Tianjin University and National University of Singapore at TRECVID 2017: Video to Text Description

An-An Liu[1], Yurui Qiu[1], Yongkang Wong[2], Ning Xu[1], Yuting Su[1], Mohan S. Kankanhalli[3]

[1]School of Electrical and Information Engineering, Tianjin University, China
[2]Smart Systems Institute, National University of Singapore
[3]School of Computing, National University of Singapore

ABSTRACT

During TRECVID 2017 our group participated in the video to text description (VTT) task. We explore and leverage the heterogeneous data by various configurations depending on training-validation splits to boost video captioning task. Particularly, our framework builds on the LSTM unit for sentence generation with one attention layer. The model is optimized with the combination of heterogeneous pair-wise examples. Totally, we submitted four runs for VTT task and the performance is significantly improved than baseline.

Table 1: Overview of various video caption datasets.

Dataset	Domain	Videos	Captions	Vocabulary	Duration(s)
TRECVID VTT [1]	open	3,795	3,830	2,534	$\leqslant 6$
MSVD [3]	open	1,970	80,839	13,010	$\leqslant 18$
MSR-VTT [13]	open	10,000	200,000	28,528	$\leqslant 30$
ActivityNet Captions [7]	open	19,994	100,000	-	-
M-VAD [12]	movie	48,986	55,905	18,269	$\leqslant 7$
MPII-MD [10]	movie	68,337	68,375	24,549	$\leqslant 6$
TACos [8, 11]	cooking	7,206	18,227	-	-
TACos M-L [9]	cooking	14,105	52,593	-	-

1 INTRODUCTION

In the task of video captioning, a system is required to automatically generate a natural language description based on video content.This task has a variety of potential real-world applications, such as assistance to a visually impaired person. The video captioning task is challenging since it not only determines which objects are in the video, but also captures complex interactions of actors and objects that evolve over time. Specifically, an ideal system should be able to identify the core semantic context (e.g. objects, actions, places) and express their relationships in a natural language. In the task of TRECVID 2017 VTT [1], all 1915 videos from TRECVID 2016 VTT is available as the training data for this task, where each video has two sets of single sentence ground truth description. The TRECVID 2017 VTT evaluation set consists of 1880 Vine videos. All the videos provided for this task are within approximately 6 seconds of duration.

Currently, there exits several video captioning datasets, such as MSVD [3], MSR-VTT [13], ActivityNet Captions [7], M-VAD [12], MPII-MD [10], TACos [8, 11], TACos M-L [9] and so on. In comparison, the video data in the TRECVID VTT task has some difference compared with them. As shown in Table 1, not only the number of the videos in TRECVID 2017 VTT task is much lower than most of the existing datasets, it also consists of lower number of captions where each video is described only by two captions. In addition, the total vocabulary is much less than others, as shown, the vocabulary of TRECVID VTT is 2,534, while for MSR-VTT, MSVD and MPII-MD, the number comes to 28,528, 13,010 and 24,549. In this case, it may lead to the phenomenon that some content in the video to be processed cannot be described, because the task can only generate captions by using the words that exist in the vocabulary. Furthermore, the dataset of TRECVID VTT, MSVD, MSR-VTT and ActivityNet Captions are open domain, the others are either movie or cooking domain. So we only considered the datasets which are of the same domain with TRECVID VTT in experiments, and we didn't take ActivityNet Captions into experiment because of its number of videos is too large. In our experiment, we set the proposal for only using the data provided in TRECVID VTT as the baseline. We use various combinations of datasets to explore the impact of multiple training data, and observe the effectiveness of selection of validation.

The remaining of the paper is structured as follows. Section 2 briefly overview the selected video captioning model. The detailed of dataset configuration, results (quantitative and qualitative) and discussion are shown in Section 3. Section 4 conclude this report.

2 FRAMEWORK

With the rapid development of natural language processing, the encoder-decoder-based model [4] has been applied for the generation of video descriptions. In the video captioning task the role of the encoder is to encode the videos as a set of specific size of the vector in a certain way, and the decoder decode the output of the encoder into a sequence of words. The combination of CNN and RNN as encoder-decoder performs well in video captioning task. Hence, the model we used is based on the encoder-decoder framework consisting of CNN [5] and RNN [14].

2.1 Encoder: Convolutional Neural Network

Since the success of the residual neural network at large-scale object recognition, we used a pretrained ResNet200 [5] model to obtain 2,048 dimensional frame-wise visual features, which were extracted from the 'pool5' layer. We selected 50 equally-spaced frames out of each video. In this case, every video clip was encoded into a equally-space representation $V = \{v_1, ..., v_n\}$, where n is 50 frames, and each vector v is 2,048 dimensions.

2.2 Decoder: Attention-LSTM Network

In the encoder-decoder framework, the decoder network needs to generate the corresponding output y from the encoder representation V. In the case of this video to a sequence of natural language

Table 2: Overview of various data configuration in this work.

	Training Data			Validation Data			Test Data		
	TRECVID VTT	MSR VTT	MSVD	TRECVID VTT	MSR VTT	MSVD	TRECVID VTT	MSR VTT	MSVD
Configuration 1	1,200	-	-	100	-	-	615	-	-
Configuration 2	1,200	8,000	-	100	1,000	-	615	1,000	-
Configuration 3	1,200	-	1,200	100	-	-	615	-	670
Configuration 4	1,200	-	1,200	-	-	100	615	-	670
Configuration 5	1,200	-	1,200	100	-	100	615	-	670

representation automatically, similar to the machine translation, Recurrent Neural Network is a more suitable choice as the decoder [2]. In the video caption task, since the attention mechanism does better in combining sentences or fragments with different importance to generate captions, we use attention-based long memory networks in challenge tasks. A basic LSTM [6] unit consists of a single memory cell, an input activation function, and three gates (input i_t, forget f_t and output o_t). i_t allows incoming signal to alter the state of the memory cell or block it. f_t controls what to be remembered and what to be forgotten by the cell and somehow can avoid the gradient from vanishing or exploding when back propagating through time. Finally, o_t allows the state of the memory cell to have an on other neurons or prevent it. The LSTM also has two states: memory state and hidden state. The memory state c_t depends on the previous memory state c_{t-1} and the new memory content update $\tilde{c_t}$, and the hidden state h_t equals to the element-wise multiplication of the o_t and the c_t. Attention based LSTM is fed by the input V extracted from the encoder as follows:

$$\varphi_t(V) = \sum_{i=1}^{n} \alpha_i^{(t)} v_i \tag{1}$$

$$\alpha_i^{(t)} = \frac{\exp(e_i^{(t)})}{\sum_{j=1}^{(t)} \exp(e_j^{(t)})}, s.t, \sum_{n}^{i=1} \alpha_i^{(t)} = 1 \tag{2}$$

$$e_i^{(t)} = w^{\mathrm{T}} \tanh(W_a h_{t-1} + U_a v_i + b_a) \tag{3}$$

where $\varphi_t(V)$ is computed as the dynamic weighted sum of the temporal feature vectors for the video input at time t. In addition, $\alpha_i^{(t)}$ is the attention weights at time t describing the relevance of the feature vector v_i in the input V. When given all the previously generated words, i.e. $y_1, \ldots y_{t-1}$ and the feature vector v_i of the i-th temporal feature and returns the unnormalized relevance score $e_i^{(t)}$. Where w, W_a , U_a and b_a are the parameters that are estimated together with all the other parameters of the LSTM network.

In the challenge, we are mainly concerned about evaluating the impact of different data training sets on the accuracy of the caption generation.

3 EXPERIMENT

In the task we submitted four run files, the details and the results will be discussed in the following parts.

3.1 Implementation Details

In order to verify whether additional data can boost the video captioning task, we added the MSVD dataset and the MSR-VTT dataset as the additional data. We did several configurations for the experiment to train the model, details are shown in the Table 2.

In the experiment, Configuration 1 was set as the baseline for the whole experiment, which train the model only using the data provided in the TREVID VTT task. In order to explore the impact of multiple training data, we set Configuration 2 and Configuration 3 as a contrast. To observe the effectiveness of heterogeneous validation sets, Configuration 3 to Configuration 5 perform model training on different validation, the validation sets of them are of these three proposals: data only from TRECVID VTT or MSVD, or the combination of them.

With these configurations, we want to verify whether the additional datasets and the validation crossing datasets will improve the results and enhance the accuracy of the description. We submitted four of them: Configuration 1, Configuration 3, Configuration 4 and Configuration 5 as the Runfile 1 to Runfile 4. Where the numbers in the table means the number of the videos of each dataset we used. The test set here is used to help us confirm which settings are better. When we train the model, we shuffle the videos. In the experiment, we found that the results of Configuration 2 is not as good as others, the video descriptions of the TREVID VTT test data from this model seem always composed of these words: 'a', 'man', 'someone', 'is', 'doing', 'something'. We think the reason is the fact that the amount of data in another dataset MSR-VTT is too large. So we didn't submit this run file. In the Configuration 3 and Configuration 4, we used different validation set to validate the training model. The purpose is to observe the effect of cross dataset validation on the training model when using the same set of training sets.

The additional training details: In the experiment each word in the sentence was represented as a 'one-hot' vector. The word embedding dimensionality was set to 468 and the dimension of hidden layers in attention-LSTM was 3,698. We optimized the hyperparameters with random search to maximize the log-probability on the validation set.

3.2 Results and Analysis

In the TRECVID 2017 video to text description task, 13 teams submitted the results for the description generation subtask, a total of 43 run files. Results of our description task is presented in Figure 1, according to the evaluation results. As shown in Figure 1, according to the of reference set (rf1 to rf5), the Runfile 4 achieves the competing performances on the four evaluation metrics. In this case, this points out that the Configuration 5 in Table 2 seems to be the best. While the other two configurations Configuration 3 and

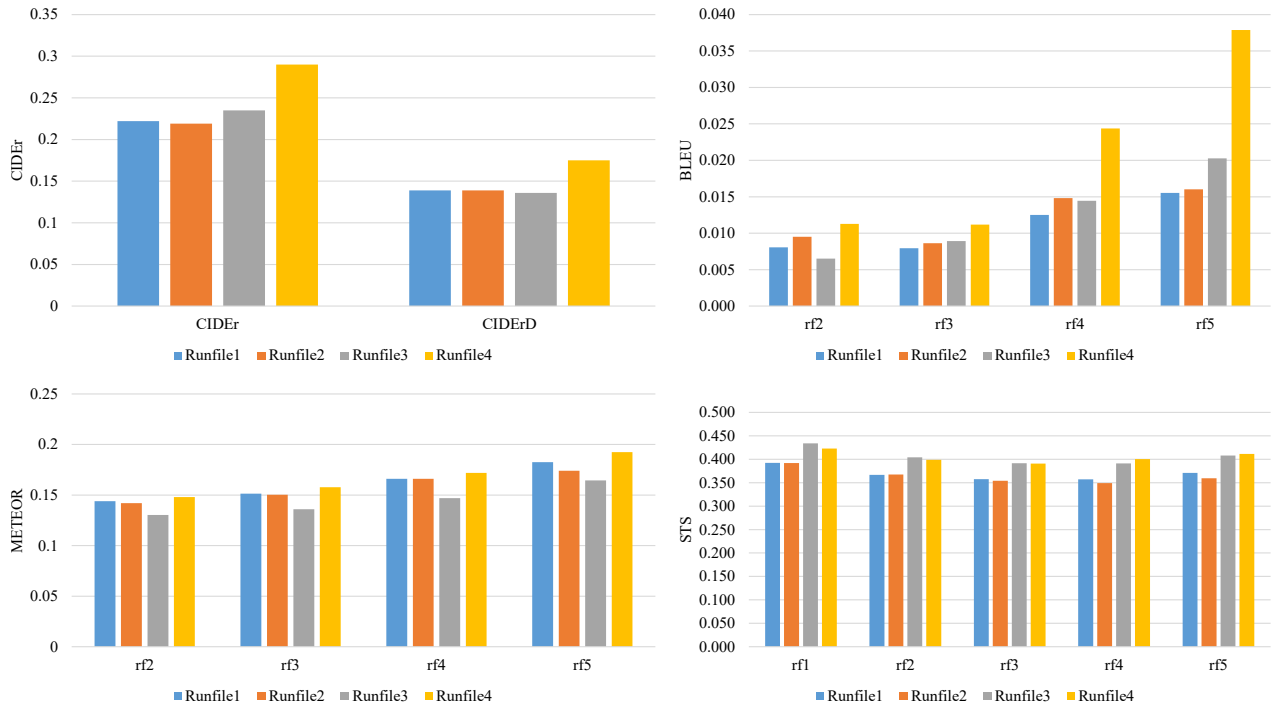

Figure 1: Evaluation results on four metrics provided by the organizer.

GT1: Basketball player jumps over man to score basket
GT2: Basketball player leaps over standing man to dunk ball in basket

Runfile1: a basketball player scores a goal on a basketball field
Runfile2: a man doing a slam dunk in a sports hall
Runfile3: a basketball game is being played
Runfile4: a basketball player is scoring a goal on a basketball court

GT1: Men are playing on a soccer field.
GT2: Soccer player running entire field and making goal

Runfile1: soccer player scoring a ball on a soccer field
Runfile2: a soccer player is scoring a goal on a soccer field
Runfile3: people are playing football
Runfile4: a soccer player scoring a goal on a soccer field

Figure 2: Examples of the video caption results.

Configuration 4 added MSVD dataset as additional data, either the use of MSVD data or TRECVID VTT data alone as the validation set, their performances are significantly worse than the use of both validation sets. This indicates that the additional data improves the accuracy of the description only when both datasets are involved in the training set and the validation set.

We show two examples of video caption results in Figure 2. The videos are used for demonstration and three frames are extracted from each video.

4 CONCLUSION

In this paper we presented our experiments performed in the TRECVID 2017 video to text description task. With the use of ResNet features and Attention-LSTM network, we mainly did several groups of experiment on whether the use of additional data can help to improve the description of videos. The participation rewarded us an experience in our researches on video caption task.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (61772359, 61472275, 61525206, 61502337, 61572356), the Tianjin Research Program of Application Foundation and Advanced Technology (15JCYBJC16200).

REFERENCES

[1] George Awad, Asad Butt, Jonathan Fiscus, David Joy, Andrew Delgado, Martial Michel, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges Qunot, Maria Eskevich, Roeland Ordelman, Gareth J. F. Jones, and Benoit Huet. 2017. TRECVID 2017: Evaluating Ad-hoc and Instance Video Search, Events Detection, Video Captioning and Hyperlinking. In *Proceedings of TRECVID 2017*. NIST, USA.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

[3] David L Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. 190–200.

[4] Kyunghyun Cho, Bart van Merriënboer, Çalar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *EMNLP*. 1724–1734.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[6] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[7] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-Captioning Events in Videos. In *ICCV*.

[8] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *TACL* 1 (2013), 25–36.

[9] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. 2014. Coherent multi-sentence video description with variable level of detail. In *GCPR*. Springer, 184–195.

[10] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *CVPR*. 3202–3212.

[11] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. 2013. Translating video content to natural language descriptions. In *ICCV*. 433–440.

[12] Atousa Torabi, Christopher J. Pal, Hugo Larochelle, and Aaron C. Courville. 2015. Using Descriptive Video Services to Create a Large Data Source for Video Annotation Research. *CoRR* abs/1503.01070 (2015).

[13] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*. 5288–5296.

[14] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *ICCV*. 4507–4515.