Image Captioning with RNN/LSTM and Attention Mechanisms

Katie Greed and Mike Keohane
https://github.com/kgreed4/Image-Captioning

April 28, 2024

Contents

1 Introduction

This paper explores the use of recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, as well as attention mechanisms, for zero-shot image captioning tasks. We discuss the theoretical background, implementation details, and experimental results of both approaches. Image captioning extends traditional classification by using the feature embedding of the CNN to generate a corresponding textual caption.

Image captioning is the task of generating textual descriptions for images automatically. It has various applications such as aiding visually impaired individuals, improving image retrieval systems, and enhancing human-computer interaction. In recent years, deep learning techniques, particularly RNNs and LSTMs, have shown promising results in generating coherent and descriptive captions for images. Moreover, attention mechanisms have been introduced to further improve the performance of image captioning models by focusing on relevant regions of the image while generating captions.

In this paper, we present an in-depth exploration of implementing image captioning using encoder-decoder models along with attention. We provide a comprehensive overview of the underlying concepts, discuss the implementation details, and compare the performance of these approaches using the consistent test set.

2 Data

The Flickr8k dataset is a widely used benchmark dataset for image captioning tasks. It contains a collection of 8,000 images, each of which is paired with five human-annotated captions. The images cover a wide range of topics and scenes, providing diverse examples for training and evaluation of image captioning models. The scenes often consist of people and animals performing sports or other activities.

The captions in the Flickr8k dataset are typically descriptive and provide detailed information about the content of the images. This richness in annotations makes the dataset suitable for training and evaluating image captioning models, allowing researchers to assess the quality and diversity of the generated captions. A few examples from this data set are shown below.1.

We used this data set for both training and testing the image captioning model. It was split with 85% used for training and 15% used for testing.

In order to use these Image / Caption pairs for training, we need to tokenize the words. We establish a vocabulary and tokenize each word. If words are out of the vocabulary, they are represented by <UNK>. Also <SOS> is start of sentence and <EOS> is end of sentence so that the RNN can lean how to end its caption. These tokens allow the RNN to generate correct sentences and learn the length of text generally correlated with an image.

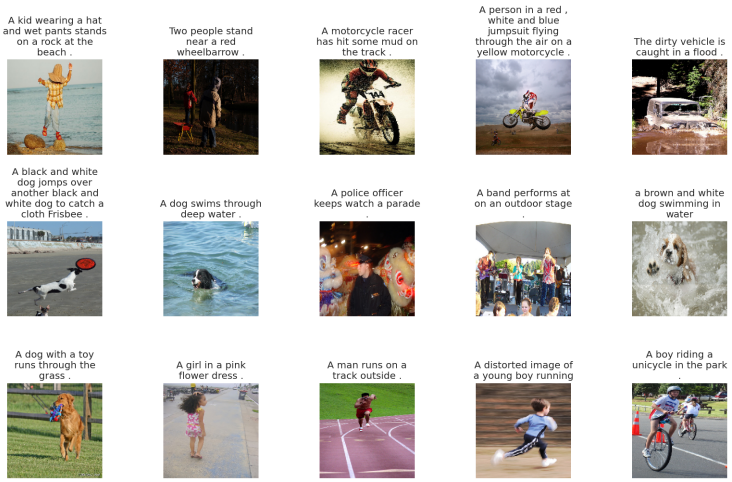

Figure 1: Flickr Examples

3 Image Captioning Architecture

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are widely used for sequential data processing tasks, including natural language processing and time-series analysis. In the context of image captioning, RNNs and LSTMs are employed to generate captions by sequentially predicting words based on the visual features extracted from the input image using a CNN.

The architecture of our image captioning model consists of two main components: an encoder and a decoder. The encoder is responsible for extracting visual features from the input image, while the decoder generates captions based on these features.

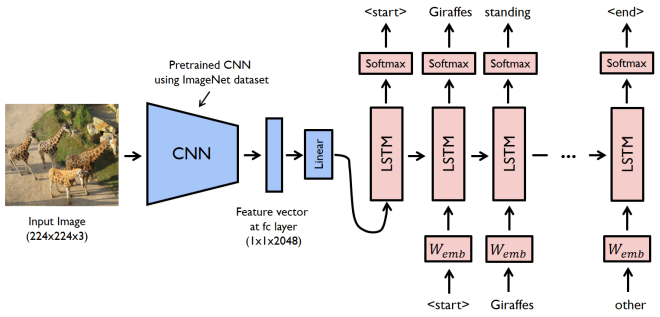

Figure 2: Model Diagram

3.1 Encoder

The encoder module utilizes a pre-trained convolutional neural network (CNN), ResNet50, to extract high-level visual features from the input image. Instead of passing these features to the last linear layers for classification, the encoder model returns the image features in a 2048 dimensional space. These extracted features are then passed to the decoder module for caption generation.

3.2 Decoder

The decoder module consists of an Recurrent Neural Network utilizing LSTM blocks. This network can learn captions based on the visual features provided by the encoder. At each time step, the decoder predicts the next word in the caption sequence conditioned on the previously generated words and the visual features of the image. For our basic approach we used a simple RNN with LSTM blocks but then adapted it to include an attention mechanism. An example of the images going through this training process is shown in the the Figure .

Figure 3: LSTM/RNN Training Examples

3.3 Attention

While basic RNN/LSTM-based decoder models have shown promising results in image captioning tasks, they often struggle to effectively capture long-range dependencies and focus on relevant regions of the image. Attention mechanisms address these limitations by dynamically attending to different parts of the image when generating captions.

In our model we included Bahdanau Attention to address these issues. This attention is used often for sequence tasks for the model to focus on various parts of the input during the sequential decoding phase. The attention layer calculates scores for each embedding dimension. Then during the RNN decoding, these scores are concatenated with the input feature embedding. Because this attention is calculated during each step of the sequential caption generation,

each word can be determined by using certain parts of the image. Check out the results from the trained model showing visually how the attention mechanism works.4 As shown in this example, it weighs certain sections of the image more when generating the sentence. The children are identified and then the fountain around them.

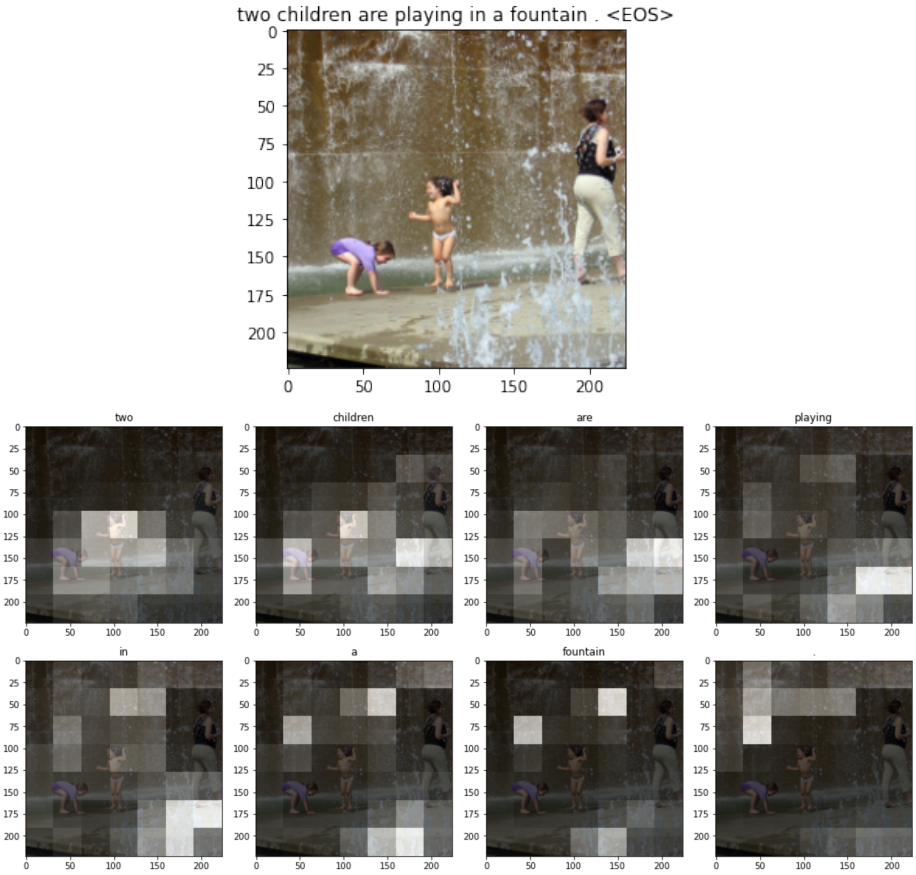

Figure 4: Visualization of Attention Mechanism

4 Methodology

4.1 Training

During training, the image-caption pairs from the training dataset are used to optimize the parameters of the model. The visual features are extracted using the encoder, and the decoder is trained to generate captions that accurately describe the corresponding images. The training objective involves minimizing

the loss function, cross-entropy loss, using an Adam optimizer, and a learning rate of 3e-4.

The key is that while the pretrained ResNet encoder weights are frozen, the RNN-LSTM decoder is updated during the training steps allowing the model to learn this multimodal space. The decoder learns how to properly decode the feature space based on these image-caption pairs.

4.2 Evaluation Metrics

To evaluate the performance of the RNN/LSTM-based image captioning model, we utilized BLEU (Bilingual Evaluation Understudy). The BLEU metric assesses the quality of the generated captions by comparing them to human-annotated references.

We can calculate this BLEU score between the generated captions and the ground truth, test-set captions.

In our evaluation, we used a smoothing function for segment-level scores, as present in Boxing chen and Collin Cherry (2014) A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU. Specifically, we used three different methods within the smoothing function to allow for robust comparison. A description of each method is found below:

Smoothing Method 1: Add ϵ counts to precision with 0 counts.

Smoothing Method 2: Adds 1 to both numerator and denominator from Chin-Yew Lin and Franz Josef Och (2004) ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation.

Smoothing Method 3: NIST geometric sequence smoothing. The smoothing is computed by taking $\frac{1}{2^k}$, instead of 0, for each precision score whose matching n-gram count is null. k is 1 for the first 'n' value for which the n-gram match count is null.

5 Results

For both model pipelines, we evaluated them using both the Bleu score and eye test on the test set. This test set is a subset of the flickr dataset that was not trained on.

5.1 BLEU Score

Using the BLEU score to evaluate the quality of generated captions on the test set, we found that the basic Encoder - Decoder model without attention model had average performance values of 0.02495, 0.13787, and 0.049580 with methods

1, 2, and 3 respectively. For attention, we got values of 0.02305, 0.12875, 0.0460, for methods 1, 2, and 3.

Comparing specifically method 2, the without attention model preformed slightly better by BLEU score (0.138 vs. 0.129) than the attention model for a couple of reasons. The first is because BLEU score measures the similarity between the generated text and the reference text(s) by counting overlapping n-grams. Generally, non-attention based caption models tend to produce shorter, simpler sentences that might align better with reference captions, hence potentially yielding higher BLEU scores. However, attention models tend to focus on different parts of the image (as shown below), which can lead to generating more detailed and contextually relevant captions that are more human-like and diverse–so they may not exactly match the reference captions. Therefore, the more diverse output can lead to lower BLEU scores.

While BLEU score is a great evaluation tool, it is not all encompassing. Higher BLEU scores can indicate better performance in terms of similarity to reference captions, but they may not reflect overall quality or human-likeness of generated captions, especially when using attention.

5.2 Visualizations

The eye test also shows that the model with attention generally performs better than the base version. Below in figure 5, we can see that the attention based model captions the clear images very well for the most part. It is able to determine colors, sports, people and dogs in these images. The words are not as creative as the human annotations, but it is able to pick on basic patterns.

We can also see how the attention mechanism picks up on various parts of the image as it generates the caption. Also shown in figure 6, the dogs are identified first and then the attention shifts to determine the red frisbee as what they are playing with.

6 Conclusion

In this paper, we have explored a encoder-decoder approach for image captioning utilizing RNN/LSTM-based models. While we used a consistent ResNet encoder, we experimented with adding attention to the RNN decoder. Both approaches have shown promising results in generating descriptive captions for images, with attention-based models often outperforming traditional RNN/LSTM-based models. However, the choice between these approaches depends on various factors such as the complexity of the task, computational resources, and available datasets.

Future research directions may include exploring and comparing other state of the art models such as vision transformers for this task. Also we could investigate alternative attention mechanisms and evaluating the performance of these models on diverse datasets across different domains.

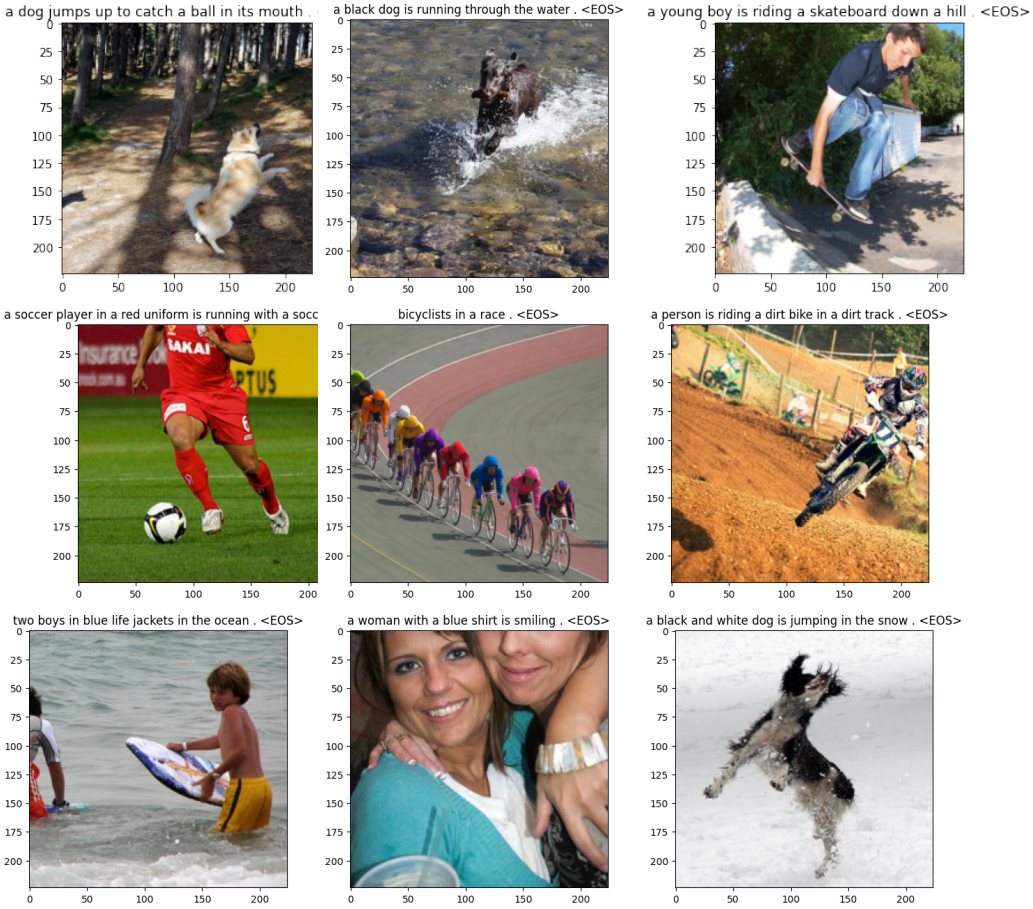

Figure 5: Results from Image Captioning w/ Attention on Test Set

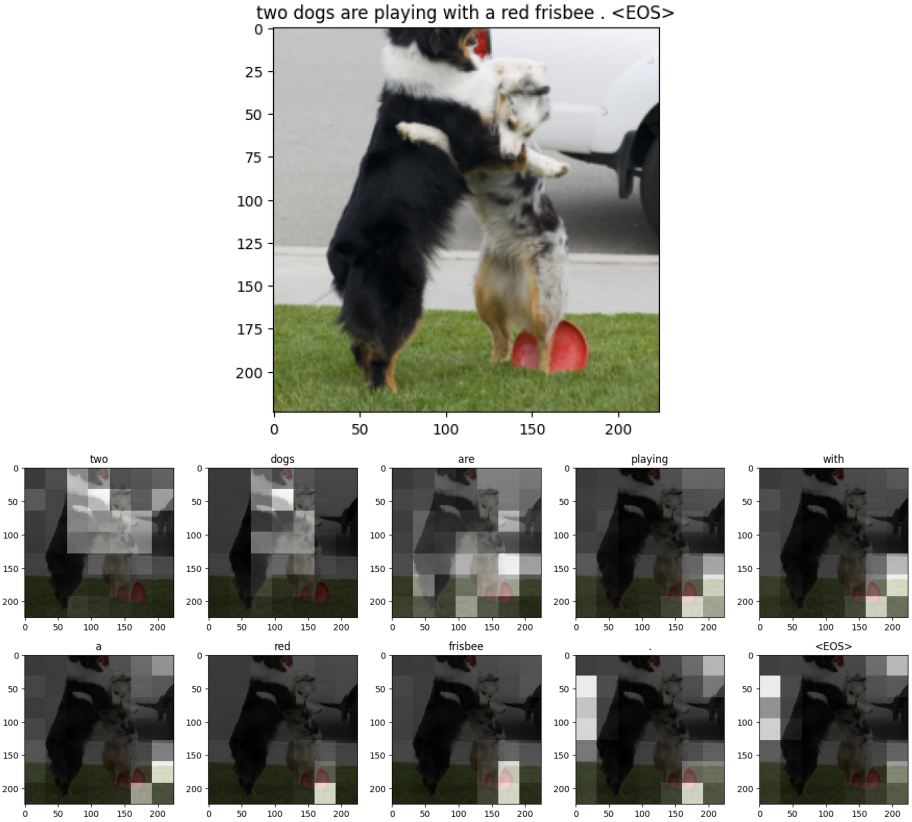

Figure 6: Visualization of Attention Mechanism