Image Captioning using Deep Learning

MAJOR PROJECT SYNOPSIS

BACHELOR OF TECHNOLOGY

DEPARTMENT OF COMPUTER ENGINEERING

Submitted By:

Uzma Afzal (17BCS085)

Newtan Kumar Rai (17BCS086)

Azzam Jafri (17BCS092)

Submitted To:

Mr Danish Raza Rizvi

JAMIA MILLIA ISLAMIA,NEW DELHI

March 2021

Abstract

The objective of our project is to build a model that can generate a descriptive caption for an image we provide it.
This model learns where to look.
As we generate a caption, word by word, you can see the model's gaze shifting across the image.
This is possible because of its Attention mechanism, which allows it to focus on the part of the image most relevant to the word it is going to utter next.
Our first will work in three phase.
Object Detection, Semantic Segmentation, Image Captioning.

Introduction

Image Captioning is a problem in computer vision where the algorithm detects objects in an image, understands the relationship among those objects and outputs a grammatically consistent sentence summarizing the key information in the image.

Encoder-Decoder architecture. Typically, a model that generates sequences will use an Encoder to encode the input into a fixed form and a Decoder to decode it, word by word, into a sequence.

Attention. The use of Attention networks is widespread in deep learning, and with good reason. This is a way for a model to choose only those parts of the encoding that it thinks is relevant to the task at hand. The same mechanism you see employed here can be used in any model where the Encoder's output has multiple points in space or time. In image captioning, we consider some pixels more important than others. In sequence to sequence tasks like machine translation, we consider some words more important than others.

Transfer Learning. This is when we borrow from an existing model by using parts of it in a new model. This is almost always better than training a new model from scratch (i.e., knowing nothing). As we will see, you can always fine-tune this second-hand knowledge to the specific task at hand. Using pretrained word embeddings is a dumb but valid example. For our image captioning problem, we will use a pretrained Encoder, and then fine-tune it as needed.

Beam Search. This is where you don't let your Decoder be lazy and simply choose the words with the best score at each decode-step. Beam Search is useful for any language modeling problem because it finds the most optimal sequence.

Proposed Method / Algorithm

Encoder

=>We use Convolutional Neural Networks(CNNs) for encoding input images.

=>We will use the 101 layered Residual Network trained on the ImageNet classification task, available in PyTorch.

Decoder

=>The Decoder's job is to look at the encoded image and generate a caption word by word. We will use Recurrent Neural Network (RNN) LSTM for decoding.

=>We want the Decoder to be able to look at different parts of the image at different points in the sequence, we will use Attention mechanism for that.

=> We use a linear layer to transform the Decoder's output into a score for each word I in the vocabulary.

Programming Environment & Tools used:

Platform used : Jupyter notebook

Framework used : PyTorch

Dataset used : MS COCO 2014

Language : Python

References:

Image Captioning with Attention (cs231n Stanford cl;ass).
http://cs231n.stanford.edu/reports/2016/pdfs/362_Report.pdf

Attention-Based Deep Learning Model for Image Captioning: A Comparative Study.
https://www.researchgate.net/publication/333564842_Attention-Based_Deep_Learning_Model_for_Image_Captioning_A_Comparative_Study

Learning a Recurrent Visual Representation for Image Caption Generation.
https://arxiv.org/pdf/1411.5654.pdf