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

Wu Liu[1], Tao Mei[2], Yongdong Zhang[1], Cherry Che[2], Jiebo Luo[3],
[1]Institute of Computing Technology, Chinese Academy of Sciences, [2]Microsoft Research, [3]University of Rochester,

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. 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. 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.

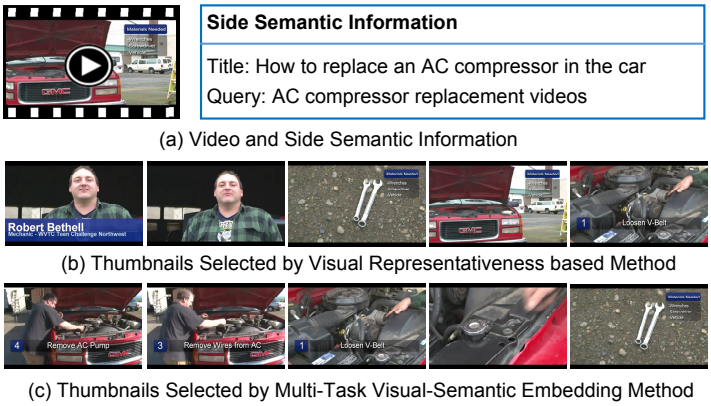
To this end, we propose a multi-task deep visual-semantic embedding method which serves as a bridge between the diverse side semantic information and visual content. The embedding model has a wide variety of real world applications such as video thumbnail selection, online video summarization, video tag localization and video captioning. 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 [1], 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.

The structure of the deep visual-semantic embedding model is described as follow: for the input textual query, the model directly employ "GloVe" word embedding [5] to map the query into the latent semantic space. The "GloVe" word embedding model is pre-trained on a corpus of 840 billion tokens of web data and used here as its comprehensive performance. For the input candidate thumbnails, we leverage the deep convolutional neural network (CNN) architecture in the model by adapting the released C++ implementation in [3]. 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. Aiming to map the thumbnails into the latent semantic space, we change the 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. As the model's performance highly depends on the massive public datasets, we train the deep visual-semantic embedding model on a click-through-based video and image dataset to exploit the relevance between a query and the clicked thumbnail or image. 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 image and video. 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 [4] to improve each individual task's performance. By following the multi-task learning setting with K learning tasks ($K=2$ in our setting), we can redefine the goal of learning deep visual-semantic embedding model as Equation (1).

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

where, M_0 and M_k indicate the projection layers in the CNN. Differently, M_0 picks up general trends across multiple datasets and $M_k = M_0 + \triangle M_k$ specializes each particular task. The minimization of $\|M_0 - I\|_F^2$ and $\|\triangle M_k\|_F^2$ ensure that the learning algorithm does not put too much emphasis onto the shared or individual data. 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 semantic vector representation of the clicked query-thumbnail/image pairs. The $\tau_k \geq 0, k = 0, 1, 2$ are trade-off parameters that control the regularization of M_k. The training of multi-task deep visual-semantic embedding model contains two processes: we first train M_0 on the common dataset, then fine-tune M_1 to the specific query-dependent thumbnail selection task on the click-through video dataset. Consequently, the learned embedding model avoids overfitting on the click-through video dataset and adequately exploits more query-thumbnail relationship from the image and video datasets.



(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 the end, the learned multi-task deep visual-semantic embedding model is employed to select the query-dependent thumbnail. First, we extract eight different video representative attributes [2] 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 calculate their relevance. Next, the visual representative and query relevance scores are fused to select the final thumbnails. Finally, the experiments on a collection of 1,000 query-video pairs labeled by 191 workers on Amazon Mechanical Turk show that 74.83% thumbnails selected by our approach achieve user satisfaction, which is nearly 6% higher than the baseline .

[1] 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.

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

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

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

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

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

The dataset is released in "http://mcg.ict.ac.cn/mcg-vts.html" as a benchmark for video thumbnail selection in this community.