Shandong Normal University in the VTT Tasks at TRECVID 2018

Chengcheng Liu, Min Gao, Jun Wang, Liming Zou, En Yu, Jiande Sun, Lingchen Gu, Xiaoxi Liu, Yu Lu, Yu Zhang, Xueqi Zhao

Abstract

The SDNU_MMSys from Shandong Normal University participated the Video to Text(VTT) task in TRECVID 2018 [1] including both Matching and Description Generation sub-tasks. In the Matching sub-task, we used two datasets, i.e., Wikipedia and Pascal Sentence, for training and used the pre-trained cross-modal retrieval method to match each video to the most relevant text description. We combined the Inception-ResNet V2 [2] and a two-layer LSTM model [3] to generate the description in the Description Generation sub-task. The MSR-VTT dataset was used for training and our model achieved good results.

1 Introduction

A team of master students from Shandong Normal University (SDNU_MMSys@Multi-Media System Lab) took part in the VTT task of TRECVID 2018 and completed the two sub-tasks, i.e., Matching and Description Generation in VTT. Specifically, in the Matching and Ranking sub-task, a ranked list of the most likely text description for each video is required to be fed back, and the text description should correspond (was annotated) to the video from each of the ground truth sets. In the Description Generation sub-task, a textual description (one sentence) is requires to be generated for each video independently without taking the existence of the ground truth sets into consideration. And we submitted three runs for each sub-task. These will be detailed in this paper.

2 Matching and Ranking

2.1 Dataset

Wikipedia and Pascal Sentence datasets are used for training. We extracted 1536 dimensional CNN features for images and 100 dimensional sentence2vector [4] features for texts. Similarly, as for the VTT test dataset, we first selected out one keyframe every two seconds from each video, and then extracted the corresponding features from the keyframe and did the same training as the training dataset.

2.2 Method

The cross-modal retrieval method aims to learn a couple of mapping matrices and projects different modality features into a common latent subspace [5][6][7], where the similarity between them can be measured directly. If we denote the feature matrices of images and texts as $X = [x_1, \dots, x_n] \in R^{p \times n}$ and $T = [t_1, \dots, t_n] \in R^{q \times n}$, respectively, the objective function can be defined as:

$$\min_{U,V} f(U,V) = C(U,V) + L(U,V) + N(U,V) \quad (1)$$

The first term is a linear regression term for keeping the closeness of the image data with the same semantics in the common latent subspace. The second term is a correlation analysis term for keeping closeness of pair-wise in the common latent space, and the third one is the $l_{2,1}$-norm regularization term for discriminative and informative feature selection [8]. In details, formula (1) can be described as:

$$\begin{cases} \min_{U,V} f(U,V) = \beta \|X^T U - Y\|_F^2 + (1-\beta)\|X^T U - T^T V\|_F^2 + \lambda_1 Tr(U^T R_1 U) + \lambda_2 Tr(V^T R_2 V) \\ 0 \le \beta \le 1 \end{cases} \quad (2)$$

Where U and V are the mapping matrices, Y is the semantic matrix, and λ_1 and λ_2 are the parameters for balancing the two regularization terms.

Given the cross-modal retrieval model, we can use the processed VTT dataset as the input to obtain the rank list, and the framework is shown in Fig 1.

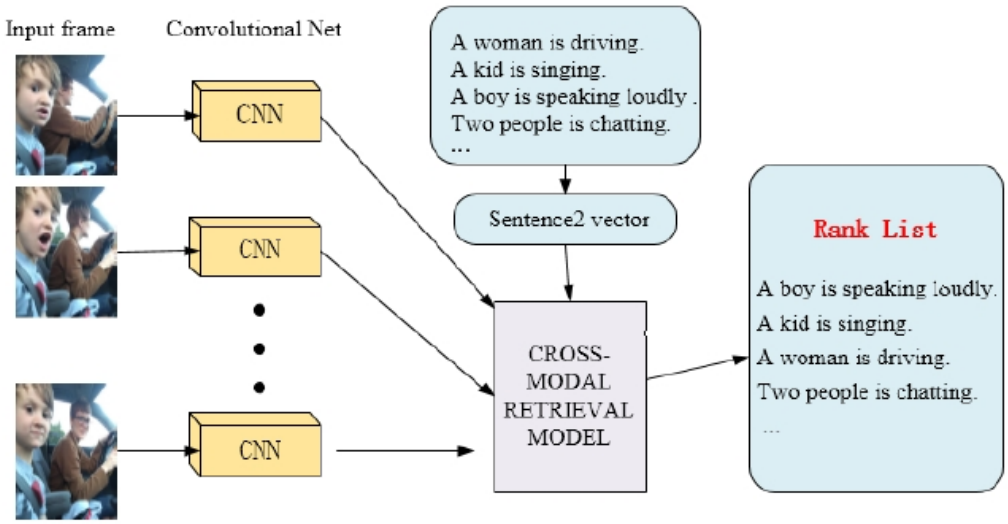

Fig 1. The frame of Matching and Ranking model

2.3 Evaluation Results

There are the results based on “A”, “B”, "C", “D” and “E” textual descriptions across 1903 test video URLs. The result consists of two parts:

1. The Mean inverted rank of our method

   Matching and Ranking sub-task mean Inverted Rank Results are shown in Table 1.

Table 1. Matching and Ranking sub-task mean Inverted Rank Results of our submissions

Run ID	Mean inverted rank
N_MMsys_CCMIP.run.A.txt	0.003
N_MMsys_CCMIP.run.B.txt	0.003
N_MMsys_CCMIP.run.C.txt	0.003

N_MMsys_CCMIP.run.D.txt	0.003
N_MMsys_CCMIP.run.E.txt	0.003

2. The rank of found matches by URL_id

The most matching item is ranked 1, and the larger the number of items ranked, the worse the matching effect. The rank results are shown in Fig 2.

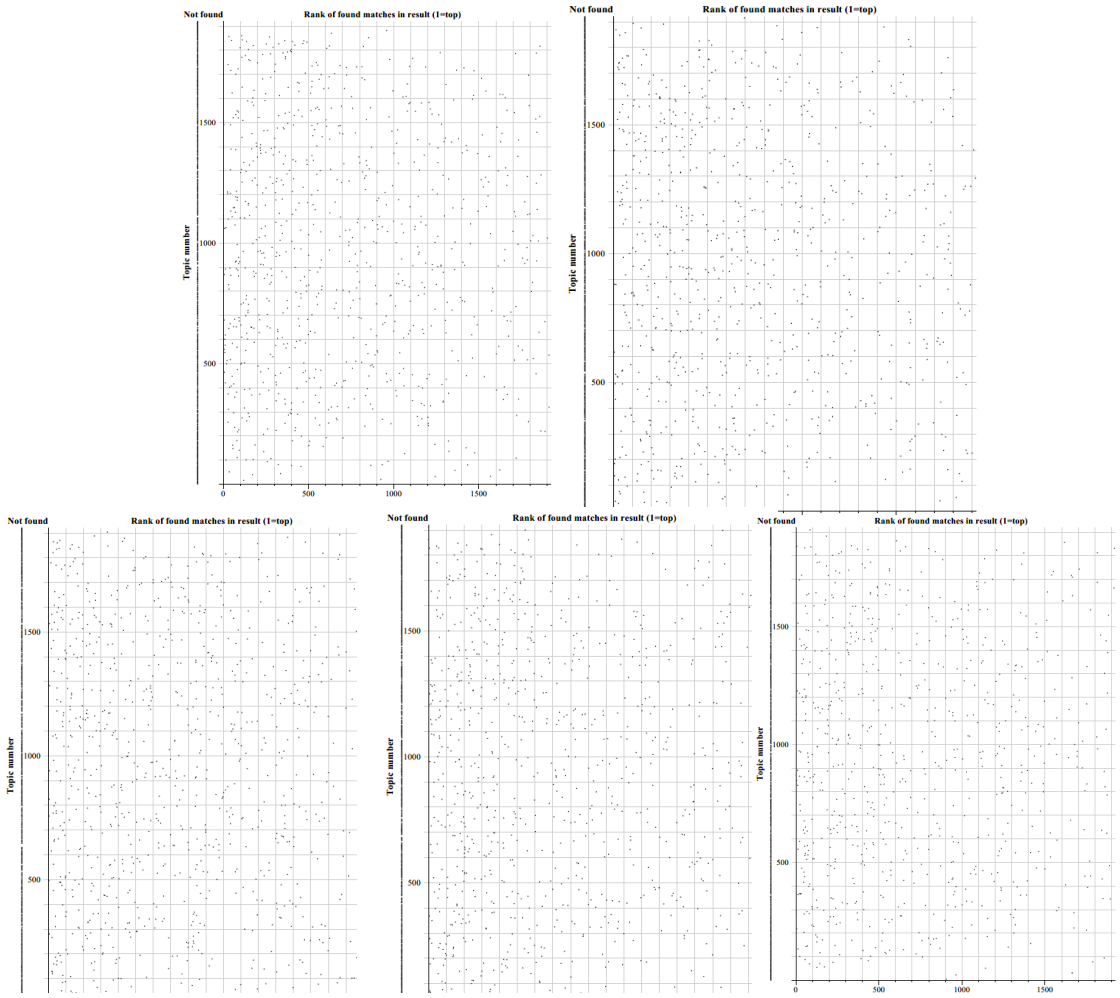

Fig 2. Matching and Ranking sub-task rank of found matches by URL_id. The above sub-figures, from top to bottom, from left to right, are “A”, “B”, “C”, “D” and “E”.

3 Description Generation

3.1 Dataset

MSR-VTT dataset [9]: The dataset is provided by Microsoft Research and provides 10K web video clips with 41.2 hours and 200K clip-sentence pairs in total. It is one of the largest datasets in terms of sentence and vocabulary, which covers the most comprehensive categories and diverse visual contents. Each clip in MSR-VTT is annotated with about 20 natural sentences by 1327 AMT workers. In addition, the MSR-VTT also provides the category information for each video (totally 20 categories).

The category information is the priori knowledge and can be referred in the test set.

3.2 Method

We trained video captioning models with MSR-VTT dataset. We extracted one keyframe per second for each video. Then we used the pre-trained Inception-ResNetV2 network to extract the features of these keyframes. Meanwhile we extracted the sen2vec features for the descriptions. Finally, we trained the model with the frame features and text features, just as shown in Fig 3.

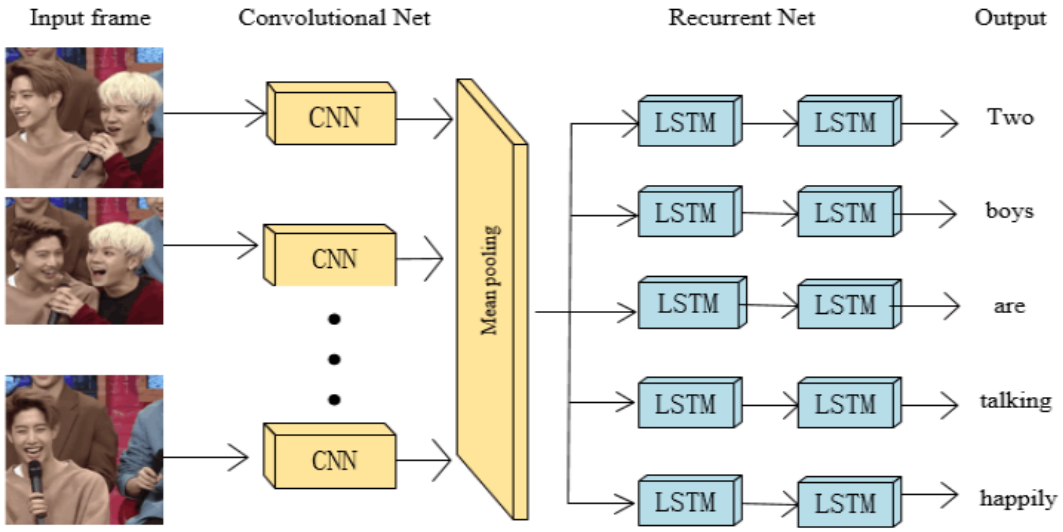

Fig 3. The training of Description Generation Model

3.3 Evaluation Results

The review results of the Description Generation sub-task are shown in Table 2.

Table 2. Description Generation sub-task Evaluation Results of our submissions

Run ID	Final score
MMsys_CCMIP.run1.txt	0.1215508777846736
MMsys_CCMIP.run2.txt	0.12012859738833576
MMsys_CCMIP.run3.txt	0.11632938865584741
MMsys_CCMIP.run4.txt	0.11385118054388724

4 Conclusions

We tested our ideas in cross-modal retrieval and video caption generation through the VTT tasks. We found some potential improvements in the future work. The task-driven semantic description can be our next focus, through which we hope to promote the performance in VTT task.

References

[1] George Awad, Asad Butt, Keith Curtis, Jonathan Fiscus, Afzal Godil, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges Quénot, Joao Magalhaes, David Semedo, Saverio Blasi. TRECVID 2018: Benchmarking Video Activity Detection, Video Captioning and Matching, Video Storytelling Linking and Video Search. 2018.

[2] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//AAAI. 2017, 4: 12.

[3] Venugopalan S, Rohrbach M, Donahue J, et al. Sequence to sequence-video to text[C]//Proceedings of the IEEE international conference on computer vision. 2015: 4534-4542.

[4] Saha T K, Joty S, Al Hasan M. C on-S2V: A Generic Framework for Incorporating Extra-Sentential Context into Sen2Vec[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2017: 753-769.

[5] En Yu, Jiande Sun, Jing Li, Xiaojun Chang, Xianhua Han, Alexander G. Hauptmann. Adaptive Semi-supervised Feature Selection for Cross-modal Retrieval[J]. IEEE Transactions on Multimedia, 2018, DOI: https://doi.org/10.1109/TMM.2018.2877127

[6] Wang K, He R, Wang L, et al. Joint feature selection and subspace learning for cross-modal retrieval[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 38(10): 2010-2023.

[7] Gong Y, Ke Q, Isard M, et al. A multi-view embedding space for modeling internet images, tags, and their semantics[J]. International journal of computer vision, 2014, 106(2): 210-233.

[8] He R, Tan T, Wang L, et al. l 2, 1 regularized correntropy for robust feature selection[C]//Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012: 2504-2511.

[9] Xu J, Mei T, Yao T, et al. Msr-vtt: A large video description dataset for bridging video and language[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 5288-5296.