ETH-CVL @ MediaEval 2016: Textual-Visual Embeddings and Video2GIF for Video Interestingness

Arun Balajee Vasudevan
CVLab, ETH Zurich
arunv@student.ethz.ch

Michael Gygli
CVLab, ETH Zurich
gygli@vision.ee.ethz.ch

Anna Volokitin
CVLab, ETH Zurich
anna.volokitin@vision.ee.ethz.ch

Luc Van Gool
CVLab, ETH Zurich
vangool@vision.ee.ethz.ch

ABSTRACT

This paper presents the methods that underly our submission to the *Predicting Media Interestingness Task* at MediaEval 2016. Our contribution relies on two main approaches: (i) A similarity metric between image and text and (ii) a generic video highlight detector. In particular, we develop a method for learning the similarity of text and images, by projecting them into the same embedding space. This embedding allows to find video frames that are both, canonical and relevant w.r.t the title of the video. We present the result of different configurations and give insights into when our best performing method works well and where it has difficulties.

1. INTRODUCTION

The number of online video uploads has been growing for many years[1]. In the contemporary fast moving world, it is clearly observable that social media trends a shortened or compressed form of videos than their complete versions, as they are more easily consumable. This increases the importance of extracting attractive keyframes or automatically finding the best video segments from the videos. Such an condensed form of videos may improve the viewer experience [1] as well as video search [2].

In the following we will detail our approach for tackling this difficult prediction problem and present our results on the MediaEval 2016 challenge on *Predicting Media Interestingness* [3]. The goal of this task is to predict the frame and segment interestingness of Hollywood like movie trailers. This, in turn, helps a user to make a better decision about whether he or she might be interested in a movie. The dataset provided for this task consists of a development set of 52 trailers and a test set of 26 trailers. More information on the task can be found in [3].

There are many conventional works for extracting frames based on the visual content [5, 9, 13, 16]. More recently, several works have presented models that rely on semantic information associated with the videos such as the title of the video [14] or a user query [12] to find relevant and interesting frames. The use of semantic side information allows to build a strong, video specific interestingness model [11, 14]. Liu *et al.* [11], for example, use the title of a video to retrieve photos from Flickr. Then, the video frame interestingness is measured by computing the visual similarity between the frame and retrieved photo set.

In this work, we rely on two models: (i) a frame-based model that uses textual side information and (ii) a generic predictor for finding video highlights in the form of segments [6]. For our frame-based model, we follow the work of [12] and learn a joint embedding space for images and text, which allows to measure relevance of a frame w.r.t. some text such as the video title. For the video segment selection based on its interestingness, we use the work of Gygli *et al.*[6] which trained a Deep RankNet to rank the video segments of a video based upon on their suitability as animated GIFs.

[1]https://www.youtube.com/yt/press/statistics.html

MediaEval 2016 Workshop October 20-21, 2016, Hilversum, Netherlands

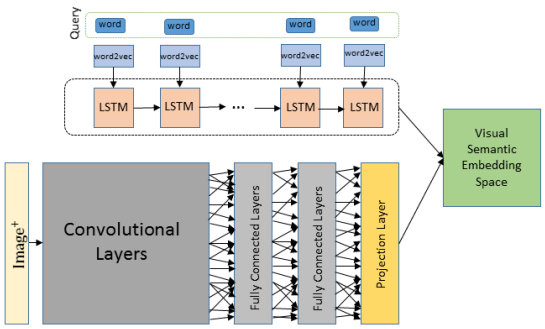

Figure 1: Visual Semantic Embedding Model

2. VISUAL-SEMANTIC EMBEDDING

The structure of our Visual Semantic Embedding model is shown in Figure 1. In our model, we have two parallel networks for the images and texts separately, which are jointly trained with a common loss function. The network is built in an end-to-end fashion for training and inference and trained on the MSR Clickture dataset [7]. The aim of our model is to map images and queries into the same textual-visual embedding space. In this space, semantic proximity between texts and images can be easily computed by the cosine similarity of their representations [4, 10]. We train the network with positive and negative examples of query-image pairs from the MSR dataset and learn to score the positive pair higher than the negative one, *i.e.* we pose it as a ranking problem. Thus, we optimize an objective that re-

Figure 2: Qualitative Results of 3 pairs of highly ranked keyframes followed by ground truth for following videos with titles: Captives, After Earth, Stonewall (from left). Blue color text depicts Prediction score while Green depicts ground truth score.

Tasks	Run types	mAP
Image	Run-1	0.1866
	Run-2	**0.1952**
	Run-3	0.1858
Video	Run-1	0.1362
	Run-2	**0.1574**

(a) Runs Comparison

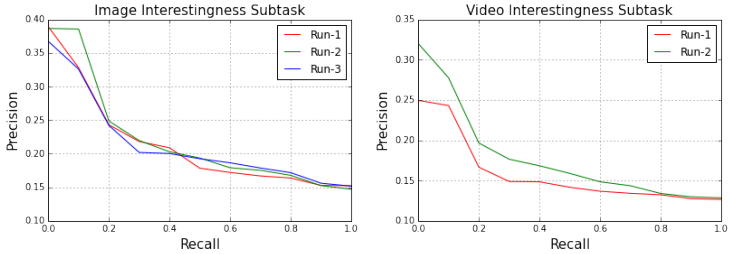

Figure 3: Tabulated Results in (a) and Precision-Recall curve of two subtasks

quires the query embedding with respect to a related image to have a higher cosine similarity compared to the embedding w.r.t. to the randomly selected image. Let $h(q, v)$ be the score from the model for a text query q for some image v. Let v^+ be an image of text-relevant image (positive) and v^- be image embedding of non-relevant image. Then, our objective function is as follows:

$$h(q, v^+) > h(q, v^-). \quad (1)$$

We use a huber rank loss [8] to optimize this objective, similar to [6].

In the inference stage, for a given movie title and given keyframes, we embed the title and the keyframes into the same space. Then, we rank the list of keyframes based on the proximity of the frame embeddings to the text embedding.

3. VIDEO HIGHLIGHT DETECTOR

We use the work of Video2GIF [6] as a generic video highlight detector. To capture the spatio-temporal visual features of video segments, 3D convolutional neural networks (C3D)[15] are used. The model comprises of C3D followed by two fully connected layers and finally outputs a score. The model is trained on the Video2GIF dataset [6] to learn to score segments that were used for GIFs higher than the non-selected segments within a video. Thus, it also uses a ranking loss for training. The scores given by the model are not absolute but are ordinal i.e. a segment with better score is more interesting than a low scored segment. These scores pave the way for ranking of segments for the interestingness. Given the segments of a video, the model ranks all the segments based on their suitability as a GIF which is generally a short segment of a video which is appealing to a viewer.

4. EXPERIMENTS

For the *Image Interestingness subtask*, we use Visual Semantic Embedding Model and we then fine-tune the model using the dev set of MediaEval for domain adaptation. We submit three runs for this task 1) **Run-1** Visual Semantic Embedding Model trained on 0.5M query(text)-image pairs of MSR Clickture dataset 2) **Run-2** Run-1 model finetuned on development set 3) **Run-3** Run-1 model but trained on 8M query-image pairs. For video interestingness task, we use Video2GIF [6]. However, Video2GIF does not consider any meta information for scoring and ranking the video segments. Hence, we propose to combine Visual Semantic Embedding model scores with Video2GIF scores. For this, we extract the middle frame from each video segment and score that frame using Visual Semantic Embedding model. Then, we combine its score with the score from Video2GIF for the same segment by averaging. We submit two runs for *Video Interestingness subtask*: 1) **Run-1** Video2GIF [6] 2) **Run-2** Averaging the prediction scores of Run-1 of video subtask and Run-2 of image subtask. The combined score seems to rank the segments better than Video2GIF model alone as seen in Figure 3.

5. RESULTS AND DISCUSSION

We evaluate our models on the MediaEval 2016 *Predicting Media Interestingness Task* [3].

Figure 3 represents the Precision-Recall curves of image and video interestingness subtasks using our models. We observe that Run-2 of image interestingness subtask performs better than the other two runs. Our initial model is trained on images which differ from video frames in quality and content. Thus, fine-tuning on the development set for adapting to video domain improves mAP. Qualitatively, in Figure 2, we observe that the first two pairs have the model selected keyframes quite close to the ground truth. This is because the movie titles (Captives, After Earth) give a clear visual hint on what an appealing frame should contain. However, the third is a failure case as the title (Stonewall) is misleading: It is about a protest movement, not a wall. Thus, our model has difficulties picking the right keyframes in this case. In the case of video interestingness subtask, we observe that Run-2 performs better than Run-1. Combining the prediction scores of Video2GIF (Run-1) and Run-2 of image interestingness subtask significantly improves the performance of video interestingness subtask. This is because Video2GIF does not take into account the relevance of movie titles for scoring the segments in contrast to query relevant scoring of keyframes of the Visual Semantic Embedding model. Hence, the combination of both models outperforms Video2GIF alone (Run-1).

6. REFERENCES

[1] S. Bakhshi, D. Shamma, L. Kennedy, Y. Song, P. de Juan, and J. Kaye. Fast, Cheap, and Good: Why Animated GIFs Engage Us. In *ACM Conference on Human Factors in Computing Systems*, 2016.

[2] L. Ballan, M. Bertini, G. Serra, and A. Del Bimbo. A data-driven approach for tag refinement and localization in web videos. *Computer Vision and Image Understanding*, 2015.

[3] C.-H. Demarty, M. Sjöberg, B. Ionescu, T.-T. Do, H. Wang, N. Q. K. Duong, and F. Lefebvre. MediaEval 2016 Predicting Media Interestingness Task. In *MediaEval 2016*, 2016.

[4] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, 2013.

[5] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Van Gool. The Interestingness of Images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2013.

[6] M. Gygli, Y. Song, and L. Cao. Video2GIF: Automatic Generation of Animated GIFs from Video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[7] X.-S. Hua, L. Yang, J. Wang, J. Wang, M. Ye, K. Wang, Y. Rui, and J. Li. Clickage: towards bridging semantic and intent gaps via mining click logs of search engines. In *Proceedings of the 21st ACM international conference on Multimedia*, 2013.

[8] P. J. Huber et al. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[9] Y.-G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yang. Understanding and predicting interestingness of videos. In *AAAI*, 2013.

[10] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[11] F. Liu, Y. Niu, and M. Gleicher. Using Web Photos for Measuring Video Frame Interestingness. In *IJCAI*, 2009.

[12] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo. Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[13] M. Soleymani. The quest for visual interest. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, 2015.

[14] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. TVSum: Summarizing Web Videos Using Titles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[15] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. *arXiv preprint arXiv:1412.0767*, 2014.

[16] Y. Wang, Z. Lin, X. Shen, R. Mech, G. Miller, and G. W. Cottrell. Event-specific image importance. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.