A Short Review on Image Caption Generation with Deep Learning

Soheyla Amirian[*], *Khaled Rasheed*[†], *Thiab R. Taha*[‡], *Hamid R. Arabnia*[§]

The University of Georgia
Athens, Georgia, USA

Abstract

Methodologies that utilize Deep Learning offer great potential for applications that automatically attempt to generate captions or descriptions about images. Image captioning is considered to be one of the intellectually challenging problems in imaging science. The application domains include: automatic caption (or description) generation for images for people who suffer from various degrees of visual impairment; the automatic creation of metadata for images (indexing) for use by search engines; general purpose robot vision systems; and many others. Each of these application domains can positively and significantly impact many other task-specific applications. This paper is not meant to be a comprehensive review of image captioning; rather, it is a concise review of image captioning methodologies based on deep learning, strengths and limitations, the datasets and the evaluation metrics used in automatic image captioning. Finally, a quick discussion about the software and hardware requirements for implementing an image captioning method is presented.

Index Terms— Deep Learning, Image Captioning, Long Short Term Memory (LSTM), Graphics Processing Unit (GPU), Tensor Processing Unit (TPU).

1. Introduction

Image processing has played and will continue to play an important role in science and industry. Its applications spread to many areas, including visual recognition [1] and scene understanding [2], to name a few. Before the advent of Deep Learning, most researchers used imaging methods that worked well on rigid objects in controlled environments with specialized hardware [3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. In recent years, deep learning based convolutional neural networks has positively and significantly impacted the field of image captioning allowing a lot more flexibility. In this paper, we attempt to highlight recent advances in the field of image captioning in the context of deep learning. Since 2012, many researchers have participated in advancing the deep learning model design [13], applications and interpretation [14]. The science and methodology behind deep learning have been in existence for decades, but an increasing abundance of digital data and the involvement of powerful GPUs has accelerated the development of deep learning research in recent years. Convenient development libraries such as TensorFlow and PyTorch, the open source community, large labeled datasets (e.g. MSCOCO, Flicker, and...) [15, 16], and splendid demonstrations simulate the explosive growth of the deep learning field.

Describing a scene in an image is a highly demanding task for humans. To create machines with this capability, computer scientists have been exploring methods to connect the science of understanding human language with the science of automatic extraction and analysis of visual information. Image captioning captioning need more effort than image recognition, because of the additional challenge of recognizing the objects and actions in the image and creating a succinct meaningful sentence based on the contents found. The advancement of this process opens up enormous opportunities in many application domains in real life, such as aid to people who suffer from various degrees of visual impairment, self-driving vehicles, sign language translation, human-robot interaction, and more. This paper surveys the state of the art approaches with a focus on deep learning models for image captioning. The models and generated captions are evaluated by using BLEU, METEOR, CIDEr [17, 18, 19], and other metrics.

This paper is a concise review of image captioning methodologies based on deep learning. This review begins by introducing the Image Captioning in Section 2. Then, a few recent methods of Image Captioning, the Datasets and Metrics are discussed in Section 3. Finally, Required Software and Hardware Platforms for implementing a model are mentioned in Section 4.

2. Image Captioning

Image captioning is the process of generating a concise description of an input picture/ image (See Figure 1). Typically, such functions are done manually. Automating this process

[*]Ph.D. Candidate, Department of Computer Science.
[†]Director, Institute of Artificial Intelligence; Professor, Department of Computer Science.
[‡]Professor and Head, Department of Computer Science.
[§]Professor and Graduate Program Director, Department of Computer Science.