Homework 5

Carlos Sancini

1. What is TensorFlow? Which company is the leading contributor to TensorFlow?

   Tensorflow is a framework for deep learning and Google is the leading contributor.

2. What is TensorRT? How is it different from TensorFlow?

   TensorRT is platform for running deep learning frameworks in NVIDIA's CUDA hardware. While Tensorflow is an API for programming deep learning models, TensorRT is a runtime environment that optimizes the underlying computations in the GPU(s).

3. What is ImageNet? How many images does it contain? How many classes?

   ImageNet is a database of labeled images that aimed to foster the developments of deep learning models and techniques. The web page of says that it contains 14,197,122 images and 21,841 synsets indexed.

4. Please research and explain the differences between MobileNet and GoogleNet (Inception) architectures.

   GoogleNet contains several inception modules that compute in parallel multiple transformations over the same layer and their results are concatenated into a single output. Then, the next layer of the model defines how to use this information.

   MobileNet authors that it uses an operation known called "depth wise separable convolution, i.e., a depth wise convolution (a spatial convolution performed independently for each channel) followed by a pointwise convolution (a 1x1 convolution across channels). This is similar to first looking for correlations across a 2D and second looking for correlations across a 1D. The 2D + 1D mapping is less computationally intensive than a full 3D mapping.

5. In your own words, what is a bottleneck?

   When a layer of the neural network has less neurons than the layers below or above, it forces the network to compress or select the features, retaining the most important information.

6. How is a bottleneck different from the concept of layer freezing?

   With layer freezing, in the context of transfer learning, the weights of a frozen layer kept unchanged during backpropagation. This means that during fine tuning only the top classification layer or additional custom layers are changed.

7. In part one this lab, you trained the last layer (all the previous layers retain their already-trained state). Explain how the lab used the previous layers (where did they come from? how were they used in the process?)

   The previous layers are part of a pre-trained model that was downloaded. A classification layer was added on top of the pre-trained layers, which were kept frozen during the fine-tuning process, i.e., the training of the classification layer.

8. Why is the batch size important? What happens if you try running with a batch size of 32? What about a batch size of 4?

   The batch size is import because it saves memory and the model converges faster. If we choose a batch size of 32, the parameters will be update much less than of 4.

9. Find another image classification feature vector from tfhub.dev and rerun the notebook. Which one did you pick and what changes, if any did you need to make?

   It picked the Inception_v3 model from

   https://tfhub.dev/google/tf2-preview/inception_v3/feature_vector/4

   The only change to the notebook was the URL of the pre-trained model.

10. How long did the training take in part 2?

    It took 3h to train the model on the TX2.

11. In part 2, you can also specify the learning rate using the flag --learning_rate. How does a low --learning_rate (part 2, step 4) value (like 0.001) affect the precision? How much longer does training take?

    Precision increases but training time gets longer.

12. How about a --learning_rate (part 2, step 4) of 1.0? Is the precision still good enough?

    Precision diminishes but training time is shorter.

13. For part 2, step 5, How accurate was your model? Were you able to train it using a few images, or did you need a lot?

    No, just a few thousands.