Predicting Micro-video Popularity via Multi-modal Retrieval Augmentation

Ting Zhong
zhongting@uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
Kash Institute of Electronics and Information Industry
Kashgar, Xinjiang, China

Jian Lang
Yifan Zhang
Zhangtao Cheng*
jian_lang@std.uestc.edu.cn
yifanzhang@std.uestc.edu.cn
zhangtao.cheng@outlook.com
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Kunpeng Zhang
kpzhang@umd.edu
University of Maryland
College Park, Maryland, United States

Fan Zhou
fan.zhou@uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
Intelligent Terminal Key Laboratory of Sichuan Province
China

ABSTRACT

Accurately predicting the popularity of micro-videos is crucial for real-world applications such as recommender systems and identifying viral marketing opportunities. Existing methods often focus on limited cross-modal information within individual micro-videos, overlooking the potential advantages of exploiting vast repository of past videos. We present MMRA, a multi-modal retrieval-augmented popularity prediction model that enhances prediction accuracy using relevant retrieved information. MMRA first retrieves relevant instances from a multi-modal memory bank, aligning video and text through transformation mechanisms involving a vision model and a text-based retriever. Additionally, a multi-modal interaction network is carefully designed to jointly capture cross-modal correlations within the target video and extract informative knowledge through retrieved instances, ultimately enhancing the prediction. Extensive experiments conducted on the real-world micro-video dataset demonstrate the superiority of MMRA when compared to state-of-the-art models. The code and data are available at https://github.com/ICDM-UESTC/MMRA.

CCS CONCEPTS

• **Information systems** → **Retrieval tasks and goals**; • **Social and professional topics** → *User characteristics.*

KEYWORDS

Multi-modal retrieval, micro-video popularity, representation augmentation.

ACM Reference Format:
Ting Zhong, Jian Lang, Yifan Zhang, Zhangtao Cheng, Kunpeng Zhang, and Fan Zhou. 2024. Predicting Micro-video Popularity via Multi-modal Retrieval Augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3626772.3657929

*Corresponding author.

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657929

1 INTRODUCTION

Micro-video platforms like TikTok and Instagram are experiencing a surge in popularity, attracting the attention of millions of users who actively create, share, and propagate micro-videos. The burgeoning interest in these platforms has given rise to a vital field known as *micro-video popularity prediction* (MVPP). The primary objective of MVPP is to predict the future viewership or engagement levels of a specific micro-video over a defined time frame. This area of research has gained significant attention from scholars and experts alike, due to its potential benefit in a wide spectrum of real-world applications, such as advertising [16], social link prediction/recommendation [3, 25, 28, 33, 35, 37], and misinformation detection [2, 9, 31].

Prior Work. Researchers have investigated MVPP using two distinct categories of methods: (1) *Feature-engineering methods* [13, 15] involve the creation of hand-crafted micro-video features with the explicit goal of predicting popularity. These functions are designed using carefully constructed functions [18, 24, 29]. However, it is important to note that these methods heavily rely on expert knowledge and the availability of high-quality features. This reliance on expertise and specific feature quality can limit the scalability of such models. (2) *Deep learning-based methods* [6, 32, 34] have leveraged the expressive capabilities of various neural networks to model multi-modal data effectively. For example, researchers usually integrate visual models like Resnet [8] and ViT [30], along

with textual models such as BERT [12] and AnglE [20], to capture and learn cross-modal correlations for popularity prediction.

Although significant progress has been made in MVPP, prior research has primarily focused on exploiting limited cross-modal correlations within individual micro-videos. This approach has often overlooked rich collaborative information that exists across different micro-videos. For example, the distribution of followers varies among source users on micro-video platforms, leading to significant variations in social feedback for identical micro-videos, depending on the viewing user group [36]. Hence, solely modeling semantic information within an individual video falls short of providing a comprehensive solution to the MVPP task. This limitation has spurred our motivation to enhance the knowledge retrieval stage, allowing the model to explicitly access relevant micro-videos stored in the memory bank to assist popularity prediction of the target video. This idea draws inspiration from the way human specialize to achieve better generalization. Instead of memorizing all concepts, humans acquire specialized skills and retrieve relevant knowledge when required [10].

Challenge. Enhancing MVPP by leveraging retrieval-augmented knowledge poses significant challenges, primarily due to two obstacles. **First**, the complexity arises from the need to assess the similarity between the target video and instances stored in the memory bank. This necessitates evaluating both visual and textual similarities to accurately identify relevant instances. Regrettably, existing retrieval methods primarily encode and retrieve single-modal information [17, 21] from the memory bank, failing to make use of valuable resource of multi-modal knowledge. **Second**, the abundance of noise within micro-video data further compounds the challenge. The noise originates from the inherent variability and irregularity of user behaviors on these platforms. Instances of inconsistency between textual descriptions and micro-video content are commonplace, rendering the direct extraction of relevant instances from the memory bank a daunting task.

Present work. We propose *MMRA*, a pioneering Multi-Modal Retrieval-Augmented micro-video popularity prediction framework for enhancing MVPP. Technically, our approach reimagines the popularity prediction process as a "retrieve-and-predict" paradigm. More specifically, we introduce a multi-modal memory bank that encodes both video frames and textual descriptions with a set of reference *<frames, text>* pairs. During the retrieval phase, we address challenges posed by noisy video-text pairs and the inherent cross-modal gap by employing a visual captioner (e.g., BLIP [19]) to generate synthetic captions based on video frames. These synthetic captions, in conjunction with the textual descriptions of the videos, serve as text prompts for retrieving the most relevant *<frames, text>* pairs from the memory bank. Next, this retrieved knowledge is incorporated as supplementary model inputs to guide the popularity prediction for the target video. In the prediction stage, we devised a multi-modal interaction network to capture both multi-modal feature interactions within the target videos and inter-sample feature interactions among relevant instances. Extensive experiments conducted on real-world micro-video dataset demonstrated the superiority of our MMRA over existing state-of-the-art baselines.

2 METHODOLOGY

Problem Definition. Let $\mathcal{V} = \{\mathcal{V}_1, \cdots, \mathcal{V}_N\}$ denotes the set of micro-videos available on online video platforms, where N is the number of micro-videos. Each video $\mathcal{V}_i$ consists of K modality content $\mathbf{M}_i = \{m_1, \cdots, m_K\}$, where $K \geq 2$. The goal of MVPP aims to predict the cumulative views y_i of a given micro-video $\mathcal{V}_i$ during a specific future period via utilizing all modalities that significantly contribute to the prediction of its popularity trend after its release. An overview of our proposed MMRA is shown in Figure 1.

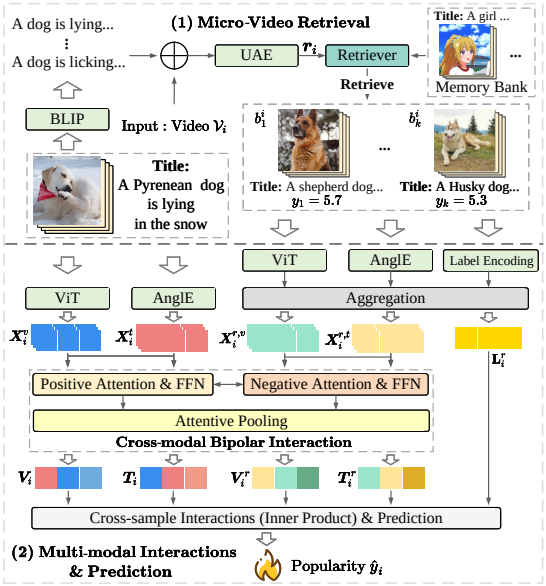

Figure 1: Overview of our proposed framework MMRA.

2.1 Micro-Video Retrieval

To retrieve relevant instances that provide useful guidance for popularity prediction of the target video, we carefully design the text prompts. Specifically, we have developed a video-to-text transformation process that leverages advanced vision models to generate captions for videos. Subsequently, we merge these synthetic video captions with the original textual descriptions associated with the video. This combined text serves as the text prompt for large language models (LLMs), which encode it into a retrieval vector representing the corresponding videos. With such a design, we successfully align the visual and textual modalities of micro-videos while addressing potential inconsistencies between textual descriptions and the actual video content.

2.1.1 Retrieval Vector Generation. The micro-video memory bank $\mathcal{B}$ is defined as a set of reference <frames, text> pairs, where video frames and textual descriptions are encoded. To use this resource, for a micro-video $\mathcal{V}_i$, we first introduce a visual captioner (i.e., BLIP [19]) to excavate the video content and generate descriptive image captions for its corresponding frames, i.e., $C_i = [c_1^i, \cdots, c_L^i]$, where L denotes the number of frames for the video $\mathcal{V}_i$. Then, we combine synthetic caption C_i with the original textual descriptions $\mathcal{T}_i$ as a text prompt $\mathcal{P}_i = C_i \oplus \mathcal{T}_i$. $\oplus$ denotes the concatenation operation. Finally, we feed this text prompt $\mathcal{P}_i$ into a pre-trained semantic extraction model (i.e., UAE-Large [20]) and generate the retrieval

vector $\boldsymbol{r}_i$ representing the corresponding video $\mathcal{V}_i$. Analogously, each raw micro-video in the memory bank obtains its corresponding retrieval vector via the same generation process.

2.1.2 Retriever. The retriever is to search Top-k nearest videos from the memory bank by calculating the similarity scores between the target video and instances in the memory bank. Specifically, the retriever takes a query $\mathcal{V}_q$ as input and retrieves from the memory $\mathcal{B}$ of frame-text pairs. Then, we use the maximum inner product search (MIPS [7]) over all memory candidates $b \in \mathcal{B}$ to find Top-k nearest neighbors $\text{Top}_K(\mathcal{B}|\mathcal{V}_q) = [b_1^q, \cdots, b_k^q]$ based on similarity scores $\mathcal{S} = [s_{q,b_1}, \cdots, s_{q,b_k}]$. This process can be defined as follows:

$$\text{Top}_K(\mathcal{B} \mid \mathcal{V}_q) = \arg\max_{b \in \mathcal{B}} \boldsymbol{r}_q \cdot \boldsymbol{r}_b \tag{1}$$

2.2 Multi-modal Interactions & Prediction

2.2.1 Multi-modal Feature Extraction. Given a micro-video $\mathcal{V}_i$, we sample video frames with a temporal stride of τ and feed these frames into a pre-trained visual model (i.e., vision transformer (ViT) [30]) to extract visual representations $\mathbf{E}_i^v \in \mathbb{R}^{T^o \times d_v}$, where $T^o = \lfloor \frac{T}{\tau} \rfloor$ and d_v denote the dimensions of visual features. Next, the generated frame features $\mathbf{E}^v$ are passed through a linear layer $\boldsymbol{W}_v \in \mathbb{R}^{d_v \times d}$ equipped with the ReLU activation function, creating vision input token $\mathbf{X}^v \in \mathbb{R}^{T^o \times d}$: $\mathbf{X}_i^v = \text{ReLU}(\mathbf{E}_i^v \boldsymbol{W}_v)$. Moreover, for textual content $\mathcal{T}_i$, we feed the textual descriptions into a pre-trained language model (i.e., AnglE [20]) and obtain a sequence of word embeddings $\mathbf{E}_i^t \in \mathbb{R}^{n \times d_t}$, where d_t denotes the embedding dimension and n is the length of words in the textual descriptions. Then the word embeddings $\mathbf{E}_i^t$ are passed through a linear layer $\boldsymbol{W}_t \in \mathbb{R}^{d_t \times d}$, generating textual input token $\mathbf{X}^t \in \mathbb{R}^{n \times d}$: $\mathbf{X}_i^t = \text{ReLU}(\mathbf{E}_i^t \boldsymbol{W}_t)$. Analogously, visual and textual representations of relevant instances are generated by the same process.

2.2.2 Cross-modal Bipolar Interaction. Aligning visual and textual modalities becomes challenging due to the presence of inconsistent information between textual descriptions and video content in real-world micro-videos. Inspired by the cross-attention mechanisms [4, 11, 27], we introduce a bipolar attention mechanism to construct a cross-modal bipolar interaction network, consisting of a positive attention and a negative attention, for addressing this issue. Specifically, the positive attention aims to identify the most consistent features across different modalities, while the negative attention is used to discriminate inconsistent or contradictory information.

In the positive attention, the most similar features between modalities can be calculated by the cross-modal attention vectors. Formally, given a micro-video $\mathcal{V}_i$, the visually guided positive textual features $\boldsymbol{T}_i^{\mathcal{P}}$ can be computed as follows:

$$\boldsymbol{T}_i^{\mathcal{P}} = \text{ATT}_{\mathcal{P}}\left(\mathbf{X}_i^v \boldsymbol{W}_{\mathcal{P}}^Q, \mathbf{X}_i^t \boldsymbol{W}_{\mathcal{P}}^K, \mathbf{X}_i^t \boldsymbol{W}_{\mathcal{P}}^V\right) = \text{Softmax}\left(\alpha \frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}\right)\boldsymbol{V}, \tag{2}$$

where $\boldsymbol{Z}_c \boldsymbol{W}_{\mathcal{P}}^Q, \boldsymbol{Z}_c \boldsymbol{W}_{\mathcal{P}}^K, \boldsymbol{Z}_c \boldsymbol{W}_{\mathcal{P}}^V$ denote the query, key, and value, respectively. $\mathbf{W}_{\mathcal{P}}^Q, \mathbf{W}_{\mathcal{P}}^K, \mathbf{W}_{\mathcal{P}}^V \in \mathbb{R}^{d \times d}$ denote the query, key, and value projection matrices, respectively. α is a adjustable parameter to control the balance between positive and negative attention proportions. Analogously, the textually guided positive visual features $\mathbf{V}_i^{\mathcal{P}}$ can be obtained through the same process.

During the negative attention, it focuses on extracting the inconsistent modal information across distinct modalities. Specifically, the negative attention scores can be calculated by the scaled dot-product attention [27] with a negative constant before applying Softmax, which can be summarized as follows:

$$\boldsymbol{T}_i^{\mathcal{N}} = \text{ATT}_{\mathcal{N}}\left(\mathbf{X}_i^v \boldsymbol{W}_{\mathcal{N}}^Q, \mathbf{X}_i^t \boldsymbol{W}_{\mathcal{N}}^K, \mathbf{X}_i^t \boldsymbol{W}_{\mathcal{N}}^V\right) = \text{Softmax}\left(\beta \frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}\right)\boldsymbol{V}, \tag{3}$$

where $\beta = -(1-\alpha)$. $\boldsymbol{T}_i^{\mathcal{N}}$ represents the visually guided negative textual features. $\mathbf{W}_{\mathcal{N}}^Q, \mathbf{W}_{\mathcal{N}}^K, \mathbf{W}_{\mathcal{N}}^V$ denote the query, key, and value projection matrices, respectively. Analogously, the textually guided negative visual features $\mathbf{V}_i^{\mathcal{N}}$ is generated via the same process.

Inspired by the findings in [5] that the FFN layer learns task-specific information, we propose to incorporate visual hidden states into textual hidden states to generate a comprehensive textual modal representation $\tilde{\boldsymbol{T}}_i$ in FFN layers, which modify the calculation of the FFN process as follows: $\tilde{\boldsymbol{T}}_i = \text{ReLU}\left(\mathbf{X}_i^t + \left(\boldsymbol{V}_i^{\mathcal{P}} \oplus \boldsymbol{V}_i^{\mathcal{N}}\right)\boldsymbol{W}_1\right)\boldsymbol{W}_2$, where $\oplus$ denotes the concatenation operation. $\boldsymbol{W}_1 \in \mathbb{R}^{2d \times d}, \boldsymbol{W}_2 \in \mathbb{R}^{d \times d}$ represent the learnable weights. Moreover, the comprehensive visual modal features $\tilde{\boldsymbol{V}}_i$ are generated in the same way. Finally, we exploit expressive representations $\boldsymbol{V}_i, \boldsymbol{T}_i \in \mathbb{R}^{1 \times d}$ from the sequences of fused features via the attentive pooling strategy [26].

2.2.3 Retrieval Interaction Enhancement. To capture meaningful knowledge from retrieved relevant instances, we explore retrieved feature- and label-level knowledge for enhancing MVPP. Specifically, we firstly utilize the aggregation function to generate comprehensive representations of k relevant instances from the memory bank. For the aggregation function, we use attention mechanism, where the attention score is computed based on normalized similarity scores in the retrieval process. The underlying intuition behind this idea is: instances with higher similarity scores are more relevant. The similarity scores indicate the importance of the corresponding relevant instance for the target micro-video. Specifically, given a micro-video $\mathcal{V}_i$ and retrieved instances $\mathcal{B}_i^r = [b_1^i, \cdots, b_k^i]$, the aggregation process of retrieved visual representations $\mathbf{X}_i^{r,v}$ can be summarized as follows:

$$\alpha_{i,b_j} = \frac{exp(s_{i,b_j})}{\sum_{j=1}^k exp(s_{i,b_j})}, \quad \mathbf{X}_i^{r,v} = \sum_{j=1}^k \alpha_{i,b_j} \mathbf{X}_{b_j}^v. \tag{4}$$

Analogously, retrieved textual representations $\boldsymbol{X}_i^{r,t}$ can be obtained in the same process. After feeding the cross-modal bipolar interaction network, we can obtain expressive aggregated multi-modal representations, i.e., $\boldsymbol{V}_i^r$ and $\boldsymbol{T}_i^r$. For popularity trends of retrieved instances, we encode the label information via a linear layer and then aggregate label information of relevant instances via the aggregation function, creating the aggregated label representations $\mathbf{L}_i^r$. Finally, MMRA interacts all the features for the $\mathcal{V}_i$ for modeling cross-sample interactions. Thus the feature interactions are constructed as: $\mathcal{I} = [\text{inter}(\boldsymbol{V}_i, \boldsymbol{V}_i^r), \cdots, \text{inter}(\boldsymbol{T}_i, \mathbf{L}_i^r)]$, where $\text{inter}(\cdot)$ denotes the inner product [23].

2.2.4 Prediction Network. For the micro-video $\mathcal{V}_i$, the output layer is fed with the concatenated vector of the former components

as: $\mathbf{H} = \text{concat}([\boldsymbol{V}_i, \boldsymbol{T}_i, \boldsymbol{V}_i^r, \boldsymbol{T}_i^r, \mathbf{L}_i^r, \mathcal{I}])$. The output layer is multi-layer perceptrons (MLPs) with one final output unit to predict the popularity of the target video $\hat{y}_i$. During training, we use the mean square error (MSE) as the loss function.

3 EXPERIMENTS

Table 1: Statistics of dataset.

Dataset	# Video	# User	# Train	# Val	# Test	V	T
MicroLens	19,738	100,000	15,790	1,974	1,974	768	768

Dataset. To evaluate the effectiveness of MMRA, we conduct experiments on the publicly available micro-video dataset: MicroLens [22]. Its descriptive statistics is summarized in Table 1. **V** and **T** represent the dimension of visual and textual features, respectively. **MicroLens** [22] consists of 19,738 unique micro-videos viewed by 100,000 users from various online video platforms.

Baselines. To evaluate the model superiority, we conduct experiments with seven competitive baselines, which can be grouped into two categories: (1) *Feature-engineering methods*: SVR [13] and HyFea [15]. (2) *Deep-learning methods*: CLSTM [6], TMALL [1], MASSL [34], CBAN [4] and HMMVED [32].

Metrics & Parameter Settings. For experimental results, we run each model on MicroLens dataset five times, and report the mean values. During training, model parameters are updated by Adam optimizer [14] and the learning rate is set to 0.0001. We averagely extract 10 key frames of micro-videos on MicroLens. α is set to 0.6. As for baselines, we employ the parameter settings specified in original papers. We utilize three widely used metrics to evaluate model performance: normalized mean square error (**nMSE**), mean absolute error (**MAE**) and Spearman's Rank Correlation (**SRC**).

Performance Comparison. The overall performance of MMRA and baselines are reported in Table 2. The best results are in bold font and the second underlined. The experimental result indicates that our MMRA consistently outperforms all baselines on MicroLens dataset, demonstrating its superiority. These results verify the effectiveness of constructing a retrieval-augmented pipeline for enhancing MVPP. Specifically, retrieving relevant instances from the memory bank via text prompting enables MMRA to better obtain meaningful knowledge. The interaction network can guide prediction via extracting expressive representations and useful popularity information through retrieved instances.

Ablation Study. We analyze how MMRA benefits from each key component. Experimental results are reported in Table 3. (1) *The effectiveness of retrieval.* We remove the retrieval module (w/o RE) and only utilize multi-modal features for prediction. We observe that removing the retrieval module leads to large performance degradation. It verifies the effectiveness of retrieval-augmented strategy. (2) *The effectiveness of multi-modal interactions.* We design several variant models without the negative attention (w/o Neg), and the positive attention (w/o Pos). Experimental results show that removing either component leads to performance degradation, demonstrating that all components in MMRA are effective and necessary. (3) *The effectiveness of modal content.* We design various variants without the visual modality (w/o V) or the textual modality (w/o T). These results indicate that multi-modal content is beneficial for the final popularity prediction.

Table 2: Performance comparison. Lower values of nMSE and MAE, and higher values of SRC, indicate better performance.

Model	MicroLens		
Metrics	nMSE	MAE	SRC
SVR	0.8132	1.2176	0.4288
HyFea	0.8106	1.2321	0.4345
CLSTM	0.7966	1.2117	0.4573
TMALL	0.9373	1.2990	0.3817
MASSL	1.0797	1.4136	0.3875
CBAN	<u>0.7727</u>	<u>1.1900</u>	<u>0.4746</u>
HMMVED	0.8632	1.2524	0.3716
MMRA	**0.7530**	**1.1787**	**0.4900**

Table 3: Ablation study on key components of MMRA.

Dataset		MicroLens		
Module	Variant	nMSE	MAE	SRC
MMRA	**All**	**0.7530**	**1.1787**	**0.4900**
Retrieval	w/o RE	0.7745	1.1908	0.4755
Multi-modal Interaction	w/o Neg	0.7595	1.1838	0.4856
	w/o Pos	0.7595	1.1829	0.4870
Modal Content	w/o V	0.7915	1.2071	0.4536
	w/o T	0.7629	1.1844	0.3716

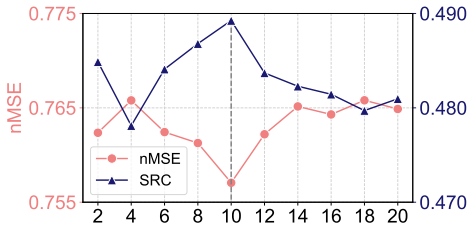

Figure 2: Performance *vs.* Numbers of retrieved videos (k).

Parameter Analysis. To analyze the influence of the number k of retrieved instances on MMRA, we conduct experiments on MicroLens dataset. The experimental results are shown in Figure 2. We observe that the model performance initially improves with an increase in the number k and subsequently declines when the number is too large. When we set k to 10 on MicroLens, the model performance is optimal. The observed performance degradation with larger numbers suggests that retrieved large number of instances from memory bank will influence the representation learning of the target micro-video.

4 CONCLUSION

This work presented MMRA, the first retrieval-augmented framework for the MVPP task. We propose to align the visual and textual modalities and generate the retrieval vectors to search Top-k nearest videos. We also introduced a cross-modal bipolar interaction to address the presence of inconsistent information between texts and video content, as well as a retrieval interaction enhancement method to capture meaningful knowledge from relevant instances. Experiments on the real-world micro-video dataset demonstrate the effectiveness of our method.

5 ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (Grant No.62176043 and No.62072077) and Grant SCITLAB-30002 of Intelligent Terminal Key Laboratory of Sichuan Province.

REFERENCES

[1] Jingyuan Chen, Xuemeng Song, Liqiang Nie, Xiang Wang, Hanwang Zhang, and Tat-Seng Chua. 2016. Micro tells macro: Predicting the popularity of micro-videos via a transductive model. In *ACM International Conference on Multimedia (MM)*. 898–907.

[2] Xueqin Chen, Fan Zhou, Goce Trajcevski, and Marcello Bonsangue. 2022. Multi-view learning with distinguishable feature fusion for rumor detection. *Knowledge-Based Systems* 240 (2022), 108085.

[3] Zhangtao Cheng, Wenxue Ye, Leyuan Liu, Wenxin Tai, and Fan Zhou. 2023. Enhancing Information Diffusion Prediction with Self-Supervised Disentangled User and Cascade Representations. In *ACM International Conference on Information and Knowledge Management (CIKM)*. 3808–3812.

[4] Tsun-hin Cheung and Kin-man Lam. 2022. Crossmodal bipolar attention for multimodal classification on social media. *Neurocomputing* 514 (2022), 1–12.

[5] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 5484–5495.

[6] 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).

[7] Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. 2020. Accelerating large-scale inference with anisotropic vector quantization. In *International Conference on Machine Learning (ICML)*. PMLR, 3887–3896.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[9] Zhenyu He, Ce Li, Fan Zhou, and Yi Yang. 2021. Rumor detection on social media with event augmentations. In *Proceedings of the international ACM SIGIR conference on research and development in information retrieval*. 2020–2024.

[10] Douglas R Hofstadter. 1995. *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*. Basic books.

[11] Tao Jiang, Jiahai Wang, Zhiyue Liu, and Yingbiao Ling. 2020. Fusion-extraction network for multimodal sentiment analysis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. Springer, 785–797.

[12] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Vol. 1. 2.

[13] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. 2014. What makes an image popular?. In *Proceedings of the ACM Web Conference (WWW)*. 867–876.

[14] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[15] Xin Lai, Yihong Zhang, and Wei Zhang. 2020. HyFea: Winning solution to social media popularity prediction for multimedia grand challenge 2020. In *ACM International Conference on Multimedia (MM)*. 4565–4569.

[16] Himabindu Lakkaraju and Jitendra Ajmera. 2011. Attention prediction on social media brand pages. In *ACM International Conference on Information and Knowledge Management (CIKM)*. 2157–2160.

[17] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems (NeuIPS)* 33 (2020), 9459–9474.

[18] Haitao Li, Xiaoqiang Ma, Feng Wang, Jiangchuan Liu, and Ke Xu. 2013. On popularity prediction of videos shared in online social networks. In *ACM International Conference on Information and Knowledge Management (CIKM)*. 169–178.

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*. PMLR, 12888–12900.

[20] Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871* (2023).

[21] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. 2022. Retrieval augmented classification for long-tail visual recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6959–6969.

[22] Yongxin Ni, Yu Cheng, Xiangyan Liu, Junchen Fu, Youhua Li, Xiangnan He, Yongfeng Zhang, and Fajie Yuan. 2023. A Content-Driven Micro-Video Recommendation Dataset at Scale. *arXiv preprint arXiv:2309.15379* (2023).

[23] Yanru Qu, Han Cai, Kan Ren, Weinan Zhang, Yong Yu, Ying Wen, and Jun Wang. 2016. Product-based neural networks for user response prediction. In *IEEE International Conference on Data Mining (ICDM)*. IEEE, 1149–1154.

[24] Suman Deb Roy, Tao Mei, Wenjun Zeng, and Shipeng Li. 2013. Towards cross-domain learning for social video popularity prediction. *IEEE Transactions on Multimedia (TMM)* 15, 6 (2013), 1255–1267.

[25] Lifeng Sun, Xiaoyan Wang, Zhi Wang, Hong Zhao, and Wenwu Zhu. 2016. Social-aware video recommendation for online social groups. *IEEE Transactions on Multimedia (TMM)* 19, 3 (2016), 609–618.

[26] Xiaobing Sun and Wei Lu. 2020. Understanding attention for text classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 3418–3428.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems (NeuIPS)* 30 (2017).

[28] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *ACM International Conference on Multimedia (MM)*. 1437–1445.

[29] 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 AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 30.

[30] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. 2020. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677* (2020).

[31] Lianwei Wu, Yuan Rao, Xiong Yang, Wanzhen Wang, and Ambreen Nazir. 2021. Evidence-aware hierarchical interactive attention networks for explainable claim verification. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 1388–1394.

[32] Jiayi Xie, Yaochen Zhu, and Zhenzhong Chen. 2023. Micro-Video Popularity Prediction Via Multimodal Variational Information Bottleneck. *IEEE Transactions on Multimedia (TMM)* 25 (2023), 24–37.

[33] Xovee Xu, Fan Zhou, Kunpeng Zhang, Siyuan Liu, and Goce Trajcevski. 2021. Casflow: Exploring hierarchical structures and propagation uncertainty for cascade prediction. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 35, 4 (2021), 3484–3499.

[34] Zhuoran Zhang, Shibiao Xu, Li Guo, and Wenke Lian. 2022. Multi-modal Variational Auto-Encoder Model for Micro-video Popularity Prediction. In *Proceedings of the International Conference on Communication and Information Processing (ICCIP)*. 9–16.

[35] Fan Zhou, Bangying Wu, Yi Yang, Goce Trajcevski, Kunpeng Zhang, and Ting Zhong. 2018. Vec2link: Unifying heterogeneous data for social link prediction. In *ACM International Conference on Information and Knowledge Management (CIKM)*. 1843–1846.

[36] Fan Zhou, Xovee Xu, Goce Trajcevski, and Kunpeng Zhang. 2021. A survey of information cascade analysis: Models, predictions, and recent advances. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–36.

[37] Fan Zhou, Xovee Xu, Kunpeng Zhang, Goce Trajcevski, and Ting Zhong. 2020. Variational information diffusion for probabilistic cascades prediction. In *IEEE INFOCOM 2020-IEEE conference on computer communications*. IEEE, 1618–1627.