Streaming Data Explanation with MacroBase

Kai Sheng Tai

in collaboration with

Peter Bailis, Edward Gan, Kexin Rong, Sahaana Suri, Firas Abuzaid, Jialin Ding, Vatsal Sharan, Greg Valiant

Stanford DAWN Project
DAWN Project: Making ML More Accessible

Pls: Peter Bailis, Kunle Olukotun, Chris Ré, Matei Zaharia
dawn.cs.stanford.edu

Data Acquisition: Snorkel, Babble, Labble, Coral
Feature Engineering: DeepDive
Model Training: MacroBase (Streaming Data)
Productionizing: ModelQA

Data Fusion: MacroBase (Streaming Data)
          : NoScope (Video)
          : YellowFin (DL)
          : *Headed, Mulligan (SQL+graph+ML)
          : AutoRec, SimDex (Recommendation)

Compilers: Weld, Spatial, Sparser, Delite
Hardware: Plasticine CGRA, FuzzyBit

Hardware: Plasticine CGRA, FuzzyBit
CPU
GPU
FPGA
...
DAWN Project: Making ML More Accessible

PIs: Peter Bailis, Kunle Olukotun, Chris Ré, Matei Zaharia
dawn.cs.stanford.edu

Data Acquisition
- Snorkel, Babble Labble, Coral
- DeepDive

Feature Engineering

Model Training
- AutoML
- ModelQA

Productionizing

Data Fusion

AutoRec, SimDex (Recommendation)

MacroBase (Streaming Data)
- *Headed, Mulligan (SQL+graph+ML)

Compilers: Weld, Spatial, Sparser, Delite

Hardware: Plasticine CGRA, FuzzyBit

Interfaces

Algorithms

Systems

Hardware:
- Plasticine CGRA
- FuzzyBit

CPU

GPU

FPGA

Mobile
Continued Growth of Streaming Data Volumes

- Telemetry from mobile devices
  - >2B smartphones worldwide
- Application logs from web services
- Visual features from video streams
  - 1000s of dashcams, security cameras

**MacroBase**: prioritizing *human attention* via *feature selection*
MacroBase: Example Use Case

Input: stream of logs from mobile app (based on a real application)

<table>
<thead>
<tr>
<th>Errors</th>
<th>Non-Errors</th>
<th>Explain error class to analyst with [location = Canada]</th>
</tr>
</thead>
<tbody>
<tr>
<td>{iPhone7, USA}</td>
<td>{iPhone8, USA}</td>
<td></td>
</tr>
<tr>
<td>{iPhone7, Canada}</td>
<td>{iPhone7, USA}</td>
<td></td>
</tr>
<tr>
<td>{iPhone8, Canada}</td>
<td>{iPhoneX, USA}</td>
<td></td>
</tr>
<tr>
<td>{iPhone7, USA}</td>
<td>{iPhone7, USA}</td>
<td></td>
</tr>
<tr>
<td>{iPhone8, Canada}</td>
<td>{iPhone7, USA}</td>
<td></td>
</tr>
<tr>
<td>{iPhone7, USA}</td>
<td>{iPhone8, USA}</td>
<td></td>
</tr>
<tr>
<td>{iPhone7, USA}</td>
<td>{iPhone7, USA}</td>
<td></td>
</tr>
</tbody>
</table>

Challenges

• **Throughput:**
  streams with millions of events/sec

• **Resource constraints:**
  limited computation and memory

• **Dimensionality:**
  high-order feature combinations (# phone models) x (# locations) x …
MacroBase Stream Analytics

**CLASSIFY**
- Identify data in tails

**EXPLAIN**
- Find disproportionately correlated attributes

**TRANSFORM**
- Extract domain-specific signals

- Other projects:
  - Kernel density estimation
  - Dimensionality reduction
  - Faster CNN queries on video
  - Method-of-moments for quantile estimation
  - Time series visualization

- In production at:
  - Major web service provider
  - Mobile app company
  - Video streaming service

Papers and links:
- In production at:
  - Major web service provider
  - Mobile app company
  - Video streaming service

macrobse.stanford.edu
MacroBase Stream Analytics

TRANSFORM

identify data in tails

extract domain-specific signals

CLASSIFY

find disproportionately correlated attributes

EXPLAIN

Outliers
{iPhone6, Canada}
{iPhone6, USA}
{iPhone5, Canada}

Inliers
{iPhone6, USA}
{iPhone6, USA}
{iPhone5, USA}

In production at:
- major web service provider
- mobile app company
- video streaming service

Other projects:
- Kernel density estimation
- Dimensionality reduction
- Faster CNN queries on video
- Method-of-moments for quantile estimation
- Time series visualization

This talk:
Online feature selection on streams

macrobase.stanford.edu
MacroBase: Streaming Feature Selection

Setup: online learning of a linear classifier (e.g. logistic regression)
Goal: return top-$k$ most discriminative features to the user
Track most frequent features? Not necessarily the most discriminative
Sparsity-inducing regularization? Hard to tune a priori to satisfy memory constraints

Weight-Median Sketch [Tai, Sharan, Bailis, Valiant. arXiv 1711.02305]
Maintain a compressed version (a sketch) of a linear classifier…
• ... that supports fast updates
• ... that supports queries for estimates of each weight
• ... with $(\epsilon, \delta)$-approximation guarantee vs. uncompressed classifier

Track (approximation of) $k$ most heavily-weighted features
Sketched Linear Classifiers

• Sketch of $x$: random projection of $x$ to low dimension

$$(x_t, y_t) \rightarrow \nabla \hat{L}_t \rightarrow \text{sketched classifier} \rightarrow \text{estimates of largest weights}$$

location = Canada 2.5
model = iPhoneX -1.9
version = 2.1.1 1.8

streaming data
gradient estimates
sketched classifier
estimates of largest weights
Accurate weight recovery in practice

Online logistic regression on Reuters RCV1 with 4KB memory budget

- feature hashing
- hard thresholding
- frequent features
- WM-Sketch (our method)

(lower is better)
Sketched Linear Classifiers

- Sketch of $x$: random projection of $x$ to low dimension

$$\begin{array}{ccc}
(x_t, y_t) & \nabla \hat{L}_t & \text{location} = \text{Canada} \quad 2.5 \\
\text{query} & \text{model} = \text{iPhoneX} & -1.9 \\
\text{streaming data} & \text{version} = 2.1.1 \quad 1.8 \\
\text{gradient estimates} & \text{sketched classifier} & \text{estimates of largest weights}
\end{array}$$

Takeaways

- **Count-Sketch** data structure can be adapted to streaming feature selection
- Essentially **feature hashing** with *highest-magnitude features in heap*
- Need only space **logarithmic** in original dimension
DAWN Stack

Data Acquisition
- Snorkel, Babble, Labble, Coral
- DeepDive

Feature Engineering
- Data Fusion
- YellowFin (DL)
- NoScope (Video)
- *Headed, Mulligan (SQL+graph+ML)
- AutoRec, SimDex (Recommendation)

Model Training
- AutoML

Productionizing
- ModelQA

Hardware
- Compilers: Weld, Spatial, Sparser, Delite
- Hardware: Plasticine CGRA, FuzzyBit

Systems
- Spark Cluster
- Mobile
- CPU
- GPU
- FPGA
DAWN Stack

Data Acquisition
- Snorkel, Babble Labble, Coral
- DeepDive

Feature Engineering
- MacroBase (Streaming Data)
- Data Fusion

Model Training
- AutoML
- YellowFin (DL)
- AutoRec, SimDex (Recommendation)

Productionizing
- ModelQA
- NoScope (Video)

Hardware
- Hardware: Plasticine CGRA, FuzzyBit

Compilers: Weld, Spatial, Sparser, Delite

Find out more @ dawn.cs.stanford.edu/blog
Recap

MacroBase: making sense of the firehose

This talk: Online feature selection by sketching linear classifiers

Check out other DAWN projects: hardware + systems + ML

macrobase.stanford.edu

dawn.cs.stanford.edu

Kai Sheng Tai / kst@cs.stanford.edu