Ergo™: Perceive’s chip – data center-class inference in edge devices at ultra-low power

Steve Teig
Perceive
September 2020
The problem

Today’s consumer camera/doorbell streams everything to the cloud

Significant costs to user & provider

- Cloud
- Latency
- Power
- Security
- Privacy
Transmitting data is expensive!

More data
Faster data
Farther data
Wireless data

HIGH RISK AND COST

Energy + Money + Privacy + Security

© 2020 Perceive
Previous Edge solutions had severe limitations

Too big
- Tens of $
- 2-5 Watts

Too limited
- <1 Watt
- Poor accuracy

Requires re-thinking, from first principles, how to do Edge inference
Ergo™: Cloud-quality inference... but running at the Edge

- High Accuracy
- Ultra-High Power Efficiency
- Privacy and Security
- High Performance
- Flexibility
- Capability

© 2020 Perceive
Datacenter-class analysis inside a security camera

- Detect interesting motion and ignore false alerts
- Recognize faces, voices, and people
- Detect relevant objects – animals, packages, vehicles, etc.
- Use voice for local commands
- Detect important sounds – alarms, people, glass breaking, etc.
- Describe people, vehicles, or even the actions in a scene
Datacenter-class analysis inside a wearable

- Detect important sounds around the user
- Use local voice commands and advanced wake words to simplify device UI
- Recognize faces, people, voice, and emotions
- Detect relevant objects around the user
- Integrate data across multiple sensors
Cloud-quality inference... but running at the Edge

**High Accuracy**
Large neural networks → datacenter-like accuracy
- Full YOLOv3 → 64 M parameters
- M2Det → 73 M parameters

**High Performance**
Large neural networks >4 sustained GPU-equivalent TOPS
- Full YOLOv3 → 250 fps
- M2Det → 150 fps
Accuracy through Capacity and Performance

Bianco et al., 2018
Cloud-quality inference... but running at the Edge

**Flexibility**
Support for a wide range of neural network architectures

- CNNs (1x1, 3x3, 5x5, 7x7, dilated)
- Residual and Inception
- RNNs, LSTMs, GRUs, etc.

**Capability**
Multiple large neural networks running in parallel on a single device
Running large networks simultaneously on Ergo

Today: Ergo imaging

- **Multi-object detection**
  - M2Det – 73 M weights or YoloV3 – 64 M weights
- **FFD**
  - Proprietary NN – 0.5 M weights
- **FR**
  - Resnet28 – 11 M weights

Multiple CNNs + RNNs

>85 M weights!

Today: Ergo audio

- **Vocative wake word**
  - Proprietary LSTM – 1.5 M weights
- **Speaker ID**
  - Proprietary LSTM – ~1.4 M weights
- **Keyword spotting**
  - Proprietary LSTM – ~1.5 M weights

Today: Ergo imaging + audio

- **Multi-object detection**
  - 73 M weights
- **FFD**
  - 0.5 M weights
- **FR**
  - 11 M weights
  
- **Keyword spotting**
  - ~1.5 M weights

- **Audio event ID**
  - Proprietary LSTM – 0.7 M weights
Cloud-quality inference... but running at the Edge

Ultra-High Power Efficiency
20-100x improvement vs. alternatives

Privacy and Security
Sensor data need not leave the chip
Encryption of neural networks, CPU boot ROM, and chip access
20-100x advantage in computation per watt
Hardware Overview

- GlobalFoundries 22FDX
- Low-power 22nm FDSOI
Summary

Transform Sensing into *Perceiving*

Replace (or accompany) raw sensory info with inferences + comprehension

- Support for multiple image, audio and other I/Os
- Solutions that provide advanced features inside consumer devices
- Shipping in customer products in 2020

Cloud-quality inference... but running at the Edge

- **Accuracy and capacity:** large neural networks for datacenter-like accuracy
- **Performance:** large neural networks at frame rate; >4 sustained GPU-equivalent TOPS
- **Flexibility and capability:** wide range of network architectures; multiple networks at once
- **Power efficiency:** 20-100x improvement vs. alternatives
- **Privacy and security:** sensor data can stay on-chip; encryption of networks
More info

Perceive website

https://www.perceive.io

2020 Embedded Vision Summit

“Accuracy: beware of red herrings and black swans”

Tuesday, September 15, 11:00 AM PDT