UTS_CAI submission at TRECVID 2017 Video to Text Description Task

Guang Li, Pingbo Pan, Yi Yang
University of Technology Sydney
15 Broadway, Ultimo NSW 2007, Australia
{Gauang.Li,Pingbo.Pan}@student.uts.edu.au, Yi.Yang@uts.edu.au

Abstract

In this paper, we aim at summarizing our experiments related to the video to text description task of TRECVID 2017 [1]. The task consists of two subtasks, i.e., Matching and Ranking, and Description generation. Our approach for description generation is based three main phases: the extraction of the high-level image feature, the aggregation of multiple image features and the sentence generation based on a probabilistic language model. For every phase, we tried several state-of-the-art techniques, and obtain the optimal combination according to the experiment results. In the matching and ranking task, we use the generated descriptions as the ground truth and rank the candidate descriptions with the similarity computed by two metrics: BLEU [9] and METEOR [3].

1. Data Collection

In the TRECVID 2017 VTT task, 1900 video clips are provided as test data. Thus, We need to collect the training data and validation data to train model in a supervised way. When it comes to the video description, there are two benchmark datasets, i.e., MSVD [4] and MSR-VTT 16 [13], we also add TRECVID 2016[2] VTT data to our collection. For every standard dataset, we randomly select 10% as evaluation set, and the rest are treated as the training set. When combining two datasets, for example, MSVD and TRECVID 2016 VTT, we just merge the corresponding parts directly.

2. Our Framework

Our system consists of three submodules for one of each in charge of high-level image feature extraction, video feature aggregation and sentence generation separately. As the advance of ConvNet, we use ResNet-152, ResNet-200 and C3D [11] as our feature extractor. In order to aggregating clip-level features, we explore three kinds of sequence model, i.e.: the plain RNN model, hierarchical RNN model as well as multi-rate RNN model. We also use RNN model as our probabilistic language model to generate the final description based on the aggregated video feature.

2.1. Clip-Level Feature Extraction

For every keyframe, we test two main architectures, ResNet [6] and C3D [11], in our experiments. The ResNet has been proved to be an excellent network in many visual tasks, such as large-scale image recognition, object detection, etc. Besides, C3D shows great potential for action recognition by leveraging the temporal information in videos through 3D convolution. In order to measure the performance of this two architectures, we built a naive sequence to sequence model as our evaluation system which only constructed by the joint of RNN encoder and RNN decoder. And the experiments show that ResNet has more stable performance than the C3D network. And intuitively, the deeper of the network, the better of the performance. In our submissions, we use the ResNet-200 trained by Facebook AI Research [1] as our basic feature extractor.

2.2. Video Feature Aggregation

Video can be treated as a sequence of keyframes. Integrating this point of view with the advance in sequence model, we tried to use Recurrent Neural Networks to leverage the temporal information between frames. The great capacity of RNN model in aggregation temporal information has been proved in [10], we investigated two advanced architecture i.e., HRNN [8] and MVRM [14].

Hierarchical Recurrent Neural Encoder The Hierarchical Recurrent Neural Encoder (HRNE) [8] investigates the temporal information trough stacking multi-layer RNN in different granularities. In the higher layer, the Recurrent Neural Network has fewer units, thus, it is able to exploit video temporal structure in a lower sampling rate which helps in long-range information propagation as well as efficiency.

Multi-rate Visual Recurrent Model In video processing, frame sampling rate should vary in accordance with

[1] https://github.com/facebook/fb.resnet.torch

Table 1. The performances on test set

	CIDER	CIDER-D	STAT	METEOR
RUN_1	0.239	0.144	0.400	0.159
RUN_2	0.272	0.158	0.402	0.152
RUN_3	0.270	0.158	0.398	0.162
RUN_4	0.161	0.112	0.319	0.160

different motion speed. The fast motion should have a slow sampling rate to obtain the accurate information easily. The Multi-rate visual recurrent model (MVRM) [14] obtains the capability of dealing with motion speed variance trough encoding frames of a clip with different intervals.

2.3. Probabilistic Language Model

After grasping the uniform gist of the input video, we then input the aggregated feature into a language model, which model the conditional probability through Recurrent Neural Network whose unit is GRU [5]. For the sake of simplicity, we didn't apply attention mechanism to our model. As in many previous works [7] about word dense representation, we set the word embedding width to be 300.

3. Submitted Runs

We submitted four runs on the Video To Text Description task. The methods are ranked by their performances on corresponding evaluation set. And the performances on test set are listed in Table. 1

In *RUN_1*, the model is trained on MSR VTT 16 only, and HRNE is used to aggregate frame-level features. It achieves CINDER [12] of 0.239, and METEOR of 0.159.

In *RUN_2*, the model is trained on the combined dataset of MSVD and TRECVIDVTT16. MVRM is applied to aggregate frame-level features. It achieves CINDER of 0.272, and METEOR of 0.152.

In *RUN_3*, the model is trained on the combined dataset of MSR VTT 16 and TRECVID VTT 16. MVRM is applied to aggregate frame-level features. It achieves CINDER of 0.270, and METEOR of 0.162.

In *RUN_4*, the model is different from the former ones. We train an image captioning model using MSCOCO, then the model is employed to generate caption for every each of the frame in a video. Finally, we use LexRank to retrieve the highest ranking sentence as the caption for the video.

References

[1] 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.

[2] G. Awad, J. Fiscus, M. Michel, D. Joy, W. Kraaij, A. F. Smeaton, G. Quénot, M. Eskevich, G. J. Jones, R. Ordelman, et al. Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking. *Proceedings of TRECVID 2016*, 32:14, 2016.

[3] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, page 65, 2005.

[4] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200. Association for Computational Linguistics, 2011.

[5] J. Chung and y. v. aglar Gülehre and Kyunghyun Cho and Yoshua Bengio, journal=CoRR. Empirical evaluation of gated recurrent neural networks on sequence modeling.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[7] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[8] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1029–1038, 2016.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[10] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[11] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. C3d: Generic features for video analysis. *CoRR*, abs/1412.0767, 2014.

[12] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[13] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language supplementary material. 2016.

[14] L. Zhu, Z. Xu, and Y. Yang. Bidirectional multirate reconstruction for temporal modeling in videos. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.