SUTD 2022 50.039 Homework 2

James Raphael Tiovalen / 1004555

Introduction

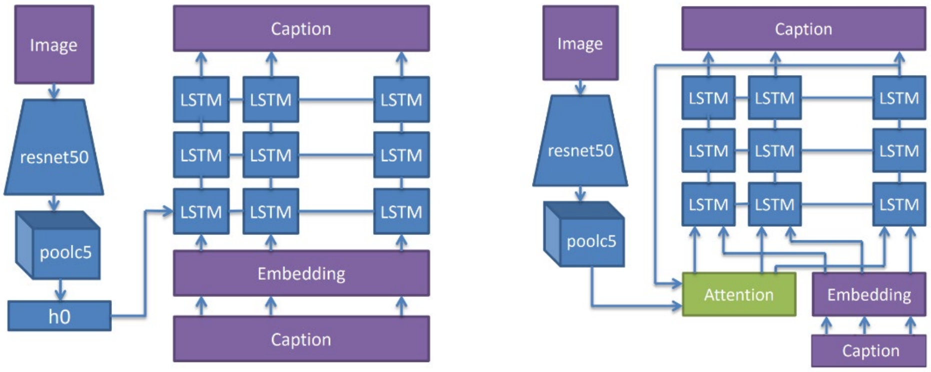
The authors of the paper "Show and Tell: A Neural Image Caption Generator" discussed a method to automatically describe the contents of an image in the form of captions. The authors proposed a generative model called the Neural Image Caption (NIC), which consists of a vision CNN followed by a language generating LSTM. Evaluating the paper, we deem that while image captioning is indeed a much more difficult task compared to image classification/recognition, there are still potential improvements that could be made to the Google NIC model proposed in the paper.

Summary

The paper first described the overall architecture of their model. They suggested a neural net that is fully trainable using stochastic gradient descent as an end-to-end system. The model integrates state-of-the-art sub-networks of CNN as a vision model with LSTM as a language model. The inspiration for this model architecture comes from machine translation. The CNN vision model serves as the encoder, while the LSTM language model serves as the decoder. Images are fed into the CNN, which would output an encoded image vector that serves as the feature map representation. The CNN model is pre-trained using the ImageNet dataset and utilizes transfer learning to be adapted into the Google NIC model. The encoded representation of the image is then processed by a recurrent network consisting of a stack of LSTM layers fed by an embedding layer. The decoded sequence is then processed by a sentence generator (which can be done by either using sampling or beam search) to generate the actual final image caption. The BeamSearch method is used in the study, with a beam size of 20. The paper then compared the results of the NIC model with other state-of-the-art models at that time in terms of the BLEU-1 and the BLEU-4 scores. On the Pascal dataset, the model achieves a BLEU-1 score of 59, which is significantly higher than the state-of-the-art score of 25 at the time and comparable to human performance of roughly 69. On the Flickr30k and SBU datasets, the model improved its BLEU-1 score from 56 to 66 and from 19 to 28 respectively, and on the COCO dataset, it obtained the state-of-the-art BLEU-4 score of 27.7. The paper then went on to discuss the results of the model, several limitations of the BLEU metric, as well as some analysis and findings on data quality, ranking, and embeddings.

Critique

While the performance of the Google NIC model is quite impressive, some improvements can be made to the proposed model. For one, in the Google NIC model, the spatial dimensions of the CNN image features were "averaged" together. Instead, if we use an "attention mechanism" to weigh those spatial locations according to their perceived importance, the overall model can perform better. This can be accomplished by adding an "attention gate" to the LSTM architecture. In fact, this better method was described by the paper "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" (Xu et al., 2015).



Given that the paper was released in 2014, a year before the "Show, Attend and Tell" paper, it's understandable if it was a little obsolete and behind current best practices. A more recent paper (Liu et al., 2021) also outlined an improved method to conduct image captioning by using transformers instead, which would replace the LSTM models.

Different architectures besides the “encoder-decoder” (or “inject”), the “encoder-decoder with attention”, and the “encoder-decoder with transformers” architectures have also been proposed. The “multi-modal” (or “merge”) architecture, whereby the CNN network only processes the image and the LSTM network operates only on the sequence generated so far (i.e., the two components operate independently of each other and do not mix), can sometimes produce better results. Other architectures include using an object detection CNN backbone instead of an image classification CNN backbone, doing dense captioning (multiple captions for different regions of an image), as well as a sentence generator with beam search. Image captioning is not limited to just the “encoder-decoder” approach since there are various approaches to do image captioning, each with its own advantages and limitations.

Conclusion

In conclusion, while more subsequent works have produced greater image captioning performance, this paper highlighted an essential piece of research work that contributed to the momentum in the field of image captioning, especially with its significant BLEU-1 score improvement on the Pascal dataset, allowing us to further expand our grasp of machine learning. The general suggestion of using optimal state-of-the-art sub-networks is an important idea that can still be used to improve the performance of machine learning models in real-life applications. Authors of more recently published papers can achieve what they have achieved by building upon the shoulders of these giants and iterating on their works. Science is a process and we live in an exciting time as we witness the ever-upward and ever-onward improvements in the latest machine learning techniques.

References

Liu, W., Chen, S., Guo, L., Zhu, X., & Liu, J. (2021). *CPTR: Full Transformer Network for Image Captioning*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015). *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*.