Towards Micro-video Understanding by Joint Sequential-Sparse Modeling

Meng Liu[†], Liqiang Nie[†], Meng Wang[‡], Baoquan Chen[†]
[†] School of Computer Science and Technology, Shandong University, Jinan, China
[‡]School of Computer and Information, Hefei University of Technology, Hefei, China
{mengliu.sdu,nieliqiang,eric.mengwang,baoquan.chen}@gmail.com

ABSTRACT

Like the traditional long videos, micro-videos are the unity of textual, acoustic, and visual modalities. These modalities sequentially tell a real-life event from distinct angles. Yet, unlike the traditional long videos with rich content, micro-videos are very short, lasting for 6-15 seconds, and they hence usually convey one or a few high-level concepts. In the light of this, we have to characterize and jointly model the sparseness and multiple sequential structures for better micro-video understanding. To accomplish this, in this paper, we present an end-to-end deep learning model, which packs three parallel LSTMs to capture the sequential structures and a convolutional neural network to learn the sparse concept-level representations of micro-videos. We applied our model to the application of micro-video categorization. Besides, we constructed a real-world dataset for sequence modeling and released it to facilitate other researchers. Experimental results demonstrate that our model yields better performance than several state-of-the-art baselines.

CCS CONCEPTS

•Information systems → Multimedia information systems;

KEYWORDS

Micro-Video Understanding; Parallel LSTMs; Dictionary Learning; Convolutional Neural Network

1 INTRODUCTION

Recent years have witnessed the proliferation of micro-video platforms, such as Vine[1], Instagram[2], and Facebook[3]. On the one hand, similar to the traditional long videos, micro-videos are also the unity of visual frames, surrounding textual descriptions, and audio channels. Hereafter, we name them as visual, textual, and acoustic modalities, respectively. These three modalities complementarily describe a real-life event from distinct angles in different media forms. And an event is usually related to a temporal expression, which is sequentially encoded into the three modalities of the given video, i.e., a set of ordered frames, textual sentences with correct syntactic and sematic orderings, as well as a sequential audio clips with the rise and fall of the waveform amplitude envelope. In a sense, video understanding requires to model the sequential structures of three modalities separately and then properly fuse them. On the other hand, unlike the traditional long videos with rich content, micro-videos are very short, only lasting for 6-15 seconds, and they hence usually contain one or a few high-level concepts [5]. We thus need to learn their sparse and conceptual representations for a better discrimination.

Regarding sequence modeling, Recurrent Neural Networks (RNNs) using Long Short Term Memory (LSTM) have been successfully applied to various sequence tasks, such as speech recognition [14], machine translation [42], and caption generation for images [46]. Besides, they have been employed to analyze videos, such as action recognition [45] and natural language description generation [9]. However, existing methods are mono-modal and they are not capable of capturing the sequences of multiple modalities simultaneously. We argue that the sequential structures of the textual, visual, and acoustic modalities carry different information and hence require sequence-dependent LSTMs. For example, the textual modality may give a high-level description of the given event, and the description order is hence not necessary to completely meet the event timeline. And although the acoustic modality is usually aligned with the visual one over the time axis, they may highlight different aspects. Considering this case, the visual modality is about the food color; whereas the acoustic one may describe food taste. In the light of this, we propose a parallel LSTMs method which can capture the triple sequences independently. It is worth mentioning that a pioneering work in [34] presents a novel LSTM model to jointly consider the sequences of the visual and acoustic modalities for the task of speaker identification. They assume that the sequence information of the video and audio is closely related, namely the face of the speaker appears in the video whenever she/he speaks, and there is only one speaker at one time during the voice over. It forces the visual and acoustic modalities to share the same LSTM. Ours

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

MM'17, October 23–27, 2017, Mountain View, CA, USA.
 ISBN 978-1-4503-4906-2/17/10...$15.00
DOI: http://dx.doi.org/10.1145/3123266.3123341

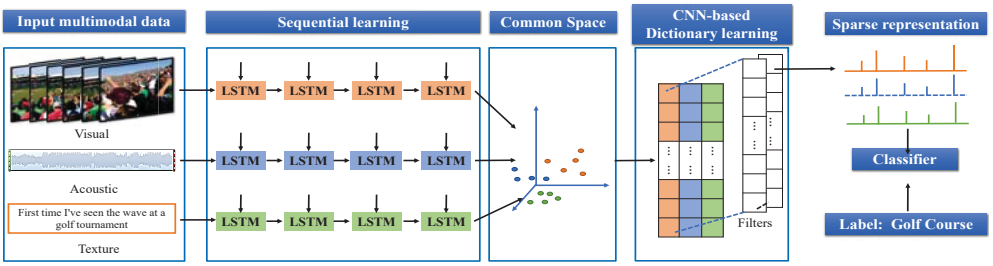

Figure 1: An illustration of our proposed EASTERN model.

is different. Instead, we have separate LSTMs for distinct modalities.

As to the sparse and conceptual representations, one intuitive thought is to use the dictionary learning approach [28], which has been widely used in the field of computer vision with theoretical and practical success. Dictionary learning aims to find a sparse representation of the input data in the form of a linear combination of basic elements as well as those basic elements themselves. These elements are called atoms and they compose a dictionary. However, the existing dictionary learning methods are in shallow settings and they are hard to learn discriminative and high-level concepts. The work in [10] has justified that a deep convolutional neural network is equivalent to the sparse dictionary learning pipeline, whereby each convolutional filter can be seen as a dictionary atom that we aim to learn, and the sparse coding can be seen as the activation value of the filtered results. We are hence inspired to utilize a convolutional neural network to replace the traditional shallow dictionary learning model for sparse and conceptual representation learning.

Regarding the aforementioned two components, we build up an end-to-end deep model, the so-called dEep pArallel Sequence wiTh sparsE constRaiNt, EASTERN for short. The framework of EASTERN is illustrated in Figure 1. In particular, we first leverage three independent LSTMs to characterize the sequential structures of three modalities in parallel. We then project their outputs into a common space by three distinct mapping functions. After that, we input the three projected vectors with the same length into a convolutional neural network to learn their sparse and conceptual representations, whereby the K filters serve as the K atoms in a dictionary. We finally adopt a softmax function for further classification tasks.

We apply our model to one application: inferring the venue categories of micro-videos. In particular, the salient feature of a micro-video is that it is usually recorded by a GPS-enabled mobile device within a few seconds at one specific place. Therefore, micro-video platforms are able to encourage users to associate their micro-videos with location information to indicate their recording places, such as "Disneyland Park in California", which will benefit several location-oriented services, such as footprints recording, local restaurants suggestion, and regional weather alert. Despite their value and significance, the majority of users are inactive to share their location information to avoid privacy leakage. Specifically, as reported in [50], around 98.78% micro-videos do not have geo-information. This motivates us to infer the missing location information of micro-videos. However, we have to figure out that it is hard to infer the specific location information, such as "American Airlines Arena in Mam Florda USA". Instead, we turn to infer the venue category of a given micro-video, such as "Basketball Court". We carried out experiments on a real-world micro-video dataset collected from Vine, whereby each micro-video is labeled with one venue category. The results demonstrate that our proposed EASTERN model significantly outperforms several state-of-the-art baselines. It is worth mentioning that our model is also applicable to other micro-video analysis tasks, such as popularity prediction.

We summarize the contributions of our work as follows:

1) We analyze the parallel sequential structures and sparse properties of micro-videos and build up an end-to-end deep model accordingly for micro-video understanding. This model is capable of jointly capturing the sequential structures of three modalities and sparsity of micro-videos.

2) We apply our proposed EASTERN model to a real-world micro-video application, i.e., venue category estimation. As a side contribution, we released the data, source codes, and parameters to facilitate other researchers[4].

The rest of the paper is organized as follows. In Section 2, we review the related work. Section 3 and 4 detail our data collection and our proposed EASTERN model, respectively. Experimental results and analysis are introduced in Section 5, followed by the concluding remarks in Section 6.

2 RELATED WORK

2.1 LSTM Recurrent Neural Network

LSTM introduced in [17] is one of the popular variations of RNN, which is designed to mitigate the gradient vanish problem of RNN [4]. In addition, it has been very successful in variety of temporal sequence tasks, such as language modeling [12, 29], translation [26], dialog system [11], time series prediction [37], rhythm learning [13], visual question answering [27], handwriting recognition [15], and protein homology detection [16]. Recently, it has been extended to extract significant features from video sequences, since it is able to preserve information over long periods of time. Baccouche et al. [2] proposed a LSTM based model to recognize

[4]https://acmmm17.wixsite.com/eastern.

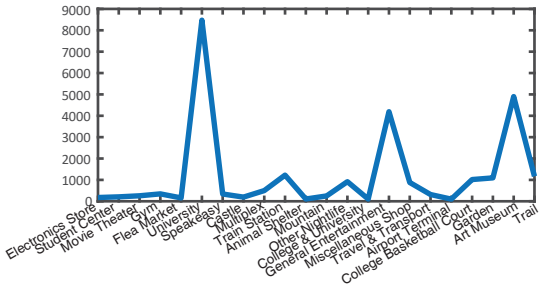

Figure 2: Distribution of micro-video counts with respect to the venue categories in our dataset.

human actions in videos. And Donahue et al. [9] combined a LSTM with a CNN network to recognize the activities in videos. Srivastava et al. [41] used a LSTM model to predict the future frames of the given videos.

Most of the above studies rely on a single LSTM to process mono-modal temporal sequences. They thereby can only capture information from one modality of the input data. In our work, we argue that the visual, textual, and acoustic modalities in micro-videos convey complementary information in heterogeneous sequential manners. So the existing methods are not suitable for our task. To address this problem, we proposed a parallel LSTMs which can effectively extract multi-modal features from multi-modal temporal sequences, simultaneously.

2.2 Convolutional Neural Networks

The CNN framework, proposed by LeCun et al. [23], has obtained promising performance in many computer vision tasks, including but not limited to image classification [7, 22], object detection [21, 33, 35], face recognition [38, 43], image denoising [18], image segmentation [19, 44], retrieval and pedestrian detection [1, 32]. They were mainly designed for images. The key enabling factors are techniques to scale up the networks to tens of millions of parameters and massive labeled dataset that can support the learning process. Under these conditions, CNNs have been proven to be robust in learning powerful and interpretable image features [39]. Motivated by the successful applications of CNNs in image domains, some researchers extended CNN models to analyze videos. Karpathy et al. [20] studied the performance of CNNs in large-scale video classification, where the networks have access to not only the appearance information present in single and static images, but also their complex temporal evolution. Simonyan et al. [40] proposed a two-stream deep Convolutional Networks (ConvNets) architecture which jointly incorporates spatial and temporal information to recognize actions in videos. Based upon a new encoding strategy, Xu et al. [49] introduced a discriminative video representation for event detection over a large-scale video dataset when only limited hardware resources are available.

Compared to the extensive studies on images, there is relatively sparse work on implementing CNNs to different video applications. This is attributable to the sequential structures of videos. CNN models indeed capture spatial features but not the temporal variance of video frames. Another reason may be that video data contain multiple modalities, while images only have mono-modality. To address the first problem, some studies combined CNNs with different RNN networks. For example, the work in [48] presents a hybrid deep learning framework for video classification, which is able to model static spatial information, short-term motion, as well as long-term temporal cues in videos. The spatial and short-term motion features are extracted separately by two CNNs. In addition, LSTM networks are applied on top of the two features to further model longer-term temporal cues. Their experimental results demonstrate that the sequence-based LSTM is highly complementary to the traditional classification strategy without considering the temporal frame orders.

However, there is little work focus on the second issue, namely the video data contain multi-modal sequential structures. In this paper, we propose a parallel LSTMs method to model the triple sequences in micro-videos. Micro-videos are very short, usually lasting for 6-15 seconds, and conveying only one or a few concepts, i.e., sparsity. In the light of this, we present a CNN model based dictionary learning procedure to fuse the multi-modal information and learn sparse representations of given micro-videos. In this part, the convolutional filters to be learned are equivalent to dictionary atoms, and the activation function applied to the filtered results can be treated as sparse coding.

3 DATA COLLECTION

In this section, we detail our dataset comprising of data preprocessing and feature extraction.

3.1 Dataset

We are aware that there are two publicly available datasets on the Web, namely, [31] and [50]. However, the authors did not release the original data and the released features were not designed for sequence information. They are thus not suitable for our sequence modeling task. To tackle this problem, we crawled micro-videos from Vine through its public API[5]. We only retained the micro-videos with three modalities[6], venue information, and exactly 6 seconds (According to our statistics, more than 83% of 24,236 micro-videos are 6 seconds exactly.) Meanwhile, we eliminated the venue categories with less than 100 labeled micro-videos. We ultimately gathered 20,093 micro-videos distributed over 22 Foursquare venue categories, as illustrated in Figure 2. Figure 3 shows two micro-videos and they were both selected from the "Garden" venue category. From Figure 3, it can be seen that the topics or concepts expressed by micro-videos are very sparse. It further reveals that we need sparse and conceptual representations for micro-videos.

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

[6] We observed that some micro-videos do not have acoustic or textual modality. As our model is a multi-modal learning model simultaneously manipulating the visual, acoustic, and textual modalities, we filtered out the micro-videos without acoustic modality.

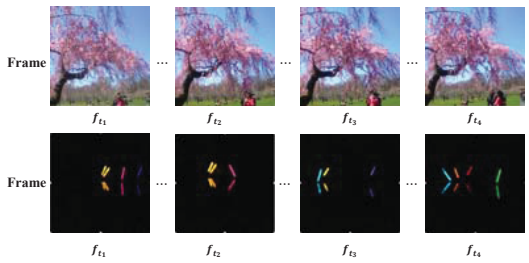

Figure 3: An illustration of two micro-videos from the dataset and the venue category of them are the "Garden".

3.2 Sequence Features

In this part, we introduce the features that we extracted from visual, acoustic, and textual modalities, respectively.

3.2.1 Sequence in the Visual Modality. First of all, we extracted frames from micro-videos with the help of FFmpeg[7], one frame per 0.5 second. We then extracted high-level semantics to represent each visual frame. As analyzed before, CNNs have been recognized as a powerful model to capture the visual concepts of images. We thus employed the AlexNet model to extract the visual features through the publicly available Caffe[8]. The model was pre-trained on a set of 1.2 million clean images from ILSVRC12[9] and it hence can provide a robust initialization for recognizing semantics. We finally obtained 12 frames from each micro-video and a 4,096 dimensional feature vector for each frame.

3.2.2 Sequence in the Acoustic Modality. The acoustic modality in the micro-videos contain useful cues or hints on places. For example, within a restaurant, acoustic modality capture employees welcoming customers to the restaurant and answering their questions to menus or specialties. Considering the stadium as another example, the audio clips may signal the cheers. Meanwhile, the acoustic information is especially useful for the cases where the visual features are too diverse or cannot carry satisfied information. To extract the acoustic sequences, we first segmented each audio channel into 6 clips. We then used Librosa[10] to extract 512 dimensional features from each of these audio clips.

3.2.3 Sequence in the Textual Modality. The textual descriptions of micro-videos, including user generated text and hashtags, can provide strong cues on micro-video venue estimation. For instance, this description "Vining the # Dancing at the Disney Park" clearly indicates that the venue category is a theme park. In particular, we first eliminated the non-English characters, followed by removing the stop words. We then employed Word2Vector[11] tool to generate a 100 dimensional feature vector for each word in the textual description.

[7]https://www.ffmpeg.org/tool.
[8]https://github.com/BVLC/caffe.
[9]http://www.image-net.org/challenges/LSVRC/2012/.
[10]https://github.com/bmcfee/librosa.
[11]https://github.com/klb3713/sentence2vec.

4 OUR PROPOSED EASTERN MODEL

Our proposed end-to-end EASTERN model is comprised of three components: 1) characterizing the sequential structures of three modalities via a parallel LSTMs method; 2) mapping the learned three sequential features into a common space to generate the same length vectors; and 3) finally learning the sparse and conceptual representation via a CNN. In this section, we detail them one by one.

4.1 Notations

For notations, we use bold capital letters (e.g., $\mathbf{X}$) and bold lowercase letters (e.g., $\mathbf{x}$) to denote matrices and vectors, respectively. We employ non-bold letters (e.g., x) to represent scalars. If not clarified, all vectors are in column forms.

Suppose we have a set of N micro-video samples. Each has M modalities ($M = 3$ in this work) and is associated with one of T venue categories. In this work, we treat each venue category as a task. We utilize $\mathbf{X}^m = [\mathbf{x}_1^m, \ldots, \mathbf{x}_N^m] \in \mathbb{R}^{D_m \times N}$ to denote the representations of N samples within a D_m dimensional feature space from the m^{th} modality.

4.2 Sequential Feature Learning

In this subsection, we will first review the structure of single RNN and LSTM models. We then introduce the multi-modal sequential feature leaning method by parallel LSTMs.

4.2.1 RNN and LSTM Models. RNN is a type of neural networks particularly suited for modeling sequential data. At each time step t, RNN takes the input vector $\mathbf{x}_t \in \mathbb{R}^{n_1}$ and the hidden state vector $\mathbf{h}_{t-1} \in \mathbb{R}^{m_1}$ as input elements. It produces the next hidden state $\mathbf{h}_t$ by applying the following recursive operation:

$$\mathbf{h}_t = f(\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{h}_{t-1} + \mathbf{b}), \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{m_1 \times n_1}$, $\mathbf{U} \in \mathbb{R}^{m_1 \times m_1}$, and $\mathbf{b} \in \mathbb{R}^{m_1}$ are parameters of an affine transformation; and f is an element-wise nonlinearity. Technically, RNN can summarize all historical information up to time t with the hidden state $\mathbf{h}_t$. In practice, learning long-range dependencies with a RNN is difficult due to vanishing gradients, which occurs as a result of the Jacobian's multiplicativity w.r.t time.

LSTM addresses the problem of learning long range dependencies by augmenting the RNN with a memory cell vector $\mathbf{c}_t \in \mathbb{R}^{m_1}$ at each time step. Concretely, one step of LSTM takes $\mathbf{x}_t$, $\mathbf{h}_{t-1}$, and $\mathbf{c}_{t-1}$ as input and produces $\mathbf{h}_t$ and $\mathbf{c}_t$ via the following intermediate calculations:

$$\begin{cases} \mathbf{i}_t = \sigma(\mathbf{W}_i\mathbf{x}_t + \mathbf{U}_i\mathbf{h}_{t-1} + \mathbf{b}_i), \\ \mathbf{f}_t = \sigma(\mathbf{W}_f\mathbf{x}_t + \mathbf{U}_f\mathbf{h}_{t-1} + \mathbf{b}_f), \\ \mathbf{o}_t = \sigma(\mathbf{W}_o\mathbf{x}_t + \mathbf{U}_o\mathbf{h}_{t-1} + \mathbf{b}_o), \\ \mathbf{g}_t = \tanh(\mathbf{W}_g\mathbf{x}_t + \mathbf{U}_g\mathbf{h}_{t-1} + \mathbf{b}_g), \\ \mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t, \\ \mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t), \end{cases} \tag{2}$$

where $\sigma(\cdot)$ and $\tanh(\cdot)$ are the element-wise sigmoid and hyperbolic tangent functions; $\odot$ is the element-wise multiplication operator; and $\mathbf{i}_t$, $\mathbf{f}_t$, and $\mathbf{o}_t$ are respectively treated

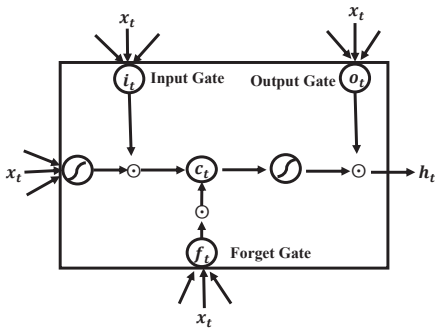

Figure 4: Schematic illustration of a LSTM unit.

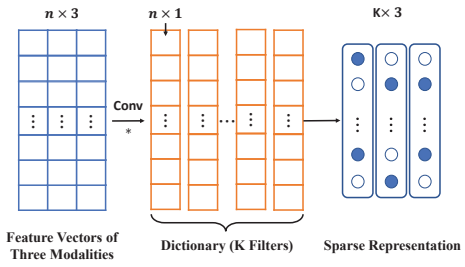

Figure 5: Illustration of CNN-based dictionary learning (each filter can be seen as a dictionary atom).

as input, forget, and output gates. At $t = 1$, $\mathbf{h}_0$ and $\mathbf{c}_0$ are initialized to zero vectors. Parameters of the LSTM are $\mathbf{W}_j$, $\mathbf{U}_j$, and $\mathbf{b}_j$ for $j \in \{i, f, o, g\}$. Figure 4 shows an unit of a LSTM network. Memory cells in the LSTM are additive w.r.t time, alleviating the gradient vanishing problem.

4.2.2 Parallel LSTMs. As analyzed before, the visual, textual, and acoustic modalities of micro-videos can be treated as sequential data. To capture the embedded sequential cues across three modalities, we devise a parallel LSTMs network as follows,

$$\begin{cases} \mathbf{i}_t^m = \sigma(\mathbf{W}_i^m \mathbf{x}_t^m + \mathbf{U}_i^m \mathbf{h}_{t-1}^m + \mathbf{b}_i^m), \\ \mathbf{f}_t^m = \sigma(\mathbf{W}_f^m \mathbf{x}_t^m + \mathbf{U}_f^m \mathbf{h}_{t-1}^m + \mathbf{b}_f^m), \\ \mathbf{o}_t^m = \sigma(\mathbf{W}_o^m \mathbf{x}_t^m + \mathbf{U}_o^m \mathbf{h}_{t-1}^m + \mathbf{b}_o^m), \\ \mathbf{g}_t^m = \tanh(\mathbf{W}_g^m \mathbf{x}_t^m + \mathbf{U}_g^m \mathbf{h}_{t-1}^m + \mathbf{b}_g^m), \\ \mathbf{c}_t^m = \mathbf{f}_t^m \odot \mathbf{c}_{t-1}^m + \mathbf{i}_t^m \odot \mathbf{g}_t^m, \\ \mathbf{h}_t^m = \mathbf{o}_t^m \odot \tanh(\mathbf{c}_t^m), \\ \quad m \in \{v, a, e\}, \end{cases} \tag{3}$$

where $\mathbf{x}_t^v \in \mathbb{R}^{D_v}$, $\mathbf{x}_t^a \in \mathbb{R}^{D_a}$, and $\mathbf{x}_t^e \in \mathbb{R}^{D_e}$ respectively represents the visual, acoustic, and textual input sequences at time t. They are not required to have the same length, namely the three LSTM networks are independent and they have different hidden units. $\mathbf{W}_j^m$, $\mathbf{U}_j^m$, and $\mathbf{b}_j^m$ for $j \in \{i, f, o, g\}$ are the parameters of the m^{th} modality LSTM. We extract the final hidden representation from the last LSTM step as the output of the parallel LSTMs.

4.3 Feature Embedding

For each input micro-video, the parallel LSTMs output three heterogeneous feature vectors with different lengths. Although these feature vectors encode distinct sequential cues, they capture the characteristics of the same micro-video. That is to say they are closely related and we can project them into a common space [47],

$$\begin{cases} \widetilde{\mathbf{x}}_v = \mathbf{W}_v \mathbf{h}_v + \mathbf{b}_v, \\ \widetilde{\mathbf{x}}_a = \mathbf{W}_a \mathbf{h}_a + \mathbf{b}_a, \\ \widetilde{\mathbf{x}}_t = \mathbf{W}_t \mathbf{h}_t + \mathbf{b}_t, \end{cases} \tag{4}$$

where $\widetilde{\mathbf{x}}_v$, $\widetilde{\mathbf{x}}_a$, and $\widetilde{\mathbf{x}}_t \in \mathbb{R}^n$ are the visual, acoustic, and textual embedding, respectively; $\mathbf{h}_v$, $\mathbf{h}_a$, and $\mathbf{h}_t$ are the hidden representation from the last LSTM step of the three parallel LSTMs, respectively; n is the dimension of the features in this learned common space; $\mathbf{W}_v$, $\mathbf{W}_a$, and $\mathbf{W}_t$ are embedding matrices; and $\mathbf{b}_v$, $\mathbf{b}_a$, and $\mathbf{b}_t$ are the bias vectors for the three modalities, respectively.

4.4 Sparse Conceptual Representation Learning

Different from the traditional shallow dictionary learning methods, in this part, we present a CNN model to learn the sparse and concept-level representations of micro-videos, whose inputs are the projected sequential features. In addition, we argue that the visual, acoustic, and textual modalities of micro-videos are not independent but highly correlated. We will show how our presented CNN uses correlations of different modalities.

Figure 5 illustrates the sparse conceptual presentation learning procedure. In particular, assuming that we obtain a $n \times 3$ matrix from the preceding common space whereby each column in the matrix denotes one sequential feature vector of a modality. We aim to learn a sparse and concept-level representation based upon a dictionary. Representation learning over a dictionary with K atoms is equivalent to applying K linear filters $n \times 1$ to each input feature vector (each column in the matrix). The sparse coding solver will then iteratively process K coefficients. The outputs are $K \times 3$ coefficients, namely applying the K filters to the $n \times 3$ matrix. This is illustrated at the right part of Figure 5. As a result, we can treat the convolutional neural network as a dictionary learning method.

In our model, we adopt a shared dictionary to generate sparse coding for the three modalities. Therefore, the learned sparse conceptual representations capture the correlations among different modalities. Formally, our multi-modal sparse representation learning is expressed as an operation F:

$$F(\widetilde{\mathbf{x}}_j) = \max(0, \mathbf{W}_1 * \widetilde{\mathbf{x}}_j + \mathbf{b}_1), j \in \{v, a, t\}, \tag{5}$$

where $\mathbf{W}_1$ and $\mathbf{b}_1$ respectively represents the filters and bias, and the symbol $*$ denotes the convolution operation. Here, $\mathbf{W}_1$ corresponds to K filters of support $n \times 1$, where n is the spatial size of a filter. Intuitively, $\mathbf{W}_1$ applies K convolutions to the multi-modal features, and each convolution has a kernel size $n \times 1$. The output is composed of K feature maps, and

the size of each feature map is 1×1. $\mathbf{b}_1$ is a K-dimensional vector, whereby each entry is associated with a filter. We apply the Rectified Linear Unit (ReLU, max(0; x)) [30] to the filter responses.

After getting the sparse codes of these three modalities based on a shared dictionary, we cascade the three feature vectors into one and then feed it into a softmax classifier. Our method is trained as an end-to-end deep learning model.

5 EXPERIMENT

In this section, we applied our proposed model to the application of micro-video categorization.

5.1 Experimental Settings

To thoroughly measure our model and the baselines, we employed Macro-F [24] and Micro-F [36] as the evaluation metrics to measure the model performance from different angles. Macro-F is the average of the F values from all classes, while Micro-F can be calculated by regarding all classes as the same class and then calculate its F value [6, 25]. Both Macro-F and Micro-F metrics reach their best at 1 and worst at 0.

The reported experimental results in this paper were based on our dataset mentioned above. We randomly sampled 18,000 micro-videos from our dataset for training and 2,093 ones for testing. It is worth emphasizing that we repeated the sampling procedure ten times and reported the average experimental results over the ten sets. Besides, we carried out experiments with the help of Tensorflow[12], selecting function AdamOptimizer as our optimizer, and function softmax_cross_entropy_with_logits as the loss. And we trained it over a server equipped with 16 Tesla K80s.

5.2 Baselines

To demonstrate the effectiveness of our proposed EASTERN model, we compared it with several state-of-the-art baselines:

- Only Three Parallel LSTMs (LSTMs): This is a simple end-to-end baseline that utilizes three parallel LSTMs to extract modality-specific sequential features from given micro-videos and then feed the learned features into a softmax classifier.
- Combining a Shared LSTM with a Dictionary learning method (LSTM+DL): This model forces the three modalities to share the same LSTM and then combines it with a CNN-based dictionary learning model.
- Task-driven Multi-modal Dictionary Learning (TMDL) [8]: It is a task-driven multi-modal dictionary learning method which ensures that samples in each class share a class-specific dictionary and combines a linear classifier into a model.
- Data-driven Multi-modal Dictionary Learning (DMDL) [3]: It is a data-driven multi-modal dictionary learning method which minimizes the reconstruction error and constrains different modalities to share the same concepts.
- Tree-guided Multi-task Learning (TRUMANN) [50]: It is a tree-guided multi-task multi-modal learning model. This model intelligently learns a common feature space from multi-modal heterogeneous spaces and simultaneously learns a classifier based on the representation of each micro-video over the learned common space.

[12] https://www.tensorflow.org/.

Table 1: Performance comparison between our proposed model and several state-of-the-art baselines in terms of Micro-F, Macro-F, and significance test value. (p-value*: p-value over Macro-F)

Method	Micro-F	Macro-F	p-value*
LSTMs	57.08%	28.73%	7.71e-03
LSTM+DL	32.13%	2.24 %	1.41e-20
TMDL	52.75%	24.60%	1.10e-07
DMDL	52.15%	23.44%	3.28e-08
TRUMANN	53.32%	18.23%	1.36e-13
EASTERN	**59.51%**	**30.57%**	-

5.3 Performance Comparison among Models

We trained our model and the baselines over the training set and verified them over the testing one. The results are summarized in Table 1. From Table 1, we have the following observations: 1) Task-driven dictionary learning model TMDL outperforms the data-driven one DMDL. This demonstrates that encoding label information is able to learn more discriminative dictionaries. 2) TRUMANN achieves relatively better results than that of the multi-modal dictionary learning paradigms TMDL and DMDL. This reveals that there are correlations among different modalities of micro-videos in our dataset and these correlations can benefit micro-video classification. 3) The shallow models, TMDL, DMDL, and TRUMANN, substantially outperform the deep one LSTM+DL. LSTM+DL implicitly assumes that the three modalities share the same sequential structure, which may not always be the case in the real-world scenarios. For example, birds chirping at time t in the acoustic modality may not guarantee a bird appearing at the same time t in the visual modality. 4) Our proposed model substantially surpasses LSTMs. This verifies that CNN-based dictionary learning can generate discriminate representations that are sparse and in concept-level, and such representations are beneficial to venue category classification. 5) Comparing our proposed EASTERN model with the LSTM+DL, we find that ours performs better. It demonstrates that shared LSTM structure is not effective for micro-video classification, since the sequential structures of the three modalities may be inconsistent.

5.4 Performance Comparison on Epochs

To justify the robustness of our proposed model, we comparatively explored the performance of our model and the

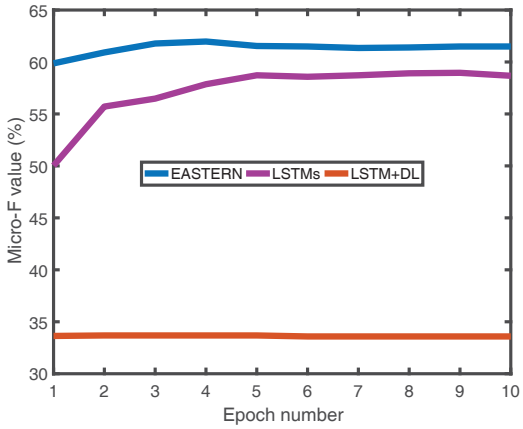

Figure 6: Performance of deep learning based models versus epoch number.

Table 2: Performance comparison between our proposed model and a baseline without considering the sequential information. (p-value*: p-value over Macro-F)

Method	Micro-F	Macro-F	p-value*
CONV	46.10%	12.37%	9.51e-14
EASTERN	**59.51%**	**30.57%**	-

baselines by varying the number of epochs. Figure 6 shows the results. From Figure 6, it can be seen that : 1) The performance of all these methods except LSTM+DL model rises fast as the number of epochs linearly increases. Their curves then gradually go up to a steady state. LSTM+DL model nearly does not change along with the number of epochs. This implies that it reached the stable state at the first epoch. 2) The dictionary learning based method EASTERN consistently and remarkably outputs a higher accuracy as compared to that of non-dictionary learning LSTMs. And 3) our method outperforms other baselines regardless of the number of epochs. This shows the robustness of our model.

5.5 Comparison on Sequential Features

To shed light on the effectiveness of the sequential features generated by the involved LSTM network, we devised a baseline CONV by eliminating the first part of our model, namely by removing the three parallel LSTMs. They were trained on the training set and justified on the testing set. The averaged experiments results are listed in Table 2. From this table, it can be seen that the performance of CONV is not satisfactory. This may be due to that the CONV model only considers the modality correlations to represent each micro-video and overlooks the discriminative sequential structures. In addition, the correlation information among different modalities is insufficient to classify micro-videos. In the light of this, sequential features come more important to classify micro-videos.

Table 3: Performance of our proposed EASTERN model with different modality combinations. (p-value*: p-value over Macro-F)

Modality	Micro-F	Macro-F	p-value*
Visual	56.15%	26.85%	2.00e-05
Audio	43.68%	11.11%	1.46e-17
Text	40.85%	7.25%	2.43e-18
Visual+Audio	58.00%	28.74%	1.95e-02
Visual+Text	57.95%	28.63%	1.16e-02
Audio+Text	48.21%	14.17%	1.68e-15
All	**59.51%**	**30.57%**	-

Table 4: Parameter settings of our proposed model.

Feature	Component	Attributes
Visual	LSTM	500 hidden units
Audio	LSTM	300 hidden units
Text	LSTM	80 hidden units
Fusion	Embedding & Dictionary	150 units & 100 filters

5.6 Comparison on Modality Combination

We also studied the performance of our model with different modality combinations. The results are summarized in Table 3. It can be seen that: 1) The visual modality performs better than the textual and acoustic ones. This is due to the fact that the visual modality is more intuitive to signal venue information than that of acoustic and textual ones. In addition, it reveals that the CNN features are capable of capturing the prominent visual characteristics of venue categories. 2) For most of the micro-videos, a single modality is insufficient to estimate the venue category, but combining them can largely enhance the performance. 3) From the results of combining visual modality with one of the other two modalities, we notice that the acoustic one conveys more important cues on venue categories than the textual modality w.r.t Micro-F and Macro-F metrics. This is because the textual descriptions are of low quality, noisy, missing, sparse, and even irrelevant to the venue categories. And 4) our proposed EASTERN achieves the best performance over three modalities. This further justifies the old saying "two heads are better than one" and shows that multi-modalities is complementary instead of conflicting.

5.7 Parameter Settings

We also carried out experiments to elaborate the parameter settings of our proposed model EASTERN and other deep baseline models. The parameters of our proposed EASTERN include the hidden numbers of three LSTMs, dimension of the common space, and the number of CNN filters. We performed grid search to seek the optimal settings of these parameters by varying one and fixing the others. And the parameters corresponding to the best Micro F-score were used to report the final results. We show the parameters of our model in Table 4. In particular, the hidden units in visual, acoustic, and textual LSTMs are 500, 300, and 80, respectively. The dimension of the common space is 150 and

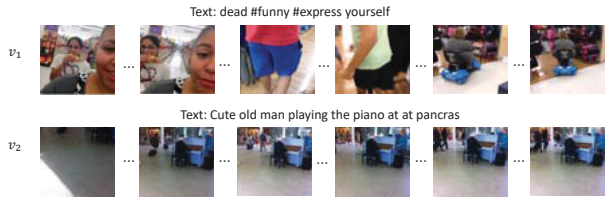

Figure 7: Exemplars of successfully classified micro-videos.

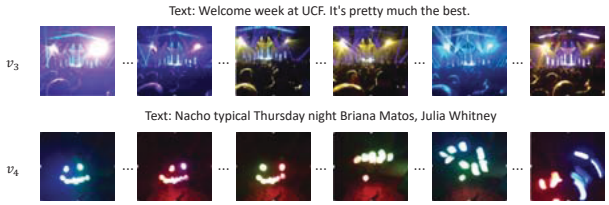

Figure 8: Exemplars of wrongly classified micro-videos.

the number of convolutional filters is 100 (i.e., the dictionary has 100 atoms). In this paper, to reduce the search space, the numbers of units for all parts in the deep baseline models are restricted to the same as ours.

5.8 Case Study

To gain the deep insights into our proposed EASTERN model, we illustrated a few successful and failed classification results of micro-videos. In particular, the successful and the failed cases are shown in Figure 7 and 8, respectively. In addition, we also display the classification results of the four micro-videos (v_1-v_4 in Figure 7 and 8) by their baselines in Table 5.

Micro-video v_1 in Figure 7 contains supermarket shelves, snacks, and other goods. Obviously, the venue category of this one is "Miscellaneous Shop". Similarly, from the micro-video v_2 in Figure 7, it can be seen that many pedestrians pulling a suitcase went by the piano. And we can hear sound of trains from its acoustic channel. Therefore, it was captured at "Train Station". However, from the estimation results of v_1 and v_2 in Table 5, we observe that: 1) LSTM+DL, DMDL, and TRUMANN models predict the venue categories of these two examples to be "University". The failure reason of the LSTM+DL model may be that it forces all the three modalities to share the same sequential structure, and hence it loses some discriminative sequential features. As for the latter two, DMDL and TRUMANN, they both overlook the discriminative sequential features. What is more, DMDL is an unsupervised learning model, so it cannot learn discriminative dictionaries to represent micro-videos. 2) For the other three supervised learning methods, TMDL, CONV, and LSTMs, the former two ignore the discriminative sequential features and the latter overlooks the correlations among different modalities. 3) Our proposed model not only considers the discriminative sequential features but also the correlations among different modalities, so it can generate more discriminative and high-level sparse representations. In the light of this, our model performs better than other baselines.

For micro-videos v_3 and v_4 in Figure 8, all the models fail to estimate their venue categories. Because some micro-videos from different venues are very similar. Taking v_3 as an example, it looks like some "General Entertainment" places rather than the "College Basketball Court". Therefore, all the models except LSTM+DL and CONV ones classify it into "General Entertainment". Since LSTM+DL and CONV ignore the discriminative sequential features, they failed to classify v_3 into a correct venue. Regarding the micro-video v_4, it is very similar to some micro-videos in "Garden" category (e.g., the second micro-video in Figure 3). Therefore, our model mistakenly classifies it into "Garden". And other ones ignore either discriminative sequential features or modality correlations, so they also classify it into an inaccurate venue.

Table 5: The classification results of different models for the four selected micro-video samples in Figure 7 and 8.

Video	v_1	v_2	v_3	v_4
LSTM+DL	University	University	University	University
LSTMs	Gym	Art Museum	General Entertainment	General Entertainment
CONV	University	Art Museum	University	University
TMDL	General Entertainment	Art Museum	General Entertainment	Garden
DMDL	University	University	General Entertainment	General Entertainment
TRUMANN	University	University	General Entertainment	University
EASTERN	Miscellaneous Shop	Train Station	General Entertainment	Garden
Ground Truth	Miscellaneous Shop	Train Station	College Basketball Court	Speakeasy

6 CONCLUSION AND FUTURE WORK

In this paper, we present a deep parallel sequence with sparse constraint approach to categorizing venues of micro-videos. This approach is able to jointly model the sequential structures of different modalities and sparse concept-level representations at the same time. This is accomplished by an end-to-end scheme encapsulating three components, namely, triple parallel LSTMs, common space projection, and CNN-based dictionary learning. To justify our scheme, we constructed a large-scale dataset based on Vine. Experimental results demonstrate that our approach is superior to several state-of-the-art baselines. It is worth emphasizing that our model is also applicable to other micro-video analysis tasks, such as popularity prediction. As a side contribution, we have released the source codes and data to facilitate other researchers.

In the future, we plan to transfer the external knowledge to strengthen the sequential structure modeling.

7 ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their valuable comments. This work was supported by the Joint NSFC-ISF Research Program (No.61561146397), jointly funded by the National Natural Science Foundation of China and the Israel Science Foundation. It is also supported in part by the National Basic Research grant (973) (No.2015CB352501) and the One Thousand Talents Plan of China (No.11150087963001).

REFERENCES

[1] Grigory Antipov, Sid-Ahmed Berrani, Natacha Ruchaud, and Jean-Luc Dugelay. 2015. Learned vs. hand-crafted features for pedestrian gender recognition. In *ACM MM*. 1263–1266.

[2] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. 2011. Sequential deep learning for human action recognition. In *HBU*. 29–39.

[3] Soheil Bahrampour, Nasser M Nasrabadi, Asok Ray, and William Kenneth Jenkins. 2016. Multimodal task-driven dictionary learning for image classification. *IEEE TIP* 25, 1 (2016), 24–38.

[4] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE NN* 5, 2 (1994), 157–166.

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

[6] Ken Chen, Bao-Liang Lu, and James T Kwok. 2006. Efficient classification of multi-label and imbalanced data using min-max modular classifiers. In *IEEE IJCNN*. 1770–1775.

[7] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. 2012. Multi-column deep neural networks for image classification. In *IEEE CVPR*. 3642–3649.

[8] Cheng Deng, Xu Tang, Junchi Yan, Wei Liu, and Xinbo Gao. 2016. Discriminative dictionary learning with common label alignment for cross-modal retrieval. *IEEE MM* 18, 2 (2016), 208–218.

[9] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE CVPR*. 2625–2634.

[10] Chao Dong, Change Loy Chen, Kaiming He, and Xiaoou Tang. 2016. Image super-resolution using deep convolutional networks. *IEEE PAMI* 38, 2 (2016), 295–307.

[11] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *ACM MM*. 835–838.

[12] Felix A Gers and E Schmidhuber. 2001. LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE NN* 12, 6 (2001), 1333–1340.

[13] Felix A Gers, Nicol N Schraudolph, and Jürgen Schmidhuber. 2002. Learning precise timing with LSTM recurrent networks. *Journal of Machine Learning Research* 3, Aug (2002), 115–143.

[14] Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *ICML*, Vol. 14. 1764–1772.

[15] Alex Graves and Jürgen Schmidhuber. 2009. Offline handwriting recognition with multidimensional recurrent neural networks. In *NIPS*. 545–552.

[16] Sepp Hochreiter, Martin Heusel, and Klaus Obermayer. 2007. Fast model-based protein homology detection without alignment. *Bioinformatics* 23, 14 (2007), 1728–1736.

[17] Sepp Hochreiter and Jürgen Schmidhuber. 1997. LTSM can solve hard time lag problems. In *NIPS*. 473–479.

[18] Viren Jain and Sebastian Seung. 2009. Natural image denoising with convolutional networks. In *NIPS*. 769–776.

[19] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*. 675–678.

[20] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *IEEE CVPR*. 1725–1732.

[21] Markus Koskela and Jorma Laaksonen. 2014. Convolutional network features for scene recognition. In *ACM MM*. 1169–1172.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*. 1097–1105.

[23] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[24] Bruno Lepri, Nadia Mana, Alessandro Cappelletti, and Fabio Pianesi. 2009. Automatic prediction of individual performance from thin slices of social behavior. In *ACM MM*. 733–736.

[25] David D Lewis. 1991. Evaluating text categorization. In *HLT*. 312–318.

[26] Guang Li, Shubo Ma, and Yahong Han. 2015. Summarization-based video caption via deep neural networks. In *ACM MM*. 1191–1194.

[27] Yehao Li, Ting Yao, Tao Mei, Hongyang Chao, and Yong Rui. 2016. Share-and-chat: Achieving human-level video commenting by search and multi-view embedding. In *ACM MM*. 928–937.

[28] Lie Lu, Hao Jiang, and HongJiang Zhang. 2001. A robust audio classification and segmentation method. In *ACM MM*. 203–211.

[29] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, Vol. 2. 3–3.

[30] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*. 807–814.

[31] Phuc Xuan Nguyen, Gregory Rogez, Charless Fowlkes, and Deva Ramanan. 2016. The open world of micro-videos. *arXiv preprint arXiv:1603.09439* (2016).

[32] Wanli Ouyang and Xiaogang Wang. 2013. Joint deep learning for pedestrian detection. In *IEEE ICCV*. 2056–2063.

[33] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *IEEE CVPR*. 779–788.

[34] Jimmy Ren, Yongtao Hu, Yu-Wing Tai, Chuan Wang, Li Xu, Wenxiu Sun, and Qiong Yan. 2016. Look, listen and learn-a multimodal LSTM for speaker identification. *arXiv preprint arXiv:1602.04364* (2016).

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*. 91–99.

[36] Chris Sanden and John Z Zhang. 2011. Enhancing multi-label music genre classification through ensemble techniques. In *ACM SIGIR*. 705–714.

[37] Jürgen Schmidhuber, Daan Wierstra, and Faustino Gomez. 2005. Evolino: Hybrid neuroevolution optimal linear search for sequence learning. In *IJCAI*. 853–858.

[38] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE CVPR*. 815–823.

[39] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In *IEEE CVPR*. 806–813.

[40] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *NIPS*. 568–576.

[41] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. 2015. Unsupervised learning of video representations using LSTMs. In *ICML*. 843–852.

[42] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*. 3104–3112.

[43] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *IEEE CVPR*. 1701–1708.

[44] Srinivas C Turaga, Joseph F Murray, Viren Jain, Fabian Roth, Moritz Helmstaedter, Kevin Briggman, Winfried Denk, and H Sebastian Seung. 2010. Convolutional networks can learn to generate affinity graphs for image segmentation. *Neural computation* 22, 2 (2010), 511–538.

[45] Vivek Veeriah, Naifan Zhuang, and Guo-Jun Qi. 2015. Differential recurrent neural networks for action recognition. In *IEEE ICCV*. 4041–4049.

[46] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. 2016. Image captioning with deep bidirectional LSTMs. In *ACM MM*. 988–997.

[47] Xiang Wang, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. 2017. Item Silk Road: Recommending Items from Information Domains to Social Users. In *ACM SIGIR*.

[48] Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue. 2015. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *ACM MM*. 461–470.

[49] Zhongwen Xu, Yi Yang, and Alex G Hauptmann. 2015. A discriminative CNN video representation for event detection. In *IEEE CVPR*. 1798–1807.

[50] Jianglong Zhang, Liqiang Nie, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat Seng Chua. 2016. Shorter-is-better: Venue category estimation from micro-video. In *ACM MM*. 1415–1424.