Md. Musfiqur Rahman

ID: 1620277042

Paper name: A Short Review on Image Caption Generation with Deep Learning.

Author Name: Soheyla Amirian, Khaled Rasheed, Thiab R. Taha, Hamid R. Arabnia

Abstract: 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.

Methodology:

Automatically generating natural language sentences describing an image generally has two components:

Extracting the visual information,

Describing it in a grammatically correct natural language sentence.

Critical point:

Convolutional Neural Network (CNN):

There is a need for a model with a large learning capacity to learn about thousands of objects from a large number of images. Deep learning presents computational models that are composed of multiple processing layers to learn representations of data in images. 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 in the field of CNNs based on object detection and segmentation.

Recurrent Neural Networks (RRN):

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

Gated Recurrent Unit (GRU):

Gated recurrent unit (GRU) is a gating mechanism in RNN, introduced in 2014 by Cho et al. 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 dependency. 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):

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. 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.

Image Captioning Evaluation Metrics:

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

*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 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.:

Conclusion: 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.