Image Captioning using Deep Learning

Arnav Arnav, Hankyu Jang, Pulkit Maloo

Indiana University - School of Informatics, Computing, and Engineering

Objectives

- Understand state of the art image captioning models
- Apply transfer learning in training phase

Introduction

Image captioning is one of the major areas of AI research since it aims to mimic the human ability to compress enormous amount of visual information in a few sentences. Recent developments in deep learning and the availability of image caption datasets such as COCO and Flickr have encouraged important research in the area.

Data and Preprocessing

The Flickr8k dataset contains 8000 images, 5 captions corresponding to each image.
We extract image features using pre-trained Convolutional Neural Networks (CNN) models and pass these (512 or 2048 dimensional vectors) as an input to the *image_input* layer of the model, and use a dense layer to obtain a lower dimensional embedding of 300 dimensions.
We prepare one-hot encoded vectors for each of the captions in the data, which are used by the *caption_input* layer.
We use the Show and Tell model [1] to learn mapping between images and their captions. The weights for the embeddings are also learned while training the model.

LSTM Network

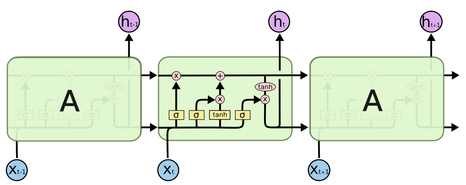


Figure: A simple LSTM network[2]

A Long Short Term Memory (LSTM) Network can learn dependencies from long sequences and is a key part to this image captioning approach.

Model Architecture

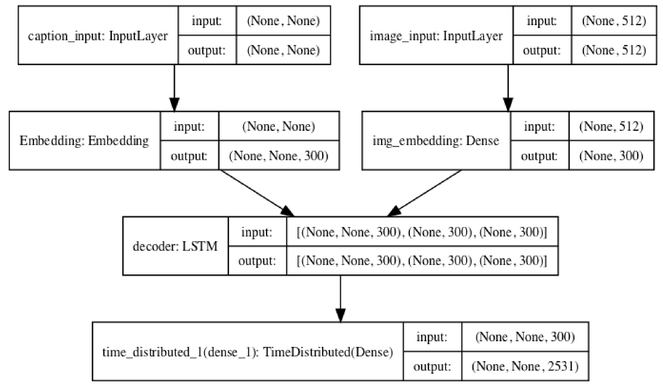


Figure: Model architecture using VGG16 image features

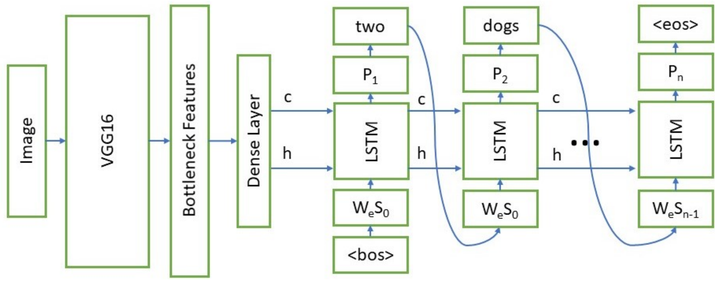
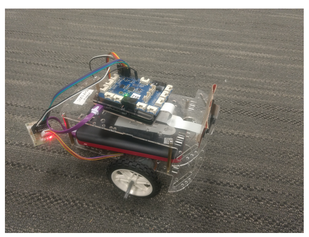


Figure: Inference: Generating captions once the LSTM is trained

Results

(a) a man and woman are sitting in a room with a man in a white shirt

(b) a group of people are sitting in the sand on a beach



(c) a man in a yellow jacket riding a yellow motorcycle

(d) a man in a blue shirt and jeans is walking down a sidewalk

(e) a group of people are sitting in the water

(f) a young boy with a long hair and a shirt is standing in front of a large tree

Figure: Generated captions from the above model using VGG16 features

Method

- We implement CNN-RNN in an encoder-decoder scheme from scratch using Keras.
- CNN: Use VGG16, VGG19, and ResNet50 models to generate bottleneck features from images, using pre-trained weights.
- Recurrent Neural Networks: Use LSTM network as a decoder to generate sentences using word embedding as input. We train the LSTM and word embeddings to learn a mapping between image features and training captions.

Evaluation

Corpus level BLEU scores for different CNN models

Metric	VGG16	VGG19	ResNet50
BLEU1	51.26	52.64	51.60
BLEU2	21.41	21.95	22.71
BLEU3	8.32	8.24	8.99
BLEU4	3.31	3.26	3.94

Future Work

Future experimentation can be done by adding attention layer to the model, training on different datasets such as MSCOCO, and Flickr30k, and implemetning beam search for generating captions.

References

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan.
Show and tell: A neural image caption generator.
In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE, 2015.

[2] Christopher Olah.
Understanding lstm networks.
colah's blog, August 2015.

[3] Pranoy Radhakrishnan.
Image captioning in deep learning.
Medium: Towards Data Science, September 2017.

[4] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio.
Show, attend and tell: Neural image caption generation with visual attention.
In *International Conference on Machine Learning*, pages 2048–2057, 2015.