CSI5386 NLP Project Proposal

Lingfeng Zhang 300134245 Ottawa University
Yu Sun 8472921 Carleton University

February 2020

1 Project Topic: Automated Image Captioning Web System

2 Introduction and Background

Currently, natural language processing(NLP) and computer vision(CV) are two main branches in artificial intelligence. With the development of deep learning, variants of convolutional neural network(CNN) and recurrent neural network(RNN) are used in these two research fields. Combining NLP and CV is a hot research area recently, and image captioning as well as video story telling are two specific research topics.

In real out life, after posting a video or a picture in social media, we also eager to make some descriptions for it. It costs a lot of time to choose the proper descriptions, so we plan to implement a tool(more specific, a web system) to generate natural language descriptions for visual content including pictures and videos. This work has brought a great challenge to both computer vision and natural language processing communities recently. This work is related how machine understand an image or a video.

In this project, we are going to build the model for extracting image features and generating description for images. Then we do some experiments and use some evaluation metrics to prove the efficiency of our model. Firstly, we plan to implement image captioning. After that, we will try to improve it to video storytelling.

3 Case study

About image captioning, there are two well-studied approaches to automatic image captioning: retrieval of existing human-written captions, and generation of novel captions. In terms of the retrieval model, Hodosh[1] establish a ranking based framework for sentence-based image description, and Ordonez[2] develop an automatic image description method using a large captioned photo

collection. However, retrieval-based methods always return well-formed human-written captions, and these captions may not be able to describe the image properly. So, some generation method like seq-2-seq model has been used to solve this problem. Fang[3] presents a novel approach for automatically generating image descriptions, learning directly from a dataset of image captions. You[4] proposes a new image captioning approach that combines the top-down and bottom-up approaches through a semantic attention model. Since deep learning structure is the black box and we cannot understand and control its processing, Marcella[5] proposes a novel image captioning method which can generate diverse descriptions by allowing both grounding and controllability. Moreover, if people want to get a paragraph which contains more than one sentence, above methods cannot generate diverse sentences, which meaning generating repeated sentences. Hence, Luck[6] raises a method that can generate sentences with diversity and can combine these sentences into a whole paragraph organically.

About video storytelling, the methods need to model the video dynamic temporal structure and then properly integrating that information into a natural language description, comparing with the image task. Venugopalan[7] translates videos directly to sentences using a unified deep neural network with both convolutional and recurrent structure. Yao[8] proposes an approach with attention mechanism that successfully takes into account both the local and global temporal structure of videos to produce descriptions. Since many video frames have the majority of similar scenes, which will tend to generate repeated sentences, Junnan[9] proposes a novel video storytelling model that can generate diverse sentences from videos without repetition, which can build up a continuous story line automatically. Marc[10] proposes a Egocentric video description model which can generated sentences to show what the user is doing.

Basically, the baseline models of image captioning and video storytelling are encoder and decoder. Encoder(mostly are CNN) can extract features from images, which can act as the input of the decoder. Decoder(mostly are RNN) uses these features to generate natural language sentences. So the whole model can get natural language sentences directly from visual data.

4 Project plan

Firstly, we plan to implement a basic image captioning task, generating a single-sentence factual description for an image or a corresponding segment, which supposes to be CNN and RNN with attention. Next, we are going to build a more complex model for generating comprehensive and fine-grained image and video descriptions with multi-sentence paragraphs like a summary. For the video material, we plan to divide it to frame by a certain time duration, then automatically pick important ones and treat them as a multi-picture task or use a complicated model to generate sentences directly from the video streaming.

About the implementation of this project, we plan to implement the image captioning model based on Python programming language with TensorFlow,

Keras or PyTorch deep learning Frameworks. In addition, we will build up a Django web system which can enable users to upload an image and get automatically generated sentences. Moreover, we can use Amazon Web Services(AWS) to deploy our web system on the cloud, so that all users can access it online directly. More complicated, we try to improve our system which can generated sentences directly from videos.

About the quality of generated sentences, we will use at least three main evaluation metrics: BLEU, CIDEr and METEOR to evaluate our generated natural language sentences.

5 Datasets

- Flickr30k
- Common objects in Context(COCO)
- Visual Genome

6 GitHub Sources

1. Show,Control and Tell: A Framework for Generating Controllable and Grounded Captions. Paper GitHub

2. Training for Diversity in Image Paragraph Captioning. Paper GitHub

3. Egocentric Video Description based on Temporally-Linked Sequences. Paper GitHub

References

[1] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[2] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Advances in neural information processing systems*, pp. 1143–1151, 2011.

[3] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, *et al.*, "From captions to visual concepts and back," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1473–1482, 2015.

[4] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4651–4659, 2016.

[5] M. Cornia, L. Baraldi, and R. Cucchiara, "Show, control and tell: a framework for generating controllable and grounded captions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8307–8316, 2019.

[6] L. Melas-Kyriazi, A. M. Rush, and G. Han, "Training for diversity in image paragraph captioning," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 757–761, 2018.

[7] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," *arXiv preprint arXiv:1412.4729*, 2014.

[8] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *Proceedings of the IEEE international conference on computer vision*, pp. 4507–4515, 2015.

[9] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Video storytelling," *arXiv preprint arXiv:1807.09418*, 2018.

[10] M. Bolaños, Á. Peris, F. Casacuberta, S. Soler, and P. Radeva, "Egocentric video description based on temporally-linked sequences," *Journal of Visual Communication and Image Representation*, vol. 50, pp. 205–216, 2018.