Curriculum Learning for Wide Multimedia-Based Transformer with Graph Target Detection

Weilong Chen
chenweilong1995@std.uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu
WeChat, Tencent, Beijing

Feng Hong
Chenghao Huang
University of Electronic Science and Technology of China
Chengdu

Shaoliang Zhang, Rui Wang
Ruobing Xie, Feng Xia
Leyu Lin
WeChat, Tencent,Beijing

Yanru Zhang, Yan Wang*
yanruzhang@uestc.edu.cn
yanbo1990@uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu

ABSTRACT

The social media prediction task is aiming at predicting content popularity which includes social multimedia data such as photos, videos, and news. The task can not only help make better decisions for recommendation, but also reveals the public attention from evolutionary social systems. In this paper, we propose a novel approach named curriculum learning for wide multimedia-based transformer with graph target detection (**CL-WMTG**). The curriculum learning is designed for the transformer to improve the efficiency of model convergence. The mechanism of wide multimedia-based transformer is to make the model capable of learning cross information from text, pictures and other features(e.g. categories, location). Moreover, the graph target detection part can extract different features in the picture by pretrained model and reconstruct the features with a homogeneous graph network. We achieved third place in the SMP Challenge 2020.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; **Supervised learning by regression**; *Object detection.*

KEYWORDS

Multimedia; Graph Network; Curriculum Learning; Transformer

ACM Reference Format:
Weilong Chen, Feng Hong, Chenghao Huang, Shaoliang Zhang, Rui Wang, Ruobing Xie, Feng Xia, Leyu Lin, and Yanru Zhang, Yan Wang. 2020. Curriculum Learning for Wide Multimedia-Based Transformer with Graph Target Detection. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3394171.3416275

1 INTRODUCTION

In the past decade, social media has become one of the increasingly popular components of our everyday lives. Social media platforms, such as Twitter, Facebook and Flickr, have provided a context where people can share the information, ideas and other forms of expression across the world. However, among the billions of posts, only a few of them have become popular, which raises the question of how to predict if one post can go virus online. In the social media prediction (SMP) task of ACM multimedia 2020 challenge, the goal is to predict the popularity of each posts and the data is provided by Flickr[18–20].

Features are important on social media prediction. User profile reflects the user's popularity on social media, including their posted images, timezone and whether it is belong to pro number. Categories and subcategories are also important features to the popularity. In this competition, there are 11 classes of the first level category, 77 classes in the second level category or subcategory, 668 different classes in third level category or concepts. We propose a novel approach named curriculum learning for wide multimedia-based transformer with graph target detection (CL-WMTG). Besides, we also use the external data as pretrained DenseNet[9] and Bottom-Up and Top-Down object detection model[1] for extracting image features. We use relatively simple features for prediction, however, some sophisticated approaches are adopted for social media prediction as well.

The main contributions of this work are as follows:

1. The graph part with multimedia helps improve the understanding ability of the model.

2. We are the first one to propose a curriculum learning way to enhance the performance of the multimedia-based transformer.

The rest of this paper is organized as follows. Section 2 includes related works about prediction methods of social network and deep learning models. Section 3 shows the method in detail. The experiment about our method has been conducted and presented in Section 4. Finally, a summary is presented in Section 5.

[*] All the corresponding to Yan Wang.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3416275

2 RELATED WORK

In recent years, a variety of studies have been conducted on how to make predictions based on the data from Flickr social media, such as predicting the popularity. In this paper, we mainly focus on the popularity-related prediction tasks on social media. According to the main forms of the dataset, which using Flickr social media data in text and image forms, we design our model by investigating some popular deep learning models and related methods.

2.1 Conventional prediction methods

There are a lot of conventional methods which can be used on social media prediction. For example, regression methods are used for predicting some trends such as popularity by analyzing relationship between the dependent variable and one or more independent variables. Besides, Bayes classifiers are used to distinguish features. K-nearest neighbor classifiers perform a function by clustering the similar features. For the sake of briefness, other frequently used methods for social media prediction such as decision tree will not be stated here in details[24].

2.2 Popularity-related prediction methods

The prediction of popularity is the key task in the realm of social media analytic. The data provided by Flickr offers photo-sharing services with text descriptions, which has drawn many scholars' research interesting, such as event detection [22, 23] and popularity prediction. Kim et al. proposed a joint photo stream and blog post framework based on support vector machine[10], which improves the performance on exploration and summarization tasks after using both posted texts and photo streams.

Noisy and a small amount of data are another two issues, which may weaken the performance of prediction models. To overcome these difficulties, Chen et al. proposed a popularity prediction model which incorporated attention mechanism to focus on more informative parts and suppress noisy [4]. Gayberi et al. combined various types of features such as enriched user and post, using more visual features of images to produce a significantly larger dataset compared to previous studies. [5]

Besides,to make the sequential prediction of popularity, Wu et al. proposed a novel prediction framework called deep temporal context networks (DTCN) by incorporating both temporal context and temporal attention into account.[19]

2.3 Deep learning models

Recently, deep learning has become a frequently-used technique to address complicated tasks in a variety of areas. For popularity prediction, recurrent neural networks (RNN) can be utilized to capture temporal dependence and make more precise predictions by taking advantage of its modeling in sequence data[14]. In addition, adopting neural networks as a transformer to leverage various features is also an effective approach for popularity prediction, including cascade path[8], cascade graph[3], and multi-modality information[19]. In addition, the transformer can also be used to extract the sequential features[21].

It's worth mentioning that the emergence and development of convolution neural networks(CNN) have promoted image recognition accuracy for popularity prediction on social media pictures. Krizhevsky et al.[11] trained a deep CNN to classify the 1.2 million high-resolution images in ImageNet, one of the most popular and largest database for image recognition, and perform well by reducing the error rate to a relatively low level. Furthermore, Szegedy designed a deep CNN architecture named as Inception[16], which achieved a higher accuracy than the Regions with convolutional neural networks (R-CNN) proposed by Girshick et al. for detection and classification tasks in ImageNet.

3 METHODOLOGY

3.1 Graph Aggregation

We can obtain the image features by using the target detection pretrained model [1]. Using the extracted features is possible to be similar from an object image to the another image. However, even the same objects in different photos have complete different meanings. Since the context will affect the meaning associated with object. The meaning of the car in the picture which contains trees and river can be totally different from the picture which contains city bridge and buildings. Effectively illustrating different meanings of an object is the main challenge in this work.

A graph-based structure is designed for learning the relationship between an object and surroundings, which can update the features during learning process. Let $F \in \mathbb{R}^{K \times d}$ denote K objects which have d dimensional feature. The object image feature is defined as $F_o = \{\mathbf{f}_1^o, \mathbf{f}_2^o, ..., \mathbf{f}_{\mathbf{k_o}}^o\}$. Graph weight $G \in \mathbb{R}^{K \times K}$ shows the relation matrix between K objects. The graph is a fully connected graph, showing the global relationship among different relations. The graph weight G is defined as:

$$G = \phi(F_o) \cdot \psi(F_o)^T \tag{1}$$

$$\phi(F_o) = F_o \cdot W_i + b_i \tag{2}$$

$$\psi(F_o) = F_o \cdot W_j + b_j \tag{3}$$

Consider $W_i, W_j \in \mathbb{R}^{d \times d}$ and the $b_i, b_j \in \mathbb{R}^d$ as graph pre-parameters. With the graph weights G for enhancing the features by aggregating related objects. The enhanced feature is defined as $\hat{F}$:

$$\hat{F_o} = G \cdot F_o \cdot W_l + b_l, \tag{4}$$

where $W_l \in \mathbb{R}^{d \times d}$ and $b_l \in \mathbb{R}^d$. The object image feature $\hat{F_o}$ not only contains its own information, but also the sourroudings information. The graph-based combination of the relation features can explain the difference of an object in different pictures.

3.2 Wide Multimedia-based Transformer

The structure of the transformer is widely known in concept among the field of natural language processing. Transformer [17] is the first transduction model which entirely relies on self-attention to compute representations of its input and output without using sequence aligned RNNs or convolution. The multi-head attention structure in the transformer can model the correlations between different feature fields and expand the model's ability to focus on different cross attention fields. The feed forward neutral network with residual structure can map the result of attention to a more complex feature. It is just as two convolutions with kernel size 1. The ReLU active function can give the ability of non-linear transformation to the multi-head attention. Since the dimension of the

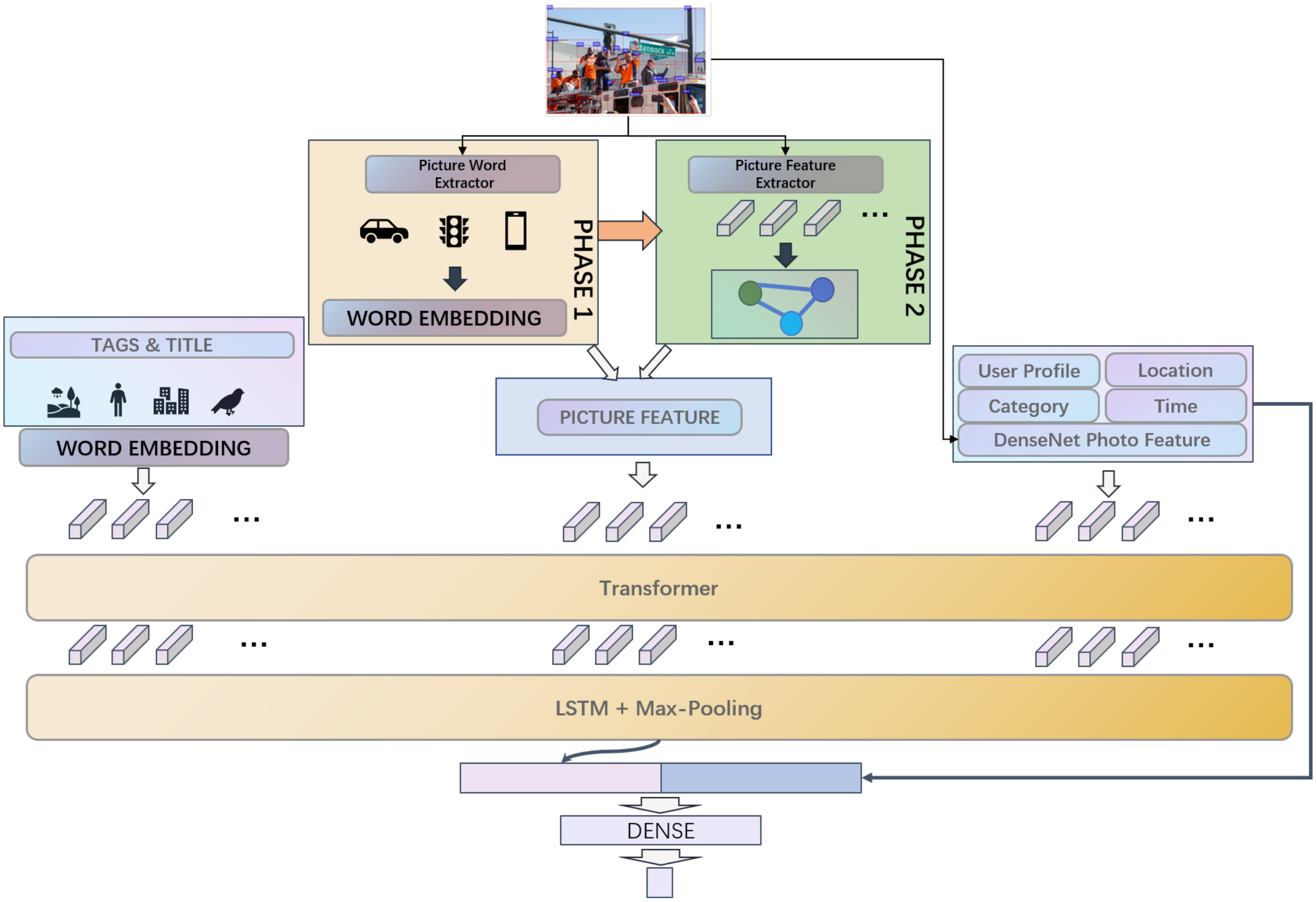

Figure 1: The structure of the Curriculum Learning for Wide Multimedia-Based Transformer with Graph Target Detection

feature should be the same when input to the transformer, the output of the graph-based text feature $\hat{F}_t$ should be the same as the text embedding dimension. We not only take the text and graph information into consideration but also the continuous features and categorical features. These features can be represented into low-dimensional spaces (e.g., word embeddings). Specifically, we represent all features with a low-dimensional vector, i.e.,

$$\mathbf{f_i^{\mathit{l}}} = \mathbf{embedding\_matrix}^{l}(\mathbf{x_i^{\mathit{l}}}), \tag{5}$$

where $\mathbf{f_i^{\mathit{l}}} \in \mathbb{R}^d$ is the representation of the $\mathbf{x_i}^l$, $\mathbf{embedding\_matrix}^l$ is an embedding matrix for field l, and $\mathbf{x_i^{\mathit{l}}}$ is the value in the field. We concatenate all of the field feature $\{\mathbf{f^1}, \mathbf{f^2}, ..., \mathbf{f^n}\}$ and stack all the features as $\mathbf{F} = \{\mathbf{f_1^{\mathit{t}}}, \mathbf{f_2^{\mathit{t}}}, ..., \mathbf{f_{k_t}^{\mathit{t}}}, \mathbf{\hat{f}_1^{\mathit{o}}}, \mathbf{\hat{f}_2^{\mathit{o}}}, ..., \mathbf{\hat{f}_{k_o}^{\mathit{o}}}, \mathbf{f^1}, \mathbf{f^2}, ..., \mathbf{f^n}\}$, where $\mathbf{f}^t$ is the word embedding of the tag, $\mathbf{\hat{f}}^o$ is the graph-aggregated object image features. The length of the tags is $\mathbf{k_t}$ and the number of the objects is $\mathbf{k_o}$. The transformer can form high-order features and determine which feature combinations are meaningful.

$$\hat{\mathbf{F}} = \mathrm{Transformer}\left(\mathbf{F}\right), \hat{\mathbf{F}} \in \mathbb{R}^{n \times d}. \tag{6}$$

The multimedia-based transformer combines the different multimedia information and form the high-order features. Moreover, we also design a wide part which directly input the $\{\mathbf{x}^l\}$ to the output dense layer. It can help the model to learn the low-order features, in order to prevent some low-dimensional information from being lost in the high-dimensional transformation. We also use the $LSTM$[6] to aggregate the output of transformer. The output is defined as:

$$popularity = W\{max\_pooling(LSTM(\hat{\mathbf{F}})), \mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^l\} + b. \tag{7}$$

3.3 Curriculum Learning Transformer

Bengio et al. proposed a new learning paradigm called curriculum learning (CL), in which a model is learned by gradually including from easy to complex samples during training process in order to increase the entropy of training samples[2]. We found that during the training processing of the model, it is difficult for wide multimedia-based transformer to learn the features from multimedia. However, the tag and object classification are all belong to text filed, and their spaces of feature are same. According to the tag and object classification, we design a curriculum learning phase where the classification and image feature of the object share the same parameters of the transformer. Compared with the object image features, the object classification words possess the same feature space to the tag words but contain less information.

Table 1: The abalation study on our method and report results. The results of our model compared with VL-BERT[15] and SCAN[12]

Method	Hyperparameter	Wide Feature	Offline SR/MSE	Online SR/MSE
SCAN	–	False	0.53/1.84	–
VL-BERT	head=1,layer=1	False	0.53/1.82	–
CLSTM	–	False	0.54/1.80	–
DTCN	–	False	0.63/1.52	0.62/1.50
CatBoost	–	True	0.62/1.48	0.62/1.56
MT	head=1,layer=1	False	0.56/1.61	0.56/1.60
MT	head=3,layer=3	False	0.55/1.62	0.55/1.63
MT	head=6,layer=12	False	0.52/1.68	–
MTG	head=1,layer=1	False	0.57/1.59	0.58/1.58
WMTG	head=1,layer=1	True	0.60/1.51	0.60/1.51
CL-WMT	head=1,layer=1	True	0.64/1.44	0.63/1.43
CL-WMTG	**head=1,layer=1**	**True**	**0.65/1.41**	**0.65/1.39**

We have two training phases. In the first a few epochs, we replace the object class with object image features, trying to learn the combination of the image features and the tag features. Next, we restore the object image features and continue to train the model. In the first phase, our input feature is:

$$\mathbf{F}^{phase1} = \{\mathbf{f}_1^t, \mathbf{f}_2^t, ..., \mathbf{f}_{\mathbf{k}_t}^t, \mathbf{f}_1^{ow}, \mathbf{f}_2^{ow}, ..., \mathbf{f}_{\mathbf{k}_o}^{ow}, \mathbf{f}^1, \mathbf{f}^2, ..., \mathbf{f}^\mathbf{n}\}. \tag{8}$$

where $\mathbf{f}^{ow}$ is the embedding of the object class words. At the second phase, the input feature is the same as **F** which means:

$$\mathbf{F}^{phase2} = \mathbf{F} \tag{9}$$

Our training process can be concluded as a curriculum learning which learn the easier first and learn the harder next. Although the effectiveness of the curriculum learning decays as the the number of training epochs raises, the score obtained is better than the model only employing with single phase

4 EXPERIMENT

We explore the spearman's rho (SR) and the mean absolute error (MSE) to measure our model performance. Here we compare the performance of our method with baselines. The results are listed in Table 1. According to the results from top to bottom, we have the following observation.

Firstly, we tried some methods like CLSTM[7] and DTCN[19] as baselines. The proposed multimedia-based transformer (MT) method consistently outperforms the VL-BERT[15] and the SCAN[12]. We found the less parameters the model has, the better the model performs. Only one head and one layer of the transformer mechanism can obtain higher SR than the three head and three layer. We attribute this phenomenon to the small amount of data. The bigger parameters can be learned better with bigger data.

Secondly, we analysis the graph part and wide part. We gain a relative improvements of about 3% over multimedia-based transformer with graph target detection (MTG) and 13% over wide multimedia-based transformer with graph target detection (WMTG). The wide part obtain the most benefit. We also try to put features of wide part into CatBoost[13] and also got a good score. The features are informative and wide part with transformer can exploit them deeply.

At last, the curriculum learning was adopted to the WMTG. The curriculum learning for wide multimedia-based transformer with graph target detection(CL-WMTG) obtain the best performance over all models. Curriculum learning solve the problem that the model is difficult to converge. Meanwhile, curriculum learning help the model learn better than directly learning the high-ordered feature between text and image.

5 CONCLUSION AND FEATURE WORK

The WMTG method proposed by us can significantly improve the accuracy of the social media prediction. By adding the curriculum learning, the CL-WMTG model learns better with the multimedia information, and has brought us to won the third place on SMP 2020 challenge. In the future, we will try to build more effective heterogeneous graph with both image objects and tags. Meanwhile, we would like to find a more useful way to learn the cross information.

ACKNOWLEDGMENTS

This work is supported by UESTC Funds for Outstanding Talents A1098531023601183.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*. 41–48.

[3] Qi Cao, Huawei Shen, Cen Keting, Wentao Ouyang, and Xueqi Cheng. 2017. DeepHawkes: Bridging the Gap between Prediction and Understanding of Information Cascades. 1149–1158. https://doi.org/10.1145/3132847.3132973

[4] Guandan Chen, Qingchao Kong, Nan Xu, and Wenji Mao. 2019. NPP: A neural popularity prediction model for social media content. *Neurocomputing* 333 (2019), 221–230.

[5] Mehmetcan Gayberi and Sule Gunduz Oguducu. 2019. Popularity Prediction of Posts in Social Networks Based on User, Post and Image Features. In *Proceedings of the 11th International Conference on Management of Digital EcoSystems*. 9–15.

[6] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with LSTM. (1999).

[7] Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry Heck. 2016. Contextual lstm (clstm) models for large scale nlp tasks. *arXiv preprint arXiv:1602.06291* (2016).

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In

Proceedings of the IEEE conference on computer vision and pattern recognition. 580–587.

[9] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. 2014. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869* (2014).

[10] Gunhee Kim, Seungwhan Moon, and Leonid Sigal. 2015. Joint photo stream and blog post summarization and exploration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3081–3089.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[12] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–216.

[13] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. In *Advances in neural information processing systems*. 6638–6648.

[14] Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison W Cottrell. 2017. A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction. In *IJCAI*.

[15] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *International Conference on Learning Representations*.

[16] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

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

[18] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, and Tao Mei. 2016. Time Matters: Multi-scale Temporalization of Social Media Popularity. In *Proceedings of the 2016 ACM on Multimedia Conference (ACM MM)* (Amsterdam, The Netherlands).

[19] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, Huang Qiushi, Li Jintao, and Tao Mei. 2017. Sequential Prediction of Social Media Popularity with Deep Temporal Context Networks. In *International Joint Conference on Artificial Intelligence (IJCAI)* (Melbourne, Australia).

[20] Bo Wu, Tao Mei, Wen-Huang Cheng, and Yongdong Zhang. 2016. Unfolding Temporal Dynamics: Predicting Social Media Popularity Using Multi-scale Temporal Decomposition. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)* (Phoenix, Arizona).

[21] Qidi Xu, Haocheng Xu, Weilong Chen, Chaojun Han, Haoyang Li, Wenxin Tan, Fumin Shen, and Heng Tao Shen. 2019. Time-aware Session Embedding for Click-Through-Rate Prediction. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2617–2621.

[22] Zhenguo Yang, Qing Li, Wenyin Liu, Yun Ma, and Min Cheng. 2017. Dual graph regularized NMF model for social event detection from Flickr data. *World Wide Web* 20, 5 (2017), 995–1015.

[23] Zhenguo Yang, Qing Li, Zheng Lu, Yun Ma, Zhiguo Gong, and Wenyin Liu. 2017. Dual structure constrained multimodal feature coding for social event detection from flickr data. *ACM Transactions on Internet Technology (TOIT)* 17, 2 (2017), 1–20.

[24] Sheng Yu and Subhash Kak. [n.d.]. A Survey of Prediction Using Social Media. ([n. d.]).