Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge

BEING able to identify the substance of a person automatically an picture that uses properly constructed It's a really difficult mission, but it could have a huge effect, because of the quality of web-based photos. This role is really critical. The bulk of past attempts have recommended sewing together Current solutions to the above-mentioned sub-problems with a view to moving from an image to its description[2],[3].

The primary inspiration for our study derives from the new Progress in computer translation, where the job is to convert Sentence written in its translation into the source. Using CNN as an icon is normal.By pre-training it first for an image classification, "encoder" Task to use the last secret layer as an RNN input Decoder, which creates sentences.

The problem of generating natural language descriptions from visual data has long been studied in computer vision, but mainly for video [8], [9]. Traditionally, this has led to complex systems composed of visual primitive recognizers combinedwith a structured formal language, e.g., And-Or Graphs or logic systems, which are further converted to natural language via rule-based systems.

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 stateof- the-art results by directly maximizing the probability of the correct translation given an input sentence in an “endto- 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.

The choice of f in (3) is governed by its ability to deal with vanishing and exploding gradients [25], the most common challenge in designing and training RNNs. To address this challenge, a particular form of recurrent nets, called LSTM, was introduced [25] and applied with great success to translation [4], [6] and sequence generation [37].

Since this model is data driven and trained end-to-end, and given the abundance of datasets, we wanted to answer questions such as “how dataset size affects generalization”, “what kinds of transfer learning it would be able to achieve”, and “how it would deal with weakly labeled examples”. As a result, we performed experiments on five different datasets, explained in Section 4.2, which enabled us to understand our model in depth.