2020 embedded VISION summit®

Lessons Learned from the Deployment of Deep Learning Applications in Edge Devices

Orr Danon, CEO

September, 2020

HAILO
Empowering Intelligence

Video Analytics Platforms

Deep Learning for Vision … Domain Convergence?

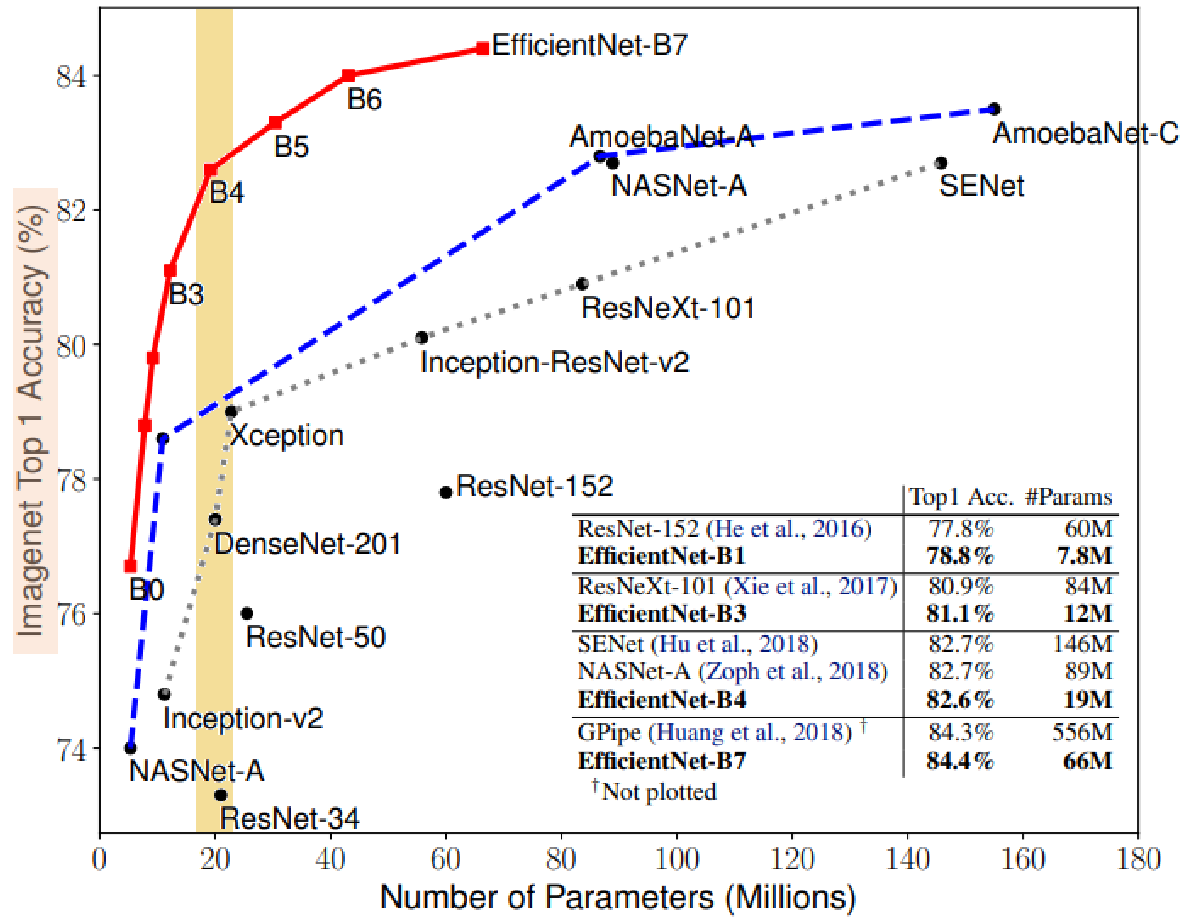
Nearly a decade of intense research activity in this domain

Yielding major achievements in the performance vs. accuracy trade-off

Recent survey activity (e.g. Google's NAS) clearly indicate this model convergence trend (in terms of model size; no. of operations and accuracy)

Can we derive hardware requirements for DL-based vision processing?

Accuracy vs. Model Size



	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**78.8%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.1%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**82.6%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.4%**	**66M**

† Not plotted

Source: Google EfficientNet (1905.11946)

~20M

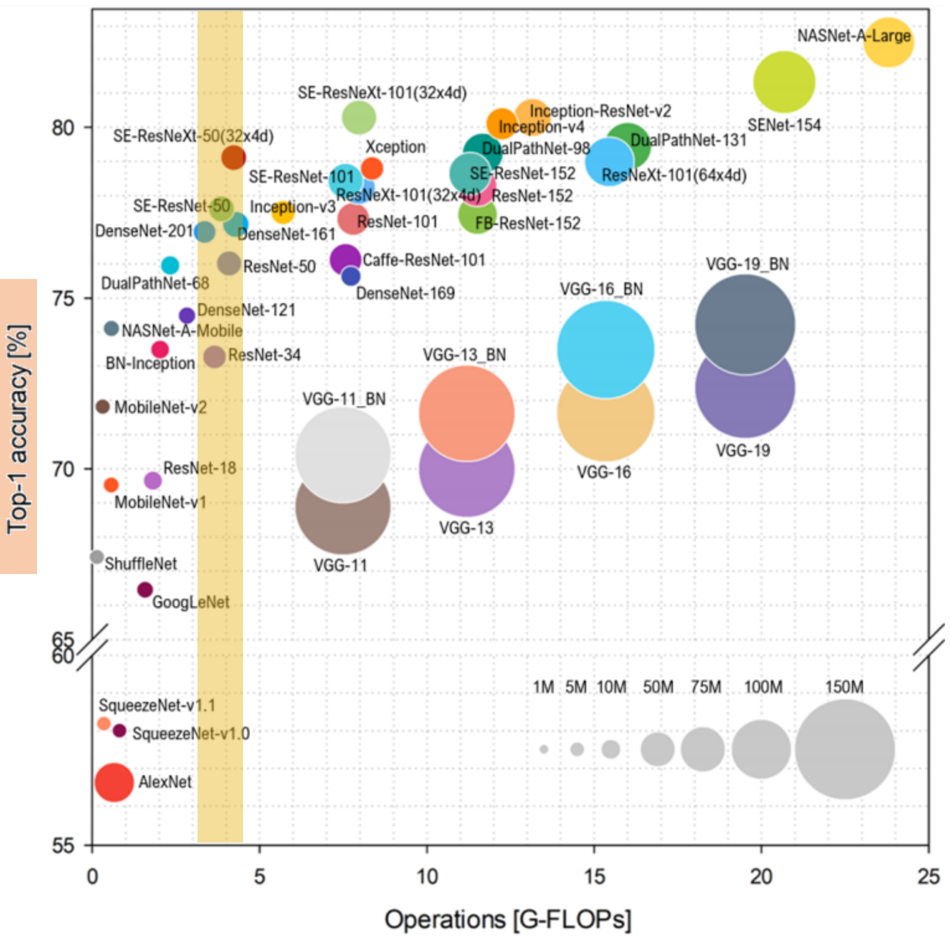
Model-size knee-point

~86%

Asymptotic Classification Accuracy

HAILO

Accuracy vs. Compute Capacity



[Benchmark Analysis of Representative DNNs]

~80K

MACs per pixel knee-point

~86%

Asymptotic Classification Accuracy

Video Analytics: System-level Challenges – Energy

How low **CAN** we go?

Driven by required performance

Determined by processing efficiency

How low **SHOULD** we go?

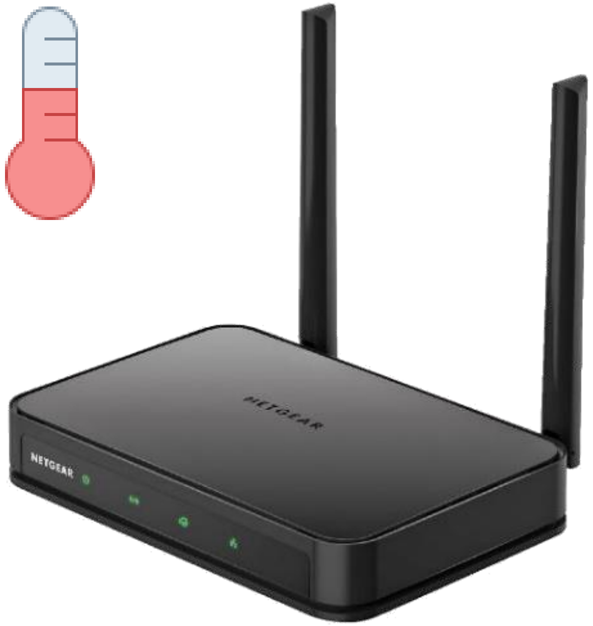
Driven by what's "good enough"

Determined by platform's power budget

Video Analytics: Upper bound (I) – Heat Constraint

Exposed box : Case cannot go >10% above human-body temperature (References: ASTM C1055; ISO 13732)	**Limits temperature based on thermal resistance**	**Results in a limited operational capacity**

Example	
Ambient Temperature	**25 °C**
Case temperature	**40 °C**
Junction Temperature (typ.)	**85 °C**
Junction-Case Thermal Resistance (typ.)	**9 °C/W**
Max allowable dissipation	**5 W**
Neural net @ 2xFHD; 30fps (or 8xVGA)	**10 TOPS**
Required efficiency	**2 TOPS/W**



Video Analytics: Upper bound (II) – Power Constraint

Allow adding AI-capability without breaking product boundaries	For instance, keep well within adapter rating

Example	
Typical adapter rating	**2A @ 5V**
Headroom, conversion loss	**25%**
Total consumption	**<7.5W**
Current content (e.g. sensor, CPU, modem)	**5W**
Budget for new content	**2.5 W**
Neural net @ FHD; 30fps	**5 TOPS**
Required efficiency	**2 TOPS/W**

HAILO

`

Industrial Machines

Industry 4.0 – A revolution in the making

Industry 1.0

Mechanization and the introduction of steam and water power

Industry 2.0

Mass production assembly lines using electrical power

Industry 3.0

Automated production, computers, IT-systems and robotics

Industry 4.0

The Smart Factory, Autonomous systems, IoT, machine learning

Source: https://www.spectralengines.com/articles/industry-4-0-and-how-smart-sensors-make-the-difference

Industry 4.0: Passive to Active

Greater **Automation**
Tighter **process control**

Higher **operational efficiency**
Down-time reduction

Increased **productivity**

~8.5%

Automation Control CAGR (2018-2026)

~14%

Industrial Robots CAGR (2018-2026)

https://loupventures.com/industrial-robotics-outlook-2025/ ; https://www.smart2zero.com/news/machine-vision-market-shows-12-cagr

Industry 4.0: Typical Flow & Node Capability

Sense → Analyze → Act

Benefits

↑	Agility	Up-time	Privacy
↓	Latency	Energy	Comm. bandwidth

Affordable Sensing + **Intelligent Processing** + **Precise Actuation**

Better visibility at the edge

Source: ABI Research, "Business for IIOT edge intelligence"

Case study: Pick & Place Machine (1/2)

Assembly line pick-and-place machine

Assembly time is dominated by the perception speed

Goal: Minimize assembly time

- Fastest pick-and-place
- 100% parts placed (guaranteed)

Source: https://youtu.be/IfojHo9cVOk

Case study: Pick & Place Machine (2/2)

Processing Pipeline

Sense
- Frame grab

Analyze I
- Detection
- Quality classification
- Grip orientation

Analyze II
- Decision Logic

Act
- Pick & place

Example for latency budget & performance derivation

Target to meet 50 ops/s	**20 ms**
Frame grabber	5 ms
Decision logic	5 ms
Pick & place	5 ms
Available latency for all perception tasks	**5 ms**

Implied frame rates are 100s to 1,000s fps (depending on meta-architecture)

Smart City

The city of the future

2020 embedded VISION summit

Highest potential benefit from AI-based computer vision

Government Efficiency
Governance effectiveness and ease of doing business

Sustainability
Environmental, energy, water and air quality Management

Health and Wellness
Mental, physical and social care and well being

Mobility
Transportation, transit and traffic management

Economic Development
Business, employment and productivity

Public Safety
Welfare and protection from Crime, hazards, and disasters

Quality Of Life
Standard of living, satisfaction and happiness

Source: https://strategyofthings.io/

HAILO

Scale challenge – In numbers

Large number of cameras

Complex data centers

Motivations for decentralization

- Privacy
- Bandwidth
- Scaling (lower infrastructure overhead)

City	Cameras [#]	Per-1000ppl	Per-km^2
London	600K	68	380
Chicago	35K	13	58
Sydney	60K	12	4.85
Shanghai	3M	113	473
Berlin	30K	11	34

Cross-road Coverage – A case study

Platform

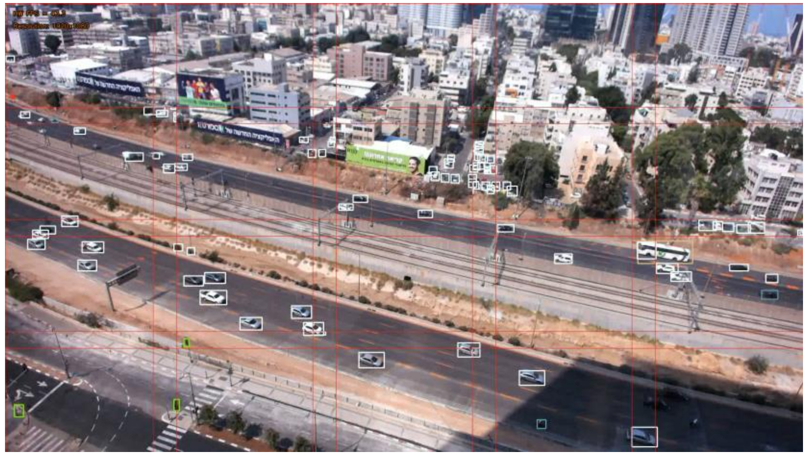
- Intelligent camera monitoring a freeway segment

Applications

- Congestion monitoring
- Policing

Technical Requirements

- Monitor the whole segment of road → camera position
- Identify cars, trucks and pedestrians → resolution
- Track all cars at max traffic density → # of objects
- Track all cars at their max speed → frame rate



Cross-road Coverage – Cont.

Targeting minimal number of cameras

Object footprint

- Cars – 2m
- Pedestrians – 0.5m

Coverage limited to freeway

- ~2MPix per frame
- Can be tiled and trained on lower-res

Example – Position and resolution	
Installation height	**60m**
Coverage distance	**150m**
Camera horizontal FOV	**70°**
Road segment length	**210m**
Min. object footprint for NN detection	8 pix
Min. object width	0.5 m
Min. Camera resolution	**3200 pix**

Cross-road Coverage – Cont.

Method
- Identify all objects
- Avoid overlap between frames for tracking

Object speed
- Cars (max) – 33 m/s (120kmph)
- Pedestrians –5 m/s

Potential for adjacent areas coverage
- Parking monitoring
- Traffic lights management
- Etc.

Example – Capacity and throughput	
Number of Lanes	8
Max cars per lane	80
Max. number of objects per frame	**640 obj/frame**
Min object distance @ max speed (car)	3 m
Max sample interval	45 msec
Required frame rate per camera	**~22 fps**
• Equivalent to **300x300** frame rate	**~900 fps**

HAILO

Machine learning for **visual perception** is well scoped

Industrial applications require **throughput, latency** & **accuracy**

System are limited by **thermal** and **power** constraints

System processing efficiency required is **>2 TOPS/W**

System processing capacity required is **>10 TOPS**

Processing latency should be **<5 msec**

Closing Remarks

Thank you

https://www.hailo.ai

HAILO
Empowering Intelligence