Spatio-Temporal Attention-Based Neural Network for Credit Card Fraud Detection

Dawei Cheng, Sheng Xiang, Chencheng Shang, Yiyi Zhang, Fangzhou Yang, Liqing Zhang*
MoE Key Lab of Artificial Intelligence, Department of Computer Science and Engineering,
Shanghai Jiao Tong University, Shanghai, China
{dawei.cheng, yi95yi, lake_titicaca}@sjtu.edu.cn, zhang-lq@cs.sjtu.edu.cn

Abstract

Credit card fraud is an important issue and incurs a considerable cost for both cardholders and issuing institutions. Contemporary methods apply machine learning-based approaches to detect fraudulent behavior from transaction records. But manually generating features needs domain knowledge and may lay behind the modus operandi of fraud, which means we need to automatically focus on the most relevant patterns in fraudulent behavior. Therefore, in this work, we propose a spatial-temporal attention-based neural network (STAN) for fraud detection. In particular, transaction records are modeled by attention and 3D convolution mechanisms by integrating the corresponding information, including spatial and temporal behaviors. Attentional weights are jointly learned in an end-to-end manner with 3D convolution and detection networks. Afterward, we conduct extensive experiments on real-word fraud transaction dataset, the result shows that STAN performs better than other state-of-the-art baselines in both AUC and precision-recall curves. Moreover, we conduct empirical studies with domain experts on the proposed method for fraud post-analysis; the result demonstrates the effectiveness of our proposed method in both detecting suspicious transactions and mining fraud patterns.

Introduction

Credit card fraud is a general term for the unauthorized use of funds in a transaction typically through a credit or a debit card (Bhattacharyya et al. 2011). Global card fraud losses amounted to over 25 billion US dollars in 2018 and is forecast to continue to increase (Wang, Chen, and Chen 2019). This huge amount of losses has increased the importance of fraud-fighting. Figure 1 shows a typical fraud detection framework deployed in a commercial system. The card alliance or banks, such as VISA, MasterCard or Citibank, assess each transaction with an online predictive model once it has passed card checking. Unlike a simple card checking system, which focuses on card blacklists, budget checking, etc., the predictive model is designed to detect fraud patterns automatically and produces a fraud risk score. Investigators can thereby focus on the high-risk transactions effectively and feedback the analysis results to the predictive model for model updating.

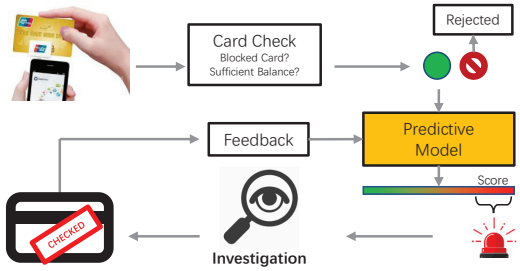

Figure 1: The framework of credit card fraud detection.

As attacking strategies from potential fraudsters change, it is essential that a well-behaved system can adapt to the evolving strategies (Randhawa et al. 2018; Jiang et al. 2018). We summarize the following two major observations from real-world fraud transactions: 1). *Temporal aggregation.* Fraudsters are subject to the limited time of the activities. As the cardholder will freeze the card as soon as possible once suspicious transactions have been detected, fraudsters are required to reach the credit limit in a short time. That means the behaviors of the fraud transaction would be exposed in a limited time. 2). *Spatial aggregation.* Fraudsters are subjected to cost on the devices and merchants of transactions. That is, due to the economic constraints, fraudsters will use the card frequently with only a small number of merchants, which are spatially different from normal transactions.

Many existing models to deal with fraud transactions have been extensively studied (Patidar, Sharma, and others 2011; Bahnsen et al. 2016; Carneiro, Figueira, and Costa 2017). They mainly split into one of two directions: 1). *Rule-based methods* directly generate sophisticated rules by domain experts for identification; for example, (Seeja and Zareapoor 2014) proposed an association rules method for mining frequent fraud rules. 2). *Machine learning-based methods* learn static models by exploring large amounts of historical data. For example, (Fiore et al. 2017) extracted features based on neural networks and built supervised classifiers for detecting

*Corresponding Author

fraudulent transactions. (Fu et al. 2016) advanced the usage of automatic feature engineering in a convolutional neural network (CNN). (Randhawa et al. 2018) applied AdaBoost and majority voting on fraud records. (Jurgovsky et al. 2018) researched on this task by a sequence LSTM model. However, all these methods require manually constructing features before feeding into a classification model, which fails to automatically learn the joint impact on spatial and temporal patterns, as the spatio-temporal patterns have been observed as the main weaknesses of fraudsters, also reported by (Gómez et al. 2018).

Recently developed attention mechanisms have shown the benefit on automatic feature learning (Vaswani et al. 2017; Cheng et al. 2019b). The superior performance of 3-dimensional (3D) convolution on spatio-temporal feature learning is also demonstrated in a wide range of prediction tasks (Allamanis, Peng, and Sutton 2016). In credit card fraud detection task, it is important to jointly consider the "temporal aggregation" and "spatial aggregation" together and then drive them into a representative and deep classifier which well-suited for spatio-temporal feature learning.

Therefore, in this paper, we present the STAN model for credit card fraud detection, a novel deep learning-based method, which jointly considers "temporal aggregation" and "spatial aggregation" in an attention network. Our proposed approach first construct raw transaction recodes into spatio-temporal based feature slices, then we use an attention mechanism to adaptively learn the importance of different slices. To uncover the hidden fraud patterns, we introduce a 3D convolution layer to capture intrinsic relationships among spatio-temporal patterns. During experiments, we show that the results of the proposed method significantly outperform the results from other state-of-the-art baselines.

In brief, the main contributions of this paper include:

- We present a novel attention-based 3D convolution neural network for credit card fraud detection by jointly capturing two weaknesses displayed by fraudsters, summarized as "temporal aggregation" and "spatial aggregation". To the best of our knowledge, this is the first time that a fraud detection problem has been addressed by spatio-temporal attention neural network approaches with a 3D convolutional mechanism.
- Our approach is extensively evaluated in a real-world credit card fraud post analysis system, hosted by a major financial institution. The experimental results demonstrate the superiority of our proposed methods, which could detect more fraud transactions with relatively high precision compared with state-of-the-art baselines.

Preliminaries

In this section, we first briefly present some data analysis to support our intuitions and then present the problem definition of our work.

Spatio-temporal Analysis

Figure 2 visualizes the scaled spatio-temporal feature slices, where the left part shows fraud transactions and the right part shows legitimate ones. We will describe the detailed

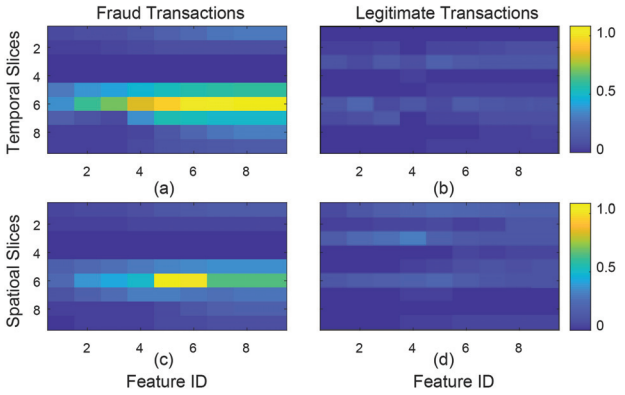

Figure 2: Heat maps of spatio-temporal feature slices from both fraudulent and legitimate transactions.

feature extraction steps in the next section. It can be seen that in temporal analysis, fraud features (shown in Figure 2a) change abruptly across different slices, while legitimate ones are much more slight (shown in Figure 2b). It confirms our original assumption of "temporal aggregation".

In spatial analysis, we encode transaction merchants into their location codes and aggregate features according to the codes within a fixed time window (we set it to days here). Figures 2c to Figure 2d show the heat map of features in spatial slices. As we can see, fraud transactions are obviously located in only a small number of zones, which means fraudsters would use the card frequently under the constraint of locations or devices, while for the normal transactions, there are no noteworthy patterns for user consuming behavior in given time windows. As a result, this set of data analysis validates our original assumption.

Problem Definition

Transaction A transaction means the use of a credit card by a consumer u to purchase commodities or services. The purchase price is sent through a processor for authorization; if the amount a is approved it is automatically submitted to the merchant m in location l.

Transaction Record A transaction record r can be defined as a tuple of attributes in a transaction payment process $r = \{u, t, l, a\}$, where u denotes the user, t and l is the time stamp and location of the transaction, and m means the amount of this payment.

Fraud Event A fraud event d in this paper refers to a transaction which is not authorized by its cardholder. A fraud event is a special type of transaction, which means it also preserves $\{u, t, l, a\}$ attributes.

The complete real-world fraud event data provided by our collaborating institution offers us the unique opportunity to tackle the problem of fraud detection. In conclusion, we now formalize our credit card fraud detection problem as follows:

Given a set of transaction records $\mathcal{R} = \{\mathcal{U}, \mathcal{T}, \mathcal{L}, \mathcal{A}\}$, a set of fraud events $\mathcal{D}$, which are a subset of the transaction collection $\{\mathcal{D}|\mathcal{D} \subset \mathcal{R}\}$, and time period t_i & t_{i+1}, for each

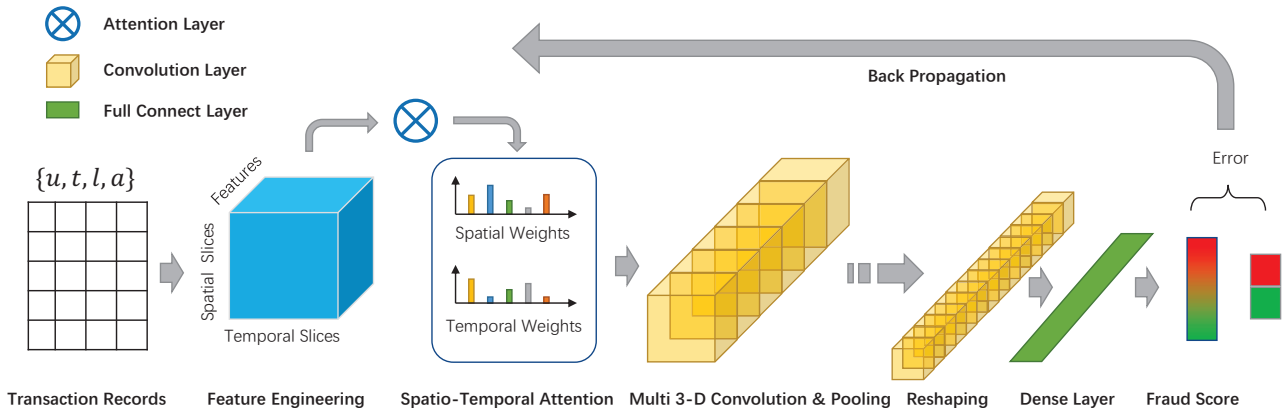

Figure 3: The illustration of the proposed spatio-temporal attention-based neural network (STAN) model. Raw transaction records are processed by feature engineering, spatio-temporal attention, and multiple 3D ConvNet to learn high-level representations. Afterward, the learned representations are reshaped to vectors and fed into a detection network for fraud estimation. Attentional weights are jointly optimized in an end-to-end mechanism with 3D convolution and detection networks.

transaction, we want to infer the possibility of whether it is a fraud event, based on the transaction records and fraud events from t_1 to t_i. The objective is to achieve a high accuracy of fraud prediction, as well as to explore the fraud patterns of credit card transactions.

The Proposed Approaches

In this section, we first introduce the framework of a spatio-temporal attention-based neural model. After that, we present the process of feature engineering, the spatio-temporal attention layer, the 3D convolution network (3D ConNet) and the detection layer. Lastly, we introduce the optimization strategy of the proposed methods.

Model Architecture

Figure 3 shows the general network architecture of STAN. The model takes users' transaction records as input and transfers them into high-order tensor spaces in spatial, temporal and feature orders. Then, we apply spatio-temporal attention and the 3D convolution layer to obtain a transaction representation vector. Spatio-temporal attention helps to draw information from tensor features by different weights and the 3D convolution layer helps to model hidden patterns of transactions. Finally, we reshape the learned feature representation from tenors to vectors for the fraud detection task by a detection network. We will first introduce each component of the model, then discuss the settings of detection layer and optimization in the following sections.

Feature Extraction

For given transaction records $\mathbf{r} = (r_1, r_2, \cdots, r_n)$, each record $r_i = \{u, t, l, a\}$, contains: user ID, timestamp, location code and transaction amount. In preprocessing, we combine users who maintain multiple credit cards into a user ID and filter out inactive users that have less than 10 records within one month. As the number of users who have never been charged with unauthorized transactions, is much larger than the number of users who are affected, we adopt user-level downsampling of normal users instead of transaction-level sampling, to maintain the fraud patterns during preprocessing.

Afterward, we construct the feature representation of each record into tensor format $\mathcal{X} \in \mathcal{R}^{N_1 \times N_2 \times N_3}$, where N_1, N_2, N_3 denote the dimensions of temporal, spatial and feature slices.

Temporal Slices $\mathcal{X}(t, :, :)$. Each temporal feature represents a vector generated in a given time window. The number and diversity in temporal slices reflect its activeness and hence are related to the consuming behavior of the user. We thus extract features in temporal slices including (1) features in the latest 1 second, minute, hour, day, week, month and quarter; (2) features in the last 1, 10, 100, 1000 transactions.

Spatial Slices $\mathcal{X}(:, l, :)$. Based on the observation that fraudsters are affected by the constraints of location, we collect zip codes, the business center from the State Postal Bureau and AliTrip, and divide the features into four levels manually according to the location. They are a one-hot representation of the location ID in the nation, state, city and business center levels and we concatenate them in the spatial slice.

Features $\mathcal{X}(:, :, f)$. Inspired by Fu's work (Fu et al. 2016) that the transaction entropy is one of the important patterns in fraud detection, we identify the extracted features including current amount, average amount, total amount, transaction times and the most recent location.

Spatio-temporal Attention Net

The attention network aims to perform proper credit assignment to the spatial and temporal slices according to their importance in the current transaction. It contains two self-attention layers targeting temporal and spatial slices respectively.

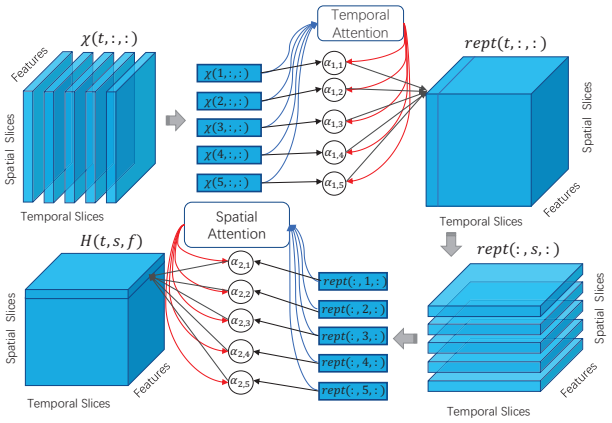

Figure 4: Illustration of spatio-temporal attention neural networks.

Temporal Attention Layer Formally, given the extracted feature tensor $\mathcal{X}$ as described above, the temporal attention layer represents the transaction by a weighted sum of the matrix representation of all the temporal slices. Mathematically, it takes the form as follows:

$$rept = \sum_{t=1}^{N_1} a_{1,t}\mathcal{X}(t,:,:) \tag{1}$$

$$a_{1,t} = \frac{\exp\left((1-\lambda_1)\cdot g_t(W_t, \mathcal{X}(t,:,:))\right)}{\sum_{t=1}^{N_1}\exp\left((1-\lambda_1)\cdot g_t(W_t, \mathcal{X}(t,:,:))\right)} \tag{2}$$

where: $a_{1,t}$ is the weight for each temporal slice, and $g_t(\cdot)$ is a fully connected layer with ReLU activation and parameters W_t; $\lambda_1 \in [0,1]$ is the temporal penalty factor to control the importance of temporal attention; $rept$ is the output of the temporal attention layer. It should be noted that we unfold matrices $\mathcal{X}(t,:,:)$ to row vectors for computational convenience and reshape the output $rept$ into tensor format $rept \in \mathbb{R}^{N_1\times N_2\times N_3}$.

Spatial Attention Layer Given the output from the temporal net $rept$, we then apply a spatial attention mechanism on the top of the temporal net. It is formulated as follows:

$$\mathcal{H}^a = \sum_{s=1}^{N_2} a_{2,s} rept(:,s,:) \tag{3}$$

$$a_{2,s} = \frac{\exp\left((1-\lambda_2)\cdot g_s(W_s, rept(:,s,:))\right)}{\sum_{s=1}^{N_2}\exp\left((1-\lambda_2)\cdot g_s(W_s, rept(:,s,:))\right)} \tag{4}$$

where W_s is the weight of spatial network g_s; $\mathcal{H}^a$ is the output of attention layer, we reshape it into tensor format with the same order as $\mathcal{X}$; $a_{2,s}$ is the weight for each spatial slices; $\lambda_2 \in [0,1]$ is the spatial penalty factor to control the importance of spatial attention.

3D Convolutional Layers

For our mission, CNN is an attractive option for three main reasons. First, they can clearly exploit the spatial features of our problem. In particular, they can learn local spatial filters that are useful for classification tasks. Second, by stacking multiple layers, the network can learn more complex features from input spatial spaces. Finally, the optimization of CNN could be learned by SGD based methods, which can be performed efficiently with commercial graphics hardware.

Compared to 2D convolution networks, 3D ConvNet is ideal for spatio-temporal learning of features. Due to 3D convolution and 3D pool operations, 3D ConvNet works temporally and spatially, whereas in 2D ConvNet it is only spatially executed. In general, the following equation represents a 3D convolution operation:

$$repc_i^c(t,l,f) = \sum_{m,n,o} \mathcal{H}^{c-1}(t-m, l-n, f-o)\mathcal{W}_i^c(m,n,o) \tag{5}$$

in which $\mathcal{W}_i^c$ is the 3D kernel in the $c-th$ layer and $i-th$ kernel which convolves over the feature $\mathcal{H}^{c-1}$, and $\mathcal{W}_i^c(m,n,o)$ is the element-wise weight in the 3D convolution kernel. Thus, the feature $\mathcal{H}^c$ is obtained by different 3D convolution kernels:

$$\mathcal{H}^c = \sigma\left(\sum_i repc_i^c + b^c\right) \tag{6}$$

where σ denotes the sigmoid function.

Then we hierarchically build a deep 3D ConvNet model by stacking convolutional layers (represented as C) and pooling layers (represented as P). In particular, multiple 3D feature volumes are generated in the C layer. In the P layer, the maximum pool operation is also performed in 3D, that is, the feature volume is subsampled based on the cube neighborhood. In the fully connected layer, the 3D feature volume is flattened into a vector as input.

Detection Layer and Optimization

The fraud detection task takes the transaction representation rep, which is the tensor flatted vector learned by attention and convolution networks, and aims to learn the probability of whether it's a fraudulent trade. The loss function is the likelihood defined as follows:

$$\begin{aligned}\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\text{detect}(rep_i:\theta)) \\ +\lambda_3(1-y_i)\log(1-\text{detect}(rep_i:\theta))]\end{aligned} \tag{7}$$

where: rep_i denotes the representation of the $i-th$ transaction record, which is the output of 3D ConvNet, and λ_3 indicates the sample weight according to the biased distribution of fraud and legitimate records; y_i denotes the label of $i-th$ records, which is set to 1 if the record is fraud and 0 other wise; $\text{detect}(rep_i)$ is the detection function that maps rep_i to a real valued score, indicating the probability of whether the current transaction is fraudulent. We implement $\text{detect}(rep_i:\theta)$ with two-layer ReLu and one-layer sigmoid network.

The proposed STAN can be optimized through the standard SGD-based algorithms. In this paper, we used the Adam Optimizer to learn the parameters. We set the initial learning rate to 0.001, and the batch size to 256 by default.

Experiments

Experiment Settings

Datasets We collected fraud transactions from a major commercial bank, which comprises real-word credit card transaction records spanning twelve months, from Jan 1 to Dec 31, 2016. The ground truth labels are reported by consumers and confirmed by domain experts in the financial institution. We first filtered suspicious records by domain experts and user reports. Based on this process, the large amount of user records with both legitimate and inactive transactions were excluded. If a trade is reported by a cardholder or identified by financial experts as fraudulent, we label it as 1; otherwise, it is labeled as 0. Finally, the dataset includes 236,706 transaction records, by 1021 users, across 1160 location codes, which were affected by fraud.

In preprocessing, each transaction includes four attributes. We simplify them as: user ID, timestamp, location code and transaction amount. We encode categorical data, such as user ID and location code, into one-hot representations. We round the time record from the millisecond level to a standard DataTime format (yyyy-MM-dd HH:mm:ss). For the amount attribute, like many other financial signals, it performs a distribution with a long tail. We first cut off the outliers by the three-sigma rule (Friedrich Pukelsheim et al. 1994) and then perform a log transform on the amount value.

Compared Methods We employ the following state-of-the-art methods on our benchmark dataset to highlight the effectiveness of the proposed STAN. In these experiments, the tasks are learned independently. These baseline includes: LR (Logistic Regression) (McMahan 2011), GBDT (Ke et al. 2017), MLP (Tang, Deng, and Huang 2015), Deep & Wide (Cheng et al. 2016), CNN-max (Fu et al. 2016), AdaBM (Randhawa et al. 2018), LSTM-seq (Jurgovsky et al. 2018). STAN-notemp/nospat/no3d denotes sub-models of STAN, in which the temporal attention, spatial attention are not used, utilizing the 2D convolution layer instead of this paper's proposed 3D ConvNet. STAN-all is the full proposed spatial-temporal atention-based neural network model in this paper.

Parameter Settings and Evaluation Metrics In this experiment, we apply the preferred parameters for each of the baseline methods as they were originally proposed. For STAN, we employ 2 convolution layers, each of them is set to $4 \times 4 \times 4$ convolution kernel, followed by a max-pooling layer. Two full connected layers are added on the top of 3D ConvNet, each of them consisting of 32 neurons. We set the temporal and spatial parameters (λ_1 and λ_2) by cross validation. The sample weight λ_3 is set by the training distribution.

We evaluated the detected results by precision, recall, and F-Score. In our implementation, we tried all possible threshold probabilities in our KS-test from 0 to 1 with the step size of 0.01. To determine the most effective threshold, we tested the detection result with the ground truth labels. We also report the AUC (area under the ROC curve) in our experiments.

Table 1: Performance comparison with baselines.

	AUC (Oct)	AUC (Nov)	AUC (Dec)
LR	0.7247	0.7163	0.7199
GBDT	0.7868	0.7949	0.7864
MLP	0.7803	0.8012	0.7891
Deep & Wide	0.8210	0.8197	0.8108
CNN-max	0.8352	0.8367	0.8267
AdaBM	0.8243	0.8249	0.8232
LSTM-seq	0.8368	0.8353	0.8290
STAN-notemp	0.8467	0.8395	0.8406
STAN-nospat	0.8602	0.8531	0.8583
STAN-no3d	0.8569	0.8562	0.8506
STAN-all	**0.8832***	**0.8789***	**0.8865***

Fraud Detection

We evaluated the performance of different models for the fraud detection task. Records of the first nine months were used as training data and then we predicted the fraud transactions in the following three months (Oct, Nov and Dec). We repeated the experiment 10 times and report the average AUC in Table 1.

The first seven lines of Table 1 contain the results of some of the latest baselines. In all baselines, CNN-max and LSTM-seq proved to be competitive, demonstrating the necessity of deep models for fraud detection. Lines 8-11 show the results of STAN and some of its submodels. STAN-notemp's performance is close to CNN-max, the spatial-attention layer prove to be effective. STAN-nospat and STAN-no3d perform much better than the baselines and STAN-notemp and we validated the essentials of each submodule. STAN-all outperforms all the other models.

Precision-recall Curves

In Figure 5 we present the precision-recall curves for the lastest baselines. As shown, our proposed STAN performs better than baselines with respect to the area under the precision-recall curves. The results of AdaBM and Deep & Wide are quite similar, both of them are much better than LR. Essentially, this might because that fraud patterns in credit card transaction records are too complex for a simple liner model like LR to address. With the help of deep structures, LSTM-seq perform a slightly promotion compared with AdaBM. In all baselines, LSTM-seq and CNN-max are shown to be the most competitive. The reason might be that they preserve deep representation of raw features and explicitly makes use of the spatial features of our problem, while Deep & Wide and AdaBM are not optimized for local spatial patterns.

Our method, STAN, consistently outperforms other state-of-the-art baselines. The reason is twofold: (1) STAN deals with both spatial and temporal features and integrates them into an attention network, contrasted with CNN-max which only deals with spatial ones that cannot address temporal patterns of transaction records; (2) STAN uses a 3D convolutional network for tensor features instead of 2D convolution so that it can better model spatio-temporal feature learning. Specifically, our methods work as well as, or even better, at

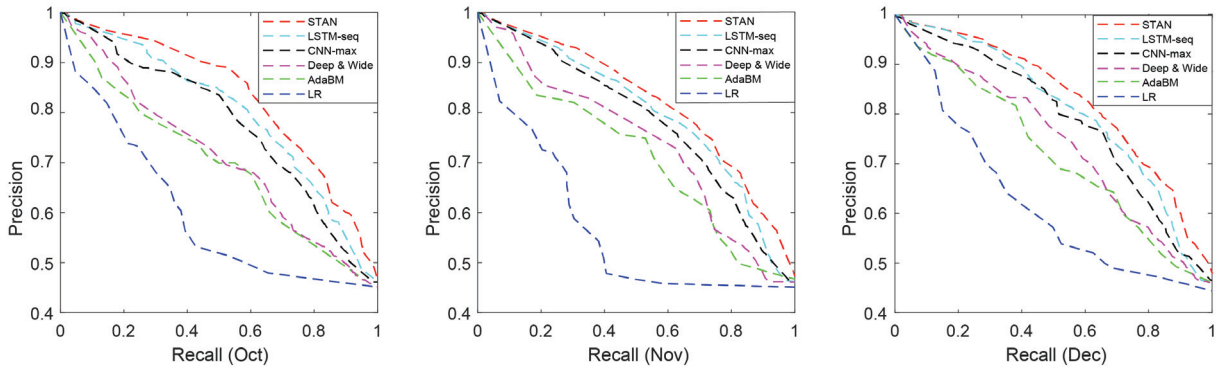

Figure 5: Precision recall curve of STAN compared with baseline methods.

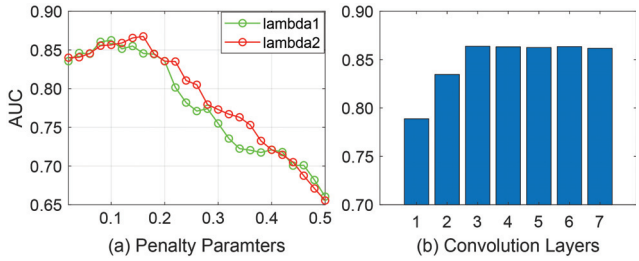

Figure 6: Parameter sensitivity experiment on temporal, spatial penalty parameters and the depth of convolution layers.

Table 2: The value of attention coefficients.

Temporal	Coefficients	Spatial	Coefficients
Seconds	0.2625	#13	0.1468
Minutes	0.0509	#21	0.1092
Hours	0.2053	#36	0.0371
Days	0.0969	#39	0.0308
Weeks	0.2971	#42	0.0227
Months	0.0161	#47	0.0214
Quarters	0.0003	#48	0.0148

the very beginning of the curve compared to the compared methods. More importantly, our methods can accurately detect many more fraud transactions (high recall) with a relatively high precision, which is quite promising.

Parameter Sensitivity

In this section, we study the model generalization which includes penalty parameters, the depth of hidden convolution layers and their impact on our task.

We vary the temporal and spatial penalty parameters (λ_1 and λ_2) from 0 to 1 with a step of 0.02. As shown in Figure 6a, it can be easily found that the parameters, λ, has a great influence to the model performance. Our model performs better by increasing λ from 0 to 0.1, and the AUC reaches the peak when $\lambda_1 = 0.1$ and $\lambda_2 = 0.15$. The performance is degradaded if we keep on increasing the value of λ. The reason might be that varying λ could balance the model consider a proper spatio-temporal window. When we increase λ from 0 to 1, our proposed model could consider features in a different spatio-temporal range and reach a performance peak around $\lambda_1 = 0.1$ and $\lambda_2 = 0.15$.

Figure 6b shows the influence of the depth of hidden convolution layers on the AUC. With the deeper hidden convolution layers, the model tends to aggregate the temporal and spatial information from a neighborhood into a wider range. As we have seen, the AUC with a depth of 1 hidden layer does not work well because the information we have is mixed. Our model needs to "swap" information in terms of temporal, spatial and feature aspects, which requires a convolution of at least two hops to display.

Case Studies

Table 2 shows the learned coefficients of spatial and temporal attention layers, in which "Weeks", "Seconds" and "Hours" weights are noticeable. This is because user behavior normally shows a periodic distribution on a weekly basis, but the fraudulent trades are concentrated in an instant until exceeding the user's credit limit (there could be more than 100 transactions in one second). This phenomenon is also reported by (Lepoivre et al. 2016). In spatial studies, we present the top seven attention weights in Table 2.

In order to uncover the fraud patterns from learned attention weights, we adopted an empirical study on infected accounts with our collaborating domain experts. We firstly randomly selected 1000 fraud transactions and then backtracked other records within one week before the fraud event in the infected account. Finally, we collected the records from infected users into hours and aggregated them by summarizing the spending amount and times of trade location ID, as shown in Figure 7. We get the following observations: *Temporally*: on average, fraud transactions account for over 70% of a user's credit limit, illustrated in hour 166 of the x-axis (fraud event time) of Figure 7a, which means an average of 70% loss for each fraud event. We notice that a small equal number of trades are usually issued in 1-3 hours (between hour 162-165 on the x-axis) before the event. Domain experts have demonstrated they are trade attempts by fraudsters, after analysis of the records. This small number of attempts is important for fraudsters: 1) if successful, the card

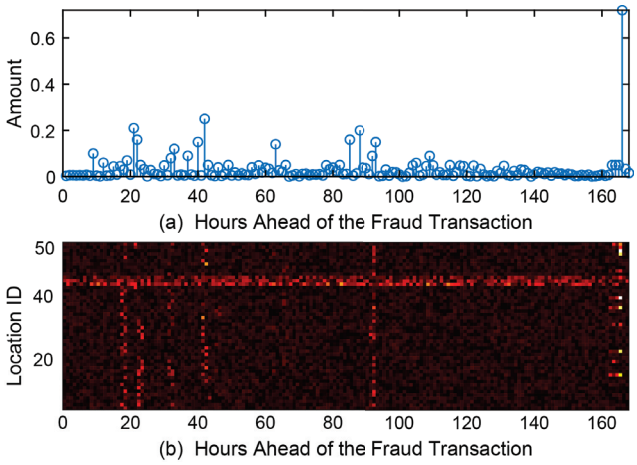

Figure 7: Case studies of attention weights. We randomly extract 1000 fraud transactions and backtracking records in one week before the fraud occurrence. (a) shows the hourly aggregated trading amount. (b) displays the heatmap of transaction locations in an hourly summary.

will be transferred for a large number of fraud transactions; 2) if failed, the cardholder might not notice the tiny amount of failed trade attempts. We also observe that the number of trades in hours 140-160 is low compared with hours 0-140, which means the card user may have missed the card one whole day before the fraud event.

Spatially: users obviously have a location propensity as shown in Figure 7b, where the brighter color (red) means a higher frequency. We observe the two most popular trade merchants are located in ID #42 and #43, which are two popular online payment systems. It should be noted that fraud transactions are concentrated in limited locations, such as #13, #21, etc., which are generally different from user's historically frequent trading locations. This study confirms our intuition of spatial aggregation and learned spatial attention coefficients.

Related work

We summarize the related work in two main areas: 1) attentional convolutional neural networks and 2) credit card fraud detection.

Attentional Convolution Neural Network Many recent works have shown the benefit of combining an attention mechanism in convolutional neural networks for a wide range of prediction tasks (Allamanis, Peng, and Sutton 2016; Vaswani et al. 2017), such as depth estimation (Xu et al. 2018), default prediction (Cheng et al. 2019a) or language understanding (Shen et al. 2018). For instance, pervasive attention are employed on 2D convolutional neural networks for sequence-to-sequence prediction (Elbayad, Besacier, and Verbeek 2018). Attention-gated networks have been considered for integrating multi-scale information in (Xu et al. 2017). In (Chen et al. 2016) an attention model is employed for combining multi-scale features in the context of semantic segmentation and object contour detection. Our approach develops from a similar intuition but further integrates an attention model in both spatial and temporal aspects which significantly improves the accuracy of the detection. To our knowledge this is the first paper exploiting joint learning attention mechanisms with 3D convolutional networks in the context of credit card fraud detection.

Credit Card Fraud Detection Several machine learning techniques have been used in the literature to approach the credit card fraud detection problem. (Maes et al. 2002) tried Bayesian Belief Networks (BBN) and Artificial Neural Networks (ANN) on a real dataset obtained from Europay International. In (Zaki, Meira Jr, and Meira 2014) neural network based models and decision tree models are compared, and the authors found that neural networks outperforms decision trees. The authors in (Fu et al. 2016) prove that using a convolution model to extract spatial patterns can achieve higher accuracy compared with neural networks, SVMs and decision trees. (Randhawa et al. 2018) applied AdaBoost and majority voting on fraud records. (Jurgovsky et al. 2018) researched on this task from sequence classify perspective by improved LSTM model. These methods, however, feed manually generated features into a classification model directly, which ignores the joint feature learning on spatial and temporal patterns. As a result, they may not be appropriate for real-world large scale fraud detection systems with complex and unpredictable fraud patterns. The approach we present in this paper is radically different, as we employ a structured attention model which is jointly learned within a 3D CNN framework.

Conclusion

In this paper, we present a novel attentional 3D convolution neural network for credit card fraud detection. In particular, we uncover the weakness of fraudsters, called "temporal aggregation" and "spatial aggregation", and propose a 3D convolutional neural network approach based on a spatio-temporal attention mechanism. This is the first work in which attentional 3D ConNet has ever been employed to the credit card fraud detection problem. Our methods achieve promising AUC and precision-recall curves compared with other state-of-the-art baseline methods. Furthermore, we explore to uncover fraud patterns by the observation of learned attention weights in case studies. The proposed method is extensively evaluated in an online transaction post-analysis system. The result demonstrates that our methods can effectively detect fraudulent transactions. In the future, we are interested in building a real-time in-process fraud detection system based on an online learning mechanism instead of the offline training approach.

Acknowledgments

The work is supported by the National Key R&D Program of China (2018AAA0100704), the China Postdoctoral Science Foundation, the National Basic Research Program of China (2015CB856004), and the Key Basic Research Program of Shanghai Science and Technology Commission, China (16JC1402800).

References

Allamanis, M.; Peng, H.; and Sutton, C. 2016. A convolutional attention network for extreme summarization of source code. In *International Conference on Machine Learning*, 2091–2100.

Bahnsen, A. C.; Aouada, D.; Stojanovic, A.; and Ottersten, B. 2016. Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications* 51:134–142.

Bhattacharyya, S.; Jha, S.; Tharakunnel, K.; and Westland, J. C. 2011. Data mining for credit card fraud: A comparative study. *Decision Support Systems* 50(3):602–613.

Carneiro, N.; Figueira, G.; and Costa, M. 2017. A data mining based system for credit-card fraud detection in e-tail. *Decision Support Systems* 95:91–101.

Chen, L.-C.; Yang, Y.; Wang, J.; Xu, W.; and Yuille, A. L. 2016. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3640–3649.

Cheng, H.-T.; Koc, L.; Harmsen, J.; Shaked, T.; Chandra, T.; Aradhye, H.; Anderson, G.; Corrado, G.; Chai, W.; Ispir, M.; et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, 7–10. ACM.

Cheng, D.; Tu, Y.; Ma, Z.; Niu, Z.; and Zhang, L. 2019a. Risk assessment for networked-guarantee loans using high-order graph attention representation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 5822–5828. AAAI Press.

Cheng, D.; Zhang, Y.; Yang, F.; Tu, Y.; Niu, Z.; and Zhang, L. 2019b. A dynamic default prediction framework for networked-guarantee loans. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2547–2555. ACM.

Elbayad, M.; Besacier, L.; and Verbeek, J. 2018. Pervasive attention: 2d convolutional neural networks for sequence-to-sequence prediction. *arXiv preprint arXiv:1808.03867*.

Fiore, U.; De Santis, A.; Perla, F.; Zanetti, P.; and Palmieri, F. 2017. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*.

Fu, K.; Cheng, D.; Tu, Y.; and Zhang, L. 2016. Credit card fraud detection using convolutional neural networks. In *International Conference on Neural Information Processing*, 483–490. Springer.

Gómez, J. A.; Arévalo, J.; Paredes, R.; and Nin, J. 2018. End-to-end neural network architecture for fraud scoring in card payments. *Pattern Recognition Letters* 105:175–181.

Jiang, C.; Song, J.; Liu, G.; Zheng, L.; and Luan, W. 2018. Credit card fraud detection: A novel approach using aggregation strategy and feedback mechanism. *IEEE Internet of Things Journal* 5(5):3637–3647.

Jurgovsky, J.; Granitzer, M.; Ziegler, K.; Calabretto, S.; Portier, P.-E.; He-Guelton, L.; and Caelen, O. 2018. Sequence classification for credit-card fraud detection. *Expert Systems with Applications* 100:234–245.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, 3146–3154.

Lepoivre, M. R.; Avanzini, C. O.; Bignon, G.; Legendre, L.; and Piwele, A. K. 2016. Credit card fraud detection with unsupervised algorithms. *Journal of Advances in Information Technology* 7(1).

Maes, S.; Tuyls, K.; Vanschoenwinkel, B.; and Manderick, B. 2002. Credit card fraud detection using bayesian and neural networks. In *Proceedings of the 1st international naiso congress on neuro fuzzy technologies*, 261–270.

McMahan, H. 2011. Follow-the-regular ized-leader and mil-ror descent: Equivalence theorems and 11 regularization. *Journal of Machine Learning Research Proceedings Trade* 15:525–533.

Patidar, R.; Sharma, L.; et al. 2011. Credit card fraud detection using neural network. *International Journal of Soft Computing and Engineering (IJSCE)* 1(32-38).

Randhawa, K.; Loo, C. K.; Seera, M.; Lim, C. P.; and Nandi, A. K. 2018. Credit card fraud detection using adaboost and majority voting. *IEEE access* 6:14277–14284.

Seeja, K., and Zareapoor, M. 2014. Fraudminer: A novel credit card fraud detection model based on frequent itemset mining. *The Scientific World Journal* 2014.

Shen, T.; Zhou, T.; Long, G.; Jiang, J.; Pan, S.; and Zhang, C. 2018. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Tang, J.; Deng, C.; and Huang, G.-B. 2015. Extreme learning machine for multilayer perceptron. *IEEE transactions on neural networks and learning systems* 27(4):809–821.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, D.; Chen, B.; and Chen, J. 2019. Credit card fraud detection strategies with consumer incentives. *Omega* 88:179–195.

Xu, D.; Ouyang, W.; Alameda-Pineda, X.; Ricci, E.; Wang, X.; and Sebe, N. 2017. Learning deep structured multi-scale features using attention-gated crfs for contour prediction. In *Advances in Neural Information Processing Systems*, 3961–3970.

Xu, D.; Wang, W.; Tang, H.; Liu, H.; Sebe, N.; and Ricci, E. 2018. Structured attention guided convolutional neural fields for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3917–3925.

Zaki, M. J.; Meira Jr, W.; and Meira, W. 2014. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.