7-8-2023

A Hierarchical Attention-based Contrastive Learning Method for Micro Video Popularity Prediction

Tan Cheng
Fudan University, chengt21@m.fudan.edu.cn

Chenghong Zhang
School of Management, Fudan University, chzhang@fudan.edu.cn

Gang Chen
Zhejiang University, chengang050970@foxmail.com

Shuaiyong Xiao
Tongji University, syxiao@tongji.edu.cn

Zongxiang Zhang
Fudan University, zongxiangzhang21@m.fudan.edu.cn

See next page for additional authors

A Hierarchical Attention-based Contrastive Learning Method for Micro Video Popularity Prediction

Completed Research Paper

Tan Cheng
Fudan University
Shanghai, China
chengt21@m.fudan.edu.cn

Chenghong Zhang
Fudan University
Shanghai, China
chzhang@fudan.edu.cn

Gang Chen
Zhejiang University
Hangzhou, China
chengang050970@foxmail.com

Shuaiyong Xiao
Tongji University
Shanghai, China
syxiao@tongji.edu.cn

Zongxiang Zhang
Fudan University
Shanghai, China
zongxiangzhang21@m.fudan.edu.cn

Xulei Jin
Fudan University
Shanghai, China
xuleijin21@m.fudan.edu.cn

Abstract

Micro videos popularity prediction (MVPP) has recently attracted widespread research interests given the increasing prevalence of video-based social platforms. However, previous studies have overlooked the unique patterns between popular and unpopular videos and the interactions between asynchronous features different data dimensions. To address this, we propose a novel hierarchical attention contrastive learning method named HACL, which extracts explainable representation features, learns their asynchronous interactions from both temporal and spatial levels, and separates the positive and negative embeddings identities. This reveals video popularity in a contrastive and interrelated view, and thus can be responsible for a better MVPP. Dual neural networks account for separate positive and negative patterns via contrastive learning. To obtain the temporal-wise interaction coefficients, we propose a Hadamard-product based attention approach to optimize the trainable attention-map matrices. Results from our experiments on a TikTok micro video dataset show that HACL outperforms benchmarks and provides insightful managerial implications.

Keywords: Micro video popularity, multimodal learning, contrastive learning

Introduction

Recent years have witnessed rapid development in micro video industry. For example, TikTok, one of the largest worldwide social video platforms, has recently realized its maximum monthly active users of about 1 billion[1]. The advantages held by these platforms, e.g., being easy-to-learn, convenient-to-use, and rapid-

Corresponding author: Chenghong Zhang (chzhang@fudan.edu.cn)
1 https://www.demandsage.com/tiktok-user-statistics/

to-separate, accelerate the generation of numerous micro videos. This brings benefits to multi-parties, such as enriching the entertainment life of users, broadening the marketing channels of sellers, and increasing the user traffic on the platform. Nonetheless, evaluating the business value of the booming micro videos poses considerable challenges. For users, the explosion of micro videos aggravates information overload, increasing the cost of searching. It is difficult for platforms to connect a large number of videos to their potential viewers, which is key to maintain stickiness and commercial values of users. Sellers also get stuck in the problem of selecting suitable micro videos for advertisement. All the above challenges point to the requirement micro video popularity prediction (short for MVPP hereafter), which aims to identify the potential extent to which the micro video will attract the attention and interest of users. For example, the MVPP offers supplementary user common interests, thereby providing complementary information to recommendation algorithms that rely on historical interaction data to learn individual preferences. This is particularly beneficial in addressing the cold start problem, where limited or no user data is available. By incorporating additional interests from MVPP, recommendation algorithms can gain a broader understanding of user preferences, enhancing their ability to generate accurate and relevant recommendations. This helps alleviate the issue of information overload for users and addresses the platform's challenge of matching the right content to the right users. Moreover, for merchants, the process of making advertising decisions is intricately linked to the exposure and visibility of their products. To effectively plan and strategize their advertising campaigns, merchants require accurate predictions of micro video popularity. This information plays a critical role in guiding their decision-making process and allows them to allocate resources effectively for optimal advertising outcomes. Micro video popularity (Chen et al., 2016) has been recently quantified and predicted by a growing number of researchers in order to mitigate the overloaded micro videos (Jing et al., 2018; Xie et al., 2021). Consequently, how to better predict the micro video's popularity is of considerable business value and has attracted widespread attention from the academia and industrial community.

The MVPP task can be approached as either a regression or classification problem, depending on how micro video popularity is defined. Popularity can be measured as a continuous value, such as the average number of comments, likes, reposts, and loops/views, leading to a regression problem (Chen et al., 2016; Jing et al., 2018). Alternatively, it can be represented as a binary label, indicating whether the number of likes exceeds a predefined threshold, resulting in a classification problem (Xie et al., 2020). Since a micro video can be deemed as a superposition of three information modalities, i.e., visual, acoustic, and textual information (i.e., multimodality), the quality of the representation learning and the rationality of the comprehensive modeling across multiple data modalities determine the effectiveness of MVPP. In this regard, extant research efforts have been made to develop either representation learning networks, e.g., variational-encoder-decoder-based MVPP methods (Xie et al., 2021; Xie et al., 2020), or multimodal fusion frameworks, e.g., attention-based (Wang et al., 2022) and regularization-based (Jing et al., 2018) MVPP methods.

However, these methods suffer from inherent defects in representing and fusing multimodal video information for MVPP. First, the deep representation features learned by existing methods from video data are the combination of positive and negative samples and thus make it difficult to reveal video popularity in an efficient way, resulting in an unstable performance of MVPP. Second, although a number of MVPP methods have provided approaches for exploring the temporal variation patterns (i.e., temporal modeling) and multimodal comparative patterns (i.e., multimodal fusion) in the video data, these methods are incompetent for modeling the complicated interrelations among asynchronous features[2], which are pivotal for the video popularity in real MVPP scenarios. For example, in an entertainment micro video, the background music affects users' attention (Shih et al., 2012), deciding whether the micro videos will be popular. Considering a charming and beautiful nature scenery, like massive glaciers, along with creepy nursery rhymes as background music, the inharmony between the visual modality and acoustic modality makes the micro video neither fish nor fowl, resulting in less popularity. It is the same for interaction between other modalities. Besides, there are complex spatial and temporal properties, along with their complex interactions. For users, their interest fluctuates with time steps, experiencing various emotions within a micro video. For micro videos, the visual, acoustic, and textual information varies with time steps simultaneously. Thus, the spatial and temporal properties cannot be learned separately.

2 Since there are three aspects for describing a feature extracted from video data, i.e., the feature dimension, the modality dimension, and the temporal dimension, we refer to asynchronous features as features that do not exist in a same data dimension, e.g., two features from different modalities and time steps.

In response to the aforementioned difficulties, we develop an innovative method, named hierarchical-attention-based contrastive learning (HACL), which exploits the complicated interactions across modalities and time steps in a micro video, and adaptively leverages the interacted features to construct positive and negative views in terms of popularity for a rational MVPP. Specifically, HACL has three components. First, we extract multimodal features from micro videos with pre-trained models and then construct feature-wise representations to enhance the characterization. Second, we reproduce these representations and capture the joint correlations at the time step and modality levels. The Hadamard- product-based attention is used to calculate temporal weights, followed by a self-attention layer extracting the modality-wise interactions. Third, for the classification task, we use the contrastive loss for the training of the dual deep neural networks. We summarize our main contributions as follows:

- We propose an end-to-end hierarchical contrastive learning model for MVPP. We apply dual deep networks to video samples and jointly leverage the positive and negative embeddings based on contrastive learning for the MVPP task.
- We combine Hadamard-product-based attention and self-attention to attain the interactions within a modality at a time step and then measure the across-level interaction coefficients via a feedforward process. Besides, we propose the contrastive loss between positive and negative networks to improve the performance of our model.
- We conduct extensive experiments to evaluate the MVPP performance of our proposed model based on a practical dataset from TikTok. From the numerical experiments, we find that HACL outperforms other state-of-the-art benchmarks.

In the rest of this paper, we first summarize the related research and point out the research gap and our motivation in Section 2. Then, in Section 3, we formulate the MVPP problem, propose our HACL method, and describe the implementation in detail. In order to evaluate the performance of our proposed HACL, we design multiple comparison experiments followed by sensitive analysis and ablation studies and visualize the results, especially the interaction coefficients, in Section 4. Finally, we draw a conclusion on our paper in Section 5.

Related Work

Multimodal Deep Learning

Multimodal deep learning has experienced a surge in popularity in recent years, driven by the proliferation of various data types and formats. Unstructured data from multiple sources, with diverse forms and distributions, often contains valuable information. Consequently, researchers have focused on leveraging multimodal deep learning techniques to extract complementary information from each modality in a learning task. This approach aims to create a comprehensive representation that harnesses the strengths of different modalities, resulting in improved performance compared to relying on a single modality alone. By integrating information from multiple modalities, the potential for achieving superior results is greatly enhanced.

Deep learning is a multi-level abstract representation of data that is learned via a hierarchical computational model (LeCun et al., 2015). Multimodal learning based on deep learning provides several benefits over classic machine learning approaches, particularly in the field of high-dimensional unstructured data. Research has shown that multimodal fusion strategies are crucial for the predictive performance of multimodal deep learning models (Zhang et al., 2020). For example, Ngiam et al. (Ngiam et al., 2011) studied numerous multimodal fusion techniques, including simple concatenation of inputs and shared representation learning, as well as cross-modality learning, after which many researches devoted to this area. Averaging (Shutova et al., 2016), voting (Morvant et al., 2014) and weighting (Ramirez et al., 2011) are the major strategies used in the feature fusion.

Recent research has made significant strides in the field of multi-view learning, aiming to uncover correlations between different modalities and enhance learning outcomes. With the rise of self-attention mechanisms, researchers have employed attention mechanisms to achieve superior fusion representations of modalities by training importance weights for each modality (Gu et al., 2018). In line with this progress, Yan et al. (Yang et al., 2021) propose an innovative multimodal emotion analysis model called the Multi-view Attentional Network (MVAN). MVAN takes into account the cross-modal relationships and employs a continuously updated memory network to extract deep semantic aspects from image-text pairs. However,

due to the input dimension limitations of the attention mechanism, this approach faces challenges when dealing with scenarios involving temporal and spatial fusion. To address these limitations, Cheng et al. (Cheng et al., 2020) introduced the Spatial-Temporal Attention-based Neural Network (STAN). STAN utilizes two distinct self-attention mechanisms and measures the importance of both temporal and spatial dimensions. This enables effective handling of scenarios involving temporal and spatial fusion. However, it is important to note that STAN still falls short in capturing the holistic interaction between time steps.

Another important multimodal fusion method is contrastive learning. Contrastive learning is a discriminative deep learning technique that employs specific criteria to compare embedded features within positive and negative sample pairs during the representation learning process of multiple modalities. Its objective is to train similar samples, represented by positive sample pairs, to be closer together, while ensuring that dissimilar samples, represented by negative sample pairs, are separated in the learned feature space (Jaiswal et al., 2021). Drawing inspiration from contrastive learning, Ding et al. (Ding et al., 2015) introduced the Distance Loss, a novel multimodal fusion approach rooted in the concept of comparative loss. By simultaneously training multiple neural networks through the maximization of relative distances (which can be viewed as a specific instance of contrastive learning), this method aims to produce distinctive feature representations.

Given the widespread adoption of graph neural networks (GNNs), they have found utility in the realm of multimodal deep learning. In this paradigm, each modality and feature assume the role of a node, while the interaction among them is represented by edges. Addressing this context, Mai et al. (Mai et al., 2020) introduced the Adversarial Representation Graph Fusion, a comprehensive framework for multimodal fusion. By employing adversarial learning, this method facilitates the collective embedding of diverse modalities into a unified representation space, which is subsequently fused using a hierarchical graph network. Nonetheless, it is important to note that this approach encounters a similar limitation as that of attention mechanisms.

Given multimodal deep learning's superior performance in processing unstructured data, researchers tried to predict the popularity of online content with multimodal deep learning methods. For instance, Abousaleh et al. (Abousaleh et al., 2021) inspired by multimodal learning and CNN, combined social and visual information, predicting the popularity of online images with two CNN. Gu et al. (Gu, 2020) designed an attention mechanism, using the CNN and LSTM to extract features from images and text, respectively, solving the problem of tweets popularity prediction. To predict the popularity of social media content, Chen et al. (Chen et al., 2019) analyzed and fused a collection of rich information from texts, users, and videos. However, existing studies do not give sufficient consideration to the temporal feature, which is of great importance in micro-video popularity prediction.

Micro Video Popularity Prediction

Due to its significant commercial implications in various domains such as recommendation systems, advertising, and bandwidth allocation, the prediction of popularity has garnered considerable attention from researchers. This encompasses popularity prediction across different media types, including text, images, and videos. In the context of videos specifically, researchers have focused on extracting crucial elements from video content to understand the patterns of video propagation, thereby enabling subsequent studies on popularity prediction. In the current state of the art, two distinct research directions have emerged in the field of Multi-View Popularity Prediction (MVPP).

On one hand, MVPP has been approached as a time series analysis, acknowledging the temporal variations in popularity. Li et al. (Li et al., 2013) developed a model that incorporates both video attractiveness and social context to describe video propagation and predict view counts on online social networks. Similarly, Ma et al. (Ma et al., 2017) proposed a lifetime-aware regression model utilizing time series analysis for long-term video popularity prediction in complex networks. Although these methods outperform traditional approaches like time series analysis and multiple linear regression, they primarily rely on structured data, overlooking valuable content-related information.

On the other hand, MVPP has been recognized as an instantaneous task with a focus on multimodal aspects. Chen et al. (Chen et al., 2016) introduced the Transductive Multimodal Learning Model (TMALL), which harnesses visual, acoustic, textual, and social features to effectively leverage heterogeneous multimodal data and identify key factors influencing micro video popularity. Building upon Chen et al.'s work, Jing et

al. (Jing et al., 2018) addressed internal noise by incorporating low-rank representation and multi-graph regularized least squares, refining the regression framework. Trzcinski et al. (Trzciński & Rokita, 2017) employed temporal and visual features, utilizing support vector regression with Gaussian radial basis functions to predict popularity, highlighting the crucial role of social features in video popularity prediction. Additionally, recurrent neural networks have been employed in popularity prediction research (Trzciński et al., 2017).To overcome internal and external uncertainties, Liao et al. (Liao et al., 2019) proposed the Deep Fusion of Temporal Process and Content Features method, combining recurrent neural networks and convolutional neural networks while incorporating temporal attention fusion. Moreover, Xie et al.(Xie et al., 2020) introduced a Hierarchical Multimodal Variational Encoder-Decoder designed for macro-video popularity prediction, which incorporates a deep information bottleneck constraint to control predictive information within hidden representations. Inspired by the multimodal extension of variational information bottleneck theory, this approach aims to mitigate uncertainties and enhance prediction accuracy.

As with the case of multimodal deep learning, current MVPP techniques often simplify the complexity of the task by overlooking the intricate and nuanced meta-interactions among features.

Research Gaps and Motivations

Overall, our review of related works suggests their deficiencies in dealing with MVPP. Regardless of the wide usage of multimodal deep learning, the temporal connection among different modalities has not been fully developed. Moreover, it is crucial to consider positive and negative embedding separately. Nevertheless, current researches regard micro video embedding as a whole, integrating the different patterns. Consequently, a capable MVPP method is supposed to satisfy the following needs: (1) understanding a series of interactions: inter-modality interaction, temporal interaction, and hierarchically across interactions; (2) identifying the key points of positive and negative patterns. Given the excellent performance of multimodal deep learning and the suitable representation of interactions with contrastive learning, we proposed a novel hierarchical Hadamard-product attention contrastive learning network that constructs dual deep networks to trace the complicated interactions, learn the distinct identification of positive and negative samples, and enable the prediction of micro video popularity. A related work summary is shown in Table 1. Compared with them, the main contributions of our work are the dual networks and Hadamard-product-based attention, which conform to a better understanding of users' interests.

Study	Multimodal deep learning	Temporal analysis	Dual networks	Across-level interactions
(Li et al., 2013)		√(non-video)		
(Ma et al., 2017)		√(non-video)		
(Chen et al., 2016)	√	√(video)		
(Jing et al., 2018)	√	√(video)		
(Trzciński & Rokita, 2017)	√	√(non-video)		
(Trzciński et al., 2017)	√	√(video)		
(Liao et al., 2019)	√	√		
(Xie et al., 2021)	√	√		
Our work	√	√(video)	√	√

Table 1. Comparison of HACL with Existing Relevant Methods.

The Proposed Model

In this section, we first introduce the definitions and notations used in this paper, then give an overview of the proposed method, followed by a detailed illustration. Finally, we demonstrate the implementation of our method.

Problem Formulation and Framework Overview

Given that N micro videos are labeled with popularity classification identification $\mathbf{y} = [y_i, \cdots, y_i, \cdots, y_n] \in \mathbb{R}^n$, where $y_i \in 0,1$, we cut each video into T time steps, extract F types of features from M modalities, and then obtain the representation of video i at time step t from modality m as $\mathbf{X}_{it}^m = \left[x_{it}^{1m}, \cdots, x_{it}^{fm}, \cdots, x_{it}^{FM}\right] \in \mathbb{R}^F$. Consequently, an input sample with a batch equaling b can be written as $\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_t, \cdots \mathbf{X}_T] \in \mathbb{R}^{b\times T\times M\times F}$, where $\mathbf{X}_t = [\mathbf{X}_t^1, \cdots, \mathbf{X}_t^m, \cdots, \mathbf{X}_t^M] \in \mathbb{R}^{b\times M\times F}$ and $\mathbf{X}_t^m = \left[\mathbf{x}_t^{1m}, \cdots, \mathbf{x}_t^{fm}, \cdots, \mathbf{x}_t^{Fm}\right] \in \mathbb{R}^{b\times F}$ indicate the representation matrixes at the t^{th} time step and of the m^{th} modality over b (batch size) instances, separately. Our goal is to predict the micro video popularity and obtain interaction coefficients based on integrated information from all time steps and modalities.

Our proposed HACL is an end-to-end MVPP-oriented method that acquires a comprehensive understanding by incorporating hierarchical attention mechanisms that operate across various time steps and modalities, while also accounting for the unique interactions between positive and negative samples. Figure 1 depicts the framework of the proposed model HACL, which includes five major components: multimode embedding, Hadamard-attention-based temporal-wise interaction, self-attention-based modality-wise interaction, contrastive-loss-based interaction separation, and popularity prediction.

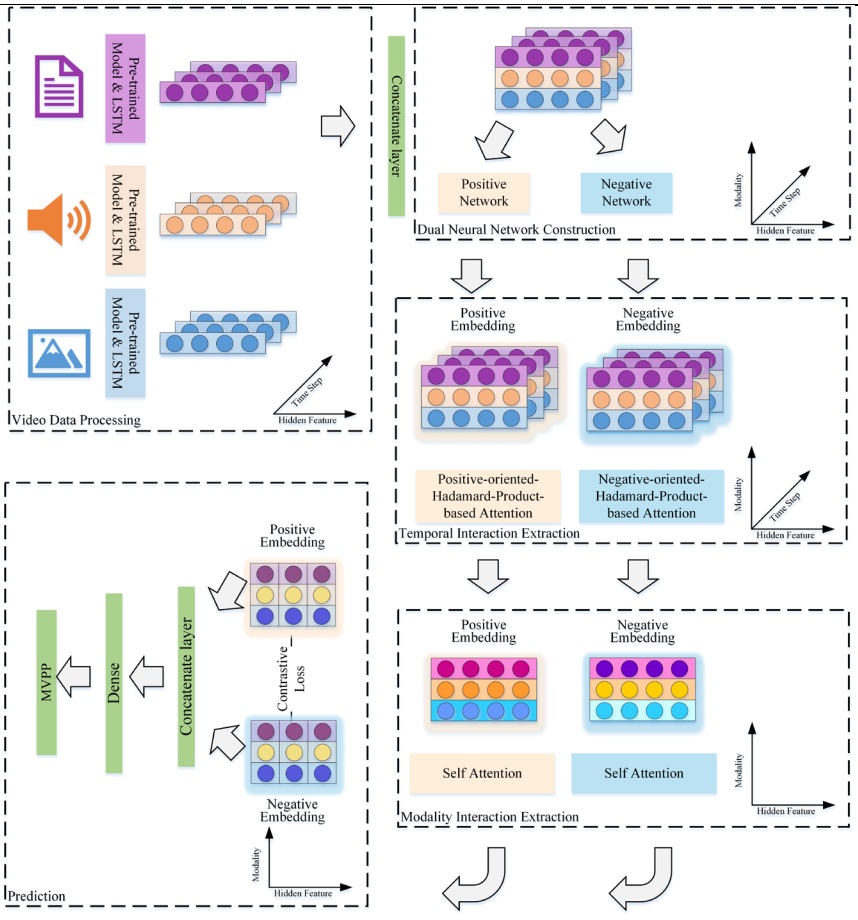


Figure 1. The Framework of HACL

First, it extracts multimodal features and gains input matrices through fine-tuned existing models. The input matrices are then fed into *M* independent LSTM networks with the same structure depending on modalities. Then, the latent representations are concatenated and duplicated. Both copies are passed into two parallel deep networks to distill and recognize positive (popular) and negative (unpopular) patterns. Specifically, in each pipeline, it uses Hadamard-attention-based techniques and self-attention to obtain interaction coefficients within time steps and modalities separately. The outputs of these two pipelines will be utilized to compute the contrastive loss and optimize the trainable parameters for the optimization of positive and negative embeddings. Finally, via this contrastive learning objective, it leverages multi-level interactions to facilitate MVPP performance.

Multimodal Embedding

Multimodal embedding seeks to extract essential features from micro videos. Each micro video is divided into ten pieces, each of which has three different modalities. For textual features, we first turn audio into text using a voice-to-text application, and then get content and sentiment features using ENRIE[3] and SENTA[4], respectively. Previous research has shown that suitable acoustic features are necessary for video popularity and has exploited acoustic features to improve prediction performance. Following previous studies, we extract twenty-one-dimensional features from the audio channel to characterize the acoustic modality of micro-videos. These features include mel-frequency cepstral coefficients (MFCCs), energy entropy, signal energy, zero crossing rate, spectral roll off, spectral centroid, and spectral flux. The ranges of the acoustic features consist of continuous real values. As for the fundamental elements in a micro video, visual features, like color histograms and objects have been harnessed in previous studies. To be more precise, we employ a frame difference algorithm with local maxima criteria to extract key frames at each time step for each video, and smoothing the average difference value prior to computing the local maximum may effectively eliminate noise and prevent the repeated extraction of frames from similar scenarios. Next, we get a 150-dimensional vector at each keyframe by classifying the color into 50 unique hues on a single RGB channel. The "AlexNet" ImageNet model is used to extract 1000-class labels for object recognition (Krizhevsky et al., 2017). In order to maintain the same dimensions for extracted features from various modalities, we use an MLP to transform each extracted feature into a unit hidden dimension. All feature vectors are then normalized to the length unit L2-norm.

Hadamard Product-based Attention Mechanism

The HPA mechanism aims to extract temporal interactions among modalities. First, given the n^{th} sample from the interaction layer's feature-time-step, we have:

$$\mathbf{Z} = \left[\mathbf{Z}^1, \cdots, \mathbf{Z}^T \right] \in \mathbb{R}^{T \times F \times d} \tag{1}$$

The classic self-attention approach calculates weights via the similarity of features:

$$a_i = \frac{e^{\left(\mathbf{q}^T \mathbf{k}_i / \sqrt{d_k}\right)}}{\sum_{j=1}^{T} e^{\left(\mathbf{q}^T \mathbf{k}_j / \sqrt{d_k}\right)}} \tag{2}$$

where **q** represents the query and **k** represents the key. Both are derived from the linear transformation of **z**. Despite this, the high-dimensionality and sparsity of the representation render the self-attention method computationally expensive, particularly for the matrix multiplication of high-dimensional matrices while computing **Q**, **K**, and **V** matrices. By transforming matrix multiplication into a Hadamard product and weighting input matrices from each time step (or feature), HPA alleviates this issue. To be more specific, the HPA weights can be derived as:

$$e_i = \| \mathbf{W} \circ \mathbf{Z}_i \|_F \tag{3}$$

3 https://github.com/PaddlePaddle/ERNIE/
4 https://github.com/baidu/Senta/

W is the trainable parameter that is adaptable to the loss function and aids in the distillation of the time-step-wise (or feature-wise) weights. $||\cdot||_F$ stands for the F norm, which reduces high-dimensional matrices to a single real number. In addition, we use the SoftMax operation to get the normalized weight:

$$a_i = \frac{e^{e_i}}{\sum_{j=1}^{T} e^{e_j}} \tag{4}$$

Via Eq.3 and Eq.4, HPA extracts the time-step-wise (or feature-wise) weights. The mechanism scheme is shown in Figure 2. The output of HPA is a weighted sum representation of all time steps (or features).

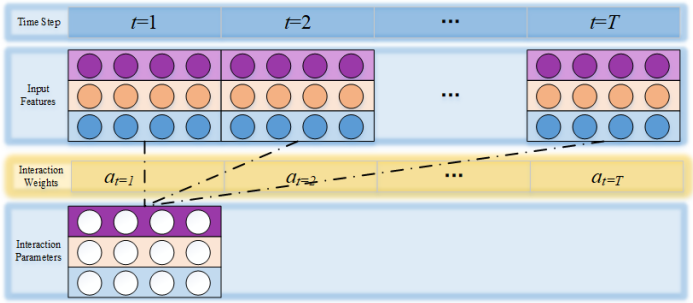


Figure 2. The mechanism of HAP

Model Implementation

Considering that positive (popular) and negative (unpopular) samples hold distinct patterns, HACL constructs dual deep neural networks and adopts two independent pipelines to isolate unique representations. In particular, the outputs of LSTM networks, $\mathbf{Z}_m \in \mathbb{R}^{b\times T\times d}$, where d is the dimension of the hidden layer in LSTM, are the concatenated-on modality and changed into $\mathbf{Z} \in \mathbb{R}^{b\times T\times M\times d}$. Then, we duplicate **Z** for positive and negative embeddings, which are denoted as $\mathbf{Z}^P$ and $\mathbf{Z}^N$ further.

Besides, the interaction among features fluctuates with time steps on account of the diverse user perceptions. HACL then applies HPA to focus on temporal-wise interaction, in which the positive embeddings and negative embeddings at each time step are weighted by the HPA coefficient and summed on time steps separately. Since the elements that go into the attention mechanism are no longer vectors but matrices, HPA uses a query matrix $\mathbf{W}_{HPA}^{(\cdot)} \in \mathbb{R}^{M\times d}$ instead of the traditional query vector. After following the HPA proposed above, we have the positive and negative embeddings fused in time steps.

$$\mathbf{Z}_b^{\left(l+1_{(P)}\right)} = \sum_{t=1}^{T} a_{bt^{(P)}} \mathbf{Z}_{bt}^{\left(l_{(P)}\right)} \in \mathbb{R}^{M\times d} \tag{5}$$

$$\mathbf{Z}_b^{\left(l+1_{(N)}\right)} = \sum_{t=1}^{T} a_{bt}^{(N)} \mathbf{Z}_{bt}^{\left(l_{(N)}\right)} \in \mathbb{R}^{M\times d} \tag{6}$$

where $\mathbf{Z}_b^{\left(l+1_{(\cdot)}\right)}$ denote the positive and negative embeddings of the b^th at the l^{th} layer; $a_{bt}^{(\cdot)}$ a transformation of $\mathbf{W}_{HPA}^{(\cdot)}$ and $\mathbf{Z}_{bt}^{\left(l_{(\cdot)}\right)}$, indicating the interaction coefficients among time steps and revealing the temporal-wise correlations for video b's popularity.

Based on the temporal-wise interaction embeddings, we not only identify the temporal-wise interaction, but also reduce the dimensions and attain the representation as $\mathbf{Z}^{(l_{(\cdot)})} \in \mathbb{R}^{b\times M\times d}$. Since the representations have been transformed into vectors, we follow Vaswani, Shazeer, and Parmar et al. (Vaswani et al., 2017) adopt the classic self-attention mechanism to capture modalities-wise interaction. Specifically, the deep representations from positive and negative embeddings are linearly transformed to obtain the query, key

and value matrix, denoted as $\mathbf{Q} \in \mathbb{R}^{b\times M\times d}$, $\mathbf{K} \in \mathbb{R}^{b\times M\times d}$ and $\mathbf{V} \in \mathbb{R}^{b\times M\times d}$ separately. And the outputs of the self-attention layer can be calculated as:

$$\mathbf{Z}^{\left(l+1_{(P)}\right)} = \frac{\mathbf{Q}^{(P)}\mathbf{K}^{(P)T}}{\sqrt{d}}\mathbf{V}^{(P)} \in \mathbb{R}^{b\times M\times d} \tag{7}$$

$$\mathbf{Z}^{\left(l+1_{(N)}\right)} = \frac{\mathbf{Q}^{(N)}\mathbf{K}^{(N)T}}{\sqrt{d}}\mathbf{V}^{(N)} \in \mathbb{R}^{b\times M\times d} \tag{8}$$

where $\sqrt{d}$ is for normalizing consideration. The weights for **Q**, **K** and **V** are trainable and adaptive to the loss function, reflecting the interaction among modalities. Consequently, the output of attention layer is the weighted sum of the modality representation and contains the interrelationship. Considering that the feature interaction of short video has the character of across-time step, a simple single-layer interaction will lose important information. Thus, the total weight coefficient of modality m at time step t can be written as

$$a_{mt} = a_m \cdot a_t \tag{9}$$

Concatenating the positive and negative embeddings that are generated by two different processes and keeping $\mathbf{Z}_{Final} \in \mathbb{R}^{b\times 6d}$ as the output allows us to encapsulate multi-level interaction as well as an across-level pattern. In the end, in order to attain the definitive value for prediction, we utilize a linear layer, followed by a sigmoid activation function.

In order to make the final prediction, we adopt binary-cross entropy. Given a dataset with b samples, the prediction loss can be written as:

$$L_1 = \frac{1}{b}\sum_{i=1}^{b}\left[y_i log\left(\hat{y}_i\right) + \left(1-y_i\right) log\left(1-\hat{y}_i\right)\right] \tag{10}$$

Furthermore, to better leverage the positive and negative information for MVPP, HACL enhances the performance by adding a contrastive loss. Unlike standard contrastive losses, whose calculation is at the sample level, we manage to realize a global contractiveness:

$$L_2 = \frac{1}{b}\sum_{i=1}^{b}\frac{\mathbf{z}_i^{(P)T}\mathbf{z}_i^{(N)}}{||\mathbf{z}_i^{(P)}||\,||\mathbf{z}_i^{T}||} \tag{11}$$

, where $\mathbf{z}_i^{(\cdot)}$ is the positive or negative embeddings of the i^{th} sample.

In summary, the total learning objective of HACL can be expressed as:

$$L_{Total} = L_1 + \lambda L_2 \tag{12}$$

Empirical Evaluation

In this section, we conduct numerical experiments to evaluate our proposed HACL model on binary popularity classification problem. To further justify the effectiveness of the proposed model, comparative experiments are delivered, followed by sensitivity analysis and ablative experiments.

Data

We have collected a micro video dataset from one of the most prominent micro-video sharing platforms, TikTok. In total, this dataset contains 8,790 user-generated micro videos, uploaded by 1,592 users. The length of all micro-videos is no longer than 60 seconds, with approximately 75 percent of the videos being 35 seconds. Since our goal is to predict the popularity, each of the videos is assigned a label equaling 0 or 1, indicating the video is not popular or popular.

Experiment Settings

In experiments evaluation, we adopt Accuracy, Precision, and AUC score to measure the consistency.

To even further demonstrate the efficacy of the proposed model, we compare proposed HACL to nine existing benchmarks approaches for predicting the popularity of micro-videos.

- Machine learning methods:
  - SVM (Cortes & Vapnik, 1995): Support vector machine is a classical classification method with a maximum margin criterion. We combined all the features from all modalities at all time steps as a unified input together and make a prediction of given micro videos.
  - MLP: multi-layer perception is a standard algorithm for supervised tasks. Exactly as the input used in SVM, distinct features are concatenated as one matrix, transiting into three linear layers with tanh function for activation.
- Graph neural network methods:
  - ARGF (Mai et al., 2020): Adversarial Representation Graph Fusion is a multimodal fusion framework. In this method, various modalities are jointly embedded into the representation space through adversarial learning, and the modalities are fused with the hierarchical graph network.
  - MAGNN (Cheng et al., 2022): The Multi-modality Graph Neural Network is originally used for financial prediction. We harness its graph structure and attention mechanism for MVPP as a benchmark.
- Other methods:
  - BBFN (Han et al., 2021): Bi-Bimodal Fusion Network is a framework that can fusion and separate modalities to gain better representations for the downstream tasks.
  - HMMVED (Xie et al., 2021): Hierarchical Multimodal Variational Encoder-Decoder adopt variational encoder-decoder framework encoding the input modalities to a lower dimensional stochastic variable. Nonetheless, on account of the absence of users' information in our datasets, we degenerate HMMVED into HMMVED without users' embeddings in our comparison.
  - MAG (Rahman et al., 2020): Multimodal Adaptation Gate provide fine-tuning operations for pre-trained models, especially the natural language processing tasks. We utilize its multimodal fusion ability as our benchmarks.
  - STAN (Cheng et al., 2020): The Spatial-Temporal Attention-based Neural Network measures the importance of time and space dimensions in neural networks based on spatiotemporal attention.
  - Soft-HGR (Wang et al., 2019): Soft-Hirschfeld-Gebelein-Rényi is a framework for extracting useful features from multiple data modalities based on the modality correlation.

Traditional classification methods such as SVM and MLP have been widely utilized. Besides, with the emergence and popularity of graph neural networks (GNN), researchers have started exploring their application in multimodal fusion. In this context, each modality or feature is treated as a node, while the edges represent the interactions between them. Notably, the adjacency matrix, formed by the product of query and key in attention mechanisms, leads to GNN yielding outcomes similar to attention mechanisms. Additionally, our comparison incorporates other methods like spatial-temporal attention (double self-attention), Variational Autoencoder (VAE), and various benchmarks to ensure comprehensive evaluation.

In addition to the primary experiments, we executed three types of exploratory experiments using HACL in order to get a deeper knowledge of its MVPP performance. First, we conducted ablation experiments by altering the input modalities to visual-acoustic, visual-text, and visual-only data. In ablation experiments, the impact of eliminating essential components such as the self-attention layer, Hadamard-product-based attention layer, and regulation item are also examined, shedding light on the predictive skills of each component. Second, we performed a sensitivity study on the main parameters in HACL λ to demonstrate how the regulatory item would improve MVPP performance. Thirdly, we conducted a visual analysis to give explanatory insights about the attractiveness of our suggested HACL and micro videos.

Experiment implementations, using Python 3.7 under TensorFlow-GPU 2.10 and Keras 2.10, were executed on a server with NVIDIA-RTX-2080Ti, 64-GB-RAM, 3.00-GHz, and Inter-Core-i7-9700-CPU.

Experiment Results

Prediction Performance

Table 2 summarizes the prediction performance of our proposed HACL versus the nine benchmarks versus the nine benchmarked multimodal deep learning methods.

Means of accuracy, precision, and AUC are reported along with the standard deviation shown in brackets.

Method	Accuracy	Precision	AUC
SVM	0.634 (0.017)	0.200 (0.400)	0.500 ((0.001)
MLP	0.634 (0.017)	0.058 (0.175)	0.500 (0.002)
ARGF	0.616 (0.016)	0.475 (0.043)	0.568 (0.011)
BBFN	0.608 (0.020)	0.460 (0.026)	0.567 (0.022)
HMMVED	0.491 (0.040)	0.366 (0.027)	0.493 (0.016)
MAG	0.615 (0.013)	0.470 (0.034)	0.570 (0.016)
MAGNN	0.610 (0.016)	0.463 (0.031)	0.566 (0.012)
SoftHGR	0.366 (0.018)	0.366 (0.028)	0.500 (0.001)
STAN	0.620 (0.016)	0.476 (0.029)	0.572 (0.014)
HACL	0.645 (0.023)	0.557 (0.089)	0.610 (0.017)

Table 2. Prediction Performance

Compared with SOTA, HACL, the method proposed in this study, has achieved stable prediction advantages. HACL holds the best accuracy, precision, and AUC performance among the given methods; they are increased by 1.7%, 17.0%, and 7.4% separately in comparison to the second-best methods (SVM(MLP), STAN, and ARGF). Although both SVM and MLP have the second highest accuracy, they may be overfitted which is deduced from their relatively low precision.

In order to further explore whether the performance results of the ten groups of experiments have significant differences, we conducted the Tukey test on the 10-fold cross-validation results. Table 3 reports the Tukey test results for this method and the comparison method. Taking AUC as an example, the differences between HACL and the other nine comparison methods are significant at 0.001 or below. The results show that the method proposed in this study is significantly better than the benchmarks in MVPP.

Method1	Method2	MeanDiff	p-value
HACL	SVM	-0.1093	0.001
	MLP	-0.1088	0.001
	ARGF	-0.0416	0.001
	BBFN	-0.0421	0.001
	HMMVED	-0.1165	0.001
	MAG	-0.0393	0.001
	MAGNN	-0.0436	0.001
	SoftHGR	-0.0376	0.001
	STAN	-0.1095	0.001

Table 3. Tukey Test Results

Ablation Experiments

Table 4 and Figure 3 demonstrate the ablation experimental results in MVPP. The ablation of HACL includes input modalities and model components and generates five ablated HACLs: (1) HACL-V: only visual modality is used in MVPP; (2) HACL-VT: visual and textual modalities are used in MVPP; (3) HACL-VA: visual and acoustic modalities are used in MVPP; (4) HACL-SA: the latent representations from all modalities are no longer fed into the Hadamard-product attention, but are first concatenated and then sent into the self-attention layer; and (5) HACL-D: the two independent pipelines degenerate single sequence of layers. As can be observed, although they have similar accuracy, but HACL is more stable and less likely to be overfitted. Consequently, HACL outperformed all of its ablated variants in MVPP, demonstrating the indispensable roles of all components in better leveraging multimodal data for prediction.

Method	Accuracy	AUC
HACL-V	0.645 (0.020)	0.551 (0.029)
HACL-VT	0.641 (0.024)	0.604 (0.038)
HACL-VA	0.648 (0.013)	0.568 (0.021)
HACL-SA	0.648 (0.015)	0.605 (0.014)
HACL-D	0.636 (0.019)	0.595 (0.016)

Table 4. Ablation Experiment Results

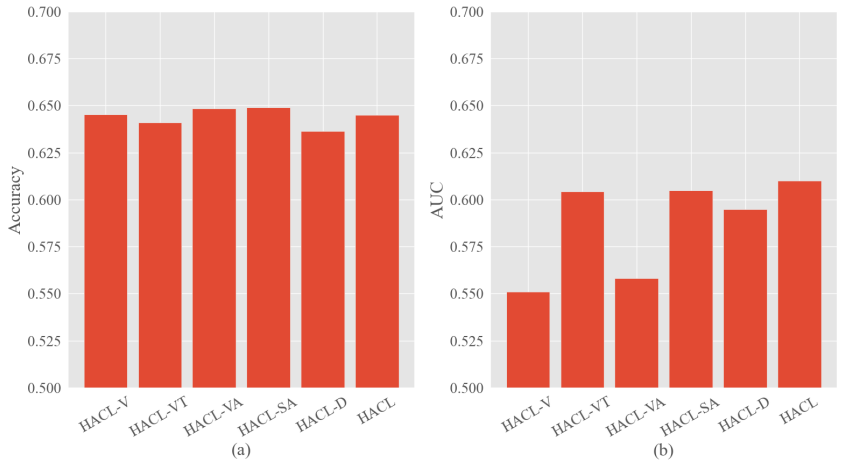


Figure 3. The Ablation Experimental Results

Sensitivity Analysis

Figure 4 shows the variation tendencies of HACL's prediction performance along with its contrastive regulation item hyper-parameter λ. The overall trend of HACL's sensitivity to λ is partially inverted "U-Shaped". Nevertheless, performance varies from metrics to metrics. The performance increased till a peak and then dropped as λ increased. To be more specific, HACL achieved the best prediction performance in the given settings, when λ equals to 0.16 based on Accuracy and Precision. Such result implies that our contrastive regulation item is influential to facilitate the MVPP task and verify that popular and unpopular micro videos hold different key latent representations for identification, emphasizing the importance of dual networks in the MVPP task.

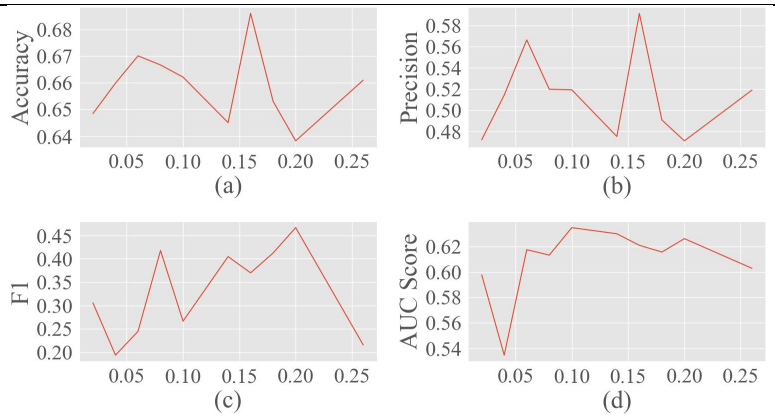


Figure 4. The Sensitivity Analysis

Visualization Analysis

In order to examine whether our proposed dual neural network works, we plot the attention coefficients at HPA, self-attention, and the tensor at the contrastive layer. Figure 5 demonstrates the interaction within time steps, identifying that users' interests vary from time to time. Y-label indicates which kinds of samples from which neural networks. For instance, TP is short for "total samples from positive network". Although there is limited difference of the coefficients between positive and negative in the middle stage. Middle stage is much more crucial for popular (positive) samples than that of unpopular (negative) one. Meanwhile, for negative samples, the fluctuation over time steps is relatively narrow, indicating that unpopular micro videos cannot appeals viewers' interests from time to time. In other words, users do not need to watch the whole micro video before deciding whether or not they enjoy it, which is consistent with reality.

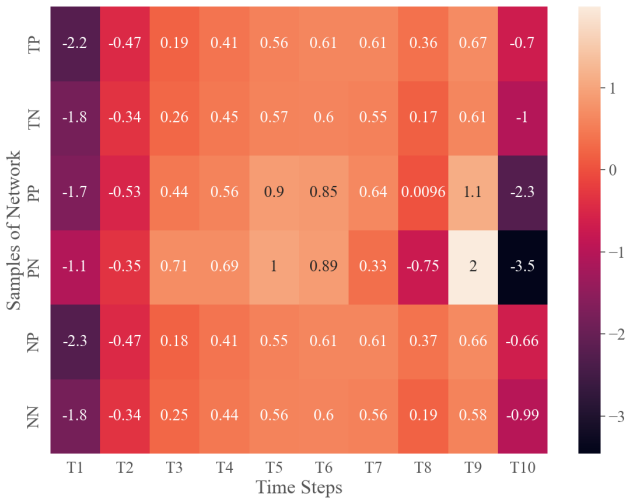
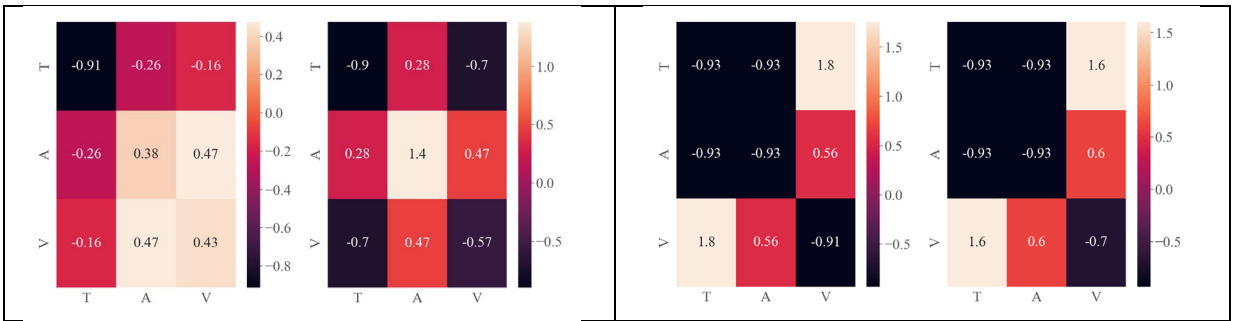


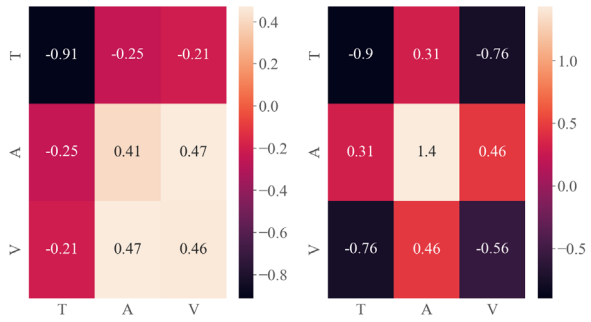
Figure 5. The Scaled Attention Coefficients (HPA)

Figure 6 illustrates the interaction among modalities in positive and negative neural network. Overall, positive neural network places a greater emphasis on visual information, whereas a negative neural network prioritizes acoustic information (see Figure 6(a)). For positive samples, the coefficients from the positive and negative neural network coefficients are not statistically different, but the interaction between textual and visual information is more active in positive neural network (see Figure 6(b)). In comparison to positive samples, the disparity of negative samples between the dual neural network is more pronounced. Interactions between modalities on the negative neural network are more powerful when compared to the positive neural networks (see Figure 6(c)). Besides, interaction coefficients across modalities vary between positive and negative samples. For example, in popular latent representations, the correlation between visual and textual information is higher than that of the unpopular. Furthermore, inconsistent visual and linguistic information will also contribute to the decline in popularity of video.



(a) Total samples from positive and negative neural networks

(b) Positive samples from positive and negative neural networks



(c) Negative samples from positive and negative neural networks

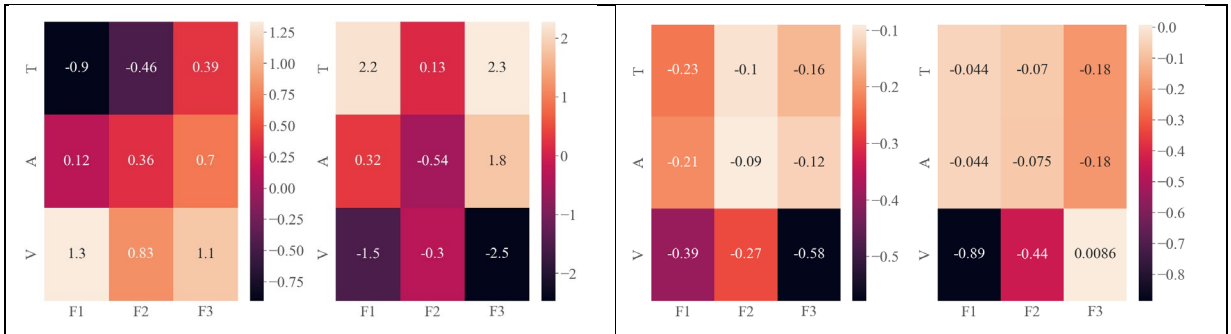
Figure 6. The Scaled Attention Coefficients (self-attention)

Figure 7 directly shows the latent representations of the dual neural network at the contrastive layer. In order to facilitate the display of visualized results, we use principal component analysis (PCA) to reduce the dimension of hidden layer features to 3. The explained variance ratios of generated features are deceits in Table 5, where the proportion of interpreted data exceeds 95%, implying that the reduced dimension data can be used to represent the latent representations.

Hidden representation	Modality	Explained variance ratio (%)			
		F1	F2	F3	Total
Positive neural network	T	99.500	0.235	0.067	99.802
	A	99.558	0.184	0.066	99.808
	V	98.349	0.695	0.047	99.091
Negative neural network	T	99.674	0.105	0.067	99.846
	A	99.682	0.105	0.063	99.850
	V	98.862	0.562	0.293	99.717

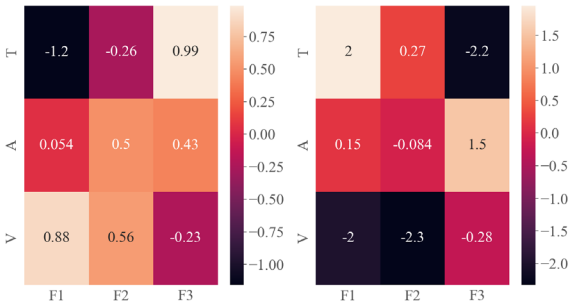
Table 5. The Results of PCA

It is obviously that positive sample embeddings vary from modality to modality, whereas negative embeddings tent to be the same among modalities. Consequently, we can find difference between positive and negative network, which justifies the effectiveness of our proposed dual neural network in HACL.



(a) Total samples from positive and negative neural networks

(b) Positive samples from positive and negative neural networks



(c) Total samples from positive and negative neural networks

Figure 7. The Tensor of Positive and Negative Networks

Contributions and Implications

This study's primary contribution is methodological. It offers the unique deep learning technique HACL, which draws power from two novel learning methods, namely Hadamard-product-based attention and dual network contrastive framework, to utilize the heterogeneous interaction across modalities and time steps. HACL provides a methodological instrument for aggregating temporal and spatial aspects to the MVPP research stream. Furthermore, the empirical findings in a real micro video dataset broaden our outlook about (1) on average in which time step of a micro video users are affected most, (2) to what extent popularity is determined by the modalities, (3) how the modalities are interacted with each other along with time steps.

Our experiments and explanatory analysis provide managerial insights. Implications from temporal and spatial interactions learned by HACL. The popularity characteristics for micro videos learned by HACL provide insights helping micro video generators consternate on crucial modalities and time steps, as well as better leverage the interaction among modalities. For instance, when filming new micro videos, the interaction generated by HACL offers micro video generators the opportunities to grasp whether their contents will be popular and to comprehend the mechanism behind micro video popularity. Implications from dual neural network learned by HACL. HACL considers the MVPP task in two different channels in HACL, separating the characteristics of popular and unpopular micro videos. This process guarantees that the numerical results can not only explains the reason of popularity, but also point out that why a certain micro video is not popular, which facilities the production of generators.

Conclusion

In this paper, we propose a Hierarchical Hadamard-product-based Attention Contrastive Learning (HACL) for the micro video popularity prediction task. The crucial dual neural network designed for contrastive learning guarantee our proposed HACL is capable of isolating positive and negative embeddings. Specifically, the similarity based contrastive loss is added to the objective function to separate the latent representations and gain better informative embeddings. Based on the dimensions of matrices at temporal-wise interaction, we design a Hadamard-product based attention to identify the coefficients among time steps. The learned hidden representations are then fed into classic self-attention for weighting the interaction among modalities. Finally, the positive and negative embeddings are concatenated and linearly projected to the output layer followed by a Sigmoid activation. Extensive experiments show the effectiveness of our proposed HACL in the field of MVPP. Our work considering micro video popularity as two problems, providing innovative insights for MVPP. Besides, it is not limited to MVPP, but transferable for other downstream tasks involving multimodal time series data. Nonetheless, social network information and video operation data are not involved in this paper with the limitation of the dataset. In the future, we will consider more external features and attempt to extend our model.

Acknowledgements

The authors would like to extend their heartfelt appreciation to the associate editor, and three anonymous reviewers for their valuable feedback, which greatly contributed to enhancing the quality of this paper. The authors would like to acknowledge the support received from the National Nature Science Foundation of China (grant numbers 71971067).

References

Abousaleh, F. S., Cheng, W. H., Yu, N. H., & Tsao, Y. (2021). Multimodal Deep Learning Framework for Image Popularity Prediction on Social Media. *IEEE Transactions on Cognitive and Developmental Systems*, *13*(3), 679-692.

Chen, G., Kong, Q., Xu, N., & Mao, W. (2019). NPP: A neural popularity prediction model for social media content. *Neurocomputing*, *333*, 221-230.

Chen, J., Song, X., Nie, L., Wang, X., Zhang, H., & Chua, T.-S. (2016). Micro Tells Macro: Predicting the Popularity of Micro-Videos via a Transductive Model. Proceedings of the 24th ACM international conference on Multimedia (pp. 898–907). Association for Computing Machinery.

Cheng, D., Xiang, S., Shang, C., Zhang, Y., Yang, F., & Zhang, L. (2020). Spatio-temporal attention-based neural network for credit card fraud detection. Proceedings of the AAAI Conference on Artificial Intelligence (pp. 362-369).

Cheng, D., Yang, F., Xiang, S., & Liu, J. (2022). Financial time series forecasting with multi-modality graph neural network. *Pattern Recognition*, *121*, 108218.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297.

Ding, S., Lin, L., Wang, G., & Chao, H. (2015). Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, *48*(10), 2993-3003.

Gu, J. (2020, 21-23 Oct. 2020). MMSPP: Multimodal Social Media Popularity Prediction. 2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS) (pp. 1-5).

Gu, Y., Li, X., Huang, K., Fu, S., Yang, K., Chen, S., . . . Marsic, I. (2018). *Human Conversation Analysis Using Attentive Multimodal Networks with Hierarchical Encoder-Decoder* Proceedings of the 26th ACM international conference on Multimedia (pp. 537–545). Association for Computing Machinery.

Han, W., Chen, H., Gelbukh, A., Zadeh, A., Morency, L.-p., & Poria, S. (2021). Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. Proceedings of the 2021 International Conference on Multimodal Interaction (pp. 6-15).

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., & Makedon, F. (2021). A Survey on Contrastive Self-Supervised Learning. *9*(1), 2.

Jing, P., Su, Y., Nie, L., Bai, X., Liu, J., & Wang, M. (2018). Low-Rank Multi-View Embedding Learning for Micro-Video Popularity Prediction. *IEEE Transactions on Knowledge and Data Engineering*, *30*(8), 1519-1532.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

Li, H., Ma, X., Wang, F., Liu, J., & Xu, K. (2013). *On popularity prediction of videos shared in online social networks* Proceedings of the 22nd ACM international conference on Information & Knowledge Management (pp. 169–178). Association for Computing Machinery.

Liao, D., Xu, J., Li, G., Huang, W., Liu, W., & Li, J. (2019). Popularity Prediction on Online Articles with Deep Fusion of Temporal Process and Content Features. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 200-207.

Ma, C., Yan, Z., & Chen, C. W. (2017). LARM: A Lifetime Aware Regression Model for Predicting YouTube Video Popularity. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 467–476). Association for Computing Machinery.

Mai, S., Hu, H., & Xing, S. (2020). Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. Proceedings of the AAAI Conference on Artificial Intelligence (pp. 164-172).

Morvant, E., Habrard, A., & Ayache, S. (2014). Majority Vote of Diverse Classifiers for Late Fusion. In P. Fränti, G. Brown, M. Loog, F. Escolano, & M. Pelillo, *Structural, Syntactic, and Statistical Pattern Recognition* (pp. 153-162). Springer Berlin Heidelberg.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). *Multimodal deep learning* Proceedings of the 28th International Conference on International Conference on Machine Learning (pp. 689–696). Omnipress.

Rahman, W., Hasan, M. K., Lee, S., Bagher Zadeh, A., Mao, C., Morency, L.-P., & Hoque, E. (2020, July). Integrating Multimodal Information in Large Pretrained Transformers.*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2359-2369). Association for Computational Linguistics.

Ramirez, G. A., Baltrušaitis, T., & Morency, L.-P. (2011). Modeling Latent Discriminative Dynamic of Multi-dimensional Affective Signals. In S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin, *Affective Computing and Intelligent Interaction* (pp. 396-406). Springer Berlin Heidelberg.

Shih, Y.-N., Huang, R.-H., & Chiang, H.-Y. (2012). Background music: Effects on attention performance. *Work*, *42*, 573-578.

Shutova, E., Kiela, D., & Maillard, J. (2016, June). Black Holes and White Rabbits: Metaphor Identification with Visual Features.*Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 160-170). Association for Computational Linguistics.

Trzciński, T., Andruszkiewicz, P., Bocheński, T., & Rokita, P. (2017). Recurrent Neural Networks for Online Video Popularity Prediction. In M. Kryszkiewicz, A. Appice, D. Ślęzak, H. Rybinski, A. Skowron, & Z. W. Raś, *Foundations of Intelligent Systems* International Symposium on Methodologies for Intelligent Systems (pp. 146-153). Springer International Publishing.

Trzciński, T., & Rokita, P. (2017). Predicting Popularity of Online Videos Using Support Vector Regression. *IEEE Transactions on Multimedia*, *19*(11), 2561-2570.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (pp. Curran Associates, Inc.

Wang, B., Huang, X., Cao, G., Yang, L., Wei, X., & Tao, Z. (2022). Attention-enhanced and trusted multimodal learning for micro-video venue recognition. *Computers and Electrical Engineering*, *102*, 108127.

Wang, L., Wu, J., Huang, S.-L., Zheng, L., Xu, X., Zhang, L., & Huang, J. (2019). An efficient approach to informative feature extraction from multimodal data. Proceedings of the AAAI Conference on Artificial Intelligence (pp. 5281-5288).

Xie, J., Zhu, Y., & Chen, Z. (2021). Micro-video Popularity Prediction via Multimodal Variational Information Bottleneck. *IEEE Transactions on Multimedia*, 1-1.

Xie, J., Zhu, Y., Zhang, Z., Peng, J., Yi, J., Hu, Y., . . . Chen, Z. (2020). *A Multimodal Variational Encoder-Decoder Framework for Micro-video Popularity Prediction* Proceedings of The Web Conference 2020 (pp. 2542–2548). Association for Computing Machinery.

Yang, X., Feng, S., Wang, D., & Zhang, Y. (2021). Image-Text Multimodal Emotion Classification via Multi-View Attentional Network. *IEEE Transactions on Multimedia*, *23*, 4014-4026.

Zhang, Y.-D., Dong, Z., Wang, S.-H., Yu, X., Yao, X., Zhou, Q., . . . Gorriz, J. M. (2020). Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation. *Information Fusion*, *64*, 149-187.