BLUE WATERS

SUSTAINED PETASCALE COMPUTING

1/16/18

Blue Waters User Monthly Teleconference

Distributed TensorFlow

GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

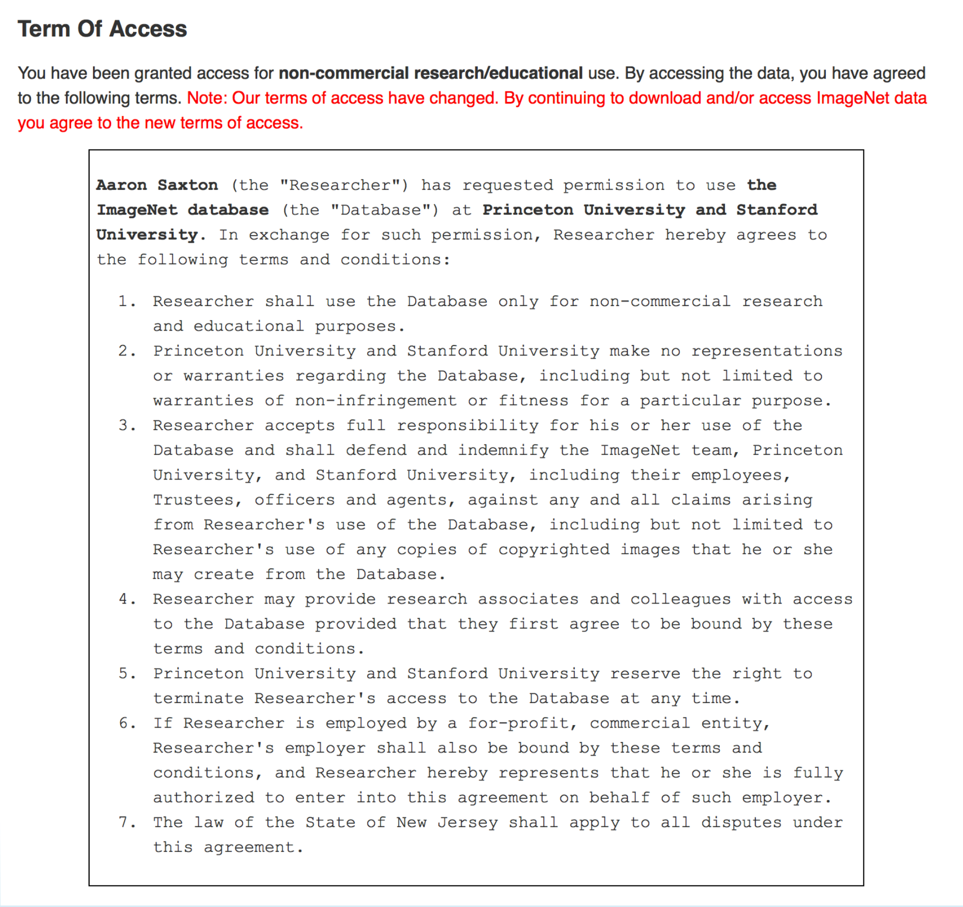
Quick Facts

- GitHub
  - https://github.com/asaxton/ncsa-bluewaters-tensorflow
- Febuary 7, 2018
  - Webinar about CNNs, TensorFlow, and BlueWaters
  - For details, keeps your eyes open at: https://bluewaters.ncsa.illinois.edu/webinars/data-analytics
- ImageNet (http://www.image-net.org)
  - Create user and login.
  - Accept Terms of Access
  - Email saxton@illinois.edu screenshot of Terms of Access

The Data:
ImageNet (http://www.image-net.org)

- Industry standard for quality images with bounding by and synset annotations
  - synset: https://wordnet.princeton.edu
- Started by Li Fei-Fei of the Stanford Vision Lab
- Total number of non-empty synsets: 21841
- Total number of images: 14,197,122
- Number of images with bounding box annotations: 1,034,908
- Free for noncommercial researchers
  - Go sign Terms of access
  - BW location: /sw/unsupported/mldata/ImageNet/

The Data: ImageNet (http://www.image-net.org)



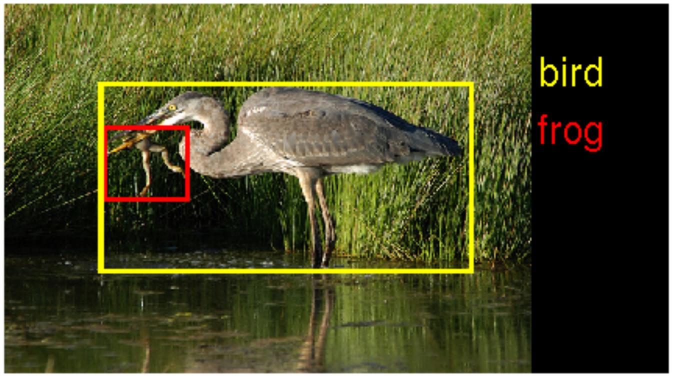
Term Of Access

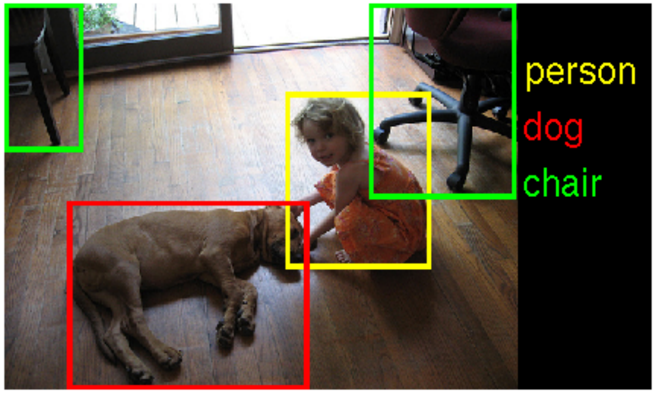
You have been granted access for **non-commercial research/educational** use. By accessing the data, you have agreed to the following terms. Note: Our terms of access have changed. By continuing to download and/or access ImageNet data you agree to the new terms of access.

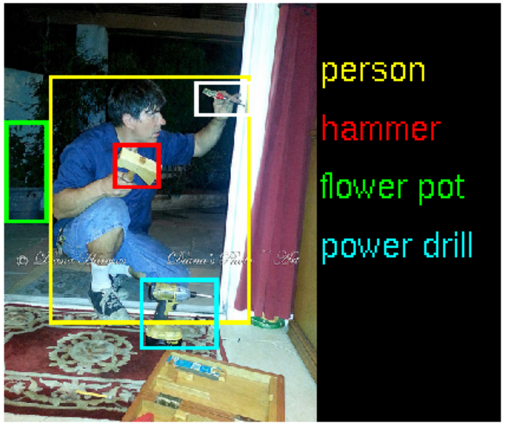
Aaron Saxton (the "Researcher") has requested permission to use **the ImageNet database** (the "Database") at **Princeton University and Stanford University**. In exchange for such permission, Researcher hereby agrees to the following terms and conditions:

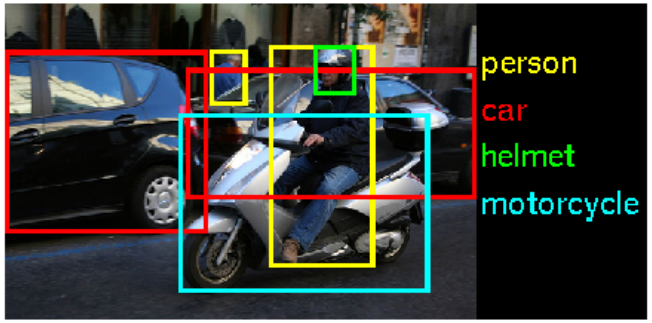
1. Researcher shall use the Database only for non-commercial research and educational purposes.
2. Princeton University and Stanford University make no representations or warranties regarding the Database, including but not limited to warranties of non-infringement or fitness for a particular purpose.
3. Researcher accepts full responsibility for his or her use of the Database and shall defend and indemnify the ImageNet team, Princeton University, and Stanford University, including their employees, Trustees, officers and agents, against any and all claims arising from Researcher's use of the Database, including but not limited to Researcher's use of any copies of copyrighted images that he or she may create from the Database.
4. Researcher may provide research associates and colleagues with access to the Database provided that they first agree to be bound by these terms and conditions.
5. Princeton University and Stanford University reserve the right to terminate Researcher's access to the Database at any time.
6. If Researcher is employed by a for-profit, commercial entity, Researcher's employer shall also be bound by these terms and conditions, and Researcher hereby represents that he or she is fully authorized to enter into this agreement on behalf of such employer.
7. The law of the State of New Jersey shall apply to all disputes under this agreement.

The Data:
ImageNet (http://www.image-net.org)









*http://image-net.org/challenges/LSVRC/2013/index

Tools We Are Providing

- https://github.com/asaxton/ncsa-bluewaters-tensorflow
- datasets/imagenet/extract_data_from_archive.pbs
  - Extracts images and annotations from local archive
- datasets/imagenet/build_imagenet_data.pbs
  - Creates tf_record from extracted image/annotation
- run_scripts/distributed_tf_launch.pbs
  - Launches a distributed TensorFlow cluster on BW

Metrics With Single Process Train

- Train with 3200 images: 8 min
  - Validation Precision @ 1 = 0.0009
- Train with 32000 images: 57 min
  - Validation Precision @ 1 = 0.0012
- Train with 96000 images: 3 hour
  - Validation Precision @ 1 = ?
- Expect decent percition when training with ~ 300k
  - Precision ~ 0.7 – 0.8
- This problem is well suited for ditributed Tensorflow

The Bad New, Then the Good

- Distributed TensorFlow as a problem saving the state of graph in distributed mode
- Team is actively working on solutions
  - TensorFlow has new MPI features
  - Directly fixing parameter server bug
- Cray has released a Machine Learning Plugin
  - The team has been testing and profiling. We're excited to make it available soon.

Thank You

Questions?

Code Tour?