SciForum
MOL2NET

Video Description with Spatio-temporal Feature and Knowledge Transferring

Xin Xu[1], Haibin Liu[1], Yi Ji[1], Xin Lin[1], Chunping Liu[1,2,3]

1.School of Computer Science and Technology, Soochow University, Suzhou, 215006, China
2.Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, 130012, China
3.Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, 210046, China

* Corresponding author email: cpliu@suda.edu.cn

Abstract:

Describing open-domain video with natural language sequence is a major challenge for computer vision. In this paper, we investigate how to use temporal information and learn linguistic knowledge for video description. Traditional convolutional neural networks (CNN) can only learn powerful spatial features in the videos, but they ignored underlying temporal features. To solve this problem, we extract SIFT flow features to get temporal information. Sequence generator of recent work are solely trained on text from video description datasets, so the sequence generated tend to show linguistic irregularities associated with a restricted language model and small vocabulary. For this, we transfer knowledge from large text corpora and employ word2vec to be the word representation. The experimental results have demonstrated that our model outperforms related work.

Conclusions

In this paper, we propose a model which contains a visual extractor and a sequence generator. First, SIFT flow features are extracted to get temporal information. Through shallow fusion with CNN feature, we can get video visual feature representation. Second, we consider two stacked LSTMs to generate natural language sequence with variable length. In order to integrate linguistic information into the sequence generator, transferring knowledge from large text corpora can generate natural language grammatically. Besides, the experiments show that word2vec is a comparatively better word representation than “one-hot” vector. We evaluate our model on Youtube2Text dataset for METEOR metric. The experiments show that our model can achieve higher METEOR scores than other methods proposed recently. In the future, we would like to exploit a more efficient visual extractor, which contains visual attention mechanism.

References

[1] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[2] Sak H, Senior A W, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling[C]//INTERSPEECH. 2014: 338-342.

[3] Liu C, Yuen J, Torralba A. Sift flow: Dense correspondence across scenes and its applications[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2011, 33(5): 978-994.

[4] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[C]//Proceedings of the ACM International Conference on Multimedia. ACM, 2014: 675-678.
[5] Tran D, Bourdev L D, Fergus R, et al. C3D: generic features for video analysis[J]. CoRR, abs/1412.0767, 2014, 2: 7.
[6] Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2625-2634.
[7] Thomason J, Venugopalan S, Guadarrama S, et al. Integrating Language and Vision to Generate Natural Language Descriptions of Videos in the Wild[C]//COLING. 2014, 2(5): 9.
[8] Venugopalan S, Xu H, Donahue J, et al. Translating videos to natural language using deep recurrent neural networks[J]. arXiv preprint arXiv:1412.4729, 2014.
[9] Horn B K, Schunck B G. Determining optical flow[C]//1981 Technical symposium east. International Society for Optics and Photonics, 1981: 319-331.
[10] Venugopalan S, Rohrbach M, Donahue J, et al. Sequence to sequence-video to text[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 4534-4542.
[11] Xu H, Venugopalan S, Ramanishka V, et al. A Multi-scale Multiple Instance Video Description Network[J]. arXiv preprint arXiv:1505.05914, 2015.
[12] Ballas N, Yao L, Pal C, et al. Delving Deeper into Convolutional Networks for Learning Video Representations[J]. arXiv preprint arXiv:1511.06432, 2015.
[13] Yao L, Torabi A, Cho K, et al. Describing videos by exploiting temporal structure[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 4507-4515.
[15] Yu H, Wang J, Huang Z, et al. Video Paragraph Captioning using Hierarchical Recurrent Neural Networks[J]. arXiv preprint arXiv:1510.07712, 2015.
[16] Pan Y, Mei T, Yao T, et al. Jointly modeling embedding and translation to bridge video and language[J]. arXiv preprint arXiv:1505.01861, 2015.
[17] Torabi A, Pal C, Larochelle H, et al. Using descriptive video services to create a large data source for video annotation research[J]. arXiv preprint arXiv:1503.01070, 2015.

Acknowledgements

This work was partially supported by National Natural Science Foundation of China (NSFC Grant No. 61272258, 61170124, 6130129, 61272005), Provincial Natural Science Foundation of Jiangsu (Grant No. BK20151254, BK20151260), Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University (Grant No. 93K172016K08), and Collaborative Innovation Center of Novel Software Technology and Industrialization.