Shorter-is-Better: Venue Category Estimation from Micro-Video

Jianglong Zhang[†,‡*], Liqiang Nie[§,‡], Xiang Wang[‡], Xiangnan He[‡], Xianglin Huang[†], Tat-Seng Chua[‡]
[†] Faculty of Science and Technolgy, Communication University of China, Beijing, China
[‡] School of Computing, National University of Singapore, Singapore
[§] Department of Computer Science and Technology, Shandong University, Shandong, China
{ajiang.zh, nieliqiang, xiangwang.nus, xiangnanhe, exl.huang}@gmail.com, dcscts@nus.edu.sg

ABSTRACT

According to our statistics on over 2 million micro-videos, only 1.22% of them are associated with venue information, which greatly hinders the location-oriented applications and personalized services. To alleviate this problem, we aim to label the bite-sized video clips with venue categories. It is, however, nontrivial due to three reasons: 1) no available benchmark dataset; 2) insufficient information, low quality, and information loss; and 3) complex relatedness among venue categories. Towards this end, we propose a scheme comprising of two components. In particular, we first crawl a representative set of micro-videos from Vine and extract a rich set of features from textual, visual and acoustic modalities. We then, in the second component, build a tree-guided multi-task multi-modal learning model to estimate the venue category for each unseen micro-video. This model is able to jointly learn a common space from multi-modalities and leverage the predefined Foursquare hierarchical structure to regularize the relatedness among venue categories. Extensive experiments have well-validated our model. As a side research contribution, we have released our data, codes and involved parameters.

Keywords

Micro-Video Analysis; Multi-Modal Multi-Task Learning; Venue Category Estimation.

1. INTRODUCTION

The popularity of the traditional online video sharing platforms, like Youtube[1], has changed everything about the Internet [39, 18]. Servers like Youtube have enabled users to capture high-quality and long videos, upload and share them socially with everyone. But the late 2012 has seen a dramatic shift in the way Internet users digest videos: micro-videos spread rapidly across various online flagship platforms, such as Viddy[2], Vine[3], Instagram[4] and Snapchat[5]. Considering Vine as an example, as of December 2015, it has experienced an exponential explosion in its user and video base, reaching approximately 200 million active users monthly and 1.5 billion video loops daily[6]. One reason that such bite-sized videos are gaining popularity is because users can conveniently shoot and instantly share videos via their smartphones without the need for professional camera work, editing, and, therefore, significant budgets. Besides, it takes seconds rather than minutes or even hours to view. The micro-video trends confirms this saying: "every good comes in small packages".

In addition to their value in brevity, authenticity and low cost, micro-video platforms have started encouraging users to associate spatial contexts (venues) to videos. The venue information is usually selected manually by the user relying on a GPS enabled device, and each venue is automatically aligned with a venue category via the Foursquare API[7]. This new feature benefits multifaceted aspects: 1) Footprints recording. It facilitates users to vividly archive where they were and what they did. 2) Personalized applications. Such people-centric location data enables precise personalized services, such as suggesting local restaurants, alerting regional weather, and spreading business information to nearby customers. And 3) other location-based services. Location information is helpful for the inference of users' interests, the improvement of activity prediction, and the simplification of landmark-oriented video search. Despite its significance, users of micro-video platforms have been slow to adopt this geospatial feature: in a random sample over 2 million Vine videos, we found that only 1.22% of the videos are associated with venue information. It is thus highly desired to infer the missing geographic cues.

As a preliminary research, we aim to first infer the venue categories from micro-videos, rather than the exact venues. This task is, however, non-trivial due to the following challenges: 1) **Insufficient information**. The most prominent attribute of micro-video platforms is that they are thriving heavily in the realm of shortness and instant. For example, Vine allows users to upload about

*Jianglong Zhang is a visiting student of National University of Singapore, supervised by Dr. Liqiang Nie and Prof. Tat-Seng Chua.

[1] https://www.youtube.com.

[2] http://www.fullscreen.com/.

[3] https://vine.co/.

[4] https://www.instagram.com/.

[5] https://www.snapchat.com/.

[6] http://tinyurl.com/zttwt6u.

[7] https://github.com/mLewisLogic/foursquare.

[9] https://developer.foursquare.com/categorytree.

MM '16, October 15-19, 2016, Amsterdam, Netherlands
© 2016 ACM. ISBN 978-1-4503-3603-1/16/10...$15.00
DOI: http://dx.doi.org/10.1145/2964284.2964307

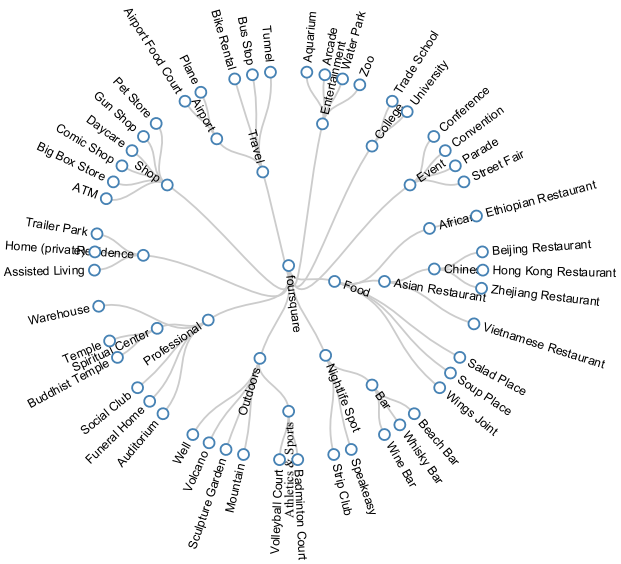


Figure 1: The hierarchical structure of the venue categories. We only illustrate part of the structure due to the limited space. The entire tree can be viewed here[9].

six-second videos online; Snapchat offers its users the option to create 10-second micro-videos; and Viddy limits the length of its upload videos to 30 seconds. Persuasively, short length makes video production and broadcasting easily, downloading timely, and playing fluently on portable devices. However, in contrast to the traditional long videos, the visual information conveyed by micro-videos is somehow inadequate and it is thus unable to provide rich contexts for effective similarity measurement. Besides, as aforementioned, only around 1.22% of videos have venue labels, which brings in a challenge for training data collection. 2) **Low quality**. Most portable devices have nothing to offer for video stabilization. Some videos can thus be shaky or bumpy, which greatly hinders the visual expression. Furthermore, the audio track that comes along with the video, can be in different types of distortion and noise, such as buzzing, hums, hisses, and whistling, which is probably caused by the poor microphones or complex surrounding environments. 3) **Information loss**. Apart from acoustic and visual modalities, micro-videos are, more often than not, uploaded with textual descriptions, which express some useful cues that may be not available in the other two modalities. However, the textual information may be not well correlated with visual and acoustic cues. Moreover, according to our statistics upon 276,624 Vine videos, more than 11.4% of them do not have such texts, probably the results of users' casual habits. This serious information missing problem greatly reduces the usability of textual modality. And 4) **hierarchical structure**. The venues of micro-videos are organized into hundreds of categories, which are not independent but hierarchically correlated. Part of this structure is shown in Figure 1. How to explore such structure to guide the venue category estimation is largely untapped.

To address the aforementioned challenges, we propose a scheme consisting of two components, as illustrated in Figure 2. In the first component, we work towards data preparation. In particular, we first collect a representative set of micro-videos with their associated venue IDs and map these venue IDs to venue categories via Foursquare API. We then extract discriminant features from the textual, visual and acoustic modalities of each video, respectively. In a sense, appropriate fusion of multi-modal features are able to comprehensively and complementarily represent micro-videos, which somehow alleviates the data insufficiency and is robust to the low quality of some modalities. Thereafter, we complete the missing modalities for some videos via matrix factorization techniques, which well solves the information loss problem. The second component aims to label micro-videos with venue categories. Towards this end, we present a TRee-guided mUlti-task Multi-modal leArNiNg model, **TRUMANN** for short. This model intelligently learns a common feature space from multi-modal heterogeneous spaces and utilizes the learned common space to represent each micro-video. Meanwhile, the **TRUMANN** treats each venue category as a task and leverages the pre-defined hierarchical structure of venue categories to regularize the relatedness among tasks via a novel group lasso. These two objectives are accomplished within a unified framework. As a byproduct, the tree-guided group lasso is capable of learning task-sharing and task-specific features. Extensive experiments on our collected real-world dataset have demonstrated the advantages of our model.

The main contributions are in threefold:

1. As far as we know, this is the first work on venue category estimation for micro-videos. We conducted an in-depth analysis of the challenges of this research problem.
2. We proposed a novel tree-guided multi-task multi-modal learning approach. This approach is able to jointly fuse multi-modal information, capture the task relatedness constrained by a pre-defined tree structure.
3. We built a large-scale micro-video datasets. Meanwhile, we have released our data, codes, and involved parameter settings to facilitate the research communities[10].

The remainder of the paper is organized as follows. Section 2 reviews the related work. Section 3 and 4 detail our data preparation and our proposed **TRUMANN** model, respectively. Experimental settings and results are reported in Section 5, followed by the conclusion and future work in Section 6.

2. RELATED WORK

Our work is related to a broad spectrum of multimedia location estimation, and multi-modal multi-task learning.

2.1 Multimedia Venue Estimation

Nowadays, it has become convenient to capture images and videos on the mobile end and associate them with GPS tags. Such a hybrid data structure can benefit a wide variety of potential multimedia applications, such as location recognition [19], landmark search [6], augmented reality[4], and commercial recommendations [44]. It hence has attracted great attention from the multimedia community. Generally speaking, prior efforts can be divided into two

[10]http://acmmm16.wix.com/mm16.

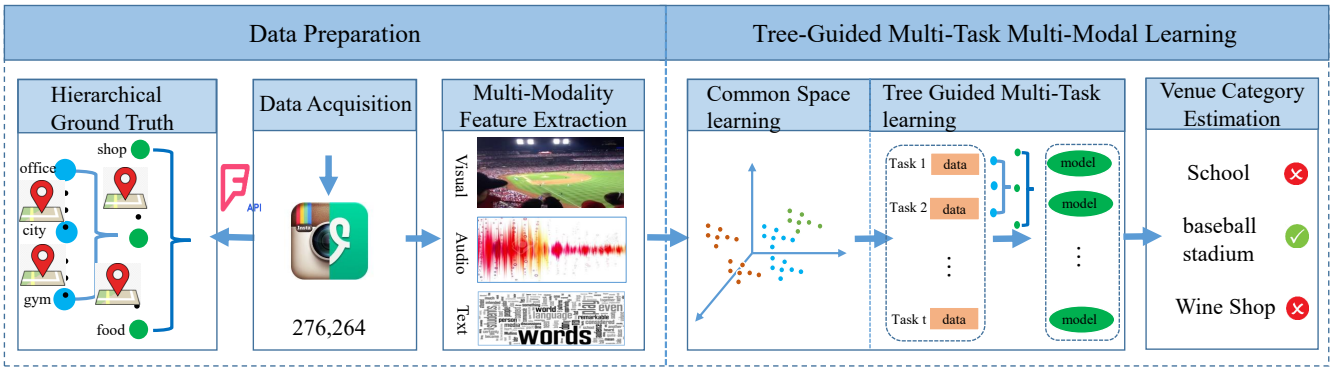


Figure 2: Graphical representation of our framework.

categories: mono-modal venue estimation [4, 6] and multi-modal venue estimation [19, 13, 8]. Approaches in the former category extract a rich set of visual features from images and leverage the visual features to train either shallow or deep models to estimate the venues of the given images. As reported in [19], the landmark identification [6] and scene classification [4] of images are the key factors to recognize the venues. The basic philosophy behind these approaches is that certain visual features in images correlate strongly with certain geographies even if the relationship is not strong enough to specifically pinpoint a location coordinate. Beyond the mono-modal venue estimation which only takes the visual information into consideration, multi-modal venue estimation works by inferring the geo-coordinates of the recording places of the given videos by fusing the textual metadata and visual or acoustic cues [12, 8]. Friendland et al. [12] determined the geo-coordinates of the Flickr videos based on both textual metadata and visual cues. Audio tracks from the Placing Task 2011 dataset videos were also used to train a location estimation models and it achieved reasonable performance [27]. The main idea is that the integration of multiple modalities can lead to better results, and it is consistent to the old saying "two heads are better than one". However, multi-modal venue estimation is still at its infant stage, and more efforts should be dedicated to improve this line of research.

Noticeably, the venue granularity of the targeted multimedia entities in the aforementioned literature varies significantly. Roughly, the spatial resolutions are in three-levels: city-level [19, 13], within-city-level [35, 6, 28] and close-to-exact GPS level [12]. City-level and within-city-level location estimation can be applied to multimedia organization [9], location visualization [6], and image classification [37]. However, their granularities are large, which may be not suitable for some application scenarios, such as business venue discovery [5]. The granularity of close-to-exact GPS level is finer; nevertheless, it is hard to estimate the precise coordinates, especially for the indoor cases. For example, it is challenging to distinguish an office on the third floor and a coffee shop on the second floor within the same building, since the GPS is not available indoors.

Our work differs the above methods from the following two aspects: 1) We focus on the estimation of venue category which is neither city-level nor the precise location. This is because, venue category is more abstract concept than single venue name, which can help many applications for personalized and location-based services/marketing [5]. And 2) micro-videos are the medium between images and traditional long videos, which pose tough challenges.

2.2 Multi-Modal Multi-Task Learning

Venue category estimation of micro-videos exhibits dual-heterogeneities. In particular, a single learning task have features from multiple modalities and multiple learning tasks could be related to each other via their commonalities [41]. In the light of this, our proposed **TRUMANN** falls into the community of multi-view multi-task learning.

The literature on the multi-task problem with multi-modal data is relatively sparse. He et al.[20] proposed a graph-based iterative framework for multi-view multi-task learning (**IteM^2**) and applied it to text classification. However, it can only deal with problems with non-negative feature values. In addition, it is a transductive model. Hence it is unable to generate predictive models for independent and unseen testing samples. To address the intrinsic limitations of transductive models, Zhang et al.[47] proposed an inductive multi-view multi-task learning model (**regMVMT**). **regMVMT** penalizes the disagreement of models learned from different sources over the unlabeled samples. However, without prior knowledge, simply restricting all the tasks to be similar is inappropriate. As an extension of **regMVMT**, an inductive convex shared structure learning algorithm for multi-view multi-task problem (**CSL-MTMV**) was developed in [23]. Compared to **regMVMT**, **CSL-MTMV** considers the shared predictive structure among multiple tasks.

However, none of the methods mentioned above can be applied to venue category estimation directly. This is due to the following reasons: 1) **IteM^2**, **regMVMT** and **CSL-MTMV** are all binary classification models, of which the extension to multi-class or regression problem is nontrivial, especially when the number of classes is large; and 2) the tasks in venue category prediction are pre-defined as a hierarchical structure.

3. DATA PREPARATION

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

3.1 Dataset and Ground Truth Construction

We crawled the micro-videos from Vine through its public API[11]. In particular, we first manually chose a small set

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

Table 1: Number of micro-videos in each of the top ten layer categories.

Top-layer Category	Num.	Top-layer Category	Num.
Outdoors & Recreation	93,196	Shop & Service	10,976
Arts & Entertainment	88,393	Residence	8,867
Travel & Transport	24,916	Nightlife Spot	8,021
Professional & Other	18,700	Food	6,484
College & Education	12,595	Event	1,047

Table 2: Leaf categories with the most and the least of micro-videos.

Leaf Category with the Most Videos	Num.	Leaf Category with the Least Videos	Num.
City	30,803	Bakery	53
Theme Park	16,383	Volcano	51
Neighborhood	15,002	Medical	51
Other Outdoors	10,035	Classroom	51
Park	10,035	Toy & Games	50

of active users from Vine as our seed users. We then adopted the breadth-first strategy to expand our user sets via gathering their followers. We terminated our expansion after three layers. For each collected user, we crawled his/her published videos, video descriptions and venue information if available. In such way, we harvested 2 million micro-videos. Thereinto, only about 24,000 micro-videos contain Foursquare check-in information. After removing the duplicate venue IDs, we further expanded our video set by crawling all videos in each venue ID with the help of vinepy API. This eventually yielded a dataset of 276,264 videos distributed in 442 Foursquare venue categories. Each venue ID was mapped to a venue category via the Foursquare API, which serves as the ground truth. As shown in Figure3, 99.8% of videos are shorter than 7 seconds.

Foursquare organizes its venue categories into a four-layer hierarchical structure[12], with 341, 312 and 52 leaf nodes in the second-layer, third-layer and fourth-layer, respectively. The top-layer of this structure contains ten non-leaf nodes (coarse venue categories). To visualize the coverage and representativeness of our collected micro-videos, we plotted and compared the distribution curves over the number of leaf categories between our dataset and the original structure, as shown in Figure 4. It is worth mentioning that, the number of leaf categories distributed in Foursquare is extremely unbalanced. For instance, the 'Food' category has 253 leaf nodes; while the 'Residence' only contains five leaf nodes. Accordingly, the distribution of our crawled videos over the top-layer categories also shows such unbalance, as displays in Table 1.

On the other hand, we observed that some leaf categories contain only a small number of micro-videos. For instance, 'Bank/Financial' only consists of 3 samples in our dataset, which is hard to train a robust classifier. We hence removed the leaf categories with less than 50 micro-videos. At last, we obtained 270,145 micro-videos distributed in 188 Foursquare leaf categories. Table 2 lists the top five leaf categories with the most and the least micro-videos, respectively.

3.2 Feature Extraction

We extracted a rich set of features from visual, acoustic and textual modalities, respectively.

[12]https://developer.foursquare.com/categorytree.

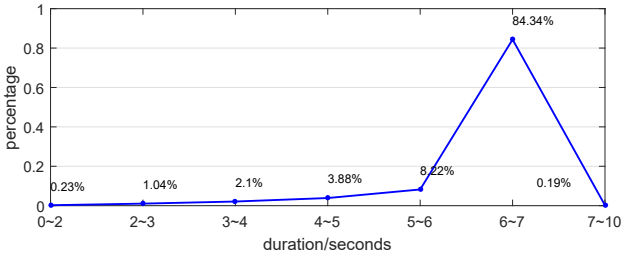


Figure 3: Duration distribution in our dataset.

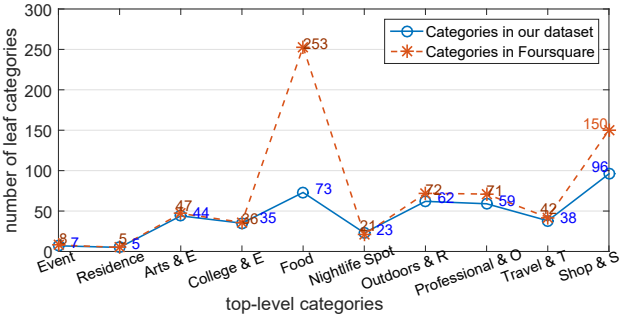


Figure 4: Top-level venue category distribution in terms of the number of their leaf nodes.

3.2.1 *Features in Visual Modality*

Scenes or visual concepts conveyed by the visual modality are intuitive signals of the venue category. For instance, dishes, drinks, waiter and tables are the main elements of restaurants. This motivates us to extract high-level semantics from the visual modality to represent the micro-videos. Deep convolutional neural networks (CNNs) have been established as a powerful model to capture the visual concepts of images [45, 43]. We employed the AlexNet [26] model to extract the visual features through the publicly available Caffe [22]. The model was pre-trained on a set of 1.2 million clean images of ILSVRC-2012[13] and it hence provides a robust initialization for recognizing semantics. Before feature extraction, we first extracted the key frames from each micro-video by using OPENCV[14], and then employed the AlexNet to get CNN features of each frame. Following that, we took the mean pooling strategy over all the key frames of one video, and generated a single 4,096 dimensional vector for each micro-video.

3.2.2 *Features in Acoustic Modality*

The audio clips embedded in the micro-videos contain useful cues or hints about the places. For example, within the clothing stores, audio clips capture employees answering customers questions as to clothing sizes or colors, and welcoming them to the store. Considering the nightclub as another example, the audio clips can reflect the mixed noise but young sounds. 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 features, we first separated audio tracks from micro-videos with the help of FFmpeg[15]. Hereafter, the audio tracks were transformed into a uniform format: 22,050Hz, 16bits, mono-channel and pulse-code modulation signals.

[13]http://www.image-net.org/challenges/LSVRC/2012/.
[14]http://opencv.org/.
[15]https://www.ffmpeg.org/.

Table 3: List of the 10 representative hashtags and their frequencies in our dataset.

ID	Hashtag	Num	ID	Hashtag	Num
1	NYC	1688	6	Paris	352
2	LosAngles	715	7	Food	345
3	Beach	694	8	Hollywood	335
4	Chicago	467	9	London	323
5	Travel	436	10	Disney	309

We then performed a spectrogram with a 46ms window and 50% overlap via librosa[16]. After getting the shallow representation of each audio track with 512 dimensional features, we adopted theano [2] to learn the deep learning features. In particular, the stack Denosing AutoEncoder (DAE) [38], which has been successful applied in speech recognition and speaker recognition [48, 11], was employed to extract acoustic features. The DAE was pre-trained on an external Vine micro-videos, which contains 120,000 samples. The deep model contains three hidden layers, with 500, 400, and 300 nodes on each layer. We ultimately obtained 200 dimensional acoustic features for each micro-video.

3.2.3 Features in Textual Modality

The textual descriptions of micro-videos, including user generated text and hashtags, can provide strong cues for micro-video venue estimation. For instance, this description *"Vining the #beach while tanning the thighs on a glorious Anzac Day"* clearly indicates that the venue category is beach. According to our statistics, 27.7% of our collected 276,264 micro-videos have hashtags, and the total number of hashtags is 253,474. Table 3 lists ten representative hashtags with their frequencies. It can be seen that, these hashtags are strongly related to venue categories. These textual data are, however, very sparse and lack of sufficient contexts. Therefore, the traditional approaches such as topic-level features [3] or n-grams may be unsuitable in such scenario. Instead, we utilized the Paragraph Vector method [33], which has been found to be effective to alleviate the semantic problems of word sparseness [14]. In particular, we first eliminated the non-English characters, followed by removing the stop words. We then employed Sentence2Vector tool[17] to extract the textual features. We finally extracted 100 dimensional features for each micro-video description.

Apart from the deep learning features mentioned above, we also extracted some traditional features such as color histogram and Mel-frequency spectrogram (MFS) to enrich the feature set, which are summarized in Table 4.

3.3 Missing Data Completion

We observed that the acoustic and textual modalities are missing in some micro-videos. More precisely, there are 169 and 24,707 micro-videos with missing acoustic and textual modality, respectively. Information missing is harmful for most machine learning performance [10, 46], including the models for venue category estimation. To alleviate such problem, we cast the data completion task as a matrix factorization problem [50]. In particular, we first concatenated the features from three modalities in order, which naturally constructed an original matrix. We then applied the matrix factorization technique to factorize this original matrix into two latent matrices with 100 latent features, such that the empirical errors between the production of these two latent matrices and the original matrix are as small as possible. The entries in the two latent matrices are inferred by the observed values in the original matrix only, and over-fitting is avoided through a regularized model.

[16] https://github.com/bmcfee/librosa.

[17] https://github.com/klb3713/sentence2vec.

Table 4: Feature summarization of three modalities.

Modality	Extracted features
Visual	4,096-D CNN, 96-D color histogram
Acoustic	200-D DAE, 256-D MFS
Textual	100-D paragraph vector

4. OUR TRUMANN MODEL

4.1 Notations and Assumptions

We first declare some notations. In particular, 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, and Greek letters (e.g., β) as parameters. We denote the Frobenius norm and the group lasso (i.e., $\ell_{2,1}$-norm) matrix $\mathbf{X}$ as $\|\mathbf{X}\|_F$ and $\|\mathbf{X}\|_{2,1}$, respectively. Moreover, let X_{ij} denote the entry in row i and column j. If not clarified, all vectors are in column forms.

Suppose we have a set of N micro-video samples. Each has S modalities and is associated with one of T venue categories. In this work, we treat each venue category as a task. We utilize $\mathbf{X}^s = [\mathbf{x}_1^s, \mathbf{x}_2^s, \ldots, \mathbf{x}_N^s]^T \in \mathbb{R}^{N \times D^s}$ to denote the representation of N samples with a D^s dimensional feature space from the s-th modality, and utilize $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N]^T \in \mathbb{R}^{N \times T}$ to denote the labels of the N samples over the T pre-defined tasks $\{t_1, t_2, \ldots, t_T\}$. Our objective is to jointly learn the mapping matrix $\mathbf{A}^s$ from the individual space $\mathbf{X}^s$ to the common space $\mathbf{B} \in \mathbb{R}^{N \times K}$, and learn the optimal coefficient matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w_T}] \in \mathbb{R}^{K \times T}$. Based on $\mathbf{A}^s$ and $\mathbf{W}$ we are able to estimate the venue categories for the unseen videos.

To intuitively demonstrate our proposed model, we introduce TWO assumptions first:

1. We assume that there exists a common discriminative space for micro-videos, originating from their multi-modalities. Micro-videos can be comprehensively described in this common space and the venue categories are more distinguishable in this space. The space over each individual modality can be mathematically mapped to the common space with a small difference.

2. The tasks (venue categories) are organized into a tree structure. We assume that such structure encodes the relatedness among tasks and leveraging this prior knowledge is able to boost the learning performance.

To consider the aforementioned assumptions simultaneously, we devise a tree-guided multi-task multi-modal learning method, which will be detailed in a stepwise way.

4.2 Common Space Learning

Common space learning [42, 16] over multiple modalities or views has been well studied. Theoretically, it can capture the intrinsic and latent structure of data, which preserveps information from multiple modalities [29, 40]. It is thus able to alleviate the fusion and disagreement problems of the classification tasks over multiple modalities [16, 30]. Based upon our first assumption, we propose a joint optimization framework which minimizes the reconstruction errors over multiple modalities of the data, and avoids overfitting using Frobenius norm on the transformation matrices. It is formally defined as,

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

where $\mathbf{B}\in\mathbb{R}^{N\times K}$ is the representation matrix in the common space learned from all modalities, and K is the latent feature dimension. $\mathbf{A}^s\in\mathbb{R}^{D^s\times K}$ is the transformation matrix from the original feature space over the s-th modality to the common space; and λ_1 and λ_2 are nonnegative tradeoff parameters.

4.3 Tree-guided Multi-task Learning

Although the existing multi-task learning methods, such as graph-regularized [49] and clustering-based [21], achieve sound theoretical underpinnings and great practical success, the tree-guided method [25] is more suitable and feasible for our problem. This is because the relatedness between the venue categories are naturally organized into a hierarchical tree structure by experts from Foursquare. As Figure1 shows, the relatedness among different tasks can be characterized by a tree τ with a set of nodes $\mathcal{V}$, where the leaf nodes and internal nodes represent tasks and groups of the tasks, respectively. Intuitively, each node $v\in\mathcal{V}$ of the tree can be associated with a corresponding group $\mathcal{G}_v=\{t_i\}$, which consists of all the leaf nodes t_i belonging to the subtree rooted at the node v. To capture the strength of relatedness among tasks within the same group $\mathcal{G}_v$, we assign a weight e_v to node v according to an affinity function, which will be detailed in the next subsection. Moreover, the higher level the internal node locates at, the weaker relatedness it controls, and hence the smaller weight it obtains. Therefore, we can formulate such tree-guided multi-task learning as follows,

$$\min_{\mathbf{W},\mathbf{B}}\Gamma=\frac{1}{2}\|\mathbf{Y}-\mathbf{B}\mathbf{W}\|_F^2+\frac{\lambda_3}{2}\sum_{v\in\mathcal{V}}e_v\|\mathbf{W}_{\mathcal{G}_v}\|_{2,1}, \quad (2)$$

where $\mathbf{W}_{\mathcal{G}_v}=\{\mathbf{w}_i:t_i\in\mathcal{G}_v\}\in\mathbb{R}^{K\times|\mathcal{G}_v|}$ is the coefficient matrix of all the leaf nodes rooted at v, where each column vector is selected from $\mathbf{W}$ according to the members within the task group $\mathcal{G}_v$; $||\mathbf{W}_{\mathcal{G}_v}||_{2,1}=\sum_{k=1}^{K}\sqrt{\sum_{t_i\in\mathcal{G}_v}w_{ki}^2}$ is the $\ell_{2,1}$-norm regularization (i.e., group lasso) which is capable of selecting features based on their strengths over the selected tasks within the group $\mathcal{G}_v$, and in this way, we can simultaneously learn the task-sharing features and task-specific features. Lastly, the nonnegative parameter λ_3 regulates the sparsity of the solution regarding $\mathbf{W}$. By integrating the common space learning function in Eqn.(1) and the tree-guided multi-task learning framework in Eqn.(2), we reach the final objective function as follows,

$$\min_{\mathbf{W},\mathbf{B},\mathbf{A}^s}\Gamma=\frac{1}{2}\|\mathbf{Y}-\mathbf{B}\mathbf{W}\|_F^2+\frac{\lambda_1}{2}\sum_{s=1}^{S}\|\mathbf{X}^s\mathbf{A}^s-\mathbf{B}\|_F^2+\frac{\lambda_2}{2}\sum_{s=1}^{S}\|\mathbf{A}^s\|_F^2+\frac{\lambda_3}{2}\sum_{v\in\mathcal{V}}e_v\|\mathbf{W}_{\mathcal{G}_v}\|_{2,1}. \quad (3)$$

Algorithm 1 optimization of **TRUMANN** model

Input: $\mathbf{X}^s$, $\mathbf{Y}$, λ_1, λ_2, λ_3, $\mathbf{e}$, $\mathcal{V}$,
K: dimension of the desired common space.
Output: $\mathbf{A}^s$, $\mathbf{B}$, $\mathbf{W}$.
Initialize $\mathbf{A}^s$, $\mathbf{B}$, $\mathbf{W}$ and $\mathbf{Q}$
while not converge **do**
  Fixing $\mathbf{B}$ and $\mathbf{W}$, update $\mathbf{A}^s$ according to $\mathbf{A}^s\leftarrow(\lambda_1\mathbf{X}^{sT}\mathbf{X}^s+\lambda_2\mathbf{I})^{-1}(\lambda_1\mathbf{X}^s\mathbf{B})$.
  Fixing $\mathbf{A}$ and $\mathbf{W}$, update $\mathbf{B}$ according to $\mathbf{B}\leftarrow(\mathbf{Y}\mathbf{W}^T+\lambda_1\sum_{s=1}^{S}\mathbf{X}^s\mathbf{A}^s)(\lambda_1 S\mathbf{I}+\mathbf{W}\mathbf{W}^T)^{-1}$
  while not converge **do**
    Fixing $\mathbf{W}$, update $q_{k,v}$ according to $q_{k,v}\leftarrow\frac{e_v\|\mathbf{W}_{\mathcal{G}_v}^k\|}{\sum_{k=1}^{K}\sum_{v\in\mathcal{V}}e_v\|\mathbf{W}_{\mathcal{G}_v}^k\|}$
    Updating $\mathbf{Q}^t$ according to $\mathbf{Q}^t\leftarrow\sum_{v\in\mathcal{V}}\frac{e_v^2}{q_{k,v}}$
    Fixing $\mathbf{A}$, $\mathbf{Q}^t$ and $\mathbf{B}$, update $\mathbf{w}^t$ according to $\mathbf{w}^t\leftarrow(\mathbf{B}^T\mathbf{B}+\lambda_3\sum_{t=1}^{T}\mathbf{Q}^t)^{-1}\mathbf{B}^T\mathbf{y}^t$
  end while
end while

4.4 Optimization

We adopt the alternating optimization strategy to solve the three variables $\mathbf{A}^s$, $\mathbf{B}$ and $\mathbf{W}$ in Eqn.(3). To be more specific, we optimize one variable while fixing the others in each iteration. We keep this iterative procedure until the objective function converges.

4.4.1 Computing $\mathbf{A}_s$ with $\mathbf{B}$ and $\mathbf{W}$ fixed

We first fix $\mathbf{B}$ and $\mathbf{W}$, and take derivative of Γ with respect to $\mathbf{A}^s$. We have,

$$\frac{\partial\Gamma}{\partial\mathbf{A}^s}=\lambda_1(\mathbf{X}^s\mathbf{A}^s-\mathbf{B})\mathbf{X}^s+\lambda_2\mathbf{A}^s. \quad (4)$$

By setting Eqn.(4) to zero, it can be derived that,

$$\mathbf{A}^s=(\lambda_1\mathbf{X}^{sT}\mathbf{X}^s+\lambda_2\mathbf{I})^{-1}(\lambda_1\mathbf{X}^{sT}\mathbf{B}), \quad (5)$$

where $\mathbf{I}\in\mathbb{R}^{D^s\times D^s}$ is an identity matrix. The first term in Eqn.(5) can be easily proven to be positive definite and hence invertible according to the definition of positive-definite matrix.

4.4.2 Computing $\mathbf{B}$ with $\mathbf{A}^s$ and $\mathbf{W}$ fixed

With $\mathbf{A}^s$ and $\mathbf{W}$ fixed, we compute the derivative of Γ regarding $\mathbf{B}$ as follows,

$$\frac{\partial\Gamma}{\partial\mathbf{B}}=\lambda_1\sum_{s=1}^{S}(\mathbf{B}-\mathbf{X}^s\mathbf{A}^s)+(\mathbf{B}\mathbf{W}\mathbf{W}^T-\mathbf{Y}\mathbf{W}^T). \quad (6)$$

By setting Eqn.(6) to zero, we have,

$$\mathbf{B}=(\mathbf{Y}\mathbf{W}^T+\lambda_1\sum_{s=1}^{S}\mathbf{X}^s\mathbf{A}^s)(\lambda_1 S\mathbf{I}+\mathbf{W}\mathbf{W}^T)^{-1}, \quad (7)$$

where $(\lambda_1 S\mathbf{I} + \mathbf{W}\mathbf{W}^T)$ can be easily proven to be invertible according to the definition of positive-definite matrix.

4.4.3 Computing $\mathbf{W}$ with $\mathbf{A}^s$ and $\mathbf{B}$ fixed

Considering that the last term in Eqn.(3) is not differentiable, we use an equivalent formulation of it, which has been proven by [1], to facilitate the optimization as follows,

$$\frac{\lambda_3}{2}(\sum_{v\in\mathcal{V}}\|\mathbf{W}_{\mathcal{G}_v}\|)^2. \tag{8}$$

Still, it is intractable. We thus further resort to another variational formulation of Eqn.(8). According to the Cauchy-Schwarz inequality, given an arbitrary vector $\mathbf{b} \in \mathbb{R}^M$ such that $\mathbf{b} \neq \mathbf{0}$, we have,

$$\begin{aligned}\sum_{i=1}^{M}|b_i| &= \sum_{i=1}^{M}\theta_i^{\frac{1}{2}}\theta_i^{-\frac{1}{2}}|b_i| \\ &\leq (\sum_{i=1}^{M}\theta_i)^{\frac{1}{2}}(\sum_{i=1}^{M}\theta_i^{-1}b_i^2)^{\frac{1}{2}} \leq (\sum_{i=1}^{M}\theta_i^{-1}b_i^2)^{\frac{1}{2}},\end{aligned} \tag{9}$$

where θ_i's are the introduced variables that should satisfy $\sum_{i=1}^{M}\theta_i = 1$, $\theta_i > 0$ and the equality holds for $\theta_i = |b_i|/\|\mathbf{b}\|_1$. Based on this preliminary, we can derive the following inequality,

$$(\sum_{v\in\mathcal{V}} e_v\|\mathbf{W}_{\mathcal{G}_v}\|)^2 \leq \sum_{k=1}^{K}\sum_{v\in\mathcal{V}}\frac{e_v^2\|\mathbf{w}_{\mathcal{G}_v}^k\|_2^2}{q_{k,v}}, \tag{10}$$

where $\sum_k\sum_v q_{k,v} = 1$, $q_{k,v} \geq 0, \forall k, v$; $\mathbf{w}_{\mathcal{G}_v}^k$ denotes the k-th row vector of the group matrix $\mathbf{W}_{\mathcal{G}_v}$. It worth noting that the equality holds when

$$q_{k,v} = \frac{e_v\|\mathbf{w}_{\mathcal{G}_v}^k\|_2^2}{\sum_{k=1}^{K}\sum_{v\in\mathcal{V}} e_v\|\mathbf{w}_{\mathcal{G}_v}^k\|_2^2}. \tag{11}$$

Thus far, we have theoretically derived that minimizing Γ with respect to $\mathbf{W}$ is equivalent to minimizing the following convex objective function,

$$\begin{aligned}\min_{\mathbf{W},q_{k,v}} \frac{1}{2}\|\mathbf{Y} - \mathbf{B}\mathbf{W}\|_F^2 + \frac{\lambda_1}{2}\sum_{s=1}^{S}\|\mathbf{X}^s\mathbf{A}^s - \mathbf{B}\|_F^2 + \\ \frac{\lambda_2}{2}\sum_{s=1}^{S}\|\mathbf{A}^s\|_F^2 + \frac{\lambda_3}{2}\sum_{k=1}^{K}\sum_{v\in\mathcal{V}}\frac{\|e_v\mathbf{w}_{\mathcal{G}_v}^k\|^2}{q_{k,v}}.\end{aligned} \tag{12}$$

To facilitate the computation of the derivative of objective function Γ with respect to $\mathbf{w}_t$ for the t-th task, we define a diagonal matrix $\mathbf{Q}^t \in \mathbb{R}^{K\times K}$ with the diagonal entry as follows,

$$Q^t{}_{kk} = \sum_{\{v\in\mathcal{V},|t\in v\}}\frac{e_v^2}{q_{k,v}}. \tag{13}$$

We ultimately have the following objective function,

$$\begin{aligned}\min_{\mathbf{W},\mathbf{Q}} \frac{1}{2}\sum_{t=1}^{T}\|\mathbf{y}_t - \mathbf{B}\mathbf{w}_t\|_F^2 + \frac{\lambda_1}{2}\sum_{s=1}^{S}\|\mathbf{X}^s\mathbf{A}^s - \mathbf{B}\|_F^2 + \\ \frac{\lambda_2}{2}\sum_{s=1}^{S}\|\mathbf{A}^s\|_F^2 + \frac{\lambda_3}{2}\sum_{t=1}^{T}\mathbf{w}_t^T\mathbf{Q}^t\mathbf{w}_t.\end{aligned} \tag{14}$$

The alternative optimization strategy is also applicable here. By fixing $\mathbf{Q}^t$, taking derivative of the above formulation regarding $\mathbf{w}_t$, and setting it to zero, we reach,

$$\mathbf{w_t} = (\mathbf{B}^T\mathbf{B} + \lambda_3\mathbf{Q}^t)^{-1}(\mathbf{B}^T\mathbf{y_t}). \tag{15}$$

Once we obtain all the $\mathbf{w}_t$, we can easily compute $\mathbf{Q}^t$ based on Eqn.(11)

4.5 Task Relatedness Estimation

According to our assumption, the hierarchical tree structure of venue categories plays a pivotal role to boost the learning performance in our model. Hence, the key issue is how to precisely characterize and model the task relatedness in the tree, namely, how to estimate the reasonable weight e_v for each node v in the tree appropriately. Although the existing tree-guided multi-task learning approaches [36, 17] have addressed this issue by exploring the geometric structure, they do not consider the semantic relatedness among tasks. To remedy this problem, we aim to model the intrinsic task relatedness based on the feature space. Towards this goal, we introduce the affinity measurement of the node group proposed in [31]. A high affinity value e_v of the node group $\mathcal{G}_v$ indicates the dense connections and compact relations among the leaf nodes within the given group. We hence can employ the affinity measurement to characterize the task relatedness in the tree.

To facilitate the affinity measurement of each node group $\mathcal{G}_v$, we need to obtain the pairwise similarity between all leaf nodes. For simplicity, we utilize the adjacency matrix $\mathbf{S} \in \mathbb{R}^{T\times T}$ to denote the pairwise similarity matrix and the entry S_{ij} to capture the non-negative relatedness between the i-th and j-th leaf nodes, which can be formulated as,

$$S_{ij} = \exp\left(-\frac{||\bar{\mathbf{x}}_\mathbf{i} - \bar{\mathbf{x}}_\mathbf{j}||^2}{\theta^2}\right), \tag{16}$$

where $\bar{\mathbf{x}}_\mathbf{i}$ represents the mean feature vector of the samples belonging to the i-th venue category which can be extracted from the training dataset; θ is radius parameter that is simply set as the median of the Euclidean distances of all node pairs.

For ease of formulation and inspired by the work in [31], we define a scaled assignment vector $\mathbf{u}_v \in \mathbb{R}^T$ for each node of the tree over all the T leaf nodes which can be stated as,

$$u_{vt} = \begin{cases}\frac{1}{\sqrt{|\mathcal{G}_v|}}, & \text{if} \quad t \in \mathcal{G}_v \\ 0, & \text{otherwise}\end{cases}. \tag{17}$$

Based on the scaled assignment $\mathbf{u}_v$ and the pairwise similarity matrix $\mathbf{S}$, we can further formulate the affinity e_v for the node v as follows,

$$e_v = \mathbf{u}_v^T\mathbf{S}\mathbf{u}_v. \tag{18}$$

Since the characteristics of the affinity definition, the value of the e_v are limited within the range of [0, 1]. More importantly, such affinity measurement can guarantee that higher nodes correspond to weaker relatedness, and vice versa.

4.6 Complexity Analysis

In order to analyze the complexity of our proposed **TRUMANN** model, we have to estimate the time complexity for constructing $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{W}$ as defined in Eqn.(5), Eqn.(7) and Eqn.(15). The computational complexity of the training process is $O(M\times(O_1+O_2+O_3))$, where O_1, O_2 and O_3 respectively equal to $((D^s)^2N +$

$(D^s)^3 + (D^s)^2K)S$, $(NK^2 + NDKS + K^3 + K^2T)$ and $(2NK^2 + K^3)T$. Thereinto, M is the iteration times of the alternative optimization, which is a small value less than 10 in our above analysis. N, T, S, K and D respectively refer to the number of micro-videos, venue categories, modalities, latent dimension and the total feature dimensions over all the modalities. Usually, we consider only a few number of modalities. S is hence very small. In our experimental settings, K and T are in the order of a few hundreds. Meanwhile, the number of feature dimension is about 5,000. Therefore, D^2 is greater than K^2T. In the light of this, we can reduce the time complexity to be ND^2, which is faster than **SVM**, in terms of $O(N^3)$.

5. EXPERIMENTS

All the experiments were conducted over a server equipped with Inter(R) Core(TM) CPU i7-4790 at 3.6GHz on 32Gb RAM, 8 cores and 64-bit Windows 10 operation system.

5.1 Experimental Settings

To thoroughly measure our model and the baselines, we employed multiple metrics, namely, macro-F1 and micro-F1[15]. The averaging macro-$F1$ gives equal weight to each class-label in the averaging process; whereas the averaging micro-$F1$ gives equal weight to all instances in the averaging process. Both macro-$F1$ and micro-$F1$ metrics reach their best value at 1 and worst score at 0.

The experimental results reported in this paper were based on 10-fold cross-validation. In particular, the stratified cross-validation [34] was adopted to ensure all categories contain approximately the same percentage between training and testing samples. In each round of the 10-fold cross-validation, we split our dataset into three chunks: 80% of the micro-videos (i.e., 194,505 videos) were used for training, 10% (i.e., 24,313 videos) were used for validation, and the rest (i.e., 24,313 videos) were held-out for testing. The training set was used to adjust the parameters, while the validation set was used to avoid overfitting, i.e., verifying that any performance increase over the training dataset actually yields an accuracy increase over a dataset that has not been shown to the model before. The testing set was used only for testing the final solution to confirm the actual predictive power of our model with optimal parameters. Grid search was employed to select the optimal parameters with small but adaptive step size.

5.2 Performance Comparison among Models

We carried out experiments to compare the overall effectiveness of our proposed **TURMANN** model with several state-of-the-art baselines:

- **SRMTL**: The Sparse Graph Regularization Multi-Task Learning method can capture the relationship between task pairs and further impose a sparse graph regularization scheme to enforce the related pairs close to each other [32].
- **regMVMT**: This semi-supervised inductive multi-view multi-task learning model considers information from multiple views and learns multiple related tasks simultaneously [47]. Besides, we also compared our model with the variant of **regMVMT** method, dubbed **regMVMT+**. **regMVMT+** can achieve better performance by modeling the non-uniformly related tasks.

Table 5: Performance comparison between our model and the baselines on the venue category estimation. (p-value*: p-value over micro-F1.)

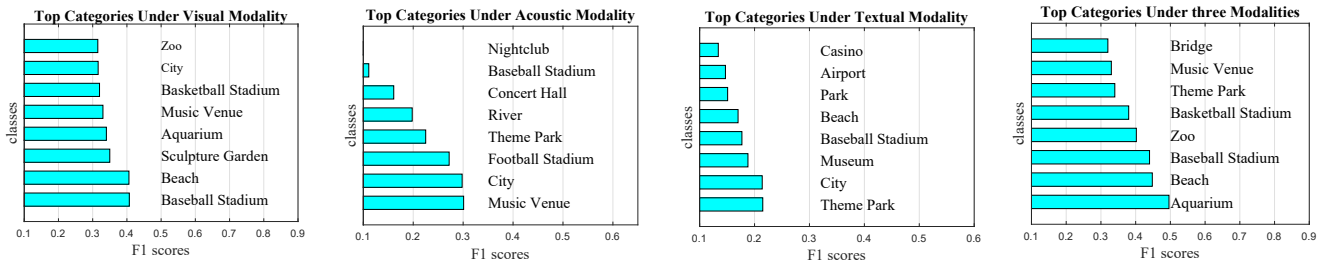
Models	Macro-$F1$	Micro-$F1$	p-value*
SRMTL	2.61±0.19%	15.71±0.21%	1.1e-3
regMVMT	4.33±0.41%	17.16±0.28%	7.0e-3
regMVMT+	4.53±0.31%	18.35±0.13%	9.1e-3
MvDA+RMTL	2.46±0.18%	17.28±1.67%	1.0e-3
TRUMANN-	3.75±0.17%	24.01±0.35%	1.0e-2
TRUMANN	**5.21±0.29%**	**25.27±0.17%**	-

- **MvDA+RMTL**: This baseline is the combination of Multi-view Discriminant Analysis [24] and Robust Multi-Task Learning [7]. In particular, **MvDA** seeks for a single discriminant common space for multiple views by jointly learning multiple view-specific linear transforms. Meanwhile, the **RMTL** is able to capture the task relationships using a low-rank structure via group-sparse lasso.
- **TURMANN-**: This baseline is the variant of our proposed model by setting all e_v in Eqn.(3) to be 1. In other words, this baseline does not incorporate the knowledge of the pre-defined hierarchical structure.

The comparative results are summarized in Table 5. From this table, we have the following observations: 1) **TRUMANN** achieves better performance, as compared to other multi-task learning approaches, such as **SRMTL**. This is because, the **SRMTL** cannot capture the prior knowledge of task relatedness in terms of tree structure. On the other hand, it reflects that micro-videos are more separable in the learnt common space. 2) Multi-modal multi-task models, such as **regMVMT** and **TRUMANN** remarkably outperform pure multi-task learning models, such as **SRMTL**. This again demonstrates that the relatedness among multi-modalities can boost the learning performance. 3) The joint learning of multi-modal multi-task models, including **regMVMT** and **TRUMANN**, shows their superiors to the sequential learning of multi-view multi-task model, **MvDA+RMTL**. This tells us that multi-modal learning and multi-task learning can mutually reinforce each other. 4) We can see that **TRUMANN** outperforms **TRUMANN-**. This demonstrate the usefulness of the pre-defined hierarchical structure, and reveals the necessity of tree-guided multi-task learning. And 5) we conducted the analysis of variance (known as ANOVA) micro-$F1$. In particular, we performed paired t-test between our model and each of the competitors over 10-fold cross validation. We found that all the p-value are substantially smaller than 0.05, which shows that the improvements of our proposed model are statistically significant.

5.3 Representativeness of Modalities

We also studied the effectiveness of different modality combination. Table 6 shows the results. From this table, we observed that: 1) The visual modality is the most discriminant one among visual, textual and acoustic modalities. This is because, the visual modality contains more location-specific information than acoustic and textual modality. On the other hand, it signals that the CNN features are capable of capturing the prominent visual characteristics of venue categories. 2) The acoustic modality provide important cues for venue categories as compared to the textual modality across micro-$F1$ and



(a) Visual Modality (b) Acoustic Modality (c) Textual Modality (d) Modality Combination

Figure 5: Categories with best classification performance under visual, acoustic, textual modality and their combination, respectively.

macro-$F1$ metrics. But only given the acoustic modality, it is hard to estimate the venue categories for most of the videos, while the combination of visual and acoustic modality get an improvement than visual modality. 3) Textual modality is the least descriptive for venue category estimation. This is due to that the textual descriptions are noisy, missing, sparse, and even irrelevant to the venue categories. And 4) the more modalities we incorporate, the better performance we can achieve. This implies that the information of one modality is insufficient and multi-modalities is complementary to each other rather than mutually conflicting. This is a consensus to the old saying "two heads are better than one".

5.4 Case Studies

In Figure 5, we respectively list the top eight categories with best performance in only visual modality, acoustic modality, textual modality, and their combination. From this figure, we have the following observations: 1) For visual modality, our model achieves stable and satisfactory performance on many venue categories, especially on those with discriminant visual characteristic, such as the micro-videos related to 'Zoo' and 'Beach'. 2) Regarding the acoustic modality, our model performs better on those with regular sounds or noisy noise, such as, 'Music Venue' and 'Concert Hall', which have discriminate acoustic signals as compared to other venue categories. 3) When it comes to the textual modality, we found that the top eight best performing categories are with high frequencies in micro-video descriptions. For instance, the terms of 'Park' and 'Beach' occur 2,992 and 3,882 times in our dataset, respectively. It is worth noting that not all the textual descriptions are correlated with the actual venue category, which in fact decreases the performance. For example, the textual description of one micro-video is *'I love my city'*. Nevertheless, its venue category is 'Park'. And 4) unsurprisingly, we obtained a significant improvement for 'Aquarium' category, which is hard to recognize with only one modality. Moreover, compared to the performance over visual modality, the 'Basketball Stadium' and 'Zoo' categories are also improved about 8% in micro $F1$. Besides, the more training samples one venue category contains, the higher probability of this category will yield, such as 'Theme Park' and 'City'.

5.5 Parameter Tuning and Sensitivity

We have four key parameters as shown in Eqn.(3): K, λ_1, λ_2 and λ_3. The optimal values of these parameters were carefully tuned with 10-fold cross-validation in the training data. In particular, for each of the 10-fold, we chose the optimal parameters by grid search with a small but adaptive step size. Our parameters were searched in the range of [50, 500], [0.01,1], [0,1] and [0,1], respectively. The parameters corresponding to the best micro $F1$-score were used to report the final results. For other competitors, the procedures to tune the parameters are analogous to the ensure fair comparison.

Table 6: Representativeness of different modalities. (p-value*: p-value over micro-F1.)

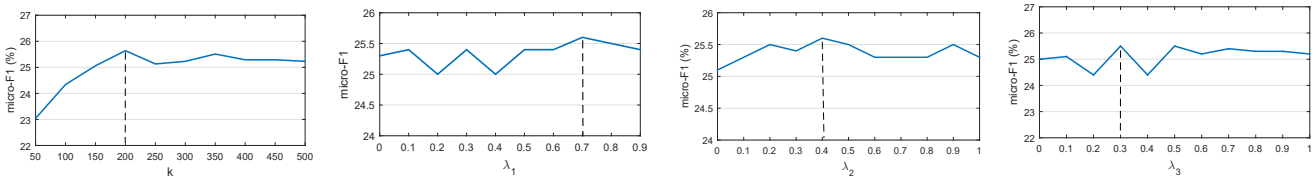
Modality	Macro-$F1$	Micro-$F1$	p-value*
Visual	4.49±0.09%	22.56±0.10%	2.3e-2
Acoustic	2.79±0.01%	16.25±0.46%	2.9e-4
Textual	1.44±0.29%	12.36±0.38%	5.4e-4
Acoustic+textual	2.87±0.16%	16.86±0.06%	6.4e-3
Visual+acoustic	4.61±0.08%	23.85±0.20%	1.8e-2
Visual+textual	4.52±0.11%	23.54±0.17%	1.1e-2
All	**5.21±0.29%**	**25.27±0.17%**	-

Take the parameter tuning in one of the 10-fold as an example. We observed that our model reached the optimal performance when K=200, $\lambda_1 = 0.7$, $\lambda_2 = 0.4$ and $\lambda_3 = 0.3$. We then investigated the sensitivity our model to these parameters by varying one and fixing the others. Figure 6 illustrates the performance of our model with respect to K, λ_1, λ_2 and λ_3. We can see that: 1) When fixing λ_1, λ_2, λ_3 and tuning K, the micro $F1$ score value increases first and then reaches the peak value at K=200. And 2) the micro $F1$ score value changes in a small range, when varying λ_1, λ_2 and λ_3 from 0 to 1. The slight change demonstrates that our model is non-sensitive to parameters.

At last, we recorded the value of micro $F1$ along with the iteration time using the optimal parameter settings. Figure7 shows the convergence process with respect to the number of iterations. From this figure, it can be seen that our algorithm can converge very fast.

6. CONCLUSION AND FUTURE WORK

In this paper, we present a novel tree-guided multi-task multi-modal learning method to label the bite-sized video clips with venue categories. This model is capable of learning a common feature space from multiple and heterogonous modalities, which preserves the information of each modality via disagreement penalty. In this common space, the venue categories of micro-videos are more distinguishable. Meanwhile, the proposed method intelligently leverages the pre-defined Foursquare hierarchical structure to regularize the relatedness among venue categories. We seamlessly integrate the common space learning and multi-task



(a) parameter k (b) parameter λ_1 (c) parameter λ_2 (d) parameter λ_3

Figure 6: Performance of TRUMANN with regards to varying parameters.

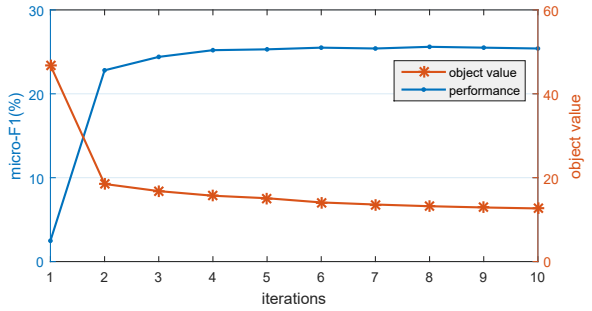


Figure 7: Performance of TRUMANN with the iteration times

classification into a unified model. To validate our model, we crawl a representative set of micro-videos from Vine and extract a rich set of features from textual, visual and acoustic modalities. Based upon the extensive experiments on this dataset, we have shown that our model is superior to several state-of-the-art baselines.

In the future, we plan to study the relatedness among multiple modalities, such as conflict, complementary and consistent relatedness, to boost the learning performance.

7. ACKNOWLEDGMENTS

This work is supported by National Key Technology Research and Development Program of the Ministry of Science and Technology of China (No.2012BAH17B02), CUC Engineering Project (No.3132014XNG1429), and China Scholarship Council (CSC). It is also supported by Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

8. REFERENCES

[1] F. R. Bach. Consistency of the group lasso and multiple kernel learning. *JMLR*, 9(Jun):1179–1225, 2008.

[2] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, D. Warde-Farley, and Y. Bengio. Theano: new features and speed improvements. *arXiv preprint arXiv:1211.5590*, 2012.

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

[4] S. Cao and N. Snavely. Graph-based discriminative learning for location recognition. In *IEEE CVPR*, pages 700–707, 2013.

[5] B.-C. Chen, Y.-Y. Chen, F. Chen, and D. Joshi. Business-aware visual concept discovery from social media for multimodal business venue recognition. In *AAAI*, pages 61–68, 2016.

[6] D. M. Chen, G. Baatz, K. Köser, S. S. Tsai, R. Vedantham, T. Pylvä, K. Roimela, X. Chen, J. Bach, M. Pollefeys, et al. City-scale landmark identification on mobile devices. In *IEEE CVPR*, pages 737–744, 2011.

[7] J. Chen, J. Zhou, and J. Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *ACM KDD*, pages 42–50, 2011.

[8] J. Choi, G. Friedland, V. Ekambaram, and K. Ramchandran. Multimodal location estimation of consumer media: Dealing with sparse training data. In *IEEE ICME*, pages 43–48, 2012.

[9] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *ACM WWW*, pages 761–770, 2009.

[10] P. Cui, Z. Wang, and Z. Su. What videos are similar with you?: Learning a common attributed representation for video recommendation. In *ACM MM*, pages 597–606, 2014.

[11] X. Feng, Y. Zhang, and J. Glass. Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition. In *IEEE ICASSP*, pages 1759–1763, 2014.

[12] G. Friedland, J. Choi, H. Lei, and A. Janin. Multimodal location estimation on flickr videos. In *ACM SIGMM*, pages 23–28, 2011.

[13] G. Friedland, O. Vinyals, and T. Darrell. Multimodal location estimation. In *ACM MM*, pages 1245–1252, 2010.

[14] D. Ganguly, D. Roy, M. Mitra, and G. J. Jones. Word embedding based generalized language model for information retrieval. In *ACM SIGIR*, pages 795–798, 2015.

[15] S. Gopal and Y. Yang. Recursive regularization for large-scale classification with hierarchical and graphical dependencies. In *ACM KDD*, pages 257–265, 2013.

[16] Y. Guo. Convex subspace representation learning from multi-view data. In *AAAI*, pages 2–9, 2013.

[17] L. Han and Y. Zhang. Learning tree structure in multi-task learning. In *ACM KDD*, pages 397–406, 2015.

[18] Z. Hanwang, W. Meng, H. Richang, N. Liqiang, and C. Tat-Seng. Play and rewind: Optimizing binary representations of videos by self-supervised temporal hashing. In *ACM MM*, October 2016.

[19] J. Hays and A. A. Efros. Im2gps: estimating geographic information from a single image. In *IEEE CVPR*, pages 1–8, 2008.

[20] J. He and R. Lawrence. A graph-based framework for multi-task multi-view learning. In *ICML*, pages 25–32, 2011.

[21] L. Jacob, J.-p. Vert, and F. R. Bach. Clustered multi-task learning: A convex formulation. In *NIPS*, pages 745–752, 2009.

[22] 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*, pages 675–678, 2014.

[23] X. Jin, F. Zhuang, S. Wang, Q. He, and Z. Shi. Shared structure learning for multiple tasks with multiple views. In *MLKDD*, pages 353–368, 2013.

[24] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. *IEEE TPAMI*, 38(1):188–194, 2016.

[25] S. Kim and E. P. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *ICML*, pages 1–8, 2010.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[27] H. Lei, J. Choi, and G. Friedland. Multimodal city-verification on flickr videos using acoustic and textual features. In *IEEE ICASSP*, pages 2273–2276, 2012.

[28] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays. Learning deep representations for ground-to-aerial geolocalization. In *IEEE CVPR*, pages 5007–5015, 2015.

[29] A. Liu, W. Nie, Y. Gao, and Y. Su. Multi-modal clique-graph matching for view-based 3d model retrieval. *IEEE TIP*, 25(5):2103–2116, 2016.

[30] A. Liu, Z. Wang, W. Nie, and Y. Su. Graph-based characteristic view set extraction and matching for 3d model retrieval. *Inf. Sci.*, 320:429–442, 2015.

[31] H. Liu, X. Yang, L. J. Latecki, and S. Yan. Dense neighborhoods on affinity graph. *IJCV*, 98(1):65–82, 2012.

[32] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient l2, 1-norm minimization. In *UAI*, pages 339–348, 2009.

[33] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *JMLR*, 12:2825–2830, 2011.

[35] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *IEEE CVPR*, pages 1–7, 2007.

[36] X. Song, L. Nie, L. Zhang, M. Liu, and T.-S. Chua. Interest inference via structure-constrained multi-source multi-task learning. In *AAAI*, pages 2371–2377, 2015.

[37] K. Tang, M. Paluri, L. Fei-Fei, R. Fergus, and L. Bourdev. Improving image classification with location context. In *IEEE CVPR*, pages 1008–1016, 2015.

[38] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR*, 11:3371–3408, 2010.

[39] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, and Y. Song. Unified video annotation via multigraph learning. *IEEE TCSVT*, 19(5):733–746, 2009.

[40] M. Wang, H. Li, D. Tao, K. Lu, and X. Wu. Multimodal graph-based reranking for web image search. *IEEE TIP*, 21(11):4649–4661, 2012.

[41] M. Wang, X. Liu, and X. Wu. Visual classification by-hypergraph modeling. *IEEE TKDE*, 27(9):2564–2574, 2015.

[42] M. White, X. Zhang, D. Schuurmans, and Y.-l. Yu. Convex multi-view subspace learning. In *NIPS*, pages 1673–1681, 2012.

[43] Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative cnn video representation for event detection. In *IEEE CVPR*, pages 1798–1807, 2015.

[44] M. Ye, P. Yin, and W.-C. Lee. Location recommendation for location-based social networks. In *ACM AGIS*, pages 458–461, 2010.

[45] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *IEEE CVPR*, pages 4694–4702, 2015.

[46] H. Zhang, X. Shang, W. Yang, H. Xu, H. Luan, and T.-S. Chua. Online collaborative learning for open-vocabulary visual classifiers. In *IEEE CVPR*, June 2016.

[47] J. Zhang and J. Huan. Inductive multi-task learning with multiple view data. In *ACM KDD*, pages 543–551, 2012.

[48] Z. Zhang, L. Wang, A. Kai, T. Yamada, W. Li, and M. Iwahashi. Deep neural network-based bottleneck feature and denoising autoencoder-based dereverberation for distant-talking speaker identification. *EURASIP*, 2015(1):1–13, 2015.

[49] J. Zhou, J. Chen, and J. Ye. Malsar: Multi-task learning via structural regularization. In *Arizona State University*, pages 1–50, 2011.

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