Existing Work

This approach reconciles classical slot filling approaches that are generally better grounded in images with modern neural captioning approaches that are generally more natural sounding and accurate. This approach first generates a sentence `template' with slot locations explicitly tied to specific image regions. These slots are then filled in by visual concepts identified in the regions by object detectors. The entire architecture sentence template generation and slot filling with object detectors is end-to-end differentiable. It verifies the effectiveness of our proposed model on different image captioning tasks. On standard image captioning and novel object captioning, our model reaches state-of-the-art on both COCO and Flickr30k datasets.[1]

The issue of generating natural language descriptions from visual data has long been studied in computer vision. This has led to development of complex systems composed of visual primitive recognizers combined with a structured formal language, e.g. And-Or Graphs or logic systems, which are further converted to natural language via rule-based systems. Such systems are heavily hand-designed, relatively brittle and have been demonstrated only on limited domains, e.g. traffic scenes or sports. Leveraging recent advances in recognition of objects, their attributes and locations, allows us to drive natural language generation systems, although these are limited in their expressivity.[2]

A group of researchers at Google presented a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. Their model is trained to maximize the likelihood of the target description sentence given the training image. Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descriptions. The model is one of the most accurate system developed for image captioning, which was verified both qualitatively and quantitatively.[3]

This paper reviews deep learning-based image captioning methods. It discusses different evaluation metrics and datasets with their strengths and weaknesses. A brief summary of experimental results is also given. It briefly outlines potential research directions in this area. Although deep learning-based image captioning methods have achieved a remarkable progress in recent years, a robust image captioning method that is able to generate high quality captions for nearly all images is yet to be achieved. With the advent of novel deep learning network architectures, automatic image captioning will remain an active research area for some time.[4]

With the rapid development of deep learning technology, CNN and LSTM have become two of the most popular neural networks. This paper combines CNN and LSTM or its variant and makes a slight change. Unlike the typical CNN, which contains convolution operation and activation function, this paper constructs two text classification models called NA-CNN-LSTM and NA-CNN-COIF-LSTM by combining CNN without activation function and LSTM, and one of its variants COIF-LSTM. Through comparative experiments, it is proved that the combination of CNN without activation function and LSTM or its variant has better performance. The experimental results on

the subjective and objective text categorization dataset show that the proposed model has better performance than the standard CNN or LSTM.[5]

Some existing systems use detections to infer a triplet of scene elements which is converted to text using templates. Similarly, Li et al. piece together a final description using phrases containing detected objects and relationships by starting off with detections. A more complex graph of detections beyond triplets is done but with template-based text generation. The mentioned approaches have been able to describe images “in the wild”, but they are heavily hand designed and not flexible when it comes to text generation.[6]

Top-down visual attention mechanisms have been used extensively in image captioning and visual question answering VQA to enable deeper image understanding through fine-grained analysis and even multiple steps of reasoning. In this work, the authors propose of a combined bottom-up and top-down attention mechanism that enables attention to be calculated at the level of objects and other salient image regions. This is the natural basis for attention to be considered. With this approach, the bottom-up mechanism based on Faster R-CNN proposes image regions, each with an associated feature vector, while the top-down mechanism determines feature weightings.[7]

It has been shown that policy-gradient methods for reinforcement learning can be utilized to train deep end-to-end systems directly on non-differentiable metrics for the task at hand. In this paper, the authors consider the problem of optimizing image captioning systems using reinforcement learning and show that by carefully optimizing our systems using the test metrics of the MSCOCO task, significant gains in performance can be realized. These systems are built using a new optimization approach that is called self-critical sequence training.[8]

Automatically describing an image with a sentence is a long-standing challenge in computer vision and natural language processing. Due to recent progress in object detection, attribute classification, action recognition, etc., there is renewed interest in this area. However, evaluating the quality of descriptions has proven to be challenging. This paper proposes a novel paradigm for evaluating image descriptions that uses human consensus. This paradigm consists of three main parts: a new triplet-based method of collecting human annotations to measure consensus.[9]

Deep neural networks have achieved great successes on the image captioning task. However, most of the existing models depend heavily on paired image-sentence datasets, which are very expensive to acquire. In this paper, the first attempt to train an image captioning model in an unsupervised manner is tried. Instead of relying on manually labeled image-sentence pairs, the proposed model merely requires an image set, a sentence corpus, and an existing visual concept detector. The sentence corpus is used to teach the captioning model how to generate plausible sentences.[10]

References

[1] Adil Ali Alaziz, M. and Adnan Abdul Kareem, S. (2019). A Novel Method for Image Captioning Based on Attributes and External Knowledge. *Journal of Engineering and Applied Sciences*, 14(2), pp.610-614.

[2] "Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences - IEEE Conference Publication", *Ieeexplore.ieee.org*, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/561027/.

[3] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and Tell: A Neural Image Caption Generator", *arXiv.org*, 2020. [Online]. Available: https://arxiv.org/abs/1411.4555.

[4] M. Hossain, F. Sohel, M. Shiratuddin and H. Laga, "A Comprehensive Survey of Deep Learning for Image Captioning", *arXiv.org*, 2020. [Online]. Available: https://arxiv.org/abs/1810.04020.

[5] "Research on Text Classification Based on CNN and LSTM - IEEE Conference Publication", *Ieeexplore.ieee.org*, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8873454.

[6] Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL '11). Association for Computational Linguistics, USA, 220–228.

[7] P. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering", *arXiv.org*, 2020. [Online]. Available: https://arxiv.org/abs/1707.07998.

[8] S. Rennie, E. Marcheret, Y. Mroueh, J. Ross and V. Goel, "Self-critical Sequence Training for Image Captioning", *arXiv.org*, 2020. [Online]. Available: https://arxiv.org/abs/1612.00563.

[9] R. Vedantam, C. Zitnick and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation", *arXiv.org*, 2020. [Online]. Available: https://arxiv.org/abs/1411.5726.

[10] Y. Feng, L. Ma, W. Liu and J. Luo, "Unsupervised Image Captioning", *arXiv.org*, 2020. [Online]. Available: https://arxiv.org/abs/1811.10787.