黎金宁 lijinning@sjtu.edu.cn

KDD2017

论文鉴赏

End-to-end Learning for Short Text Expansion*

Jian Tang[1], Yue Wang[2], Kai Zheng[3], Qiaozhu Mei[1,2]
[1]School of Information, University of Michigan
[2]Department of EECS, University of Michigan
[3]Department of Informatics, University of California, Irvine
jiant,raywang,qmei@umich.edu,kai.zheng@uci.edu

Jian Tang (唐建)

Assistant Professor,

HEC Montreal & Montreal Institute for Learning Algorithms (MILA)

Email: tangjianpku at gmail.com, Weibo@chuckpku, Twitter@tangjianpku, CV

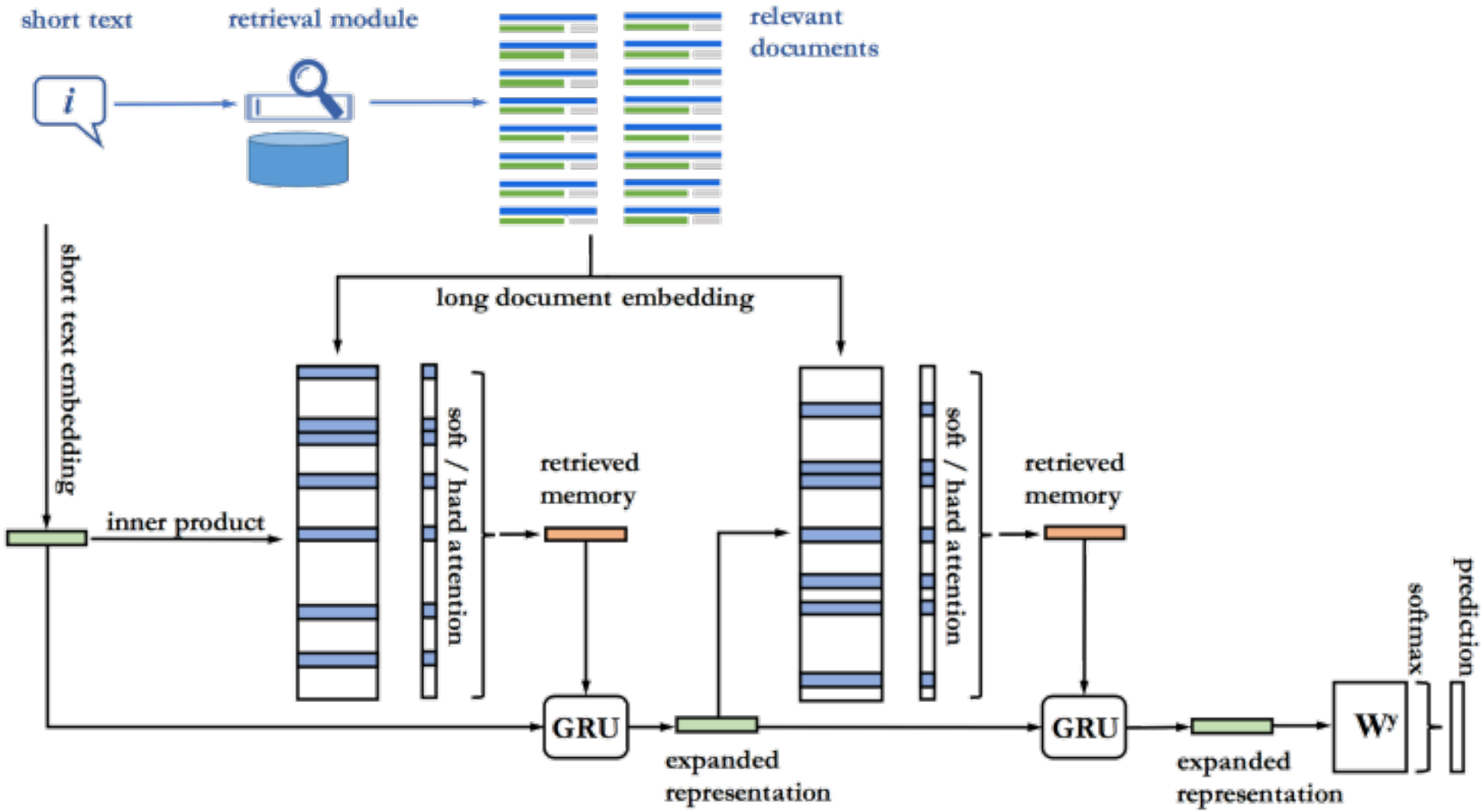
- Meng Qu, **Jian Tang**, Jingbo Shang, Xiang Ren, Ming Zhang, Jiawei Han. An Attention-based Collaboration Framework for Multi-View Network Representation Learning, in Proc. of 2017 ACM Int. Conf. on Information and Knowledge Management (CIKM'17), Singapore, Nov. 2017
- **Jian Tang,** Cheng Li and Qiaozhu Mei. Learning representations of large-scale networks. KDD'17 Tutorial. slides.
- **Jian Tang,** Yue Wang, Kai Zheng and Qiaozhu Mei. End-to-end learning for short text expansion. To appear in KDD'17.
- Xuanzhe Liu*, Wei Ai*, Huoran Li, **Jian Tang**, Gang Huang, and Qiaozhu Mei, "Derive User Preferences of Mobile Apps from their Management Activities," in ACM Transactions on Information Systems (**TOIS**) , in press, 2017.

End-to-end Learning for Short Text Expansion.

- 有效地理解**短文本**是许多真实世界应用的关键任务，例如搜索引擎，社交媒体服务和推荐系统。
- 任务特别具有挑战性，因为短文本包含非常**稀疏**的信息，对于机器学习算法来说，很难提高性能。
- 分析简短文本的一种常见做法是首先用**外部信息**进行扩展，外部信息通常是从大量较长的文本中收集的。
- 我们提出了一个end-to-end的解决方案，可以自动学习如何扩展短文本以**优化特定的学习任务**。

比如我们的微博、推特数据就是短文本
以后我们用自己模型之前可以先跑一下这个模型，假装做了个超强的优化

retrieval module 检索模块



Definition 3.1. **(Problem Definition.)** *Given a collection of long documents C, we aim to learn a function f that expands a short text q into a richer representation q', i.e., $q' = f(q, C)$. Based on the richer representation q', we can accurately classify the short text into one of the predefined categories $\mathcal{Y}$.*

3.2 Short Text Representation Module

We represent each short text $q = w_1, \ldots, w_n$ as a d-dimensional vector $\vec{q}$ in a continuous space. Each word in the vocabulary is represented as a d-dimensional vector, and then the entire short text is represented as the average vector of words in the short text, i.e.,

$$\vec{q} = \frac{\sum_{i=1}^{n} \mathbf{A}_{w_n}}{n} \tag{1}$$

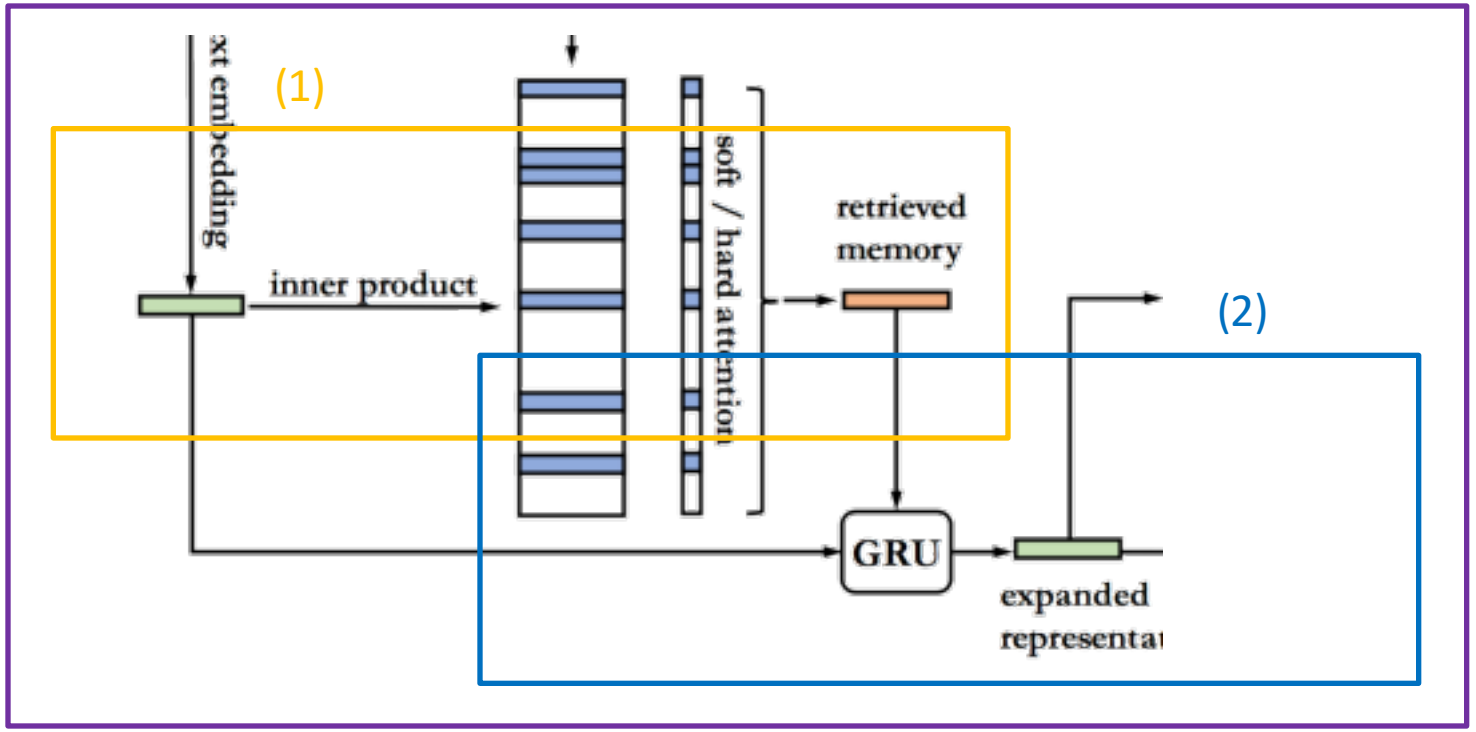
Similarly, Long Text: $\vec{d}_i = \frac{\sum_{i=1}^{n} \mathbf{B}_{w_n}}{n}$,

The expansion module is the core part of ExpaNet. The goal is to expand the continuous representation of input short text $\vec{q}$ by incorporating the information in the memory $M = \{\vec{d}_i\}_{i=1}^{K}$, where K is the number of documents in the memory. The expansion process can be divided into two different components: (1) given the query representation $\vec{q}$, what information should we read from the memory? (2) how to integrate the information from the memory with the original query representation $\vec{q}$?

$$\vec{q} = \frac{\sum_{i=1}^{n} \mathbf{A}_{w_n}}{n}$$

Integrate 整合

Expansion module



(1) Memory Reading

Soft attention: Soft attention is widely used in existing memory networks. We use the same mechanism as [29]. The relevance between the query $\vec{q}$ and each document $\vec{d}_i$ is calculated as their inner product, and a softmax function is used to define the attention probability over each document i in the memory, i.e.,

$$a_i = \text{Softmax}\left(\vec{q}^\top \vec{d}_i\right) , \tag{3}$$

where $\text{Softmax}(z_i) = e^{z_i}/\sum_j e^{z_j}$. In this way, the a_i's define a probability distribution over the long documents in memory M, and the information read from the document is defined as:

$$\vec{o} = \sum_{i=1}^{K} a_i \vec{d}_i . \tag{4}$$

$$\vec{d}_i = \frac{\sum_{i=1}^{n} \mathrm{B}_{w_n}}{n},$$

Hard attention: Instead of looking at each document with some probability, a human searcher often picks a document that seems relevant and focus on it. Therefore, we also investigate using hard attention here [22], which is achieved by randomly sampling a document from the probability distribution $\vec{a} = (a_1, \ldots, a_K)$ defined in the soft attention, i.e.,

$$\vec{p} \sim \text{multinomial}(\vec{a}) , \tag{5}$$

where $\vec{p}$ is a one-hot vector. Then the information read from the memory is defined as:

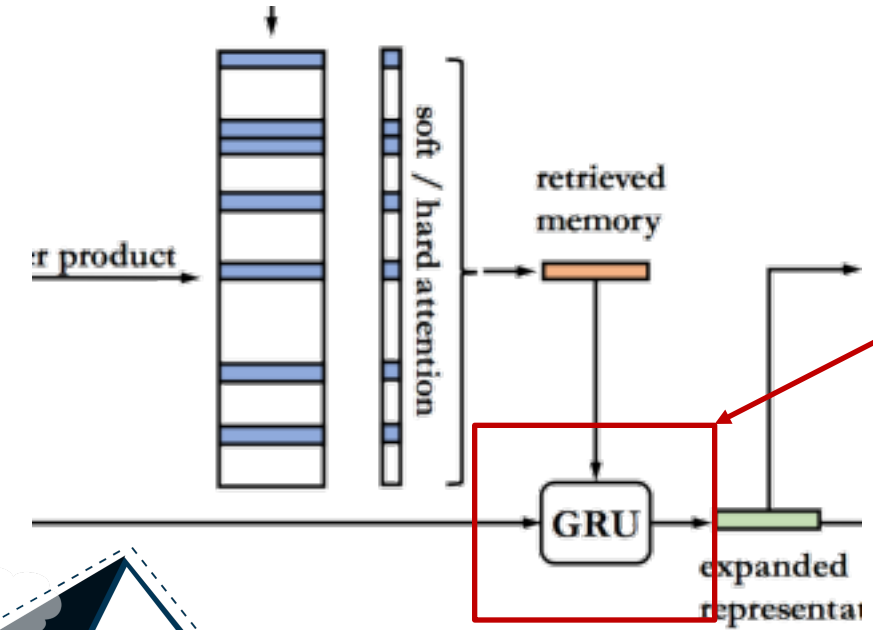
$$\vec{o} = \sum_{i=1}^{K} p_i \vec{d}_i . \tag{6}$$

REINFORCE [32] algorithm), and complicated variance reduction methods [33] must be used. In this paper, we use a recent technique, the Gumbel-Softmax [11], for backpropagating through samples,

$$p_i = \text{Softmax}\left(\frac{\vec{q}^\top \vec{d}_i + g_i}{\tau}\right) ,$$

(2) Short Text Expansion

Here, we use a principled method to integrate the two sources of information. We use a gating mechanism, the Gated Recurrent Unit (GRU) [3], to combine the information, which is able to automatically determine the weight of the two sources of information. Specifically, the two sources of information are integrated as follows:



$$\vec{z} = \sigma\left(\mathbf{W}^{(z)}\vec{q} + \mathbf{U}^{(z)}\vec{o}\right) ; \quad (8)$$

$$\vec{r} = \sigma\left(\mathbf{W}^{(r)}\vec{q} + \mathbf{U}^{(r)}\vec{o}\right) ; \quad (9)$$

$$\vec{o}' = \tanh\left(\mathbf{W}\vec{q} + \vec{r} \circ \mathbf{U}\vec{o}\right) ; \quad (10)$$

$$\vec{q}' = (\mathbf{1} - \vec{z}) \circ \vec{q} + \vec{z} \circ \vec{o}' , \quad (11)$$

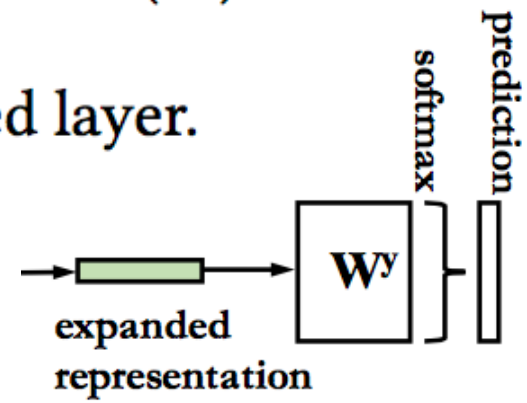
where $\circ$ denotes elementwise multiplication and $\sigma(x) = 1/(1 + \exp(x))$, $\tanh(x) = (1 - \exp(-2x))/(1 + \exp(-2x))$ are both element-wise operations. $\vec{o}'$ is the new information from the memory, which is determined by both sources of information $\vec{q}$ and $\vec{o}$. $\vec{z}$ is the weighting vector between the original information $\vec{q}$ and the new information $\vec{o}'$. The output $\vec{q}'$ is the expanded representation of the input short text q.

Classification Module

As in classical methods for query expansion, we keep the original short text representation $\vec{q}$ and represent the final short text representation as a concatenation of $\vec{q}$ and the expanded representation $\vec{q}'$, i.e., $\vec{q}_{\text{final}} = [\vec{q}, \vec{q}']$, which is then used to predict the category of the short text. A fully connected layer is first applied to the short text representation and then followed by a Softmax transformation, yielding a distribution over the categories, i.e.,

$$p(y|\vec{q}_{\text{final}}) = \text{Softmax}(\mathbf{W}^y \vec{q}_{\text{final}}), \tag{12}$$

where $\mathbf{W}^y \in \mathbb{R}^{|\mathcal{Y}| \times 2d}$ is the parameter for fully connected layer.



Experiments

WIKIPEDIA. Titles of Wikipedia articles represent short texts in the general domain. The length of Wikipedia titles is on average 3.12 words, similar to that of search queries [1]. We take a recent snapshot of English Wikipedia[1] to construct this data set. We use

DBLP. Titles of computer science literature represent short texts in formal communication. We choose 6 diverse research fields for

TWITTER. The 140-character microblog data represent informal short texts widely used in social media. We use a large corpus of tweets for positive/negative sentiment classification.[4] We randomly

Tripoles: A New Class of Relationships in Time Series Data

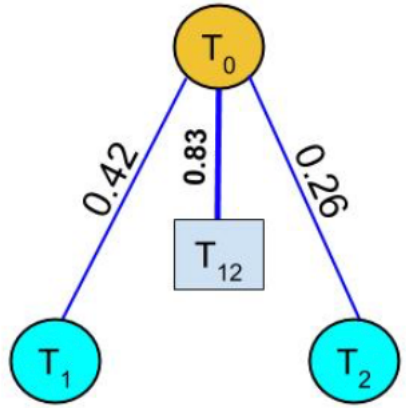
Saurabh Agrawal
agraw066@umn.edu
University of Minnesota

Gowtham Atluri
atlurigm@ucmail.uc.edu
University of Cincinnati

Anuj Karpatne, William Haltom, Stefan Liess, Snigdhansu Chatterjee, Vipin Kumar
karpa009,halto004,liess,chatt019,
kumar001@umn.edu
University of Minnesota

- **时间序列数据**中的挖掘关系对神经科学，气候科学和交通等几个学科有极大的应用。
- 传统的挖掘关系的方法着重于发现数据中的成对关系。在这项工作中，我们定义了一种新颖的关系，它涉及**三个相互作用的时间序列**，我们称之为三极。
- 我们表明，三极捕获数据中的有趣关系，这些数据不可能使用传统研究的成对关系捕获。
- 我们展示了来自各个领域（包括气候科学和神经科学）的多个真实世界数据集中三极的效用。 具体而言，我们的方法能够发现三态，这些三态在统计学意义上是重要的，可以在多个独立的数据集上重现，并导致新的领域见解。

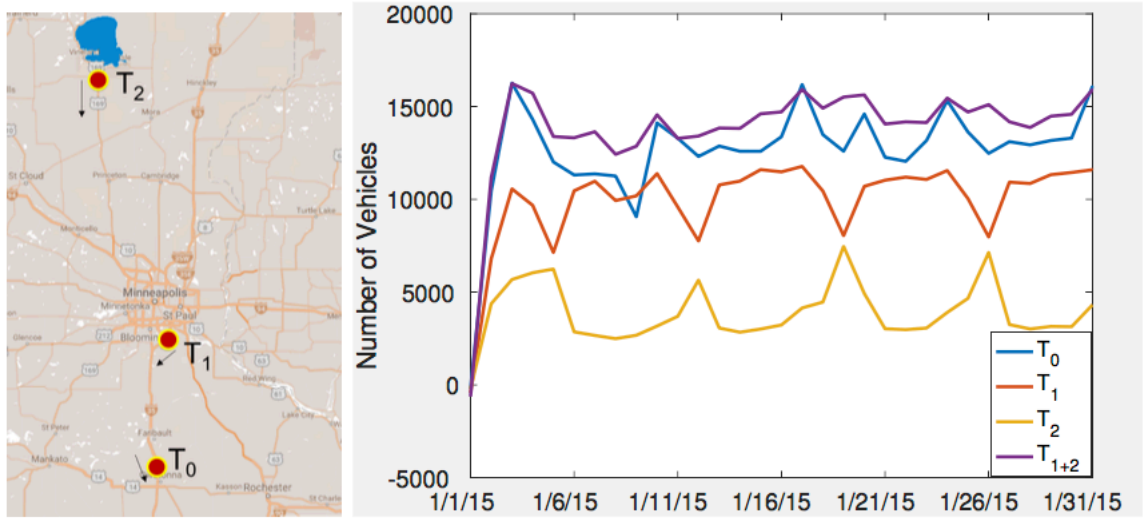
*Definition 2.1 (**Tripole**).* A tripole, $\Delta \equiv (T_0 : T_1, T_2)$, is a collection of three time-series, T_0, T_1, and T_2, where T_0 is referred to as the *root* of the tripole, while T_1 and T_2 are referred to as the *leaves* of the tripole.



*Definition 2.2 (**Strength**).* The strength of a tripole $\Delta \equiv (T_0 : T_1, T_2)$, denoted by σ_Δ, measures the correlation between the time-series at the root T_0, with the sum time-series of the leaves, $T_{1+2} = (T_1 + T_2)$, as follows

$$\sigma_\Delta = corr(T_0, T_{1+2}) \tag{1}$$

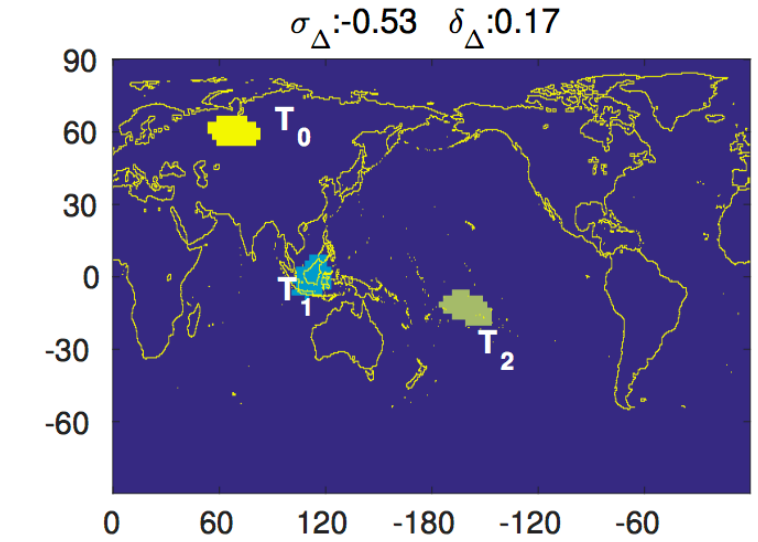
.....



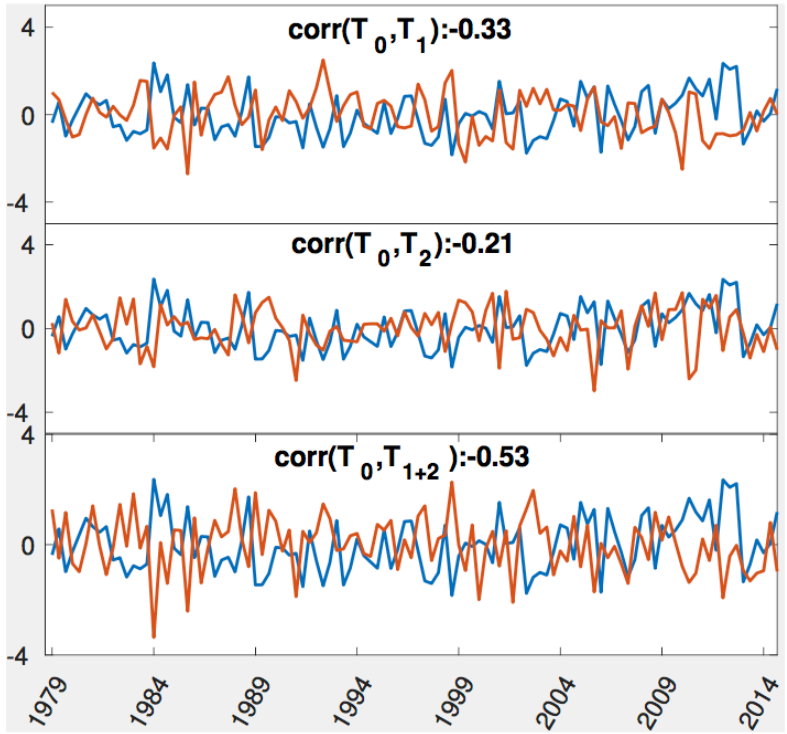
(a) Spatial Map **(b) Time-series**

Figure 2: Example of a tripole in transportation data showing different modes of traffic on a highway near Minneapolis.

As a real-world example of the tripole pattern, consider the traffic data set from the Minnesota Department of Transportation [1] where the volume of traffic crossing a road section is represented as a daily time series. Using this data, one may be interested in finding non-trivial relationships among the traffic activity at three road sections. Figure 2 shows an example of such a tripole where the January 2015. Time series T_1 and T_2 indicate the traffic volume at the roads that contribute to the traffic at the main highway where T_0 is being observed. This is evident from the high correlation (0.83) of their sum ($T_{1+2} = T_1 + T_2$, shown as magenta curve in Figure 2(b))



(a) Three regions that of a tripole



(b) Time series of regions

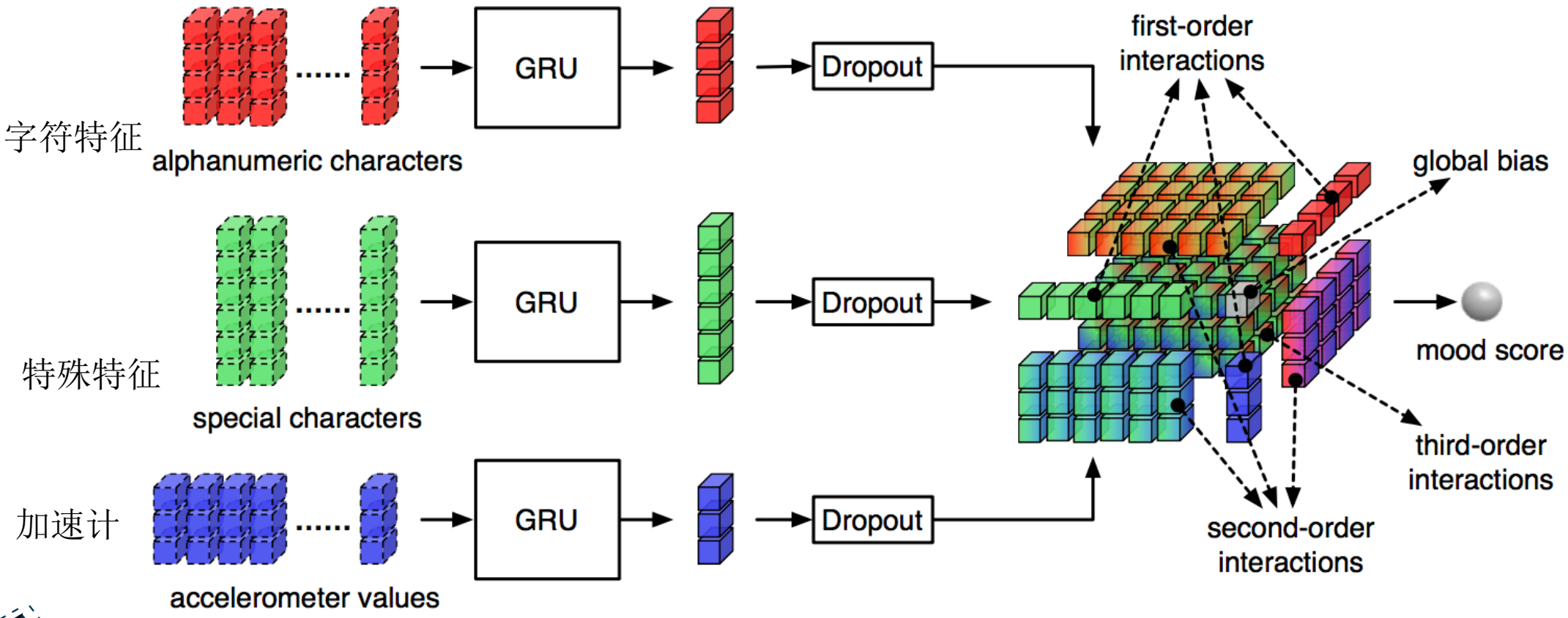
As another example, consider the tripole in Sea level Pressure (SLP) data between SLP time series of three regions that are shown in Figure 4(a) using differently colored patches on the world map.

We can see from Figure 4(b) that T_0 shows a correlation of -0.53 with T_{1+2}, which is significantly stronger than the correlation it shows with either of the two leaves, T_1 and T_2 (-0.33 and -0.21 respectively). This tripole indeed represents a physically relevant but previously unknown phenomenon (atmospheric waves that flow from Siberian Region (root) towards the two leaves in the Pacific Ocean) that was discovered using the approach described in this paper[11].

DeepMood: Modeling Mobile Phone Typing Dynamics for Mood Detection

Bokai Cao[1], Lei Zheng[1], Chenwei Zhang[1], Philip S. Yu[1,2], Andrea Piscitello[1], John Zulueta[3], Olu Ajilore[3], Kelly Ryan[4], and Alex D. Leow[1,3,5]
[1]Department of Computer Science, University of Illinois at Chicago
[2]Institute for Data Science, Tsinghua University
[3]Department of Psychiatry, University of Illinois at Chicago
[4]Department of Psychiatry, University of Michigan
[5]Department of Bioengineering, University of Illinois at Chicago
caobokai,lzheng21,czhang99,psyu,apisci2@uic.edu
jzulueta,oajilore,aleow@psych.uic.edu
karyan@med.umich.edu

- 电子通信形式的使用越来越多，为研究精神卫生提供了新的机会，包括能够无情地调查精神疾病的表现以及患者日常生活的情况。进行了一项前瞻性研究，探讨双相障碍和手机使用之间的可能联系。
- 在这项研究中，为志愿者提供了一部手机作为他们的主要手机。这款手机加载了一个自定义键盘，用于收集**由按键输入时间和重力加速计移动**组成的元数据。
- 我们提出了一种基于后期融合的**端到端深度架构**，名为DeepMood，用于对用于预测情绪分数的多视图元数据进行建模。
- 实验结果表明，基于会话级别的手机打字动态（通常小于1分钟），可以实现90.31％的抑郁评分预测准确度。**它展示了使用手机元数据来推断情绪障碍和严重程度的可行性**。


字符特征
alphanumeric characters
GRU
Dropout
特殊特征
special characters
GRU
Dropout
加速计
accelerometer values
GRU
Dropout
first-order interactions
global bias
mood score
third-order interactions
second-order interactions

Dataset

The data used in this work were collected from the BiAffect[3] study which is the winner of the Mood Challenge for ResearchKit[4]. During a preliminary data collection phase, for a period of 8 weeks, 40 individuals were provided a Galaxy Note 4 mobile phone which they were instructed to use as their primary phone during the study. This phone was loaded with a custom keyboard that replaced the standard Android OS keyboard. The keyboard collected metadata consisting of keypress entry time and accelerometer movement and uploaded them to the study server. In order to protect participants' privacy, individual character data with the exceptions of the backspace key and space bar were not collected.

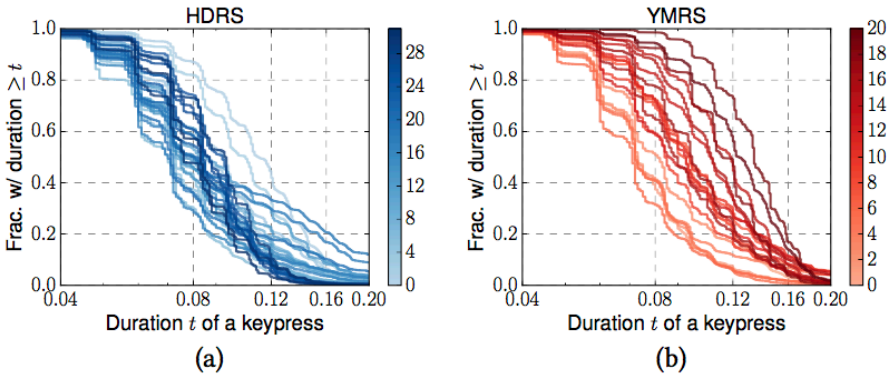


Figure 3: CCDFs of duration of a keypress.

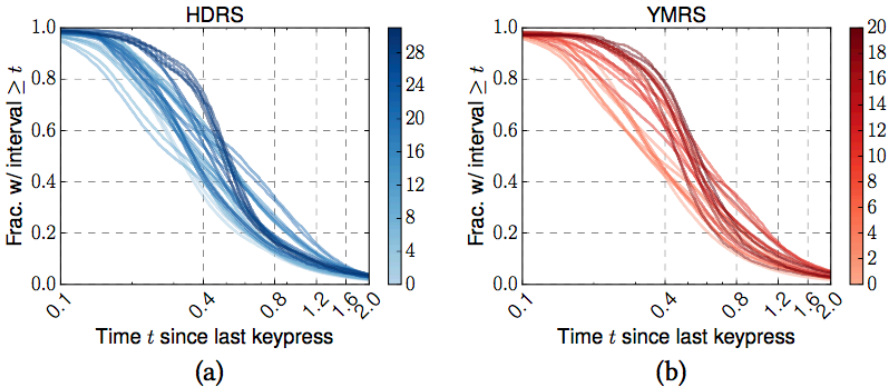
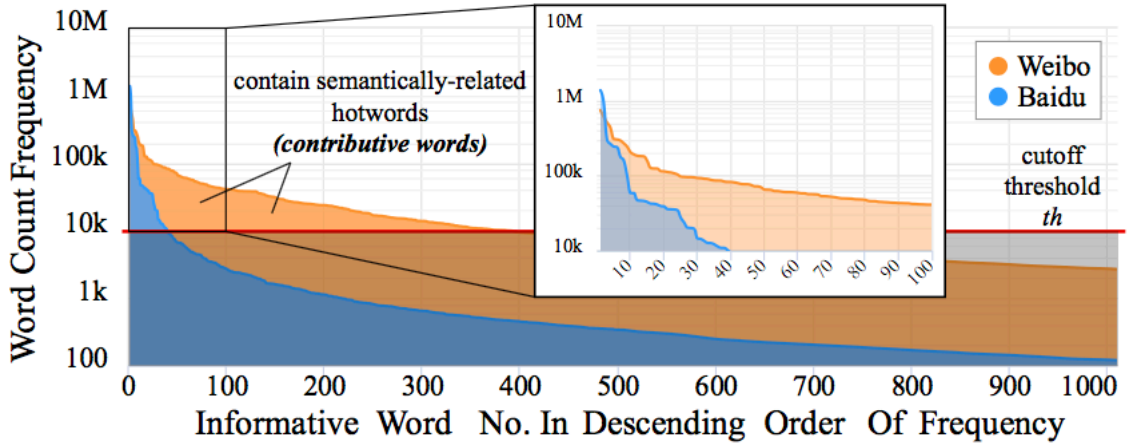


Figure 4: CCDFs of time since last keypress.

are dealing with a heavy-tailed distribution: (1) most keypresses are very fast with median 85ms, (2) but a non-negligible number have longer duration with 5% using more than 155ms. Interestingly, samples with mild depression tend to have shorter duration than normal ones, while those with severe depression stand in the middle. Samples in manic symptoms seem to hold a key longer than normal ones.

Alphanumeric Characters

Dataset



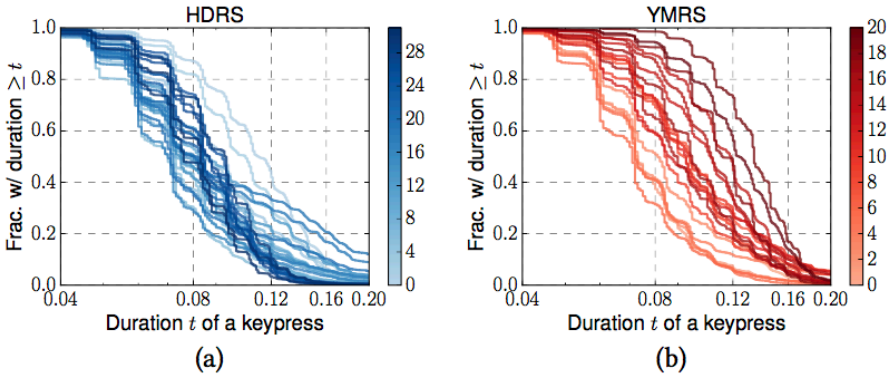


Figure 3: CCDFs of duration of a keypress.

在机率论中，重尾分布（英语：Heavy-tailed distribution）是一种机率分布的模型，它的尾部比指数分布还要厚。在许多状况中，通常右边尾部的分布会比较受到重视，但左边尾部比较厚，或是两边尾部都很厚的状况，也会被认为是一种重尾分布。

重尾分布之中，又有两个子类型，分别称为长尾分布（long-tailed distributions）以及次指数分布（subexponential distributions）。

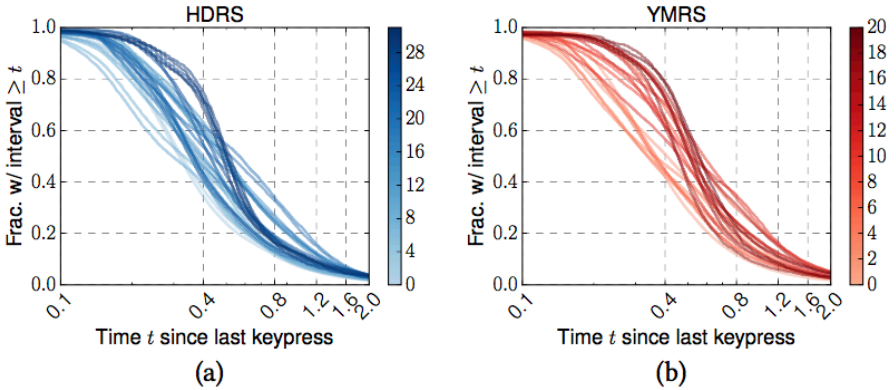


Figure 4: CCDFs of time since last keypress.

Alphanumeric Characters

1.422s. We can observe that the values of time since last keypress from the normal group (with light blue/red) approximate a uniform distribution on the log scale in the range from 0.1s to 2.0s. On the contrary, this metric from samples with mood disturbance (with dark blue/red) shows a more skewed distribution with a few values on the two tails and majority centered between 0.4s and 0.8s. In other words, healthy people show a good range of reactivity that gets lost in mood disturbance where the range is more restricted.

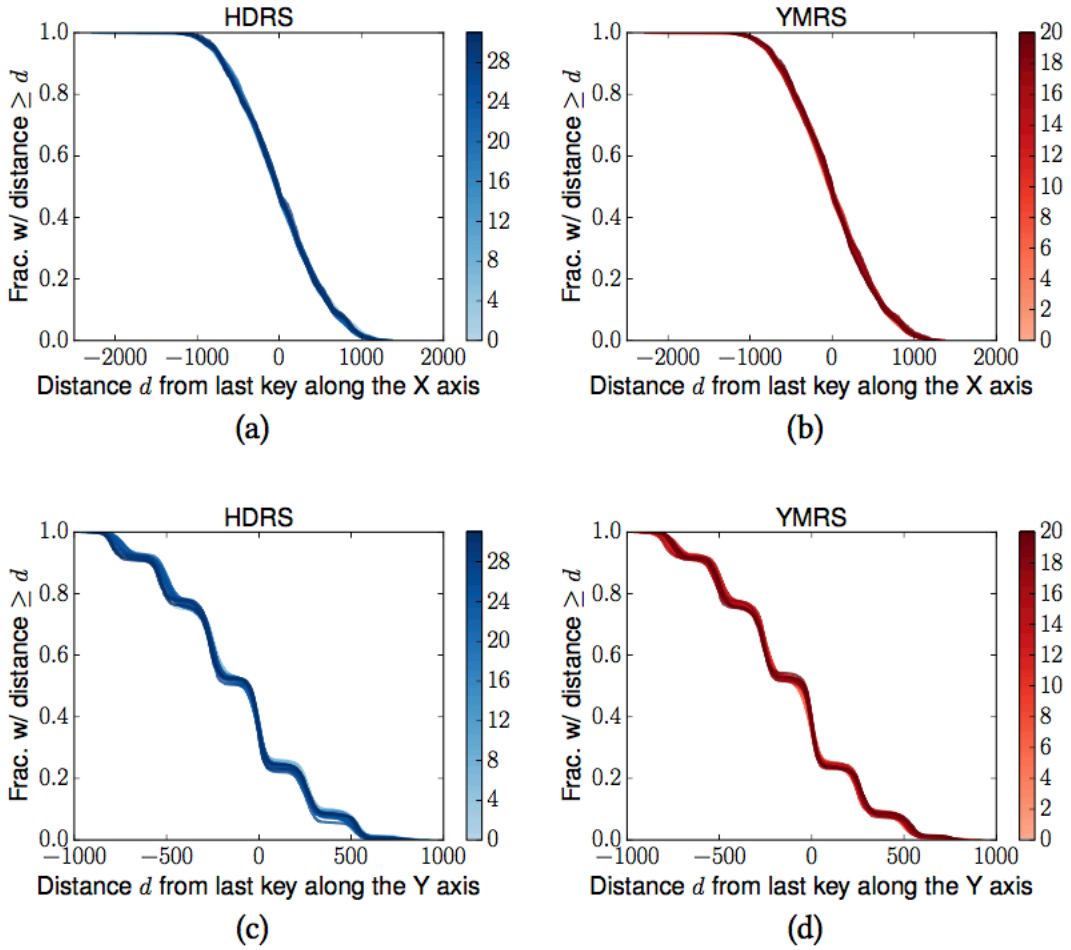


Figure 5: CCDFs of distance from last key along two axises. Note that lines are almost identical.

2.2 Special Characters

In this view, we use one-hot-encoding for typing behaviors other than alphanumeric characters, including *auto-correct*, *backspace*, *space*, *suggestion*, *switching-keyboard* and *other*. They are usually sparser than alphanumeric characters. Figure 6 shows the scatter

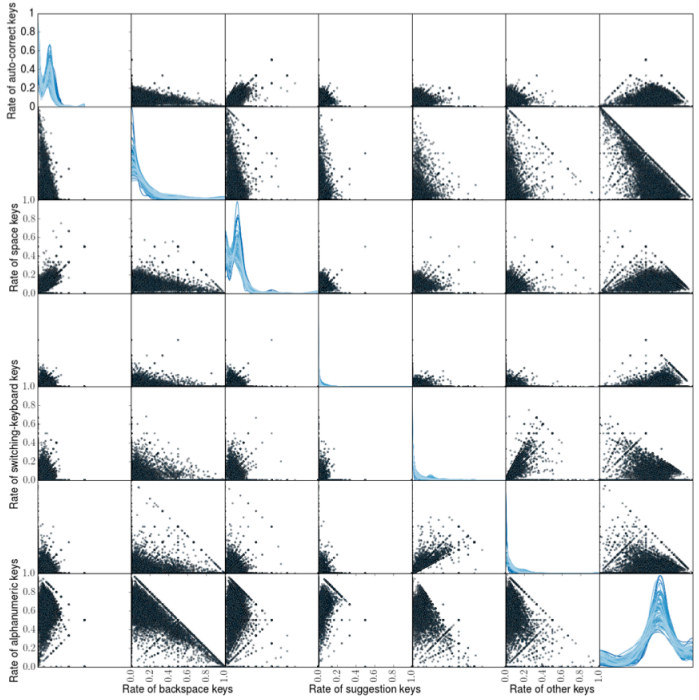
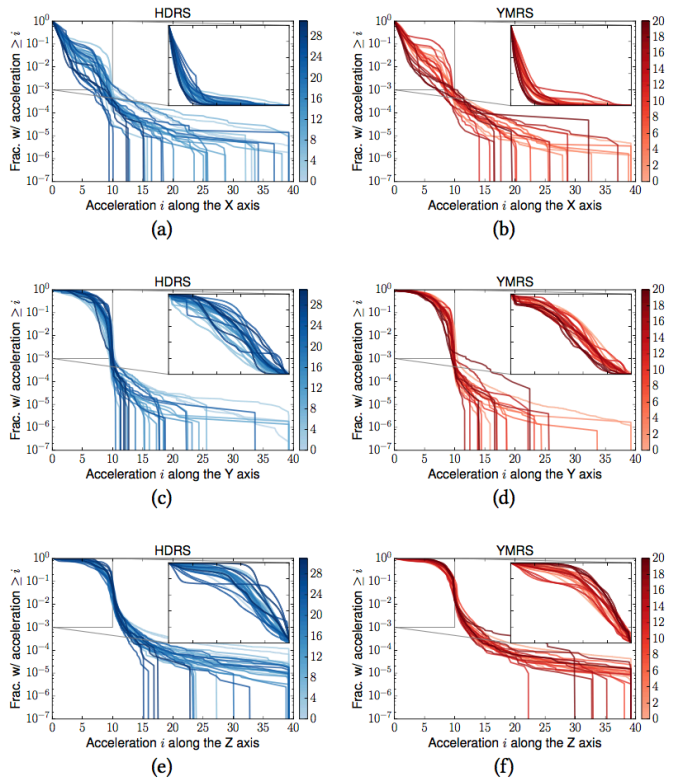


Figure 6: Scatter plot between rates of different keys.

2.3 Accelerometer Values

Accelerometer values are recorded every 60ms in the background during an active session regardless of a person's typing speed, thereby making them much denser than alphanumeric characters.

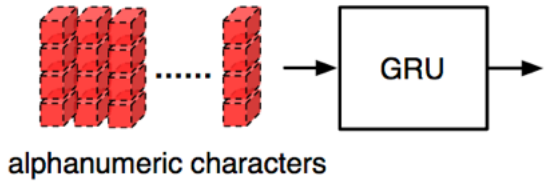


Architecture

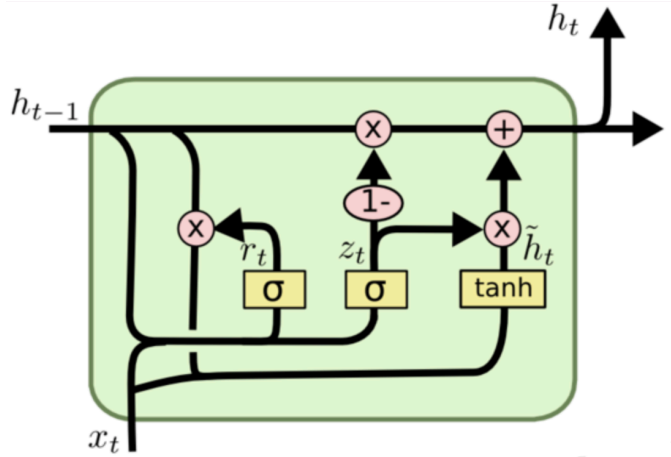
To make the learning procedure more effective over long sequences, the GRU [12] is proposed as a variation of the LSTM unit [24]. The GRU has been attracting great attentions since it overcomes the vanishing gradient problem in traditional RNNs and is more efficient than the LSTM in some tasks [14]. The GRU is designed to learn from previous timestamps with long time lags of unknown size between important timestamps via memory units that enable the network to learn to both update and forget hidden states based on new inputs.

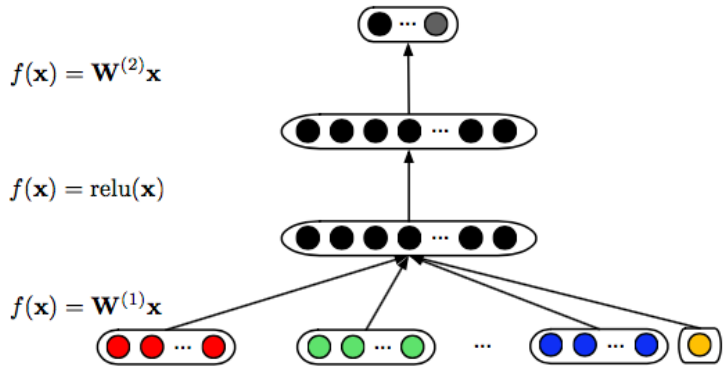
A typical GRU is formulated as:

$$\begin{aligned} \mathbf{r}_k &= \text{sigmoid}(\mathbf{W}_r \mathbf{x}_k + \mathbf{U}_r \mathbf{h}_{k-1}) \\ \mathbf{z}_k &= \text{sigmoid}(\mathbf{W}_z \mathbf{x}_k + \mathbf{U}_z \mathbf{h}_{k-1}) \\ \tilde{\mathbf{h}}_k &= \tanh(\mathbf{W}\mathbf{x}_k + \mathbf{U}(\mathbf{r}_k \odot \mathbf{h}_{k-1})) \\ \mathbf{h}_k &= \mathbf{z}_k \odot \mathbf{h}_{k-1} + (1 - \mathbf{z}_k) \odot \tilde{\mathbf{h}}_k \end{aligned} \tag{3}$$

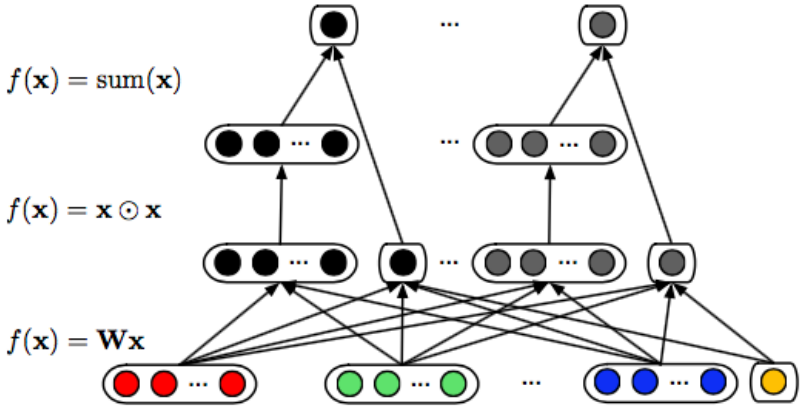


late fusion. These include not only the straightforward approach based on adding a fully connected layer to concatenate the features from different views, but also novel approaches to capture interactions among the features across multiple views by exploring the concept of Factorization Machines [39] to capture the second-order interactions as well as the concept of Multi-view Machines [9] to capture higher order interactions as shown in Figure 8.

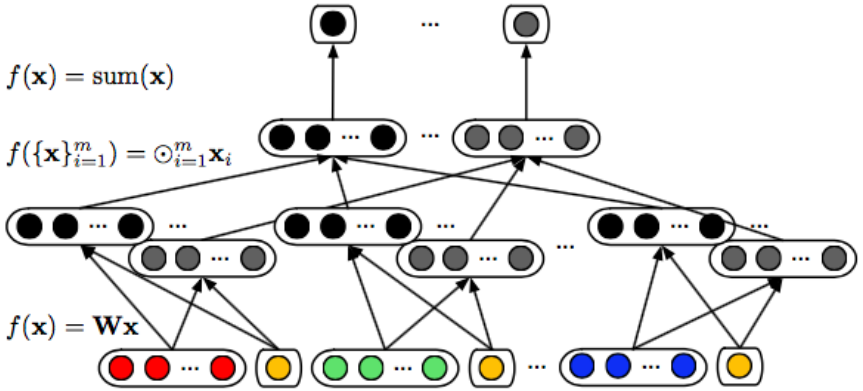




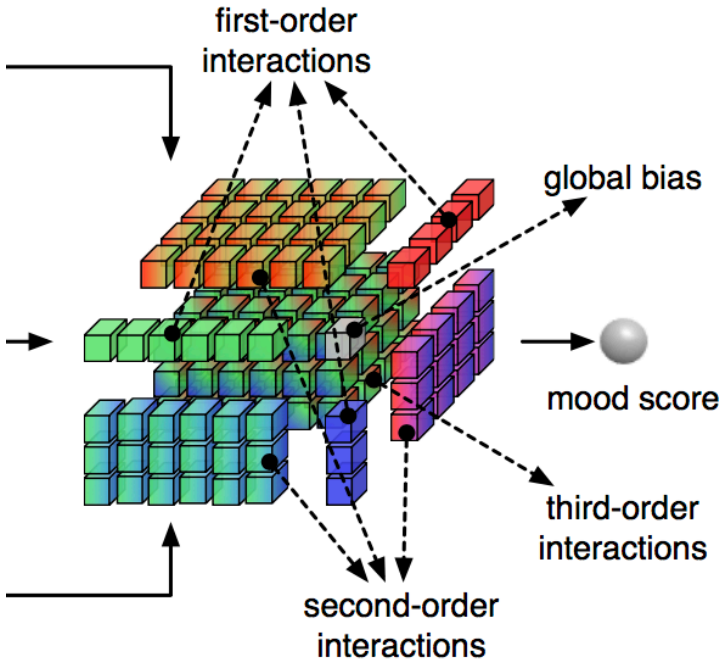
(a) Fully connected layer.



(b) Factorization Machine layer.



(c) Multi-view Machine layer.



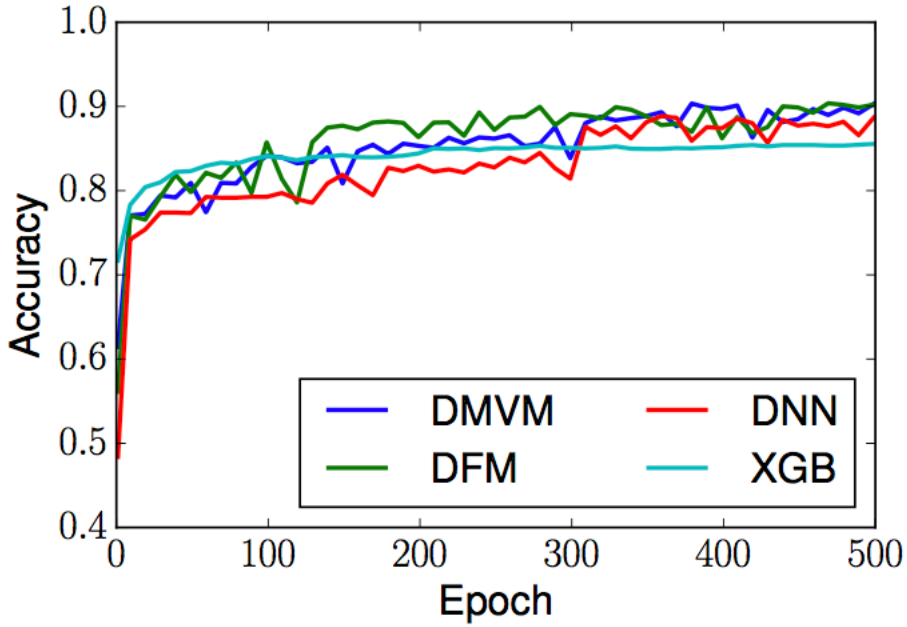
Prediction Performance

The compared methods are summarized as follows:

- **DMVM**: The proposed DeepMood architecture with a Multi-view Machine layer for data fusion.
- **DFM**: The proposed DeepMood architecture with a Factorization Machine layer for data fusion.
- **DNN**: The proposed DeepMood architecture with a conventional fully connected layer for data fusion.
- **XGB**: The implementation of a tree boosting system from `XGBoost`[6] [10] is used. We concatenate the sequence data with the maximum length 100 (padding 0 for short ones) of each feature as the input.
- **SVM** and **LR**: These are two linear models. With the same input setting as XGB, the implementations of Linear Support Vector Classification/Regression and Logistic/Ridge Regression from `scikit-learn`[7] are used for Classification/Regression tasks.

Task	Classification		Regression
Metric	Accuracy	F-score	RMSE
DMVM	0.9031	0.9070	3.5664
DFM	0.9021	0.9029	3.6767
DNN	0.8868	0.8929	3.7874
XGB	0.8555	0.8562	3.9634
SVM	0.7323	0.7237	4.1257
LR	0.7293	0.7172	4.1822

Convergence Efficiency



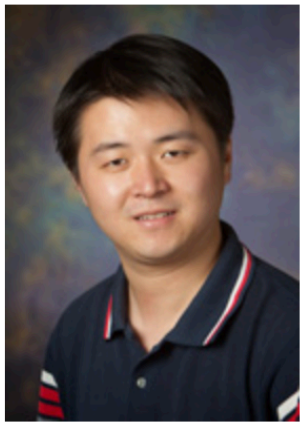
Stock Price Prediction via Discovering Multi-Frequency Trading Patterns

Liheng Zhang
University of Central Florida
4000 Central Florida Blvd.
Orlando, Florida 32816
lihengzhang1993@knights.ucf.edu

Charu Aggarwal
IBM T. J. Watson Research Center
New York 10598
charu@us.ibm.com

Guo-Jun Qi*
University of Central Florida
4000 Central Florida Blvd.
Orlando, Florida 32816
guojun.qi@ucf.edu

Dr. Guo-Jun Qi



Assistant Professor

Department of Computer Science

Email:
guojun.qi at ucf dot edu | guojunq at gmail dot com
Phone: (407) 823-2764
FAX: (407) 823-5835

Address:
University of Central Florida
Department of Computer Science
4328 Scorpius HEC 318
Orlando, FL 32816

Laboratory:

Machine **P**erception and **Le**arning (MAPLE) [url | github]

Jun Ye[§], Hao Hu[§], **Guo-Jun Qi***, Kien Hua. A Temporal Order Modeling Approach to Human Action Recognition from Multimodal Sensor Data, in **ACM Transactions on Multimedia Computing, Communications, and Applications** (TOMM), Volume 13 Issue 2, Article No. 14, May 2017. [**pdf**]

Kai Li[§], **Guo-Jun Qi***, Jun Ye, Kien Hua. Linear Subspace Ranking Hashing for Cross-modal Retrieval, in **IEEE Transactions on Pattern Analysis and Machine Intelligence** (**T-PAMI**), Volume 39, Issue 9, September 2016. [**pdf**] [**code**]

Guo-Jun Qi, Wei Liu, Charu Aggarwal, and Thomas Huang. Joint Intermodal and Intramodal Label Transfers for Extremely Rare or Unseen Classes, in **IEEE Transactions on Pattern Analysis and Machine Intelligence** (**T-PAMI**), Volume 39, Issue 7, July 2016. [**pdf**]

Jinhui Tang, Xiangbo Shu, **Guo-Jun Qi**, Zechao Li, Meng Wang, Shuicheng Yan, and Ramesh Jain. Tri-clustered Tensor Completion for Social-Aware Tag Refinement, in **IEEE Transactions on Pattern Analysis and Machine Intelligence** (**T-PAMI**), Vol. 39, No. 8, pp. 1662 - 1674, August 2017. [**pdf**]

- 股票价格基于反映不同频率交易模式的短期和/或长期商业和交易活动形成。
- 然而，这些模式往往**难以捉摸**，因为它们受到现实世界中许多不确定的政治经济因素的影响，例如企业绩效，政府政策，甚至是跨市场的突发新闻。
- 此外，**股票价格的时间序列是非平稳的和非线性的**，使得对未来价格趋势的预测非常具有挑战性。
- 为了解决这些问题，我们提出了一种新颖的状态频率记忆（SFM）**循环网络**，以捕捉过去市场数据中的多频交易模式，以便随时间进行长期和短期预测。

data vector

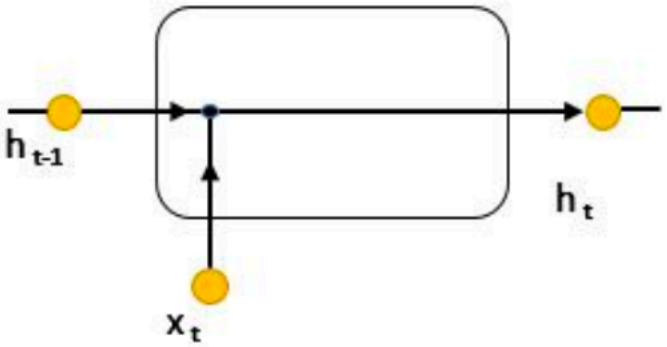
data matrix

element-wise multiplication

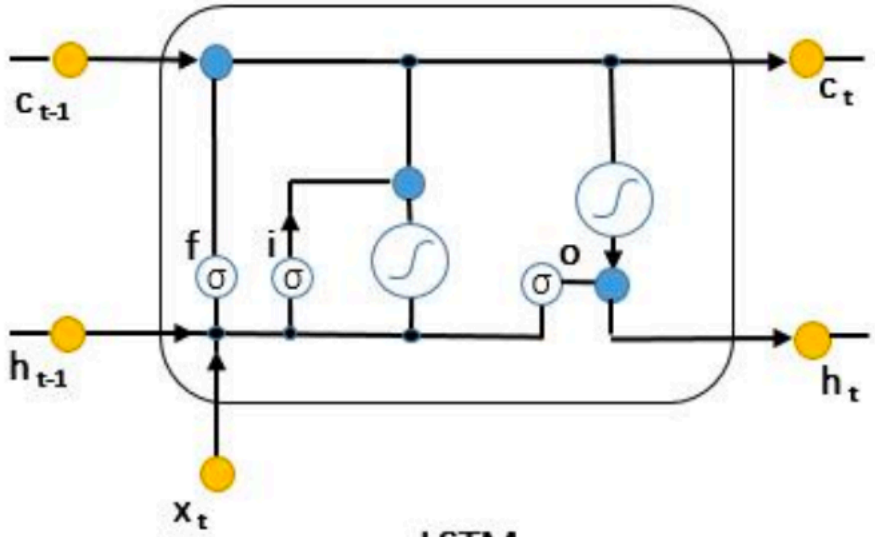
sigmoid

activation function

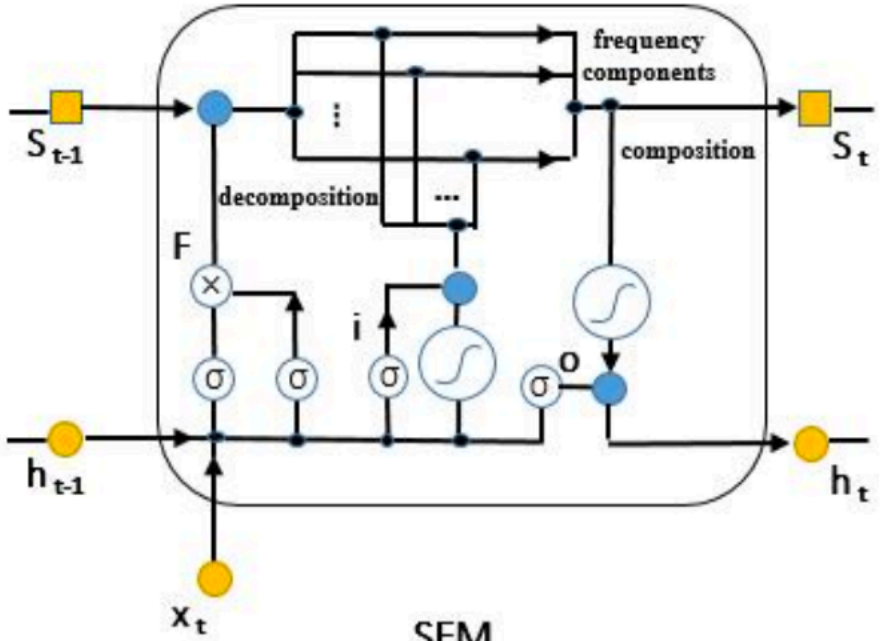
outer product



RNN



LSTM



SFM

Architecture

$$\boldsymbol{S_t} = \boldsymbol{F_t} \circ \boldsymbol{S_{t-1}} + (\boldsymbol{i_t} \circ \tilde{\boldsymbol{c}}_{\boldsymbol{t}}) \begin{bmatrix} e^{j\omega_1 t} \\ e^{j\omega_2 t} \\ \dots \\ e^{j\omega_K t} \end{bmatrix}^T \in \mathbb{C}^{D \times K} \qquad (7)$$

where $j = \sqrt{-1}$ and $\left[e^{j\omega_1 t}, e^{j\omega_2 t}, ..., e^{j\omega_K t}\right]$ are the Fourier basis of K frequency components of the state sequence.

LSTM

Formally, the LSTM can be formulated as follows. At each time t, $\boldsymbol{x_t}$ is an input vector (e.g., stock prices), $\boldsymbol{c_t}$ denotes the memory state vector, and $\boldsymbol{h_t}$ is the hidden state vector output from $\boldsymbol{c_t}$. Then we have:

$$\boldsymbol{i_t} = \text{sigmoid}(\boldsymbol{W_i x_t} + \boldsymbol{U_i h_{t-1}} + \boldsymbol{b_i}) \qquad (1)$$

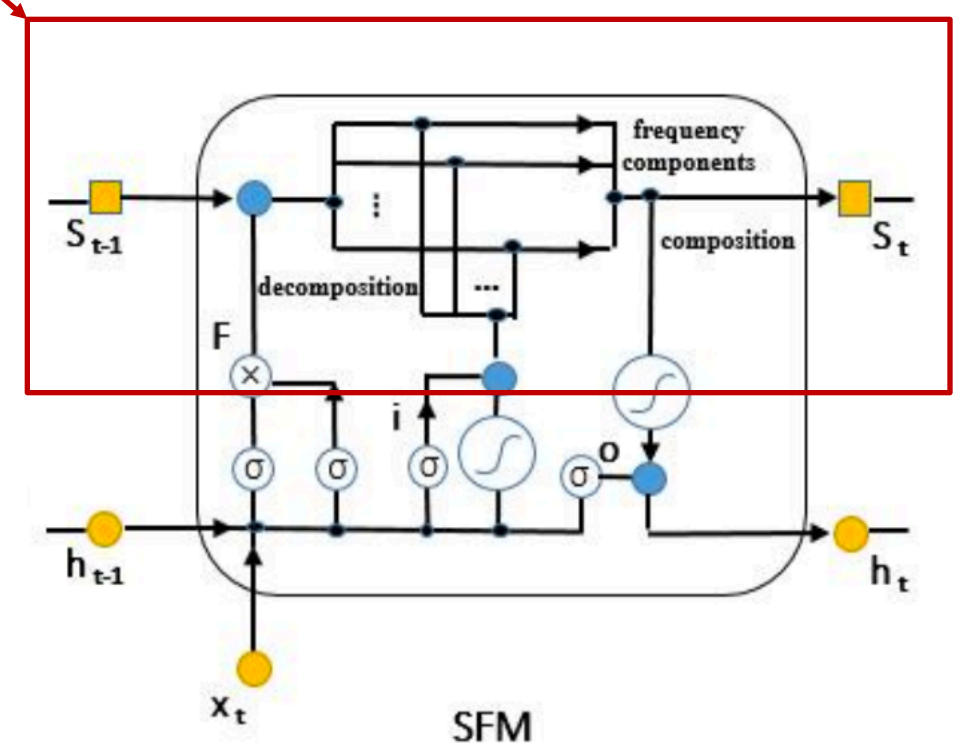
$$\boldsymbol{f_t} = \text{sigmoid}(\boldsymbol{W_f x_t} + \boldsymbol{U_f h_{t-1}} + \boldsymbol{b_f}) \qquad (2)$$

$$\tilde{\boldsymbol{c}}_{\boldsymbol{t}} = \tanh(\boldsymbol{W_c x_t} + \boldsymbol{U_c h_{t-1}} + \boldsymbol{b_c}) \qquad (3)$$

$$\boldsymbol{c_t} = \boldsymbol{i_t} \circ \tilde{\boldsymbol{c}}_{\boldsymbol{t}} + \boldsymbol{f_t} \circ \boldsymbol{c_{t-1}} \qquad (4)$$

$$\boldsymbol{o_t} = \text{sigmoid}(\boldsymbol{W_o x_t} + \boldsymbol{U_o h_{t-1}} + \boldsymbol{V_o c_t} + \boldsymbol{b_o}) \qquad (5)$$

$$\boldsymbol{h_t} = \boldsymbol{o_t} \circ \tanh(\boldsymbol{c_t}) \qquad (6)$$



The updating rule can be separated into the real and imaginary parts of the state-frequency matrix $\boldsymbol{S_t}$:

$$Re\boldsymbol{S_t} = \boldsymbol{F_t} \circ Re\boldsymbol{S_{t-1}} + (\boldsymbol{i_t} \circ \tilde{\boldsymbol{c}}_{\boldsymbol{t}}) \left[cos\omega_1 t, ...cos\omega_K t\right] \quad (8)$$

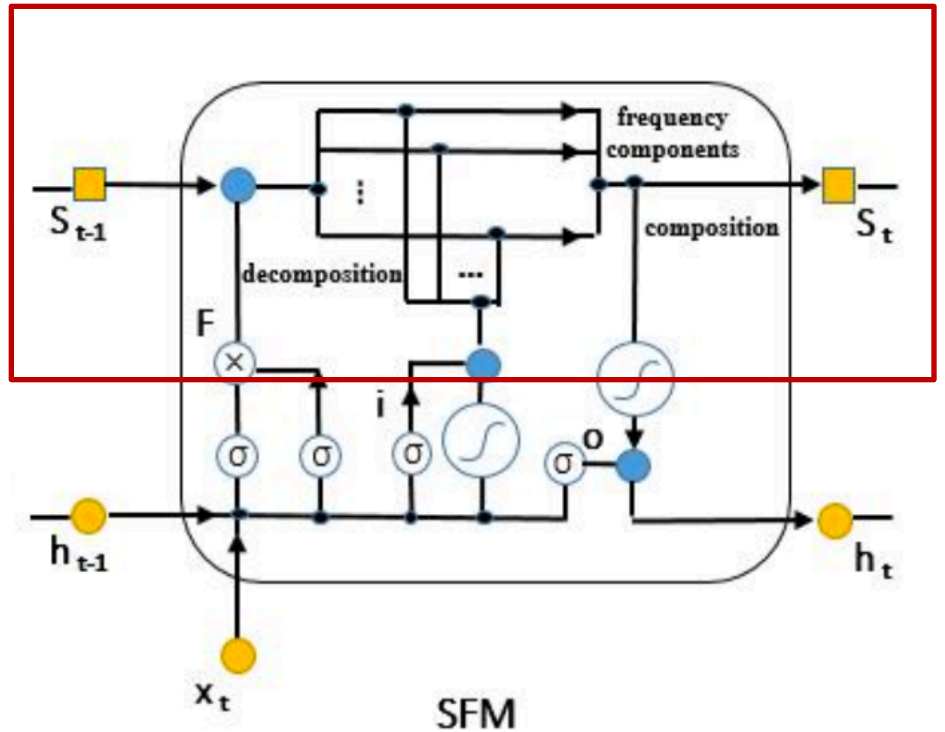
$$Im\boldsymbol{S_t} = \boldsymbol{F_t} \circ Im\boldsymbol{S_{t-1}} + (\boldsymbol{i_t} \circ \tilde{\boldsymbol{c}}_{\boldsymbol{t}}) \left[sin\omega_1 t, ...sin\omega_K t\right] \quad (9)$$

It is well known that complex numbers can be uniquely represented by its amplitude and phase. To encode the state-frequency matrix $\boldsymbol{S_t}$, we represent its amplitude $\boldsymbol{A_t}$ and the phase $\angle \boldsymbol{S_t}$ as:

$$\boldsymbol{A_t} = |\boldsymbol{S_t}| = \sqrt{(Re\boldsymbol{S_t})^2 + (Im\boldsymbol{S_t})^2} \in \mathbb{R}^{D \times K} \quad (10)$$

$$\angle \boldsymbol{S_t} = \arctan(\frac{Im\boldsymbol{S_t}}{Re\boldsymbol{S_t}}) \in [-\frac{\pi}{2}, \frac{\pi}{2}]^{D \times K} \quad (11)$$

The amplitude will be fed into the memory cell gate and its frequency components will be composed to obtain the output hidden state $\boldsymbol{h_t} \in \mathbb{R}^D$. We ignore the phase $\angle \boldsymbol{S_t}$ as we found it has no significant impact on the results in



SFM

Architecture

忘记

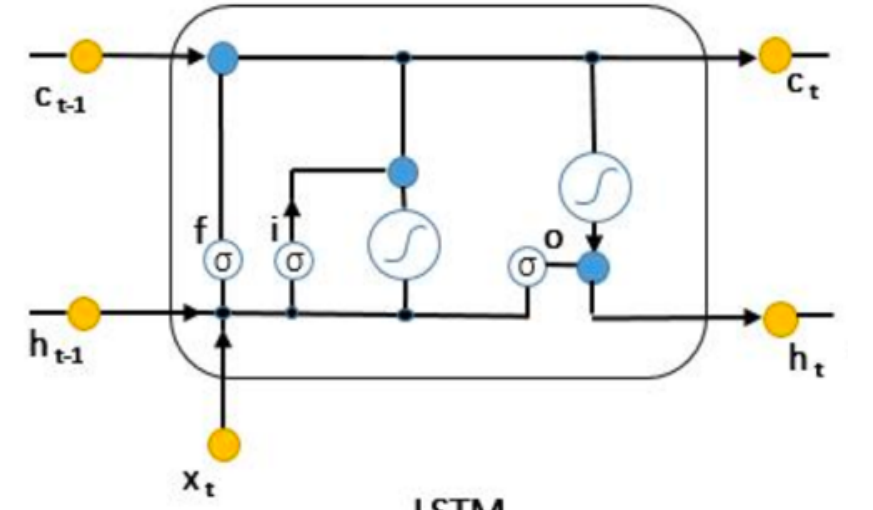
To control how much of the past information should be kept in the memory cell, we define a state forget gate $\boldsymbol{f_t^{ste}}$ and a frequency forget gate $\boldsymbol{f_t^{fre}}$ to regulate the information on multi-states and multi-frequencies respectively. They are formulated as:

$$\boldsymbol{f_t^{ste}} = \text{sigmoid}(\boldsymbol{W_{ste}x_t} + \boldsymbol{U_{ste}h_{t-1}} + \boldsymbol{b_{ste}}) \in \mathbb{R}^D \quad (12)$$

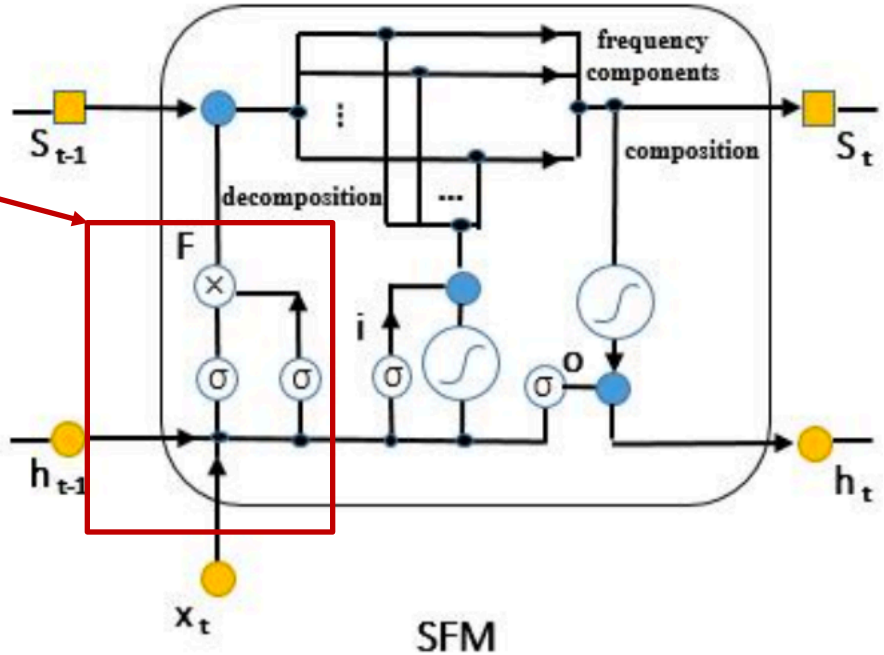
$$\boldsymbol{f_t^{fre}} = \text{sigmoid}(\boldsymbol{W_{fre}x_t} + \boldsymbol{U_{fre}h_{t-1}} + \boldsymbol{b_{fre}}) \in \mathbb{R}^K \quad (13)$$

Then a state-frequency forget gate $\boldsymbol{F_t}$ is defined as the outer product $\otimes$ between $\boldsymbol{f_t^{ste}}$ and $\boldsymbol{f_t^{fre}}$ to jointly regulate the state and frequency information:

$$\boldsymbol{F_t} = \boldsymbol{f_t^{ste}} \otimes \boldsymbol{f_t^{fre}} \in \mathbb{R}^{D \times K} \quad (14)$$



LSTM



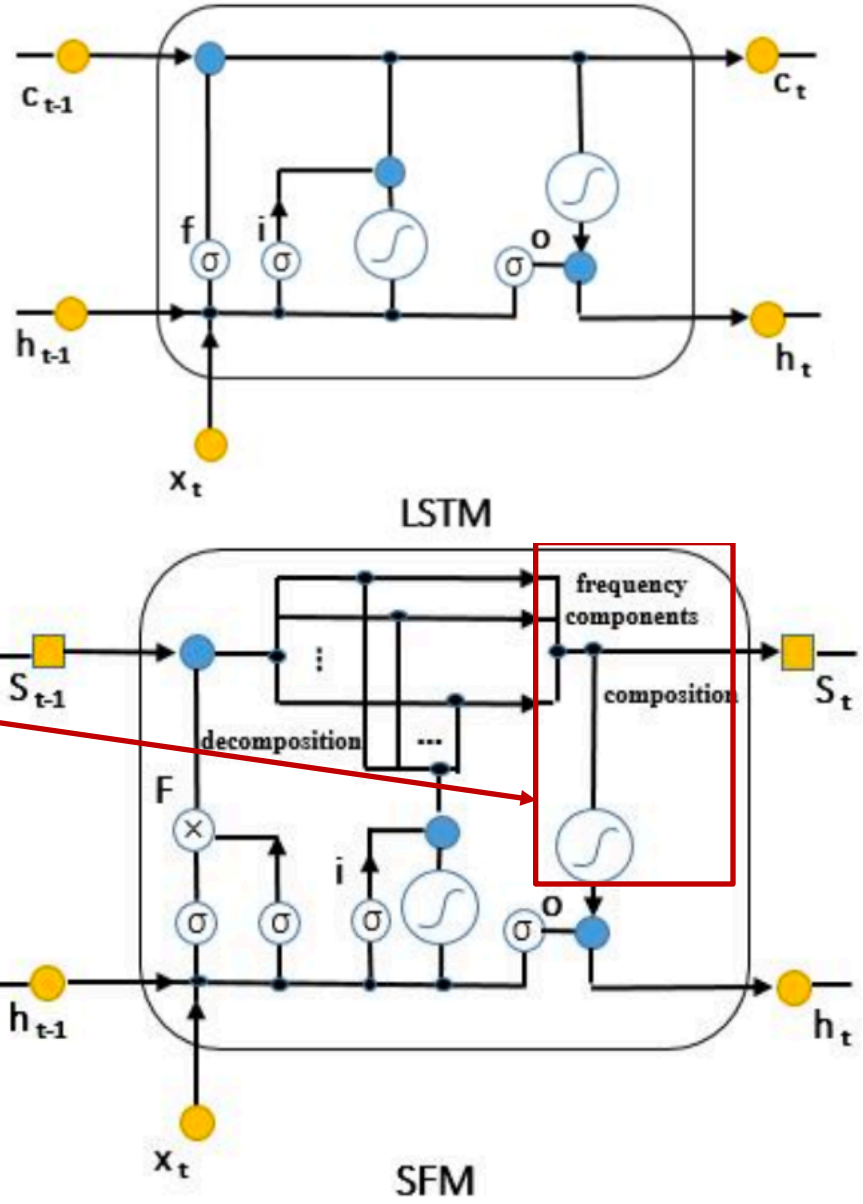
SFM

Architecture

To obtain the output hidden state $\boldsymbol{h_t}$, a state-only memory state $\boldsymbol{c_t}$ is reconstructed to aggregate the information over various frequencies on the state amplitude $\boldsymbol{A_t}$:

$$\boldsymbol{c_t} = \tanh(\boldsymbol{A_t u_a} + \boldsymbol{b_a}) \tag{17}$$

where $\boldsymbol{u_a} \in \mathbb{R}^K$ is a inverse transform vector. The vector composites the frequency components of the memory state. Then the state-only state $\boldsymbol{c_t}$ is obtained as a non-linear mapping of the composition. This process is like the Inverse Fourier Transformation (IFT) which recovers the original signal by combining the frequency components. Rather than adopting

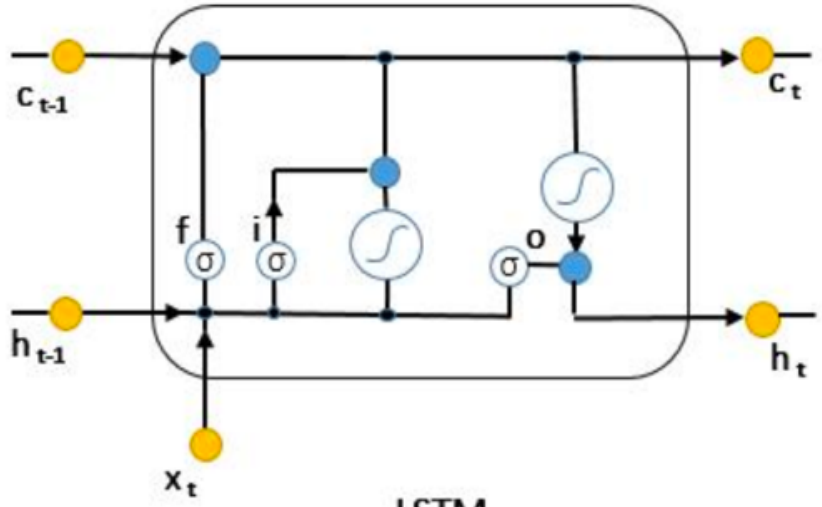


With the output gate $\boldsymbol{o_t}$ regulating the information allowed to output from the memory cell, the output hidden state $\boldsymbol{h_t}$ is computed as:

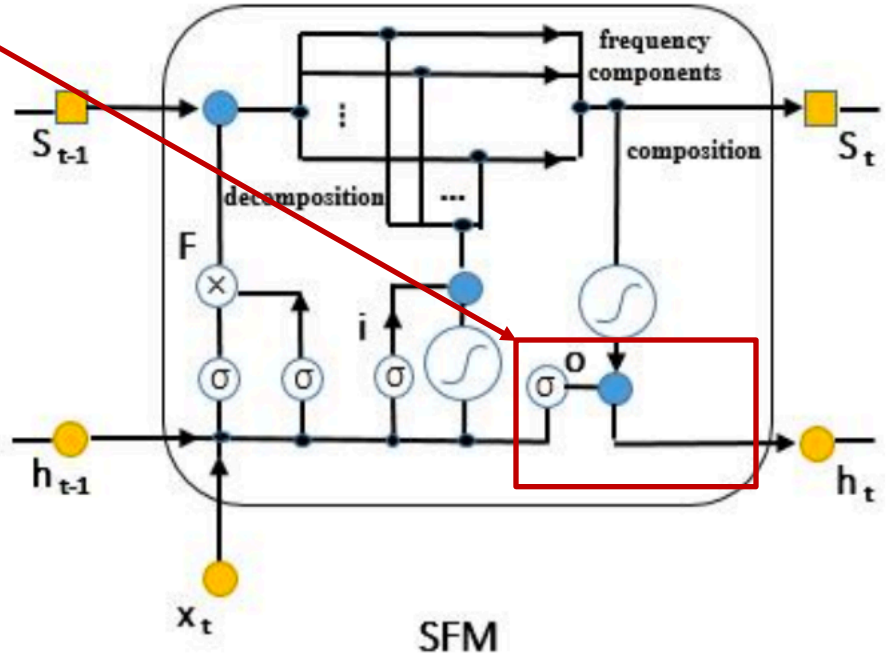
$$\boldsymbol{o_t} = \text{sigmoid}(\boldsymbol{W_o x_t} + \boldsymbol{U_o h_{t-1}} + \boldsymbol{V_o c_t} + \boldsymbol{b_o}) \quad (18)$$

$$\boldsymbol{h_t} = \boldsymbol{o_t} \circ \boldsymbol{c_t} \quad (19)$$

记住



LSTM

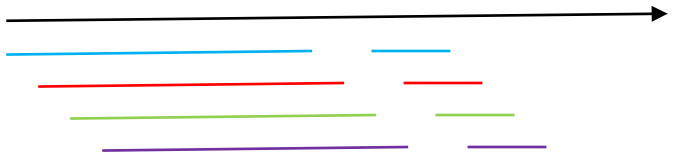


SFM

Definition 3.1. (n-step prediction)
Given prices $\{p_t|t = 1, 2, ..., T\}$, n-step prediction on the price p_{t+n} at time $t + n$ can be seen as a function:

$$\hat{p}_{t+n} = f(p_t, p_{t-1}, ..., p_1) \tag{20}$$

where f denotes the model mapping from the history prices to the price of n-step ahead.



$$\hat{v}_{t+n} = \boldsymbol{w_p h_t} + b_p \tag{21}$$

where $\boldsymbol{w_p}$ is a weight vector, and b_p is the bias. Note that although this is a linear transformation, the nonlinearity of this price predictor arises from the nonlinear hidden vector $\boldsymbol{h_t}$.

$$\mathcal{L} = \sum_{m=1}^{M} \sum_{t=1}^{T} (v_{t+n}^m - \hat{v}_{t+n}^m)^2 \tag{22}$$

The RNN layer in the regression network is flexible to be replaced by any RNN variant. We replace the RNN with the LSTM and the SFM for comparison. In both architectures,

A Hybrid Framework for Text Modeling with Convolutional RNN

Chenglong Wang, Feijun Jiang, Hongxia Yang
Alibaba Group
969 West Wenyi Road
Hangzhou, China 310000
{chenglong.cl,feijun.jiangfj,yang.yhx}@alibaba-inc.com

- 在本文中，我们引入了一种用于文本语义建模的卷积递归神经网络（conv-RNN）的通用推理混合框架，无缝集成了从提取和回归神经的语言信息的不同方面的优点网络结构，从而增强新框架的语义理解能力。
- 此外，基于conv-RNN，我们还提出了一个新的句子分类模型和一个基于句子的答案选择模型，分别为句子匹配和分类提供了强化能力。
- 我们在各种数据集上验证所提出的模型，其中包括两个具有挑战性的答案选择（AS）和用于句子分类（SC）的基准数据集。就我们所知，它是迄今为止在AS和SC中最完整的比较结果。我们在这些不同的具有挑战性的任务和基准数据集中凭经验显示了conv-RNN的优异表现，并总结了其他一些最新方法的表现。

Architecture

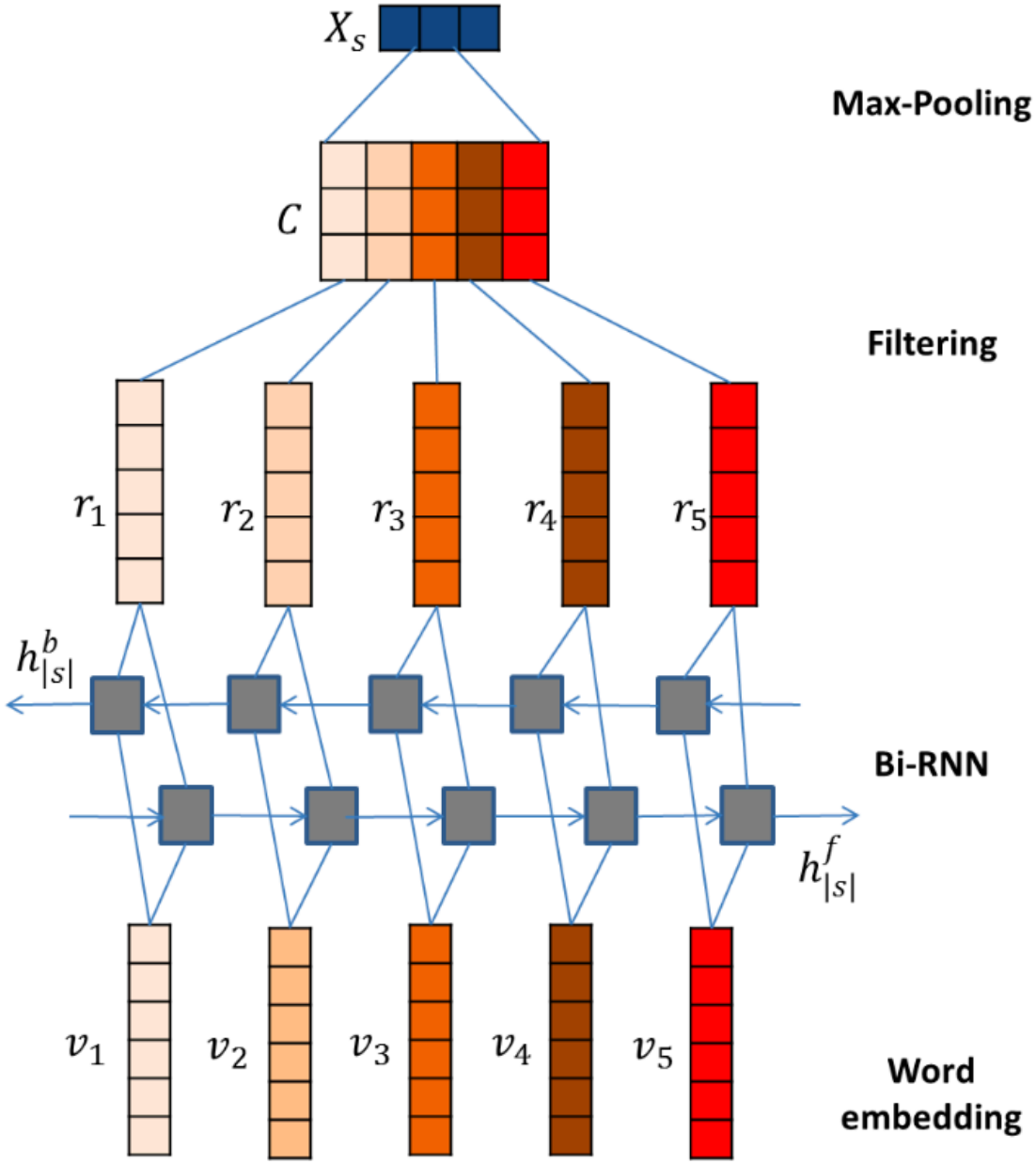


Figure 1: *conv – RNN*

conv−RNN for Sentence Classification

This joint layer concatenates the output of $\boldsymbol{conv - RNN}$, X_q, and two final hidden states from the forward and backward RNN units respectively into $X_{join} = \left[h^f_{|s|}{}', X'_q, h^b_{|s|}{}'\right]'$, which is used as the final representation of input texts. This model includes an additional hidden layer on top of the joint layer to allow for modeling interactions between the components of intermediate representations.

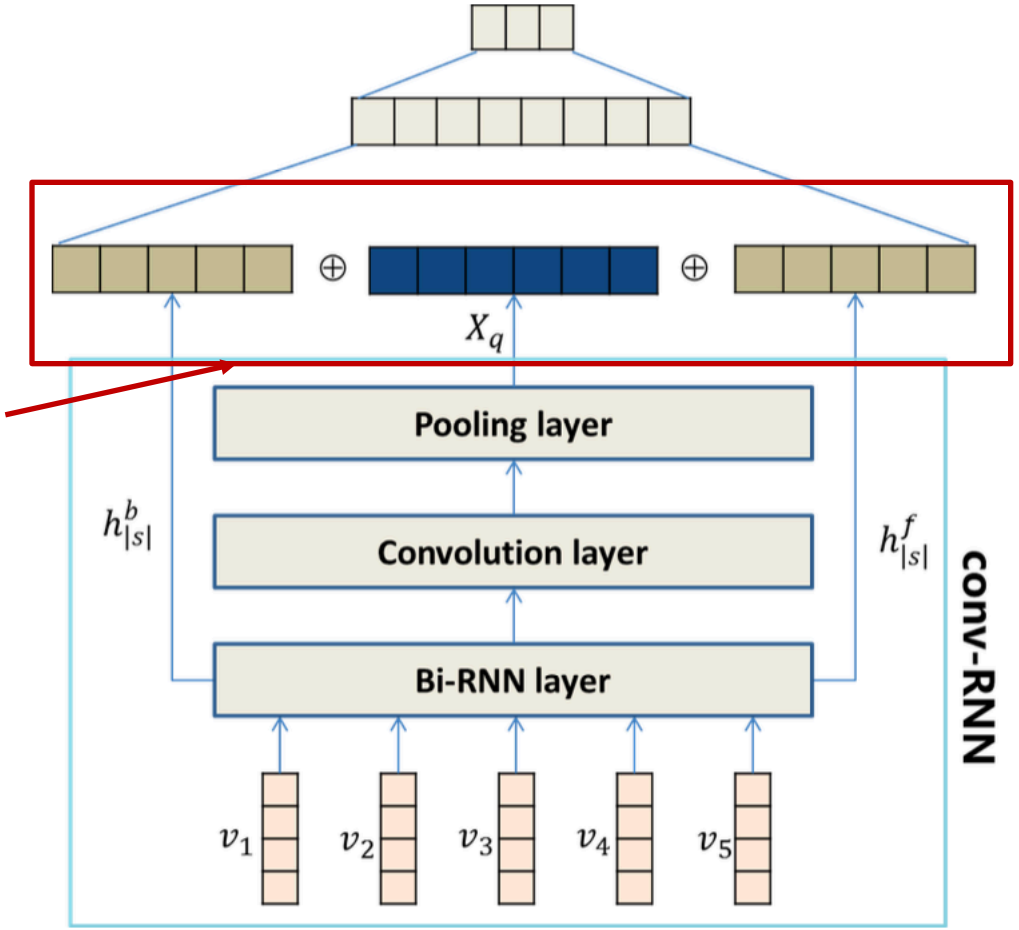


Figure 2: *conv − RNN* based sentence classification.

Attention Based conv–RNN for Answer Selection

Given the resulting vector representations X_q and X_a, the Geometric mean of Euclidean and Sigmoid Dot (GESD)[4] is used to measure the relatedness between the two representations:

$$X_{sim} = \frac{1}{1 + \|x - y\|} \times \frac{1}{1 + exp(-\gamma(xy^T + c))}. \quad (18)$$

It has been proved that GESD could achieve superior performance than simple cosine similarity. On top of the GESD layer and two blocks, there is a joint layer which concatenates X_q, X_a and X_{sim} into a single vector: $X_{join} = [X'_q, X'_a, X'_{sim}]'$. This vector is then passed through two layers of full-connected neural networks, which generates a distribution over the class labels.

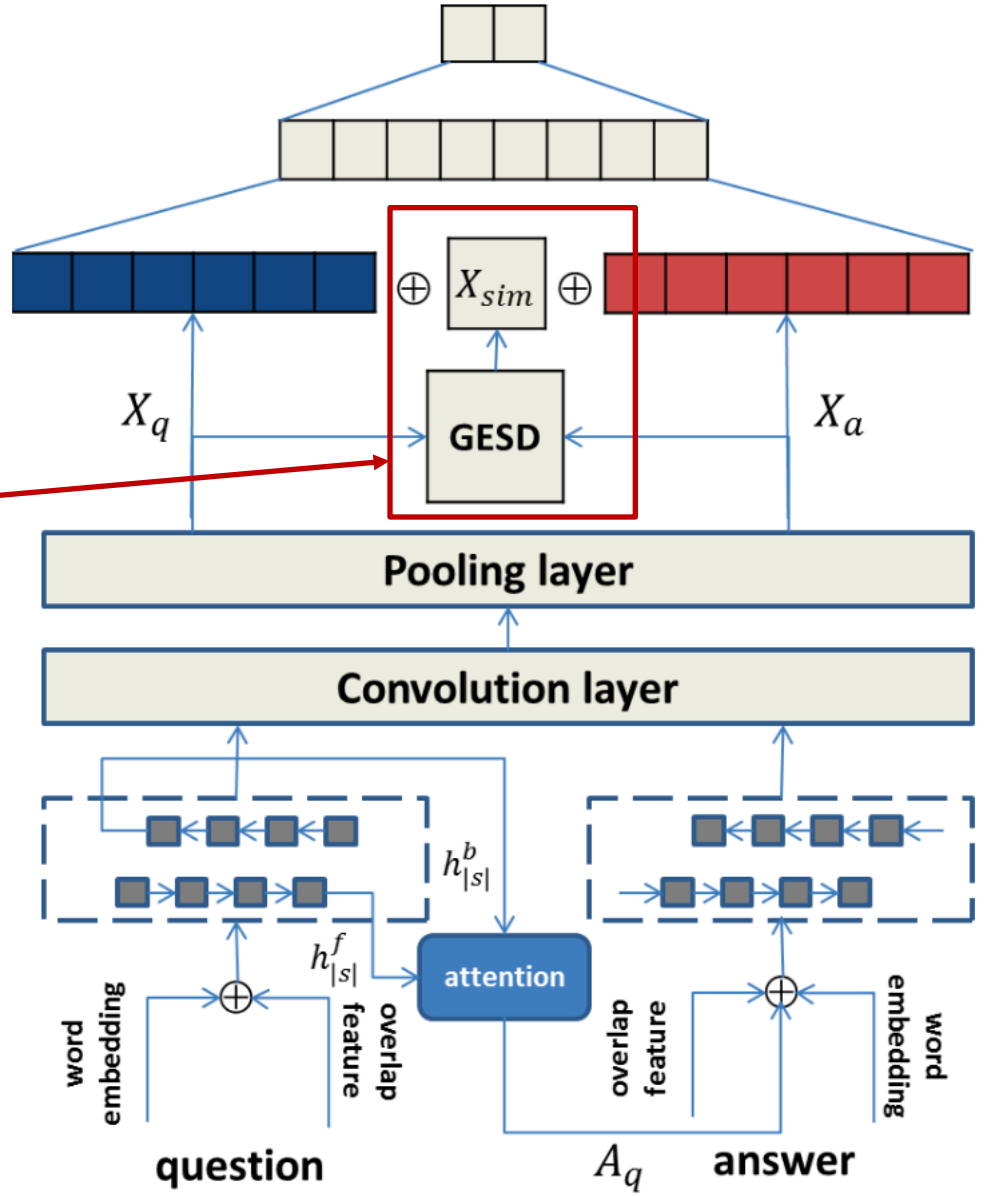


Figure 3: *conv – RNN* based question-answer matching network.

STAR: A System for Ticket Analysis and Resolution

Wubai Zhou, Wei Xue
Computer Science
Florida International University
Miami, USA
{wzhou005,wxue004}@cs.fiu.edu

Ramesh Baral
Computer Science
Florida International University
Miami, USA
rbara012@cs.fiu.edu

Qing Wang
Computer Science
Florida International University
Miami, USA
qwang028@cs.fiu.edu

Chunqiu Zeng
Computer Science
Florida International University
Miami, USA
czeng001@cs.fiu.edu

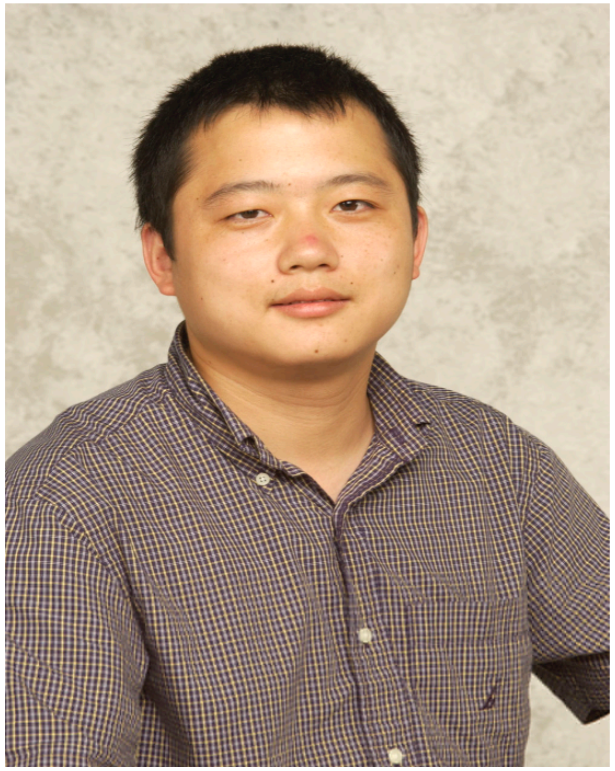
Tao Li
Florida International University
Nanjing University of Posts and Telecommunications
taoli@cs.fiu.edu

Jian Xu
Computer Science & Engineering
Nanjing University of Science and Technology
dolphin.xu@njust.edu.cn

Zheng Liu
Nanjing University of Posts and Telecommunications
zliu@njupt.edu.cn

Larisa Shwartz
IBM T.J. Watson Research Center
New York, USA
lshwart@us.ibm.com

Genady Ya. Grabarnik
Dept. Math & Computer Science
St. John's University, Queens
grabarng@stjohns.edu



Tao Li, PhD (李涛)

Professor
School of Computer Science
Florida International University

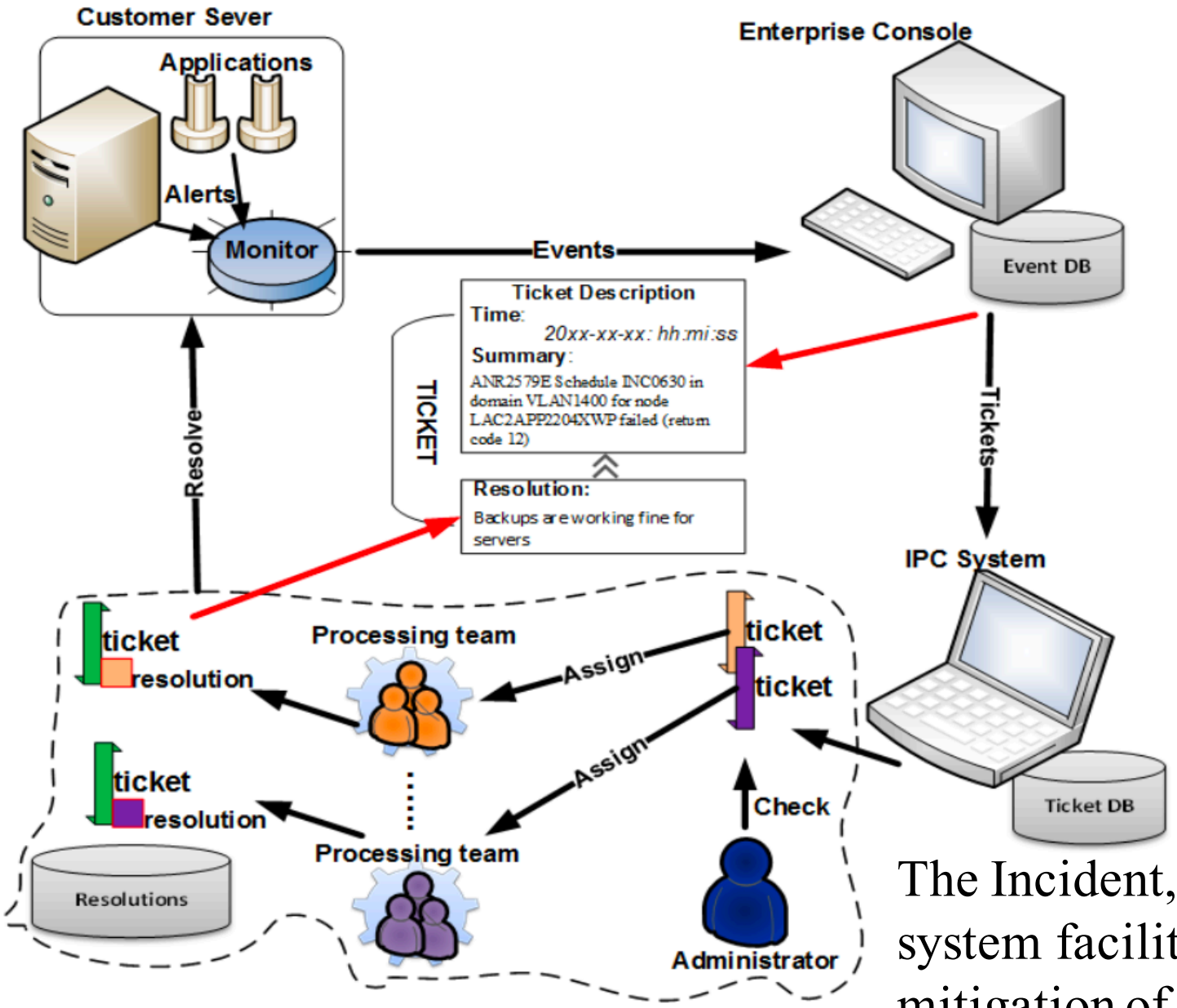
Office: ECS 365
Lab: ECS 251
11200 SW 8th Street

Dr Tao Li's research explores two related topics on learning from data---how to efficiently discover useful patterns and how to effectively retrieve information. The interests lie broadly in data mining and machine learning studying both the algorithmic and application issues.

2015

- Liang Tang, Yexi Jiang, Lei Li, Chunqiu Zeng, and Tao Li. **Personalized Recommendation via Parameter-Free Contextual Bandits**. In Proceedings of the 38th International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR 2015), to appear, 2015.
- Wubai Zhou, Liang Tang, Tao Li, Larisa Shwartz, Genady Ya. Graharnik. **Resolution Recommendation for Event Tickets in Service Management**. In Proceeding of the 14th IFIP/IEEE Symposium on Integrated Network and Service Management (IM 2015), to appear, 2015.
- Longhui Zhang, Lei Li, Chao Shen, and Tao Li. **PatentCom: A Comparative View of Patent Document Retrieval**. In Proceedings of 2015 SIAM International Conference on Data Mining (SDM 2015), to appear, 2015.
- Qifeng Zhou, Hao Zhou, Yimin Zhu, Tao Li. **Data-Driven Solutions for Building Environmental Impact Assessment.** In Proceedings of the 9th IEEE International Conference on Sematic Computing (ICSC 2015), pages. 316--319, 2015.
- Lingwei Chen, Tao Li, Melih Abdulhayoglu, Yanfang Ye. **Intelligent Malware Detection Based on File Relation Graphs.** Proceedings of the 9th IEEE International Conference on Sematic Computing (ICSC 2015), pages. 85--92, 2015.

- 在大规模和复杂的IT服务环境中，有问题的事件会记录为故障单并包含**故障单**摘要（系统状态和问题描述）。系统管理员在解决此类故障单时记录逐步解决方案说明。
- 重复服务事件很可能通过推断类似的历史**故障单**来解决。凭借合理的大量故障单数据集的可用性，我们可以使用自动化系统为给定票据摘要推荐最佳匹配解决方案。
- 在本文中，我们首先确定实际故障单分析中的挑战，并开发一个集成框架来有效地处理这些挑战。该框架首先使用基于精心设计的功能的回归模型来量化故障单解决方案的质量。
- 故障单以及从故障单方案质量量化获得的质量分数，然后用于训练一个深度神经网络排名模型，输出故障单汇总和分辨率对的匹配分数。



The Incident, Problem, and Change (IPC) system facilitates the tracking, analysis and mitigation of problems and is a requirement for organizations adapting the ITIL framework.

Table 1: A sample ticket

SEVERITY	FIRST-OCCURRENCE	LAST-OCCURRENCE
0	2014-03-29 05:50:39	2014-03-31 05:36:01
SUMMARY	ANR2579E Schedule INC0630 in domain VLAN1400 for node LAC2APP2204XWP failed (return code 12)	
RESOLUTION	Backups are working fine for the server.	
CAUSE	**ACTIONABLE**	**LAST-UPDATE**
Maintenance	Actionable	2014-04-29 23:19:25

described in the ticket is negligible. Based on our long preliminary study [39], we've found that for a typical ticket, the ticket resolution quality is driven by the 33 features that can be broadly divided into following four groups:

- **Character-level features:** A low-quality ticket resolution might include a large number of unexpected characters, such as space, wrong or excessive capitalization, and special characters.
- **Entity-level features:** A high-quality ticket resolution is expected to provide information on IT-related entities, such as server name, file path, IP address, and so forth. Because the ticket resolutions are expected to guide system administrators to solve the problem specified in the ticket summary, the presence of the context-relevant entities makes the resolution text more useful.
- **Semantic-level features:** A high-quality ticket resolution typically includes *Verb* and *Noun*, which explicitly guides system administrators on the actions taken to diagnose the problem and to resolve the ticket.
- **Attribute-level features:** A high-quality ticket resolution usually is lengthy enough to carry sufficient information relevant to the problem described in the ticket summary.

The ticket resolution quality quantifier uses these 4 groups of features and operates on the historical tickets to output a set of triplets $\{< s_1, r_1, q_1 >, < s_2, r_2, q_2 >, \ldots, < s_n, r_n, q_n >\}$ where s_i and r_i are ticket summary and ticket resolution for the i^{th} ticket, and q_i is the quality score assigned by the quantifier.

Representation for ticket summary and resolution pair

Having the output of our sentence model for processing ticket summary and resolution, respectively, the resulting representation vectors x_s and x_r, can be used to compute the ticket summary and resolution similarity score as follows:

$$sim(x_s, x_r) = x_s^T M x_r \quad (2)$$

Where $M \in \mathbb{R}^{d \times d}$ is a similarity matrix, it acts as a model of noisy channel approach for machine learning, which has been commonly adopted as a scoring model in information retrieval and question answer [8]. It can also be viewed as a process of learning similarity metric on two vectors drawing from different feature spaces [14]. The similarity matrix M is a parameter of the network and is optimized during the training

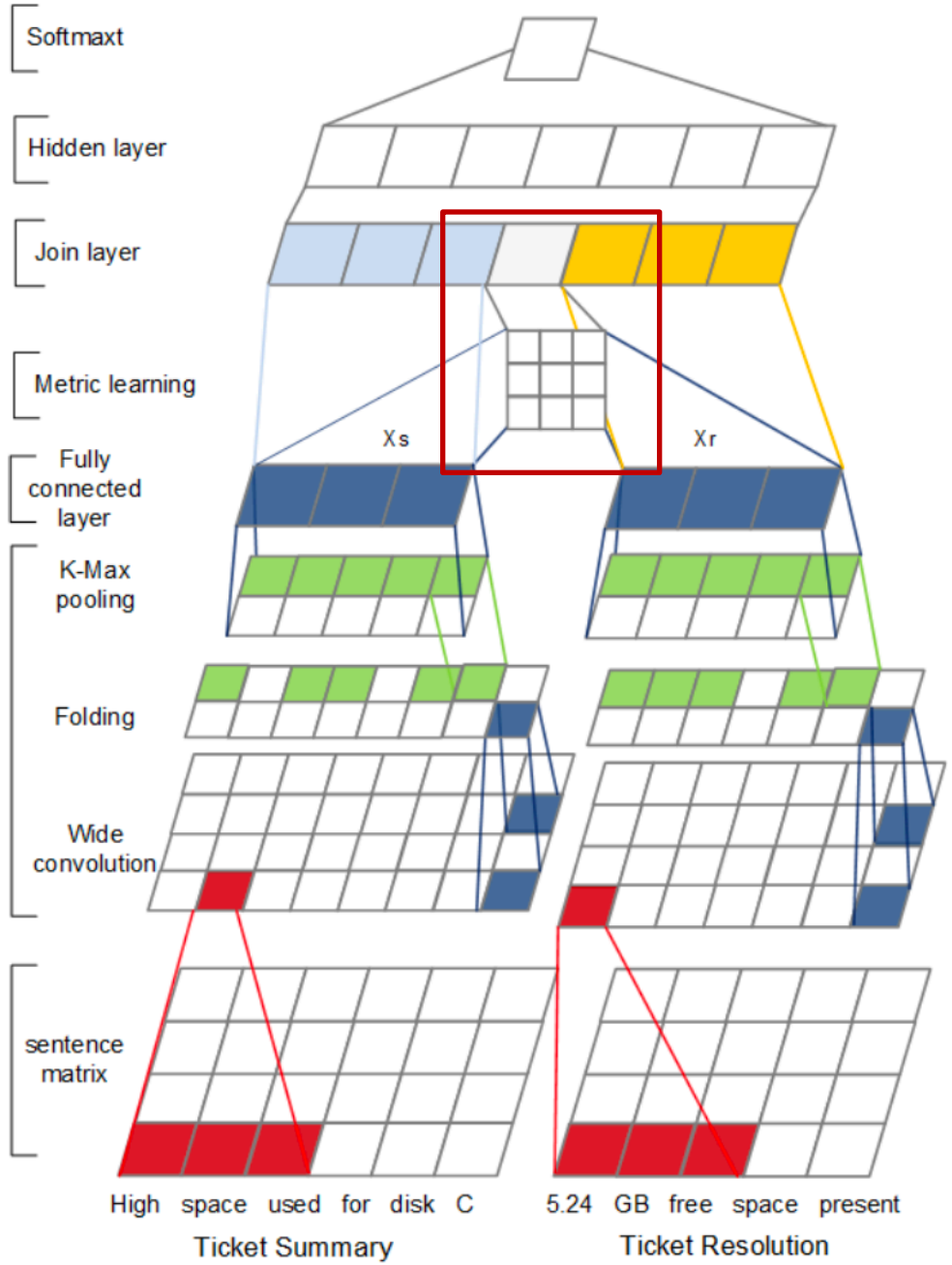


Table 6: Overall performance comparison.

System	p1	MAP	nDCG5	nDCG10
SMT	0.421	0.324	0.459	0.501
LSTM-RNN	0.563	0.367	0.572	0.628
Random Shuffle	0.343	0.273	0.358	0.420
CombinedLDAKNN	0.482	0.347	0.484	0.536
Our method	**0.742**	**0.506**	**0.628**	**0.791**

Structural Deep Brain Network Mining

Shen Wang
University of Illinois at Chicago
Chicago, USA
swang224@uic.edu

Lifang He*
Shenzhen University
Shenzhen, China
lifanghescut@gmail.com

Bokai Cao
University of Illinois at Chicago
Chicago, USA
caobokai@uic.edu

Chun-Ta Lu
University of Illinois at Chicago
Chicago, USA
clu29@uic.edu

Philip S. Yu
University of Illinois at Chicago
Chicago, USA
Tsinghua University
Beijing, China
psyu@cs.uic.edu

Ann B. Ragin
Northwestern University
Chicago, USA
ann-ragin@northwestern.edu

[ICDM 2017]
MNE: Emerging Network Embedding with Aligned Autoencoder, Jiawei Zhang, Congying Xia, Chenwei Zhang, Limeng Cui, Yanjie Fu and Philip S. Yu.
HitFraud: A Broad Learning Approach for Collective Fraud Detection in Heterogeneous Information Networks, Bokai Cao, Mia Mao, Siim Viidu and Philip S. Yu.
Collaborative Inference of Coexisting Information Diffusions, Yanchao Sun, Cong Qian, Ning Yang and Philip S. Yu.
Multi-view Graph Embedding with Hub Detection for Brain Network Analysis, Guixiang Ma, Chun-Ta Lu, Lifang He, Philip S. Yu and Ann B. Ragin.
A Broad Learning Approach for Context-Aware Mobile Application Recommendation, Tingting Liang, Lifang He, Chun-Ta Lu, Liang Chen, Philip S. Yu and Jian Wu.
[CIKM 2017]
ECD: Enterprise Social Community Detection via Hierarchical Structure Fusion, Jiawei Zhang, Limeng Cui, Philip S. Yu, Yuanhua Lv and Yanjie Fu.
Multi-Source Collaborative Recommendation, Junxing Zhu, Jiawei Zhang, Lifang He, Quanyuan Wu, Bin Zhou, Chenwei Zhang and Philip S. Yu.
Unsupervised Feature Selection with Heterogeneous Side Information, Xiaokai Wei, Bokai Cao and Philip S. Yu.
Multi-view Clustering via Graph Embedding for Connectome Analysis, Guixiang Ma, Lifang He, Chun-Ta Lu, Weixiang Shao, Philip S. Yu, Alex D. Leow and Ann B. Ragin.
Coupled Sparse Matrix Factorization for Response Time Prediction in Logistics Services, Yuqi Wang, Jiannong Cao, Lifang He, Wengen Li, Lichao Sun and Philip S. Yu.
[KDD 2017]
Structural Deep Brain Network Mining, Shen Wang, Lifang He, Bokai Cao, Chun-Ta Lu, Philip S. Yu and Ann B. Ragin.
DeepMood: Modeling Mobile Phone Typing Dynamics for Mood Detection, Bokai Cao, Lei Zheng, Chenwei Zhang, Philip S. Yu, Andrea Piscitello, John Zulueta, Olu Ajilore, Kelly Ryan and Alex D. Leow.
[ICML 2017]
Kernalized Support Tensor Machines, Lifang He, Chun-Ta Lu, Guixiang Ma, Shen Wang, Linlin Shen, Philip S. Yu and Ann B. Ragin.
[CVPR 2017]
Multi-way Multi-level Kernel Modeling for Neuroimaging Classification, Lifang He, Chun-Ta Lu, Hao Ding, Shen Wang, Linlin Shen, Philip S. Yu and Ann B. Ragin.
[IJCAI 2017]
SEVEN: Deep Semi-supervised Verification Networks, Vahid Noroozi, Lei Zheng, Sara Bahaadini, Sihong Xie and Philip S. Yu.
[SDM 2017]
t-BNE: Tensor-based Brain Network Embedding, Bokai Cao, Lifang He, Xiaokai Wei, Mengqi Xing, Philip S. Yu, Heide Klumpp and Alex D. Leow.
[ICDE 2017]
Link Prediction across Aligned Networks with Sparse Low Rank Matrix Estimation, Jiawei Zhang, Jianhui Chen, Shi Zhi, Yi Chang, Philip S. Yu and Jiawei Han.
Enterprise Social Community Detection, Jiawei Zhang, Philip S. Yu and Yuanhua Lv.
[WWW 2017]
Cross View Link Prediction by Learning Noise-resilient Representation Consensus, Xiaokai Wei, Linchuan Xu, Bokai Cao and Philip S. Yu.
[WSDM 2017]
Embedding of Embedding (EOE) : Embedding for Coupled Heterogeneous Networks, Linchuan Xu, Xiaokai Wei, Jiannong Cao and Philip S. Yu.
Enterprise Employee Training via Project Team Formation, Jiawei Zhang, Philip S. Yu and Yuanhua Lv.
Joint Deep Modeling of Users and Items Using Reviews for Recommendation, Lei Zheng, Vahid Noroozi and Philip S. Yu.
Link Prediction with Cardinality Constraint, Jiawei Zhang, Jianhui Chen, Junxing Zhu, Yi Chang and Philip S. Yu.
Multilinear Factorization Machines for Multi-Task Multi-View Learning, Chun-Ta Lu, Lifang He, Weixiang Shao, Bokai Cao and Philip S. Yu.

Philip S. Yu

UIC Distinguished Professor and Wexler Chair in Information Technology
Department of Computer Science
University of Illinois at Chicago

- 由于**神经影像学数据**在医疗保健和生物信息学领域的挖掘越来越受欢迎，因为它有可能发现有助于理解和诊断神经系统疾病和神经系统疾病的有临床意义的结构组织。
- 最近的研究集中在应用子图挖掘技术来发现大脑网络中的连通子图模式。但是，潜在的大脑网络结构是复杂的。作为浅层线性模型，子图模型不能捕获高度非线性的结构，导致次优的模型。因此，如何学习能够捕捉大脑网络的高度非线性并保持底层结构的表示是一个关键问题。
- 在本文中，我们提出了一种结构深部脑网络分割方法，即SDBN，以学习高度非线性和保持结构的大脑网络表示。
- 具体而言，我们首先引入一种基于模块标识的新型图形重新排序方法，它重新排列节点的顺序以保留图形的模块化结构。接下来，我们执行结构增强以进一步增强重新排序的图的空间信息。
- 我们提出了一种深层特征学习框架，通过将**卷积神经网络**（CNN）与解码路径进行重建，将小监视下的监督学习和无监督学习结合起来。
- 在多层非线性映射的帮助下，所提出的SDBN方法可以捕获脑网络的高度非线性结构。此外，它具有高维脑网络的泛化能力，即使对于小样本学习也能很好地工作。从CNN的面向任务的学习风格来看，学习的分层表示对于临床任务是有意义的。

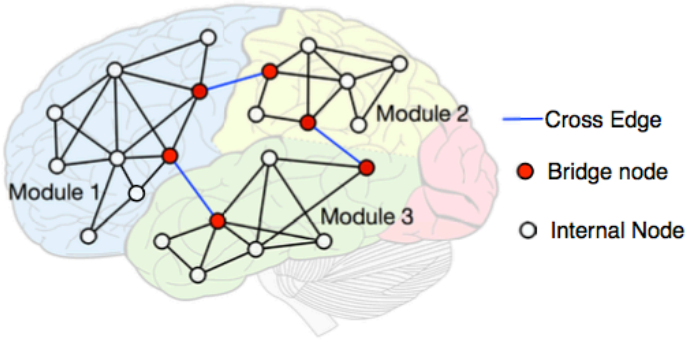
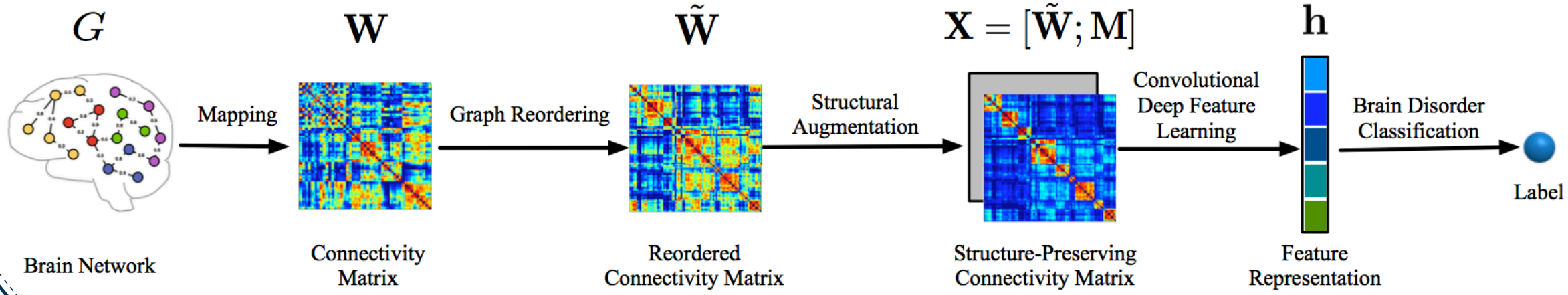


Figure 1: A brain network example, associated with modules, cross edges, internal and bridge nodes

Definition 2.1. **(Brain Network)** A brain network (or connectome), is a weighted graph $G = (V, E, \mathbf{W})$ with $|V|$ nodes and $|E|$ edges reflecting brain regions of interest (ROIs) and connectivities between ROIs, respectively. The weights or adjacency matrix in $\mathbf{W}$ represent the connectivity degree, where a larger weight corresponds to a higher connectivity degree, reflecting stronger functional correlations in fMRI and tighter fiber connections in DTI.



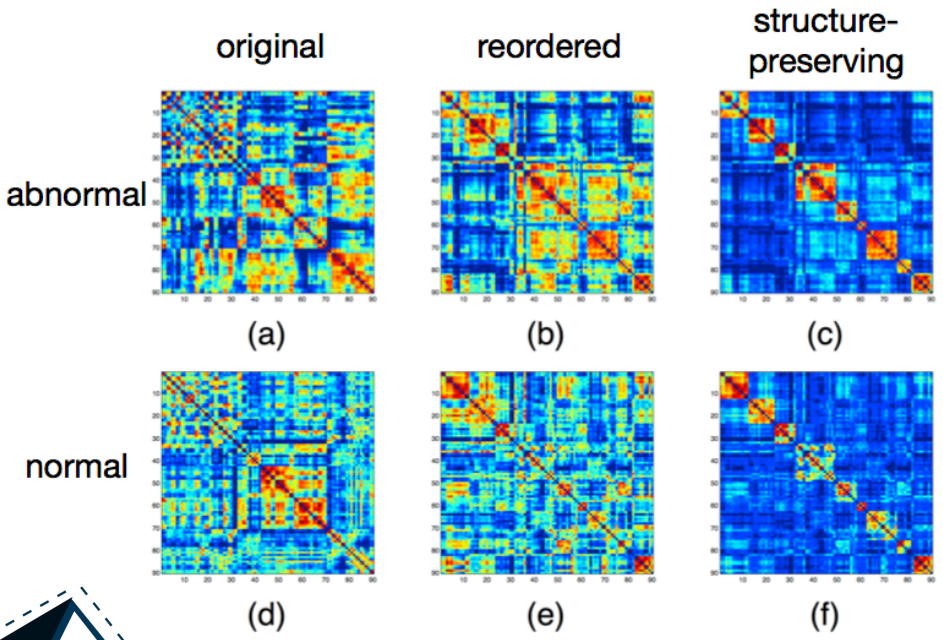
Definition 2.2. **(Module)** A module (also called community or group) in a graph is a subset of nodes which are densely connected to each other, but sparsely connected to the nodes in other modules. The node with all its neighboring nodes in the same module is internal node. The node with neighboring nodes belong to the different modules is bridge node.

Definition 2.3. **(Graph Reordering)** Given a collection of unlabeled graphs $\mathcal{G} = \{G_1, G_2, \cdots, G_N\}$, the goal of graph reordering is to find a labeling ℓ such that for any two graphs $G_i, G_j \in \mathcal{G}$ drawn uniformly at random from $\mathcal{G}$, the expected difference between the distance of the graph connectivity matrices based on ℓ and the distance of the graphs in graph space is minimized. Let $\mathbf{d_G}$ be a distance measure on graphs $\mathcal{G}$, and $\mathbf{d_W}$ be a distance measure on connectivity matrices $\mathcal{W}$. It can be formulated as the following optimization problem:

$$\arg\min_{\ell} \mathbb{E}_{\mathbf{G}}[\|\mathbf{d_W}(\mathbf{W}_i^{\ell}, \mathbf{W}_j^{\ell}) - \mathbf{d_G}(G_i, G_j)\|] \tag{1}$$

Graph Reordering

Alternatively, we borrow the idea from graph compression [2] to address this problem.



form module identification. The key idea of spectral clustering is to convert a clustering problem into a graph partitioning problem and then solve this problem by means of matrix theory. Let K be the number of modules to be identified, and $\hat{\mathbf{W}} \in \mathbb{R}^{N\times N}$ be the average connectivity matrix across all graphs. Then, spectral clustering can be formulated as follows:

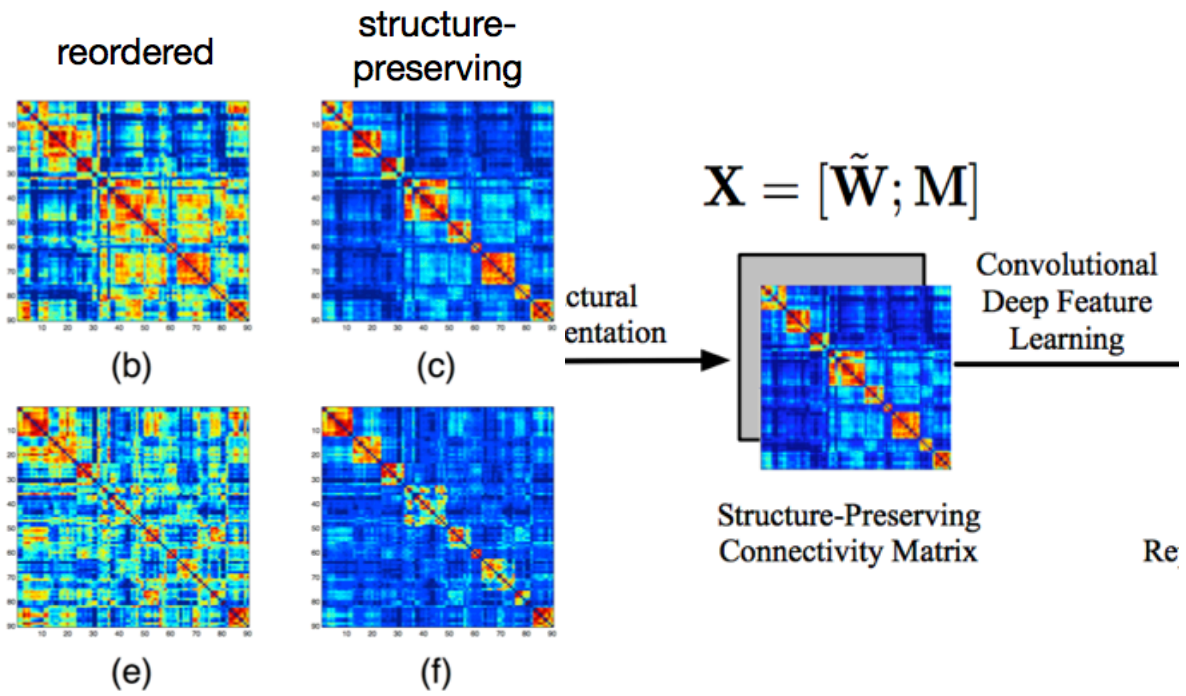
$$\min_{\mathbf{F}} \sum_{i,j=1}^{N} \hat{w}_{ij} \left\|\mathbf{f}_i - \mathbf{f}_j\right\|_2^2 = \mathrm{tr}\left(\mathbf{F}^{\mathrm{T}}\mathbf{L}\mathbf{F}\right)$$
$$\text{s.t. } \mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{I}_K \tag{2}$$

where $tr(\cdot)$ denotes the trace function, a superscript T denotes transposition, $\mathbf{I}_K$ denotes the identity matrix with size K, and $\mathbf{L}$ is the Laplacian matrix [45] obtained from $\hat{\mathbf{W}}$.

By applying K-means to the eigenvectors of the Laplacian matrix $\mathbf{L}$, we can obtain K modules $\mathcal{M} = \{M_1, M_2, \cdots, M_K\}$, with $V = M_1 \cup M_2 \cup \cdots \cup M_K$ and $M_i \cap M_j = \varnothing$ for every pair i, j with $i \neq j$. We treat the module sequence $1, 2, \cdots, K$ as a permutation of nodes. Based on this, the reordered connectivity matrix $\tilde{\mathbf{W}}$ for each brain network can be established. Since all reordered connectivity matrices are obtained according to the global module structure, different subjects are more comparable. An example of the reordered connectivity matrices for abnormal and normal subjects is shown in the second column (b) and (e) of Figure 3.

3.2 Structural Augmentation

According to the graph reordering procedure above, we obtain the reordered connectivity matrix for each brain network. This representation is more meaningful and beneficial to CNN architecture. However, there are still two problems affecting the use of CNN. The first one is the spatial information loss caused by CNN itself. The second one is the noise in brain network. To address these problems, we propose a structural augmentation approach to further enhance the spatial information of the reordered graph for structure-preserving and noise-robust feature learning.



To tackle this problem, we augment the reordered connectivity matrix with an additional channel. Specifically, we define a module identification matrix $\mathbf{M}$ to further encode the module structure information, whose element is:

$$m_{ij} = \begin{cases} k & \text{for} \quad v_i, v_j \in M_k, i, j = 1, \ldots, N \\ 0 & \text{for} \quad otherwise \end{cases} \tag{3}$$

On the other hand, brain networks usually suffer from noise, which is introduced by the error in the acquisition and in the image analysis. For the reordered connectivity matrix obtained from graph reordering, the intra-module neighbor edges reside in the block-diagonal region and the cross edges lie in the off-diagonal region. The intra-module neighbor edges preserve the structure within each module, which are more reliable to learn structure-preserving graph representation, while the cross edges are more complicated which may have negative effects to study the modular structure and some of them may be invalid connections [18]. To alleviate the effects of cross edges, we further augment the reordered connectivity matrix $\tilde{\mathbf{W}}$ by lowering the weights of the off-block diagonal regions such that

$$\tilde{w}_{ij} = \begin{cases} 1 & \text{for} \quad i, j \in M_k, k = 1, \ldots, K \\ \epsilon & \text{for} \quad i, j \notin M_k, k = 1, \ldots, K \end{cases} \tag{4}$$

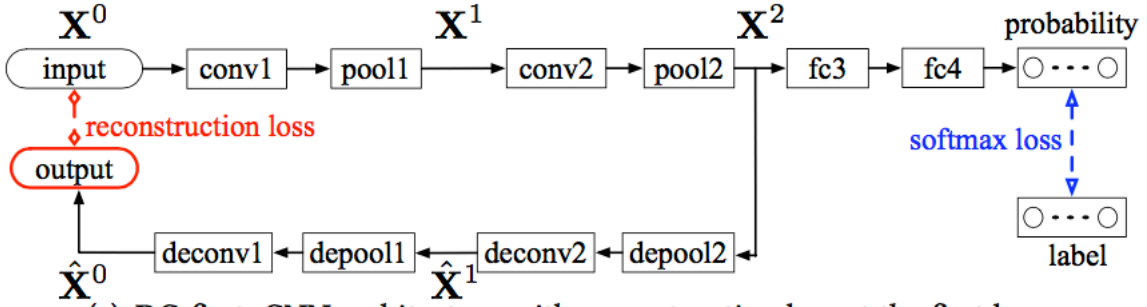
Unsupervised Learning Augmented CNN

However, CNN usually requires a large amount of labeled data for training, which is infeasible for current brain health research, due to the limited number of available patients of interest and the high costs of acquiring the data. It could potentially limit the performance of CNN.

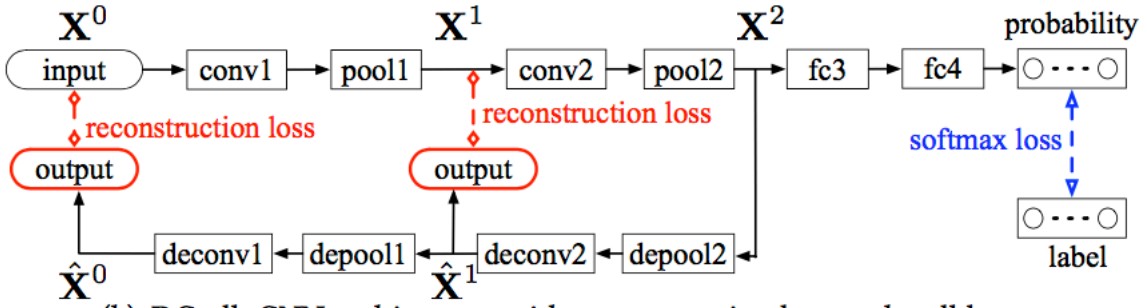
Motivated by the recent attempts [28, 35, 41, 49] to couple the unsupervised and supervised learning in the same phase, making the unsupervised objective being able to impact the network training after supervised learning took place, we augment the auxiliary unsupervised learning objective function to the supervised learning objective function to address the above problem. The joint objective function is as follows

$$\frac{1}{N}\sum_{i=1}^{N}(C(\mathbf{X}_i, y_i) + \lambda U(\mathbf{X}_i)) \tag{8}$$

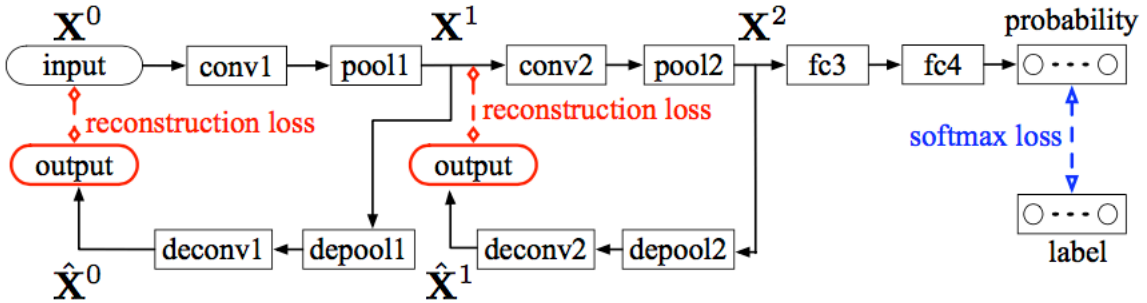
Marc'Aurelio Ranzato and Martin Szummer. Semi-supervised learning of com- pact document representations with deep networks. In Proceedings of the 25th International Conference on Machine Learning, pages 792–799, 2008.



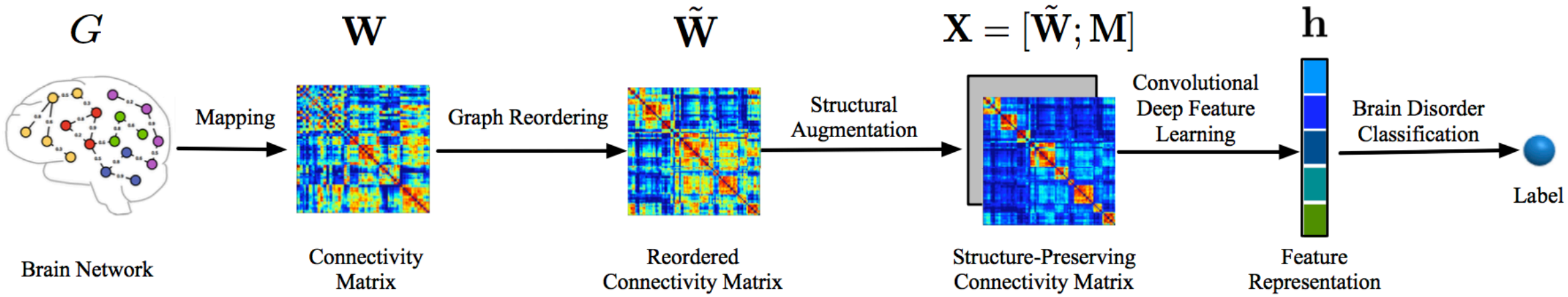
(a) DC-first: CNN architecture with reconstruction loss at the first layer



(b) DC-all: CNN architecture with reconstruction loss at the all layers



(c) DC-layerwise: CNN architecture with layerwise reconstruction loss


G
$\mathbf{W}$
$\tilde{\mathbf{W}}$
$\mathbf{X} = [\tilde{\mathbf{W}}; \mathbf{M}]$
$\mathbf{h}$
Mapping
Graph Reordering
Structural Augmentation
Convolutional Deep Feature Learning
Brain Disorder Classification
Label
Brain Network
Connectivity Matrix
Reordered Connectivity Matrix
Structure-Preserving Connectivity Matrix
Feature Representation

Thanks