Image Captioning System

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

1. Introduction

We will be attempting to implement an Image Captioning System which will recognize the important objects, their attributes and their relationships in an image. This attempt will also make sure to generate sentences that are syntactically and semantically correct.

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, but also because it is a major challenge in the combined fields of Natural Language Processing and Computer Vision.

2. Related Work

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 about 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, at 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 [1].

Template-based image captioning have a fixed template with a number of blank slots to generate captions. However, templates are predefined and cannot generate variable captions [1]. 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 [1].

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 multilayer deep convolutional neural network is used to create a semantic representation of an

image, which is followed by a LSTM network which is a special kind of recurrent neural network, capable of learning long-term dependencies [2]. [3] proposes a new algorithm that combines both top down approach 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 lead to improved performance.

3. Project Topic and Proposed Solution

For the purpose of this project we will be implementing the successful deep learning approach for image captioning system which uses an architecture of multilayer CNN and LSTM-RNN [2].

- **Encoder**
  The project will be divided into two main parts as visual feature extraction from the given set of images and generation of captions for images based on these extracted features. The CNN will act as an encoder which compresses the information in the original image into a smaller representation. We will use the VGG-CNN model due to its performance in object identification. This VGG model is trained to identify all possible objects in an image.

- **Sequence processor:**
  The text input in the form of captions will passed to the sequence processor which will extract features from the text. This module will perform the task of word embedding. Word embedding captures the context of a word in a document semantic and syntactic similarity, relation with other words, etc. Words are mapped in a vocabulary of vectors of numerical values This enables words with similar context occupy close spatial positions.

- **Decoder:**
  The extracted visual features and output from the sequence processor will be used as input to the decoder which is a RNN. We will implement the LSTM architecture of RNN which will be trained to predict the next word in the sentence depending upon the previous word. The sentence (caption) generated will be variable depending upon encountering the stop word where it marks the end of the sentence.

- **Dataset:**
  We will be leveraging the Flickr 8K dataset which is an annotated list of day-to-day images where each image is described by 5 description sentences, which to introduce a

degree of variance in the way images can be described and satisfy the dynamic nature of images. The data (8000 images) is partitioned into train data (6000 images), test data and development data (1000 images each).

- **Evaluation:**
  The evaluating parameter for this system will be BLEU which is short for Bilingual Evaluation Understudy Score. This is a metric for evaluating a generated sentence to a reference sentence. It ranges between 0 and 1 where 1 corresponds to a perfect match and 0 corresponds to an imperfect match. This particular architecture generated captions with an average BLEU score of 60.1[2] which is at par with a human generated sentence. We will train our model and generate image captions which will aim to achieve this score.

4. Project Timeline

Sr. No.	Date	Milestone achieved
1	September 9th	Project Proposal submitted
2	September 25th	Train CNN model (Encoder)
3	October 7th	Construct sequence processor.
4	October 14th	Project Progress due
5	October 28th	Train RNN model (Decoder)
6	November 18th	Testing, Evaluation, Fine tuning model
7	November 25th	Project Complete, final report due

5. Team Roles and Contributions

Sr. No.	Team Member Name (800 id)	Responsible For
1	801074059	Training CNN model, Training RNN model, Project reports, testing evaluation and fine tuning models
2	801077504	Constructing sequence processor, training RNN model, project reports, testing evaluation and fine tuning models

6. References

[1] 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.

[2] Lakshminarasimhan Srinivasan, Dinesh Sreekanthan, Amutha A.L, Image Captioning - A Deep Learning Approach.

[3] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo, Image Captioning with Semantic Attention.