UTS_CETC_D2DCRC Submission at the TRECVID 2018 Video to Text Description Task

Guang Li
University of Technology Sydney, Australia
Gauang.Li@student.uts.edu.au

Ziwei Wang
Information Science Academy, CETC, CHINA
wangziwei26@126.com

Yi Yang
University of Technology Sydney, Australia
Yi.Yang@uts.edu.au

Abstract

In this paper, we report our methods on the video to text description task of TRECVID 2018[1]. The task consists of two subtasks, i.e., Description generation and Matching & Ranking. In the description generation subtask, because no standard training data provided, we principally focused on saturating the generalization ability of our model. Instead of exploring complex models, we investigated the widely used LSTM based sequence to sequence model[10] and some of its variants, which are simple yet robust enough. Besides, we also reviewed some training strategies to expand the generalization ability of our model. In the matching and ranking subtask, we designed a two-branch deep model[6] to embed visual content and semantic content respectively. The model helps to project the information from different modalities into the common embedding space. Further, we examined some metric learning losses, like triplet loss and its variants, in our experiments.

1. Data Collection

In the TRECVID 2018 VTT task, there are 1921 video clips were provided as testing data and no standard training data. The testing data were collected from Twitter Vine service. Each has an average duration of 6 seconds and has been annotated Y times (where $Y \leq 5$) by different annotators. The annotators were asked to include four facets (who, what, where, when) of the video in one sentence.

We collected our training data from five datasets: MSVD[5], MSR-VTT 2016[12], TGIF[9] and the 2016/2017 testing data of VTT[2][3]. According to the language style and video collection source, we divide the datasets into three groups.

1. The MSVD and MSRVTT 2016 construct the first group, and we refer it as **YouTube dataset**, because they all collected from YouTube and shared similar language style. 2. **TGIF dataset**, which consists of a large number of GIF images and whose annotations followed similar collection rules as the TRECVID data, forms the second group. 3. We refer the old VTT testing data in 2016 and 2017 as the **Vine dataset**.

To compare the effect of training data, we used the TRECVID VTT 2017 data as the testing set, and for each dataset, 10% data was randomly selected to perform model selection. In our experiments, it shows that the TGIF dataset had the best generalization ability. Although the Vine data is of the same domain with the 2018 VTT testing data, our model had difficulties in grasping the gist of the data, and the generated captioning is far from being a natural sentence. What's more, we also tried to fine-tune the well-trained models on YouTube dataset and TGIF dataset with the 2016 VTT data, and they show a certain kind of degeneration in 2017 VTT data. Consequently, in our final submission, we only ensembled the models trained on Non-Vine datasets.

2. Our Framework

Because no standard training data was provided, and the public available large video captioning datasets are of different sources to the VTT testing data, and the captions have different locution. The essential aspect is to improve the generalization ability of the model or learn a good model transferring from the source domain to the target domain. In this work, we focused on enhancing the generalization ability on unseen data. Thereby, we selected the efficient yet straightforward sequence to sequence model as our primary model. We will introduce some variants of the primary model structure and training strategies in the follow-

ing sections.

2.1. Feature Extraction

Video can naturally be decomposed into spatial and temporal components. For the spatial feature, we investigated ResNet[7] based network structure, such as ResNet200, DenseNet[8], and ResNeXt[11]. Although the RNN encoder network is considered to aggregating the long-range temporal information, we still need the 3D convolutional network to extract short-term temporal information as the supplement. In temporal part, we examined the aligned RGB and optical flow feature of the I3D Network[4]. In our experiments, we find that I3D-RGB feature outperforms the 2D convolutional features and I3D-Flow feature. In our final submissions, we used several combinations of the spatial and temporal features and ensembled them together.

2.2. Model Structures

For the sequence model, we mainly tested Mean-Pooling sequence model and seq2seq model. In Mean-Pooling model, we directly applied average-pooling operation on the sequence of video features to gain a single representation. In the basic seq2seq model, the RNN encoder provides a more efficient encoder to extract long-term information. Further, we implemented a bidirectional RNN encoder to model the frame sequence. To improve the capacity of the sequence model, we evaluated multi-layer RNNs and applied residual connection between layers to contribute the gradient propagation. Although the deeper model has better performance in the validation set, it didn't show equivalent performance in VTT 2017 testing data. Because the model complexity may hurt the generalization ability, we only use the single-layer models in our final submission.

3. Model Ensemble and Submissions

We use SGD with momentum to train our model, and in most case, SGD optimizer tends to have better performance than ADAM optimizer. To ensemble the models in different structures and different hyper-parameters, we tried to aggregate the predictions of candidate models in every single step. We submitted four runs on the Video To Text Description task. The methods are ranked by their performances on the testing set VTT 2017. And the performances on VTT 2018 testing set are listed in Table.1. We can find the primary run got the best performance in 2017 didn't rank first in 2018.

Table 1. The performances on test set

RUN ID	Mode	CIDEr	CIDEr-D	BLEU	METEOR
1	N	0.351	0.130	0.01504	0.2213
2	N	0.351	0.132	0.01729	0.2217
3	N	0.354	0.134	0.01609	0.2244
4	N	0.340	0.128	0.01643	0.2210

References

[1] G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, D. Joy, A. Delgado, A. F. Smeaton, Y. Graham, W. Kraaij, G. Qunot, J. Magalhaes, D. Semedo, and S. Blasi. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In *Proceedings of TRECVID 2018*. NIST, USA, 2018.

[2] 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.

[3] 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.

[4] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4724–4733. IEEE, 2017.

[5] 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.

[6] J. Dong, X. Li, and C. G. Snoek. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, 2018.

[7] 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.

[8] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks.

[9] Y. Li, Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016.

[10] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.

[11] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5987–5995. IEEE, 2017.

[12] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5288–5296, 2016.