UCR-VCG @ TRECVID 2018: VIDEO TO TEXT RETRIEVAL

Niluthpol C. Mithun, Amit K. Roy-Chowdhury

Department of Electrical and Computer Engineering, University of California, Riverside, USA

ABSTRACT

We participated in the video to text description: matching and ranking task in TRECVID 2018. The goal of this task is to return a ranked list of the most likely text descriptions that correspond to each video in the test set. We trained joint visual-semantic embedding models using image-text pairs from an image-captioning dataset and applied to the video-text retrieval task utilizing key frames of videos extracted by a sparse subset selection approach. Our retrieval system performed reasonably across all the testing sets. Our best system, which uses a late-fusion of similarity scores obtained from the key frames of a video, achieved mean inverted ranking score of 0.225 on the testing set C, and we ranked the 4th overall on this task.

Index Terms— Video to Text Retrieval, Joint Embedding, Ranking Loss, Subset Selection

1. INTRODUCTION

Joint embedding has a wide use case in multimedia data analysis and retrieval as it can bridge the gap between different modalities [1, 2, 3, 4, 5, 6]. Joint embeddings are learned by projecting semantically associated inputs from two or more domains into a common space (e.g., images and text) so that the embedding tends to represent the underlying correspondence of multiple domains. In this work, we focus on solving cross-modal video-text retrieval task utilizing joint image-text embeddings. In this work, we capitalized on the weighted pair-wise ranking loss mentioned in [5] for training joint image-text embeddings. The performance of the approach is evaluated using mean inverted rank (MIR) at which the annotated item is found or equivalent.

Existing video-text datasets are very small considering the diversity visual world have, and the enormous amount of rich description human can compose. Our retrieval approach is based on joint embeddings trained on image-captioning datasets, which has a significantly larger size and variety compared to video-captioning datasets. We believe that models trained on image captioning sets are more likely to show higher cross-dataset generalization performance on the TRECVID 2018 test set, compared to training with video captioning datasets consisting of a smaller number of examples. Moreover, the TRECVID test set contains short Vine videos and a few key frames are often enough to summarize most of the videos. In this work, we utilize a fixed number of key frames extracted from each of the videos and employed joint image-text embedding model for the retrieval task using the frames.

2. SYSTEM OVERVIEW

We consider the problem as matching key frames from video and text descriptions in a joint image-text embedding space following [7]. We adopt the approach proposed in [5] to learn the joint embedding using image captioning dataset MSCOCO [8]. At the time of retrieval, given key frames from a query video, we calculate similarity score for each of the frames with the all the sentences in the dataset using the joint embedding model and use a fusion of the similarity scores for the final ranking. The key frames from the videos are extracted following dissimilarity based subset selection approach [9].

2.1. Training Joint Embedding

Joint visual-semantic embedding models are trained to project visual and textual features into a common space [3, 10, 5, 11]. The embedding is learned such that the similarity in the joint space is reflective of semantic closeness between images and their corresponding text. In this work, we followed a pair-wise ranking loss based approach for training joint space following [5]. The network is trained by minimizing a weighted ranking loss that emphasizes on hard negatives and tries to maximize the similarity between an image embedding $x^{(v)}$ and its corresponding text embedding $x^{(t)}$, and minimize similarity to the non-matching one with the highest similarity score. The optimization problem can be written as follows,

$$\min_{\theta} \sum_{x^{(v)}} L(r_v)[\alpha - S(x^{(v)}, x^{(t)}) + S(x^{(v)}, x_n^{(t)})]_+ + \sum_{x^{(t)}} L(r_t)[\alpha - S(x^{(t)}, x^{(v)}) + S(x^{(t)}, x_n^{(v)})]_+ \quad (1)$$

Here, in the Eqn.1, $[f]_+ = max(0, f)$. $L(.)$ is a weighting function. For an image embedding $x^{(v)}$, r_v is the rank of

matching sentence $x^{(t)}$ among all compared sentences. Similarly, for a text embedding $x^{(t)}$, r_t is the rank of matching image embedding $x^{(v)}$ among all compared images in the batch. The weighting function is defined as $L(r) = (1 + \beta/(N - r + 1))$, where N is the number of compared images and β is the weighting factor. Here, for a positive pair $(x^{(v)}, x^{(t)})$, the hardest negative text sample $x_n^{(t)}$ can be identified as the negative text having the highest similarity score with image embedding $x^{(v)}$ in the batch. Similarly, the hardest negative image sample $x_n^{(v)}$ can be identified as the negative image sample having the highest similarity score with $x^{(t)}$ in the batch. α is the margin value for the loss function. $S(x^{(v)}, x^{(t)})$ is defined as the similarity function to measure the similarity between the images and text in the embedding.

The embedding model is trained using pairs from MS-COCO dataset [12] using a two-branch network. One of the branches of this network takes in visual features and the other one takes in text features. In this work, Resnet152 is used for visual feature encoding [13] and a GRU-based text encoder for caption encoding [14]. To calculate the similarity between the embedded vectors, cosine similarity is used.

2.2. Key frame Extraction

Key frame extraction is another major step in our retrieval pipeline. The goal of this step is to find a small subset of representative frames from a video. The selected frames should represent the entire video and have enough variety between each other. Recently, sparse coding based techniques have been shown to be highly successful in finding an informative subset of a large number of data points [15, 9]. In this work, we adopt the approach proposed in [9], which uses a sparse coding based approach to find a representative subset of the source set to describe the target set, given pairwise relationships between the sets. Here, we consider a special case where the source and target sets are same and consider the problem of finding representatives of a set X, given pairwise dissimilarity D between the elements of X.

The problem of subset selection is formulated as a row-sparsity regularized trace minimization problem following [9], where the regularization parameter puts a trade-off between the number of representatives and the encoding cost of the original set via representatives. The algorithm ultimately finds a small set of representative frames. It also returns the confidence score, which indicates how all the frames in the original video are associated with the representative set. For the frames in a video, we extracted features using pre-trained Alexnet CNN [16]. To calculate dissimilarity score, we use Euclidean distance based measure. As we are dealing with small videos, in this work, we choose to limit the number of representatives to four.

Table 1. Model Performance on TRECVID VTT Test Sets

Method	SetA	SetB	SetC	SetD	SetE
Method-1	0.218	0.214	0.225	0.212	0.216
Method-2	0.198	0.195	0.204	0.199	0.205
Method-3	0.169	0.170	0.180	0.166	0.171
Method-4	0.168	0.163	0.166	0.164	0.165

3. RESULTS

3.1. Dataset

The TRECVID test dataset [17] contains randomly selected 1921 Vine videos. The videos are short and the duration is less than 10 seconds in most cases. Each video is annotated with sentences by 5 different annotators. We did not use the vine videos provided by NIST for training our joint embedding model. We utilize MS-COCO dataset to train our joint image-text embedding model [12]. The MSCOCO training set contains about 82K images and each image in MSCOCO comes with 5 captions.

3.2. Video-Text Retrieval Performance

We submitted four runs for each matching task. Our four submitted runs were based on results obtained using the key frames extracted from the videos. Method-1 uses scores obtained by averaging similarity scores from the key frames of a video for ranking. Method-2 uses the maximum similarity score obtained by the key frames for ranking a video. Method-3 reports MIR obtained by key frame 2 and Method-4 reports MIR obtained by key frame 4. Table 1 reports the performance of our approach on the TRECVID video to text (VTT) dataset on annotation set A, B, C, D, and E. We observe that our method performs consistently across test sets. We also observe that Method-1 performs best across training sets, where we use the average of scores obtained by all key frames of a video for the final ranking.

4. CONCLUSION

This work focused on utilizing visual-semantic embedding models for TRECVID video to text matching and ranking task. We propose an approach that employs a joint image-text embedding model for the task utilizing a few key frames extracted from the videos. Experiments on TRECVID 2018 test sets demonstrate that our simple yet efficient approach is promising as it consistently achieves performance comparable to the state-of-the-art methods.

5. REFERENCES

[1] Y. Li, H. Su, Charles Ruizhongtai Qi, Noa Fish, Daniel Cohen-Or, and Leonidas J Guibas, "Joint embeddings of shapes and images via cnn image purification.," *ACM Trans. Graphics*, vol. 34, no. 6, pp. 234–1, 2015.

[2] W. Y. Zou, R. Socher, D. Cer, and C. D. Manning, "Bilingual word embeddings for phrase-based machine translation," in *EMNLP*, 2013, pp. 1393–1398.

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

[4] N. C. Mithun, R. Panda, E. Papalexakis, and A. K. Roy-Chowdhury, "Webly supervised joint embedding for cross-modal image-text retrieval," in *ACM Multimedia*, 2018.

[5] N. C. Mithun, J. Li, F. Metze, and A. K. Roy-Chowdhury, "Learning joint embedding with multimodal cues for cross-modal video-text retrieval," in *ACM ICMR*, 2018.

[6] N. C. Mithun, S. Paul, and A. K. Roy-Chowdhury, "Weakly supervised video moment retrieval from text queries," in *CVPR*, 2019.

[7] N. C. Mithun, J. Li, F. Metze, A. K Roy-Chowdhury, and S. Das, "Cmu-ucr-bosch trecvid 2017: Video to text retrieval," in *TRECVID 2017 Workshop*, 2017, vol. 4.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[9] E. Elhamifar, G. Sapiro, and S S. Sastry, "Dissimilarity-based sparse subset selection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 38, no. 11, pp. 2182–2197, 2016.

[10] F. Faghri, D. J. Fleet, R. Kiros, and S. Fidler, "VSE++: improved visual-semantic embeddings," *CoRR*, vol. abs/1707.05612, 2017.

[11] N. C. Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury, "Joint embeddings with multimodal cues for video-text retrieval," *International Journal of Multimedia Information Retrieval*, pp. 1–16, 2019.

[12] X. Chen, H. Fang, T. Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[14] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[15] N. C. Mithun, R. Panda, and A. K Roy-Chowdhury, "Generating diverse image datasets with limited labeling," in *ACM Multimedia*, 2016, pp. 566–570.

[16] A. Krizhevsky, I. Sutskever, and G E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[17] George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, David Joy, Andrew Delgado, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges Qunot, Joao Magalhaes, David Semedo, and Saverio 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.