THGFormer: Time-Aware Hypergraph Learning for Multimodal Social Media Popularity Prediction (Student Abstract)

Jienan Zhang[1], Jie Liu[1], Zhangtao Cheng[1,4], Xovee Xu[1,4], Fang Liu[* 2], Ting Zhong[1,4], Kunpeng Zhang[3]

[1]University of Electronic Science and Technology of China, Chengdu, Sichuan 610054, China
[2]Civil Aviation Flight University of China, Guanghan, Sichuan 618307, China
[3]University of Maryland, College park, MD 20742, USA
[4]Kashi Institute of Electronics and Information Industry, Kashgar 844000, China
{eroicazhang, uestc_liujie, zhangtao.cheng}@outlook.com, xovee@live.com, fangliu@cafuc.edu.cn, zhongting@uestc.edu.cn, kpzhang@umd.edu

Abstract

Social media popularity prediction of multimodal user-generated content (UGC) is a crucial task for many real-world applications. However, existing efforts are often limited by missing inter-instance correlations and UGC temporal patterns. To address these issues, we propose a novel time-aware hypergraph Transformer framework, THGFormer. It fully represents inter-instance and intra-instance relations by hypergraphs, captures the temporal dependencies with a time encoder, and enhances UGC's representations via a neighborhood knowledge aggregation.

Extensive experiments conducted on two real-world datasets demonstrate that THGFormer outperforms state-of-the-art popularity prediction models across several settings.

Introduction

Multimodal social media popularity prediction (MSMPP) aims to infer the future number of interactions between users and UGCs via learning and aggregating multimodal contents. It is beneficial for aiding users in sifting from information overload and improving various applications from recommendation to rumor detection. Existing works on MSMPP can be summarized into two categories: (1) Feature-based methods (Khosla, Das Sarma, and Hamid 2014; Lai, Zhang, and Zhang 2020) emphasize designing and incorporating hand-crafted UGC features; (2) Deep learning-based methods exploit end-to-end frameworks to capture more comprehensive multimodal representations (Zhang et al. 2018; Wang et al. 2023).

Challenges. Despite their successes, the usability of current works is limited by the following aspects: **(1)** Existing works treat a single UGC independently to learn its representation, which fail to consider the inter-relations among different UGCs. The auxiliary information existing in related UGCs are also ignored to assist UGC reasoning. For instance, different UGCs posted by the same user might be interacted with similar user groups and produce similar UGC popularity. **(2)** Existing works neglect the temporal dependencies among UGCs. For example, before and after the Thanksgiving Day, the popularity of pumpkin content rises and decays in a few days.

Present Work. To tackle above issues, we propose a **T**ime-aware **H**yper**G**raph trans**Former** (**THGFomer**) framework. First, to preserve temporal correlations among different UGCs, we retrieve the top $\mathcal{K}$ relevant instances from a time-centered perspective. Subsequently, using UGC attributes, we connect all relevant instances to construct a hypergraph of the target UGC. Second, we design a time-aware hypergraph transformer to capture the intra- and inter-modal correlations, and meanwhile employ a time encoder to inject the temporal information into the information mixture process. This dual focus ensures fine-grained and aligned UGC representations that are crucial for multimodal popularity prediction. Finally, we use a 2-layer feed-forward neural network for the target UGC's popularity prediction.

Methodology

Hypergraph Construction. We design two steps to build a hypergraph for MSMPP. First, given the target UGC $\mathcal{C}_t$ and a post time t, we retrieve the $\mathcal{K}$ most related instances from user-post sequence $\mathbf{S}$ before time t and construct a temporal context sequence $\mathbf{P}_t$. $\mathbf{P}_t$ can be expressed by a triple sequence: $\mathbf{P}_t = \{< u_1, \mathcal{C}_1, t_{p_1} >, \ldots, < u_\mathcal{K}, \mathcal{C}_\mathcal{K}, t_{p_\mathcal{K}} >\}$, $t_{p_1} \leq t_{p_2} \cdots \leq t_{p_\mathcal{K}} \leq t$, where $\mathcal{K}$ denotes the instance count. Second, after obtaining $\mathbf{P}_t$, the resulting instances set $\mathbf{P}_t$ is transformed into an adaptive hypergraph $\mathcal{G}_t = \{\mathcal{V}_t, \mathcal{E}_t\}$ of the target instance $\mathcal{C}_t$, where each data instance forms a node, i.e., $\mathcal{V}_t = \{\mathcal{C}_t, \mathcal{C}_1, \ldots, \mathcal{C}_\mathcal{K}\}$. The attributes of $\mathcal{C}_t$ (i.e., posted user, category, topic) constructs a hyperedge that represents the inter-instance relations.

Time-aware Hypergraph Transformer. For MSMPP, the main challenge is how to jointly incorporate the temporal information, and model the intra- and inter-modal correlations. We design a time-aware hypergraph Transformer, which integrates a time encoder to inject temporal information into the multimodal mixture process of the hypergraph Transformer. First, the time encoder (T-encoder) (Xu et al. 2019) is used to model the continuous temporal information, which maps scalar timestamps into d_T-dimensional vector space. The process is summarized as: $\Psi(t) \rightarrow$

*corresponding author.

$\sqrt{\frac{1}{d_T}}\left[\cos\left(\omega_1 t\right), \ldots, \cos\left(\omega_{d_T} t\right), \sin\left(\omega_{d_T} t\right)\right]$, where $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_{d_T}]^\top$ are learnable parameters. Second, the hypergraph Transformer learns intra- and inter-modal correlations via the multimodal mixture process, including *node-to-hyperedge* and *hyperedge-to-node*. For the *node-to-hyperedge*, we distill visual (textual) information from the visual (textual) hypergraph to the textual (visual) hypergraph to reduce the modality heterogeneity. The process from nodes to k-th hyperedge is defined as: $\tilde{\mathbf{e}}^v_{k,h} = \sum_{i=1}^{\mathcal{I}} \mathbf{v}^v_{i,h} \mathbf{k}^{v\top}_{i,h} \mathbf{q}^v_{k,h}$, $\tilde{\mathbf{e}}^t_{k,h} = \sum_{i=1}^{\mathcal{I}} \mathbf{v}^t_{i,h} \mathbf{k}^{t\top}_{i,h} \mathbf{q}^t_{k,h}$, where v and t denote the visual and textual modality, respectively. $\tilde{\mathbf{e}}_{k,h} \in \mathbb{R}^{d/\mathcal{H}}$ denotes k-th hyperedge features. Here $\mathcal{H}$ is the number of attention heads. $\mathbf{q}_{k,h}, \mathbf{k}_{i,h}, \mathbf{v}_{i,h} \in \mathbb{R}^{d/\mathcal{H}}$ present the query, key and value vectors for node i and hyperedge k, which is computed through linear transformations and slicing: $\mathbf{q}_{k,h} = \mathbf{E}_{k,p_h:p_{h-1}}$, $\mathbf{k}_{i,h} = \mathbf{K}_{p_{h-1}:p_h}(\mathbf{z}_i \oplus \Psi(t_i))$, $\mathbf{v}_{i,h} = \mathbf{V}_{p_{h-1}:p_h}(\mathbf{z}_i \oplus \Psi(t_i))$. $\mathbf{E} \in \mathbb{R}^{K\times d}$ is the embedding matrix for K hyperedges. $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{(d+d_\mathcal{T})\times d}$ represent the key and the value transformation of all $\mathcal{H}$ heads. p_{h-1} and p_h denote the start and end indices of the h-th slice.

To mine inter-modal correlations and reduce the effect of modality heterogeneity, we implement a gating mechanism to control the information flow from one modality to another. The cross-modal interactive process can be summarised as: $\tilde{\mathbf{e}}^{v\to t}_{k,h} = \sum_{i=1}^{\mathcal{I}} \mathbf{v}^v_{i,h} (\mathbf{k}^v_{i,h})^\top \mathbf{q}^t_{k,h}$. Specifically, the gating process can be summarized as:

$$\mathbf{e}^t_{k,h} = (1-\lambda) \sum_{i=1}^{\mathcal{I}} \mathbf{v}^t_{i,h} \mathbf{k}^{t\top}_{i,h} \mathbf{q}^t_{k,h} + \lambda \sum_{i=1}^{\mathcal{I}} \mathbf{v}^v_{i,h} \mathbf{k}^{v\top}_{i,h} \mathbf{q}^t_{k,h},$$

where λ denotes the scalar for the sum of normalized attention weights on the visual key and value vectors:

$$\lambda = \frac{\sum_{i=1}^{\mathcal{I}} \mathbf{k}^{v\top}_{i,h} \mathbf{q}^t_{k,h}}{\sum_{i=1}^{\mathcal{I}} \mathbf{k}^{t\top}_{i,h} \mathbf{q}^t_{k,h} + \sum_{i=1}^{\mathcal{I}} \mathbf{k}^{v\top}_{i,h} \mathbf{q}^t_{k,h}}. \quad (1)$$

Afterward, we employ the concatenation operation to fuse $\mathcal{H}$ head-specific representations of the hyperedge k: $\mathbf{e}^t_k = \|^H_{h=1} \mathbf{e}^t_{k,h}$. Analogously, $\mathbf{e}^t_{k,h}$ can be replaced with $\mathbf{e}^v_{k,h}$ to obtain $\mathbf{e}^v_k$. In the process of *hyperedge-to-node*, we propagate the information from hyperedges to nodes through a similar but reverse process. Finally, we will obtain the node i's representations $\mathbf{z}^v_i$ and $\mathbf{z}^t_i$.

Prediction. The final popularity prediction is made by a 2-layer feed-forward neural network:

$$\hat{y} = \mathbf{W}^2_P \left(\mathrm{ReLU}\left(\mathbf{W}^1_P(\mathbf{z}^t_i \oplus \mathbf{z}^v_i) + \mathbf{b}^1_P\right)\right) + \mathbf{b}^2_P, \quad (2)$$

where $\mathbf{W}^1_P$, $\mathbf{W}^2_P$, $\mathbf{b}^1_P$, and $\mathbf{b}^2_P$ are the parameters of the prediction network. $\oplus$ is the concatenation operation. In addition, we use mean square error (MSE) as the optimization loss to train the model's parameters.

Experiments

Datasets and Baselines. We select two public datasets, i.e., SMPD and ICIP, and compare our model THGFormer with five baselines: **SVR** (Khosla, Das Sarma, and Hamid 2014), **Hyfea** (Lai, Zhang, and Zhang 2020), **DTCN** (Wu et al. 2017), **UHAN** (Zhang et al. 2018) and **MHF** (Wang et al. 2023). We select three evaluation metrics: Spearman ranking correlation (SRC), mean absolute error (MAE), and MSE.

Model	SMPD			ICIP		
	MSE	MAE	SRC	MSE	MAE	SRC
SVR	4.9886	1.6749	0.5312	2.0942	1.0552	0.3723
Hyfea	4.9297	1.6623	0.5518	1.9813	0.9935	0.3641
DTCN	4.2523	1.4998	0.5432	2.8361	1.3432	0.3893
UHAN	<u>3.8471</u>	<u>1.4833</u>	<u>0.5541</u>	2.7492	1.2824	0.3981
MHF	3.9297	1.5433	0.5419	<u>1.8736</u>	<u>0.9132</u>	<u>0.4041</u>
Ours	**3.5673**	**1.4138**	**0.5741**	**1.5615**	**0.8326**	**0.4759**
(improves)	7.27%	4.69%	3.48%	16.66%	8.83%	17.77%

Table 1: Performance comparison on two real-world datasets. The best results are in bold font and the second underlined. Lower values of MSE and MAE, and higher values of SRC, indicate better performance.

Main Results. The evaluation results are shown in Table 1. Our proposed THGFormer model exhibits significant improvements in MSMPP performance when compared to the baselines. Moreover, these results validate our motivation for utilizing the neighborhood knowledge for assisting UGC reasoning. The enhanced model performance can be attributed to the exploitation of both intra- and inter-modal correlations through the hypergraph Transformer approach.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Grant No.62176043 and No.62072077) and Kashgar Science and Technology Bureau (Grant No.KS2023025).

References

Khosla, A.; Das Sarma, A.; and Hamid, R. 2014. What makes an image popular? In *WWW*, 867–876.

Lai, X.; Zhang, Y.; and Zhang, W. 2020. Hyfea: winning solution to social media popularity prediction for multimedia grand challenge 2020. In *ACM MM*, 4565–4569.

Wang, J.; Yang, S.; Zhao, H.; and Yang, Y. 2023. Social media popularity prediction with multimodal hierarchical fusion model. *Computer Speech & Language*, 80: 101490.

Wu, B.; Cheng, W.-H.; Zhang, Y.; Huang, Q.; Li, J.; and Mei, T. 2017. Sequential prediction of social media popularity with deep temporal context networks. In *IJCAI*, 3062–3068.

Xu, D.; Ruan, C.; Korpeoglu, E.; Kumar, S.; and Achan, K. 2019. Self-attention with functional time representation learning. *NeurIPS*.

Zhang, W.; Wang, W.; Wang, J.; and Zha, H. 2018. User-guided hierarchical attention network for multi-modal social image popularity prediction. In *WWW*, 1277–1286.