Shandong Normal University in the VTT Tasks at TRECVID 2017

En Yu, Min Gao, Yafei Li, Xiao Dong, Jiande Sun

School of Information Science and Engineering,
Shandong Normal University,
Jinan 250014, Shandong Province, China

Abstract

SDNU_MMSys from Shandong Normal University participated the Video to Text(VTT) task in TRECVID 2017 including both Matching and Description Generation sub-tasks. We used our cross-modal retrieval model in the Matching sub-task. In the Description Generation sub-task, we combined the Inception V3 [1] and a two-layer LSTM [2] model to generate the description for each video.

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 2017 for first time and completed the two sub-tasks, i.e., Matching and Description Generation in VTT[3]. 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 text description (one sentence) is requires to be generated for each video independently without taking into consideration the existence of the ground truth sets. And we submitted three runs for each sub-task. These will be described in this paper.

2 Video to Text(VTT)

2.1 Matching and Ranking

2.1.1 Dataset

Wikipedia and Pascal Sentence are used for training. There are 10 categories, and totally 2866 image-text pairs in Wikipedia. And There are 20 categories in Pascal Sentence with 50 image-text pairs in each category. We extracted 2048 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 per second from each video, and then extract the corresponding features from the keyframe and ground-truth sentence as what is done in training datasets respectively.

2.1.2 Method

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

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

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 one 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 coupled feature selection [8]. In details, (1) can be:

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

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.

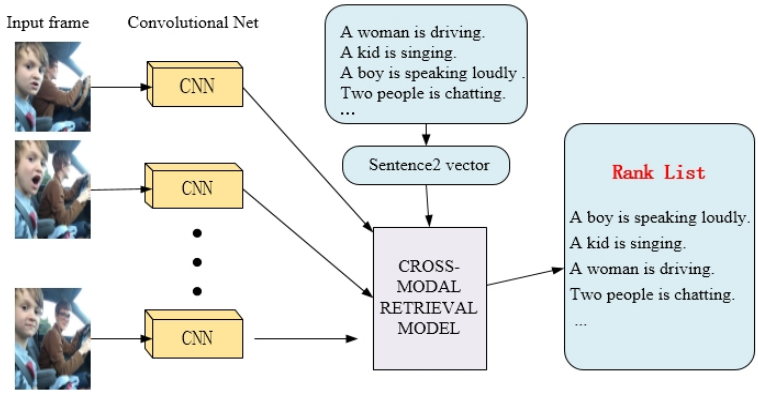

Fig 1. The frame of Matching and Ranking model

2.2 Description Generation

2.2.1 Dataset

MSR-VTT dataset [9]: MSR-VTT 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 the largest dataset in terms of sentence and vocabulary, which covers the most comprehensive categories and diverse visual content. 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 known in the test set. At the same time, each video contains audio information, though we didn't use audio in our method.

2.2.2 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 V3 CNN 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 as shown in Fig 2.

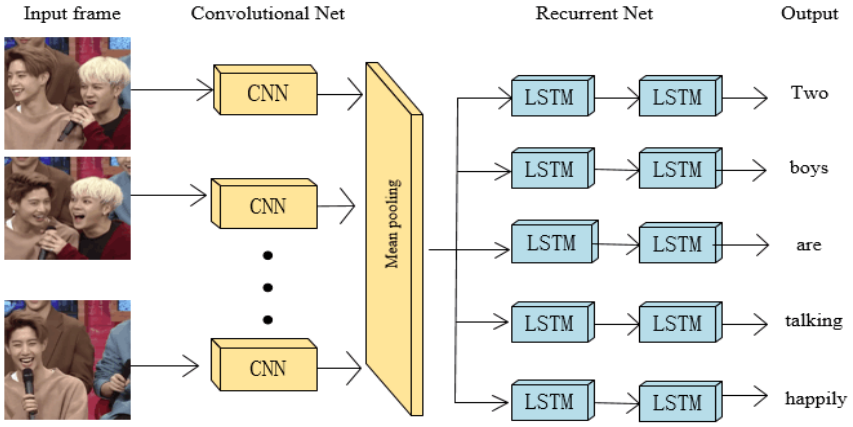

Fig 2. The Training of Description Generation Model

3 Evaluation Result

3.1 Matching and Ranking

Test.2	Run.A	Run.B
Set.A	0.008	0.007
Set.B	0.005	0.005

Test.3	Run.A	Run.B
Set.A	0.007	0.011
Set.B	0.007	0.005
Set.C	0.008	0.007

Test.4	Run.A	Run.B
Set.A	0.014	0.017
Set.B	0.016	0.013
Set.C	0.018	0.012
Set.D	0.016	0.013

Test.4	Run.A	Run.B
Set.A	0.032	0.030
Set.B	0.045	0.043
Set.C	0.045	0.043
Set.D	0.029	0.043
Set.E	0.038	0.044

Table 1. Matching and Ranking sub-task meanInvertedRank Results of our submissions

3.2 Description Generation

run.A	BELU	METEOR	CIDEr	CIDErD
Test2	0.0033	0.1362	0.144	0.094
Test3	0.0047	0.1456		
Test4	0.0061	0.1554		
Test5	0.0115	0.1729		

run.B	BELU	METEOR	CIDEr	CIDErD
Test2	0.00330067236697962	0.14164882752844563	0.127	0.087
Test3	0.00412502340256523	0.15257858970629867		
Test4	0.00486457917407028	0.1562971826569037		
Test5	0.00894347912022264	0.17384566971785426		

run.C	BELU	METEOR	CIDEr	CIDErD
Test2	0.00499217141180668	0.13126911726465373	0.180	0.110
Test3	0.00788647405500112	0.14367205907915007		
Test4	0.0139808374567146	0.15611684638626638		
Test5	0.0217061101547659	0.16985316837525477		

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

4 Conclusions and Future Work

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.

Acknowledgement: This work is supported by Natural Science Foundation for Distinguished Young Scholars of Shandong Province (JQ201718), Key Research and Development Foundation of Shandong Province (2016GGX101009), the Natural Science Foundation of China (61572298, 61603225), Shandong Provincial Key Research and Development Plan (2017CXGC1504), Natural Science Foundation of China (No. 61601268), Natural Science Foundation of Shandong Province (ZR2016FB12). And we gratefully acknowledge the support of NVIDIA corporation with the donation of the TITAN X GPU used for this research.

References

[1] Szegedy C, Liu W, Jia Y, et al. Going Deeper with Convolutions[J]. 2014:1-9.

[2] Venugopalan S, Rohrbach M, Donahue J, et al. Sequence to Sequence -- Video to Text[J]. 2015.

[3] George A, Asad B, Jonathan F, et al. TRECVID 2017: Evaluating Ad-hoc and Instance Video

Search, Events Detection, Video Captioning and Hyperlinking. Proceedings of TRECVID 2017. 2017

[4] Saha T K, Joty S, Hasan M A, et al. CON-S2V: A Generic Framework for Incorporating Extra-Sentential Context into Sen2Vec[C]// ECML PKDD. 2017.

[5] Wei Y, Zhao Y, Zhu Z, et al. Modality-Dependent Cross-Media Retrieval[J]. Acm Transactions on Intelligent Systems & Technology, 2016, 7(4):57.

[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 & 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] Zheng W S, Wang L, Tan T, et al. l2, 1 Regularized correntropy for robust feature selection[J]. 2012, 157(10):2504-2511.

[9] Xu J, Mei T, Yao T, et al. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language[C]. Computer Vision and Pattern Recognition. IEEE, 2016:5288-5296.