PROJECT SUMMARY: ATTENTION BASED IMAGE CAPTIONING

Aman Goel (Roll No: 11806)
aman.goel185@gmail.com

Nirupam Gunwal (Roll No: 11725)
nirupamgunwal@ducic.ac.in

Chappa Sri Vinay (Roll No: 11812)
vinay07rn@gmail.com

ABSTRACT

The problem we are focusing on is attention based image captioning, which essentially means we work on the problem of automatically describing the content of the image but, different from previous work we focus on particular parts of the image which are of more importance to us expecting more accurate captions. We work on a encoder-decoder model for image captioning with the encoder as a CNN model and the decoder as the language model RNN.

1 AIM

In this paper, we intend to work on the problem of Image Captioning, which basically involves automatically generating a natural description of an image. We will particularly focus on adding attention mechanism to the domain of image captioning using attention models.

2 HISTORY AND PREVIOUS WORK

2.1 ENCODER-DECODER MODEL

The Encoder-Decoder model for RNNs was introduced in the paper, i.e., Sutskever et al. (2014). The paper developed the technique to address the sequence-to-sequence nature of machine translation where the input and output sequences both differ in length. The model consists of two sub-models: an Encoder and a Decoder. Encoder: The encoder is responsible for stepping through the input time steps and encoding the entire sequence into a fixed length vector called a context vector. Decoder: The decoder is responsible for stepping through the output time steps while reading from the context vector. Key to the model is that the entire model, including encoder and decoder, is trained end-to-end, as opposed to training the elements separately.

2.2 ATTENTION MODEL

Attention was presented in the paper Bahdanau et al. (2014). Attention is proposed as a solution to the limitation of the Encoder-Decoder model encoding the input sequence to one fixed-length vector from which to decode each output time step. This issue is believed to be more of a problem when decoding long sequences. Attention is proposed as a method to both align and translate. As with the Encoder-Decoder paper Sutskever et al. (2014), the technique is applied to a machine translation problem and uses GRU units rather than LSTM memory cells. In this case, a bidirectional input is used where the input sequences are provided both forward and backward, which are then concatenated before being passed on to the decoder.

3 Baseline and Proof of Concept

In related work, in You et al. (2016), the problem of image captioning is solved using a encoder-decoder model using CNN and RNN. Particularly in Vinyals et al. (2014), CNN's are used as encoder which extracts feature maps from the image and we get better spatial information from the image. Features are extracted in a CNN either from the fully connected layer or other lower-level layers. We, get fixed length feature vector representation of an image through this encoder. We then use the output of this encoder as an input to the decoder model which consists of an RNN to generate sentences word by word. Although, this method proposed gave a benchmark at its time in the BELU scores, but with time we have other state of the art method. Usually, we are achieving better scores by using an attention mechanism You et al. (2016), It was earlier used in language translation models giving attention to some part of a sentence. In a similar fashion, attention mechanism is introduced Xu et al. (2015), in a encoder-decoder model focusing on some particular parts of an image at every time step by calculating probability scores matrices. In Xu et al. (2015), two types of attention mechanism are used soft-attention and hard-attention. Both the type of attention mechanisms give state of the art belu scores and similar performances. But, most of the recent work uses soft attention because it is differentiable and easier to implement than hard attention. Hard attention does not have differentiable functions and we cannot train it with backpropagation. In our work, we intend to implement soft attention based image captioning and implement some novel methods to improve performance metrics for image captioning.

4 Project Image

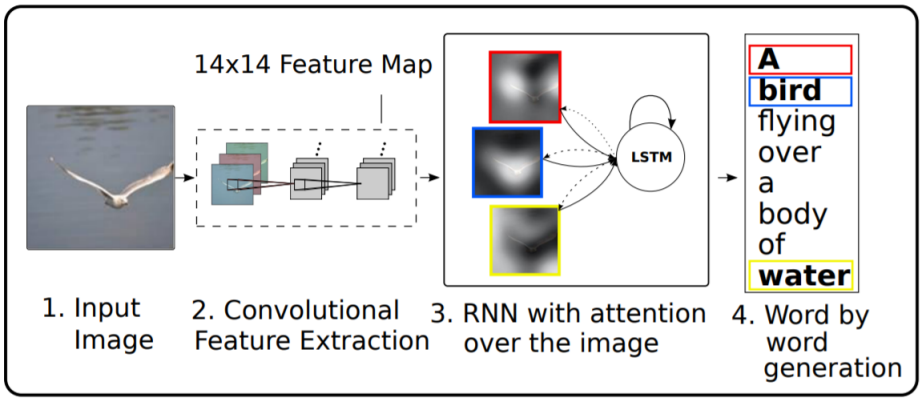

Figure 1: Pipeline for our project.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014. URL `http://arxiv.org/abs/1411.4555`.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015. URL `http://arxiv.org/abs/1502.03044`.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. *CoRR*, abs/1603.03925, 2016. URL `http://arxiv.org/abs/1603.03925`.