Micro Tells Macro: Predicting the Popularity of Micro-Videos via a Transductive Model

Jingyuan Chen[†], Xuemeng Song[‡,†*], Liqiang Nie[‡,†], Xiang Wang[†], Hanwang Zhang[†], Tat-Seng Chua[†]
[†] National University of Singapore, [‡] Shandong University
{jingyuanchen91, sxmustc, nieliqiang, xiangwang.nus, hanwangzhang}@gmail.com,
dcscts@nus.edu.sg

ABSTRACT

Micro-videos, a new form of user generated contents (UGCs), are gaining increasing enthusiasm. Popular micro-videos have enormous commercial potential in many ways, such as online marketing and brand tracking. In fact, the popularity prediction of traditional UGCs including tweets, web images, and long videos, has achieved good theoretical underpinnings and great practical success. However, little research has thus far been conducted to predict the popularity of the bite-sized videos. This task is non-trivial due to three reasons: 1) micro-videos are short in duration and of low quality; 2) they can be described by multiple heterogeneous channels, spanning from social, visual, acoustic to textual modalities; and 3) there are no available benchmark dataset and discriminant features that are suitable for this task. Towards this end, we present a transductive multi-modal learning model. The proposed model is designed to find the optimal latent common space, unifying and preserving information from different modalities, whereby micro-videos can be better represented. This latent space can be used to alleviate the information insufficiency problem caused by the brief nature of micro-videos. In addition, we built a benchmark dataset and extracted a rich set of popularity-oriented features to characterize the popular micro-videos. Extensive experiments have demonstrated the effectiveness of the proposed model. As a side contribution, we have released the dataset, codes and parameters to facilitate other researchers.

Keywords

Micro-Videos; Popularity Prediction; Multi-View Learning

1. INTRODUCTION

The last couple of years have witnessed the unprecedented growth of smart mobile devices, enabling users to record life stories vividly with short videos and then instantly upload them to social media websites, such as Snapchat[1] and Vine[2]. Since micro-videos, acting more like 'live action photographs', are usually short in length, they need little bandwidth and hence gain tremendous user enthusiasm. The limits regarding the maximum length of micro-videos on Snapchat and Vine, are set as 10 and 6 seconds, respectively. Considering Vine as an example, its video length distribution over our collected 303,242 Vine videos is illustrated in Figure 1. Micro-videos, representing a new form of user generated contents (UGCs), can be viewed, discussed and even reposted by users once they are uploaded, which leads to their rapid rise. It is reported that Vine hit 1.5 billion daily loops[3] in 2015[4] and Snapchat hit 7 billion daily views in 2016[5]. As a video messaging platform, it is hard to crawl micro-videos from Snapchat. Hence, we focus on Vine here, which can be gathered more easily for research.

Interestingly, among the tremendous volume of micro-videos, some popular ones will be widely viewed and spread by users, while many only gain little attention. This phenomena is similar to many existing social media sites, such as Twitter[6]. For example, the micro-video about the explosion that interrupted during the France-Germany soccer match in 2015 has been successfully looped by over 330 million times. Obviously, if we can identify the hot and popular micro-videos in advance, it will benefit many applications, such as online marketing and network reservation. Regarding online marketing, the accurate early prediction of popular micro-videos can facilitate companies' planning of advertising campaigns and thus maximizing their revenues. For network service providers, they can timely reserve adequate distributed storage and bandwidth for popular ones, based on the prediction. Therefore, it is highly desirable to develop an effective scheme to accurately predict the popularity of micro-videos.

However, the popularity prediction of micro-videos is non-trivial due to the following challenges. First of all, due to the short duration of micro-videos, each modality can only provide limited information, the so-called modality limitation. Fortunately, micro-videos always involve multiple modalities, namely, social, visual, acoustic

*Xuemeng Song is the corresponding author.

MM '16, October 15-19, 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2964284.2964314

[1] https://snapchat.com.
[2] https://vine.co.
[3] Loops refer to the times a micro-video has been viewed.
[4] http://blog.vine.co.
[5] http://goo.gl/YYnBbd.
[6] https://twitter.com.

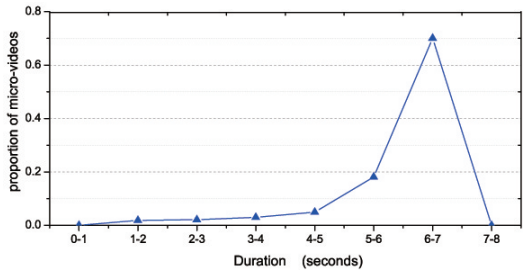


Figure 1: Duration distribution over our collected 303, 242 micro-videos.

and textual[7] modalities. In a sense, these modalities are co-related rather than independent and essentially characterize the same micro-videos. Therefore, the major challenge lies on how to effectively fuse micro-videos' heterogeneous clues from multiple modalities [31, 30, 36]. The most naive strategies are early fusion and late fusion [27]. They, however, fail to account for the relatedness among multiple modalities. Therefore, it is important to take modality relatedness into consideration. Secondly, due to certain external factors, such as camera shaking and lighting condition, some modalities of the micro-videos, such as visual or acoustic ones, may be of poor quality, which is another kind of modality limitation. Hence, learning directly from the original feature spaces of modalities, which was adopted by most multi-modal learning methods may be imprecise. Consequently, to improve the learning performance, how to compensate the noisy modality with reliable ones poses a crucial challenge for us. The last challenge we are facing is the lack of benchmark dataset to support our research. We found that both the contents and social influence of micro-video publishers would account for their popularity. As far as we know, the only available micro-video dataset [24] does not contain the important textual and social modalities and is lack of popularity indicators as the ground truth, which makes it unsuitable for our research. It is thus necessary to build a comprehensive micro-video dataset and further extract a rich set of discriminant features to enhance the learning performance.

To address the aforementioned challenges, we present a novel **T**ransductive **M**ulti-mod**AL** **L**earning approach, TMALL for short, to predicting the popularity of micro-videos. As illustrated in Figure 2, we first crawl a representative micro-video dataset from Vine and develop a rich set of popularity-oriented features from multi-modalities. We then perform multi-modal learning to predict the popularity of micro-videos, which seamlessly takes the modality relatedness and modality limitation into account by utilizing a common space shared by all modalities. We assume that there exists an optimal common space, which maintains the original intrinsic characteristics of micro-videos in the original spaces. In the light of this, all modalities are forced to be correlated. Meanwhile, micro-videos with different popularity can be better separated in such optimal common space, as compared to that of each single modality. In a sense, we alleviate the modality limitation problem. Extensive experiments on this real-world dataset have well-validated our work.

Our main contributions can be summarized in threefold:

- We approached the popularity prediction of micro-videos by proposing a TMALL model, which is able to simultaneously model the modality relatedness and handle the modality limitations by introducing a common space shared among all modalities. Moreover, we have derived its closed-form solution rigorously.
- We developed a rich set of popularity-oriented features from multiple modalities to comprehensively characterize the popular micro-videos. Apart from numerical results, we also provided several deep insights based on the experimental results.
- We constructed a large-scale micro-video dataset, comprising of 303, 242 micro-videos, 98, 166 users and 120, 324 following relationships. We have released our compiled dataset, codes and parameters[8] to facilitate other researchers to repeat our experiments and verify their proposed approaches.

The remainder of this paper is structured as follows. Section 2 reviews the related work. Data preparation is introduced in Section 3. Sections 4 and 5 detail the proposed TMALL model and the feature extraction, respectively. Section 6 presents the experimental results and analysis, followed by our concluding remarks in Section 7.

2. RELATED WORK

Popularity predication and multi-view learning are both related to this work.

2.1 Popularity Prediction

Due to its enormous commercial potential, popularity prediction of UGCs has attracted great attention from both the industry and academia [12, 20, 5, 29, 11]. Hong et al. [12] explored the popularity prediction of tweets, which is measured by the number of future retweets, by introducing a rich set of features, such as topological features, temporal features and meta features. Beyond the text popularity prediction, McParlane et al. [20] focused on the popularity of images. They extracted some sophisticated features and cast the task of image popularity prediction as a problem of binary classification, where the given image would be classified as popular or not. Later, Cappallo et al. [5] proposed a latent ranking method for the image popularity prediction solely based on the visual content. The proposed method was evaluated on several image datasets collected from micro-blogging and photo-sharing websites. Although huge success has been achieved by these approaches, limited efforts have thus far been dedicated to the problem of video popularity prediction, where multiple modalities coexist. Noting this gap, Trzcinski et al. [29] shifted from images to videos, and studied the problem of video popularity prediction utilizing both the visual clues and the early popularity pattern of the video once it is released. However, this approach suffered from two limitations. First, the proposed approach can only work on videos that have been published over a certain period. Second, the authors only used the traditional machine learning method—Support Vector Regression (SVR), which failed to make full use of the relationship among modalities. As a complement, we aim to timely predict the popularity of a given micro-video even before it get published by proposing a novel multi-modal learning scheme.

[7]Micro-videos are usually associated with certain textual data, such as video descriptions given by the video owners.

[8]The dataset can be accessible via http://acmmm2016.wix.com/micro-videos.

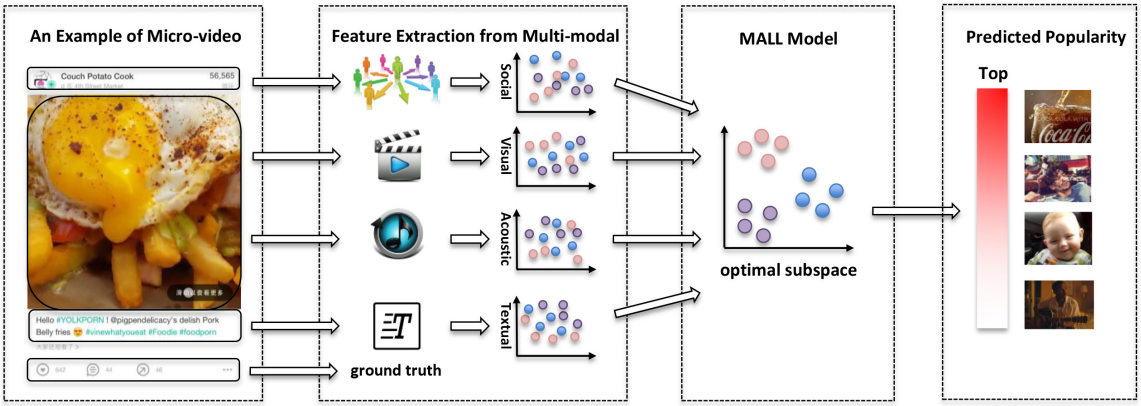


Figure 2: Micro-video popularity prediction via our proposed TMALL model.

2.2 Multi-View Learning

To deal with data containing multiple modalities, multi-view learning is a highly feasible paradigm. Multi-view learning is designed to improve the learning performance by introducing a function to model each view and jointly optimizing all functions. Existing work follows this line can be roughly classified into two categories: co-training and subspace learning. Co-training algorithms usually train separate learners on distinct views, which are then imposed to be consistent across views. Sindhwani et. al. [26] introduced a co-regularization framework for multi-view semi-supervised learning, as an extension of supervised regularization algorithms. Noticing that corruption may exist among different views, Christoudias et al. [7] proposed an approach for multi-view learning taking the view disagreement into consideration. In contrast, subspace learning approaches hold the general assumption that different views are generated from a latent view. Chaudhuri et al. [6] first employed canonical correlation analysis (CCA) to learn an efficient subspace, on which traditional machine learning algorithms can be applied. Gao et al. [9] later introduced a novel multi-view subspace clustering method, which is able to simultaneously perform clustering on the subspace of each view and guarantee the consistency among multiple views by a common clustering structure.

Overall, compelling success has been achieved by multi-view learning models on various problems, such as categorization [26, 28], clustering [6, 9] and multimedia retrieval [18, 19]. However, to the best of our knowledge, limited efforts have been dedicated to applying multi-view learning in the context of micro-video popularity prediction, which is the major concern of our work.

3. DATA COLLECTION

This section details the dataset setup, which covers the crawling strategy, and ground truth construction.

3.1 Crawling Strategy

Our micro-video dataset was crawled from one of the most prominent micro-video sharing social networks, Vine. Beside the historical uploaded micro-videos, Vine also archives users' profiles and their social connections.

In particular, we first randomly selected 10 active Vine users from Rankzoo[9], which provides the top 1,000 active users on Vine, as the seed users. We then adopted the breadth-first crawling strategy to expand the seed users by crawling their followers. Considering that these seed users may have millions of followers, we practically only retained the first 1,000 returned followers for each seed user to improve the crawling efficiency. After three layers of crawling, we harvested a densely connected user set consisting of 98,166 users as well as 120,324 following relationships among users. For each user, his/her brief profile was crawled, containing full name, description, location, follower count, followee count, like count, post count and loop count of all post videos. Besides, we also collected the timeline (the micro-video posting history, including the repostings from others.) of each user between July 1st and October 1st, 2015. Finally, we obtained 1.6 million video postings, including a total number of 303,242 unique micro-videos with a total duration of 499.8 hours.

[9]https://rankzoo.com.

3.2 Ground Truth Construction

We employed four popularity-related indicators, namely, the number of comments (n_comments), the number of likes (n_likes), the number of reposts (n_reposts), and the number of loops/views (n_loops) to measure the popularity of micro-videos. Figure 3 illustrates the proportion of micro-videos regarding each of the four indicators in our dataset. It is noticed that the distributions of them are different, and each of them measure one aspect of the popularity. In order to comprehensively and precisely measure the popularity of each micro-video, y_i, we linearly fuse all the four indicators:

$$y_i = \frac{(n\_reposts + n\_comments + n\_likes + n\_loops)}{4}. \quad (1)$$

4. OUR PROPOSED TMALL MODEL

4.1 Notation

We first declare several notations. We employ bold capital letters (e.g., $\mathbf{X}$) and bold lowercase letters (e.g., $\mathbf{x}$) to denote matrices and vectors, respectively. We use non-bold letters (e.g., x) to represent scalars, and Greek letters (e.g., β) as parameters. If not clarified, all vectors are in column form.

Without loss of generality, suppose we have N labeled samples and M unlabeled samples with $K \geqslant 2$ modalities. It is worth noting that the unlabeled samples also serve as testing samples. Z_k stands for the number of features generated from the k-th modality. Then the k-th modality can be represented as $\mathbf{X_k} \in \mathbb{R}^{(N+M)\times Z_k}$. The popularity of

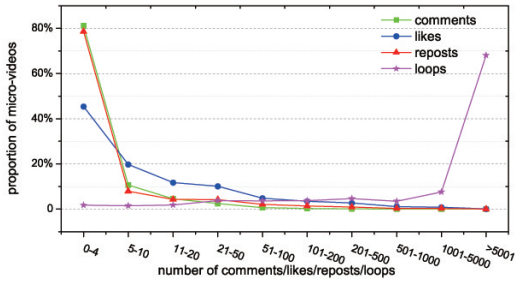


Figure 3: Distribution of the number of comments, likes, reposts and loops of micro-videos in our dataset.

all the videos are denoted by $\mathbf{y} = \{y_1, y_2, \cdots, y_N\}^T \in \mathbb{R}^N$. Let $\mathbf{f} = \{f_1, f_2, \cdots, f_N, f_{N+1}, f_{N+2}, \cdots, f_{N+M}\}^T \in \mathbb{R}^{N+M}$ stand for the predicted results regarding popularity for all samples, including the labeled and unlabeled ones. We aim to jointly learn the common space $\mathbf{X_0} \in \mathbb{R}^{(N+M)\times Z_0}$ shared by multiple modalities and the popularity for the M unlabeled micro-videos.

Our proposed model targets at reasoning from observed training micro-videos to testing ones. Such prediction belongs to transductive learning, in which both labeled samples as well as unlabeled samples are available for training. It hence obtains better performance. In contrast, inductive model is reasoning from observed training cases to general rules, which are then applied to the test cases.

4.2 Problem Formulation

It is apparent that different modalities may contribute distinctive and complementary information about micro-videos. For example, textual modality gives us hints about the topics of the given micro-video; acoustic and visual modalities may respectively convey location and situation of micro-videos, and user modality demonstrates the influence of the micro-video publisher. These clues jointly contribute to the popularity of a micro-video. Obviously, due to the noise and information insufficiency of each modality, it may be suboptimal to conduct learning directly from each single modality separately. In contrast, we assume that there exists an optimal latent space, in which micro-videos can be better described. Moreover, the optimal latent space should maintain the original intrinsic characteristics conveyed by multi-modalities of the given micro-videos. Therefore, we penalize the disagreement of the normalized Laplacian matrix between the latent space and each modality. In particular, we formalize this assumption as follows. Let $\mathbf{S}_k \in \mathbb{R}^{(N+M)\times(N+M)}$ be the similarity matrix[10], which is computed by the Gaussian similarity function as follows,

$$S_k(i,j) = \begin{cases} exp\Big(-\frac{\left\|\mathbf{x}_k^i - \mathbf{x}_k^j\right\|^2}{2\sigma_k^2}\Big) & , \quad \text{if } i \neq j; \\ 0 & , \quad \text{if } i = j. \end{cases} \tag{2}$$

where $\mathbf{x}_k^i$ and $\mathbf{x}_k^j$ are the micro-video pairs in the k-th modality space. Thereinto, the radius parameter σ_k is simply set as the median of the Euclidean distances over all video pairs in the k-th modality. We then derive the corresponding normalized Laplacian matrix as follows,

$$\mathbf{L(S_k)} = \mathbf{I} - \mathbf{D}_k^{-\frac{1}{2}}\mathbf{S}_k\mathbf{D}_k^{-\frac{1}{2}}, \tag{3}$$

[10]To facilitate the illustration, k ranges from 0 to K.

where $\mathbf{I}$ is a $(N+M)\times(N+M)$ identity matrix and $\mathbf{D}_k \in \mathbb{R}^{(N+M)\times(N+M)}$ is the diagonal degree matrix, whose (u,u)-th entry is the sum of the u-th row of $\mathbf{S}_k$. Since $S_k(i,j) > 0$, we can derive that $tr(\mathbf{L(S_k)}) > 0$. We thus can formulate the disagreement penalty between the latent space and the original modalities as,

$$\sum_{k=1}^{K} \| \frac{1}{tr(\mathbf{L(S_0)})}\mathbf{L(S_0)} - \frac{1}{tr(\mathbf{L(S_k)})}\mathbf{L(S_k)} \|_F^2, \tag{4}$$

where $tr(\mathbf{A})$ is the trace of matrix $\mathbf{A}$ and $\|\cdot\|_F$ denotes the Frobenius norm of matrix. In addition, inspired by [32], considering that similar micro-videos attempt to have similar popularity in the latent common space, we adopt the following regularizer,

$$\begin{aligned} &\frac{1}{2}\sum_{m=1}^{N+M}\sum_{n=1}^{N+M}\Big(\frac{f(\mathbf{x}_0^m)}{\sqrt{D_0(\mathbf{x}_0^m)}} - \frac{f(\mathbf{x}_0^n)}{\sqrt{D_0(\mathbf{x}_0^n)}}\Big)^2 S_0(m,n) \\ &= \mathbf{f}^T\mathbf{L(S_0)}\mathbf{f}. \end{aligned} \tag{5}$$

Based upon these formulations, we can define the loss function that measures the empirical error on the training samples. As reported in [22], the squared loss usually yields good performance as other complex ones. We thus adopt the squared loss in our algorithm for simplicity and efficiency. In particular, since we do not have the labels for testing samples, we only consider the squared loss regarding the N unlabeled samples to guarantee the learning performance. We ultimately reach our objective function as,

$$\begin{aligned} &\min_{\mathbf{f},\mathbf{L(S_0)}} \sum_{i=1}^{N}(y_i - f_i)^2 + \mu\mathbf{f}^T\mathbf{L(S_0)}\mathbf{f} \\ &+ \lambda\sum_{k=1}^{K} \| \frac{1}{tr(\mathbf{L(S_0)})}\mathbf{L(S_0)} - \frac{1}{tr(\mathbf{L(S_k)})}\mathbf{L(S_k)} \|_F^2, \end{aligned} \tag{6}$$

where λ and μ are both nonnegative regularization parameters. To be more specific, λ penalizes the disagreement among the latent space and modalities, and μ encourages that similar popularity will be assigned to similar micro-videos.

4.3 Alternative Optimization

To simplify the representation, we first define that,

$$\begin{cases} \tilde{\mathbf{L}} = \frac{1}{tr(\mathbf{L(S_0)})}\mathbf{L(S_0)}, \\ \tilde{\mathbf{L}}_\mathbf{k} = \frac{1}{tr(\mathbf{L(S_k)})}\mathbf{L(S_k)}. \end{cases} \tag{7}$$

Therefore, the objective function can be transformed to,

$$\begin{aligned} &\min_{\mathbf{f}} \sum_{i=1}^{N}(y_i - f_i)^2 + \lambda\sum_{k=1}^{K}\left\|\tilde{\mathbf{L}} - \tilde{\mathbf{L}}_\mathbf{k}\right\|_F^2 + \mu\mathbf{f}^T\tilde{\mathbf{L}}\mathbf{f}, \\ &\text{subject to} \quad tr(\mathbf{L(S_0)}) = 1. \end{aligned} \tag{8}$$

Furthermore, to optimize $\tilde{\mathbf{L}}$ more efficiently, inspired by the property that $tr(\tilde{\mathbf{L}}_\mathbf{k}) = 1$, we let,

$$\mathbf{L(S_0)} = \sum_{k=1}^{K}\beta_k\tilde{\mathbf{L}}_\mathbf{k}, \qquad \text{subject to} \qquad \sum_{k=1}^{K}\beta_k = 1. \tag{9}$$

Consequently, we have,

$$\tilde{\mathbf{L}} = \frac{1}{tr(\mathbf{L}(\mathbf{S_0}))}\mathbf{L}(\mathbf{S_0}) = \sum_{k=1}^{K}\beta_k\tilde{\mathbf{L_k}},$$
$$\text{subject to} \quad \sum_{k=1}^{K}\beta_k = 1. \tag{10}$$

Interestingly, we find that β_k can be treated as the co-related degree between the latent common space and each modality. It is worth noting that we do not impose the constraint of $\beta \geq 0$, since we want to keep both positive and negative co-relations. A positive coefficient indicates the positive correlation between the modality space and the latent common space, while a negative coefficient reflects the negative correlation, which may be due to the noisy data of the modality. The larger the β_k is, the higher correlation between the latent space and the k-th modality will be. In the end, the final objective function can be written as,

$$\min_{\mathbf{f},\boldsymbol{\beta}} \sum_{i=1}^{N}(y_i - f_i)^2 + \lambda\sum_{k=1}^{K}\left\|\sum_{i=1}^{K}\beta_i\tilde{\mathbf{L}}_\mathbf{i} - \tilde{\mathbf{L}}_\mathbf{k}\right\|_F^2 +$$
$$\mu\mathbf{f}^T\sum_{k=1}^{K}\beta_k\tilde{\mathbf{L}}_\mathbf{k}\mathbf{f} + \theta\left\|\boldsymbol{\beta}\right\|^2, \quad \text{subject to} \quad \mathbf{e}^T\boldsymbol{\beta} = 1, \tag{11}$$

where $\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_K]^T \in \mathbb{R}^K$ and $\mathbf{e} = [1, 1, \cdots, 1]^T \in \mathbb{R}^K$. θ is the regularization parameter, introduced to avoid the overfitting problem. We denote the objective function of Eqn.(11) as $\mathbf{\Gamma}$. We adopt the alternating optimization strategy to solve the two variables $\mathbf{f}$ and $\boldsymbol{\beta}$ in $\mathbf{\Gamma}$. In particular, we optimize one variable while fixing the other one in each iteration. We keep this iterative procedure until the $\mathbf{\Gamma}$ converges.

4.3.1 Computing β_j with $\mathbf{f}$ fixed

We first fix $\mathbf{f}$ and transform the objective function $\mathbf{\Gamma}$ as,

$$\min_{\boldsymbol{\beta}} \lambda\sum_{k=1}^{K}\sum_{t=1}^{N+M}\left\|\mathbf{M}^{(t)}\boldsymbol{\beta} - \tilde{\mathbf{l}}_\mathbf{k}^{(t)}\right\|_F^2 + \mu\mathbf{g}^T\boldsymbol{\beta} + \theta\left\|\boldsymbol{\beta}\right\|^2,$$
$$\text{subject to} \quad \mathbf{e}^T\boldsymbol{\beta} = 1, \tag{12}$$

where $\mathbf{g} = [\mathbf{f}^T\tilde{\mathbf{L}}_\mathbf{1}\mathbf{f}, \mathbf{f}^T\tilde{\mathbf{L}}_\mathbf{2}\mathbf{f}, \cdots, \mathbf{f}^T\tilde{\mathbf{L}}_\mathbf{K}\mathbf{f}]^T \in \mathbb{R}^K$, $\mathbf{M}^{(t)} = [\tilde{\mathbf{l}}_\mathbf{1}^{(t)}, \tilde{\mathbf{l}}_\mathbf{2}^{(t)}, \cdots, \tilde{\mathbf{l}}_\mathbf{K}^{(t)}] \in \mathbb{R}^{(N+M)\times K}$ and $\tilde{\mathbf{l}}_\mathbf{k}^{(t)} \in \mathbb{R}^{N+M}$ denotes the t-th column of $\tilde{\mathbf{L}}_\mathbf{k}$. For simplicity, we replace $\tilde{\mathbf{l}}_\mathbf{K}^{(t)}$ with $\tilde{\mathbf{l}}_\mathbf{k}^{(t)}\mathbf{e}^T\boldsymbol{\beta}$, as $\mathbf{e}^T\boldsymbol{\beta} = 1$. With the help of Lagrangian, $\mathbf{\Gamma}$ can be rewritten as follows.

$$\min_{\boldsymbol{\beta}}\lambda\sum_{k=1}^{K}\sum_{t=1}^{N+M}\left\|(\mathbf{M}^{(t)} - \tilde{\mathbf{l}}_\mathbf{k}^{(t)}\mathbf{e}^T)\boldsymbol{\beta}\right\|_F^2 + \mu\mathbf{g}^T\boldsymbol{\beta} + \delta(1 - \mathbf{e}^T\boldsymbol{\beta})$$
$$+ \theta\left\|\boldsymbol{\beta}\right\|^2, \tag{13}$$

where δ is a nonnegative Lagrange multiplier. Taking derivative of Eqn.(13) with respect to $\boldsymbol{\beta}$, we have,

$$\frac{\partial\mathbf{\Gamma}}{\partial\boldsymbol{\beta}} = \mathbf{H}\boldsymbol{\beta} + \mu\mathbf{g} - \delta\mathbf{e}, \tag{14}$$

where,

$$\mathbf{H} = 2\Big[\Big(\lambda\sum_{k=1}^{K}\sum_{t=1}^{N+M}(\mathbf{M}^{(t)} - \tilde{\mathbf{l}}_\mathbf{k}^{(t)}\mathbf{e}^T)^T(\mathbf{M}^{(t)} - \tilde{\mathbf{l}}_\mathbf{k}^{(t)}\mathbf{e}^T)\Big) + \theta\mathbf{I}\Big], \tag{15}$$

and $\mathbf{I}$ is a $K \times K$ identity matrix. Setting Eqn.(14) to zero, we have,

$$\boldsymbol{\beta} = \mathbf{H}^{-1}(\delta\mathbf{e} - \mu\mathbf{g}). \tag{16}$$

Substituting Eqn.(16) into $\mathbf{e}^T\boldsymbol{\beta} = 1$, we have,

$$\begin{cases}\delta &= \dfrac{1 + \mu\mathbf{e}^T\mathbf{H}^{-1}\mathbf{g}}{\mathbf{e}^T\mathbf{H}^{-1}\mathbf{e}},\\ \boldsymbol{\beta} &= \mathbf{H}^{-1}\Big[\dfrac{\mathbf{e} + \mu\mathbf{e}^T\mathbf{H}^{-1}\mathbf{g}\mathbf{e}}{\mathbf{e}^T\mathbf{H}^{-1}\mathbf{e}} - \mu\mathbf{g}\Big].\end{cases} \tag{17}$$

According to the definition of positive-definite matrix, $\mathbf{H}$ is always positive definite and hence invertible. Therefore, $\mathbf{H}^{-1}$ is also positive definite, which ensures $\mathbf{e}^T\mathbf{H}^{-1}\mathbf{e} > 0$.

4.3.2 Computing $\mathbf{f}$ with β_j fixed

With fixed β_j, taking derivative of $\mathbf{\Gamma}$ with respect to f_i, where $1 \leq i \leq N$, we have,

$$\frac{\partial\mathbf{\Gamma}}{\partial f_i} = 2(f_i - y_i) + 2\mu\sum_{j=1}^{N+M}\tilde{L}(i,j)f_j. \tag{18}$$

We then take derivative of the $\mathbf{\Gamma}$ with respect to f_i, where $N + 1 \leq i \leq N + M$. We reach,

$$\frac{\partial\mathbf{\Gamma}}{\partial f_i} = 2\mu\sum_{j=1}^{N+M}\tilde{L}(i,j)f_j. \tag{19}$$

In a vector-wise form, we restate the solution of $\mathbf{f}$ as follows,

$$\mathbf{f} = \mathbf{G}^{-1}\hat{\mathbf{y}}, \tag{20}$$

where $\mathbf{G} = \hat{\mathbf{I}} + \mu\sum_{k=1}^{K}\beta_k\tilde{\mathbf{L}}_\mathbf{k}$, $\hat{\mathbf{y}} = \{y_1, y_2, \cdots, y_N, 0, 0, \cdots, 0\}$ and $\hat{\mathbf{I}} \in \mathbb{R}^{(N+M)\times(N+M)}$ is defined as follows,

$$\hat{I}(i,j) = \begin{cases}1 & \text{if } i = j, \text{ and } 1 \leq i \leq N,\\ 0 & \text{otherwise.}\end{cases} \tag{21}$$

5. FEATURE EXTRACTION

It is apparent that both the publisher influence and content influence contribute to the popularity of UGCs. In particular, we characterized the publisher influence via the social modality, and the content influence via visual, acoustic and textual modalities. For content influence, we first examined the popular micro-videos in our dataset and propose three common characteristics of online micro-videos. For each characteristic, we then explained the insights, and transformed it into a set of features for video representation. Finally, we developed a rich set of popularity-oriented features from each modality.

5.1 Observations

Universal Appeal. The subjects of widely popular micro-videos cannot be something that can only be appreciated by a small group of people. Therefore, the topics and objects contained in micro-videos should be something common so that to be interpreted the same way across people and cultures. To capture this characteristic, we extracted Sentence2Vector feature from the textual modality and deep object feature from the visual one.

Emotional Content. People are naturally drawn to things that arouse their emotions. Micro-videos showing funny animals or lovely babies make people feel urge to

share them to express the same emotions. As a result, micro-videos that are highly emotional are more likely to be shared. Therefore, we extracted textual sentiment, visual sentiment features for each video as well as several acoustic features, which is widely used in emotion recognition in music [33].

High Quality and Aesthetic Design. When people share information on social networks, people are actually showing a little piece of themselves to their audience. Therefore, high quality and aesthetic design of the content, which could reflect the taste of people, is another important characteristic of popular micro-videos. Color histogram, aesthetic feature and visual quality feature were thus extracted to encode such characteristic. In addition, the acoustic features we extracted are frequently used in music modeling, which could help to detect music in the audio track of micro-videos [17].

5.2 Social Modality

It is intuitive that micro-videos posted by users, who has more followers or has a verified account, are more likely to be propagated, and thus tend to receive a higher number of audiences. To characterize the influence of micro-video publishers, we developed the following publisher-centric features for micro-videos.

- **Follower/Followee Count**. The number of followers and followees of the given micro-video publisher.
- **Loop Count**. The total number of loops received by all the posts of the publisher.
- **Post Count**. The number of posts generated by the publisher.
- **Twitter Verification**. A binary value indicating whether the publisher has been verified by Twitter[11].

5.3 Visual Modality

Due to the short-length of micro-videos, the visual content is usually highly related to a single theme, which enables us to only employ a few key frames to represent the whole micro-video. Inspired by this, we extracted the visual features from certain key frames. The mean pooling was performed across all the key frames to create a fixed-length vector representation of each micro-video.

5.3.1 Color Histogram

It has been found that most basic visual features (i.e., intensity and the mean value of different color channels in HSV space) except color histogram, have little correlation with popularity [15]. Color histogram has outstanding correlation due to the fact that striking colors tend to catch users' eyes. Therefore, we only extracted color histogram as the basic visual feature to characterize popular micro-videos. To reduce the size of color space, we grouped the color space into 50 distinct colors, which results in a 50-D vector for each frame.

5.3.2 Object Features

It has been studied that popular UGCs are strongly correlated with the objects contained in the videos [10]. We believe that the presence of certain objects affect micro-videos' popularity. For example, micro-videos with 'cute dogs' or 'beautiful girls' are more likely to be popular than those with 'desks' and 'stones'. We thus employed the deep convolutional neural networks (CNNs) [16], a powerful model for image recognition problems [35], to detect objects in micro-videos. Specifically, we applied the well-trained AlexNet DNN provided by the Caffe software package [13] to the input key frames. The output of the fc7 layer and the final $1,000$-way softmax layer in AlexNet is a probability distribution over the $1,000$ class labels predefined in ImageNet. We treat them as our feature representation of each frame. In the end, a mean pooling was performed over the frames to generate a single $4,096$-D vector and $1,000$-D vector for each micro-video.

5.3.3 SentiBank Features

We performed the sentiment analysis of the visual modality due to that the sentiment of UGCs has been proven to be strongly correlated with their popularity [10]. In particular, we extracted the visual sentiment features based on the deep CNNs model which was trained on the SentiBank dataset[4]. SentiBank contains $2,089$ concepts and each of them invokes specific sentiments such as 'cute girls' and 'funny animals'. Therefore, after mean pooling among keyframes, each micro-video is represented by a $2,089$-D vector.

5.3.4 Aesthetic Features

Aesthetic features are a set of handful selected features related to the principles of the nature and appreciation of beauty, which have been studied and found to be effective in popularity prediction [8]. Intuitively, micro-videos that are objectively aesthetic are more likely to be popular. We employed the released tool[12] [3] to extract the following aesthetic features: a) dark channel feature; b) luminosity feature; c) s3 sharpness; d) symmetry; e) low depth of field; f) white balance; g) colorfulness; h) color harmony, and i) eye sensitivity, at 3×3 grids over each key frame. We then calculated: a) normalized area of dominant object; and b) normalized distances of centroid of dominant objects with respect to four stress points at frame level. In the end, we obtained 149-D aesthetic features for each micro-video.

5.3.5 Visual Quality Assessment Features

It is important that the visual quality of popular contents are maintained at an acceptable level, given rising consumer expectations of the quality of multimedia content delivered to them [25]. In particular, we employed the released tool[13] to extract the micro-videos quality features based on the motion and spatio-temporal information, which have been proven to correlate highly with human visual judgments of quality. This results in a 46-D features.

5.4 Acoustic Modality

Acoustic modality usually works as an important complement to visual modality in many video-related tasks, such as video classification [34]. In fact, audio channels embedded in the micro-videos may also contribute to the popularity of micro-videos to a large extent. For example, the audio channel may indicate the quality of a given micro-video and convey rich background information about the emotion as well as the scene contained in the

[11] A Vine account can be verified by Twitter, if it is linked to a verified Twitter account.

[12] http://www.ee.columbia.edu/~subh/Software.php.

[13] http://live.ece.utexas.edu/.

micro-video, which significantly affects the popularity of a micro-video. The acoustic information is especially useful for the cases where the visual features could not carry enough information. Therefore, we adopted the following widely-used acoustic features, i.e., Mel-Frequency Cepstral Coefficients (MFCC) [17] and Audio-Six (i.e., Energy Entropy, Signal Energy, Zero Crossing Rate, Spectral Rolloff, Spectral Centroid, and Spectral Flux [33]). These features are frequently used in different audio-related tasks, such as emotion detection and music recognition. We finally obtained a 36-D acoustic feature vector for each micro-video.

5.5 Textual Modality

Micro-videos are usually associated with textual modality in the form of descriptions, such as "when Leo finally gets the Oscar" and "Puppy dog dreams", which may precisely summarize the micro-videos. Such summarization may depict the topics and sentiment information regarding the micro-videos, which has been proven to be of significance in online article popularity prediction [2].

5.5.1 Sentence2Vector

We found that the popular micro-videos are sometimes related to the topics of the textual descriptions. This observation propels us to conduct content analysis over the textual descriptions of micro-videos. Considering the short-length of descriptions, to perform content analysis, we employed the state-of-the-art textual feature extraction tool Sentence2Vector[14], which was developed on the basis of work embedding algorithm Word2Vector [21]. In this way, we extracted 100-D features for video descriptions.

5.5.2 Textual Sentiment

We also analyze the sentiments over texts, which has been proven to play an important role in popularity prediction [1]. With the help of the Sentiment Analysis tool in Stanford CoreNLP tools[15], we assigned each micro-video a sentiment score ranging from 0 to 4 and they correspond to *very negative*, *negative*, *neutral*, *positive*, and *very positive*, respectively.

6. EXPERIMENT

In this section, we conducted extensive experiments to comparatively verify our model.

6.1 Experiment Settings

The remaining experiments were conducted over a cluster of 50 servers equipped with Intel Xeon(2x) CPU E5-2620 v3 at 2.40 GHz on 64 GB RAM, 24 cores and 64-bit Linux operating system. Regarding the deep feature extraction, we deployed Caffe framework [13] on a server equipped with a NVIDIA Titan Z GPU. The experimental results reported in this paper were based on 10-fold cross-validation. In each round of the 10-fold cross-validation, we split our dataset into two chunks: 90% of the micro-videos were used for training, 10% were used for testing. We report performance in terms of normalised Mean Square Error (nMSE) [22] between the predicted popularity and the actual popularity. The nMSE is an estimator of the overall deviations between predicted and measured values. It is defined as,

$$nMSE = \frac{\sum_{i=1}(p_i - r_i)^2}{\sum_{i=1} r_i^2}, \tag{22}$$

where p_i is the predicted value and r_i is the target value in ground truth.

We have three key parameters as shown in Eqn.(8). The optimal values of these parameters were carefully tuned with the training data in each of the 10 fold. We employed the grid search strategy to obtain the optimal parameters between 10^{-5} to 10^2 with small but adaptive step sizes. In particular, the step sizes were 0.00001, 0.0001, 0.001, 0.01, 0.1, 1 and 10 for the range of [0.00001,0.0001], [0.0001,0.001], [0.001,0.01], [0.01,0.1], [0.1,1], [1,10] and [10,100], respectively. The parameters corresponding to the best nMSE were used to report the final results. For other compared systems, the procedures to tune the parameters are analogous to ensure the fair comparison. Considering one fold as an example, we observed that our model reached the optimal performance at $\lambda = 1$, $\mu = 0.01$ and $\theta = 100$.

6.2 On Model Comparison

To demonstrate the effectiveness of our proposed TMALL model, we carried out experiments with several state-of-the-art multi-view learning approaches:

- **Early_Fusion**. The first baseline concatenates the features extracted from the four modalities into a single joint feature vector, on which traditional machine learning models can be applied. In this work, we adopted the widely used regression model—SVR, and implemented it with the help of scikit-learn [23].
- **Late_Fusion**. The second baseline first separately predicts the popularity of micro-videos from each modality via SVR model, and then linearly integrates them to obtain the final results.
- **regMVMT**. The third baseline is the regularized multi-view learning model [37]. This model only regulates the relationships among different views within the original space.
- **MSNL**. The fourth one is the multiple social network learning (MSNL) model proposed in [27]. This model takes the source confidence and source consistency into consideration.
- **MvDA**. The fifth baseline is a multi-view discriminant analysis (MvDA) model [14], which aims to learn a single unified discriminant common space for multiple views by jointly optimizing multiple view-specific transforms, one for each view. The model exploits both the intra-view and inter-view correlations.

Table 1 shows the performance comparison among different models. From this table, we have the following observations: 1) TMALL outperforms the Early_Fusion and Late_Fusion. Regarding the Early_Fusion, features extracted from various sources may not fall into the same semantic space. Simply appending all features actually brings in a certain amount of noise and ambiguity. Besides, Early_Fusion may lead to the curse of dimensionality since the final feature vector would be of very high dimension. For the Late_Fusion, the fused result however might not be reasonably accurate due to two reasons. First, a single modality might not be sufficiently descriptive to represent

[14] https://github.com/klb3713/sentence2vec.
[15] http://stanfordnlp.github.io/CoreNLP/.

Table 1: Performance comparison between our proposed TMALL model and several state-of-the-art baselines in terms of nMSE.

Methods	nMSE	p-value
Early Fusion	59.931 ± 41.09	9.91E-04
Late Fusion	8.461 ± 5.34	3.25E-03
regMVMT	1.058 ± 0.05	1.88E-03
MSNL	1.098 ± 0.13	1.42E-02
MvDA	0.982 ± 7.00E-03	9.91E-04
TMALL	0.979 ± 9.42E-03	–

the complex semantics of the videos. Separate results would be thus suboptimal and the integration may not result in a desired outcome. Second, it is labor-intensive to tune the fusion weights for different modalities. Even worse, the optimal parameters for one application cannot be directly applied to another one. 2) TMALL achieves better performance, as compared with regMVMT and MSNL. This could be explained that linking different modalities via a unified latent space is better than imposing disagreement penalty directly over original spaces. 3) The less satisfactory performance of MvDA indicates that it is necessary to explore the consistency among different modalities when building the latent space. And 4) as compared to the multi-view learning baselines, such as regMVMT, MSNL, and MvDA, our model stably demonstrates its advantage. This signals that the proposed transductive models can achieve higher performance than inductive models under the same experimental settings. This can be explained by the fact that TMALL leverages the knowledge of testing samples.

Moreover, we performed the paired t-test between TMALL and each baseline on 10-fold cross validation. We found that all the p-values are much smaller than 0.05, which shows that the performance improvements of our proposed model over other baselines are statistically significant.

6.3 On Modality Comparison

To verify the effectiveness of multi-modal integration, we also conducted experiments over different modality combinations of the four modalities. Table 2 summarizes the multi-modal analysis and the paired t-test results. It is obvious that the more modalities we incorporated, the better performance we can obtain. This implies the complementary relationships rather than mutual conflicting relationships among the different modalities. Moreover, we found that removing features from any of these four modalities suffers from a decrease in performance. In a sense, this is consensus with the old saying "two heads are better than one". Additionally, as the performance obtained from different combinations are not the same, this validates that incorporating β which controls the confidence of different modalities is reasonable. Interestingly, we observed that the combination without social modality obtains the worst result which indicates that the social modality plays a pivotal role in micro-video propagation, as compared to visual, textual or acoustic modality. This also validates that the features developed from social modality are much discriminative, even though they are with low-dimensions. On the other hand, the textual modality contributes the least among all modalities, as the performance of our model without textual modality still achieves good performance. This may be caused by the sparse textual description, which is usually given in one short sentence.

Table 2: Performance comparison among different modality combinations with respect to nMSE. We denote T, V, A and S as textual, visual, acoustic and social modality, respectively.

View combinations	nMSE	p-value
T+V+A	0.996 ± 4.20E-03	2.62E-05
T+A+S	0.982 ± 4.27E-03	2.59E-05
T+V+S	0.982 ± 4.13E-03	3.05E-04
V+A+S	0.981 ± 5.16E-03	2.16E-05
T+V+A+S	0.979 ± 9.42E-03	–

Table 3: Performance comparison among different visual features with respect to nMSE.

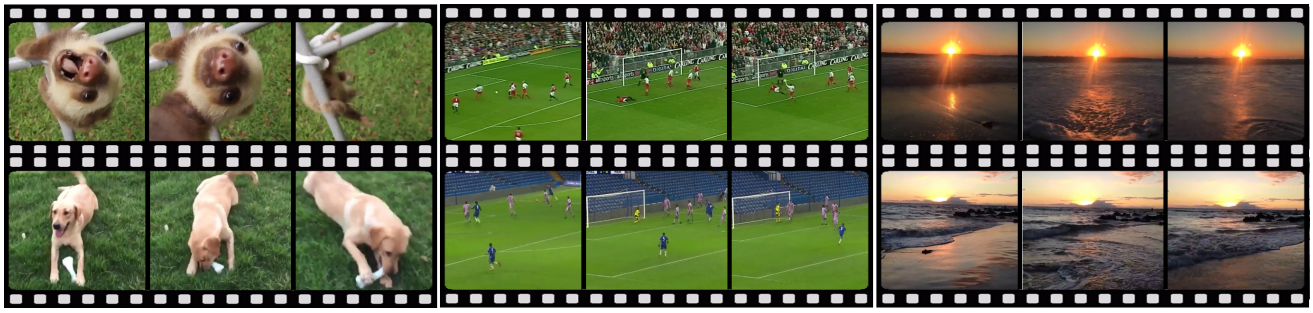
Features	nMSE	p-value
Color Histogram	0.996 ± 6.88E-03	1.94E-04
Object Feature	0.994 ± 6.71E-03	2.47E-04
Visual Sentiment	0.994 ± 6.72E-03	2.49E-04
Aesthetic Feature	0.984 ± 6.95E-03	4.44E-01
ALL	0.979 ± 9.42E-03	–

6.4 On Visual Feature Comparison

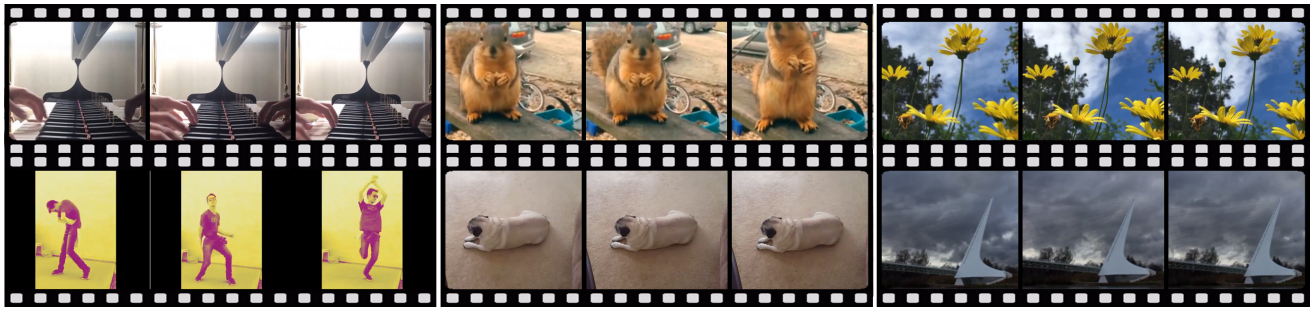
To further examine the discriminative visual features we extracted, we conducted experiments over different kinds of visual features using TMALL. We also performed significant test to validate the advantage of combining multiple features. Table 3 comparatively shows the performance of TMALL in terms of different visual feature configurations. It can be seen that the object, visual sentiment and aesthetic features achieve similar improvement in performance, as compared to color histogram features. This reveals that micro-videos' popularity is better reflected by their content, sentiment and design, including what objects they contain, which emotion they convey and what design standards they follow. This is highly consistent with our oberservations and also implies that micro-videos which aim to gain high popularity need to be well designed and considered more from the visual content.

6.5 Illustrative Examples

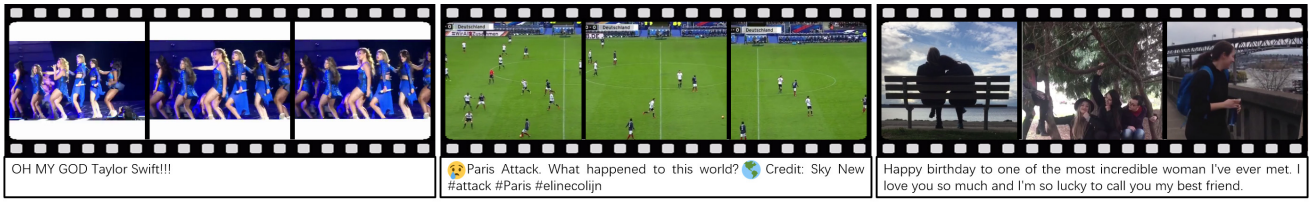
To gain the insights of the influential factors in the task of popularity prediction of micro-videos, we comparatively illustrate a few representative examples in Figure 4. From this figure, we have the following observations: 1) Figure 4(a) shows three micro-video pairs. Each of the three micro-video pairs describes the similar semantics, i.e., animals, football game and sunset, respectively, but they were published by different users. The publishers of the videos in top row are much more famous than those of the bottom. We found that the corresponding popularity of micro-videos in the second row are much lower than those in the first row, although they have no significant difference from the perspective of video contents, which clearly justifies the importance of social modality. 2) Figure 4(b) illustrates three micro-video pairs, where each pair of micro-videos were published by the same user. However, the micro-videos in the first row achieve much higher popularity than those in the second row, which demonstrates that the contents of micro-videos also contribute to their popularity. In



(a) Illustration of three micro-video pairs, and each pair was published by two distinct users. The publishers of the videos in top row are much more famous than those of the bottom.



(b) Illustration of three micro-video pairs, and each pair was published by the same user. The videos in the first row are much more acoustically comfortable, visually joyful, and aesthetically beautiful than those in the second row.



(c) Illustration of three popular micro-videos with different textual descriptions, which contains superstar names, hot events, and detail information, respectively.

Figure 4: Comparative illustration of video examples. They respectively justify the importance of social, acoustic as well as visual, and textual modalities. We use three key frames to represent each video.

particular, the comparisons in Figure 4(b), from left to right, are i) the existence of 'skillful pianolude' compared with 'noisy dance music', ii) 'funny animals' compared with 'motionless dog', and iii) 'beautiful flowers' compared with 'gloomy sky'. These examples indicate the necessity of developing acoustic features, visual sentiment and visual aesthetic features for the task of micro-video popularity. And 3) Figure 4(c) shows a group of micro-videos, whose textual descriptions contain either superstar names, hot hashtags, or informative descriptions. These micro-videos received a lot of loops, comments, likes and reposts. These examples thus reflect the value of textual modality.

6.6 Complexity Analysis

To theoretically analyze the computational cost of our proposed TMALL model, we first compute the complexity in the construction of $\mathbf{H}$ and $\mathbf{g}$, as well as the inverse of matrices $\mathbf{H}$ and $\mathbf{G}$. The construction of $\mathbf{H}$ has the time complexity of $\mathbf{O}(K^2(N+M))$. Fortunately, $\mathbf{H}$ keeps the same in each iteration, and thus can be computed by offline. The computation of $\mathbf{g}$ needs the time cost $\mathbf{O}(K(N+M)^2)$. In addition, computing the inverse of $\mathbf{H}$ and $\mathbf{G}$ has the complexity of $\mathbf{O}(K^3)$ and $\mathbf{O}((N+M)^3)$, respectively. The computation cost of $\boldsymbol{\beta}$ in Eqn.(17) is $\mathbf{O}(K^2)$. Therefore, the speed bottleneck lies in the computation of the inverse of $\mathbf{G}$. In practice, the proposed TMALL model converges very fast, which on average takes less than 10 iterations. Overall, the learning process over $9,720$ micro-videos can be accomplished within 50 seconds.

7. CONCLUSION AND FUTURE WORK

This paper presents a novel transductive multi-modal learning method (TMALL), to predict the popularity of micro-videos. In particular, TMALL works by learning an optimal latent common space from multi-modalities of the given micro-videos, in which the popularity of micro-videos are much more distinguishable. The latent common space is capable of unifying and preserving information from different modalities, and it helps to alleviate the modality limitation problem. To verify our model, we built a benchmark dataset and extracted a rich set of popularity-oriented features to characterize micro-videos from multiple perspectives. By conducting extensive experiments, we draw the following conclusions: 1) the optimal latent common space exists and works; 2) the more modalities we incorporate to learn the common space, the more

discriminant it is; and 3) the features extracted to describe the social and content influence are representative. As a side research contribution, we have released the dataset, codes and parameters to facilitate other researchers. In the future, we plan to incorporate the cross-domain knowledge, such as the hot topics on Twitter, to enhance the performance of popularity prediction.

8. ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IRC@SG Funding Initiative.

9. REFERENCES

[1] J. Berger. Arousal increases social transmission of information. *Psychological science*, 22(7):891–893, 2011.

[2] J. Berger and K. L. Milkman. What makes online content viral? *Journal of Marketing Research*, 49(2):192–205, 2012.

[3] S. Bhattacharya, B. Nojavanasghari, T. Chen, D. Liu, S.-F. Chang, and M. Shah. Towards a comprehensive computational model foraesthetic assessment of videos. In *Proceedings of the ACM Multimedia Conference*, pages 361–364. ACM, 2013.

[4] D. Borth, R. Ji, T. Chen, T. M. Breuel, and S. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the ACM Multimedia Conference*, pages 223–232. ACM, 2013.

[5] S. Cappallo, T. Mensink, and C. G. M. Snoek. Latent factors of visual popularity prediction. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*, pages 195–202. ACM, 2015.

[6] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the International Conference on Machine Learning*, pages 129–136. ACM, 2009.

[7] C. M. Christoudias, R. Urtasun, and T. Darrell. Multi-view learning in the presence of view disagreement. *CoRR*, abs/1206.3242, 2012.

[8] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1657–1664. IEEE, 2011.

[9] H. Gao, F. Nie, X. Li, and H. Huang. Multi-view subspace clustering. In *IEEE International Conference on Computer Vision*, pages 4238–4246. IEEE, 2015.

[10] F. Gelli, T. Uricchio, M. Bertini, A. D. Bimbo, and S. Chang. Image popularity prediction in social media using sentiment and context features. In *Proceedings of the ACM Multimedia Conference*, pages 907–910. ACM, 2015.

[11] X. He, M. Gao, M. Kan, Y. Liu, and K. Sugiyama. Predicting the popularity of web 2.0 items based on user comments. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 233–242. ACM, 2014.

[12] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *Proceedings of the ACM International Conference on World Wide Web*, pages 57–58. ACM, 2011.

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM Multimedia Conference*, pages 675–678. ACM, 2014.

[14] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):188–194, 2016.

[15] A. Khosla, A. D. Sarma, and R. Hamid. What makes an image popular? In *Proceedings of the ACM International Conference on World Wide Web*, pages 867–876, 2014.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 1106–1114. NIPS Foundation, 2012.

[17] Z. Li, J. Wang, J. Cai, Z. Duan, H. Wang, and Y. Wang. Non-reference audio quality assessment for online live music recordings. In *Proceedings of the ACM Multimedia Conference*, pages 63–72. ACM, 2013.

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

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

[20] P. J. McParlane, Y. Moshfeghi, and J. M. Jose. "nobody comes here anymore, it's too crowded"; predicting image popularity on flickr. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, page 385. ACM, 2014.

[21] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 3111–3119. NIPS Foundation, 2013.

[22] L. Nie, L. Zhang, Y. Yang, M. Wang, R. Hong, and T.-S. Chua. Beyond doctors: future health prediction from multimedia and multimodal observations. In *Proceedings of the ACM Multimedia Conference*, pages 591–600. ACM, 2015.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[24] M. Redi, N. O'Hare, R. Schifanella, M. Trevisiol, and A. Jaimes. 6 seconds of sound and vision: Creativity in micro-videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4272–4279. IEEE, 2014.

[25] M. A. Saad, A. C. Bovik, and C. Charrier. Blind prediction of natural video quality. *IEEE Transactions on Image Processing*, 23(3):1352–1365, 2014.

[26] V. Sindhwani, P. Niyogi, and M. Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of the International Conference on Machine Learning*, pages 74–79. ACM, 2005.

[27] X. Song, L. Nie, L. Zhang, M. Akbari, and T. Chua. Multiple social network learning and its application in volunteerism tendency prediction. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 213–222, 2015.

[28] X. Song, L. Nie, L. Zhang, M. Liu, and T. Chua. Interest inference via structure-constrained multi-source multi-task learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2371–2377. AAAI Press, 2015.

[29] T. Trzcinski and P. Rokita. Predicting popularity of online videos using support vector regression. *CoRR*, abs/1510.06223, 2015.

[30] M. Wang, X. Hua, R. Hong, J. Tang, G. Qi, and Y. Song. Unified video annotation via multigraph learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(5):733–746, 2009.

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

[32] M. Wang, X. Liu, and X. Wu. Visual classification by $_1$-hypergraph modeling. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2564–2574, 2015.

[33] B. Wu, E. Zhong, A. Horner, and Q. Yang. Music emotion recognition by multi-label multi-layer multi-instance multi-view learning. In *Proceedings of the ACM Multimedia Conference*, pages 117–126. ACM, 2014.

[34] Z. Wu, Y. Jiang, J. Wang, J. Pu, and X. Xue. Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In *Proceedings of the ACM Multimedia Conference*, pages 167–176. ACM, 2014.

[35] H. Zhang, X. Shang, W. Yang, H. Xu, H. Luan, and T. Chua. Online collaborative learning for open-vocabulary visual classifiers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016.

[36] H. Zhang, M. Wang, R. Hong, and T. Chua. Play and rewind: optimizing binary representations of videos by self-supervised temporal hashing. In *Proceedings of the ACM Multimedia Conference*. ACM, 2016.

[37] J. Zhang and J. Huan. Inductive multi-task learning with multiple view data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 543–551. ACM, 2012.