embedded VISION SUMMIT 2018

At the Edge of AI at the Edge

Ultra efficient AI on low-power devices

XNOR.AI

Mohammad Rastegari

mohammad@xnor.ai

May 2018

Problem

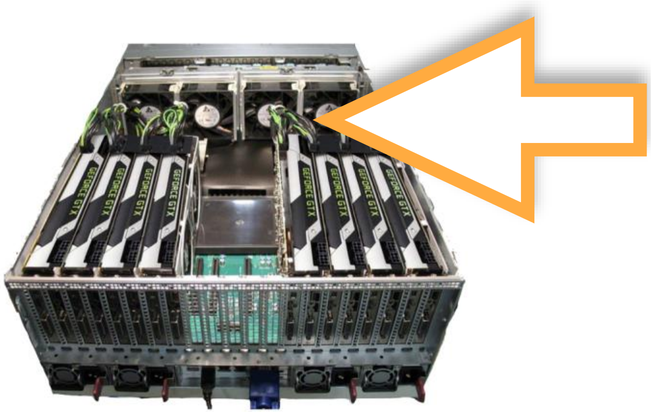
AI is confined to the cloud

far from the users at the edge

Goal

XNOR.AI bridges the growing divide between AI models dependent on the cloud and devices running at the edge

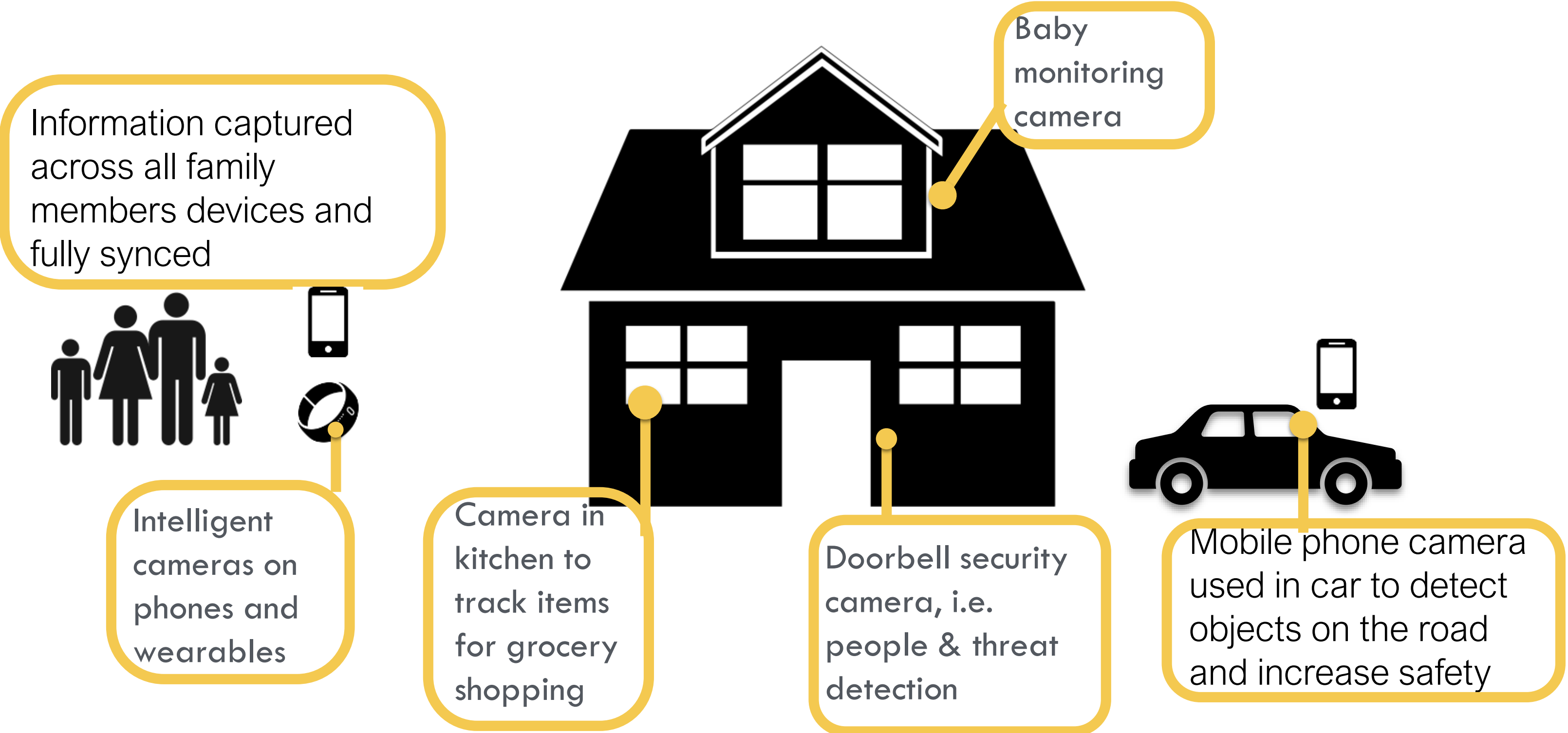
Deep learning models reliant on the cloud



Growing demand for edge devices

Opportunities

Intelligent cameras that preserve privacy, security and bandwidth at home



Information captured across all family members devices and fully synced

Baby monitoring camera

Intelligent cameras on phones and wearables

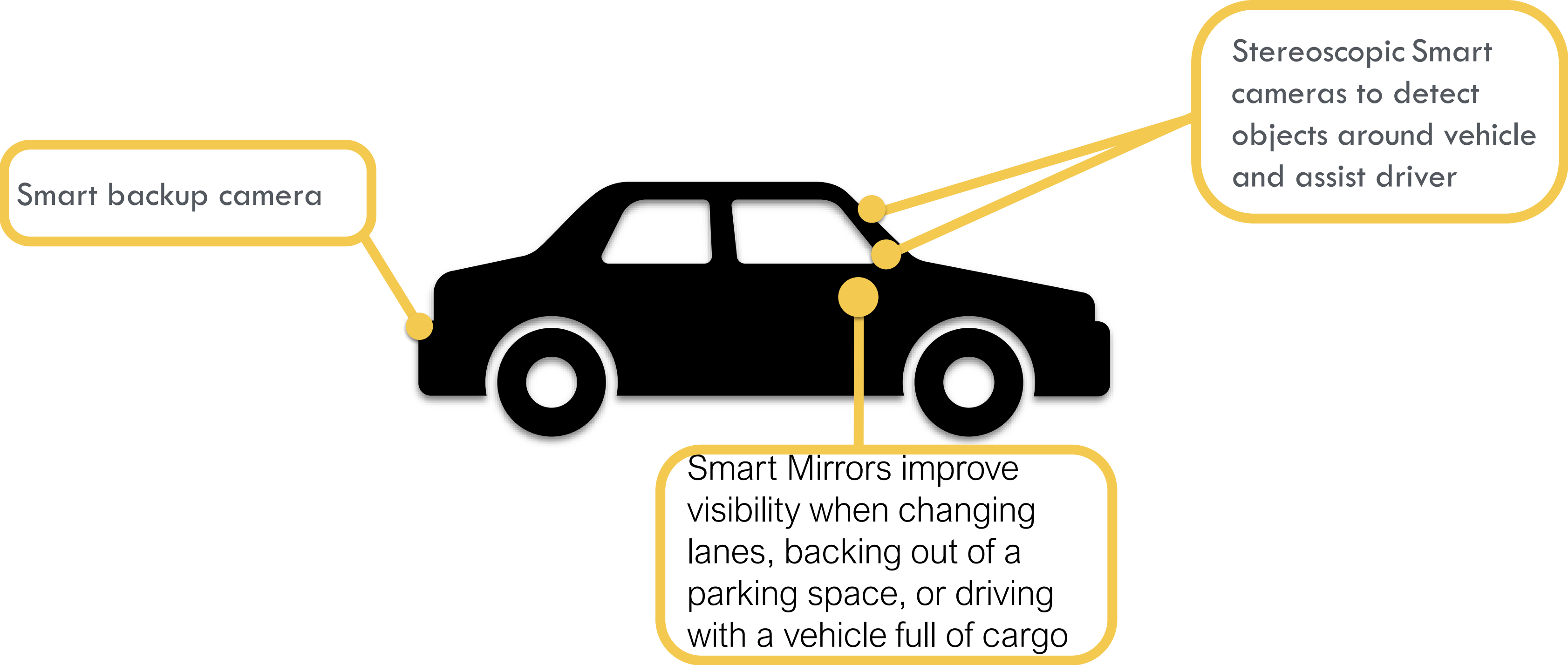
Camera in kitchen to track items for grocery shopping

Doorbell security camera, i.e. people & threat detection

Mobile phone camera used in car to detect objects on the road and increase safety

Opportunities

Intelligent cameras for Advanced Driver Assistance Systems (ADAS)



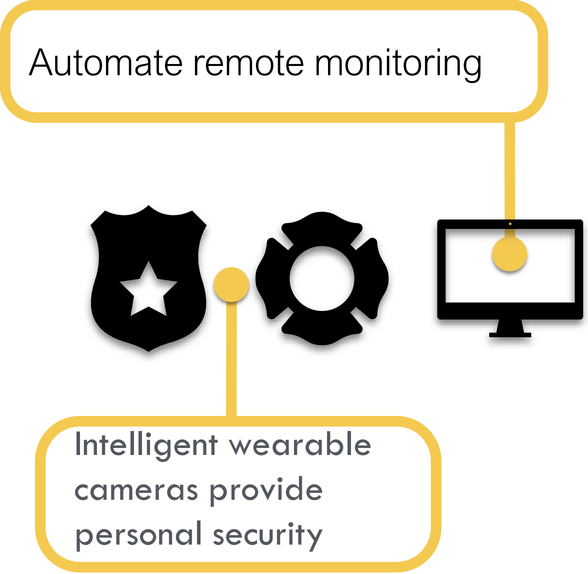
Smart backup camera

Stereoscopic Smart cameras to detect objects around vehicle and assist driver

Smart Mirrors improve visibility when changing lanes, backing out of a parking space, or driving with a vehicle full of cargo

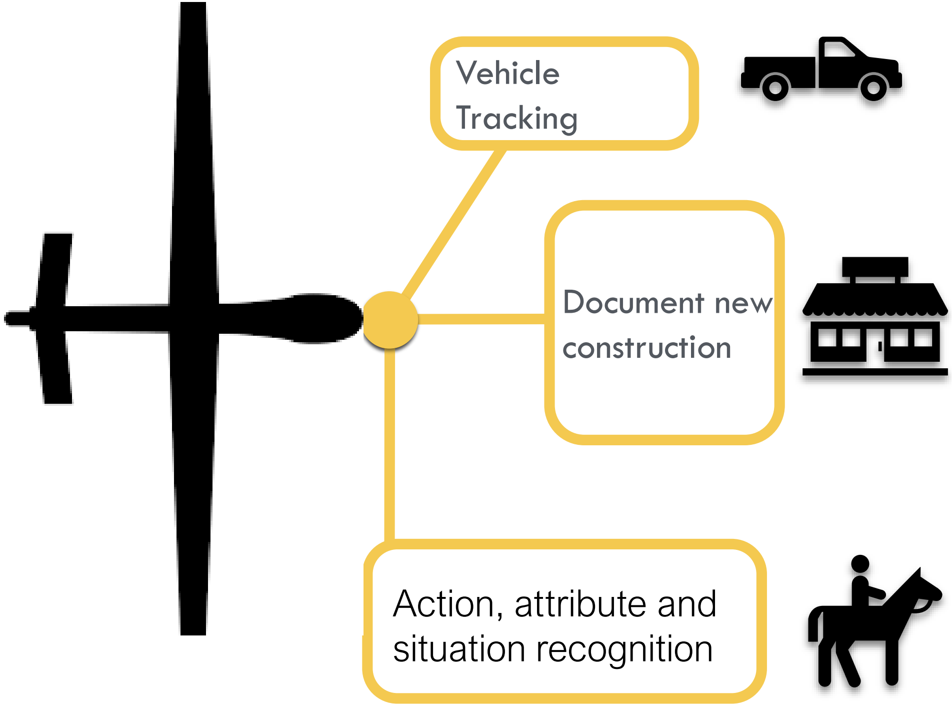
Opportunities

Intelligent cameras provide real-time tracking, recognition & detection on device



Automate remote monitoring

Intelligent wearable cameras provide personal security



Vehicle Tracking

Document new construction

Action, attribute and situation recognition

What we do

Image Tagging

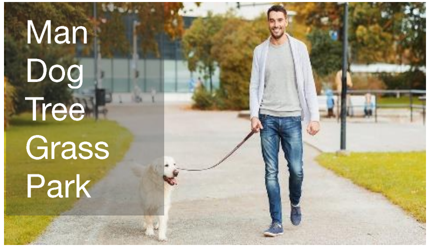
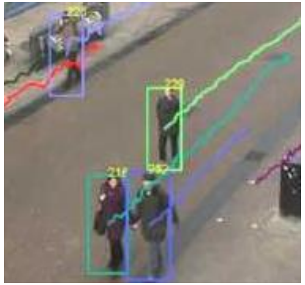
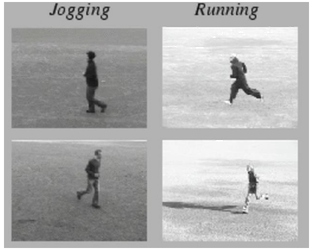


Image Enhancement

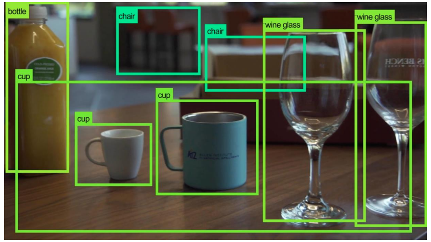
Tracking



Action Recognition

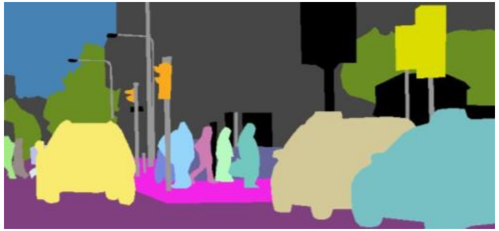


Object Detection



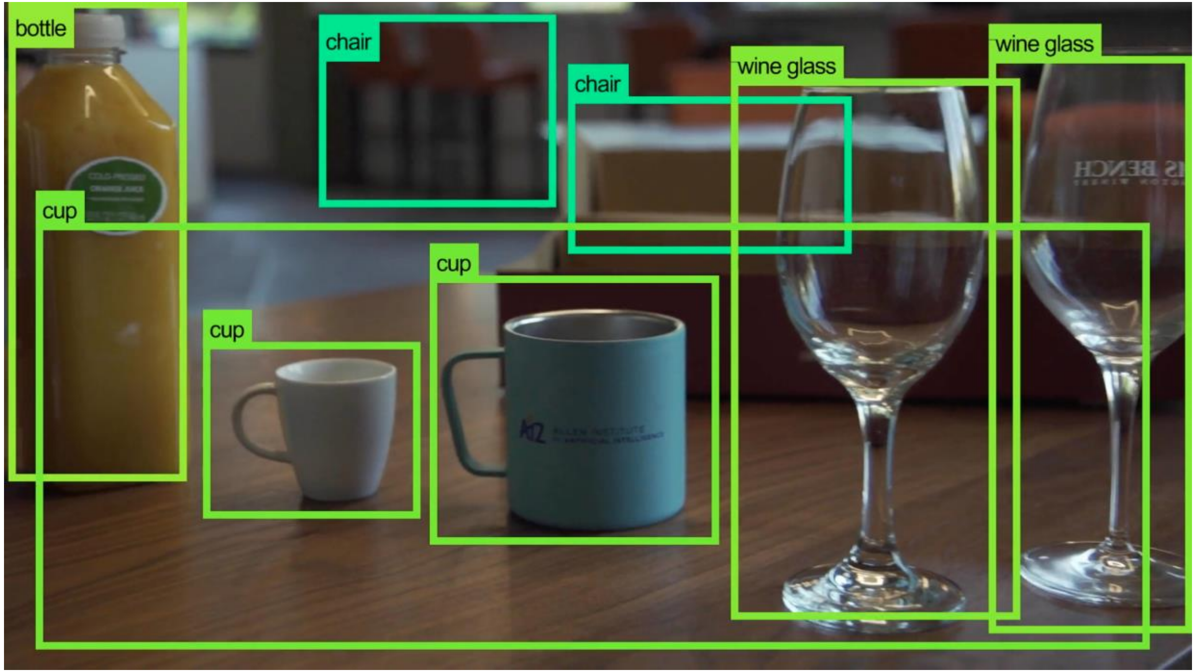
Scene Recognition

Segmentation

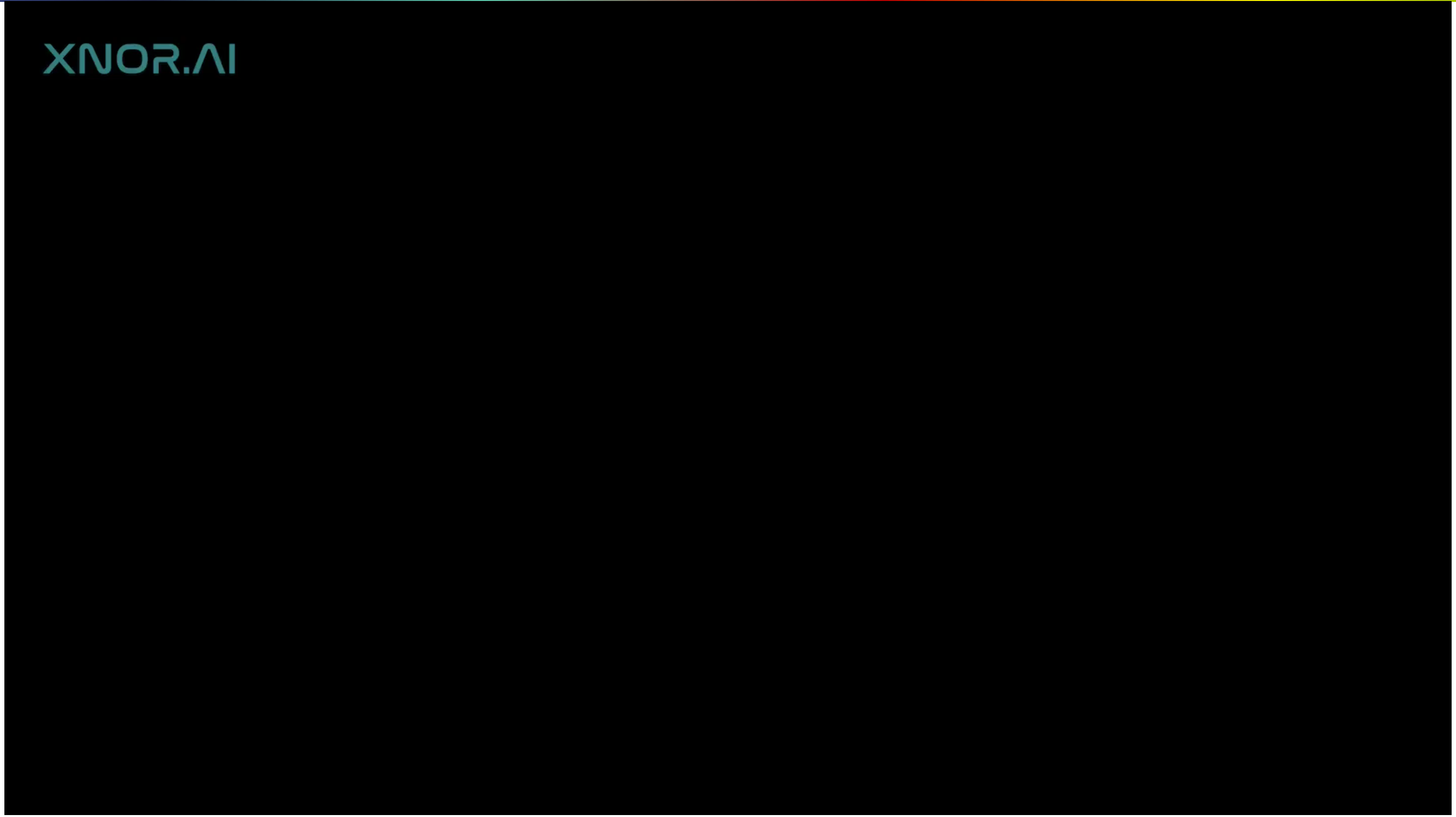


What we do

Object Detection: An Expensive Task in AI

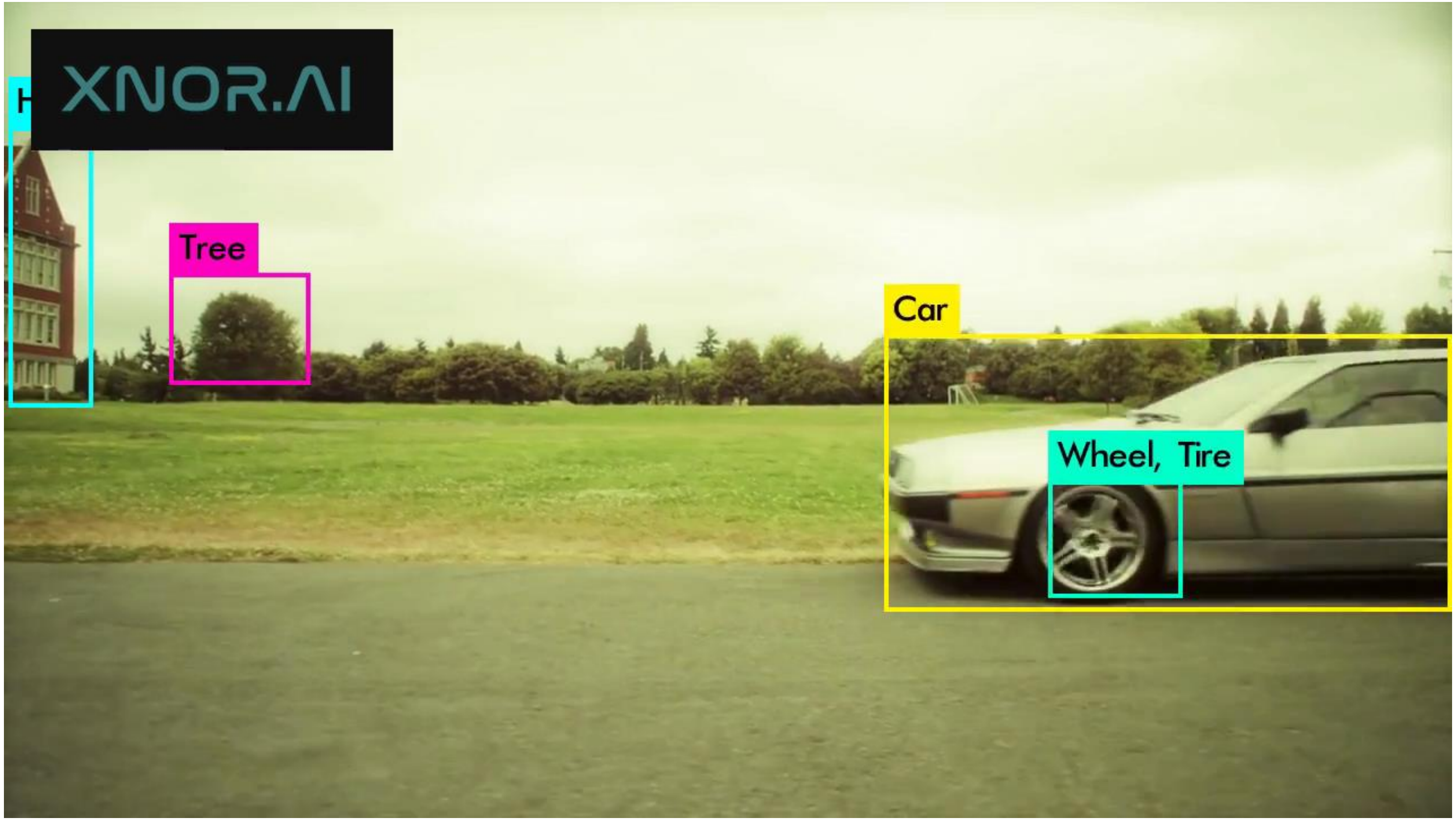


Dash Cam



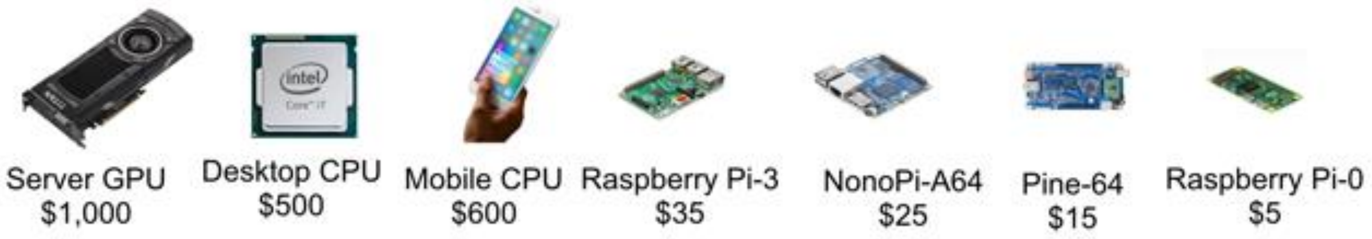
Security Camera

Fine Grain Categories



Hardware Platforms

XNOR.AI provides fast and efficient AI at the edge

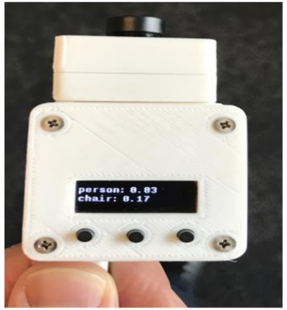


Low Power AI

State-of-the-Art AI: all the way to Pi Zero

XNOR $5 deep learning machine…
… on Raspberry Pi Zero

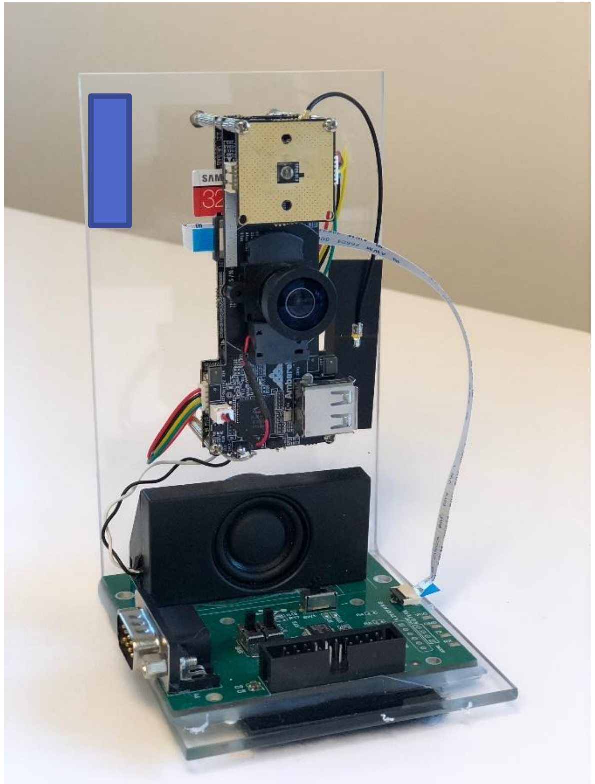
Modular AI at Edge



Low Power AI

Ambarella S5L

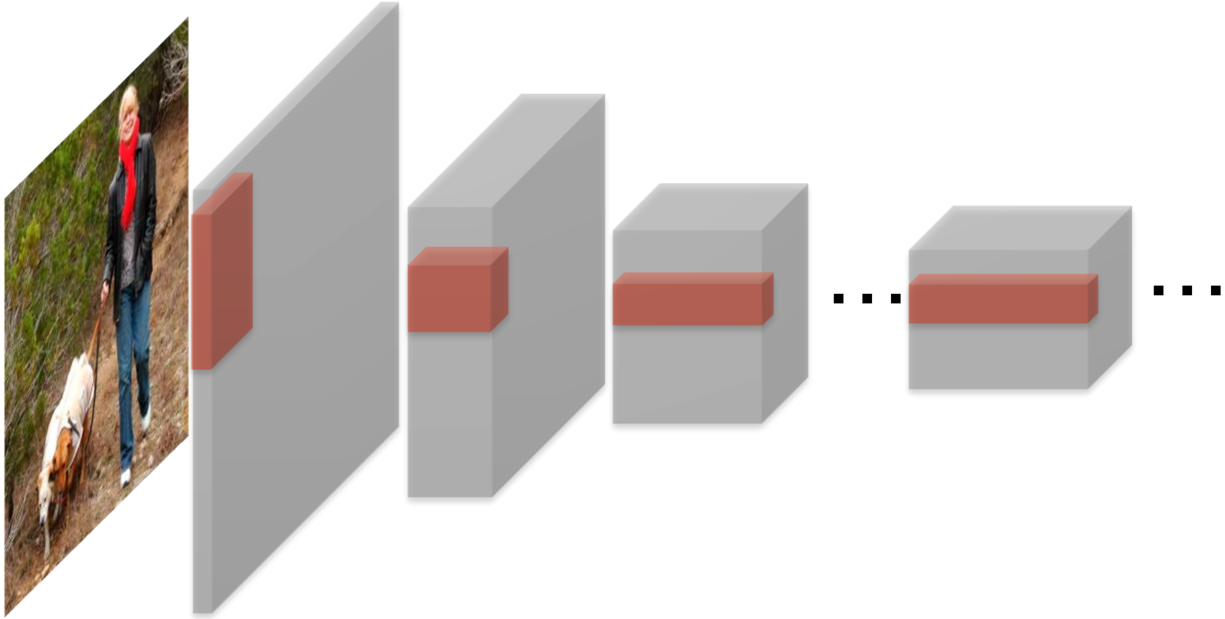
- Very low power (~2x lower than Pi Zero)
- Standard AI model for object detection
  - 1 fps
- XNOR AI Model for object detection
  - 17 fps



XNOR.AI Technology

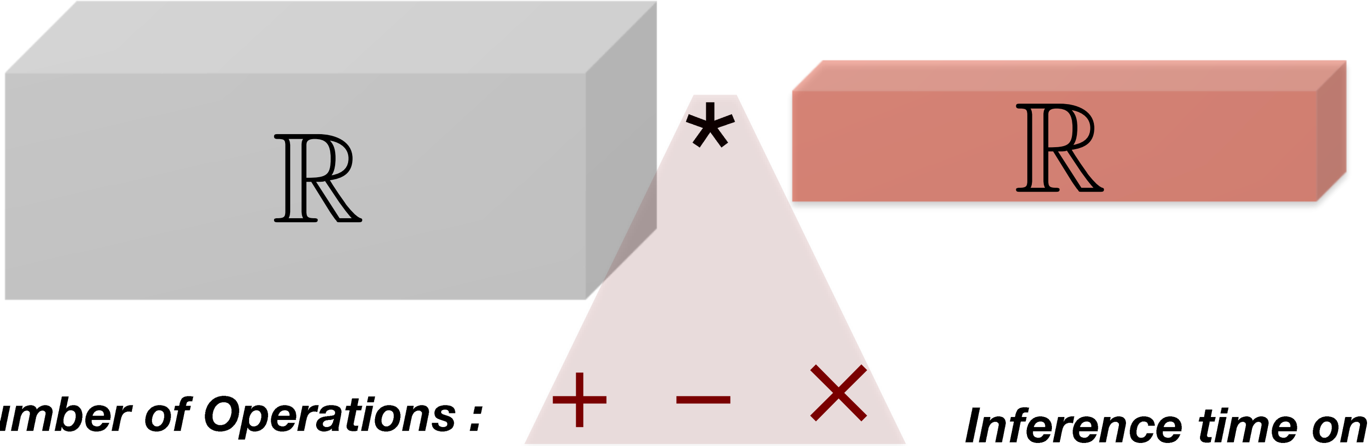
How our technology works

Convolutional Neural Network



XNOR.AI Technology

GPU !



Number of Operations :

- AlexNet →1.5B FLOPs
- VGG → 19.6B FLOPs

Inference time on CPU :

- AlexNet →~3 fps
- VGG → ~0.25 fps

XNOR.AI Technology

Lower Precision

Reducing Precision

- Saving Memory
- Saving Computation

$$\mathbb{R} \longrightarrow \mathbb{I} \longrightarrow \mathbb{B} \in \{-1, +1\}$$

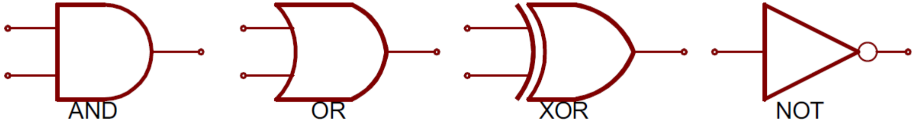
32-bit 8-bit 1-bit

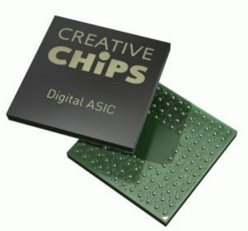
{-1,+1}	{0,1}
MUL	XNOR
ADD, SUB	Bit-Count (popcount)

XNOR.AI Technology

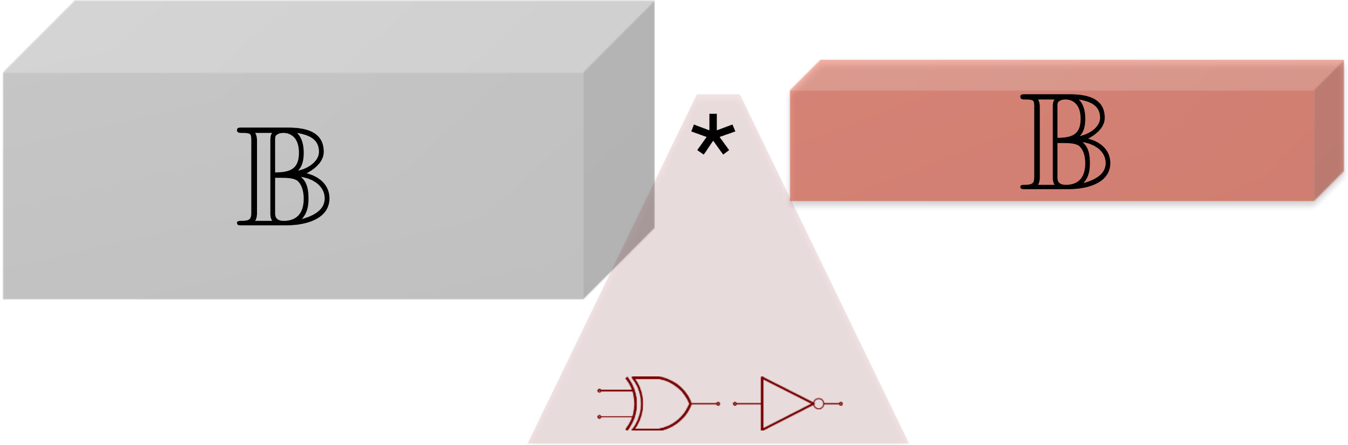
Why Binary?

- Binary Instructions
  - AND, OR, XOR, XNOR, PopCount (Bit-Count)
- Low Power Device
- Easy to Implement in hardware





XNOR.AI Technology

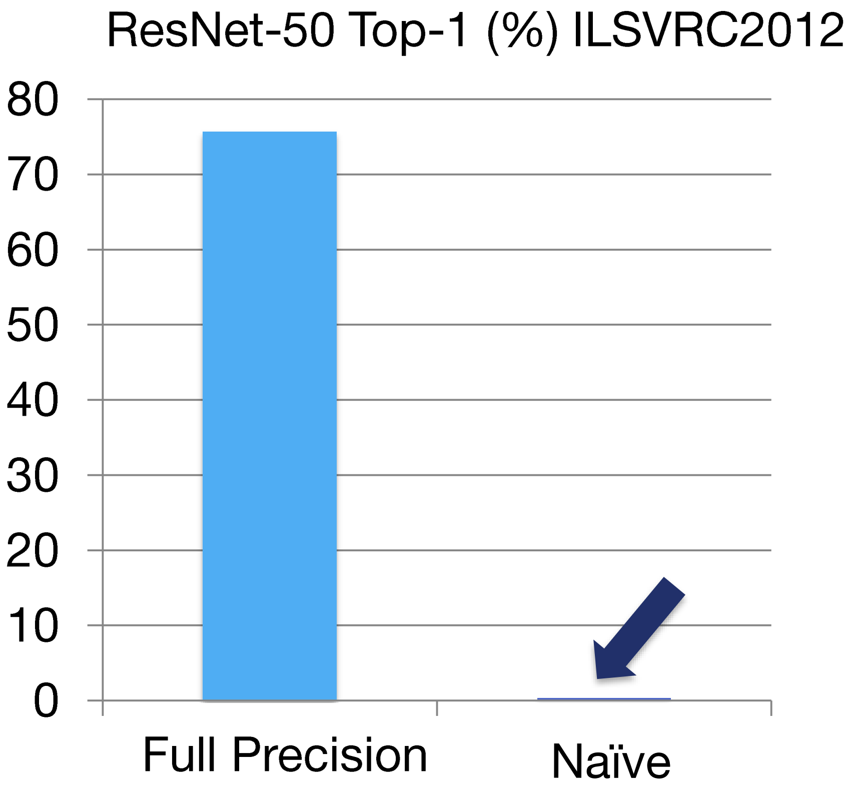


Training Binary Weight Networks

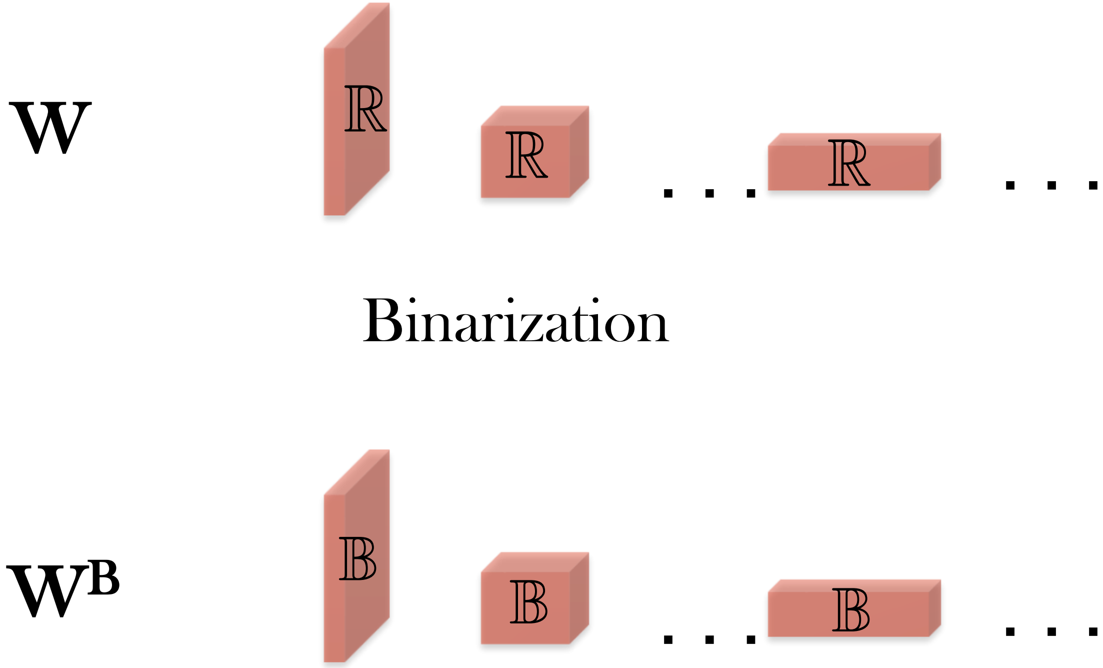
Naive Solution:

1. Train a network with real value parameters
2. Binarize the weight filters

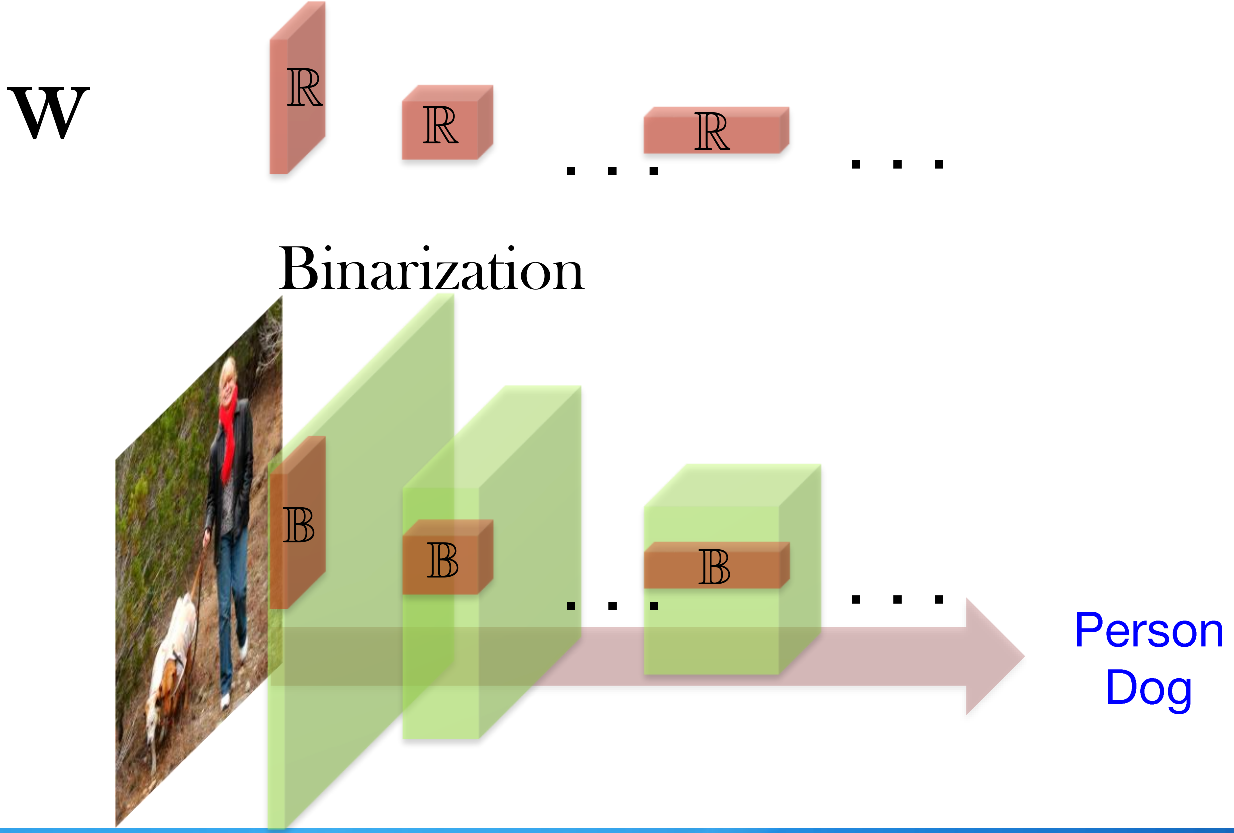
XNOR.AI Technology



XNOR.AI Technology



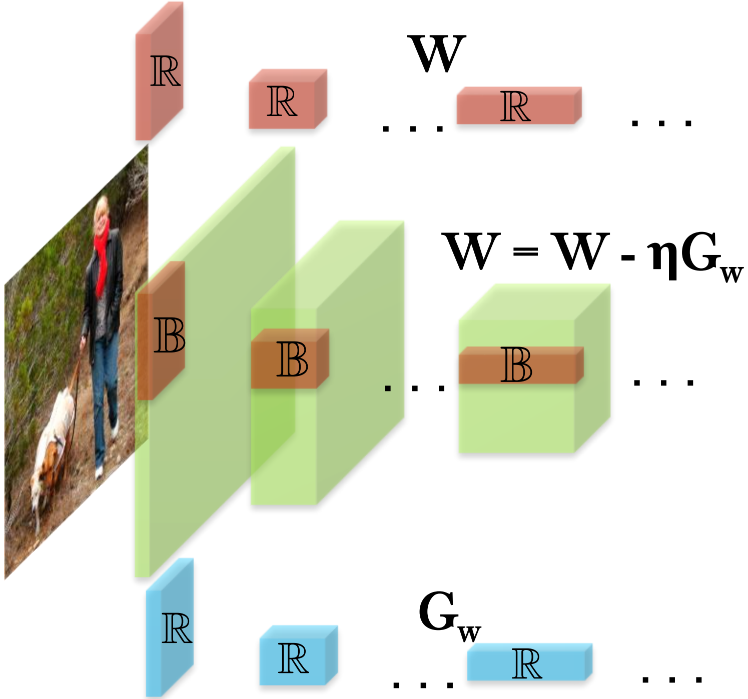
XNOR.AI Technology



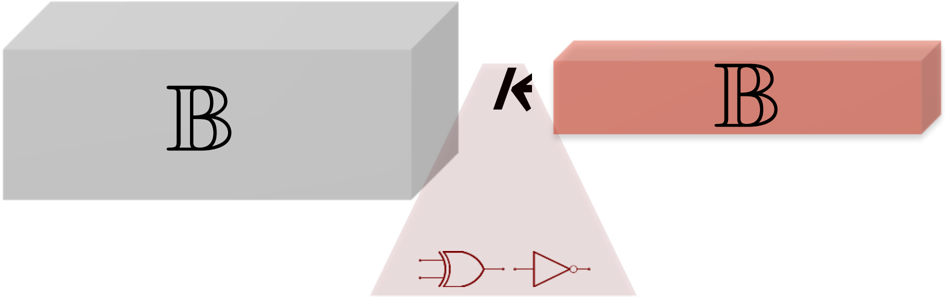
XNOR.AI Technology

Train for binary weights:

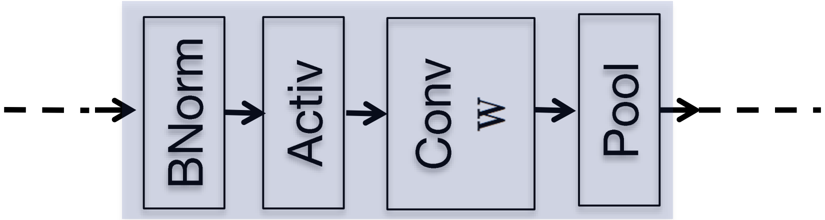
```
1. Randomly initialize $\mathbf{W}$
2. For $iter = 1$ to $N$
3.     Load a random input image $\mathbf{X}$
4.     $\mathbf{W}^{\mathbf{B}} = \text{sign}(\mathbf{W})$
5.     $\alpha = \frac{\|W\|_{\ell 1}}{n}$
6.     Forward pass with $\alpha, \mathbf{W}^{\mathbf{B}}$
7.     Compute loss function $\mathbf{C}$
8.     $\frac{\partial \mathbf{C}}{\partial \mathbf{W}} =$ Backward pass with $\alpha, \mathbf{W}^{\mathbf{B}}$
9.     Update $\mathbf{W}$ $(\mathbf{W} = \mathbf{W} - \frac{\partial \mathbf{C}}{\partial \mathbf{W}})$
```



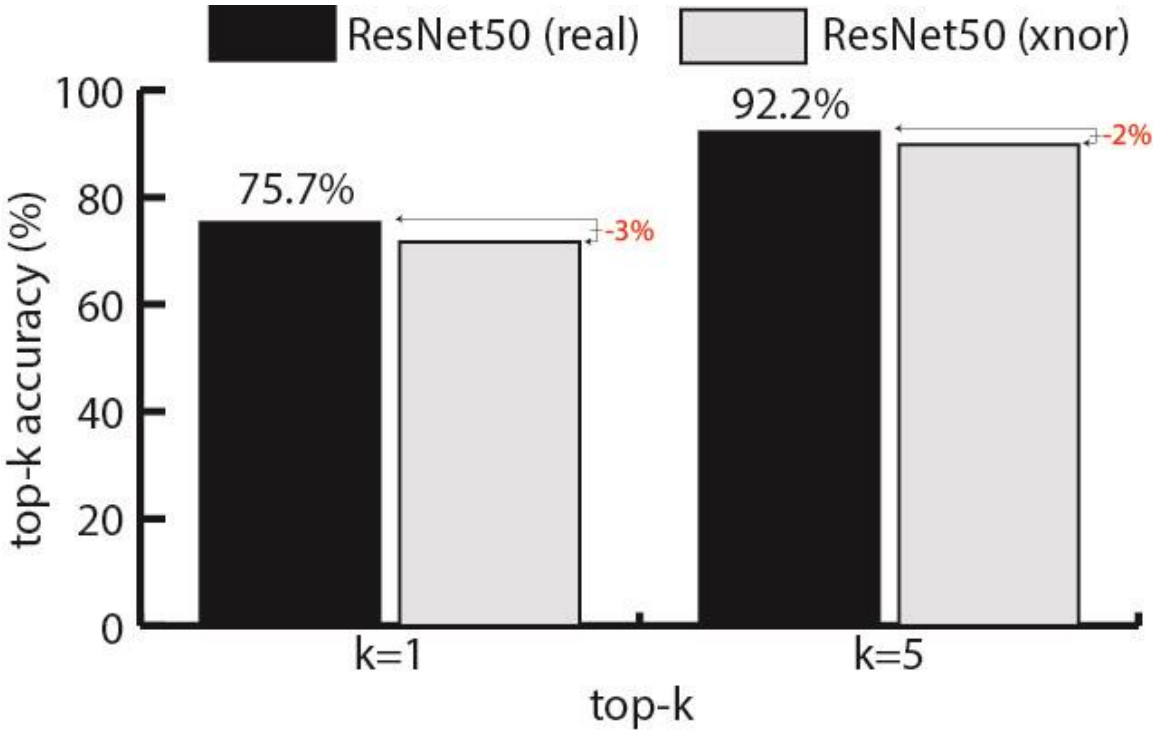
XNOR.AI Technology



```
1. Randomly initialize $\mathbf{W}$
2. For $iter = 1$ to $N$
3.     Load a random input image $\mathbf{X}$
4.     $\mathbf{W}^{\mathbf{B}} = \text{sign}(\mathbf{W})$
5.     $\alpha = \frac{\|W\|_{\ell 1}}{n}$
6.     Forward pass with $\alpha, \mathbf{W}^{\mathbf{B}}$
7.     Compute loss function $\mathbf{C}$
8.     $\frac{\partial \mathbf{C}}{\partial \mathbf{W}}$ = Backward pass with $\alpha, \mathbf{W}^{\mathbf{B}}$
9.     Update $\mathbf{W}$ $(\mathbf{W} = \mathbf{W} - \frac{\partial \mathbf{C}}{\partial \mathbf{W}})$
```



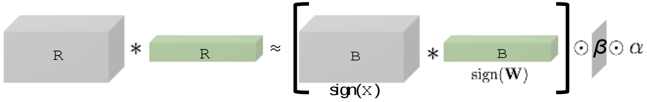
XNOR.AI Technology



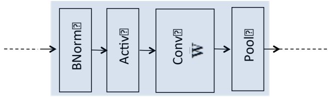
- 15x Smaller
- 10x Faster
- 200% power efficiency

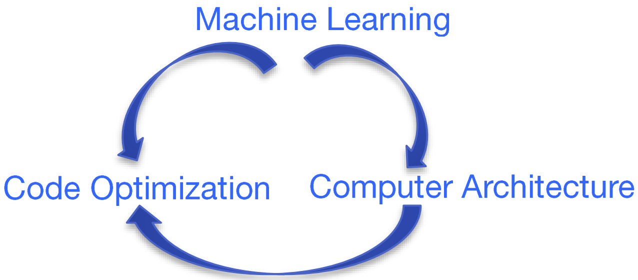
XNOR.AI Technology

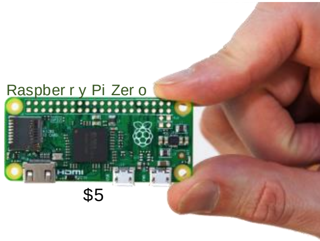
XNOR.AI



```
1. Randomly initialize W
2. For iter = 1 to N
3.    Load a random input image X
4.    $W^B$ = sign(W)
5.    ↵ = $\frac{kWk`1}{n}$
6.    Forward pass with ↵, $W^B$
7.    Compute loss function C
8.    $\frac{@C}{@W}$ = Backward pass with ↵, $W^B$
9.    Update W (W = W − $\frac{@C}{@W}$)
```



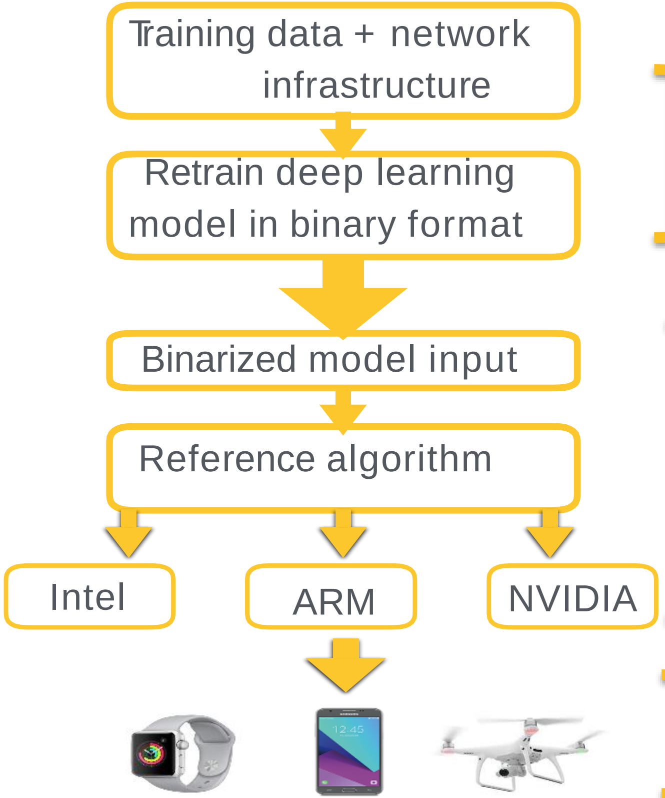




Integrate with XNOR.AI

How to integrate with XNOR.AI?

Enterprise



CUSTOMER INPUT

Training data + network infrastructure

XNOR-NET

Retrain deep learning model in binary format

XNOR CORE

Binarized model input

Reference algorithm

HARDWARE PLATFORMS

Intel

ARM

NVIDIA

CUSTOM INTEGRATIONS

XNOR.AI deep learning Deep learning platform for model creation and optimization

XNOR.AI CORE

A proprietary code base that automatically translates binarized deep learning models into highly optimized programs that run on the leading hardware platforms

Product integrations for devices

Custom work done as part of our services agreement

Integrate with XNOR.AI

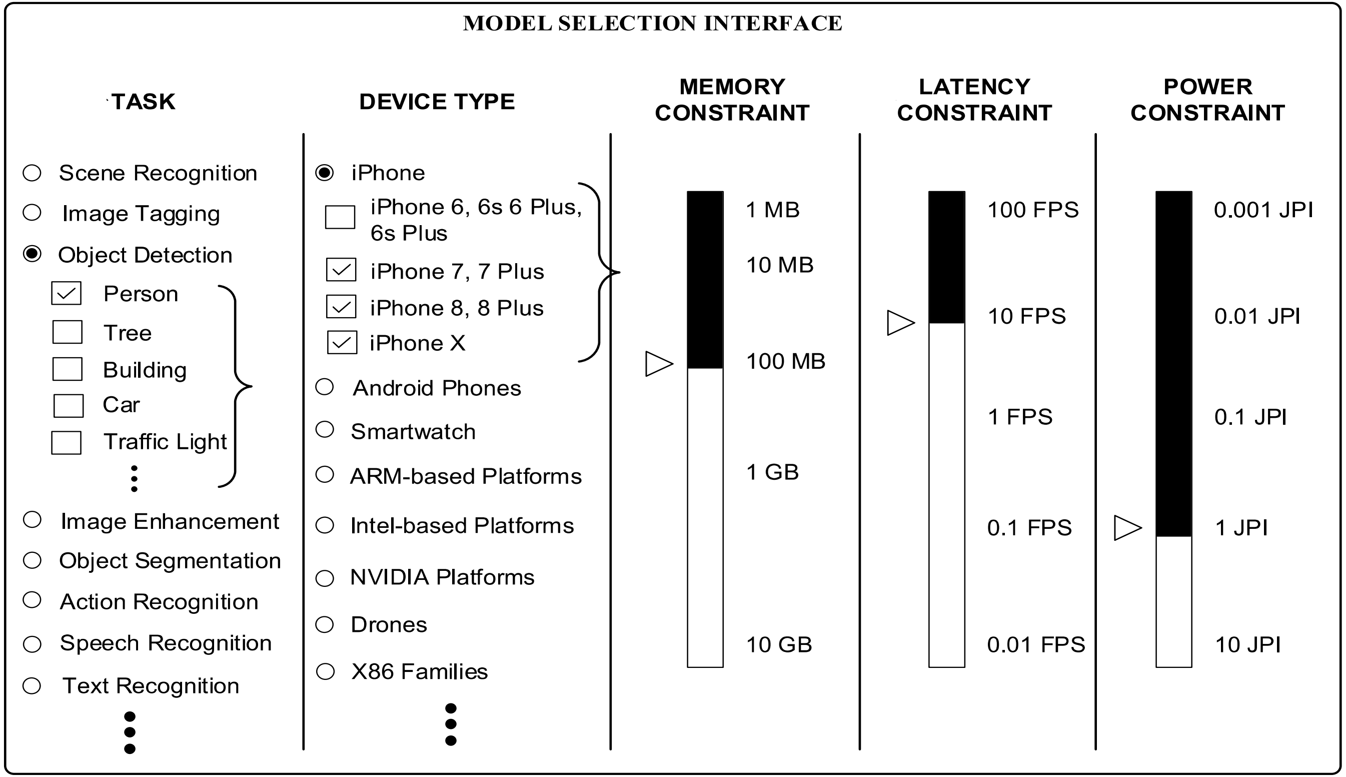
Developers:

AI Everywhere

For Everyone

Integrate with XNOR.AI

Developers Platform



MODEL SELECTION INTERFACE

TASK

- ○ Scene Recognition
- ○ Image Tagging
- ● Object Detection
  - ☑ Person
  - □ Tree
  - □ Building
  - □ Car
  - □ Traffic Light
  - ⋮
- ○ Image Enhancement
- ○ Object Segmentation
- ○ Action Recognition
- ○ Speech Recognition
- ○ Text Recognition
- ⋮

DEVICE TYPE

- ● iPhone
  - □ iPhone 6, 6s 6 Plus, 6s Plus
  - ☑ iPhone 7, 7 Plus
  - ☑ iPhone 8, 8 Plus
  - ☑ iPhone X
- ○ Android Phones
- ○ Smartwatch
- ○ ARM-based Platforms
- ○ Intel-based Platforms
- ○ NVIDIA Platforms
- ○ Drones
- ○ X86 Families
- ⋮

MEMORY CONSTRAINT

LATENCY CONSTRAINT

POWER CONSTRAINT

Application Domains

XNOR.AI technology powers multiple domains

- ✓ Aerospace & Surveillance
- ✓ Driver Assisted Systems
- ✓ Retail
- ✓ Consumer Mobile

Company Background

AI Everywhere

Founded:
2017 by Professor Ali Farhadi
and Dr. Mohammad Rastegari

Intellectual Property:
Highly strategic patent portfolio covering efficient AI at the edge

Press:
New York Times
Tech Crunch

Board members:
Ali Farhadi (CEO), Oren Etzioni (CEO of Allen Institute for AI), Matt McIlwain (Madrona Venture Group)

XNOR Innovations in AI:
XNOR-Net, Yolo, Yolo9000, LCNN, Neural Speed Reading, understanding actions (imSitu), question answering (BiDAF)

Awards:
Best paper at CVPR 2017
CVPR 2016 People’s Choice Award

Thank you !!!

Learn more

Visit our table #809

www.xnor.ai

Mohammad Rastegari | Chief Technology Officer | mohammad@xnor.ai