Efficient Deep Learning is the Key To Privacy (and Security)

Zhen Dong, Tianren Gao, Ravi Krishna, Suresh Krishna, Ani Nrusimha
Sheng Shen, Zhewei Yao, Bohan Zhai,
Amir Gholami, Shanghang Zhang,
Joey Gonzalez, Kurt Keutzer, Michael Mahoney,
with Forrest Iandola (self), Albert Shaw (Tesla), Bichen Wu (FB), Flora Xue (DeepMind)
Artificial intelligence: definition is always evolving
- Machine learning: well defined
- Deep Learning is a relatively small subset of Machine Learning approaches
State-of-the-art solutions for all these problems (and more) rely on deep learning.
State-of-the-Art Solutions Typically Rely on one DNN (or a few)

- **Image Classification**
- **Object Detection**
- **Image Segmentation**
- **Convolutional NN**
- **Audio Enhancement**
- **Call-center Sentiment Analysis**
- **Speech Recognition**
- **Recurrent NN**
- **Video Sentiment Analysis**
- **Music Recommendation**
- **Ad Recommendation**
- **DLRM**
- **Translation**
- **Question answering**
- **Document Understanding**
- **Transformer**
Losing privacy and security in our modern world

• Gaining convenience, retaining privacy, in our personal world
  – Home
  – Car
  – Office
  – Personal assistant

• Local processing is the key to personal privacy
  – Leveraging federated data while retaining personal privacy
  – Efficiency is the key to local processing
  – Computer vision
  – Audio and Speech
  – NLU
  – Recommendations

• Challenges for the future
We Are Losing Our Privacy Nearly Everywhere:
Most Public Spaces

- Outdoor Surveillance
- Drones
- Automatic License Plate Reader (ALPR)
- Retail stores
- Gym
No Privacy in our Back Yard
May 2020: US Removes Restrictions on Commercial Satellite Resolution

2020
25 cm resolution

2021
10 cm resolution

photo: https://www.albedo.space
Outline

• Losing privacy and security in our modern world

Gaining convenience, retaining privacy, in our personal world
  – Home
  – Car
  – Office
  – Personal assistant

• Local processing is the key to personal privacy
  – Leveraging federated data while retaining personal privacy
  – Efficiency is the key to local processing
  – Computer vision
  – Audio and Speech
  – NLU
  – Recommendations

• Challenges for the future
There is a fundamental tension between the convenient features that we would like to have in our private spaces and preserving our privacy.

Conveniences:
- Voice commands
- Intelligent vision
- Natural language understanding
- Personal recommendations

Privacy in our:
- Speech (pattern and intonation)
- Conversations
- Personal visual spaces
- Personal preferences
In the Home
Convenience vs privacy

Conveniences:
• Simple voice commands
• Interactive commands (How many eggs in this recipe?)
• Visual analysis of our living space (where are my glasses?)
• Entertainment recommendations

With privacy:
• No eavesdropping on conversations
• No “peeping toms”
• Personal entertainment preferences stay private
In the Car
Convenience vs privacy

Conveniences:
• Local voice commands (Play music or select a radio station)
• Control basic car functions: roll down a window; open the trunk
• Ask for directions or navigation tips
• Find a gas station or restaurant
• Drowsy?

With privacy:
• No eavesdropping on conversations
• Car is a private space
• Location and destination private
• Don’t report driver status or driving errors not reported
Conveniences:
• A personal digital assistant may be the dashboard of every capability we have described so far.

With privacy:
• Because the PA will be with us everywhere, *all* of the prior privacy concerns are only amplified
• Our PA may know us better than any other human.

“If I get one more productivity improving time saving device my productivity will go to 0.”
• Kurt Keutzer
Just What Could Go Wrong?

Privacy Failure

- Deep Fakes
- Surveilled in our Home
- Located without Permission
- Preferences Exploited
Outline

• Losing privacy and security in our modern world
• What conveniences from Deep Learning applications do we want?
  – Home
  – Car
  – Personal assistant

How do we get these, but retain privacy (and security)?
  – Privacy vs security
  – Leveraging federated data while retaining personal privacy
• Privacy at the Edge: Efficiency is the key to local processing
  – Computer vision
  – Audio and Speech
  – NLU
  – Recommendations
• Summary
First: Privacy vs Security

• Security:
  – Data only accessed by authorized agents (but could include FB or Amazon)

• Privacy:
  – Allowing the user to completely determine who (if anyone) has access to the data

• User privacy has become a mainstream concern with the General Data Protection Regulation in Europe and the California Consumer Privacy Act
GDPR – Relevance to Privacy

• Privacy programs:
  – GDPR: General Data Protection Regulation
  – California Consumer Privacy Act

• Users control access to data before its collected
  – Who will be given the data?
  – What the data will be used for?
  – How long the data will be stored?

• Users must be assured that data is deleted at their request

• Individuals and corporations interests are aligning
Approaches to Providing Conveniences and Privacy

- Cloud hosted data, training, and inference
- Federated Learning
- Differential privacy
- Full /partial Training as well as Full Inference at Edge
Cloud hosted applications: data, inference, and training

- Flow:
  - Nominal case is that the user data (e.g. speech, photo) is sent to cloud
  - Application (e.g. automatic speech recognition, image classification) is run in cloud
  - Result is sent back to user
- Users data may be used for future training
- Problem: Users data may not be secure, certainly no longer private to user
Federated Learning

• Protects: users private data
• Approach
  – Local training is performed on the local computer/phone
  – Only local updates (gradients) sent to server
  – New global model periodically trained
  – Global models returned to user
• Problem: Some user information may be leaked through the gradients
• E.g. Movie viewing behavior might be inferred based non-zero gradients

This workshop:
“Attack Resistant Federated Learning with Residual-Based Reweighting” Song, Fu, Xie, Li, and Chen

• Protects: any information about the user, with high probability
  – Like federated learning (local training, gradients passed up), but …
  – Adds noise during local training to obfuscate what data was used
    • Fundamental trade-off between information leakage and accuracy
    • This approach gives mathematical guarantees on user-data loss

\[
\Pr(Y \text{ in Unknown Data}) \approx \Pr(Y \text{ not in Unknown Data})
\]
Full/Partial Edge Training and All Inference at the Edge

• Protects: privacy of all sensitive user data
  – No server communication
  – Requires local compute capability
  – Requires local or mobile DNN efficiency
  – May require capture of local data for personalization

• Premise of this talk: if you really want privacy you need *inference* at the edge and then your choice of edge-training, federated-learning plus/minus differential privacy
Outline

• Losing privacy and security in our modern world
• What conveniences from Deep Learning applications do we want?
  – Home
  – Car
  – Personal assistant
• How do we get these, but retain privacy (and security)?
  – Privacy vs security
  – Leveraging federated data while retaining personal privacy
Privacy at the Edge: Efficiency is the key to local processing
  – Computer vision
  – Audio and Speech
  – NLU
  – Recommendations
• Summary
Challenges Moving to the Edge

TPU Pod
125,000 TFLOPS

~10,000 x >>>>

Edge Client
5 – 15 TOPS

Samsung S21 Ultra
We Want to Operate Across a Broad Range of Hosts at the Edge

Typical handset
- 32g, 13cc, 5.5Wh = 19.8 kJ

Typical usage
- 5kJ active + 12kJ standby = 1 battery charge

Per Ljung – Nokia, 2012

**Average Power Dissipation**

<table>
<thead>
<tr>
<th>Device</th>
<th>Power Dissipation</th>
<th>Battery Life</th>
</tr>
</thead>
<tbody>
<tr>
<td>iWatch Series 3</td>
<td>0.72Wh, 0.12 W</td>
<td>30m/day</td>
</tr>
<tr>
<td>OZO Digital Pedometer</td>
<td>80μW, 1Wh</td>
<td>1 year</td>
</tr>
<tr>
<td>2017 iPhone 8</td>
<td>6.96Wh, 1.54 W</td>
<td>14h</td>
</tr>
<tr>
<td>iPad Pro</td>
<td>41Wh, 1.07W</td>
<td>1.5h</td>
</tr>
<tr>
<td>Eee PC 1000HE</td>
<td>49Wh, 1.476W</td>
<td>9.5h</td>
</tr>
<tr>
<td>Kindle Oasis</td>
<td>0.91Wh, 0.312W</td>
<td>7.5h</td>
</tr>
<tr>
<td>Asus 9.5h</td>
<td>5.2 W</td>
<td>9.5h</td>
</tr>
<tr>
<td>Apple 10h use</td>
<td>4.1 W</td>
<td>10h</td>
</tr>
<tr>
<td>15 inch Macbook Pro</td>
<td>76Wh, 273.6kJ</td>
<td>10h</td>
</tr>
<tr>
<td>13 inch Macbook Air</td>
<td>54Wh, 194.4kJ</td>
<td>12h</td>
</tr>
<tr>
<td>13 inch MacBook Pro</td>
<td>54Wh, 25kJ</td>
<td>10h</td>
</tr>
<tr>
<td>11W</td>
<td>2.4 Hours</td>
<td></td>
</tr>
<tr>
<td>11W</td>
<td>2.5W, 8.4 Hours</td>
<td></td>
</tr>
<tr>
<td>2.5W</td>
<td>8.4 Hours</td>
<td></td>
</tr>
<tr>
<td>1-1000 mW</td>
<td>8-240 Hours</td>
<td></td>
</tr>
</tbody>
</table>

1 Wh = 3.6 kJ
Other Commercial Pushes to the Edge

Offload computations to the edge:
- Teenagers are impatient ➔ low latency
- Hate speech detection
- Porn detection
- $0.86 for 1M inferences not a lot, unless you have 1 billion users

Sell more chips
Sell more phones and gadgets
Outline

• Losing privacy and security in our modern world
• What conveniences from Deep Learning applications do we want?
  – Home
  – Car
  – Office
  – Shopping and recommendations
  – Personal assistant
• How do we get these, but retain privacy (and security)?
  – Privacy vs security
  – A variety of approaches for providing privacy and security
• Privacy at the Edge:
  Efficiency is the key to local processing
  – Computer vision
  – Audio and Speech
  – NLU
  – Recommendations
• Challenges for the future
Efficient Deep Learning Technologies at the Edge
Enable Applications at the Edge

Computer Vision and Core ML
- Image Classification
- Object Detection
- Image Segmentation

Audio Analysis
- Audio Enhancement
- Call-center Sentiment Analysis
- Speech Recognition

Multimedia and Rec Systems
- Video Sentiment Analysis
- Music Recommendation
- Ad Recommendation

Natural Language Processing
- Translation
- Question answering
- Document Understanding
Last five years of research have nicely matured CV at the edge

Top-1 Image Classification 75%+; <300MOPS; 1-5M model params
Automatic Speech Recognition at the Edge

End-to-end Deep Learning models have brought on-device ASR to the edge

<table>
<thead>
<tr>
<th>Model</th>
<th>Size</th>
<th>Voice Search WER (%)</th>
<th>EOU Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional Server</td>
<td>87.2GB</td>
<td>6.6%</td>
<td>870ms</td>
</tr>
<tr>
<td>On-Device End-to-End</td>
<td>0.18GB</td>
<td>6.1%</td>
<td>780ms</td>
</tr>
</tbody>
</table>


Others:
- Jasper, Nvidia
- Conformer, Google
- Amazon, Apple

- End-to-end Deep Learning models have brought on-device ASR to the edge
Neural Network Architecture | GLUE score | Model Params (Million) | GFLOPs per seq | Latency Google Pixel 3 | Speedup
---|---|---|---|---|---
BERT-base | 78.3 | 109 | 22.5 | 1.7 (sec) | 1x
MobileBERT | 78.5 | 25.3 | 5.36 | 0.57 | 3.0x
SqueezeBERT | 78.1 | 51.1 | 7.42 | 0.39 | 4.3x
Recommendation Systems at the Edge: Inference

- Less computation than CV/NLP/ASR
- But … large embedding tables that encode products (e.g. retail products) and user behavior
  - Exceed size of on-device memory
- Solution 1: Only deploy low parameter models to the edge
  - recipe choices, recent TV series, recent movies
Split Architecture for Rec Systems

- **Solution 2:** split rec model between cloud and edge
  - Cloud model narrows selection to k candidates
  - Local user chooses best of k using local data

- **Example:** Alibaba EdgeRec
  - All models trained on cloud
  - Low latency
  - Lacks full privacy

- **Interesting future direction**
Summary: Three Elements of Efficiency at the Edge

- New DNN Models
- Optimizations: Pruning, Quantization, Distillation
- New Processors And DNN Accelerators
Rapid Improvements in Edge Processors is Going to Help
Summary and Conclusions

• We’re losing our privacy in the public world, let’s not lose it in our private world
  – Home, car, office
• We want the convenience of applications built from Deep Learning systems
  – Command and control in home or car
  – Natural language understanding in more complex question-answer situations: cooking, recipes, everyday questions
• But we don’t’ want
  – Auditory or visual eavesdropping (aka peeping tom)
• The key to balancing convenience and privacy is efficient Deep Learning at the edge
• We’ve made a lot of research progress, commercial availability of integrated applications are still to come
• Still many problems to be solved to improve accuracy, latency, and efficiency
Thank You For Your Attention