No Stock is an Island: Learning Internal and Relational Attributes of Stocks with Contrastive Learning

Shicheng Li[1*], **Wei Li**[2*], **Zhiyuan Zhang**[1], **Ruihan Bao**[3†], **Keiko Harimoto**[3] and **Xu Sun**[1]
[1]MOE Key Lab of Computational Linguistics, School of Computer Science, Peking University
[2]Institute of Information Science, Beijing Language and Culture Univeristy
[3]Mizuho Securities Co., Ltd.
lisc99@pku.edu.cn, liweitj47@blcu.edu.cn, zzy1210@pku.edu.cn {ruihan.bao, keiko.harimoto}@mizuho-sc.com, xusun@pku.edu.cn

Abstract

Previous work has demonstrated the viability of applying deep learning techniques in the financial area. Recently, the task of stock embedding learning has been drawing attention from the research community, which aims to represent the characteristics of stocks with distributed vectors that can be used in various financial analysis scenarios. Existing approaches for learning stock embeddings either require expert knowledge, or mainly focus on the textual part of information corresponding to individual temporal movements. In this paper, we propose to model stock properties as the combination of internal attributes and relational attributes, which takes into consideration both the time-invariant properties of individual stocks and their movement patterns in relation to the market. To learn the two types of attributes from financial news and transaction data, we design several training objectives based on contrastive learning to extract and separate the long-term and temporary information in the data that are able to counter the inherent randomness of the stock market. Experiments and further analyses on portfolio optimization reveal the effectiveness of our method in extracting comprehensive stock information from various data sources.

1 Introduction

With the prosperity of machine learning, a whole new range of powerful data analysis tools has been introduced to applied fields such as health and economics. One of the areas that benefit most from this revolution is the area of financial technologies, where machine learning has been widely used in tasks including stock trend prediction [Li *et al.*, 2020; Zhao *et al.*, 2021] and optimal execution [Ning *et al.*, 2018].

In this work, we focus on the task of stock embedding learning. Similar to the well-studied task of word embeddings, stock embedding learning aims to represent the characteristics of a stock with a densely distributed vector. Stock embeddings that capture the comprehensive properties of stocks accurately can provide valuable stock information for downstream financial analysis.

Previous methods for learning stock embeddings broadly fall into two categories. Methods in the first category propose to combine stock representation learning with traditional technical analysis and learn the intrinsic properties or indicators based on the investment behavior of fund managers [Li *et al.*, 2019; Chen *et al.*, 2019]. However, the professional knowledge of human experts is usually difficult to access for the public, limiting the scope of application. Also, due to the vast amount of data in the stock market, even experts cannot take a comprehensive view of all available information.

The second category consists of methods that are more focused on data-driven deep learning techniques and bear more resemblance to our approach [Du and Tanaka-Ishii, 2020]. These methods make use of financial news and stock price data by employing learnable stock embeddings in the stock movement prediction or classification task. The main drawback of these approaches is that they only focus on the textual part of information corresponding to the movement of stock. Consequently, they either neglect the intrinsic properties of a stock or fail to explicitly model the relations between stocks, both of which carry valuable information for the process of financial analysis.

In view of the downsides of existing work, we propose a method to extract comprehensive stock information solely from news and price data. To be specific, we model the characteristics of a stock from two aspects: the internal attributes and the relational attributes. For example, consider the following excerpt from a financial news article:

> ...Carmakers Toyota [7203.T] and Nissan [7201.T], for instance, have both underperformed the Nikkei's 5.6 percent gain this year, posting losses of 11 percent and 6.6 percent respectively. ...

Here, "*Carmakers*" conveys information about the internal attributes of Toyota (7203.T) and Nissan (7201.T), i.e., both are in the industrial sector of "Transportation Equipment". On the other hand, "*both underperformed*" reflects the resemblance in the market performance of these two stocks, an example of what we define as the relational attributes .

However, information regarding the two types of attributes is implicit, largely blended in the data and subject to the ran-

[*]These authors contributed equally to this work.
[†]Contact Author.

Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP), pages 147 - 153

domness of the market. To address these issues, we propose to disentangle the long-term information and the temporary information in the data for mining the stable properties of individual stocks and stock relations. Inspired by the success of contrastive learning, we design several contrastive training objectives that extract long-term and temporary information from data to learn internal and relational attributes, respectively. Compared to previous methods, our stock embeddings are able to capture more comprehensive information contained in the text and price data, thus modeling the inherent properties and relations of stocks more accurately.

To testify the effectiveness of our approach, we apply our learned stock embeddings to the task of portfolio optimization. The portfolio yielded by our method achieves the highest return and the lowest risk of all tested approaches, demonstrating the superiority of our approach.

Our contributions can be summarized as follows:

- We propose to model stock properties as the combination of internal attributes and relational attributes and to learn these attributes from the long-term information and temporary information in the data.
- We design several contrastive objectives to counter the effect of randomness in the market and extract long-term and temporary information into stock embeddings.
- Experiments on portfolio optimization and further analyses show the effectiveness of our method in learning internal and relational attributes of the stocks.

2 Method

In this section, we describe the model architecture and the designed contrastive objectives to capture the internal and relational attributes of stocks from textual news and transaction data in our embeddings, as shown in Figure 1.

2.1 Overview

We argue that the stock embeddings should contain the intrinsic attributes of stocks from two aspects, internal and relational. **Internal attributes** refer to the attributes that are inherent in a stock and remain stable over time, such as the sector of the stock, while **relational attributes** encode the relationship between stocks in the market. Whereas internal attributes contain information about individual stocks, relational attributes can tell us about their positions in the stock market and allow us to infer knowledge of one stock from other stocks. Our goal is to capture these two types of attributes into stock embeddings from textual financial news and time series transaction data.

However, learning these attributes is non-trivial as they are not explicitly available in the data. To make matters worse, we observe that both news articles and price history are subject to large temporal variations rooted in the randomness of the market. To solve these problems, we propose to view the relevant information in the data as the combination of two independent parts: long-term information and temporary information. **Long-term information** is the time-invariant part of stock information that usually encodes the internal attributes of a stock. **Temporary information** corresponds to the temporal variations specific to a short time period. Despite its randomness with respect to individual stocks, patterns exist in the relative fluctuation between different stocks and are informative of their relationship. By treating the two types of information separately and focusing on the stable, invariant elements, we are able to alleviate the negative effect of random market fluctuations and capture the attributes of the stocks.

2.2 Preliminaries: Contrastive Learning

The key idea behind contrastive learning is that the representations of similar inputs should be close to each other, while the representations of dissimilar inputs should lie far apart. To be more specific, the representation of each input example is treated as an anchor point. Several positive examples and negative examples are constructed using heuristic rules. Then the model tries to reduce the distance between the anchor and the positive examples while enlarging the distance between the anchor and the negative examples by minimizing a contrastive loss such as InfoNCE [van den Oord *et al.*, 2018] or the triplet loss [Schroff *et al.*, 2015].

Previous work on contrastive learning has explored various ways to construct positive and negative examples. Methods to construct positive examples include applying different data augmentations or transformation to the input [Ye *et al.*, 2019] and using different views of the same object such as different channels of the image [Tian *et al.*, 2020]. Negative examples are usually randomly sampled from the dataset, within the same mini-batch or from a memory bank of previously computed representations [He *et al.*, 2020].

In our work, we regard the stock embedding and data from different sources on different days as multiple views of the same piece of information. We use the triplet loss which encourages the anchor to be at least closer to the positive example than the negative examples by a distance $\mathcal{D}(\cdot, \cdot)$ of 1, i.e.,

$$\mathcal{L}_{\text{cont}} = \frac{1}{N} \sum_{i=1}^{N} \max(1 + \mathcal{D}(x, x^+) - \mathcal{D}(x, x_i^-), 0) \quad (1)$$

where x is the anchor, x^+ is the positive example, and $\{x_i^-\}_{i=1}^N$ are the negative examples.

2.3 Notations

In this work, we assume access to two types of data: the financial news of a given set of stocks over a certain time period, and the numerical transaction data of these stocks over the same period. We also assume each news article is annotated with the stock codes it concerns. Datasets that satisfy these requirements are readily available as they are provided by news organizations such as Reuters.

We denote the set of stocks as $\mathcal{S}$ and the trading days as $\mathcal{T} = \{1, 2, \ldots, T\}$. Let e_s be the embedding of stock $s \in \mathcal{S}$. For each stock $s \in \mathcal{S}$ and each trading day $\tau \in \mathcal{T}$, we aim to extract information from two data sources: the news article concerning stock s on day τ denoted by $n_{s,\tau}$, and the sequence of transaction data of stock s leading up to day τ denote by $p_{s,\tau}$. We use h to denote the vector representing the information in the data. Superscripts $^{\text{l}}/^{\text{t}}$ indicate whether the

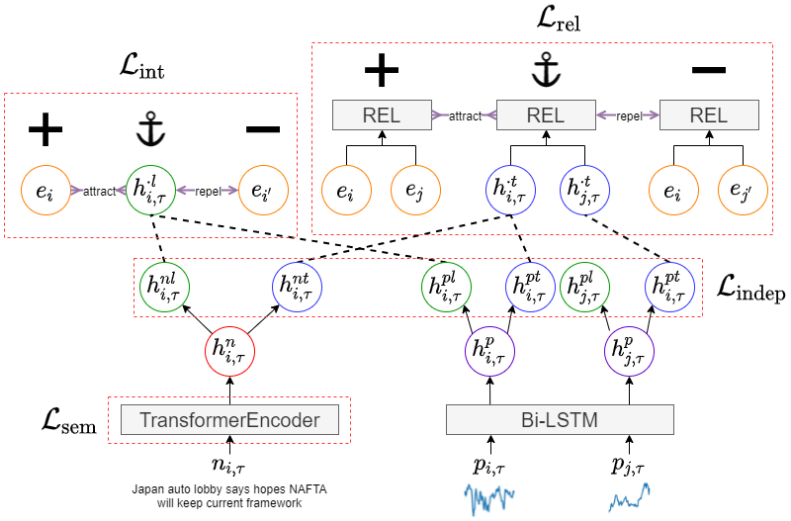

Figure 1: The model architecture and training objectives of our stock embedding learning method.

representation corresponds to long-term or temporary information and $^{\text{n}}/_{\text{p}}$ indicate whether the information comes from the news article or the price time series, while subscripts $_{s,\tau}$ indicate the stock and trading date associated with the data.

2.4 Encoder Architecture

We use a transformer [Vaswani *et al.*, 2017] encoder and a bidirectional LSTM to encode the textual news and transaction time series, respectively. After obtaining the news representation $h^{\text{n}}_{s,\tau}$ and price representation $h^{\text{p}}_{s,\tau}$, we apply separate linear transformations for $h^{\text{n}}_{s,\tau}$ (or $h^{\text{p}}_{s,\tau}$) to extract the long-term representations $h^{\text{nl}}_{s,\tau}$ (or $h^{\text{pl}}_{s,\tau}$) and temporary representations $h^{\text{nt}}_{s,\tau}$ (or $h^{\text{pt}}_{s,\tau}$) with different transformation weights. These long-term and temporary representations serve as the basis of our contrastive learning objectives.

2.5 Training Objectives

We design the following training objectives to guide the training of stock embeddings: the internal contrastive loss $\mathcal{L}_{\text{int}}$, the relational contrastive loss $\mathcal{L}_{\text{rel}}$, the semantic loss $\mathcal{L}_{\text{sem}}$ and the independence loss $\mathcal{L}_{\text{indep}}$.

Internal Contrastive Loss $\mathcal{L}_{\text{int}}$

The internal contrastive loss is designed to capture the internal attributes of the stocks. In most cases, the internal attributes of a stock should remain invariant over different time periods despite the temporal fluctuations of the market. Therefore, information concerning such attributes should appear consistently in the financial news and transaction data associated with the same stock on different days, which we define as long-term information.

Based on this observation, our internal contrastive loss aims at reducing the distance between the long-term representation and the embedding of its associated stock while enlarging the distance between the long-term representation and other stock embeddings. We choose the long-term representation vector as the anchor and stock embeddings as positive and negative examples:

$$\begin{aligned} &\mathcal{L}_{\text{int}} = \mathcal{L}_{\text{cont}}(x_{\text{int}}, x^{+}_{\text{int}}, \{x^{-}_{\text{int},i}\}_{i=1}^{N}), \\ &x_{\text{int}} = h^{\text{l}}_{s,\tau}, \quad x^{+}_{\text{int}} = e_{s}, , \quad x^{-}_{\text{int},i} = e_{s'_i} \end{aligned} \tag{2}$$

where $\{s'_i\}_{i=1}^{N}$ are N randomly sampled stocks.

By minimizing the internal contrastive loss of these representations with respect to the same positive example (i.e., the corresponding stock embedding), the encoders are encouraged to extract from the data the long-term information that is consistent over time. Furthermore, since all these long-term representation vectors are encouraged to be close to their corresponding stock embeddings, we implicitly tell the model to capture the information concerning internal attributes in the stock embeddings.

Relational Contrastive Loss $\mathcal{L}_{\text{rel}}$

Besides regarding each stock as an individual entity, it is also crucial to consider its position in the market and its relation to other stocks. One rationale behind this is that the stocks in a stock market are inter-correlated and information about one stock may reveal some information about other stocks. Also, some financial analytic methods such as portfolio optimization explicitly call for the modeling of stock correlation. This motivates us to design the relational contrastive loss.

Whereas the computation of internal contrastive loss only concerns a single stock, for the relational contrastive loss we need to represent the relationship between two stocks or the temporary information with a vector. Inspired by Socher *et al.* [2013], we design the following REL module which produces a relational vector for any two d-dimensional vectors h_1 and h_2 as follows,

$$\text{REL}(h_1, h_2) = \tanh(h_1^T W^{[1:k]} h_2 + V \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} + b) \tag{3}$$

where $W^{[1:k]} \in \mathbb{R}^{d \times d \times k}$ is a tensor, and the result of $h_1^T W^{[1:k]} h_2$ is a k-dimensional vector with the i-th dimen-

sion being $h_1^T W^{[i]} h_2$. $W^{[1:k]} \in \mathbb{R}^{d\times d\times k}$, $V \in \mathbb{R}^{k\times 2d}$, $b \in \mathbb{R}^k$ are all learnable parameters.

For simplicity, we view the relation between any stock pair as a stable quantity that remains unchanged throughout despite the temporal fluctuations of the market. Reflected in the data, this means that for the temporary information of the two stocks on the same day τ, their relational vector is consistent over different τs. Our relational contrastive loss strives to mine this stable pattern by comparing the relational vector computed using the temporary information of different stocks and the relational vector of their stock embeddings,

$$\begin{aligned} \mathcal{L}_{\text{rel}} &= \mathcal{L}_{\text{cont}}(x_{\text{rel}}, x_{\text{rel}}^+, \{x_{\text{rel},i}^-\}_{i=1}^N) \\ x_{\text{rel}} &= \text{REL}(h_{s,\tau}^{\cdot\text{t}}, h_{\tilde{s},\tau}^{\cdot\text{t}}) \\ x_{\text{rel}}^+ &= \text{REL}(e_s, e_{\tilde{s}}), \quad x_{\text{rel},i}^- = \text{REL}(e_s, e_{s_i'}) \end{aligned} \tag{4}$$

where $\{s_i'\}_{i=1}^N$ are N randomly sampled stocks.

Similar to the internal contrastive loss, the relational contrastive loss tries to push the temporary information relational vector towards the stock embedding relational vector of the corresponding stock pair, while pushing it away from that of other pairs. By assigning the same positive example (i.e., the relational vector of the corresponding stock embeddings) to the temporary information of the same stock pair, the model is encouraged to extract the stable relational attributes implied in the data and capture them in the stock embeddings.

Semantic Loss $\mathcal{L}_{\text{sem}}$

The contrastive losses serve the purpose of guiding the model to filter out relevant information that benefits stock embedding learning from the deluge of extracted information. However, due to the complexity of natural language, such high-level guidance alone is insufficient for the extraction of useful information from the news. Inspired by the success of masked language modeling (MLM) in pretrained language models, we incorporate an MLM-based semantic loss to inject our training objective with low-level supervision signals that endow the text encoder with the capability of understanding natural language and extracting semantic information from news articles. In this way, the text encoder not only learns the semantics of textual data, but also aligns the stock embeddings with the embeddings of natural language, providing extra supervision signals for capturing long-term information into the internal attributes of the stocks.

Independence Loss $\mathcal{L}_{\text{indep}}$

Ideally, the relational contrastive loss should only capture the attributes that are not inherent to a stock and extract the temporary information that is only invariant when considered in relation to the temporary information of another stock. In reality, however, since the long-term information and temporary information are intertwined with each other in the data, it is difficult for the model to tell them apart from each other, leading to the degeneration of representations. To alleviate this problem, we incorporate an independence loss to encourage statistical independence between the long-term representations and temporary representations by training the model to minimize the mutual information (MI) between them.

As calculating the MI between continuous random vectors is intractable, following Belghazi *et al.* [2018], we leverage a neural network T_θ to estimate the MI between the long-term representation random vector $H^{\cdot l}$ and the temporary representation random vector $H^{\cdot t}$ by maximizing

$$\hat{I}(H^{\cdot l}, H^{\cdot t}) = \mathbb{E}_{\mathbb{P}_{H^{\cdot l}H^{\cdot t}}}[T_\theta] - \log(\mathbb{E}_{\mathbb{P}_{H^{\cdot l}}\otimes\mathbb{P}_{H^{\cdot t}}}[e^{T_\theta}]) \tag{5}$$

where $\mathbb{P}_{H^{\cdot l}H^{\cdot t}}$ denotes the joint distribution of $H^{\cdot l}$ and $H^{\cdot t}$, and $\mathbb{P}_{H^{\cdot l}}\otimes\mathbb{P}_{H^{\cdot t}}$ denotes the product of marginal distributions.

In the meantime, our encoder serves as an adversary for the estimator T_θ, trained to minimize the MI between $H^{\cdot l}$ and $H^{\cdot t}$. To allow the entire model to be trained in an end-to-end manner, a gradient reversal layer [Ganin and Lempitsky, 2015] is inserted between the representations and the estimator T_θ. Through the independence loss, we are able to separate the long-term and temporary information in the data and prevent potential representation degeneration.

Final Objective Our final training objective is a linear combination of all the aforementioned losses,

$$\mathcal{L} = \lambda_{\text{int}}\mathcal{L}_{\text{int}} + \lambda_{\text{rel}}\mathcal{L}_{\text{rel}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{indep}}\mathcal{L}_{\text{indep}} \tag{6}$$

where the λs are manually chosen hyper-parameters to balance the effect of different objectives.

3 Experiments

3.1 Dataset

We carry out our experiments on the 500 largest stocks of Tokyo Stock Exchange known as TOPIX 500. Our data is composed of hourly transaction data (open, high, low, close, volume) and Reuters news articles of these stocks from *2013-01-01* to *2018-09-30*. We use the data from *2017-10-01* to *2018-03-31* as the validation set and the data from *2018-04-01* to *2018-09-30* as the test set.

3.2 Model

We set the dimension of the stock embeddings to 512. Our price encoder is a 2-layer bidirectional LSTM with a hidden size of 512. The news encoder is a randomly initialized 6-layer transformer encoder with a vocabulary size of 50000.

We use the Adam [Kingma and Ba, 2015] optimizer with a batch size of 64 and a weight decay of 1e-4. The learning rate is warmed up linearly to 1e-4 for the first 10,000 steps and then decays on a cosine annealing schedule. For stability, the model is trained only with the semantic loss during the warmup stage before we add other objectives into the training process.

3.3 Portfolio Optimization

To demonstrate that the information captured in our stock embeddings is useful in real-world financial analysis, we follow Du and Tanaka-Ishii [2020] and evaluate the quality of our stock embeddings on the task of portfolio optimization.

The goal of portfolio optimization is to decide the proportion of capital to invest in each stock within a stock list to maximize the expected return or minimize the risk. Based on the intuition that risk can be reduced by investing money in uncorrelated or negatively correlated stocks, the problem is formulated by Markowitz [1959] as

$$\begin{aligned} &\min_{w_j\in[0,1], 1\le j\le J} w^T\Sigma w \\ &\text{s.t.} \quad w^T r = E, \quad w^T\mathbf{1} = 1 \end{aligned} \tag{7}$$

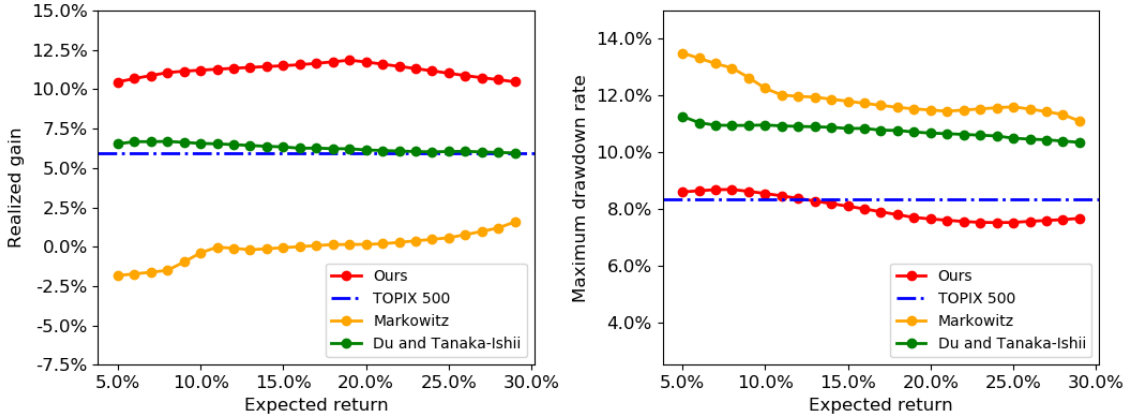

Figure 2: Results on portfolio optimization. Our model achieves higher gain (income) and lower drawdown rate (risk) than baseliness.

where $\Sigma \in \mathbb{R}^{J \times J}$ is the risk matrix measuring the correlation between each two stocks; w is a weight vector that sums to 1 denoting how much capital to invest in each stock; r is the vector indicating the historical returns for each stock; and E is a parameter set by the investor denoting the expected portfolio return. In other words, the goal is to minimize the correlation between invested stocks contingent on a given amount of expected return.

Following Du and Tanaka-Ishii [2020], we define the correlation between two stocks i and j as the cosine similarity between their stock embeddings, i.e.

$$\Sigma_{i,j} = \cos(e_i, e_j) \tag{8}$$

The expected return is set to different values in $\{0.05, 0.06, \ldots, 0.29\}$. The optimization problem is then solved using quadratic programming.

We measure the quality of our stock embeddings primarily based on the realized gain of the portfolio, which is the profit rate over the test period. To assess the risk of the portfolio induced by our embeddings, we also compute the maximum drawdown rate, i.e., the maximum loss rate from a historical peak at any time point over the period.

We compare our method against three baselines that restrict available external data to news articles and transaction data:

- **TOPIX 500 Market Index:** The captalization-weighted portfolio of all TOPIX 500 stocks;
- **Markowitz:** The portfolio computed by the original Markowitz model where the risk matrix is the covariance matrix of stock returns;
- **Du and Tanaka-Ishii:** Our re-implementation of the stock embedding-based portfolio propose by Du and Tanaka-Ishii [2020].

The results are presented in Figure 2. As can be seen, our method achieves higher realized gain (higher income) and lower maximum drawdown rate (lower risk) compared to all the baselines. The original Markowitz model constructs the risk matrix only from the correlation between stock price series. This approach neither utilizes information from textual data, nor takes the randomness of the market into account, and therefore performs considerably worse even compared to the market baseline of the TOPIX 500 Index. On the other hand, although taking advantage of both price and news data, Du and Tanaka-Ishii [2020] only focuses the textual information concerning stock movements and therefore performs only slightly over the TOPIX 500 Index. In contrast, our method is able to make use of the comprehensive information regarding both internal and relational attributes in the data, thus modeling the relationship between different stocks more accurately and obtaining the best-performing portfolio out of all considered approaches.

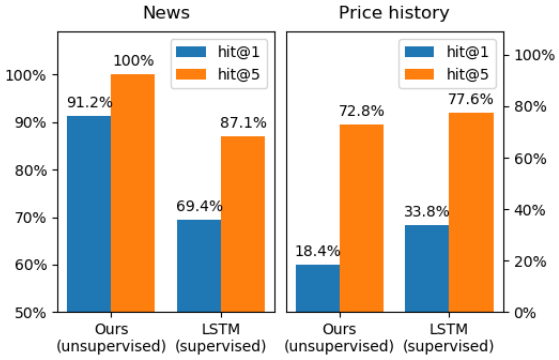

Figure 3: Results on news/price history classification

3.4 News and Price History Classification

To verify that our stock embeddings can capture the internal attributes of a stock by extracting stock-specific long-term information from data, we apply our model and stock embeddings to the task of news and price history series classification, where the model is asked to predict which stock corresponds to a given news article or price time series.

We directly encode the news article or price history with our model and make predictions based on the distances between the stock embeddings and the long-term representation of the input data. Note that no further training or fine-tuning is required for our model, which means that our model works in a purely unsupervised way. For simplicity, we only consider the 30 stocks from TOPIX Core 30 Index. We compare our results with two LSTMs trained on the news article or price

history from the training data and report hit@1 and hit@5 on the test set. As shown in Figure 3, although not trained on the classification task, our model achieves comparable or even better results than the baseline methods. This demonstrates that our method is able to encode the long-term information in the data and capture stock attributes in its embedding.

3.5 Clustering

To examine whether our stock embeddings indeed encode internal attributes of the stocks, we perform spectral clustering [Shi and Malik, 1997] on the learned representations of the stocks. Two examples of the resulting clusters are shown in Table 1. Cluster 1 is primarily composed of stocks from the industrial sector of "Electric Power & Gas", while the second cluster contains several of Japan's largest carmakers. This result lends credence to our statement that our stock embeddings manage to capture internal attributes of the stocks such as the industrial area of the company.

3.6 Ablation Study

We conduct ablation studies to verify the effect of the designed objectives on portfolio optimization.

For simplicity, we average the realized gains and maximum drawdown rates over different expected returns. As shown in Table 2, although using long-term information or temporary information alone leads to positive results compared to the baselines, their performance lags far behind our whole model. This supports our motivation of jointly learning the internal and relational attributes by utilizing the two types of information. Removing the independence loss also causes slight performance degradation, which may be the consequence of representation degeneration.

We further remove all training objectives related to one of the data sources. Unsurprisingly, the removal of information from either financial news or transaction data leads to a significant drop in the realized gain. This substantiates the importance to leverage information from both data sources.

Cluster 1	
Hokuriku Electric Power	Electric Power & Gas
Osaka Gas	Electric Power & Gas
Daiwa House Industry	Construction
Tokyo Gas	Electric Power & Gas
Sumitomo Forestry	Construction
Chugoku Electric Power	Electric Power & Gas
Chubu Electric Power	Electric Power & Gas
Shikoku Electric Power	Electric Power & Gas
Cluster 2	
Hino Motors	Transportation Equipment
Honda Motor	Transportation Equipment
Subaru Corporation	Transportation Equipment
Toyota Motor Corporation	Transportation Equipment
Tohoku Electric Power	Electric Power & Gas
Skylark Holdings	Retail Trade
Kawasaki Heavy Industries	Transportation Equipment
Mitsubishi Gas Chemical	Chemicals
Suzuki Motor Corporation	Transportation Equipment
JGC Holdings Corporation	Construction
Japan Tobacco	Foods
Daicel Corporation	Chemicals

Table 1: Examples of the clusters acquired from our embeddings.

Method	**Gain(+)**	**Drawdown(-)**
Ours	**11.20%**	8.02%
w/o $\mathcal{L}_{\text{indep}}$	10.01%	**7.65%**
w/o $\mathcal{L}_{\text{int}}$	7.16%	9.11%
w/o $\mathcal{L}_{\text{rel}}$	7.10%	9.50%
w/o price	9.42%	8.27%
w/o news	7.27%	9.53%

Table 2: Ablation study. (+): the higher the better; (-): the opposite.

4 Related Work

4.1 Distributed Representations

Representing the semantic meaning of tokens with distributed vectors has been studied for a long time. Mikolov *et al.* [2013a; 2013b] first propose to learn the semantic meaning of words from their context. Apart from natural language, efforts have also been made to represent the entities and relations in knowledge bases [Bordes *et al.*, 2013; Wang *et al.*, 2014; Lin *et al.*, 2015] or the nodes and edges in graphs [Perozzi *et al.*, 2014; Tang *et al.*, 2015; Grover and Leskovec, 2016]. In this work, we focus on the topic of learning stock embeddings where we introduce both textual financial news and time series transaction data to learn both internal and relational attributes of stocks.

4.2 Contrastive Learning

Contrastive learning is a promising approach in unsupervised representation learning. Early attempts in this area include Mikolov *et al.* [2013b] which designs a negative sampling method based on noise-contrastive estimation [Gutmann and Hyvärinen, 2010] to learn word embeddings. Recent years have witnessed a flourishing of literature concerning contrastive learning. CPC [van den Oord *et al.*, 2018; Hénaff, 2020] proposes to predict subsequent inputs based on previous inputs to learn representations for any data that is serializable on the dimension of time or space. MoCo [He *et al.*, 2020] formulates contrastive learning as a dictionary lookup task and propose to use a momentum encoder to improve the consistency between the key-value pairs. In this work, we first introduce contrastive learning into stock representation learning to help extract information from news and price data and learn expressive stock embeddings.

5 Conclusion

In this paper, we propose to model the properties of a stock from two aspects: its internal attributes as an individual stock, and its relational attributes relative to other stocks. We propose to extract long-term information and temporary information from financial news and transaction data to learn these two types of attributes. To capture these attributes in the stock embeddings, we design several training objectives based on contrastive learning that are able to counter the randomness of the stock market. Comprehensive empirical evidence demonstrates that our stock embeddings are able to model stock properties and relations more accurately.

References

[Belghazi *et al.*, 2018] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. Mutual information neural estimation. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 530–539. PMLR, 2018.

[Bordes *et al.*, 2013] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795, 2013.

[Chen *et al.*, 2019] Chi Chen, Li Zhao, Jiang Bian, Chunxiao Xing, and Tie-Yan Liu. Investment behaviors can tell what inside: Exploring stock intrinsic properties for stock trend prediction. In *KDD*, pages 2376–2384. ACM, 2019.

[Du and Tanaka-Ishii, 2020] Xin Du and Kumiko Tanaka-Ishii. Stock embeddings acquired from news articles and price history, and an application to portfolio optimization. In *ACL*, pages 3353–3363. Association for Computational Linguistics, 2020.

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1180–1189. JMLR.org, 2015.

[Grover and Leskovec, 2016] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *KDD*, pages 855–864. ACM, 2016.

[Gutmann and Hyvärinen, 2010] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, volume 9 of *JMLR Proceedings*, pages 297–304. JMLR.org, 2010.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735. IEEE, 2020.

[Hénaff, 2020] Olivier J. Hénaff. Data-efficient image recognition with contrastive predictive coding. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 4182–4192. PMLR, 2020.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

[Li *et al.*, 2019] Zhige Li, Derek Yang, Li Zhao, Jiang Bian, Tao Qin, and Tie-Yan Liu. Individualized indicator for all: Stock-wise technical indicator optimization with stock embedding. In *KDD*, pages 894–902. ACM, 2019.

[Li *et al.*, 2020] Wei Li, Ruihan Bao, Keiko Harimoto, Deli Chen, Jingjing Xu, and Qi Su. Modeling the stock relation with graph network for overnight stock movement prediction. In *IJCAI*, pages 4541–4547. ijcai.org, 2020.

[Lin *et al.*, 2015] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*, pages 2181–2187. AAAI Press, 2015.

[Markowitz, 1959] Harry Markowitz. Portfolio selection, 1959.

[Mikolov *et al.*, 2013a] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR (Workshop Poster)*, 2013.

[Mikolov *et al.*, 2013b] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[Ning *et al.*, 2018] Brian Ning, Franco Ho Ting Ling, and Sebastian Jaimungal. Double deep q-learning for optimal execution. *CoRR*, abs/1812.06600, 2018.

[Perozzi *et al.*, 2014] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: online learning of social representations. In *KDD*, pages 701–710. ACM, 2014.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823. IEEE Computer Society, 2015.

[Shi and Malik, 1997] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. In *CVPR*, pages 731–737. IEEE Computer Society, 1997.

[Socher *et al.*, 2013] Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. Reasoning with neural tensor networks for knowledge base completion. In *NIPS*, pages 926–934, 2013.

[Tang *et al.*, 2015] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. LINE: large-scale information network embedding. In *WWW*, pages 1067–1077. ACM, 2015.

[Tian *et al.*, 2020] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV (11)*, volume 12356 of *Lecture Notes in Computer Science*, pages 776–794. Springer, 2020.

[van den Oord *et al.*, 2018] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[Wang *et al.*, 2014] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, pages 1112–1119. AAAI Press, 2014.

[Ye *et al.*, 2019] Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *CVPR*, pages 6210–6219. Computer Vision Foundation / IEEE, 2019.

[Zhao *et al.*, 2021] Liang Zhao, Wei Li, Ruihan Bao, Keiko Harimoto, Yunfang Wu, and Xu Sun. Long-term, short-term and sudden event: Trading volume movement prediction with graph-based multi-view modeling. In *IJCAI*, pages 3764–3770. ijcai.org, 2021.