Neural Multimodal Cooperative Learning Towards Micro-video Understanding

Yinwei Wei, Xiang Wang, Weili Guan, Liqiang Nie , *Member, IEEE*,
Zhouchen Lin , *Fellow, IEEE*, and Baoquan Chen , *Senior Member, IEEE*

Abstract—The prevailing characteristics of micro-videos result in the less descriptive power of each modality. The micro-video representations, several pioneer efforts proposed, are limited in implicitly exploring the consistency between different modality information but ignore the complementarity. In this paper, we focus on how to explicitly separate the consistent features and the complementary features from the mixed information and harness their combination to improve the expressiveness of each modality. Towards this end, we present a **N**eural **M**ultimodal **C**ooperative **L**earning model (**NMCL**) to split the consistent component and the complementary component by a novel relation-aware attention mechanism. Specifically, the computed attention score can be used to measure the correlation between the features extracted from different modalities. And then, a threshold is learned for each modality to distinguish the consistent and complementary features, according to the score. Thereafter, we integrate the consistent parts to enhance the representations and supplement the complementary ones to reinforce the information in each modality. As to the problem of redundant information, which may cause overfitting and is hard to distinguish, we devise an attention network to dynamically capture the features which closely related the category and output a discriminative representation for prediction. Experimental results on a real-world micro-video dataset show that NMCL outperforms state-of-the-art methods. Further studies verify the effectiveness and cooperative effects brought by the attentive mechanism.

Index Terms—Cooperative Learning, Venue Category Estimation, Attention Model, Consistency and Complementarity.

1 Introduction

The proliferation of Web 2.0 and portable devices contributes to the success of micro-videos. As a new media type, micro-videos allow the users to record their daily life within a few seconds and share over social media platforms (e.g., Vine[1], Instagram[2], and Kwai[3]). The properties of easy-to-operate, instant sharing, and down-to-the-earth contents make the platforms unexpectedly popular, especially among the grassroots. Considering Kwai as an example, as of September 2017, it had attracted over 600 million registered users, and 87 million of them are active for approximately 60 minutes per day, producing around 10 million micro-videos[4].

Different from the traditional long videos, a micro-video is usually recorded at one specific spot without any post-edit. As such, users can associate each micro-video with the specific geo-location tag (e.g., Beijing's Olympic Basketball Stadium) to indicate where it is captured. As known, geographic information benefits many services, such as location-based search, recommendation, and social networking. However, in real-world scenarios, few users tag their micro-videos with specific geographic information owing to privacy concerns. According to our statistics [1], only about 1.22% of two million micro-videos in Vine were labelled with locations.

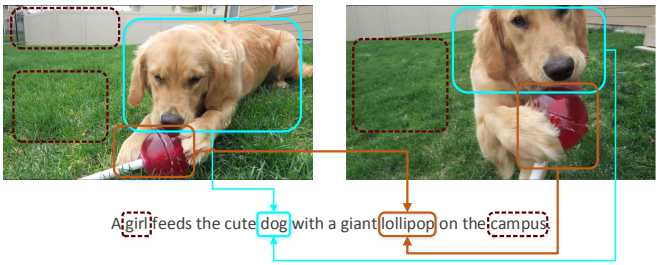


Fig. 1: Exemplar demonstration of the correlation between the visual modality and textural modality. The blue and brown boxes show the consistent information and the red dashed boxes show the complementary ones, respectively.

Despite its significance, we have to mention that it is hard, if not impossible, to infer the specific location information, such as "American Airlines Arena in Florida, USA". Instead, we turn to infer the venue category of a given micro-video, such as "Basketball Court". Technically speaking, venue category estimation of micro-videos is usually treated as a multimodal fusion problem and solved by integrating the geographic cues from visual, acoustic, and textual modalities of micro-videos. Several pioneer efforts have been dedicated to the task. Lazebnik *et al.* [2]

- *Y. Wei, L. Nie, and B. Chen are with the College of Computer Science and Technology, Shandong University, Qingdao, Shandong 266237, China. (e-mail: weiyinwei@hotmail.com; nieliqiang@gmail.com; baoquan@sdu.edu.cn).*
- *X. Wang is with the School of Computing, National University of Singapore, Singapore. (e-mail: xiangwang@u.nus.edu).*
- *W. Guan is with the Hewlett Packard Enterprise Singapore, Singapore. (e-mail: honeyguan@gmail.com).*
- *Z. Lin is with the Key Laboratory of Machine Perception (Ministry of Education), School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, P.R. China, and also with the Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai 200240, P.R. China (e-mail: zlin@pku.edu.cn).*

1. https://vine.co/.
2. https://www.instagram.com/.
3. https://www.kuaishou.com/.
4. http://tech.china.com/article/20170904/2017090455657.html.

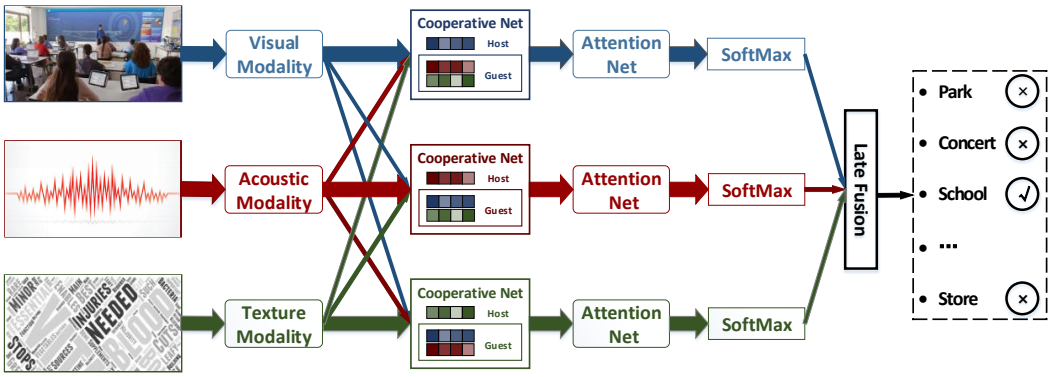


Fig. 2: An illustration of our framework. It separates the consistent features from the complementary ones and enhances the expressiveness of each modality via the proposed cooperative net. Then, it selects the features to generate a discriminative representation in the attention network towards venue category estimation.

proposed the spatial pyramid matching to incorporate the spatial layout into bag-of-word for predicting the category. Singh *et al.* [3] discovered and leveraged the concept of mid-level discriminative parts to classify the venue. In addition, deep convolutional networks are exploited for the classification by Zhou *et al.* [4], and a large-scale Places dataset is introduced as well. Towards the micro-video venue prediction, Zhang *et al.* [1] projected three modalities into a common space, whereby a tree-guided multi-task scheme is employed to capture and model the hierarchical relatedness among venue categories. To better characterize the sequential structure and sparse concepts, Liu *et al.* [5] presented a model to first jointly sew up the parallel Long Short-Term Memory (LSTM) models and then integrate a Convolutional Neural Network (CNN) model. For different modalities, they mapped the features into a common space considering the consistency in the time sequence. More recently, Nie *et al.* [6] proposed a deep transfer model (DARE) to alleviate the low-quality of the acoustic by harnessing the external sound knowledge and fuse the multimodal information by utilizing the consistency among multiple modalities, leading to better performance. In all of these methods, to obtain a joint representation for each micro-video, the authors introduced a common space and projected the heterogeneous data into this space. However, the projection inevitably causes some modal-specific information loss and affects the expression of the micro-video.

Upon further analysis, the current methods are restricted to only fusing the common (a.k.a., consistent) cues over multiple modalities. In fact, beyond the consistency, the relations among multiple modalities are much more sophisticated. For instance, complementarity is another equally important relation among modalities. Moving one step forward, in this work, we shed light on the cooperative relations, comprising the or comprised of the consistent and complementary components. We refer to the consistent component as the same information appearing in more than one modality in different forms. As shown in Figure 1, a red candy displaying in the visual modality and the text of "lollipop" describe the consistency. By contrast, the complementary component represents the exclusive information appearing only in one modality. For instance, it is hard to find the equivalent in other modalities in Figure 1 of the textual concept of "girl" or the visual concept of "grass". To supercharge a multimodal prediction scheme with such cooperative relations, the multimodal cooperation shall be able to: 1) enhance the confidence of the same evidence from various views via consistent regularization, and 2) provide a comprehensive representation from the exclusive perspective of complementary component. Nevertheless, characterizing and modeling multimodal cooperation is non-trivial due to the following challenges: 1) Consistent and complementary information are often mixed. How to separate it from different modalities is largely untapped. And 2) after separation, it is difficult to associate them with each other, since they are orthogonal.

In recently, several approaches [7, 8, 9, 10, 11, 12] have been proposed to integrate the multimodal information. For instance, Wang *et al.* [8] leveraged a hybrid multimodal fusion strategy to integrate the learned visual and text features, in which the early fusion part concatenates the multimodal features and the late fusion part is used to classification. Yang *et al.* [9] proposed a multilayer and multimodal fusion framework for video classification. In this framework, a robust boosting model is employed to learn the optimal combination of multiple layers and modalities. However, these methods cannot distinguish and represent the correlations (e.g., consistency and complementarity) between different modalities, which benefit the expressiveness of each modality. Therefore, we present an deep multimodal cooperative learning approach which explicitly models the correlations between different modalities and enhances the representation of each modality to estimate the venue categories of micro-videos. As illustrated in Figure 2, the features are firstly extracted from each modality and fed into three cooperative peer nets. In each cooperative net, we respectively treat one modality as the host and the rest as the guests. Then we obtain the augmented feature vectors as the output of the cooperative nets. Following that, each vector is fed into an attention net followed by a late fusion over the prediction results from different softmax functions. Stepping into the cooperative net as demonstrated in Figure 3, the structure is symmetric. In particular, on the left hand side, we first concatenate the guest modalities and estimate the relevance between each dimension of the combining vector and the host vector. As to the combined vector, a gate with the

learned threshold is used to separate its consistent part and complementary part. An analogous process is applied to the right hand side. Thereafter, two consistent parts are fused with a deep neural network model, and the fusion result is ultimately concatenated with the two complementary parts. We validate our model on a publicly accessible benchmark dataset and compare it with several state-of-the-art models.

The main contributions of this work are threefold:

- To the best of our knowledge, this is the first work on multimodal cooperative learning. We clearly define that the cooperative relationship among multimodalities is comprised of consistent and complementary components.
- We devise a cooperative network to automatically distinguish and fuse the consistent and the complementary information among multiple modalities.
- We apply our proposed deep multimodal cooperative learning approach to estimating the venue categories of micro-videos. In addition, we released our codes, parameters, and involved baselines to facilitate other researchers[5].

The remaining of this paper is structured as follows. In Section 2, we briefly review the related work. Section 3 and 4 detail our proposed model and the data collection, respectively. Experimental settings and results analysis are presented in Section 5 followed by the conclusion and future work in Section 6.

2 RELATED WORK

Our work is closely related to multimodal fusion and micro-video understanding.

2.1 Multimodal Fusion

Technically speaking, traditional multimodal fusion approaches consist of early fusion and late fusion.

Early fusion approaches, such as [13, 14], typically concatenate the unimodal features extracted from each individual modality into a single representation to adapt to the learning setting. Following that, one can devise a classifier, such as a neural network, treating the overall representation as the input. However, these approaches generally overlook the obvious fact that each view has its own specific statistical property and ignore the relatedness among views. Hence, it fails to explore the modal correlations to strengthen the expressiveness of each modality and further improve the capacity of the fusion method.

Late fusion performs the learning directly over unimodal features, and then the prediction scores are fused to predict the venue category, such as averaging [15], voting [16] and weighting [17]. Although this fusion method is flexible and easy to work, it overlooks the correlation in the mixed feature space.

In contrast to the early and late fusion, as a new paradigm, multi-view learning exploits the correlations between the representations of the information from multiple modalities to improve the learning performance. It can be classified into three categories: co-training, multiple kernel learning, and subspace learning.

5. https://nicemodel.wixsite.com/nice.

2.1.1 Co-training

Co-training [18] is a semi-supervised learning technique which first learns a separate classifier for each view using the labeled examples. It maximizes the mutual agreement on two distinct views of the unlabeled data by alternative training. Many variants have since been developed. Instead of committing labels for the unlabeled examples, Nigam *et at.* [19] proposed a co-EM approach to running EM in each view and assigned probabilistic labels to the unlabeled examples. To resolve the regression problems, Zhou and Li [20] employed two k-nearest neighbor regressors to label the unknown instances during the learning process. More recently, Yu *et at.* [21] proposed a Bayesian undirected graphical model for co-training through the Gaussian process. The success of the co-training algorithms relies on three assumptions: (a) each view is sufficient to estimate on its own; (b) it is probable that a function predicts the same labels for each view feature; and (c) the views are conditionally independent of the given label. However, these assumptions are too strong to satisfy in practice, especially for the micro-videos with different modalities, whereby the information in each modality is insufficient to generate the same label prediction.

2.1.2 Multiple Kernel Learning

Multiple Kernel Learning [22] leverages a predefined set of kernels corresponding to different views and learns an optimal linear or non-linear combination of kernels to boost the performance. Lanckriet *et at.* [23] constructed a convex Quadratically Constrained Quadratic Program by conically combining the multiple kernels from a library of candidate kernels and applied the method to several applications. To extend this method to a large-scale dataset, Bach *et at.* [24] took the dual formulation as a second-order cone programming problem and developed a sequential minimal optimization algorithm to obtain the optimal solution. Further, Ying and Campbell [25] used the metric entropy integrals and pseudo-dimension of a set of candidate kernels to estimate the empirical Rademacher chaos complexity.

2.1.3 Subspace Learning

Subspace learning [26] obtains a latent subspace shared by multiple views by assuming that the input views are generated from this subspace. The dimensionality of the subspace is lower than that of any input view, so the subspace learning alleviates the "curse of dimensionality". The canonical correlation analysis (CCA) [27] is straightforwardly applied to select the shared latent subspace through maximizing the correlation between the views. Since the subspace is linear, it is impossible to apply CCA to the real-world datasets exhibiting non-linearities. To compensate for this problem, Akaho [28] proposed a kernel variant of CCA, namely KCCA. Diethe *et at.* [29] proposed the Fisher Discriminant Analysis using the label information to find the informative projections, more informative in the supervised learning settings. Recently, Zhai *et at.* [30] studied the multi-view metric learning by constructing embedding projections from multi-view data to a shared subspace. Although

the subspace learning approaches alleviate the "curse of dimensionality", the dimensionality of subspace changes along with the task.

To address the problems mentioned above, it is crucial to subtly leverage the correlation information among modalities. We can roughly divide it into consistent and complementary components, which to our knowledge, has never been studied before. The consistent component mainly refers to the common or shared information among various modalities, such as the visual feature of "red candy" and the textual feature of "lollipop". On the contrary, the complementary one presents typically the exclusive information appearing only in one modality, like the textual feature of "girl". However, none of the models has thoroughly exploited consistent and complementary information. However, the consistent information can be leveraged to alleviate the low-quality of the micro-video, and the complementary information is able to make up for the information insufficiency of the micro-video. Inspired by this, we proposed a cooperative net to capture the consistent and complementary information explicitly.

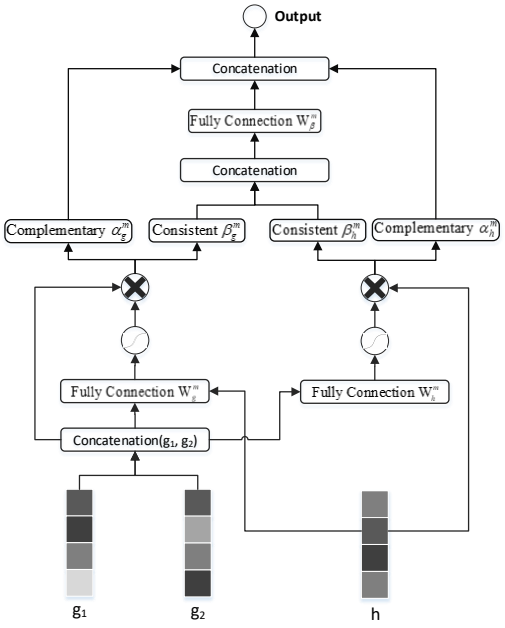


Fig. 3: Illustration of Cooperative Net. The cooperative nets separate the consistent components from the complementary ones, and yield an augmented feature vector comprised of the enhanced consistent vector and complementary vectors.

2.2 Micro-video Understanding

Micro-video, as a new form of medium, has attracted much attentions in recent years, spanning from degree-of-loop assessment [31], hashtag labeling [32], popularity prediction [33], to venue category estimation [1, 5, 6].

2.2.1 Degree-of-loop Assessment

Degree-of-loop is used to evaluate the smoothness of the connection between the first and the last frames of a video. To address this problem, Sano *et at.* [31] proposed a preliminary study by analyzing the spatial and temporal statistics of visual features to classify the loop and non-loop micro-videos.

2.2.2 Hashtag Labeling

Labeling micro-videos with semantic tags can facilitate micro-video understanding, archiving and searching. Towards this end, Chen *et at.* [32] designed a viewpoint-specific and temporally-evolving model to label the micro-videos with hashtags by leveraging the motion and visual features.

2.2.3 Popularity Prediction

Micro-video popularity prediction enables the advertisers to inject or bind their products into/with the to-be popular micro-videos in advance and hence maximizes their profits. Zhang *et at.* [1] proposed a transductive fusion model to integrate social, acoustic, visual and textual modalities for the popularity estimation.

2.2.4 Venue Category Estimation

Given a micro-video, estimating its venue category has enormous commercial potential in many location-based services, such as targeted advertising and location-based organization.

Prior literature studies how to exploit the consistent information encoded in the visual, acoustic, and textual modalities to construct the representations of micro-videos and accordingly identifies their specific venue category. On the basis of the state-of-the-art review, we can roughly categorize the existing efforts to shallow and deep fusion models. The representative work in the shallow learning group is introduced by [1]. This work projects the visual, acoustic, and textual features into a common space via simple mapping, and then adopts a tree-guided multi-task learning classifier to encode the hierarchical structure of the venue categories to eventually identify the leaf category. Owing to the linearity of mappings, the shallow model hardly captures the complex relations among modalities which may further lead to negative transfer and suboptimal prediction performance. To better enrich the expressiveness of micro-videos, researchers turn to deep models. For example, Liu *et at.* [5] employed parallel LSTMs to capture the modality-wise sequential representations, projected the representations into a common space, and adopted CNNs to obtain the sparseness over the underlying concepts. Considering compensating the original acoustic modality, Nie *et at.* [6], more recently, proposed a neural transfer model to enrich the acoustic features via harnessing the external sound knowledge, and automatically learn the nonlinear and complex relations among modalities via a deep neural network.

3 Neural Multimodal Cooperative Model

We first formally define the problem in this section. Assume that we are given a set of N micro-videos $\mathcal{X} = \{x_i\}_{i=1}^N$. For each micro-video $x \in \mathcal{X}$, we segment it into three modalities $\{\mathbf{x}_v, \mathbf{x}_a, \mathbf{x}_t\}$, where v, a, and t denote the visual, acoustic, and textual modality indices, respectively. Let $m \in \mathcal{M} = \{v, a, t\}$ denote the modality indicator, and $\mathbf{x}^m \in \mathbb{R}^{D_m}$ denote the D_m-dimensional feature vector over the m-th modality. In

our work, each micro-video is associated with one of K predefined venue categories, namely a one-hot label vector $\mathbf{y} \in \mathbb{R}^K$, where K refers to the number of venue category.

3.1 Multimodal Early Fusion

In this work, we argue that the information across modalities can be categorized into two parts: the consistent component and the complementary component. For example, let certain features of $\mathbf{x}^v$ indicate the visual concepts of "sunshine" and "crowd"; and some of $\mathbf{x}^a$ describe the acoustic concepts of "wind" and "crowd cheering". From the angle of consistency, the visual concept of "crowd" is consistent with the acoustic concept of "crowd cheering". For complementarity, the visual concept of "sunshine" provides the exclusive signals, as compared to the acoustic one of "wind".

Uncovering the underlying modality relations in micro-videos is already challenging, not to mention different types of relations to the final prediction. To the best of our knowledge, most existing efforts only implicitly model the modality relations during the learning process, leaving the explicit exhibition of relations untouched. Specifically, the deep learning based methods, which feed multimodal features together into a black box multi-layer neural network and output a joint representation, are widely used to characterize the multimodal data. With the deep neural network, the correlations between different features are involved in the new representations. However, the corresponding features cannot be captured and filtered from the vectors. Towards this end, we aim to propose a novel cooperative learning mechanism to leverage the uncovered relations and boost the prediction performance.

3.2 Cooperative Networks

Our preliminary consideration is to explicitly model the relations comprised of the consistent and complementary parts. A viable solution [28, 34] is to project the representations of different modalities into a common latent space, formally as

$$\min_{A^m, B} \frac{\lambda_1}{2} \sum_{m=1}^{M} \|X^m A^m - B\|_F^2 + \frac{\lambda_2}{2} \sum_{m=1}^{M} \|A^m\|_F^2, \quad (1)$$

where $B \in \mathbb{R}^{N \times K'}$ is the representation matrix in the common space learned from all modalities, and K' is the latent feature dimension. $A^m \in \mathbb{R}^{m \times K'}$ is the transformation matrix from the original feature space over the m-th modality to the common space; λ_1 and λ_2 are nonnegative tradeoff parameters. In this solution, the consistent cues should be close to each other since they show the same evidence, whereas the complementary cues in the common space should be distant due to the fact they have no overlapping information. To map the heterogeneous information extracted from a micro-video to the same coordinate, some information, especially the modality-specific information, probably lose during the projection. We term it as the common-specific method. Hence, such direct mapping will lead to suboptimal expressiveness of the method. Although through careful parameter tuning, we can control the loss to a certain extent, it requires extensive experiments which are not easily adapted to other applications.

To avoid such information loss, we devised a novel solution named the cooperative network, in which each modality information was overall retained and augmented by the other modalities. Specifically, this network assigns each dimension of features with a relation score and consequently divides the features into the consistent part and the complementary part. Here the relation score for each feature reflects how consistent the information is derived from the other modalities. The use of relation score endows our model with strong expressiveness and benefits further cooperative learning. In what follows, we elaborate on the key ingredients of the cooperative network.

3.2.1 Relation Score

The goal of the relation score is to select features from each modality, where the underlying information is consistent among modalities. As shown in Figure 3, we treated one specific modality m as the host represented as $\mathbf{h}^m$; the other modalities as the guests denoted as $\mathbf{g}_1^m$ and $\mathbf{g}_2^m$, respectively. Intuitively, we can explicitly capture the varying consistency of the host and guest features by assigning an attentive weight for each feature dimension. The weights are considered as relation scores. Therefore, given the representations of the host and guest modalities, we presented a novel relation-aware attention mechanism to score each feature.

Considering that the consistency should be the correlation between host features and whole guest information, we concatenated all the guest vectors together as follows,

$$\mathbf{g}^m = [\mathbf{g}_1^m, \mathbf{g}_2^m], \quad (2)$$

where the $\mathbf{g}^m$ encapsulates all the features from the guest modalities.

We took each dimension in the host modality as a feature and used the attention mechanism to score the correlation between each feature and the guest information. A higher score suggests that the correlation is more consistent. In contrast, a lower score means that the corresponding feature is independent of the guest modal and can be treated as the complementary features. It is formulated as,

$$\mathbf{s}_{h,i}^m = \sigma(\mathbf{h}_i^m, \mathbf{g}^m), \quad (3)$$

where the $\mathbf{h}_i^m$ and $\mathbf{g}^m$ denote the i-th feature in the host vector and the guest vector, respectively; and σ is the nonlinear function to score the correlation, taking the concatenation of the host and guest vectors as inputs. Moreover, the host vector and the guest vector can be concatenated to evaluate the relation scores for efficiency. Specifically, we fed the guest vector $\mathbf{g}^m$ and the host vector $\mathbf{h}^m$ into the attention scoring function, which is a neural network composed of a single hidden layer and a softmax layer. The output of this function is a host score vector, where the value of each dimension reflects the degree of a host feature derived from the whole guest features. The degree reaches the highest at 1 and the lowest at 0. It is formally defined as,

$$\mathbf{s}_h^m = \text{softmax}(\mathbf{W}_h^m \cdot [\mathbf{h}^m, \mathbf{g}^m]), \quad (4)$$

where $\mathbf{W}_h^m \in \mathbb{R}^{D_h * D}$ and $\mathbf{s}_h^m \in \mathbb{R}^{D_h}$ denote the learnable weight matrix and relation score vector corresponding to each dimension of the host vector, respectively; the D_h denotes the dimension of the host vector, and D is the dimension of the overall vector. For simplicity, we omit the bias terms.

For the guest modality, we analogously scored the feature dimensions to measure the degree of a guest feature derived from the host features, defined as follows,

$$\mathbf{s}_g^m = \text{softmax}(\mathbf{W}_g^m \cdot [\mathbf{g}^m, \mathbf{h}^m]), \tag{5}$$

where $\mathbf{W}_g^m \in \mathbb{R}^{D_g \times D}$, the D_g and $\mathbf{s}_g^m \in R^{\mathbb{R}_g}$ denote the trainable weight matrix, the dimension of guest vector and the relation score vector corresponding to each dimension of the guest vector, respectively.

3.2.2 Consistency and Complementary Components

Having established the attentive relation scores, we can easily locate the consistent and complementary features from each modality. Towards this end, we set a trainable threshold denoted as ξ_o^m, in which we use $o \in \mathbb{O} = \{h, g\}$ as the host and guest indicator. This threshold divides the relation score vector into two parts: consistent vector and complementary vector, namely $\boldsymbol{\gamma}_o^m$ and $\boldsymbol{\delta}_o^m$. The element in the consistent vectors is defined as follows,

$$\boldsymbol{\gamma}_o^m[i] = \begin{cases} \mathbf{s}_o^m[i], & \text{if } \mathbf{s}_o^m[i] \geq \xi_o^m; \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where $\boldsymbol{\gamma}_o^m[i]$ is the value of i-th the dimension in the consistent weight vector $\boldsymbol{\gamma}_o^m$, indicating the degree of the consistency. For the complementary weight vector $\boldsymbol{\delta}_o^m$, we formulated its element as,

$$\boldsymbol{\delta}_o^m[i] = \begin{cases} 1 - \mathbf{s}_o^m[i], & \text{if } \mathbf{s}_o^m[i] < \xi_o^m; \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where $\boldsymbol{\delta}_o^m[i]$ is the value of the i-th dimension in the complementary weight vector $\boldsymbol{\delta}_o^m$, reflecting the degree of the complementary relation.

Particular, since the original functions are not continuous, we introduced a sigmoid function to make them differentiable, as follow,

$$\begin{cases} \boldsymbol{\gamma}_o^m[i] = \frac{\mathbf{s}_o^m[i]}{1+e^{-w*(\mathbf{s}_o^m[i]-\xi_o^m)}}, \\ \boldsymbol{\delta}_o^m[i] = 1 - \boldsymbol{\gamma}_o^m[i], \end{cases} \tag{8}$$

where w denotes a scalar weighting the difference between $\mathbf{s}_o^m[i]$ and ξ_o^m to make the output $\boldsymbol{\gamma}_o^m[i]$ as close as possible to 0 or $\mathbf{s}_o^m[i]$. Through experiments, the best results are obtained with a weight of 50.

After that, we gained four correlation weight vectors from each host-guest pair, namely $\boldsymbol{\delta}_h^m$, $\boldsymbol{\delta}_g^m$, $\boldsymbol{\gamma}_h^m$, and $\boldsymbol{\gamma}_g^m$. Based on these weight vectors, we separated the consistent features and the complementary features from the mixed information, which are the element-wise products of the original feature vector and each weight vector, as,

$$\begin{cases} \boldsymbol{\alpha}_h^m = \mathbf{h}^m \otimes \boldsymbol{\delta}_h^m, \\ \boldsymbol{\alpha}_g^m = \mathbf{g}^m \otimes \boldsymbol{\delta}_g^m, \\ \boldsymbol{\beta}_h^m = \mathbf{h}^m \otimes \boldsymbol{\gamma}_h^m, \\ \boldsymbol{\beta}_g^m = \mathbf{g}^m \otimes \boldsymbol{\gamma}_g^m, \end{cases} \tag{9}$$

where two complementary vectors and two consistent vectors of host and guest are denoted as $\boldsymbol{\alpha}_h^m$, $\boldsymbol{\alpha}_g^m$, $\boldsymbol{\beta}_h^m$ and $\boldsymbol{\beta}_g^m$, respectively.

With the separated consistent and complementary components, we can reconstruct the representations with better expressiveness. We employed different strategies on distinct components. To adequately exploit the correlations between the consistent component pairs, we concatenated these vectors and feed them into a neural network to learn an enhanced consistent vector,

$$\widetilde{\boldsymbol{\beta}}^m = \varphi(\mathbf{W}_\beta^m \cdot [\boldsymbol{\beta}_h^m, \boldsymbol{\beta}_g^m]), \tag{10}$$

where $\mathbf{W}_\beta^m$, $\varphi(\cdot)$, and $\widetilde{\boldsymbol{\beta}}^m$ denote the trainable weight matrix, activation function, and the enhanced consistent vector in the modality m, respectively.

To supplement the exclusive information from other modalities, we integrated the enhanced consistent components and the complementary components to generate a feature vector with powerful expressiveness as,

$$\hat{\mathbf{x}}^m = [\boldsymbol{\alpha}_h^m, \widetilde{\boldsymbol{\beta}}^m, \boldsymbol{\alpha}_g^m]. \tag{11}$$

Meanwhile, to guarantee the consistency, the diversity of the consistent component pairs should be minimized. However, the dimension of each vector is different, and the number of consistent features is dynamic. We hence failed to capture the diversity of these features directly. Towards this end, we proposed to compute the probability distributions of venue categories represented by consistent vectors, and further leverage the Kullback-Leibler divergence (KL divergence) [35] to encourage them to be close.

Particularly, the probability distribution over categories is defined as follows,

$$\mathbf{p}_o^m = \text{softmax}(\mathbf{U}_o^m \cdot \boldsymbol{\beta}_o^m), \tag{12}$$

where $\mathbf{U}_o^m \in \mathbb{R}^{K * D_o}$ and $\mathbf{p}_o^m \in \mathbb{R}^K$ denote the learnable weight matrix and the probability distribution of the venue categories represented by the consistent vector $\boldsymbol{\beta}_o^m$, respectively.

Following that, we computed the KL divergence between the two probability distributions $\mathbf{p}_h^m$ and $\mathbf{p}_g^m$, formally as,

$$\mathcal{L}_1^m = \sum_{x \in \mathcal{X}} (\mathbf{p}_g^m \log \mathbf{p}_h^m - \mathbf{p}_h^m \log \mathbf{p}_g^m), \tag{13}$$

where $\mathbf{p}_h^m$ and $\mathbf{p}_g^m$ both denote the probability distribution of the venue categories. Based upon this, we calculated the sum of the KL divergences from all modalities as,

$$\mathcal{L}_1 = \sum_{m \in \mathcal{M}} \mathcal{L}_1^m. \tag{14}$$

3.3 Attention Networks

Given the augmented representations above, a straightforward way to estimate the venue category is to adopt a classifier. However, we argue that the rich information within the augmented representations is redundant for the prediction task, and hence the simple classifier can hardly select the discriminative features. Several efforts have been paid to achieve a discriminative representation from massive features, like Principal

Component Analysis [36] and sparse representation [37]. These approaches, however, have many hyper-parameters to tune. More principle components, for instance, can lead to the suboptimal performance.

With the advance of the attention mechanism, we employed an attention network to evaluate the attention scores for each feature towards different venue categories. These scores can measure the relevance and significance of the features to the venue category. In addition, continuous attention scores make the feature selection flexible. Thereafter, we obtained the scored features and leveraged them to learn a discriminative representation to estimate the venue category in each modality.

3.3.1 Attention Score

Given a feature vector, we assigned an attention score to each feature according to the venue category and yielded the scored feature to learn a discriminative representation.

Instead of computing the importance of each feature to categories, we constructed a trainable memory matrix to store the attention score of them. The matrix is denoted as $\mathbf{\Omega}^m \in \mathbb{R}^{D_m \times K}$ in the modality m and the entry in row i and column j represents the importance of $i-th$ feature towards $j-th$ venue category. For each category, the scored feature vector is obtained by calculating the element-wise product of the feature vector and the corresponding row vector in matrix $\mathbf{\Omega}^m$. It is formulated as,

$$\boldsymbol{\psi}_j^m = \boldsymbol{\omega}_j^m \otimes \hat{\mathbf{x}}^m, \tag{15}$$

where $\hat{\mathbf{x}}^m \in \mathbb{R}^D$ is the augmented vector in the modality m; $\boldsymbol{\omega}_j^m \in \mathbb{R}^D$ denotes the feature attention scores of the venue category j and $\boldsymbol{\psi}_j^m \in \mathbb{R}^D$ denotes the scored feature vector towards venue category j.

To yield the discriminative representation, we feed the scored feature vector into a fully connected layer as follows,

$$\boldsymbol{\theta}_j^m = \phi(\mathbf{W}^m \cdot \boldsymbol{\psi}_j^m), \tag{16}$$

where $\mathbf{W}^m$, $\phi(\cdot)$, and $\boldsymbol{\theta}_j^m$ denote the trainable weight matrix, the activation function and the discriminative representation of j-th venue category in the modality m, respectively.

3.3.2 Multimodal Estimation

After obtaining the discriminative representations, we passed them into a fully connected softmax layer. It computes the probability distributions over the venue category labels in each modality, mathematically stated as,

$$p(\hat{y}_k^m|\boldsymbol{\theta}_k^m) = \frac{\exp(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\theta}_k^m)}{\sum_{k'=1}^{K}\exp(\mathbf{z}_{k'}^{\mathrm{T}}\boldsymbol{\theta}_{k'}^m)}, \tag{17}$$

where $\mathbf{z}_k$ is a trainable weight vector of the k-th venue category, and $\boldsymbol{\theta}_k^m$ can be viewed as the discriminative representation of k-th venue category in the modality m. Thereafter, we gained the probabilistic label vector $\hat{\mathbf{y}}^m = [\hat{y}_1^m, \cdots, \hat{y}_K^m]$ over K venue categories.

For multiple modalities, the probabilistic label vector over three modalities are fused, defined as follows,

$$\hat{\mathbf{y}} = \sum_{m\in\mathcal{M}}(\hat{\mathbf{y}}^m). \tag{18}$$

Following that, we adopted a function to minimize the loss between the estimated label vector and its target values, as

$$\mathcal{L}_2 = -\sum_{x\in\mathcal{X}}\sum_{k=1}^{K}\mathbf{y}_k\log(\hat{\mathbf{y}}_k). \tag{19}$$

Ultimately, this function and the KL divergence of consistent representation pairs are combined as the objective function of our proposed method, as follows,

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_1 + \mathcal{L}_2 \\ &= \sum_{m\in\mathcal{M}}\sum_{x\in\mathcal{X}}(\mathbf{p}_g^m\log\mathbf{p}_h^m - \mathbf{p}_h^m\log\mathbf{p}_g^m) \\ &- \sum_{x\in\mathcal{X}}\sum_{k=1}^{K}\mathbf{y}_k\log(\sum_{m\in\mathcal{M}}(\frac{\exp(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\theta}_k^m)}{\sum_{k'=1}^{K}\exp(\mathbf{z}_{k'}^{\mathrm{T}}\boldsymbol{\theta}_{k'}^m)})).\end{aligned} \tag{20}$$

3.4 Training

We adopted the stochastic gradient descent (SGD) to train our model in a mini-batch mode and updated the model parameters using back propagation until convergence. In particular, a training instance x is iteratively selected and used to optimize the parameters towards the direction of its negative gradient:

$$\epsilon = \epsilon - \eta\frac{\partial\mathcal{L}}{\partial\epsilon}, \tag{21}$$

where ϵ and η denote the trainable parameter and learning rate that control the step size of gradient descent, respectively. As the model is a multi-layered neural network model, the gradient of each model parameter can be derived using the chain rule. In our model, we initialed the attention matrix and stored it in the memory during the training phase. The attention vector of each category can be used directly during the testing instead of calculating with input vectors.

While deep neural networks are powerful in representation learning, they easily lead to the overfitting on the limited training data. To alleviate this issue, we employed dropout [38] to improve the regularization of our model. The idea is to randomly drop part of neurons during training and update only part of the model parameters that contributes to the prediction. It is important to note that in the testing phase, dropout must be disabled and the whole model is used for estimating. Therefore, dropout can be treated as an approximate model averaging. Moreover, we also conducted dropout on each hidden layer of our model to prevent the overfitting.

4 Data Preparation

In this section, we detail the data preparation, namely dataset collection, feature extraction and missing data completion.

4.1 Data Collection

We crawled the micro-videos from Vine through its public API[6]. In particular, we first manually chose a small set of active users as the seeds. We expanded the user sets through incrementally gathering the seed users' followers.

6. https://github.com/davoclavo/vinepy.

TABLE 1: Features extracted from three modalities.

Modality	Extracted Features
Textual	100-D paragraph vector
Acoustic	200-D DAE vector
Visual	2048-D CNN vector

With the user set, we then crawled the published videos, descriptions and venue information if available from the collected users. We picked out about 24,000 micro-videos containing Foursquare check-in information from the overall crawled micro-video set. After removing the duplicated venue IDs, we further expanded our video set by crawling all videos in each venue ID with the help of API. Thereafter, we obtained a dataset of 276,264 videos distributed in 442 Foursquare venue categories and served the corresponding ID as the ground truth. Furthermore, we observed that the category distribution is heavily unbalanced. Thereinto, several categories contain a limit number of micro-videos to train a robust classifier. We hence removed the leaf categories with less than 50 micro-videos. At last, we achieved 270,145 micro-videos distributed in 188 Foursquare venue categories.

4.2 Feature Extraction

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

4.2.1 Visual Features

The information conveyed by the visual modality provides intuitive and efficient signals to estimate the venue category. For instance, if we observe "tables", "desks", "coffee cups" and "employees" from a micro-video, we can easily and accurately predict that the micro-video is recorded in a coffee shop. This prompts us to extract rich features from the visual modality to represent the micro-videos. Deep CNN has been proved as an excellent model to represent the images [39]. We applied the ResNet [40] model to extract the visual features through the publicly available Caffe [41]. Before extracting features, we first selected the keyframes from each micro-video by using OpenCV[7], and then employed the ResNet to get features from each frame. Following that, we took the mean pooling strategy over all keyframes of the micro-video and generated a single 2,048-dimensional vector for each micro-video.

4.2.2 Acoustic Features

The audio clips embedded in the micro-videos contain useful cues or hints about the locations. For example, within the coffee shops, audio clips capture that "the employees are answering customers' questions, and welcoming them to the shop". For the situation where the visual features contain little information to predict the location, the acoustic modality takes the upper hand to supplement the exclusive information. To leverage the acoustic features, the audio tracks were separated from each micro-video by FFmpeg[8]. Then, we transformed the tracks into a uniform format: 22,050Hz, 16 bits, mono-channel with pulse-code modulation signals and performed a spectrogram with a 46ms window and 50% overlap via librosa[9]. Thereafter, we adopted theano to extract the acoustic features with a stack Denoising AutoEncoder (DAE) [42]. The DAE model was pre-trained on an external set of 120,000 micro-videos crawled from Vine containing three hidden layers, with 500, 400, and 300 neurons on each layer. We ultimately obtained 200-dimensional acoustic features for each micro-video.

4.2.3 Textual Features

The textual descriptions of micro-videos, including user-generated text and hashtags, can provide strong cues for micro-video venue estimation. For instance, the hashtag from the "Vining the #beach while tanning the thighs on a glorious Anzac Day" clearly indicates that the venue category is "beach". However, only around 27.7% of our collected micro-videos have hashtags, and the total number of hashtags is 253,474. Therefore, the traditional approaches such as topic-level features [43] and n-grams [44] may be unsuitable. Instead, we utilized the Paragraph Vector method [45] proven to be effective to alleviate the semantic problems of word sparseness [46]. To accomplish this, we applied Sentence2Vector tool[10] to extract a textual set of 100-d features for each micro-video description.

We ultimately obtained a feature set from three modalities as summarized in Table 1.

4.3 Missing Data Completion

Different from the visual modality, we found some micro-videos lack acoustic and textual information. Statistically, 169 and 24,707 micro-videos miss the acoustic and textual modality, respectively. Information missing deteriorates the performance of most machine learning methods [47], including the models of venue category estimation. To alleviate the problem, we exploited the low-rank matrix factorization method [48], which is a commonly used technique in data compression, recommendation system, and matrix completion, to complete the missing data. In particular, we concatenated the multimodal features extracted from micro-videos and formed all of the feature vectors as an original feature matrix. And then, we factorized this matrix into two low dimension matrices with 100 latent features. Following this, the missing features are inferred through minimizing the empirical error between the product of these two matrices and the original matrix, and over-fitting is avoided through a regularized model.

5 EXPERIMENT

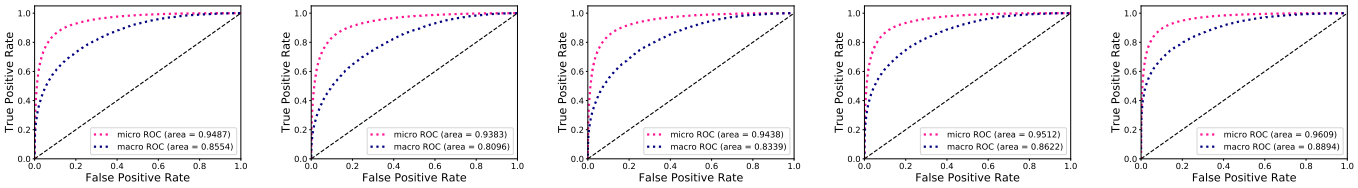
In this section, we validate our proposed model and its components over micro-video understanding.

5.1 Experiment Settings

5.1.1 Metric

In this work, Macro-F1 and Micro-F1[11] are adapted to measure the performance of the micro-video venue category

7. http://opencv.org/.
8. https://www.ffmpeg.org/.
9. https://github.com/bmcfee/librosa.
10. https://github.com/klb3713/sentence2vec.
11. The F1 scores will be the harmonic mean of precision and recall. In the multi-classification, the macro one computes the metric independently for each class and then take the average, whereas the micro one aggregates the contributions of all classes to compute the average metric.



(a) ROC_EarlyAtt (b) ROC_LateAtt (c) ROC_TRUMANN (d) ROC_DARE (e) ROC_NMCL

Fig. 4: ROC curves and AUC scores of methods.

TABLE 2: Performance comparison between our model and the baselines (p-value*: p-value[12]over micro-F1.).

	Micro-F1	Macro-F1	p-value*
Early Fusion	11.39±0.01%	0.12±0.01%	1.31e-8
Late Fusion	12.57±0.23%	0.20±0.04%	4.29e-9
Early+Att	31.24±0.37%	14.03±0.19%	2.48e-8
Late+Att	30.00±0.31%	13.71±0.51%	1.52e-7
TRUMANN	27.38±0.21%	10.87±0.05%	8.71e-8
DARE	34.40±0.32%	20.21±0.35%	5.94e-7
NMCL	**40.04±0.37%**	**26.78±0.42%**	-

estimation. Both of them reach the best score at 1 and the worst at 0. The averaging macro-F1 gives equal weight to each class-label; whereas the averaging micro-F1 gives equal weight to all instances. Besides, we provide the Receiver Operating Curves (ROC) of our method and four baselines and use Areas Under Curve (AUC) scores to evaluate the results. AUC is the area under the ROC curve, which is created by plotting the true positive rate against the false positive rate. We divided our dataset into three chunks: $132,370$ for training, $56,731$ for validation, and $81,044$ for testing. The training set is used to adjust the parameters, while the validation one provides an unbiased evaluation of a model fit on the training dataset and tunes the model's hyperparameters. The testing one is used only to report the final solution to confirm the actual predictive power of our model with the optimal parameter settings.

5.1.2 Baselines

We compare the performance of our proposed model with several state-of-the-art baselines:

- **Early Fusion** [49]: For any given micro-video, we concatenated multimodal features into one vector, and then learned a model consisting of three fully connected layers to estimate the venue category over the concatenation vectors.
- **Late Fusion** [49]: To calculate the categories distribution, we devised the classifiers which are respectively implemented by a neural network with one, two and three hidden layers for the textual, acoustic and visual modality. And we fused these distributions to yield a final prediction venue category.
- **Early+Att** [50]: This baseline is the combination of the early fusion and attention model. In particular, the attention model gives different attention weights to all features integrated from multiple modalities according to different venue categories. Here, the attention weights are calculated by a scoring function of the concatenated features and venue category. After that, a neural network is devised with three fully connected layers to categorize the unsee micro-videos over the attended feature vectors.
- **Late+Att** [50]: For various venue categories, features in each modality have varying contributions to the final prediction. Therefore, this baseline introduces the attention mechanism into classifiers of each modality to obtain the venue category representations and then fuses these representations to yield a final venue category.
- **TRUMANN** [1]: This is a tree-guided multi-task multi-modal learning method, which is the first one towards the micro-video venue category estimation. This model is able to jointly learn a common space from multiple modalities and leverage the predefined Foursquare hierarchical structure to regularize the relatedness among venue categories.
- **DARE** [6]: This work is a deep transfer model which harnesses the external knowledge to enhance the acoustic modality and regularizes the representation learning of micro-videos of the same venue category to alleviate the sparsity problem of unpopular categories.

5.1.3 Parameter Settings

We implemented our model with the help of Tensorflow[13]. Particularly, we applied the Xavier approach to initialize the model parameters, which has been proved as an excellent initialization method for the neural network models. The mini-batch size and learning rate are respectively searched in $\{128, 256, 512\}$ and $\{0.001, 0.005, 0.01, 0.05, 0.1\}$. The optimizer is set as Adaptive Moment Estimation (Adam) [51]. Moreover, we empirically set the size of each hidden layer as 256 and the activation function as ReLU. Without special mention, all the models employ one hidden layer and one prediction layer. For a fair comparison, we initialized other competitors with an analogous procedure. The average results over five-round predictions are illustrated in the testing set.

5.2 Performance Comparison

The comparative results are shown in Table 2 and Figure 4. From this table, we have the following observations:

1) In terms of the Micro-F1, **Early Fusion** and **Late Fusion** achieve the worst performance, since these standard fusion approaches rarely exploit the correlations between different modalities.
2) Integrating the attention model to the standard fusion is able to improve the performance obviously. Taking the advantages of the attention mechanism, **Early+Att**

12. In statistical hypothesis testing, the probability value (p-value) is the probability for a given statistical model that, when the null hypothesis is true, the statistical summary would be greater than or equal to the actual observed results.

13. https://www.tensorflow.org.

TABLE 3: Representativeness of different modalities (p-value*: p-value over micro-F1.).

	Micro-F1	Macro-F1
Textual	13.40±0.14%	2.23±0.1
Acoustic	14.21±0.12%	3.40±0.02
Visual	28.16±0.23%	11.22±0.41
Acoustic+Textual	20.57±0.41%	7.08±0.09
Visual+Textual	38.45±0.34%	23.83±0.34
Visual+Acoustic	37.07±0.35%	23.34±0.11
All	**40.04±0.37%**	**26.78±0.42**

TABLE 4: Performance of each enhanced modality in different modality pairs. (V-MicroF1, A-MicroF1, and T-MicroF1 denote Micro-F1 score on the visual, acoustic and textual modality, respectively.)

	V-MicroF1	A-MicroF1	T-MicroF1
Acoustic+Textual	-	20.12±0.15%	20.13±0.14%
Visual+Textual	**37.46±0.26%**	-	**35.75±0.36%**
Visual+Acoustic	35.09±0.15%	34.8±0.16%	-
All	36.07±0.28%	**35.27±0.17%**	33.73±0.51%

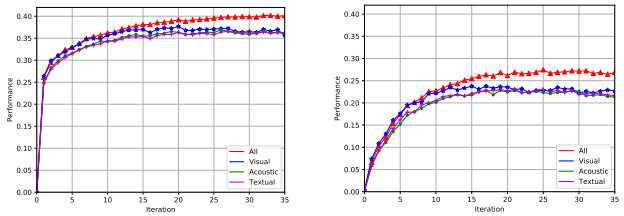
and **Late+Att** can dynamically select the discriminative features, which are tailored to the prediction task. This verifies the feasibility of revising the weight of each feature.

3) When performing the estimation task, **TRUMANN** outperforms **Early Fusion** and **Late Fusion**. It is reasonable since it considers the hierarchical structure of venue categories and employs the multi-task learning, whereas **Early+Att** and **Late+Att** outperform the **TRUMANN**. It again admits the effectiveness of assigning the attentive weights to the features.
4) The performance of **DARE** exceeds the others except ours, indicating that **DARE** benefits from the enhanced audio modality via an external dataset and alleviates the sparse problem of unpopular categories by regularizing the similarity among the categories.
5) Our proposed model achieves the best w.r.t. micro-F1 and macro-F1. By exhibiting the consistency and complementary of features, our model achieves a better expressiveness compared to all baselines. While **DARE** and **TRUMANN** treat all features linearly independently and equally, our model can capture and leverage the correlation between different modalities, as well as employ the attention networks to identify the tailored attention of each feature. We further conducted a pair-wise significant test to verify that all improvements are statistically significant with p-value < 0.05.
6) As shown in Figure 4, **NMCL** achieves an AUC score of more than 96% and is superior to the baselines, further demonstrating the effectiveness of our proposed method. Despite **DARE** yields the AUC score of 95.12% and ranks the second-best performance among all the methods, our proposed method outperforms it by a gain of about 1%. Besides, in terms of the macro-average ROC curve, **NMCL** gets an AUC score of about 89%, which increases by $3\% \sim 10\%$ than the baselines.

5.3 Study of NMCL Model

5.3.1 Representativeness of Modalities

In this section, we studied the effectiveness of combining different modalities. Table 3 and Table 4 show the performance of different modality pairs and each enhanced modality with our proposed model, respectively. In addition, we plotted the Macro-F1 and Micro-F1 $w.r.t.$ the number of iterations in Figure 5 to illustrate the convergence and efficiency of our model in each modality. From these tables and the figure, we observe that:



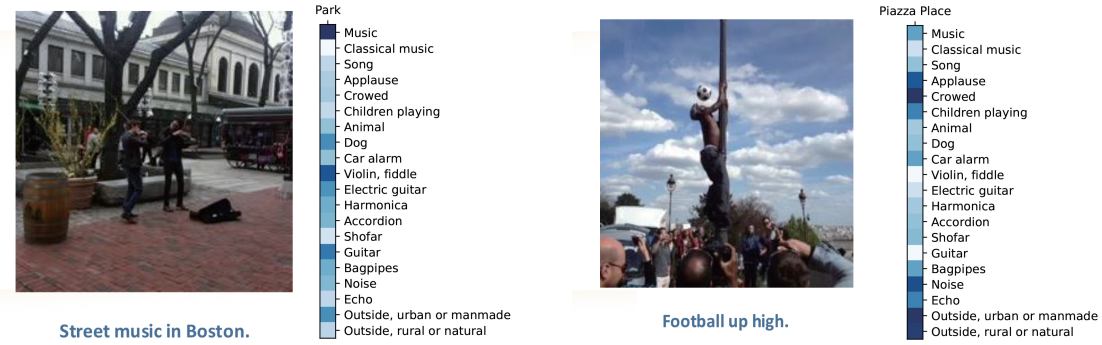
(a) Micro-F1 (b) Macro-F1

Fig. 5: Convergence and effective study of the NMCL.

1) On the first row in Table 3, solely considering the visual modality achieves the best performance compared to the other mono-modal estimation methods. This is consistent with the finding in [1, 6, 52], verifying the rich geographic information conveyed by the visual features. In addition, the CNN features are capable of capturing the prominent visual characteristics of the venue categories.
2) The acoustic modality and textual modality perform similarly in estimating the venue categories, which are listed on the second row and the third row in Table 3, respectively. Only using one modality, however, is insufficient to estimate the categories for most micro-videos, since the textual and acoustic information is noisy, sparse, and even irrelevant to the venue categories.
3) The more modalities we incorporate, the better performance we can achieve, as the display on last three rows in Table 3 and in Table 4. This implies that the information of one modality is insufficient and multiple modalities are complementary to each other rather than conflicting. This is a consensus to the old saying "two heads are better than one".
4) Table 4 shows that the performance of each modality enhanced by our proposed approach is improved obviously, especially when the acoustic and textual modalities are combined with the visual modality. This improvement validates that each modality can be enforced by the other modalities in our model.
5) Comparing each row in Table 4 to the first three rows in Table 3, the performance of each modality, which is enhanced by the other two modalities with our model, is better than that of the early fusion integrating the attention model. It indicates that our model can capture the correlations between different modalities.
6) Jointly analyzing the curves in Figure 5, we find that utilizing our proposed cooperative learning to seamlessly integrate multiple modalities can boost the performance effectively. This demonstrates the rationality of our model. And the performance tends to be stable at around 30 iterations. This signals the convergence property of our model and also indicates its efficiency.

5.3.2 Study of Components

In this section, we list several variants based on our proposed cooperative net. These methods group the features of each modality into consistent and complementary parts,



(a) Park Example (b) Piazza Place Example

Fig. 6: Visualization of the correlation scores between the same acoustic concept-level features and different visual and textual features.

TABLE 5: Performance of variants (p-value*: p-value over micro-F1.).

	Micro-F1	Macro-F1
Variant-I	39.17±0.27%	25.05±0.28
Variant-II	39.01±0.37%	23.70±0.19
Variant-III	38.11±0.40%	22.78±0.10
Variant-IV	38.48±0.33%	24.49±0.18
NMCL	**40.04±0.37%**	**26.78±0.42**

and then we adopted different fusing strategies to leverage the consistent and complementary features, including:

- **Variant-I**: In this model, Eq. 10 is removed. In other words, we integrated the guest complementary information into the host modality without enhancing the consistent parts, while the guest consistent part is retained to calculate the KL-diversity for keeping the consistency.
- **Variant-II**: This variant discards Eq. 11 and merely harnesses the consistent vector pairs to learn an enhanced feature vector for each modality and categorize the venue with these enhanced feature vectors.
- **Variant-III**: After obtaining the consistent and complementary features from each host and modality pair, Eq. 10 is replaced, and a new enhanced consistent vector is learned by integrating all host consistent vectors. Afterthat, the category is estimated by fusing the predictions of the newly enhanced consistent vector and each complementary vector.
- **Variant-IV**: In this variant model, we respectively concatenated all complementary parts and all consistent parts, instead of Eq. 10 and Eq. 11. Finally, we estimated the venue category of the two concatenated parts and fused them to gain the result.

From Table 5, we have the following observations:

1) In terms of Macro-F1, Variant-I and Variant-IV outperform Variant-II and Variant-III, respectively. This may because combining the complementary information can involve more information, strengthening the expressiveness of the representations.
2) The accuracy of the first two variants is comparatively higher than the other variants. This benefits from capturing the correlation between the host and the guest features which is ignored by the Variant-III and the Variant-IV.
3) Our proposed method outperforms its all variants, justifying the rationality and effectiveness of cooperative learning. Different from several variants, the original one considers the consistency between each host and guest modality pairs and supplements the exclusive signals from the guest modalities.
4) We observe that Variant-I, which discards one of the consistent parts, does not cause a significant reduction in accuracy. It shows that the information contained in the two consistent vectors is almost the same, and it also demonstrates that our model can correctly distinguish and capture the consistent features.
5) Comparing the method with Variant-II, we observe that the improvement in terms of Micro-F1 is not significant. For further analysis, we believe that the main reason is that the concepts contained in micro-videos are sparse. Moreover, the information contained in any single modality is almost covered by the other two modalities. In other words, the complementary parts contain little external information. Therefore, the removal of the complementary parts barely affects the performance.

5.4 Visualization

Apart from achieving more accurate prediction, the key advantage of NMCL over other methods is that it exhibits the consistent and complementary features. Towards this end, we show examples drawn from our model to visualize two representation components.

Since the acoustic modality is the hardest one to be visualized among the multiple modalities, we utilized the concept-level features to present the acoustic one. To extract the concept from fine-grained acoustic features, we leveraged an external dataset namely AudioSet, which is a large-scale dataset released by Google[14].

The AudioSet consists of an expanding ontology of 632 audio event classes and a collection of 2,084,320 human labelled 10-second sound clips drawn from YouTube[15] videos. The ontology is specified as a hierarchical graph of event categories, covering a wide range of human and animal sounds, musical instruments and genres, and everyday common sounds from the environment, like "Speech", "Laughter" and "Guitar".

14. The external audio dataset was just used for the visualization. https://research.google.com/audioset/.
15. https://youtube.com.

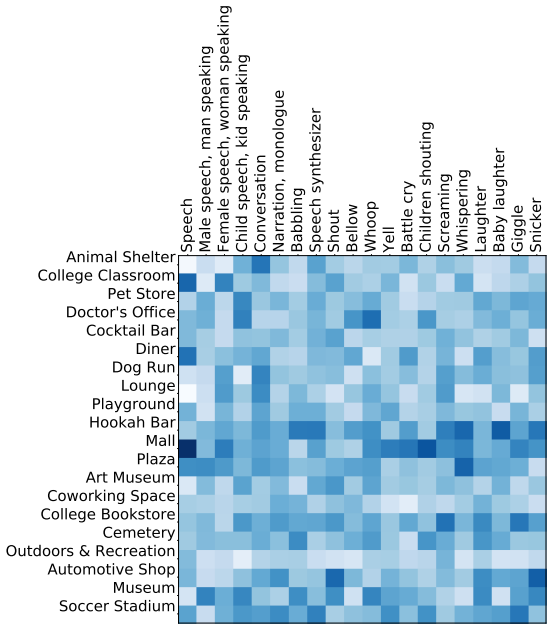


Fig. 7: Visualization of the attention scores of acoustic concept-level features and venue category pairs.

To estimate the concepts in the audio, we employed a VGG-like model [53] and trained it over the AudioSet. According to the input format of the CNN model, we regenerated the acoustic features of the micro-videos. The extracted audios are divided into non-overlapping 960 ms frames, and then the spectrogram transformed from the frames are integrated into 64 Mel-spaced frequency bins. Finally, we took the mean pooling strategy over all the frames of the micro-video to yield a new acoustic feature vector.

With the new acoustic conceptual features, we conducted experiments to shed some light on the correlation between the acoustic modality and the other modalities. In addition, we visualized the attention score matrix between the acoustic concepts and venue categories to validate our proposed model intuitively.

- To visualize the consistent and complementary parts among modalities, we selected exemplary demonstrations of two micro-videos categorized as "Park" and "Piazza place", as shown in Figure 6(a) and Figure 6(b). For these demonstrations, we treated the acoustic modality as the host part, the visual and the textual modalities as the guests. And we showed a heat map to illustrate the correlation between the host and guest feature pairs, where the darker color indicates that the host feature is consistent with the guest modalities and vice versa. From the Figure 6(a), we observe that several acoustic concepts are consistent with the visual and textual modalities, such as "Music" and "Violin", and some are exclusive ones hardly revealed from the other modalities, such as "Applause", "Noise" and "Car alarm". In contrast, given the Figure 6(b), we find that the correlation score distribution is totally different. The concepts, such as "Applause", "Crowed" and "Noisy", can be represented by the guest features, and the "Music" and the "Violin" are barely captured in the other modalities. However, these "lighter-colored" features provide the exclusive and discriminative information to predict the venue category. In our proposed model, we can explicitly capture the exclusive information as a supplement, rather than omitting it during the learning produce. These observations verify the assumption that the information from different modalities is complementary to each other and demonstrate that our proposed model can explicitly separate the consistent information from the complementary one.
- To save the space, we performed the part of the attention matrix via a heat map, where lighter color indicates weak attention and vice versa, as shown in Figure 7. We can see that every selected venue category has various relations to each acoustic concept. For instance, the micro-videos with the venue of "Mall" have strong correlations with "Speech" and "Children shouting"; the correlation with "Babbling" is loose. In addition, for the venue of "Pet store", the colors representing "Kid speaking" and "Whoop" are dark, and the color representing "Battle cry" is lighter. These observations agree with our common sense and demonstrate that the attention score can select the discriminative features towards the venue category.

6 Conclusion and Future Work

In this paper, we shed light on characterizing and modeling the correlations between modalities, especially the consistent and complementary relations. The consistent part is to strengthen the confidence and the complementary one is able to supplement much exclusive information. We argue that explicitly parsing these two kinds of correlations and treating them separately within a unified model can boost the representation discrimination for multimodal samples. Towards this goal, we devised a cooperative learning model. In this model, we introduced a novel relation-aware attention mechanism to split the consistent information from the complementary one. Following that, we integrated the consistent information to learn an enhanced consistent vector and supplemented the complementary information to enrich this enhanced vector. To learn a discriminative representation from this richer information, we devised an attention network to score the features. To validate the proposed model, we applied it to the application scenario: venue category estimation of micro-videos. And the results outperformed several state-of-the-art baselines, verifying the efficiency of the method. As a side contribution, we have released our codes, parameter settings, and the involved baselines to facilitate other researchers.

In the future, we expect to capture more complex correlations among multiple modalities, such as conflict. In addition, we plan to leverage the extracted correlations to various applications, such as image caption and frame recommendation.

7 Acknowledgments

This work is supported by the National Basic Research Program of China (973 Program), No.: 2015CB352502; National Natural Science Foundation of China, No.: 61772310, No.:61702300, and No.:61702302; the Project of Thousand Youth Talents 2016; and the Tencent AI Lab Rhino-Bird Joint Research Program, No.:JR201805.

REFERENCES

[1] J. Zhang, L. Nie, X. Wang, X. He, X. Huang, and T. S. Chua, "Shorter-is-better: Venue category estimation from micro-video," in *ACM MM*, 2016, pp. 1415–1424.

[2] C. S. S. Lazebnik and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006, pp. 2169–2178.

[3] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," pp. 73–86, 2012.

[4] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *NIPS*, 2014, pp. 487–495.

[5] M. Liu, L. Nie, M. Wang, and B. Chen, "Towards micro-video understanding by joint sequential-sparse modeling," in *ACM MM*, 2017, pp. 970–978.

[6] L. Nie, X. Wang, J. Zhang, X. He, H. Zhang, R. Hong, and Q. Tian, "Enhancing micro-video understanding by harnessing external sounds," in *ACM MM*, 2017, pp. 1192–1200.

[7] V. Radu, C. Tong, S. Bhattacharya, N. D. Lane, C. Mascolo, M. K. Marina, and F. Kawsar, "Multimodal deep learning for activity and context recognition," *ACM IMWUT*, vol. 1, no. 4, p. 157, 2018.

[8] D. Wang, K. Mao, and G.-W. Ng, "Convolutional neural networks and multimodal fusion for text aided image classification," in *FUSION*. IEEE, 2017, pp. 1–7.

[9] X. Yang, P. Molchanov, and J. Kautz, "Multilayer and multimodal fusion of deep neural networks for video classification," in *ACM MM*. ACM, 2016, pp. 978–987.

[10] B.-K. Bao, C. Xu, W. Min, and M. S. Hossain, "Cross-platform emerging topic detection and elaboration from multimedia streams," *TOMM*, vol. 11, no. 4, p. 54, 2015.

[11] W. Min, B.-K. Bao, S. Mei, Y. Zhu, Y. Rui, and S. Jiang, "You are what you eat: Exploring rich recipe information for cross-region food analysis," *TMM*, vol. 20, no. 4, pp. 950–964, 2018.

[12] W. Min, B.-K. Bao, C. Xu, M. S. Hossain *et al.*, "Cross-platform multi-modal topic modeling for personalized inter-platform recommendation." *TMM*, vol. 17, no. 10, pp. 1787–1801, 2015.

[13] S. K. D'Mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys*, vol. 47, no. 3, pp. 1–36, 2015.

[14] A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *ACM MM*, 2005, pp. 399–402.

[15] E. Shutova, D. Kiela, and J. Maillard, "Black holes and white rabbits: Metaphor identification with visual features," in *NAACL*, 2016, pp. 160–170.

[16] E. Morvant, A. Habrard, S. Ayache, and phane, *Majority Vote of Diverse Classifiers for Late Fusion*. Springer, 2014.

[17] G. A. Ramirez, T. Baltrušaitis, and L. P. Morency, "Modeling latent discriminative dynamic of multi-dimensional affective signals," in *ACLL*, 2011, pp. 396–406.

[18] C. M. Christoudias, R. Urtasun, A. Kapoorz, and T. Darrell, "Co-training with noisy perceptual observations," in *CVPR*, 2016, pp. 2844–2851.

[19] K. Nigam and R. Ghani, "Analyzing the effectiveness and applicability of co-training," in *CIKM*, 2000, pp. 86–93.

[20] Z. H. Zhou and M. Li, "Semi-supervised regression with co-training," in *IJCAI*, 2005, pp. 908–913.

[21] S. Yu, B. Krishnapuram, R. Rosales, and R. B. Rao, "Bayesian co-training," *JMLR*, vol. 12, no. 3, pp. 2649–2680, 2011.

[22] L. Duan, D. Xu, I. W. Tsang, and J. Luo, "Visual event recognition in videos by learning from web data." *TPAMI*, vol. 34, no. 9, pp. 1667–1680, 2012.

[23] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *JMLR*, vol. 5, no. 1, pp. 27–72, 2002.

[24] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *ICML*, 2004, pp. 6–13.

[25] Y. Ying and C. Campbell, "Generalization bounds for learning the kernel: Rademacher chaos complexity," vol. 75, no. 4, 2009, pp. 247–254.

[26] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *TPAMI*, vol. 38, no. 10, pp. 2010–2024, 2016.

[27] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[28] S. Akaho, "A kernel method for canonical correlation analysis," *IMPS*, vol. 40, no. 2, pp. 263–269, 2006.

[29] T. Diethe, D. R. Hardoon, and J. Shawe-Taylor, "Multiview Fisher discriminant analysis," in *NIPS*, 2008, pp. 1–8.

[30] D. Zhai, H. Chang, S. Shan, X. Chen, and W. Gao, "Multiview metric learning with global consistency and local smoothness," *TIST*, vol. 3, no. 3, pp. 1–22, 2012.

[31] S. Sano, T. Yamasaki, and K. Aizawa, "Degree of loop assessment in microvideo," in *IEEE ICIP*, 2015, pp. 5182–5186.

[32] P. X. Nguyen, G. Rogez, C. C. Fowlkes, and D. Ramanan, "The open world of micro-videos," *arXiv preprint arXiv:1510.03519*, 2016.

[33] 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 *ACM MM*, 2016, pp. 898–907.

[34] N. Quadrianto and C. H. Lampert, "Learning multi-view neighborhood preserving projections," in *ICML*, 2011, pp. 425–432.

[35] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[36] D. Wang, S. C. H. Hoi, Y. He, J. Zhu, T. Mei, and J. Luo, "Retrieval-based face annotation by weak label regularized local coordinate coding," *TPAMI*, vol. 36, no. 3, pp. 1–14, 2013.

[37] X. Zhou, M. Zhu, S. Leonardos, and K. Daniilidis, "Sparse representation for 3D shape estimation: A convex relaxation approach," *TPAMI*, vol. PP, no. 99, pp. 1–14, 2015.

[38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[39] S. Ren, R. Girshick, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *TPAMI*, vol. 39, no. 6, pp. 1137–1151, 2017.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2015, pp. 770–778.

[41] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM MM*, 2014, pp. 675–678.

[42] R. Memisevic, "Gradient-based learning of higher-order image features," in *ICCV*, 2011, pp. 1591–1598.

[43] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[44] C. Y. Suen, "n-gram statistics for natural language understanding and text processing," *TPAMI*, vol. 1, no. 2, pp. 164–172, 1979.

[45] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *NIPS*, vol. 26, pp. 3111–3119, 2013.

[46] D. Ganguly, D. Roy, M. Mitra, and G. J. F. Jones, "Word embedding based generalized language model for information retrieval," in *SIGIR*, 2015, pp. 795–798.

[47] H. Zhang, X. Shang, W. Yang, H. Xu, H. Luan, and T. S. Chua, "Online collaborative learning for open-vocabulary visual classifiers," in *CVPR*, 2016, pp. 2809–2817.

[48] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan, "Large-scale parallel collaborative filtering for the netflix prize," in *AAIM*, 2008, pp. 337–348.

[49] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011, pp. 689–696.

[50] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *ICCV*. IEEE, 2017, pp. 4203–4212.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *NIPS*, pp. 1–1, 2015.

[52] P. Zhao, X. Xu, Y. Liu, V. S. Sheng, K. Zheng, and H. Xiong, "Photo2trip: Exploiting visual contents in geo-tagged photos for personalized tour recommendation," in *ACM MM*, 2017, pp. 916–924.

[53] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, and B. Seybold, "CNN architectures for large-scale audio classification," in *ICASSP*, 2017, pp. 131–135.

Yinwei Wei received the MS degree in computer science and technology from Tianjin University, Tianjin, China, in 2014. He is currently pursuing the Ph.D. degree at School of Computer Science and Technology, Shandong University, Shandong, in 2017. His current research interests include multimedia computing and information retrieval.

Zhouchen Lin (M'00-SM'08) received the Ph.D. degree in applied mathematics from Peking University in 2000. Currently, he is a professor at the Key Laboratory of Machine Perception (MOE), School of Electronics Engineering and Computer Science, Peking University. He is also a chair professor at Northeast Normal University. He was a guest professor at Shanghai Jiaotong University, Beijing Jiaotong University and Southeast University. He was also a guest researcher at the Institute of Computing Technology, Chinese Academic of Sciences.
His research interests include computer vision, image processing, machine learning, pattern recognition and numerical optimization. He is an area chair of CVPR 2014, ICCV 2015, NIPS 2015, AAAI 2016, CVPR 2016 and IJCAI 2016. He is an associate editor of IEEE T. Pattern Analysis and Machine Intelligence and International J. Computer Vision.

Xiang Wang is a research fellow with NExT++, National University of Singapore (NUS). He received his Ph.D. in Computer Science from NUS in 2019. His research interests span recommender systems, information retrieval, data mining, and explainable AI. Various parts of his work have been published in top forums, such as SIGIR, WWW, KDD, TOIS, AAAI, and MM. Moreover, he has served as the PC member for several top-tier conferences, such as SIGIR, MM, and CIKM, and the regular reviewer for TKDE, TOIS, and TIST.

Weili Guan received her bachelor degree from Huaqiao University in 2009. She then obtained her graduate diploma and master degree from National University of Singapore in 2011 and 2014, respectively. After that, she joined Hewlett Packard enterprise Singapore as a software engineer and works there for around five years. She is good at commercializing the multimedia and data mining techniques and use them to boost the business revenue.

Baoquan Chen is now a Professor of School of Computer Science and Technology, Shandong University. He received his MS degree from Tsinghua University, Beijing, and PhD degree from the State University of New York at Stony Brook. He is the recipient of the NSF CAREER award 2003, IEEE Visualization Best Paper Award 2005, and NSFC Outstanding Young Researcher program in 2010. His research interests generally lie in computer graphics, visualization, and human-computer interaction.

Liqiang Nie is currently a professor with the School of Computer Science and Technology, Shandong University. Meanwhile, he is the adjunct dean with the Shandong AI institute. He received his B.Eng. and Ph.D. degree from Xi'an Jiaotong University in July 2009 and National University of Singapore (NUS) in 2013, respectively. After PhD, Dr. Nie continued his research in NUS as a research follow for more than three years. His research interests lie primarily in information retrieval and multimedia computing. Dr. Nie has co-/authored more than 100 papers, received more than 2,500 Google Scholar citations as of Aug 2018. He is an AE of Information Science, an area chair of ACM MM 2018, a special session chair of PCM 2018, a PC chair of ICIMCS 2017. Meanwhile, he is supported by the program of "Thousand Youth Talents Plan 2016" and "Qilu Scholar 2016".