New Developments in Large Data that have *Immediate* Application in Industry

(but you haven’t heard of yet)

Joseph Turian

@turian

MetaOptimize

#strataconf
perhaps you should close your laptops
How do you get a competitive advantage with data?
How do you get a competitive advantage with data?

- More data
How do you get a competitive advantage with data?

• More data

• Better algorithms
When big data gives diminishing returns, you need better algorithms
When big data gives diminishing returns, you need better algorithms

@turian #strataconf
When should you use better algorithms?
When should you use better algorithms?

- If they are really cool algorithms
When should you use better algorithms?

• If they are really cool algorithms
When should you use better algorithms?

• If they are really cool algorithms
• If you have a lot of time on your hands
When should you use better algorithms?

• If they are really cool algorithms
• If you have a lot of time on your hands
Only use better algorithms if they will *qualitatively* improve your product
Only use better algorithms if they will *qualitatively* improve your product

@turian #strataconf
Who am I?
Who am I?

• Engineer with 20 years coding experience

• Ph.D. 10 yrs exp in large-scale ML + NLP
What is MetaOptimize?
What is MetaOptimize?

optimizing the process of
What is MetaOptimize?

optimizing the process of optimizing the process of
What is MetaOptimize?

optimizing the process of optimizing the process of optimizing the process of optimizing the process of optimizing the process of optimizing the process of optimizing the process of optimizing the process of optimizing the process of optimizing the process of optimizing the process of optimizing the process of optimizing the process of optimizing the process of optimizing the process of optimizing the process.
What is MetaOptimize?

• Consultancy on:

• Large scale ML + NLP

• Well-engineered solutions
“Both NLP and ML have a lot of folk wisdom about what works and what doesn't. [This site] is crucial for sharing this collective knowledge.” - @aria42

http://metaoptimize.com/qa/
Outline

- Deep Learning
  - Semantic Hashing
- Graph parallelism
- Unsupervised semantic parsing
Outline

• Deep Learning
  – Semantic Hashing
• Graph parallelism
• Unsupervised semantic parsing
Opportunity with Deep Learning

• Machine learning that’s
  – Large-scale (>1B examples)
  – Can use all sorts of data
  – General purpose
  – Highly accurate
Deep Learning
Deep Learning

• Artificial intelligence???
Natural Intelligence
Natural Intelligence
Works!
Artificial Intelligence
Artificial Intelligence

- Still far from the goal!
- Why?
Where does intelligence come from?
Intelligence comes from knowledge
How can a machine get knowledge?

Human input
Intelligence comes from knowledge.
Knowledge comes from learning.
Intelligence comes from knowledge.
Knowledge comes from learning.

@turian #strataconf
Statistical Learning

- New multi-disciplinary field
- Numerous applications
Memorize?  or  Generalize?

- Easy for machines
- Harder for humans

- Mathematically: fundamentally difficult
- Easier for humans
How do we build a learning machine?
Deep learning architecture

Output: is bob?

Highest-level features:
- Faces

Abstract features:
- Shapes

Primitive features:
- Edges

Input: Raw pixels
Shallow learning architecture
Why deep architectures?
“Deep” computer program
subroutine1 includes subsub1 code and subsub2 code and subsubsub1 code

subroutine2 includes subsub2 code and subsub3 code and subsubsub3 code and ...

main

“Shallow” computer program
“Deep” circuit
“Shallow” circuit

input

output

$2^n$
Insufficient Depth

Insufficient depth =
May require exponential-size architecture

Sufficient depth =
Compact representation

\[ \text{O}(n) \]
What’s wrong with a fat architecture?
Overfitting!

bad generalization
Occam’s Razor

ALWAYS CHOOSE THE SIMPLEST EXPLANATION.
Other motivations for deep architectures?
Learning Brains

- $10^{11}$ neurons, $10^{14}$ synapses
- Complex neural network
- Learning: modify synapses
Visual System
Deep Architecture in the Brain

Area V1
- Edge detectors
- Primitive shape detectors
- Higher level visual abstractions

Area V2
- Primitive shape detectors

Area V4
- Higher level visual abstractions

Retina
- Pixels
Deep architectures are Awesome!!!

• Because they’re compact

but...
Why not deep architectures?

- How do we train them?
Before 2006

Failure of deep architectures
Mid 2006

Breakthrough!
Signal-to-noise ratio

- More signal!
Deep training tricks

- Unsupervised learning
Deep training tricks

• Create one layer of features at a time
(I did my postdoc here)
Deep learning a success!

Since 2006

Deep learning breaks records in:

- Handwritten character recognition
- Component of winning NetFlix entry
- Language modeling

Interest in deep learning:

- 300 AI researchers attend workshop
- NSF and DARPA
Opportunity with Deep Learning

• Machine learning that’s
  - Large-scale (>1B examples)
  - Can use all sorts of data
  - General purpose
  - Highly accurate
Outline

- Deep Learning
  - Semantic Hashing
- Graph parallelism
- Unsupervised semantic parsing
Opportunity with Semantic Hashing

- Fast semantic search
What’s wrong with keyword search?
Keyword search

• Search for tweets on “Hadoop”
Keyword search

- Search for tweets on “Hadoop”
- Misses the following tweets:
  - “Just started using HBase”
  - “I really like Amazon Elastic Map-Reduce”
What’s wrong with keyword search?
What’s wrong with keyword search?

Misses relevant results!
Standard search: Inverted Index

- my
- awesome
- fantastic
- blog
- post
- entry

My awesome blog post
My fantastic blog entry
Hashing

• Another technique for search
Hashing

• FAST!

Address Space

Similar Documents

Hashing Function

Document
Hashing

• **Compact!**

• Without hashing:
  - Billions of images => 40 TB

• With 64-bit hashing:
  - Billions of images => 8GB
“Dumb” hashing

• Typically no learning, not data-driven
• Examples:
  - Random Projections
  - Count-Min Sketch
  - Bloom filters
  - Locality Sensitive Hashing
“Smart” Hashing

• As fast as “dumb” hashing
• Data-driven
• Examples:
  – Semantic Hashing (2007)
  – Kulis (2009)
  – Kumar, Wang, Chang (2010)
  – Etc.
Semantic Hashing
Semantic Hashing

= Smart hashing + deep learning

Salakhutdinov + Hinton (2007)
Semantic Hashing architecture

2000

\[ W_1 + \epsilon_6 \]

500

\[ W_2 + \epsilon_5 \]

500

\[ W_3 + \epsilon_4 \]

32

Code Layer

Gaussian Noise

\[ W_3^T + \epsilon_3 \]

500

\[ W_2^T + \epsilon_2 \]

500

\[ W_1^T + \epsilon_1 \]

2000

Fine-tuning
Semantic Hashing architecture

- LSA/LSI, LDA
- TF*IDF

The diagram shows a flow of data through various layers, starting with 2000 inputs, reducing to 500, then 500, and finally 32 for the code layer.

Gaussian Noise is indicated at the point where the data is reduced from 500 to 32.
Opportunity with Semantic Hashing

Semantic search that is:

- General purpose
- Fast
- Compact
Opportunity with Semantic Hashing

Semantic search that is:

• **General purpose**
  - Search text, images, videos, audio, etc.
• Fast
• Compact
Opportunity with Semantic Hashing

Semantic search that is:

- General purpose
- **Fast**
  - Indexing: few weeks for 1B docs, using 100 cores
  - Retrieval: 3.6 ms for 1 million docs, scales sublinearly
- Compact
Opportunity with Semantic Hashing

Semantic search that is:

• General purpose
• Fast
• Compact

- 1B docs, 30-bit hashes => 4GB
- 1B images, 64-bit hashes => 8GB (vs. 40 TB naïve)
Prediction

Smart hashing will revolutionize search
Prediction

Smart hashing will revolutionize search

@turian #strataconf
Outline

• Deep Learning
  – Semantic Hashing
• Graph parallelism
• Unsupervised semantic parsing
The rise of Graph stores

- Neo4J, HyperGraphDB, InfiniteGraph, InfoGrid, AllegroGraph, sones, DEX, FlockDB, OrientDB, VertexDB
Opportunity with graph-based parallelism

- Scale sophisticated ML algorithms
- Larger data sets
- Higher accuracy
Useful machine learning algorithms

- Gibbs sampling
- Matrix factorization
- EM
- Lasso
- Etc.

Have graph-like data dependencies
Machine learning in Map-Reduce
Machine learning in Map-Reduce
Machine learning in Map-Reduce

Map-Abuse

-Carlos Guestrin
There are too many graph-like dependencies in many ML algorithms
Parallel abstractions for graph operations

- Pregel (Malewicz et al, 2009, 2010)
  - Erlang implementation called Phoebus
- GraphLab (Low et al, 2010)
  - Source code available
When should you prefer GraphLab over MapReduce? And vice-versa?

GraphLab is an abstraction for parallel programming. (Low et al, 2010) It is higher level than MPI and Pthreads, but lower level and more expressive than MapReduce.

In what circumstances is MapReduce preferable? In what circumstances is GraphLab preferable?

http://metaoptimize.com/qa/

MapReduce is good for single-iteration and embarrassingly parallel distributed tasks like feature processing, while GraphLab is good for iterative algorithms with computational dependencies or complex asynchronous schedules. For instance, we find that GraphLab is highly suitable for Gibbs sampling, EM-style algorithms and some classes of optimization algorithms. Programs which fit well on a systolic abstraction (such as PageRank on Pregel) will also work well with GraphLab. There are probably a lot more algorithms that will fit well in the GraphLab and we are still exploring the capabilities and implications of the abstraction (and whether further extensions will be needed).

We are in the process of releasing our (shared memory) multi-core implementation of GraphLab and are in the process of designing a (distributed-memory) cluster implementation. Currently we do not provide fault tolerance so if fault tolerance is a critical requirement, we recommend the use of Hadoop/MapReduce. The GraphLab abstraction does not preclude fault tolerance though due to the computational / state dependencies, distributed fault tolerance is slightly more difficult and is currently ongoing research.

The current implementation is a C++ API which can easily work with other tools like Hadoop and SQL (we use some of it ourselves in our cluster implementation)
The current implementation is a C++ API which can easily work with other tools like Hadoop and SQL (we use some of these for logging already) and does not require any additional language support. It is important to keep in mind that while we provide a reference implementation, we are describing an abstraction that could be integrated into other software tools like Hadoop.

We have posted the initial public beta release of the GraphLab API at http://www.graphlab.ml.cmu.edu and welcome any comments or suggestions. The GraphLab webpage provides tutorials and detailed descriptions of the GraphLab API features. We are in the processing of migrating to a Google code project to give contributors access to the API and the code as it evolves.

edited Aug 02 '10 at 12:13

Joseph Gonzalez
answered Jul 03 '10 at 16:22

1 I don't see fault tolerance to an issue with small scale clusters. Where it really becomes important is with >100 nodes participating in a computation.

Is Graphlab susceptible to an abstraction layer like FlumeJava? That was substantially ease the introduction of a new computational paradigm.

Ted Dunning (Jul 03 '10 at 16:45)

What are some algorithms in which GraphLab cannot be parallelized? In particular, can it parallelize matrix-multiplication?

Joseph Turian ** (Jul 03 '10 at 19:34)
Opportunity with graph-based parallelism

- Scale sophisticated ML algorithms
- Larger data sets
- Higher accuracy
Prediction

Map-Reduce for simple algorithms, graph parallelism for sophisticated ML
Prediction

Map-Reduce for simple algorithms, graph parallelism for sophisticated ML

@turian #strataconf
Outline

• Deep Learning
  – Semantic Hashing
• Graph parallelism
• Unsupervised semantic parsing
Opportunity with Semantic Parsing

• Simply reads texts and understands them
• Applicable to general domains
• Applications
  – Question Answering (cf. Wolfram Alpha)
  – Natural language search (cf. Powerset)
  – Spam generation / Spam detection
  – Knowledge Extraction from Wikipedia, web, etc.
Question-Answer: Example

Q: What does IL-2 control?
A: ???
Interestingly, the DEX-mediated IkappaBalpha induction was completely inhibited by IL-2, but not IL-4, in Th1 cells, while the reverse profile was seen in Th2 cells.
Q: What does IL-2 control?
A: The DEX-mediated IkappaBalpha induction

Sentence:
Interestingly, the DEX-mediated IkappaBalpha induction was completely inhibited by IL-2, but not IL-4, in Th1 cells, while the reverse profile was seen in Th2 cells.
**Challenge:**

**Same Meaning, Many Variations**

**IL-4 induces CD11b**

**Protein IL-4 enhances the expression of CD11b**

**CD11b expression is induced by IL-4 protein**

**The cytokin interleukin-4 induces CD11b expression**

**IL-4’s up-regulation of CD11b, ...**

......
Semantic Parsing

Microsoft buys Powerset.
Semantic Parsing

*Microsoft buys Powerset.*

BUYS (MICROSOFT, POWERSET)
Where does intelligence come from?

Knowledge!
Extracting Knowledge From Text

\[
\text{INDUCE}(e_1) \land \text{IL-4}(e_2) \land \text{CD11B}(e_3) \\
\land \text{INDUCER}(e_1, e_2) \land \text{INDUCED}(e_1, e_3)
\]

\[
\begin{array}{c}
\text{REGULATE} \\
\quad \text{regulate, control, govern, modulate}
\end{array}
\]

\[
\begin{array}{c}
\text{induce, enhance, trigger, augment, up-regulate} \\
\text{inhibit, block, suppress, prevent, abolish, abrogate, down-regulate} \\
\text{activate, activate}
\end{array}
\]

\[
\begin{array}{c}
\text{INDUCE} \\
\text{INHIBIT}
\end{array}
\]
Ontology

REGULATE

regulate, control, govern, modulate

ISA

induce, enhance, trigger, augment, up-regulate

ISA

inhibit, block, suppress, prevent, abolish, abrogate, down-regulate

ISA

activate

ACTIVATE

INDUCE

INHIBIT
How do we extract knowledge from text?
How do we extract knowledge from text?

Hire a handful of Ph.D. linguists to write a grammar
How do we extract knowledge from text?
Manual approach

Costly!
Ineffective!
Inflexible!
Manual approach

• Challenge: Same meaning can be expressed in many different ways

  - Microsoft buys Powerset
  - Microsoft acquires semantic search engine Powerset
  - Powerset is acquired by Microsoft Corporation
  - The Redmond software giant buys Powerset
  - Microsoft’s purchase of Powerset, ...

  ...

• Manual encoding of variations?
Manual approach

• **Challenge:** Domain specific
  – Grammar for newspaper articles !=
  – Grammar for biomed articles !=
  – Grammar for tweets
  – Etc.
Knowledge extraction that is:

• Large-scale,
• Open-domain,
• Automatic,
• End-to-end
Interestingly, the DEX-mediated IkappaBalpha induction was completely inhibited by IL-2, but not IL-4, in Th1 cells, while the reverse profile was seen in Th2 cells.

Q: What does IL-2 regulate?

A: The DEX-mediated IkappaBalpha induction
Ontology Learning

• Step 1: Induction
• Step 2: Population
• Limitations in existing approaches
  – Require heuristic patterns or existing KBs
  – Pursue each task in isolation
Unsupervised Semantic Parsing with Ontologies

Jointly conducts:
• Ontology induction,
• Ontology population,
• and knowledge extraction

Why is this so cool??

Poon + Domingos (2009, 2010)
Intuition

• Cluster syntactic or lexical variations of the same meaning

\[
\text{BUYS} (-, -, -) = \{ \text{buys, acquires, 's purchase of, ...} \}
\]

= Cluster of various expressions for acquisition

\[
\text{MICROSOFT} = \{ \text{Microsoft, the Redmond software giant, ...} \}
\]

= Cluster of various mentions of Microsoft
Microsoft buys Powerset
Microsoft acquires semantic search engine Powerset
Powerset is acquired by Microsoft Corporation
The Redmond software giant buys Powerset
Microsoft’s purchase of Powerset, ...
Microsoft buys Powerset

Microsoft acquires semantic search engine Powerset

Powerset is acquired by Microsoft Corporation

The Redmond software giant buys Powerset

Microsoft’s purchase of Powerset, ...
Microsoft buys Powerset

Microsoft acquires semantic search engine Powerset

Powerset is acquired by Microsoft Corporation

The Redmond software giant buys Powerset

Microsoft's purchase of Powerset, …
Microsoft buys Powerset
Microsoft acquires semantic search engine Powerset
Powerset is acquired by Microsoft Corporation
The Redmond software giant buys Powerset
Microsoft’s purchase of Powerset...
Microsoft buys Powerset

Microsoft acquires semantic search engine Powerset

Powerset is acquired by Microsoft Corporation

The Redmond software giant buys Powerset

Microsoft’s purchase of Powerset, …
Clusters And Compositions

• Clusters in core forms
  \{ investigate, examine, evaluate, analyze, study, assay \}
  \{ diminish, reduce, decrease, attenuate \}
  \{ synthesis, production, secretion, release \}
  \{ dramatically, substantially, significantly \}

  ...

• Compositions
  amino acid, t cell, immune response, transcription factor, initiation site, binding site ...
Experiments

- Evaluate on Question answering
- Evaluation: Number of answers and accuracy
- GENIA dataset: 1999 Pubmed abstracts
- 2000 questions e.g.:
  - *What does anti-STAT1 inhibit?*
  - *What regulates MIP-1 alpha?*
Total vs. Correct Answers

- KW-SYN
- TextRunner
- RESOLVER
- DIRT
- USP
- OntoUSP
Total vs. Correct Answers

Five times as many correct answers as TextRunner

Highest accuracy of 91%
Opportunity with Semantic Parsing

• Simply reads texts and understands them
• Applicable to general domains
• Applications
  – Question Answering (cf. Wolfram Alpha)
  – Natural language search (cf. Powerset)
  – Spam generation / Spam detection
  – Knowledge Extraction from Wikipedia, web, etc.
Prediction

Automated knowledge extraction will become widespread
Prediction

Automated knowledge extraction will become widespread

@turian #strataconf
Outline

- Deep Learning
  - Semantic Hashing
- Graph parallelism
- Unsupervised semantic parsing
Take-home points
When big data gives diminishing returns, you need better algorithms

@turian #strataconf
Only use better algorithms if they will qualitatively improve your product

@turian strataconf
Intelligence comes from knowledge.
Knowledge comes from learning.

@turian #strataconf
Prediction

Smart hashing will revolutionize search

@turian #strataconf
Prediction

Map-Reduce for simple algorithms,
graph parallelism for sophisticated ML

@turian #strataconf
Prediction

Automated knowledge extraction will become widespread

@turian #strataconf
Thanks to the following people for letting me adapt their slides

- Hoifung Poon
- Yoshua Bengio
- Geoff Hinton
- Ruslan Salakhutdinov
Questions?
Joseph Turian
@turian
MetaOptimize
http://metaoptimize.com/qa/
Why not deep architectures?

- How do we train them?
Supervised Training Example

Input $X$

Output $f(X)$

Target $Y$

six

two!
Gradient descent

Input X → Output f(X) → Target Y

six = ?

two!
Gradient descent

\[ f(X) \text{ six} = \text{Target } Y = \text{two!} \]

Input X
Problem on deep architectures
Before 2006

Failure of deep architectures
Mid 2006
Breakthrough!
(I did my postdoc here)
Signal-to-noise ratio

- More signal!
Deep training tricks

• Unsupervised learning
Deep training tricks

• Create one layer of features at a time
Deep training
Deep training

features

input
Deep training

reconstruction of input
features
input

\[ \text{input} = \text{reconstruction of input} \]
Deep training

features

input
Deep training

More abstract features
features
input
Deep training

reconstruction of features

More abstract features

features

input
Deep training

More abstract features

features

input
Deep training

Even more abstract features

More abstract features

features

input
Deep training

Output: $f(X)$

Target: $Y$

six = two!

Even more abstract features

More abstract features

features

input