Language Model Methods used for Image Captioning

By Dani Richmond

Image captioning is a difficult undertaking as it combines the need for accurate computer vision with accurate natural language generation. With social media becoming increasingly image-laden rather than text-laden, it has become an even more pressing need for visually impaired individuals to be able to use image captioning to fully participate in social media. Unfortunately, the leading methods of image captioning are not ready yet to take on the variety of images shared on social media, but they have made significant improvements in the past six years. Broadly speaking, the task of image captioning involves taking an image as input, encoding the image's pixels as data points, then sending that data into a decoder that uses a language model to generate a sequence of words. For the purposes of this review, I will focus on the three main language model methods that have been used to accomplish the decoding part of this undertaking: LSTM, Transformers, & BERT-like.

The first breakthrough for image captioning occurred in April 2015 when the seminal paper "Show and Tell: A Neural Image Caption Generator" by Vinyals *et al.* proposed using a Convolution Neural Network (CNN) as an encoder to represent the input image as a feature vector and a Recurrent Neural Network (RNN) as a decoder to generate the captions as output [1]. More specifically, a Long-Short Term Memory (LSTM) recurrent neural network was used for the language model. It took two inputs – the vector of image features generated by the CNN and the prior words generated for the caption sentence and computed the probabilities for all words in the vocabulary given those inputs to determine what word to output next. Vinyals *et al.* also employed BeamSearch [1] to further improve their results. Instead of just generating the one word with the highest probability at each time-step (known as greedy search), they kept a set of the top *k* sentences and generated a new word based on probability for each of these sentences. Once all the top k sentences are complete, the one with the highest overall probability is output as the caption sentence.

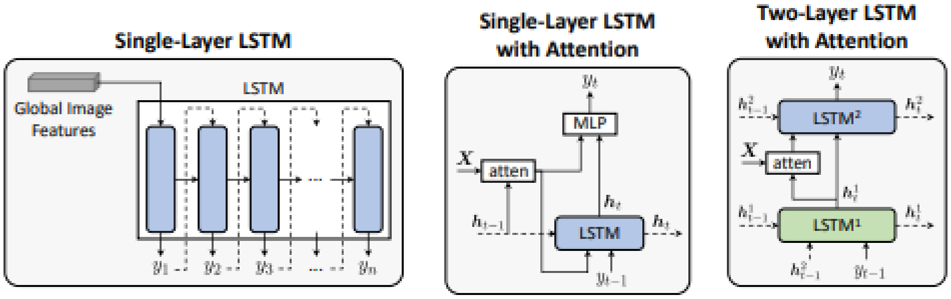

Figure 1: Examples of LSTM language models; left – version proposed in [1]; middle – version proposed in [2]; right – version proposed in [3]. "X represents either a grid of CNN features or image region features extracted by an object detector" [4].

The LSTM language model method saw another jump in efficacy with the introduction of attention mechanisms. The first to introduce using attention for image captioning was Xu *et al.* in "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" [2]. Instead of using a static feature vector representation of the image as input to the LSTM, a new version of the image's feature vector is generated for each time-step. The attention component computes a weight for each pixel as an indicator of that pixel's importance for generating the next word. These weights are then applied to the original image feature vector so that the language model focuses on a subset of the image which led to better associations between objects in an image and words in the vocabulary.

There have been many more modifications proposed for the LSTM method but the most noteworthy is the introduction of two-layer LSTMs. Anderson *et al.* were able to set the new standard for state-of-the-art image captioning with their two-layer LSTM [3]. The first LSTM layer functions as a top-down visual attention model that receives as input the output of the second LSTM layer at the last time-step, the image's feature vector, and a separate encoding of the prior words generated. "These inputs provide the attention LSTM with maximum context regarding the state of the language LSTM, the overall content of the image, and the partial caption output generated so far, respectively." [3] With this context, an attention weight is then computed and applied to the image's feature vector. The second LSTM layer, the language LSTM, receives this weighted image feature vector and the output of the first LSTM layer to determine the probability of words in the vocabulary and select the next word.

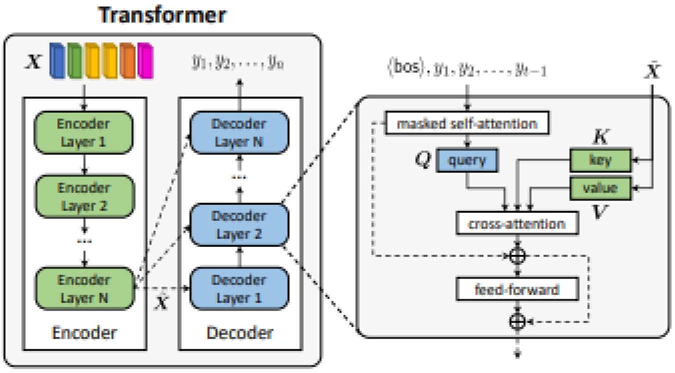

Figure 2: "Schema of the Transformer-based language model. The caption generation is performed via masked self-attention over previously generated tokens and cross-attention with encoded visual features" [4].

Building upon the demonstrated efficacy of using attention mechanisms, Vaswani *et al.* introduced the idea of self-attention in their paper "Attention Is All You Need" [4]. This represented a big departure from the previous works as they introduced a new architecture called the Transformer that did not use any kind of RNN for language generation. The Transformer's decoder still functions as the basis for the language model but operates in a very different way than the LSTM. First, it performs self-attention in parallel which computes an attention score for every word in the sentence generated thus far compared to every other word. This self-attention operation (which is performed in both the encoder and decoder components) produces a much richer representation of how the words in a sentence relate to one another (and for the encoder, how the parts of an image relate). Second, it performs cross-attention that computes attention scores denoting how the output of the decoder's self-attention correlate to the output of the image encoder component. With their unique method of performing both self-attention and cross-attention, Transformers became the new definition for what was considered state-of-the-art. While there have been some proposed modifications to the Transformer architecture to improve image captioning, the majority have been related to improving aspects of the encoder component.

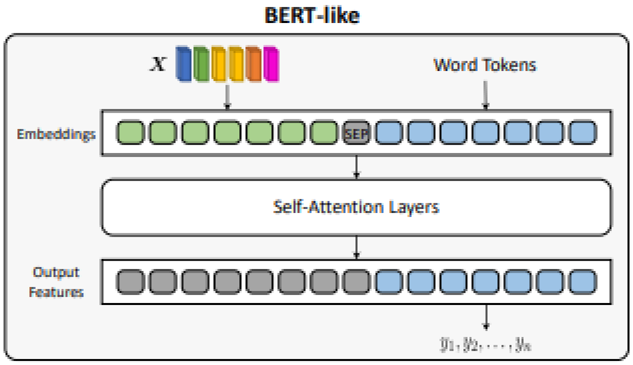

Figure 3: "Schema of a BERT-like language model. A single stream of attentive layers processes both image regions and word tokens and generates the output caption" [4].

Another paradigm shift in image captioning approaches happened when the Bidirectional Encoder Representations from Transformers (BERT) language model was introduced [5]. BERT language models rely on large amounts of pre-training to build "deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context" [5]. A defining component of BERT-like image captioning methods is that there is no longer a sharp delineation between the image encoder component and the language model decoder component because the Transformers that make up the backbone of this architecture are used for both encoding and decoding. Zhou *et al.* introduced the idea of a unified Vision-Language Pre-training (VLP) model as the first method to use BERT for image captioning [6]. By using unified encoder-decoder Transformers, richer connections are created in the model between images and captions which leads to better caption generation.

Although the progression of approaches for image captioning have been building on each other over the past 6 years, the newer ones have not replaced the older models and much research continues to improve the three main language model approaches as well as to develop novel ones. The BERT-like methods currently have the highest CIDEr accuracy scores, but they also have three main downsides: they require a lot of pre-training, they need a lot of training examples (which is not always feasible), and they are resource intensive (they require far more model parameters than the LSTM and Transformer approaches). LSTM methods have continued to hold their own compared to Transformer methods due to continued improvements. For example, the AutoCaption LSTM model has a CIDEr score of 135.8 while the RSTNet Transformer model's score is 135.6 [7]. The main drawbacks for the LSTM methods are that these models continue to be slow to train and they struggle to maintain connections between words that are far apart in the sentence being generated. In contrast, one of the key benefits of Transformers is that the distance between words has no impact on its ability to calculate the relationship between them as part of the self-attention operation. Unfortunately, generalization beyond a set of well-crafted training and testing sets continues to be an issue for the state-of the-art version of all the approaches [7]. Generalization of the model is a necessary first step to producing captions for the breadth of image content currently shared on social media but there is exciting progress being made in pre-training BERT-like methods on poorly curated data that may provide the insights needed for generalization to social media.

References

[1] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and tell: A neural image caption generator," *CVPR,* 2015.

[2] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel and Y. Bengio, "Show, attend and tell: Neural image caption generator," *ICML,* 2015.

[3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," *CVPR,* 2018.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need," *NeurIPS,* 2017.

[5] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," *NAACL,* 2018.

[6] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso and J. Gao, "Unified Vision-Language Pre-Training for Image Captioning," *AAAI,* 2019.

[7] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni and R. Cucchiara, "From Show to Tell: A Survey on Image Captioning," *arXIV,* 2021.