Multi-Task Deep Visual-Semantic Embedding for Video Thumbnail Selection

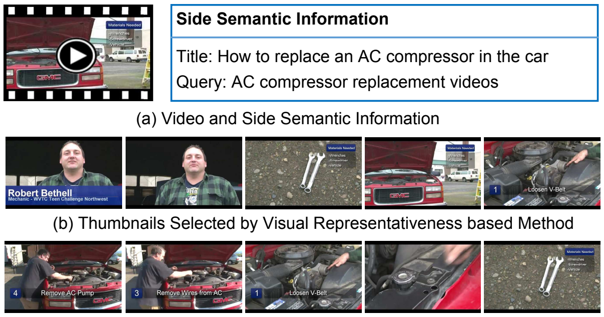
Wu Liu [1,3], Tao Mei [2], Yongdong Zhang [1], Cherry Che [2], and Jiebo Luo [4]
[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[2] Microsoft Research, Beijing 100080, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China
[4] University of Rochester, Rochester, NY 14627, USA
liuwu@live.cn; {tmei, cherc}@microsoft.com; zhyd@ict.ac.cn; jluo@cs.rochester.edu

Abstract

*Given the tremendous growth of online videos, video thumbnail, as the common visualization form of video content, is becoming increasingly important to influence user's browsing and searching experience. However, conventional methods for video thumbnail selection often fail to produce satisfying results as they ignore the side semantic information (*e.g.*, title, description, and query) associated with the video. As a result, the selected thumbnail cannot always represent video semantics and the click-through rate is adversely affected even when the retrieved videos are relevant. In this paper, we have developed a multi-task deep visual-semantic embedding model, which can automatically select query-dependent video thumbnails according to both visual and side information. Different from most existing methods, the proposed approach employs the deep visual-semantic embedding model to directly compute the similarity between the query and video thumbnails by mapping them into a common latent semantic space, where even unseen query-thumbnail pairs can be correctly matched. In particular, we train the embedding model by exploring the large-scale and freely accessible click-through video and image data, as well as employing a multi-task learning strategy to holistically exploit the query-thumbnail relevance from these two highly related datasets. Finally, a thumbnail is selected by fusing both the representative and query relevance scores. The evaluations on 1,000 query-thumbnail dataset labeled by 191 workers in Amazon Mechanical Turk have demonstrated the effectiveness of our proposed method.*

1. Introduction

As the most widely adopted representation of video content, a video thumbnail provides a vivid yet condensed preview of the entire video. Most conventional methods for video thumbnail selection have focused on learning visual representativeness purely from visual content [5, 12, 13]. However, these methods ignore the abundant semantic information associated with the video. For example, a video is usually associated with a title, a short description, a piece of transcript, or a textual query for searching this video. This side information is important for thumbnail selection and often overlooked in previous research. As shown in Figure 1, a video with the title "AC compressor replacement" should be represented by the thumbnails which are truly related to the action of "compressor replacement."


Side Semantic Information
Title: How to replace an AC compressor in the car
Query: AC compressor replacement videos

(a) Video and Side Semantic Information

(b) Thumbnails Selected by Visual Representativeness based Method

(c) Thumbnails Selected by Multi-Task Visual-Semantic Embedding Method

Figure 1. Examples of query-dependent thumbnails. (a) is the video and its side semantic information, (b) shows the thumbnails selected by the visual representativeness based method, and (c) contains the query-dependent thumbnails selected by multi-task deep visual-semantic embedding. Compared with (b), the thumbnails in (c) are more representative and semantically meaningful.

In this paper, we study the problem of video thumbnail selection with side semantic information, which is overlooked in previous research. We investigate how to embed this side semantic information (*e.g.*, title, description, query, and transcript) with visual content to select semantically meaningful thumbnails. The problem has a wide variety of real world applications such as online video summarization [24] where a video is usually associated with a title or a description, and video search where the video thumbnails returned by a search engine are expected to be semantically relevant to a given query.

To this end, we propose a multi-task deep visual-semantic embedding method which serves as a bridge between the diverse side semantic information and visual content. Our main idea is to learn a deep visual-semantic embedding model which directly maps the two views (textual and visual) to a latent semantic embedding space, where the relevance between two incomparable views can be computed through their projections. Different from existing works [2], we employ a large-scale click-through based video and image data to learn a robust embedding model, as well as close the domain gap between video and image by a multi-task learning strategy. We demonstrate the effectiveness of this method in the query-dependent thumbnail selection task. To the best of our knowledge, this paper represents one of the first attempts towards visual-semantic embedding for selecting video thumbnails.

Compared with other multi-view embedding methods [4, 16], visual-semantic embedding model has its unique advantage. It can leverage the semantic similarity to correctly predict the relevance between unseen textual or visual information, and overcome the category limitations of the conventional N-way discrete classifiers models [2]. In our application, we train the visual-semantic embedding model on a click-through video dataset to exploit the relevance between a query and the clicked thumbnail. As the model's performance highly depends on the massive public datasets, we also expand our dataset with user click-through based image data. Compared with artificially labeled data, such click-through data are large-scale, freely accessible, and more useful for understanding the relevance of query-visual information.

However, directly training model on the fusion dataset neglects the gap between images and videos. To solve the problem, we adopt the multi-task learning strategy, which refers to the joint training of multiple tasks, while enforcing a common intermediate parameterization or representation to improve each individual task's performance [17]. Here, we consider the model trained on click-through based image and video datasets as two different but highly related tasks. Thus, we first train one metric that is shared between the two tasks, and then fine-tune the metric to the specific query-dependent thumbnail selection task. Consequently, the learned multi-task visual semantic embedding model avoids overfitting on the click-through based video dataset and adequately exploits more query-thumbnail relationship from the two datasets.

The multi-task deep visual-semantic embedding model is employed to select the query-dependent thumbnail, which is used to supply representative and semantic relevant thumbnails. First, we extract eight different video representative attributes to select 20 most visual representative keyframes as candidate thumbnails. Then, we leverage the trained embedding model to map the side semantic information (*i.e.*, query and title) and visual content into a latent embedding space to compute their relevance. Finally, the visual representative and query relevance scores are fused to select the final thumbnails. One such example can be found in Figure 1. Compared with thumbnails selected by a conventional method in (b), the thumbnails selected by our proposed method in (c) can not only well represent the video content, but also help video browsers or searchers quickly find their interested videos. Furthermore, the experiments on a collection of 1,000 query-video pairs labeled by 191 workers on Amazon Mechanical Turk (AMT) show that 74.83% thumbnails selected by our approach achieve user satisfaction, which is nearly 6% higher than the baseline.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 introduces the deep visual-semantic embedding model trained on video dataset. The proposed multi-task deep visual-semantic embedding is described in Section 4. Section 5 showcases the application on query-dependent thumbnail selection. Experiments and evaluations are presented in Section 6, followed by the conclusion and future work in Section 7.

2. Related Work

Most conventional methods for video thumbnail selection have focused on learning visual representativeness purely from visual content. For instance, Kang *et al*. define the concept of the "representativeness" and divide the criterion into four main video attributes: frame quality, visual details, content dominance, and attention measurement [5]. In addition, Luo *et al*. segment the video clip into homogeneous parts based on major types of camera motion and utilize them to select representative keyframes [13]. Furthermore, high-level features like important object, people and subjects are also utilized to summarize the video [8, 19]. Besides, Lu *et al*. propose a saliency based video summarization method, which trains a linear regression model to predict the importance score for each frame in egocentric videos [12]. Admittedly, the above representativeness-based methods can choose the visual representative frame to represent the video's visual content, but they neglect the semantic information and user's search intent, which may not be adequate to satisfy the users with widest kinds information need.

Accordingly, more recent researchers start to research on how to choose the query-dependent thumbnails to supply specific thumbnails for different queries. The existed works can be classified into search based and learning based. After receiving query information, the search based methods firstly use input queries to search the related images, then to calculate the relevance between query and thumbnail by the similarity between the searched images and thumbnail [1, 6, 10, 23]. However, the images search process in the online stage is too time-consuming for real-world applications. Consequently, learning based methods are employed by more researchers [9, 22]. For example, Wang *et al*. adopt a multiple instance learning approach to localize the tags into video shots and select query-dependent thumbnail according to the tags [22]. Although the method can achieve satisfy performance on limited query-video dataset (*e.g*., 60 queries in [22]), scaling such N-way discrete classifiers methods beyond a limited number of discrete query categories remains an unsolved problem.

In the image search field, aiming to address the above query and image relevance calculation problem, the multi-view embedding methods have been proposed. For instance, Pan *et al*. propose a click-through-based cross view learning strategy, which directly calculate the multi-view distance between a textual query and an image by learning a latent common subspace with the ability in computing the query-image relevance [16]. Furthermore, the deep visual-semantic embedding model leverages textual data to learn semantic relationships between labels, and explicitly mapped images into a rich semantic embedding space [2]. According to their evaluation, they can correctly predict object category labels for unseen categories. Although the above methods can effectively compute the relevance between query and images, we cannot directly use them as the gap between image recognition/search tasks and video thumbnail selection. Differently, we employ a freely click-through based image and video dataset to train the multi-task deep visual-semantic embedding model. In order to eliminate the gap between the two different tasks, we employ multi-task learning strategy to holistically exploit the query-thumbnail relevance from the two high-related datasets. In this way, the query-dependent thumbnails can be effectively and efficiently selected.

3. Deep Visual-Semantic Embedding

The deep visual-semantic embedding was first introduced in [2] to leverage semantic knowledge learned in the text domain, and transfer it to a model trained for visual object recognition. By both mapping into the latent semantic space, they can directly compute the relevance between the label and image by the semantic vector representations of the image and label. More important, for unseen labels, the visual-semantic embedding model can also correctly find the similar images by the labels' semantic representations. Hence, we can leverage the deep visual-semantic embedding to calculate the relevance between the widest unpredictable queries and thumbnails.

The structure of the visual-semantic embedding model can be seen in Figure 2. For the input textual query, the model needs a neural language model which is well-suited for learning semantically-meaningful dense vector representation of words to map the query into the latent semantic space. Here we directly employ the "GloVe" word embedding model [18] which is pre-trained on a corpus of 840 billion tokens of web data [1]. The "GloVe" is used here as its comprehensive performance on word analogy, word similarity, and named entity recognition tasks. Nevertheless, our algorithm does not depend on one fixed word embedding method, any efficient models can also be used here [15]. As suggested in [2], we choose the 300-D embeddings as the good compromise between training speed, semantic quality and ultimate performance.

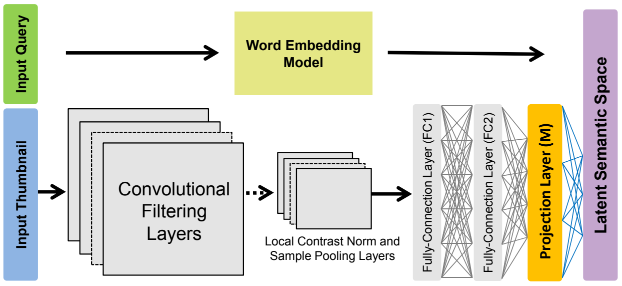

Input Query
Word Embedding Model
Latent Semantic Space
Input Thumbnail
Convolutional Filtering Layers
Local Contrast Norm and Sample Pooling Layers
Fully-Connection Layer (FC1)
Fully-Connection Layer (FC2)
Projection Layer (M)

Figure 2. The structure of the deep visual-semantic embedding model.

For the input thumbnails, we leverage the deep convolutional neural network (CNN) architecture as discussed in [7] by adapting their publicly released C++ implementation [2]. This model has been successfully trained on the ILSVRC-2012 dataset and achieved winning performance on 1,000 categories. The original CNN consists of two parts: 1) the input layers, five convolution layers and maxpooling layers, and 2) two fully connection layers "FC1" and "FC2", and the output layers which produces a distribution over the 1,000 class labels. Here, aiming to map the thumbnails into the latent semantic space, we change the 1,000 label output to the semantic vector representations of the query related to the thumbnail. Meanwhile, the softmax prediction layer is replaced by a projection layer M. Furthermore, we change the original loss function to Equation (1),

$$loss(v, t^{\vec{+}}, t^{\vec{-}}) = \sum max[0, \gamma - t^{\vec{+}} M \vec{v} + t^{\vec{-}} M \vec{v}], \quad (1)$$

where $t^{\vec{+}}$ is the semantic vector representation of the query and $\vec{v}$ is the output of FC2 in CNN network for the thumbnail in the given clicked query-thumbnail pairs, $t^{\vec{-}}$ is the semantic vector representations of other text which is independent of v, M is the matrix of trainable parameters in the projection layer, γ is the parameter of margin (set to 0.1). The loss function is a combination of dot-product similarity and hinge rank loss such that the model was trained to produce a higher dot-product similarity between the vector rep-

[1] "GloVe," http://nlp.stanford.edu/projects/glove/.

[2] "Cuda-convnet," https://code.google.com/p/cuda-convnet/.

resentation of the clicked query-thumbnail pairs than similarity between randomly generated query-thumbnail pairs.

As M is set as one layer of the CNN, we train the visual-semantic embedding model by the stochastic gradient descent. For training set, we use the click-through based video dataset collected from Bing, which consists of 0.5 million $\{query, URL, click\}$ triples, where $query$ is a textual phrase, URL is the video's hyperlink and $click$ is an integer indicating the total clicked number. Compared with artificially labeled data, we use click-through data not only for large-scale and freely accessible, but also as they directly capture the relevance between the query and the clicked thumbnails. From the URL, we download video's title and thumbnail where $click \geq 5$. Then we choose the thumbnail as v, all words in its query and title as t^+, and randomly choose other words as t^-. In addition, Natural Language Toolkit is used to remove the non-English words, stop words, and the morphological affixes [3]. Finally, we get 0.41 million $\{v, t^+, t^-\}$ triples. Training and validate data are randomly chosen from the triples according to the proportion of $10 : 1$. Before training, we init the CNN with the parameters trained well in [21]. In the training process, we firstly trained the parameters of M while holding both the above layers in CNN and the text representation fixed. In the later stages, the derivative of the loss function was back-propagated into all CNN layers to fine-tune the output. However, as the visual-semantic embedding model directly learns the visual feature and visual-semantic relevance from the big data, to train a reliable model, the scale of the training dataset (*i.e.*, 0.41 million) is far from enough. In order to solve the problem, we expand the dataset with click-through based image dataset and describe the proposed multi-task learning strategy in the next section.

4. Multi-Task Deep Visual-Semantic Embedding

The multi-task learning strategy learns related tasks simultaneously by extracting and utilizing appropriate shared information across tasks, which can effectively increase the sample size for each task, and improve their prediction performance. Thus, multi-task learning is especially beneficial when the training sample size is small for each task [3]. As a result, we employ the multi-task learning strategy in our algorithm to expand the training dataset with Clickture to improve the performance of the deep visual-semantic embedding model for query-thumbnail relevance calculation[4]. The Clickture is a large-scale click based image dataset collected from one year click-through data, which is commonly used for image search task. Although the two tasks refer different source medias, video and image, they are highly related as they both provide strong connections between semantics and visual information, as well as connections between users' search intents and queries. Hence, we can employ the Clickture to improve our multi-task deep visual-semantic embedding model. In our experiments, we use the training set in Clickture which consists of 23.1 million $\{query, image, click\}$ triads, *i.e.*, 11.7 million distinct queries and 1.0 million unique images. After the same preprocessing as described in Section 4, we obtain 10.0 million $\{v, t^+, t^-\}$ triples.

By following the multi-task learning setting with K learning tasks (*i.e.*, $K = 2$ in our setting), we can redefine the goal of deep visual-semantic embedding model learning as Equation (2).

$$\begin{aligned}\min_{M_k} \quad & \tau_0 \|M_0 - I\|_F^2 \quad + \\ & \sum_{k=1}^{2} \{\tau_k \|\triangle M_k\|_F^2 + max[0, \gamma - S(t_k^+, v) + S(t_k^-, v)]\},\end{aligned} \tag{2}$$

where M also indicates the projection layer in the CNN. Differently, M_0 picks up general trends across multiple datasets and $M_k = M_0 + \triangle M_k$ specialize each particular task. As an important aspect of multi-task learning is the appropriate coupling of the multiple learning tasks, the minimization of $\|M_0 - I\|_F^2$ and $\|\triangle M_k\|_F^2$ ensures that the learning algorithm does not put too much emphasis onto the shared or individual data. Similar to Equation (1), the minimization of $max[0, \gamma - S(t_k^+, v) + S(t_k^-, v)], S(t_k, v) = \vec{t_k} M_k \vec{v_k}$ make sure that the model is trained to produce a higher dot-product similarity between the vector representation of the clicked query-thumbnail/image pairs. The $\tau_k \geq 0, k = 0, 1, 2$ is a trade-off parameter that controls the regularization of M_t.

The training of multi-task deep visual-semantic embedding model contains two processes. First, we train the M_0 on the common dataset, just like the process in Section 3. Differently, the t and v are both selected from the click-through based image and video dataset. The loss function can be changed to Equation (3).

$$\begin{aligned}& loss(v, \vec{t^+}, \vec{t^-}) = \\ & \tau_0 \|M_0 - I\|_F^2 + \sum max[0, \gamma - \vec{t^+} M_0 \vec{v} + \vec{t^-} M_0 \vec{v}].\end{aligned} \tag{3}$$

We can regard the top seven layers as extracting the common features of the images and thumbnails. Simultaneously, the M_0 layer is employed as the transforming of the visual feature into the semantic space by extracting the common relationship of query-image and query-thumbnail.

After trained M_0, we need to fine-tune the M_1 for query-dependent thumbnail selection. In this stage, we lock the top seven layers and only train the M_1 layer on the click-through based video dataset. In particular, the loss function

[3] "Natural Language Toolkit," http://www.nltk.org/.

[4] "Clickture," http://research.microsoft.com/en-us/projects/clickture/.

is changed to Equation (4).

$$loss(v, \vec{t^+}, \vec{t^-}) =$$
$$\tau_1 \|M_1 - M_0\|_F^2 + \sum max[0, \gamma - \vec{t^+} M_1 \vec{v} + \vec{t^-} M_1 \vec{v}], \tag{4}$$

where $\|M_1 - M_0\|_F^2$ ensures that the learning algorithm does not put too much emphasis onto the video dataset to avoid the overfitting, τ_1 is the trade-off parameter. In the training, we can initialize the τ_1 with 1.0 and modify it according to the test error. As we mainly focus on the query-dependent thumbnail selection task, in this stage, we only fine-tune M_1 on click-through based video dataset and neglect M_2. In the future, we will try to simultaneously fine-tune more parallel projection layers M_k for different tasks in one CNN network.

5. Query-dependent Thumbnail Selection

In this section, we will introduce how to utilize the above multi-task deep visual-semantic embedding model to select the query-dependent thumbnails. The process can be divided into two stages: offline and online.

Although we aim to select the query-dependent thumbnail, we cannot ignore thumbnail's primary role is to represent the video content properly. Hence in the offline stage, we firstly need to select the visual representative and comprehensive keyframes as the candidate thumbnails by the video representation attributes based method. Inspired by [5], we firstly segment the videos into scenes, shots and sub-shots by the color histogram. Then for keyframes in each sub shots, two kinds of video attributes are extracted to compute the keyframes' representative score. For high image visual quantity, we extract 1) the durations of sub-shot, neighbor sub-shot, shot and scene; 2) the keyframes' position, color entropy, motion blur and edge sharpness attributes; and 3) successive keyframe's similarities as video representation attributes. For high image's attractiveness, we extract the human face and skin ratio as video representation attributes. After that, the final representative score is computed by the linear fusion of these attributes. The top 20 keyframes are selected as the candidate thumbnails by their representative scores. Additionally, their representative scores are also saved to be fused with query relevance scores in the online stage. Finally, we will use the trained multi-task deep visual-semantic embedding model to map all the candidate images into the latent space to get their semantic vector representations.

In the online video search stage, after receiving user's query, we also need to map it into the semantic space by the word embedding model. Then the query-thumbnail relevance is directly computed in the latent semantic space by their semantic representation vectors. As the query may contain several words, we compute one thumbnail's cosine similarity with each word in the query. Then the highest similarity will be used as each thumbnail's query relevance score. Finally, we fuse each thumbnail's visual representative score and query relevance score through the average late fusion and send thumbnail with the highest score to the user as the final selected query-dependent thumbnail. As most of the complicated works (*e.g.*, model training, video keyframes extraction and representative score computation) are processed offline, in the online, it only takes $21ms$ to select the query-dependent thumbnail for each query-video pair on an Ubuntu 14.10 server with Intel Xeon CPU E5-2650, 64 GB Memory and NVidia Tesla K20Xm GPU.

6. Experiments

6.1. Evaluation Dataset

To evaluate the performance of our proposed algorithms, we select 1,000 query-video pairs from the click-through based video dataset, which are not used in the training process. The videos contain nine categories and the average length is 331 seconds. To reduce the workload of labeling, for each video, we use video representativeness attributes based method described in Section 5 to extract 20 most keyframes as candidate thumbnails. In the end, we totally get 17,480 candidate thumbnails. As the thumbnail selection is somewhat subjective task, and for each video more than one frame could be chosen as a thumbnail, we ask the workers in the AMT to label each candidate thumbnail according to the query and video [5]. To ensure that the labels are consistent across viewers, we totally publish 5,000 hit tasks on AMT to make each video be labeled by five diffident workers. Each hit contains one query-video pair, 20 candidate thumbnails. The workers must firstly read the query and watch the video, then label all the candidate thumbnails by five different scores: Very Good (VG), Good (G), Fair (F), Bad (B) and Very Bad (VB) [6].

Furthermore, to control the label quality, the follow requirements are also added: 1) Only the workers who have already been approved more than 100 hits and the hit approval rates are higher than 80% can join the project. 2) If more than one thumbnail is missed in one hit, we will reject it to make sure most of the candidate thumbnails are labeled by five different people. 3) For 65% videos, we choose at least one exactly VB or VG thumbnail as seed. If a worker gives more than three obviously wrong scores for the seeds (*e.g.*, label higher than B for VB or lower than G for VG), we will block him and reject all of his hits. 4) If the worker gives the same score to 90% candidate thumbnails in one hit, we will check the video to decide whether to reject the

[5] "Amazon Mechanical Turk," https://www.mturk.com.

[6] The detailed score criteria and labeled dataset are released in http://mcg.ict.ac.cn/mcg-vts.html as a benchmark for video thumbnail selection in this community.

hit. In the label processing, nearly 12% hits are rejected and all the rejected hits are published to be labeled again. At last, more than 191 workers join the label project and averagely spend 355 seconds for each hit. For each candidate thumbnail, we select the score labeled by the most workers as its final score. If there is more than one such score, we choose the low one. Some example videos and labeled results can be found in Figure 5.

6.2. Experimental Settings

In order to evaluate the multi-task visual-semantic embedding for query-dependent thumbnail selection, we compare the following seven methods on the labeled query-thumbnail dataset:

(1) **RANDOM.** The method randomly selects one image from the candidate thumbnails as final thumbnail.

(2) **Video Representation Attributes based Method (ATTR).** The method selects the most visual representative video frame as thumbnail by the video representation attributes [5] described in Section 5. We select it as the state-of-the-art for video representativeness based method.

(3) **Canonical Correlation Analysis (CCA).** A classical and successful query-image similarity computation approach to map visual and textual features into a latent subspace where the correlation between the two views is maximized [4, 16]. We retrain the CCA model on the click-through based image and video dataset by the implementation in [4], and select it as the state-of-the-art for traditional query-dependent method.

(4) **VSEM-VIDEO.** Query-dependent thumbnail selection method in which the deep visual-semantic embedding model is only trained on click-through based video dataset.

(5) **VSEM-ALL.** Query-dependent thumbnail selection method in which the deep visual-semantic embedding model is trained on click-through based image and video dataset without multi-task learning strategy.

(6) **MTL-VSEM.** The proposed query-dependent thumbnail selection method with the multi-task visual-semantic embedding model trained on click-through based image and video dataset.

(7) **FUSION.** The proposed query-dependent thumbnail selection method considers both the visual representativeness (ATTR) and user's search intent (MTL-VSEM) by averagely fusing their scores.

As the video search engineer can supply one or multi thumbnails to the user, we evaluate the methods by two criteria: HIT@1 computing the hit ratio for the first selected thumbnail and Mean Average Precision (MAP) computing the precision for all the candidate thumbnails. The MAP is computed by

Method	HIT@1 (VG)	HIT@1 (VG & G)
RANDOM	26.95%	55.68%
ATTR [5]	40.21%	68.89%
CCA [4]	31.66%	60.06%
VSEM-VIDEO	32.96%	62.48%
VSEM-ALL	39.88%	67.84%
MTL-VSEM	**43.03%**	**71.70%**
FUSION	**47.13%**	**74.83%**

Table 1. Comparison of the thumbnail selection methods (in terms of HIT@1).

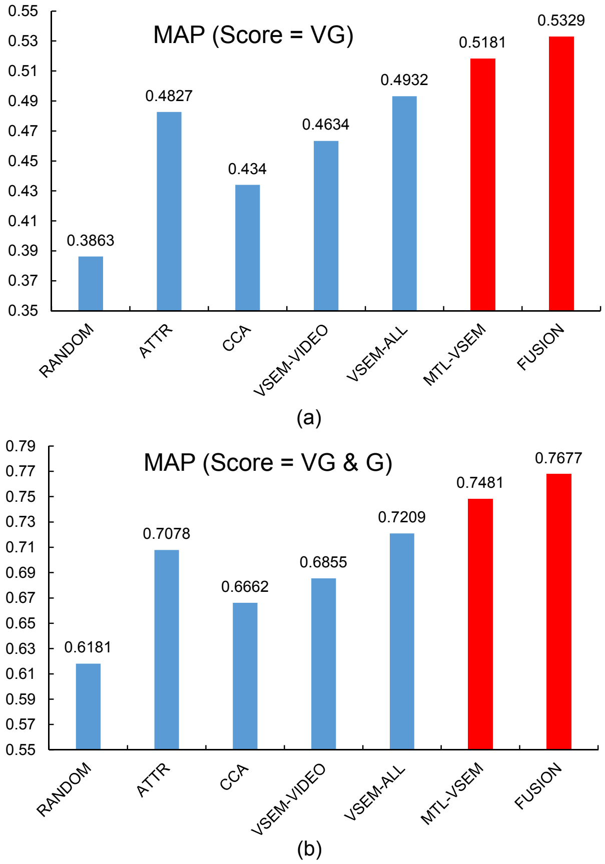
$$MAP = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{jk}), \qquad (5)$$

where Q is the query set, m_j is the number of positive thumbnails in each query-video pairs, $Precision(R_{jk})$ is the average precision at the position of returned kth positive thumbnails. As we label the thumbnails by five different scores, we separately compute the HIT@1 and MAP in two different situations: set positive score equals VG and set positive scores as VG and G.

6.3. Evaluation on Entire Videos

We first test the seven methods with all the 1,000 query-video pairs. The hit@1 results can be seen in Table 1. The MAP results for different positive thumbnail selection standards can be found in Figure 3. In order to evaluate the gap between proposed methods and human labeling, we also compute the degree of agreement among AMT labels, which is 68.34% and 83.04% for VG and VG&G. From the results, we find that whether only selecting one thumbnail or giving several thumbnails, our method both achieves the highest accuracy among the seven methods. Compared to ATTR, our method can obviously improve the video thumbnail selection accuracy, which demonstrates that the proposed query-thumbnail relevance calculation method definitely refers the user's search intent and gives them more satisfactory thumbnails. Furthermore, compared with CCA, our method also achieves much better accuracy. The reason is that the CCA only trains a transformation matrix which linearly maps the designed visual and textual features into one latent subspace. However, the designed features maybe not well linear correlation with each other. Differently, in the training of the visual-semantic embedding model, the derivative of the loss function was back-propagated into the core visual model to fine-tune the generated visual features.

More important, among all the query-dependent thumbnail selection methods based on visual-semantic embed-



(a)

(b)

Figure 3. Comparison of the thumbnail selection methods (in terms of MAP). (a) positive score equals VG; (b) positive scores equal VG and G.

ding model, *i.e*., VSEM-VIDEO, VSEM-ALL and MTL-VSEM, the proposed MTL-VSEM also achieves the highest accuracy. The reason is that VSEM-VIDEO only trains the visual-semantic embedding model on the limited click-through based video dataset, which is too small to exploit the visual-semantic relevance. Although adding the image dataset, the VSEM-ALL method does not consider the gap between the image search and video thumbnail selection. The trained model put too much emphasis onto the shared information across the two tasks without fine-tuned on the individual task. Instead, the MTL-VSEM effectively holistically exploits the appropriate shared information across the two dataset and avoids the overfitting on the limited video dataset. The results demonstrate the effectiveness of the multi-task learning strategy. Finally, the proposed FUSION method achieves the best performance, which demonstrates that the representative and query-dependent information are both very important for the video thumbnail selection.

6.4. Evaluation on Different Video Categories

In order to further evaluate the effectiveness of our approach, we test the thumbnail selection accuracy on nine different video categories: Education and Technology (Educ.), Entertainment (Enter.), Film, Games and Cartoon (Games), Music, News and Politics (News), Objects, People and Blogs (People), and Sports. We use MAP as criterion and the positive thumbnails are set as scores equal VG and G. As shown in Figure 4, the performance improvements of MTL-VSEM and FUSION are consistent and stable, *i.e*., all the video categories improved compared to other methods. In particular, for "Objects" and "Education," the improvements are very significant. It demonstrates that when searching these kind videos, the users are more purposeful to find the specific video. In this case, the query-dependent based methods will play a greater role. By contrast, for "Sports" and "Film," the representativeness based method can give similar effects with the query-dependent methods. The reason is that in this case, most of the video contents are related to the query. For instance, the football video always contains football match views from beginning to the end. Hence the selected representative thumbnail is also highly related to the query. However, as it is hard to capture the match's wonderful moment or the film's specific actor in the query, all methods cannot give very satisfy results for "Sports" and "Film" videos. Besides, compared with VSEM-ALL, the MTL-VSEM also achieves better performance for most categories, especially for "Object" and "Games," which further demonstrates the effectiveness of the multi-task learning strategy.

Figure 5 shows some selected thumbnails by different approaches for different queries and video categories to give the reader a visual sense. It is clear that the proposed MTL-VSEM and FUSION methods produce the most satisfactory thumbnails. Specifically, compared to other baselines, the selected thumbnails by the proposed method can not only provides a vivid yet condensed preview of the entire video, but also be very related to the queries, which can help users quickly find the relevant videos. As supplements, we also provide two failed examples in the last rows for discussion in Figure 5 (b). For the first example, FUSION and MTL-VSEM methods prefer the American Football match screen instead of Lance Briggs's close-up. For the second example, the two proposed methods also fail to capture "Matt Damon."

7. Conclusions and Future Work

In this paper, we have investigated the issue of directly learning the multi-view distance between the video side semantic information and visual content by training the multi-task deep visual-semantic embedding model on the click-through image-video data. The large scale and freely acces-

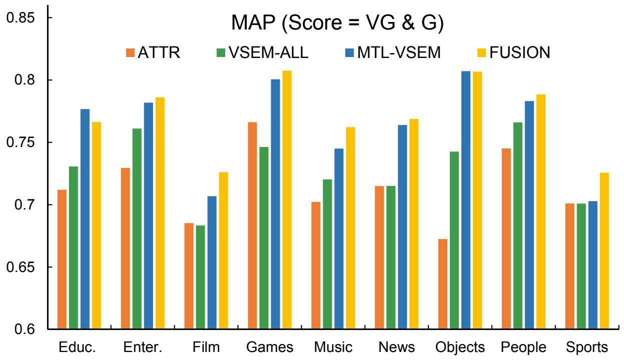


Figure 4. Comparison of the thumbnail selection methods on different video categories (in terms of MAP, the positive scores equal VG and G).

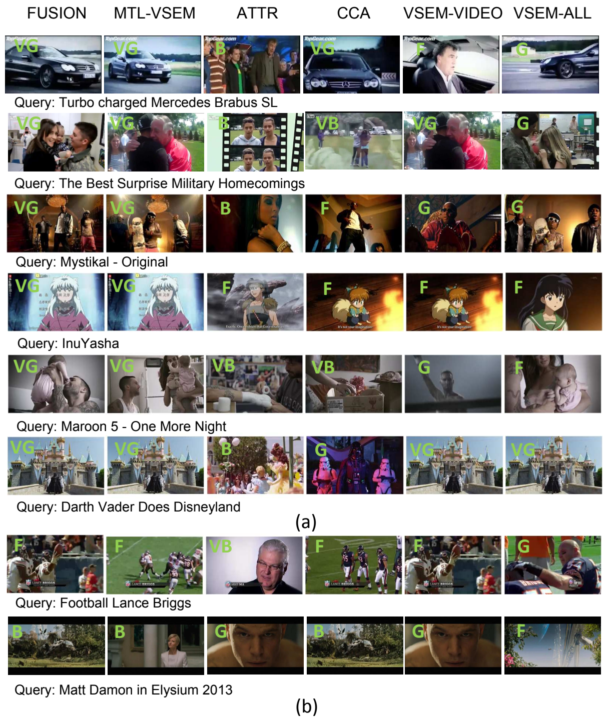

FUSION
MTL-VSEM
ATTR
CCA
VSEM-VIDEO
VSEM-ALL
VG VG B VG F G
Query: Turbo charged Mercedes Brabus SL
VG VG B VB VG G
Query: The Best Surprise Military Homecomings
VG VG B F G G
Query: Mystikal - Original
VG VG F F F F
Query: InuYasha
VG VG VB VB G F
Query: Maroon 5 - One More Night
VG VG B G VG VG
Query: Darth Vader Does Disneyland
(a)
F F VB F F G
Query: Football Lance Briggs
B B G B G F
Query: Matt Damon in Elysium 2013
(b)

Figure 5. Examples of selected thumbnails by different methods for different queries and video categories (better viewed in color). The labeled score is provided at the top left corner for each thumbnail. (a) six successful examples; (b) two failure examples for discussion.

sible click-through data supply a massive open dataset to train the visual-semantic embedding model. The multi-task learning strategy proves effective by holistically exploiting the click-through between queries and thumbnails. Finally, the proposed algorithm is evaluated for the query-dependent video thumbnail selection task. The results demonstrate that the user search experience has been significantly improved with our proposed method due to the selected representative and personalized thumbnails.

Although the multi-task deep visual-semantic embedding model is only used for query-dependent video thumbnail selection in this paper, it can also be easily applied to other applications, such as video tag localization [20], video search reranking [14], mobile video search [11] and so on. Furthermore, we will try to simultaneously fine-tune more different tasks with more parallel projection layers M_k to improve each other. More modality such as video tags, subtitle, automatic speech recognition and face recognition will also be investigated to select more effective thumbnails.

Acknowledgement

The authors would like to thank Dr. Ji Wan for his valuable comments in preparing this paper. This work was partially supported by the National High Technology Research and Development Program of China (2014AA015202), and the National Natural Science Foundation of China (61173054).

References

[1] L. Ballan, M. Bertini, A. D. Bimbo, and G. Serra. Enriching and localizing semantic tags in internet videos. In *ACM Multimedia*, pages 1541–1544, 2011.

[2] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013.

[3] P. Gong, J. Zhou, W. Fan, and J. Ye. Efficient multi-task feature learning with calibration. In *KDD*, pages 761–770, 2014.

[4] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision*, 106(2):210–233, 2014.

[5] H. W. Kang and X. S. Hua. To learn representativeness of video frames. In *ACM Multimedia*, pages 423–426, 2005.

[6] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan. Large-scale video summarization using web-image priors. In *CVPR*, pages 2698–2705, 2013.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.

[8] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, pages 1346–1353, 2012.

[9] H. Li, L. Yi, B. Liu, and Y. Wang. Localizing relevant frames in web videos using topic model and relevance filtering. *Machine Vision and Applications*, 25(7):1661–1670, 2014.

[10] C. Liu, Q. Huang, and S. Jiang. Query sensitive dynamic web video thumbnail generation. In *ICIP*, pages 2449–2452, 2011.

[11] W. Liu, T. Mei, and Y. Zhang. Instant mobile video search with layered audio-video indexing and progressive transmission. *IEEE Transactions on Multimedia*, 16(8):2242–2255, 2014.

[12] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *CVPR*, pages 2714–2721, 2013.

[13] J. Luo, C. Papin, and K. Costello. Towards extracting semantically meaningful key frames from personal video clips: From humans to computers. *IEEE Trans. Circuits Syst. Video Techn.*, 19(2):289–301, 2009.

[14] T. Mei, Y. Rui, S. Li, and Q. Tian. Multimedia search reranking: A literature survey. *ACM Computing Surveys*, 46(3):38, 2014.

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

[16] Y. Pan, T. Yao, T. Mei, H. Li, C.-W. Ngo, and Y. Rui. Click-through-based cross-view learning for image search. In *SIGIR*, pages 717–726, 2014.

[17] S. Parameswaran and K. Q. Weinberger. Large margin multi-task metric learning. In *NIPS*, pages 1867–1875, 2010.

[18] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[19] A. Rav-Acha, Y. Pritch, and S. Peleg. Making a long video short: Dynamic video synopsis. In *CVPR*, pages 435–441, 2006.

[20] K. D. Tang, R. Sukthankar, J. Yagnik, and F. Li. Discriminative segment annotation in weakly labeled video. In *CVPR*, pages 2483–2490, 2013.

[21] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li. Deep learning for content-based image retrieval: A comprehensive study. In *ACM Multimedia*, pages 157–166, 2014.

[22] M. Wang, R. Hong, G. Li, Z. Zha, S. Yan, and T. Chua. Event driven web video summarization by tag localization and key-shot identification. *IEEE Transactions on Multimedia*, 14(4):975–985, 2012.

[23] W. Zhang, C. Liu, Z. Wang, G. Li, Q. Huang, and W. Gao. Web video thumbnail recommendation with content-aware analysis and query-sensitive matching. *Multimedia Tools Appl.*, 73(1):547–571, 2014.

[24] B. Zhao and E. P. Xing. Quasi real-time summarization for consumer videos. In *CVPR*, pages 2513–2520, 2014.