Graduation Project

Image Captioning

Abstract

Automatically describe the content of an image using properly formed English sentences

Mohamed Khaled Ahmed Atyaa
Hesham Mostafa Mohamed Sherif

Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. In this final project we will define and train an image-to-caption model that can produce descriptions for real world images. To do that we will use CNN encoder and RNN decoder. First of all because of the data is huge as we will work on MSCOCO dataset which has 82783 training set and 40504 validation set. So we find that coursera website prepare some files which shorten the way of our model.

Download data

Takes 10 hours and 20 GB. coursera downloaded necessary files for us.

Relevant links (just in case):

- train images http://msvocds.blob.core.windows.net/coco2014/train2014.zip
- validation images http://msvocds.blob.core.windows.net/coco2014/val2014.zip
- captions for both train and validation http://msvocds.blob.core.windows.net/annotations-1-0-3/captions_train-val2014.zip

Extract image features¶

We will use pre-trained InceptionV3 model for CNN encoder (https://research.googleblog.com/2016/03/train-your-own-image-classifier-with.html) and extract its last hidden layer as an embedding.

As features extraction takes too much time on CPU:

- Takes 16 minutes on GPU.
- 25x slower (InceptionV3) on CPU and takes 7 hours.

- 10x slower (MobileNet) on CPU and takes 3 hours.

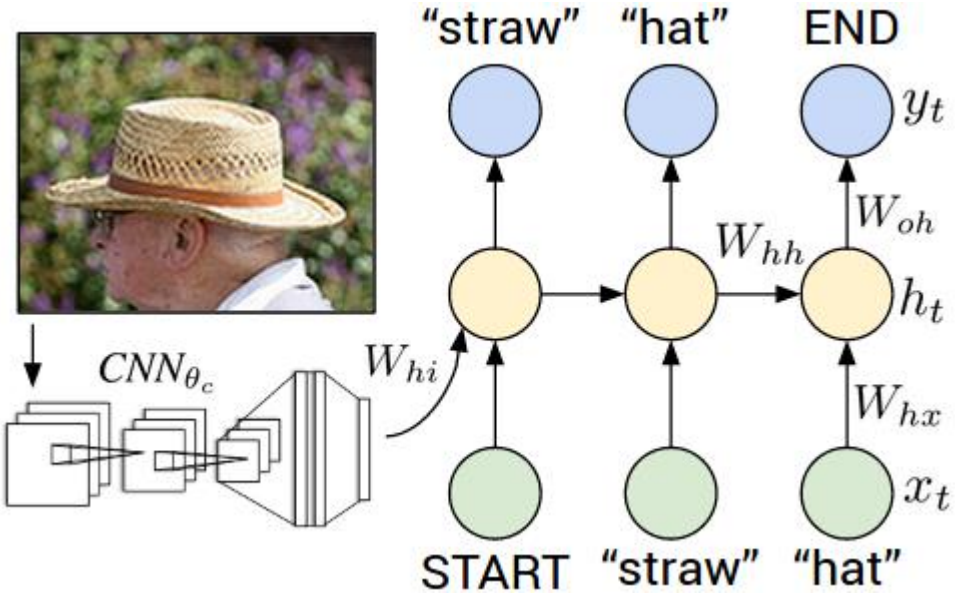
So coursera done it for us and save it in a pickle files. Then we Extract captions for images.

Training

Define architecture

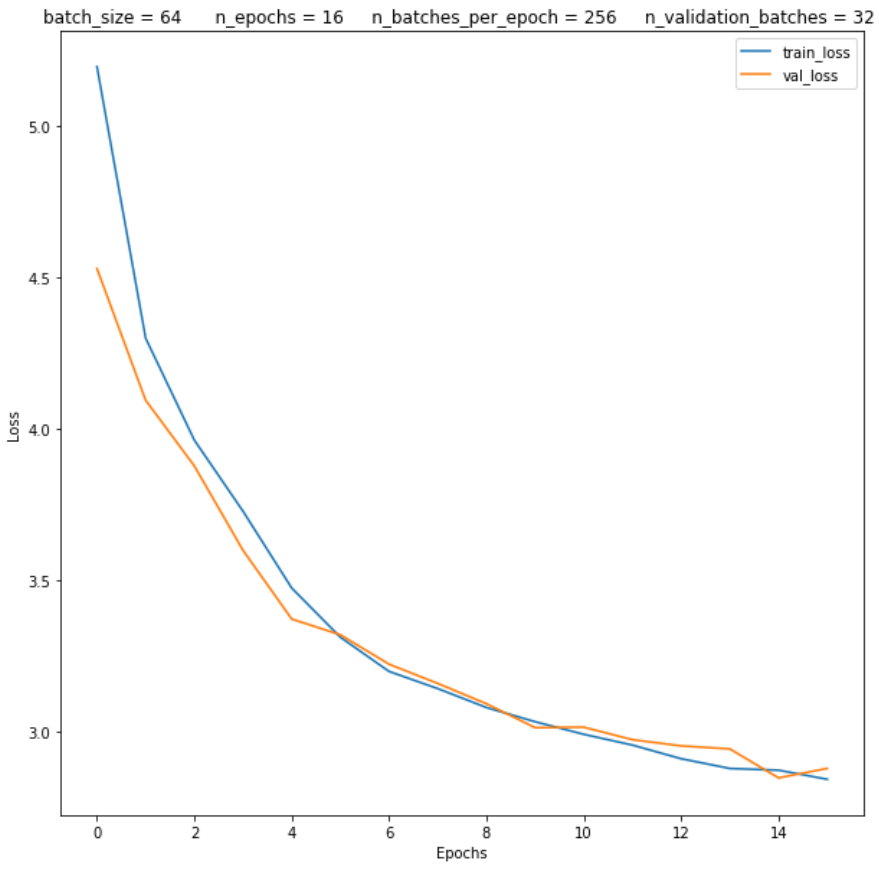
Since our problem is to generate image captions, RNN text generator should be conditioned on image. The idea is to use image features as an initial state for RNN instead of zeros. We approach that based on the idea of co-embedding of images and text in the same vector space. For an image query, descriptions are retrieved which lie close to the image in the embedding space. Most closely, neural networks are used to co-embed images and sentences together [29] or even image crops and sub sentences but do not attempt to generate novel descriptions.

During training we will feed ground truth tokens (start & end) into the lstm to get predictions of next tokens.



As the model is need heavily computational power we use coursera machine and split the training into to trails as the session on this machine expire in a certain time before we can train the whole epochs so we divided the training and use the

weights from the first trial to feed it to the second trail we can see that error in the first trial decreased this way

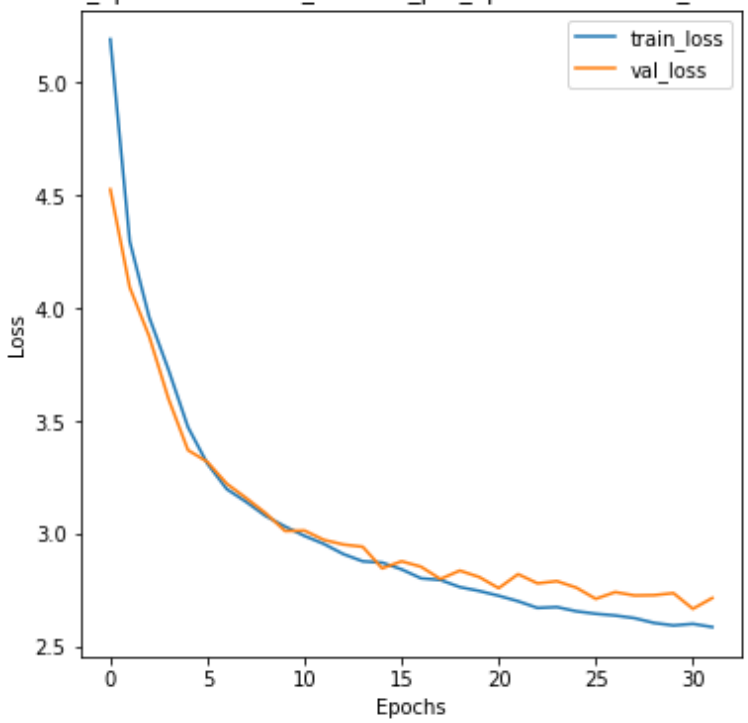


After we train the first trial which use:

- batch_size = 64
- n_epochs = 16
- n_batches_per_epoch = 256
- n_validation_batches = 32

We use the weights to initialize the second trial and the error decreases more

batch_size = 64 n_epochs = 32 n_batches_per_epoch = 256 n_validation_batches = 32



Actually we act as we train the model on 32 epoch but divided into to trail we can see the accuracy enhancement on the image description

From the first trail :

```
a man in a snow board on a snowboard
```

The same image from the second trail:

```
a skier is skiing down a snowy mountain
```

We see the caption enhances from first trail to the second trial through fine tuning the second trail with the weights of first trial.

Conclusion

We have presented an end-to-end neural network system that can automatically view an image and generate a reasonable description in plain English. The model is based on a convolution neural network that encodes an image into a compact representation, followed by a recurrent neural network that generates a corresponding sentence. The model is trained to maximize the likelihood of the sentence given the image.