Project Proposal

CUNY 698 - MS Research Project

Walt Wells, 2018

Problem

A significant byproduct of Moore's law is how large scale computation is now considerably less expensive and more accessible to everyone. As a result, computationally expensive modeling techniques like Deep Learning that can accurately model highly dimensional data like images, video, and audio are becoming increasingly popular.

I propose to use the CUNY MS Research Project as an opportunity to learn the basics of this broad and important field. The project paper will focus on training and optimizing neural networks (ANN, CNN, DNNs) to make classification predictions over the well established ImageNet dataset, and provide methods to evaluate and improve performance.

In order to accomplish this task, I will need to gain acumen in Deep Learning models, Optimization and Regularization techniques for training, and hardware and distributed computing techniques for Deep Learning over larger datasets.

While I do not expect to make any new or novel contributions to the state of the art, by tackling this problem using a diverse set of resources and texts across multiple statistical programming languages, and working with a major dataset in the field, I expect to examine the problem from many different angles and fully cement my understanding. A semester devoted to studying Deep Learning will prove a powerful catalyst for my career and establish a strong foundation for a lifetime studying ML techniques.

Preparation

I am excited to undertake this course of study, but there's a lot to learn in order to accomplish the tasks set out in this proposal. At a very high level, I will need to do a broad survey of Deep Learning techniques in order to understand which tools are right for the job, how to optimize and train Deep Learning models, and how to judge model performance. In addition, I will need to review and implement resource techniques to manage and train over some of these larger datasets using modern microchips optimized for Deep Learning tasks.

To that end, I have prepared a syllabus that will keep me on track throughout the semester, and I hope will provide the Deep Learning primer necessary for this project and beyond. This includes texts and MOOCs including:

- Make Your own Neural Network; Rashid
- Deep Learning; Goodfellow, Bengio, Courville
- Machine Learning with R; Lantz
- Hands on Machine Learning with Scikit-Learn & TensorFlow; Geron
- Introduction to Deep Learning Using R; Beysolow
- Coursera - Geoffrey Hinton; Neural Networks for Machine Learning

The Syllabus (and my progress) can be found at: https://github.com/wwells/CUNY_DATA_698/blob/master/mySyllabus.xlsx?raw=true

Additional work will need to be done to find resources on GPUs or TPUs and other optimal resources to manage and train deep learning models.

Hypothesis

There will be differences in methods to train and optimize different deep learning models using different languages over different hardware. We will use the classification task and dataset from the 2012 ImageNet competition to benchmark and learn more about deep learning modeling.

"For each image, algorithms will produce a list of at most 5 object categories in the descending order of confidence. The quality of a labeling will be evaluated based on the label that best matches the ground truth label for the image. The idea is to allow an algorithm to identify multiple objects in an image and not be penalized if one of the objects identified was in fact present, but not included in the ground truth."

For more information visit: http://image-net.org/challenges/LSVRC/2012/index

Dataset

For this project I propose to use a major dataset for image classification and Deep Learning benchmarking, the ImageNet dataset. We will use the 2012 edition, and only concern ourselves with parts of the dataset that will help us with the classification task.

"The validation and test data will consist of 150,000 photographs, collected from flickr and other search engines, hand labeled with the presence or absence of 1000 object categories. The 1000 object categories contain both internal nodes and leaf nodes of ImageNet, but do not overlap with each other. A random subset of 50,000 of the images with labels will be released as validation data included in the development kit along with a list of the 1000 categories."

For more information visit: http://www.image-net.org/challenges/LSVRC/2012/nonpub-downloads

- Training images - 138GB
- Validation images - 6.3GB
- Test images - 13GB

For development / testing: Tiny Imagenet has 200 classes. Each class has 500 training images, 50 validation images, and 50 test images. ~.5GB uncompressed

https://tiny-imagenet.herokuapp.com/

Resources

To manage the data and models, I will leverage Google Compute Platform. They have a very generous "first time customer" incentive of $300 credit, which should go a long way towards running the clusters and training the models I'll need. I have experience with AWS, but am already very excited by the ease of use I've observed over the last few months of testing GCP.

I have setup snapshots for managing Jupyter Notebooks and RStudio Server in GCP. This includes options for running them in the head-node of a cluster to do distributed computation.

I expect to also use GCP object storage and DataProc clusters to manage and train over the medium/large sized dataset. In addition, GCP has recently made their TPUs available to the public for rental. This may be a great opportunity to learn how to use these chips or GPUS. https://techcrunch.com/2018/02/12/googles-custom-tpu-machine-learning-accelerators-are-now-available-in-beta/

Additionally, I expect to explore using deep learning libraries in Python (Caffe, Scikit-learn, Lasagne) R (Keras, darch, deepnet, caret), and explore back-end connections to hardware like Tensor-flow and Theano.

I will use Github to manage the work and have a repository already prepared: https://github.com/wwells/CUNY_DATA_698