Ahmed Mostafa
MLND Capstone Project Proposal

Domain Background

This project is my attempt to solve a problem I'm fascinated by to generate a meaningful caption from an image. It is a challenging problem because of the technical complexity for its implementation, and also because an image can have multiple captions/descriptions which makes the decision making process to specify captions hard.

Problem Statement

The main problem of this challenge is behind the proverb: *"A picture is worth a thousand words"*; because a single picture can be explained/captioned in so many ways depending on how people sees it.

It requires both Computer Vision and NLP to properly identify the objects in an image, then describe it properly with their relationships in a meaningful natural language.

Take an example of this picture below; it can be captioned in so different ways, for example:

- People sitting at a beach
- People sitting under an umbrella
- A bird is approaching group of people
- A child is playing in the beach.
- … and many others…….

Some of these captions are more descriptive (accurate) than others; we can't say any of them is wrong because each of them describes a single aspect of them.

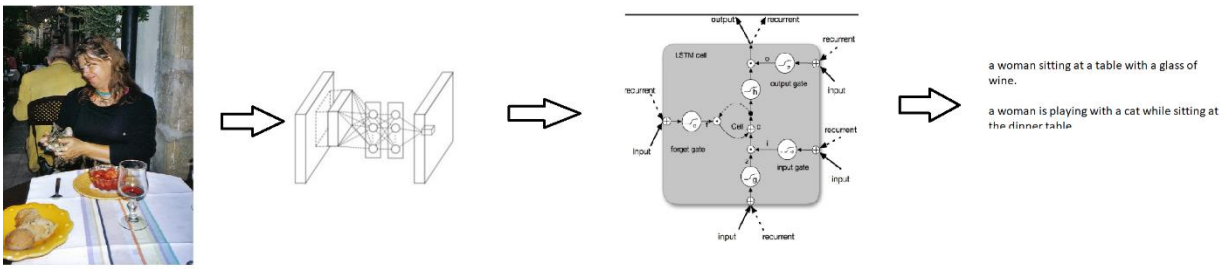
Datasets and Inputs

I'm going to build a deep learning pipeline based on the COCO dataset (and probably reuse one or more of the existing models to try TransferLearning) and see how far can I go describing an image. The COCO dataset can be downloaded directly from http://cocodataset.github.io. The dataset contains more than 91 common object categories [paper reference] with 82 of them having more than 5000 labeled instances. In total the dataset has more than 2,500,000 colored labeled instances in more than 328,000 images. The 2017 training dataset contains 17,887 files while the validation dataset contains 5000 files.

DESCLAIMER: Because of the limitation in resources (time, money, hardware, etc), I will be using a fractional subset of the dataset to perform this project. I estimate starting with 10% of the dataset and growing from there if needed.

Solution Statement

The solution at heart has a pipeline of convolutional neural network(s) CNN for features extraction and objects identification, and recurrent neural network(s) for building a word-by-word generation.



Benchmark Model

The benchmark model for this project will be the Google Neural Image Caption Generator. GNIC is a generative model based on a recurrent deep neural network that combines recent implementations in both computer vision and machine translation which can be used to generate natural sentences to describe images.

Evaluation Metrics

The evaluation metrics used in this project is the Bilingual Evaluation Understudy (BLEU). The idea behind BLEU is the closer a machine translation is to a professional human translation, the better it is, by measuring the difference between human and machine translation output.

Project Design

The project pipeline will be designed in multiple modules performing certain tasks, starting by data acquisition, data processing, training, evaluation and optimization. I still don't have a clear design, but I intend starting by an existing model (apply Transfer Learning) for data preprocessing as well as initial training (using VGG16/19 or ResNet50, or InceptionResNetV2, etc), I believe I will also need to apply word embedding and use LSTM RNN to generate sentences (similar to what we tried to do in DLFND). The usage of python scripts or Notebooks will be used to provide convenience as much as possible.