Information Cascades Modeling via Deep Multi-Task Learning

Xueqin Chen*, Kunpeng Zhang‡, Fan Zhou*†, Goce Trajcevski§, Ting Zhong*, Fengli Zhang*
University of Electronic Science and Technology of China*, University of Maryland, College park‡
Iowa State University, Ames§
nedchen0728@gmail.com, kzhang@rhsmith.umd.edu, gocet25@iastate.edu,
{fan.zhou, zhongting, fzhang}@uestc.edu.cn

ABSTRACT

Effectively modeling and predicting the information cascades is at the core of understanding the information diffusion, which is essential for many related downstream applications, such as fake news detection and viral marketing identification. Conventional methods for cascade prediction heavily depend on the hypothesis of diffusion models and hand-crafted features. Owing to the significant recent successes of deep learning in multiple domains, attempts have been made to predict cascades by developing neural networks based approaches. However, the existing models are not capable of capturing both the underlying structure of a cascade graph and the node sequence in the diffusion process which, in turn, results in unsatisfactory prediction performance. In this paper, we propose a deep multi-task learning framework with a novel design of shared-representation layer to aid in explicitly understanding and predicting the cascades. As it turns out, the learned latent representation from the shared-representation layer can encode the structure and the node sequence of the cascade very well. Our experiments conducted on real-world datasets demonstrate that our method can significantly improve the prediction accuracy and reduce the computational cost compared to state-of-the-art baselines.

CCS CONCEPTS

• **Information systems → Social networks**; • **Computing methodologies → Multi-task learning**;

KEYWORDS

multi-task learning, information cascades, cascade graph embedding, diffusion process embedding, shared-Gate

ACM Reference Format:

Xueqin Chen, Kunpeng Zhang, Fan Zhou, Goce Trajcevski, Ting Zhong, Fengli Zhang. 2019. Information Cascades Modeling via Deep Multi-Task Learning In *Proceedings of the 42nd Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIRâĂŹ19), July 21–25, 2019, Paris, France.* ACM, NY, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331288

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331288

1 INTRODUCTION

Online social platforms, such as Twitter, Weibo, Wechat, Instgram, etc. have a significant impact on our daily life. Their dramatic growth has facilitated fast propagation of information in various contexts, e.g., the spread of rumors in news, the propagation of marketing campaigns, the diffusion of innovative technological achievements, etc. – spurring the ubiquitous phenomenon of information cascades. Modeling and predicting such cascades is one of the fundamental components for understanding information propagation which, in turn, is beneficial for variety of downstream applications like, for example, fake news detection and viral marketing identification.

Existing studies on modeling information cascades mainly focus on two aspects: (1) On the one hand, macro-level tasks focused on estimating cascade growth [3, 6, 12], and forecasting outbreak [5, 16], which are rough estimations and not suitable for micro-level tasks. (2) On the other hand, micro-level tasks always studying the local patterns of social influence – which pay more attention to user-level modeling instead of the cascade-level (e.g., inferring the action status of a user [12, 14]).

Complementary to these, the conventional methods studying the information diffusion problem can be summarized into the following four categories: (1) *Diffusion-based Approaches* [8, 13] make a strong assumption that the underlying diffusion model follows a known prior distribution – which often is not quite appropriate for cascade prediction; (2) *Feature-based Approaches* [1, 9] focus on identifying and incorporating complicated hand-crafted features, which requires extensive domain knowledge and thus is hard to be generalized to new domains; (3) *Generative Modeling-based Approaches* [3, 7] focus on modeling the intensity function of the arrival process for each message independently. These methods demonstrate an enhanced interpretability but are still unable to fully leverage the information encoded in the cascade for a satisfactory prediction; and (4) *Deep Learning-based Approaches*, especially Recurrent Neural Networks (RNN) based sequential models [3, 6, 12, 14]. While automatically learning temporal characteristics, all proposed methods fail short in integrating structural information of cascades, which are essential for their prediction [4].

While some of the methods mentioned above can achieve certain improvements in cascade modeling, they still exhibit several drawbacks. What motivates this work is the lack of methodology to jointly model cascades from both a micro (user) and a macro (overall cascade estimate) level. To capture both the underlying structure of a cascade graph and node sequence in the diffusion process, we take a full advantage of the modeling from both levels. Inspired by the great success of multi-task learning, we propose a Deep

Multi-Task Learning based Information Cascades model (*DMT-LIC*), which explicitly models and predicts cascades through a multi-task framework with a novel design of a shared-representation layer. In summary, the main contributions of our work are:

- We propose a novel, deep multi-task learning-based method to learn latent semantics of cascades for prediction in an end-to-end manner. In addition, our method does not involve massive and complex feature engineering, which makes our model more generalizable to new domains.
- We design a shared-representation layer based on the attention mechanism and gated mechanism to capture both the underlying structure of a cascade graph and node sequence in the diffusion process.
- We conduct extensive evaluations on several publicly available benchmark datasets, demonstrating that *DMT-LIC* can significantly improve the prediction accuracy and reduce the computational cost on both level tasks compared to state-of-the-art baselines.

2 PROBLEM AND METHODOLOGIES

In this paper, we formulate a deep multi-task model that jointly learns the micro-level task and macro-level task of information cascade modeling, specifically, we focus on the task of activation prediction and cascade size prediction.

Activation prediction Given a sequence of previously infected users $U^{t_i} = \{u_1, u_2, ..., u_i\}$ before the observation time t_i, the task of activation prediction aims to predict the next infected user u_{i+1} at time t_{i+1}.

Cascade Size Prediction Given a cascade graph G regarding some specific information (e.g. a post/news) within an observation time window t_i, we formulate this micro-level task as a regression problem that aims at predicting the incremental size ΔS after a fixed time interval Δt, where $\Delta S = \left|U^{t_i+\Delta t}\right| - \left|U^{t_i}\right|$.

We now proceed with elaborating on the proposed DMT-LIC model. The overall structure and the main components of DMT-LIC are depicted in Figure 1. The three basic components are: (1) Embedding layer – embedding the task-specific input into a low-dimensional space via various embedding methods to represent the cascade-level and user-level embedding, respectively. (2) Shared-representation layer – feeding the task-specific embedding to learn a shared latent representation via attention and gated mechanism. (3) Multi-Task layer – concatenating the shared-representation with task-specific embeddings, to form new representations for different tasks. At last, the representations are connected by different dense layers to predict the results, i.e., cascade size and next infected user. In the sequel, we present detailed discussions of the respective modules.

2.1 Task-specific Embedding Layer

We assume that the inputs of the two tasks are a cascade graph and user sequence in the cascading process. We employ a graph representation model and an RNN model to learn embeddings for these two inputs, respectively.

Cascade Graph embedding The cascade size prediction task is a macro-level problem, which takes a cascade graph G as input. Since the cascade graph is a directed acyclic graph and its adjacency matrix is not symmetric, we use a multi-layer Graph Attention Network (GAT [11]) with multi-head attention to model the cascade graph. The layer-wise propagation rule is $H^{(l+1)} = \sigma(\sum_{k=1}^{K} A^k W^k H^{(l)})$.

Here, W^k is a set of independent trainable weight matrices and K is the number of single attention. $\sigma(\cdot)$ denotes activation function, i.e., ReLU$(\cdot)$. $H^{(l)} \in \mathbb{R}^{N\times F}$ is the matrix of activations in the l^{th} layer, where N is the number of nodes in cascade graph, and F is the number of features.. The input of our fist layer is $H^0 = Adj + I_N$, where $Adj \in \mathbb{R}^{N\times N}$ is an adjacency matrix and I_N is an identity matrix. $A^k = [a_{ij}]^k_{N\times N}$ is attention matrix through a self-attention mechanism defined as follows:

$$a_{ij} = \frac{exp(LeakyReLU(c^T[Wh_i||Wh_j]))}{\sum_{k\in N} LeakyReLU(c^T[Wh_i||Wh_k])} \tag{1}$$

After the attention-based GCN layer, the cascade graph G is represented as a vector matrix $H^{cas} \in \mathbb{R}^{N\times d_{cas}}$.

Diffusion Process Embedding For the activation prediction, the input is an ordered user sequence with timestamps in a diffusion. We represent each user in the sequence via a one-hot vector $q \in \mathbb{R}^M$, where M denotes the total number of users in a dataset and all users are associated with a specific embedding matrix $E \in \mathbb{R}^{M\times D}$, where D is an adjustable dimension. E converts each user into its representation vector $x = qE$. Then we employ a bi-directional LSTM to model this diffusion process sequentially, in which a hidden state is used to memorize the diffusion history. At each step t_i, user embedding and previous hidden state are taken as inputs. Bi-directional LSTM computes the updated hidden state as follows: $\overleftrightarrow{h_i} = \text{Bi-LSTM}(x_i, h_i - 1), \overleftrightarrow{h_i} \in \mathbb{R}^{2d_{seq}}$. Thus, the user sequence is represented as $H^{user} \in \mathbb{R}^{N\times 2d_{seq}}$.

2.2 Shared-representation Layer

The task-specific embeddings are fed into the shared-representation layer, which includes a user importance scoring function and a shared gate.

User importance Learning Each row of H^{cas}, a user-level vector, can be treated as a structural diffusion context for each potentially infected user. We define a scoring function via an attention mechanism, which measures the importance of a user based on structural contexts – $A_{user} = \text{softmax}(\tanh(H^{cas}W_{attn} + b_{attn}) \cdot U_{attn})$,

where $W_{attn} \in \mathbb{R}^{d_{cas}\times d_{attn}}$, $b_{attn} \in \mathbb{R}^{d_{attn}}$ and $U_{attn} \in \mathbb{R}^{d_{attn}\times 1}$ are attention parameters, and $A_{user} \in \mathbb{R}^{N\times 1}$ is the user importance matrix.

Shared-Gate Given the user importance matrix A_{user} and user sequence embedding H^{user}, the new representation for each infected user can be calculated as $H^{new} = A_{user}H^{user}$. Inspired by the gated mechanisms used in LSTM and GRU, we design a novel shared-gate as shown in Figure 1, that takes H^{new}, H^{cas}, H^{user} as the input. The detailed process is described as follows:

$$\begin{aligned}
f_t &= \sigma\left(h^{new}_{t-1}W_f + h^{cas}_{t-1}U_f + h^{user}_{t-1}V_f + b_f\right)\\
r_t &= \sigma\left(h^{new}_{t-1}W_r + h^{cas}_{t-1}U_r + h^{user}_{t-1}V_r + b_r\right)\\
c_t &= f_t \odot c_{t-1} + (1 - f_t) \odot (h^{new}_{t-1}W_c + h^{cas}_{t-1}U_c + h^{user}_{t-1}V_c)\\
h_t &= r_t \odot \tanh c_t + (1 - r_t) \odot (h^{new}_{t-1}W_h + h^{cas}_{t-1}U_h + h^{user}_{t-1}V_h)
\end{aligned} \tag{2}$$

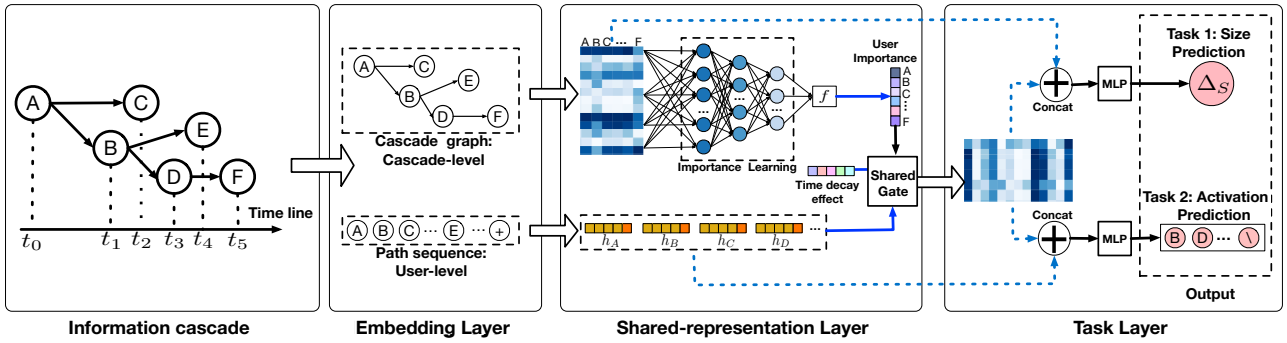


Figure 1: Overview of DMT-LIC.

where σ denotes the sigmoid function, $W_* \in \mathbb{R}^{d_{new} \times d_{share}}, U_* \in \mathbb{R}^{d_{cas} \times d_{share}}, V_* \in \mathbb{R}^{2d_{seq} \times d_{share}}$ and $b_f, b_r \in \mathbb{R}^{d_{share}}$. f_t is a forget gate aiming to forget the irrelevant part of previous information and update the cell state c_t. The reset gate r_t is used to control the influence of h_{t-1} and compute the output state h_t based on c_t and the linear combination of h_{t-1}^{new}, h_{t-1}^{cas} and h_{t-1}^{user}. After obtaining the output $H^{share} \in \mathbb{R}^{N \times d_{share}}$ from the shared gate, we calculate the shared representation using a weighted sum pooling mechanism with a non-parametric time decay function $f(T - t_i) = l$, if $t_{l-1} \leq T - t_i < t_l$, and $H^{share} = \sum \lambda_{f(T-t_i)} h_i$.

2.3 Multi-Task Layer

We concatenate the task-specific representation H^{cas} and H^{user} for each task with the shared-representation H^{share}, respectively, and ultimately feed into different output layers for prediction.

For the activation prediction task, our model predicts the next infection probability for each user: $\hat{p}(u_{i+1}|h^{user}, h^{share}) = \text{softmax}(\text{MLP}(\text{concat}(h^{user}, h^{share}))$, where the learning objective is to maximize the infection likelihood of all users in a diffusion sequence, i.e., $\hat{P}(U^{t_i}) = \prod_{i=1}^{l-1} p(u_{i+1}|h^{user}, h^{share})$. This task is trained by minimizing the cross-entropy loss between the predicted $\hat{P}$ and true probability P of sequence U^{t_i}:

$$\ell_1\left(\hat{P}, P\right) = \frac{1}{P}\sum_{i=1}^{P} cross\_entropy(\hat{P}, P) \tag{3}$$

While for the increment size prediction task, our goal is to predict the incremental cascade size for a fixed time interval in the future, which can be done by minimizing the following loss function:

$$\ell_2\left(\Delta S_i, \Delta\widetilde{S_i}\right) = \frac{1}{P}\sum_{i=1}^{P}\left(\Delta S_i - \Delta\widetilde{S_i}\right)^2 \tag{4}$$

where P is the information volume (e.g., the number of posts), $\Delta S_i = \text{MLP}(\text{concat}(h^{cas}, h^{share}))$ is the predicted incremental size for information p_i, and $\Delta\widetilde{S_i}$ is the ground truth.

Our overall loss function is $L = \gamma\ell_1 + (1-\gamma)\ell_2$, where $\gamma \in [0, 1]$ is a learning parameter balancing ℓ_1 and ℓ_2.

3 EXPERIMENTS

We now discuss our experiments and present the empirical evaluations for the following research-related questions:

- **RQ1** How does DMT-LIC perform compared with the state-of-the-art baselines on both tasks?
- **RQ2** Is the shared-representation layer helpful for learning a good representation for information cascades both structurally and temporally?

3.1 Experimental Settings

Datasets. To demonstrate the performance of DMT-LIC and the comparison with some existing methods, we conduct our experiments on two publicly available real-world datasets: Weibo [3] and APS [10]. The descriptive statistics are shown in Table 1.

Table 1: Descriptive statistics of two datasets.

	Weibo	APS
# Nodes	10,077	13,945
# Edges	11,956	15,508
# Cascades	306	509
Avg. cascade length	61.2	84.8

Baselines. We compare our proposed model with the following state-of-the-art baselines: EIC [2], DeepCas [6], CYAN-RNN [14], DeepHawkes [3], Topo-LSTM [12], and SNIDSA [15]. Note that the original application of Topo-LSTM, CYAN-RNN and SNIDSA is to predict node activations, and we replace the logistic classifier in these models with a diffusion size regressor to predict the size of cascades. In addition, for DeepCas and DeepHawkes, we replace the diffusion size regressor at the end of their process with a logistic classifier to predict node activation.

Evaluation Protocols. For the activation prediction, we cast this task as a retrieval problem. Hence, we use two widely ranking metrics for evaluation. They are Mean Reciprocal Rank (*MRR*) and Accuracy on top K (*A@K*). The larger values in *A@K* and *MRR* indicate the better performance. However, for the task of size prediction, we following existing works to choose standard evaluation metric *MSE* (mean square error) [3, 6, 12] – the smaller *MSE*, the better prediction performance.

Parameter Setups. The parameter settings are as follows. For all embedding-based and deep learning-based methods, we set the dimensionality of node embedding to 50, the hidden layer of each RNN to 32 units, and the hidden dimensions of the one-layer MLP

to 32, respectively. The learning rate for user embedding is 1×10^{-4} and 1×10^{-3} for other variables. The batch size in each iteration is 32 and the training will stop when the loss on the validation set does not decline for 10 consecutive iterations. Other parameters follow default settings in the corresponding papers.

3.2 Performance Comparison (RQ1)

Table 2 shows the results of the performance comparison to the existing state-of-the-art methods in information cascades modeling and prediction, from which we can clearly observe that our proposed DMT-LIC model performs the best across almost all metrics on the two datasets for both macro-level and micro-level tasks. The overall superiority of DMT-LIC over the baselines stems from the fact that it leverages multi-task learning.

Table 2: Performance comparison: DMT-LIC vs. baselines.

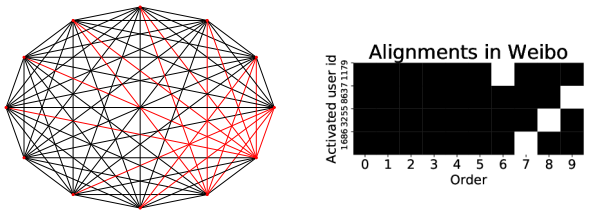
Datasets	Weibo				APS			
Metric / Model	MRR (%)	A@5 (%)	A@10 (%)	MSE	MRR (%)	A@5 (%)	A@10 (%)	MSE
EIC	53.27	51.26	58.94	—	63.15	70.23	76.48	—
DeepCas	80.12	71.56	81.22	1.017	81.35	80.21	86.54	0.873
DeepHawkes	83.21	77.22	84.15	0.328	84.57	79.53	86.47	0.276
CYAN-RNN	81.25	79.38	86.14	0.892	87.55	84.13	86.14	1.126
Topo-LSTM	84.23	81.37	94.77	0.411	87.64	86.31	94.72	0.386
SNIDSA	**90.15**	84.62	97.84	0.364	93.67	87.36	98.63	0.427
DMT-LIC	87.69	**86.73**	**98.72**	**0.196**	**94.53**	**89.15**	**98.75**	**0.177**

3.3 Model Analysis (RQ2)

To investigate the performance of the shared-representation layer in our model, we attempt to infer network structure and diffusion sequence using latent representations of nodes from this layer. Specifically, we conduct link prediction using latent learned node representations. As shown in Figure 2(a), the structure of the cascade graph is relatively well captured. For the cascade sequence, we first take the latent representations and perform dimension reduction with L_1 regularization for sparsity promotion, where the largest value in the reduced dimensional vector represents its activation time point. In Figure 2(b) the rows represent the reduced dimensional vectors for 4 randomly sampled users from one cascade on the Weibo dataset. The position with the largest value in the reduced dimension vector is highlighted in white, which is consistent with the true cascade sequence. Therefore, both results indicate that our shared-representation layer is able to capture both latent structure and sequence in cascades.

4 CONCLUSIONS

Existing research has investigated the information cascades problem from separate perspectives on micro and macro levels. In this work, we presented DMT-LIC – the first multi-task learning based approach for information cascade modeling, which is able to jointly optimize the two-level tasks. Specifically, we designed a shared representation layer with graph attention and a novel gated mechanism. The experimental results based on two real-world datasets demonstrate that DMT-LIC outperforms the state-of-the-art baselines on both tasks, and the shared-representation layer can well learn the latent representations that reflect both structural and sequential patterns. This, in turn, indicates a promising direction that training and optimizing cascade-related tasks with multi-task learning. In the future, we plan to extend DMT-LIC to more specific applications (e.g., fake news detection) and investigate the problem of correlating cascade from multiple contexts and spatio-temporal scales (e.g., spread of viral marketing followed by tweets on social media) .



(a) Network structure inference. (b) Cascade sequence learning.

Figure 2: Visualization of network inference and an example of user activation in one cascade on the Weibo dataset. (a) Edges in black are the correct inferred edges, while edges highlighted in red are either missed or predicted incorrectly. (b) Each row is the vector from shared-representation layer after dimension reduction. The white cell refers to the largest value in the vector, corresponding to the position in the cascade sequence.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (Grant No.61602097 and No.61472064), NSF grants III 1213038 and CNS 1646107, and ONR grant N00014-14-10215.

REFERENCES

[1] Peng Bao, Hua Wei Shen, Junming Huang, and Xue Qi Cheng. 2013. Popularity prediction in microblogging network: a case study on sina weibo. In *WWW*.
[2] Simon Bourigault, Sylvain Lamprier, and Patrick Gallinari. 2016. Representation learning for information diffusion through social networks: an embedded cascade model. In *WSDM*.
[3] Qi Cao, Huawei Shen, Keting Cen, Wentao Ouyang, and Xueqi Cheng. 2017. DeepHawkes: Bridging the Gap between Prediction and Understanding of Information Cascades. In *CIKM*.
[4] Justin Cheng, Lada Adamic, P. Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted?. In *WWW*.
[5] Maximilian Jenders, Gjergji Kasneci, and Felix Naumann. 2013. Analyzing and predicting viral tweets. In *WWW*.
[6] Cheng Li, Jiaqi Ma, Xiaoxiao Guo, and Qiaozhu Mei. 2017. DeepCas: An End-to-end Predictor of Information Cascades. In *WWW*.
[7] Swapnil Mishra, Marian Andrei Rizoiu, and Lexing Xie. 2016. Feature Driven and Point Process Approaches for Popularity Prediction. In *CIKM*.
[8] Naoto Ohsaka, Tomohiro Sonobe, Sumio Fujita, and Ken-ichi Kawarabayashi. 2017. Coarsening Massive Influence Networks for Scalable Diffusion Analysis. In *SIGMOD*.
[9] Daniel M. Romero, Chenhao Tan, and Johan Ugander. 2013. On the Interplay between Social and Topical Structure. In *AAAI*.
[10] Hua-Wei Shen, Dashun Wang, Chaoming Song, and Albert-László Barabási. 2014. Modeling and Predicting Popularity Dynamics via Reinforced Poisson Processes.. In *AAAI*.
[11] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph Attention Networks. In *ICLR*.
[12] Jia Wang, Vincent W. Zheng, Zemin Liu, and Chen Chuan Chang. 2017. Topological Recurrent Neural Network for Diffusion Prediction. In *ICDM*.
[13] Yongqing Wang, Huawei Shen, Shenghua Liu, and Xueqi Cheng. 2015. Learning user-specific latent influence and susceptibility from information cascades. In *AAAI*.
[14] Yongqing Wang, Huawei Shen, Shenghua Liu, Jinhua Gao, and Xueqi Cheng. 2017. Cascade Dynamics Modeling with Attention-based Recurrent Neural Network. In *IJCAI*.
[15] Zhitao Wang, Chengyao Chen, and Wenjie Li. 2018. A Sequential Neural Information Diffusion Model with Structure Attention. In *CIKM*.
[16] Lilian Weng, Filippo Menczer, and Yong Yeol Ahn. 2014. Predicting Successful Memes using Network and Community Structure. In *AAAI*.