UTS_ISA Submission at the TRECVID 2019 Video to Text Description Task

Qi Rao, Guang Li, Yi Yang
The ReLER Lab
University of Technology Sydney, Australia
{rao, guang.li}@student.uts.edu.au, yi.yang@uts.edu.au

Feng Zhang, Ziwei Wang
Information Science Academy, China

Abstract

In this paper, we summarize the technical details applied in our submission of TRECVID 2019[1] video to text task. The main effective improvements include three parts: Several efficient and comprehensive high-level features to gain expressive visual feature encodings, the algorithms in regulating and optimizing a robust language model, the expandable strategy to ensemble the well-trained single models. Besides, we conducted a meticulous evaluation of these techniques, and a comprehensive comparison of the experiments indicated the effectiveness of these techniques in video captioning.

1. Data Collection

The Trecvid VTT 2019 testing set contains videos from two sources: vine videos from Twitter Vine videos collected by NIST, where each video is about 6 seconds long. And videos from Flicker Creative Commons. Approximately 2000 videos were randomly selected and annotated. Each video was annotated five times by five different annotators.

For the training set, we collected from six open datasets: MSVD[7], MSR-VTT[15], TGIF[10] and the 2016-2018 testing data of VTT[4][3][2]. We combined the MSVD and MSR-VTT together, VTT 2016-2018 together because these two groups shared similar language style within the group. Therefore, six datasets are separated into three training sets.

Models trained on different training sets were tested and selected with uniform metrics by their performance. These metrics included: BLEU[12], METEOR[5], CIDER[14], CIDER-D and validation loss of our model. For each dataset, we split 10% of the data as validation set, which was used to select the model with relatively better performance in its own domain. And cross-domain validation was conducted to verify the generalization ability of the models. We selected a model list with a balance of domain performance and generalization ability in our later ensemble procedure.

2. Our framework

Our overall framework contained three parts: Firstly, high level visual and action features were extracted from video frames to represent the video content. Secondly, we utilized a simple but efficient LSTM[9] based encoder-decoder framework to handle features fusion and learning. Lastly, a controllable beam search with different expectations of sequence length and an expandable ensemble strategy were applied to generate sentences. Besides, we have also tried several strategies in training to boost the performance, including embedding tying, weight-drop LSTM, label smoothing, and sequence level criterion.

2.1. Feature Extraction

We extracted both visual and action features as the representation of video content. For the spatial vision feature, we utilized two models with state-of-the-art performance on Image-Net classification: ResneXt-WSL[11] and EfficientNet[13]. Besides, we utilized Kinect-i3d[6] feature as the representation of action and video temporal. We used single features or concated these features together for training to fusion spatial and temporal information of the video content. Different combinations of features will obtain different video content representation effect. We tried to use more combinations for training to increase model diversity, which was helpful in the later model ensemble procedure.

2.2. Model Structures

Video content can be regarded as a sequence of visual data. And the caption output is also a sequence. Based on these characteristics, we used Recurrent Neural Networks to bridge vision sequence and language sequence.

The great capability of RNN model in aggregating sequential information can help to leverage the temporal information through video frames. We utilized a simple and robust LSTM model with attention, and an encoder-decoder framework for captioner's training and generation. For the encoder, we implemented a random sampler and a bidirectional RNN model to enrich the training data. For the decoder, the aggregated encoded features are fed into a language model, which model the conditional probability through Recurrent Neural Network composed with GRU[8]. We have explored the limit of deeper RNNs and residual connections between layers. Experiments showed that a 512-layer RNNs performed better in separate validation set but not evident improvement when applying to a different dataset. Due to its worse generalization ability, we only selected few 512-layer RNNs models in the ensemble phase.

3. Model Ensemble and Submission

Having selected a list of well-performed single caption models, we implemented an expandable ensemble module to gather predictions from all models to vote for a better prediction. Since different model can be trained on different datasets, or have a different features combination, predictions generated by these models can be varied a lot. Ensemble these predictions can absorb the advantages of all different models and output a more precise prediction. A controllable beam search strategy is utilized to generate sentences with different sequence length expectations. In the experiments, we found generating longer sentences is helpful to boost the CIDER and METEOR scores, but caused a reduction of BLEU4 and CIDER-D scores. To cope with the trade-off, we conducted a series of hyper-parameters search and finally set the expectation of our sentences to around length of 20 as primary submission, whose performance is relatively balanced and accurate when testing on VTT 2017 set.

References

[1] G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, A. Delgado, A. F. Smeaton, Y. Graham, W. Kraaij, and G. Quénot. Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval. In *Proceedings of TRECVID 2019*. NIST, USA, 2019. 1

[2] G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, D. Joy, A. Delgado, A. F. Smeaton, Y. Graham, W. Kraaij, G. Quénot, 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. 1

[3] G. Awad, A. Butt, J. Fiscus, D. Joy, A. Delgado, M. Michel, A. F. Smeaton, Y. Graham, W. Kraaij, G. Quénot, 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. 1

[4] 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. 1

[5] S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. 1

[6] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[7] D. Chen and W. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. 1

[8] J. Chung, Çaglar Gülçehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *ArXiv*, abs/1412.3555, 2014. 2

[9] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997. 1

[10] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[11] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. Exploring the limits of weakly supervised pretraining. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1

[12] 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*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. 1

[13] M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2019. cite arxiv:1905.11946Comment: Published in ICML 2019. 1

[14] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726, 2014. 1

[15] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016. 1