Interpretable Multimodal Learning for Intelligent Regulation in Online Payment Systems

Shuoyao Wang[1*†] , **Diwei Zhu**[2] ,
[1]College of Electronic and Information Engineering, Shenzhen University, China
[2]Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong
w.shuoy@gmail.com, zd115@ie.cuhk.edu.hk

Abstract

With the explosive growth of transaction activities in online payment systems, effective and real-time regulation becomes a critical problem for payment service providers. Thanks to the rapid development of artificial intelligence (AI), AI-enable regulation emerges as a promising solution. One main challenge of the AI-enabled regulation is how to utilize multimedia information, i.e., multimodal signals, in Financial Technology (FinTech). Inspired by the attention mechanism in nature language processing, we propose a novel cross-modal and intra-modal attention network (CIAN) to investigate the relation between the text and transaction. More specifically, we integrate the text and transaction information to enhance the text-trade joint-embedding learning, which clusters positive pairs and push negative pairs away from each other. Another challenge of intelligent regulation is the interpretability of complicated machine learning models. To sustain the requirements of financial regulation, we design a CIAN-Explainer to interpret how the attention mechanism interacts the original features, which is formulated as a low-rank matrix approximation problem. With the real datasets from the largest online payment system, WeChat Pay of Tencent, we conduct experiments to validate the practical application value of CIAN, where our method outperforms the state-of-the-art methods.

1 Introduction

As a new term in the financial industry, FinTech has become a popular term that describes novel technologies adopted by the financial service institutions. At the juncture of these phenomena, the risk of online transaction and shopping becomes increasingly prominent. Establishing a reliable intelligent regulation infrastructure, e.g., to detect whether the transaction flow of a merchant is beyond the scope of its licensed business, is essential to accommodating more FinTech startups and e-commerce. Thanks to the advance of personal computers, smart phones, and internet, the quantity of digitized multimedia contents has increased dramatically [YouTube, 2020]. Consequently, despite of the efforts utilizing subtle feature engineering and classifiers for AI-enabled regulation [Cao *et al.*, 2019], the efficient utilization of digitized multimedia contents is still a core research challenge for intelligent regulation.

Inspired by the recent advances in deep learning, multimodal machine learning has been widely employed to interpret such multimedia information during the past few years. These techniques can be roughly classified into four categories: multimdedia speech recognition [Afouras *et al.*, 2018], multimdedia content analysis (e.g., automatic shot boundary detection [Lienhart, 1998], video summarization [Zhang *et al.*, 2016]), human behavior understanding (e.g., emotion recognition [Chu *et al.*, 2016]), and multimodal matching (e.g., visual question-answering [Antol *et al.*, 2015], image-text matching [Wang *et al.*, 2019]).

In this paper, we focus on the general line regulation problem, i.e., to detect whether the transaction flow of a merchant is beyond the scope of its licensed business, which can be formulated as a multimodal matching problem. In the past few years, many research efforts have been devoted to multimodal matching. For instance, [Cheng *et al.*, 2019] and [Verma *et al.*, 2019] concatenated features from different modalities via a fusion layer for sentiment analysis and crowdfunding success prediction, respectively. Inspired by the great success of attention in nature language processing (NLP), many studies have validated the attention is helpful to model a more reliable relationship between image and text [Yang *et al.*, 2019; Ma *et al.*, 2019; Wang *et al.*, 2019]. Most recently, a popular framework to model the multimodal relationship is the two-branch embedding network, where one branch encodes the first modality information and another modules the other one [Wang *et al.*, 2018; Gu *et al.*, 2018; Wang *et al.*, 2019]. It is desired if the cross-modal and intra-modal relationships could be jointly investigated in a unified two-branch embedding network for AI-enabled regulation problem. Unfortunately, the existing multimodal learning methods and attention mechanisms mostly focus on the textual and visual information. To the best of our knowledge, transaction flow information as the most important modality in FinTech has

[*] Work done while Shuoyao was a senior researcher with Financial Technology Group, Tencent, China.

[†] Corresponding author.

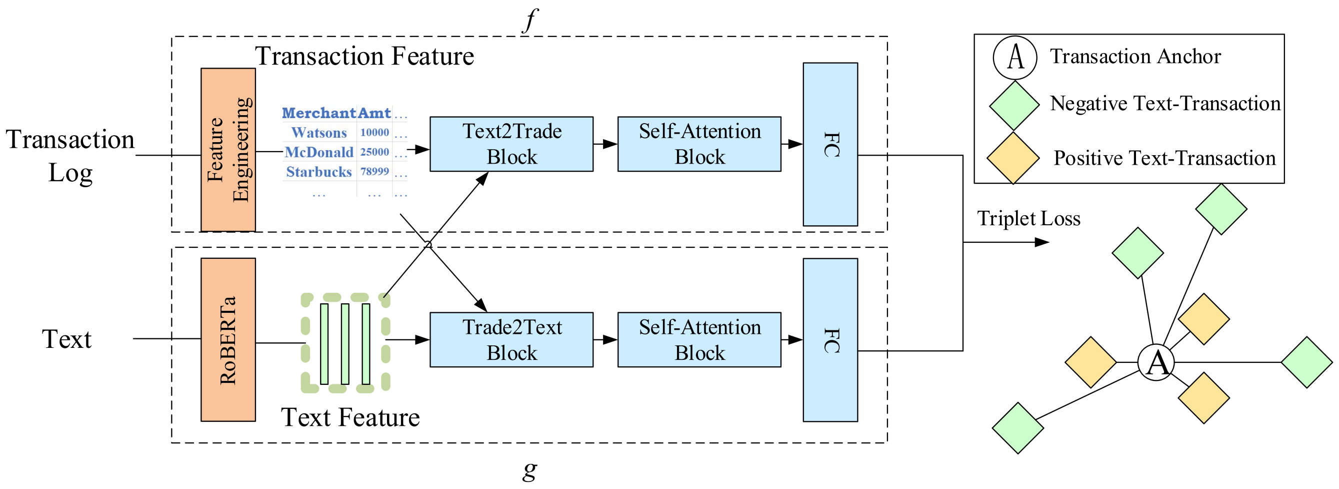


Figure 1: CIAN: (i) the feature extraction module which extracts text and transaction features using their corresponding backbone architectures; (ii) the cross-modal and inter-modal attention projection that match the feature distributions originating from the same identity.

not been studied yet.

In particular, the intra-modality relations within text and transaction flow are complementary to the inter-modality relations between text and transaction flow, which were mostly ignored by existing mechanisms. For example, if the description of a merchant indicates it is a restaurant, the meal-time transaction flows express the main semantics of the merchant with high probability, while the gender of the customers may not be that important. Moreover, on one hand, fully exploiting interaction relations leads to complex models and thus becomes hard to explain. On the other hand, "need to explain" is already enforced in many industries, especially in FinTech. To tackle the gap, we propose a novel cross-modal and intra-modal attention network (CIAN) to investigate the relation between the textual and transaction flow views. Moreover, we also propose a CIAN-Explainer and formulate it as an optimization problem, i.e., a low-rank matrix approximation problem. Consequently, CIAN-Explainer identifies a small subset of transaction flow features that have a crucial role in transaction-text matching. The major contributions are summarized as follows:

- To the best of our knowledge, this paper is among the first to propose an attention mechanism to text and transaction flow for AI-enabled regulation.
- We propose both intra-modal and cross-modal attention to capture the correspondences inside the financial behavior, and between the financial behavior and text description.
- We also investigate a novel CIAN-Explainer, which provides a variety of benefits, from the ability to visualize semantically relevant features to interpretability, to saving a huge amount of human-resources for manually auditing.
- We validate the performance and evaluate the application value on a practical e-payment dataset, which is with heavy noise. The results show that the performance of the proposed method is significantly better than all the benchmark methods with regard to different evaluation metrics.

2 Problem and Methodology

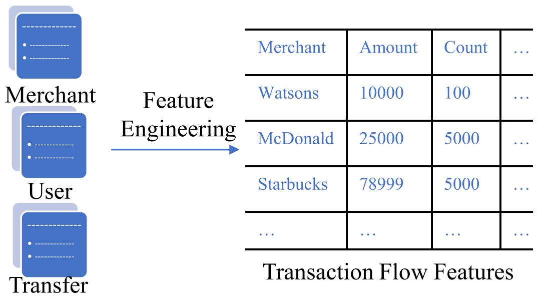
In this section, we first define the problem accurately and then elaborate the details of CIAN: a cross-modal matching approach that learns to match the feature representations from the two modalities in order to perform both text-to-transaction and transaction-to-text retrieval.

2.1 Motivation

In a multimedia e-commerce system (e.g., WeChat Pay), there is a set of resident merchants $\mathcal{V}$. Each merchant $i \in \mathcal{V}$ is composed of its transaction flow logs T_i (transaction information) and descriptions D_i (text information). Without confusion, we would use the term of transaction and text interchangeably in this paper. In the e-commerce system, each merchant is only allowed to carry out the designated business. In order to avoid potential harm to consumer and small business borrowers, merchants should be appropriately regulated at their licensed business scope. As a result, if the descriptions of a merchant are conflict with its transaction logs, the merchant will be suspended or limited. However, the merchants could fake the descriptions to match the designated business, which are at high-risk of regulation. To tackle this problem, we propose CIAN to investigate the transaction-text matching problem and figure out the dismatching merchants, which could significantly enhance the FinTech regulation.

2.2 Joint Feature Learning

In our multimodal matching problem, one of the main objectives is to learn discrimiative transaction and text feature representations that accurately retrieve transaction/text from text/transaction. Fig. 1 shows the the training procedure. Specifically, the CIAN framework consists of a transaction encoder and a text encoder, which are described in detail below.



Merchant	Amount	Count	…
Watsons	10000	100	…
McDonald	25000	5000	…
Starbucks	78999	5000	…
…	…	…	…

Figure 2: Transaction flow features are extracted from user profile, merchant profile, transfer environment, and etc.

Transaction Encoder

As the problem of AI-enabled regulation is vital to a financial business, efforts have been spent for years, where about 415 features are carefully engineered in WeChat Pay system. We call such features as transaction flow features. Specifically, given transaction flow logs $T_i \in \mathcal{T}$, we obtain the transaction features $\boldsymbol{t}_i = [t_{i,1}, t_{i,2}, ..., t_{i,415}] \in \mathbb{R}^{415}$ (Fig. 2).

Text Encoder

Similarly, the corresponding description of a merchant is also represented by features through a text encoder. Thanks to the recent advance in NLP, we adopt RoBERTa [Liu *et al.*, 2019] as the text encoder, where all sentences are padded and truncated to the same length 128. Given the description of merchant $D_i \in \mathcal{D}$, we obtain the text features $\boldsymbol{d}_i = [d_{i,1}, d_{i,2}, ..., d_{i,768}] \in \mathbb{R}^{768}$.

Objective

We denote the transaction and text encoder as function $f : \mathcal{T} \rightarrow \mathbb{R}^{1024}$ and $g : \mathcal{D} \rightarrow \mathbb{R}^{1024}$, which map transaction and text to vectors with the same dimension 1024, respectively (Fig.1). For a transaction-text pair (T_i, D_i) of merchant i, the similarity S_i is measured by cosine similarity:

$$S_i = \left\langle \frac{f(T_i)}{||f(T_i)||_2}, \frac{g(D_i)}{||g(D_i)||_2} \right\rangle : \mathcal{T} \times \mathcal{D} \rightarrow \mathbb{R}. \quad (1)$$

According to the similarity, the network is trained by the incremental-margin triplet loss [Zhang *et al.*, 2019], which clusters positive pairs and push negative pairs away from each other. There are mainly two kinds of blocks inside f and g: cross-modal attention blocks and intra-modal attention blocks, which will be introduced in the next section.

2.3 Attention Mechanism

Intuitively, the transaction behavior and description of a merchant are highly correlated and coupled. For example, if the description of a merchant indicate it is a restaurant, the meal-time transaction flows may express the main semantics of the merchant with higher probability, while the genders of the customers may not be significant. However, if the description is highly related to makeup, the genders of the customers become very essential. To tackle this issue, we propose the cross-modal attention blocks and inter-modal attention blocks as shown in Fig. 1. In the remaining of this section, we explain the underlying attention mechanisms employed at the blocks.

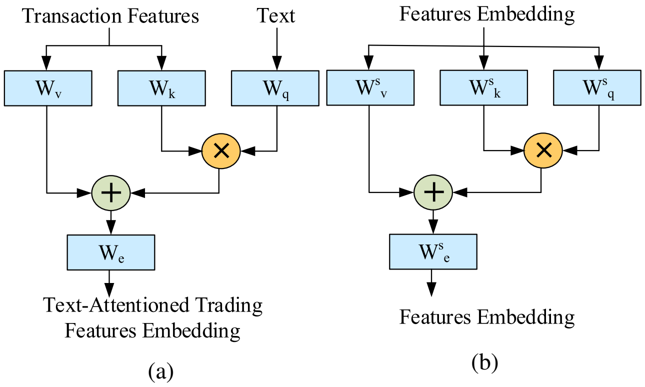


Figure 3: (a) Text to Transaction Attention Block. (b) Intra-modal Attention Block. $\times$ denotes the inner product operation; $+$ denotes the element-wise multiplication operation.

Cross-modal Attention

The cross-modal attention blocks learn to capture the interaction between different modalities. In particular, we adopt the text to transaction attention block (Fig. 3a) as an example to illustrate how the cross-modal attention works. Then, the transaction to text attention block could be easily understood by swapping the transaction features and text features. Each merchant transaction and text features are first transformed into value, key, and query features following [Vaswani *et al.*, 2017; Gao *et al.*, 2019], where the transformed features are denoted as $\boldsymbol{v}_i$, $\boldsymbol{k}_i$, $\boldsymbol{q}_i \in \mathbb{R}^{\text{dim}}$. Notation dim represents the common dimension of transformed features from both modalities[1]. Given merchant i, the block first captures the value $\boldsymbol{v}_i$ and key $\boldsymbol{k}_i$ from the transaction features:

$$\begin{aligned} \boldsymbol{v}_i &= \boldsymbol{W}_v \times \boldsymbol{t}_i + \boldsymbol{b}_v, \\ \boldsymbol{k}_i &= \boldsymbol{W}_k \times \boldsymbol{t}_i + \boldsymbol{b}_k, \end{aligned} \quad (2)$$

and the query $\boldsymbol{Q}_i$ from the text features:

$$\boldsymbol{q}_i = \boldsymbol{W}_q \times \boldsymbol{d}_i + \boldsymbol{b}_q, \quad (3)$$

where $\boldsymbol{W}_v$, $\boldsymbol{W}_k$, $\boldsymbol{W}_q$, and $\boldsymbol{b}_v$, $\boldsymbol{b}_k$, $\boldsymbol{b}_q$ are the weights and bias, respectively. The block then packs $\boldsymbol{k}_i$ and $\boldsymbol{q}_i$ together and calculate their attention to $\boldsymbol{v}_i$:

$$\boldsymbol{a}_i = \text{softmax}\left(\frac{\boldsymbol{k}_i \boldsymbol{q}_i^T}{\sqrt{\text{dim}}} \right), \quad (4)$$

where $\boldsymbol{a}_i \in \mathbb{R}^{\text{dim}}$ is the set of attention weights. The dot product values are proportional to the dimension of the common feature space. Consequently, the product is normalized by $\sqrt{\text{dim}}$. The softmax non-linearity function is applied row-wisely. Notice that the package operation could be Gaussian $e^{\boldsymbol{k}_i \boldsymbol{q}_i^T}$, Kernel Gaussian $e^{\phi(\boldsymbol{k}_i)\phi(\boldsymbol{q}_i)^T}$, dot product $\boldsymbol{k}_i \boldsymbol{q}_i^T$, and etc. Interestingly, we will show by experiments that CIAN is not sensitive to the choices and thus choose the dot product operation for its simplicity and interpretability. With the attention weights, the block outputs the embedding features simply by element-wise multiplication as shown in Fig. 3a:

$$\boldsymbol{o}_i = \boldsymbol{a}_i \circ \boldsymbol{v}_i, \quad (5)$$

where $\boldsymbol{o}_i$ is the cross-modal attention output of merchant i.

[1] In our simulation, we set $\text{dim} = 1024$.

Intra-modal Attention

On top of the cross-modal attention, we also observe the relationships within each modality are complementary to the cross-modal relations. For instance, we should pay more attention to the gender related features when we figure out the merchant is related to makeup. Meanwhile, if the gender related features are essential, we should also focus on the repurchase rate due to the difference of purchasing habits among different genders.

Therefore, we also design intra-modal attention block in f and g. The implementation of intra-modal attention block is illustrated at Fig.3b, which differs from the cross-modal attention block only on the input and thus we omit the details due to the page limitation.

3 CIAN-Explainer

To sustain the requirements of financial regulation, we should not only detect whether the transaction flow of a merchant is beyond the scope of its licensed business but also figure out the reason of the dis-match. Consequently, we also design a CIAN-Explainer to interpret how the attention mechanism interact the original features. In this section, we propose a CIAN-Explainer, an approach for explaining matching made by CIAN. The proposed CIAN-Explainer takes a trained CIAN and its match results as input, and returns an explanation in the form of a small subset of transaction flow features that are most crucial for the match (Fig. 4). CIAN-Explainer specifies an explanation as a rich subset of the entire transaction flow information, such that the subset minimizes the distance between the approximate embedding and real embedding of CIAN match. Based on this, we formulate CIAN-Explainer as a low-rank approximation problem:

$$\min_{\hat{a}} \ ||W_v\left(t_i \circ \hat{a}\right) + b_v - \boldsymbol{a}_i \circ \boldsymbol{v}_i||_2^2 \tag{6a}$$

$$\text{s.t.} \ ||\hat{a}||_1 \leq 10, \tag{6b}$$

$$||\hat{a}||_2 = 1. \tag{6c}$$

Unfortunately, the constraint $||\hat{a}||_1 \leq 10$ describes a non-convex set, which makes Problem (6) hard to solve. Thanks to the duality principle, we could re-formulate CIAN-Explainer as the dual problem of Problem (6):

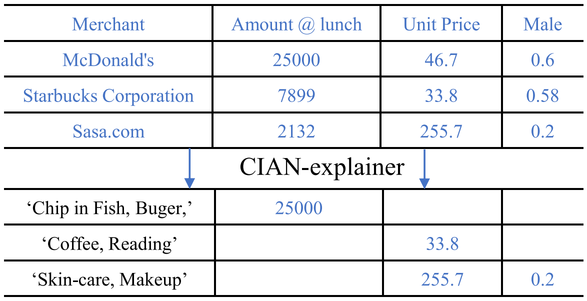
$$\min_{\hat{a}} \ ||W_v\left(t_i \circ \hat{a}\right) + b_v - \boldsymbol{a}_i \circ \boldsymbol{v}_i||_2^2 + \lambda||\hat{a}|| \tag{7a}$$

$$\text{s.t.} \ ||\hat{a}||_2 = 1, \tag{7b}$$

where λ is the Lagrange multiplier and the problem could be solved by alternating direction method of multipliers (ADMM). As a result, the explainer learns a real-valued feature mask $\hat{a}$ which selects the important transaction flow features and masks out unimportant ones. In our experiments, we evaluate CIAN-Explainer and the results show that CIAN-Explainer provides consistent and concise explanations of CIAN match. The visualization of CIAN-Explainer will be shown in Section 4.3.

4 Experiments

In order to demonstrate the effectiveness of our proposed methods, we conduct CIAN on a practical e-payment dataset: WeChat Pay. We systematically make comparisons with several latest start-of-the-art methods and thoroughly investigate the performance of the proposed CIAN framework. As for the performance measure criterion for transaction-text matching, we apply the commonly used precise, recall, and F1-Score as the performance evaluation metrics.



Merchant	Amount @ lunch	Unit Price	Male
McDonald's	25000	46.7	0.6
Starbucks Corporation	7899	33.8	0.58
Sasa.com	2132	255.7	0.2

CIAN-explainer

‘Chip in Fish, Buger,’	25000		
‘Coffee, Reading’		33.8	
‘Skin-care, Makeup’		255.7	0.2

Figure 4: CIAN-Explainer specifies an explanation as a rich subset of the entire transaction flow information.

4.1 Implement Details

WeChat Pay Dataset

For a merchant, its transaction flow records make up a basic data identification, and the remarks and comments are regarded as the description of this transaction flow. From the real-world datasets in WeChat Pay, we extract two sub-datasets for the evaluation. In the first sub-dataset, namely 31-D dataset, we collect 50,000 merchants from 2019/07/01 to 2019/07/31 for training, 1,000 merchants from 2019/08/01 to 2019/08/31 for validating, and 1,000 merchants from 2019/07/01 to 2019/07/31 for testing, from WeChat Pay system. In the second dataset, namely 7-D dataset, we collect 50,000 merchants from 2019/07/01 to 2019/07/07 for training, 1,000 merchants from 2019/08/01 to 2019/08/07 for validating, and 1,000 merchants from 2019/07/01 to 2019/07/07 for testing. Through the hardest negative sampler [Hermans *et al.*, 2017], there are total 500,0000 training pairs, 10,000 pairs for validating, and 10,000 for testing. In our datasets, we define the positive pairs as the merchants' transaction flow stats and their corresponding descriptions, and the negative pairs as the merchants' transaction flow stats and the descriptions of the merchants who belong to other categories.

Training Details

All of our experiments are conducted on a machine with an Intel Xeon E5-2630 CPU, two NVIDIA GTX 1080 Ti GPUs, and 64GB RAM. The text embedding is achieved with the pre-trained RoBERTa [Liu *et al.*, 2019].[2] After text embedding and feature engineering, the dimension of transaction and text information are 415 and 768, respectively. For information interaction, the transaction and text information are projected into the same dimension, which is fixed as 1024 in our experiment. The networks are constructed using Pytorch for computational speed boost. The model parameters are initialized with kaiming normal initializer and Adam optimization algorithm is used to train the overall network. Moreover,

[2] https://github.com/brightmart/roberta_zh

Model	Component	Output Layer	7-D Dataset			31-D Dataset		
			Precise	Recall	F1-Score	Precise	Recall	F1-Score
FC	NAN	Fusion Layer+Classification[Cheng *et al.*, 2019]	53.0	57.8	55.3	55.3	60.3	57.7
		Triplet Loss[Wang *et al.*, 2018]	50.0	50.1	50.1	51.2	50.3	50.7
Bi-Linear Attention	NAN	Fusion Layer+Classification [Kim *et al.*, 2018]	57.6	58.1	57.8	60.0	60.2	60.1
		Triplet Loss	65.8	64.1	64.9	69.1	66.7	67.9
CIAN	Intra-modal Attention	Fusion Layer+Classification	57.8	59.1	58.4	59.8	61.3	60.5
		Triplet Loss	69.0	67.0	68.0	71.9	70.0	70.9
	Cross-modal Attention	Fusion Layer+Classification	59.4	58.8	59.1	62.2	61.9	62.0
		Triplet Loss	77.0	68.5	72.5	80.9	**71.4**	75.9
	Both	Fusion Layer+Classification	59.0	58.9	58.9	61.7	61.5	61.6
		Triplet Loss	79.5	69.9	74.4	**85.9**	**71.4**	**78.0**

Table 1: Comparisons of transaction-text matching on WeChat Pay dataset with the competing methods.

we set the batch size to 256, the initial learning rate to 0.01 and the regularizer parameter as 0.01 to prevent over-fitting.

4.2 Embedding Visualization

We use t-SNE [Maaten and Hinton, 2008] to visualize our transaction features before and after cross-modal embedding, coloring each merchant according to its category in Fig.5.

In the two-dimensional visualization, the left figure shows that the transaction features before cross-modal embedding are vaguely clustered but the boundaries merge together and some categories are diffuse. Meanwhile, in the right figure, the transaction features after cross-modal embedding are less noisy and clearly clustered corresponding to the merchant category. The above observations corroborate our motivation of combining intra-modal and cross-modal attention mechanism in our CIAN design. Through cross-model embedding, more attentions are paid to the information that helps identify the scope of business.

4.3 Performance Evaluation

Comparison versus Different Package Operations

We observe through our experiments that CIAN is not sensitive to the choices of the package operation in Eq. (4) and thus choose the dot product operation for its simplicity and interpretability. To validate the observation, we exhibit the comparison among different types of the package operation in Table 2. We can see that the Gaussian, Kernel Gaussian, and Dot product versions achieve the similar performance, up to some random variations (77.5 to 78.3). Due to the simplicity, we adopt dot product in the proposed CIAN framework. Thanks to the simplicity, we proposed the CIAN-Explainer in section III, which provides inspiring insights and save a huge amount of human-resources for manually auditing.

Operation	7-D Dataset			31-D Dataset		
	Precise	Recall	F1-Score	Precise	Recall	F1-Score
Gaussian	**80.5**	**70.3**	**75.1**	85.5	70.9	77.5
Kernel Gaussian	80.1	69.8	74.6	**75.0**	70.5	**78.3**
Dot product	79.5	69.9	74.4	85.9	**71.4**	78.0

Table 2: Comparisons of the choices of package operation in Eq. (4).

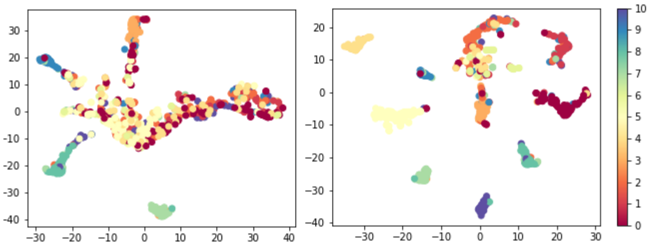


Figure 5: Two-dimensional visualization of transaction features before and after cross-modal embedding. Colors represent : Restaurant, Shopping Mall, Travel, Express, Education, Technology, Healthy, Jewellery, Monopoly, Clothes, and Makeup from 0 to 10.

Comparison with Competing Methods

Fig. 6 shows the curves of the training procedure of our proposed CIAN framework *vs* the Bi-Linear Attention baseline [Kim *et al.*, 2018] with 31-D dataset. Under both 7-D dataset and 31-D dataset, our proposed CIAN model is consistently better than the Bi-Linear Attention baseline throughout the training procedure, in both training and validation er-

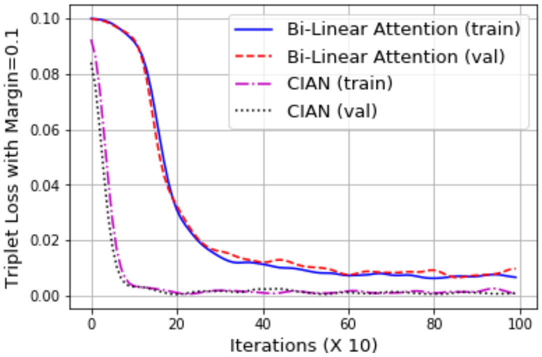
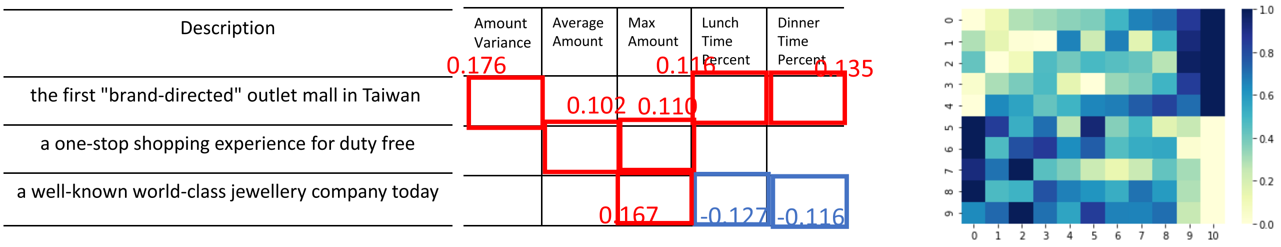


Figure 6: Curves of the training procedure for CIAN *vs* the Bi-linear Attention baseline with 31-D dataset. We show the training and validation triplet loss, i.e., the objective function during the training procedure. The matching precise and recall results are in Table 1.



(a) Examples of the relationship between the description and CIAN-Explainer.

(b) Projected attention of different categories.

Figure 7: Visualization of CIAN-Explainer. In (b), the columns represent 11 categories: Restaurant, Shopping Mall, Travel,Express, Education, Technology, Healthy, Jewellery, Monopoly, Clothes, and Makeup from 0 to 10; the rows stand for 10 selected features, which is not elaborated due to the non-disclosure agreement.

ror. Moreover, the convergence rate of CIAN is much faster than the baseline and the converged performance of CIAN also outperforms the baseline.

Table 1 exhibits the performances of all methods on WeChat Pay dataset. The major findings from the experimental results can be summarized as follows:

- Our model outperforms all the baselines, which indicates our model adopts a more principled way to leverage intra-modal information and cross-modal relations for improving match performance. The performance of all the methods under 31-D dataset is notably better than the one under 7-D dataset. Our model achieves 85.9% precise, which is a remarkable performance under the real-world data with randomness and noise.
- Among these baselines, we can find that the overall performance order is as follows: (attention+triplet loss) based methods, (fully-connected layers+triplet loss) based methods, (attention+fusion layer+classification) based methods, (fully-connected layers+fusion layer+classification) based methods. It indicates that the better performances can be achieved through attention mechanism and two-branch embedding structure with triplet loss.
- Comparing the only cross-modal attention network and only intra-modal attention network, we can find that the cross-modal attention network outperforms the intra-modal attention one, which further demonstrates the importance of capturing the correspondences inside the trading behavior and between the trading behavior and the meaning of the sentences.

4.4 Explainer Visualization

To sustain the requirements of financial regulation, we design a CIAN-Explainer to interpret how the attention mechanism interact the original features in Section III. In Fig. 7, we visualize the CIAN-Explainer to analyze transaction and text matching for WeChat Pay. An exemplary visualization result is shown in Fig. 7a. There are two major observations:

- The correspondences between the text and transaction are satisfied, most of the phrases can attend to their related transaction features, like the phrases “outlet mall”, “duty free”, “Jeweller”, and so on.
- The sign of the attention indicates the positive or negative text-transaction correlation. The absolute value of attention could filtered out unrelated information flows, like the average transaction amount of a outlet mall.

Intuitively, the merchants with similar text should focus on relative features. In order to see if the learned attention preserves the similarity, we cluster the merchants with their categories and then plot the average attentions for selected 10 features in Fig. 7b. We observe that the closer categories share higher attention similarity in most cases. For example, merchants of Restaurant and Shopping Mall, i.e., column 0 and 1 in Fig. 7b, both pay more attentions on the features related to the transaction time distribution and pay less attentions on the features related to the gender distribution. However, merchants of Clothes and Makeup, i.e., column 9 and 10 in Fig. 7b, focus on the features related to the repurchase rate and gender. As a consequence, CIAN-Explainer could not only visualize relevant features, but also save a huge amount of human-resources for manually auditing, which is quiet important in financial business.

5 Conclusion

In this paper, we develop a novel cross-modal and intra-modal attention network (CIAN) for the transaction-text matching task. Our proposed CIAN framework can be adapted to any cross-modal matching problem, e.g., image-text, image-transaction, transaction-text. We show the CIAN framework alternatively passes information within and across transaction and text based on the proposed attention mechanism. To sustain the requirements of financial regulation, we also design a CIAN-Explainer to interpret how the attention mechanism interacts the original features. Furthermore, we collect a dataset from a practical online payment (WeChat Pay) and make the first attempt to evaluate the application value of our transaction-text model, which provides solid improvement over baselines. We hope CIAN as well as CIAN-Explainer provides a big insight for both the AI-enabled FinTech and interpretable multimodal learning.

Acknowledgments

This work is supported in part by General Research Funding (project number 14200315) established by Hong Kong Research Grant Council.

References

[Afouras *et al.*, 2018] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018.

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[Cao *et al.*, 2019] Shaosheng Cao, XinXing Yang, Cen Chen, Jun Zhou, Xiaolong Li, and Yuan Qi. Titant: online real-time transaction fraud detection in ant financial. *Proceedings of the VLDB Endowment*, 12(12):2082–2093, 2019.

[Cheng *et al.*, 2019] Chaoran Cheng, Fei Tan, Xiurui Hou, and Zhi Wei. Success prediction on crowdfunding with multimodal deep learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 2158–2164. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[Chu *et al.*, 2016] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. Selective transfer machine for personalized facial expression analysis. *IEEE transactions on pattern analysis and machine intelligence*, 39(3):529–545, 2016.

[Gao *et al.*, 2019] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6639–6648, 2019.

[Gu *et al.*, 2018] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7181–7189, 2018.

[Hermans *et al.*, 2017] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[Kim *et al.*, 2018] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1571–1581. Curran Associates Inc., 2018.

[Lienhart, 1998] Rainer W Lienhart. Comparison of automatic shot boundary detection algorithms. In *Storage and Retrieval for Image and Video Databases VII*, volume 3656, pages 290–301. International Society for Optics and Photonics, 1998.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Ma *et al.*, 2019] HaoJie Ma, Wenzhong Li, Xiao Zhang, Songcheng Gao, and Sanglu Lu. Attnsense: Multi-level attention mechanism for multimodal human activity recognition. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3109–3115. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[Maaten and Hinton, 2008] L.J.P. Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. 2008.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[Verma *et al.*, 2019] Sunny Verma, Chen Wang, Liming Zhu, and Wei Liu. Deepcu: Integrating both common and unique latent information for multimodal sentiment analysis. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3627–3634. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[Wang *et al.*, 2018] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018.

[Wang *et al.*, 2019] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. Position focused attention network for image-text matching. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3792–3798. AAAI Press, 2019.

[Yang *et al.*, 2019] Yang Yang, Ke-Tao Wang, De-Chuan Zhan, Hui Xiong, and Yuan Jiang. Comprehensive semi-supervised multi-modal learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4092–4098. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[YouTube, 2020] YouTube. Youtube statistics. https://www.youtube.com/yt/press/statistics.html, 2020. Accessed: 2020-01-01.

[Zhang *et al.*, 2016] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *European conference on computer vision*, pages 766–782. Springer, 2016.

[Zhang *et al.*, 2019] Yingying Zhang, Qiaoyong Zhong, Liang Ma, Di Xie, and Shiliang Pu. Learning incremental triplet margin for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9243–9250, 2019.