KeystoneML: Optimizing Pipelines for Large-Scale Advanced Analytics

Vincent Li, Tina Lu, Vincent Viego
Machine Learning

Knowledge
Profit
Happiness
Smart Factory

Sensors + Equipment

ML Pipeline

Feature Extraction

Dimensionality Reduction

Data Transformation

Model

OK

FAIL
Domain Specific Feature Extractors

General Purpose Numeric Optimizers

Errors

Change in Data Volume

Atlas
Wouldn’t it be great if my ML pipeline could:
1.
2.
3.

KeystoneML

Pipeline 0

Pipeine *
Existing Solutions are not...

scalable and end-to-end

Vowpal Wabbit and scikit-learn

systemML
Databases!

optimizable logical operators

What about sparsity?
Desired accuracy?
Computation vs communication?
Why is it hard?

How many iterations?

\[ x_1 \quad x_2 \quad x_3 \]

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>52</td>
<td>9</td>
</tr>
<tr>
<td>65</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>42</td>
<td>15</td>
<td>32</td>
</tr>
<tr>
<td>2</td>
<td>57</td>
<td>65</td>
</tr>
</tbody>
</table>

\[ X \xrightarrow{\text{ML Operator}} f(X) + \epsilon \]
Why is it hard?
Goals

Logical Operators

Dynamic Scaling

Automatic Optimization
Logical Operators

Estimator → Transformer

Optimizable → CostModel → Iterative
What other logical operators could KeystoneML be extended to support?
Relational DB Cost Optimizer

SELECT *
FROM ...
\[ c(\text{PO, Data, Cluster}) = (K_{\text{node}}) c_{\text{node}}(\text{PO, Data, N}) + (K_{\text{network}}) c_{\text{network}}(\text{PO, Data, N}) \]
How might KeystoneML benefit from a different implementation of cost models (e.g., OtterTune)?
Physcial Operator Optimization
End-to-end Optimization

Knapsack: Memory Budget

Costs: Estimated Runtimes

Goal: Min Total Runtime

Greedy Approximation
What are other methods for end-to-end optimization?
Evaluation

- Speech
- Image
- Text

Task
- Memory
- Nodes
- Features

keystoneML
- Cache Management
- Apache Spark

Runtime
- Accuracy
TextClassifier

Trim

andThen

LowerCase

andThen

Tokenizer

Tokenize

andThen

NGramsFeaturizer

Feature Generation

andThen...

LinearSolver

Estimator

Pipeline

TextClassifier
Data Scalability

![Bar chart showing time in seconds for different feature sizes and systems: KeystoneML, Vowpal Wabbit, SystemML.](image)
Scalability

Optimization
Unequal Performance Comparisons

VOC Image Classification

- 12.4x speedup
- 16x cores
- 15.25x RAM

ImageNet Image Classification

- 21x speedup
- 50x cores
Type-safety?

High Level ML Operators

ML Pipeline Specification

Training

1010010
0000001
1000001
End-to-end Optimization Alternatives

Maximize
\[ P = p_1 x_1 + p_2 x_2 + \cdots + p_k x_k \]
Subject to:
\[ a_{11} x_1 + a_{12} x_2 + \cdots + a_{1k} x_k \leq q_1 \]
\[ a_{21} x_1 + a_{22} x_2 + \cdots + a_{2k} x_k \leq q_2 \]
\[ \vdots \]
\[ a_{n1} x_1 + a_{n2} x_2 + \cdots + a_{nk} x_k \leq q_n \]
\[ x_1, x_2, \ldots, x_k \geq 0 \]

<table>
<thead>
<tr>
<th>(F(i, k))</th>
<th>(k=0)</th>
<th>(k=1)</th>
<th>(k=2)</th>
<th>(k=3)</th>
<th>(k=4)</th>
<th>(k=5)</th>
<th>(k=6)</th>
<th>(k=7)</th>
<th>(k=8)</th>
<th>(k=9)</th>
<th>(k=10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i=0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(i=1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>(i=2)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>50</td>
<td>70</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>(i=3)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>(i=4)</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
</tr>
</tbody>
</table>
backpropagation vs communication at large scale?