Week 8 paper summary

Sina Mehdinia

05/21/2021

Summary

Vinyals et al. [1] present an end to end deep learning framework from an input image to a caption descriptor. They use a CNN to encode the features of the input image and then train a recurrent neural network to generate sentences. The RNN used is a LSTM model that generates the output by seeing the input image and all the previous words. The goal of the network is to maximize the probability of a correct sentence given an input image. Their CNN embedder uses a novel batch normalization technique. The LSTM model is a special type of RNN that does not have the problem of vanishing/exploding gradients. It has a memory cell that retains knowledge at each step in time and it is controlled by different gates of input, output and forget. They used ImageNet weights for initializing the weights which can reduce overfitting. For training, they used stochastic gradient descent with a constant learning rate and no momentum. In order to evaluate the results, they used different methods, human raters rated the results from 0 to 4 and also some metric called BLEU. They tested the results on different datasets and showed that their method is superior to other existing approaches of that time.

References

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.