Image Captioning

Introduction

- Image captioning is a popular research area of Artificial Intelligence (AI) that deals with image understanding and a language description for that image. Image understanding needs to detect and recognize objects. It also needs to understand scene type or location, object properties and their interactions. Generating well-formed sentences requires both syntactic and semantic understanding of the language.
- Image captioning is important for many reasons. For example, they can be used for automatic image indexing. Image indexing is important for Content-Based Image Retrieval (CBIR) and therefore, it can be applied to many areas, including biomedicine, commerce, the military, education, digital libraries, and web searching. Social media platforms such as Facebook and Twitter can directly generate descriptions from images.
- Understanding an image largely depends on obtaining image features. The techniques used for this purpose can be broadly divided into two categories: (1) Traditional machine learning based techniques and (2) Deep machine learning based techniques.
- In this project we will use a Deep learning based technique using Encoder-Decoder Architecture.

Data

- **MS COCO Dataset.**
- Microsoft COCO Dataset [1] is a very large dataset for image recognition, segmentation, and captioning. There are various features of MS COCO dataset such as object segmentation, recognition in context, multiple objects per class, more than 300,000 images, 80 object categories, and 5 captions per image. Many image captioning Methods use the dataset in their experiments.
- url: http://cocodataset.org/#download

Approach - Encoder-Decoder Architecture-Based Image captioning

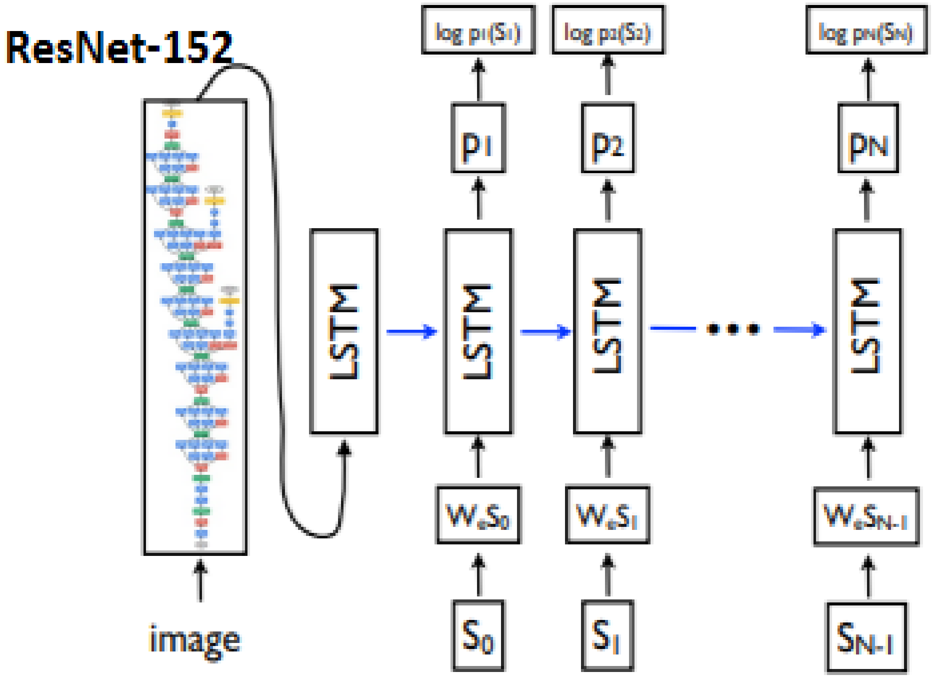
- We use a Deep learning based technique using Encoder-Decoder Architecture. The neural network-based image captioning methods are very similar to the encoder-decoder framework-based neural machine translation. In this network, global image features are extracted from the hidden activations of CNN and then fed them into an LSTM to generate a sequence of words.

A typical method of this category has the following general steps:

- A vanilla CNN is used to obtain the scene type, to detect the objects and their relationships.
- The output of Step 1 is used by a language model to convert them into words, combined phrases that produce an image captions.

Approach - Architecture

CNN-RNN (EncoderCNN-DecoderRNN) model is based on the model proposed in the paper "Show and Tell: A Neural Image Caption Generator" (https://arxiv.org/pdf/1411.4555.pdf). Below figure adapted from the paper shows a LSTM model combined with a CNN (ResNet152) image embedder and word embeddings. The unrolled connections between the LSTM memories are in blue and they correspond to the recurrent connections. All LSTMs share the same parameters.



Approach: Evaluation Metrics

BLEU (Bilingual evaluation understudy) is a metric that is used to measure the quality of machine generated text. Individual text segments are compared with a set of reference texts and scores are computed for each of them. In estimating the overall quality of the generated text, the computed scores are averaged. However, syntactical correctness is not considered here. The performance of the BLEU metric is varied depending on the number of reference translations and the size of the generated text. BLEU is popular because it is a pioneer in automatic evaluation of machine translated text and has a reasonable correlation with human judgements of quality. However, it has a few limitations such as BLEU scores are good only if the generated text is short.

A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0.

Conclusions and Future Work

- In this Image Captioning project, we implemented a Deep learning based technique using Encoder-Decoder Architecture. We have shown our calculated Evaluation Metric - BLEU scores, outperformed the reference BLEU scores.
- In their paper - BLEU: a Method for Automatic Evaluation of Machine Translation, the authors quote -
- The BLEU metric ranges from 0 to 1. Few translations will attain a score of 1 unless they are identical to a reference translation. For this reason, even a human translator will not necessarily score 1. […] on a test corpus of about 500 sentences (40 general news stories), a human translator scored 0.3468 against four references and scored 0.2571 against two references.

Conclusions and Future Work

Future improvements to investigate further:

- 1) Beam Search as an approximate search (often works better than the greedy approach)
- 2) Implement other models like attention based model (Show, Attend and Tell: Neural Image Caption Generation with Visual Attention)

Recommendations to the Client

Facebook can use Image captioning functionality to automatically generate captions for photos in the News Feed of people who can't see them. This can be used with text-to-speech engines that allow blind people to use Facebook in other ways