CMU-UCR-BOSCH @ TRECVID 2017: VIDEO TO TEXT RETRIEVAL

Niluthpol C. Mithun[¶][*], *Juncheng B Li*[†][‡][*], *Florian Metze*[‡], *Amit K. Roy-Chowdhury*[¶], *Samarjit Das*[†]

[†] Robert Bosch LLC, Research and Technology Center, USA
[‡]Language Technology Institute, Carnegie Mellon University, Pittsburgh, USA
[¶]Electrical and Computer Engineering, University of California, Riverside, USA

ABSTRACT

We participated in the matching and ranking subtask in TRECVid challenge 2017. The task here was to return a ranked list of the most likely text descriptions that correspond to each video. We adopted a joint visual semantic embedding approach for image-text retrieval and applied to the video-text retrieval task utilizing key-frames extracted by dissimilarity-based sparse subset selection approach. We trained our system on the MS-COCO dataset and tested on the TRECVid dataset. Our approach got an average mean inverted ranking score of 0.255 across 4 sets of testing data, and we ranked the 3rd overall in the challenge on this task.

Index Terms— Video to Text Retrieval, Visual-Semantic Embedding, Cross-Modal Deep Learning

1. INTRODUCTION

Joint embedding has wide use case in multimedia data mining. It enables us to combine the understanding of different modalities together. There are lots of previous work done on this topic; for instance, [1] used shape and image together, while [2] combined embeddings from multiple languages. Joint embeddings are usually done by mapping semantically associated inputs from two or more domains into a common vector space (e.g., images and text). Thus, the joint embedding space tends to better represent the underlying correspondence of multiple domains.

In this work, we mainly focus on learning visual-semantic embeddings, which is crucial to the video-text retrieval task [3]. As is common in information retrieval, we measure performance by mean inverted ranking (MIR)[1] - the fraction of queries for which the correct item is retrieved in the closet point to the query. We capitalized on the performance gain by using the pair-wise ranking loss mentioned in [4], and we also adopted the more powerful image encoder in this work.

Along with the visual-semantic embedding model, we take advantage of the video data which contains dynamic information compared to the static images. We attempt to use multiple frames from the same video to reflect more interactions of the objects in the video rather than focusing on an instant moment. This is proven to be effective empirically and it gave us average 0.04 points gain in MIR benchmark when using different key-frames from the videos.

2. SYSTEM OVERVIEW

Our approach is developed mainly based on two major observations. First, both the size and the variety of image-captioning datasets is significantly larger and richer compared to the video captioning datasets. Hence, it is highly likely that models trained on image captioning sets will have comparatively higher cross-dataset generalization capability. Second, when retrieving a matched sentence from short videos, it is often the case that only a few key frames are enough to summarize the entire video. Hence, it may be possible to use image-text embedding in the video-to-text matching task with high accuracy utilizing a suitable method for selecting a few representative frames from the videos.

Motivated by above, we consider the problem as matching key frames from video and text descriptions in a joint visual-semantic embedded space. We adopt the approach proposed in [4] to learn the joint image-text embedding using image captioning datasets. The key frames from the videos are extracted using dissimilarity based subset selection approach proposed in [5]. A brief illustration of our proposed framework is shown in Fig. 1.

2.1. Key Frame Extraction

The goal of this step is to find an optimal subset of the frames in a video. In particular, we are trying to represent a video by selecting only a few frames which represent the entire video and still 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 [6, 5], and they also show great robustness against outliers, which not only makes them fit well in our scenario but also show superiority over other clustering methods such as K-means. In this work, we adopt a sparse coding based approach [5], which finds a representative subset of the

[*] Co-first Author

[1] http://www-nlpir.nist.gov/projects/tv2017/Tasks/vtt/

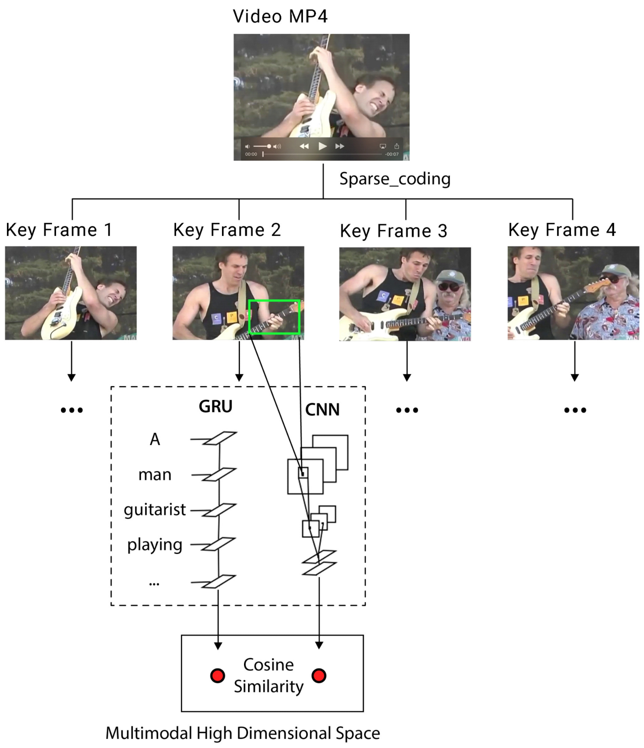

Fig. 1. A brief illustration of our proposed video to text matching method. Given a video, our method extracts a few key frames from the video. Based on these key frames, we apply joint image-text embedding to retrieve best matching text descriptions. Please see Section 2 for details.

source set to describe the target set, given pairwise relationships between two sets.

We consider a special case of the approach [5], where the source and target sets are identical and consider the problem of finding a representative subset of a set X, given pairwise dissimilarity D between the elements of X. The problem is formulated as a row-sparsity regularized trace minimization problem on Z.

$$\begin{aligned} &\min_{Z} \; tr(D^T Z) \\ &\text{s.t. } ||Z||_{2,1} \leq \tau, \; 1^T Z = 1^T, \; Z \geq 0 \end{aligned} \tag{1}$$

Here, $tr(.)$ denotes the trace operator. $X \in \mathbb{R}^{B \times N}$ is the feature matrix of all frames in a video, where $X = \{x_i \in \mathbb{R}^B, i = 1, \cdots, N\}$. Each x_i represents the feature descriptor of a frame in a video in B-dimensional feature space. N denotes the number of frames in the video. $Z \in \mathbb{R}^{N \times N}$ is the sparse coefficient matrix, where $Z = \{z_{ij}\}_{i=1,\cdots,N}^{j=1,\cdots,N}$. $\|Z\|_{2,1} \triangleq \sum_{i=1}^{N} \|z_i\|_2$ is the row sparsity regularizer, i.e., sum of l_2 norms of the rows of Z. $D \in \mathbb{R}^{N \times N}$ is the dissimilarity matrix, where $D = \{d_{ij}\}_{i=1,\cdots,N}^{j=1,\cdots,N}$. Here, d_{ij} indicates how well x_i represents x_j and a smaller value of d_{ij} indicates that x_i can represent x_j well. In Eq.1, unknown variable z_{ij} is associated with dissimilarity score d_{ij}. The regularization parameter $\tau(\tau \geq 0)$ puts a trade-off between the number of representatives and the encoding cost of the original set via representatives [5].

In this work, we extracted features from the frames in videos using pre-trained CNN model Alexnet [7]. To calculate dissimilarity score, we use Euclidean distance based measure. Minimization of Eq. 1 leads to a sparse solution for Z in terms of rows, i.e., Z contains few nonzero rows which constitute the representative set. As the TRECVid dataset contains mostly short videos of 4 seconds long, in this work, we choose to fix the number of representatives as 4 for all videos.

2.2. Visual Semantic Embedding

Joint visual-semantic embedding models project visual and textual features into a common space [3, 4]. It is expected that in the joint space, the similarity is reflective of semantic closeness between images and their corresponding text. In this work, we followed pair-wise ranking loss based joint image-text embedding approach proposed in [4]. The network is trained by minimizing a 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

$$\begin{aligned} \min_{\theta} \sum_{x^{(v)}} &[\alpha - S(x^{(v)}, x^{(t)}) + S(x^{(v)}, x_n^{(t)})]_+ + \\ \sum_{x^{(t)}} &[\alpha - S(x^{(t)}, x^{(v)}) + S(x^{(t)}, x_n^{(v)})]_+ \end{aligned} \tag{2}$$

where, $[f]_+ = max(0, f)$. Here, for a positive pair $(x^{(v)}, x^{(t)})$, the hardest negative text sample $x_n^{(t)}$ can be identified as the negative text sample having the highest similarity score with $x^{(v)}$ 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 a batch. α is the margin value for the pairwise ranking loss. The scoring function $S(x^{(v)}, x^{(t)})$ is defined as the similarity function to measure the similarity between the embedded images and text.

To encode image and text, the embedding model is trained using image-text pairs from MS-COCO dataset [8]. One of the branches of this network takes in visual features and the other one takes in text features. In this work, we used the trained joint embedding model [4], where Resnet152 model is used for visual feature encoding [9] and a GRU-based text encoder for caption encoding [10]. We used cosine similarity to calculate similarity between the embedded vectors of frames and text descriptions.

3. RESULTS

3.1. Dataset & Model Parameters

There was no training data provided by NIST for the Video to Text matching task. Hence, we utilize MS-COCO dataset to train our joint embedding model [8]. Following [11], we use their splits in MS-COCO. In this split, the training set contains 82,783 images, 5000 validation, and 5000 test images. However, there are also 30,504 images that were originally in the validation set of MS-COCO but have been left out in this split. Each image comes with 5 captions. The hyper-parameters are chosen following [4].

The TRECVid dataset [12] contains randomly selected 1880 Vine videos. The videos are short and about 6 seconds in duration. Each video is annotated with sentences by 2-5 different annotators. We use the TRECVid dataset for testing. The organizers provide four test sets for this task, denoted as set 2, set 3, set 4 and set 5. The test sets have 1,613, 795, 388 and 159 videos respectively. The sets are named based on the number of captions associated with the videos in the test set.

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 keyframes extracted from the videos. Table 1 shows the performance of our approach on the TRECVid VTT test datasets. Here, we report only the best MIR result achieved on the 4 test sets for the keyframes. Note that, the keyframes are named based on their relative appearance in the video as shown in Fig. 1. The performance can be improved further by an ensemble of the four models.

Model	TestSet2	TestSet3	TestSet4	TestSet5
KeyFrame1	0.130	0.190	0.265	0.431
KeyFrame2	0.129	0.192	0.265	0.430
KeyFrame3	0.127	0.186	0.261	0.406
KeyFrame4	0.125	0.190	0.256	0.396

Table 1: Model performance on TRECVid VTT17 dataset.

4. CONCLUSION

This work focused on learning visual-semantic embedding for cross-modal, video-caption retrieval. We propose an approach that employs a joint image-text embedding model for the task utilizing key-frames extracted from the videos. Experiments on four TRECVid test dataset demonstrate that our proposed approach can consistently achieve state-of-the-art performance.

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

[5] 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.

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

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

[8] 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.

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

[10] 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.

[11] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *CVPR*, 2015, pp. 3128–3137.

[12] G. Awad, A. Butt, J. Fiscus, D. Joy, A. Delgado, M. Michel, A. F. Smeaton, Y. Graham, W. Kraaij, G. Qunot, M. Eskevich, R. Ordelman, G. J. F. Jones, and B. Huet, “Trecvid 2017: Evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking,” in *Proceedings of TRECVID 2017*. NIST, USA, 2017.