CAP 5516
Medical Image Computing
(Spring 2022)

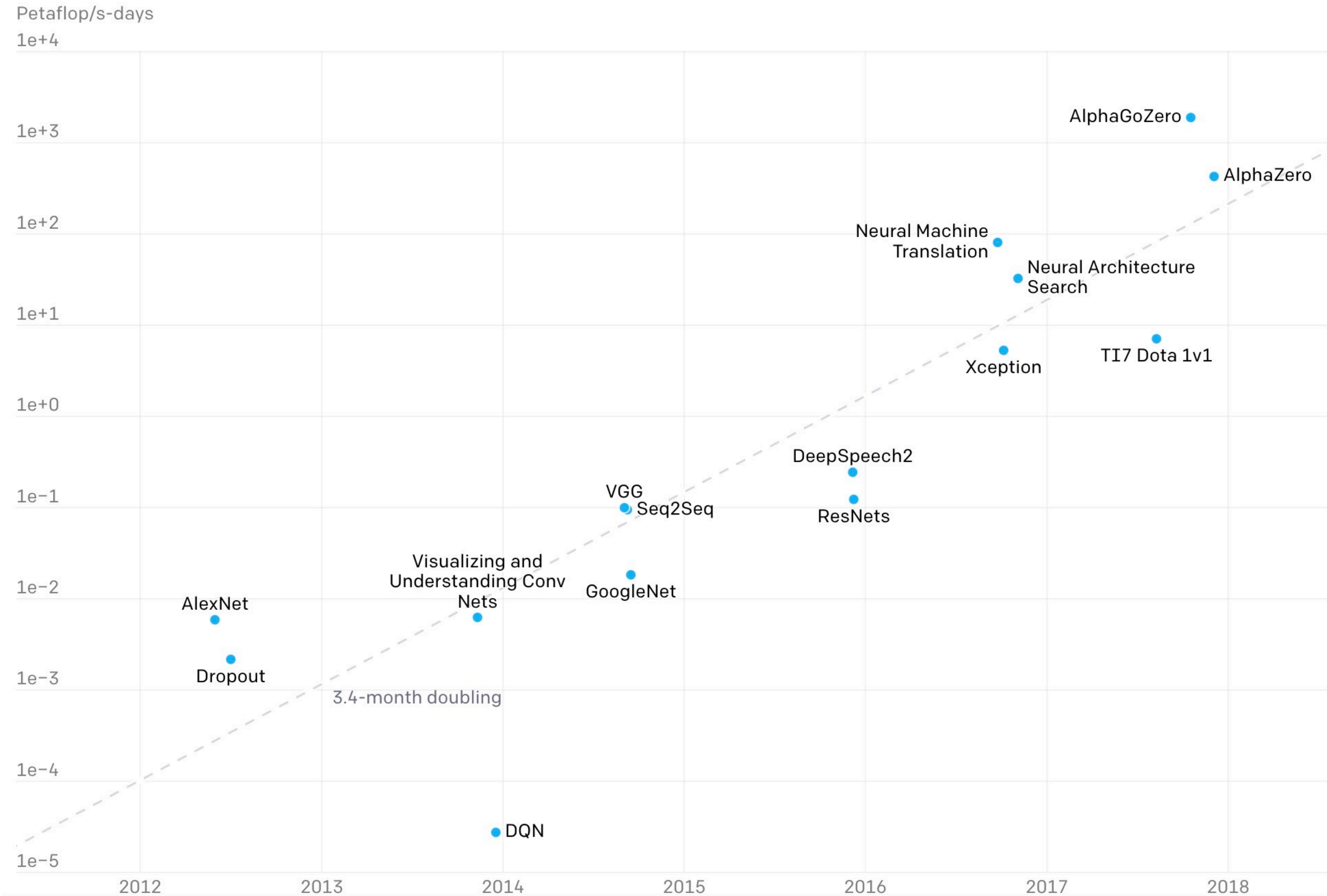
Dr. Chen Chen
Center for Research in Computer Vision (CRCV)
University of Central Florida
Office: HEC 221
Address: 4328 Scorpius St., Orlando, FL 32816-2365
Email: chen.chen@crcv.ucf.edu
Web: https://www.crcv.ucf.edu/chenchen/

UCF Center for Research in Computer Vision

Lecture 14
Deep Learning Model Efficiency

Compute Demands for Deep Neural Networks

AlexNet to AlphaGo Zero: A 300,000x Increase in Compute



https://openai.com/blog/ai-and-compute/

Popular DNN Models

Metrics	LeNet-5	AlexNet	VGG-16	GoogLeNet (v1)	ResNet-50	EfficientNet-B4
Top-5 error (ImageNet)	n/a	16.4	7.4	6.7	5.3	3.7*
Input Size	28x28	227x227	224x224	224x224	224x224	380x380
# of CONV Layers	**2**	**5**	**16**	**21 (depth)**	**49**	**96**
# of Weights	2.6k	2.3M	14.7M	6.0M	23.5M	14M
# of MACs	283k	666M	15.3G	1.43G	3.86G	4.4G
# of FC layers	**2**	**3**	**3**	**1**	**1**	**65****
# of Weights	58k	58.6M	124M	1M	2M	4.9M
# of MACs	58k	58.6M	124M	1M	2M	4.9M
Total Weights	**60k**	**61M**	**138M**	**7M**	**25.5M**	**19M**
Total MACs	**341k**	**724M**	**15.5G**	**1.43G**	**3.9G**	**4.4G**
Reference	**Lecun,** *PIEEE* 1998	**Krizhevsky,** *NeurIPS* 2012	**Simonyan,** *ICLR* 2015	**Szegedy,** *CVPR* 2015	**He,** *CVPR* 2016	**Tan,** *ICML* 2019

multiply and accumulate (MAC)

DNN models getting **larger** and **deeper**

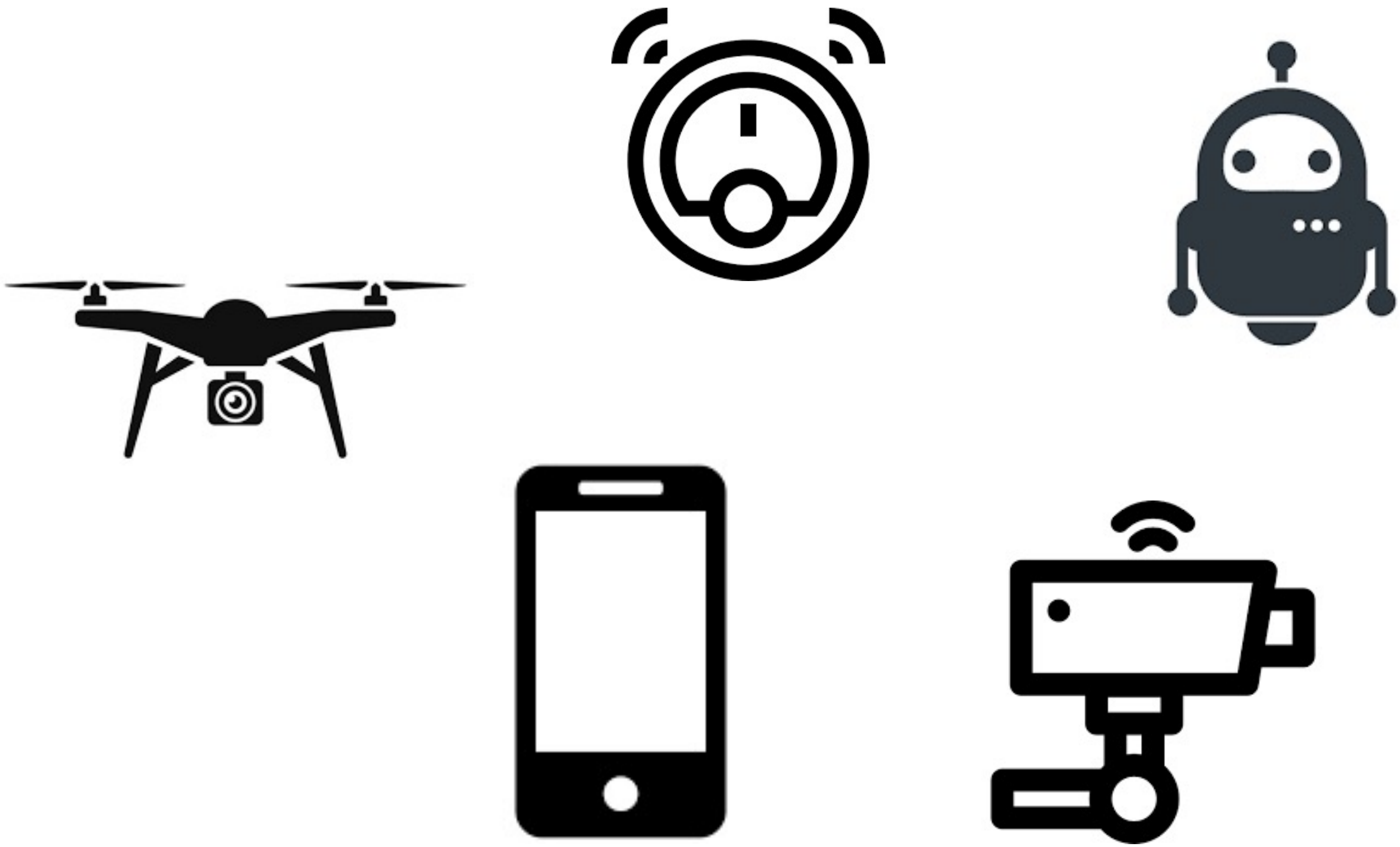
Large memory and computational cost

* Does not include multi-crop and ensemble

** Increase in FC layers due to squeeze-and-excitation layers (much smaller than FC layers for classification)

Credit: Vivienne Sze

Need Efficient Neural Networks for Real-World Applications

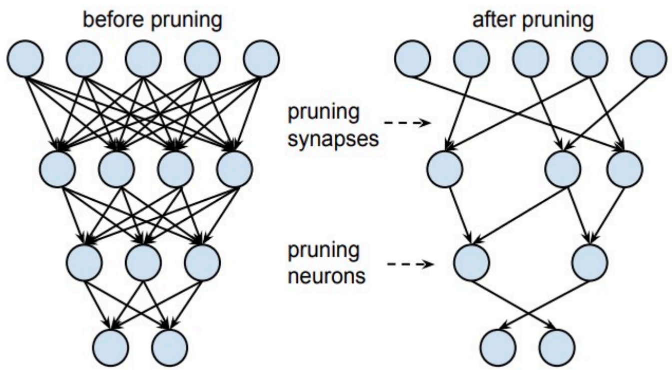


Smart edge devices with limited resources (e.g., memory and computation)

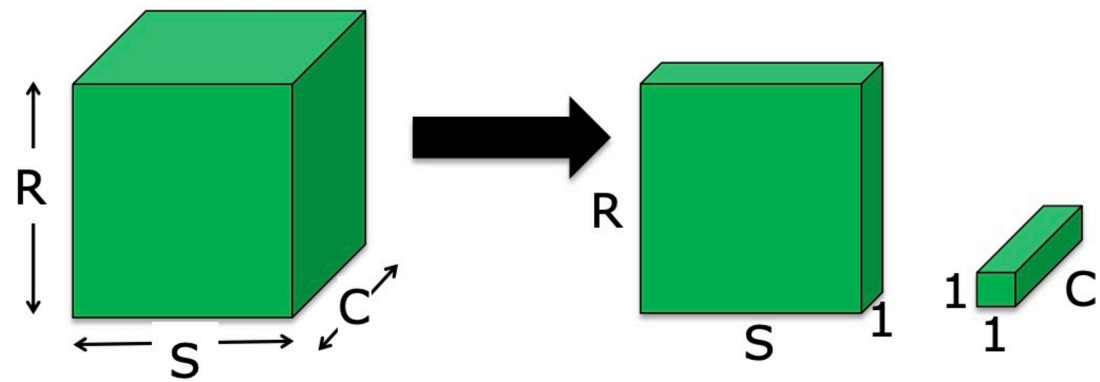
Efficient Neural Networks Design

Credit: Vivienne Sze

Network Pruning



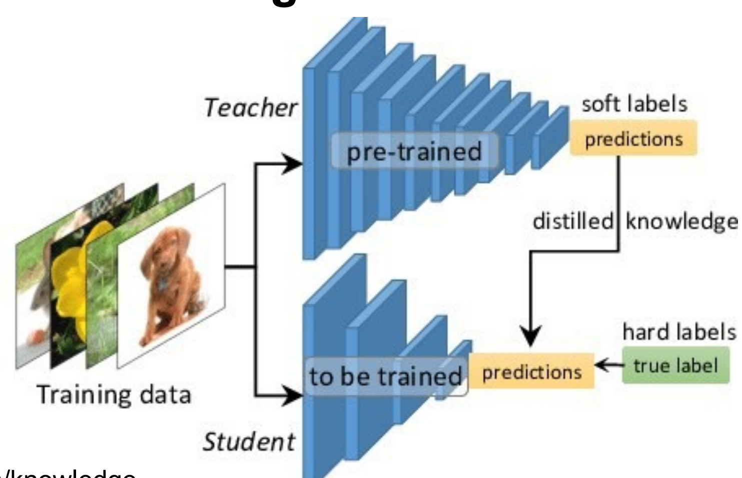
Efficient Network Architectures



[MobileNets, ShuffleNets, AdderNet]

Reduce Precision

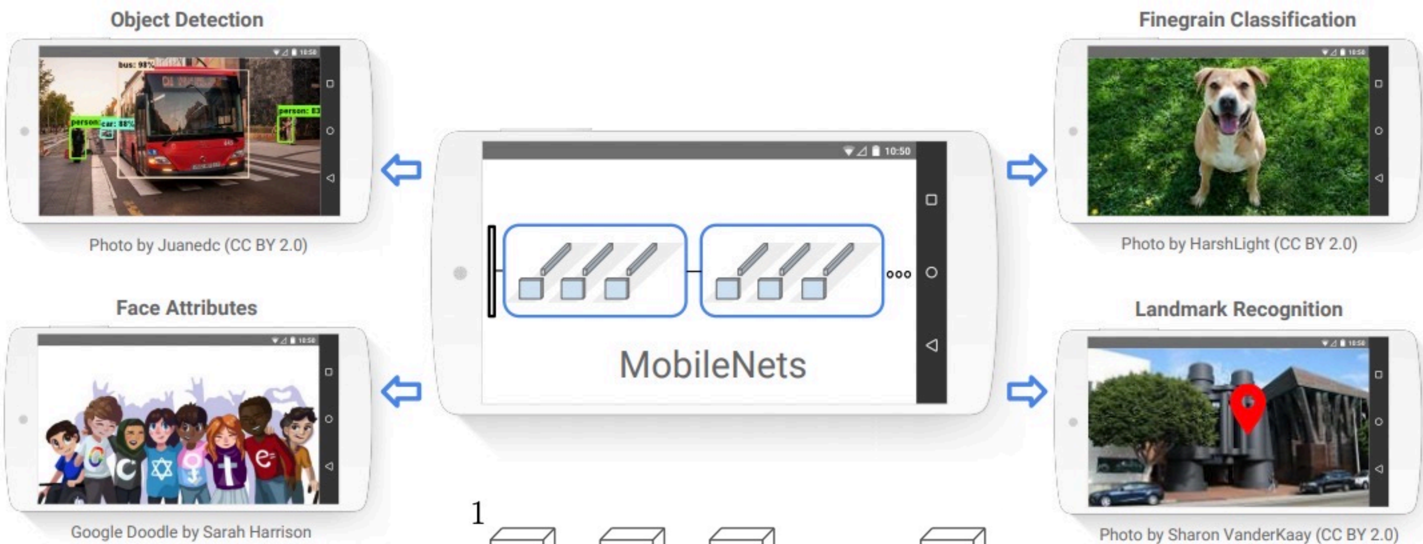
Knowledge Distillation

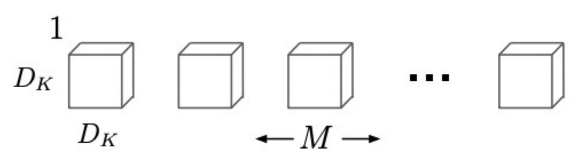


Source:
https://towardsdatascience.com/knowledge-distillation-simplified-dd4973dbc764

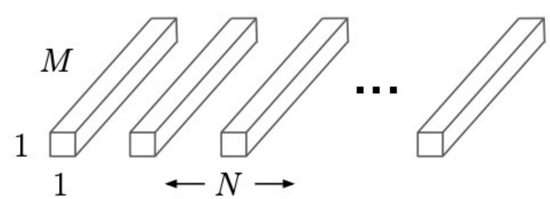
Efficient Network Architectures

- MobileNet





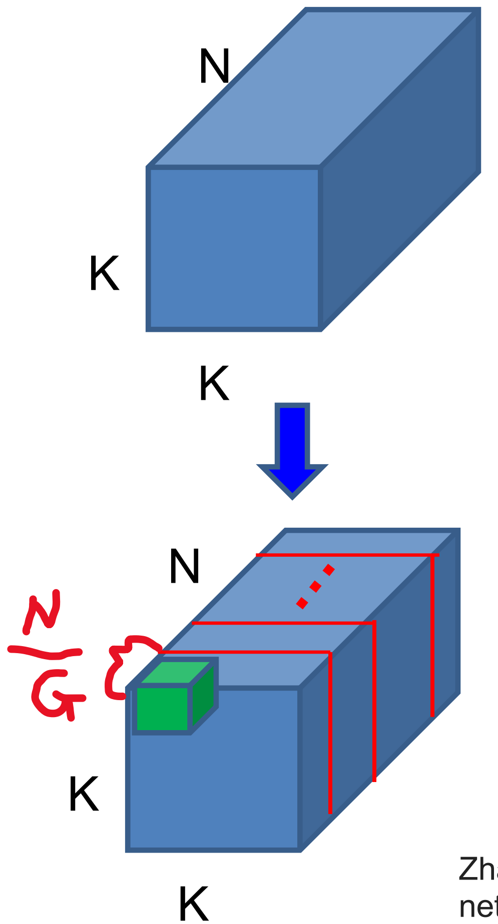
(b) Depthwise Convolutional Filters



Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." arXiv preprint arXiv:1704.04861 (2017).

Efficient Network Architectures

- ShuffleNet
  - Group convolution



M filters/kernels are also divided into G groups

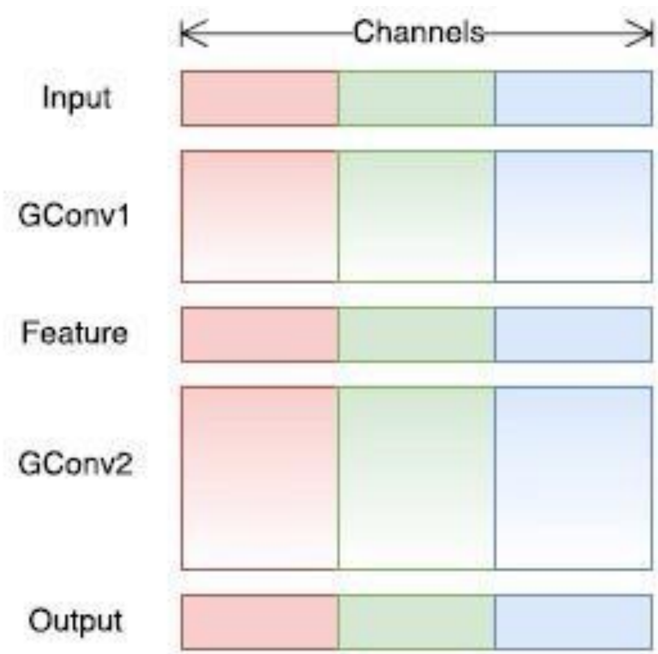
Each group has M/G filters

In each group, the filter has size: m x m x (N/G)

Zhang, Xiangyu, et al. "Shufflenet: An extremely efficient convolutional neural network for mobile devices." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

Efficient Network Architectures

- Group convolution



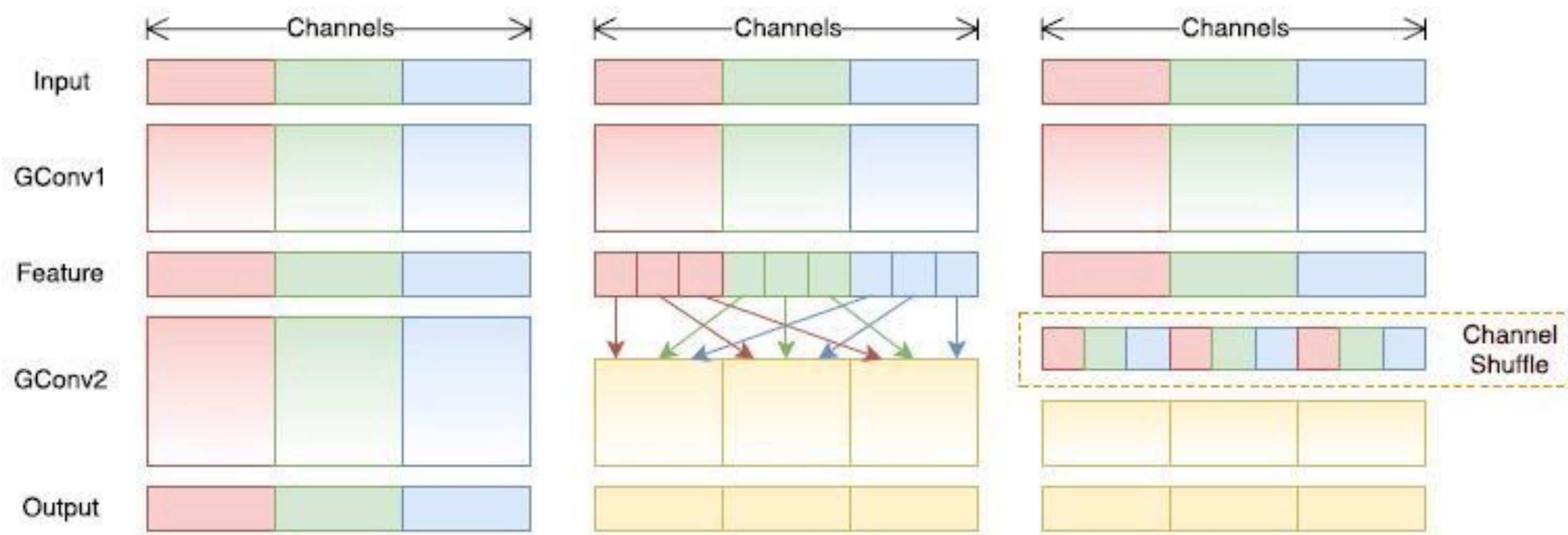
If multiple group convolutions stack together, there is one side effect!

Outputs from a certain group only relate to the inputs within the group.

No information exchange across groups.

Efficient Network Architectures

- Shuffled Group convolution



Zhang, Xiangyu, et al. "Shufflenet: An extremely efficient convolutional neural network for mobile devices." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

Efficient Network Architectures

- GhostNet

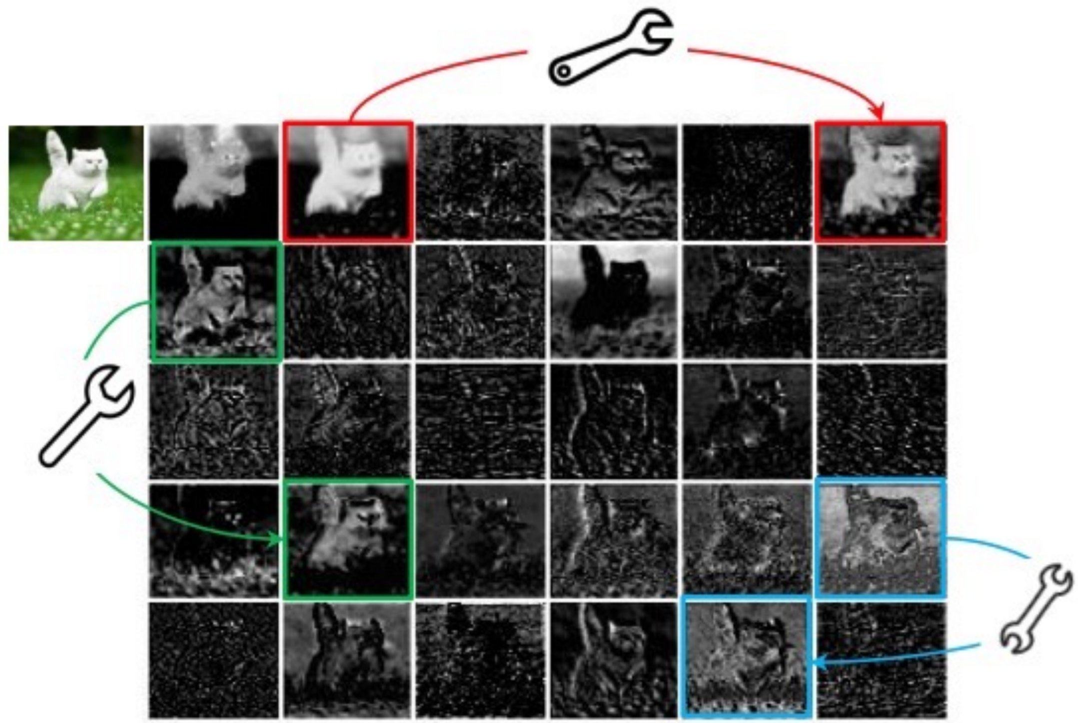
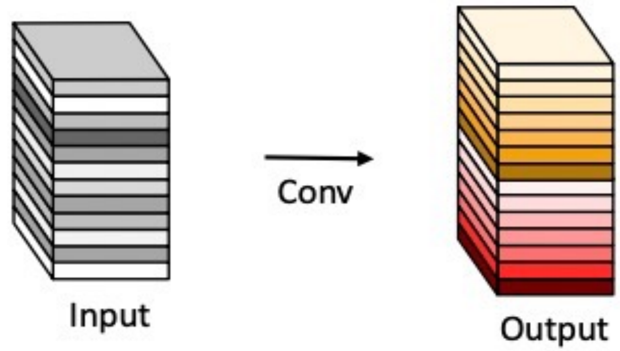


Figure 1. Visualization of some feature maps generated by the first residual group in ResNet-50, where three similar feature map pair examples are annotated with boxes of the same color. One feature map in the pair can be approximately obtained by transforming the other one through cheap operations (denoted by spanners).

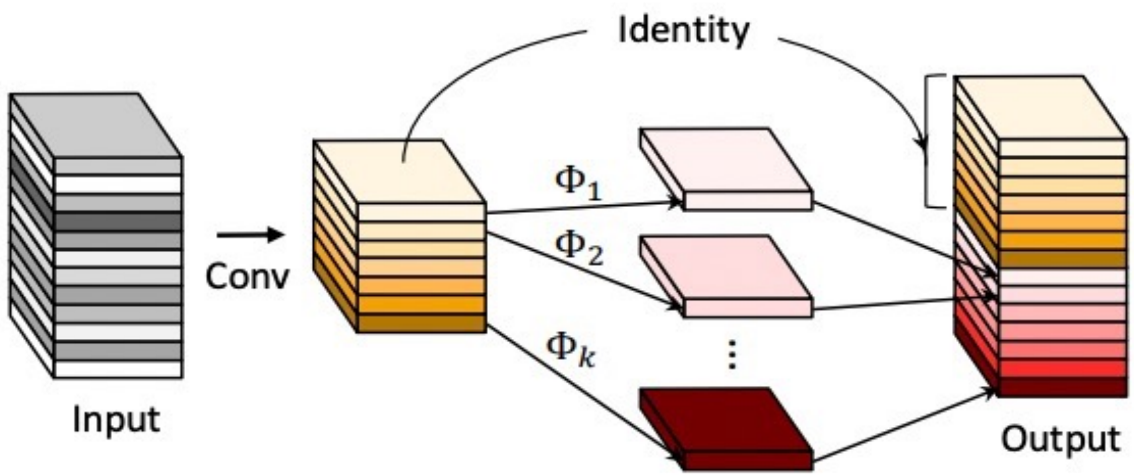
Han, Kai, et al. "Ghostnet: More features from cheap operations." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

Efficient Network Architectures

- GhostNet



(a) The convolutional layer.



(b) The Ghost module.

Figure 2. An illustration of the convolutional layer and the proposed Ghost module for outputting the same number of feature maps. Φ represents the cheap operation.

Han, Kai, et al. "Ghostnet: More features from cheap operations." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

Efficient Network Architectures

- GhostNet

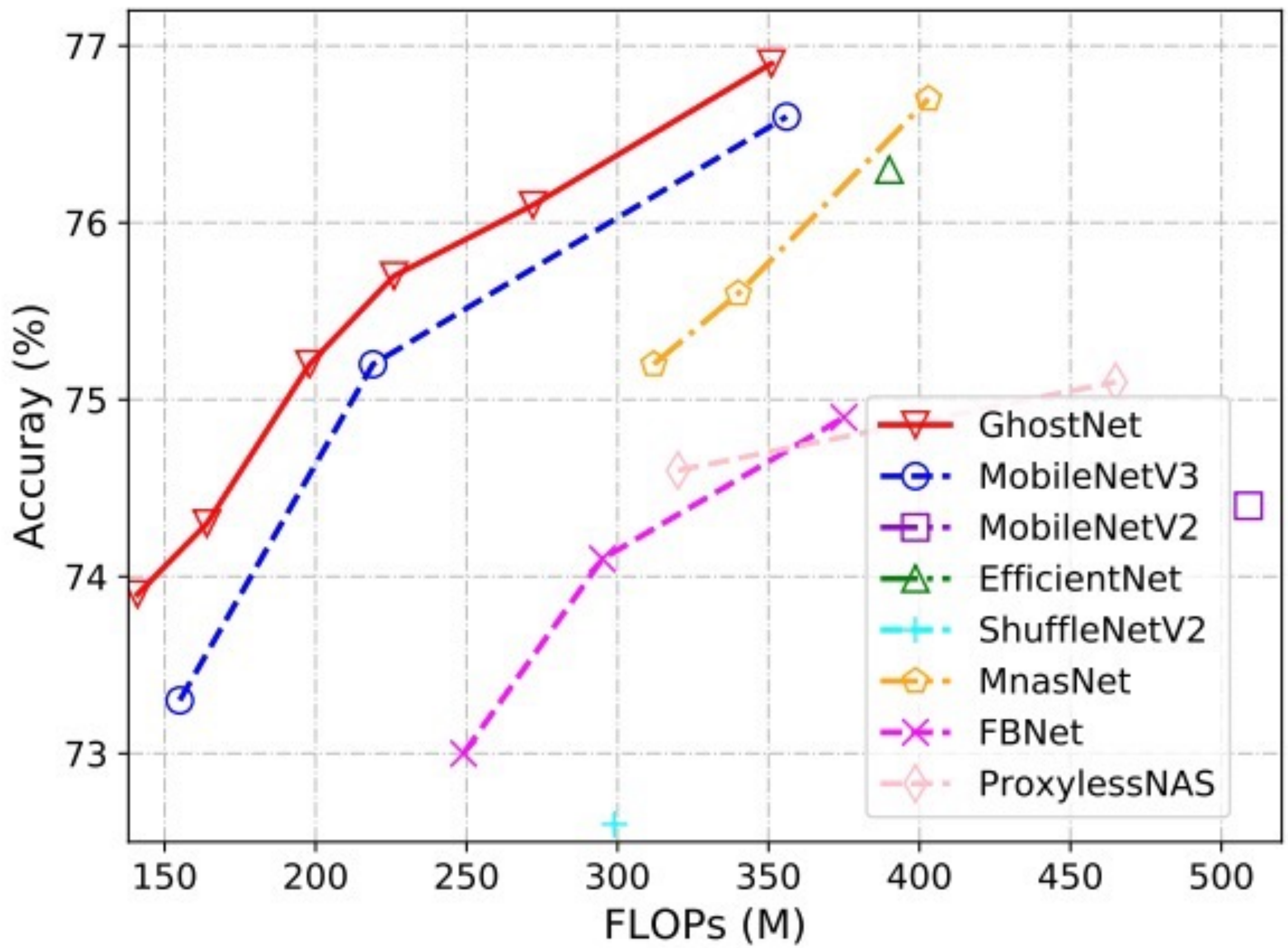
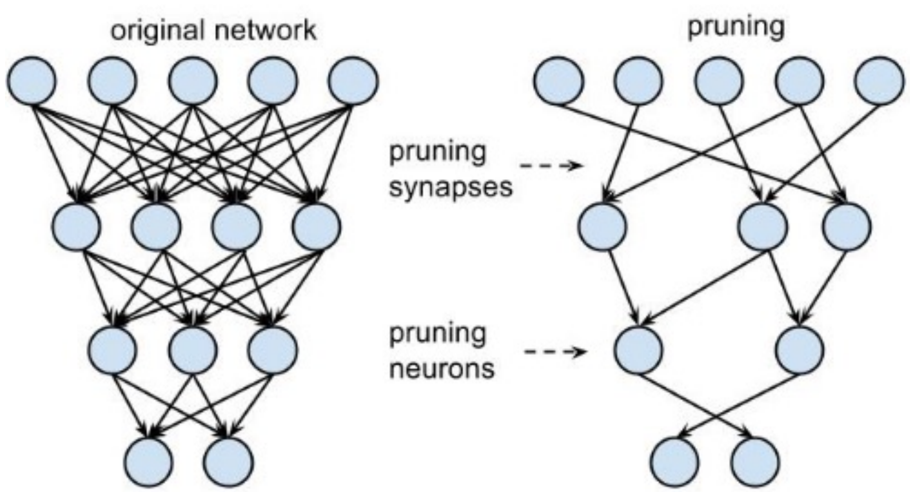


Figure 6. Top-1 accuracy *v.s.* FLOPs on ImageNet dataset.

Han, Kai, et al. "Ghostnet: More features from cheap operations." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

Network Pruning

- Remove weights/synapses “close to zero”
- **Retrain** to maintain accuracy
- Repeat



Sparse Network

Credit: Song Han

Network Pruning

- **Unstructured Pruning** methods prune individual parameters.
- Doing so produces a sparse neural network which, although smaller in terms of parameter count, may not be arranged in a fashion conducive to speed enhancements using modern libraries and hardware.
- This is also called **Weight Pruning** as we set individual weights in the weight matrix to zero.

https://blog.paperspace.com/neural-network-pruning-explained/

Network Pruning

- **Structured Pruning** methods consider parameters in groups, removing entire neurons, filters, or channels to exploit hardware and software optimized for dense computation.
- This is also called **Unit/Neuron Pruning**, as we set entire columns in the weight matrix to zero, in effect deleting the corresponding output neuron.

https://blog.paperspace.com/neural-network-pruning-explained/

Network Pruning

(a) Test Errors on CIFAR-10

Model	Test error (%)	Parameters	Pruned	FLOPs	Pruned
VGGNet (Baseline)	6.34	20.04M	-	7.97×10^8	-
VGGNet (70% Pruned)	**6.20**	2.30M	88.5%	3.91×10^8	51.0%
DenseNet-40 (Baseline)	6.11	1.02M	-	5.33×10^8	-
DenseNet-40 (40% Pruned)	**5.19**	0.66M	35.7%	3.81×10^8	28.4%
DenseNet-40 (70% Pruned)	5.65	0.35M	65.2%	2.40×10^8	55.0%
ResNet-164 (Baseline)	5.42	1.70M	-	4.99×10^8	-
ResNet-164 (40% Pruned)	**5.08**	1.44M	14.9%	3.81×10^8	23.7%
ResNet-164 (60% Pruned)	5.27	1.10M	35.2%	2.75×10^8	44.9%

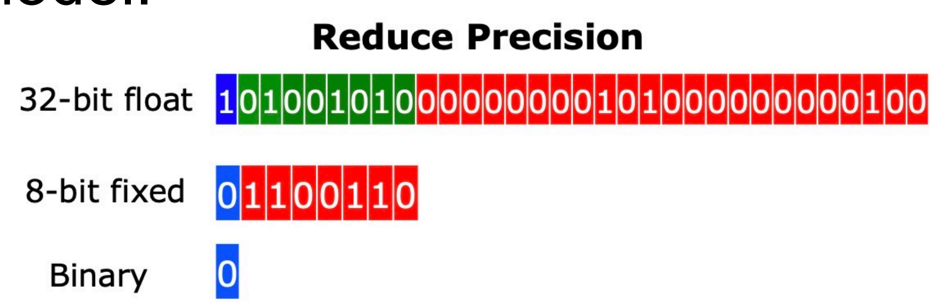
(b) Test Errors on CIFAR-100

Model	Test error (%)	Parameters	Pruned	FLOPs	Pruned
VGGNet (Baseline)	26.74	20.08M	-	7.97×10^8	-
VGGNet (50% Pruned)	**26.52**	5.00M	75.1%	5.01×10^8	37.1%
DenseNet-40 (Baseline)	25.36	1.06M	-	5.33×10^8	-
DenseNet-40 (40% Pruned)	**25.28**	0.66M	37.5%	3.71×10^8	30.3%
DenseNet-40 (60% Pruned)	25.72	0.46M	54.6%	2.81×10^8	47.1%
ResNet-164 (Baseline)	23.37	1.73M	-	5.00×10^8	-
ResNet-164 (40% Pruned)	**22.87**	1.46M	15.5%	3.33×10^8	33.3%
ResNet-164 (60% Pruned)	23.91	1.21M	29.7%	2.47×10^8	50.6%

Liu, Zhuang, et al. "Learning efficient convolutional networks through network slimming." *Proceedings of the IEEE international conference on computer vision*. 2017.

Network Quantization

- Quantization for deep learning is the process of approximating a neural network that uses floating-point numbers by a neural network of low bit width numbers.
- Network quantization dramatically reduces both the memory requirement and computational cost of using neural networks.
- We assume that we have the trained model parameters θ, stored in floating point precision. In quantization, the goal is to reduce the precision of both the parameters (θ), as well as the intermediate activation maps to low-precision, with minimal impact on the generalization power/accuracy of the model.



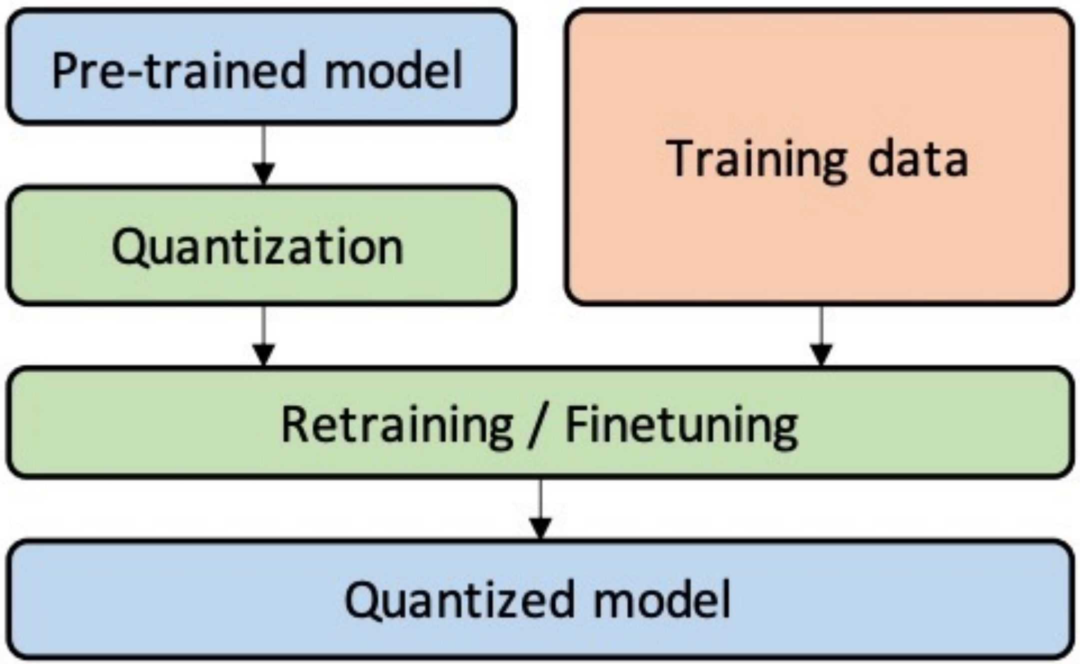
UCF CENTER FOR RESEARCH IN COMPUTER VISION

Network Quantization

- Quantization methods can be roughly divided into two categories:
  - quantization aware training (QAT)
  - post-training quantization (PTQ)

Network Quantization

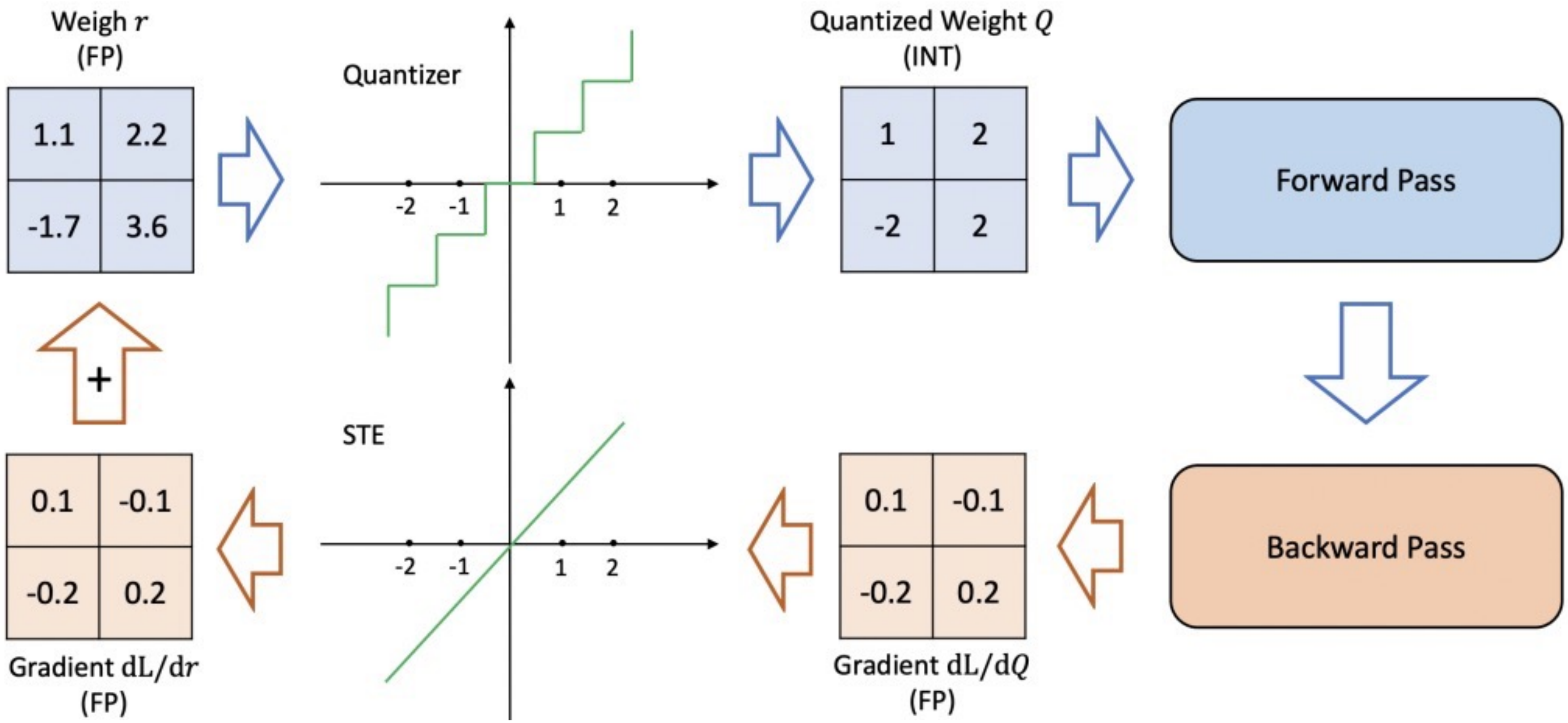
- Quantization aware training (QAT)
  - In QAT, a pre-trained model is quantized and then finetuned using training data to adjust parameters and recover accuracy degradation



Gholami, Amir, et al. "A survey of quantization methods for efficient neural network inference." arXiv preprint arXiv:2103.13630 (2021).

Network Quantization

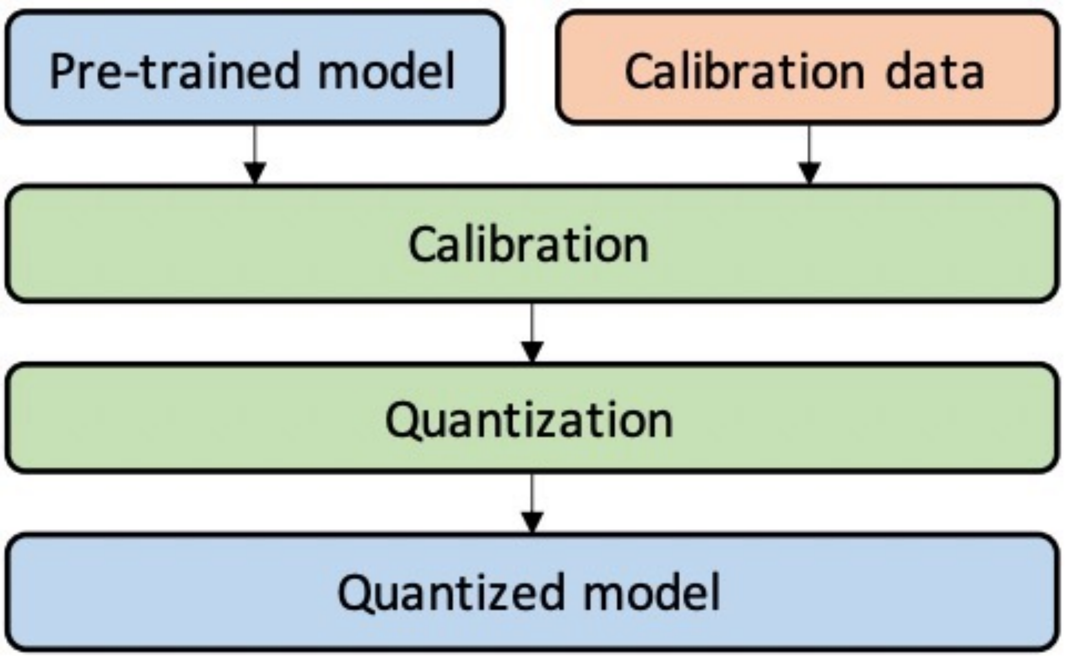
- Quantization aware training (QAT) procedure



Gholami, Amir, et al. "A survey of quantization methods for efficient neural network inference." arXiv preprint arXiv:2103.13630 (2021).

Network Quantization

- Post-Training Quantization (PTQ)
  - In PTQ, a pre-trained model is calibrated using calibration data (e.g., a small subset of training data) to compute the clipping ranges and the scaling factors. Then, the model is quantized based on the calibration result.



Gholami, Amir, et al. "A survey of quantization methods for efficient neural network inference." arXiv preprint arXiv:2103.13630 (2021).

Network Quantization

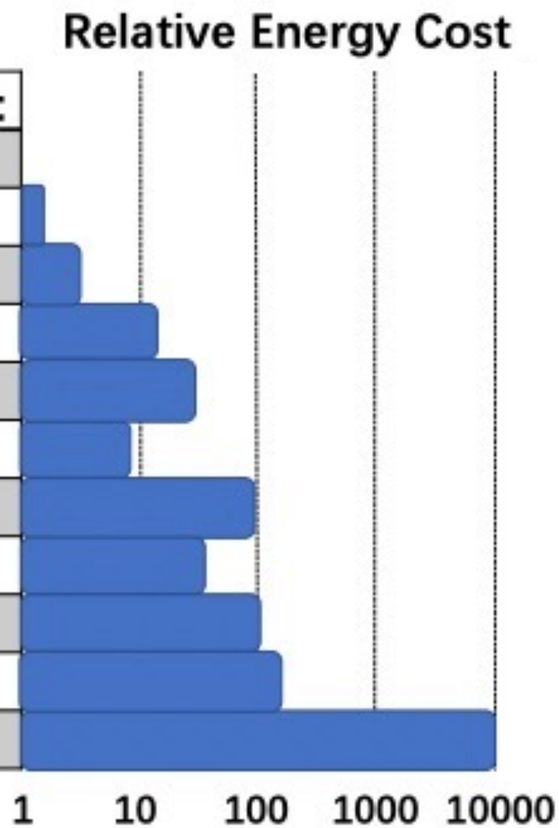
- Quantization methods can be roughly divided into two categories:
  - quantization aware training (QAT)
  - post-training quantization (PTQ)

- QAT methods usually achieve better results than PTQ methods. PTQ methods are simpler and add quantization to a given network model without any training process.

Network Quantization

- Lower precision provides exponentially better energy efficiency

Operation:	Energy(pJ):
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8kb)	5.0
32b DRAM Read	640



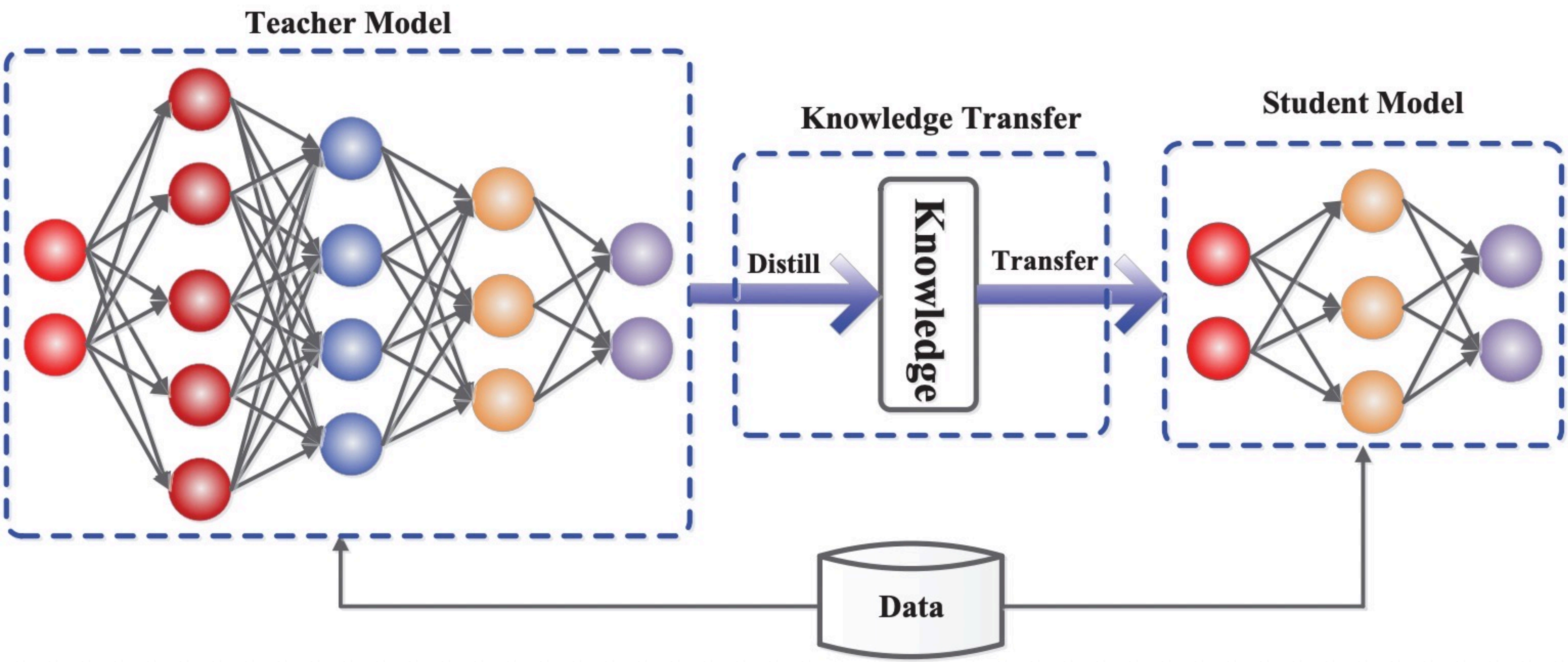
Comparison of the corresponding energy cost for different precision for 45nm technology.

Gholami, Amir, et al. "A survey of quantization methods for efficient neural network inference." arXiv preprint arXiv:2103.13630 (2021).

Knowledge Distillation

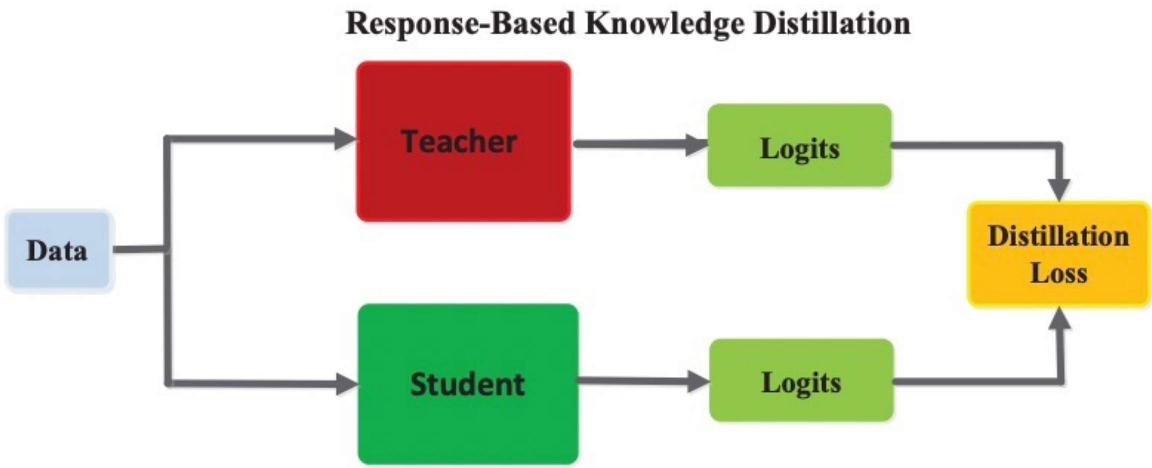
- Knowledge distillation is a process of distilling or transferring the knowledge from a (set of) large, cumbersome model(s) to a lighter, easier-to-deploy single model, without significant loss in performance.

Knowledge Distillation



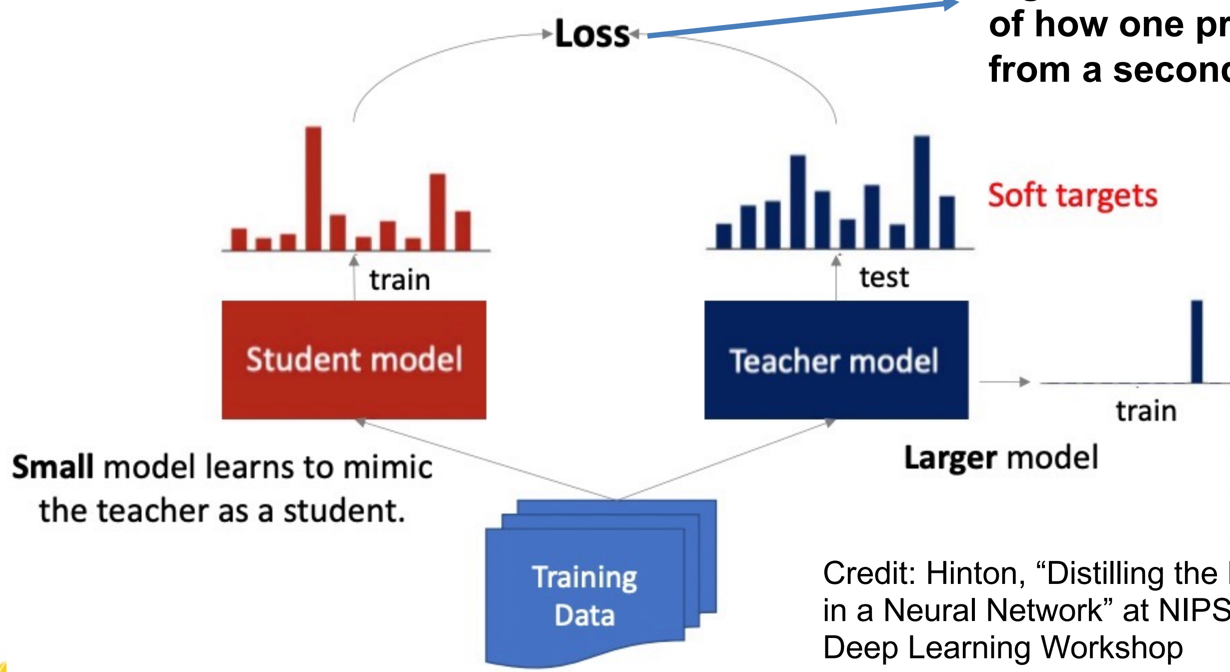
Gou, Jianping, et al. "Knowledge distillation: A survey." International Journal of Computer Vision 129.6 (2021): 1789-1819.

Knowledge Distillation



Gou, Jianping, et al. "Knowledge distillation: A survey." International Journal of Computer Vision 129.6 (2021): 1789-1819.

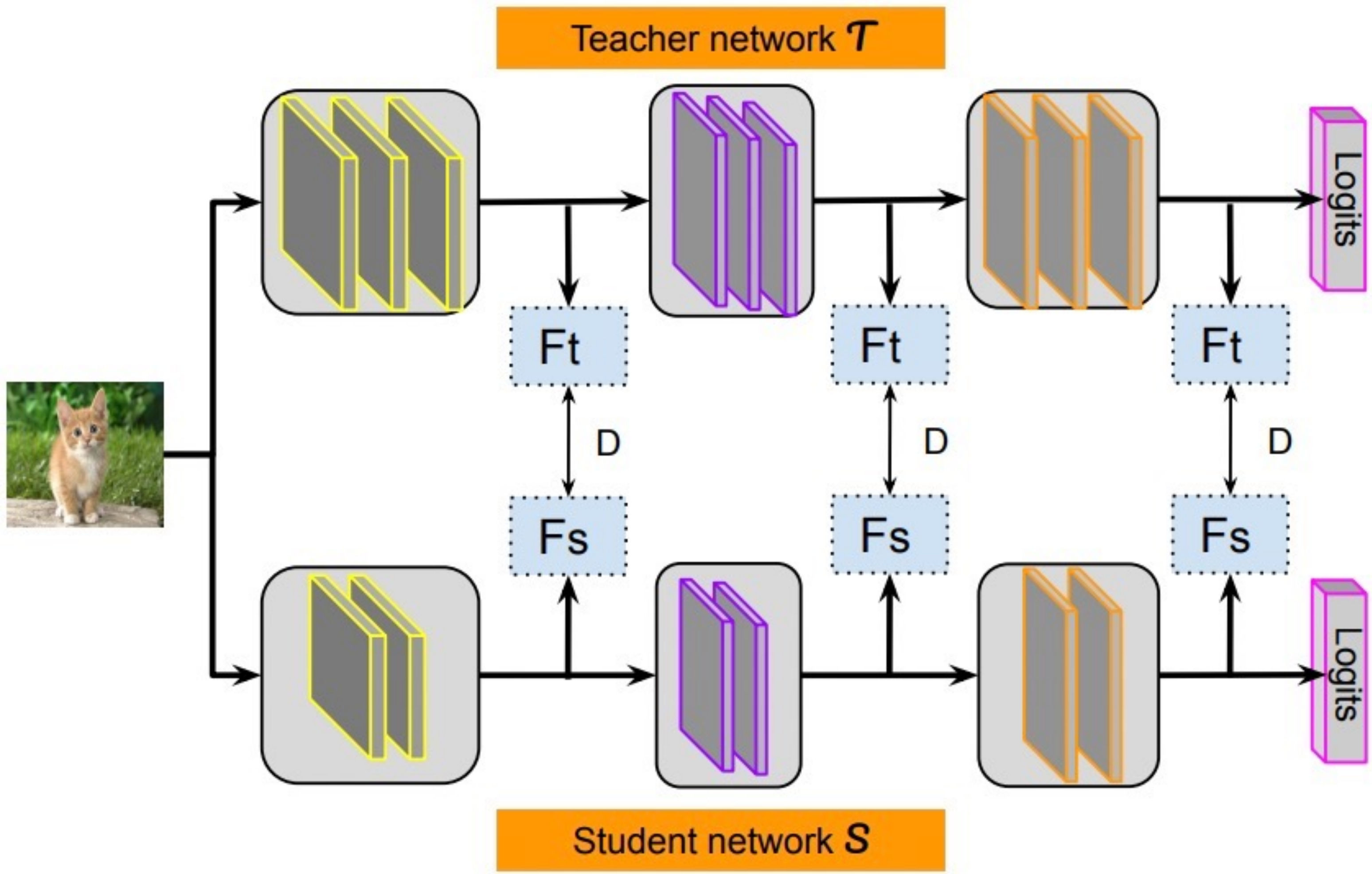
e.g., the Kullback–Leibler divergence: a measure of how one probability distribution Q is different from a second, reference probability distribution P



Credit: Hinton, “Distilling the Knowledge in a Neural Network” at NIPS 2014 Deep Learning Workshop

Knowledge Distillation

Feature-based knowledge distillation



Wang, Lin, and Kuk-Jin Yoon. "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

Knowledge Distillation

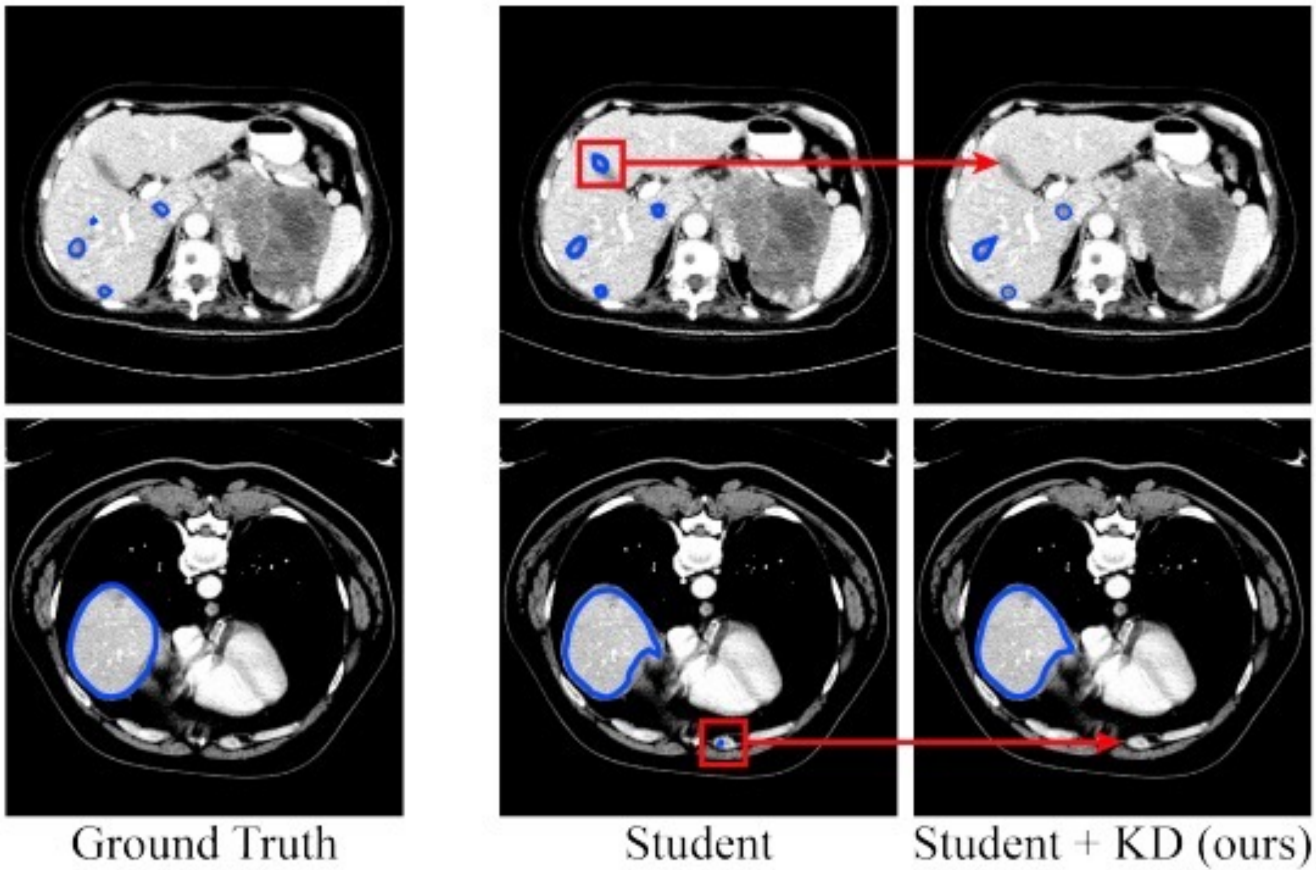
Table 5 Performance comparison of different knowledge distillation methods on CIFAR10. Note that ↑ indicates the performance improvement of the student network learned by each method comparing with the corresponding baseline model.

Offline Distillation				
Methods	Knowledge	Teacher (baseline)	Student (baseline)	Accuracies
FSP (Yim et al., 2017)	RelK	ResNet26 (91.91)	ResNet8 (87.91)	88.70 (0.79 ↑)
FT (Kim et al., 2018)	FeaK	ResNet56 (93.61)	ResNet20 (92.22)	93.15 (0.93 ↑)
IRG (Liu et al., 2019g)	RelK	ResNet20 (91.45)	ResNet20-x0.5 (88.36)	90.69 (2.33 ↑)
SP (Tung and Mori, 2019)	RelK	WRN-40-1 (93.49)	WRN-16-1 (91.26)	91.87 (0.61 ↑)
SP (Tung and Mori, 2019)	RelK	WRN-40-2 (95.76)	WRN-16-8 (94.82)	95.45 (0.63 ↑)
FN (Xu et al., 2020b)	FeaK	ResNet110 (94.29)	ResNet56 (93.63)	94.14 (0.51 ↑)
FN (Xu et al., 2020b)	FeaK	ResNet56 (93.63)	ResNet20 (92.11)	92.67 (0.56 ↑)
AdaIN (Yang et al., 2020a)	FeaK	ResNet26 (93.58)	ResNet8 (87.78)	89.02 (1.24 ↑)
AdaIN (Yang et al., 2020a)	FeaK	WRN-40-2 (95.07)	WRN-16-2 (93.98)	94.67 (0.69 ↑)
AE-KD (Du et al., 2020)	FeaK	ResNet56 (—)	MobileNetV2 (75.97)	77.07 (1.10 ↑)
JointRD (Li et al., 2020b)	FeaK	ResNet34 (95.39)	plain-CNN 34 (93.73)	94.78 (1.05 ↑)
TOFD (Zhang et al., 2020a)	FeaK	ResNet152 (—)	ResNeXt50-4 (94.49)	97.09 (2.60 ↑)
TOFD (Zhang et al., 2020a)	FeaK	ResNet152 (—)	MobileNetV2 (90.43)	93.34 (2.91 ↑)
CTKD (Zhao et al., 2020a)	RelK, FeaK	WRN-40-1 (93.43)	WRN-16-1 (91.28)	92.50 (1.22 ↑)
CTKD (Zhao et al., 2020a)	RelK, FeaK	WRN-40-2 (94.70)	WRN-16-2 (93.68)	94.42 (0.74 ↑)

Gou, Jianping, et al. "Knowledge distillation: A survey." International Journal of Computer Vision 129.6 (2021): 1789-1819.

Case Study (Medical Imaging)

- Efficient Medical Image Segmentation Based on Knowledge Distillation

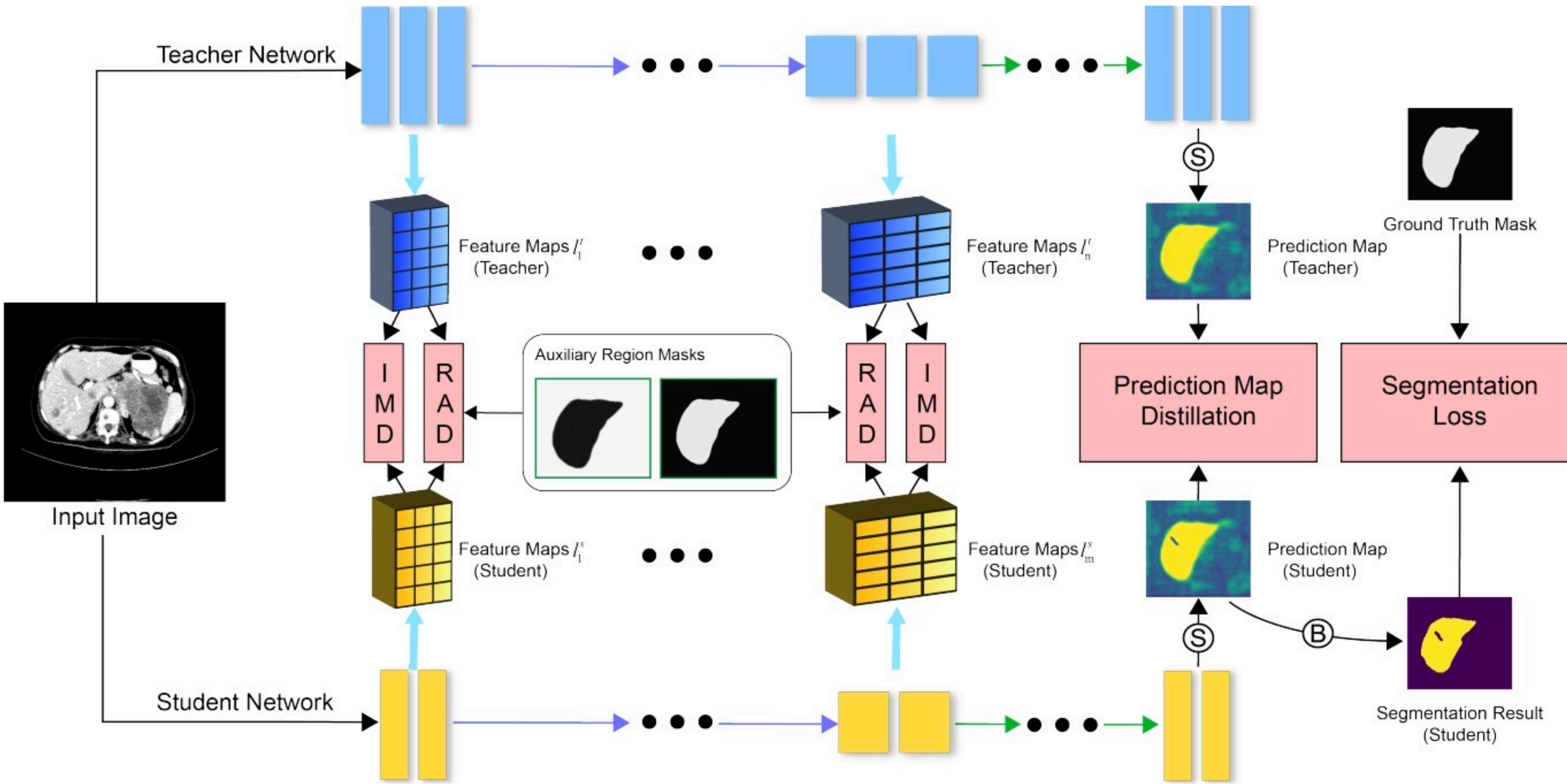


Liver segmentation

Qin, Dian, et al. "Efficient medical image segmentation based on knowledge distillation." *IEEE Transactions on Medical Imaging* 40.12 (2021): 3820-3831.

Case Study (Medical Imaging)

- Efficient Medical Image Segmentation Based on Knowledge Distillation



Qin, Dian, et al. "Efficient medical image segmentation based on knowledge distillation." *IEEE Transactions on Medical Imaging* 40.12 (2021): 3820-3831.

Case Study (Medical Imaging)

- Efficient Medical Image Segmentation Based on Knowledge Distillation

Method	Liver Tumor Dice	Liver Dice	Kidney Tumor Dice	Kidney Dice	#Params (M)
		Teachers			
T1: RA-UNet	0.685 ± 0.004	0.960 ± 0.001	0.745 ± 0.003	0.970 ± 0.001	22.1
T2: PSPNet	0.640 ± 0.005	0.959 ± 0.001	0.659 ± 0.007	0.968 ± 0.002	46.7
T3: UNet++	0.669 ± 0.003	0.949 ± 0.001	0.644 ± 0.007	0.943 ± 0.002	20.6
		Students and their performances distilled from different teachers by our approach			
ENet	0.574 ± 0.005	0.952 ± 0.001	0.521 ± 0.015	0.939 ± 0.001	0.353
ENet + T1 (ours)	$\mathbf{0.652 \pm 0.005}$	$\mathbf{0.959 \pm 0.001}$	0.676 ± 0.007	0.965 ± 0.001	
ENet + T2 (ours)	0.635 ± 0.003	0.958 ± 0.001	0.599 ± 0.009	$\mathbf{0.967 \pm 0.001}$	
ENet + T3 (ours)	0.634 ± 0.004	0.953 ± 0.001	0.648 ± 0.008	0.941 ± 0.001	
MobileNetV2	0.540 ± 0.003	0.921 ± 0.002	0.516 ± 0.009	0.945 ± 0.001	2.2
MobileNetV2 + T1 (ours)	0.595 ± 0.004	0.932 ± 0.002	$\mathbf{0.684 \pm 0.006}$	0.952 ± 0.001	
MobileNetV2 + T2 (ours)	0.590 ± 0.006	0.927 ± 0.002	0.678 ± 0.003	0.949 ± 0.001	
MobileNetV2 + T3 (ours)	0.589 ± 0.002	0.924 ± 0.001	0.679 ± 0.005	n/a	
ResNet18	0.464 ± 0.008	0.934 ± 0.001	0.435 ± 0.005	0.933 ± 0.001	11.2
ResNet18 + T1 (ours)	0.508 ± 0.004	0.943 ± 0.001	0.582 ± 0.008	0.939 ± 0.001	
ResNet18 + T2 (ours)	0.491 ± 0.004	0.946 ± 0.001	0.551 ± 0.005	0.941 ± 0.001	
ResNet18 + T3 (ours)	0.508 ± 0.006	0.935 ± 0.001	0.450 ± 0.009	0.934 ± 0.001	

Qin, Dian, et al. "Efficient medical image segmentation based on knowledge distillation." *IEEE Transactions on Medical Imaging* 40.12 (2021): 3820-3831.

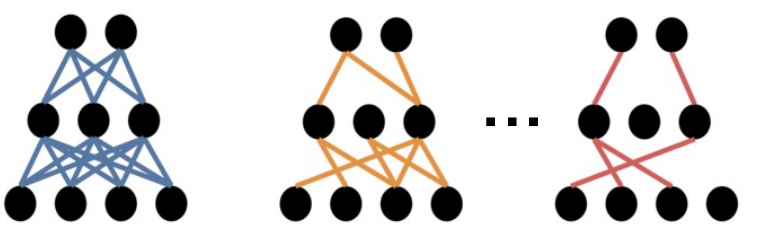
Dynamic Networks

- How to cope with dynamic resources and achieve trade-off between accuracy and efficiency?
- One possible solution: install all the possible model variants with various resource-accuracy trade-offs in the heterogeneous AI systems
  - Consumes more memory and storage
  - Not scalable

Different models with different sizes

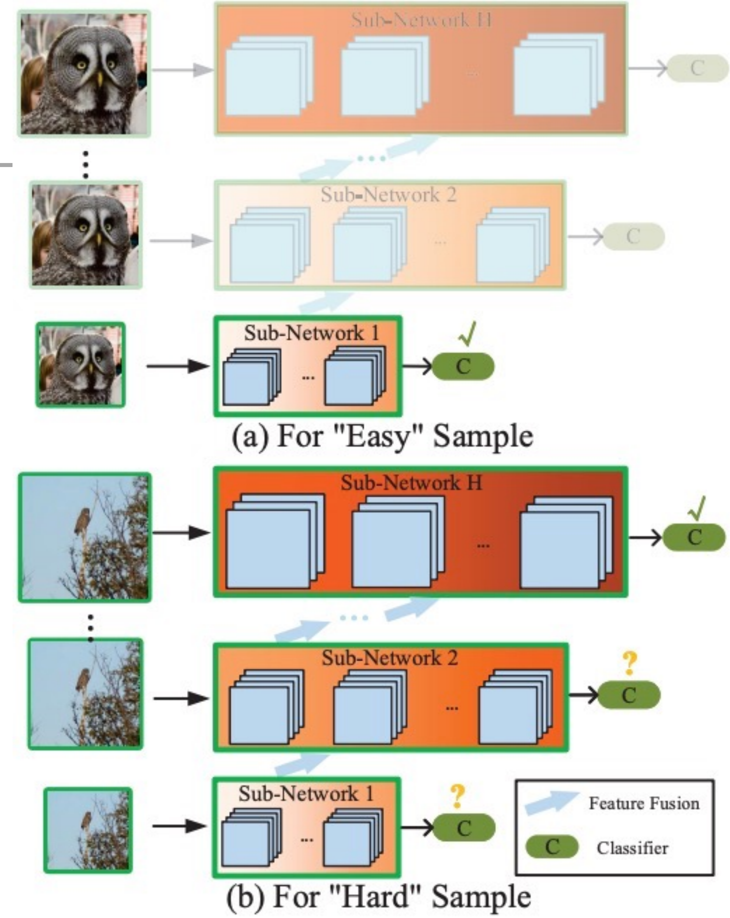
Model	Params	FLOPs
ResNet-50	25.5M	4.1G
MobileNet v1	4.2M	569M
MobileNet v2	3.5M	300M

Pruned networks with various pruning ratios

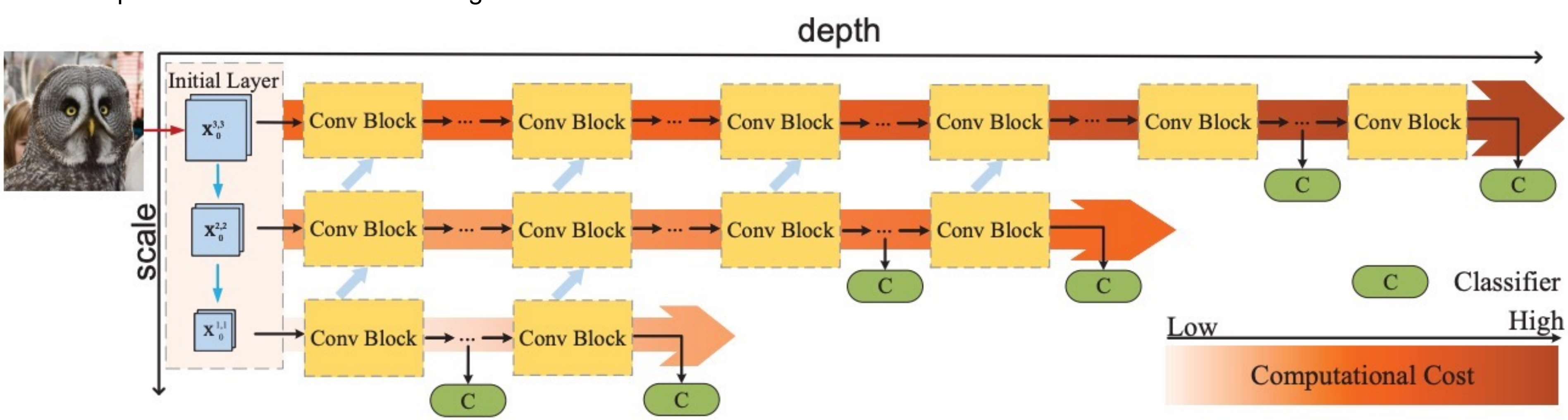


Dynamic Networks

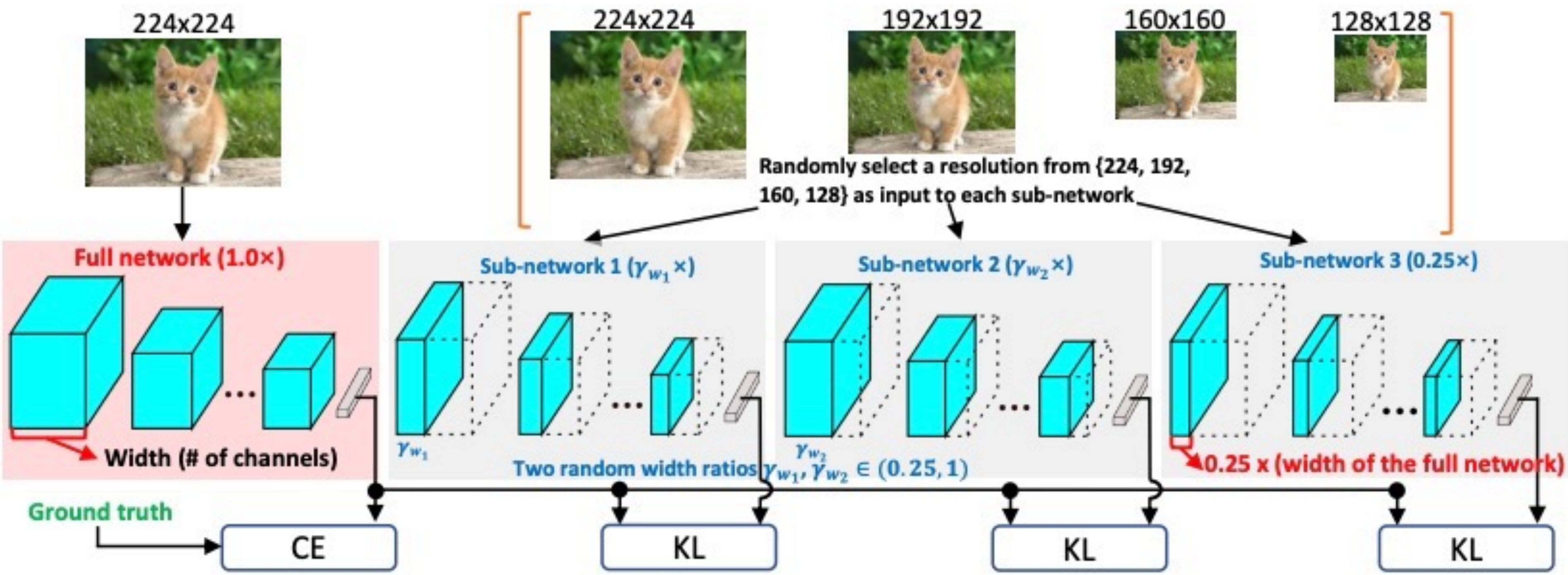
In RANet, the input images are first routed to a lightweight sub-network that efficiently extracts low-resolution representations, and those samples with high prediction confidence will exit early from the network without being further processed. Meanwhile, high-resolution paths in the network maintain the capability to recognize the “hard” samples.



Yang, Le, et al. "Resolution adaptive networks for efficient inference." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

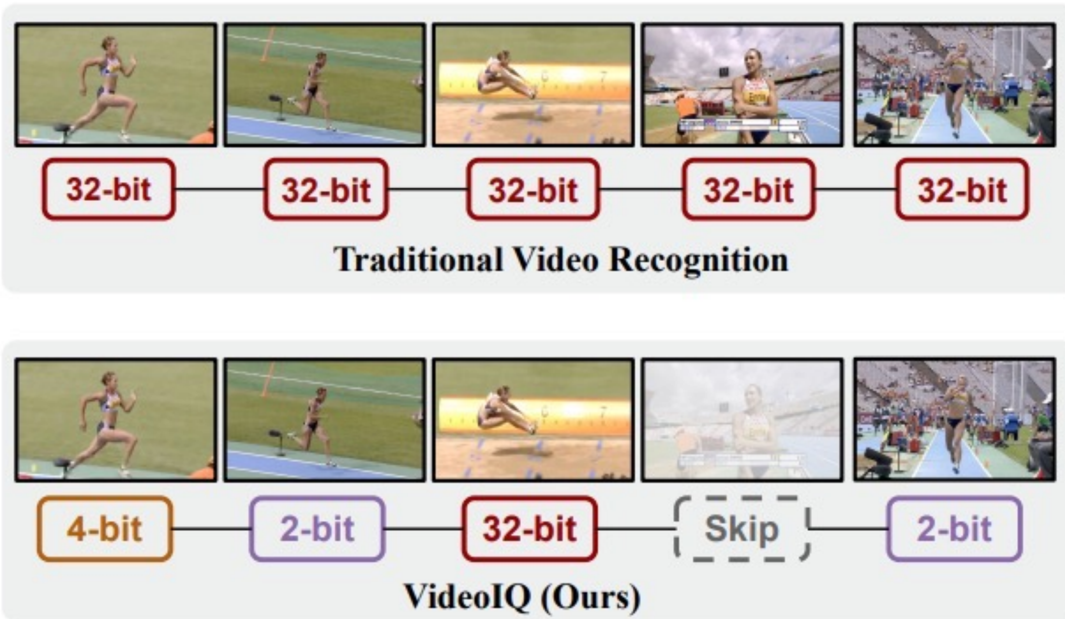


Dynamic Networks

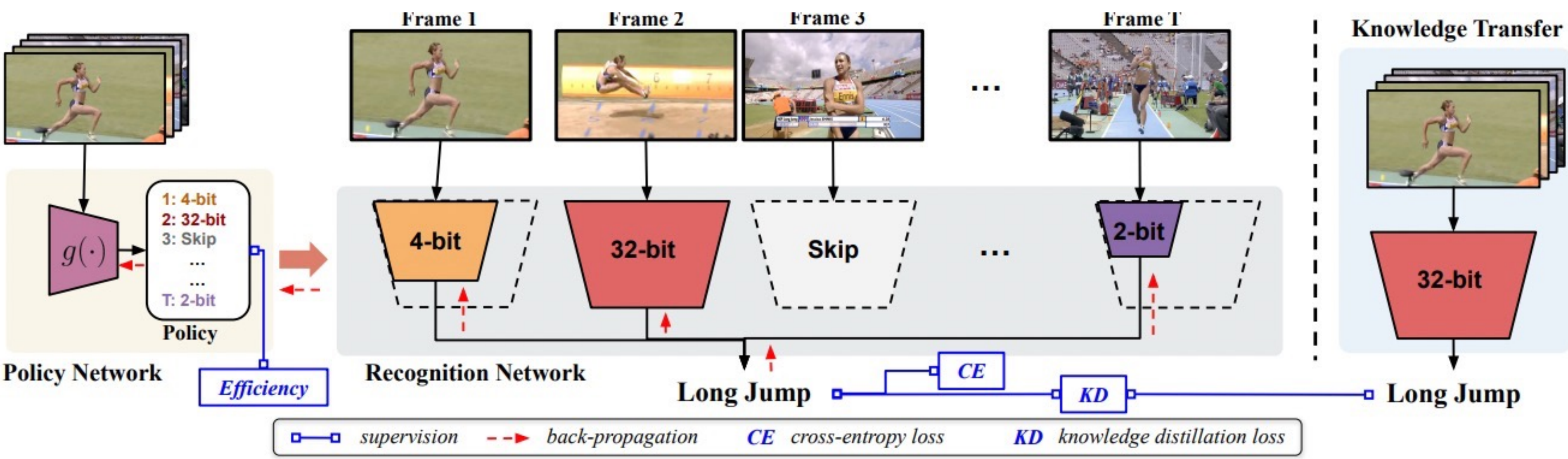


Yang, Taojiannan, et al. "MutualNet: Adaptive ConvNet via Mutual Learning from Different Model Configurations." IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).

Dynamic Networks



Sun, Ximeng, et al. "Dynamic network quantization for efficient video inference." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.



Can we design dynamic networks for medical imaging?

Paper Presentation

Student Paper Presentation

- 3/1 - Presentation 1 & 2
- 3/3 - Presentation 3 & 4
- 3/15 - Presentation 5 & 6
- 3/17 - Presentation 7 & 8
- 3/22 - Presentation 9 & 10
- 3/24 - Presentation 11 & 12
- 3/29 - Presentation 13 & 14
- 3/31 - Presentation 15 & 16
- 4/5 - Presentation 17 & 18
- 4/7 - Presentation 19 & 20

Paper Presentation

Select your paper for presentation
Due Jan 27 at 11:59pm

Presentation Slides
Due Apr 15 at 11:59pm | 10 pts

Oral Presentation (No submission, for grading purpose)
10 pts

Please submit your presentation slides here before or right after your presentation!

3/1 – Presentation 1 & 2

- **Presentation 1**
  - Paper: "Beyond COVID-19 Diagnosis: Prognosis with Hierarchical Graph Representation Learning"
  - Presenter: Kyle Beggs
- **Presentation 2**
  - Paper: "Big Self-Supervised Models Advance Medical Image Classification"
  - Presenter: Ilkin Isler

3/3 – Presentation 3 & 4

- **Presentation 3**
  - Paper: "Data augmentation using learned transformations for one-shot medical image segmentation"
  - Presenter: Joe Fioresi

- **Presentation 4**
  - Paper: "Group Shift Pointwise Convolution for Volumetric Medical Image Segmentation"
  - Presenter: Ryan Glaspey

Paper Presentation

- 20 minutes for presentation
- 15 minutes for Q&A and discussion
- Participate in discussion is important

How to present research papers (in class)? (1)

- Make good presentations
  - Know your audience: fellow graduate students with good background
  - Adapt the presentation goal: explain and discuss the paper
  - *Assume no one in the class has read the paper before*

How to present research papers? (2)

- Make good presentations
  - Title, authors (full name), authors' institutes, your name
    Motivation of the research (1—2 slides)
  - Problem statement
    - What is being solved?
    - Why is it an important problem?
  - Main contributions of the paper
    - Studied a new and important problem
    - Proposed a novel approach
    - Improved or extended existing methods
    - Compared several popular methods
    - Explored a variety of use cases (many datasets of different kinds)
    - Presented new theories
    - Presented a new dataset and benchmark results
    - Introduced new methodologies or tools to the field

How to present research papers? (3)

- Make good presentations
  - Title, authors (full name), authors' institutes, your name
  - Motivation of the research (1—2 slides)
  - Problem statement (1—2 slides)
    - It would be helpful to lay out some background about the problem
  - Main contributions of the paper
  - Approach outline (1 slide)
  - Details of the proposed approach
  - Experiments
    - Data, features, baselines, evaluation metrics, results
  - Related work (1—3 slides)
  - Conclusion: take-home message (1—2 slides)

How to present research papers? (4)

- Make good presentations
- Title, authors (full name), authors' institutes, your name and email
- Motivation of the research (1—2 slides)
- Problem statement (1—2 slides)
- Main contributions of the paper
- Approach outline (1 slide)
- Details of the proposed approach
- Experiments
- Related work (1—3 slides)
- Conclusion: take-home message (1—2 slides)
- Strengths & weaknesses of the paper (1—2 slides)
- Overall rating & why (how you weigh the strengths and weaknesses) (1 slide)
- Future directions (1—3 slides)

References and resources

1. Blalock, Davis, et al. "What is the state of neural network pruning?." *Proceedings of machine learning and systems* 2 (2020): 129-146.
2. Liang, Tailin, et al. "Pruning and quantization for deep neural network acceleration: A survey." *Neurocomputing* 461 (2021): 370-403.
3. Gholami, Amir, et al. "A survey of quantization methods for efficient neural network inference." arXiv preprint arXiv:2103.13630 (2021).
4. Gou, Jianping, et al. "Knowledge distillation: A survey." *International Journal of Computer Vision* 129.6 (2021): 1789-1819.
5. Wang, Lin, and Kuk-Jin Yoon. "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
6. https://github.com/lilujunai/Awesome-Knowledge-Distillation-for-CV

Thank you!

Question?