Neural Network Encoder-Decoder Architecture for Image Captioning

UNION COLLEGE

Emily Kern

Kristina Striegnitz, Advisor

CS

Introduction

Neural networks are computing systems that learn by example over several time steps known as epochs. Their applications vary depending on the network's architecture, which is determined by the patterns of connections between neurons. Image captioning is one such application that combines object classification and sentence generation. The encoder-decoder architecture is best suited for image captioning, as it is capable of taking input and generating output.

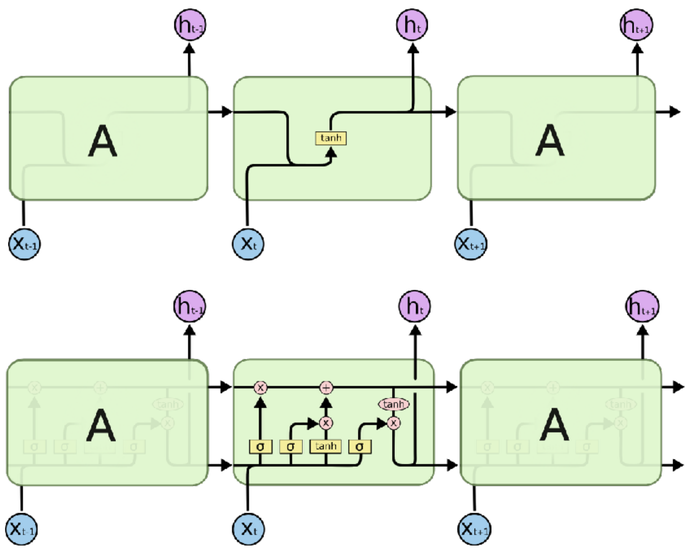


Figure 1. RNN (top) versus LSTM (bottom) architecture. LSTMs use additional gates to simulate memory, so they take more time run than RNNs

Neural Network Types

Convolutional neural networks (CNNs) can extract high-level features of images, making them useful for object classification. Recurrent neural networks (RNNs) can operate over a sequence of inputs (e.g. sentences), but have a tendency to "forget" information from several epochs ago. Long short-term memory neural networks (LSTMs) resolve that pitfall by incorporating additional gates that simulate memory (Figure 1). Gates are tanh or sigmoid functions that determine how much data from an input should be forgotten.

Architecture

The Karpathy/Fei-Fei model [1] combines a CNN with either a RNN or LSTM to construct an encoder-decoder architecture (Figure 2). The CNN encoder takes an input image and outputs a vectorization of that image. The vectorization is fed into the RNN/LSTM decoder, which outputs a caption based on the embedding. Due to the known pitfalls present in RNNs, the LSTM was selected for decoding.

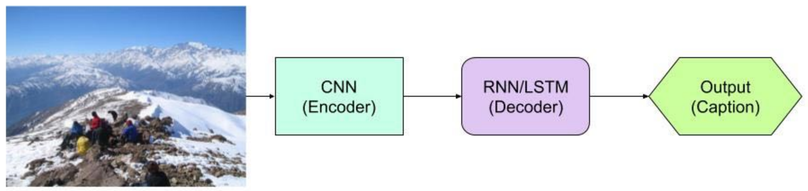
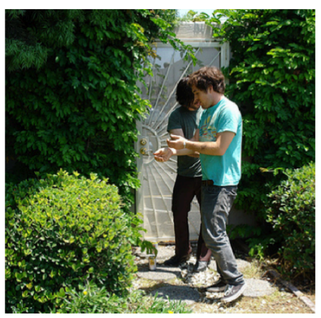


Figure 2. A simplified diagram of the encoder-decoder architecture using a CNN and RNN/LSTM.

Training and Dataset

The neural network was trained on the Flickr30k dataset [2] (Figure 3), which contains raw images and crowd-sourced captions and was split into training, validation, and testing portions. Training images were fed into and followed the encoder-decoder's layout. The network compared the outputted caption to the "real" crowd-sourced captions from the dataset and adjusted its weights accordingly. This helped the network's gates make better (weighted) decisions on what to forget and remember. Checkpoints were kept to monitor training at various epochs. Validation images were fed through to tweak the network's adjustments.



{'filename': '1000092795.jpg', 'imgid': 0, 'sentences': [{'tokens': ['two', 'young', 'guys', 'with', 'shaggy', 'hair', 'look', 'at', 'their', 'hands', 'while', 'hanging', 'out', 'in', 'the', 'yard'], 'raw': 'Two young guys with shaggy hair look at their hands while hanging out in the yard.', 'imgid': 0, 'sentid': 0}, {'tokens': ['two', 'young', 'white', 'males', 'are', 'outside', 'near', 'many', 'bushes'], 'raw': 'Two young, White males are outside near many bushes.', 'imgid': 0, 'sentid': 1}, {'tokens': ['two', 'men', 'in', 'green', 'shirts', 'are', 'standing', 'in', 'a', 'yard'], 'raw': 'Two men in green shirts are standing in a yard.', 'imgid': 0, 'sentid': 2}, {'tokens': ['a', 'man', 'in', 'a', 'blue', 'shirt', 'standing', 'in', 'a', 'garden'], 'raw': 'A man in a blue shirt standing in a garden.', 'imgid': 0, 'sentid': 3}, {'tokens': ['two', 'friends', 'enjoy', 'time', 'spent', 'together'], 'raw': 'Two friends enjoy time spent together.', 'imgid': 0, 'sentid': 4}], 'split': 'train', 'sentids': [0, 1, 2, 3, 4]}

Figure 3. Sample image and corresponding information from Flickr30k. Includes crowd-sourced captions.

Evaluation

Testing images from Flickr30k were fed into the neural network to evaluate the network's training. Test images were new to the network, so captions generated for those images were based on the trained weights the network maintained. Generated captions were compared to the "real" captions provided by Flickr30k to evaluate their accuracy and had their syntax scored by BLEU to assess the propriety of natural language in the caption. Accuracy and BLEU score helped calculate "confidence" (logprob) in the generated caption, where a confidence close to 0 is ideal. By the final epoch, the magnitude of confidence was lower than expected with signs of word bias, but showed improvement over time (Figure 4).

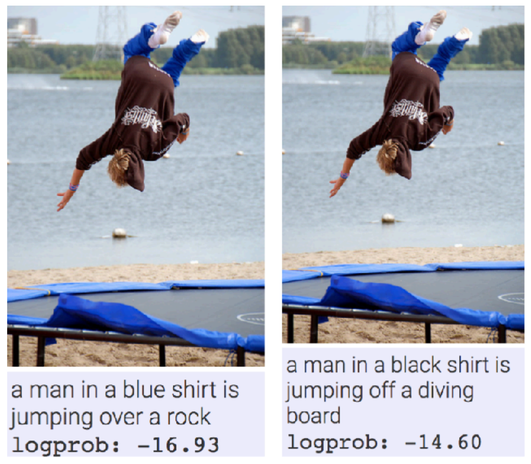


Figure 4. Generated captions for a test image. Logprob improved from first epoch (left) to last epoch (right).

Acknowledgments

I would like to thank Kristina Striegnitz, David Frey, and CROCHET Lab.

References

[1] Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3128-3137).

[2] Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. (2015, December). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Computer Vision (ICCV), 2015 IEEE International Conference on* (pp. 2641-2649). IEEE.