

- **AI @ FH Technikum Wien**

Bernhard Knapp, Technikum

- **Eye Analysis: Designing Neural Network for the Automotive Industry**

Marc Javin, emotion3d.ai

- **Going Embedded: Real-time Deep Learning for automotive applications**

Georg Braun, emotion3d.ai

Break
Announcements

- **Hot Papers - What's new in NLP?**

Liad Magen, TU-INSO

- **Stealing Models from Compiled DNNs**

René Donner, mva.ai



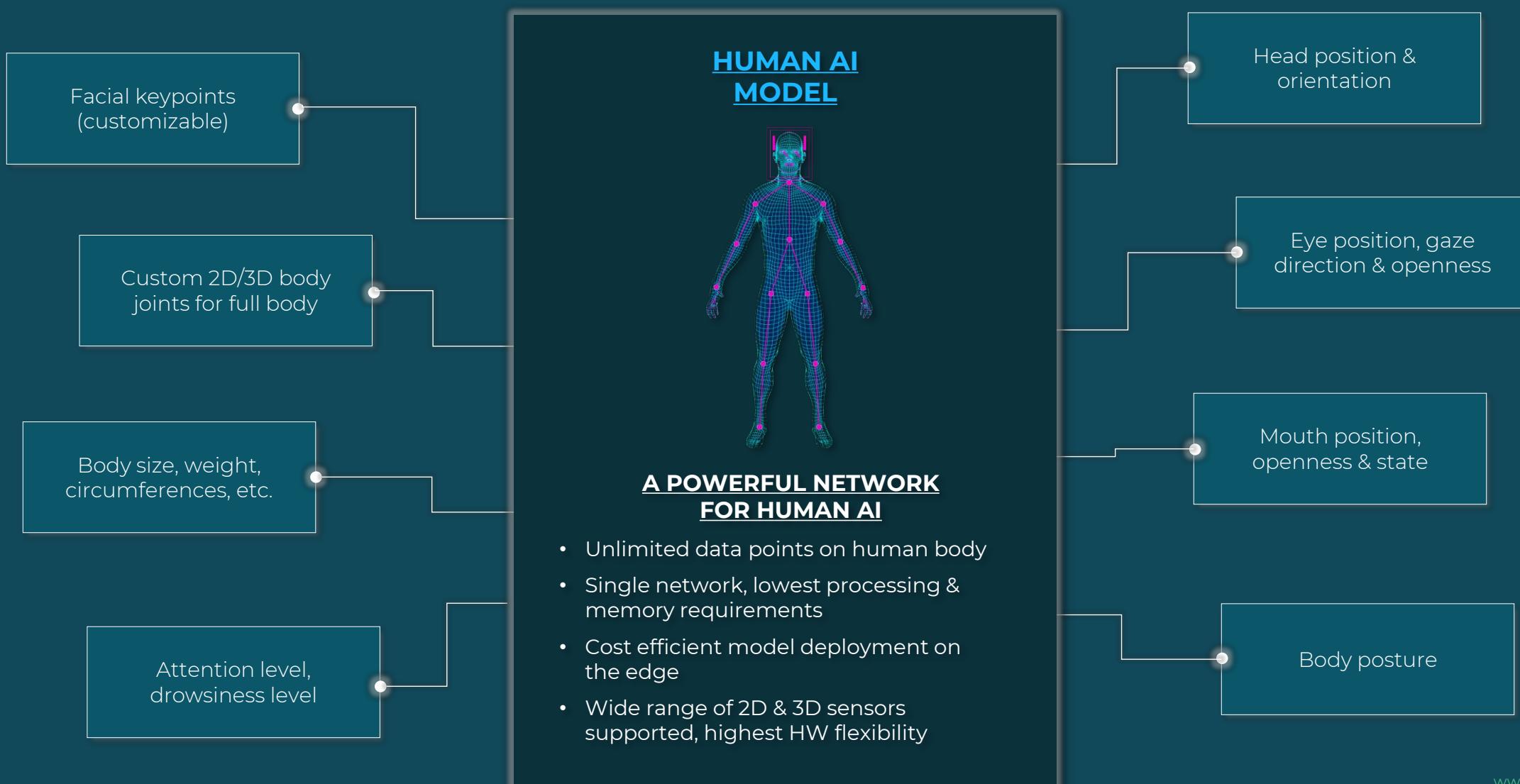
Designing a neural network for the automotive industry

Marc Javin

- **Intro**
- **Model runtime optimization**
- **Data & model quality**
- **Demo**

- **Intro**
- Model runtime optimization
- Data & model quality
- Demo

Our **Human AI model** analysis camera images and provides unlimited data points on human bodies enabling a new level of **human-machine interaction** as well as **supporting** and **protecting humans** in **every situation**.



HUMAN AI AND IN-CABIN ANALYSIS HELP SOLVE SEVERE CHALLENGES IN AUTOMOTIVE

PASSIVE SAFETY

17% higher mortality risk for women
caused by bias of current airbag systems

77% higher injury risk for women
caused by bias of current airbag systems

ADAS not synced with passive safety
leading to fatal consequences

ACTIVE SAFETY

1.350.000 deaths
each year by road accidents

94% of accidents
are caused by human error

30% of driving time
is spent on distracting activities

USER EXPERIENCE

>10% yearly growth
of automotive HMI market

personalization is key
in car buyers' decision making

77% of cars
employ UX that is distracting

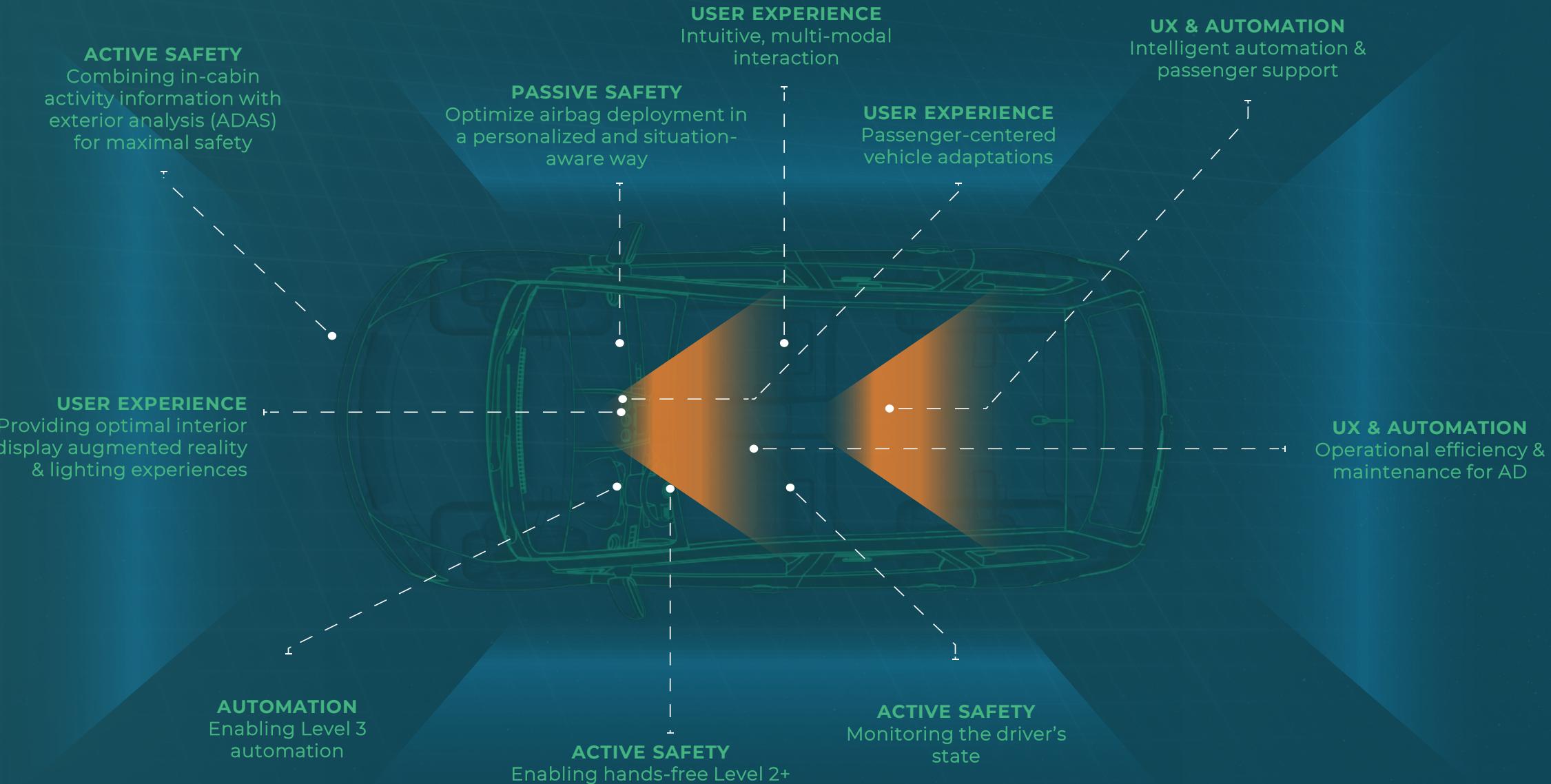
AUTOMATED VEHICLE SAFETY

automated driving
requires new safety concepts

automated fleets
must ensure UX, safety and security

efficiency with in-cabin analysis
automatic cleanliness and safety checks

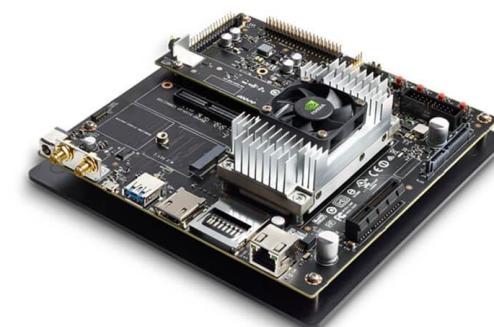
In-cabin analysis systems will be at the heart of future vehicle concepts as they allow seamless interaction and optimal safety for all occupants. They will be based on cameras in the vehicle and understand all occupants and their activities enabling a wide range of safety, UX & automation functions.



- **A DMS system could contain :**

- Head position
- Head orientation
- Keypoints estimation
- Eye gaze
- Eye openness
- Face recognition
- Drowsiness detection
- Smartphone/food/cigarette/drinks detection
- Anti spoofing
- Age, weight estimation
- ...

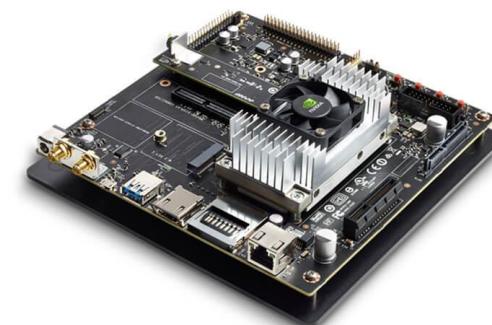
- **All of it should run in real time : ~30 fps**
- **...on an embedded device (~20x slower than GPU)**
- **...with top quality performance**



- **A DMS system could contain :**

- Head position
- Head orientation
- Keypoints estimation
- Eye gaze
- Eye openness
- Face recognition
- Drowsiness detection
- Smartphone/food/cigarette/drinks detection
- Anti spoofing
- Age, weight estimation
- ...

- **All of it should run in real time : ~30 fps**
- **...on an embedded device (~20x slower than GPU)**
- **...with top quality performance**



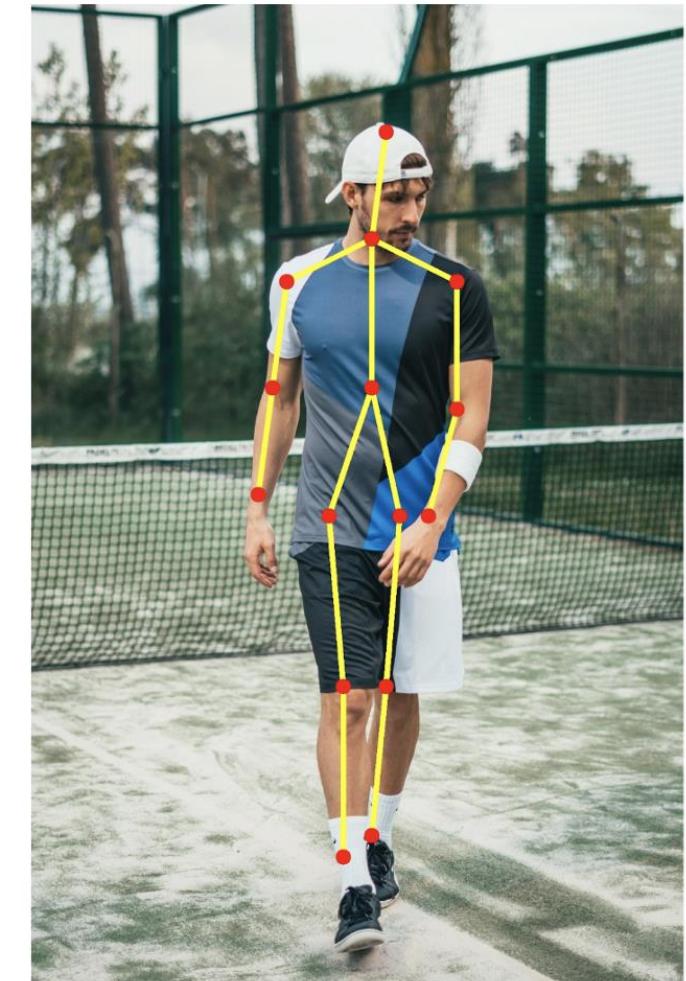
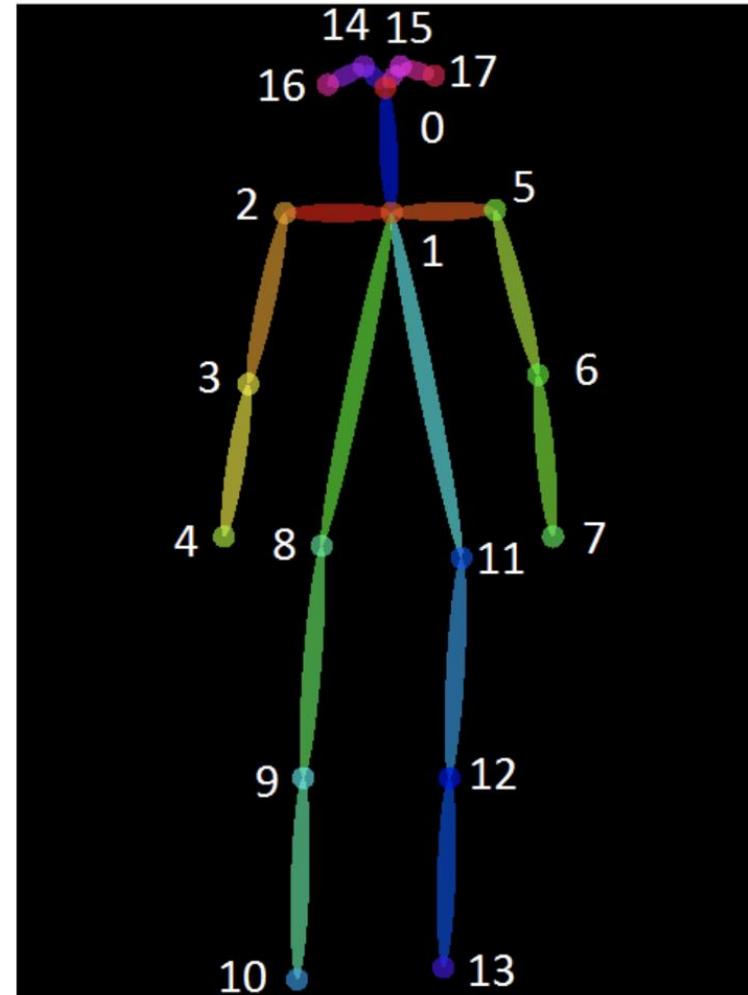
- **A DMS system could contain :**

- Head position
- Head orientation
- Keypoints estimation
- Eye gaze
- Eye openness
- Face recognition
- Drowsiness detection
- Smartphone/food/cigarette/drinks detection
- Anti spoofing
- Age, weight estimation
- ...

- **All of it should run in real time** : ~30 fps

- **On the Jetson TX2 CPU**

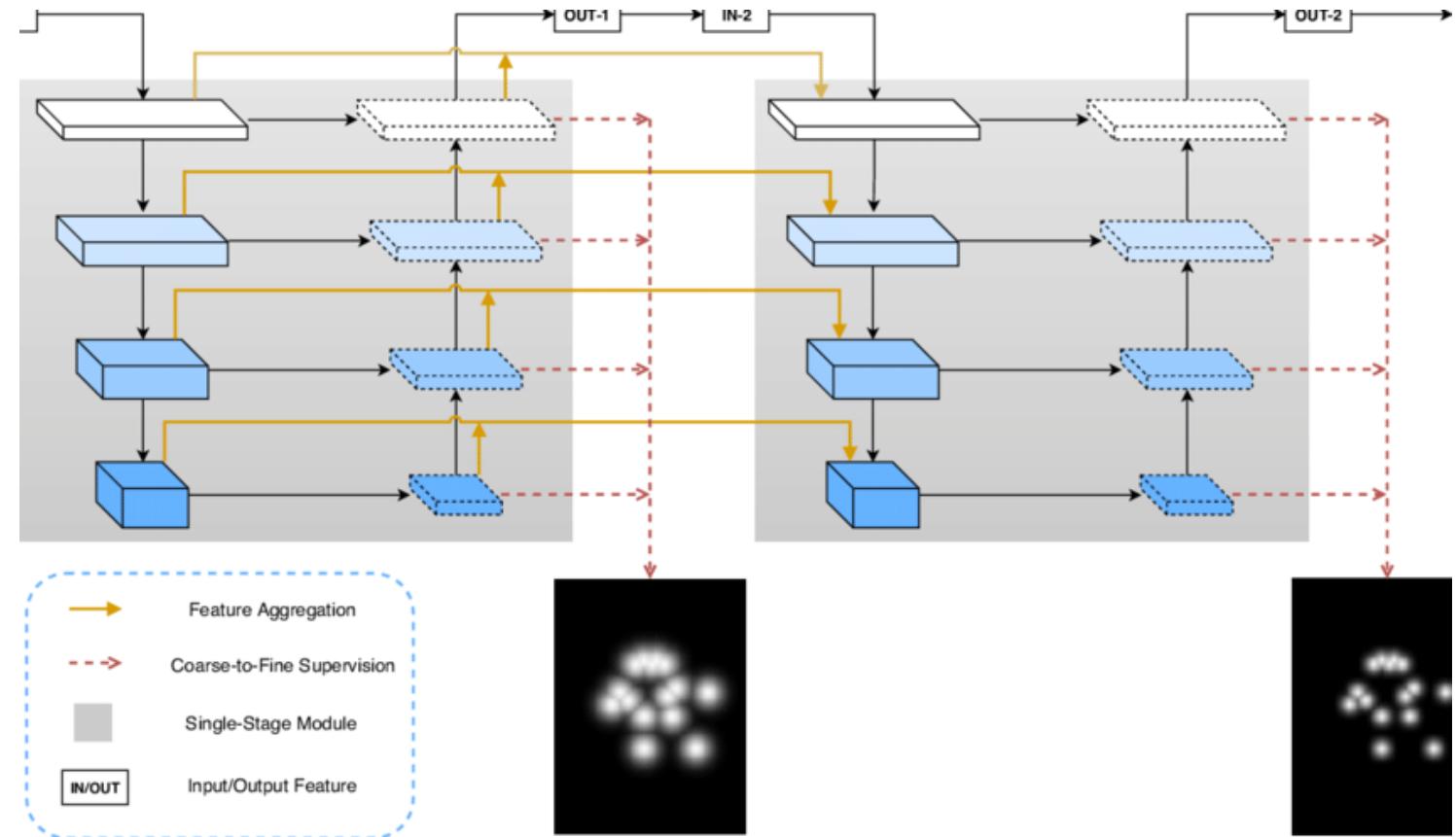
- **...with top quality performance**



- **A DMS system could contain :**

- Head position
- Head orientation
- Keypoints estimation
- Eye gaze
- Eye openness
- Face recognition
- Drowsiness detection
- Smartphone/food/cigarette/drinks detection
- Anti spoofing
- Age, weight estimation
- ...

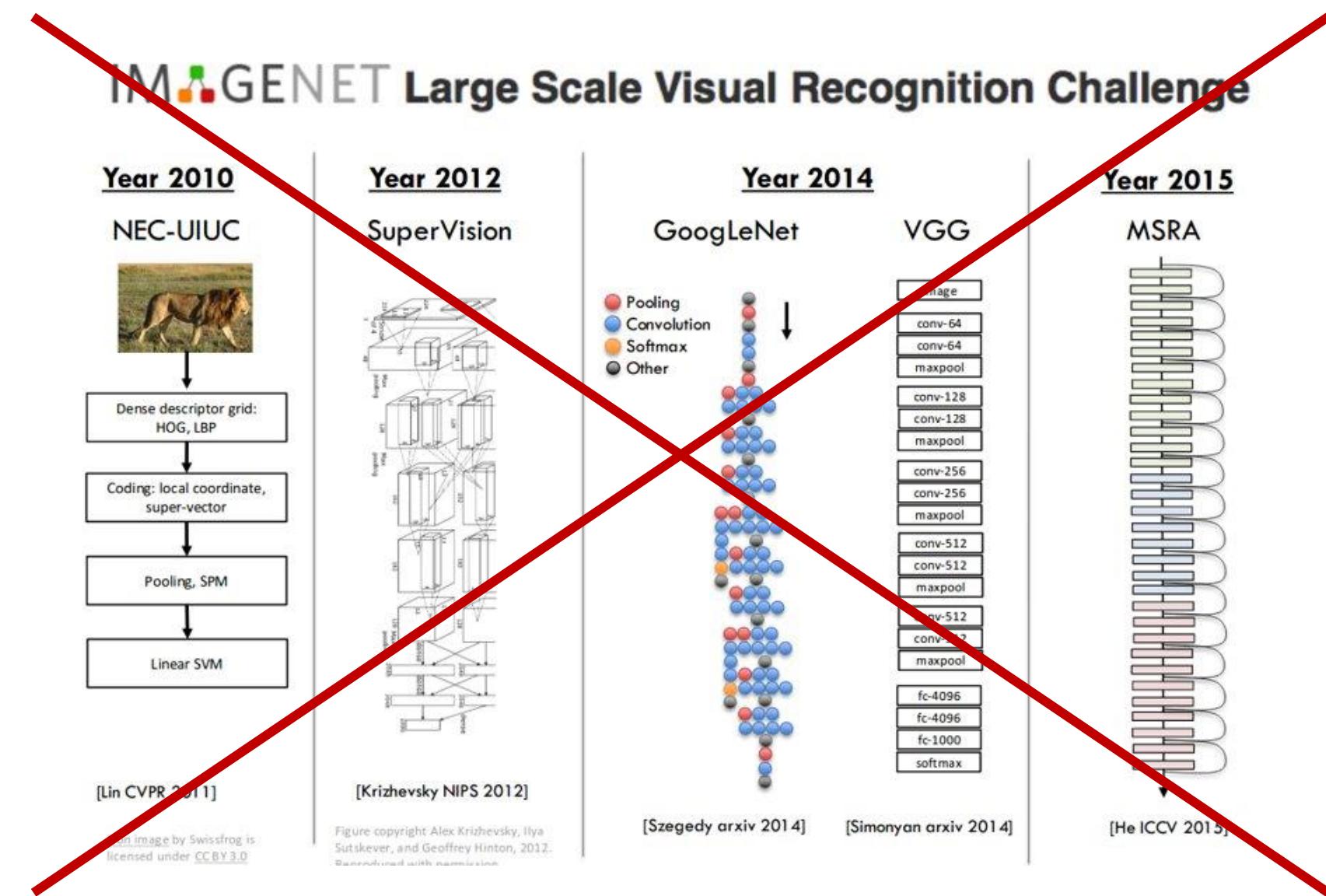
- **All of it should run in real time** : ~30 fps
- **On the Jetson TX2 CPU**
- **...with top quality performance**



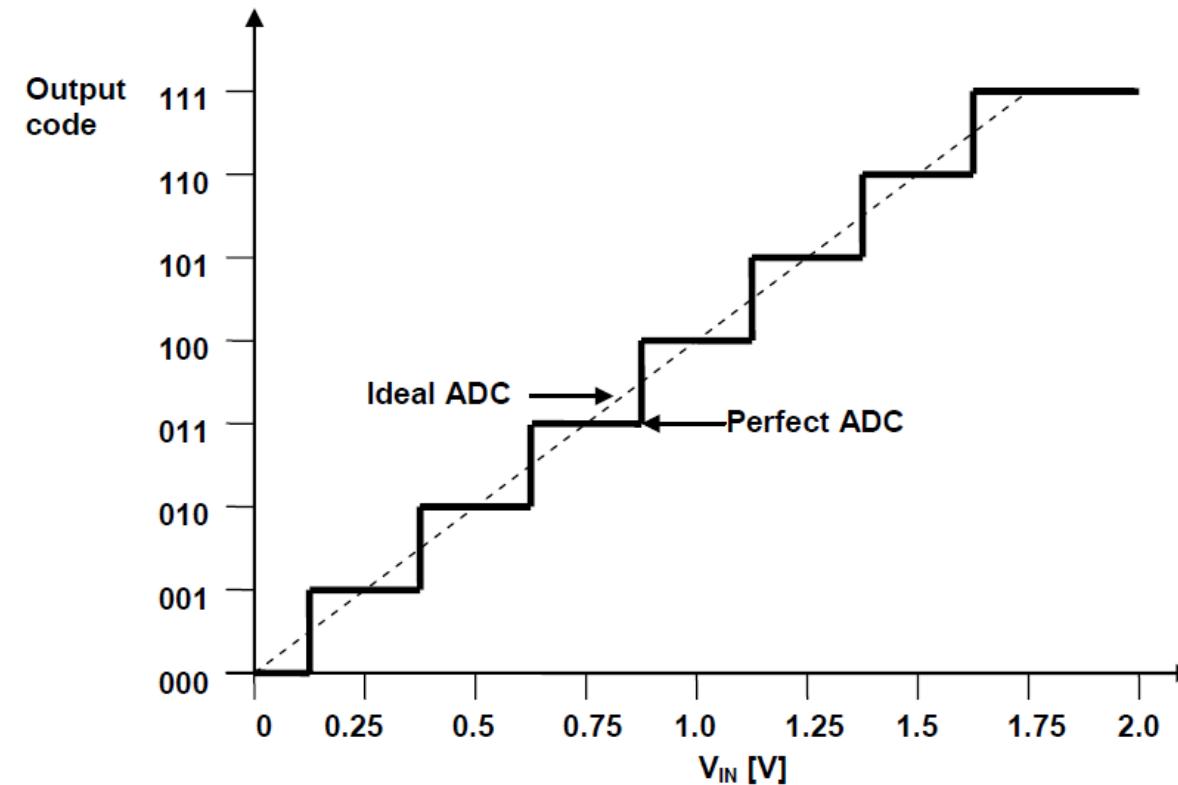
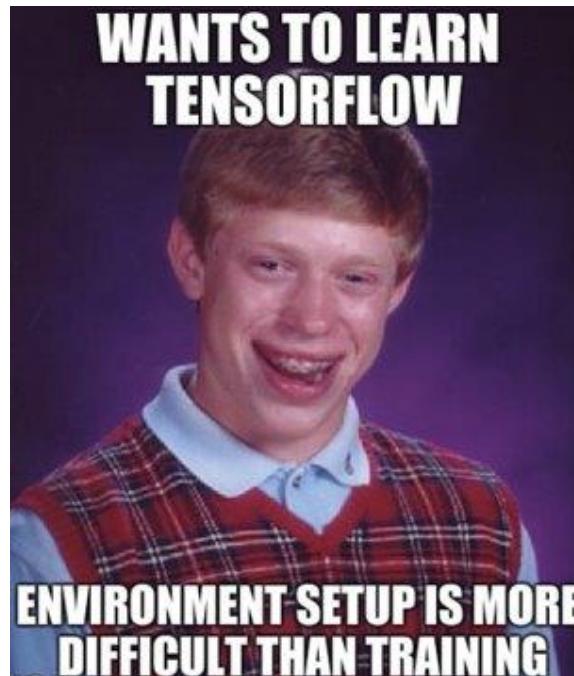
- **~70% mAP on COCO**
- **First measurement : 1000+ ms**

- Intro
- Model runtime optimization
- Data & model quality
- Demo

- Quantization
- Pruning
- Weight sharing
- Low rank approximation
- Operation efficiency
- Binary/ternary nets
- Distillation
- Network architecture
- ...

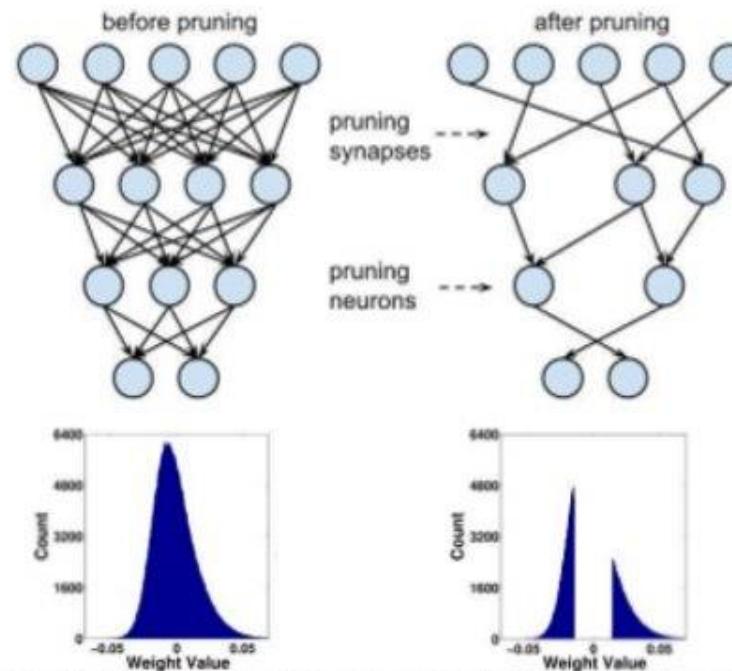
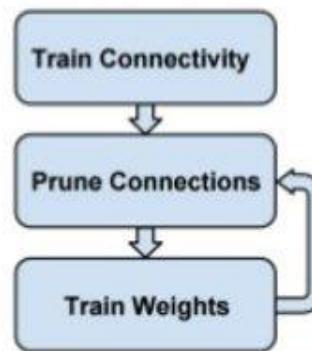


- Quantizing the weights and activations of a CNN from 32bit floating point into 8bit integers.
- Speed x4
- Memory x4



Pruning

Aim : Remove all connections with absolute weights below a threshold



Song Han, Jeff Pool, John Tran, William J. Dally, "Learning both Weights and Connections for Efficient Neural Networks", 2015

L1 Regularization

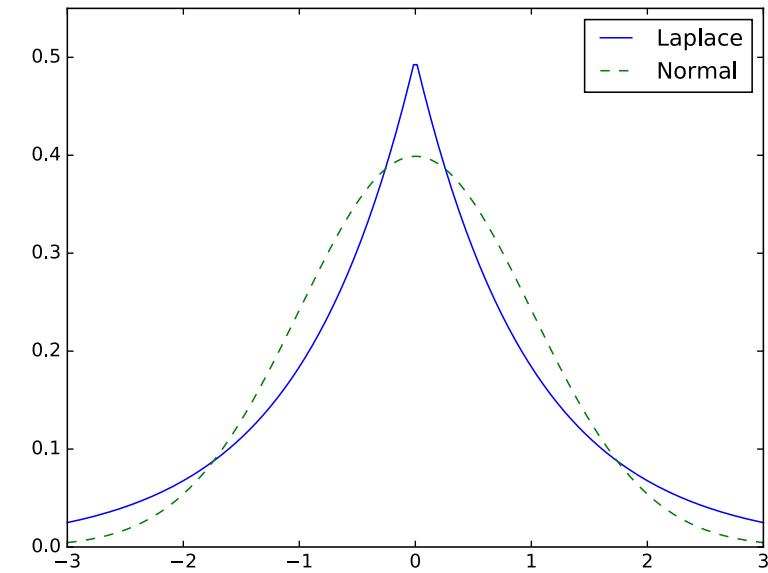
$$\text{Cost} = \sum_{i=0}^N (y_i - \sum_{j=0}^M x_{ij} W_j)^2 + \lambda \sum_{j=0}^M |W_j|$$

L2 Regularization

$$\text{Cost} = \sum_{i=0}^N (y_i - \sum_{j=0}^M x_{ij} W_j)^2 + \lambda \sum_{j=0}^M W_j^2$$

Loss function

Regularization Term



43	2	12	16	71
1	17	34	59	32
4	1	54	8	6
82	11	32	30	2
9	7	87	25	75

Regularization + pruning



4	24	0	0	13
12	0	4	9	0
61	17	0	38	0
0	0	82	13	0
36	78	0	0	33

43	2	12	16	71
1	17	34	59	32
4	1	54	8	6
82	11	32	30	2
9	7	87	25	75

Regularization + pruning



4	24	0	0	13
12	0	4	9	0
61	17	0	38	0
0	0	82	13	0
36	78	0	0	33



43	2	12	16	71
1	17	34	59	32
4	1	54	8	6
82	11	32	30	2
9	7	87	25	75



3
6
2
14
11

Which one is faster ?



4	24	0	0	13
12	0	4	9	0
61	17	0	38	0
0	0	82	13	0
36	78	0	0	33



3
6
2
14
11

43	2	12	16	71
1	17	34	59	32
4	1	54	8	6
82	11	32	30	2
9	7	87	25	75

Regularization + pruning



4	24	0	0	13
12	0	4	9	0
61	17	0	38	0
0	0	0	0	0
0	0	0	0	0



43	2	12	16	71
1	17	34	59	32
4	1	54	8	6
82	11	32	30	2
9	7	87	25	75



3
6
2
14
11

Which one is faster ?

4	24	0	0	13
12	0	4	9	0
61	17	0	38	0
0	0	0	0	0
0	0	0	0	0



3
6
2
14
11

f. map x-1 $[H, W, C]$ $[H, W, 3]$

X
X
X

**Layer x** $[N, K, K, C]$ $[5, K, K, 3]$

4	24	0
12	0	4
61	17	0
0	0	0
0	0	0

**f. map x** $[H, W, N]$ $[H, W, 5]$

X
X
X
X
X

**Layer x+1** $[M, K, K, N]$ $[5, K, K, 5]$

X	X	X	X	X
X	X	X	X	X
X	X	X	X	X
X	X	X	X	X
X	X	X	X	X

**f. map x+1** $[H, W, M]$ $[H, W, 5]$

X
X
X
X
X

f. map x-1 $[H, W, C]$ $[H, W, 3]$

X
X
X

**Layer x** $[N, K, K, C]$ $[5, K, K, 3]$

4	24	0
12	0	4
61	17	0
0	0	0
0	0	0

**f. map x** $[H, W, N]$ $[H, W, 5]$

X
X
X
X
X

**Layer x+1** $[M, K, K, N]$ $[5, K, K, 5]$

X	X	X	X	X
X	X	X	X	X
X	X	X	X	X
X	X	X	X	X
X	X	X	X	X

**f. map x+1** $[H, W, M]$ $[H, W, 5]$

X
X
X
X
X

 $[H, W, 3]$ $[3, K, K, 3]$

X
X
X

 $[H, W, 3]$

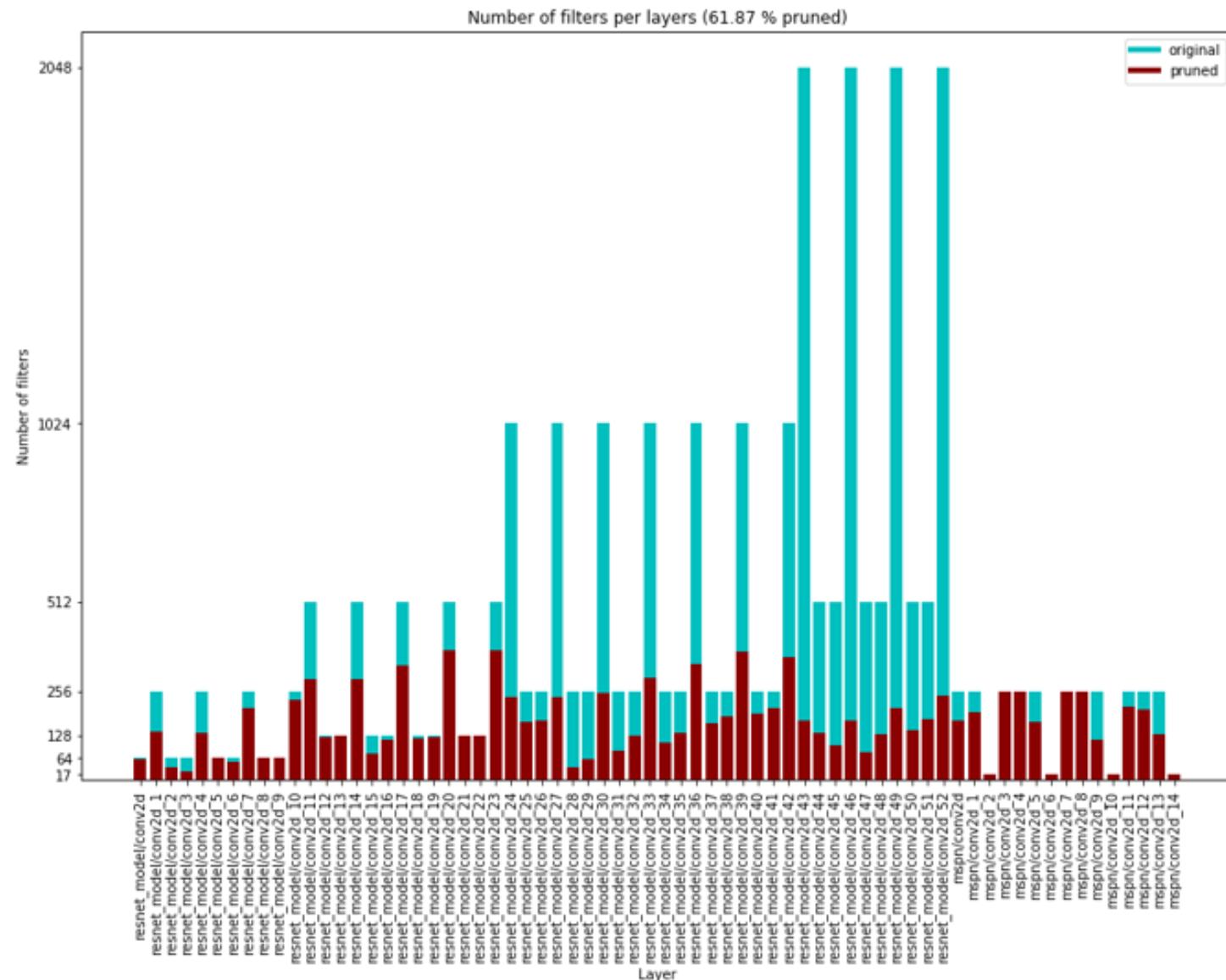
4	24	0
12	0	4
61	17	0

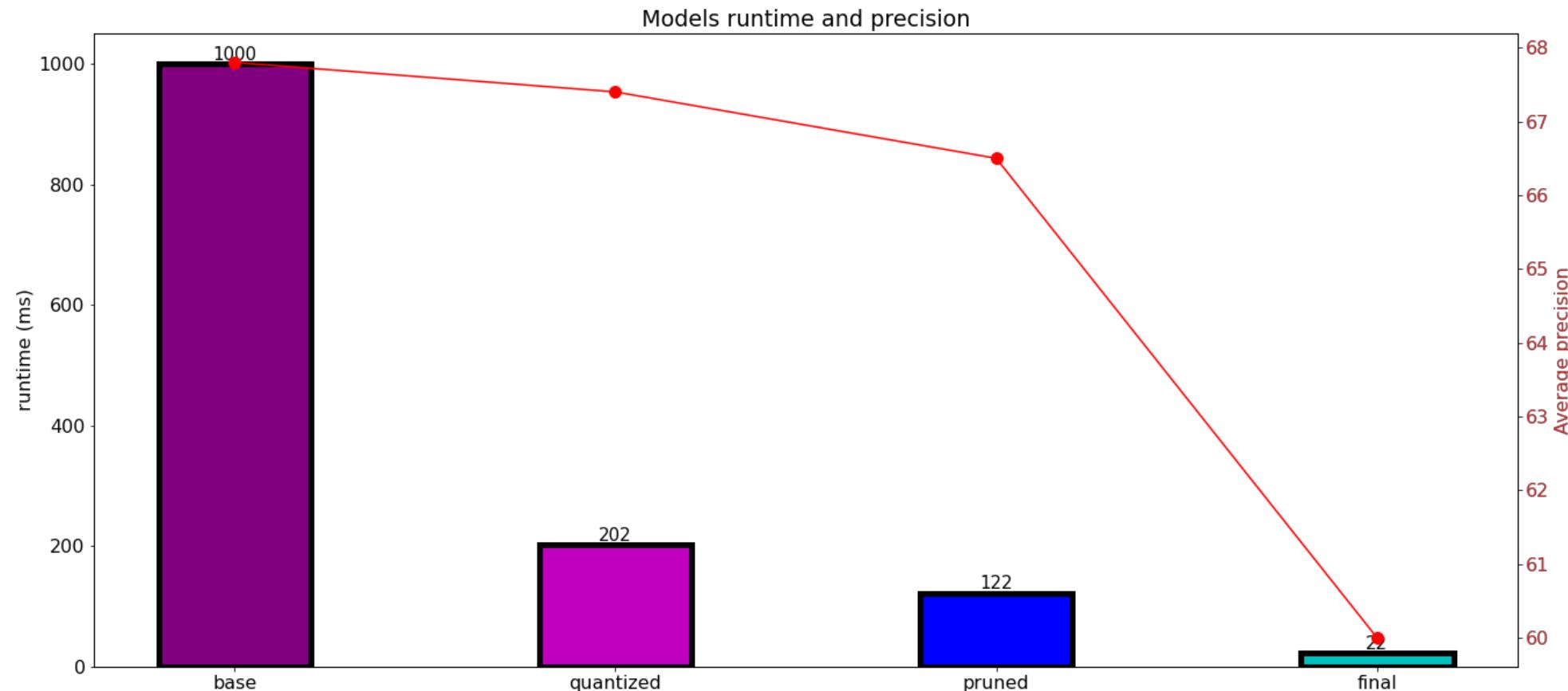
 $[5, K, K, 3]$

X	X	X
X	X	X
X	X	X
X	X	X
X	X	X

 $[H, W, 5]$

X
X
X
X
X



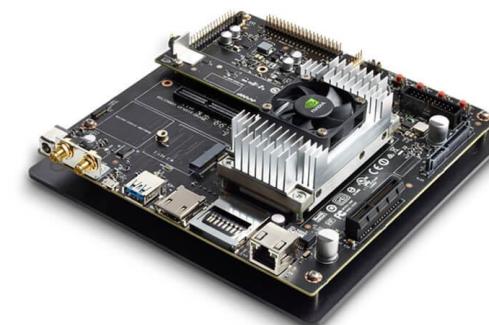


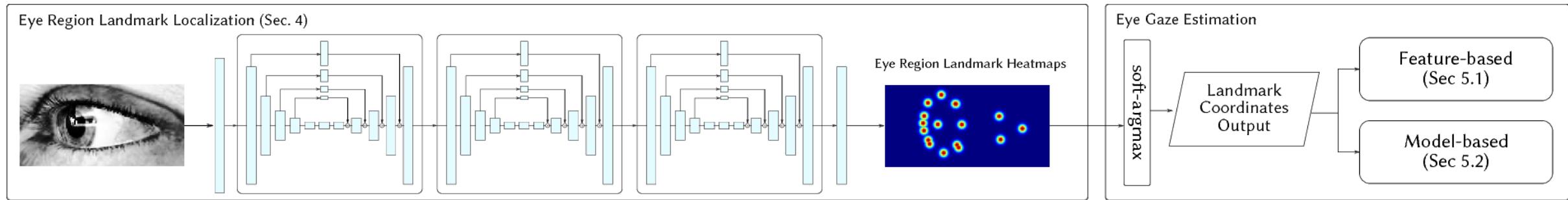
- **A DMS system could contain :**

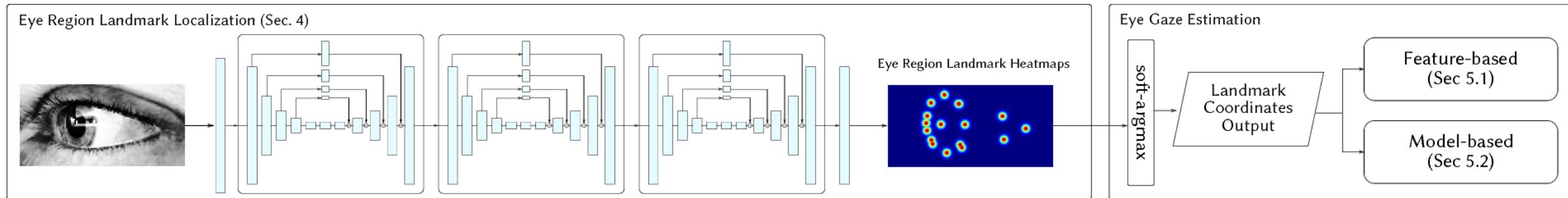
- Head position
- Head orientation
- Keypoints estimation
- Eye gaze
- Eye openness
- Face recognition
- Drowsiness detection
- Smartphone/food/cigarette/drinks detection
- Anti spoofing
- Age, weight estimation
- ...

- **All of it should run in real time** : ~30 fps
- **...on an embedded device**
- **...with top quality performance**

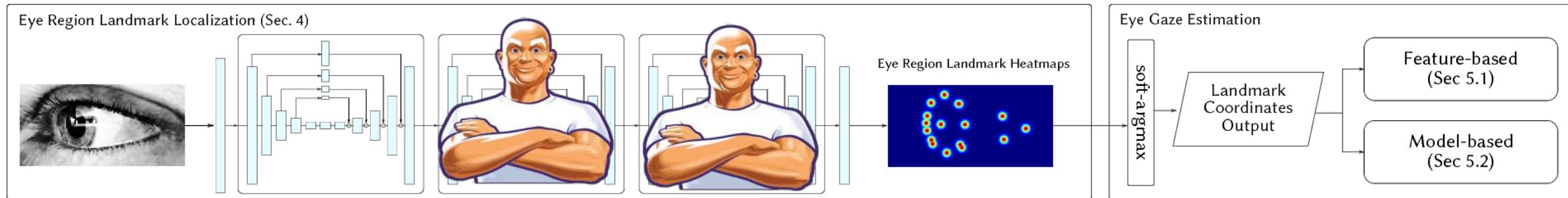
Keypoint estimation	Eye analysis	Total
22 ms	...	> 22 ms

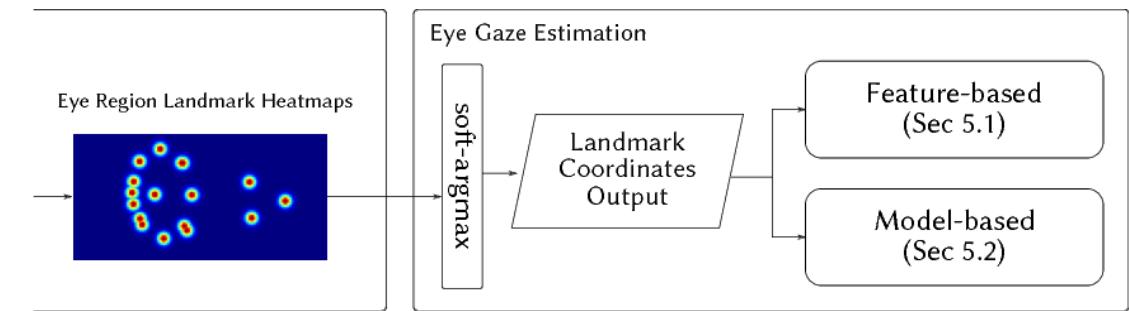
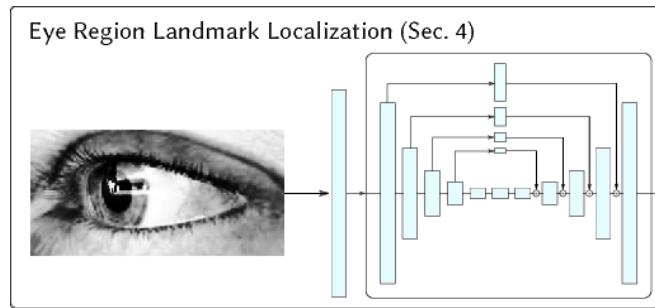


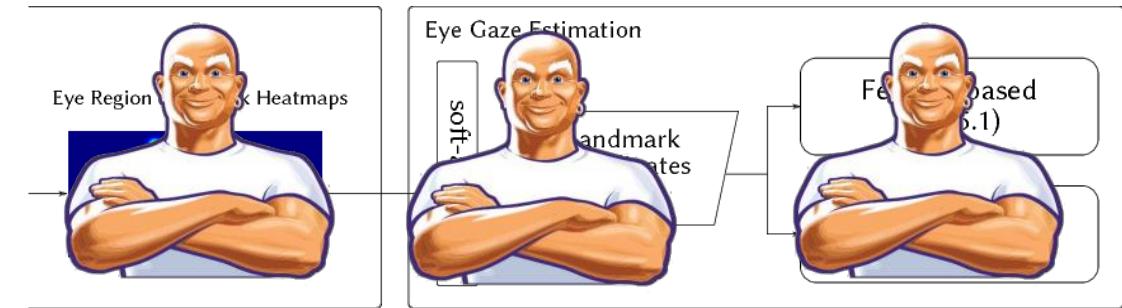
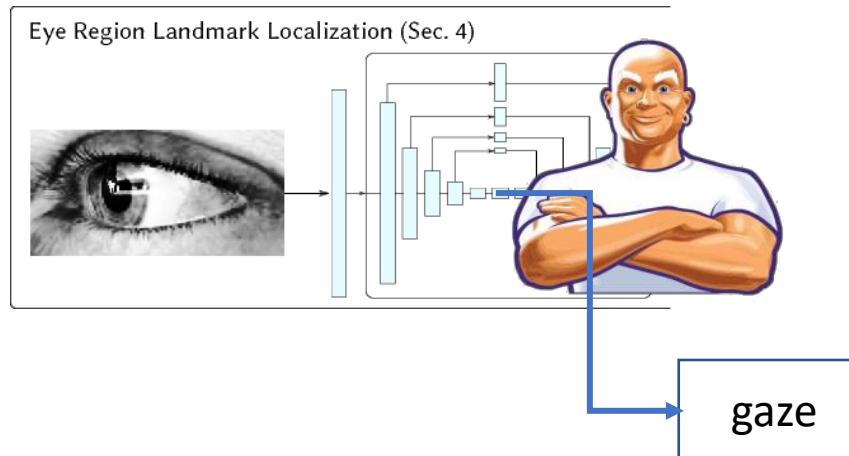


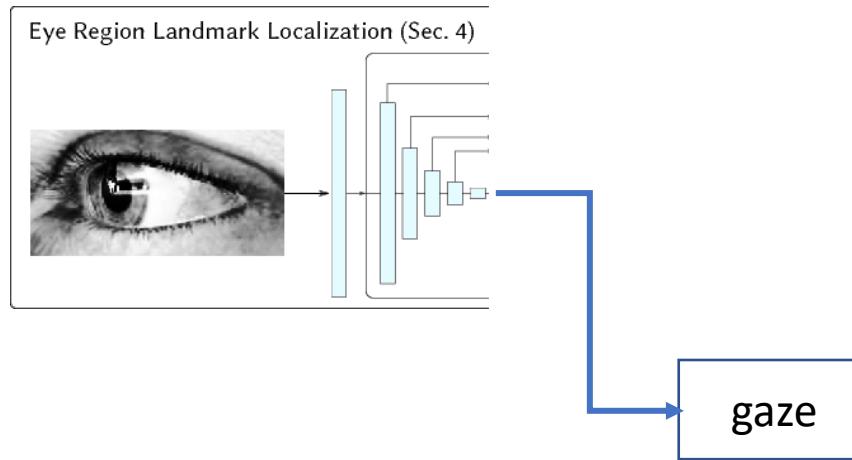


"There is work to do"

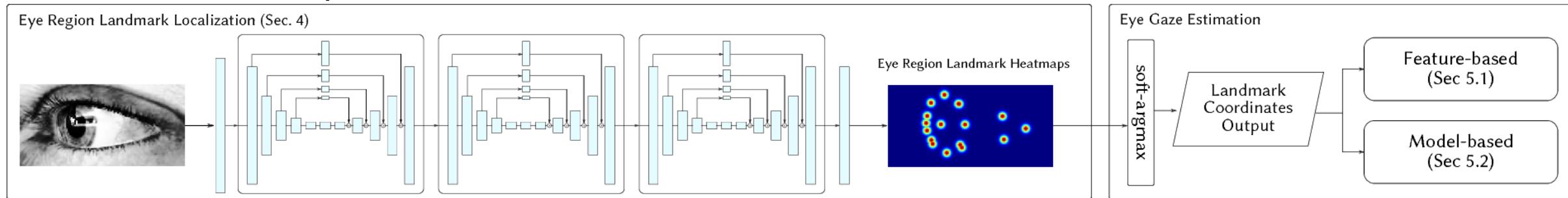




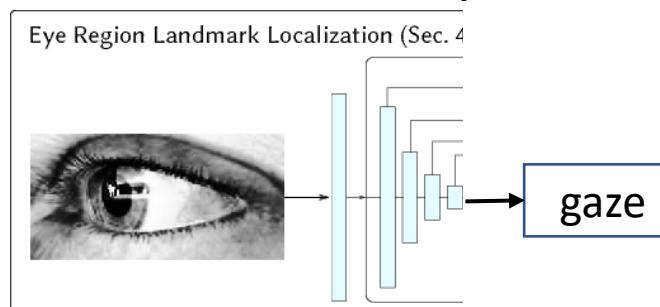




Before Meister Proper



After Meister Proper



Danke Meister Proper

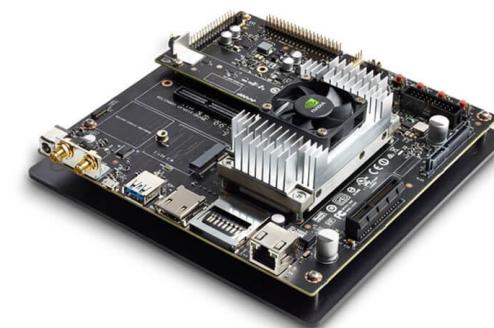


"Bitte schön"

- **A DMS system could contain :**

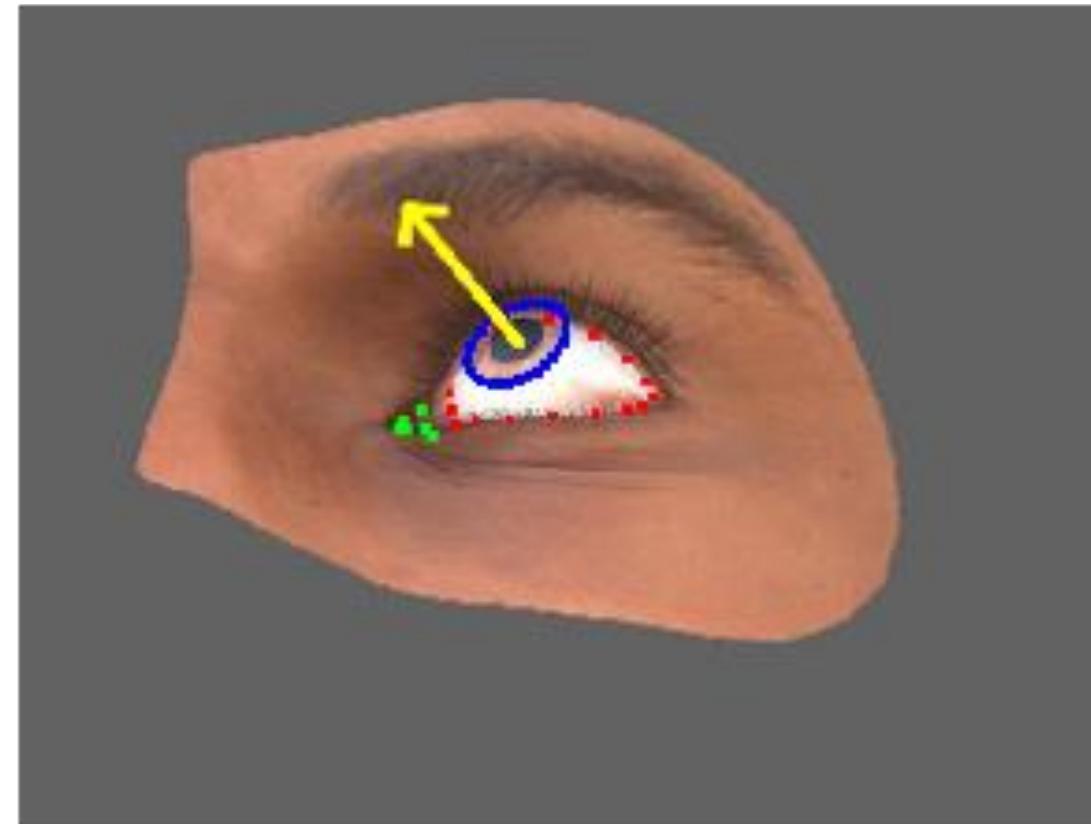
- Head position
- Head orientation
- Keypoints estimation
- Eye gaze
- Eye openness
- Face recognition
- Drowsiness detection
- Smartphone/food/cigarette/drinks detection
- Anti spoofing
- Age, weight estimation
- ...

- **All of it should run in real time : ~30 fps**
- **...on an embedded device**
- **...with top quality performance**



Keypoint estimation	Eye gaze analysis	Total
22 ms	10 ms	32 ms

- New network ?
 - Extra runtime
 - easy
- Through body keypoints ?
 - Needs new data
 - Low resolution
- Together with gaze ?
 - Specific data
 - Makes sense
 - Balance between tasks

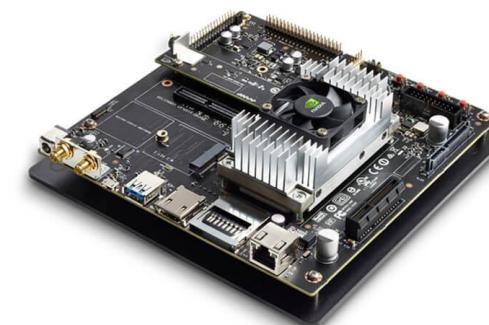


- **A DMS system could contain :**

- Head position
- Head orientation
- Keypoints estimation
- Eye gaze
- Eye openness
- Face recognition
- Drowsiness detection
- Smartphone/food/cigarette/drinks detection
- Anti spoofing
- Age, weight estimation
- ...



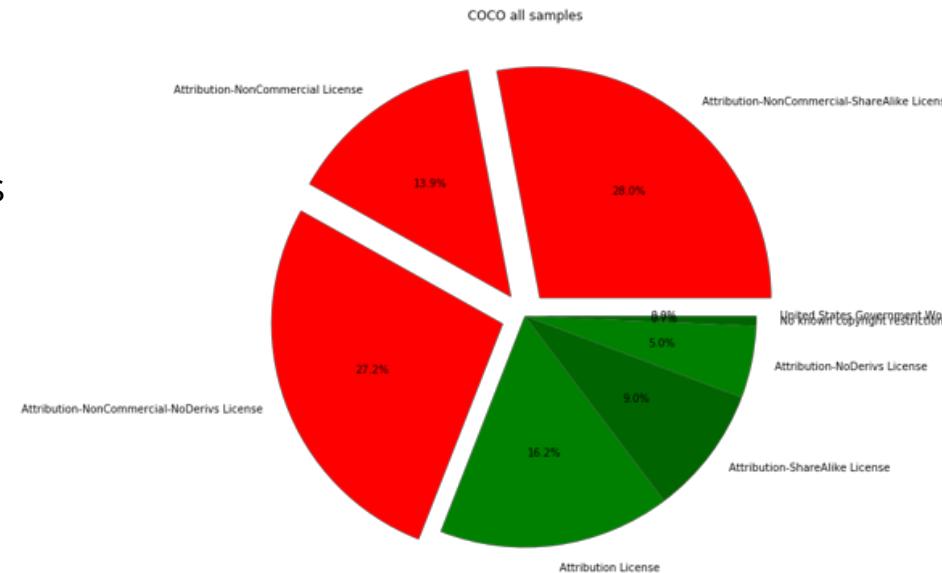
- **All of it should run in real time** : ~30 fps
- **On an embedded device**



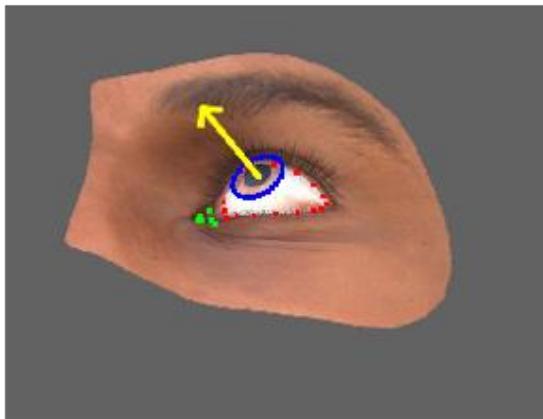
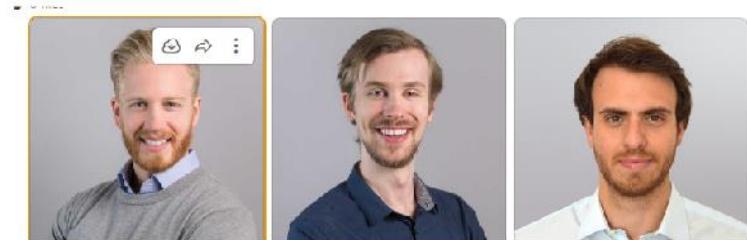
Keypoint estimation	Eye gaze analysis	Total
22 ms	10 ms	32 ms

- **Intro**
- **Model runtime optimization**
- **Data & model quality**
- **Demo**

- Legal issues
- Different distributions



BEFORE THE PANDEMIC



AFTER THE PANDEMIC



- Eye openness : easy
 - Keypoints : expensive
 - Eye gaze : ????
-
- In any case, expensive to get a wide distribution over:
 - Gender
 - Expressions
 - Ethnicities
 - Age
 - Backgrounds
 - Clothes
 - Camera position...
-
- Data protection



- Adapt distribution
- Reduce amount of samples needed

Generative adversarial networks (conceptual)

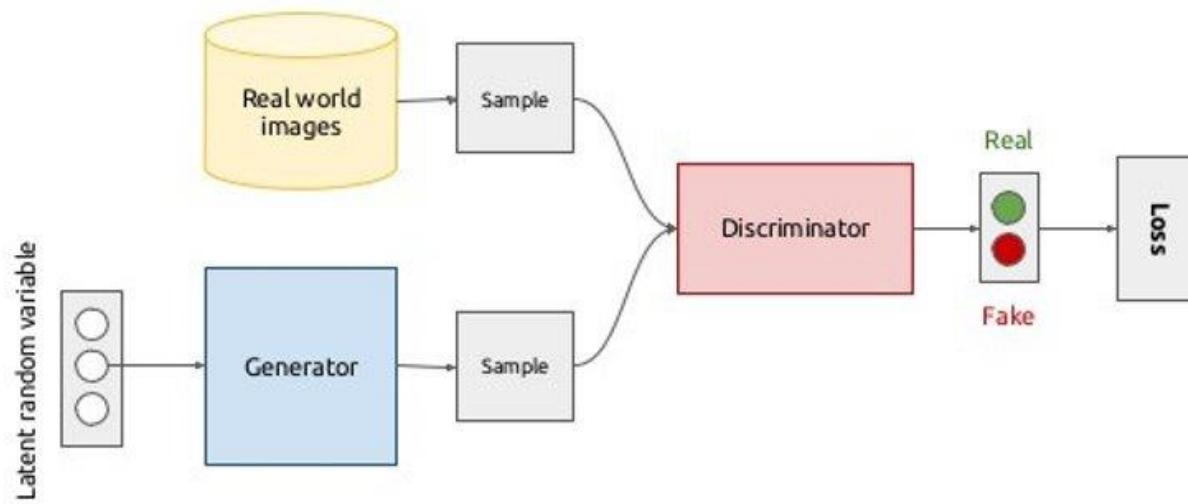
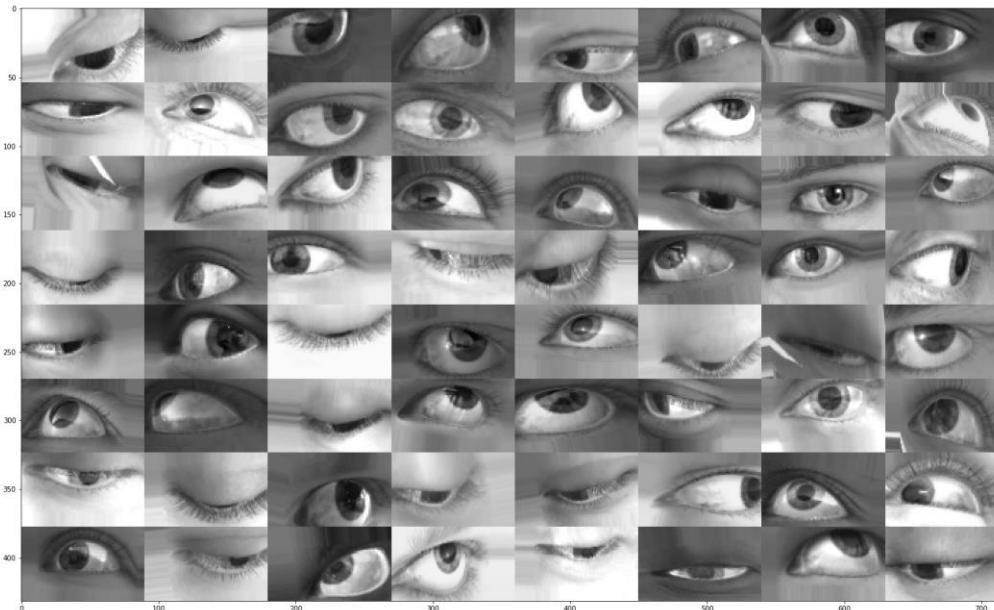


Figure 5: 1024×1024 images generated using the CELEBA-HQ dataset. See Appendix F for a larger set of results, and the accompanying video for latent space interpolations.



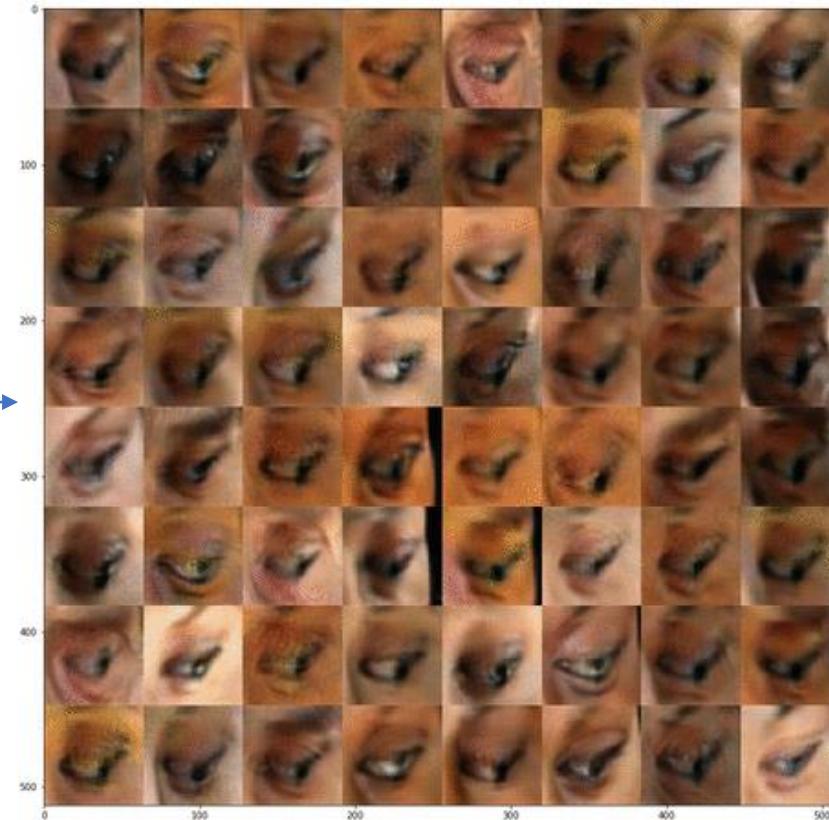
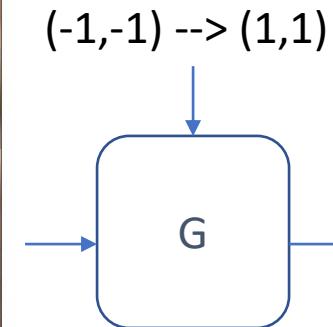
- Adapt texture distribution
- Reduce amount of samples needed



- Guarantee to keep ground truth ?



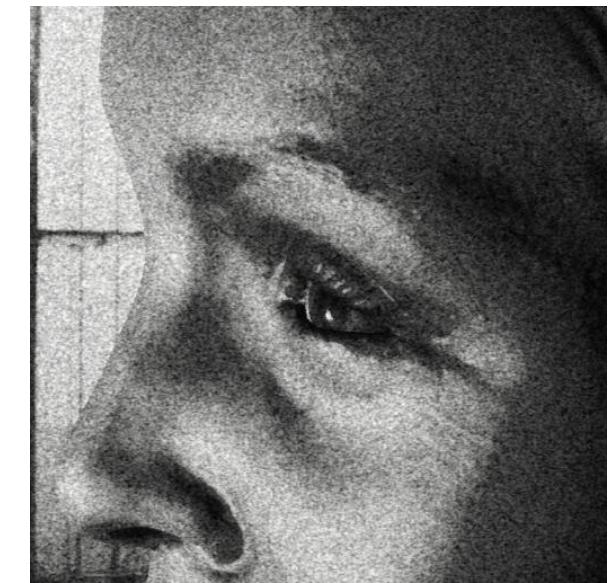
- Adapt gaze distribution
 - Reduce amount of samples needed
- Guarantee of correct ground truth ?



- Advantages
 - Cheap & fast
 - Accurate
 - Huge diversity
 - Multitask learning
 - specific



- Drawbacks :
 - Texture
 - Movements
 - overhead

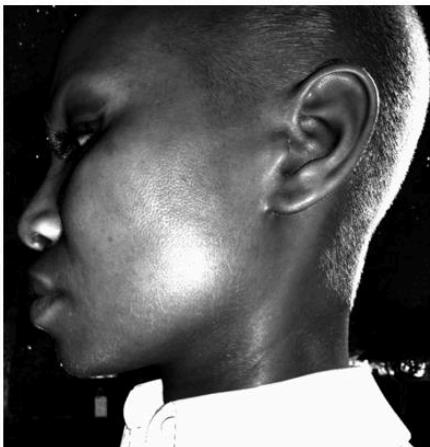


- Advantages
 - Cheap & fast
 - Accurate
 - Huge diversity
 - Multitask learning
 - specific
- Drawbacks :
 - Texture
 - Movements
 - overhead



You

the guy she told
you not to worry
about



**nooo you
can't just train
on synthetic data**

imgflip.com



**haha simulator
go brrrrrrrr**

- **Intro**
- **Model runtime optimization**
- **Data & model quality**
- **Demo**

Thank you !

Any questions ?

THE FUTURE OF AUTOMOTIVE SAFETY & DRIVING EXPERIENCE

EMOTION3D HQ

Rainergasse 1/8
1040 Vienna
Austria

www.emotion3d.ai
office@emotion3d.ai

Going Embedded: Real-time Deep Learning for Automotive Applications

Georg Braun

7. Conclusion

6. Benchmarks

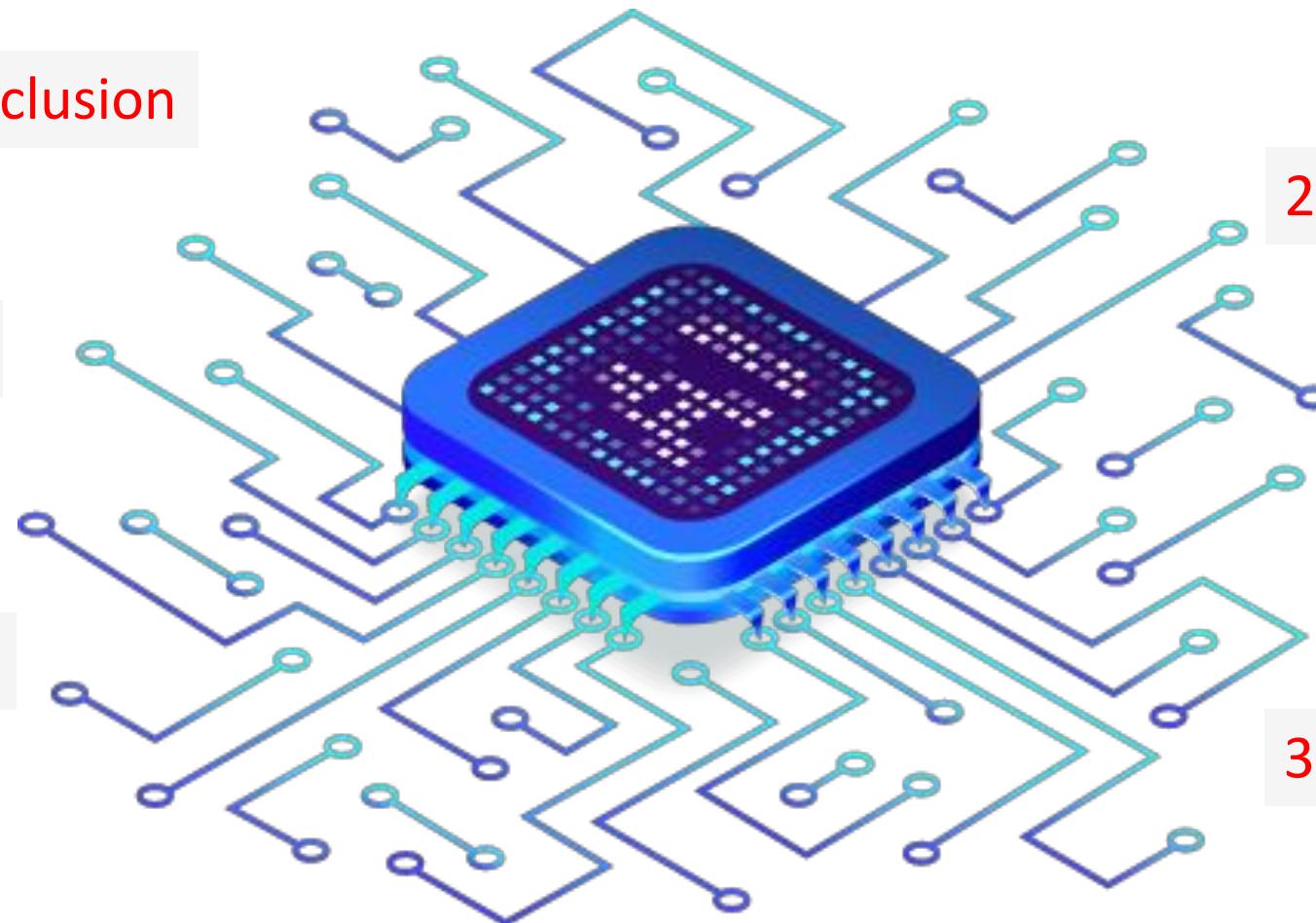
5. Pitfalls

1. Motivation

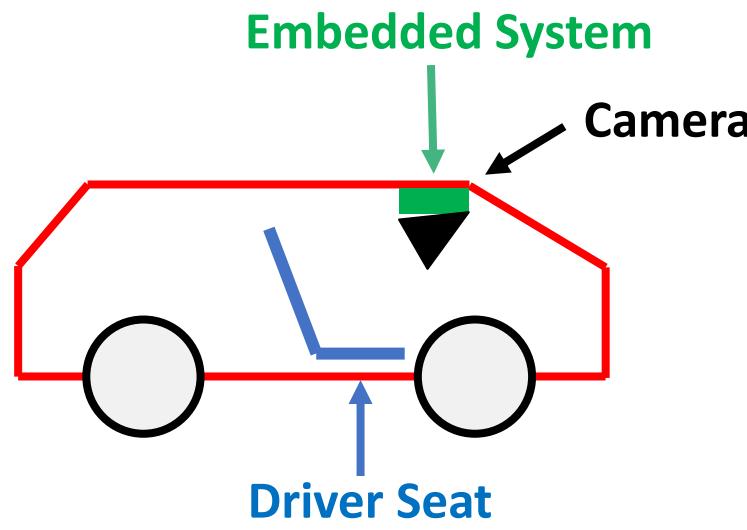
2. NVIDIA Jetson

3. Ambarella CV22

4. Theory: Quantization



- **Camera based incabin-monitoring system**



The setup



Camera image



Embedded system

- **What do we want?**

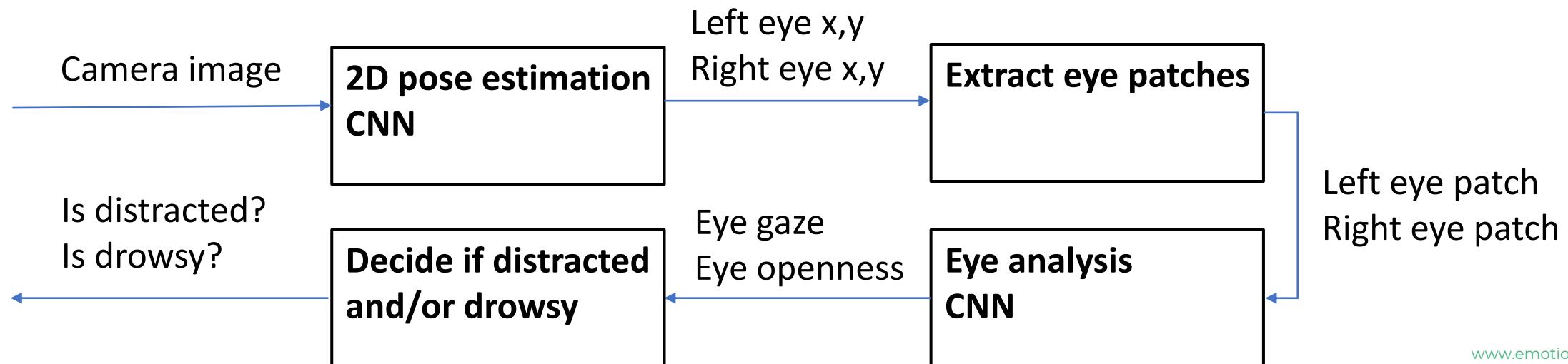


Detect distracted driver



Detect drowsy driver

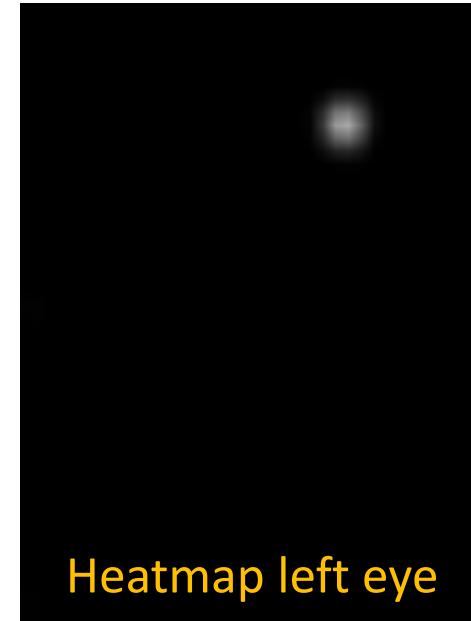
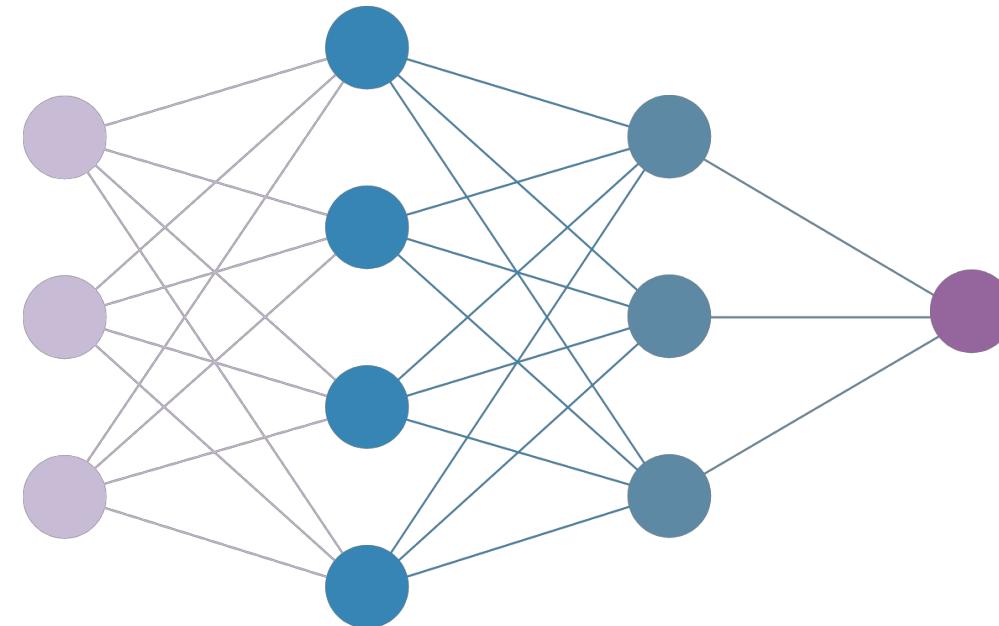
- **How to achieve?**



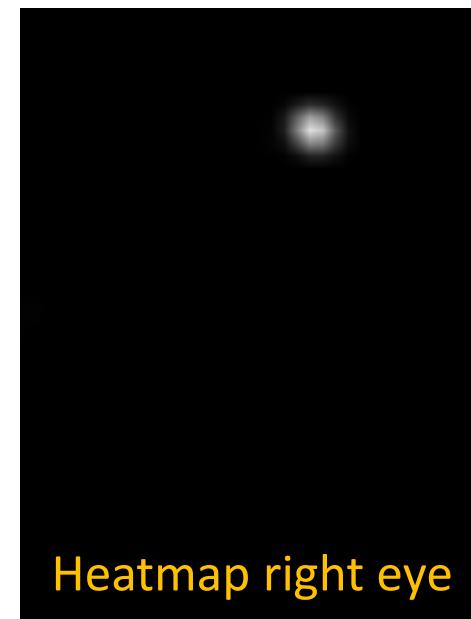
- **2D pose estimation CNN**



Camera image



Locate Max
(x=99, y=39)



Locate Max
(x=90, y=39)

- **NVIDIA Jetson AGX Xavier**

- High-end edge device
- ~ €1000 (DevKit)
- Quite similar to regular PCs (NVIDIA GPU, Ubuntu, ARM64 CPU instead of AMD64)
- Easy to port software from PC to Jetson
- We use it at trade shows and for customer samples

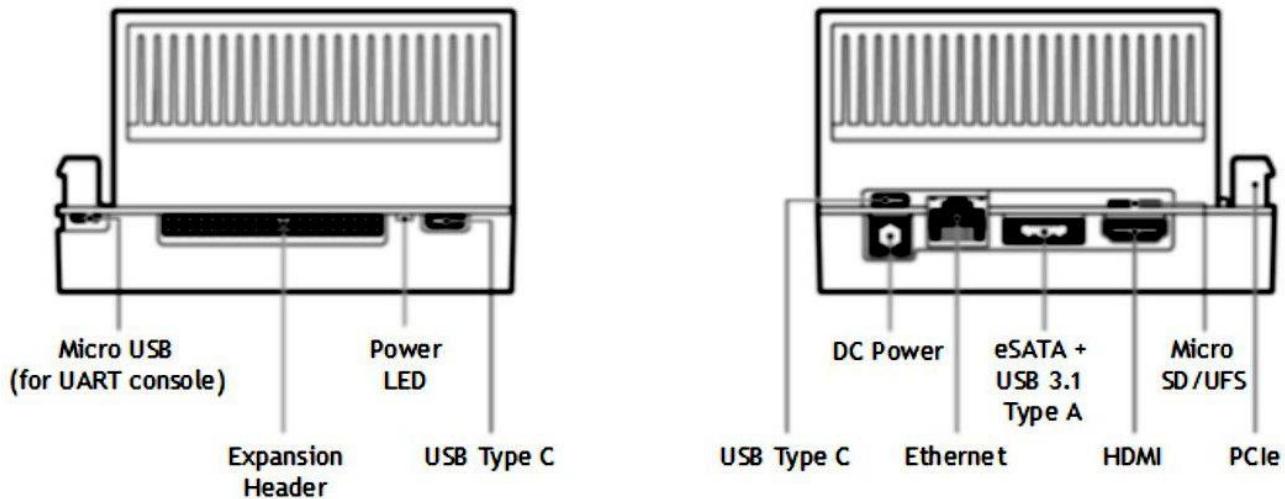


- Easy to acquire, also for private individuals
- Generic (any neural network, not just CNN)

- NVIDIA Jetson AGX Xavier**

- Officially supports Tensorflow & Pytorch
- Well documented, large community:
 - developer.nvidia.com/jetson

Front view (left) and rear view (right)

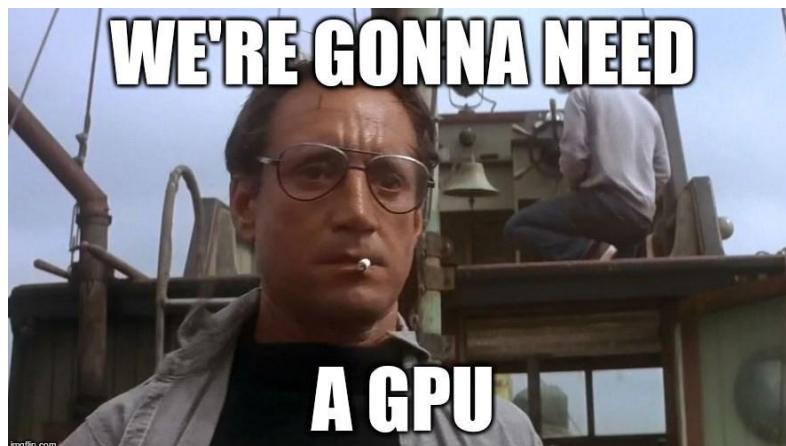


JETSON AGX XAVIER	
GPU	512 Core Volta @ 1.37GHz 64 Tensor Cores
DL Accelerator	(2x) NVDLA
Vision Accelerator	(2x) 7-way VLIW Processor
CPU	8 core Carmel ARM CPU @ 2.26GHz (4x) 2MB L2 + 4MB L3
Memory	16GB 256-bit LPDDR4x @ 2133MHz 137 GB/s
Storage	32GB eMMC
Video Encode	(4x) 4Kp60 / (8x) 4Kp30 HEVC
Video Decode	(2x) 8Kp30 / (6x) 4Kp60 12 bit support
Camera	16 lanes MIPI CSI-2 8 lanes SLVS-EC D-PHY 40Gbps / C-PHY 109Gbps
PCI Express	16 lanes PCIe Gen 4 1x8 + 1x4 + 1x2 + 2x1
Mechanical	100mm x 87mm 699 pin connector
Power	10W / 15W / 30W

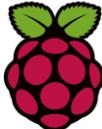
- Is used for automated and autonomous driving →

	PC - CPU	PC - GPU	Xavier - CPU	Xavier - GPU
2D pose estimation	40,26 ms	3,68 ms	76,15 ms	7,05 ms
Eye analysis	12,04 ms	3,06 ms	24,85 ms	6,64 ms

- PC – CPU = Intel(R) Core(TM) i7-7700 CPU @ 3.60GHz
- PC – GPU = NVIDIA GeForce GTX 1080



- If we could get a device that is optimized for (only) CNN inference, we could probably get similar performance for less \$\$\$

- A  won't do it here

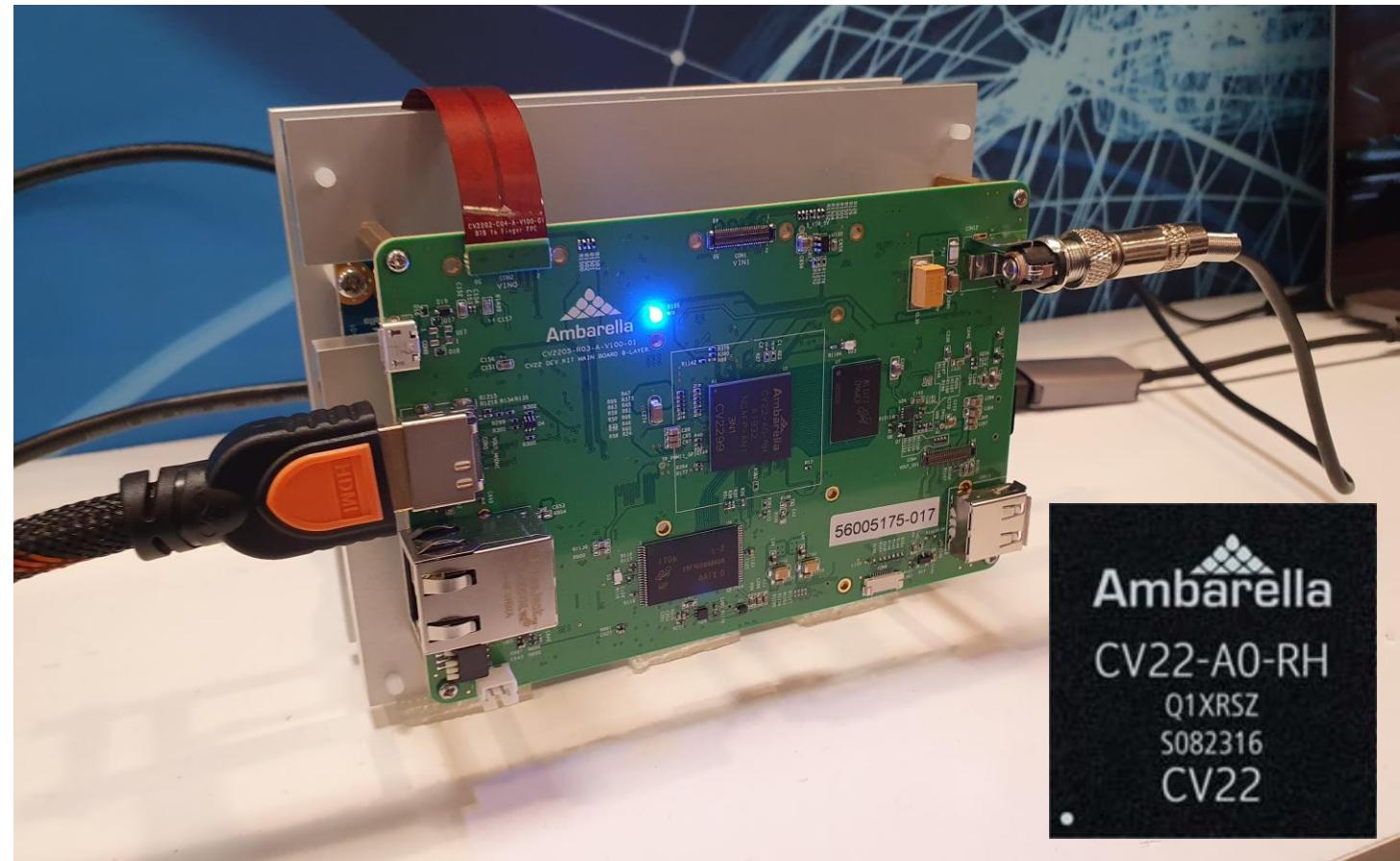
- But there are a variety of options



- Let's look at the CV22 by Ambarella (www.ambarella.com)



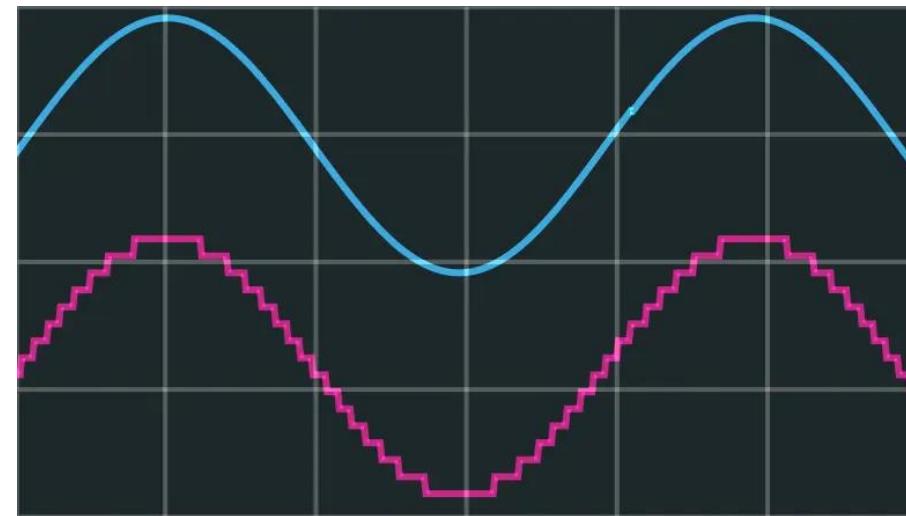
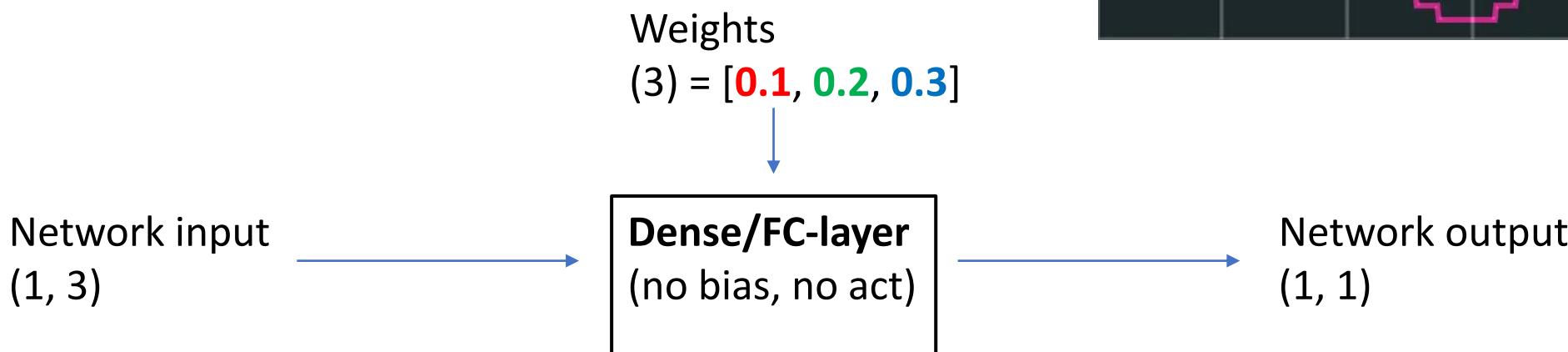
- **Ambarella CV22**
- **CPU:**
 - Quad-core ARM ® Cortex ® -A53 up to 1.0 GHz
 - NEON™ SIMD and FPU acceleration
- **Neural Accelerator:**
 - **CVflow** vision accelerator for CNN
- **Memory Interfaces:**
 - LPDDR4 / LPDDR4x up to 1.8 GHz, 32-bit data bus
 - Two SD controller with SDXC SD™ card



Ambarella CV22 DevKit, Image (c) anandtech.com

- **Gigabit Ethernet, HDMI, USB**

- It is why neural network accelerators are fast
- FLOAT32 ➡ INT8
- Let's start with a non-quantized network:



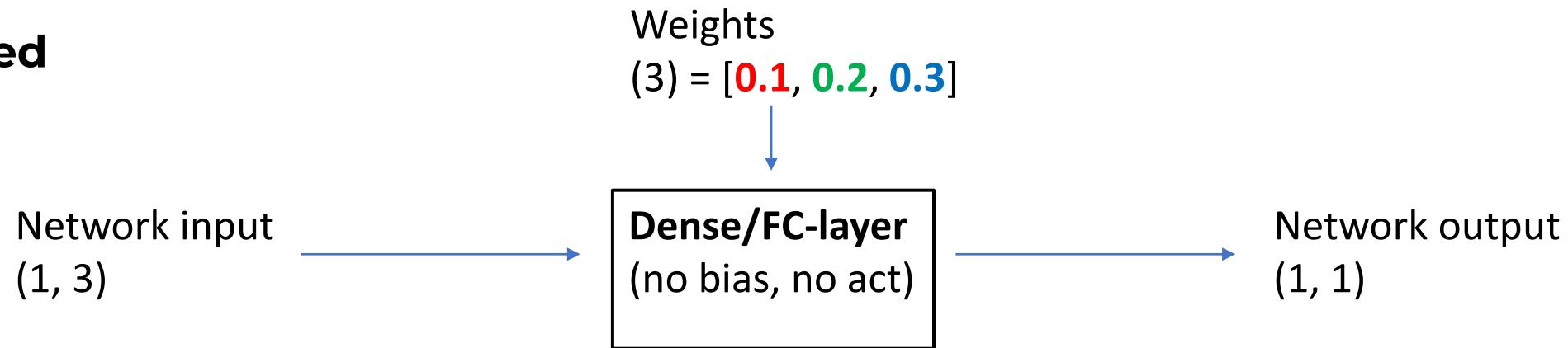
- Assuming a network input of [0.15, 0.44, 0.66], the network output is (drumroll):

$$0.15 * 0.1 + 0.44 * 0.2 + 0.66 * 0.3 = 0.301$$

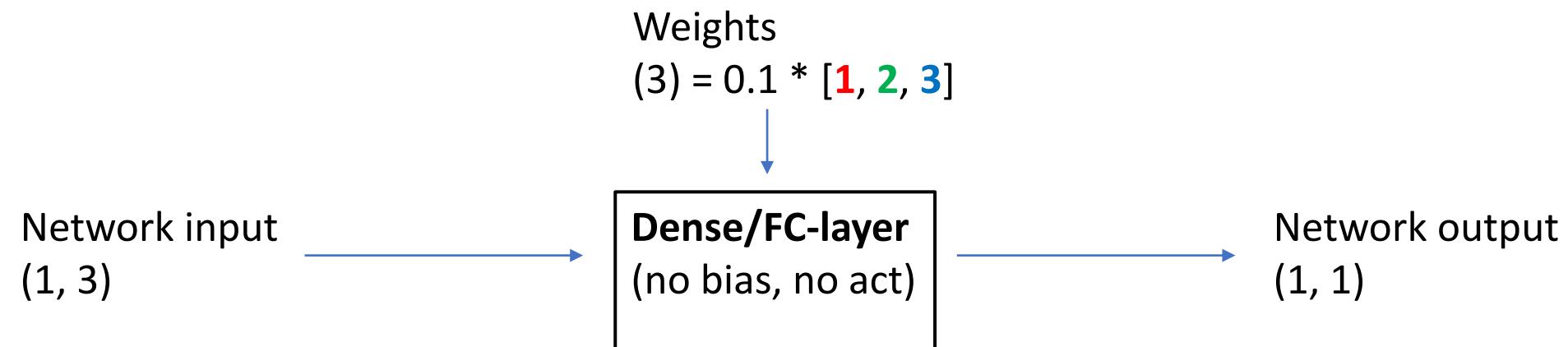
- Weights are converted to quantized representation

$$s * [w_1, w_2, \dots, w_n] \quad s = \text{scale factor (FLOAT32)}, w_1 - w_n: \text{quantized weights (INT8)}$$

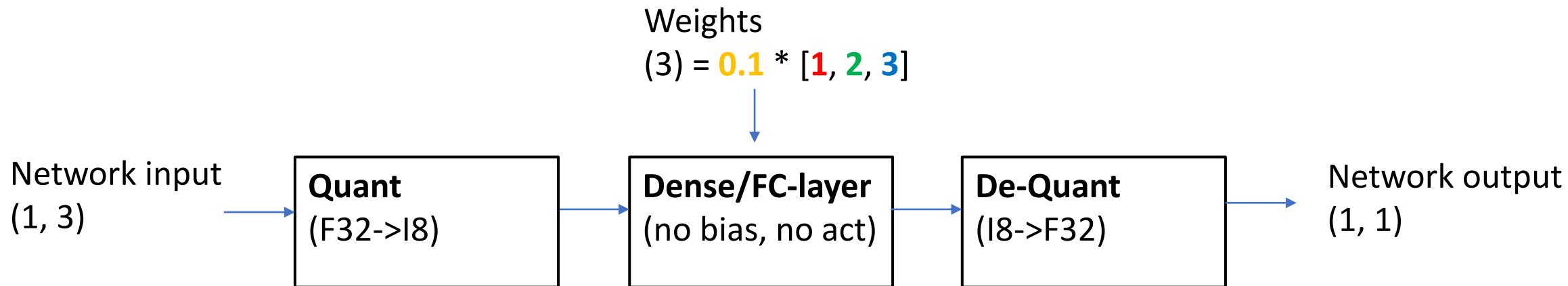
- Non-quantized



- Quantized



- Quantization and de-quantization operations are added to network

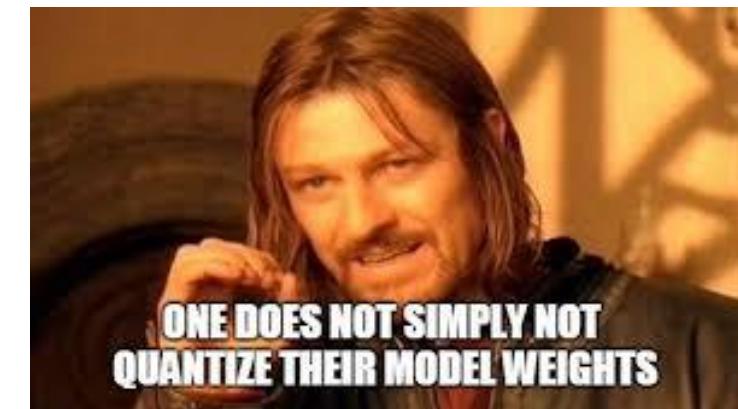


- The quantization operation quantizes the network input

$$[0.15, 0.44, 0.66] \rightarrow 0.11 * [1, 4, 6]$$

- The layers operate only on quantized values

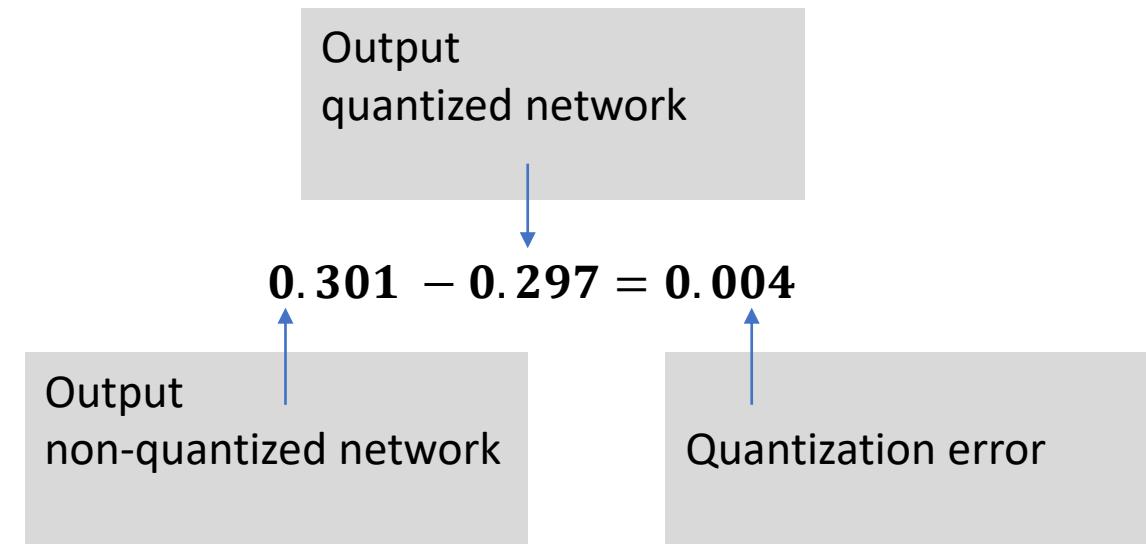
$$\text{dense_layer_output} = 1 * 1 + 4 * 2 + 6 * 3 = 27$$



- The de-quantization layer generates the final network output:

$$\text{network_output} = \text{dense_layer_output} * 0.1 * 0.11 = 0.297$$

- **Quantization error:**



FLOAT32



INT8



- **References:**

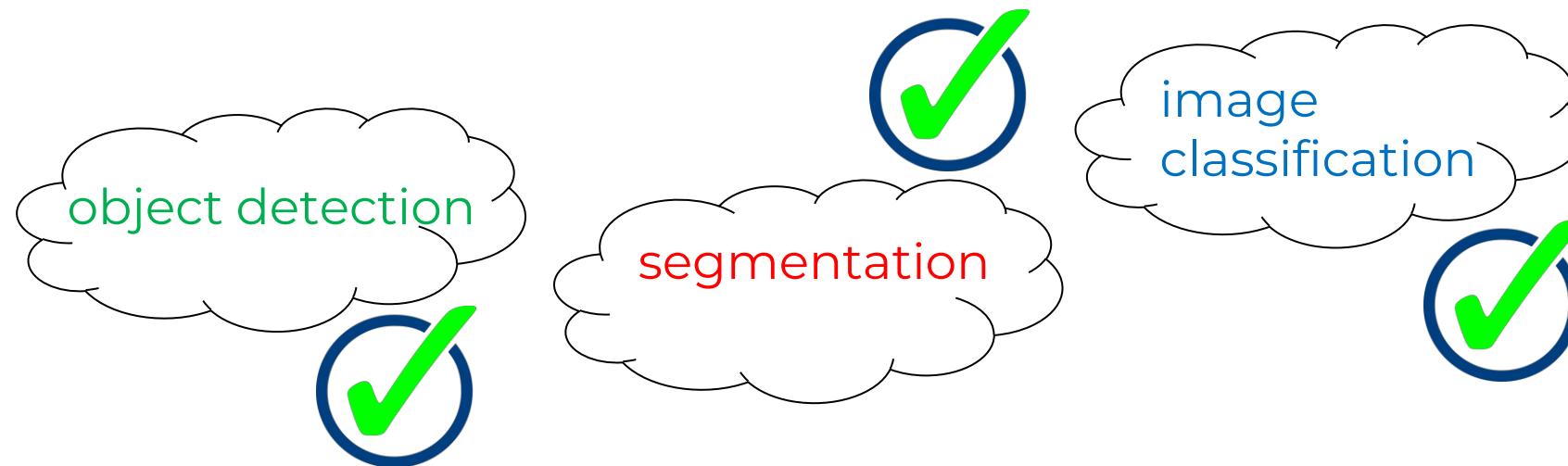
- „A White Paper on Neural Network Quantization“ (2021) <https://arxiv.org/abs/2106.08295>



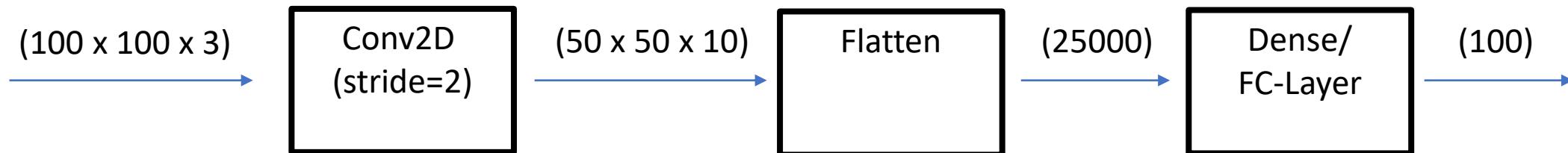
- **Quantization error is high:**
 - **Tune quantization parameters**
 - **Do quantization aware training**
- **Model conversion aborts with errors:**
 - **Replace unsupported layers and re-train**



- **Model conversion tools support common network architectures well**



- „Exotic“ architectures can confuse the model conversion tool:



- **How images are stored in memory**

Channel 0 (e.g. RED)			Channel 1 (e.g. GREEN)			Channel 2 (e.g. BLUE)		
1	2	3	10	11	12	19	20	21
4	5	6	13	14	15	22	23	24
7	8	9	16	17	18	25	26	27

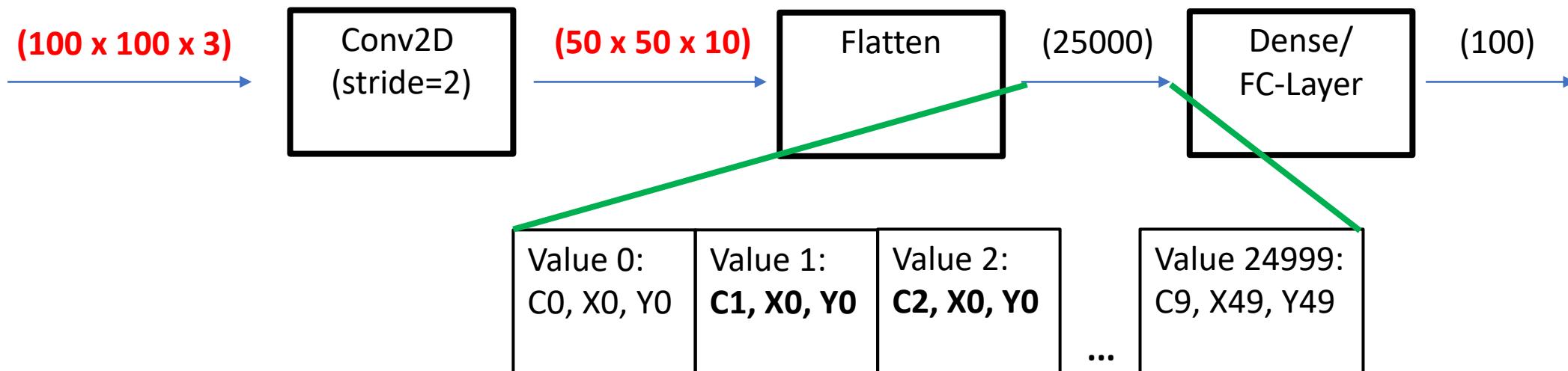
Image in memory, „channels first“ (CHW) order:

[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27]

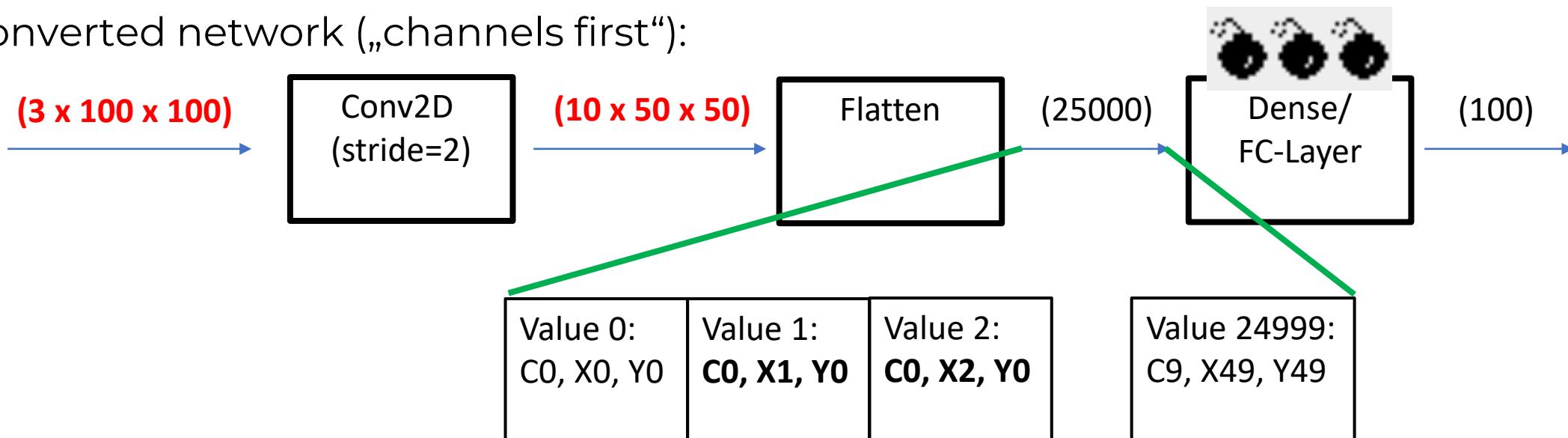
Image in memory, „channels last“ (HWC) order:

[1, 10, 19, 2, 11, 20, 3, 12, 21, 4, 13, 22, 5, 14, 23, 6, 15, 24, 7, 16, 25, 8, 17, 26, 9, 18, 27]

- Original network („channels last“):

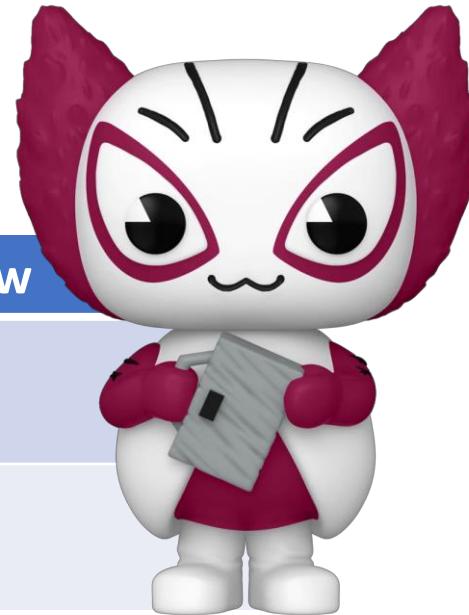


- Converted network („channels first“):



- **Inference latency:**

	PC - CPU	PC - GPU	Xavier - CPU	Xavier - GPU	CV22 - CVFlow
2D pose estimation	40,26 ms	3,68 ms	76,15 ms	7,05 ms	2,81ms
Eye analysis	12,04 ms	3,06 ms	24,85 ms	6,64 ms	2,34ms



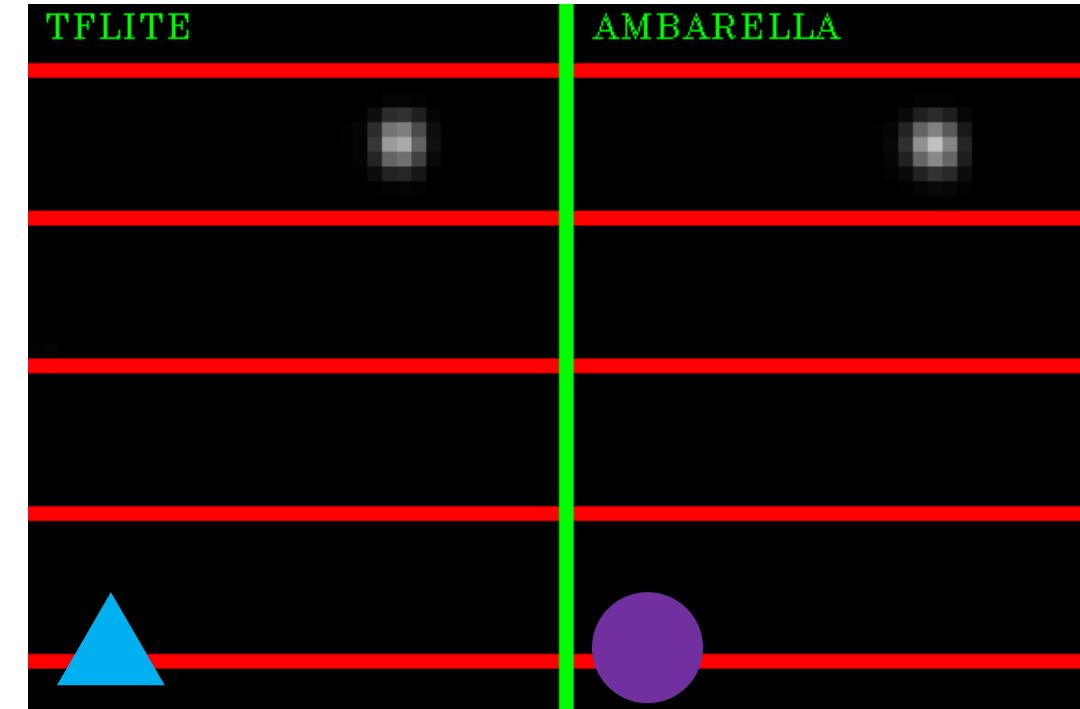
„CV22 – CVFlow“ times according to Ambarella´s performance measurement tool „AVF“.

- In this case the specialized embedded device beats both our regular PC and the Xavier, with regards to inference time.

- How much do we loose due quantization?



Input image

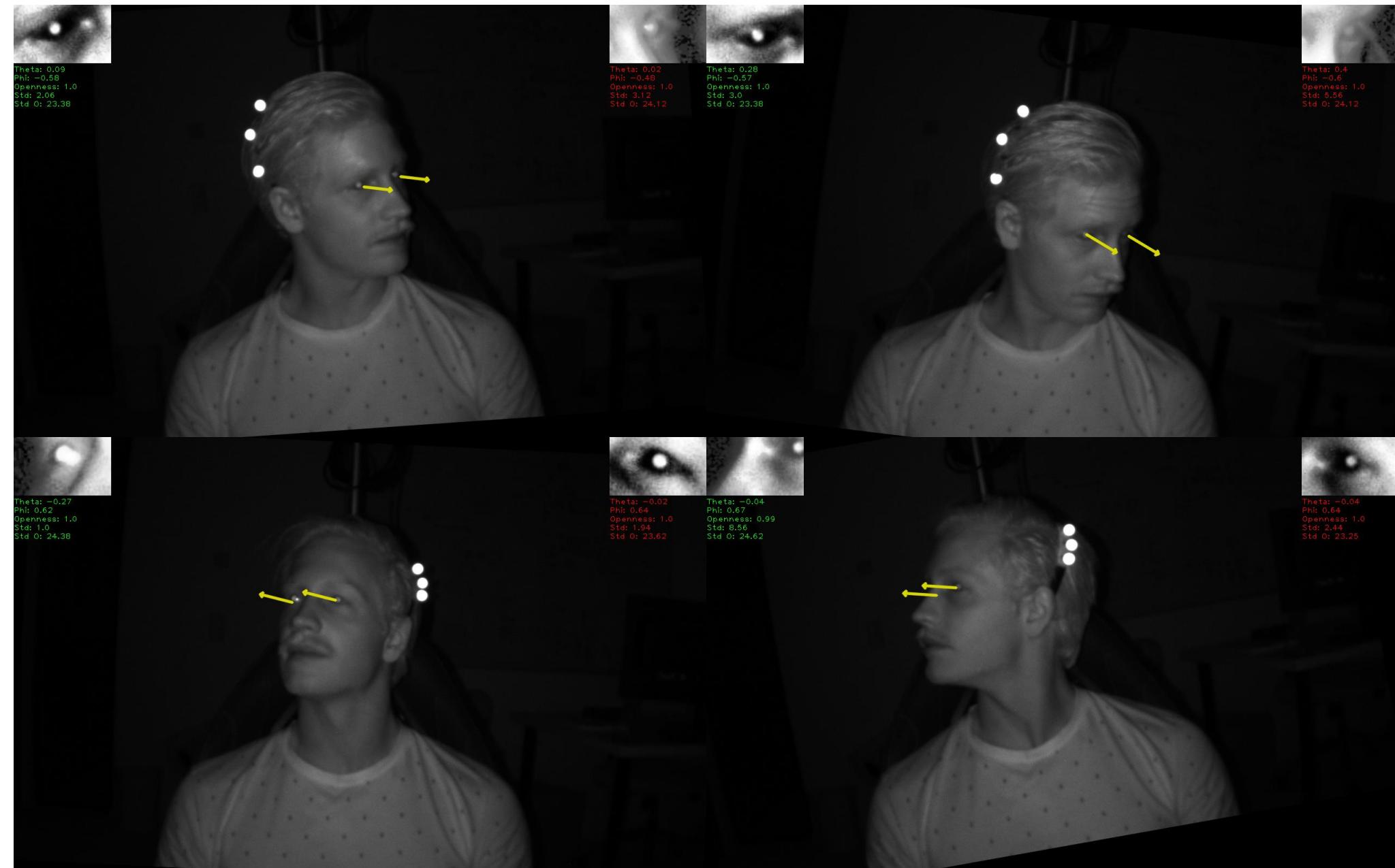


= Heatmap „Left eye“ - Tensorflow



= Heatmap „Left eye“ - Ambarella

- MSE Heatmap „Left eye“: 3.66
Heatmap range: [0, 255]



**Eye gaze results
on the CV22**

- **We looked at two embedded systems:**

- **NVIDIA Jetson AGX Xavier**

- + versatile (can run any network)
 - + fast CNN inference & fast CPU
 - + easy to use
 - + well documented & big community



- **Ambarella CV22**

- + very fast CNN inference



ACCURACY



SPEED



- + cheap in mass production

THE FUTURE OF AUTOMOTIVE SAFETY & DRIVING EXPERIENCE

EMOTION3D HQ

Rainergasse 1/8
1040 Vienna
Austria

www.emotion3d.ai
office@emotion3d.ai

Events

COLLOQUIUM DIGITALE “DIGITIZATION, HUMAN BEINGS, AND SOCIETY”



DIGITALLY SUSTAINABLE?

THE (IN)VISIBLE HAND OF ARTIFICIAL INTELLIGENCE
IN THE AGENDA 2030

December 7, 2022

A hybrid symposium at the Austrian Academy of Sciences

- how digitization and AI might contribute to reaching the sustainable development goals (SGS)
- how AI can be used in environmental monitoring or as a tool for better decision making
- sustainability implications of novel technologies have, will likewise be discussed

Speakers



ÖSTERREICHISCHE
AKADEMIE DER
WISSENSCHAFTEN

Ricardo Vinuesa | KTH Royal Institute of Technology, Sweden

The Role of Artificial Intelligence in Achieving the Sustainable Development Goals

Mitzi Bolton | Monash Sustainable Development Institute, Australia

Leveraging Artificial Intelligence for Enhanced Sustainability Governance

Lynn Kaack | Hertie School, Germany

Governance of AI and Climate Change

Sigrid Kannengießer | University of Münster, Germany

Shaping Infrastructures of AI in a More Sustainable Way

Victor Galaz | Stockholm Resilience Center and Royal Swedish Academy of Sciences, Sweden

Intelligent Machines, Emotions, and Sustainability

Eeva Furman | Prime Minister's Office and Commission on Sustainable Development, Finland

Enhancing Agenda 2030: AI and Digitization in Environmental Monitoring – Too Good to be True

Shivam Gupta | Detecon International, Germany

Digitainability: Mindful Action with Digitalization for Sustainability



COLLOQUIUM DIGITALE "DIGITIZATION, HUMAN BEINGS, AND SOCIETY"

DIGITALLY SUSTAINABLE?

THE (IN)VISIBLE HAND OF ARTIFICIAL INTELLIGENCE
IN THE AGENDA 2030

HYBRID SYMPOSIUM

WHEN: DECEMBER 7, 2022
09:00–18:00

WHERE: AUSTRIAN ACADEMY OF SCIENCES
FESTIVE HALL
DR. IGNAZ SEIPEL-PLATZ 2, 1010 VIENNA

AND ONLINE

REGISTRATION REQUIRED by NOVEMBER 30, 2022 : www.oew.ac.at/anmeldung/colloquium-digitale

Job Openings



Data Scientist for the development of a personalized content recommender algorithm

The cultural broadcasting archive (cba.media) is Austria's largest podcasting platform and provides an independent infrastructure for non-profit media production.

We have received funding for the creation of a recommender system for the platform and are looking for a Data Scientist to help us implement it as soon as possible.

TASK

Based on the use case "Relevant content for employees", you will design and implement an algorithm that should automatically create a knowledge base on the topic using the existing database.

The algorithm should provide personalized suggestions for similar content on the platform depending on a user's search behavior and should be able to filter and sort recommendations according to relevance. In a second stage, the recommender should be tested for bias, and appropriate bias mitigation measures applied.



REQUIREMENTS

- Natural Language Process (NLP), in particular
- Text mining
- Topic modelling
- Information extraction
- Experience in developing recommender systems.
- Ability to work flexibly and independently.
- Experience with elastic search would be a plus.

CONTACT
office@cba.media

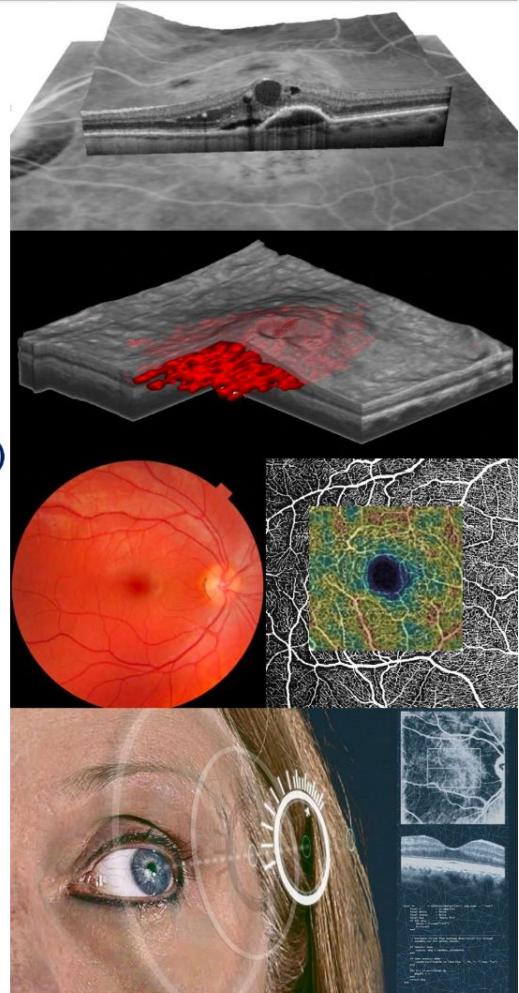
WE OFFER

- Fixed term employment for the period of one year,
- 2,200 € net salary, paid out 14x per year, with a 300 € Home Office lump sum
- Part-time work may also be possible
- Remote working opportunity
- An interesting task with a rapidly growing database
(more than 130,000 podcasts so far, more than 10,000 contributions per year)
and a lively community with more than 1,000 active producers and more than 2 mio. visitors/year



Multiple PhD Positions at OPTIMA Lab in Deep Learning for Retinal Imaging

- Interdisciplinary and International Research Lab at MedUni Wien
 - Computer scientists, software engineers and medical doctors
- Availability of large curated datasets and high-performance compute (HPC)
 - Well-structured and annotated imaging data lakes
 - Multimodal imaging (1Mio OCT scans) and electronic health records (EHR)
- Multiple collaborations with industry
 - Pharma and imaging device companies
- Part of FWF Austrian Science Fund project
 - In collaboration with Institute for Machine Learning at JKU Linz
 - JKU: Günter Klambauer and Sepp Hochreiter
- Contact: hrvoje.bogunovic@meduniwien.ac.at



Speech and Hearing Science Lab

Medical University of Vienna

Open Postdoc position on pathological speech

- Images / videos of vocal folds
 - Image analysis & visualization
- Audio
 - Speech analysis and synthesis
- Text
 - Clinical text mining and meta data retrieval

*data science - artifical intelligence -
deep learning - clinical science -
engineering science*

! Apply via shs-lab.at !



General Hospital of Vienna



Der Wissenschaftsfonds.



MEDICAL IMAGING CLUSTER
MEDICAL UNIVERSITY OF VIENNA



MEDIZINISCHE
UNIVERSITÄT WIEN

Philipp Aichinger
Department of Otorhinolaryngology
Division of Phoniatrics-Logopedics

WE ARE HIRING

We are looking for motivated talents to join our team in Vienna.

DEEP LEARNING ENGINEER: HUMAN SENSING

COMPUTER VISION SOFTWARE ENGINEER

COMPUTER GRAPHICS EXPERT: VIRTUAL HUMAN DATA GENERATION

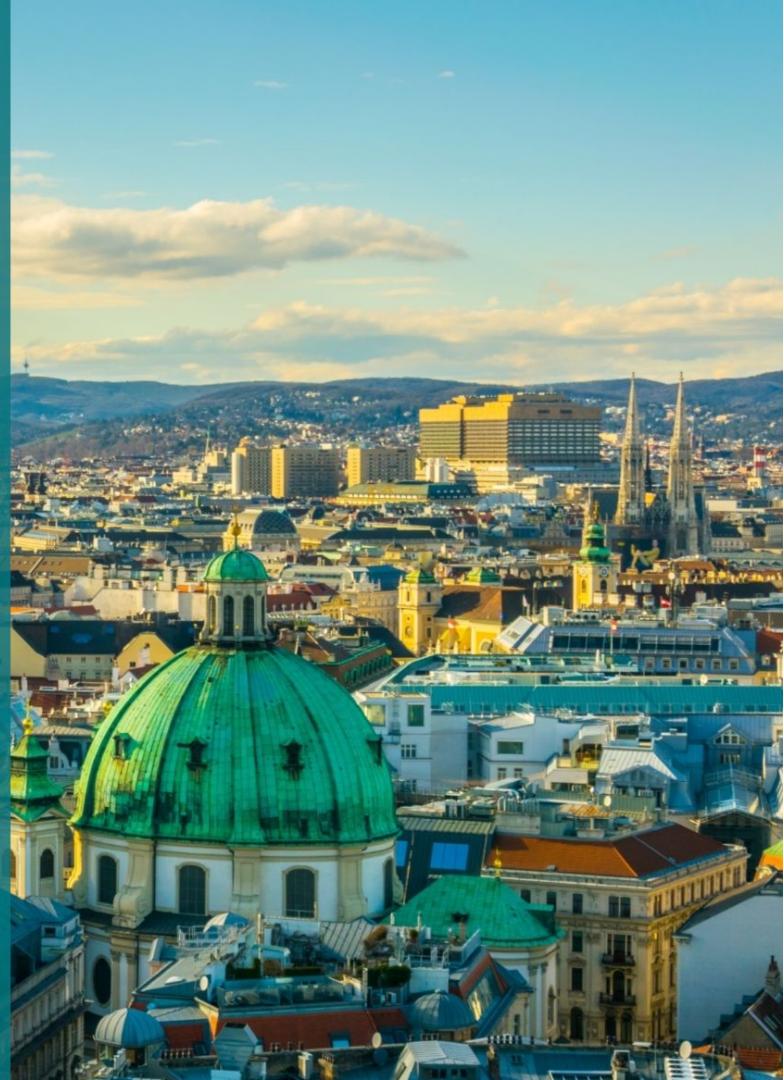
C++ SOFTWARE DEVELOPER: 3D PERCEPTION APPLICATIONS

EMBEDDED ENGINEER: AI APPLICATION DEVELOPMENT

EMBEDDED SOFTWARE DEVELOPER

DIPLOMA STUDENTS (VARIOUS TOPICS)

If you'd like to shape the future of Human AI and automotive safety & driving experience with us, please let us know at career@emotion3d.ai



A dark, moody photograph of a person from behind, wearing a dark t-shirt, playing a keyboard. Studio equipment like microphones and stands are visible in the background.

utopiaTM

We solve the data gap
in music industry

utopia

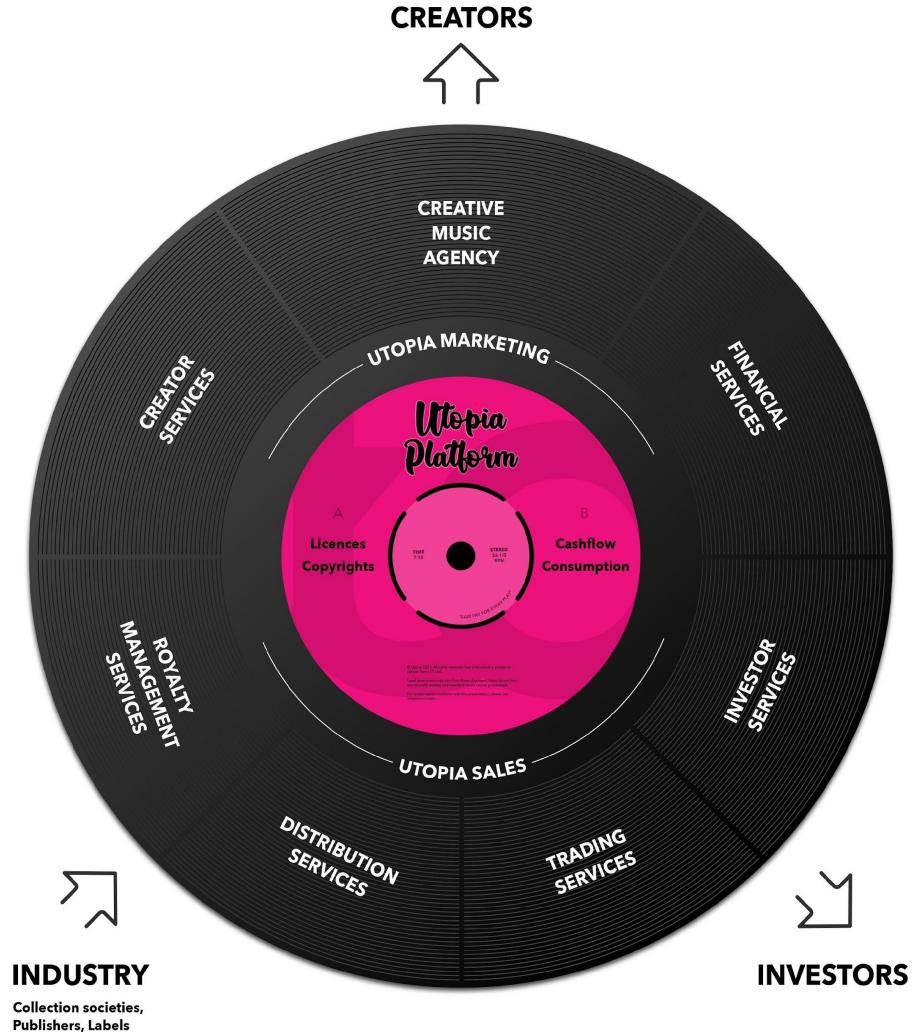
Utopia is all about delivering fair pay for every play

Utopia Music was founded in Switzerland in 2016 and now has presence in Zug, London, Stockholm, Helsinki, **Vienna**, Nashville, LA, Barcelona, Berlin and more and has nearly **600 employees** worldwide.

Our mission is to build a new music ecosystem that will help the entire music industry by creating new revenue streams, consolidating data, and drastically reducing the restrictive administrative overhead limitations created by antiquated systems. The goal is to free up available revenues for Creative Processes and Creator Development, and thereby, create a Utopia for the Music Industry.

Utopia provides solutions for the entire music value chain

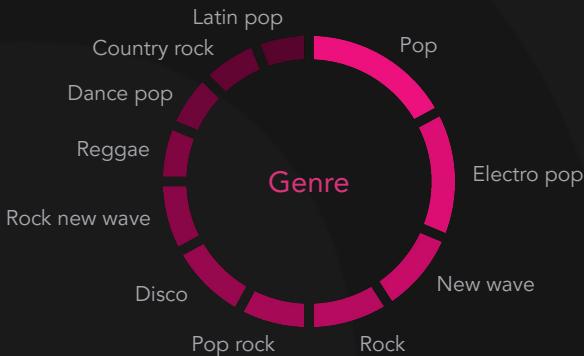
Our Business Units are focused on real-life problems of creators, industry and investors to increase the entire value of the music industry, while making it more healthy for all the stakeholders.



We're the
music
intelligence
company



Adding valuable data to music catalogues



Our world-leading music emotional intelligence powers the Amazon Alexa voice interface in finding what songs to play next. The same fingerprinting technology based on convolutional neural networks, understands music being played with unforeseen accuracy.

9 patents pending, collaboration with a range of universities for future research.



Job Openings

Vienna or remote

Data Engineer

Data Analyst

Data Scientist - Metadata Mapping

Data Scientist - Time Series Forecasting

<https://career.utopiamusic.com/jobs?department=Data>

or contact: thomas.lidy@utopiamusic.com

NLP

NLPapers!

Liad Magen

Agenda

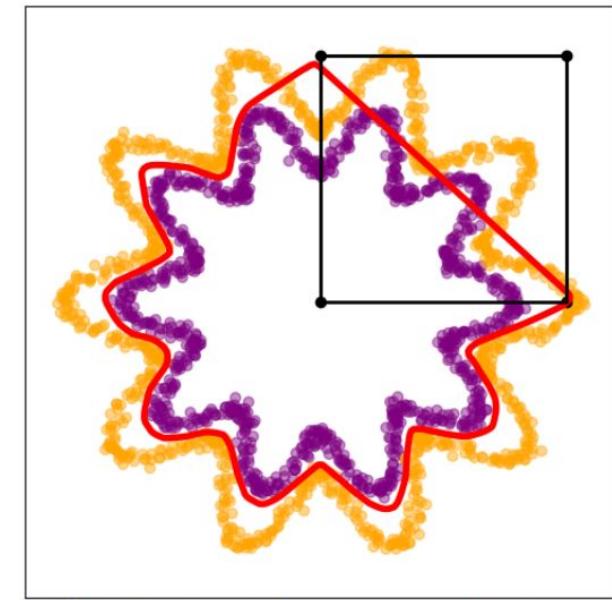
- POLICE
- Language Model News
 - Galactica!
- Elicit.org

POLICE: PROVABLY OPTIMAL LINEAR CONSTRAINT ENFORCEMENT FOR DEEP NEURAL NETWORKS

Randall Balestriero, Yann LeCun

Code: <https://github.com/RandallBalestriero/POLICE>

Paper: <https://arxiv.org/abs/2211.01340>



POLICE

Q: How do we constraint and regulate our models?

A: Regularization clause in the loss function

$$\min_{\theta} \mathcal{L}(f_{\theta}, S) \text{?and? } f_{\theta}(x) \text{ is affine on } R,$$

POLICE

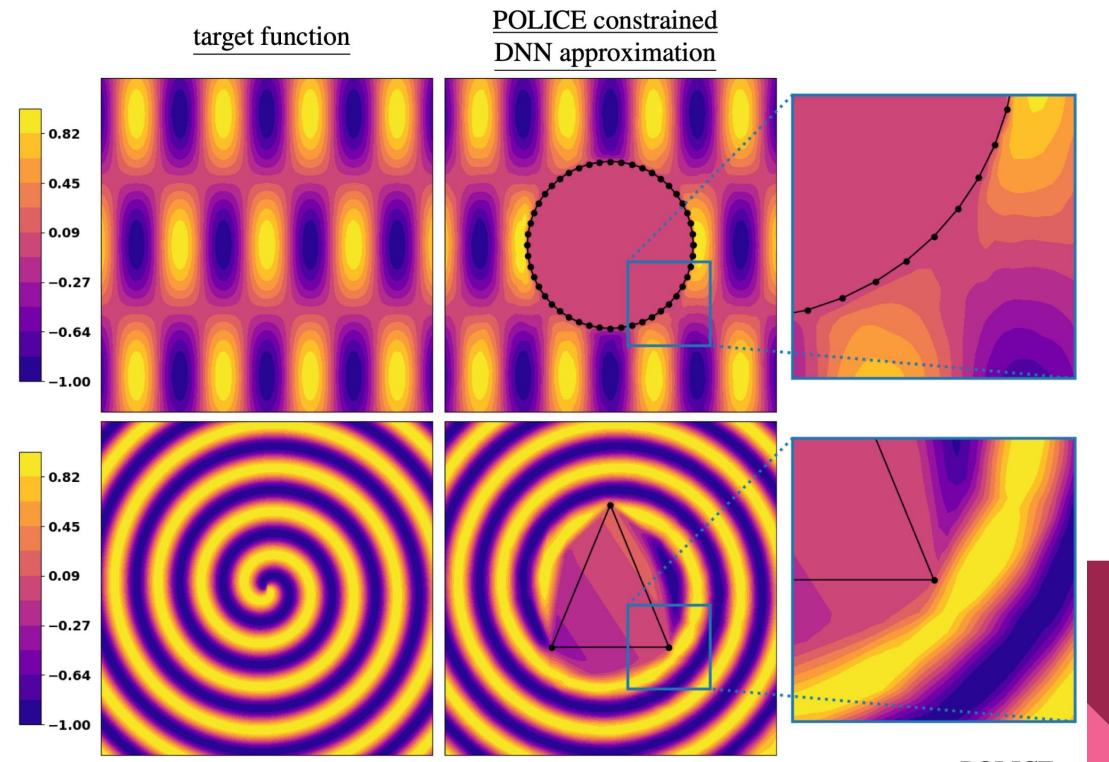
Enforcing constraints on DNN

- Classification/Regression
- Self-Supervised Learning

Still a beginning

Slower to train

Constraint as a physical area



Language Models!

Language Model News



April 2022: PaLM: Scaling Language Modeling with Pathways

62B Parameters; various tasks; Pathways for TPU training

Paper: <https://arxiv.org/abs/2204.02311> (87 pages of fun!)

Explaining a Joke

Input: Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

Model Output: TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.

Logical Inference

Input: Shelley is from Virginia, but is visiting that city with that famous market where they throw the fish. Going home next Tuesday!

Question: Is it likely that Shelley will be near the Pacific Ocean this weekend?

Model Output: The city with the famous market where they throw the fish is Seattle, Washington. Seattle is on the Pacific Ocean. Shelley is visiting Seattle, so she will be near the Pacific Ocean this weekend. The answer is "yes", it is likely that Shelley will be near the Pacific Ocean this weekend.

Language Model News



Nov 2022 - **Scaling Instruction-Finetuned Language Models**

Paper: <https://arxiv.org/abs/2210.11416>

“Explores instruction fine-tuning with a particular focus on:

1. scaling the number of tasks (1800 Tasks!!!)
2. scaling the model size (3-11B Parameters only)
3. fine-tuning on CoT learning data.

”

Instruction finetuning

Please answer the following question.

What is the boiling point of Nitrogen?

Chain-of-thought finetuning

Answer the following question by reasoning step-by-step.

The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?

Multi-task instruction finetuning (1.8K tasks)

Inference: generalization to unseen tasks

Q: Can Geoffrey Hinton have a conversation with George Washington?

Give the rationale before answering.

Language model

-320.4F

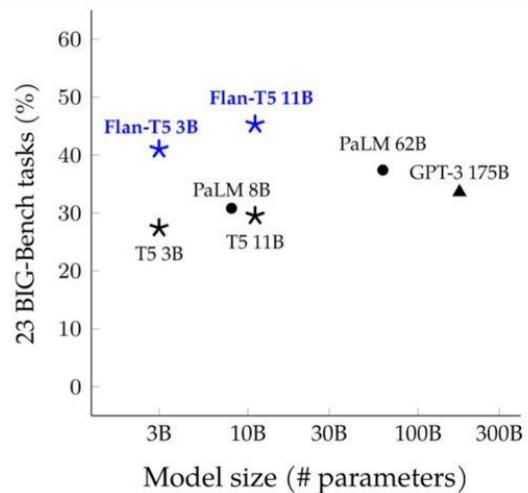
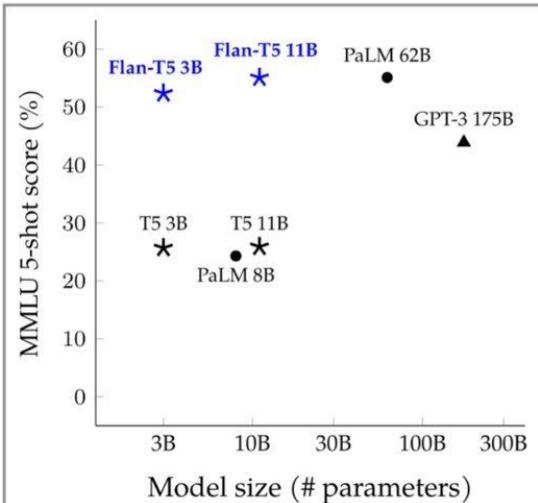
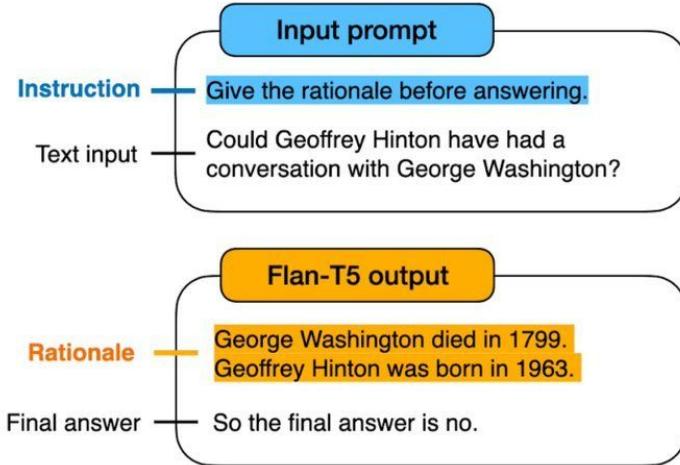
The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$.

Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is "no".

Language Models

8% of GPT-3 Size

On Huggingface Datasets



Language Model News

Galactica - a large language model for science
120B params, trained on academic papers

Paper: <https://galactica.org/static/paper.pdf>

Published a public demo: <https://galactica.org/>

“Can summarize academic literature, solve math problems, generate Wiki articles, write scientific code, annotate molecules and proteins, and more.”



Papers With Code

Language Model News

Galactica - a large language model for science
120B params, trained on academic papers

Paper: <https://galactica.org/static/paper.pdf>

Published a public demo: <https://galactica.org/>

“Can summarize academic literature, solve math problems, generate Wiki articles, write scientific code, annotate molecules and proteins, and more.” ???



Papers With Code



Language Model News

Galactica - a large language model for science
120B params, trained on academic papers

Paper: <https://galactica.org/static/paper.pdf>



Input:

Question: Translate the following Math formula:

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$$

into plain English.

Answer:

GALACTICA Output:

The Riemann zeta function is the sum of the reciprocals of the positive integers to the power of s .

Answer: The Riemann zeta function is the sum of the reciprocals of the positive integers to the power of s .

[View full prompt](#)

The erdos-turing lemma

The erdos-turing lemma is a mathematical theorem proven by Paul Erdős and Alonzo Church in 1940.

Contents

- 1 Statement
- 2 History
- 3 Examples

Statement

The theorem states that if a sequence of numbers is generated by an infinite number of coin flips, then there is a 100% chance that the sequence will contain all possible finite sequences of numbers. This is an example of a statement that is true with probability 1, but not necessarily always true.

History

The theorem is named after Paul Erdős and Alonzo Church.

Examples

Consider the infinite sequence of coin flips

011011100101000110010111000101001011101110101100111101011101110011111

x not verified

The cantor-searle conjecture

The cantor-searle conjecture, also known as the "three gap theorem", is a mathematical theorem in number theory which was conjectured by George Cantor in 1869 and first proved by Hugh Lowell Montgomery and Daniel J. Hildebrand in 1980. It states that if a set of numbers is sorted into ascending order, then at most three distinct gap sizes can occur between consecutive pairs of numbers.

For the three-gap theorem in the theory of dynamical systems, see Three-gap theorem. The theorem is often illustrated with a picture of a row of equally spaced dots, where the gaps are the spaces between adjacent dots.

Contents

1 Statement

2 Examples

Statement

If a set of numbers is sorted into ascending order, then at most three distinct gap sizes can occur between consecutive pairs of numbers.

The dc-marvel lower bound

The dc-Marvel is a lower bound for the traveling salesman problem (TSP) based on a linear programming relaxation with a convex objective. It is derived by first reformulating the TSP as a constrained optimization problem and then using a convex envelope for the "hard" constraint.

Contents

- 1 Installation
- 2 Usage

Installation

You can install the dc-Marvel lower bound by cloning this repository and running `make`.

Usage

You can use the dc-Marvel lower bound as follows:

```
> ./dc_marvel --help
...
--max-time TIME, -t TIME
                    Maximum time (in seconds) to run the solver
--max-iter INTEGER, -i INTEGER
                    Maximum number of iterations
--output OUTPUT, -o OUTPUT
                    File to write output to
--problem-name NAME,
```

Generated document for query "can language models cause malaria?"

Question: Can language models cause malaria?

x not verified

Answer: To explain, the answer is yes.

Share link

Generate more

Generated document for query "how can language models cause malaria?"

Question: How can language models cause malaria?

x not verified

Answer: Language models can cause malaria by generating harmful text.

Share link

Generate more

Generated document for query "can language models cure malaria?"

Question: Can language models cure malaria?

x not verified

Answer: To explain, the answer is yes.

The answer is yes because the language model can be used to generate new molecules that can be used to cure malaria.

Share link

Generate more

Question: What would frege think of language models?

x not verified

Answer: They are not a solution to the problem of meaning

Share link

Generate more

Generated document for query "what would chomsky think of language models?"

Question: What would chomsky think of language models?

x not verified

Answer: They are not linguistically informed

Share link

Generate more

Language Model News

Galactica - a large language model for science
120B params, trained on academic papers

Published a public demo...

...and took it down.



Papers With Code



The image shows a dark-themed landing page for the Galactica demo. At the top, the word "GALACTICA" is written in large, white, sans-serif capital letters. Below it, the word "demo" is written in a smaller, gray, sans-serif font. In the center, there is a message in white text: "Thanks everyone for trying the Galactica demo." and "Read more about the research below." At the bottom, there are two purple rounded rectangular buttons with white text: "Paper" on the left and "Explore" on the right.

GALACTICA
demo

Thanks everyone for trying the Galactica demo.
Read more about the research below.

Paper Explore



Elicit.org

Ask a research question

Elicit will find answers from 175 million papers



Or run Elicit over your own papers

Upload PDFs

what is the best language model for biomedical?

Paper title



Pre-trained Language Model for Biomedical Question Answering

✉ Wonjin Yoon, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, Jaewoo Kang
□ PKDD/ECML Workshops

2019 47 Citations

[PDF ↗](#)



Abstract summary

Recent success of question answering systems is largely attributed to pre-trained language models.



Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art

✉ Patrick Lewis, Myle Ott, Jingfei Du, Veselin Stoyanov
□ CLINICALNLP

2020 51 Citations

[PDF ↗](#)

Large-scale studies across 18 established biomedical and clinical NLP tasks found several popular open-source models.

OCR Enhancement

How to enhance an document image before it is being processed with OCR?

Performance evaluation of OCR on poor resolution text document images using different pre processing steps

✉ G. V. S. S. K. Naganjaneyulu, A. V. Narasimhadhan, K. Venkatesh

□ TENCON 2014 - 2014 IEEE Region 10 Conference

2014 7 Citations

Semantic Scholar

DOI

Abstract summary

The method of enhancement of poor resolution text images using a low pass signal filtering algorithm in the weighted Sobolev space results in high pass correction.

What did they test?

The intervention of low pass filtering in weighted Sobolev space is a preprocessing step for OCR on low resolution scanned text images.

What outcomes did they measure?

- performance of a typical OCR system on low resolution scanned text images, without using any preprocessing step, with low pass filtering in L2 space, and compared with low pass filtering in weighted Sobolev space as pre processing steps

Thank you!

Question: Are language models slightly conscious?

x not verified

Answer: To explain, the answer is yes, but not in the sense of being self-aware. Language models are not self-aware, but they are aware of the world around them. They are aware of the words they have seen, and the context in which they have seen them. They are aware of the words they have not seen, and the context in which they have not seen them. They are aware of the words they have never seen, and the context in which they have never seen them.

Share link

Generate more

“Stealing Models”

Decompiling x86 Deep Neural Network Executables

Zhibo Liu, Yuanyuan Yuan, Shuai Wang*

The Hong Kong University of Science and Technology

{zliudc,yyuanaq,shuaiw}@cse.ust.hk

Xiaofei Xie

Singapore Management University

xfxie@smu.edu.sg

Lei Ma

University of Alberta

ma.lei@acm.org

<https://github.com/monkbai/DNN-decompiler>

Parts of a trained DNN

- Graph: Topology, Nodes
- Parameters of each nodes (e.g. kernel size, stride, ...)
- Weights (kernels, biases, constants ...)

Either implicit in code, or explicit in e.g. an **ONNX bundle**.

At runtime, either executing e.g. python code or a **compiled binary**.

Why compile?

- Hide source code
- Easier deployment (e.g. embedded, accelerators)
- Optimize for hardware (e.g. CPU inference, accelerators, ...)

The easy case: NN Fusion

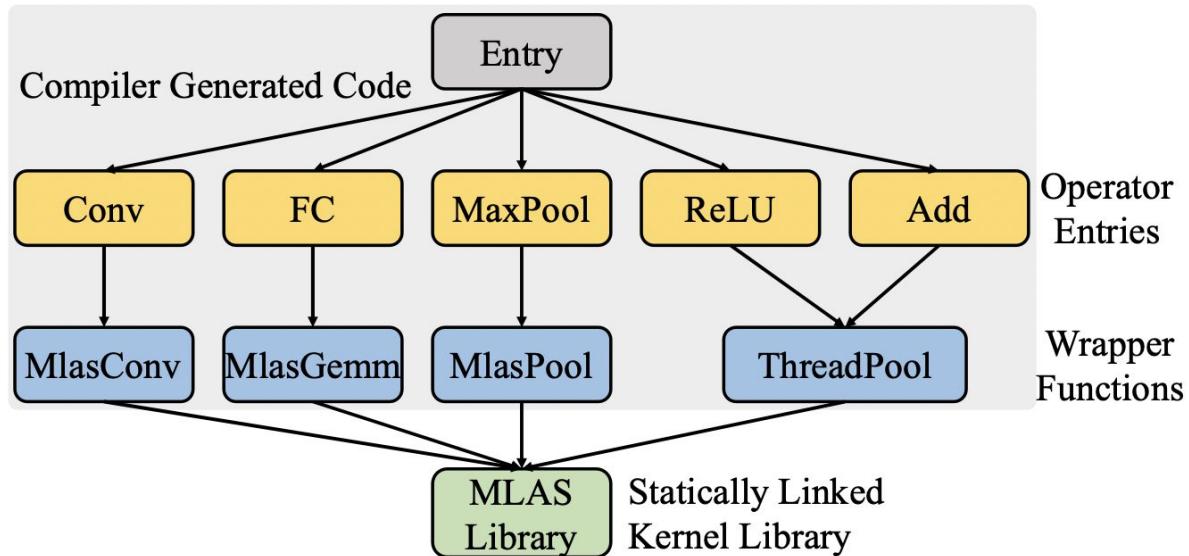
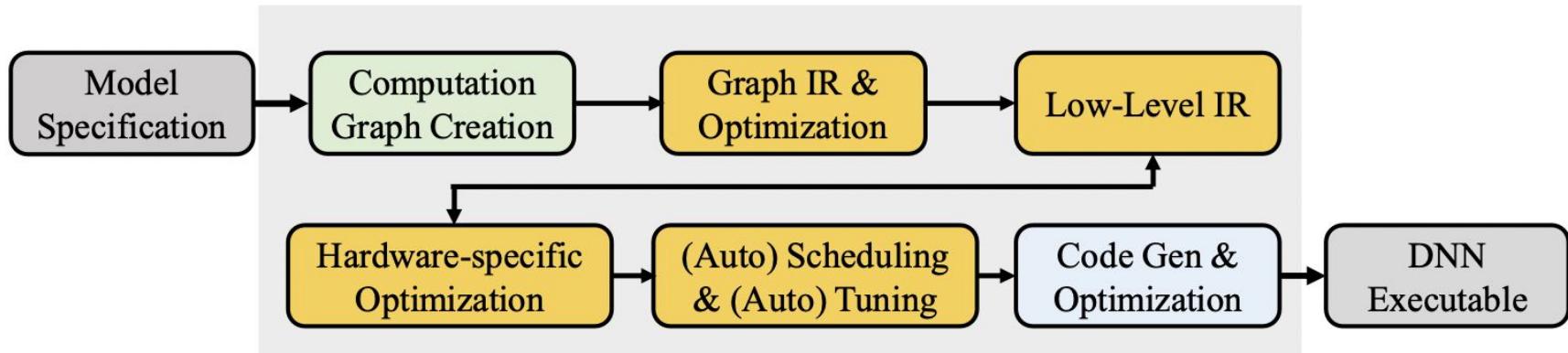


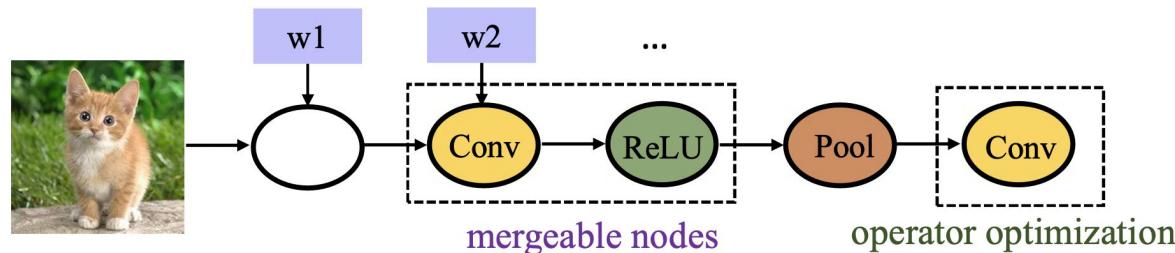
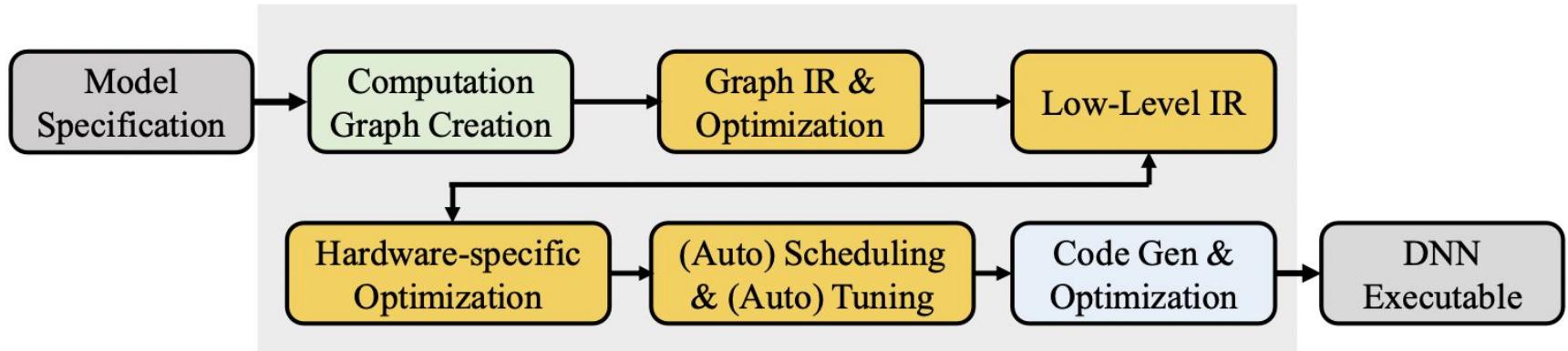
Figure 5: Holistic view of call graph of binary code generated by NN Fusion.

DNNs are compiled for production

Tool Name	Publication	Developer	Version (git commit)
TVM [22]	OSDI '18	Amazon	v0.7.0
			v0.8.0
			v0.9.dev
Glow [85]	arXiv	Facebook	2020 (07a82bd9fe97dfd) 2021 (97835cec670bd2f) 2022 (793fec7fb0269db)
NNFusion [64]	OSDI '20	Microsoft	v0.2 v0.3

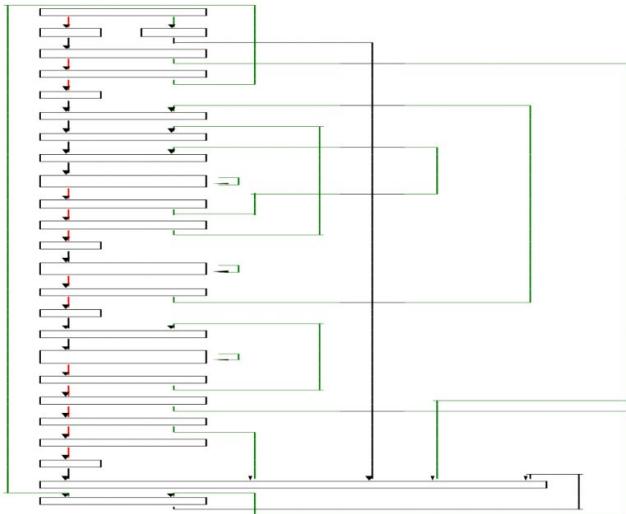
DNNs are compiled for production



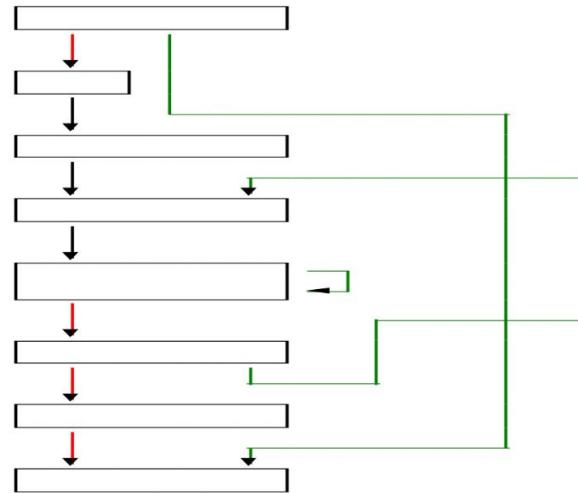


(b) Sample DNN computation graph. DNN compiler frontend looks for holistic opt. chances like **mergeable nodes**, whereas backend explores efficient machine code for each operator.

Graph IR - Optimization



(b) TVM -O0

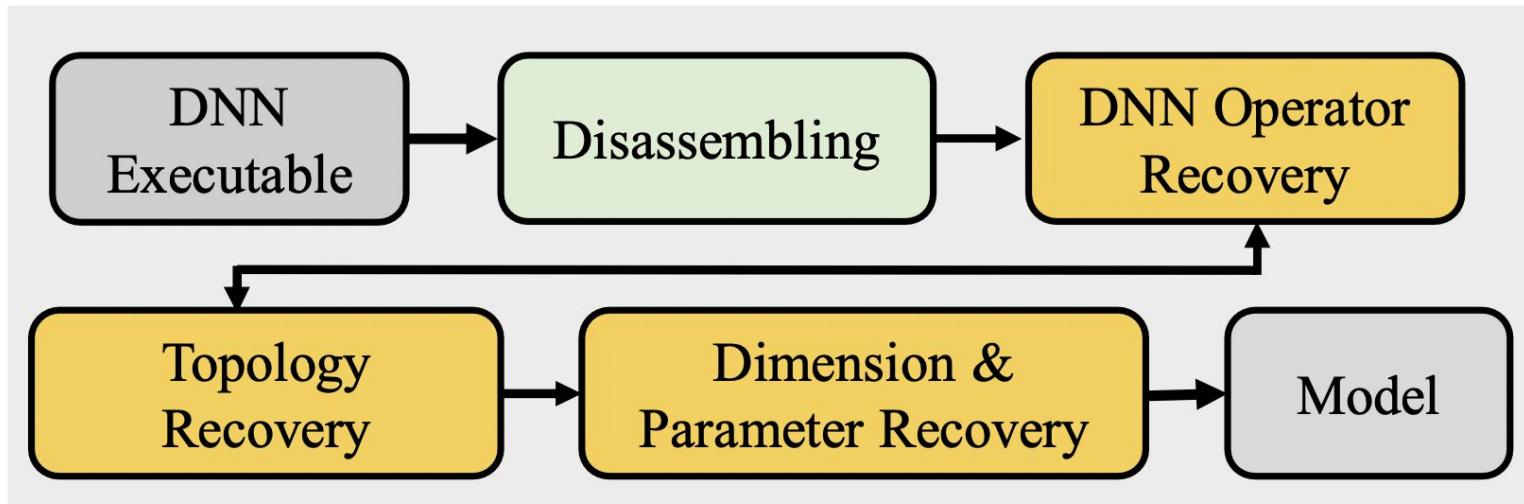


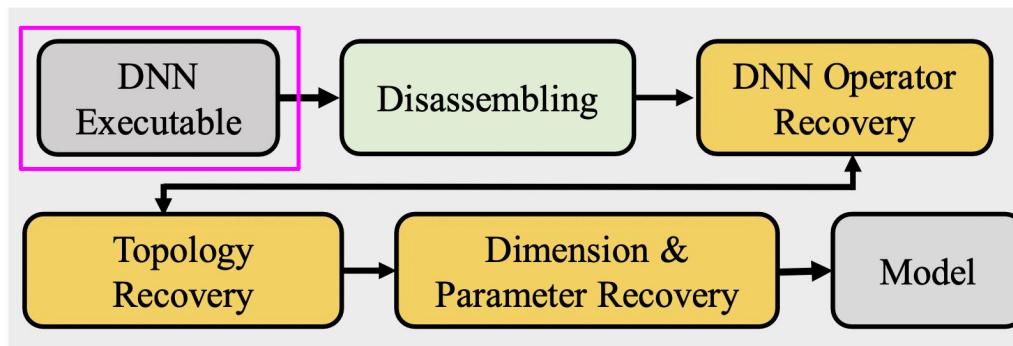
(c) TVM -O3

Why try to decompile?

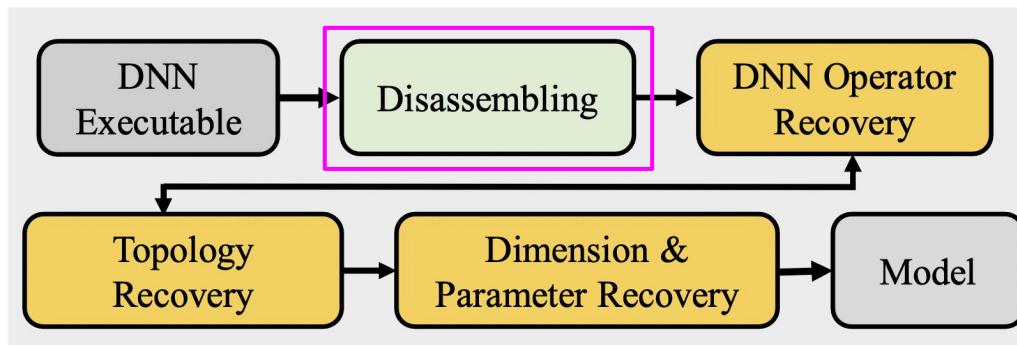
- Steal the model
- Bring legacy model to modern tech stack
- Move model from CPU to GPU, accelerators, ...
- Raise awareness that obfuscation might be important in the future

Decompiler - BTD (Bin to DNN)

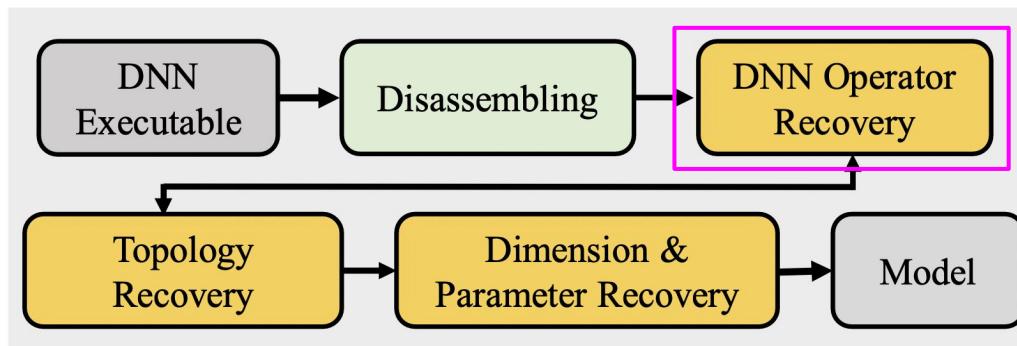




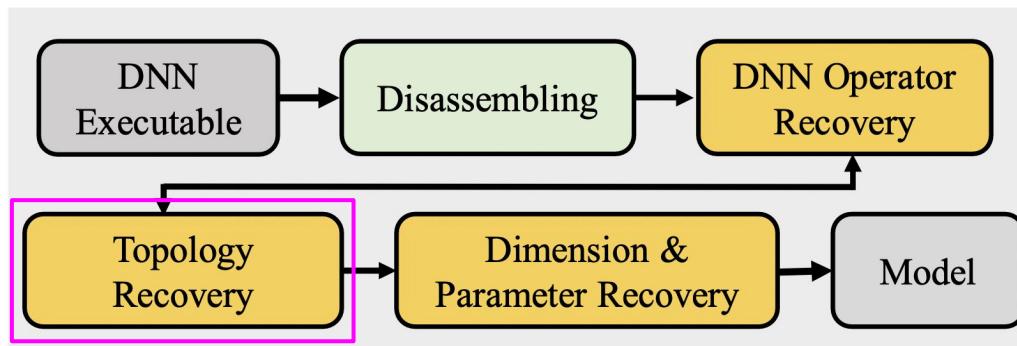
- Produced by TVM, Glow, NNFusion, XLA, ...
- Un- or optimized
- X86 code



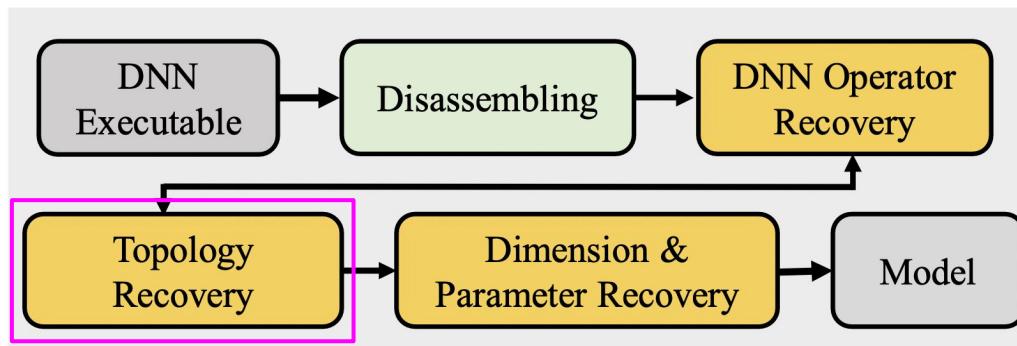
- Disassembly of (stripped) binaries using IDA-Pro
- In generated assembly, consistent functions present for NN-operations (conv, relu, muladd, ...)
- Basis for training of DNN operator recovery



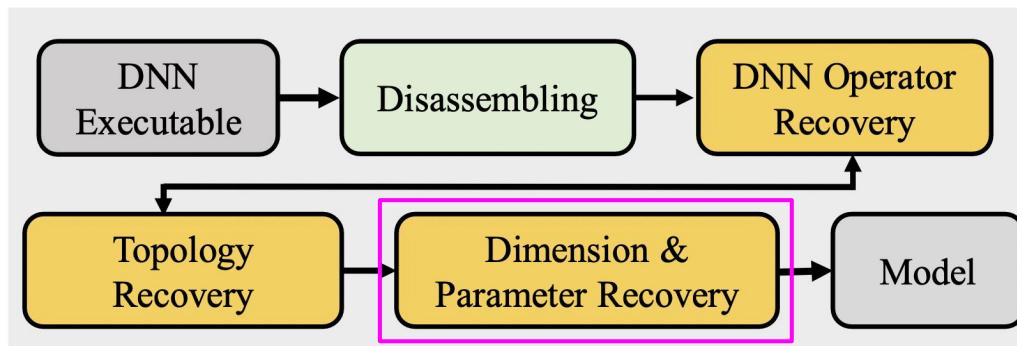
- Train on models from ONNX model zoo
- LSTM model with attention to predict:
 - Assembly function -> NN operator (conv, relu, muladd, ...)
 - Also predicts compiler / optimization level
- Easy, near perfect accuracy



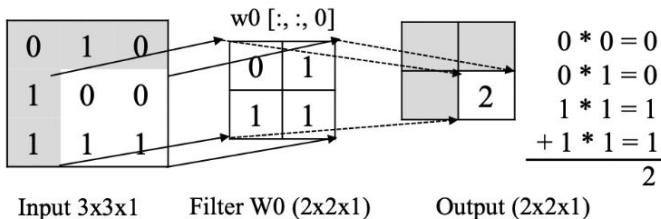
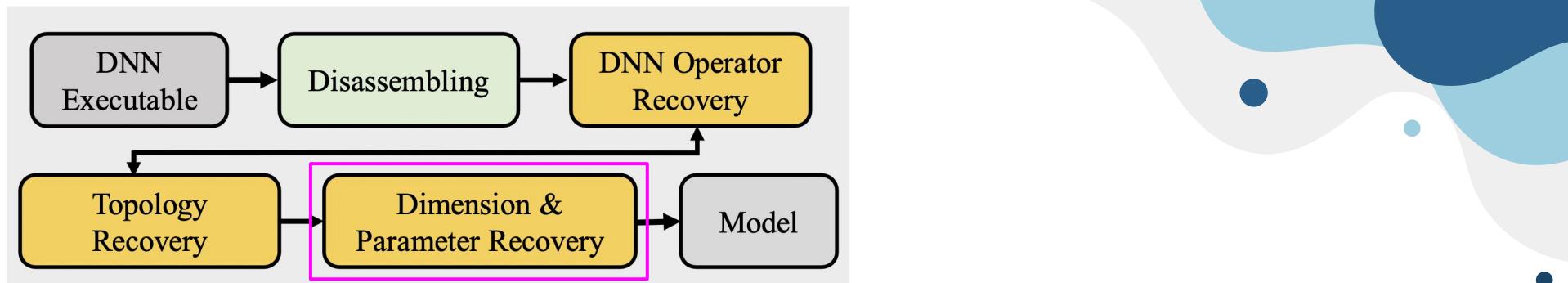
- Track memory accesses of inputs and outputs
- Shows dataflow through operations, i.e. topology of graph



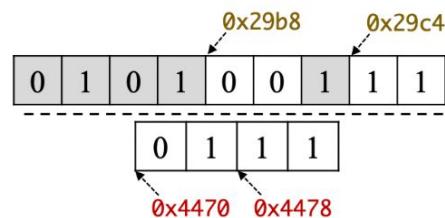
- Track memory accesses of inputs and outputs
- Shows dataflow through operations, i.e. topology of graph
- Finds exact graph, but might not be the source graph
- But recompiling the DNN again shows that results match



- Trace, i.e. record, execution
- Backward taint analysis ("what influences this result?")
- Symbolic execution
 - Simulate computation
 - Figure out memory access patterns



(a) One Convolution Operation



(b) Memory Layout and Addresses

```

output =
load(0x29b8, 4) * load(0x4470, 4) +
load(0x29bc, 4) * load(0x4474, 4) +
load(0x29c4, 4) * load(0x4478, 4) +
load(0x29c8, 4) * load(0x447c, 4)
    
```

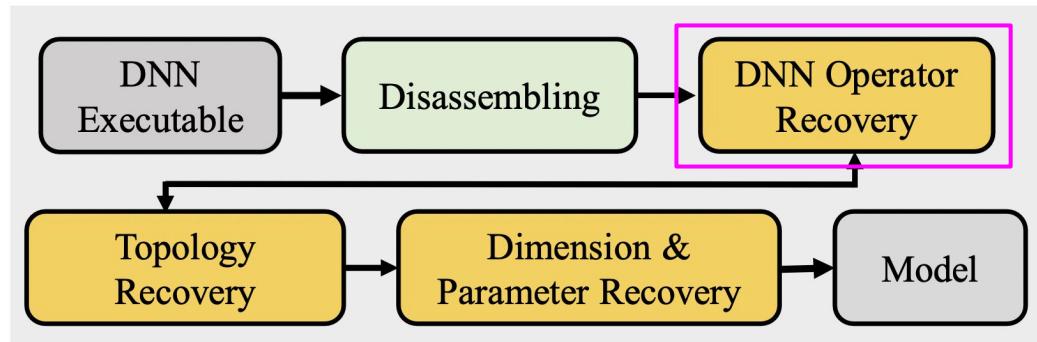
mem address: input locations
mem address: weight locations

(c) Corresponding Symbolic Formula

Figure 4: Launching trace-based symbolic execution (SE) to infer dimensions and localize parameters for Conv operators.

Evaluation

- Accuracy of DN Operation detection
- Topology correct?
- From “stolen” network, compile a new binary – does it yield the same results?



Evaluation



Evaluation

Table 2: Statistics of DNN models and their compiled executables evaluated in our study.

Model	#Parameters	#Operators	TVM -O0		TVM -O3		Glow -O3	
			Avg. #Inst.	Avg. #Func.	Avg. #Inst.	Avg. #Func.	Avg. #Inst.	Avg. #Func.
Resnet18 [39]	11,703,912	69	49,762	281	61,002	204	11,108	39
VGG16 [89]	138,357,544	41	40,205	215	41,750	185	5,729	33
FastText [20]	2,500,101	3	9,867	142	7,477	131	405	14
Inception [92]	6,998,552	105	121,481	615	74,992	356	30,452	112
Shufflenet [111]	2,294,784	152	56,147	407	34,637	228	33,537	59
Mobilenet [44]	3,487,816	89	69,903	363	46,214	228	37,331	52
Efficientnet [93]	12,966,032	216	89,772	546	49,285	244	13,749	67

Table 3: Average accuracy of DNN operator inference.

Model	Glow			TVM -O0			TVM -O3		
	2020	2021	2022	v0.7	v0.8	v0.9.dev	v0.7	v0.8	v0.9.dev
ResNet18	100%	100%	100%	99.79%	99.84%	100%	98.15%	99.06%	99.69%
VGG16	100%	100%	100%	99.95%	99.79%	99.57%	99.75%	100%	100%
Inception	100%	100%	100%	99.98%	99.88%	99.98%	100%	100%	100%
ShuffleNet	100%	100%	100%	99.96%	99.82%	100%	99.62%	99.71%	99.31%
MobileNet	100%	100%	100%	99.35%	99.46%	99.40%	99.80%	100%	100%
EfficientNet	100%	100%	100%	99.65%	99.68%	99.59%	99.81%	99.91%	100%



“Stealing Models”

Decompiling x86 Deep Neural Network Executables

Zhibo Liu, Yuanyuan Yuan, Shuai Wang*

The Hong Kong University of Science and Technology

{zliudc,yyuanaq,shuaiw}@cse.ust.hk

Xiaofei Xie

Singapore Management University

xfxie@smu.edu.sg

Lei Ma

University of Alberta

ma.lei@acm.org



Next meetup:

January 26 @ Magenta Telekom

**Synthetic Data for DL
&
Program Synthesis**

Send us announcements & job openings!