Big Data

Big data is the field that deals with analysing and extracting meaning information from very large datasets that cannot be handled efficiently by current and past software technologies. Usually this large data set will consist of many rows and columns [1]. The associated challenges are extensive and include; data capturing, storage and visualization [1].

In this coursework I will be using Google Cloud to assess the performance of the inception model applied to two datasets through Cloud-ML framework. I will test the overall performance of the inception algorithm applied to each dataset by varying server/cluster configurations with and without GPU. I will also explore the effect of dropout for each dataset whilst documenting metric development during training for each model.

Task 1:

The second data-set I have chosen is Coastline as it has > 10000 images. To successfully pre-process this data-set I had to copy the preprocess.py file and change 'flowers' to 'coastline'. Since dataset is much bigger than flowers, epoch size for inception model had to be increased to 5000. Also, evaluation set size was increased to 2343 and the label count was increased to 18, without this increase the model training would not run.

The prediction cannot be run from the coastline.sh-file or flowers.sh-file and must be typed into google-shell. The jpeg coastline image has been compressed and saved in the bucket before pre-processing and training of model. The cluster/server location are set to us-central1 by default and will not be changed. As expected the coastline dataset takes much longer to pre-process and training of model takes longer too.

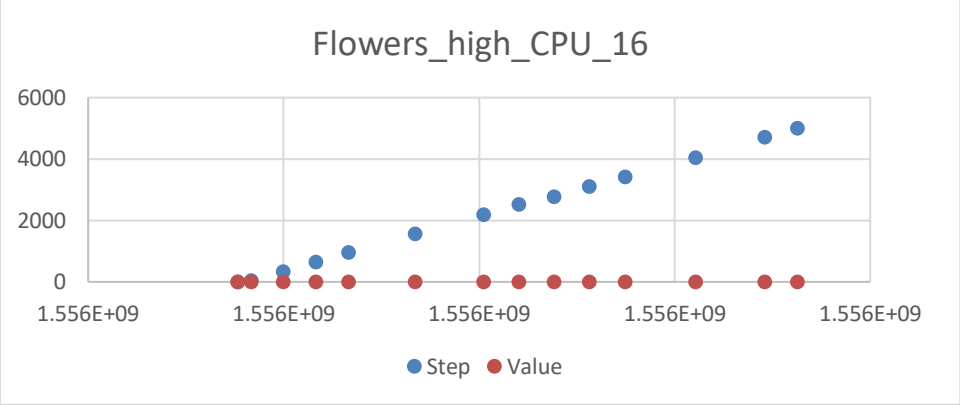
Task 2:

Experimenting with CPU for the flowers data-set, I was unable to use machine type; 'complex_model_gpu' configuration as this uses 4 accelerators, 8 virtual CPUs and 30 GB of memory, the error message given on the shell-command; 'The request for 4 K80 accelerators exceeds the allowed maximum of 1 K80, 1 P100, 1 P4, 1 V100, 8 TPU_V2'. Also, CPU was limited to only 20, error given; ': Quota failure for project level-clone-237916. The requested 32.0 CPUs exceeds the allowed maximum of 20.0.'. For the above reasons, I will use CPU counts of 8 and 16. The region us-east1 seemed to have suffered from some sort of failure as google cloud was unable to run for more than an hour, this probably caused a shortage of available/usable CPUs as I continually got the error; 'No zone in region us-east1 has accelerators of all requested types'.

For GPUs, I will experiment with 'standard_gpu' which has one NVIDIA Tesla K80 GPU, 8 Virtual-CPUs and 30 GB of memory and 'standard_p100' which has one NVIDIA Tesla P100 GPU, 8 Virtual-CPU and 30 GB.

Flowers;	**High-CPU-16**	**complex_model_s-8 CPU**	**STANDARDGPU**	**standard_p100**
Eval-Accuracy	0.9133	0.6868	0.8123	0.8949
Time-To-Train-Model	16 minutes	16 mins 34 sec	17 mins 53 sec	15 min 44 sec

Training model which uses 16 CPUs has higher evaluation accuracy but second worst training time, the figure below shows this graph with x-axis representing wall time with plotted dots for number of steps and accuracy. We can clearly see as time and steps increase so does the accuracy.



Coastline;	**High-CPU-16-No GPU**	**Coastline Tesla GPU with 8CPU**	**16 CPU with GPU**
Eval-Accuracy	0.7087	0.6861	0.6903
Time-To-Train-Model	9 min 18 sec	8 min 39 sec	8 min 45 sec

Again, 16 CPU without GPU seems to give a higher evaluation accuracy than with GPU and as with the previous experiment, it takes longer to train the model with only CPU.

Task 3;

For this section I will explore the effect of dropout rate for the inception V3 neural network, this will be done by randomly dropping out neurons from the network. This is done to regularize and limit the co-adaptability of neural network. The trained models will all run on 16 CPU as this gives the best

evaluation accuracy, dropout will be at 0.1, 0.5 and 0.9. The table below displays flowers dataset with 16-CPU and dropout rates as stated above;

Flowers;	**16-CPU D = 0.1**	**16-CPU D = 0.5**	**16-CPU D = 0.9**
Eval-Accuracy	0.9045	0.9004	0.6206
Time-To-Train-Model	15 min 50 sec	16 min 49 sec	15 min 58 sec

As the dropout rate increase the evaluation/training set accuracies decrease substantially, this can be deduced intuitively as a dropout rate of 0.9 means only 10% of neurons have been used, also completion time seems unaffected. Dropout rate is deployed to avoid overfitting but making it too higher will decrease network accuracy.

Public Access To Bucket; https://console.cloud.google.com/storage/browser?project=level-clone-237916