A Neural Image Caption Generator

Emad Abd El-Hamied Nasr & $AbdEl - MoneamHassan$ & $EsraaAbdEl - Rahman$
Menna Adel Hamed & $MennaHaziemAhmed$

May 23, 2022

Abstract

Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. In this paper, we present a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. The model is trained to maximize the likelihood of the target description sentence given the training image.

1. Introduction

Every day, we encounter a large number of images from various sources such as the internet, news articles, document diagrams and advertisements. These sources contain images that viewers would have to interpret themselves. Most images do not have a description, but the human can largely understand them without their detailed captions. However, machine needs to interpret some form of image captions if humans need automatic image captions from it. Image captioning is important for many reasons. Captions for every image on the internet can lead to faster and descriptively accurate images searches and indexing. Ever since researchers started working on object recognition in images, it became clear that only providing the names of the objects recognized does not make such a good impression as a full human-like description. As long as machines do not think, talk, and behave like humans,

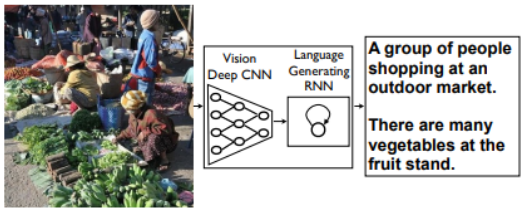

Figure 1. model, is based end-to-end on a neural network consisting of a vision CNN followed by a language generating RNN. It generates complete sentences in natural language from an input image, as shown on the example above.

1.1. MOTIVATION

Generating captions for images is a vital task relevant to the area of both Computer Vision and Natural Language Processing. Mimicking the human ability of providing descriptions for images by a machine is itself a remarkable step along the line of Artificial Intelligence. The main challenge of this task is to capture how objects relate to each other in the image and to express them in a natural language (like English).Traditionally, computer systems have been using predefined templates for generating text descriptions for images. However, 1 this approach does not provide sufficient variety required for generating lexically rich text descriptions. This shortcoming has been suppressed with the increased efficiency of neural networks. Many state of art models use neural networks for generating captions by taking image as input and predicting next lexical unit in the output sentence.

2. Related Work

The problem of generating natural language descriptions from visual data has long been studied in computer vision, This has led to complex systems composed of visual primitive recognizers combined with a structured formal language, e.g. And-Or Graphs or logic systems, which are further converted to natural language via rule-based systems. Such systems are heavily hand-designed, relatively brittle and have been demonstrated only on limited domains, e.g. traffic scenes or sports. The problem of still image description with natural text has gained interest more recently. Leveraging recent advances in recognition of objects, their attributes and locations, allows us to drive natural language generation systems, though these are limited in their expressivity. Farhadiet use detections to infer a triplet of scene elements which is converted to text using templates. Similarly, Liet start off with detections and piece together a final description using phrases containing detected objects and relationships. A more complex graph of detections beyond triplets is used by Kulkani, but with template-based text generation. More powerful language models based on language parsing The above approaches have been able to describe images "in the wild", but they are heavily hand designed and rigid when it comes to text generation. In this work we combine deep convolutional nets for image classification [12] with recurrent networks for sequence modeling [10], to create a single network that generates descriptions of images. The RNN is trained in the context of this single "end-to-end" network. The model is inspired by recent successes of sequence generation in machine translation, with the difference that instead of starting with a sentence, we provide an image processed by a convolutional net. The closest works are by Kiros et who use a neural net, but a feedforward one, to predict the next word given the image and previous words. A recent work by Mao uses a recurrent NN for the same prediction task. This is very similar to the present proposal but there are a number of important differences: we use a more powerful RNN model, and provide the visual input to the RNN model directly, which makes it possible for the RNN to keep track of the objects that have been explained by the text. As a result of these seemingly insignificant differences, our system achieves substantially better results on the established benchmarks. Lastly, Kiros propose to construct a joint multimodal embedding space by using a powerful computer vision model and an LSTM that encodes text. In contrast to our approach, they use two separate pathways (one for images, one for text) to define a joint embedding, and, even though they can.

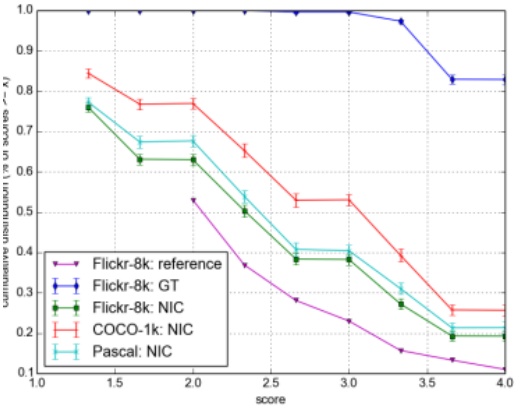

Figure 2. Datasets Accuracy

3. Model

In this paper, we propose a neural and probabilistic framework to generate descriptions from images. Recent advances in statistical machine translation have shown that, given a powerful sequence model, it is possible to achieve state-of-the-art results by directly maximizing the probability of the correct translation given an input sentence in an "end-to-end" fashion – both for training and inference. These models make use of a recurrent neural network which encodes the variable length input into a fixed dimensional vector, and uses this representation to "decode" it to the desired output sentence. Thus, it is natural to use the same approach where, given an image (instead of an input sentence in the source language), one applies the same principle of "translating" it into its description. Thus, we propose to directly maximize the probability of the correct description given the image by using the following formulation:
$\theta* = argmax \sum_{(I,S)} logp(S|I;\theta)$

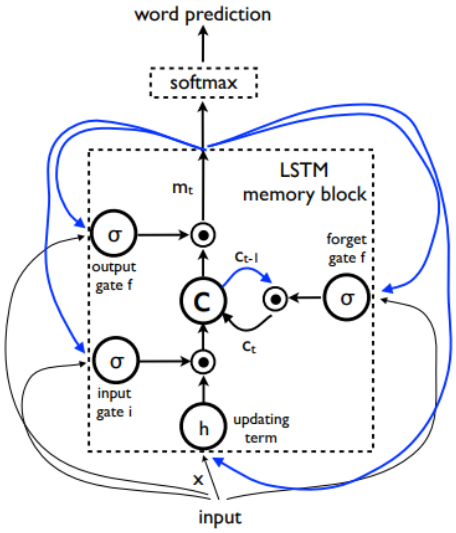

Figure 3. LSTM: the memory block contains a cell c which is controlled by three gates. In blue we show the recurrent connections – the output m at time t 1 is fed back to the memory at time t via the three gates; the cell value is fed back via the forget gate; the predicted word at time t 1 is fed back in addition to the memory output m at time t into the Softmax for word prediction.

3.1. Model Architecture

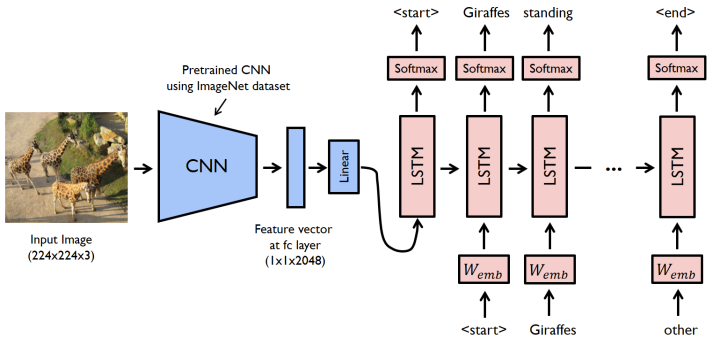

Figure 4. Model Arch.

4. Datasets

- Flicker8k
- Flickr30k
- MS COCO
- SBU
- Pascal

5. Training Details

We trained all sets of weights using stochastic gradient descent with fixed learning rate and no momentum. All weights were randomly initialized except for the CNN weights, which we left unchanged because changing them had a negative impact. We used 512 dimensions for the embeddings and the size of the LSTM memory. Descriptions were preprocessed with basic tokenization, keeping all words that appeared at least 5 times in the training set.

6. References

- Show and Tell: A Neural Image Caption Generator
- How to Develop a Deep Learning Photo Caption Generator from Scratch
- Where to put the Image in an Image Caption Generator