IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 20, 2019, accepted June 5, 2019, date of publication June 12, 2019, date of current version June 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2922430

Joint Learning of NNeXtVLAD, CNN and Context Gating for Micro-Video Venue Classification

WEI LIU[1], XIANGLIN HUANG[1], GANG CAO[1], JIANGLONG ZHANG[2], GEGE SONG[1], AND LIFANG YANG[1]

[1]School of Computer Science and Cybersecurity, Communication University of China, Beijing 100024, China
[2]State Grid Fujian Information and Telecommunication Company, Fuzhou 350003, China

Corresponding author: Xianglin Huang (huangxl@cuc.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61772539 and 61401408, and in part by the Fundamental Research Funds for the Central Universities under Grant 3132017XNG1715 and CUC2019B021.

ABSTRACT Currently, micro-videos have grown explosively on various online social platforms. Accordingly, how to encode them to yield effective representation attracts our attention. NeXtVLAD is such an effective network that aggregates frame-level features into a compact supervector. However, the discriminant capability of such a supervector is still limited due to the lack of non-linear transformation and L2 normalization at the head and tail of original NeXtVLAD network, respectively. In order to address such problems, we propose an improved neural network architecture, normalized NeXtVLAD (NNeXtVLAD), which is extended with ReLU function and L2 normalization. In the light of such a new network, we build up an end-to-end framework which jointly learns NNeXtVLAD, CNN layer, and context gating for micro-video venue classification. Specifically, we first apply NNeXtVLAD layers as three-stream architecture to aggregate visual, acoustic, and textual features. We then pack and embed the aggregated features into CNN layer for enhancing the sparse concept-level representation. Finally, context gating is used to capture the interdependency among different network activations. Extensive experimental results on a real-world micro-video dataset exhibit that our proposed model significantly outperforms the state-of-the-art baselines in terms of both Micro-F1 and Macro-F1 scores.

INDEX TERMS Micro-video venue classification, normalized NeXtVLAD (NNeXtVLAD), sparse representation, context gating.

I. INTRODUCTION

In the era of Mobile Internet, the emergence of smart mobile devices has changed people's life way. As a result, it is much more convenient for people to upload, watch and share micro-videos on various online social platforms, such as Vine[1], Instagram,[2] and Snapchat,[3] which respectively limit their micro-video length up to 6, 10, and 15 seconds. Because of the shortness, each micro-video only can record various social events at single specific venue. In turn, such venue information further assists numerous location-oriented applications and personalized services on these social platforms. This phenomenon has stimulated the evolution of micro-video venue classification and recognition approaches. However, the automatic venue classification and recognition in micro-videos are subject to the low quality of micro-videos, which are probably caused by the poor smart phones, complex surrounding environments or user operation randomness, such as illumination variations, view point changes, and camera motions. Therefore, devising an accurate, efficient and robust feature representation of the micro-video is a crucial step for venue classification and recognition tasks.

Many existing methods focus on capturing spatio-temporal dependencies from temporal frames of micro-videos to generate frame-level feature representations. Recently popular methods are recurrent models (LSTM [1] and GRU [2]). Their main principle is to regard micro-videos as time-series data and learn long-term dependencies by relying on their ability of remembering and memory accesses. Recurrent models have been widely used in video classification filed, but such recurrent models have inherent limitation. For example, since each frame of videos must be processed step by step, such recurrent models cannot take full advantage of the GPU to perform parallel computation. Arandjelovic et al. [3]

The associate editor coordinating the review of this manuscript and approving it for publication was Jafar A. Alzubi.

[1]https://vine.co/
[2]https://instagram.com/
[3]https://www.snapchat.com/

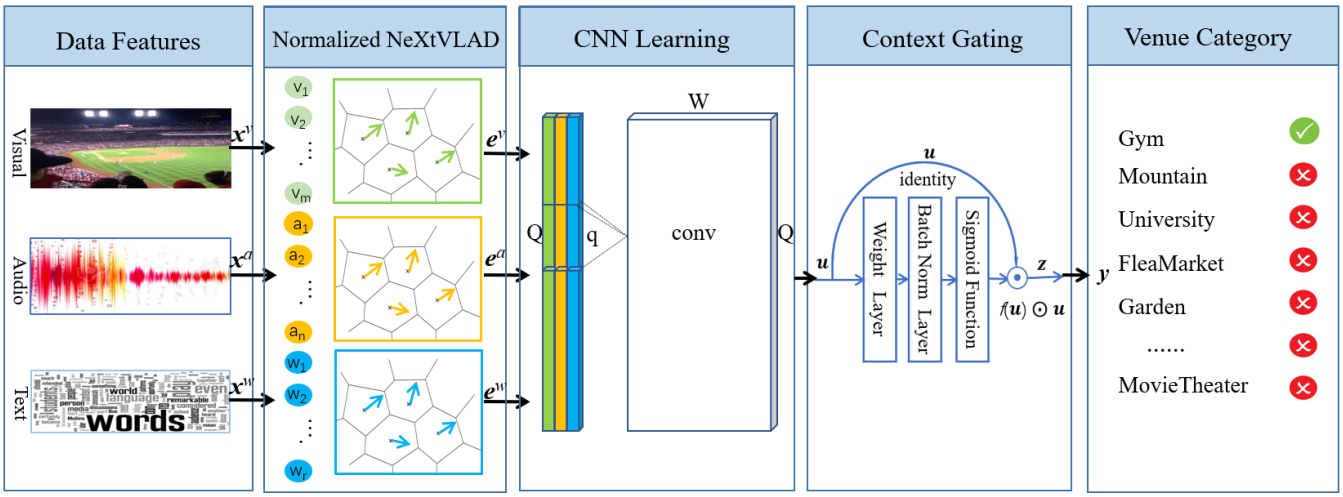


FIGURE 1. Proposed joint learning model with NNeXtVLAD, CNN and context gating.

developed a NetVLAD model which can aggregate all local descriptors to a global representation at once and produce a dominant result compared to recurrent methods. Despite the success of NetVLAD [4] for the task of temporal aggregation of visual and audio features, the encoded global features are in high dimension which generates numbers of parameters resulting in the overfitting and difficult optimization. To handle the problem, Lin et. al. [5] proposed NeXtVLAD that decompose the input features into a group of relatively lower-dimensional vectors with attention before they are encoded and aggregated over time. However, it is regrettable that non-linear transformation and L2 normalization are not considered at the head and tail of original NeXtVLAD network, making it difficult to further improve descriptive power of the compact feature.

In order to address aforementioned problems, we extend NeXtVLAD to an improved neural network architecture, Normalized NeXtVLAD (NNeXtVLAD). This extension utilizes ReLU function to realize non-linear transformation in the first fully connected layer at the head of original NeXtVLAD network and employs L2 normalization technology to normalize the final feature vector at the tail of that. In the light of such network, we build up an end-to-end learning framework which stacks the NNeXtVLAD, CNN and context gating for micro-video venue classification. The whole model is illustrated in Figure 1. In particular, we firstly leverage three independent NNeXtVLAD networks to aggregate frame-level features into a compact feature vector with common dimensions on visual, acoustic, and textual modalities in parallel. Secondly, we pack the three vectors with the same length as an input and then apply a CNN layer to extract their sparse and conceptual representations. Thirdly, context gating is introduced for modeling the dependency among labels. Finally, a softmax classifier is adopted for micro-video-level multi-label classification. The whole model can be trained efficiently and effectively. We validate our model on a publicly accessible benchmark micro-video dataset, whereby each micro-video is labeled with one venue category. Extensive experiments demonstrate that ReLU function and L2 normalization further improve NeXtVLAD encoding capability and our joint learning model obtains the state-of-the-art performance in terms of both Micro-F1 and Macro-F1 scores.

We summarize the main contributions of our work as follows:

(1) We propose an improved neural network architecture, NNeXtVLAD, which is extended with ReLU function and L2 normalization to obtain higher performance and faster convergence compared with NeXtVLAD.

(2) We build up an end-to-end joint learning model with NNeXtVLAD, CNN and context gating for micro-video venue classification. CNN can further enhance NNeXtVLAD encoding by extracting sparse and conceptual representations. Context gating can further capture the dependency among labels. This work has obtained the state-of-the-art performance on a real-world micro-video dataset.

The remainder of this paper is organized as follows. In Section 2, we first briefly review pioneering efforts related to micro-video content analysis and feature aggregation. Section 3 introduces our proposed model in more details. Experimental evaluation and analysis of our method are reported in Section 4, followed by conclusion in Section 5.

II. RELATED WORK

In this section, we review the related work including micro-video content analysis and feature aggregation.

A. MICRO-VIDEO CONTENT ANALYSIS

As a new form of user generated contents (UGCs), the micro-video content analysis has been attracted great attention from both industry and academia. Some researches on micro-videos content analysis have already begun in recent years. To support such researches, different large-scale micro-video

datasets have been constructed by [6]–[10] for different application. And, a tag refinement approach [11] for micro-videos was provided. Meanwhile, pioneers have obtained preliminary achievements in many domains, including creative micro-videos [11], popularity prediction [7], [12], venue retrieval [13] and venue estimation [10], [14]–[18]. The main principle is to complete the missing data [10], [19], [20], and then integrate high-dimensional and sparse data to achieve higher performance in [6], [7], [10], [12]–[23]. Moreover, in order to further efficiently apply temporal information, per-frame features of micro-videos are fed into LSTM models in [14], [15], since the intuition of using LSTM is to effectively capture long-range dependencies among micro-video frames which are crucial for micro-video venue classification tasks. However, the above Non-LSTM-based methods ignore temporal information of data, while LSTM-based approaches process each frame-level feature step by step. Different from such LSTM-based approaches, we propose the NNeXtVLAD to aggregate all frame-level features into a compact feature vector at the same time.

B. FEATURE AGGREGATION

Aggregating video features from individual frames or short clips plays a key role in the tasks of video understanding. After investigation, prior efforts can be divided into three categories: the first way is Convolutional Neural Network (CNN) that processes the frames to extract high-level semantic features and captures long-term relationship information among frames of the video. It has been very successful in many domains, such as video summarization [24], video action recognition [25]. The second method is Recurrent Neural Network (RNN), like LSTM, which has been extended to extract frame-level features step by step, thereby capturing the temporal structure of the video into a single representation. Great deal of work indicates that recurrent neural networks are effective. For example, Baccouche et al. [26] proposed a LSTM-based model to recognize human actions in videos. Srivastava et al. [27] used a LSTM model to predict the future frames of the given videos. The last approach captures only the distribution of local descriptor. Before the era of deep neural networks, there are many encoding methods, including bag-of-visual-words (BoW [28]), fisher vector (FV [29]), and vector of locally aggregated descriptors (VLAD [30]). Recently, metric learning [31] and fisher discriminative CNN [32], [33] have been proposed to enforce CNN features to be more discriminative. Besides, NetVLAD [3] and NeXtVLAD [5] have been further proposed by integrating VLAD into current neural networks to obtain end-to-end trainable aggregation. However, the above methods are rarely focused on learning a sparse representations of multi-modal data. Here we extend NeXtVLAD by using ReLU function and L2 normalization. In the light of this, we apply NNeXtVLAD layers as three-stream architecture to aggregate visual, acoustic and textual inputs and combine CNN model to fuse such three modal information, thereby obtaining sparse representations of micro-videos.

III. PROPOSED MODEL

Proposed end-to-end model consists of three components: (1) aggregating the sequential structures of three modalities via parallel NNeXtVLAD network; (2) learning the sparse and conceptual representation via a CNN layer; (3) finally capturing internal dependencies within the feature via context gating. In this section, we introduce them in detail.

A. NOTATIONS AND PROBLEM FORMULATION

Before going into details, we first list some notations to be used throughout the paper. Bold capital letters (e.g., $\boldsymbol{X}$) and bold lowercase letters (e.g., $\boldsymbol{x}$) are employed to denote matrices and vectors, respectively. Non-bold letters (e.g., x, X) are used to represent scalars. If not clarified, all vectors are in column form.

The problem definition is introduced formally. Considering a micro-video $\boldsymbol{x} \in \boldsymbol{X}$ with m modalities ($m = 3$ in this work), N-dimensional frame-level features $\boldsymbol{x}^m$ of such modalities are extracted (e.g. by a pretrained CNN) recursively, whereinto the modality indicator $m \in \{v, a, w\}$, v, a and w respectively represent the visual, acoustic, and textural modality. Each modality can be split into T key frames (or clips or words). We denote the key frame (or clip or word) in each modality as $\boldsymbol{x}_t^m$. We thus represent $\boldsymbol{X} = \{\{\boldsymbol{x}_{it}^m\}_{t=1}^T\}_{i=1}^M$ as a M micro-videos dataset. Each micro-video $\boldsymbol{x}$ is associated with one of the c pre-defined venue categories, namely a one-hot label vector $\boldsymbol{y}$. We aim to build a venue estimation model over the training set and produce a venue classifier that outputs the final classification scores for the new coming micro-videos.

B. NORMALIZED NEXTVLAD

For each N-dimensional frame-level features $\boldsymbol{x}^m$, we would like to aggregate these features while preserving their informative content. This is achieved by an extended NeXtVLAD encoder which is added ReLU fuction and L2 normalization at the head and tail of that, respectively. We refer to the extension as Normalized NeXtVLAD (NNeXtVLAD), which will be detailed in a stepwise way.

Firstly, as NeXtVLAD [5], we expand the input feature vector $\boldsymbol{x}^m$ using a fully connected layer. The formula is as follows:

$$\tilde{\boldsymbol{x}}^m = Max(0, \boldsymbol{W}_{fc}\boldsymbol{x}^m + \boldsymbol{b}_{fc}) \tag{1}$$

where $\boldsymbol{W}_{fc} \in R^{N\times 2N}$ and $\boldsymbol{b}_{fc} \in R^{2N}$ respectively represent the weight matrix and bias vector, and $Max(0, \cdot)$ implements $ReLU$ activation function. This operation relearns input feature $\boldsymbol{x}^m \in R^N$ to $\tilde{\boldsymbol{x}}^m \in R^{2N}$.

In order to reduce the dimension of descriptors and shrink the number of parameters, the grouping idea is realized by grouping all into G groups, each of which contains the $\frac{2N}{G}$ dimensional input feature vectors. Specifically, we apply a reshape operation to transform $\tilde{\boldsymbol{x}}^m$ with a shape of $(T, 2N)$ to $\tilde{\boldsymbol{x}}^{mg}$ with a shape of $(T, G, \frac{2N}{G})$.

Secondly, we can achieve an aggregate descriptor via aggregating the encoded vectors over time steps and groups.

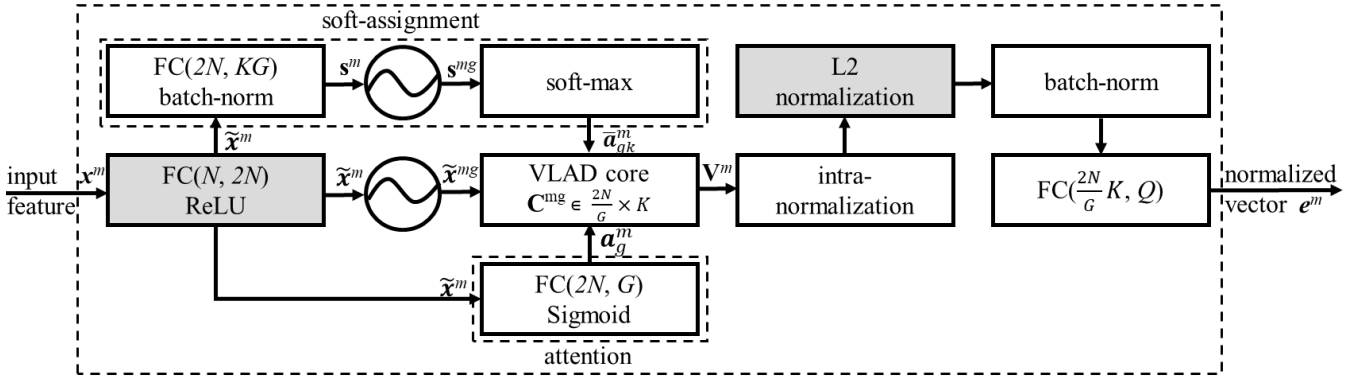


FIGURE 2. NNeXtVLAD algorithm procedure. Dimensional varieties are shown in brackets.

Formally, take the given $\tilde{\boldsymbol{x}}^m$ and the corresponding $\tilde{\boldsymbol{x}}^{mg}$ as inputs, K cluster centers $\{\boldsymbol{c}_k^{mg}\}$ as VLAD parameters over groups, and the output representation $\boldsymbol{V}^m(:, j, k)$ is a $\frac{2N}{G} \times K$ matrix which records the residual sum $(\tilde{\boldsymbol{x}}_t^{mg} - \boldsymbol{c}_k^{mg})$ of descriptors $\tilde{\boldsymbol{x}}_t^{mg}$ in all groups from learnable anchor point $\boldsymbol{c}_k^{mg}$ in the same grouped lower-dimensional space. The (j, k) element of $\boldsymbol{V}^m$ is computed as follows:

$$\boldsymbol{V}^m(:, j, k) = \Sigma_g^G \Sigma_t^T a_g^m(\tilde{\boldsymbol{x}}_t^m) \bar{a}_{gk}^m(\tilde{\boldsymbol{x}}_t^m)(\tilde{\boldsymbol{x}}_t^{mg}(j) - \boldsymbol{c}_k^{mg}(j)) \tag{2}$$

where $\boldsymbol{c}_k^{mg}$ is anchor point of cluster $\boldsymbol{C}^{mg} \in R^{\frac{2N}{G} \times K}$, which is sets of trainable parameters. $\tilde{\boldsymbol{x}}_t^{mg}(j)$ and $\boldsymbol{c}_k^{mg}(j)$ are the j-th dimensions of the t-th descriptor and k-th cluster center after grouping, respectively.

Note that the first term in formula (2) represents the attention function over groups. The formulation can be given:

$$a_g^m(\tilde{\boldsymbol{x}}_t^m) = \sigma(\boldsymbol{W}_g \tilde{\boldsymbol{x}}_t^m + \boldsymbol{b}_g) \tag{3}$$

where $\boldsymbol{W}_g \in R^{2N \times G}$ and $\boldsymbol{b}_g \in R^G$ respectively represent the weight matrix and bias vector, and $\sigma(\cdot)$ implements sigmoid function with output scale from 0 to 1. Thus, we can obtain the probability values to measure how important about each group.

The second term in formula (2) represents the soft-assignment of descriptor $\tilde{\boldsymbol{x}}_t^{mg}$ to cell k. The soft assigning weight can be given:

$$\bar{a}_{gk}^m(\tilde{\boldsymbol{x}}_t^m) = \frac{e^{BN(\boldsymbol{W}_{gk}\tilde{\boldsymbol{x}}_t^m + \boldsymbol{b}_{gk})}}{\Sigma_{s=1}^K e^{BN(\boldsymbol{W}_{gs}\tilde{\boldsymbol{x}}_t^m + \boldsymbol{b}_{gs})}} \tag{4}$$

where $\boldsymbol{W}_{gk} \in R^{2N \times GK}$ and $\boldsymbol{b}_{gk} \in R^{GK}$ respectively represent the weight matrix and bias vector. $BN(\cdot)$ implements batch normalization function. The batch normalizing transform is as follows:

$$BN(\boldsymbol{h}; \boldsymbol{\gamma}, \boldsymbol{\beta}) = \boldsymbol{\beta} + \boldsymbol{\gamma} \odot \frac{\boldsymbol{h} - Mean[\boldsymbol{h}]}{\sqrt{Var[\boldsymbol{h}] + \varepsilon}} \tag{5}$$

where $\boldsymbol{h}$ is the vector passing through fully connected layer over a complete minibatch, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are model parameters that determine the mean and standard deviation of the normalized activation, ε is a regularization hyperparameter, and $\odot$ is the element-wise multiplication operator. The statistics Mean[$\boldsymbol{h}$] and Var[$\boldsymbol{h}$] are estimated by the sample mean and sample variance of the current minibatch.

Finally, the matrix $\boldsymbol{V}^m$ is L2-normalized column-wise (intra-normalization [34]), reshaped into a vector, L2-normalized again in its entirety [30], batch-normalized again with formula (5), and finally used a linear fully connected layer to reduce vector dimension from $\frac{2N}{G}K$ to Q. At last, we can obtain a normalized vector $\boldsymbol{e}^m$ as each modal representation of the micro-video. The above entire algorithm procedure is shown in Figure 2.

Similarly, as shown in Figure 2, the number of parameters is mainly generated by the full connected layers and VLAD core. Specifically, ignoring biases, the number of parameters Num_p by only calculating weights can be obtained as follow:

$$Num_p = 2N^2 + 2NKG + 2NG + \frac{2N}{G}KQ + \frac{2N}{G}K \tag{6}$$

The data type of each parameter is float32, which takes up 4 bytes. According to experimental settings in Section IV, the model sizes of visual, acoustic and textual modalities are bout 160M, 7.2M and 2.8M, respectively. Thus, the whole model size is about 170M.

C. CNN LEARNING

Through the processing of three parallel NNeXtVLAD networks above, we obtain three feature vectors with equal length whereby each vector denotes one aggregate feature of a modality from the micro-video. Such three feature vectors come from the visual, acoustic and textual modalities of the same micro-video, so the vectors are not independent but highly correlated. Formally, for each micro-video, we can stack the three feature vectors into one feature map with the size of $Q \times 1 \times 3$ (i.e. Q height, 1 width and 3 Channels) as:

$$\tilde{\boldsymbol{E}} = [\tilde{\boldsymbol{e}}^v, \tilde{\boldsymbol{e}}^a, \tilde{\boldsymbol{e}}^w] \tag{7}$$

where $\tilde{\boldsymbol{E}}$ is a $Q \times 1 \times 3$ matrix and represents the feature map, whereinto $\tilde{\boldsymbol{e}}^v, \tilde{\boldsymbol{e}}^a, \tilde{\boldsymbol{e}}^w$ respectively denote the embedding over the visual, acoustic, and textual modalities.

In our model, we aim to learn a sparse and semantic representation using CNN. In particular, we devise

a 2-D convolutional layer with f filters of size $[q, 1]$, a stride of [1, 1] and SAME padding, where $q \leqq Q$. The feature map $\tilde{\boldsymbol{E}}$ as input data. And $ReLU$ is selected as the activation function. The formula is as follows:

$$f(\tilde{E}) = Max(0, \boldsymbol{W}_{conv} * \tilde{E} + \boldsymbol{b}_{conv}) \tag{8}$$

where $\boldsymbol{W}_{conv}$ and $\boldsymbol{b}_{conv}$ respectively represent the filters and bias, and * denotes the convolution operation. Here, we apply f convolution filters to extract the multi-modal feature, and each filter has a kernel size $p \times 1$. This is illustrated at the third part of Figure 1. The output is composed of f feature maps, and the size of each feature map is $Q \times 1$. The shape of the output is $Q \times 1 \times f$, we finally reshape the output to a vector $\boldsymbol{u}$ as a sparse representation of the three modalities.

D. CONTEXT GATING

In order to enhance description capability of the feature $\boldsymbol{u}$ and capture feature dependencies among labels $\boldsymbol{y}$, context gating module is adopted. Following prior work [4], the input feature representation $\boldsymbol{u}$ is transformed into a new representation $\boldsymbol{z}$. This is illustrated at the forth part of Figure 1. Context gating is defined as:

$$\boldsymbol{z} = \sigma(BN(\boldsymbol{W}_{gate}\boldsymbol{u} + \boldsymbol{b}_{gate})) \odot \boldsymbol{u} \tag{9}$$

where $\boldsymbol{u}$ is the input feature vector, $\boldsymbol{W}_{gate} \in R^{Qf \times Qf}$ and $\boldsymbol{b}_{gate} \in R^{Qf}$ are trainable parameters. $\sigma(BN(\boldsymbol{W}_{gate}\boldsymbol{u} + \boldsymbol{b}_{gate}))$ is a weight vector with output scale from 0 to 1. Hence, it means that a set of learned gates applied to the individual dimensions of the feature $\boldsymbol{u}$. Through the element-wise multiplication and training between the weight vector and $\boldsymbol{u}$, it is easy to get a new dependency representation $\boldsymbol{z}$, which has more powerful discriminant capability. After getting the code $\boldsymbol{z}$, we feed it into a softmax classifier.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we carried out extensive experiments to evaluate the effectiveness of our proposed approach on a public benchmark dataset.

A. DATASET AND EVALUATION METRICS

In order to validate our work, we conducted experiments on a public benchmark micro-video dataset released by Liu et al. [14]. The dataset consists of 20,093 micro-videos distributed in 22 Foursquare venue categories crawled from Vine through public API[4]. This dataset is for sequence modeling task which is to estimate venue category of the micro-video, so they firstly extracted 11 key frames and employed the AlexNet to get 4,096-D convolutional neural networks (CNN) visual feature from each key frame. And then they segmented 6 audio clips and employed Librosa[5] to generate 512-D acoustic feature from each audio clip. Finally, they separated 30 text words and employed Word2Vector to extract 100-D textual feature from each word in the textual description. We randomly shuffled the dataset and split it into two parts: 18,000 micro-videos for training and 2,093 ones for testing. We trained our model and the baselines over the same training set and verified them over the same testing one. And we repeated each experiment ten times and reported the average experimental results.

Following [10], [14], we evaluated performance of our model and the baselines in terms of both Macro-F1 [35] and Micro-F1 [36] as follow:

$$Macro - F1 = \frac{1}{C}\sum_{c=1}^{C}\frac{2p_c r_c}{p_c + r_c} \tag{10}$$

$$Micro - F1 = \frac{2\bar{p}\bar{r}}{\bar{p} + \bar{r}} \tag{11}$$

where p_c and r_c are the precision and recall values of the predicted c-th class, $\bar{p}$ and $\bar{r}$ are the precision and recall values across all classes. Macro-F1 assigns equal weight to each class-label in the averaging process, while Micro-F1 assigns equal weight to all instances in the averaging process. Both Macro-F1 and Micro-F1 metrics reach their best value at 1 and worst at 0.

B. BASELINES

To demonstrate the effectiveness of our proposed model, we chose the following several state-of-the-art methods as baselines:

1) TRUMANN [10]

This is a tree-guided multi-task multi-modal learning method, which learns a common feature space from multi-modal heterogeneous spaces and utilizes it to represent each micro-video. And then, the model can be leveraged to predict the venue category of a micro-video.

2) LSTMs [14]

This is a model with three parallel LSTMs, which extract respectively modality-specific sequential features from visual, acoustic and textual modalities of each micro-video, and then the one feature vector is cascaded via the three extracted feature vectors and fed into a softmax classifier.

3) EASTERN [14]

This is an end-to-end joint sequential-sparse model, which is capable of jointly capturing the sequential structures of visual, acoustic and textual modalities and sparsity of micro-videos.

4) LIU et al. [15]

This is also an end-to-end joint learning model. Different from EASTERN, it applies the CNN with smaller filters and the SAME type of padding, followed by the directly learning of prototypes for micro-video venue classification.

[4]https://github.com/davoclavo/vinepy
[5]https://github.com/librosa/librosa

TABLE 1. Performances of our model with different group sizes *G* on the mono-modalities.

Visual			Audio			Text		
Groups	Micro-F1	Macro-F1	Groups	Micro-F1	Macro-F1	Groups	Micro-F1	Macro-F1
2	60.17%	34.23%	2	47.67%	15.72%	2	51.17%	29.56%
4	59.65%	34.34%	4	47.83%	16.68%	**5**	**51.79%**	**32.33%**
8	60.04%	34.13%	8	48.88%	19.01%	8	51.61%	31.49%
16	**60.09%**	**34.55%**	16	49.22%	20.34%	10	51.39%	31.82%
32	59.57%	34.25%	**32**	**49.48%**	**22.30%**	20	50.81%	31.65%
64	59.37%	33.62%	64	49.40%	21.47%	25	50.70%	31.44%
128	59.03%	32.70%	128	49.23%	21.22%	50	50.75%	31.09%

5) NEXTVLAD [5]

This is a trainable model, which integrates VLAD, grouping idea and attention mechanism in a neural network to improve significantly the aggregated feature representation. Here, for the visual, acoustic and textual modalities of micro-video, the NeXtVLAD is applied to aggregate the sequential features of each modality into a single compact representation. Such compact representations of the three modalities are cascaded as a vector which can predict the venue category of the entire micro-video.

C. EXPERIMENTAL SETTINGS

Except TRUMANN model that was conducted based on Matlab2015,[6] we implemented our model and baselines with the help of Tensorflow.[7] Thus, the GPU can be used for accelerating. All the experiments above were conducted on a 64-bit Ubuntu 16.04 Server with Intel Xeon(R) CPU E5-1603 v3 at 2.80 GHz × 4 on 20 GB RAM and NVIDIA GeForce GTX 1080 GPU support.

At the NNeXtVLAD learning stage, the dimension Q of each modal output feature was set to 256. We also conducted experiments to find the best group and cluster sizes for our model and to demonstrate their effectiveness. Like NeXtVLAD, NNeXtVLAD is also based on an underlying assumption that one video frame may contain multiple objects, so the input features are decomposed into a group of relatively lower-dimensional vectors with attention. To demonstrate the effect of group size in our model for micro-video venue classification, we investigated the most appropriate group sizes for each modality in micro-video and summarized the result in Table 1. Since each frame feature dimension of each modality in micro-video is different, i.e. 1024-D, 512-D and 100-D for visual, acoustic and textual modalities, the three modal candidates for the group sizes are chosen from [2, 4, 8, 16, 32, 64, 128], [2, 4, 8, 16, 32, 64, 128] and [2, 5, 8, 10, 20, 25, 50], respectively. From table 1, we can see that the performance is the best when the group sizes are 16, 32, 5 for the visual, acoustic and textual modalities. According to this result, we determined [16, 32, 5] as group sizes of the three modalities on our model.

Similarly, the cluster is used to quantize descriptors into cluster words. The cluster size can affect final experimental performance. In order to evaluate the effect of different cluster sizes, we tested the most appropriate cluster size for each modality in micro-video. Table 2 represents our model performance when the cluster sizes varies. The three modal candidates for the cluster sizes are all chosen from [1, 2, 4, 8, 16, 32, 64, 128]. From table 2, we can see that the performance is the best when the cluster sizes are 32, 32, 64 for the visual, acoustic and textual modalities, respectively. The similarity in cluster sizes also reveals that the three modalities are not independent but highly correlated. With this experiment, we decided to choose [32, 32, 64] as cluster sizes for the three modalities on our model in the next experiments.

TABLE 2. Performances of our model with different cluster sizes *K* on the mono-modalities.

K	Visual		Audio		Text	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
1	58.73%	33.74%	47.26%	20.48%	50.40%	27.56%
2	58.91%	33.79%	48.68%	21.99%	51.44%	29.90%
4	59.09%	33.96%	48.98%	21.76%	51.45%	30.82%
8	59.77%	34.77%	49.31%	21.83%	51.48%	31.90%
16	59.72%	34.09%	49.10%	21.69%	51.26%	31.39%
32	**60.09%**	**34.55%**	**49.48%**	**22.30%**	51.61%	31.65%
64	59.53%	34.33%	49.41%	20.91%	**51.79%**	**32.33%**
128	58.92%	33.50%	49.03%	19.84%	51.62%	31.53%

At the CNN learning stage of our model, we used 1 convolutional layer with convolutional padding of the SAME type and stride of [1, 1]. The filters size [q, 1] was set to [5, 1] and the number f of convolutional filters was set to 4.

For our model, we selected Adam as the optimizer. The learning rate was set to 0.01 and the learning rate decay was searched in [0.25, 0.3, 0.5]. Training stop if there is no improvement after 10 epochs with batch size of 100.

D. PERFORMANCE COMPARISON WITH RELU AND L2 NORMALIZATION

To demonstrate the impact of ReLU function and L2 normalization for NeXtVLAD [5] network on micro-video venue classification, we carried out experiments and summarized the results in Table 3. (1) Compared to the NeXtVLAD whose Micro-F1 and Macro-F1 are 60.16% and 32.58%, adding ReLU function brings improvement by about 2.51% and 2.69%, respectively. This demonstrates that ReLU function realizes non-linear transformation in the first fully connected layer at the head of original NeXtVLAD, resulting in a better

[6]https://ww2.mathworks.cn/
[7]https://www.tensorflow.org

TABLE 3. Performance comparison concatenated with ReLU and L2 normalization for NeXtVLAD network.

Method	Micro-F1	Macro-F1
NeXtVLAD [5]	60.16%	32.58%
NeXtVLAD+ReLU	62.67%	35.27%
NeXtVLAD+L2 Norm	62.79%	36.51%
ALL (NNeXtVLAD)	**65.25%**	**38.75%**

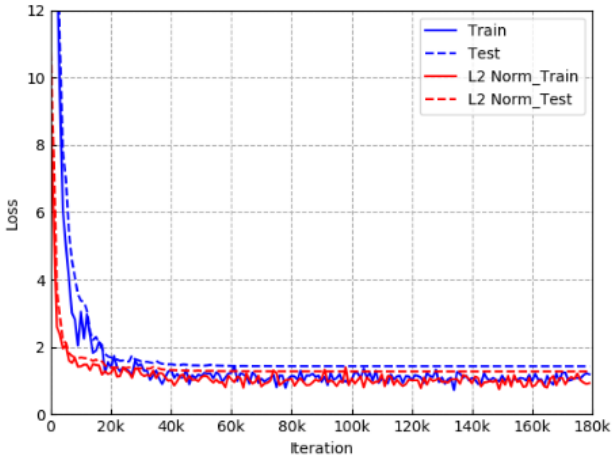


FIGURE 3. Loss trend curves on the train and test sets for original (blue) and L2 normalized (red) NeXtVLAD networks. The solid lines are the training curves and the dotted lines are the test curves.

improvement in performance. (2) Similar results can be found based on NeXtVLAD, adding L2 normalization provides improvement by about 2.63% and 3.93%, respectively. This is because the final feature vector is L2-normalized at the tail of original NeXtVLAD, thereby improving performance. (3) Without a doubt, the best result is achieved by NNeXtVLAD, which applies ReLU function and L2 normalization at the head and tail, respectively. This further shows our proposed NNeXtVLAD has greater aggregated power from frames than NeXtVLAD.

To further demonstrate the effect by L2 normalization on training and testing step for NeXtVLAD, we measured the variation in loss. From Figure 3, we can observe that loss value of the L2 normalized NeXtVLAD is almost always lower than that of the original NeXtVLAD whether in the training phase or in the testing phase. This demonstrates that L2 normalization is applied to normalize the final feature vector at the tail of original NeXtVLAD network, thereby improving the conditioning of the optimization problem.

E. PERFORMANCE COMPARISON CONCATENATED WITH CNN AND CONTEXT GATE

We carried out experiments to study the effect of CNN layer and context gating on our model. In particular, we connected CNN layer, context gating and their combination to NNeXtVLAD, respectively. The results are summarized in Table 4. We can see that whether individually or in combination, both CNN layer and context gating can enhance NNeXtVLAD in terms of both Micro-F1 and

TABLE 4. Performance comparison concatenated with CNN and context gating for NNeXtVLAD network.

Method	Micro-F1	Macro-F1
NNeXtVLAD	65.25%	38.75%
NNeXtVLAD+CNN	66.59%	41.33%
NNeXtVLAD+Gate	66.67%	40.49%
ALL (NNeXtVLAD+)	**66.87%**	**41.88%**

TABLE 5. Performance comparison between our proposed method and several state-of-the-art baselines.

Method	Micro-F1	Macro-F1	Time
TRUMANN [10]	58.58%	18.87%	**3.02s**
LSTMs [14]	61.47%	31.03%	11.27s
EASTERN [14]	62.00%	30.69%	10.80s
Liu et al. [15]	62.73%	32.93%	11.05s
NNeXtVLAD	65.25%	38.75%	19.42s
NNeXtVLAD+	**66.87%**	**41.88%**	19.47s

Macro-F1 scores. This verifies that CNN layer can further enhance the sparse concept-level representations and context gating can model the dependency among labels, thereby they improve the performance of NNeXtVLAD.

F. PERFORMANCE COMPARISON AMONG MODELS

We summarized the performance comparison between our approach and the baselines in Table 5. (1) The TRUMANN [10] model achieves the worst performance, as compared to other learning approaches. This is because it does not consider temporal sequence infomation of the micro-videos. (2) The three methods LSTMs [14], EASTERN [14] and Liu et al. [15] have similar performance. This is due to the fact that all three methods are LSTM-based. But there are also some differences. The EASTERN [14] and Liu et al. [15] outperform LSTMs [14] model. This demonstrates that CNN layer can capture sparse representation in feature and such the representation is beneficial to venue category classification. EASTERN [14] performs weaker than Liu et al. [15]. This is because Liu et al. [15] applied prototype learning to improve the robustness. (3) Compared with TRUMANN, the performance of NNeXtVLAD is better. This signals that NNeXtVLAD can capture temporal sequential patterns of micro-videos. Compared with such the three LSTM-based methods which are models of handing sequences, the performance of NNeXtVLAD is also better. This is because it integrates VLAD, attention mechanism, grouping idea, non-linear transformation and normalization operation into neural networks, resulting in greater aggregated power from temporal sequential frames of micro-videos. (4) Our proposed joint learning model, NNeXtVLAD+, outperforms NNeXtVLAD. This verifies that CNN layer can enhance the sparse concept-level representations, and context gating can capture the dependency among labels. (5) Our proposed NNeXtVLAD+ achieves the best. In particular, compared to Liu et al. [15] which is state-of-the-art baseline for micro-video venue classification, our model brings improvement by about 4.14% and 8.95% in

TABLE 6. Performance comparison between NNeXtVLAD+ and Liu et al. [15] methods with different modality combinations.

Method	Modality	Micro-F1	Macro-F1
Liu et al. [15]	Visual	58.03%	28.57%
	Audio	44.63%	12.01%
	Text	42.82%	10.87%
	Visual+Audio	60.50%	29.95%
	Visual+Text	60.72%	31.04%
	Audio+Text	50.66%	17.00%
	ALL	62.73%	32.93%
NNeXtVLAD+	Visual	**60.09%**	**34.55%**
	Audio	**49.48%**	**22.30%**
	Text	**51.79%**	**32.33%**
	Visual+Audio	**61.87%**	**32.79%**
	Visual+Text	**64.16%**	**40.19%**
	Audio+Text	**58.37%**	**36.10%**
	ALL	**66.87%**	**41.88%**

terms of Micro-F1 and Macro-F1. This justifies the effectiveness of our model for micro-video venue classification. (6) As to the efficiency over the training set, the time cost of our model is higher than that of the others. But, considering better classification performance, our model is still very effective.

G. COMPARISON ON MODALITY COMBINATION

We further compare the performance of our proposed joint learning model (NNeXtVLAD+) and Liu et al. [15] under different modality combinations. The test results shown in Table 6 indicate that our model achieves overall superiority. In particular, among the three mono-modalities, we note that the textual discriminating capability rises to the second place, which further verifies the effectiveness of our proposed model. The results also verify the following conclusions: (1) The visual modality behaves the best among the three ones. It should owe to the enriched location-specific information included in such a modality. It also reveals that AlexNet can extract more prominent characteristics of venue categories from the visual modality. (2) Modalities combination outperforms the mono-modality due to the full use of complementary information among different modalities.

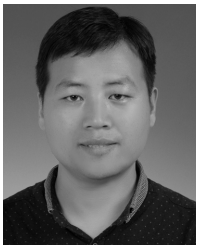
V. CONCLUSION

In this paper, we propose an improved neural network architecture, NNeXtVLAD, which is extended with ReLU function and L2 normalization to improve higher performance and faster convergence. Based on this, we combine NNeXtVLAD, CNN and context gating to generate a novel end-to-end joint learning model which is applied to recognize the category of the micro-video venue. By conducting extensive experiments, our NNeXtVLAD outperforms NeXtVLAD, as well as our joint learning model significantly outperforms baselines on a real-world micro-video dataset in terms of both Micro-F1 and Macro-F1 scores. In the future, we will try to extend attention mechanism of NeXtVLAD model from group-level to concept-level, aiming to enchane each dimensional influence of the aggregated descriptor.

REFERENCES

[1] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[2] K. Cho, B. Van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," 2014, *arXiv:1409.1259*. [Online]. Available: https://arxiv.org/abs/1409.1259
[3] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5297–5307.
[4] A. Miech, I. Laptev, and J. Sivic, "Learnable pooling with context gating for video classification," 2017, *arXiv:1706.06905*. [Online]. Available: https://arxiv.org/abs/1706.06905
[5] R. Lin, J. Xiao, and J. Fan, "NeXtVLAD: An efficient neural network to aggregate frame-level features for large-scale video classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 206–218.
[6] J. Chen, "Multi-modal learning: Study on a large-scale micro-video data collection," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 1454–1458.
[7] J. Chen, X. Song, L. Nie, X. Wang, H. Zhang, and T. S. Chua, "Micro tells macro: Predicting the popularity of micro-videos via a transductive model," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 898–907.
[8] P. X. Nguyen, G. Rogez, C. Fowlkes, and D. Ramanan, "The open world of micro-videos," 2016, *arXiv:1603.09439*. [Online]. Available: https://arxiv.org/abs/1603.09439
[9] M. Redi, N. O'Hare, R. Schifanella, M. Trevisiol, and A. Jaimes, "6 seconds of sound and vision: Creativity in micro-videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 4272–4279.
[10] J. Zhang, L. Nie, X. Wang, X. He, X. Huang, and T. S. Chua, "Shorter-is-better: Venue category estimation from micro-video," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 1415–1424.
[11] L. Huang and B. Luo, "Tag refinement of micro-videos by learning from multiple data sources," *Multimedia Tools Appl.*, vol. 76, no. 19, pp. 20341–20358, 2017.
[12] P. Jing, Y. Su, L. Nie, X. Bai, J. Liu, and M. Wang, "Low-rank multi-view embedding learning for micro-video popularity prediction," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 8, pp. 1519–1532, Aug. 2018.
[13] J. Guo, X. Nie, C. Cui, X. Xi, Y. Ma, and Y. Yin, "Getting more from one attractive scene: Venue retrieval in micro-videos," in *Proc. Pacific Rim Conf. Multimedia*. Springer, 2018, pp. 721–733.
[14] M. Liu, L. Nie, M. Wang, and B. Chen, "Towards micro-video understanding by joint sequential-sparse modeling," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 970–978.
[15] W. Liu, X. Huang, G. Cao, G. Song, and L. Yang, "Joint learning of LSTMs-CNN and prototype for micro-video venue classification," in *Proc. Pacific Rim Conf. Multimedia*, 2018, pp. 705–715.
[16] L. Nie, X. Wang, J. Zhang, X. He, H. Zhang, R. Hong, and Q. Tian, "Enhancing micro-video understanding by harnessing external sounds," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 1192–1200.
[17] M. Liu, L. Nie, X. Wang, Q. Tian, and B. Chen, "Online data organizer: Micro-video categorization by structure-guided multimodal dictionary learning," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1235–1247, Mar. 2019.
[18] S. Jiang, W. Min, and S. Mei, "Hierarchy-dependent cross-platform multi-view feature learning for venue category prediction," *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1609–1619, Jun. 2019.
[19] X. Luo, M. Zhou, Y. Xia, Q. Zhu, A. C. Ammari, and A. Alabdulwahab, "Generating highly accurate predictions for missing QoS data via aggregating nonnegative latent factor models," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 3, pp. 524–537, Mar. 2016.
[20] X. Luo, M. Zhou, S. Li, Y. Xia, Z.-H. You, Q. Zhu, and H. Leung, "Incorporation of efficient second-order solvers into latent factor models for accurate prediction of missing QoS data," *IEEE Trans. Cybern.*, vol. 48, no. 4, pp. 1216–1228, Apr. 2018.
[21] X. Luo, M. Zhou, S. Li, Z. You, Y. Xia, and Q.-S. Zhu, "A nonnegative latent factor model for large-scale sparse matrices in recommender systems via alternating direction method," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 3, pp. 579–592, Mar. 2016.
[22] X. Luo, J. Sun, Z. Wang, S. Li, and M. Shang, "Symmetric and nonnegative latent factor models for undirected, high-dimensional, and sparse networks in industrial applications," *IEEE Trans. Ind. Informat.*, vol. 13, no. 6, pp. 3098–3107, Dec. 2017.
[23] X. Luo, H. Wu, H. Yuan, and M. Zhou, "Temporal pattern-aware QoS prediction via biased non-negative latent factorization of tensors," *IEEE Trans. Cybern.*, to be published. doi: 10.1109/TCYB.2019.2903736.

[24] M. Rochan, L. Ye, and Y. Wang, "Video summarization using fully convolutional sequence networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 347–363.
[25] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7445–7454.
[26] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Proc. Int. Workshop Hum. Behav. Understand.*, 2011, pp. 29–39.
[27] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using LSTMs," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 843–852.
[28] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop Stat. Learn. Comput. Vis. (ECCV)*, vol. 1, 2004, pp. 1–2.
[29] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[30] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3304–3311.
[31] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.
[32] G. Cheng, P. Zhou, and J. Han, "Rifd-cnn: Rotation-invariant and Fisher discriminative convolutional neural networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2884–2893.
[33] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and Fisher discriminative convolutional neural networks for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 265–278, Jan. 2019.
[34] R. Arandjelovic and A. Zisserman, "All about VLAD," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1578–1585.
[35] B. Lepri, N. Mana, A. Cappelletti, and F. Pianesi, "Automatic prediction of individual performance from thin slices of social behavior," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 733–736.
[36] C. Sanden and J. Z. Zhang, "Enhancing multi-label music genre classification through ensemble techniques," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2011, pp. 705–714.

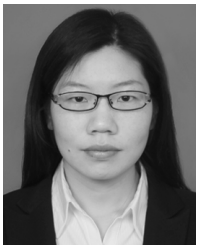


WEI LIU received the B.S. degree in computer science and technology from Xinyang Normal University, Henan, China, in 2006, and the M.S. degree in computer application and technology from China University of Geosciences, Wuhan, China, in 2009. He is currently a Ph.D. Student Member with the School of Computer Science and Cybersecurity, Communication University of China, Beijing, China. His current research interests include multimedia content analysis and deep learning.

XIANGLIN HUANG received the B.S. and M.S. degrees from Jilin University, Jilin, China, in 1990 and 1998, and the Ph.D. degree from Beijing University of Technology, Beijing, China, in 2002. He is currently a Professor with the School of Computer Science and Cybersecurity, Communication University of China. His research interests include image and video intelligent processing.

GANG CAO received the B.S. degree from Wuhan University of Technology, Hubei, China, in 2005, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University, Beijing, China, in 2013. He was with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, in 2010. He is currently an associate professor of the School of Computer Science and Cybersecurity, Communication University of China, Beijing, China. His current research interests include digital forensics, data hiding, and multimedia signal processing.

JIANGLONG ZHANG received the B.S. degree in electronic engineering from Fujian Normal University, Fuzhou, Fujian, China, in 2009, and the M.S. and Ph.D. degrees in signal and information processing from the Communication University of China, Beijing, China, in 2013 and 2017. He is currently a Faculty Member of the State Grid Corporation of China. His research interests include machine learning and multimedia content analysis. Various parts of his work have been published in top forums including ACM MM.



GEGE SONG received the B.S. degree from Shandong Technology and Business University, Shandong, China, in 2013, and the master's degree from the School of Computer Science, Communication University of China, Beijing, China, in 2016. She is currently a Ph.D. Student Member with the School of Computer Science and Cybersecurity, Communication University of China, Beijing, China. Her current research interests include image processing and natural language processing.

LIFANG YANG received the B.S. degree in electronic information from Qingdao University, Qingdao, Shandong, China, in 2005, and the M.S. and Ph.D. degrees in signal and information processing from the Communication University of China, Beijing, China. She is currently an Associate Professor with the Communication University of China. Her research interests include intelligent retrieval and high-dimensional index structure.