Multi-View Active Learning for Video Recommendation

Jia-Jia Cai[1] , **Jun Tang**[2] , **Qing-Guo Chen**[2] , **Yao Hu**[2] , **Xiaobo Wang**[2] and **Sheng-Jun Huang**[1]
[1]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, China
[2]YouKu Cognitive and Intelligent Lab, Alibaba Group, Hangzhou, China
{caijia, huangsj}@nuaa.edu.cn, {donald.tj, qingguo.cqg, yaoohu, yongshu.wxb}@alibaba-inc.com

Abstract

On many video websites, the recommendation is implemented as a prediction problem of video-user pairs, where the videos are represented by text features extracted from the metadata. However, the metadata is manually annotated by users and is usually missing for online videos. To train an effective recommender system with lower annotation cost, we propose an active learning approach to fully exploit the visual view of videos, while querying as few annotations as possible from the text view. On one hand, a joint model is proposed to learn the mapping from visual view to text view by simultaneously aligning the two views and minimizing the classification loss. On the other hand, a novel strategy based on prediction inconsistency and watching frequency is proposed to actively select the most important videos for metadata querying. Experiments on both classification datasets and real video recommendation tasks validate that the proposed approach can significantly reduce the annotation cost.

1 Introduction

Video service providers usually have a huge video corpus and billions of users. Video recommender system plays an important role in helping users discover content that they are interested in among so many videos. Traditional video recommender systems are mostly based on collaborative filtering [Baluja *et al.*, 2008], which provides users with videos that have been watched by other users whose preferences are similar. However, collaborative filtering is less suitable for online videos due to the cold-start problem. Recently, content-based video recommendation [Cui *et al.*, 2014] has attracted more research interests, where a classification model is trained to predict whether a user is interested in a video or not. After the prediction, the system will produce a list of videos for a specific user according to the predicted scores. To train the recommendation model, a large set of video-user pairs are collected and labeled, where the class label indicates whether the user watched the video or not.

Figure 1 shows an example of the video-user pairs. Each video can be represented by plentiful features from multiple views. The visual view consists of visual features extracted from video frames, and features in the text view are mainly extracted from the metadata of the video, including the title, tags, comments, etc. The user is represented with profile features including watching history, search token, demographics and so on. Finally, there is a class label recording the user action on this video. The example will be labeled as positive if the user watched this video, and negative otherwise.

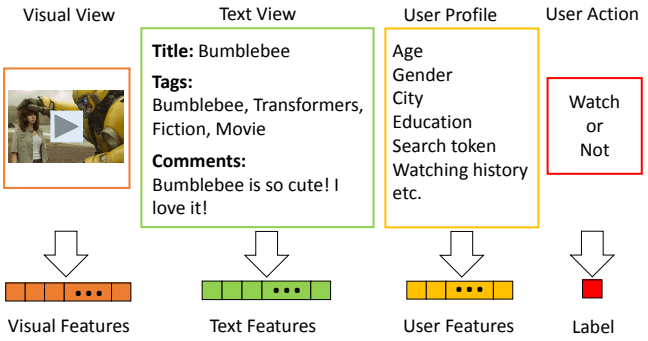


Figure 1: An example of video-user pairs with multi-view representations in the video recommendation task. The text features are usually unavailable for online videos.

It is well known that there is a semantic gap between the low-level visual features and the high-level class labels [Lee *et al.*, 2018]. Instead, the text features extracted from metadata describe the semantic information of the video and are expected to provide more effective representation. Therefore, in many real recommender systems, the text view is employed to represent the videos in the training data, while leaving the visual view less exploited. Unfortunately, the metadata is unavailable or incomplete for so many online videos [Davidson *et al.*, 2010]. On the Internet, a vast amount of user-generated-content (UGC) accounts for a large proportion of the video corpus. Users may simply upload videos, but are not willing to manually annotate their metadata in detail. This will result in serious missing of the text view, and further lead to the lack of training data for recommender systems. If we want to collect sufficient high-quality training data, the metadata of a huge number of videos needs to be annotated or refined by human experts, which will lead to high annotation cost since this process is time-consuming and laborious.

To overcome this challenge, we propose a Multi-View Active Learning (MVAL) approach to reduce the annotation cost by selectively querying the metadata for most informative examples. In traditional active learning, existing methods focus on querying the class labels to reduce labeling costs. Different from that, in video recommendation tasks, class labels can be easily obtained by recording the watching actions of users. So all examples are naturally labeled. Instead, the text view of videos is seriously missing and expensive to obtain. We thus propose a novel approach to query text features. Specifically, we first learn a mapping function from the visual view to the text view (V2T) to fully exploit the visual features. By jointly minimizing the reconstruction error and the classification loss, these two views are incorporated in a supervised way. Then based on this trained V2T mapping, a novel criterion is proposed to simultaneously consider the prediction inconsistency and watching frequency for selecting the most informative examples. By querying the text view of videos with large inconsistency and high frequency, the model performance can be effectively boosted with less annotation.

The main contributions of this work are summarized as follows:

- A multi-view active learning framework for video recommendation is proposed. It actively queries the missing view of selected examples, which is a novel setting different from existing studies that focus on querying the class labels.
- An effective algorithm MVAL is proposed to reduce the annotation cost, which on one hand fully exploits the visual view with supervised information, and on the other hand, integrate the prediction inconsistency with frequency to select the most informative examples.
- Experiments are performed on classification benchmark datasets as well as real video recommendation tasks. The results demonstrate that the proposed approach can achieve effective performance with significant lower annotation cost.

The rest of this paper is organized as follows. We review related work in Section 2 and introduce the proposed method in Section 3. Section 4 reports the experiments, followed by the conclusion in Section 5.

2 Related Work

Recommendation algorithms can be generally categorized into two groups: content-based filtering and collaborative filtering [Chen *et al.*, 2018]. Content-based method recommends items whose content may meet the users' interests [Cui *et al.*, 2014] while collaborative filtering recommends items that other similar users prefer [Baluja *et al.*, 2008]. The performance of content-based video recommendation heavily relies on user-annotated metadata, i.e., the text view of videos [Siersdorfer *et al.*, 2009]. To overcome the serious missing of metadata, some studies try to apply computer vision techniques for automatic video annotation [Siersdorfer *et al.*, 2009]. However, these methods can only predict some object-level labels instead of a detailed description and the results are far from satisfactory due to the lack of training data compared with the large variations of the video semantic concepts[Song *et al.*, 2005], as well as the large gap between low-level features and high-level semantics [Lee *et al.*, 2018].

Active learning is a primary approach to reduce labeling cost [Settles, 2012]. It progressively queries the labels of the most useful examples and tries to train an effective model with fewer queries. Traditional studies focused on designing a selection criterion such that selected instances can improve the model performance maximally [Huang *et al.*, 2014]. Although many effective criteria are proposed [Tong and Koller, 2001; Huang and Zhou, 2013], they are designed for traditional classification tasks, and cannot be applied to video recommendation tasks, where the metadata instead of class label need to be queried. Several studies try to apply active learning methods for recommendation tasks [Rubens *et al.*, 2015]. However, different from our work, previous methods focus on querying intersections between users and items to address notorious cold-start problem [Houlsby *et al.*, 2014].

The multi-view learning framework is firstly formalized by [Blum and Mitchell, 1998]. Co-Testing is a multi-view active learning method, which queries the labels of unlabeled examples on which the views predict a different label [Muslea *et al.*, 2006]. Theoretical analysis proves that the sample complexity of multi-view active learning has an exponential improvement [Wang and Zhou, 2008]. To the best of our knowledge, there is no active learning study for querying the missing features in multi-view tasks.

3 The Proposed Approach

Let $R = \{(\boldsymbol{v}_i, \boldsymbol{u}_j, y_{ij})\}_{i=1\cdots n, j=1\cdots m}$ be a set of user action records on videos, where $\boldsymbol{v}_i$ denotes the i-th video, and $\boldsymbol{u}_j$ denotes the j-th user. $y_{ij} = 1$ if the user $\boldsymbol{u}_j$ have watched the video $\boldsymbol{v}_i$, and otherwise $y_{ij} = 0$. As shown in Figure 1, a video can be represented by multiple views. We denote by $\boldsymbol{v}_i^V$ the visual view feature vector of the i-th video, and $\boldsymbol{v}_i^T$ its text view. In content-based recommender system, a classification model f is trained based on the dataset R, by taking $\boldsymbol{x}_{ij} = [\boldsymbol{v}_i^T, \boldsymbol{u}_j]$ as an input instance and y_{ij} as its output label. After training the model, given a new instance $\boldsymbol{x} = [\boldsymbol{v}^T, \boldsymbol{u}]$, the model prediction $f(\boldsymbol{x})$ estimates how likely the user $\boldsymbol{u}$ is interested in watching the video $\boldsymbol{v}$.

As discussed previously, the text view $\boldsymbol{v}_i^T$ of a video is extracted from the metadata, which is usually missing because it requires users to manually annotate. We denote by A a small set of annotated videos, where the visual and text view are both available for each video. Similarly, we have a large set U consists of unannotated videos, where only the visual view is available, while text features are missing. We denote by R^A the user action records on videos in A and R^U the user action records on videos in U respectively. In real recommender systems, the model is trained based on the text view, and leaves the visual view less exploited because it can not well reflect the semantic information of the video. Obviously, if we simply train the model f based on the annotated set A, the limited number of training examples will lead to poor performance. We thus propose to actively select the most important examples from the unannotated set U to query their text view, and then add them into the training set. In this way, the model

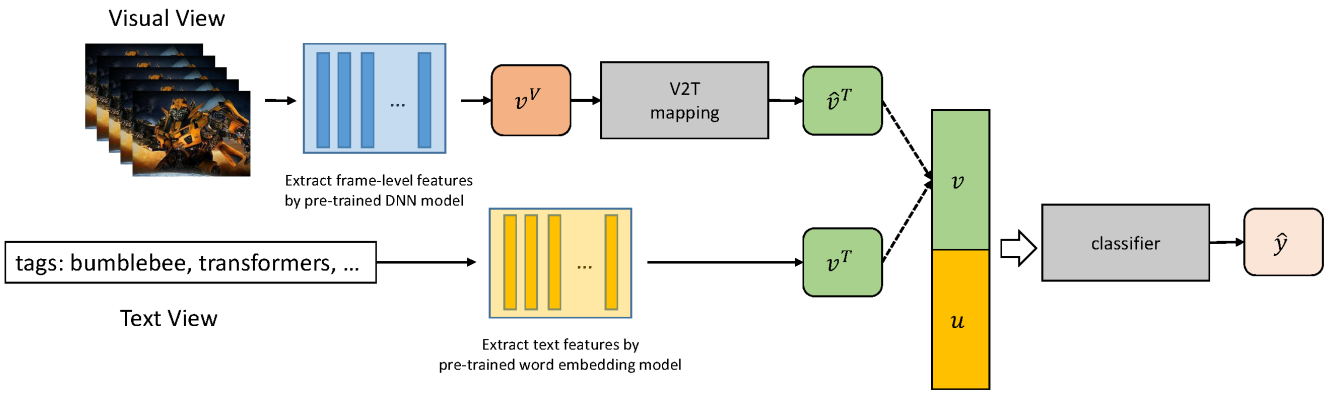


Figure 2: The framework for learning the V2T mapping.

f can be boosted effectively with less annotation cost. The key task is to decide which videos in U should be selected to query their text features.

It is well accepted in active learning that querying the most informative examples will contribute most to the classification model. The informativeness is typically estimated based on the prediction of the current model. For example, the uncertain examples with lower prediction confidence are expected to be more informative. Obviously, this strategy cannot be applied to our setting. Note that our model f is trained based on the text view, it is thus even impossible to get the prediction for the videos in U because their text view is missing. We thus need to exploit the visual view to make a bridge for estimating the informativeness of the unannotated videos.

In the rest of this section, we will first propose a visual to text (V2T) mapping to exploit the visual view, and then present an active learning criterion to select the most informative videos for text view annotation.

3.1 Visual to Text Mapping

To exploit the visual features for videos without text view, we learn a V2T mapping function e to connect the two views. Based on this mapping function, we can get the outputs as the approximated text features of unannotated videos. These approximated features, of course, are not effective enough to directly train the recommender classifier f, however, they can act as a bridge to provide the possibility of estimating the informativeness for unannotated videos.

A straightforward way to learn the mapping function e is to minimize the reconstruction error between the ground-truth text features and the transformed features. Noticing that we have an initial small set of annotated videos A with both views, then the objective function can be formalized as:

$$\min_{e} \sum_{\boldsymbol{v}_i \in A} \ell_1 \left(\boldsymbol{v}_i^T, e\left(\boldsymbol{v}_i^V\right) \right), \tag{1}$$

where ℓ_1 is the loss function to calculate the error between transformed text features and ground-truth features.

Obviously, it is less likely that the mapping function e can be well trained as in Eq. (1), given that there is only a small number of annotated videos in A. Noticing that the class labels are naturally available for all video-user pairs and inspired by [Huang *et al.*, 2018], we utilize this supervised information to guide the training of e.

Figure 2 illustrates the basic idea of learning V2T mapping. $\hat{\boldsymbol{v}}^T = e(\boldsymbol{v}_i^V)$ denotes the transformed text features by the V2T mapping. For videos in A, the ground-truth text features $\boldsymbol{v}^T$ are available, and thus e can be optimized by minimizing Eq. (1). Furthermore, the supervised information included in video-user action labels can be employed to guide the V2T mapping. As introduced previously, the classification function f is a mapping from text feature space to the user action label space and thus can be utilized to inversely transfer the user action label information for text feature optimization. For example, given a record with its visual features $\boldsymbol{v}_i^V$, user features $\boldsymbol{u}_j$ and action label y_{ij}, we denote by $\hat{\boldsymbol{v}}_i^T$ its transformed text feature vector by V2T mapping. Assuming the classifier f is reliable, if the prediction $\hat{y}_{ij} = f(\hat{\boldsymbol{v}}_i^T, \boldsymbol{u}_j)$ is far away from the ground-truth label y_{ij}, then it is more likely that the feature vector $\hat{\boldsymbol{v}}^T$ is not accurately recovered. Based on this motivation, we propose to minimize the empirical classification error for optimizing the V2T mapping. Formally, we have the following objective function:

$$\begin{aligned} \min_{e,f} & \sum_{(\boldsymbol{v}_i, \boldsymbol{u}_j, y_{ij}) \in R^A} \ell_2 \left(f(\boldsymbol{v}_i^T, \boldsymbol{u}_j), y_{ij} \right) + \\ & \sum_{(\boldsymbol{v}_i, \boldsymbol{u}_j, y_{ij}) \in R^U} \ell_2 \left(f(e(\boldsymbol{v}_i^V), \boldsymbol{u}_j), y_{ij} \right), \end{aligned} \tag{2}$$

where ℓ_2 is the loss function to calculate the classification error. Note here both the V2T mapping e and the classification model f are to be optimized. It can be observed that videos without text view can also be used, and thus significantly expand the training set for learning the mapping function.

Finally, we integrate the two objective functions into one

framework as in Eq. (3).

$$\min_{e,f} \frac{\lambda}{|A|} \sum_{\boldsymbol{v}_i \in A} \ell_1\left(\boldsymbol{v}_i^T, e\left(\boldsymbol{v}_i^V\right)\right) + \frac{1}{|R^A|} \sum_{(\boldsymbol{v}_i, \boldsymbol{u}_j, y_{ij}) \in R^A} \ell_2\left(f(\boldsymbol{v}_i^T, \boldsymbol{u}_j), y_{ij}\right) + \frac{1}{|R^U|} \sum_{(\boldsymbol{v}_i, \boldsymbol{u}_j, y_{ij}) \in R^U} \ell_2\left(f(e(\boldsymbol{v}_i^V), \boldsymbol{u}_j), y_{ij}\right), \tag{3}$$

where λ is a trade-off parameter between transformation loss and classification loss. For the videos in A that have the visual view and text view, they contribute to the transformation loss and the classification loss both. For the other videos, they are utilized along with the user action records to minimize the classification loss. In our implementation, we simply employ the quadratic loss $l(y, \hat{y}) = (y - \hat{y})^2$ for both ℓ_1 and ℓ_2.

In the training phase, the V2T mapping e and classification model f are unified into one neural network, and optimized via back propagation.

3.2 Active Selection

In this subsection, we discuss how to actively select the most informative videos from unannotated set U to query their ground-truth text features. With the mapping function e introduced in the previous subsection, we can get the transformed text features $\hat{\boldsymbol{v}}_i^T$ for each video $\boldsymbol{v}_i \in U$, and further can get the predicted label $\hat{y}_{ij} = f(\hat{\boldsymbol{v}}_i^T, \boldsymbol{u}_j)$ with regard to user u_j. In our setting, video-user labels are all recorded in advance. If the predictions of a video are more inconsistent with the ground-truth labels (i.e., $\hat{y}_{ij} \neq y_{ij}$), then the video is considered to be more uncertain, and may contribute more to the current model after annotation. Specifically, we define the error rate of the predictions over $\boldsymbol{v}_i$ as:

$$err(\boldsymbol{v}_i) = \frac{\sum_{(\boldsymbol{v}_i, \boldsymbol{u}_j, y_{ij}) \in R} I(\hat{y}_{ij} \neq y_{ij})}{n_i}, \tag{4}$$

where $I(\cdot)$ is the indicator function, and n_i denotes the number of records involves $\boldsymbol{v}_i$.

Noticing that each video is paired with different users to form multiple records, thus selecting more popular videos will influence more training samples, which will also contribute more to the model improvement. Motivated by this phenomenon, we further define the score S_i to estimate the informativeness of the video $\boldsymbol{v}_i$ as in Eq. 5:

$$S_i = freq_i * err(\boldsymbol{v}_i), \tag{5}$$

where $freq_i$ is the frequency of occurrence for $\boldsymbol{v}_i$. After evaluating the informativeness for all the videos in U, the ones with the highest S_i scores are selected to query their metadata. After that, the text features are extracted and the corresponding examples will be added into the training set to update the recommendation model. This process is repeated until the performance is satisfied or the query budget has run out. The pseudo code of the proposed multi-view active learning method for video recommendation is summarized in Algorithm 1.

Algorithm 1 The MVAL algorithm

Input:
  R: the data set of video-user action records
  A: the annotated video set
  U: the unannotated video set
Process:
  Initialize V2T mapping e and classification model f
  For: $t = 1 : T$
    For each video $\boldsymbol{v}_i \in U$
      Calculate informativeness score S_i for $\boldsymbol{v}_i$ as Eq. (5).
    End For
    Select a batch of examples Q with the highest scores from U.
    Query the text view for the videos in Q.
    Add Q to A, and remove Q from U, Update R based on A and U.
    Update V2T mapping e and classification model f by minimizing loss function in Eq. (3)
  End For

At last, we discuss a more complicated case, where the annotation costs vary among different videos. This is a common case in real tasks. For example, to give tags for a video, the human annotator need to watch the whole video. The length of videos will directly decide the time or cost for annotation. We denote the cost for acquiring text features as C_i for video $\boldsymbol{v}_i$. Then an intuitive method to balance the informativeness and annotation cost is to simply divide the S_i score by the annotation cost C_i. In this case, the videos with high utility but easy to annotate will be preferred during the active selection.

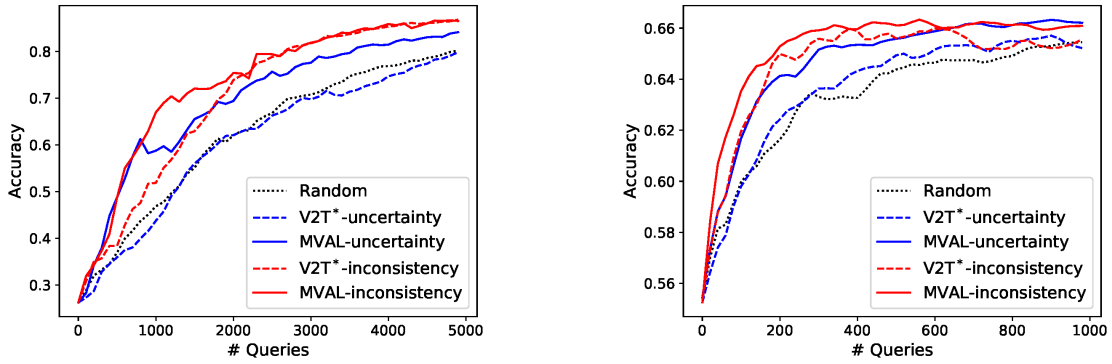
4 Experiments

To validate the effectiveness of the proposed approach, we perform experiments on both multi-view classification datasets and real video recommendation tasks. All experiments are implemented in python with scikit-learn and PyTorch.

4.1 Results on Multi-View Classification Tasks

Firstly, we test the proposed method on two multi-view classification datasets which both contain visual and text views:

- YouTube Multiview Video Games [Madani *et al.*, 2013]. This dataset contains about 30k instances spread over 30 categories. Each instance is described by 13 feature types, from 3 high-level feature families: text, visual, and auditory feature. In the experiments, we concatenate five visual feature types as visual view and take the *text_game_lda_1000* feature extracted with *Latent Dirichlet Allocation* as text view.
- Wikipedia Articles [Rasiwasia *et al.*, 2010]. This dataset contains 2,669 articles spread over 10 categories. Every article contains a single image and at least 70 words. The features of the image and the words are regarded as visual and text view respectively. In our experiments, the text view is represented by 10-dim features extracted



(a) Youtube (b) Wikipedia

Figure 3: Comparison results on two classification datasets.

from words with topic model and the visual view is represented by 128-dim SIFT features of images.

We randomly select 1000 and 100 examples as initial annotated data for two datasets respectively. And the batch size of selection is set as 100 and 20 for two datasets. We repeat the experiments for 10 times and report the average results.

Note that, there are no "users" in classification tasks, so Eq. (2) will degenerate to:

$$\min_{e,f} \sum_{(\boldsymbol{v}_i, y_i) \in A} \ell\left(f(\boldsymbol{v}_i^T), y_i\right) + \sum_{(\boldsymbol{v}_i, y_i) \in U} \ell\left(f(e(\boldsymbol{v}_i^V)), y_i\right) \quad (6)$$

where y_i is denoted by the class label of instance $\boldsymbol{v}_i$. Also, the record frequency is identical for all examples, we thus adjust the selection criterion for the classification task by using uncertainty and inconsistency. After the V2T mapping, the uncertainty criterion will select examples with lowest prediction confidence, while the inconsistency criterion selects examples that are misclassified with high confidence.

We compare the following active learning methods in the experiments:

- **Random**. Randomly select examples to query.
- **V2T*-uncertainty**. The V2T mapping is learnt without supervised information as in Eq. (1); and the examples are selected based on prediction uncertainty.
- **MVAL-uncertainty**. The proposed method with adjusted uncertainty sampling.
- **V2T*-inconsistency**. The V2T mapping is learnt without supervised information as in Eq. (1); and the examples are selected based on prediction inconsistency.
- **MVAL-inconsistency**. The proposed method with adjusted inconsistency sampling.

Figure 3 shows the classification accuracy curves of different active learning approaches with varied numbers of queries on two classification datasets. As expected, the random approach is not as effective as active learning approaches. Compared to uncertainty, the inconsistency based active selection achieves better performance on both datasets. When comparing the proposed MVAL approach with the degenerated version V2T*, it can be observed that no matter which selection criterion applied, the MVAL approach can beat the baseline. This phenomenon validates that the mapping function is better learned by utilizing the supervised information as in Eq. (1).

4.2 Results on Real Video Recommendation

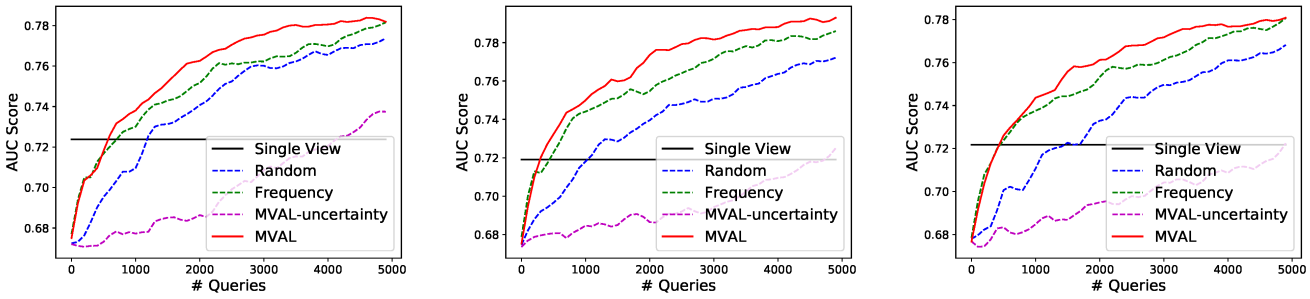
We further test our method on real video recommendation tasks. Three real datasets are collected from Youku [1], a large video service provider in China. These datasets are about three different video categories: **Blooper**, **Highlight** and **Trailer**. There are 243,219 intersections that 6,680 users have made on 69,790 videos in these datasets. Only when a video is exposed to a user and generate a watching action, a positive label is assigned to it.

For each dataset, we randomly separate it into two subsets, one with 70% examples for training, and the other one with 30% examples for testing. From the training set, we sample 1000 videos as the initial annotated set with both the visual and text view, while the text view is missing for other examples. 100 examples are actively selected for annotation at each iteration. We repeat the random partition for 5 times and report the average results. In this experiment, we use the AUC score to evaluate the performance, which is a commonly used criterion in recommendation tasks.

For the visual view, we follow [Abu-El-Haija *et al.*, 2016] to extract visual features from videos using pre-trained models. Specifically, visual features are first extracted at 1-frame-per-second using an Inception-v3 network. We then apply PCA (and whitening) to reduce feature dimensions to 2,048. These frame-level features are eventually aggregated into video-level by average pooling. Regarding the text view, features are defined by averaging word embeddings of refined tags of videos.

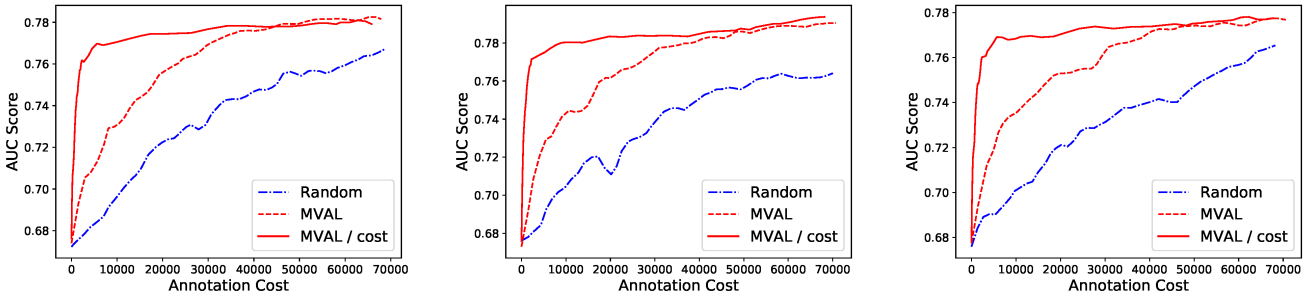
For the V2T model, we use a four-layer multi-layer perceptron (MLP). More specifically, the numbers of units in every

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



(a) Blooper (b) Highlight (c) Trailer

Figure 4: AUC curves of different methods on real video recommendation datasets.



(a) Blooper (b) Highlight (c) Trailer

Figure 5: AUC curves on real video recommendation datasets when the annotation cost varies among different videos.

layer are 2048-1024-512-32. And the model of recommendation is a four-layer MLP too, whose architecture is 64-48-32-1. The optimizer is SGD and the learning rate is 0.01. We split annotated data and unannotated data into 100 batches evenly and respectively. In each iteration, an annotated batch and an unannotated batch are used to train the model.

We compare five methods in the experiments:

- **Single View**. Use visual features of all training data.
- **Random**. Randomly select instances for query.
- **Frequency**. Select videos that are most frequent in user action records.
- **MVAL-uncertainty**. A baseline version of the proposed method, where the videos with largest average uncertainty over all records are selected for query.
- **MVAL**. The proposed method.

Figure 4 presents the AUC score for measuring different methods on three datasets. Note that when there is only one single view, i.e. visual view, even though all training data is used, the accuracy score is quite poor. Although querying most frequent videos is an intuitive strategy, it achieves significantly better results than random query strategy. When combining the V2T mapping with uncertainty sampling, the performance is worse than random query. This indicates that the frequency plays a more important role in MVAL strategy comparing to uncertainty. At last, the proposed approach MVAL achieves the best performance on all of the three datasets.

Figure 5 presents the AUC score when the differences in annotation costs among videos are considered in active selection. Here we take the video length as the annotation cost, and simply divide the criterion value by the cost for active selection. It can be observed that the cost-aware selection will further reduce the annotation cost.

5 Conclusion

In this paper, a novel multi-view active learning approach is proposed for video recommendation. Observing that the text view extracted from video metadata is seriously missing and expensive to obtain, we propose to actively query the missing text features for learning effective recommender systems with lower annotation cost. By simultaneously minimizing the reconstruction error and classification loss, a visual to text mapping function is learned to fully exploit the visual view. Further, the most informative videos are actively selected based on both inconsistency and frequency to reduce the annotation cost. Experiments demonstrate that proposed methods are effective in public datasets as well as real recommendation tasks. In the future, other strategies for active sampling will be studied.

Acknowledgments

This research was supported by NSFC(61876081, 61732006) and the Fundamental Research Funds for the Central Universities, NO.NE2019104.

References

[Abu-El-Haija *et al.*, 2016] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[Baluja *et al.*, 2008] Shumeet Baluja, Rohan Seth, D. Sivakumar, Yushi Jing, Jay Yagnik, Shankar Kumar, Deepak Ravichandran, and Mohamed Aly. Video suggestion and discovery for youtube: taking random walks through the view graph. In *Proceedings of the 17th International Conference on World Wide Web*, pages 895–904, Beijing, China, April 2008.

[Blum and Mitchell, 1998] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

[Chen *et al.*, 2018] Xusong Chen, Dong Liu, Zheng-Jun Zha, Wengang Zhou, Zhiwei Xiong, and Yan Li. Temporal hierarchical attention at category- and item-level for micro-video click-through prediction. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018*, pages 1146–1153, Seoul, Republic of Korea, October 2018. ACM.

[Cui *et al.*, 2014] Peng Cui, Zhiyu Wang, and Zhou Su. What videos are similar with you?: Learning a common attributed representation for video recommendation. In *Proceedings of the ACM International Conference on Multimedia, MM'14*, pages 597–606, Orlando, FL, USA, November 2014.

[Davidson *et al.*, 2010] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Michael Ian Lambert, Blake Livingston, and Dasarathi Sampath. The youtube video recommendation system. In *RecSys*, 2010.

[Houlsby *et al.*, 2014] Neil Houlsby, José Miguel Hernández-Lobato, and Zoubin Ghahramani. Cold-start active learning with robust ordinal matrix factorization. In *International Conference on Machine Learning*, pages 766–774, 2014.

[Huang and Zhou, 2013] Sheng-Jun Huang and Zhi-Hua Zhou. Active query driven by uncertainty and diversity for incremental multi-label learning. In *2013 IEEE 13th International Conference on Data Mining*, pages 1079–1084, 2013.

[Huang *et al.*, 2014] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(10):1936–1949, 2014.

[Huang *et al.*, 2018] Sheng-Jun Huang, Miao Xu, Ming-Kun Xie, Masashi Sugiyama, Gang Niu, and Songcan Chen. Active feature acquisition with supervised matrix completion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018*, pages 1571–1579, London, UK, August 2018. ACM.

[Lee *et al.*, 2018] Joonseok Lee, Sami Abu-El-Haija, Balakrishnan Varadarajan, and Apostol Natsev. Collaborative deep metric learning for video understanding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018*, pages 481–490, London, UK, August 2018. ACM.

[Madani *et al.*, 2013] Omid Madani, Manfred Georg, and David A. Ross. On using nearly-independent feature families for high precision and confidence. *Machine Learning*, 92:457–477, 2013.

[Muslea *et al.*, 2006] Ion Muslea, Steven Minton, and Craig A. Knoblock. Active learning with multiple views. *Journal of Artificial Intelligence Research*, 27:203–233, 2006.

[Rasiwasia *et al.*, 2010] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260. ACM, 2010.

[Rubens *et al.*, 2015] Neil Rubens, Mehdi Elahi, Masashi Sugiyama, and Dain Kaplan. Active learning in recommender systems. In *Recommender systems handbook*, pages 809–846. Springer, 2015.

[Settles, 2012] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[Siersdorfer *et al.*, 2009] Stefan Siersdorfer, Jose San Pedro, and Mark Sanderson. Automatic video tagging using content redundancy. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 395–402. ACM, 2009.

[Song *et al.*, 2005] Yan Song, Xian-Sheng Hua, Li-Rong Dai, and Meng Wang. Semi-automatic video annotation based on active learning with multiple complementary predictors. In *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 97–104. ACM, 2005.

[Tong and Koller, 2001] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

[Wang and Zhou, 2008] Wei Wang and Zhi-Hua Zhou. On multi-view active learning and the combination with semi-supervised learning. In *Proceedings of the Twenty-Fifth International Conference (ICML 2008)*, pages 1152–1159, Helsinki, Finland, June 2008. ACM.