Chapter 8
Automatic Image Captioning Using Different Variants of the Long Short-Term Memory (LSTM) Deep Learning Model

Ritwik Kundu
https://orcid.org/0000-0001-5666-8833
Vellore Institute of Technology, Vellore, India

Shaurya Singh
Vellore Institute of Technology, Vellore, India

Geraldine Amali
Vellore Institute of Technology, Vellore, India

Mathew Mithra Noel
https://orcid.org/0000-0002-3442-1642
Vellore Institute of Technology, Vellore, India

Umadevi K. S.
Vellore Institute of Technology, Vellore, India

ABSTRACT

Today's world is full of digital images; however, the context is unavailable most of the time. Thus, image captioning is quintessential for providing the content of an image. Besides generating accurate captions, the image captioning model must also be scalable. In this chapter, two variants of long short-term memory (LSTM), namely stacked LSTM and BiLSTM along with convolutional neural networks (CNN) have been used to implement the Encoder-Decoder model for generating captions. Bilingual evaluation understudy (BLEU) score metric is used to evaluate the performance of these two bi-layered models. From the study, it was observed that both the models were on par when it came to performance. Some resulted in low BLEU scores suggesting that the predicted caption was dissimilar to the actual caption whereas some very high BLEU scores suggested that the model was able to predict captions almost similar to human. Furthermore, it was found that the bidirectional LSTM model is more computationally intensive and requires more time to train than the stacked LSTM model owing to its complex architecture.

DOI: 10.4018/978-1-6684-6001-6.ch008

INTRODUCTION

Artificial Intelligence, a field in computer science that aims in giving computers the ability to mimic human-like intelligence, is being heavily deployed in building powerful and highly intelligent machines. Nowadays, Machine learning has become quite popular in the field of Artificial Intelligence; and is often used interchangeably with the term ‘Artificial Intelligence’. One of the most studied sub-domains of machine learning is Deep Learning, which provides high accuracy in its results, so its performance is high too through its output. One such field of work where artificial intelligence can be applied is image captioning. The idea of being able to explore more about perceptual tasks like image recognition or object detection has enabled researchers to take up more complex tasks which are much above and beyond image recognition (Ivašić-Kos et al., 2019). Image captioning has a huge positive impact on society, for instance it can be used for facilitating ones with visual impairment in understanding the different types of imagery data available on the internet without any external support. Image captioning entails extracting important content from an image and representing it in the form of a meaningful sequence of words. The key idea behind image captioning is to recognise and analyse objects, the relationships between them and the actions performed by the objects from a given input image (Hossain et al., 2019). In simple terms, the process of automatically describing the contents of an image by the use of deep learning and natural language sentences is called Image Captioning. This technique is used for the conversion of images, which are a sequence of pixels into a sequence of words. This can be considered as an end-to-end and sequence-to-sequence (seq2seq) problem. The authors of this chapter aim to build a model that can provide the caption for any image presented to it accurately and quickly.

In order to achieve their purpose of building a sustainable Image Captioning software, the concepts of Deep Learning will be implemented. A rapidly growing and researched domain, Deep Learning is gradually getting into all of our daily lives. It involves the use of ANN (artificial neural networks) using robust and high performing, premium and up-to-date hardware. Deep learning facilitates the development, training, and application of neural networks while keeping the time constraint required for the same as minimum. One such neural network is Convolution Neural Network (CNN). The Convolutional Neural Networks were designed for the purpose of mapping the input image data to an output variable. CNNs have proved to be very effective in these applications and thus they are very commonly used in a variety of prediction problems that involve images as an input. Long Short-Term Memory model (LSTM) is another such neural network. A variant of the Recurrent Neural Networks, it is used frequently with problems involving images. It is a distinct type of network, which has the capability of learning long-term dependencies within the data.

The primary objective of this chapter is to design a neural network that when trained can recognize real as well as synthetic images. It should be able to generate the most accurate caption for a given image in the matter of a few seconds. By combining the advantages of CNN and LSTM models, the aim has been to develop a novel Image Captioning model.

BACKGROUND

The complexity of images can vary widely from being described by a single word to requiring multiple phrases to describe a single image. The authors of this chapter have conducted a detailed analysis of different related studies conducted across the world in order to understand the different models used in