A cute little dog sitting in a heart drawn on a sandy beach.
A dog walking next to a little dog on top of a beach.

(Image source)

Automatic Image Captioning

13.10.2020

Vivek Mankar
mankarvivek2000@gmail.com
Pune, India

Overview/Domain Background

Artificial Intelligence shows us the tremendous possibility to change our way of working in different fields. Take the "Snapshot Serengeti"[1] project, Where more than 200 automatic cameras accrocs the savannah have amassed millions of photos of animals.But wildlife Biologists have had trouble figuring out how to actually deal with all that information. Means those were just photos and they needed data (example, "Two lions are eating a buffalo" rather than an Image) which is easier to store and manipulate.Turning all those photos into data has taken more than **30,000 hours**.The scientists build a computer program using AI(which usages Image classification and image captioning) that completed the task in just **one day.** In this project, we will develop and deploy an image captioning model that can automatically generate captions for the given image[2].

Problem Statement

Image captioning has various applications in virtual assistants, for image indexing, for visually impaired persons, for social media, and several other natural language processing applications.

For a scientific project Generating captions for images can become a tedious task and take years to complete.

PS: Develop a system that can automatically generate captions for a given image.

DataSets and Inputs

The dataset used for this project is "**Flickr8K**" dataset[3].

Flickr8K dataset includes images obtained from the Flickr website.

It is a labeled dataset consisting of 8000 photos.

This dataset is suitable for the project because there are 5 captions for each photo.

Solution Statement

Apply Deep Learning Techniques to develop an image captioning model that can automatically develop captions for a given image.

We will use a combination of CNN(Convolutional Neural Network) and RNN(Recurrent Neural Network) to develop this system.

First we will extract features of the image by using CNN , and then we will feed this feature vector to a LSTM language model that will generate captions.

Thus we will have a CNN Encoder and a LSTM Decoder.(LSTM is a special kind of RNN, capable of learning long-term dependencies)

We will also use pre-trained models on standard Imagenet dataset(provided in keras) to develop the CNN encoder.

Benchmark Model

In the past few years, the problem of generating descriptive sentences automatically for images has garnered a rising interest in natural language processing and computer vision research.Some of them are listed in the reference section from __ to __.All the papers are focused on increasing the performance of the image captioning model. There are various ways to measure the performance of an image captioning model like BLEU, ROUGE, CIDEr, METEOR, SPICE etc.[4] but out of these BLUE(Bilingual Evaluation Understudy) is most common and widely used in the evaluation of image annotation results, which is based on the n-gram precision. [5] achieved a effective BLEU score of 0.683 for their model.

Evaluation Metrics

As discussed above, There are various ways to measure the performance of an image captioning model. For this project we will be using BLUE score to check and compare the performance of our model. The principle of the BLEU measure is to calculate the distance between the evaluated and the reference sentences. BLEU method tends to give the higher score when the caption is closest to the length of the reference statement.

Project Design

Data Preprocessing :

Since the Flickr8K dataset contains 8000 different photos with 5 descriptions for each photo. The preprocessing will involve both processing of images as well texts.

We will first develop a feature extractor using transfer Learning(VGG19 model) to extract features from the images and save them in a .pkl file.

Then we will move towards text processing, we will start with loading and cleaning the descriptions (tokenization , converting to lowercase, removing punctuation, etc) and then save the cleaned descriptions in a .txt file. We will also create a vocabulary consisting of all the words in the descriptions.

Model Development and Training:

We will first split data into training and testing sets. Then we will develop a deep learning model that will take an image and an input_sequence(since we are training a language model) as an input and will output a sequence of words(caption).

The training will be done on google Colab or AWS sagemaker to achieve faster results.

Then we will save the model weights to use them in our final web application.

Model Deployment:

We will create a streamlit web app to deploy our DL model.Streamlit is an awesome platform to develop machine learning apps.

References

[1]https://www.zooniverse.org/projects/zooniverse/snapshot-serengeti

[2]https://www.youtube.com/watch?v=tSoqJpisKIg

[3]https://www.kaggle.com/shadabhussain/flickr8k

[4] Shuang Liu, Liang Bai,a, Yanli Hu and Haoran Wang College of Systems Engineering, National University of Defense Technology,410073 Changsha, China

[5]Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares :Image Captioning: Transforming Objects into Words,Yahoo Research San Francisco, CA, 94103

[6] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, Serge Belongie,Department of Computer Science, Cornell University https://vision.cornell.edu/se3/wp-content/uploads/2018/03/1501.pdf

[7]Lakshminarasimhan Srinivasan, Dinesh Sreekanthan,Amutha A.L : Image Captioning - A Deep Learning Approach, https://www.ripublication.com/ijaer18/ijaerv13n9_102.pdf