HOW TO COOK EMBEDDED COMPUTER VISION. PRODUCT VS TECHNOLOGY.

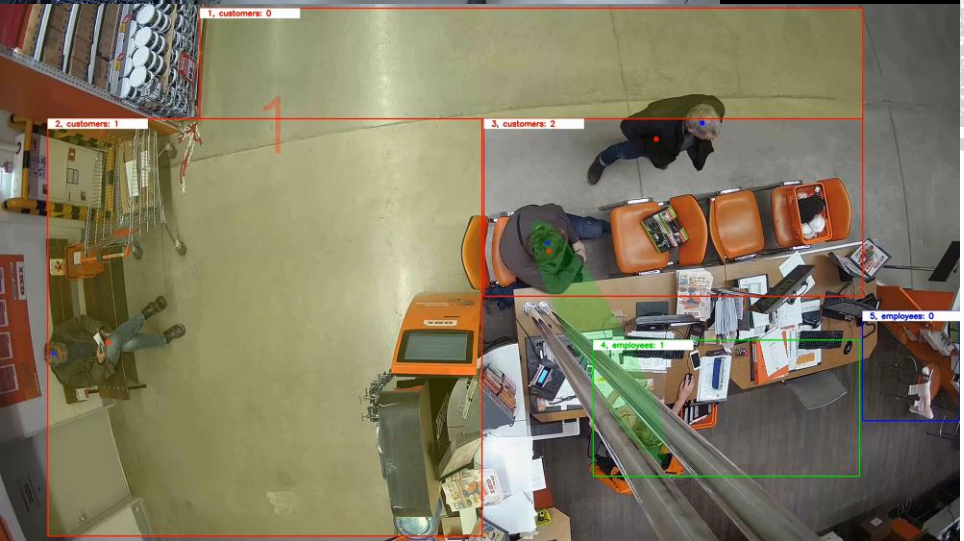
Roman Storchak

CTO @ DatAI

Assumptions:

Higher customer satisfaction ->
Higher sales ->
Higher revenue

Recent events

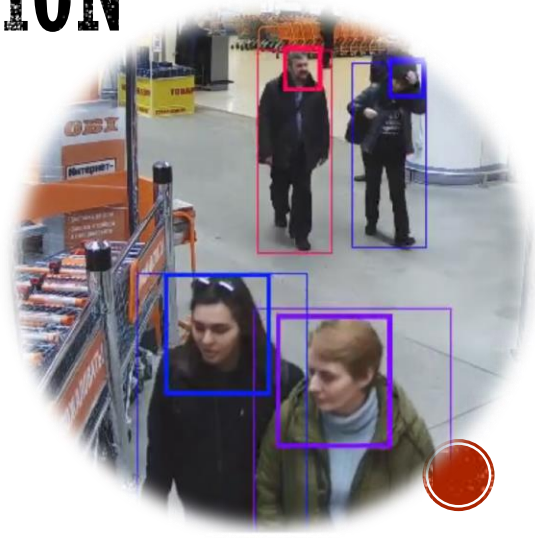


RETAIL USE CASES

- Traffic counting
- Queue analytics
- Sales Zone monitoring
- …
- …
- Targeting Ads
- Sales terminal recommenders

LET'S BUILD TRAFFIC COUNTING

STAGE 1: OBJECT DETECTION



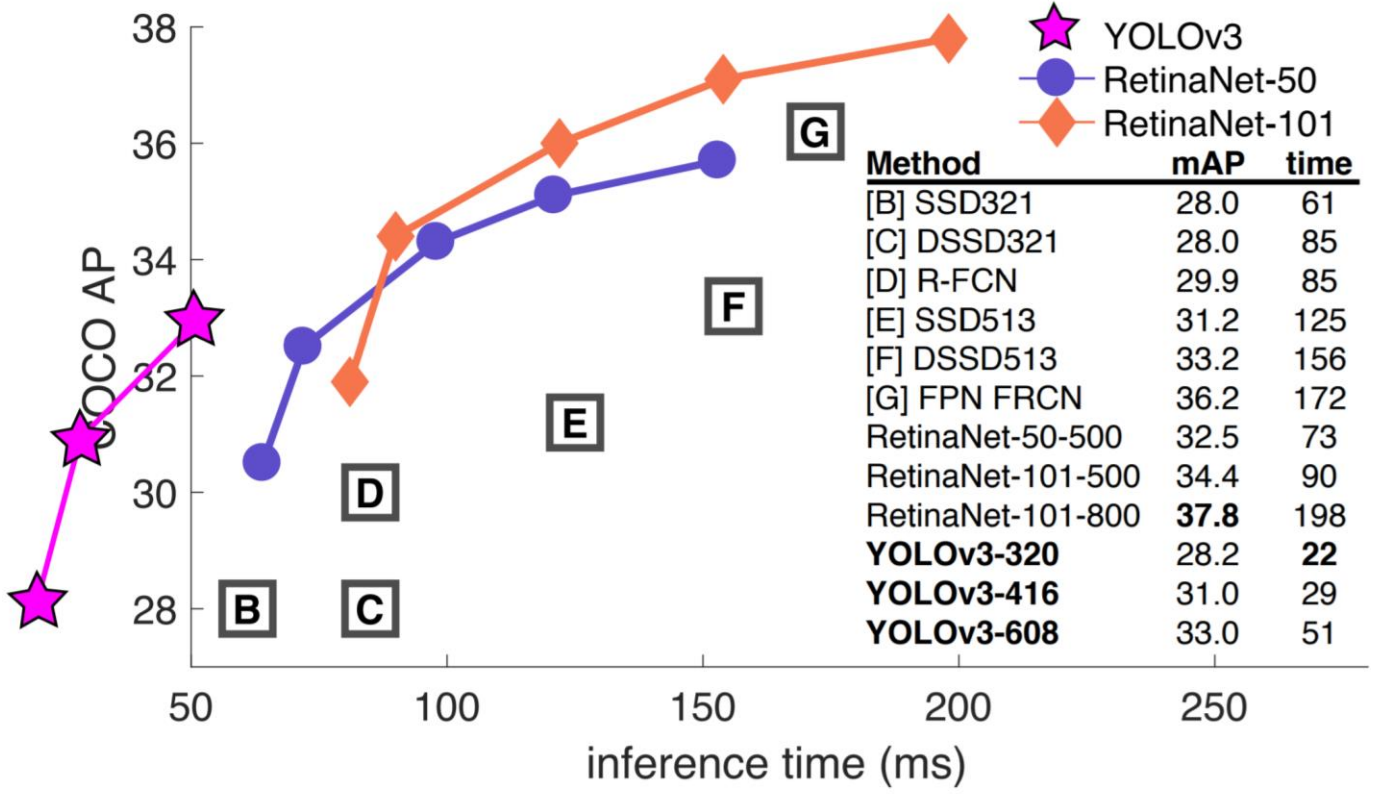
STAGE 1: OBJECT DETECTION / DATASET

- **Object segmentation**
- **Recognition in context**
- **Superpixel stuff segmentation**
- **330K images (>200K labeled)**
- **1.5 million object instances**
- **80 object categories**
- **91 stuff categories**
- **5 captions per image**
- **250,000 people with keypoints**

STAGE 1: OBJECT DETECTION / MODEL

Approach	Model	Comment
Traditional detectors	VJ (2001), HOG(2005)	
Two-stage detectors	R-CNN (2014), Fast RCNN, Faster RCNN	Region proposal
One-stage detector	Yolo (2016), SSD(2016), Retina-Net	NMS, Hard neg. mining

STAGE 1: OBJECT DETECTION / MODEL

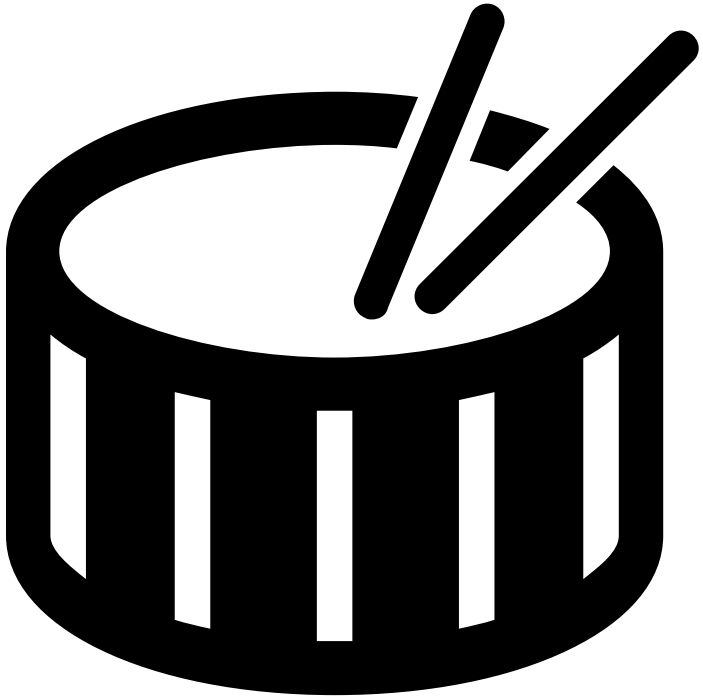


Method	mAP	time
[B] SSD321	28.0	61
[C] DSSD321	28.0	85
[D] R-FCN	29.9	85
[E] SSD513	31.2	125
[F] DSSD513	33.2	156
[G] FPN FRCN	36.2	172
RetinaNet-50-500	32.5	73
RetinaNet-101-500	34.4	90
RetinaNet-101-800	**37.8**	198
YOLOv3-320	28.2	**22**
YOLOv3-416	31.0	29
YOLOv3-608	33.0	51

YOLOv3: An Incremental Improvement
Joseph Redmon, Ali Farhadi

YOLO V3 MAKES SENSE

+ JETSON NANO



YOLO V3 - 2 FPS

HARDWARE LIMITS AI MORE THAN SOFTWARE

STAGE 2: PLATFORM SELECTION

- GOOGLE Coral edge TPU
- Nvidia Jetson Nano / TX2
- …
- Intel NCS2 (Intel® Movidius™ Myriad™ X Vision Processing Unit)
- ARM (ARM NN SDK, includes MALI GPU)
- …
- Xilinx (Out of scope)
- GPU (Nvidia or AMD, out of scope)
- CPU (OpenVino, out of scope)

STAGE 2: PLATFORM SELECTION / CAMERA SELECTION

- Compressed vs Uncompressed video
- Optical aberrations
- Rolling vs global shutter

STAGE 2: PLATFORM SELECTION / H.264 TO RGB CONVERSION

Coral beta

Tech specs

CPU	NXP i.MX 8M SoC (quad Cortex-A53, Cortex-M4F)
GPU	Integrated GC7000 Lite Graphics
ML accelerator	Google Edge TPU coprocessor
RAM	1 GB LPDDR4
Flash memory	8 GB eMMC

GC7000 Series	GC7000 UltraLite	GC7000 Lite
Vega Shader Cores*	8	16
Shader Clock Frequency in 28HPM	1GHz	1GHz
GFLOPS (Medium Precision)	32	64
GFLOPS (High Precision)	16	32
Texel Rate (GTexel/sec)	0.8	1.6
Vertex Rate (Gvert/sec)	0.5	1.0
OpenGL ES 3.1 Support	✓	✓
OpenCL 1.2 Support	✓	✓
Geometry Shader/Tessellation Shader	Optional	Optional

STAGE 2: PLATFORM SELECTION / H.264 TO RGB CONVERSION

- Hardware accelerated decoding rarely works off the shelf
- CPU load on google coral drops 10x (50%->5%)
- Use zero-copy when possible (Jetsons)
- Decoder output is often YUV (for example I420) or NV12
- Hardware acceleration for scaling and colour conversion

STAGE 2: PLATFORM SELECTION / HARDWARE FOR NN

- Production grade hardware rarely works with float32
- You have to decide if quantization aware training is ok for you.
- Use NN optimizations (tensorRT, openVino, …)
- Static input tensors

STAGE 2: PLATFORM SELECTION / HARDWARE FOR NN

Object detection

MobileNet SSD v1 (COCO)

Detects the location of 90 types objects
Dataset: COCO
Input size: 300x300

↓ Edge TPU model

↓ Labels file

↓ All model files

MobileNet SSD v2 (COCO)

Detects the location of 90 types objects
Dataset: COCO
Input size: 300x300

↓ Edge TPU model

↓ Labels file

↓ All model files

MobileNet SSD v2 (Faces)

Detects the location of human faces
Dataset: Open Images v4
Input size: 320x320
(Does not require a labels file)

↓ Edge TPU model

↓ All model files

OH WAIT! WHICH PROBLEM ARE WE SOLVING

- Beck-end & Front-end
- API for integration with client's IT infrastructure

Recent events

Someone's else tech

vs

MobileNET 2+

IOU tracker +

Google Coral edge TPU

Cutting-Edge Hardware

To achieve unparalleled accuracy, the Density DPU pairs Class 1 infrared lasers with onboard compute power for running proprietary machine learning algorithms.

Download the Datasheet

The World's Most Ad
People Counter

Accurately count people in real tim
infringing on occupant privacy

Get a Live Demo

Designed for Enterp

- Standard PoE+ connectiv
- Easy to install and configu
- 24/7 dedicated support
- Scales to 100s of devices

Funding

AMOUNT RAISED	FUNDED OVER
$16 200 000	3 rounds

16 M$

Rounds

A $12 000 000

Series A Jun 2018

A $4 000 000

Series A Jul 2016 Read press

A $200 000

Series A Feb 2016

BARRIERS TO ENTRY

Barier	Easiness
Dataset	EASY
Model	EASY
…	…

BARRIERS TO ENTRY

Barier	Easiness
Dataset	EASY
Model	EASY
Optimization	MODERATE
…	…

WE HAVE MISSED FEW THINGS

- Legal
- Supply availability
- Manufacturing
- QA
- Monitoring
- **Unit economy**

BARRIERS TO ENTRY

Barier	Easiness
Dataset	EASY
Model	EASY
Optimization	MODERATE
Sales	Extremely difficult

Porter's Five Forces Framework

Threat of new entry

- Time and cost to entry
- Economy of scale
- Specialist knowledge
- Technology protection
- Cost advantages
- Barriers to entry

Competitor Rivalry

- Number of competitors
- quality and other differences
- Switching costs
- Customer loyalty
- Costs of leaving the market

Bargaining power of suppliers

- Number of suppliers
- Size of suppliers
- Uniqueness of the service
- your Ability to substitute
- Cost of changing

Rivalry between existing competitors

Bargaining power of buyers

- Number of customers
- Size of each order
- Difference between competitors
- Price sensitivity
- Ability to substitute
- Cost of changing

Threat of Substitutes

- Substitute performance
- Cost of change

WHEN AI CAN BE A CORE OF SUCCESSFUL PRODUCT?

DEMAND +

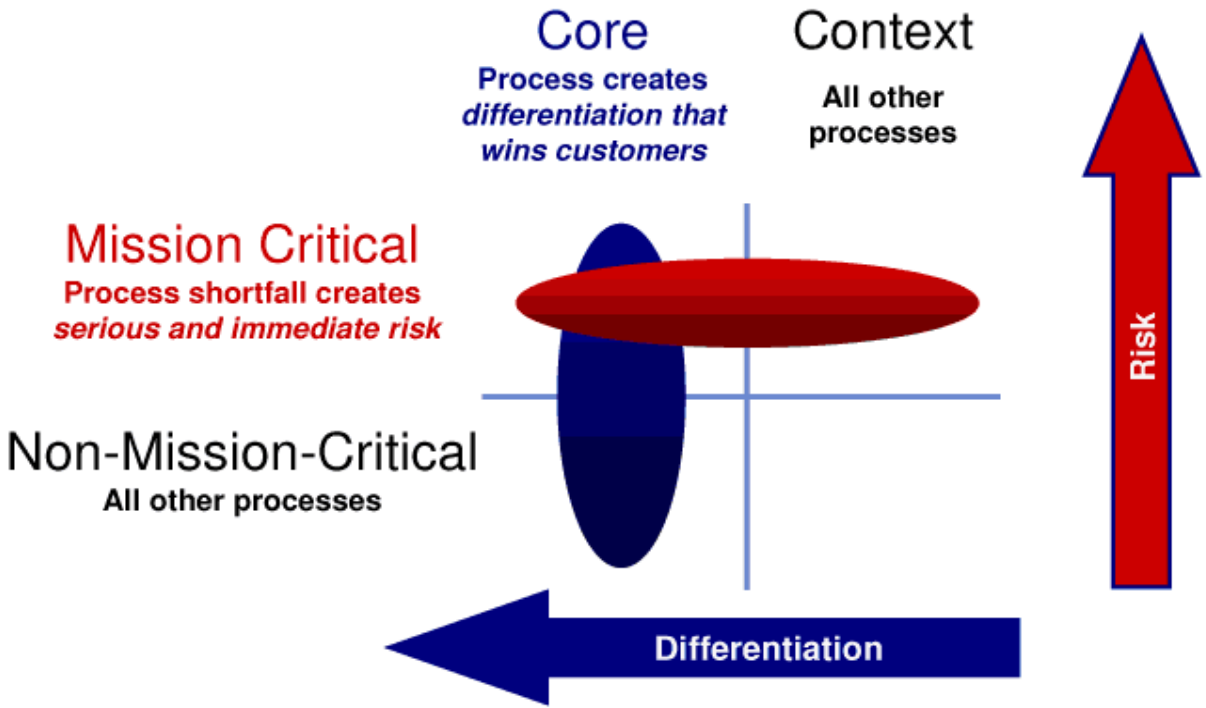
Proprietary **and** unique technology

or

Proprietary dataset

Core/Context Analysis Framework

What are we really doing?



Embedded AI is simple to start

And it's easy to miss core & mission critical processes

QUESTIONS?