Machine Learning Systems: On-Device AI and Beyond

Nicholas D. Lane

TECoSA Seminar Series
Thursday December 3rd 2020

@niclane7 | http://niclane.org
ndl32@cam.ac.uk

Efficient ML Revolution
Efficient ML Revolution

- User Privacy
- Strongest algorithms for the hardest problems
- Robust devices without a network dependency
- Low energy and latency
From Late ‘14: Efficient ML Revolution

2014
1st Proof-of-Concept DL of Audio/IMU on Mobile [HotMobile ‘15]

2015
DeepEar (1st DSP-based DL General Audio Sensing) [UbiComp ‘15]

2016
DL Smartwatch Activity Recognition [WristSense ‘16]
1st time: VGG executing directly on a commodity SmartWatch
1st time: Smartphone-scale DL on embedded processors (e.g., M0/M3) [SenSys ‘16]

2017
1st time: Multiple DL Vision Models on Wearable [MobiSys ‘17]

Community Innovations

Algorithmic & Architecture Advances
- Node Pruning, Leverage Sparsity
- SqueezeNet (50x AlexNet reduction)
- Low Precision (8/4 bit), Binarization
- MobileNet, MCDNN, Custom Nets

Hardware Innovations
- Diannao and ConvLutin2
- Front-ends e.g., SNPE - Qualcomm
- TPU, FPGAs / Hybrids
- Analog from Digital Approaches
- Spiking HW & Approx. Compute
Model Compression
Example: Node Pruning

Related Methodologies
- Leveraging Sparsity
- Low Precision Results (8-bit etc.)
- Binarization of Networks
- MobileNet, MCDNN, Custom Nets
  .. and even hardware approaches

~50x gains
tiny accuracy loss

Song Han, Jeff Pool, John Tran, William J. Dally, "Learning both Weights and Connections for Efficient Neural Networks", NIPS 2015
ML Efficiency is a fundamental crisis
Can we keep up with the upcoming leaps in ML complexity?

New Deep Learning Innovations

Can we keep up with the upcoming leaps in ML complexity?

Efficiency Gap

New Deep Learning Innovations

ML Efficiency Results
What will an Efficient ML device need look like in 2030?

(1) Rich Powerful ML Tasks
   From Classification to Open-world Weakly-supervised Reasoning

(2) 100s of ML Models per device

(3) On-Device Learning is routine
Fundamental **Efficient ML** Challenges

#1: Think (i.e., learn) Different

#2: Automated Specialization

#3: Sharing and Cooperation
Fundamental Efficient ML Challenges

#1: Think (i.e., learn) Different

#2: Automated Specialization

#3: Sharing and Cooperation

Rethinking the complete stack (and the learning algorithms)
#1 Think (i.e., learn) Different
#1 Think (i.e., learn) Different

**AMBITION:** Invent new learning algorithms with greater flexibility and diversity that fully utilize compute resources.
Inflating deep models from functions at inference: a new form of trading memory for compute

[ICAI ‘18, SysML ‘18, submitted FPL ‘20]
Inflating deep models from functions at inference: a new form of trading memory for compute

\[ f(x,y,z) \approx \begin{bmatrix} 5 & -1 & 9 \\ -6 & 3 & -2 \\ 2 & 8 & -7 \end{bmatrix} \]

[IJCAI '18, SysML '18, submitted FPL '20]

Inference is radically different...

\[ f(x,y,z) \]

\[ e.g., \text{OVSF orthogonal binary basis} \]
Inference is radically different..

### Top1 Accuracy (%)

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-18</td>
<td>91.15</td>
</tr>
<tr>
<td>ResNet-18 (OVSF)</td>
<td>91.02</td>
</tr>
<tr>
<td>ResNet-34</td>
<td>92.46</td>
</tr>
<tr>
<td>ResNet-34 (OVSF)</td>
<td>92.32</td>
</tr>
<tr>
<td>SqueezeNet</td>
<td>91.16</td>
</tr>
<tr>
<td>SqueezeNet (OVSF)</td>
<td>91.33</td>
</tr>
</tbody>
</table>

Dataset: CIFAR-10

Maximizing efficiency potential through hardware

unzipFPGA

---

[IJCAI '18, SysML '18, submitted FPL '20]
Maximizing efficiency potential through hardware

![Graph showing speedup/baseline](image)

<table>
<thead>
<tr>
<th>Compression (ResNet34)</th>
<th>Params (millions)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>21.8</td>
<td>73.3</td>
</tr>
<tr>
<td>Tay82</td>
<td>17.4</td>
<td>72.7</td>
</tr>
<tr>
<td>OVSF50</td>
<td>17.2</td>
<td>72.1</td>
</tr>
</tbody>
</table>

Dataset: ImageNet

Fundamental Efficient ML Challenges

#1: Think (i.e., learn) Different

#2: Automated Specialization

#3: Sharing Memory and Compute

[IJCAI '18, SysML '18, submitted FPL '20]
#2 Automated Specialization

Design Tools
#2 Automated Specialization

**Design Tools**

**AMBITIOn:** Automated synthesis of ML models specialized for a target chip/platform and task that exceeds expert hand-design

Joint Optimization of Hardware Operation and Neural Architecture

- **conventional approach**
  - DNN Design → Hardware Operation

- **joint optimization**
  - Co-Design of DNN and Hardware Operation

[interspeech '19, interspeech '20, eecv '20, dac '20, neurips '20]
Joint Optimization of Hardware Operation and Neural Architecture

**Dataset: Librispeech**

<table>
<thead>
<tr>
<th></th>
<th>WER (test-clean)</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoCHA</td>
<td>8.32</td>
<td>1x</td>
</tr>
<tr>
<td>AutoCAML</td>
<td><strong>8.19</strong></td>
<td>5.1x</td>
</tr>
</tbody>
</table>

**Dataset: Set5**

<table>
<thead>
<tr>
<th></th>
<th>LPIPS</th>
<th>Mult-Adds (G)</th>
<th>Params (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESRGAN</td>
<td>0.074</td>
<td>1034.1</td>
<td>16,697</td>
</tr>
<tr>
<td>FEQE</td>
<td>0.091</td>
<td>5.6</td>
<td>96</td>
</tr>
<tr>
<td>AutoCAML</td>
<td><strong>0.076</strong></td>
<td>3.6</td>
<td>61</td>
</tr>
</tbody>
</table>

[ INTERSPEECH '19, INTERSPEECH '20, EECV '20, DAC '20, NeurIPS '20 ]
Joint Optimization of Hardware Operation and Neural Architecture

vs

conventional approach

joint optimization

[ INTERSPEECH '19, INTERSPEECH '20, EECV '20, DAC '20, NeurIPS '20 ]

Joint Optimization of Hardware Operation and Neural Architecture

vs

conventional approach

joint optimization

[ INTERSPEECH '19, INTERSPEECH '20, EECV '20, DAC '20, NeurIPS '20 ]
Joint Optimization of Hardware Operation and Neural Architecture

AutoCAML

Controller
- CNN Search Space
- Accelerator Design Space

Evaluator
- Accuracy
- Latency
- Area
- Power

Propose (CNN, Accelerator)
Multiobjective Reward

CHaidDNN

[ INTERSPEECH ’19, INTERSPEECH ’20, EECV ’20, DAC ’20, NeurIPS ’20 ]
Joint Optimization of Hardware Operation and Neural Architecture

1. **AutoCAML**

2. **Controller**
   - CNN Search Space
   - Accelerator Design Space

3. **Evaluator**
   - Accuracy
   - Latency
   - Area
   - Power

4. \( \mathcal{E}(s) = \mathcal{R}(-\text{area}(s), -\text{lat}(s), \text{acc}(s)) \)

[ INTERSPEECH ’19, INTERSPEECH ’20, EECV ’20, DAC ’20, NeurIPS ’20 ]

Joint Optimization of Hardware Operation and Neural Architecture

[ INTERSPEECH ’19, INTERSPEECH ’20, EECV ’20, DAC ’20, NeurIPS ’20 ]
Joint Optimization of Hardware Operation and Neural Architecture

<table>
<thead>
<tr>
<th>prior SOA</th>
<th>AutoCAML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>92.8%</td>
</tr>
<tr>
<td>Latency</td>
<td>51ms</td>
</tr>
<tr>
<td>HW Area</td>
<td>170mm$^2$</td>
</tr>
</tbody>
</table>

Zynq Ultrascale+

Fundamental Efficient ML Challenges

#1: Think (i.e., learn) Different

#2: Automated Specialization

#3: Sharing and Cooperation
#3 Sharing and Cooperation
Federating ML Models for Mobile and Embedded Devices

**Dataset: FashionMNIST**

<table>
<thead>
<tr>
<th>Training time (in mins)</th>
<th>E=1</th>
<th>E=5</th>
<th>E=10</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU-Only</td>
<td>0</td>
<td>0</td>
<td>2000</td>
</tr>
<tr>
<td>GPU and CPU</td>
<td>100</td>
<td>500</td>
<td>1000</td>
</tr>
</tbody>
</table>

3.5x

3. Sharing and Cooperation

**AMBITION:** Maximize resource utilization through intra- and inter-device sharing and cooperation of learning algorithms
Fundamental Efficient ML Challenges

#1: Think (i.e., learn) Different

#2: Automated Specialization

#3: Sharing and Cooperation

#4: Next Steps in Hardware
Predictions for the Efficient ML Revolution

#1 On-Device ML goes far beyond *just* classification
#2 SOTA Accuracy will come from efficient ML Models
#3 Data Center is *replaced* by devices as the heart of ML

#1 Efficient ML Prediction

On-Device ML goes far beyond classification
Reasoning  
Common Sense 
Understanding  
Perception (Discriminative)  

Cognitive Embedded Stack

Peace Dividends of Efficiency
- Faster exploration & experimentation
- “Intractable approaches” become possible...
- Consuming more data – both labeled and unlabeled varieties
- Larger and larger architectures
- Able to heavily exploit automated methods like architecture search
- Brand new ML tasks

#2 Efficient ML Prediction

SOTA Accuracy will come from Efficient ML
#3 Efficient ML Prediction

Devices replace Data Centers as core of ML

---

Select publications (and submissions) referenced in this talk……

- "Flower: A Friendly Federated Learning Framework" – available on Arxiv
- "BRP-NFS: Prediction-based NAS using GCN" – NeurIPS 2020
- "unzipFPGA: Enhancing FPGA-based CNN Engines with On-the-Fly Weight Generation" – submitted FPL 2020
- "Best of Both Worlds: AutoML Co-design of a CNN and its Hardware Accelerator" – DAC 2020
- "Journey Towards Tiny Perceptual Super-Resolution" – IEEE ICME 2020
- "Iterative Compression of End-to-End ASR Model using Reinforcement Learning" – ICASSP 2020
- "UnshrinkML: End-to-End ASR Model Compression Using Reinforcement Learning" – INTERSPEECH 2019
- "An Empirical study of Binary Neural Networks' Optimisation" – ICLR 2019
- "MicMic: using cycle-consistent generative adversarial networks to overcome microphone variability in speech systems" – IPSN 2019
- "The deep (learning) transformation of mobile and embedded computing" – IEEE Computer Magazine, 51 (5), 2018
- "BinaryCond: Keyword Spotting with Deterministic Binary Basis" – SysML 2018
- "Deterministic Binary Masks for convolutional neural networks" – IJCAI 2018
- "DeepEye: Resource Efficient Local Execution of Multiple Deep Vision Models using Wearable Commodity Hardware" – MobiSys 2017
- "Low-residing multi-task audio sensing for mobile and embedded devices via shared deep neural network representations" – ICASSP / UbiComp 2017
- "From Smart to Deep: Robust Activity Recognition on Smartwatches using Deep Learning" – IEEE Pervasive 2016
- "DeepEar: robust smartphone audio sensing in unconstrained acoustic environments using deep learning" – UbiComp 2015
- "Can Deep Learning Revolutionize Mobile Sensing?" – HotMobile 2015

---

Nicholas D. Lane

@niclane7
ndl32@cam.ac.uk