Abstract

In this report, we present our solution to the Video Pentathlon Challenge 2020, which contains three main contributions. Firstly, we propose a hierarchical matching method to improve fine-grained video-text retrieval. Secondly, we enhance naive nearest neighbor search with query extension and hubness mitigation during inference. Finally, we show that it is beneficial to use additional datasets in a simple multi-task training approach. Experimental results demonstrate the effectiveness of our proposed solution.

1. Introduction

The task of the Video Pentathlon challenge is to retrieve videos using natural language queries, namely text-to-video cross-modal retrieval. In order to reduce computational burden and fairly compare different retrieval models, participants in the challenge are only allowed to use provided multimodal features. The retrieval performances are evaluated across a “pentathlon” of five video-text benchmarks.

Our contributions in the challenge are three-fold:

- **Hierarchical Video-Text Matching.** Simple embeddings are insufficient to represent complicated video and text details, such as scenes, objects, actions and their compositions. Therefore, we employ the Hierarchical Graph Reasoning model (HGR) [1] which decomposes video-text matching into hierarchical levels for fine-grained retrieval.

- **Enhanced Inference Methods.** We explore query expansion and hubness mitigation methods during inference. Query expansion is to reformulate a given query and ensemble all expanded queries to improve retrieval performance. Hubness mitigation [2] aims to alleviate hubness problem in high-dimensional embedding space to improve a naive nearest neighbor search.

- **Knowledge Transfer from Additional Datasets.** Due to cross-domain discrepancy and different dataset scales, simply training with all datasets is not helpful. We systematically evaluate the cross-domain generalization performance across the five datasets and explore a simple multi-task training approach to improve retrieval performance on each target dataset.

Experimental results on the five benchmarks demonstrate the effectiveness of the above proposed approaches. We achieved the 2nd place in the pentathlon challenge.

2. Method

2.1. Hierarchical Video-Text Matching

The proposed Hierarchical Graph Reasoning model (HGR) [1] first disentangles text into a hierarchical semantic graph, including levels of event, actions, entities and interactions across levels. The event level captures global salient semantics, while actions and entities levels capture different local details. The hierarchical textual embeddings are generated via attention-based graph reasoning. Then, different textual levels are used to guide the learning of hierarchical video embeddings. Cross-modal matchings at all levels are aggregated for the final cross-modal similarities.

2.2. Query Expansion

In MSRVTT [3] and MSVD [4], there are multiple query texts for a video, which contain similar semantics and can be treated as extensions for each other. We thus average cross-modal similarities of all extensions across inference.

2.3. Hubness Mitigation

The hubness problem [2] is that some points (hubs) have high probabilities to be nearest neighbors of many other points, which are common in high-dimensional space learning. But in our task, it is preferable to retrieve different videos rather than the hub videos to different queries. Therefore, we leverage Inverted Softmax [6] to mitigate the hubness problem. It scales down the similarity \( s(t, v) \) between text \( t \) and video \( v \) if \( v \) is also close to other texts.

\[
s'(t, v) = \frac{e^{\beta s(t, v)}}{\sum_{t \in T \setminus \{t\}} e^{\beta s(t, v)}}
\]

where \( T \) denotes all text queries and we set \( \beta = 20 \).
2.4. Multi-task Training

Due to different dataset scales and cross-domain discrepancies, combining all datasets is not helpful according to our experiments. We thus explore a simple approach to transfer knowledge from other datasets. For each dataset, we select the dataset which achieves the best cross-domain performance as the additional training set. In each training epoch, we balance the examples in the additional dataset and the target dataset as multi-task training.

3. Experiments

3.1. Experimental Setups

The five video datasets used in the challenge are MSRVTT [3], MSVD [4], DiDeMo [7], ActivityNet (ANet) [8] and YouCook2 (YC2) [9]. Geometric mean of Recall@K for K={1, 5, 10} is the evaluation metric for each dataset. We use resnext101, r2p1d, audio, speech and ocr features if available and Glove word vectors [10].

3.2. Comparison of Single Models

We compare our HGR model [1] with VSE++ [11] and DualEnc [12] models. The results on validation set are presented in Table 1. Our HGR model achieves the best performances on all datasets. Because the hierarchical matching is able to better capture semantic information from global to local, the HGR model performs better than baselines especially on DiDeMo and ANet datasets whose description lengths are relatively long.

3.3. Evaluation of Inference Methods

Table 2 shows the results of using different inference methods. The query expansion (QE) method significantly improves retrieval performance in MSRVTT and MSVD datasets. However, the extended queries are based on groundtruth grouping. It is interesting to explore other techniques for query expansion such as automatically paraphrasing in real applications. The hubness mitigation (Hub) method is also effective. We will further explore mitigating the hubness problem during training.

3.4. Evaluation of Multi-task Training

We evaluate cross-dataset performance using the best HGR model in Table 1. As shown in Table 3 the MSRVTT dataset is most beneficial to other datasets, which might result from the largest training size of the MSRVTT dataset. Therefore, we use MSRVTT as the additional dataset to train with each dataset. The results are presented in Table 4, which shows that it is beneficial to employ additional datasets when the size of training set in target domain is small. However, we do not observe much improvements on MSRVTT dataset when combining with other datasets, which suggests that more effective approaches are required.

3.5. Testing Submission

We ensemble 3 - 5 models for each dataset for the final submissions. Table 5 presents results on different partitions.

4. Conclusion

In our submission to the Video Pentathlon Challenge 2020, we validate the effectiveness of our proposed hierarchical graph reasoning (HGR) model for cross-modal retrieval. We further explore different inference approaches including query expansion and hubness mitigation, and knowledge transfer approaches using multi-task training with additional selected datasets. Our solution achieves significant improvements over baselines. In our future work, we will explore mitigating the hubness problem during training and more effective transfer learning approaches for cross-domain generalization.

\footnote{On the YouCook2 dataset, the HGR model does not benefit from additional dataset as others, which might result from special styles of YouCook2 sentences (no subjects). We thus use DualEnc model instead.}
References


