Image Captioning System

Team Members: Anagha Sarmalkar (801077504) , Shweta Patil (801074059)

1. Introduction

We have implemented an Image Captioning System which recognizes the important objects, their attributes and their relationships in an image and generates sentences that are syntactically and semantically correct.

This topic is very important for us because it provides us the opportunity to learn and understand the challenges in describing image attributes textually using Natural Language Generation. This topic is interesting because we understand a good deal about perception in terms of machines. This could be a stepping stone in generating search engines which are sentence-based.

Our main motivation in implementing this model lies in the fact that millions of images being uploaded on the internet can be utilized to create practical applications which range from helping the visually impaired to assisting auto-driving vehicles, but also because it is a major challenge in the combined fields of Natural Language Processing and Computer Vision.

The rest of the paper has been outlined as follows. Section 2 gives a brief overview of the related work in this field. Section 3 explains the exhaustive methodology that we implemented. Section 4 summarizes the findings from this project along with the analysis of results. Section 5 concludes this report with our accomplishments and future work.

2. Background

The motivation for artificial language generation has been derived from human motivation to exchange information. The ability to generate coherent descriptions can be an important step in video classification and image and text inference. The recent breakthroughs in image captioning provide good testimony.

Image captioning comes with its own set of unique challenges. First, to generate a semantically meaningful and syntactically fluent caption, the system needs to detect salient semantic concepts in the image, understand the relationships among them, and compose a coherent description of the overall content of the image, which involves language and common-sense knowledge modeling beyond object recognition. In addition, due to the complexity of scenes in the image, it is difficult to represent all fine-grained, subtle differences among them with the simple attribute of category. Moreover, unlike image classification tasks, where we can easily tell if the classification output is correct or wrong after comparing it to the ground truth, there are multiple valid ways to describe the content of an image. It is not easy to tell if the generated caption is correct or not, to what degree.

The techniques used for existing image captioning systems can be classified into three main categories. (1) Template-based image captioning, (2) Retrieval-based image captioning, and (3) Novel image caption generation [3].

Template-based image captioning has a fixed template with a number of blank slots to generate captions. However, templates are predefined and cannot generate variable captions [3]. In retrieval-based approaches, captions are retrieved from a set of existing captions. Retrieval based methods first find visually similar images with their captions from the training data set. These captions are called candidate captions. These methods, however, cannot generate specific and semantically correct captions [3].

Novel image caption generation methods employ deep machine learning-based techniques where models are trained to learn features using training data. Moreover, the models can handle a large and diverse set of images and videos. A hybrid system employing the use of a multilayer deep convolutional neural network is used to create a semantic representation of an image, which is followed by an LSTM network which is a special kind of recurrent neural network, capable of learning long-term dependencies [1]. [4] proposes a new algorithm that combines both top-down approaches through a model of semantic attention. This algorithm learns to selectively attend to semantic concept proposals and fuse them into hidden states and outputs of recurrent neural networks. Attention mechanisms at input and output layers attend to different aspects of visual attributes which provides a richer interpretation of the context and thus leads to improved performance.

[5] Described the importance of good embeddings with the quality of captions generated. The focus of their word was to generate full descriptive sentences from generated annotated regions represented as embeddings. [6] utilized attention inputs to deconstruct the image into weighted sections that represent the importance of the section.

3. Project

We propose a hybrid model that uses multilayer Convolutional Neural Network (CNN) which generates a vocabulary of objects in images and an LSTM which helps in structuring meaningful sentences using the existing vocabulary. The model has been tested on the Flickr8K dataset and the performance evaluating BLEU metrics has been reported. The Flickr8K dataset consists of 8K annotated images from Flickr.com [2]. These images are paired with five different captions, each of which describes the salient features of the images. BLEU metric helps evaluate a machine translation system by determining the quality of the caption translated from the natural language to another.

Approach:

- **Encoder**

  Images in the training dataset have been preprocessed using OpenCV and objects in the image are detected using transfer learning through the VGG model. The features of all the images in the training dataset are stored in pickle format to be used later. This helps us save time by not having to recompute the features for every image. This will enable faster training and less consumption of memory.

  The last layer of the VGG model is removed so that we get the objects/features (vector representation of the photos). Every image is of size 1 dimension, 4096 element vector.

- **Sequence Processor**

  This module performs the task of word embedding. Every image has been described in 5 sentences. These descriptions are first processed to remove punctuation, they are lowercased, and tokenized. A *startToken* and *endToken* are added to every sentence which acts as sentence limiters so that the RNN model can generate proper sentences. A vocabulary of unique words from the entire corpus of descriptions is created and saved in a pickle file. The words in each description are also tokenized which creates a vocabulary index based on word frequency. Every word gets a unique integer value where lower integer means more frequent words.

- **Model**

  The output from the sequence processor (image-wise descriptions) is encoded and padded so that all the descriptions are of uniform length (max length).

  The extracted visual features from Encoder and the description files from the Sequence Processor are then passed as input to the RNN model. The size of the vocabulary is taken into consideration (the length of the longest description) which will ensure that the generated captions do not exceed this length. This also helps in curbing long unending sequences of repeated words.

The Encoder model expects a 16 layer VGG model in the form of photo features of a vector of 4096 elements. A regularization in the form of a 50% dropout has been used. The features are then processed by a Dense layer to produce a 256 element representation of the image.

The Sequence Processor model expects padded sequences that have a specific length (max length). These sequences are fed into the Embedding layer which masks the sentences to ignore padded values. This is then fed to an LSTM layer with 256 memory units.

The Decoder takes outputs of 256 element representation from both the Encoder model and the Sequence Processor model and merges them using addition. This is then fed to a dense layer of 256 neurons and then to a final output dense layer which makes softmax predictions for the next word using the output vocabulary.

Layer (type)	Output Shape	Param #	Connected to
input_2 (InputLayer)	(None, 33)	0	
input_1 (InputLayer)	(None, 4096)	0	
embedding_1 (Embedding)	(None, 33, 256)	1860352	input_2[0][0]
dropout_1 (Dropout)	(None, 4096)	0	input_1[0][0]
dropout_2 (Dropout)	(None, 33, 256)	0	embedding_1[0][0]
dense_1 (Dense)	(None, 256)	1048832	dropout_1[0][0]
lstm_1 (LSTM)	(None, 256)	525312	dropout_2[0][0]
add_1 (Add)	(None, 256)	0	dense_1[0][0] lstm_1[0][0]
dense_2 (Dense)	(None, 256)	65792	add_1[0][0]
dense_3 (Dense)	(None, 7267)	1867619	dense_2[0][0]

Total params: 5,367,907
Trainable params: 5,367,907
Non-trainable params: 0

Figure 1. Model Summary

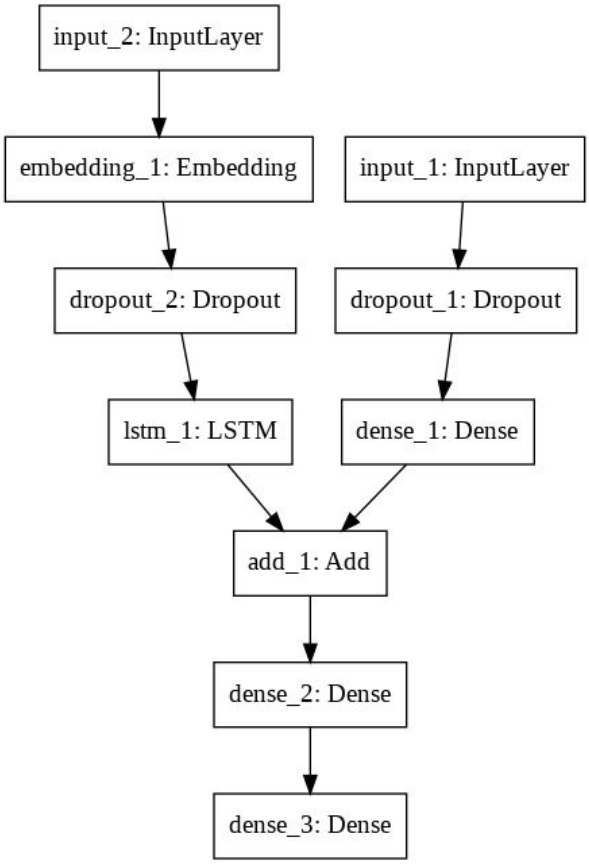

Figure 2. Model Architecture

This hybrid model is then fit on the training dataset for 15 epochs. As per the graph, the loss is minimum at epoch 14.

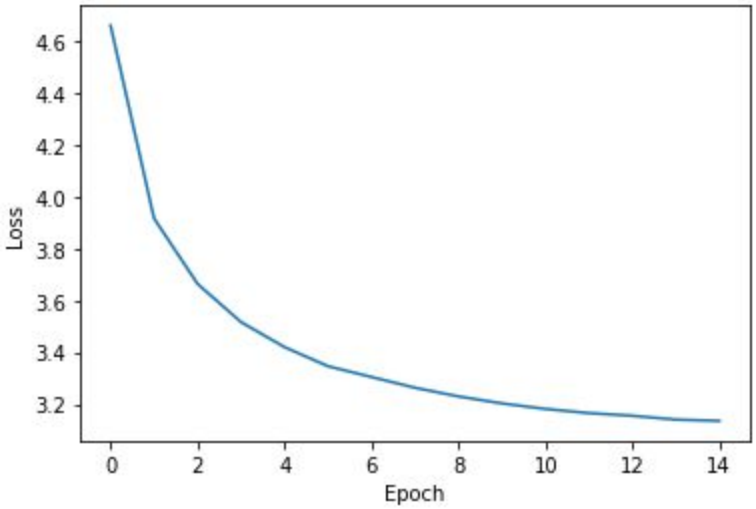

Figure 3. Loss Vs. Epoch

- **Generate Captions**

  The model with the least loss has been selected and captions are generated using this. Captions are created by mentioning a start token (*startToken)* for every caption. This word is then padded with zeros with the length of the maximum description. This is then passed to the RNN model along with the feature vector of the particular image. The RNN model then prints out a numpy array of possible words probabilities for this word and the image features. The word corresponding to the maximum probability is then extracted from the tokenizer and returned as the next word. If no word is returned then the caption generation is stopped since the word is not present in the vocabulary. If end token (*endToken)* is returned then the caption generation is stopped signifying sentence has ended. We generated captions for the test data. Following are the sample images:

Figure 4. Sample Image-Caption 1

Figure 5. Sample Image-Caption 2

Figure 6. Sample Image-Caption 3

Figure 7. Sample Image-Caption 4

- **Evaluation**

  The BLEU metric is used to evaluate the quality of the machine-translated output text generated. BLEU is short for the Bilingual Evaluation Understudy Score. It ranges between 0 and 1 where 1 corresponds to a perfect match and 0 corresponds to an imperfect match. Our model architecture generated captions with the following BLEU scores for unigram, bigram, trigram, and 4-gram. The captions generated are not accurate however they seem to be capturing the context of the image which is a good starting point.

  BLEU-1: 0.329550
  BLEU-2: 0.136576
  BLEU-3: 0.075454
  BLEU-4: 0.025752

- **Team Contribution**

Sr. No.	**Team Member Name (800 id)**	**Responsible For**
1	801074059	Exploratory Data Analysis, Training CNN model, RNN model
2	801077504	Exploratory Data Analysis, Constructing sequence processor, a model for generating captions, a model for evaluation.

4. Summary

We explained our motivation behind implementing an Image Captioning system. In recent years, biggest advances have been made in major Computer Vision tasks like object recognition, handwriting identification, etc through the use of Convolutional Neural Networks. RNNs and LSTMs have been crucial to some of the biggest advances in NLP. Implementing a task that falls at the intersection of NLP and Computer Vision was our greatest motivator.

We outlined a brief overview of the related works in this domain which gave us a broader idea of the field in general and redefined our focus on creating a prototype.

We implemented a model architecture which was proposed in [1] and were able to generate captions for images. The average unigram BLEU score of the captions is 0.32,

but we hope with access to large annotated image data, this performance can be enhanced. However, despite this BLEU score, the captions are able to identify the main feature in the image.

5. Conclusions

We were successful in implementing a deep learning approach for image captioning, with decent results. We learned about handling text and image data and designing a custom RNN-LSTM architecture for generating image captions. In the future, this system can be enhanced by using alternate pretrained photo models with more annotated data which will improve feature extraction. Similarly, training the word vectors on a larger data corpus can aid in improving performance.

6. References

[1] Lakshminarasimhan Srinivasan, Dinesh Sreekanthan, Amutha A.L, Image Captioning - A Deep Learning Approach.
[2] Micah Hodosh, Peter Young, Julia Hockenmaier, Framing Image Description as a Ranking Task:Data, Models and Evaluation Metrics.
[3] MD. ZAKIR HOSSAIN, Murdoch University, Australia FERDOUS SOHEL, Murdoch University, Australia MOHD FAIRUZ SHIRATUDDIN, Murdoch University, Australia HAMID LAGA, Murdoch University, Australia. A Comprehensive Survey of Deep Learning for Image Captioning.
[4] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo, Image Captioning with Semantic Attention.
[5] Andrej Karpathy, Li Fei-Fei, Deep Visual-Semantic Alignments for Generating Image Descriptions
[6] Kelvin Xu et al., Show, Attend and Tell: Neural Image Caption Generation with Visual Attention