Multimodal cooperative learning for micro-video advertising click prediction

Runyu Chen
*School of Information Technology and Management,
University of International Business and Economics, Beijing, China*

Received 13 July 2020
Revised 6 December 2020
3 January 2021
25 May 2021
Accepted 25 May 2021

Abstract

Purpose – Micro-video platforms have gained attention in recent years and have also become an important new channel for merchants to advertise their products. Since little research has studied micro-video advertising, this paper aims to fill the research gap by exploring the determinants of micro-video advertising clicks. We form a micro-video advertising click prediction model and demonstrate the effectiveness of the multimodal information extracted from the advertisement producers, commodities being sold and micro-video contents in the prediction task.

Design/methodology/approach – A multimodal analysis framework was conducted based on real-world micro-video advertisement datasets. To better capture the relations between different modalities, we adopt a cooperative learning model to predict the advertising clicks.

Findings – The experimental results show that the features extracted from different data sources can improve the prediction performance. Furthermore, the combination of different modal features (visual, acoustic, textual and numerical) is also worth studying. Compared to classical baseline models, the proposed cooperative learning model significantly outperforms the prediction results, which demonstrates that the relations between modalities are also important in advertising micro-video generation.

Originality/value – To the best of our knowledge, this is the first study analysing micro-video advertising effects. With the help of our advertising click prediction model, advertisement producers (merchants or their partners) can benefit from generating more effective micro-video advertisements. Furthermore, micro-video platforms can apply our prediction results to optimise their advertisement allocation algorithm and better manage network traffic. This research can be of great help for more effective development of the micro-video advertisement industry.

Keywords Micro-video, Multimodal analysis, Advertising click prediction, Cooperative learning

Paper type Research paper

1. Introduction

With the rapid development of web services and mobile devices, micro-video platforms have appeared and then gained considerable attention in recent years. Micro-video platforms allow users to record their daily lives in video clips within a couple of seconds. Because of the properties of instant sharing and interesting content, user-generated micro-videos take up users' fragmented spare time in our fast-paced modern society, which makes these platforms immensely popular (Wei *et al.*, 2020). According to the statistical data of QuestMobile [1], the number of monthly active users of Douyin [2] (one of the most influential user-generated micro-video platforms in China) reached 518 million in March 2020. As the international version of Douyin, TikTok [3] was the most downloaded app in Q1 2020, and its total downloads exceeded 315 million in the App Store and Google Play [4]. The explosive growth of micro-video platform users also attracts merchants. Many merchants explore effective ways to advertise their products on micro-video platforms. Since micro-video platforms are an emerging advertising channel, improving the advertising effect is an important challenge for both merchants and researchers.

This work was supported by "the Fundamental Research Funds for the Central Universities" in UIBE (20QD23).

Internet Research
Vol. 32 No. 2, 2022
pp. 477-495
© Emerald Publishing Limited
1066-2243
DOI 10.1108/INTR-07-2020-0388

In micro-video advertising, the commodity being sold is usually involved in a user-generated advertising micro-video, along with a purchase link at the bottom of the video. If the micro-video viewer is interested in the commodity being sold, a details page will be shown after clicking the purchase link button. Similar to traditional online advertising platforms, merchants generate their advertisements in micro-video platforms, and the service platform allocates the advertisement resources to platform users (Goldfarb and Tucker, 2011). In micro-video platforms, merchants also seek third-party partners to generate their advertising micro-videos. These cooperating people usually have a vast number of followers on the platform, and the advertising micro-video contents can be generated by either the merchant or the partner. Previous studies have demonstrated that the effect of an online advertisement depends on its creative elements and context (Liuthompkins, 2019). As a fast-growing online advertising channel, a micro-video advertisement contains textual advertising descriptions within a video that is a few seconds long. The determinants of the micro-video advertising effect can be complex but extremely important for the micro-video advertisement industry.

Since micro-videos with high popularity spread rapidly amongst users, a precise prediction method of micro-videos' popularity will benefit many applications, such as network bandwidth management and online advertising (Chen *et al.*, 2016b, 2018b; Jing *et al.*, 2018). However, unlike general micro-video content, popularity is an important but noisy indicator of advertising content. Some indicators that better represent the advertising effect (e.g., commodity clicks) are more valuable to study (Li *et al.*, 2015; Andrews *et al.*, 2016). Since a good advertising effect brings greater benefits, both scholars and entrepreneurs seriously examine the determinants of purchase intention in online advertising (Shaouf *et al.*, 2016; Cheah *et al.*, 2019; Kusumasondjaja and Tjiptono, 2019). As an important new online advertising channel, few researchers have studied the determinants of the micro-video advertising effect. To fill this research gap, we seek to answer the research question regarding how to generate a micro-video with a good advertising effect. Specifically, we proposed a micro-video advertising click prediction model that innovatively extracts multimodal information from advertisement producers, commodities being sold and micro-video content. The experimental results of the prediction model can provide us with valuable insights to generate more effective micro-video advertisements.

In summary, our study established a micro-video advertising click prediction model. There were three main contributions of our research work. First, we explored the determinants of micro-video advertising clicks based on real-world micro-video advertisement datasets. Second, in our prediction model, we extracted multimodal information, including visual, acoustic, textual and numeric features, from advertisement contents based on multiple data sources. Finally, we established a multimodal cooperative learning model based on a relation-aware attention mechanism. The empirical analysis demonstrated the effectiveness of our model in the prediction task. To the best of our knowledge, this is the first study conducting micro-video advertising effect analysis. The managerial implications of our research are as follows. With the help of our advertising click prediction model, advertisement producers (merchants or their partners) can benefit from generating more effective micro-video advertisements. Furthermore, micro-video platforms can apply our prediction results to optimise their advertisement allocation algorithm and better manage their network traffic. This research can greatly assist the more effective development of the micro-video advertisement industry.

The remainder of this paper is organised as follows. Section 2 summarises previous studies related to product advertising and micro-video multimodal analysis. Section 3 describes the proposed multimodal cooperative learning framework for micro-video advertising click prediction. Section 4 contains the experimental results and a discussion of our experiments. In the last section, we present concluding remarks and the future directions of our research work.

2. Literature review

2.1 Product advertising in an online content

Marketing managers develop plans to enhance the competitiveness of their products. Boulding *et al.* (1994) found that common marketing actions such as advertising, promotions and sales force activities increased product differentiation. Based on less negative price elasticities, these products had an advantage from future price competition. As time progresses, consumers were more sensitive to price and promotion (Mela *et al.*, 1997). In the long term, advertising has a positive effect on brand equity, while promotions have a negative effect (Jedidi *et al.*, 1999). Erdem *et al.* (2008) demonstrated that frequent price cuts had significant negative effects on brand equity. Advertising frequency was also significant in signalling brand quality, but it was less important than price. For new product sales, Chaudhuri *et al.* (2018) found that promotional strategies were effective in sales improvement. Cash rebates were more valid for mass goods, while financing incentives offered more benefits for luxury goods. The spillover effects of advertising and promotions were also well discussed (Erdem and Sun, 2002). Advertisements were found to significantly increase the sales of non-advertised restaurants (Sahni, 2016). In addition, promotions for one drug could also increase the demand for other drugs that were usually bundled with the promoted drug (Liu *et al.*, 2017a).

Compared with traditional marketing activities, personalised advertising is more accurate at addressing the target audience. Based on the big data analytics of customer behaviours, customers' requirements and references can be inferred. Therefore, companies can provide more suitable advertisements for different people (Liuthompkins, 2019). Similar to targeted advertising, targeted discounts can also be an effective promotional strategy (Sahni *et al.*, 2017). Chen and Stallaert (2014) studied the economic implications of engaging in behavioural targeted advertising. They identified a competitive effect and a propensity effect in behavioural targeting. Bleier and Eisenbeiss (2015) studied the timing and placement factors in personalised advertising effectiveness. Personalisation was demonstrated to increase ad effectiveness only on motive congruent websites. Some of the advertisements are displayed based on the content that consumers view, and these are called contextual advertisements (Zhang and Katona, 2012). It has been demonstrated that contextual advertisements enhance brand recognition. Moreover, consumers had higher recall rates when they are exposed to less complex contextual advertisements (Chun *et al.*, 2014). In our study, product advertisements are usually involved in micro-video content, which can be categorised into contextual advertisements.

System design and resource allocation in targeted advertising systems have also been studied (Bilenko and Richardson, 2011; Bimpikis *et al.*, 2016). Since mobile devices contain more location information, mobile advertising systems were usually considered separately (Li and Du, 2012; Zhang *et al.*, 2019). Based on the designed advertising system, many researchers studied customer behaviours and advertising performance. Chatterjee *et al.* (2003) analysed consumer response behaviours to banner advertisements. The results showed that click-prone consumers clicked banner advertisements faster and browsed through fewer pages than less click-prone consumers. Li *et al.* (2014) proposed a personalised advertising mechanism based on social influence measures and context embellishment. The experiment conducted on Facebook showed that the proposed method was able to improve the click-through rate (CTR) and user impressions. Goh *et al.* (2015) studied the relationships between advertisement content, information search behaviour and advertising responses on a mobile platform. The results showed that users' responses to mobile advertising were influenced by both the breadth and depth of the search and the advertisement content. Andrews *et al.* (2016) examined consumer responses to mobile ads in situations with physical crowdedness. The experimental results showed that a mobile advertisement was a welcome relief in a crowded subway environment. Therefore, crowded environments may be a crucial setting for marketing managers to improve mobile advertising effectiveness. In this paper, we study advertising effectiveness on micro-video platforms. Our study supplements the online advertising literature on micro-video platforms based on mobile devices.

2.2 Online advertising effect prediction

The advertising effect is one of the most important factors in online advertising. Therefore, researchers explored what influences the effectiveness of online product advertising. Goldfarb and Tucker (2011) found that both advertisements' conspicuousness and degree of content matching improved the effectiveness. Ching *et al.* (2013) investigated the effects of some design elements of online advertising, which include interactivity, entertainment, vividness and self-referencing. Shaouf *et al.* (2016) focused on the effectiveness of advertising's visual design. The results showed that visual cues had a significant effect on purchase intentions for males but not for females. Cheah *et al.* (2019) observed how celebrities endorse advertisements and self-promotion influence consumers' decisions. Kusumasondjaja and Tjiptono (2019) tested the effectiveness of celebrity and expert endorsers in Instagram advertising.

To accurately evaluate the advertising effect, many researchers have proposed prediction frameworks based on real-world datasets. As a pivotal indicator of the advertising effect, the CTR is usually referenced for advertisement ranking, allocation and pricing; thus, it is widely used as the prediction target (Chen and Yan, 2012). Li *et al.* (2015) presented a click-through prediction for advertising on Twitter timelines. Placing advertisements into a Twitter thread is a complex task since a suitable placement is dynamically updated. Therefore, they proposed a learning-to-rank method that can be trained and updated online. Qu *et al.* (2016) proposed a product-based neural network (PNN) to predict users' responses to online advertisements. Compared to traditional models, it performed significantly better when faced with extremely sparse data. Chen *et al.* (2016a, b) focused on the CTR prediction of image advertisements in online advertising systems. They proposed a deep neural network model that combines raw image pixels and other basic features in one step. To model the interactions between the different features of multi-field categorical data, Pan *et al.* (2018) proposed field-weighted factorisation machines that sharply reduced the number of parameters compared to field-aware factorisation machines. Ouyang *et al.* (2019) designed a deep spatiotemporal neural network (DSTN) for CTR prediction. The model is able to learn the interactions between auxiliary data and fuses these heterogeneous data in a unified framework. Pan *et al.* (2019) proposed a meta-learning-based approach that learns to generate initial embeddings for new advertisements. The proposed method can speed up the model fitting and thus solve the cold-start problem on some level. Table 1 summarises the representative studies of advertising click prediction in chronological order. Compared to previous studies, in addition to the features extracted from merchants and commodities, we innovatively extract acoustic and visual features from micro-video advertising contents. The proposed multimodal cooperative learning method also contributes to fusion methods in the field of advertising click prediction.

2.3 Micro-video multimodal analysis

Micro-video platforms usually allow users to generate their own videos within a few seconds. User-generated micro-videos take up users' fragmented spare time, which has made these platforms immensely popular in recent years. As a special type of media, micro-videos have some similar properties to traditional long videos. Technically speaking, micro-video analysis is usually treated as a multimodal fusion problem that contains visual, textual and acoustic modalities (Wei *et al.*, 2020). In micro-video analysis, the main focus of the research is

Studies	Feature types	Fusion methods
Li *et al.* (2015)	Numerical features from ads, users and ad–user interactions and textual features from stream contexts	Learning-to-rank method
Chen *et al.* (2016b)	Numerical features from ads and users and visual features from images	Novel deep learning network-based model
Qu *et al.* (2016)	Numerical features from users, publishers and ads	Product-based neural network
Guo *et al.* (2017)	Numerical features from ads, users and contexts	DeepFM
Pan *et al.* (2018)	Numerical features from users, publishers, advertisers and the context	Field-weighted factorisation machine
Zhou *et al.* (2018)	Numerical features from user profiles, user behaviours, ads, users and contexts	Deep interest network
Ouyang *et al.* (2019)	Numerical features from ads, users and queries	Deep temporal neural network
Pan *et al.* (2019)	Numerical features from ads, users and queries and textual features from ads	Meta-learning-based approach

Table 1.
A brief review of advertising click prediction

micro-video understanding. Liu *et al.* (2017b) developed an end-to-end deep learning model to classify micro-videos into 22 Foursquare venue categories. The model contains three parallel Long Short Term Memory (LSTMs) to capture the sequential structures and a convolutional neural network to represent the sparse concept. Nie *et al.* (2017) enhanced the acoustic modality by harnessing the external sound knowledge for the venue category estimation task. To alleviate the sparsity problem, they further regularised the representation learning method for the same category. Liu *et al.* (2019c) applied NNeXtVLAD layers to aggregate visual, textual and acoustic features. Then, a Convolutional Neural Network (CNN) layer was used to represent the concept level, while context gating captures the network interdependency. Based on concept-level micro-video representation, Liu *et al.* (2019a) further developed an online learning algorithm to categorise micro-videos. In addition, Guo *et al.* (2019) presented a scene retrieval method for micro-videos. A combinational fusion method that combines a neural network and supervised hash learning was proposed to better extract semantic features.

Researchers have also studied micro-video analysis application topics, such as micro-video recommendation and popularity prediction. To form a quality micro-video recommendation service, Wei *et al.* (2019b) designed a multimodal graph convolution network that captures user–item interactions. Specifically, a user–item bipartite graph in each modality was separately constructed, while the representation of each node considers both its topological structure and neighbours' features. In addition, Wei *et al.* (2019a) proposed a personalised hashtag recommendation system for micro-videos based on a graph convolution network. Liu *et al.* (2019b) proposed a user-video co-attention network that can capture multimodal features from both users and videos based on an attention mechanism. Li *et al.* (2019) considered users' diverse and dynamic interests in micro-video platforms. To address these issues, they proposed a temporal graph-guided recommendation system and trained the model using users' true negative data samples. Amongst the massive number of micro-videos, popular micro-videos received more attention and benefits. Chen *et al.* (2016a) proposed a transductive multimodal learning model to predict micro-video popularity. The optimal common space from different modalities was extracted to represent micro-videos in the model. Furthermore, the authors also constructed a large-scale micro-video dataset, which could support many research topics, including popularity prediction (Chen, 2016). Jing *et al.* (2018) proposed a transductive low-rank multi-view regression framework by jointly combining the representations of the source and target samples. The framework sought projection matrices to map multi-view features into a common subspace. In addition, a multigraph regularisation term was designed to prevent the overfitting problem. Chen *et al.* (2018b) proposed a temporal hierarchical attention network at the category and item levels to model users' historical behaviours. The click-through prediction task was then evaluated using a new dataset of 1.7 million micro-videos coming from China.

Compared to other deep learning models, the multimodal deep learning model can capture more valuable information from micro-video advertisements. The existing literature on micro-video analysis shows that multimodal features (including visual, acoustic, textual and numerical features) all contain valuable signals. The multimodal deep learning model can simultaneously process these features, thus enhancing the prediction performance. Therefore, we implement a multimodal cooperative learning model for micro-video advertising click prediction. Compared to the existing research studies, our research focuses on advertising effect analysis but not micro-video popularity. In addition, the features extracted from advertisement producers and commodities are also considered in our multimodal analysis model.

3. A micro-video advertising click prediction framework

Although previous studies have conducted some analysis on micro-video contents (Chen *et al.*, 2016a; Liu *et al.*, 2017b), little work has studied advertising micro-videos. In this paper, from the perspective of advertising click prediction, we conducted our analysis considering not only the micro-video contents but also the information on the commodities being sold and the advertisement producers' profile. After data pre-processing and feature extraction, we constructed a deep learning model that comprises a cooperative net and an attention net to fuse multimodal information. Based on our multimodal cooperative learning model, the advertising clicks of a micro-video can be predicted. The proposed micro-video advertising click prediction framework is outlined in Figure 1, and the details of the framework are expanded in the following sections.

3.1 Data collection

Previous studies have demonstrated that the features extracted from ads and users are effective for advertising click prediction (Pan *et al.*, 2019). Therefore, as Figure 1 shows, for

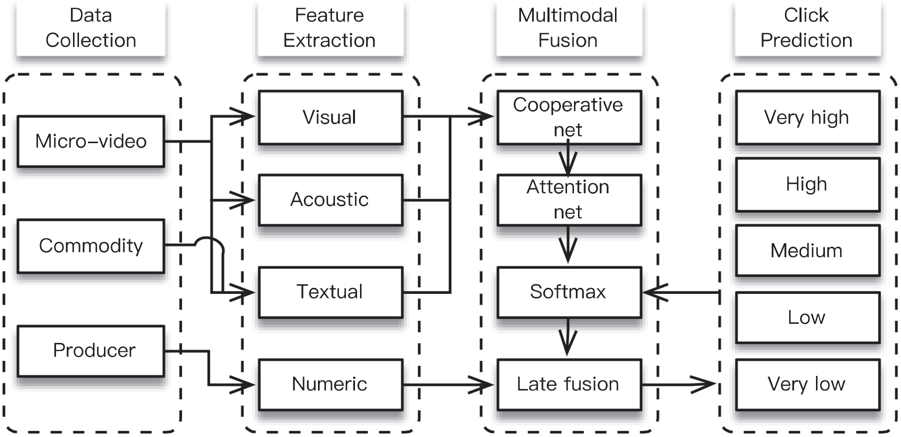

Figure 1.
Micro-video advertising click prediction framework

each advertising micro-video, we collected all of the data that a micro-video viewer can obtain. Specifically, for micro-videos, we collected the micro-video title, the micro-video poster and all multimodal information about the contents. For the commodity being sold, we collected the textual commodity description on the click button. Since we aimed to study the micro-video advertising effect estimated by the number of commodity clicks, the commodity information that can be shown only after clicking (such as the commodity type and the price) was not considered in this study. For advertisement producers, although the details about producers are not shown on the main page, we still collected producers' profile data for the following two reasons. First, a micro-video platform allocates different resources for different advertisement producers; therefore, the producers' profiles may influence the advertising effect (Liuthompkins, 2019). Second, micro-video viewers can prejudge producers' reputations based on their profiles before making a purchase decision (Chen *et al.*, 2018a); thus, producers' reputations are also likely to influence advertising clicks. In summary, we collected data about the advertisement producers, the commodities being sold and the micro-video contents. The details about the collected data and the descriptions are listed in Table 2.

3.2 Feature extraction

After data collection, we pre-processed the data and extracted multimodal features. The existing literature on micro-video analysis showed that multimodal features (including visual, acoustic, textual and numerical features) all have a significant influence on micro-video popularity (Jing *et al.*, 2018). For micro-video advertisements, a micro-video with very popular content is also more likely to have more commodity clicks. To capture more signals from micro-video contents, following the popularity prediction literature, we extracted visual, acoustic, textual and numerical features and then designed a multimodal deep learning model to predict advertising clicks. As shown in Table 2, from the data type perspective, the features extracted from advertisement producers are all numerical features. Textual features are embedded in commodity descriptions and micro-video titles. The micro-video poster is a visual image, and the micro-video content contains both visual and acoustic data. Compared to traditional online advertising platforms, the multimodal features extracted from micro-videos are unique and rarely applied in previous advertising click prediction tasks. To improve the effectiveness and efficiency of the feature

Data source	Data item	Description
Advertisement producer	Follower count	The number of followers of the advertisement producer
	Personal certification	Whether the advertisement producer has a personal certification or not
	Enterprise certification	Whether the advertisement producer has an enterprise certification or not
	Contact information	Whether the advertisement producer shows its contact information in its self-introduction
Commodity being sold	Commodity description	High-dimensional representative information of the textual description of the commodity being sold
Micro-video	Micro-video title	High-dimensional representative information of the textual description about the micro-video
	Micro-video poster	High-dimensional representative information of the picture poster of the micro-video
	Micro-video content	High-dimensional representative information of the content of the micro-video, which consists of visual data and acoustic data

Table 2.
Data item descriptions

representation, we applied some pre-trained models to extract the multimodal features. It has been fully tested that the features extracted by these pre-trained models have a good representation of the raw multimedia data (such as images, audio and textual data) (Simonyan and Zisserman, 2015; Chen *et al.*, 2016a; Mikolov *et al.*, 2013). The details about the multimodal features are introduced in the following subsections.

3.2.1 Visual features. Deep convolutional neural networks have been widely proven to have excellent performance in extracting visual features (Wei *et al.*, 2016; Ren *et al.*, 2017). As a classical deep convolutional neural network model, VGGNet has achieved outstanding performance in multiple transfer learning tasks; therefore, it has been widely applied to extract features from images (Simonyan and Zisserman, 2015). In this paper, we chose the pre-trained VGG16 model to extract our visual features. For micro-video contents, we first selected the keyframes of the micro-video by using OpenCV [5]. The poster picture was specified as the first keyframe. Then, the VGG16 model was applied to extract the features from each frame. The mean pooling strategy was applied for all keyframes of the micro-video, and the output was a 1024-dimensional feature set.

3.2.2 Acoustic features. In addition to visual data, micro-video content also contains acoustic information. In many video-related tasks, the acoustic modality was processed to enhance the experimental effect (Wu *et al.*, 2014). In this paper, we considered two types of widely used acoustic features: Mel-frequency Cepstral Coefficients (MFCCs) and psychoacoustic features (e.g., loudness, roughness, sharpness and tonality features) (Li *et al.*, 2013; Chen *et al.*, 2016a). For each micro-video, we ultimately constructed a 100-dimensional acoustic feature set.

3.2.3 Textual features. Textual descriptions usually contain a critical topic or sentiment about the micro-video (Chen *et al.*, 2016a). Considering the short length of the descriptions, we utilised the paragraph vector method to alleviate the semantic problems of word sparseness. The sentence2vector [6] tool, which was developed using the pre-trained word embedding method word2vector, was employed in this study (Mikolov *et al.*, 2013). For each textual description, we extracted 100-dimensional features in this way.

3.2.4 Numeric features. To comprehensively consider the influential factors, we also extracted the features of the advertisement producer. Except for the number of followers, the other features were all processed as binary features. Specifically, the contract information was filtered from the producer's textual self-introduction. For each advertising micro-video, four numerical features were extracted and inputted into the prediction model.

3.3 Multimodal fusion

After extracting multimodal features, we fused these features together to predict the advertising clicks. Compared to traditional models, the multimodal deep learning model can capture different information from different modalities, thus enhancing the prediction performance. Most of the previous work constructs a joint representation, which captures the common cues over multiple modalities, for all of these multimodal features. To better capture the relations between different modalities, we adopted the cooperative learning model that was proposed by Wei *et al.* (2020). In our fusion model, there were three cooperative nets, which set one modality (visual, acoustical or textual) as the host, and the others as guests. The outputs of the cooperative net are then fed into the attention net, which captures the important correlations between different modalities. Then, the prediction results of the three softmax functions were averaged with the predicted results of the numerical features. The details are described in the following subsections.

3.3.1 Cooperative networks. Classical multimodal fusion models consist of early fusion and late fusion (Ngiam *et al.*, 2011). Compared to early fusion or late fusion methods, multi-view learning can better capture the correlations between different modalities. Neural multimodal cooperative learning is a multi-view learning model proposed by Wei *et al.* (2020). It is clearly defined that the cooperative relationship amongst multimodalities comprises consistent and complementary components. The cooperative network can automatically distinguish and fuse these two components amongst multiple modalities by treating one modality as the host and the rest as the guests. The experimental results showed that the F1 score of the neural multimodal cooperative learning model outperformed that of the other baseline models by more than 6%; therefore, we adopted the cooperative networks in our prediction model.

As shown in Figure 2, in each cooperative net, one modality is treated as the host, and the other two are guests. There are three separate cooperative nets whose host modalities are the visual, acoustical and textual features in our network. The structure of the cooperative net is symmetric. On the left side of Figure 2, first, two guest modalities are concatenated together, and then the relevance between the concatenated vector and the host-vector is estimated. Based on the attention mechanism, we estimated the correlation between each host-vector dimension and the guest features. The output of the function is a score vector for each host dimension, where the value represents the correlation between a host feature and all guest information. Then, the consistent part and complementary part of the output vector are separated through the gate with a learned threshold. For the guest modality on the right side, a similar function is conducted to separate the consistent part and complementary part. Two consistent parts are concatenated together and fed into a fully connected neural network, and the outputs of the cooperative net are obtained by concatenating the complementary parts with the consistent part representations.

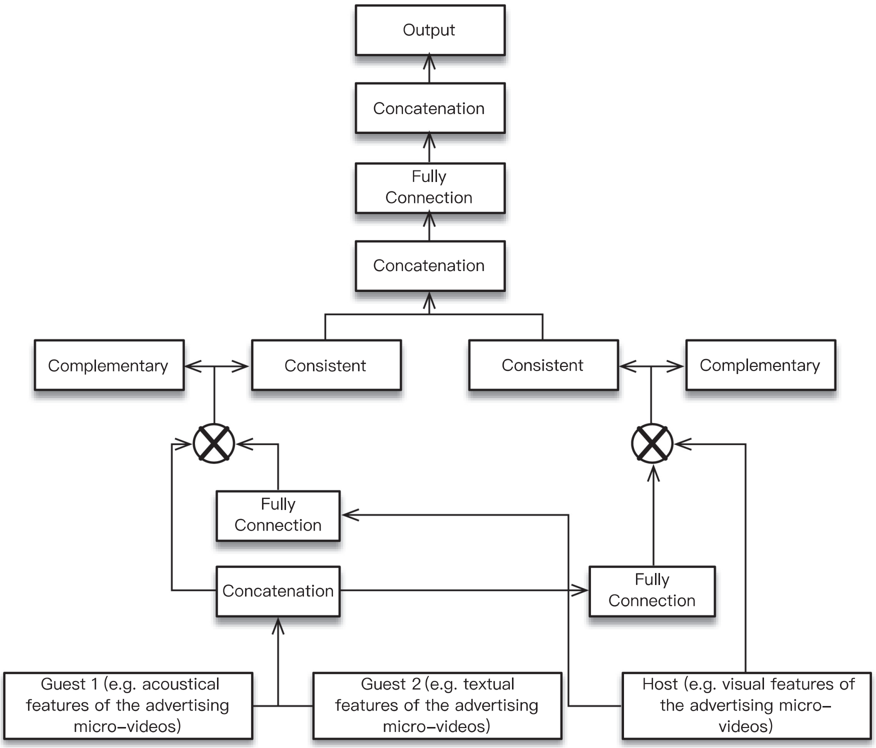
Output
Concatenation
Fully Connection
Concatenation
Complementary
Consistent
Consistent
Complementary
Fully Connection
Concatenation
Fully Connection
Guest 1 (e.g. acoustical features of the advertising micro–videos)
Guest 2 (e.g. textual features of the advertising micro–videos)
Host (e.g. visual features of the advertising micro– videos)

Figure 2.
Cooperative net architecture

3.3.2 Attention networks. The cooperative network captures rich multimodal information; however, some parts of the representations may be redundant for this specific prediction task. To remedy this issue, we designed an attention network (Vaswani *et al.*, 2017) to connect to the cooperative network. The attention network estimates a score for each representation feature towards different advertising click classifications, which makes some important features more effective in the network. Eventually, we obtained discriminative representations and then fed them into a softmax layer.

3.3.3 Late fusion. Each cooperative net has its own prediction result through the fully connected softmax layer. In other words, there are three prediction results generated by the cooperative nets whose host modalities are visual, acoustical and textual. In addition, we also used a random forest (RF) model (Breiman, 2001) to predict advertising clicks based on the numerical features extracted from advertisement producers. The advertising micro-videos were divided into five classes according to the number of commodity clicks. The result of the late fusion was the average of the four prediction results.

4. Empirical study and results

4.1 Data description and evaluation criteria

The data we used in the experiments were collected from Douyin. Based on a data analytics platform [7], we first collected the user list of those who generated the top 500 most influential advertising micro-videos on Douyin each day for one week from 9 March 2020 to 15 March 2020. The dataset contains 2007 micro-video producers, and then all of the advertising micro-videos generated by these producers in the last 90 days were collected. After analysing the incremental changes in the commodity clicks of the newly released advertising micro-videos, we removed the micro-videos generated within 7 days to reduce errors caused by different release durations. The cleaned dataset contains a total of 23,378 advertising micro-videos. According to the advertising effects, we classified the advertising micro-videos into five classes in the order of their commodity clicks. As a result, 4,677 were labelled "very high", 4,675 were labelled "high", 4,678 were labelled "medium", 4,672 were labelled "low" and 4,676 were labelled "very low". The datasets were randomly divided into three parts: 80% was used as the training set, 10% was used as the validation set and the remainder was used as the testing set.

To evaluate the effectiveness of the extracted multimodal features and the proposed prediction model, we adopted four common performance measures: accuracy, precision, recall and F1 score (Manning *et al.*, 2008).

4.2 Performance comparison amongst modality combinations

4.2.1 Data source combinations. In our proposed method, we extracted features from different data sources, including advertisement producers, commodities being sold and micro-video contents. To measure the effectiveness of the features extracted from different data sources, we conducted experiments based on different data source combinations. For each data source, the feature extraction methods were in accordance with the descriptions in Section 3.2. After feature extraction, one fully connected layer and a softmax layer were applied as the classification model. Since the number of features extracted from advertisement producers was proportionately small, we applied late fusion methods to prevent valuable information from being missed. Therefore, when combining different modal features or different data sources, we first implemented the classifier for each respective modality and then averaged the prediction results. The experimental results are shown in Table 3.

As Table 3 shows, compared to the commodity being sold and the micro-video content, the features extracted from advertisement producers are highly effective in advertising click prediction. In addition, the combinations of different data sources can significantly improve

Data sources	Accuracy	Precision	Recall	F1 score
Features extracted from advertisement producers	0.51	0.54	0.58	0.56
Features extracted from commodities being sold	0.32	0.29	0.26	0.27
Features extracted from micro-video contents	0.38	0.37	0.34	0.35
Features extracted from advertisement producers + commodities being sold	0.55	0.58	0.64	0.61
Features extracted from advertisement producers + micro-video contents	0.62	0.65	0.71	0.66
Features extracted from commodities being sold + micro-video contents	0.47	0.46	0.59	0.52
Features extracted from advertisement producers + commodities being sold + micro-video contents	0.67	0.69	0.73	0.71

Table 3. Prediction performance for data source combinations

the prediction performance. When we combine features extracted from all three data sources (advertisement producers, commodities being sold and micro-video contents), the experimental results show the best performance.

4.2.2 Multimodal feature combinations. In the proposed advertising click prediction model, we considered multimodal features extracted from different data sources. Especially for micro-video contents, we extracted visual, acoustic and textual features. To demonstrate the prediction effectiveness of these modalities, we extracted the features from these modalities and train different prediction models. The feature extraction methods for different modalities accord with the former descriptions, and then the same modality features extracted from different data sources were simply concatenated. Since the numerical features extracted from advertisement producers are important for advertisement allocations, we considered these features as the baseline model and then added other modalities. When combining different modalities, we implemented the late fusion method and finally averaged the prediction results. The prediction performance can be seen in Table 4. Amongst these three modalities, textual features show the best effectiveness in our experiments, and acoustic features perform the worst. Meanwhile, modality combinations significantly outperform a single modality in our experimental results. The F1 score can reach 0.71 when combining all three modalities together into the prediction model.

4.3 Performance comparison on baselines

To capture rich multimodal information, especially the correlations between different modalities, we adopted cooperative networks with attention layers. In this section, compared with some classical methods, we analysed the experimental results of our proposed model.

Modalities	Accuracy	Precision	Recall	F1 score
Visual	0.54	0.62	0.58	0.60
Acoustic	0.52	0.52	0.56	0.54
Textual	0.58	0.59	0.63	0.61
Visual + acoustic	0.60	0.58	0.64	0.61
Visual + textual	0.64	0.60	0.75	0.67
Acoustic + textual	0.62	0.61	0.67	0.64
All	0.67	0.69	0.73	0.71

Table 4. Prediction performance on modality combinations

4.3.1 Baseline models. In the early fusion models, we concatenated multimodal information into one vector and then added two fully connected layers to estimate the advertising clicks. In the late fusion model, first, different modality features were trained in the neural network, and then the final prediction results were obtained by averaging these distributions.

In addition to the basic fusion models, we also made some classical improvements to conduct the experiments. The attention model, which is used to estimate different attention weights for different features, was applied for both early fusion and late fusion (Vaswani *et al.*, 2017). Based on the basic fusion models, we also applied principal component analysis (PCA) to achieve dimension reduction (Hyvarinen and Oja, 2000), and the support vector machine (SVM) and RF were then applied as the classifiers to predict the advertising clicks.

In addition, some influential multimodal learning methods that were proposed in recent years were also applied as baseline models. TRUMANN (Zhang *et al.*, 2016) is a tree-guided multitask method that is able to jointly learn a common space from multiple modalities. NeXtVLAD (Lin *et al.*, 2018) is a trainable model that integrates VLAD, grouping and an attention mechanism in a neural network to achieve feature representation. NNeXtVLAD+ (Liu *et al.*, 2019c) is an end-to-end method that jointly uses the normalised NeXtVLAD, a CNN layer and context gating for micro-video classification. DARE (Nie *et al.*, 2017) is a deep transfer model that can enhance the acoustic modality and alleviate the data sparsity problem in classification.

4.3.2 Parameter settings. We implemented our model using the TensorFlow [8] framework in Python, and the Xavier approach was used to initialise the model parameters. The size of the hidden layer was set as {64, 128, 256}, and the activation function is set as the ReLU. Unless otherwise stated, the models in our study all contain two hidden layers.

During the training process, the minibatch size was set as {64, 128, 256}, and the learning rate is set as {0.001, 0.005, 0.01}. To optimise the parameters, the optimiser was set as the adaptive moment estimation (Adam). All of the compared models were initialised and trained in the same way.

4.3.3 Performance comparison. The prediction performances of the baseline models are shown in Table 5. In addition, the cooperative network without an attention layer was also tested. As the experimental results show, compared to some traditional models, such as the

Models	Accuracy	Precision	Recall	F1 score
Early fusion	0.60	0.58	0.64	0.61
Late fusion	0.67	0.69	0.73	0.71
Early fusion + attention	0.68	0.72	0.66	0.69
Late fusion + attention	0.71	0.75	0.79	0.77
Early fusion + PCA + RF	0.61	0.67	0.61	0.64
Early fusion + PCA + SVM	0.64	0.57	0.65	0.61
Late fusion + PCA + RF	0.73	0.75	0.72	0.73
Late fusion + PCA + SVM	0.72	0.74	0.78	0.76
TRUMANN	0.69	0.73	0.69	0.71
NeXtVLAD	0.72	0.72	0.74	0.73
NNeXtVLAD+	0.72	0.76	0.82	0.79
DARE	0.77	0.73	*0.84*	0.78
Cooperative networks	0.71	0.75	0.69	0.72
Cooperative networks + attention	*0.83*	*0.84*	0.81	*0.82*

Note(s): The best performance values for each evaluation criterion amongst different models are shown in italic font. The applied model that combines cooperative networks and attention mechanisms shows the best performance in accuracy, precision and F1 score, while the DARE model performs the best in recall

Table 5.
Prediction performance of the baseline models

early fusion and late fusion models, cooperative networks actually captured some valuable features about the correlations between different modalities. Since the feature dimensions are relatively high, it is useful to select important features through some additional methods. The prediction results of the proposed cooperative learning model based on the relation-aware attention mechanism outperform other baseline models in most of the evaluation criteria, which demonstrates the effectiveness of the combinations between cooperative networks and the attention model.

Table 6 shows detailed results of the proposed cooperative learning model. In our prediction task, the advertising micro-videos were divided into five levels according to the number of commodity clicks. In addition to the final prediction results for different click levels, we also tested the prediction performance for each cooperative network with different host modalities without the late fusion layer. The experimental results show that the proposed model achieves better prediction performance at extreme levels (very high and very low). Moreover, the cooperative learning networks with the visual host perform the best at the very high level, while the textual host performs the best at the very low level.

4.4 Sensitivity analysis

In our experiments, we extracted 1,024 visual features, 100 acoustic features and 200 textual features from experience based on the methods described in Section 3.2. Since the dimensions of the multimodal features may influence the prediction results, we conducted additional experiments to test the sensitivity. In the sensitivity analysis, we extracted {512, 1,024} visual features, {50, 100, 200} acoustic features and {100, 200, 400} textual features. Half of the textual features were extracted from the commodity descriptions, and the rest were extracted from the micro-video content. As the experimental results in Table 7 show, the feature dimensions chosen in our main experiments perform the best amongst these results. In addition, most of the F1 scores (Macro-F1) amongst the various feature dimensions are over 0.7, which demonstrates that our model is relatively stable.

The poster picture is an important element for advertising micro-videos. In our experiments, the visual features extracted from the poster picture were treated as one of the keyframes in micro-video contents. In this section, we further tested the effectiveness of the poster picture in the prediction task by distinguishing the features extracted from the poster picture and other keyframes in the micro-video content. As Table 8 shows, in addition to regarding the poster as one keyframe and extracting {512, 1,024} visual features from all the keyframes, we conducted additional experiments that extract half of the visual features directly from the poster picture. The respective dimensions of the acoustic features and textual features that perform the best along with 512 visual features and 1,024 visual features from the former prediction task were used in the experiments. The experimental results show that when we reinforce the importance of the poster visual features in our prediction model, the F1 score of the prediction model is relatively stable but shows no obvious improvement.

Click amount levels	Visual	Acoustical	Textual	Integrated
Very high	0.87	0.80	0.85	0.90
High	0.76	0.72	0.72	0.77
Medium	0.74	0.71	0.71	0.75
Low	0.74	0.70	0.77	0.81
Very low	0.81	0.75	0.85	0.87

Table 6. Detailed results (F1 score) of the proposed cooperative learning model

4.5 Discussion

In this study, we proposed a multimodal cooperative learning model to predict micro-video advertising clicks. Previous advertising click prediction studies mainly extract features from advertisement producers and commodities. In addition to these features, we innovatively extracted acoustic and visual features from micro-video advertising contents. In addition, to better capture the correlations between different modalities, we adopted cooperative networks to fuse multimodal features, which also contribute to the fusion methods in the advertising click prediction field. The experimental results show that all three collected data sources (advertisement producers, commodities being sold and micro-video contents) are effective in the prediction task. Meanwhile, we simultaneously extracted visual, acoustic and textual features and then feed them into the prediction model. According to the experimental results, all of the multimodal features validate improved performance. Compared to classical multimodal fusion models, the cooperative learning model that best captures the correlations between different modalities further improves the prediction performance.

The experimental results also offer some deep insights into micro-video advertising practice. The results of the data source combination experiments show that advertisement producers have the heaviest impacts on advertising clicks. Merchants should seek a suitable advertisement producer with the utmost priority. Then, an effective design for commodity descriptions and micro-video content can both significantly increase the click numbers. Regarding the multimodal feature combinations, we notice that the acoustic features have a relatively small effect on the advertisement clicks, while the correlations between different modalities actually improve the advertisement clicks. Therefore, when producing micro-video advertisements, advertisement producers should also pay attention to the correlations between different modalities and not design each part separately. In addition, according to the detailed results for different click levels, textual features are relatively useful to

Table 7.
Prediction performance on various feature dimensions

Visual	Acoustic	Textual	F1 score	Visual	Acoustic	Textual	F1 score
512	50	100	0.67	1,024	50	100	0.75
512	50	200	0.72	1,024	50	200	0.78
512	50	400	0.73	1,024	50	400	0.78
512	100	100	0.69	1,024	100	100	0.76
512	100	200	0.73	1,024	100	200	0.82
512	100	400	0.73	1,024	100	400	0.80
512	200	100	0.71	1,024	200	100	0.76
512	200	200	0.74	1,024	200	200	0.81
512	200	400	0.75	1,024	200	400	0.80

Table 8.
Prediction performance for various visual feature combinations

Poster visual	Keyframe visual	Acoustic	Textual	F1 score
256	256	100	200	0.71
0	512	100	200	0.73
512	512	100	200	0.79
0	1024	100	200	0.82
256	256	200	400	0.76
0	512	200	400	0.75
512	512	200	400	0.76
0	1024	200	400	0.80

predict click-level advertisements, while visual features are more valid for high-level advertisements. Therefore, to avoid a failed advertisement, advertisement producers should give priority to a careful design of textual descriptions of the commodity and the micro-videos. If merchants aim to produce high click-level micro-video advertisements, then visual and acoustical designs are increasingly important.

5. Conclusions and future work

This paper proposes a multimodal cooperative learning framework for predicting the advertising clicks of micro-videos. The proposed framework achieves an average F1 score of 82% in the five-classification task on a real-world dataset collected from Douyin. Although previous studies have focused on micro-video popularity, no study has analysed the advertising effect of micro-videos. In existing advertising click prediction studies, multimodal content from the advertisement content is rarely simultaneously extracted and analysed. Therefore, this study fills the aforementioned research gap with the following contributions. First, to the best of our knowledge, this is the first study to explore the determinants of micro-video advertising clicks based on real-world micro-video advertisement datasets. Second, we extract multimodal information, including visual, acoustic, textual and numeric features, from the advertisement contents based on multiple data sources. The prediction effectiveness of these features is analysed. Third, to capture the relations between different modalities, we implement a cooperative learning model based on a relation-aware attention mechanism. The empirical analysis demonstrates that the correlations of different modalities are also effective in micro-video advertising click prediction.

Since micro-video advertising is still in a preliminary stage of development, our study has important managerial implications. With the help of our advertising click prediction model, advertisement producers (merchants or their partners) can better understand the importance of different modalities in micro-video advertising, thus generating more effective micro-video advertisements. In addition, micro-video platforms can apply our prediction results to optimise their advertisement allocation algorithm and better manage network traffic. The current research makes practical contributions in that the effectiveness of micro-video advertisements can be predicted in advance. This is of great help in advancing the micro-video advertisement industry.

The study contains several limitations. First, the platforms' advertisement allocation methods and recommendation mechanisms may impact the number of advertising clicks. In our study, we collect data in a short period of time with limited producers and assume that the allocation model is unchanged in our dataset. However, this assumption may not be realistic, which may cause errors in our study. Second, the proposed method does not have a special solution for some real situations in micro-video advertising. For example, for micro-videos, some poster pictures may be completely unrelated to micro-video content. Regarding the users, users may click the advertisements before they finish watching the whole micro-video; thus, analysing the entire micro-video content is not necessary. Third, some other advanced feature extraction and prediction methods can be employed. Although we design our framework based on one of the currently most advanced multimodal learning models, there are still some ways to improve either the prediction performance or the interpretability of the results in the micro-video advertising click prediction task. In future studies, we will improve our study mainly in two ways. First, we will attempt to improve the interpretability of our prediction model. Second, we will explore more practical issues in micro-video advertising, thus providing more valuable insights for the micro-video advertisement industry.

Notes

[1] https://www.questmobile.com.cn

2. https://www.douyin.com
3. https://www.tiktok.com
4. https://sensortower.com/blog/tiktok-downloads-2-billion/
5. http://opencv.org/
6. https://github.com/klb3713/sentence2vec
7. https://www.doushangyan.com
8. https://www.tensorflow.org

References

Andrews, M., Luo, X., Fang, Z. and Ghose, A. (2016), "Mobile ad effectiveness: hyper-contextual targeting with crowdedness", *Marketing Science*, Vol. 35 No. 2, pp. 218-233.

Bilenko, M. and Richardson, M. (2011), "Predictive client-side profiles for personalized advertising", *Proceedings of International Conference on Knowledge Discovery and Data Mining*, California, USA, 2011, pp. 413-421.

Bimpikis, K., Ozdaglar, A. and Yildiz, E. (2016), "Competitive targeted advertising over networks", *Operations Research*, Vol. 64 No. 3, pp. 705-720.

Bleier, A. and Eisenbeiss, M. (2015), "Personalized online advertising effectiveness: the interplay of what, when, and where", *Marketing Science*, Vol. 34 No. 5, pp. 669-688.

Boulding, W., Lee, E. and Staelin, R. (1994), "Mastering the mix: do advertising, promotion, and sales force activities lead to differentiation?", *Journal of Marketing Research*, Vol. 31 No. 2, pp. 159-172.

Breiman, L. (2001), "Random forests", *Machine Learning*, Vol. 45 No. 1, pp. 5-32.

Chatterjee, P., Hoffman, D.L. and Novak, T.P. (2003), "Modeling the clickstream: implications for web-based advertising efforts", *Marketing Science*, Vol. 22 No. 4, pp. 520-541.

Chaudhuri, M., Calantone, R.J., Voorhees, C.M. and Cockrell, S. (2018), "Disentangling the effects of promotion mix on new product sales: an examination of disaggregated drivers and the moderating effect of product class", *Journal of Business Research*, Vol. 90, pp. 286-294.

Cheah, J., Ting, H., Cham, T.H. and Memon, M.A. (2019), "The effect of selfie promotion and celebrity endorsed advertisement on decision-making processes", *Internet Research*, Vol. 29 No. 3, pp. 552-577.

Chen, J. (2016), "Multi-modal learning: study on a large-scale micro-video data collection", *Proceedings of ACM International Conference on Multimedia*, Amsterdam, Netherlands, 2016, pp. 1454-1458.

Chen, J. and Stallaert, J. (2014), "An economic analysis of online advertising using behavioral targeting", *MIS Quarterly*, Vol. 38 No. 2, pp. 429-449.

Chen, Y. and Yan, T.W. (2012), "Position-normalized click prediction in search advertising", *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Beijing, China, pp. 795-803.

Chen, J., Song, X., Nie, L., Wang, X. and Chua, T.S. (2016a), "Micro tells macro: predicting the popularity of micro-videos via a transductive model", *Proceedings of ACM International Conference on Multimedia*, Amsterdam, Netherlands, 2016, pp. 898-907.

Chen, J., Sun, B., Li, H., Lu, H. and Hua, X.S. (2016b), "Deep CTR prediction in display advertising", *Proceedings of the 24th ACM International Conference on Multimedia*, Amsterdam, Netherlands, pp. 811-820.

Chen, R., Zheng, Y., Xu, W., Liu, M. and Wang, J. (2018a), "Secondhand seller reputation in online markets: a text analytics framework", *Decision Support Systems*, Vol. 108, pp. 96-106.

Chen, X., Liu, D., Zha, Z., Zhou, W., Xiong, Z. and Li, Y. (2018b), "Temporal hierarchical attention at category- and item-level for micro-video click-through prediction", *Proceedings of ACM International Conference on Multimedia*, Seoul, Republic of Korea, 2018, pp. 1146-1153.

Ching, R.K., Tong, P., Chen, J. and Chen, H. (2013), "Narrative online advertising: identification and its effects on attitude toward a product", *Internet Research*, Vol. 23 No. 4, pp. 414-438.

Chun, K.Y., Song, J.H., Hollenbeck, C.R. and Lee, J. (2014), "Are contextual advertisements effective", *International Journal of Advertising*, Vol. 33 No. 2, pp. 351-371.

Erdem, T. and Sun, B. (2002), "An empirical investigation of the spillover effects of advertising and sales promotions in umbrella branding", *Journal of Marketing Research*, Vol. 39 No. 4, pp. 408-420.

Erdem, T., Keane, M.P. and Sun, B. (2008), "A dynamic model of brand choice when price and advertising signal product quality", *Marketing Science*, Vol. 27 No. 6, pp. 1111-1125.

Goh, K.Y., Chu, J. and Wu, J. (2015), "Mobile advertising: an empirical study of temporal and spatial differences in search behavior and advertising response", *Journal of Interactive Marketing*, Vol. 30, pp. 34-45.

Goldfarb, A. and Tucker, C.E. (2011), "Online display advertising: targeting and obtrusiveness", *Marketing Science*, Vol. 30 No. 3, pp. 389-404.

Guo, H., Tang, R., Ye, Y., Li, Z. and He, X. (2017), "DeepFM: a factorization-machine based neural network for CTR prediction", *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, Melbourne, Australia, pp. 1725-1731.

Guo, J., Nie, X., Jian, M. and Yin, Y. (2019), "Binary feature representation learning for scene retrieval in micro-video", *Multimedia Tools and Applications*, Vol. 78 No. 17, pp. 24539-24552.

Hyvarinen, A. and Oja, E. (2000), "Independent component analysis: algorithms and applications", *Neural Networks*, Vol. 13 No. 4, pp. 411-430.

Jedidi, K., Mela, C.F. and Gupta, S. (1999), "Managing advertising and promotion for long-run profitability", *Marketing Science*, Vol. 18 No. 1, pp. 1-22.

Jing, P., Su, Y., Nie, L., Bai, X., Liu, J. and Wang, M. (2018), "Low-rank multi-view embedding learning for micro-video popularity prediction", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 30 No. 8, pp. 1519-1532.

Kusumasondjaja, S. and Tjiptono, F. (2019), "Endorsement and visual complexity in food advertising on Instagram", *Internet Research*, Vol. 29 No. 4, pp. 659-687.

Li, K. and Du, T.C. (2012), "Building a targeted mobile advertising system for location-based services", *Decision Support Systems*, Vol. 54 No. 1, pp. 1-8.

Li, Z., Wang, J., Cai, J., Duan, Z., Wang, H. and Wang, Y. (2013), "Non-reference audio quality assessment for online live music recordings", *Proceedings of the ACM Multimedia Conference*, Barcelona, Spain, 2013, pp. 63-72.

Li, Y., Lin, L. and Chiu, S. (2014), "Enhancing targeted advertising with social context endorsement", *International Journal of Electronic Commerce*, Vol. 19 No. 1, pp. 99-128.

Li, C., Lu, Y., Mei, Q., Wang, D. and Pandey, S. (2015), "Click-through prediction for advertising in twitter timeline", *Proceedings of International Conference on Knowledge Discovery and Data Mining*, Sydney, Australia, 2015, pp. 1959-1968.

Li, Y., Liu, M., Yin, J., Cui, C., Xu, X. and Nie, L. (2019), "Routing micro-videos via a temporal graph-guided recommendation system", *Proceedings of ACM International Conference on Multimedia*, Nice, France, 2019, pp. 1464-1472.

Lin, R., Xiao, J. and Fan, J. (2018), "NeXtVLAD: an efficient neural network to aggregate frame-level features for large-scale video classification", *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, pp. 206-218.

Liu, H., Liu, Q. and Chintagunta, P.K. (2017a), "Promotion spillovers: drug detailing in combination therapy", *Marketing Science*, Vol. 36 No. 3, pp. 382-401.

Liu, M., Nie, L., Wang, M. and Chen, B. (2017b), “Towards micro-video understanding by joint sequential-sparse modeling”, *Proceedings of ACM International Conference on Multimedia*, Mountain View, USA, 2017, pp. 970-978.

Liu, M., Nie, L., Wang, X., Tian, Q. and Chen, B. (2019a), “Online data organizer: micro-video categorization by structure-guided multimodal dictionary learning”, *IEEE Transactions on Image Processing*, Vol. 28 No. 3, pp. 1235-1247.

Liu, S., Chen, Z., Liu, H. and Hu, X. (2019b), “User-video co-attention network for personalized micro-video recommendation”, *Proceedings of the World Wide Web Conference*, San Francisco, USA, 2019, pp. 3020-3026.

Liu, W., Huang, X., Cao, G., Zhang, J., Song, G. and Yang, L. (2019c), “Joint learning of NNeXtVLAD, CNN and context gating for micro-video venue classification”, *IEEE Access*, Vol. 4, pp. 1-9.

Liuthompkins, Y. (2019), “A decade of online advertising research: what we learned and what we need to know”, *Journal of Advertising*, Vol. 48 No. 1, pp. 1-13.

Manning, C., Raghavan, P. and Schütze, H. (2008), *Introduction to Information Retrieval*, Cambridge University Press, Cambridge.

Mela, C.F., Gupta, S. and Lehmann, D.R. (1997), “The long-term impact of promotion and advertising on consumer brand choice”, *Journal of Marketing Research*, Vol. 34 No. 2, pp. 248-261.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013), “Distributed representations of words and phrases and their compositionality”, *Proceedings of the Annual Conference on Neural Information Processing Systems*, Lake Tahoe, USA, 2013, pp. 3111-3119.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. and Ng, A. (2011), “Multimodal deep learning”, *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, USA, 2011, pp. 689-696.

Nie, L., Wang, X., Zhang, J., He, X., Zhang, H., Hong, R. and Tian, Q. (2017), “Enhancing micro-video understanding by harnessing external sounds”, *Proceedings of ACM International Conference on Multimedia*, Mountain View, USA, 2017, pp. 1192-1200.

Ouyang, W., Zhang, X., Li, L., Zou, H., Xing, X., Liu, Z. and Du, Y. (2019), “Deep spatio-temporal neural networks for click-through rate prediction”, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA, pp. 2078-2086.

Pan, J., Xu, J., Ruiz, A.L., Zhao, W., Pan, S., Sun, Y. and Lu, Q. (2018), “Field-weighted factorization machines for click-through rate prediction in display advertising”, *Proceedings of the 2018 World Wide Web Conference*, Lyon, France, pp. 1349-1357.

Pan, F., Li, S., Ao, X., Tang, P. and He, Q. (2019), “Warm up cold-start advertisements: improving ctr predictions via learning to learn id embeddings”, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pairs, France, pp. 695-704.

Qu, Y., Cai, H., Ren, K., Zhang, W., Yu, Y., Wen, Y. and Wang, J. (2016), “Product-based neural networks for user response prediction”, *Proceedings of the 16th International Conference on Data Mining*, Barcelona, Spain, pp. 1149-1154.

Ren, S., He, K., Girshick, R. and Sun, J. (2017), “Faster R-CNN: towards real-time object detection with region proposal networks”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39 No. 6, pp. 1137-1149.

Sahni, N.S. (2016), “Advertising spillovers: evidence from online field-experiments and implications for returns on advertising”, *Journal of Marketing Research*, Vol. 53 No. 4, pp. 459-478.

Sahni, N.S., Zou, D. and Chintagunta, P.K. (2017), “Do targeted discount offers serve as advertising? evidence from 70 field experiments”, *Management Science*, Vol. 63 No. 8, pp. 2688-2705.

Shaouf, A., Lü, K. and Li, X. (2016), “The effect of web advertising visual design on online purchase intention: an examination across gender”, *Computers in Human Behavior*, Vol. 60, pp. 622-634.

Simonyan, K. and Zisserman, A. (2015), "Very deep convolutional networks for large-scale image recognition", *Proceedings of International Conference on Learning Representations*, San Diego, USA, 2015, pp. 1-14.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. and Polosukhin, I. (2017), "Attention is all you need", *Proceedings of the 31st Conference on Neural Information Processing Systems*, Long Beach, USA, 2017, pp. 5998-6008.

Wei, Y., Xia, W., Lin, M., Huang, J., Ni, B., Dong, J., Zhao, Y. and Yan, S. (2016), "HCP: a flexible CNN framework for multi-label image classification", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38 No. 9, pp. 1901-1907.

Wei, Y., Cheng, Z., Yu, X., Zhao, Z., Zhu, L. and Nie, L. (2019a), "Personalized hashtag recommendation for micro-videos", *Proceedings of ACM International Conference on Multimedia*, Nice, France, 2019, pp. 1446-1454.

Wei, Y., Wang, X., Nie, L., He, X., Hong, R. and Chua, T. (2019b), "MMGCN: multi-modal graph convolution network for personalized recommendation of micro-video", *Proceedings of ACM International Conference on Multimedia*, Nice, France, 2019, pp. 1437-1445.

Wei, Y., Wang, X., Guan, W., Nie, L., Lin, Z. and Chen, B. (2020), "Neural multimodal cooperative learning toward micro-video understanding", *IEEE Transactions on Image Processing*, Vol. 29, pp. 1-13.

Wu, Z., Jiang, Y., Wang, J., Pu, J. and Xue, X. (2014), "Exploring inter-feature and inter-class relationships with deep neural networks for video classification", *Proceedings of the ACM Multimedia Conference*, Orlando, USA, 2014, pp. 167-176.

Zhang, K. and Katona, Z. (2012), "Contextual advertising", *Marketing Science*, Vol. 31 No. 6, pp. 980-994.

Zhang, J., Nie, L., Wang, X., He, X., Huang, X. and Chua, T.S. (2016), "Shorter-is-better: venue category estimation from micro-video", *Proceedings of the 24th ACM International Conference on Multimedia*, Amsterdam, Netherlands, pp. 1415-1424.

Zhang, Y., Li, B., Luo, X. and Wang, X. (2019), "Personalized mobile targeting with user engagement stages: combining a structural hidden Markov model and field experiment", *Information Systems Research*, Vol. 30 No. 3, pp. 787-804.

Zhou, G., Zhu, X., Song, C., Fan, Y., Zhu, H., Ma, X. and Gai, K. (2018), "Deep interest network for click-through rate prediction", *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, London, United Kingdom, pp. 1059-1068.

Corresponding author
Runyu Chen can be contacted at: ry.chen@uibe.edu.cn