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.
would be a significant contribution. A lot of research has been done on image captioning [20, 21, 22, 23] that are quite impactful. A system that automatically generates image captions can be utilized in many applications. Examples include: enhancing the accuracy of search engines; recognition and vision applications; enriching and creating new image datasets; enhancing the functionality of systems similar to Google Photos; and enhancing the optical system analysis of self-driving vehicles.

Fig. 1. These are a few examples of captions that has been generated for images.

For image captioning there are some challenges in how to extract visual information from the input and how to transform the visual information into a proper and meaningful language. Captioning research started with the classical retrieval and template based [20, 24] approaches in which Subject, Verb, and Object are detected separately and then joined using a sentence template. However, the advent of Deep Learning and the tremendous advancements in Natural Language Processing have equally affected the area of captioning. Hence, latest approaches follow deep learning based architectures that encode the visual features with Convolutional Neural Networks and decode with a language based model, which translates the features and objects given by an image based model to a meaningful sentence.

3. Image Captioning Methodologies

Automatically generating natural language sentences describing an image generally has two components: extracting the visual information, and describing it in a grammatically correct natural language sentence. Figure 2 depicts a simple Encoder-Decoder deep learning based captioning framework for image captioning. By convolutional Neural Network the objects and features are extracted from the image, then we need a network to generate a natural sentence based on the information that we have.

**Convolutional Neural Network:** There is a need for a model with a large learning capacity to learn about thousands of objects from a large number of images [14]. Deep learning presents computational models that are composed of multiple processing layers to learn representations of data in images [13, 25]. Deep learning based Convolutional Neural Networks play a key role in many applications, one of which is image recognition. Image recognition is used to perform a large number of visual tasks, such as understanding the content of images. There are several well-known models [13] in the field of CNNs based on object detection [1, 26, 27] and segmentation [28].

**Recurrent Neural Networks:** Sequence models like recurrent neural network (RNN) [29] have widely been utilized in speech recognition, natural language processing, and other areas. Sequence models can address supervised learning problems like machine translation [30], name entity recognition, DNA sequence analysis, and sentiment classification.

**Gated Recurrent Unit:** Gated recurrent unit (GRU) is a gating mechanism in RNN, introduced in 2014 by Cho et al. [30]. The basic RNN algorithm runs into vanishing gradient problem (a difficulty in training artificial neural networks). The gated recurrent units are an effective solution for addressing the vanishing gradient problem. They allow neural networks to capture a much longer range dependencies [29]. The advantage of the GRU is that it is a simple model and so it is actually easy to build a big network. Also, it only has two gates, so it computes quickly.

**Long Short Term Memory:** LSTM, as a special RNN structure, has proven to be stable and powerful for modeling long-range dependencies in various studies. LSTM can be adopted as a building block for complex structures. The complex unit in Long Short Term Memory is called a memory cell. Each memory cell is built around a central linear unit with a fixed self-connection [31]. LSTM is historically proven more powerful and more effective than a regular RNN since it has three gates (forget, update, and output). Long Short Term Memory recurrent neural networks can be used to generate complex sequences with long-range structure [32, 33].

3.1. Recent Deep Learning based Models:

There are many methods for image captioning. Earlier methods, prior to deep neural networks (DNNs), were retrieved-based [20] or template-based [24] models. Recent methods are based on deep neural networks. Generating an automatic caption for describing an image has two stages. First, the information needs to be extracted from the image and put it in a feature vector. This stage focuses on visual recognition by deep learning models. Then the feature vector is fed into the
second stage. The second stage is caption generation which is describing what is extracted in a grammatically correct natural language sentence. Figure 2 depicts an overall encoder-decoder structure for image captioning methods. So, we classified DNN-based methods on the basis of the main framework into subcategories that they respectively use. Here, a review of recent deep learning based methods for automatic image captioning is discussed. All are summarized in Table 1.

A major breakthrough in image captioning occurred in 2014 through the application of encoder-decoder models. Kiros et al. introduced an encoder-decoder pipeline model in which an encoder network takes the image as an input and extracts a fixed-size feature vector that a decoder network maps to a sequence of words. They set new best results when using the 19-layer Oxford convolutional network. They were also developing an attention-based model that jointly learns to align parts of captions to images. The generated descriptions are arguably the nicest ones to date [32]. The attention model is one of the models used in deep learning got from one of the most curious facets of the human visual system. Attention-based model learns to focus on different parts of the image. This is important when there is a lot of clutter in an image. However, this may cause losing information which could be useful for richer and more descriptive captions [17].

You et al. proposed the attention based approach that gives the state of the art performance on three benchmark datasets using the BLEU and METEOR metric (See section 3.3). They also showed how the learned attention can be exploited to give more interpret ability to the model generation process, and demonstrate that the learned alignments correspond very well to human intuition. Their model encourages future work in using visual attention [17].

You et al. also proposed a model of semantic attention. The algorithm combines top-down and bottom-up strategies to extract richer information from the image, and fuses them with a RNN that can selectively attend on rich semantic attributes detected from the image. They performed their method on the Microsoft COCO and the Flickr30K, and captioning system was implemented based on LSTM network. The image feature vector is extracted from the last 1024 dimensional convolutional layer of the GoogleNet [13] CNN model. Furthermore, their framework employs attention at both input and output layers to the RNN module. Their effort was exploiting abundant fine-grain visual semantic aspects, and fusing global and local information for generating a better caption. The experimental results show that the algorithm significantly outperforms the state-of-the-art approaches consistently across different evaluation metrics [21].

Fu et al. proposed the image caption system that exploits the parallel structures between images and sentences. One contribution of this system is that it aligns the process of generating captions and the attention shifting among the visual regions. Another is that it introduces the scene-specific contexts to LSTM that adapt language models for word generation to specific scene types. The architecture is that, an image is first analyzed and represented with multiple visual regions from which visual features are extracted. The visual feature vectors are then fed into a LSTM network, which predicts both the sequence of focusing on different regions and the sequence of generating words based on the transition of visual attention. The neural network model is also governed by a scene vector, a global visual context extracted from the whole image. Intuitively, it selects a scene-specific language model for generating text. This has been tested on several popular datasets, including MSCOCO, Flickr8K, and Flickr30K. They evaluated captions in BLEU-n, METEOR, ROUGE-L and CIDEr-D metrics (See Table 1). Either region-based attention or scene-specific contexts alone improve performance, but combining the two provides a further improvement [34].

In 2018, Aneja et al. developed a convolutional image captioning technique with existing LSTM techniques and also analyzed the differences between RNN based learning and their method. This technique contains three main components. The first and the last components are input and output word embedding respectively. However, while the middle component contains LSTM or GRU units in the RNN case, masked convolutions are employed in their CNN-based ap-
Table 1. The summary of a few recent works for Image Caption (All the results have been converted to percentages).

<table>
<thead>
<tr>
<th>Model</th>
<th>Architecture</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2014, Kiros et al.)</td>
<td>CNN+LSTM encoder–decoder Attention-based</td>
<td>Image Annotation result</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R@1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>23.0</td>
</tr>
<tr>
<td>(2015, Xu et al.)</td>
<td>CNN+RNN Attention-based</td>
<td>BLEU-1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>71.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hard attention on MSCOCO</td>
</tr>
<tr>
<td>(2016, You et al.)</td>
<td>CNN+RNN Attention-based</td>
<td>BLEU-1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Using the ground-truth visual attributes on MSCOCO</td>
</tr>
<tr>
<td>(2017, Fu et al.)</td>
<td>Region-based attention and scene-specific context VGG/Alex/ResNet + LSTM Attention-based</td>
<td>ENSEMBLE result</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BLEU-1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>72.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>on MSCOCO</td>
</tr>
<tr>
<td>(2018, Aneja et al.)</td>
<td>CNN+LSTM ResNet152 Attention-based</td>
<td>BLEU-1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>72.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>on MSCOCO</td>
</tr>
<tr>
<td>(2018, Anderson et al.)</td>
<td>CNN+LSTM Faster R-CNN, ResNet101 Attention-based</td>
<td>BLEU-1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>on MSCOCO</td>
</tr>
</tbody>
</table>

Anderson et al. proposed a combined bottom-up and top-down attention mechanism that enables attention to be calculated at the level of objects and other salient image regions. The bottom-up attention uses Faster R-CNN with ResNet-101 [13], which represents a natural expression of a bottom-up attention mechanism. The top-down mechanism uses task-specific context to predict an attention distribution over the image regions. The attended feature vector is then computed as a weighted average of image features over all regions. Their results on the MSCOCO dataset present a new state-of-the-art for the task, achieving CIDER, BLEU-4 scores of 117.9 and 36.9, respectively. Demonstrating the broad applicability of the method, applying the same approach to Visual Question Answering they obtained first place in the 2017 VQA Challenge [36].

3.2. Image Captioning Datasets:

There are a few datasets that are widely used to evaluate and compare image captioning methods: Flickr8K [23], Flickr9K [17], Flickr30k [17, 23] and Microsoft COCO [17, 18].

**Flickr**: The Flickr8K, 9k, and 30k datasets contain more than 8000, 9000, and 30000 images, respectively. Each image is annotated using Amazon Mechanical Turk with 5 independent sentences. The Flickr8K dataset contains human and animal images, while the Flickr30k dataset contains humans involved in everyday activities and events. For each image, five sentences are provided [17, 23].

**COCO**: Microsoft Common Objects in Context (MS-COCO) is a large-scale object detection, segmentation, and captioning dataset that contains 91 object categories, 328K images and five assigned captions to each image [17, 18, 37].

3.3. Image Captioning Evaluation Metrics:

Captions are evaluated using the BLEU, METEOR, CIDEr, and other metrics [17, 18, 19]. These metrics are common for comparing the different image captioning models, and have varying degrees of similarity with human judgment [38].

**BLEU**: BiLingual Evaluation Understudy is a method of automatic machine translation evaluation that is a precision-based metric, correlates highly with human evaluation, and has a little marginal cost per run [17, 39]. BLEU has different n-grams based versions for candidate sentences with respect to the reference sentences.
METEOR: Metric for Evaluation of Translation with Explicit ORdering is an automatic metric that evaluates translation hypotheses. It is based on a generalized concept of unigram matching between the machine-produced translation and human-produced reference translations [17, 18, 40, 41].

CIDEr: Consensus-based Image Description Evaluation enables an objective comparison of machine generation approaches based on their human-likeness, without having to make arbitrary calls on weighing content, grammar, saliency, etc. with respect to each other [19]. CIDEr was first developed specifically for evaluating image captioning tasks, but it is also used in video captioning methods.

ROUGE: Recall-Oriented Understudy for Gisting Evaluation determines the quality of a summary by comparing it to other summaries created by humans. ROUGE, similar to BLEU, has different n-grams based versions [42].

SPICE: Anderson et al. introduced Semantic Propositional Image Captioning Evaluation, a novel semantic evaluation metric that measures how effectively image captions recover objects, attributes, and the relations between them. It correlates more with human judgment of semantic quality as compared to previously reported metrics [43].

WMD: Word Mover’s Distance measures the dissimilarity between two text documents. Therefore, the sensitivity of this metric when compared to BLUE, ROUGE, and CIDEr, is low about word order or synonym swapping, but, like CIDEr and METEOR, provides high correlation with human judgments [44].

3.4. Discussion:

In this section, we briefly reviewed a few methods, according to the common approaches that they have used. For fair comparison of the models, Table 1 shows the results of attention-based methods on the MSCOCO dataset, the common dataset that they have utilized. We could state that Anderson et al. performed better on the MSCOCO dataset. This method outperformed previous works. The reason is that it uses the attention mechanism which focuses only on relevant objects of the image. Also, We found that the performance of a technique can vary across different metrics, parameters and datasets. Here, we tried to compare them based on the common performance criteria. However, image captioning still has a long way to go in improving the accuracy of captioning the events in images (See Figure 3).

4. The Required Platform for Implementation:

Deep Learning has dramatically improved the accuracy of image recognition. Image recognition is considered to be one of the most challenging problems in image science. In recent years, deep learning based convolutional neural networks have positively and significantly impacted the field of image recognition allowing a lot of flexibility. Deep Learning is responsible for many of the recent breakthroughs in image science, such as image captioning. Despite Deep Learning’s popularity, it is difficult to accurately predict the time that it takes to train a deep learning network to solve a given problem. The training time can be seen as the product of the training time per epoch and the number of epochs which need to be performed to reach the desired level of accuracy. We define the features which could influence the prediction of execution time while performing the training. We categorize these features into layer, implementation, and hardware features. Each of these categories can contain almost an endless list of features. Layer (Algorithm or model) Features include Activation Function (ReLU, Softmax, Tanh,...), Optimizer (Gradient Descent, Momentum, Adam,...), Batch Size (the number of training samples which are processed together as part of the same batch), Number of inputs to the layer, the neurons within the layer, Matrix, Kernel, Stride, and Padding size. Hardware Features include CPU, GPU or TPU technology (memory, clock, speed, bandwidth,...) [45].

4.1. Software Requirement:

Tensorflow: TensorFlow is an end-to-end open source platform for machine learning. TensorFlow is developed by Google and has integrated the most common units in deep learning frameworks. It supports many up-to-date networks such as CNN and RNN with different settings. TensorFlow...
is designed for remarkable flexibility, portability, and high efficiency of equipped hardware [46].

**PyTorch:** PyTorch is a Python-based scientific computing package that serves two purposes: as a replacement for NumPy to use the power of GPUs and as a deep learning research platform that provides maximum flexibility and speed\(^1\) [35].

**Keras:** Keras is a high-level neural network API, written in Python and capable of running on top of TensorFlow, CNTK, or Theano. It was developed with a focus on enabling fast experimentation. Being able to go from idea to result with the least possible delay is the key to doing good research. Keras allows for easy and fast prototyping (through user friendliness, modularity, and extensibility). Keras supports both convolutional networks and recurrent networks, as well as a combination of both. Keras runs seamlessly on CPU and GPU\(^2\).

![Fig. 4. Computing time dependencies](image)

### 4.2. Hardware Requirement:

The science and methodology behind deep learning have been in existence for decades. In recent years, however, there has been a significant acceleration in the utilization of deep learning due to an increasing abundance of digital data and the involvement of the powerful hardware.

**GPU:** Compared to CPU, the performance of matrix multiplication on Graphics Processing Unit is significantly better. With GPU computing resources, all the deep learning tools mentioned achieve much higher speedup when compared to their CPU-only versions [46]. GPUs have become the platform of choice for training large, complex Neural Network-based systems because of their ability to accelerate the systems. For example, it used to take a few days to train AlexNet (the work of Krizhevsky et al. \([14]\) which outperformed all other image recognition approaches at the time \([13]\)) on the ImageNet dataset with a NVIDIA K40 machine. Now with DGX-2, NVIDIA group are able to train AlexNet in a few minutes\(^3\). Shi et al. worked to evaluate the running time performance of a set of modern deep learning software tools and see how they perform on different types of neural networks and different hardware platforms. They showed that all tested tools can make good use of GPUs to achieve significant speedup over their CPU counterparts. There is no single software tool that can consistently outperform others, however, which implies that there exist some opportunities to further optimize the performance [46].

**TPU:** Tensor Processing Unit (Domain-Specific Architecture) is a custom chip that has been deployed in Google data centers since 2015. DNNs are dominated by tensors, so the architects created instructions that operate on tensors of data rather than one data element per instruction [47]. To reduce the time of deployment, TPU was designed to be a coprocessor on the PCI Express (PCIe) I/O bus rather than be tightly integrated with a CPU, allowing it to plug into existing servers just as a GPU does. The goal was to run whole inference models in the TPU to reduce I/O between the TPU and the host CPU. Minimalism is a virtue of domain-specific processors. Jouppi et al. show in their paper that the TPU leverages its advantages to run 15 times as fast as the K80 GPU, resulting in a performance/ Watt advantage of 29 times. While future CPUs and GPUs will surely run inference faster, a redesigned TPU using circa-2015 GPU memory would go three times faster and boost the performance/ Watt advantage to nearly 70 over the K80 and 200 over Haswell CPU [33, 47].

### 5. Conclusion and Future Work

There are many models that have already been presented to generate meaningful captions for images. These models are quite good, but have some constraints. Image captioning still have a long way to go in improving the accuracy of captioning the events in images (See Figure 3). We reviewed some of the recent deep learning based works, and it is hard to compare different works due to the different combination of structures, using different parameters and implying various datasets. We also noticed that there is a lot of room for improvement in accuracy.

Researchers attempt to give sight to the machines. First, machines learn to see. Then, they help us to see better. We will not only use the machines because of their intelligence, we will also collaborate with them in ways that we cannot even imagine. By improving image captioning models, we can further aid people with hearing or sight impairments as well as improving search engines.

\(^1\)https://pytorch.org/tutorials/beginner/blitz/tensor-tutorial.html  
\(^2\)https://keras.io/  
\(^3\)https://devblogs.nvidia.com/tensor-core-ai-performance-milestones/
6. Acknowledgement

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

7. References


[18] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and...


