GPT-2 for Image Description Generation

Xavier Reyes Tejedor

Description

Transformers are widely used to convert words to meaningful vectors, as we see in the Transformer Encoder architecture (see BERT), or to generate new words from other words, as we see in a Transformer decoder or in the full Transformer architecture (see GPT).

Taking a look at the inverse task of that, we can see that the research on generating words from meaningful vectors has not been as widely investigated. It is for this reason that this project will be centered on generating descriptive sentences for images, using the power of **Convolutional Neural Networks** and the **Language Model based on Transformer Decoders GPT-2**.

Datasets

The GPT-2 model is already pretrained, which is a plus. The Convolutional Neural Network which will be used could be trained from scratch or using a pretrained one (like EfficientNet on ImageNet).

To fine-tune these joint network for the task in hand we could use the `Flickr8K`, the `Flickr30K` or the `MSCOCO 2014` datasets, which provide annotations for given images and will be useful.

Work Plan

1. Detect the best dataset for the task in hand.
2. Conceptualize possible architectures that may work correctly.
3. Download and run both the Vision and Language model.
4. Build an fine-tune the model conceptualized.
5. Fine-tune and validate the model to achieve the best possible results.

Reference paper and baseline

There has been multiple papers on the topic, being **Deep Visual-Semantic Alignments for Generating Image Descriptions** (https://cs.stanford.edu/people/karpathy/cvpr2015.pdf) or **Show and Tell: A Neural Image Caption Generator** (https://arxiv.org/pdf/1411.4555.pdf).

These papers, though, use RNN and are dated from 2015 for the most recent ones. With the new breakthoughs in the NLP area, I think those results can be improved substantially to get much richer annotations.

Evaluation

The papers provided release results on publicly available datasets. If I run the model on these datasets and circumstances, I will be able to compare the model more effectively.

It is worth to note that this model will be much more complex than previous approaches, since only the smallest version of GPT-2 alone contains more than 100 million parameters.

Some type of cross-validation (like 10-fold) will be possible, specially if we use a pre-trained CNN. This is possible due that only fine-tuning will be necessary, and the training time will be reduced (very) substantially even if the model is orders of magnitude more complex.