Open Source Machine Learning Tools Overview

All Things Open
10-26-2016

Phillip Rhodes
Fogbeam Labs

https://github.com/fogbeam/ATO2016

Goals

- Overview of what the “cutting edge” projects in the field are
- An argument **against** solely focusing on the “cutting edge”
  - FDD – Fad Driven Development
  - Nothing in AI is every really out-dated. See: ANN's
- Don't forget about GOFAI – Good Old Fashioned AI
- Some speculation in regards to uniting the AI/ML fiefdoms

Latest Entrants

- IBM / Apache – SystemML - August 27, 2015
- Google – TensorFlow - November 9, 2015
- Microsoft – DMTK – November 12, 2015
- Baidu – WarpCTC - January 14, 2016
- Microsoft – CNTK – January 25, 2016
- Yahoo – CaffeOnSpark - Feb 24, 2016
- Amazon.com – DSSTNE (“Destiny”) - May 10, 2016
- Apache PredictionIO - Jul 22, 2016
- Facebook – FastText - August 18, 2016
- Baidu – PaddlePaddle – August 31, 2016

Apache SystemML

- General purpose distributed machine learning platform
- Written in Java, but exposes functionality in a dialect of R (DML), or a dialect of Python (PyDML)
- Heavily rooted in query optimizer technology ala RDBMS's
- Allows for automatic, seamless scalability from a single core to a thousand node cluster
- Especially handy for R programmers, since R doesn't scale terribly well by default

Apache SystemML

- Includes a lot of pre-built implementations of popular ML algorithms out of the box
- Runs on top of Spark or Hadoop (Map/Reduce)
- Spark MLContext supports programming in Scala, Java or Python
- Lacks native GPU support

Google TensorFlow

- Billed as a library for “deep learning” but is much more general than that
- Really a numerical computing library
- Based on data-flow graphs (similar to Spark)
- Written in C++, primary API interface is via Python
- Wrappers can be implemented using SWIG and there are some out there already
- TF Board is a handy debugging tool for introspecting TF graphs

Google TensorFlow

- TF Learn is a simpler, friendlier API
- TensorFlow Serving – for “productionizing” TF models
- Seamless CPU/GPU support
- Supports distributed operation on compute clusters
- More Neural Network focused, at least in terms of docs and examples
- Includes many optimization algorithms out of the box
- *contrib* package includes other packaged algorithm implementations
- HDFS support

Microsoft - DMTK

- Framework for distributed computation, focusing on machine learning
- Written in C++
- Uses MPI or 0MQ for cluster communication
- Native Windows support, but also supports Linux
- Seems to cater heavily to a couple of specific algorithms.
  - LightLDA, an extremely fast and scalable topic model algorithm
  - a distributed version of (multi-sense) word embedding
- But general purpose, you can implement your own algorithms

Baidu Warp-CTC

- “A fast parallel implementation of CTC, on both CPU and GPU” (Warp-CTC README)
- “What is ~~Aleppo~~, er, CTC?”
- Connectionist Temporal Classification
  - A specific “objective function” that works well for training RNN's (Recurrent Neural Networks) for “sequence labeling” tasks.
  - Specifically, things like handwriting recognition, speech recognition, gesture recognition, etc.
- Differentiable function, so works with standard Gradient Descent and the like

Microsoft - CNTK

- “a unified deep-learning toolkit that describes neural networks as a series of computational steps via a directed graph” (CNTK README)
- Makes it easy to realize NN models including feed-forward DNNs, convolutional nets (CNNs), and recurrent networks (RNNs/LSTMs)
  - But provides a plug-in architecture allowing users to define their own computation nodes
- Includes stochastic gradient descent learning with automatic differentiation and parallelization across multiple GPUs and servers
- Custom networks are described in CNTK's custom network description language "BrainScript"
- Use models from C++ and C#

Yahoo - CaffeOnSpark

- A Spark package for deep learning
- Combines features from Caffe with Apache Spark and Hadoop
- Enables distributed deep learning on a cluster of GPU and CPU servers
- Scala API
- Tight Hadoop (HDFS) integration
- Incremental learning is supported to leverage previously trained models
  - This has the potential to be a big deal

Amazon – DSSTNE (“Destiny”)

- An open source software library for training and deploying recommendation models with sparse inputs, fully connected hidden layers, and sparse outputs
- Used at Amazon to generate personalized product recommendations
- Designed for production deployment of real-world applications which need to emphasize speed and scale over experimental flexibility
- Data must be in NetCDF format
- Definitions for the Neural Networks fed into DSSTNE are represented in a custom JSON format

Apache - PredictionIO

- An open source Machine Learning Server
- Sits on top of other ML engines and provides services
  - quickly build and deploy an engine
  - evaluate and tune multiple engine variants systematically
  - speed up machine learning modeling with systematic processes and pre-built evaluation measures
  - respond to dynamic queries in real-time
- support machine learning and data processing libraries such as Spark MLLib and OpenNLP
- unify data from multiple platforms
- implement your own machine learning models and seamlessly incorporate them into your engine

Facebook - FastText

- A library for efficient learning of word representations and sentence classification
- Builds on Mac OSX and Linux; requires a modern C++ 11 compile
- Represents sentences with bag of words or bag of n-grams
- Faster to train and test than a deep neural network
  - FastText is exclusively dedicated to text classification. This allows it to be quickly trained on extremely large datasets
- Uses a hierarchical classifier instead of a flat structure
- Besides text classification, FastText can also be used to learn vector representations of words

Baidu - PaddlePaddle

- “PArallel Distributed Deep LEarning is an easy-to-use, efficient, flexible and scalable deep learning platform”
- Neural-network / deep-learning focused
- Written in C++
- C++ and Python API
- Includes many optimization algorithms out-of-the-box
- Has built-in clustering code, but docs suggest using MPI or other cluster software for more robust operation
- Has GPU support using CUDA libraries
- Requires significantly less code than on other popular deep learning platforms?

But wait, there's more...

Apache SAMOA

- Scalable Advanced Massive Online Analysis
- Specifically oriented towards streaming scenarios
- Runs on top of Storm, S4, Flink, or Samza
- “Provides a collection of distributed streaming algorithms for the most common data mining and machine learning tasks such as classification, clustering, and regression”
- Also provides the primitives you need to implement your own algorithms

Apache Singa

- Yet another distributed deep learning framework
- Similar to TensorFlow in that Tensors (multi-dimensional arrays) are the primary data abstraction
- GPU support using CUDA or OpenCL
- Provides optimization algorithms and abstractions designed for implementing neural networks
- Has Python and C++ APIs

Caffe

- Deep learning framework by the Berkeley Vision and Learning Center
- Somewhat targeted towards computer vision applications, at least in terms of the docs, examples, etc.
- Written in C++
- C++, Python and Matlab API's
  - As of August 2015, the Matlab support requires “real” Matlab, and doesn't support Octave
- Claims to be one of the fastest DL frameworks out there

Keras

- “Keras is a high-level neural networks library, written in Python and capable of running on top of either TensorFlow or Theano”
- Developed with a focus on enabling fast experimentation
- Supports both convolutional networks and recurrent networks, as well as combinations of the two
- Runs seamlessly on CPU and GPU
- Python API
- Supports arbitrary connectivity schemes
- The core data structure of Keras is a **model**, a way to organize layers

Sci-kit Learn

- General purpose machine learning library written in Python
- Built on NumPy, SciPy, and matplotlib
- Provides many “out of the box” algorithms for:
  - Clustering
  - Classification
  - Dimensionality reduction
  - Model selection
  - Pre-processing
- Not natively a distributed / cluster-aware framework
- But the API does support “out of core” processing using a streaming model and incremental training
- No GPU support

Theano

- “Theano is a Python library that allows you to define, optimize, and evaluate mathematical expressions involving multi-dimensional arrays efficiently”
- Integrated with Numpy
- Native GPU support
- No native multi-node / cluster support
- Automatic compilation to C or C++ for performance optimization
- Has a reputation for being very fast

Torch

- “A scientific computing framework with wide support for machine learning algorithms that puts GPUs first.”
- Based on Lua/LuaJIT
- Features:
  - a powerful N-dimensional array
  - lots of routines for indexing, slicing, transposing, etc.
  - linear algebra routines
- Easy to use, fast interface to C code
- Embeddable, with ports to iOS, and Android, as well as custom FPGA backends

A Lot More!

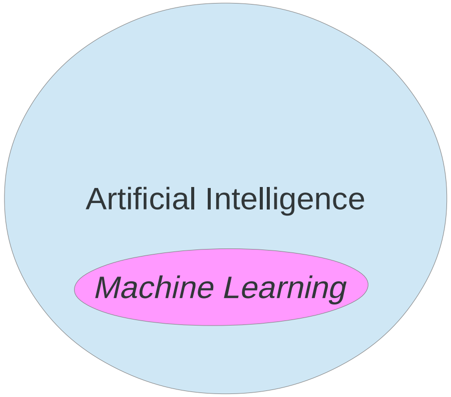
- Aerosolve
- Lasagne
- DL4J
- MLLib
- Mahout
- Weka
- MXNet
- OpenNLP
- CoreNLP
- OpenCV
- Yahoo Yamall
- Veles
- Leaf
- … see http://mloss.org

Fad Driven Development

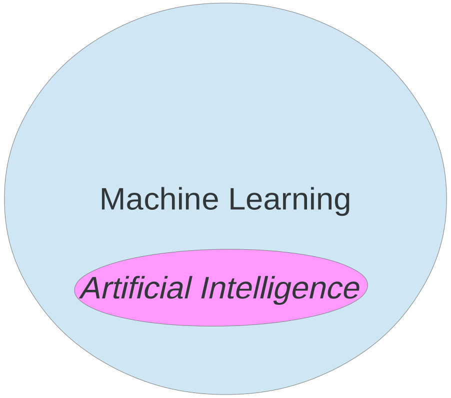
- Things in our industry tend to come in and out of fashion in cycles
  - Neural Networks may be THE canonical example of this
  - Expert Systems
  - Genetic Algorithms
  - Logic Programming
  - Most of what falls under “GOFAI”
- Use what works, not what's trendy

AI vs. ML

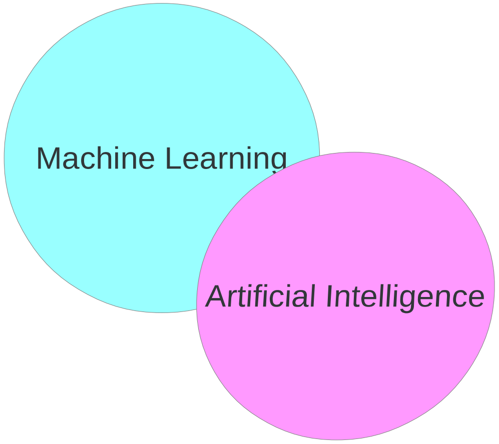
- ML is a subset of AI?



- ML is a subset of AI?



- Or maybe it's more like this?



AI vs. ML

- In either case, the point is to not “throw the baby out with the bathwater” and forget all of GOFAI just because we don't have AGI yet

GOFAI

- OpenCog
- NuPIC
- OpenCyc
- ACT-R
- CLIPS
- Racer
- LOOM
- Constraint Logic Programming
- Answer Set Programming
- Etc.

Genetic Algorithms

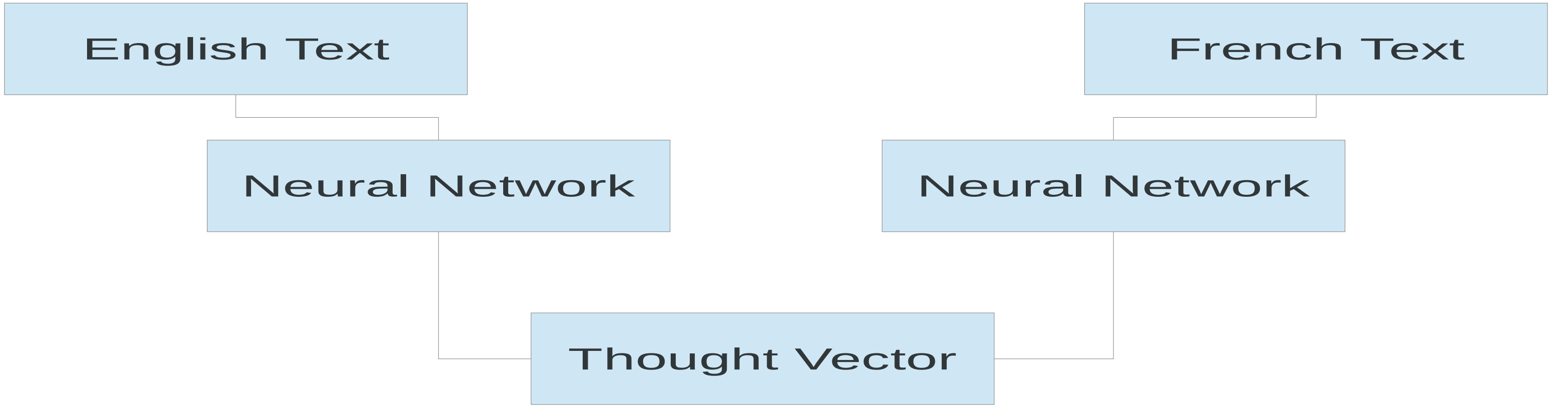
- Jenetics
- JGAP
- Watchmaker
- MOEA
- JAGA
- ECJ
- JENES 2.0

Rule Induction

- Charade
- PROGOL
- RuleX
- CN2

Thought Vectors

- Somewhat analogous to a “word vector” which is a vector of associations between one word and a group of other words
- A “thought vector” then is a “thought” and a vector of associations to other thoughts
- Language independent and heavily used in Machine Translation



Thought Vectors

- “Thoughts” are linked by a chain of reasoning, similar to how words are linked by grammar
- Common representation of a “thought”
- Possibly a route to unifying disparate approaches to AI
- Share thought vectors between different processing sub-systems or “minds”

Multiple Minds

- Blackboard Architecture
- Tuple Spaces
- Multi-Agent Systems
- Pandemonium Architecture
- Competitive Learning
- ...