Zero-shot Video Retrieval using a Large-scale Video Database

Kazuya Ueki
Meisei University, Tokyo, Japan
kazuya.ueki@meisei-u.ac.jp

Takayuki Hori
SoftBank Corp. Tokyo, Japan
Waseda University, Tokyo, Japan
takayuki.hori@g.softbank.co.jp

Abstract

Following are two mainstream approaches of video retrieval from large-scale video data using query sentences: (1) an approach to find pre-trained concepts such as objects, persons, scenes, and activities corresponding to a query sentence, and (2) an approach to map a query sentence and images/videos into the same feature space and directly search for images/videos that match the query sentence. In this study, we analyze the advantages and disadvantages of these two approaches using a large-scale video database of TRECVID benchmark and confirm whether the fusion of these approaches can improve video retrieval performance.

1. Introduction

Many researchers have been working on video retrieval technology, which can search for videos that a user wants to watch using search keywords from large-scale and miscellaneous video data uploaded on the Internet such as YouTube. In the TREC video retrieval evaluation (TRECVID) benchmark [1], which is organized by the US National Institute of Standards and Technology every year, research institutes around the world form teams and work on several video retrieval research tasks.

In this research, we aim to improve video retrieval performance in zero-shot learning, using complicated query sentences including multiple concepts. Video data of the TRECVID ad-hoc video search (AVS) task, evaluated in the 2016–2018 TRECVID benchmark, was used to determine the video retrieval performance. In the TRECVID AVS task, given a query sentence such as "Find shots of a person in front of a blackboard talking or writing in a classroom," a system needs to retrieve videos corresponding to this query sentence from a large-scale video database. The major difficulty in this task is that a system must retrieve videos under conditions where no training videos match a query sentence, i.e., zero-shot learning.

2. Related Work

There are two mainstream approaches of video retrieval; concept-based and visual-semantic embedding approaches.

The concept-based approach is a method of constructing a large-scale concept bank consisting of pre-trained concept classifiers. In this approach, a system tries to select appropriate concepts in a query sentence from the concept bank. Therefore, it is important to prepare as many concept classifiers as possible to increase the coverage of words appearing in the query sentences.

The visual-semantic embedding approach maps visual and semantic features onto a common space. Visual-semantic embedding approaches were also seen in the TRECVID benchmark and sometimes gave better results than the concept-based approach.

3. Experiments

3.1. Experimental Setup

By using the query sentences used in the TRECVID 2017 benchmark, we performed experiments to compare concept-based and visual-semantic embedding approaches and subsequently combined the two approaches. As a concept-based approach, we used a video retrieval system, which we created for the TRECVID 2017 benchmark and achieved the best performance.

For training the visual-semantic embedding, four image caption datasets, Flickr8k, Flickr30k, MS COCO, and Conceptual Captions, were used. We used the implementation[1] of VSE++ [2] for training. We used GRU for feature extraction from query sentences and the ResNet-50, ResNet-101, and ResNet-152 models for feature extraction from images.

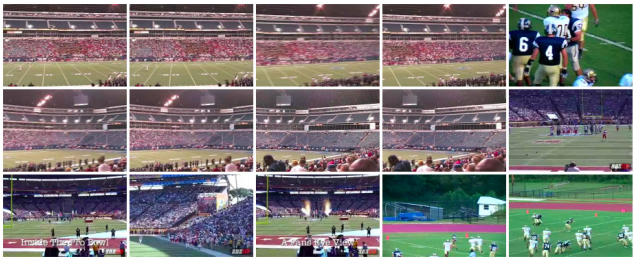
3.2. Experimental Results and Discussion

Table 1 shows the comparison results of the two methods (concept-based and visual-semantic embedding approaches) based on average precision. The concept-based approach had higher average precision than the visual-semantic embedding approach for many query sentences. On the other hand, for some query sentences, the visual-semantic embedding approach gave better results than the concept-based approach. Moreover, to check whether the concept-based and visual-semantic embedding approaches are complementary, the raking was re-computed by fusing the two approaches.

[1] https://github.com/fartashf/vsepp

Table 1. Part of 30 query sentences evaluated in the TRECVID 2017 AVS task and comparison of video retrieval performance (average precision) between concept-based and visual-semantic embedding approaches and fusion result of the two approaches.

Query ID	Query sentence	Concept	Embedding	Fusion
534	Find shots of a person talking behind a podium wearing a suit outdoors during daytime	**31.55**	2.88	24.46
535	Find shots of a person standing in front of a brick building or wall	2.21	**9.62**	8.10
538	Find shots of a crowd of people attending a football game in a stadium	26.05	10.78	**38.67**
542	Find shots of at least two planes both visible	**30.29**	12.50	19.26
546	Find shots of a male person falling down	**0.60**	0.04	0.58
548	Find shots of a chef or cook in a kitchen	28.01	36.34	**40.09**
553	Find shots of a person talking on a cell phone	1.14	**5.08**	3.60
554	Find shots of a person holding or operating a tv or movie camera	14.93	3.20	**23.07**
557	Find shots of person holding, throwing or playing with a balloon	15.00	10.82	**26.27**
558	Find shots of a person wearing a scarf	5.70	23.35	**23.59**
559	Find shots of a man and woman inside a car	61.53	72.09	**82.06**
	mAP over 30 query sentences	21.61	17.56	**24.16**



Concept-based approach

Visual-semantic embedding approach

Figure 1. Top 15 retrieval results for query ID 538.

Henceforth, we will show the results of verifying the characteristics of each approach while actually viewing the retrieved videos. There were many cases where the results after the integration of the two approaches improved, such as for query IDs 538, 548, 554, 557, and 559. From the videos retrieved at the top rank generated by each approach, we confirmed that different types of videos were retrieved by the two approaches; hence, the approaches were complementary. The result for query ID 538 is shown in Fig. 1. In the concept-based approach, concept classifiers corresponding to words such as "football" and "stadium" were selected. On the other hand, the visual-semantic embedding approach could work with the phrase "a crowd of people," which could not be acquired by the concept-based approach, and different types of videos could be retrieved.

For query IDs 534, 542, and 554, although the concept-based approach could search for appropriate videos relatively well, the visual-semantic embedding approach could not retrieve videos very well. This was because the concept-based approach could correctly detect indispensable concepts in the query sentence, while the visual-semantic embedding approach could not cover some important concepts.

Meanwhile, in some cases, the visual-semantic embedding approach was better than the concept-based approach, such as for query IDs 535, 553, and 558. For query ID 553, the concept-based method did not work well because words such as "person" and "cell phone" could be captured but phrases such as "talking on xxx" could not be handled. In the case of such a verb phrase or a query sentence, which contains positional relationship between people or objects, such as "in front of xxx," the visual-semantic embedding approach was often better than the concept-based approach.

4. SUMMARY

In this research, we compared two approaches (concept-based and visual-semantic embedding approaches) for large-scale video retrieval using query sentences and examined whether they were complementary. We revealed that the concept-based approach could accurately detect specific concepts for words appearing in a query sentence. On the other hand, in the visual-semantic embedding approach, we found that phrases including verbs, prepositions, and relationship between two objects (people and objects) were captured relatively well. We also confirmed that the video retrieval performance could be improved by integrating these two approaches because of their complementarity.

References

[1] G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, D. Joy, A. Delgado, A. F. Smeaton, Y. Graham, W. Kraaij, G. Qunot, J. Magalhaes, D. Semedo, and S. Blasi. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In *Proceedings of TRECVID 2018*. NIST, USA, 2018. 1

[2] F. Faghri, D. J. Fleet, R. Kiros, and S. Fidler. VSE++: Improved visual-semantic embeddings. *CoRR*, abs/1707.05612, 2017. 1