Adapting Deep Learning to New Data Using ORNL’s Titan Supercomputer

Steven R. Young
Travis Johnston

Oak Ridge National Laboratory
Overview

• Deep Learning for Problems of National Interest
• Challenges
• Tools
• Next Steps
Deep Learning for National Interest Problems

**Commercial Interest**

**State of the Art Results**
- Object Recognition
- Face Recognition

**Characteristics**
- Data is *easy* to collect
- Inexpensive labels

**National Interest**

**Challenging New Domains**
- Material Science
- High Energy Physics
- Remote Sensing

**Characteristics**
- Data is *difficult* to collect
- Few labels available
Problem: Adaptability Challenge

**Premise:** For every data set, there exists a corresponding neural network that performs ideally with that data

- What’s the ideal neural network architecture (i.e., hyper-parameters) for a particular data set?

- Widely-used approach: intuition
  1. Pick some deep learning software (Caffe, Torch, Theano, etc)
  2. Design a set of parameters that defines your deep learning network
  3. Try it on your data
  4. If it doesn’t work as well as you want, go back to step 2 and try again.
The Challenge

Deep Learning Toolbox

- Learning Rate
- Batch Size
- Momentum
- Weight Decay

Output
- Fully Connected
- Pooling
- Convolutional
- Pooling
- Convolutional
- Pooling
- Convolutional

Input
The Challenge

Deep Learning Toolbox

- Output
- Pooling
- Convolutional
- Fully Connected
- Input

- Learning Rate
- Batch Size
- Momentum
- Weight Decay
Adapting DL to New Data

Current Approaches to Hyper-parameter Optimization

• Use out-of-the-box network
  – Why spend time trying to create your own network when there are already so many good ones available? Surely, one of those networks will also solve your problem.

• Tune an out-of-the-box network
  – Hyper-parameter sweeps
    Assumes independence of hyper-parameters
  – Grid search
    Requires training an exponential number of networks (infeasible)
  – Random search
    Significant improvement over grid search, but doesn’t make use of information learned during training.

What can we do with Titan?

18,688 GPUs
Two Approaches

- **RAvENNA: RApidly Evolving Neural Network Architecture**
  - Optimizes hyper-parameters of a *pre-existing* network.

- **MENNDL: Multi-node Evolutionary Neural Networks for Deep Learning**
  - Constructs neural networks from *scratch*.
  - Chooses number of layers, layer types, and layer hyper-parameters.
RAvENNA: Improved Random Search

- Bad Hyperparameters
- Good Hyperparameters
RAvENNA: Does smart searching help?

Random Search

Smart Search
RAvENNA: Current status, quick stats

• Implemented in **Apache-Spark** and **Caffe**.

• Running on Titan
  – Typical jobs 1,000-4,000 nodes (1 GPU / node)
  – Have run optimizations on up to 18,000 nodes

• Applied to several datasets/problems
  – Image segmentation (cloud detection in overhead imagery)
  – Model prediction (neutron scattering data)
  – Crystal lattice structure prediction
MENNDL: Multi-node Evolutionary Neural Networks for Deep Learning

• Evolutionary algorithm as a solution for searching hyper-parameter space for deep learning
  – Focus on Convolutional Neural Networks
  – Evolve only the topology with EA; typical SGD training process
  – Generally: Provide scalability and adaptability for many data sets and compute platforms

• Leverage more GPUs; ORNL’s Titan has 18k GPUs
  – Next generation, Summit, will have increased GPU capability

• Provide the ability to apply DL to new datasets quickly
  – Climate science, material science, physics, etc.
Designing the Genetic Code

- **Goal:** facilitate complete network definition exploration
- **Each population member is a network which has a genome with sets of genes**
  - Fixed width set of genes corresponds to a layer
    - Layers contain multiple distinct parameters
  - Restrict layer types based on section
    - Feature extraction and classification
    - Minor guided design in network, otherwise we attempt to fully encompass all layer types
**MENNDL: Communication**

- **Genetic Algorithm Master**
- **Gene: Population Network Parameters**
- **Fitness Metrics: Accuracy**

**Network 1 Parameters, Model**
- Predictions
- Performance Metrics

**Network 2 Parameters, Model**
- Predictions
- Performance Metrics

... (N workers, one per node)

**Network N Parameters, Model**
- Predictions
- Performance Metrics

---

15 Adapting DL to New Data
Hyper-parameter Values and Improved Performance

• Currently T&E of latest code that changes all possible parameters (e.g., # of layers, layer types, etc)
• Using just 4 nodes
• From 27% to 65% Accuracy
Hyper-parameter Values and Improved Performance

• Improved performance over known good network
• Using just 4 nodes
• From 75% to 82%
Unusual Layers (limited training examples)
MINERvA Detector Vertex Reconstruction

Goal: Classify which segment the vertex is located in.

<table>
<thead>
<tr>
<th>Segments</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Challenge: Events can have very different characteristics.

Table 1: Class distribution

<table>
<thead>
<tr>
<th>Target</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution</td>
<td>12.9%</td>
<td>13.8%</td>
<td>11.4%</td>
<td>8.4%</td>
<td>10.8%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Segment</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segment Distribution</td>
<td>2.4%</td>
<td>4.7%</td>
<td>4.8%</td>
<td>13.5%</td>
<td>1.2%</td>
</tr>
</tbody>
</table>
Application: 3D Electron Microscopy

- St. Jude Children’s Research Hospital is interested in developing tools which will aid biologists in labeling and analyzing new image volumes for the location, density, shape, and other characteristics of sub-cellular structures such as mitochondria.

- Segmentation of 3D electron microscopy (EM) imagery is an important initial characterization task as mitochondria are relatively distinct but occur in a variety of locations, shapes, and sizes.

- MENNDL evaluated nearly 900k convolutional networks on +18k of Titan’s nodes for 24 consecutive hours.

- Achieved a classification accuracy of 93.8%, representing a 30% reduction in error vs. a human expert defined network configuration.
MENNDL Current Status

• Scaled to 18,000 nodes of Titan
• 460,000 Networks evaluated in 24 hours
• Expanding to more complex topologies
• Evaluating on a wide range of science datasets
• Preparing for Summit (6 Volta GPUs per node, 4,600 nodes)
Acknowledgements

• Gabriel Perdue (FNAL) and Sohini Upadhyay (University of Chicago)
• Adam Terwilliger (Grand Valley State University) and David Isele (University of Pennsylvania)
• Robert Patton, Seung-Hwan Lim, Thomas Karnowski, and Derek Rose (ORNL)
• Devin White and David Hughes (ORNL)
Questions