396

B Input Fundamental Factors

The fundamental data is from **Bloomberg Fundamentals**. Bloomberg covers the entire financial reporting process of companies - from earnings to preliminary releases to full fundamentals. With fundamentals data on more than 85,000 companies (both active and inactive) starting from the late 1980s, this dataset is the backbone of Bloomberg's Equity solutions. Bloomberg's Fundamental coverage includes current and normalized historical data for the balance sheet, income statement, cash flows statement and financial ratios, as well as industry-specific data for communications, consumer, energy, health care and many more.

Table 3: Input Fundamental Factors

Field Name	Field ID	Description
NET_REV	RR209	Revenue
GROSS_PROFIT	RR861	Gross Profit
IS_OPER_INC	IS033	Operating Income
PRETAX_INC	RR001	Pretax Income
NET_INCOME	IS050	Net Income, GAAP
GROSS_MARGIN	RR057	Gross Margin
OPER_MARGIN	RR026	Operating Margin
PROF_MARGIN	RR243	Net Profit Margin
BOOK_VAL_PER_SH	RR020	Book Value per share
CASH_FLOW_PER_SH	RR022	Cash Flow per share
IS_DIV_PER_SHR	IS151	Dividend per share
EBITDA	RR009	EBITDA
CUR_RATIO	RR053	Current Ratio
CF_CASH_FROM_OPER	CF105	Cash Flow from Operations
CF_CASH_FROM_INV_ACT	CF025	Cash Flow from Investing Activities
CF_CASH_FROM_FNC_ACT	CF035	Cash Flow from Financing Activities
NET_DEBT	RR208	Net Debt
IS_COMP_EPS_ADJUSTED	IS900	Estimate Comparable EPS Adjusted

C Input BEST Factors

The estimate data is from **Bloomberg Estimates**. Bloomberg provides derived consensus estimates for all companies using brokers' financial estimates for more than 18,000 companies globally. All estimates used in the consensus data are current and conform to the applicable accounting standards.

Table 4: Input BEST Factors

Field Name	Field ID	Description
BEST_EPS	BE008	Consensus estimate for adjusted EPS.
BEST_EPS_STDDEV	BE406	Standard deviation of consensus EPS.
BEST_EPS_MEDIAN	BE405	Median of BEST_EPS.
BEST_EPS_LO	BE408	Analysts' low EPS estimate.
BEST_EPS_HIGH	BE407	Analysts' high EPS estimate.
BEST_NET_INCOME	BE006	Consensus estimate for adjusted net income.
BEST_EBITDA	BE003	Consensus estimate for ebitda.
BEST_GROSS_MARGIN	BE245	Consensus estimate for gross margin.
BEST_EPS_CHG_PCT	BE401	Precentage change of BEST_EPS over prior estimate.
BEST_EPS_4WK_CHG	BE403	Change of BEST_EPS within the four weeks prior to today's date.

D Hyper-parameters Tuning

Table 5: Hyper-parameter search space for LSTM

Parameter	Search Space	Best Value
Hidden Size	8, 16, 24	24
Number of Layers	2, 3, 4	2
Learning Rate	1e-3, 2e-3	2e-3
Dropout	0.15, 0.2	0.15
Weight Decay	5e-3, 1e-2	5e-3

Table 6: Hyper-parameter search space for Attention

Parameter	Search Space	Best Value
Embedding Size	16, 24	24
Number of Layers	2	2
Number of Heads	2, 4	4
Learning Rate	1e-3, 2e-3	2e-3
Dropout	0.15, 0.2	0.2
Weight Decay	5e-3, 1e-2	5e-3
Training Epochs	30, 40, 50	40