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

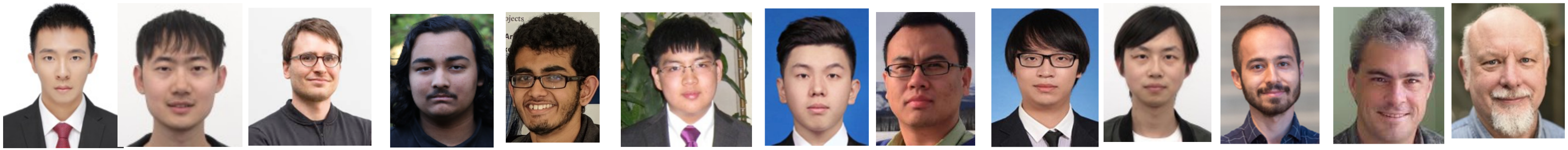
Zhen Dong, Tianren Gao, **Ravi Krishna, Suresh Krishna, Ani Nrusimha**

Sheng Shen, Zhewei Yao, Bohan Zhai,

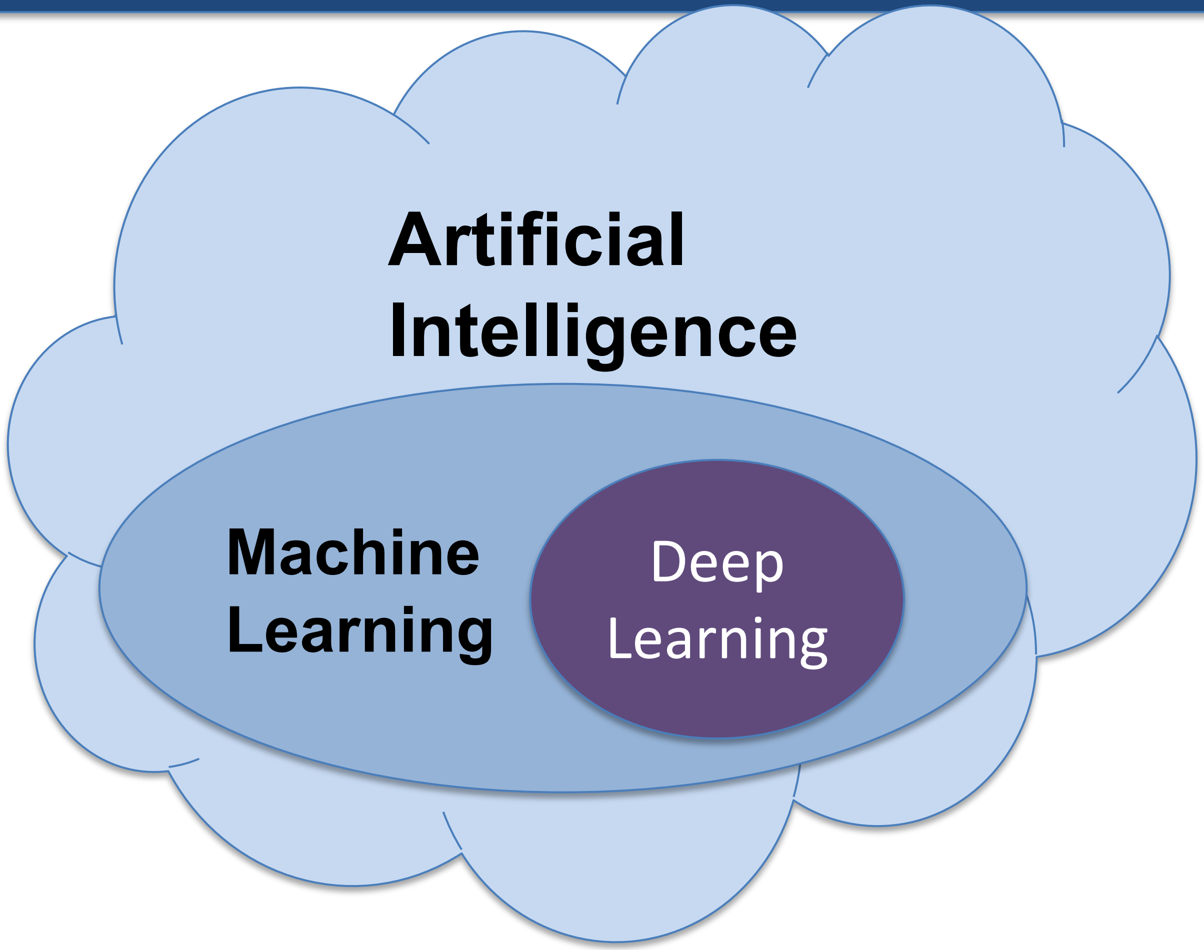
Amir Gholami, Shanghang Zhang,

Joey Gonzalez, **Kurt Keutzer**, Michael Mahoney,

with Forrest Iandola (self), Albert Shaw (Tesla), Bichen Wu (FB), Flora Xue (DeepMind)



AI/Machine Learning/Deep Learning



- Artificial intelligence: definition is always evolving
- Machine learning: well defined
- Deep Learning is a relatively small subset of Machine Learning approaches

My Group Today:
All Deep Learning All The Time

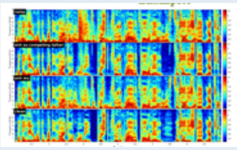
Computer Vision and Core ML

Image Classification

Object Detection

Image Segmentation

Audio Analysis


Audio Enhancement

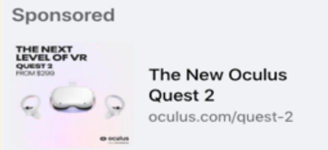

Call-center Sentiment Analysis


Speech Recognition

Multimedia and Rec Systems

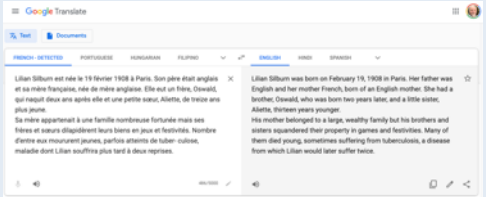

Video Sentiment Analysis


Music Recommendation



Ad Recommendation

Natural Language Processing


Translation


Question answering


Document Understanding

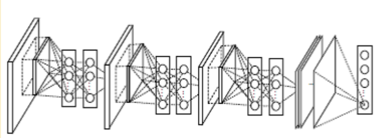
State-of-the-art solutions for all these problems (and more) rely on deep learning

State-of-the-Art Solutions
Typically Rely on one DNN (or a few)

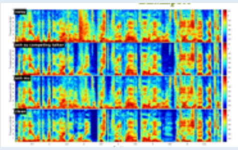
Image Classification

Object Detection

Image Segmentation



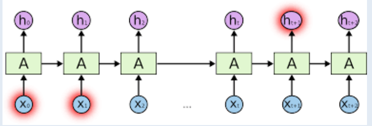
Convolutional NN



Audio Enhancement

Call-center Sentiment Analysis

Speech Recognition

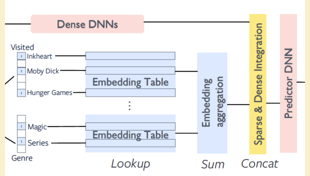


Recurrent NN

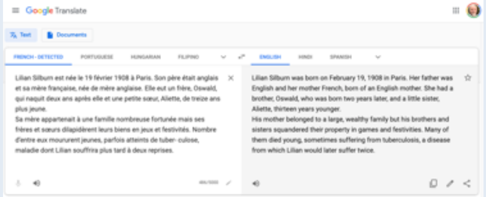
Video Sentiment Analysis

Music Recommendation

Ad Recommendation



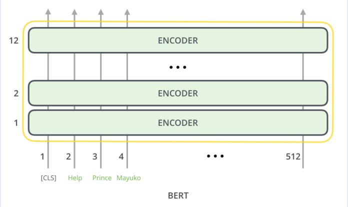
DLRM



Translation

Question answering

Document Understanding



Transformer

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

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

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

The Tightrope
Convenience vs privacy

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

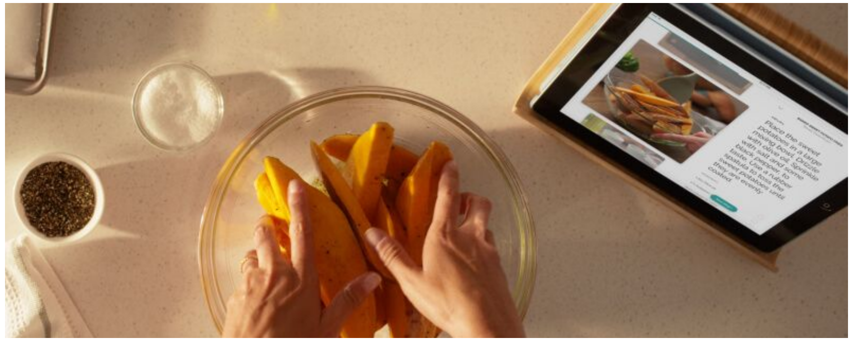
Privacy in our:
- Speech (pattern and intonation)
- Conversations
- Personal visual spaces
- Personal preferences

- There is a fundamental tension between the convenient features that we would like to have in our private spaces and preserving our privacy

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

Personal Assistant Everywhere
Convenience vs Privacy

Conveniences:

- A personal digital assistant may be the dashboard of every capability we have described so far.

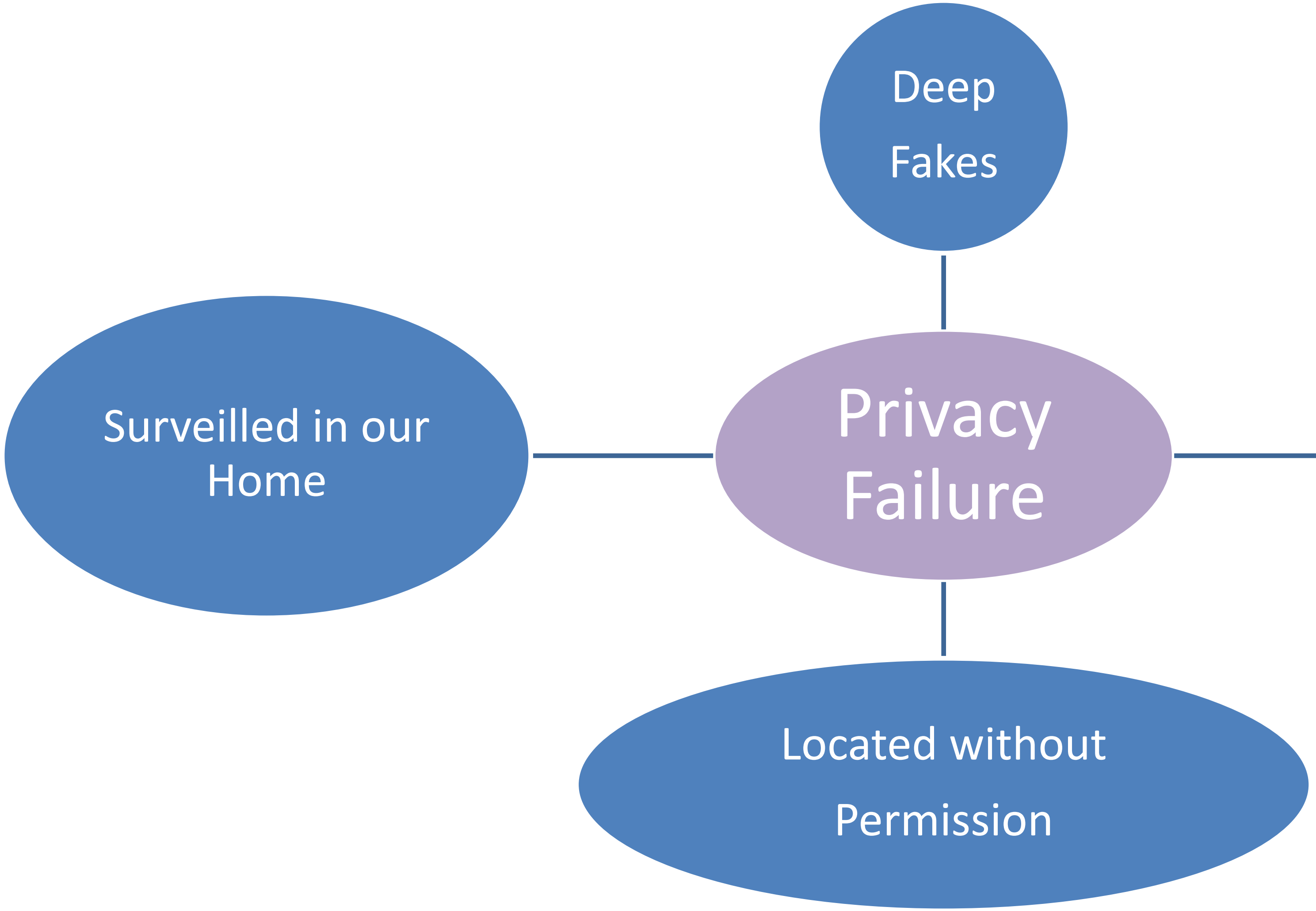
With privacy:

- 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?



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

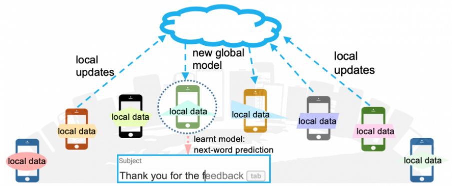
- Privacy programs:
  - GDPR: General Data Protection Regulation
  - California Consumer Privacy Act
- Users control access to data before its collected
  - Who will be given the data?
  - What the data will be used for?
  - How long the data will be stored?
- Users must be assured that data is deleted 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

Mystery Data → Training with DP → Outputs Model X ⇒

P(Data without you → Training with DP → Outputs Model X)

≈ P(Data including you → Training with DP → Outputs Model X)

- Full /partial Training as well as Full Inference at Edge

Cloud hosted applications: data, inference, and training

Application served

User Data

- 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

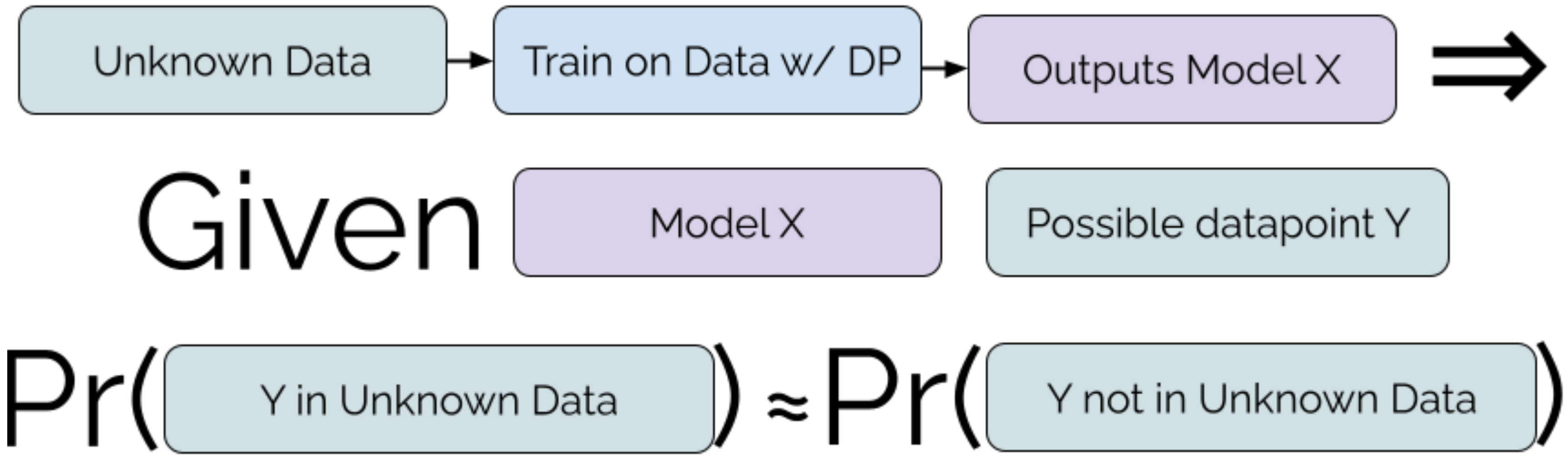
- Protects: users private data
- Approach
  - Local training is performed on the local computer/phone
  - Only local updates (gradients) sent to server
  - New global model periodically trained
  - Global models returned to user
- Problem: Some user information may be leaked through the gradients
- E.g.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

Li, Tian et al. "Federated Learning: Challenges, Methods, and Future Directions." IEEE Signal Processing Magazine 37.3 (2020): 50–60. Crossref. Web.

Differential Privacy with Federated Learning

- 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



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

~10,000 x   >>>>

TPU Pod
125,000 TFLOPS

Samsung S21 Ultra

Edge Client
5 – 15 TOPS

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

Average Power Dissipation

Battery Life

1-1000 mW
8-240 Hours

2.5 W
8.4 Hours

11 W
2.4 Hours

15 inch Macbook Pro |
76Wh = 273.6kJ
Apple: 10h = 7.6 W

13 inch Macbook Air | 54Wh = 194.4kJ
Apple: 12h = 4.5 W

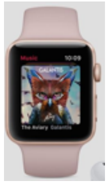
Eee PC 1000HE | 49Wh = 176kJ
Asus: 9.5h = 5.2 W

iPad Pro | 41Wh = 147kJ
Apple: 10h use = average 4.1W

Kindle Oasis | 0.91Wh = 3.276kJ
Ebook Friendly: "15days @ 30m/day" = 7.5h @ 0.12 W average

2017 iPhone8 | 6.96Wh = 25kJ
Talk time: 14h = 0.5 W
Video: 13h = 0.54 W

Typical handset
32g, 13cc, 5.5Wh = 19.8 kJ
Typical usage
5kJ active + 12kJ standby = 1 battery charge
Per Ljung – Nokia, 2012



iwatch Series 3 1.07Wh | 3.8kJ
60mW, 18 hours

OZO Digital Pedometer
80µW, 0.72Wh | 1 year

1 Wh = 3.6 kJ

Other Commercial Pushes to the Edge


Sell more chips

facebook

Sell more phones and gadgets

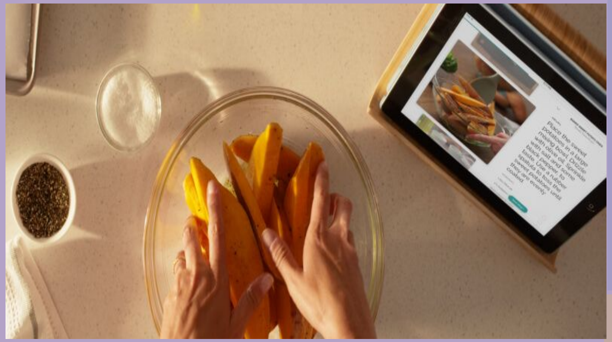
Offload computations to the edge:

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

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



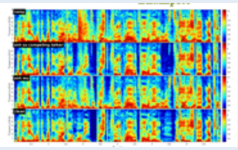
Computer Vision and Core ML

Image Classification

Object Detection

Image Segmentation

Audio Analysis



Audio Enhancement

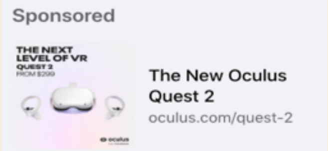
Call-center Sentiment Analysis

Speech Recognition

Multimedia and Rec Systems

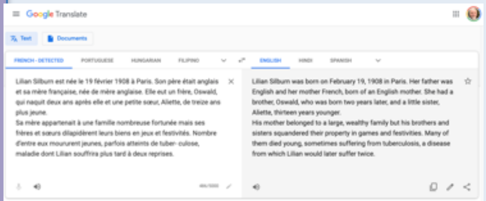
Video Sentiment Analysis

Music Recommendation

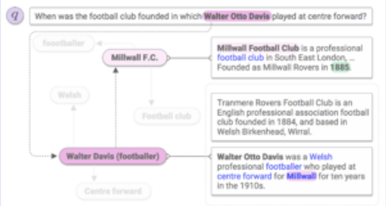


Ad Recommendation

Natural Language Processing



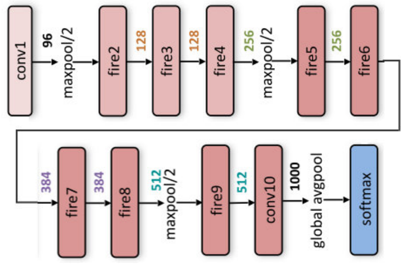
Translation

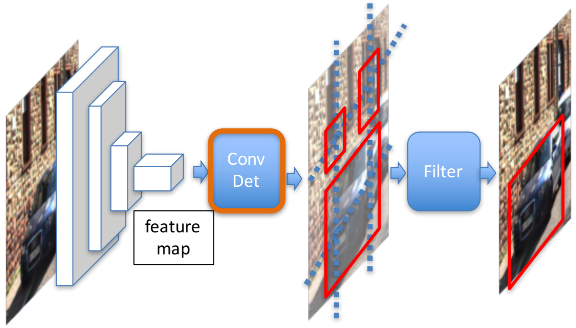


Question answering

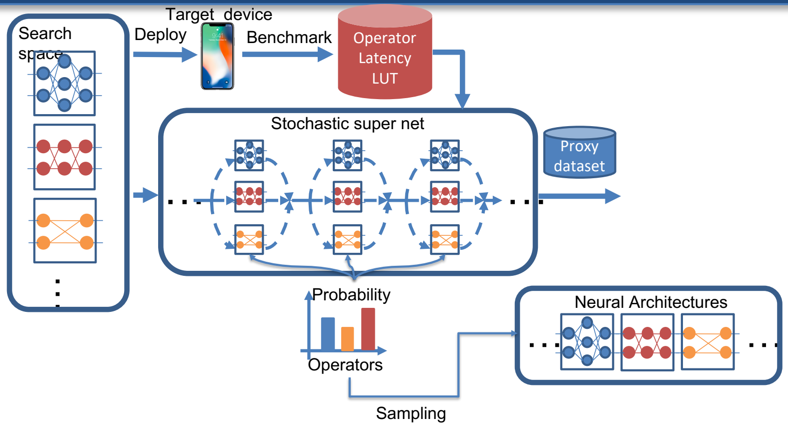
Document Understanding

Computer Vision at the Edge

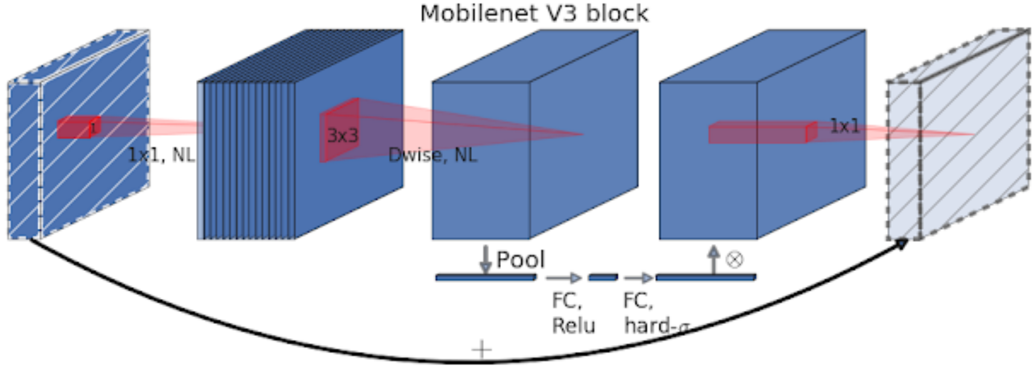




Squeeze Family

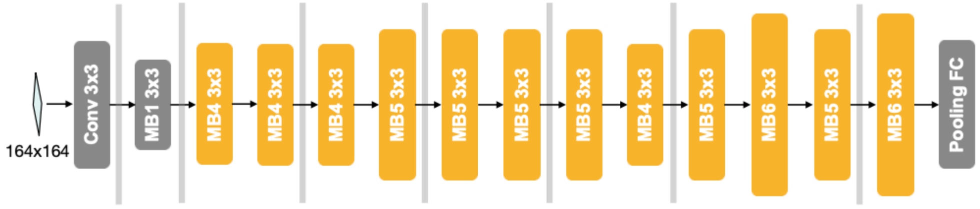


FBNet v3 Family



MobileNet v3 family

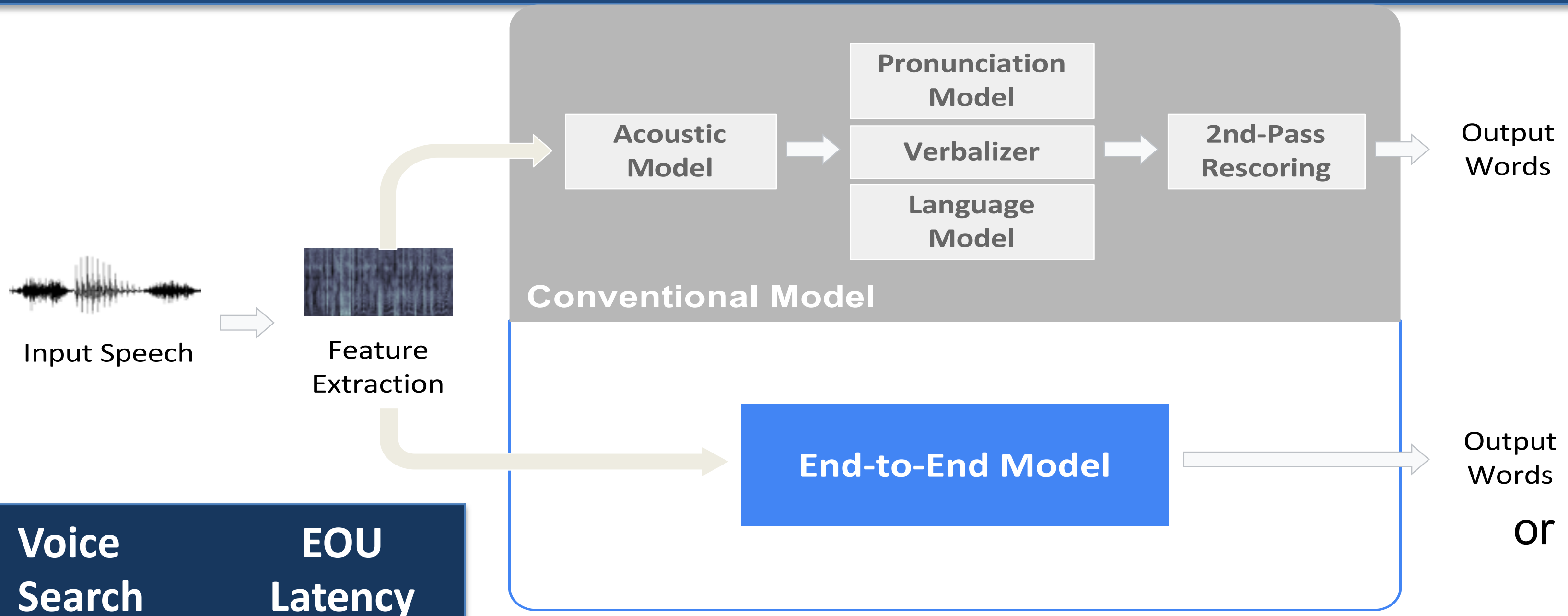
(kdnuggests.com)



Han Lab family

- 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



or commands

Model	Size	Voice Search WER (%)	EOU Latency
Conventional Server	87.2GB	6.6%	870ms
On-Device End-to-End	0.18GB	6.1%	780ms

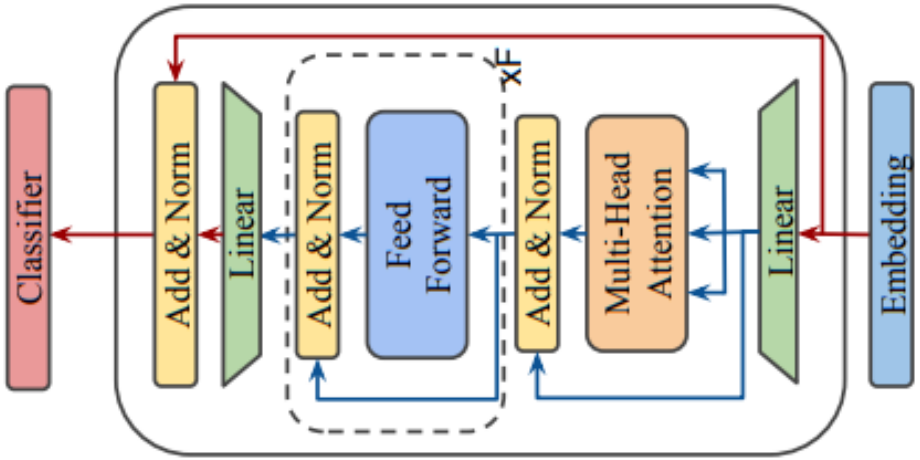
[Sainath2020] T. N. Sainath, Y. He and et. al., “A streaming on-device End-To-End model surpassing server-side conventional model quality and latency,” Proc. of ICASSP, 2020.

Others:

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

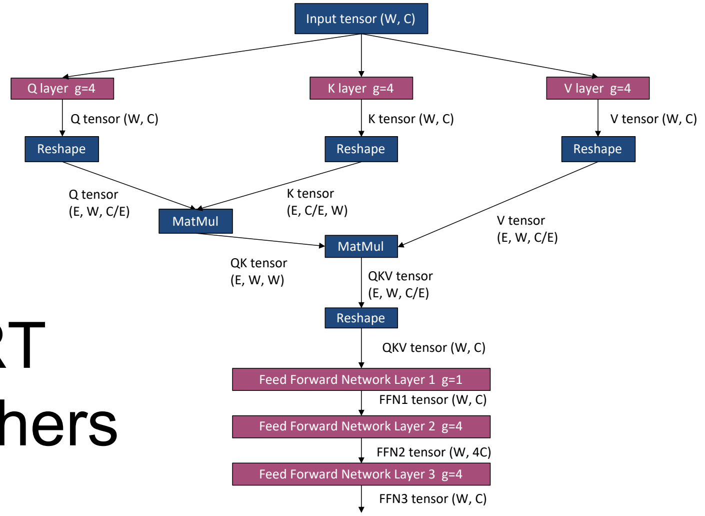
- End-to-end Deep Learning models have brought on-device ASR to the edge

Natural Language Understanding at the Edge



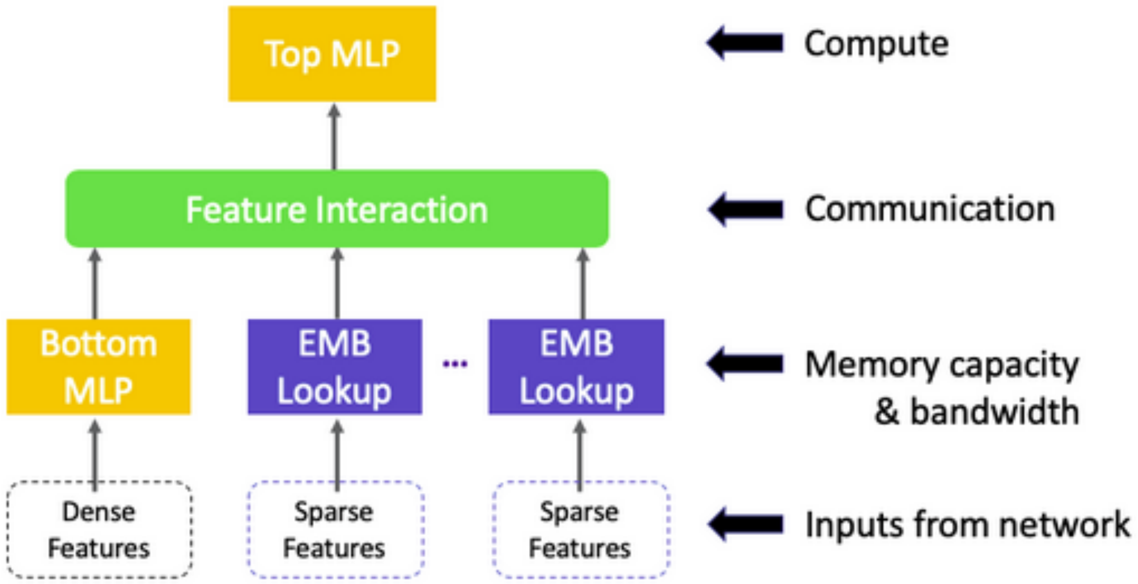
MobileBERT
Google

SqueezeBERT
Berkeley and others



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

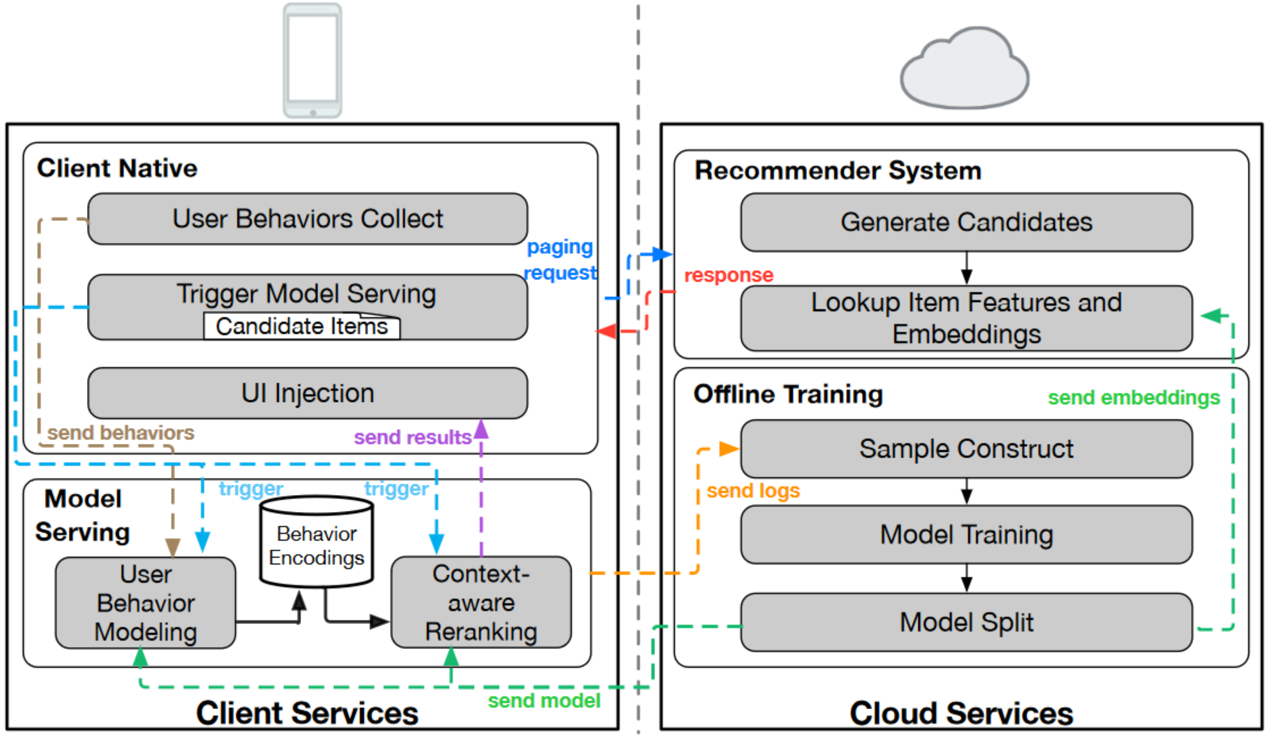


Facebook Deep Learning Recommendation Model

- 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

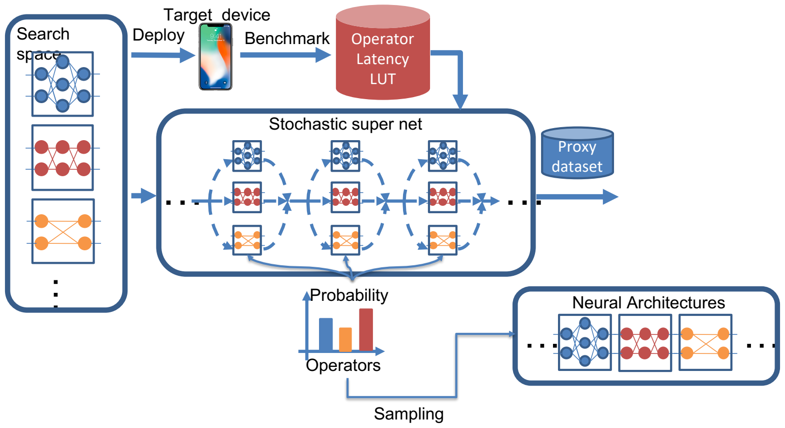
Split Architecture for Rec Systems

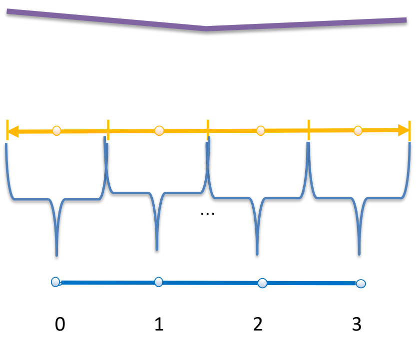
- Solution 2: split rec model between cloud and edge
  – Cloud model narrows selection to k candidates
  – Local user chooses best of k using local data
- Example: Alibaba EdgeRec
  – All models trained on cloud
  – Low latency
  – Lacks full privacy
- Interesting future direction



Summary: Three Elements of Efficiency at the Edge

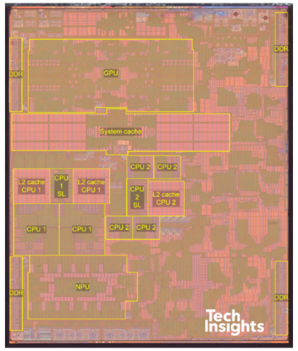
New DNN Models

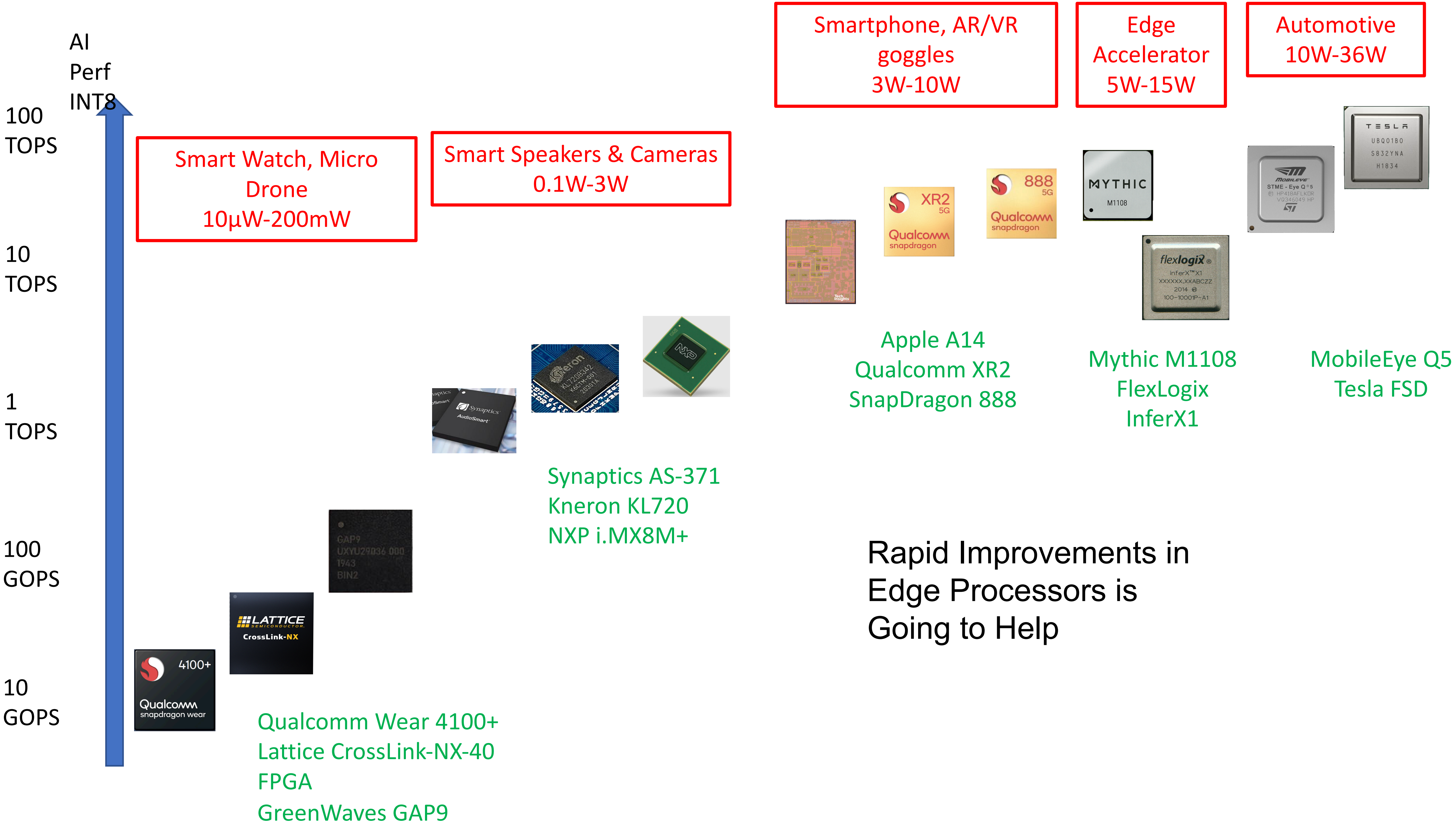




Optimizations: Pruning Quantization Distillation

New Processors And DNN Accelerators





AI Perf INT8
100 TOPS
10 TOPS
1 TOPS
100 GOPS
10 GOPS
Smart Watch, Micro Drone 10µW-200mW
Smart Speakers & Cameras 0.1W-3W
Smartphone, AR/VR goggles 3W-10W
Edge Accelerator 5W-15W
Automotive 10W-36W
Qualcomm Wear 4100+
Lattice CrossLink-NX-40 FPGA
GreenWaves GAP9
Synaptics AS-371
Kneron KL720
NXP i.MX8M+
Apple A14
Qualcomm XR2
SnapDragon 888
Mythic M1108
FlexLogix InferX1
MobileEye Q5
Tesla FSD
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 Sponsors

Cloud

Google

Alibaba.com

Facebook AI Research (FAIR)

Mobile Clients

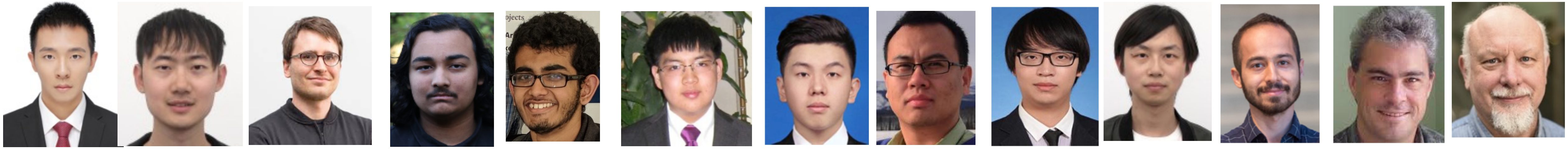
facebook

SAMSUNG

SAMSUNG ADVANCED INSTITUTE OF TECHNOLOGY

Berkeley DeepDrive

Thank You For Your Attention



Extras

-