tinyML® Summit

Miniature dreams can come true…

March 28-30, 2022 | San Francisco Bay Area

TINY ML

www.tinyML.org

SYNTIANT®

Making Edge AI a Reality

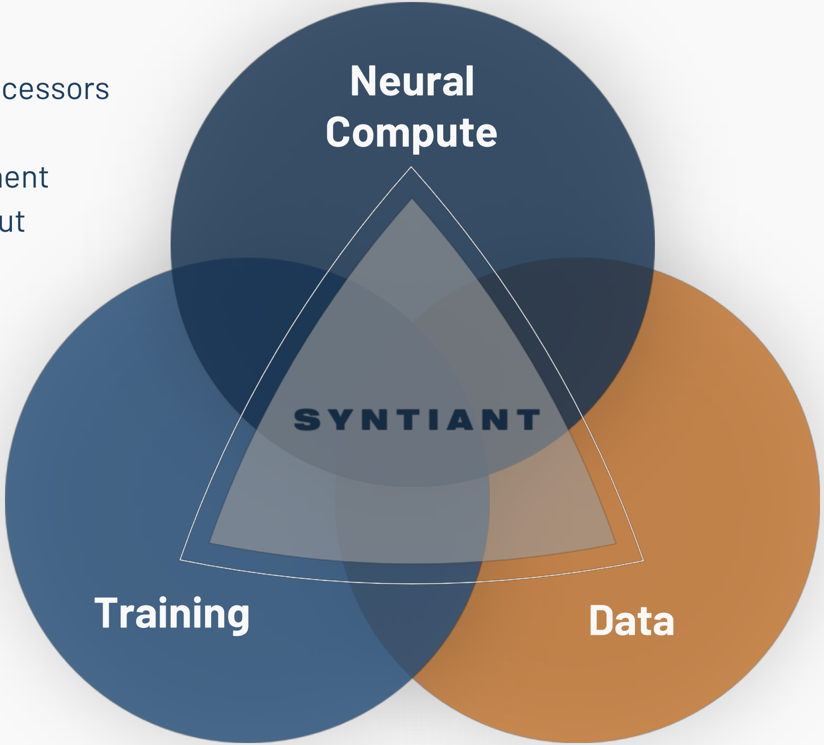
A Complete Edge ML Company
Delivering **Deep Learning** Solutions for
Always-On Devices

Syntiant Complete ML solution

Neural Compute

Syntiant Neural Decision Processors vs. current MCUs:

- ✓ 100x efficiency improvement
- ✓ 30x increase in throughput
- ✓ ½ the die size



Easy to use

All the elements are required to go to market

Training

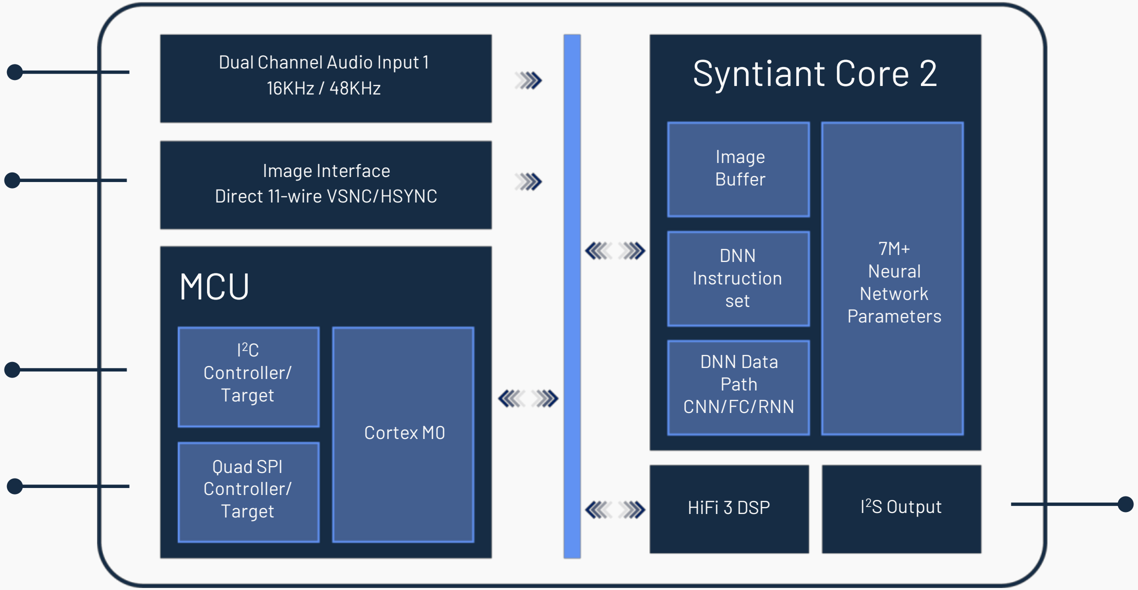
Syntiant Training Pipeline delivers turn-key models for the mass market

Data

Syntiant Data Platform automates the ingestion, labelling, aligning, cleaning, and synthetic data generation to turn raw data into training data sets

SYNTIANT

NDP200 Image, Voice & Sensor Neural Decision Processor



Dual Channel Audio Input 1
16KHz / 48KHz

Image Interface
Direct 11-wire VSNC/HSYNC

MCU

I^2C Controller/ Target

Quad SPI Controller/ Target

Cortex M0

Syntiant Core 2

Image Buffer

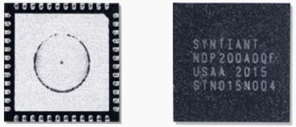
DNN Instruction set

DNN Data Path CNN/FC/RNN

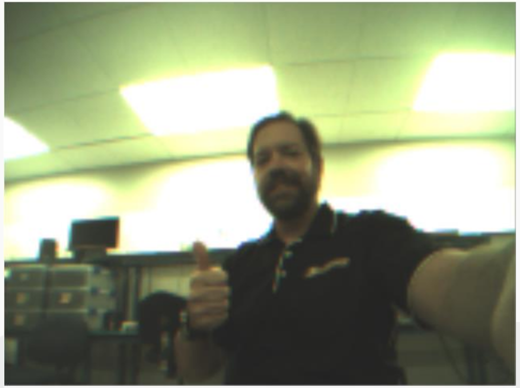
7M+ Neural Network Parameters

HiFi 3 DSP

I^2S Output



5mm x 5mm
QFN40



QQVGA (160x120) color image

Phones / Tablets

Laptops

Smart Speakers

Robotics

Security

Always-on Image processing

MobileNet + sensor in < 1 mW

SYNTIANT

Syntiant Silicon: Neural Decision Processor (NDP)

At-Memory Compute

Tightly coupled memory and MAC functions to minimize data movement

Sustained High MAC Utilization

- Syntiant architecture executes NN every clock cycle achieving over 80% utilization for common NN
- Assures maximum usage of operations per second

Native Neural Network Processing

- No need for intermediary compilers
- Networks trained in TensorFlow, etc are deployed directly to the NDPs.

SYNTIANT

Performance Comparison

	Syntiant Core 2	Arm A53	Syntiant Advantage
Inferences per Million Cycles	1.346	0.0471	~30x
μJ per Inference	166	16131	~100X

Identical MobileNet V1 0.25 int8 network on A53 & NDP200 (Syntiant Core 2)

tinyMLPerf- style test mechanics - - single, identical input vector

Syntiant Core 2 @ 32MHz

=

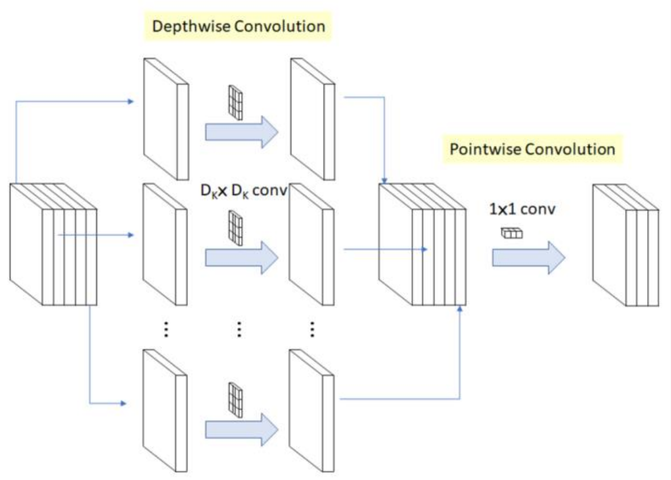
Arm A53 @ 1 GHz

at **1%**

of the energy

If an ARM A53 powered device has a battery life of 3.5 days, the NDP200 powered device has a 1-year battery life running the same neural network.

MobileNetV1



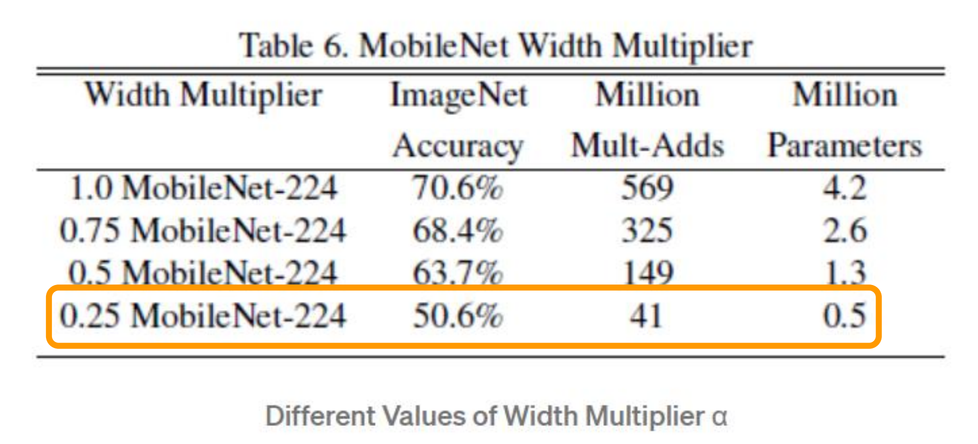


Table 6. MobileNet Width Multiplier

Width Multiplier	ImageNet Accuracy	Million Mult-Adds	Million Parameters
1.0 MobileNet-224	70.6%	569	4.2
0.75 MobileNet-224	68.4%	325	2.6
0.5 MobileNet-224	63.7%	149	1.3
0.25 MobileNet-224	50.6%	41	0.5

Different Values of Width Multiplier α

- Scalable image network architecture for reducing model complexity
  - Depthwise convolutions break complexity
  - Width Multiplier α, resolution Multiplier ρ

- 0.25 MobileNetV1 can easily fit in NDP200
- Under the 640 kB params support
  - Depthwise convolutions break complexity
  - Width Multiplier α, resolution Multiplier ρ

https://towardsdatascience.com/review-mobilenetv1-depthwise-separable-convolution-light-weight-model-a382df364b69
https://arxiv.org/pdf/1704.04861.pdf

MobileNetV1 Style Optimizations

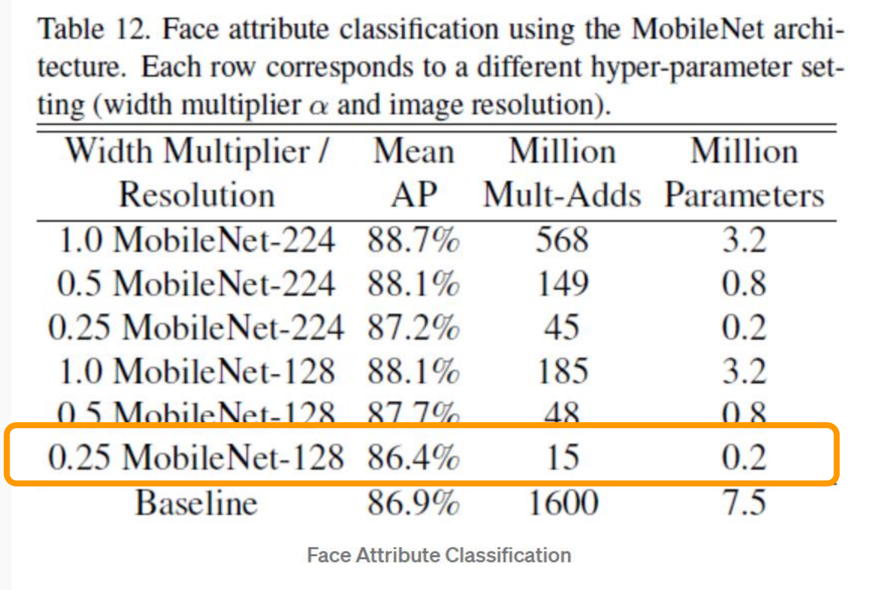


Table 12. Face attribute classification using the MobileNet architecture. Each row corresponds to a different hyper-parameter setting (width multiplier α and image resolution).

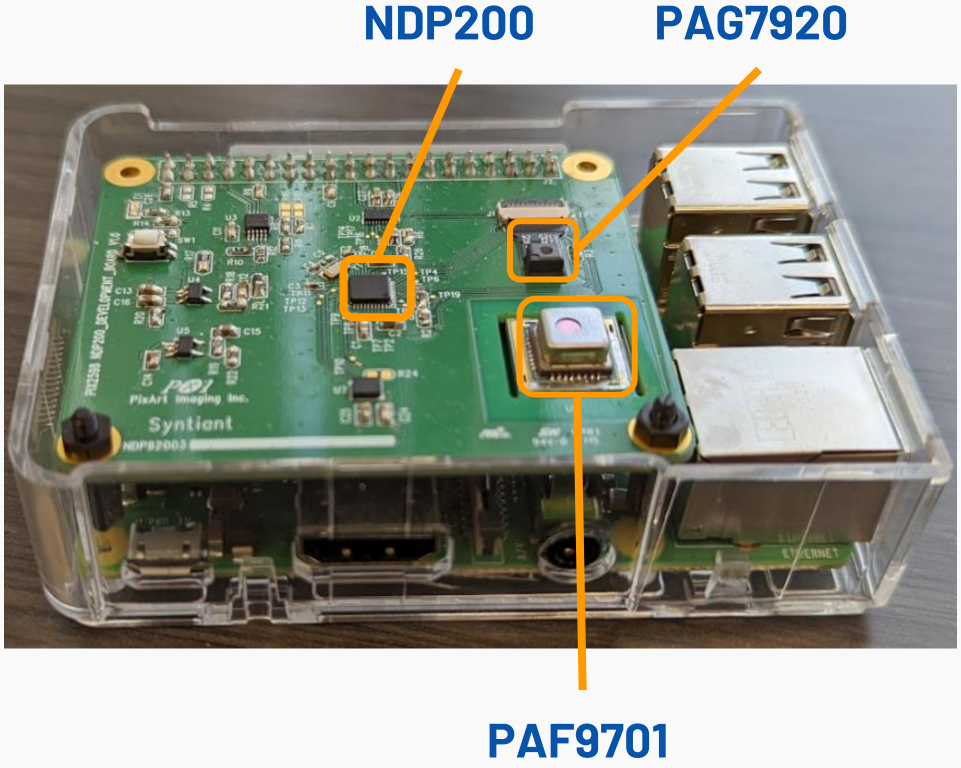
Width Multiplier / Resolution	Mean AP	Million Mult-Adds	Million Parameters
1.0 MobileNet-224	88.7%	568	3.2
0.5 MobileNet-224	88.1%	149	0.8
0.25 MobileNet-224	87.2%	45	0.2
1.0 MobileNet-128	88.1%	185	3.2
0.5 MobileNet-128	87.7%	48	0.8
0.25 MobileNet-128	86.4%	15	0.2
Baseline	86.9%	1600	7.5

Face Attribute Classification

- NDP200 is a flexible architecture that can support tuning the networks
- Example of aggressive model compression in Google's Mobilenet paper - Face attribute classification
  - Shrink to 200k params still with good performance
- Tons of work to tailor optimal solutions for the NDP200 platform
  - MobileNet V2
  - Single-shot multibox detection (SSD)

NDP200 Development Platform

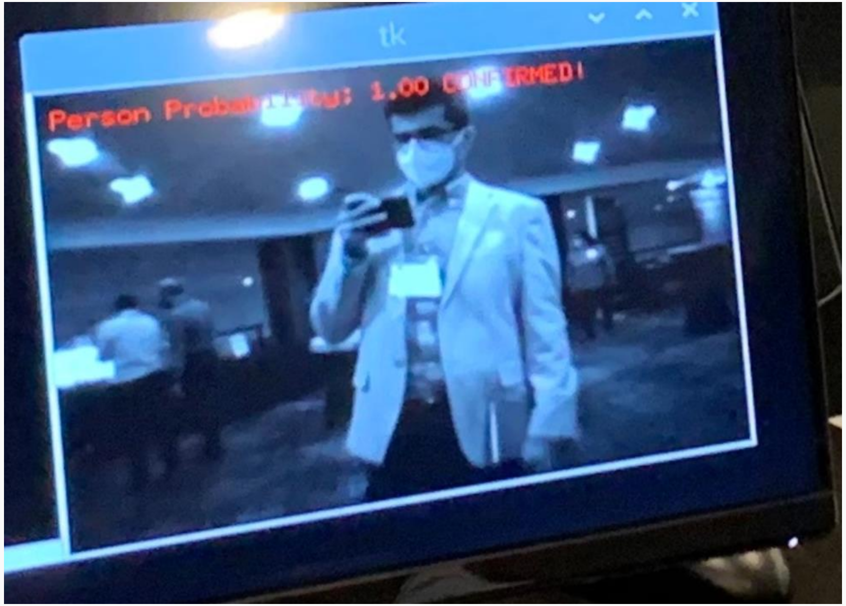
- Jointly developed image development platform with PixArt
  - Raspberry Pi form factor
  - NDP200 for ML solutions
  - 320x240 image sensor (PAG7920)
  - 8x8 thermal sensor (PAF9701)
  - 6-axis accel/gyro BMI sensor (BMI160)
- Enables both capture and model development for Computer Vision (CV) ML model solutions



SYNTIANT

NDP200 Person Detection Demo

- Trained up a MobileNetV1 0.25 person detection
  - Input size 320x240 grayscale
  - First layer was additional decimation to 240x120 image
  - 59-layer Neural network on Core 2
- Used person/non-person images from MS COCO for demo network
  - OpenCV to decimate + grayscale images to match image sensor
- Seeing 90% accuracy on the demo model
  - 4-5 hours of training



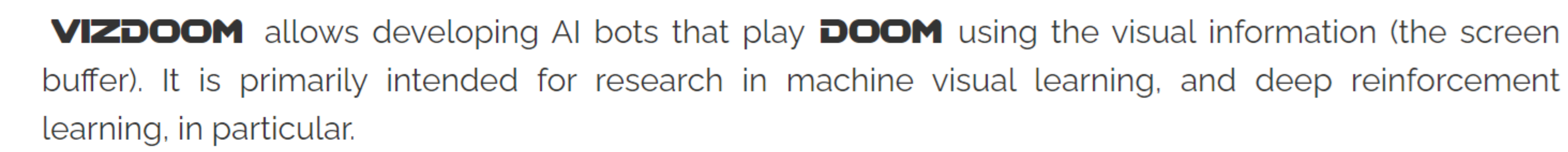
Person Probability: 1.00 CONFIRMED!

SYNTIANT

NDP200 - More than just Object detection

- Can NDP200 do more than just object detection?
- Can it observe and act on the images presented to it?
- **Goal: teach NDP200 to play DOOM!**

Doom-based AI Research Platform for Reinforcement Learning from Raw Visual Information

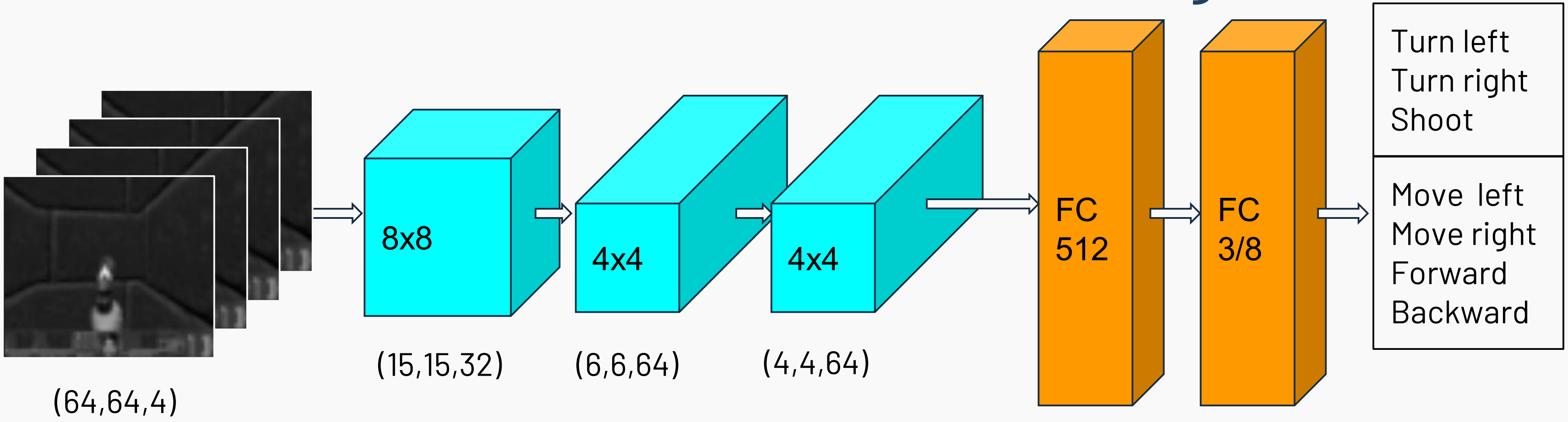


VIZDOOM allows developing AI bots that play **DOOM** using the visual information (the screen buffer). It is primarily intended for research in machine visual learning, and deep reinforcement learning, in particular.

SYNTIANT

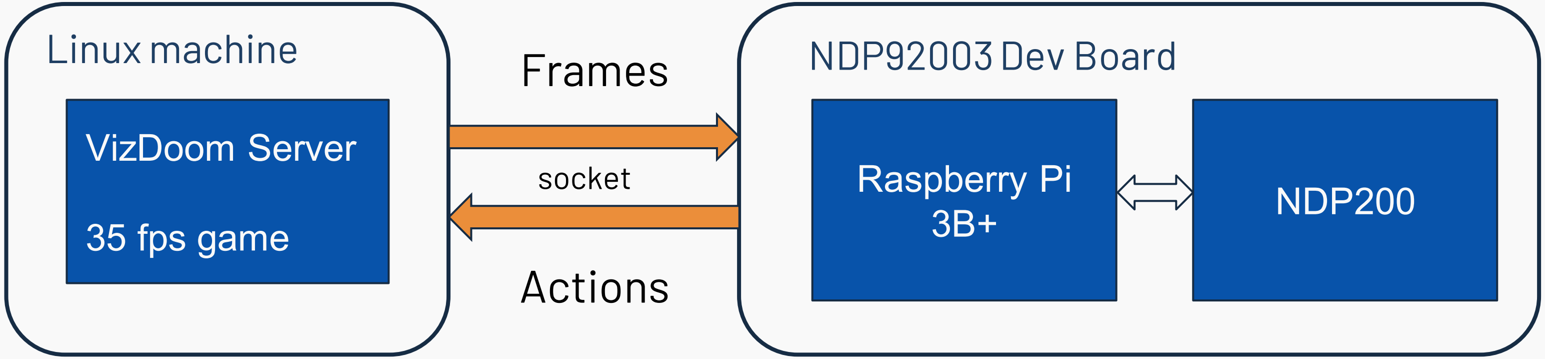
http://vizdoom.cs.put.edu.pl/

NDP200 Platform - Reinforcement Learning



- Using the VizDoom platform, able to train an NDP200 using reinforcement learning to play "DOOM"
  - Reward is killing a monster, only limited time + ammo
  - Training on Syntiant's TDK environment (Linux), inference running on the NDP200
- Convert frames into 64x64 grayscale (4 frames in time series) to the network
  - 3 Convolutional layers into 2 Dense – 606k params

Running VizDoom on the NDP200



Linux machine

VizDoom Server

35 fps game

Frames

socket

Actions

NDP92003 Dev Board

Raspberry Pi 3B+

NDP200

5 layer Deep Neural Network
606k params
1 mW @ 35 fps

15,000x improvement over my 15W brain

NDP200 Platform - Reinforcement Learning

- Start with no knowledge of the game mechanics
- Play thousands of games with rewards for killing monsters
- Learns to shoot and turn around 360 looking for approaching monsters

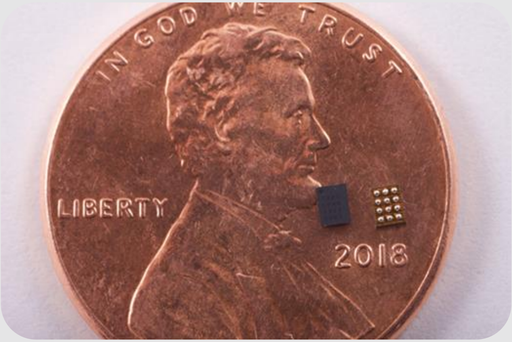
NDP200 Platform - Reinforcement Learning

- Deadly corridor expands to 7 actions (motion)
- CANNOT get past without killing all the monsters
- Roughly 3-days of training, almost the same network
  - 8 output classes

Conclusion

- NDP200 is ready for Computer Vision (CV) applications at the edge in < 1 mW
  - Always-on domain
- NDP200 Development platform is ready to explore CV and sensor-fusion applications
  - Supports data collection (to help calibrate images), and real inference for development
- More than just simple object-detection, NDP200 can use ML to observe and control in the real world

Thank you



David Garrett

Chief Architect, SVP Engineering

info@syntiant.com

www.syntiant.com

SYNTIANT

TINY ML SUMMIT

tinyML Summit 2022 Sponsors

ANALOG DEVICES
AHEAD OF WHAT'S POSSIBLE™

AONdevices

arm

ASPINITY

brainchip
The Neuromorphic Computing Company

CEVA®

Deeplite

EDGE IMPULSE

emza
visual sense

FotaHub

GREENWAVES
TECHNOLOGIES

Grovety Inc.

Himax

HOTG

imagimob

infineon

itemis

KLIKA TECH
GLOBAL IOT SOLUTIONS

LatentAI

LATTICE
SEMICONDUCTOR

Micro.ai

OmniML

NXP

PXI

Plumerai

PROPHESEE

Qeexo

Qualcomm

Rackner

RealityAI®
Engineering Solutions for the Edge

REEXEN
technology

RENESAS

SAP

seeed studio
The IoT Hardware Enabler

SensiML™

Sony
Semiconductor
Solutions
Corporation

ST
life.augmented

SA STREAM ANALYZE

synaptics®

SynSense

SYNTIANT

Tensil.ai

TensorFlow

XMOS

TINY ML SUMMIT

Copyright Notice