CAEML Research in Hardware Design and Optimization Using Machine Learning

Chris Cheng, Distinguished Technologist
Yongjin Choi, Master Technologist
Sumon Dey, Specialist Machine Learning Engineer
Content

– Introduction to CAEML
– Unique types of machine learning models for IoT and hardware system management
– Example small learning: proactive hardware failure prediction
– Example medium learning: 56G PAM SerDes performance optimization
– Example deep learning: Dynamic resources demand forecast
Vision

The center’s goal is to enable fast, accurate design and verification of microelectronic circuits and systems by creating machine learning algorithms to derive models used for electronic design automation.

By speeding up the design and verification of microelectronic circuits and systems, CAEML will reduce development cost and time-to-market for manufacturers of microelectronic products, and will enable the development of optimized products, e.g. for low-power, high-reliability or security.
IoT management model

Data Flow

Control Flow

IEC IoT white paper
**Edge inference**

**Edge tier**
- Node processor with inference and online training engine
- High frequency activity data
- Local inference
- Dynamic performance tuning, abnormal detection and frequent data caching

**Platform tier**
- Slow speed sensors and clusters signatures
- Global Inference Models, feature list
- Data Flow

**Enterprise tier**
- Big Data Collection
- Analytics
  - Feature selection
  - Classifier training
- Control Flow
## 3 Types of IOT and hardware management models

<table>
<thead>
<tr>
<th>Feature size (number of variables or complexity)</th>
<th>Prediction throughput</th>
<th>Prediction interval</th>
<th>Machine learning engine</th>
<th>Examples</th>
</tr>
</thead>
</table>
| **Big data**
**Small learning** | Less than 100 | A few thousands to millions of predictions per sec | Variable from days to ms | Ensemble classifier | Hardware failure prediction
**Software failure prediction**
**Automatic application detection**
**Storage security applications (ransomware detection)**
**System abnormally detection** |
| **Big data**
**Medium learning** | Between 100 to 500 | A few hundred predictions per secs | A few secs | Generative performance surrogate model with Bayesian learning | Dynamic system performance optimization |
| **Big data**
**Deep learning** | 100s to 1000s | 1000’s of predictions in an hours | 5-15 mins | Deep Markovian Models
**Deep learning neural networks** | High dimensional time series for resource demand prediction |
Small learning example: proactive hardware failure detection
Proactive hardware failure prediction

Trade off between detection rate and false positive

Recommended MATLAB package: Statistics and Machine Learning Tool Box
Causal inference for feature selection

- Causal inference is used to pick the sensor signals and sample window
- 30% features reduction
- 15% accuracy improvement

Predict $Y_{T_n}$ from $(X_k)_{T_n-t_{adv}}^{T_n-t_{adv}-w_k}$ for each feature $X_k$

Prediction Windows

CAEML research
Performance limitation of GPU on ensemble classifier

<table>
<thead>
<tr>
<th>Number of samples</th>
<th>Intel CPU</th>
<th>CUDA-Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>10k</td>
<td>1</td>
<td>2.2</td>
</tr>
<tr>
<td>20k</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>50k</td>
<td>1</td>
<td>1.8</td>
</tr>
</tbody>
</table>

Higher is slower execution time
GPU is slower than Intel CPU!

Control Flow Problem in GPUs/SIMD

- GPU uses SIMD pipeline to save area on control logic.
  - Group scalar threads into warps
- Branch divergence occurs when threads inside warps branch to different execution paths.

Wilson Peng et al. UBC
Medium learning example:
Surrogate models for system performance optimization
Discriminative system model

- Collect n samples of system configurations and observed performance
  - Neural network
  - Linear/non-linear regression
- In the future, given a configuration X, model predict performance X
Generative system model

- Collect n samples of system configuration and performance
- Construct surrogate model and use Bayesian learning to construct additional configurations and performance measurement
- Keep iteration until target uncertainty is hit
- Given a new target performance, what is the optimal configuration to achieve, usually tie in with a cost function and optimize for the lowest cost
Principal component analysis and surrogate models

25 controlling taps in the 56G PAM SerDes is mapped into 4 principal component vectors that can cover 92% of solutions

Surrogate models based on 5000+ measured samples

Zhu et al, DesignCon 2019

Recommended MATLAB package: Statistics and Machine Learning Toolbox
Accelerated 56G PAM channel optimization using PCA

Best fitness function versus generation.

(a) GA

(b) GA-PCA

PCA based GA algorithm converged with fewer generations, more quickly than the general GA algorithm.

Recommended MATLAB Package: Global Optimization Toolbox
Deep learning example: High dimension resource demand forecast
High dimension resources demand forecast

Many system resources show clusters of users and daily pattern

Gajowniczek et al., energies 2017
Time domain demand forecast

Recommended MATLAB package: Statistics and Machine Learning Toolbox

Gajowniczek et al., energies 2017

Hewlett Packard Enterprise
Thank you