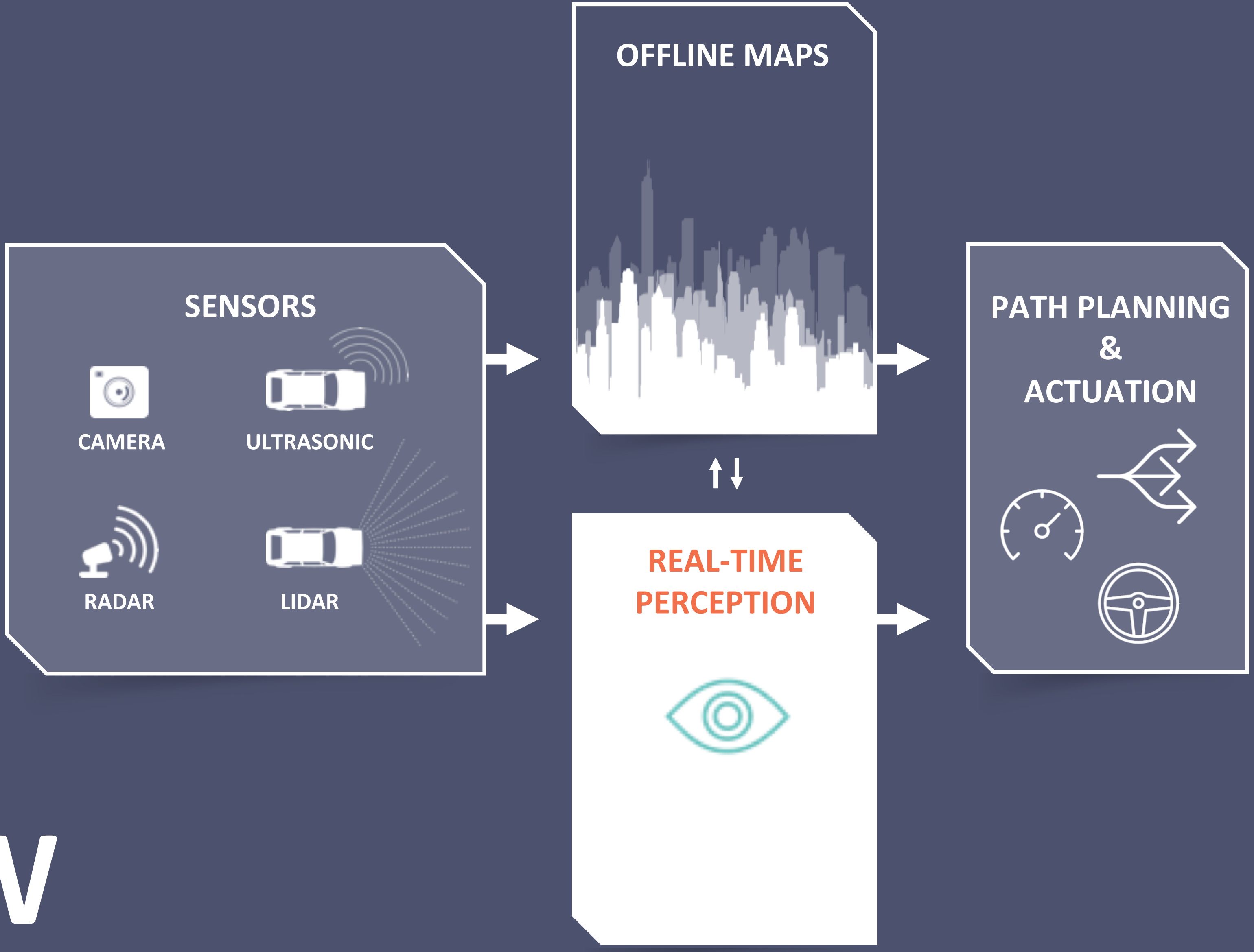
DEEPSCALE

Squeezing down the computing requirements of deep neural networks

Albert Shaw, Daniel Hunter,
Sammy Sidhu, and Forrest Iandola

Levels of automated driving

Advanced Driver Assistance (e.g. Tesla Autopilot)	LEVEL 1	Driver Assistance
	LEVEL 2	Partial Automation
	LEVEL 3	Conditional Automation
Robo-taxis, robo-delivery, …	LEVEL 4	High Automation
	LEVEL 5	Full Automation



THE FLOW

IMPLEMENTING AUTOMATED DRIVING

Deep learning is used in the best perception systems for automated driving

180x higher productivity with deep learning

Chris Urmson, CEO of Aurora: With deep learning, an engineer can accomplish in one day what would take 6 months of engineering effort with traditional algorithms.[1]

100x fewer errors with deep learning

Dmitri Dolgov, CTO of Waymo: "Shortly after we started using deep learning, we reduced our error-rate on pedestrian detection by 100x."[3]

Deep learning has become the go-to approach

Andrej Karpathy, Sr Director of AI at Tesla: "A neural network is a better piece of code than anything you or I could create for interpreting images and video."[2]

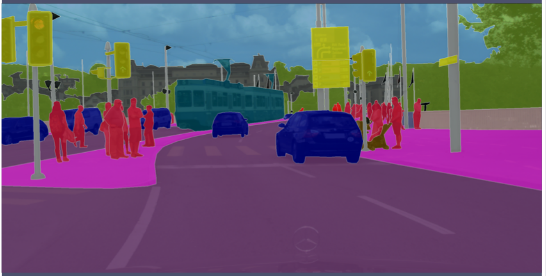
[1] https://www.nytimes.com/2018/01/04/technology/self-driving-cars-aurora.html
[2] https://medium.com/@karpathy/software-2-0-a64152b37c35
[3] https://medium.com/waymo/google-i-o-recap-turning-self-driving-cars-from-science-fiction-into-reality-with-the-help-of-ai-89dded40c63

Diverse Applications of Deep Learning for Computer Vision

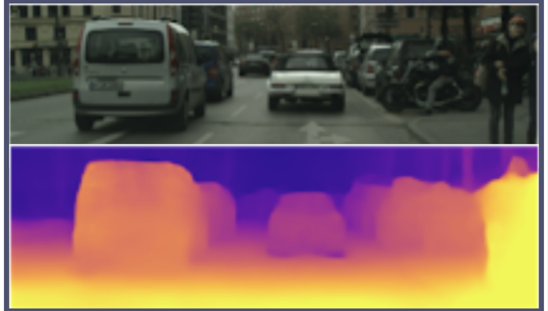
Image → Scalar or Vector

Image Classification [1]

Image → Image

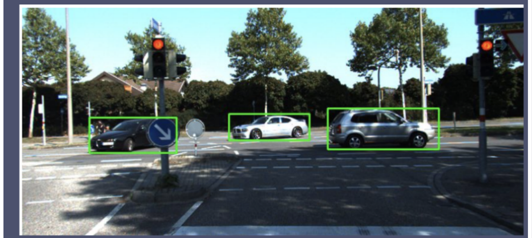


Semantic Segmentation [2]

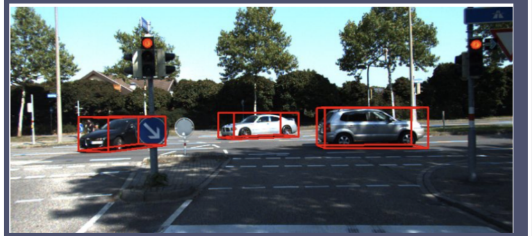


Depth Prediction [3]

Image → Boxes

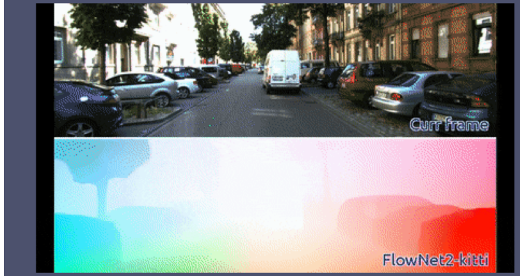


2D Object Detection [4]

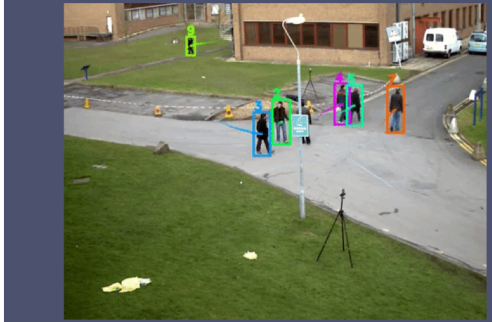


3D Object Detection [4]

Video



Optical Flow [5]



Object Tracking [6]

[1] O. Russakovsky et al. ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015.

[2] M. Cordts et al. The Cityscapes Dataset for Semantic Urban Scene Understanding. CVPR, 2016.

[3] Casser, Vincent et al. Depth Prediction Without the Sensors: Leveraging Structure for Unsupervised Learning from Monocular Videos. AAAI, 2018

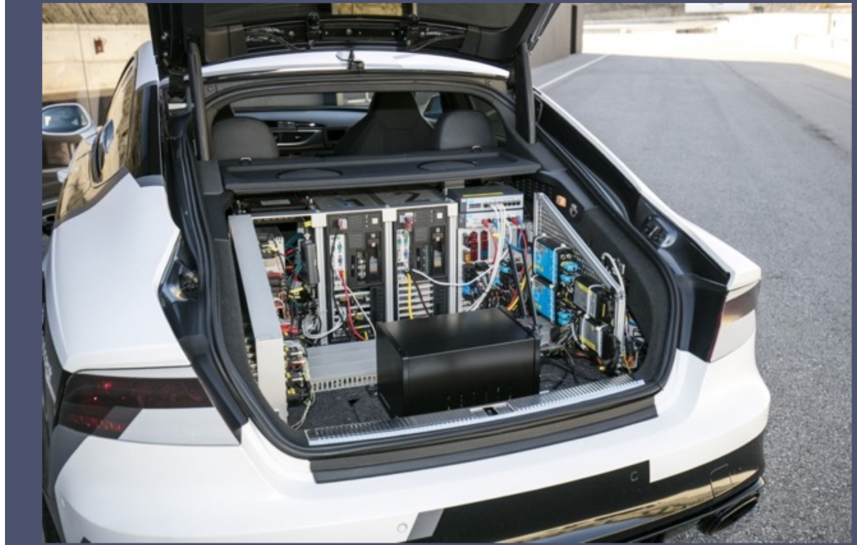
[4] Liang, Ming, et al. Multi-Task Multi-Sensor Fusion for 3D Object Detection. CVPR, 2019.

[5] Ilg, Eddy, et al. Flownet 2.0: Evolution of optical flow estimation with deep networks. CVPR. 2017.

[6] Bewley, Alex, et al. Simple online and realtime tracking. IEEE ICIP, 2016.

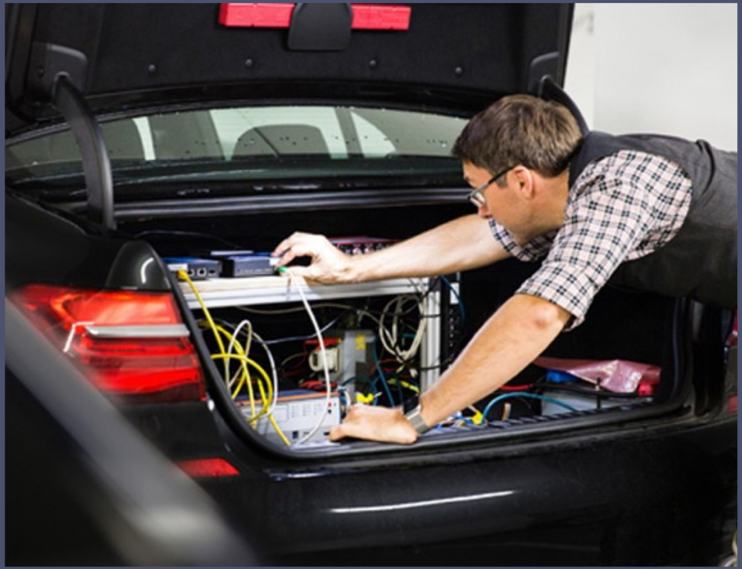
We don't just need deep learning…

We need efficient deep learning



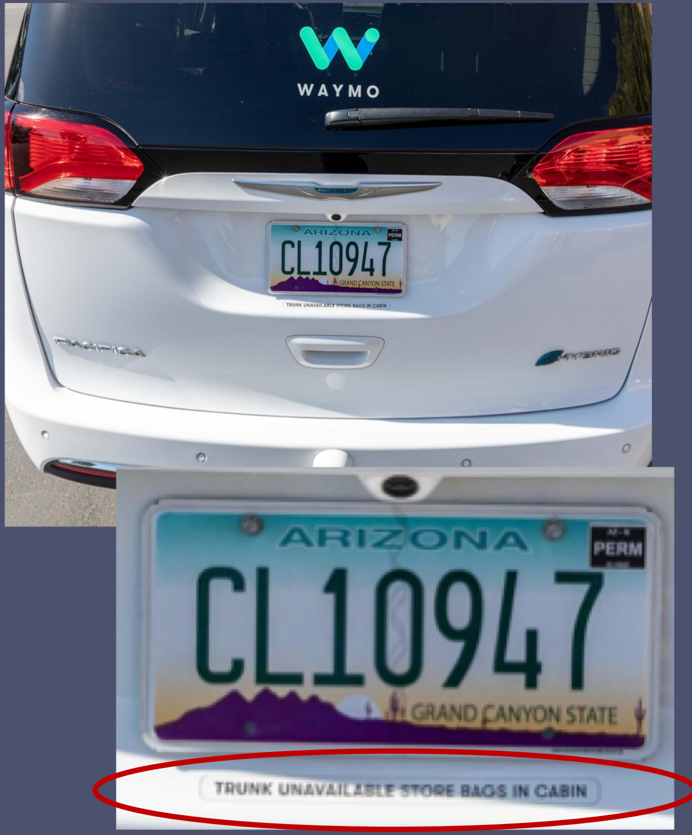
Audi

https://www.slashgear.com/man-vs-machine-my-rematch-against-audis-new-self-driving-rs-7-21415540/



BMW + Intel

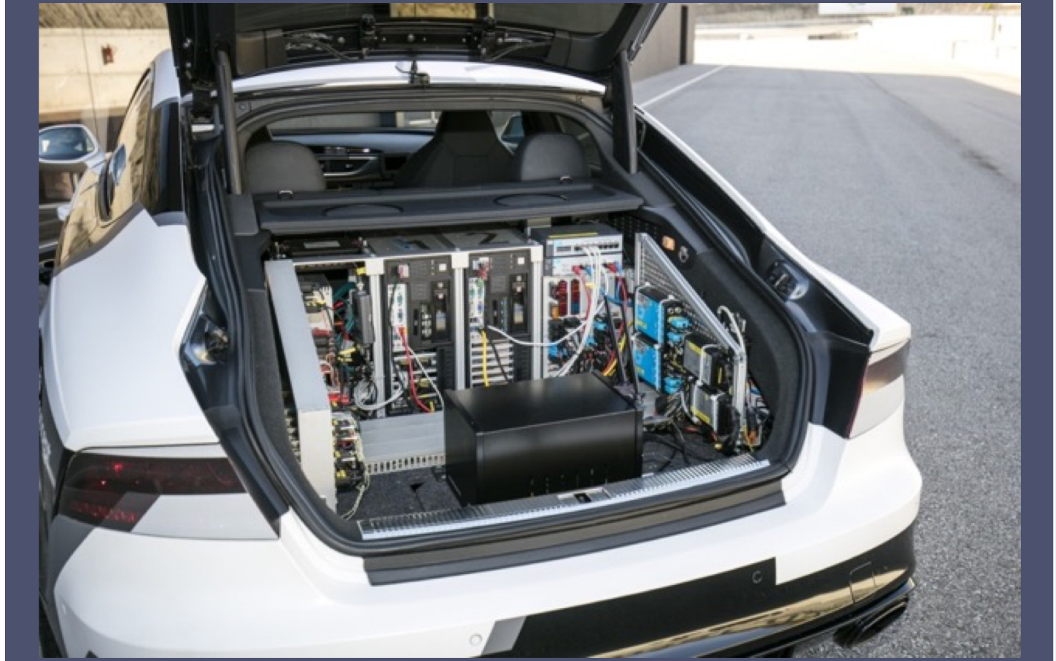
https://newsroom.intel.com/news-releases/bmw-group-intel-mobileye-will-autonomous-test-vehicles-roads-second-half-2017/



Waymo

We don't just need deep learning…

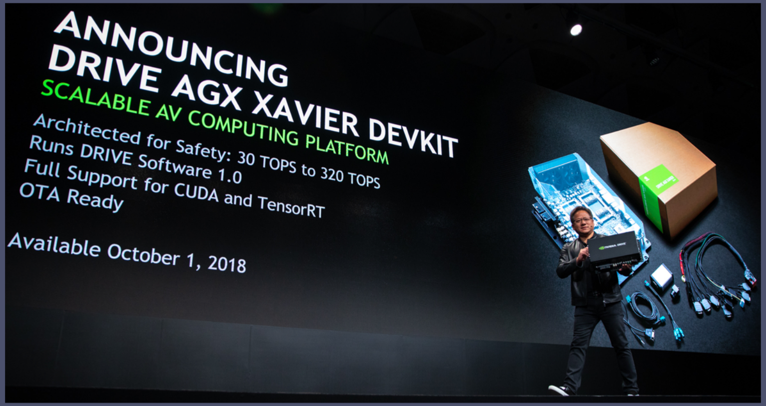
We need efficient deep learning



Trunkloads of servers cause problems:

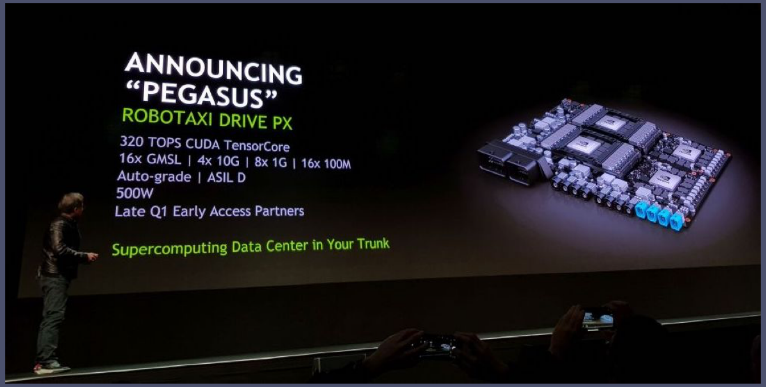
- Limited trunk space
- Cost
- Energy usage
- Reduced EV battery range
- Lower reliability
- Massive heat dissipation

From high-end hardware to affordable hardware



ANNOUNCING
DRIVE AGX XAVIER DEVKIT
SCALABLE AV COMPUTING PLATFORM
Architected for Safety: 30 TOPS to 320 TOPS
Runs DRIVE Software 1.0
Full Support for CUDA and TensorRT
OTA Ready

Available October 1, 2018



ANNOUNCING
"PEGASUS"
ROBOTAXI DRIVE PX
320 TOPS CUDA TensorCore
16x GMSL | 4x 10G | 8x 1G | 16x 100M
Auto-grade | ASIL D
500W
Late Q1 Early Access Partners

Supercomputing Data Center in Your Trunk

- 30 to 500 watts
- 500s-5000s+ of dollars
- 10s-100s of TOPS/s

NXP

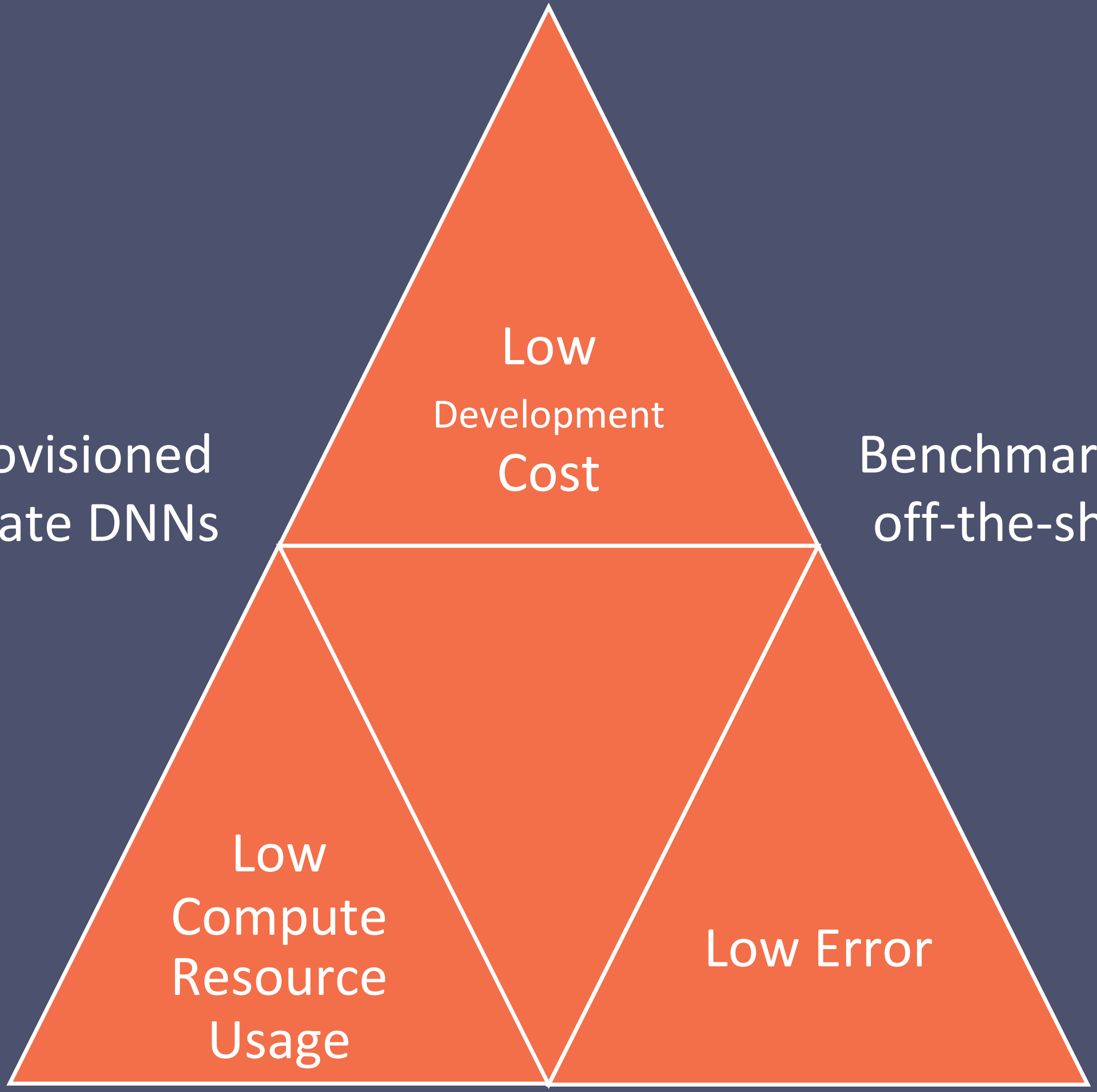
TEXAS INSTRUMENTS

RENESAS

- 1 to 30 watts (for chip + memory + I/O)
- 10s of dollars
- 1s of TOPS/s

Tradeoffs for deployable DNN models

for automotive deep learning practitioners



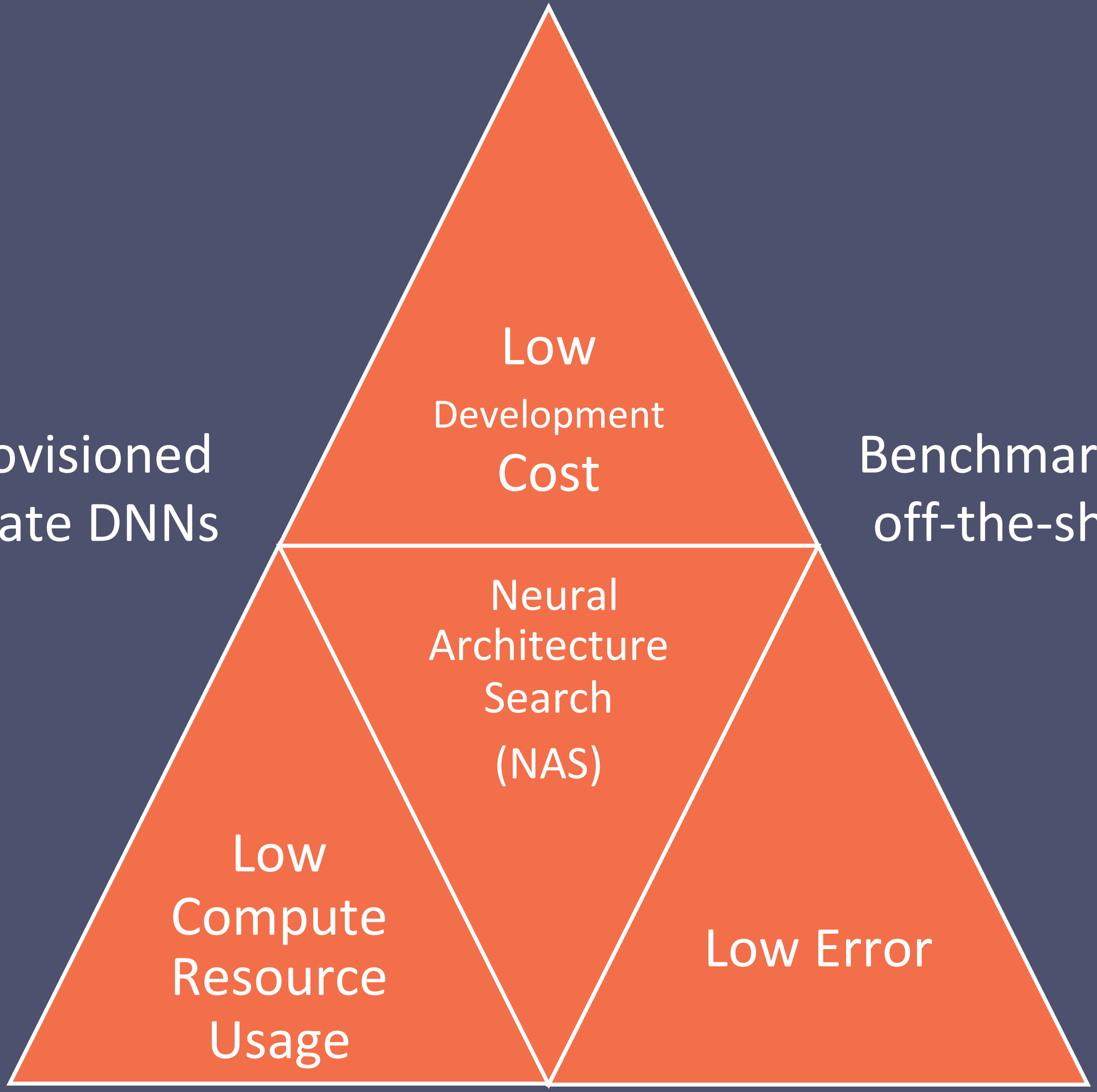
Under-provisioned less-accurate DNNs

Benchmark-winning off-the-shelf DNNs

Manually design a new DNN from scratch

Neural Architecture Search (NAS) to the Rescue

NAS can co-optimize resource-efficiency and accuracy



Low Development Cost

Neural Architecture Search (NAS)

Low Compute Resource Usage

Low Error

Under-provisioned less-accurate DNNs

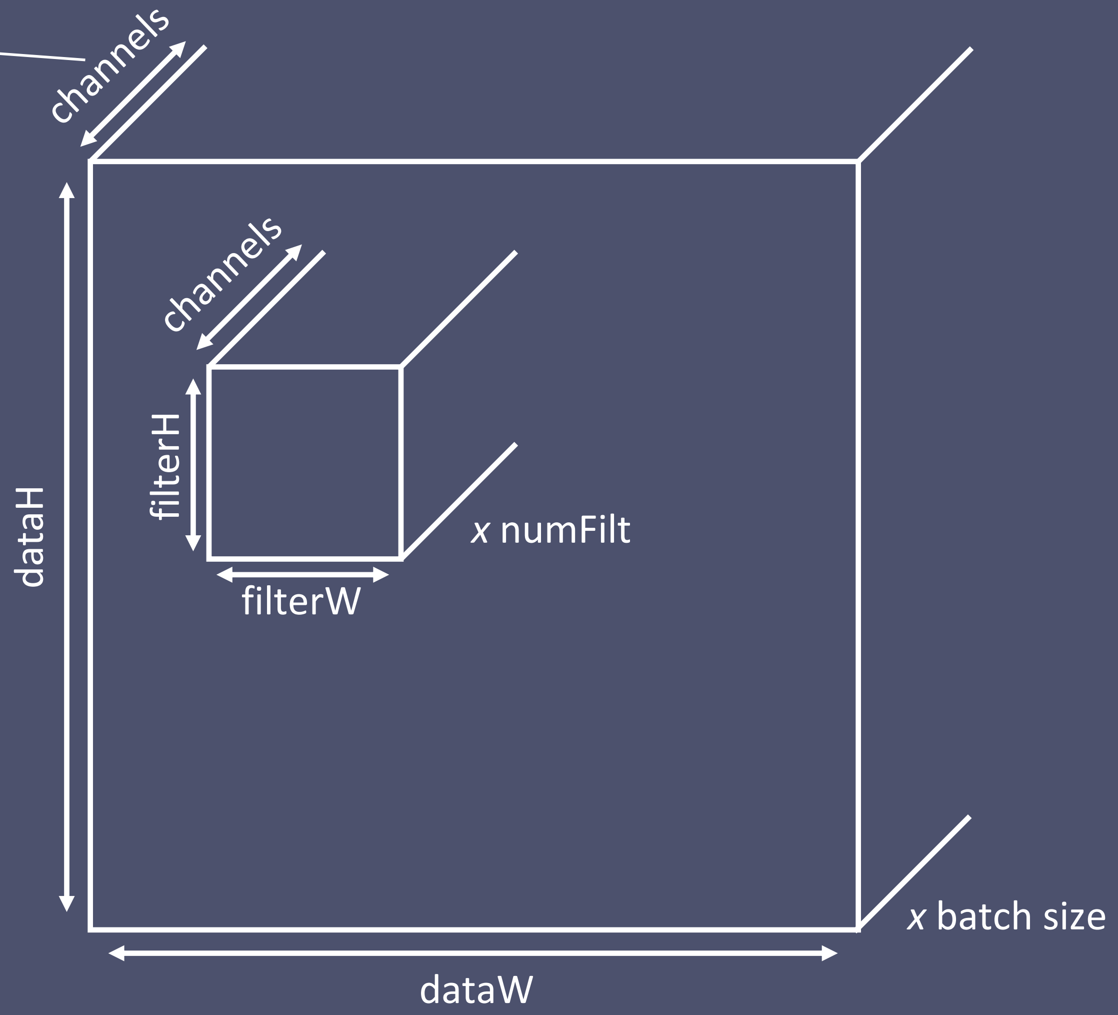
Benchmark-winning off-the-shelf DNNs

Manually design a new DNN from scratch

What's in the design space of Deep Neural Networks for computer vision?

Anatomy of a convolution layer

IMPORTANT TO KNOW: MULTIPLE CHANNELS AND MULTIPLE FILTERS



The number of channels in the current layer is determined by the number of filters (numFilt) in the previous layer.

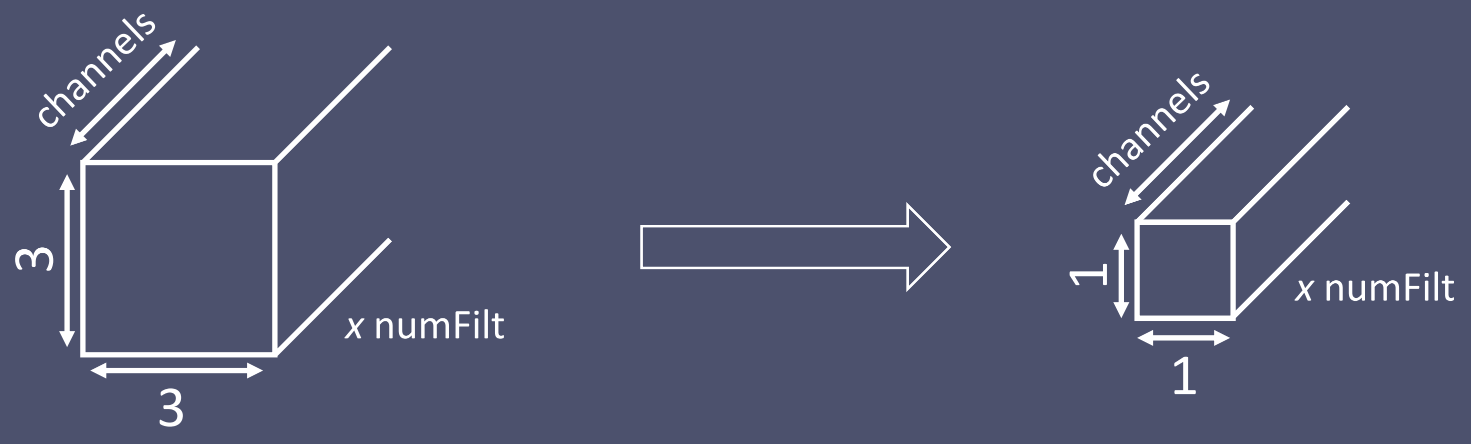
Recent history of DNN design for computer vision

DNN	Year	Accuracy* (ImageNet-1k)	Parameters (MB)	Computation (GFLOPS per frame)	Key Techniques
AlexNet	2012	57.2%	240	1.4	Applying a DNN to a hard problem; ReLU; more depth (8 layers)
VGG-19	2014	75.2%	490	19.6	More depth (22 layers)
ResNet-152	2015	77.0%	230	22.6	More depth & residual connections
SqueezeNet	2016	57.5%	4.8	0.72	Judicious use of filters and channels
MobileNet-v1	2017	70.6%	16.8	0.60	1-channel 3x3 convolutions
ShuffleNet-v1	2017	73.7%	21.6	1.05	Shuffle layers
ShiftNet	2017	70.1%	16.4	…	Shift layers
SqueezeNext	2018	67.4%	12.8	1.42	Oblong convolution filters
mNasNet-A3	2018	76.1%	20.4	0.78	Neural architecture search
FBNet-C	2018	74.9%	22.0	0.75	Really fast neural architecture search

* Top-1 single-model, single-crop accuracy

1. Kernel Reduction

REDUCING THE HEIGHT AND WIDTH OF FILTERS



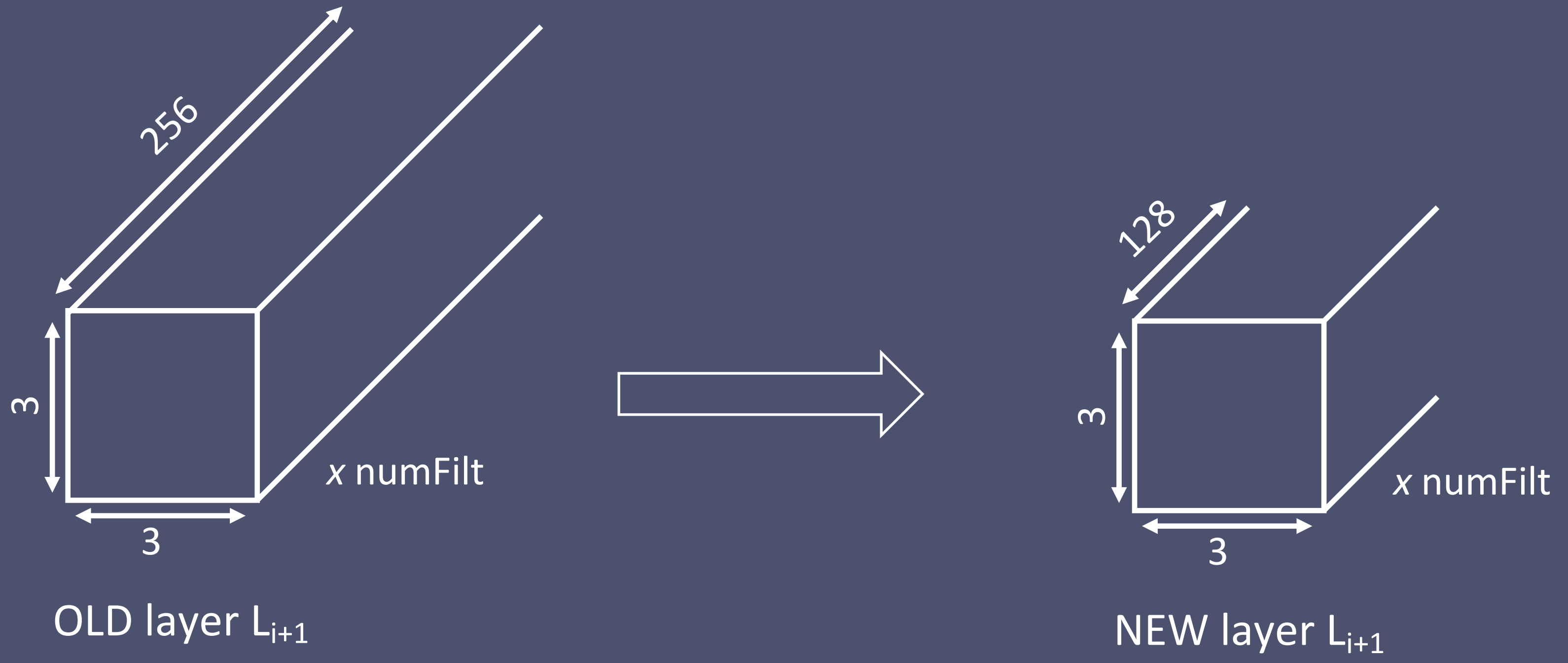
While 1x1 filters cannot see outside of a 1-pixel radius, they retain the ability to combine and reorganize information across channels.

In our design space exploration that led up to SqueezeNet, we found that we could replace half the 3x3 filters with 1x1's without diminishing accuracy

A "saturation point" is when adding more parameters doesn't improve accuracy.

2. Channel Reduction

REDUCING THE NUMBER OF FILTERS AND CHANNELS



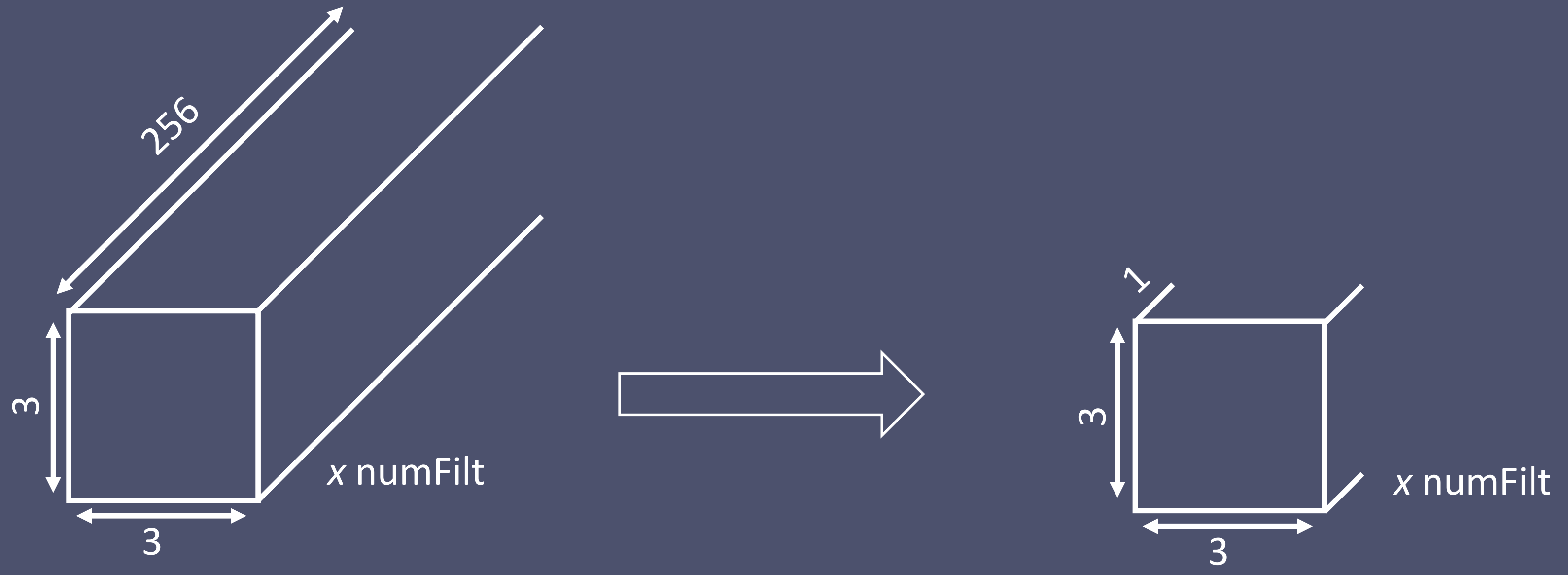
If we halve the number of filters in layer L_i

this halves the number of input channels in layer L_{i+1}

4x reduction in number of parameters

3. Depthwise Separable Convolutions

ALSO CALLED: "GROUP CONVOLUTIONS" or "CARDINALITY"



Each 3x3 filter has 1 channel
Each filter gets applied to a different channel of the input

Popularized by MobileNet and ResNeXt

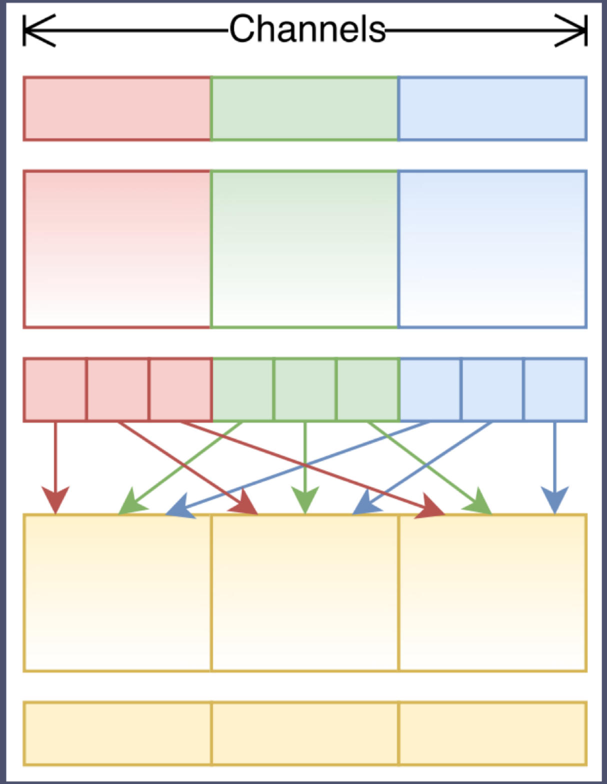
4. Shuffle Operations

After applying aggressive kernel reduction, we may have 50-90% of the parameters in 1x1 convolutions

Group-1x1 convs would lead to multiple DNNs that don't communicate

Solution: *shuffle* layer after separable 1x1 convs

"shuffle" layer



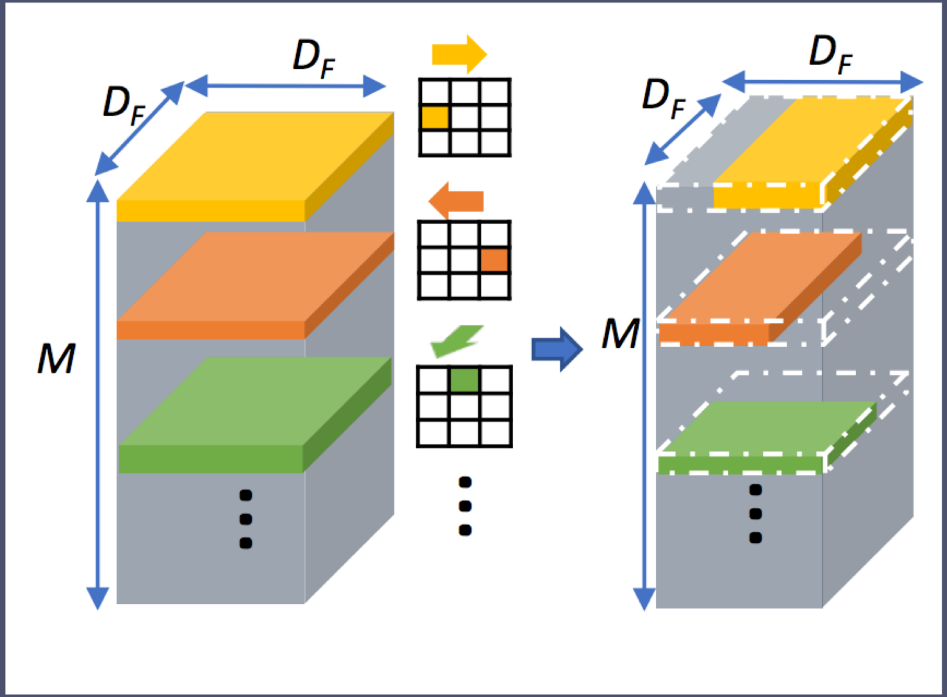
Zhang, et al. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. arXiv, 2017.

5. Shift Operations

Shift each channel's activation grid by one cell

This allows all your filters to be 1x1xChannels (and not 3x3)

"shift" layer



[1] B. Wu, et al. Shift: A Zero FLOP, Zero Parameter Alternative to Spatial Convolutions. CVPR, 2018.

Device-specific DNN design considerations

Deep Learning Processors have arrived!

THE SERVER SIDE

Platform	Computation (GFLOPS/s)	Memory Bandwidth (GB/s)	Computation-to-bandwidth ratio	Power (TDP Watts)	Year
NVIDIA K20 [1]	3500 (32-bit float)	208 (GDDR5)	17	225	2012
NVIDIA V100 [2]	112000 (16-bit float)	900 (HBM2)	124 (yikes!)	250	2018

Uh-oh… Processors are improving much faster than Memory.

[1] https://www.nvidia.com/content/PDF/kepler/Tesla-K20-Passive-BD-06455-001-v05.pdf
[2] http://www.nvidia.com/content/PDF/Volta-Datasheet.pdf (PCIe version)

Deep Learning Processors have arrived!

MOBILE PLATFORMS

Device	Cores	Computation (GFLOPS/s)	Memory Bandwidth (GB/s)	Computation-to-bandwidth ratio	System Power (TDP Watts)	Year
Samsung Galaxy Note 3	Arm Mali T-628 GPU [1]	120 (32-bit float)	12.8 (LPDDR3)	9.3	~10	2013
Huawei P20	Kirin 970 NPU [2]	1920 (16-bit float)	30 (LPDDR4X)	64 (ouch!)	~10	2018
NVIDIA Jetson Xavier [3,4]	NVIDIA Tensor Cores	30000 (8→32 int)	137	218 (yikes!)	10 to 30 (multiple modes)	2018

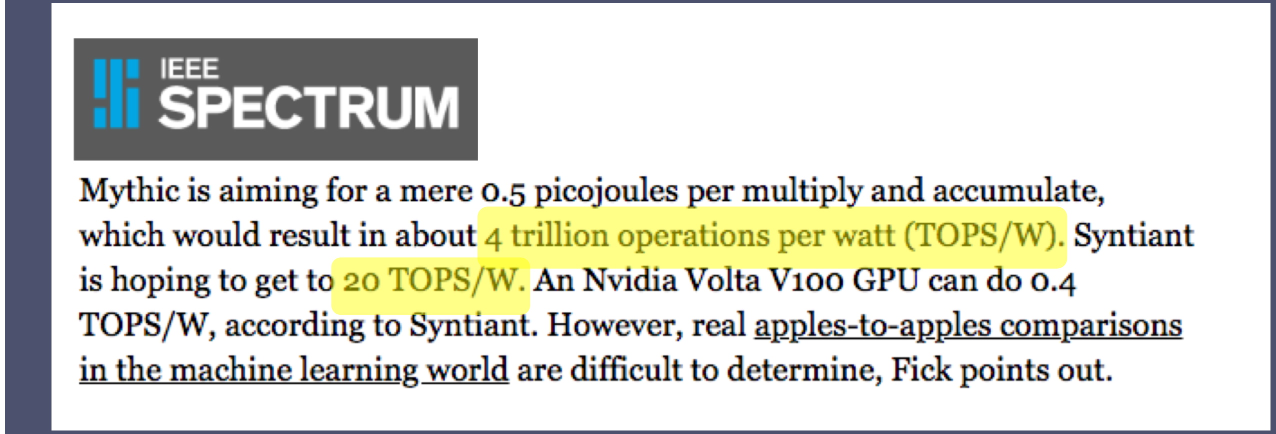
[1] https://indico.cern.ch/event/319744/contributions/1698147/attachments/616065/847693/gdb_110215_cesini.pdf

[2] https://www.androidauthority.com/huawei-announces-kirin-970-797788

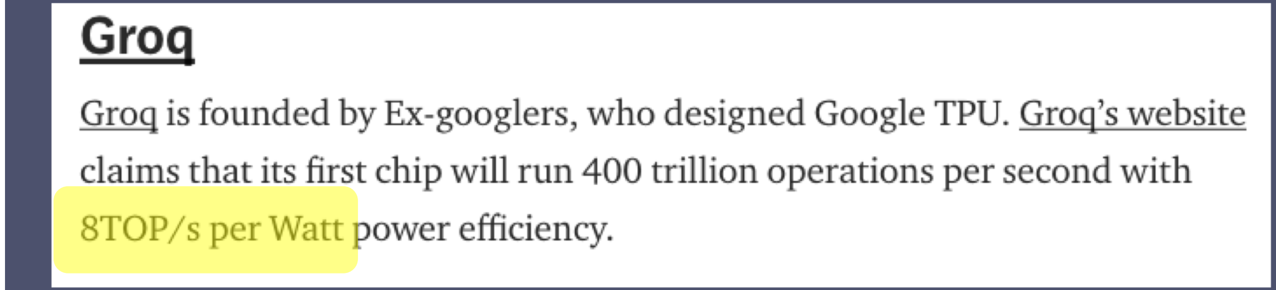
[3] https://blogs.nvidia.com/blog/2018/01/07/drive-xavier-processor/

[4] https://developer.nvidia.com/jetson-xavier

What will the next generation Deep Learning servers look like?



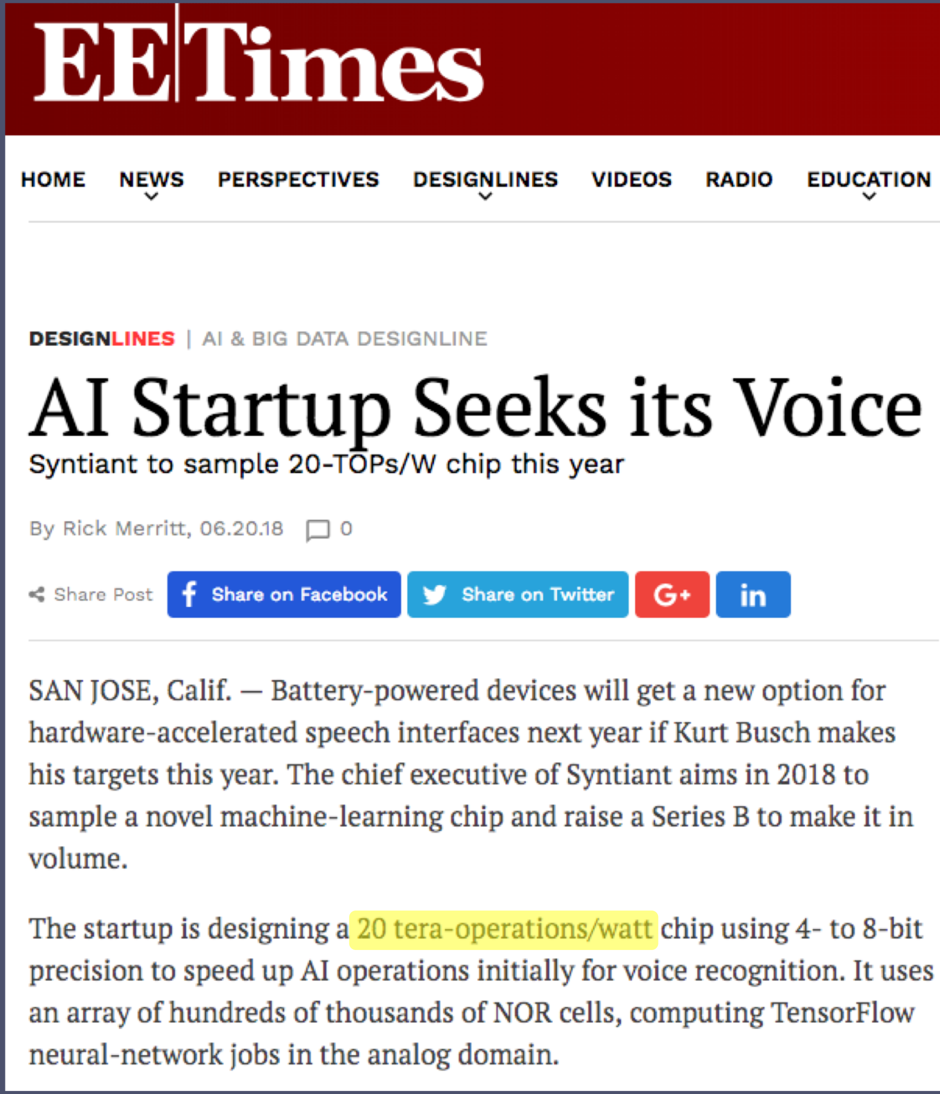
Mythic is aiming for a mere 0.5 picojoules per multiply and accumulate, which would result in about 4 trillion operations per watt (TOPS/W). Syntiant is hoping to get to 20 TOPS/W. An Nvidia Volta V100 GPU can do 0.4 TOPS/W, according to Syntiant. However, real apples-to-apples comparisons in the machine learning world are difficult to determine, Fick points out.



Groq

Groq is founded by Ex-googlers, who designed Google TPU. Groq's website claims that its first chip will run 400 trillion operations per second with 8TOP/s per Watt power efficiency.

https://medium.com/@shan.tang.g/a-list-of-chip-ip-for-deep-learning-48d05f1759ae



DESIGNLINES | AI & BIG DATA DESIGNLINE

AI Startup Seeks its Voice

Syntiant to sample 20-TOPs/W chip this year

By Rick Merritt, 06.20.18 0

SAN JOSE, Calif. — Battery-powered devices will get a new option for hardware-accelerated speech interfaces next year if Kurt Busch makes his targets this year. The chief executive of Syntiant aims in 2018 to sample a novel machine-learning chip and raise a Series B to make it in volume.

The startup is designing a 20 tera-operations/watt chip using 4- to 8-bit precision to speed up AI operations initially for voice recognition. It uses an array of hundreds of thousands of NOR cells, computing TensorFlow neural-network jobs in the analog domain.

What will the next generation Deep Learning servers look like?

20 TOP/W COMPUTATION

Platform	Efficiency (TOP/s/W)	Computation (TOP/s)	Memory Bandwidth (TB/s)	Computation-to-bandwidth ratio	Power (TDP Watts)	Year
NVIDIA K20 [1]	0.015	3.50 (32-bit float)	0.208 (GDDR5)	17	225	2012
NVIDIA V100 [2]	0.45	112 (16-bit float)	0.900 (HBM2)	124	250	2018
Next-gen: 20 TOP/W	20	2500*	1.800 (HBM3) [3]	1389 (oh no!)	250	2020 (est.)

[1] https://www.nvidia.com/content/PDF/kepler/Tesla-K20-Passive-BD-06455-001-v05.pdf
[2] http://www.nvidia.com/content/PDF/Volta-Datasheet.pdf (PCIe version)
[3] https://www.eteknix.com/gddr6-hbm3-details-emerge/

* Assuming half the power is spent on computation, and the other half is spent on memory and other devices.
20 TOP/s/W * 20W * 0.5 = 2500 TOP/s

DEEPSCALE

Summary: Device-specific DNN design considerations

- Processors have recently increased 10-100x in dense-matrix computation-per-watt.

- But, DRAM memory bandwidth is increasing slowly (2x more bandwidth-per-watt every 4 years).
    - So, we need DNNs with cache-locality that don't need frequent DRAM accesses

Related work on Neural Architecture Search

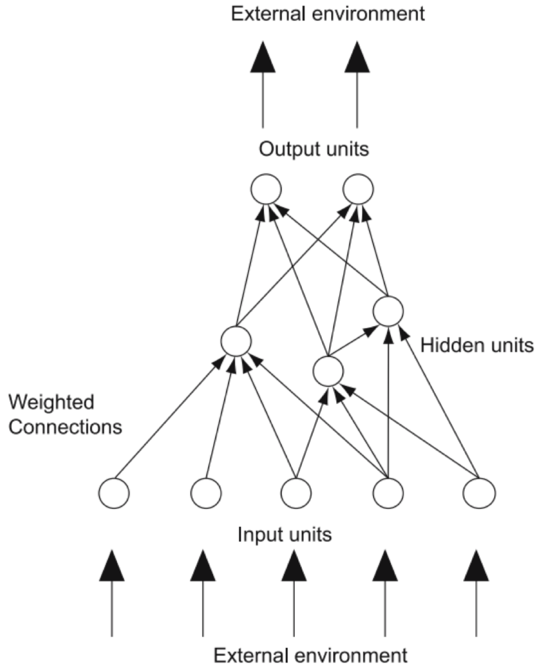
Hyperparameter Optimization Methods

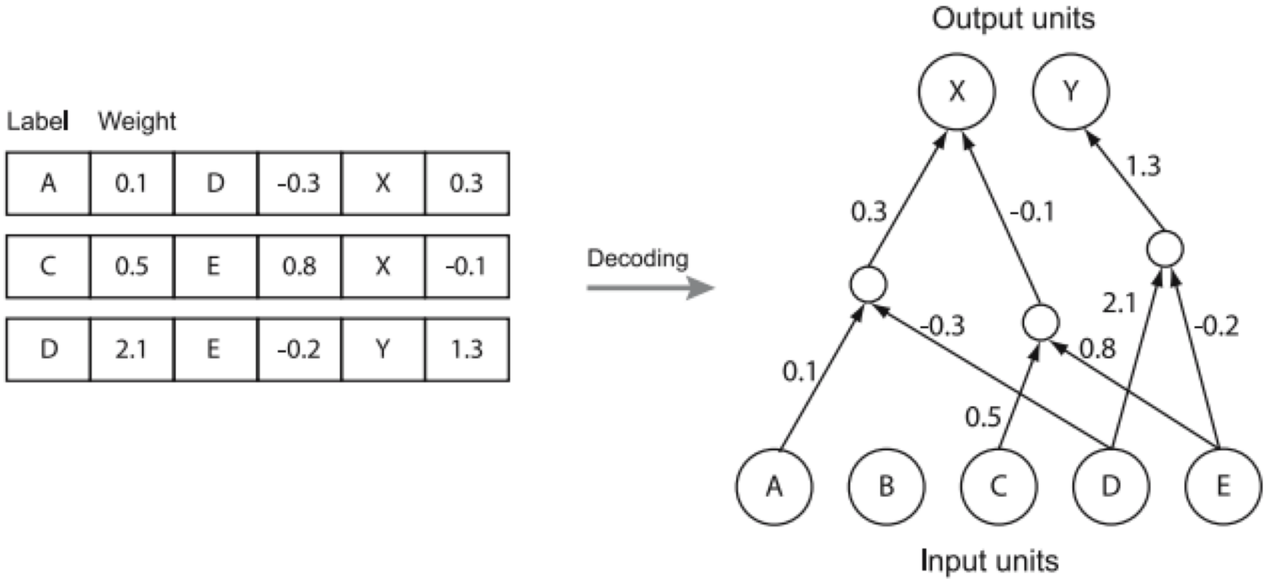
- Grid Search
  - Exhaustively search user-defined space

- Random search
  - Try random combinations

- Bayesian optimization
  - Try to infer a probabilistic model

Neuroevolution: from architectures to learning[1]





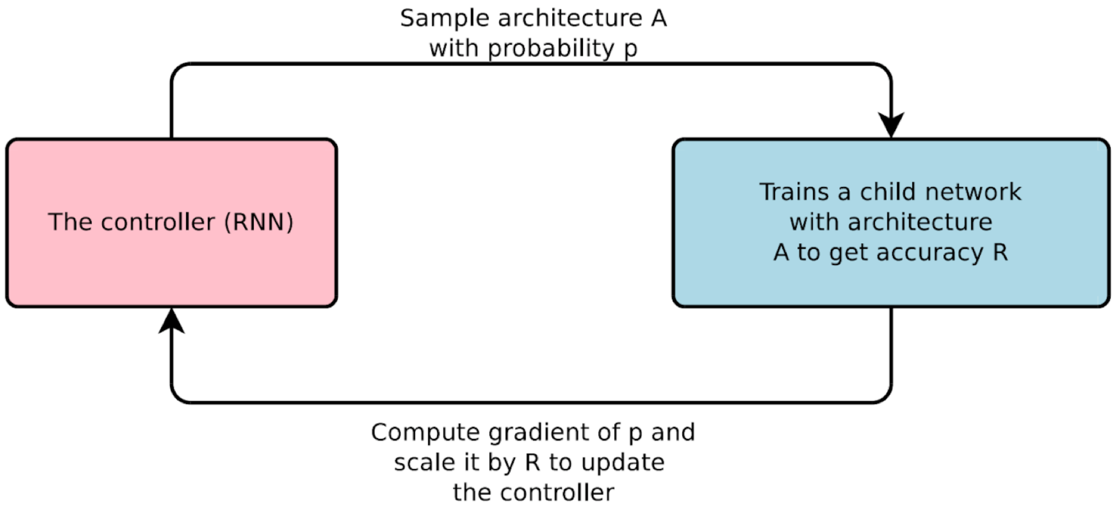
Label	Weight				
A	0.1	D	-0.3	X	0.3
C	0.5	E	0.8	X	-0.1
D	2.1	E	-0.2	Y	1.3

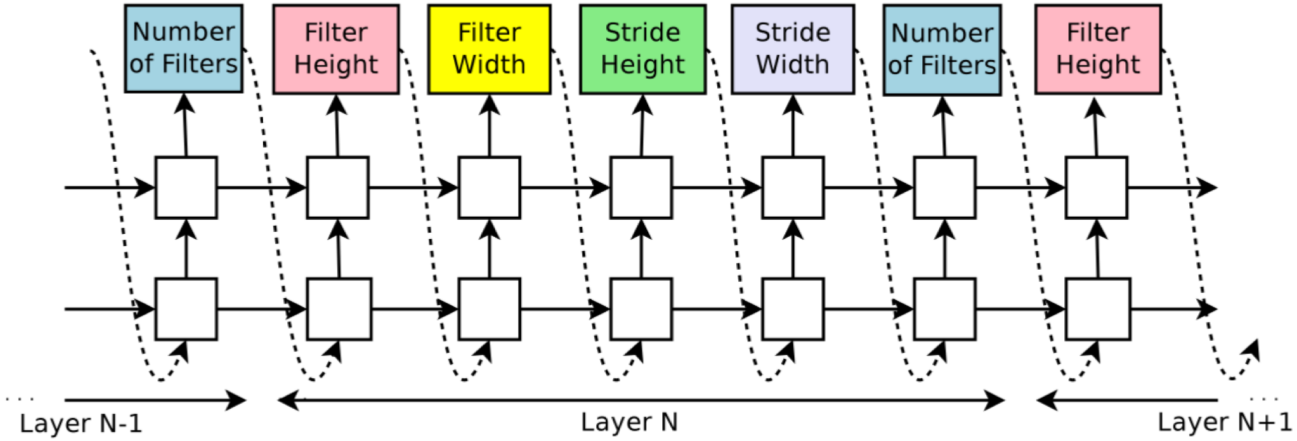
Paper from 2008 gives an overview of work on evolutionary methods for NN architecture design and initialization.

“In order to design a neural network for a particular task, the choice of an architecture (including the choice of a neuron model), and the choice of a learning algorithm have to be addressed”

“This paper gives an overview of the most prominent methods for evolving NNs with a special focus on recent advances in the synthesis of learning architectures.”

[1] Floreano, D., Dürr, P., & Mattiussi, C. (2008). Neuroevolution: from architectures to learning. Evolutionary Intelligence, 1(1), 47-62.





NAS with Reinforcement Learning

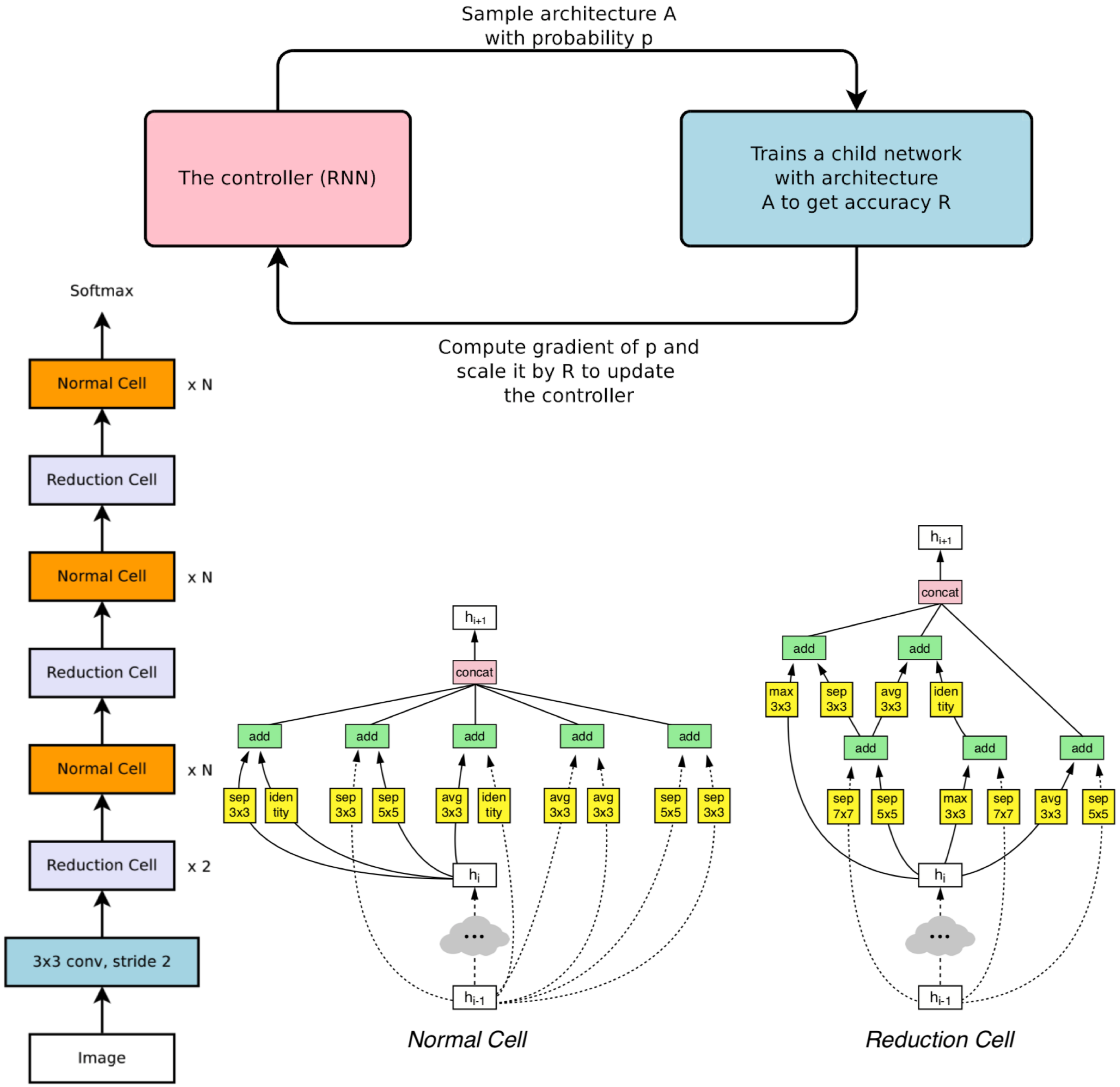
Block-level search [1]

- Use a Recurrent Neural Network in a RL loop to generate entire child network for the CIFAR dataset updating after each model has trained
- Achieved 0.09% better accuracy at the time and 1.05x faster on CIFAR-10
- 800 Nvidia K40 GPUs for 28 days = 22,400 GPU Days
- Search performed on small dataset
- Better than brute force approach but still too much compute too be practical

[1] B. Zoph, Q. Le. Neural Architecture Search with Reinforcement Learning. ICLR, 2018.

Learning Transferable Architectures

Cell-level search [2]



- Use a Recurrent Neural Network in a RL loop to generate cells using CIFAR-10 as proxy task then adapted to ImageNet
- Achieved 1.20% better accuracy while being 28% faster on ImageNet1000
- 500 Nvidia P100 GPUs for 4 days = 2,000 GPU Days
- cells are all the same (unlike [1])
- More efficient than previous method but still expensive

[1] B. Zoph, Q. Le. Neural Architecture Search with Reinforcement Learning. ICLR, 2018.

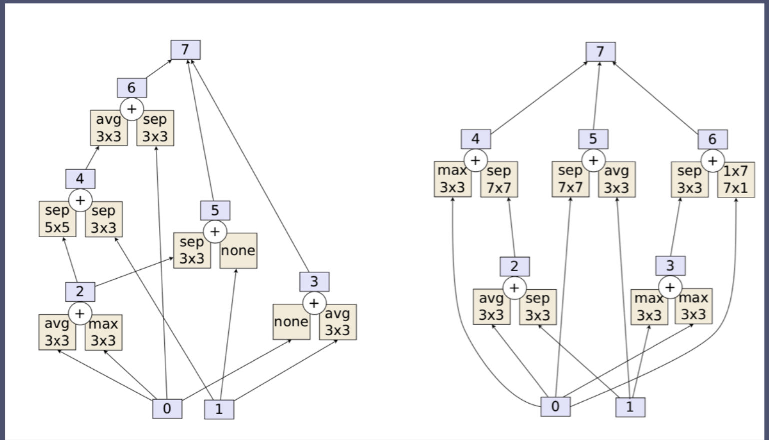
[2] B. Zoph et al. Learning Transferable Architectures for Scalable Image Recognition. CVPR, 2018.

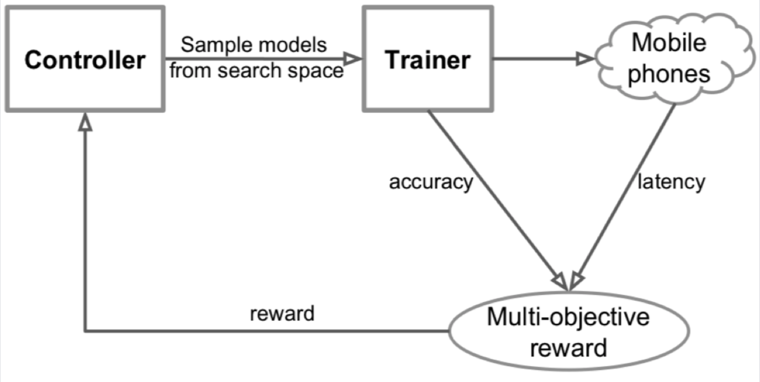
Other Related Work

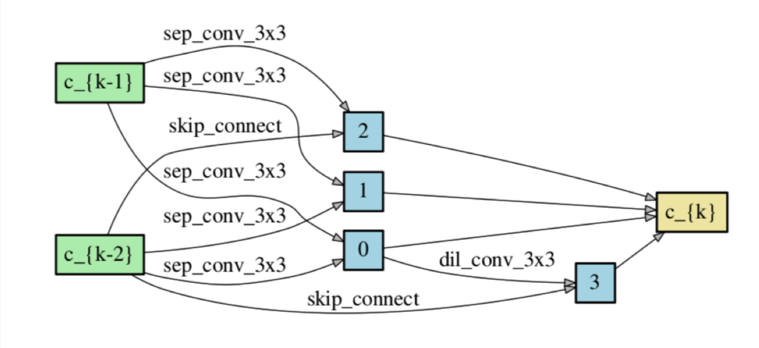
- Evolutionary Method
  - AmoebaNet[1]
    - Tournament Selection Evolutionary on Cell Space
    - 3,150 K40 GPU Days

- Latency Aware Reinforcement Learning
  - MnasNet[2]
    - Latency Aware Block level Search on proxy ImageNet
    - 288 TPUv2 Days ≈ 2,000 P100 GPU Days

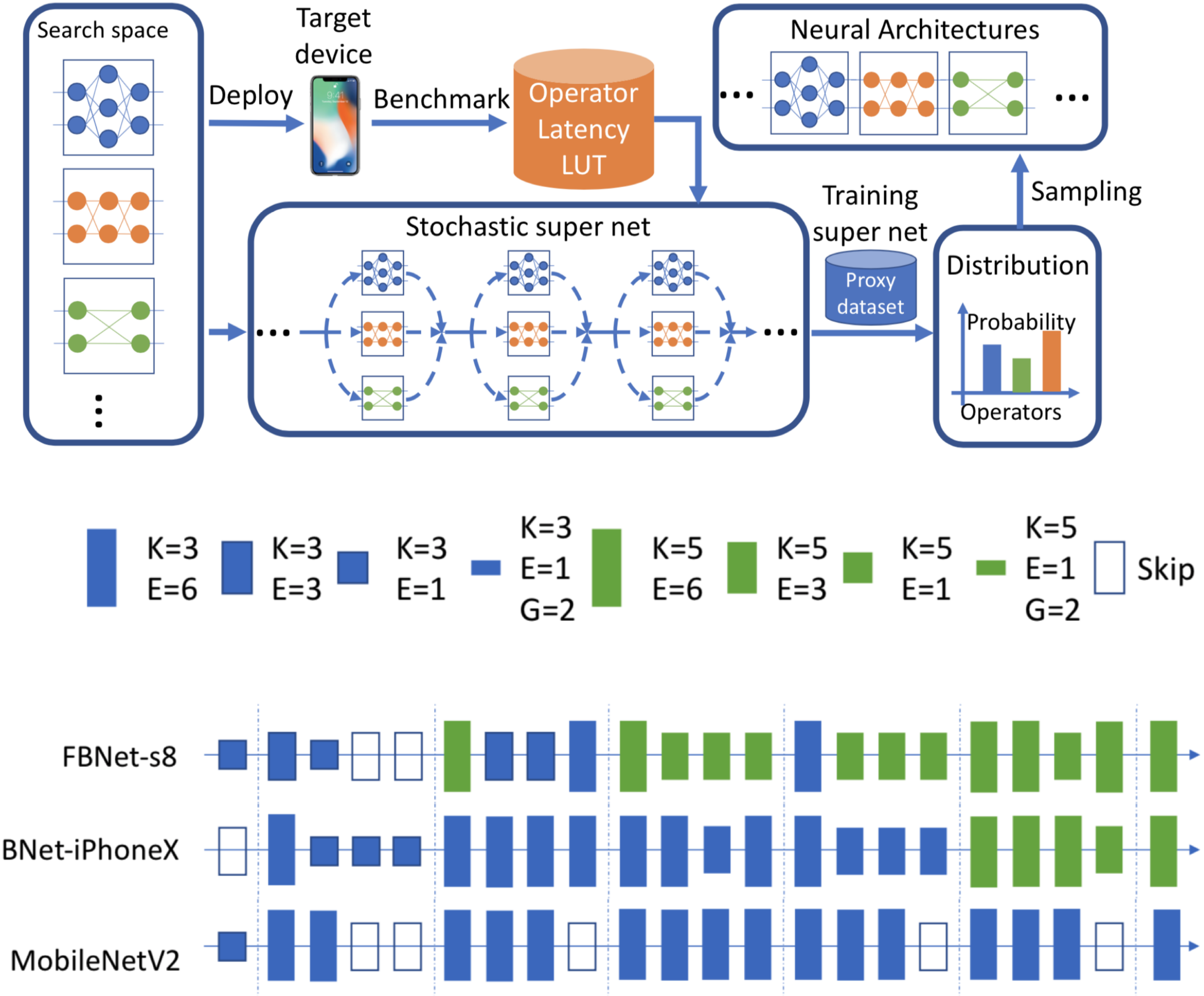
- Supernetwork - Differential Search
  - DARTS: Differential ARchitecTure Search[3]
    - Gradient Based Cell Search performed on CIFAR-10
    - 4 1080 TI GPU Days







[1] E. Real et al. Regularized Evolution for Image Classifier Architecture Search. AAAI, 2019.
[2] M. Tan et al. MnasNet: Platform-Aware Neural Architecture Search for Mobile. CVPR, 2019.
[3] H. Liu et al. DARTS: Differentiable Architecture Search. ICLR, 2019.



Stochastic Supernet Optimization

FBNet [3]

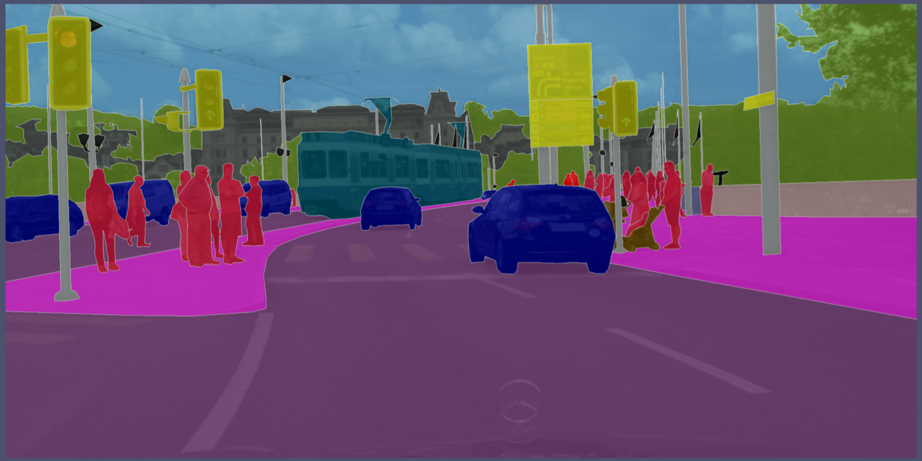
- Creates Stochastic Supernet which contains entire architecture Search space. Only has to train this one meta-network instead of many child networks.
- Uses Gumbel-Softmax to sample from categorical distribution for layer choices weighted by learnable parameters
- Uses a Latency Look Up Table(LUT) to estimate and optimize network latency
- FBNet-B achieved MobileNetV2-1.3 Accuracy while being 1.5x lower latency
- 9 P100 GPU Days Search Cost
- Search Space inspired by MobilenetV2

[3] Wu, B., Dai, X., Zhang, P., Wang, Y., Sun, F., Wu, Y., ... & Keutzer, K. (2019). FBNet: Hardware-aware efficient convnet design via differentiable neural architecture search. CVPR, 2019.

Applying NAS to design DNNs for semantic segmentation

Classification vs Semantic Segmentation tasks

Examples of image classification (ImageNet[1])



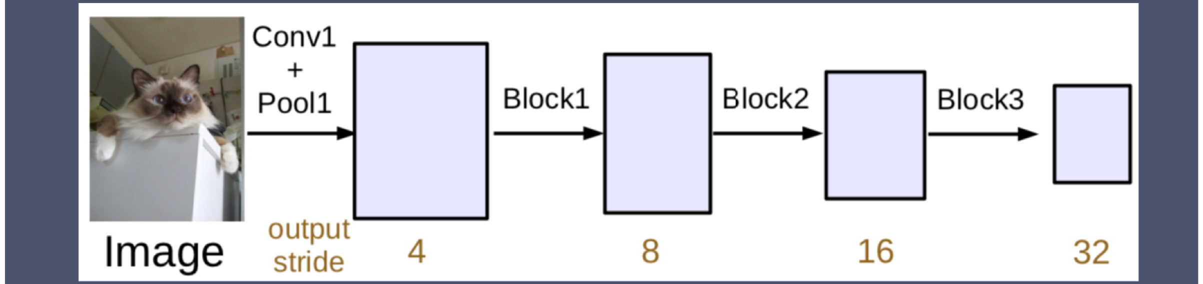
Example of Semantic Segmentation (Cityscapes[2])

• Image level prediction • Location Invariant • Low Resolution (224x224 input) • SOTA Networks compute: ~10 GFLOPs	• Pixel level prediction • Location Variant • High Resolution (1024x2048 input) • SOTA Networks range: ~1 **T**FLOPS

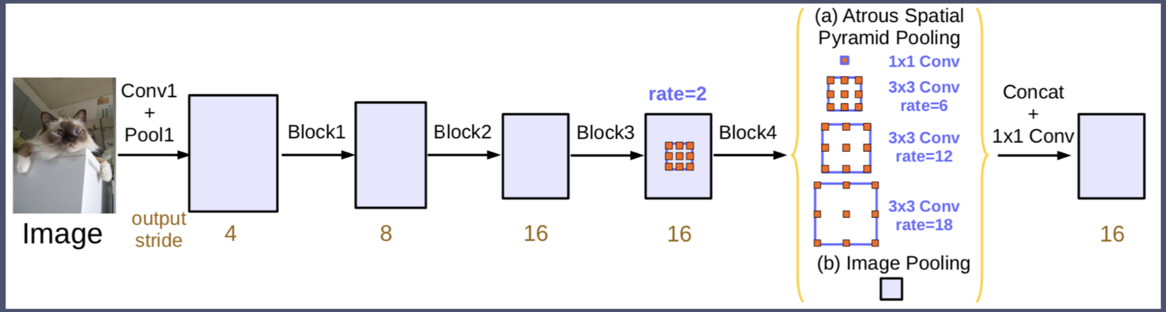
[1] O. Russakovsky et al. ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015.

[2] M. Cordts et al. The Cityscapes Dataset for Semantic Urban Scene Understanding. CVPR, 2016.

Classification vs Semantic Segmentation DNNs



Examples DNN for image classification



Example DNN for Semantic Segmentation (DeepLabV3[1])

• Networks designed for task and are trained from scratch	• SS Networks are adapted from classification networks and then retrained.

[1] LC. Chen et al. Rethinking Atrous Convolution for Semantic Image Segmentation, 2017.

Applying NAS to design DNNs for semantic segmentation

- We need a network that runs in realtime on our automotive grade platform that gets as high of a performance as we can on our target task.

- Goal: advance the frontier of accuracy/efficiency on Semantic Segmentation

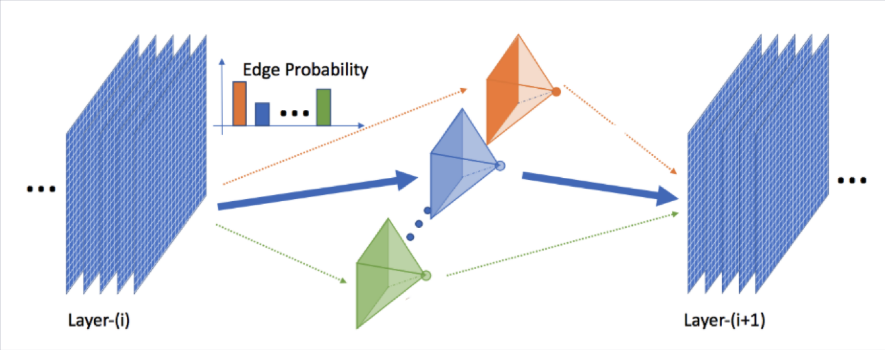
SqueezeNAS: An Adaptation of FBNet for Semantic Segmentation Search

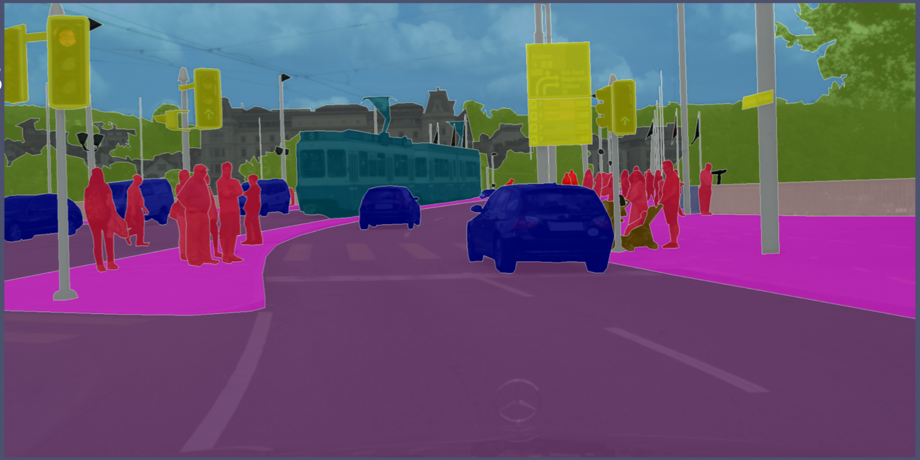
- Stochastic Super Network
  - Run all units in parallel
  - Perform weighted sum of activations where weights are sampled from Gumbel-Softmax
  - 2 types of learned parameters: Convolution parameters and architecture parameters

- Resource aware learned architecture parameter
  - A unit in the meta-network is chosen by its architecture parameter plus a random variable
  - Optimize model-parameters and architecture-parameters simultaneously

- Proxyless training
  - We train directly on Cityscapes training set
  - Training until both model-parameters and architecture-parameters converge

Figure courtesy of Bichen Wu, et al.

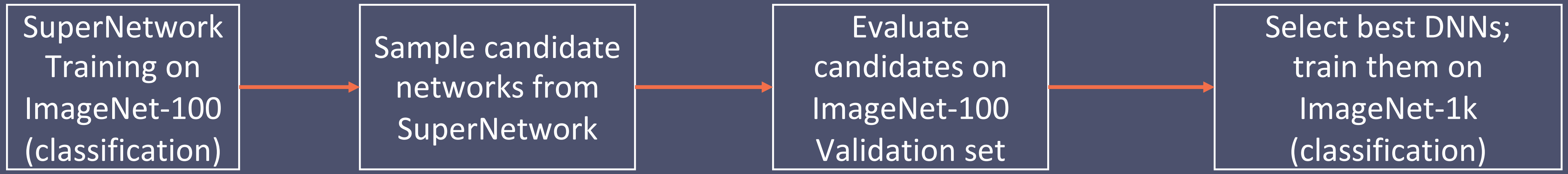




DEEPSCALE

Training scheme

FBNet training flow



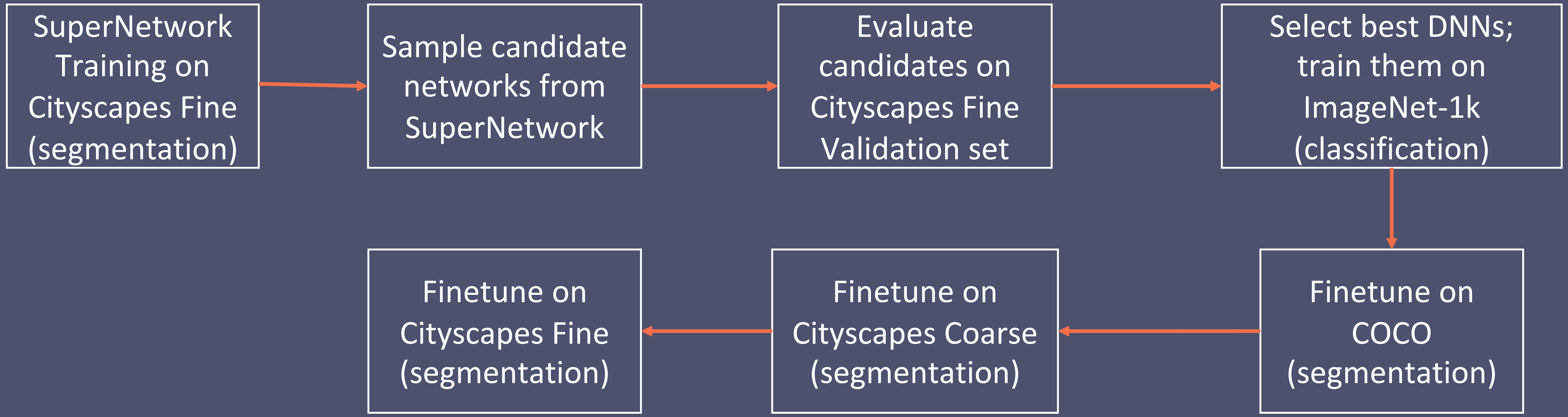
SuperNetwork Training on ImageNet-100 (classification)

Sample candidate networks from SuperNetwork

Evaluate candidates on ImageNet-100 Validation set

Select best DNNs; train them on ImageNet-1k (classification)

SqueezeNAS training flow



SuperNetwork Training on Cityscapes Fine (segmentation)

Sample candidate networks from SuperNetwork

Evaluate candidates on Cityscapes Fine Validation set

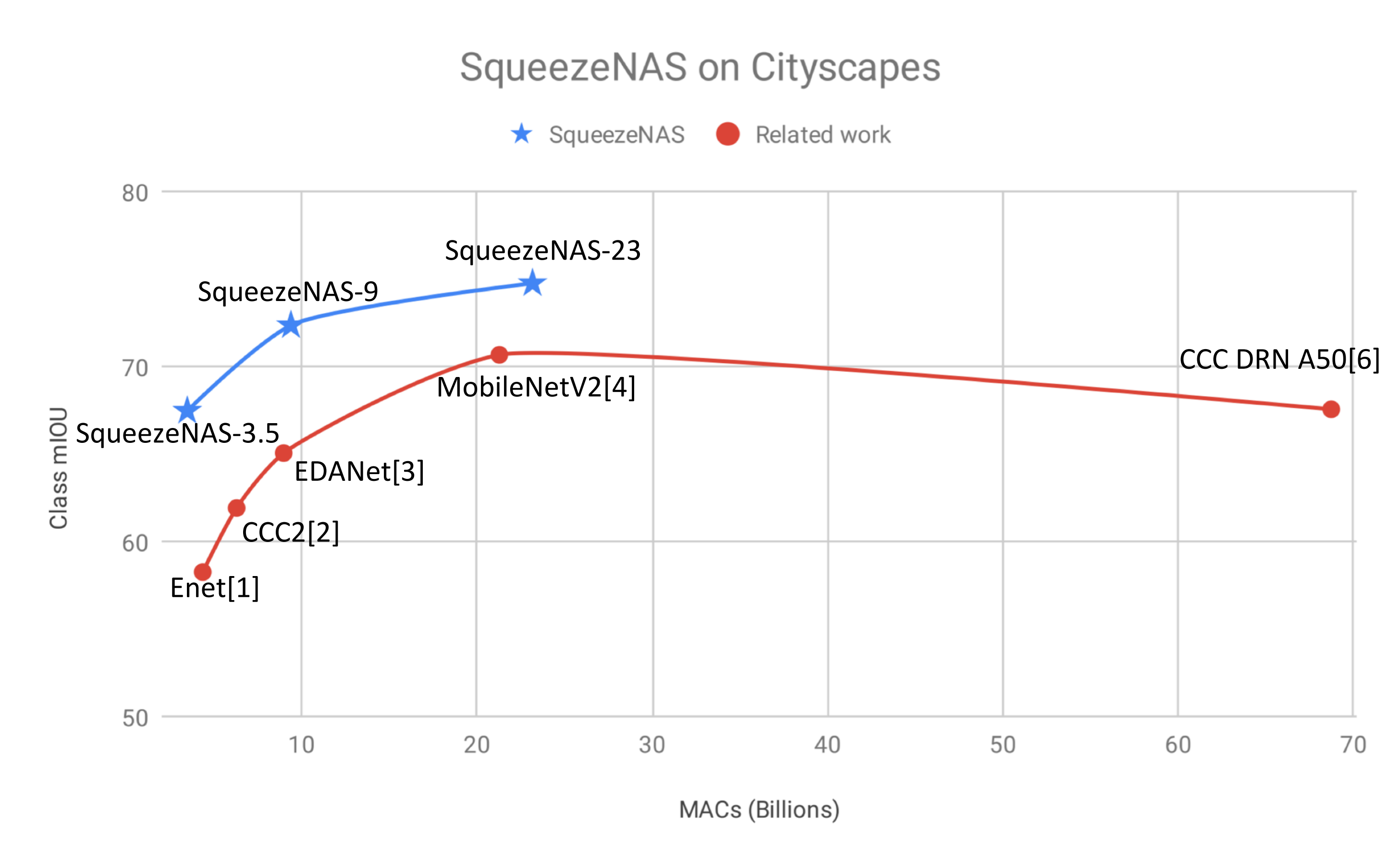
Select best DNNs; train them on ImageNet-1k (classification)

Finetune on COCO (segmentation)

Finetune on Cityscapes Coarse (segmentation)

Finetune on Cityscapes Fine (segmentation)

SqueezeNAS: Cityscapes Results



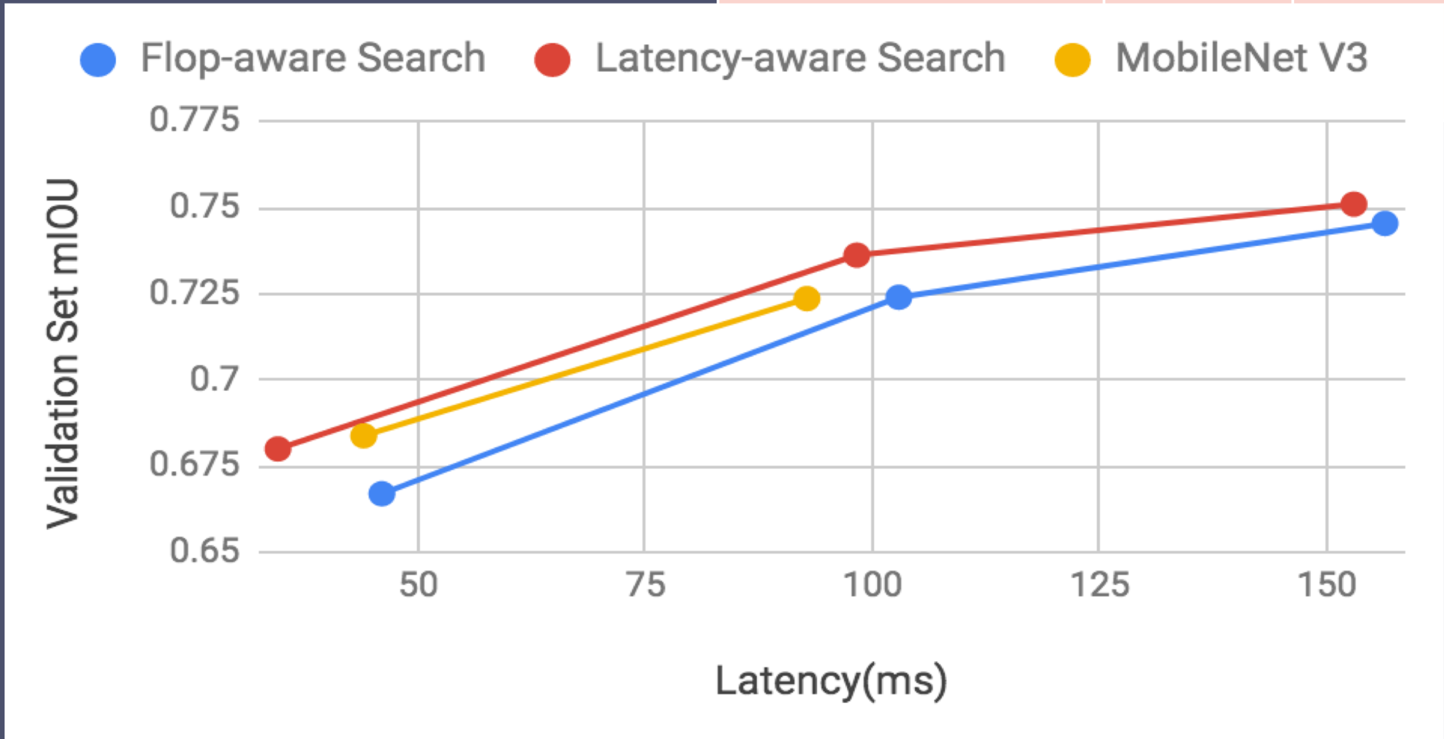
Name	MACs (Billions)	Class mIOU on Cityscapes
SqueezeNAS-3	3.0	66.7
SqueezeNAS-9	9.4	72.4
SqueezeNAS-22	21.8	74.5
Enet[1]	4.4	58.3
CCC2[2]	6.3	62.0
EDANet[3]	9.0	65.1
MobileNetV2 OS=16[4]	21.3 [5]	70.7 [5]
CCC DRN A50[6]	68.7	67.6

[1] Paszke, Adam et al. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation, 2016
[2] Park, Hyojin et al. Concentrated-Comprehensive Convolutions for lightweight semantic segmentation, 2018
[3] Lo, Shao-Yuan et al. Efficient Dense Modules of Asymmetric Convolution for Real-Time Semantic Segmentation, 2018
[4] Sandler, Mark et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks, CVPR 2018.
[5] https://github.com/tensorflow/models/blob/master/research/deeplab/g3doc/model_zoo.md
[6] Yu, Fisher et al. Dilated Residual Networks, CVPR 2017.

SqueezeNAS: Cityscapes Results

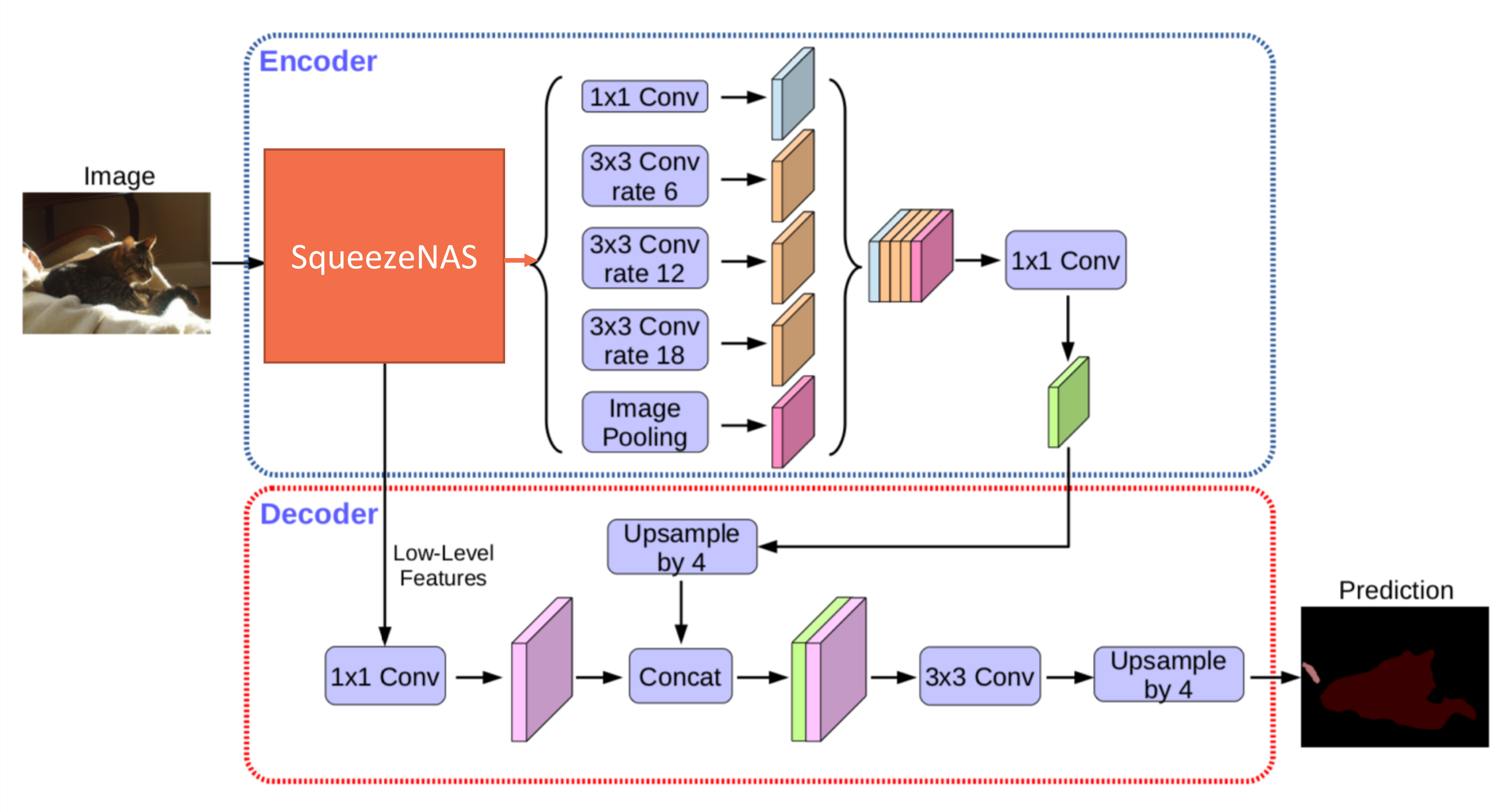
Name	Search Goal	MACs (Billions)	Latency (ms) on NVIDIA Xavier	Class mIOU on Cityscapes
SqueezeNAS-3	MACs	3.0	46.0	66.7
SqueezeNAS-9	MACs	9.4	103	72.4
SqueezeNAS-22	MACs	21.8	156	74.5

Name	Search Goal	MACs (Billions)	Latency (ms) on NVIDIA Xavier	Class mIOU on Cityscapes
SqueezeNAS-4.5 v2	Latency	4.5	34.6	68.0
SqueezeNAS-20 v2	Latency	19.6	98.3	73.6
			153	75.1



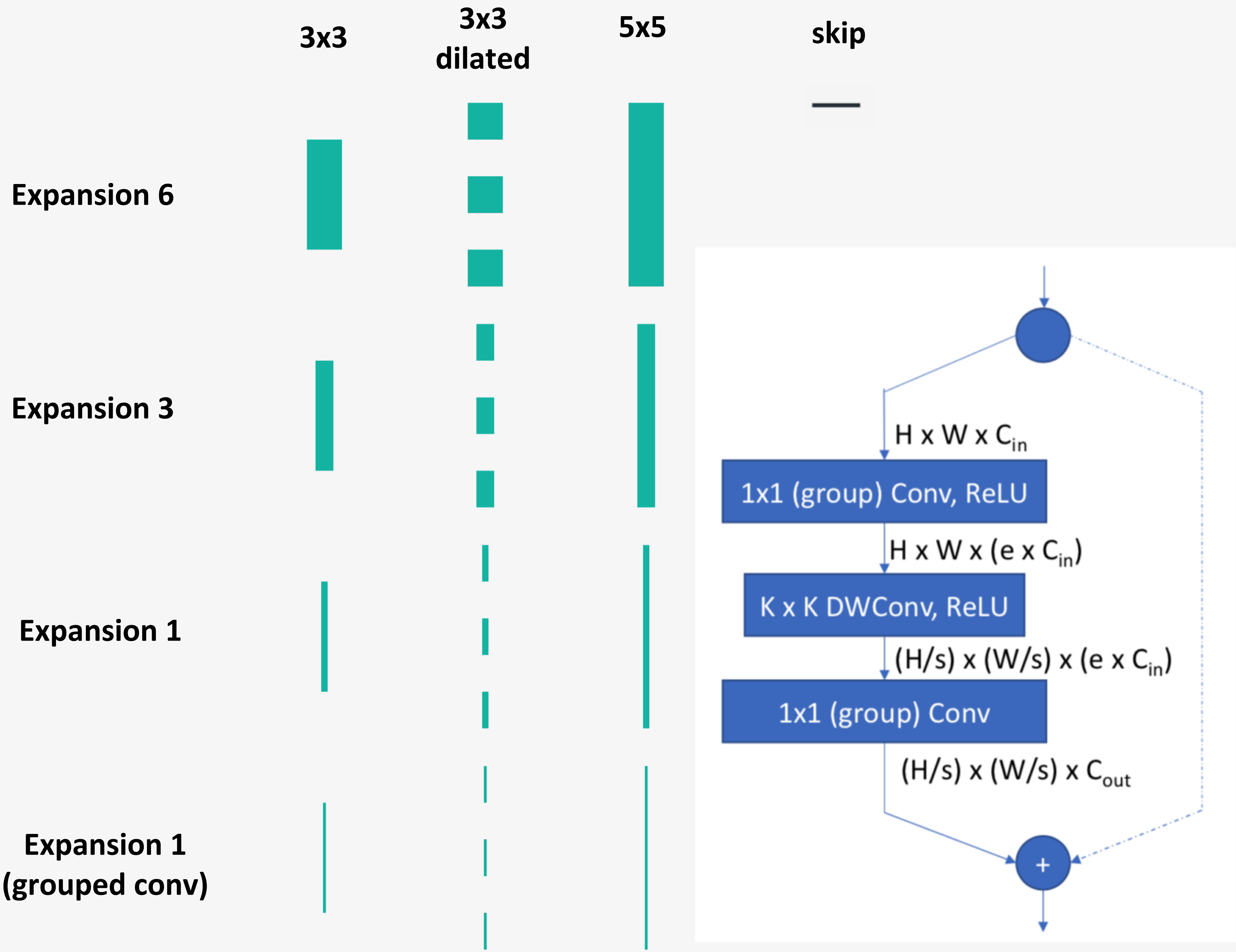
Search Space

We employ the encoder-decoder depthwise head from DeepLab V3+[1] while allowing the base network to be completely learned



[1] Chen et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, ECCV 2018

Search Space



3x3

3x3 dilated

5x5

skip

Expansion 6

Expansion 3

Expansion 1

Expansion 1 (grouped conv)

H x W x C_{in}

1x1 (group) Conv, ReLU

H x W x (e x C_{in})

K x K DWConv, ReLU

(H/s) x (W/s) x (e x C_{in})

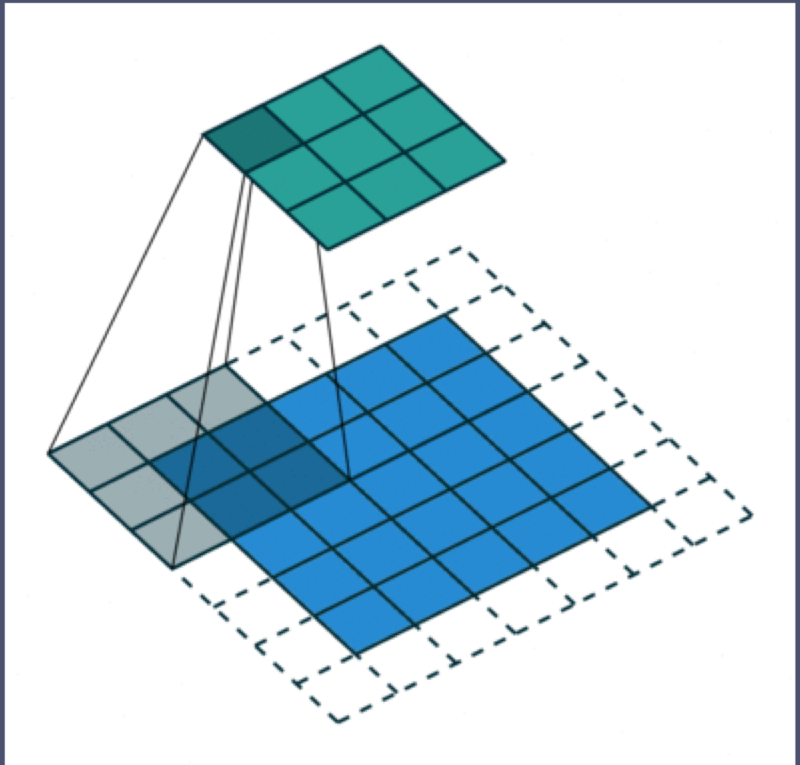
1x1 (group) Conv

(H/s) x (W/s) x C_{out}

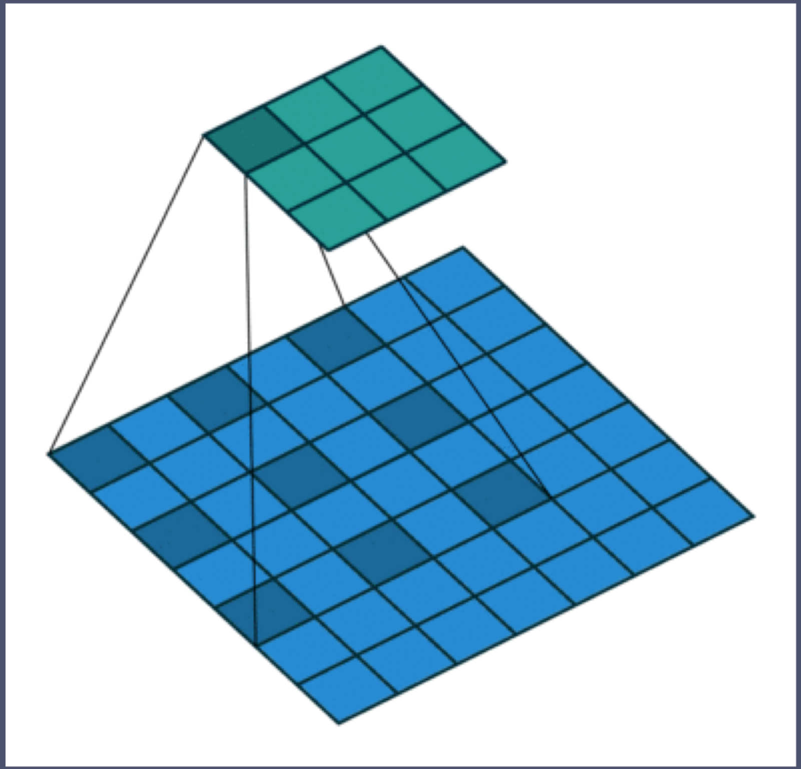
+

Dilated Convolutions

(also known as Atrous Convolution)



Normal 3x3 Convolution

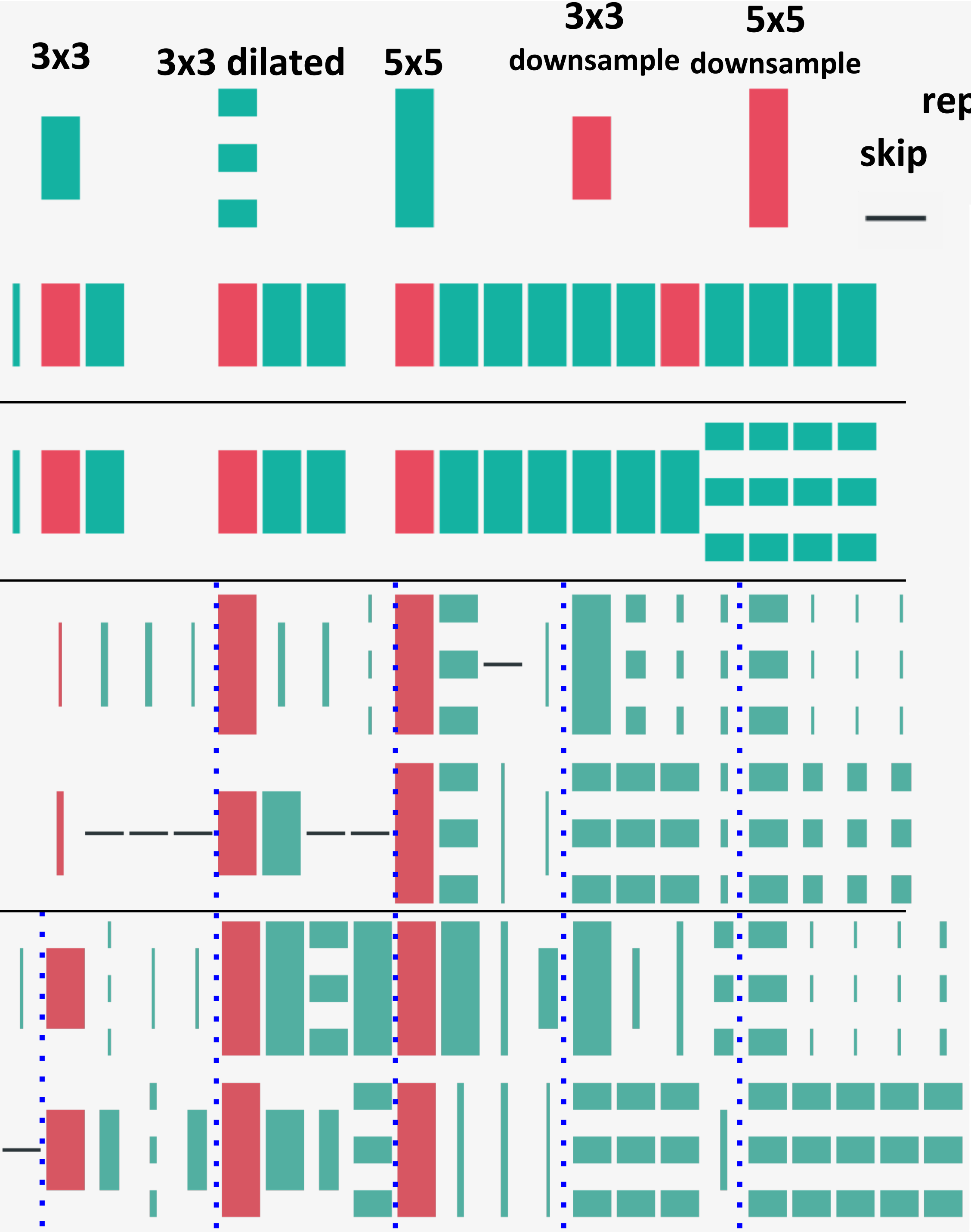


Dilated 3x3 Convolution

Graphic taken from Sik-Ho Tsang’s article https://towardsdatascience.com/review-dilated-convolution-semantic-segmentation-9d5a5bd768f5

SqueezeNAS

Resulting Networks



Legend (Unit Type) 1x1->unit->1x1

3x3

3x3 dilated

5x5

3x3 downsample

5x5 downsample

skip

Box Width represents channel expansion

	MACs (Giga)	mIOU %
MobileNetV2 Classification		
MobileNetV2 DeepLabV3	21.3	70.71
SqueezeNAS-3 (MAC Optimized)	3.0	66.7
SqueezeNAS-4.5 v2 (Latency Searched)	4.5	68.0
SqueezeNAS-22 (MAC Optimized)	21.8	74.5
SqueezeNAS-33 v2 (Latency Optimized)	32.7	75.1

SqueezeNAS: Search Time Results

Name	NAS Method	Search Time (GPU Days)	Dataset Searched on
SqueezeNAS-3	gradient	7	Cityscapes
SqueezeNAS-9	gradient	11	Cityscapes
SqueezeNAS-23	gradient	14	Cityscapes
Neural Architecture Search with Reinforcement Learning	RL	22,400	CIFAR-10
NASNet	RL	2,000	CIFAR-10
mNasNet	RL	2,000*	Proxy ImageNet
AmoebaNet	genetic	3,150	CIFAR-10
FBNet	gradient	9	Proxy ImageNet
DARTS	gradient	4	CIFAR-10

* Approximated from TPUv2 Hours

Conclusions

- Deep learning applications, and their computing platforms, are more diverse than ever, necessitating the design of many new DNNs
- Good news! Neural Architecture Search (NAS) is 100-1000x more efficient today than it was 2 years ago
- SqueezeNAS has achieved a new speed vs accuracy curve created for Semantic Segmentation on an automotive-grade platform
- Some architecture patterns follow human intuition and some don't
    - We can learn new design paradigms from NAS
- Moving up a level of abstraction: Researchers can now design Neural Architecture search spaces instead of individual networks