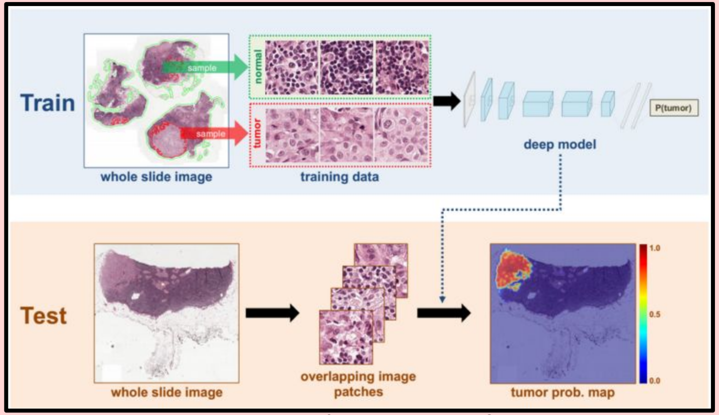
MINT: Deep Network Compression via Mutual Information-based Neuron Trimming

Madan Ravi Ganesh, Jason J. Corso, and Salimeh Yasaei Sekeh

Versatility of Deep Neural Networks

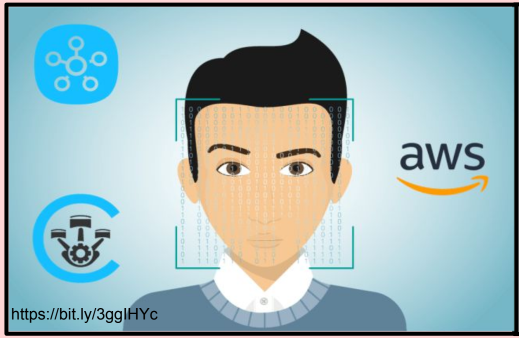
Breast Cancer Detection

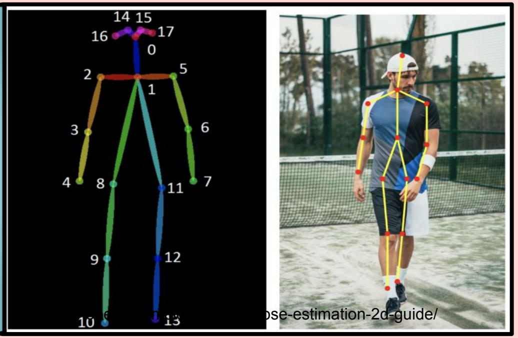


Tseng et al. Machine learning and imaging informatics in oncology. Oncology 2020.

Autonomous Driving

Face Recognition





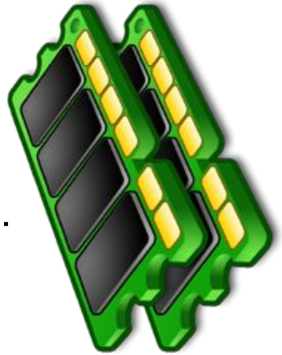
Pose Estimation

Hardware Constraints

Response Time

Memory Consumption

vs.



Performance

\+ Hardware Design Cost

\+ Costly Conversion Process

Solution: Deep Neural Network Compression

Compression Approaches

Low-rank Approximations:

- Jaderberg et al. *Speeding up Convolutional Neural Networks with Low Rank Expansions*. In BMVC 2014*.*

Quantization:

- Courbariaux et al. *Binaryconnect: Training deep neural networks with binary weights during propagations*. NeurIPS 2015.

Knowledge Distillation:

- Lu et al. March. *Knowledge distillation for small-footprint highway networks*. In ICASSP 2017.

Pruning:

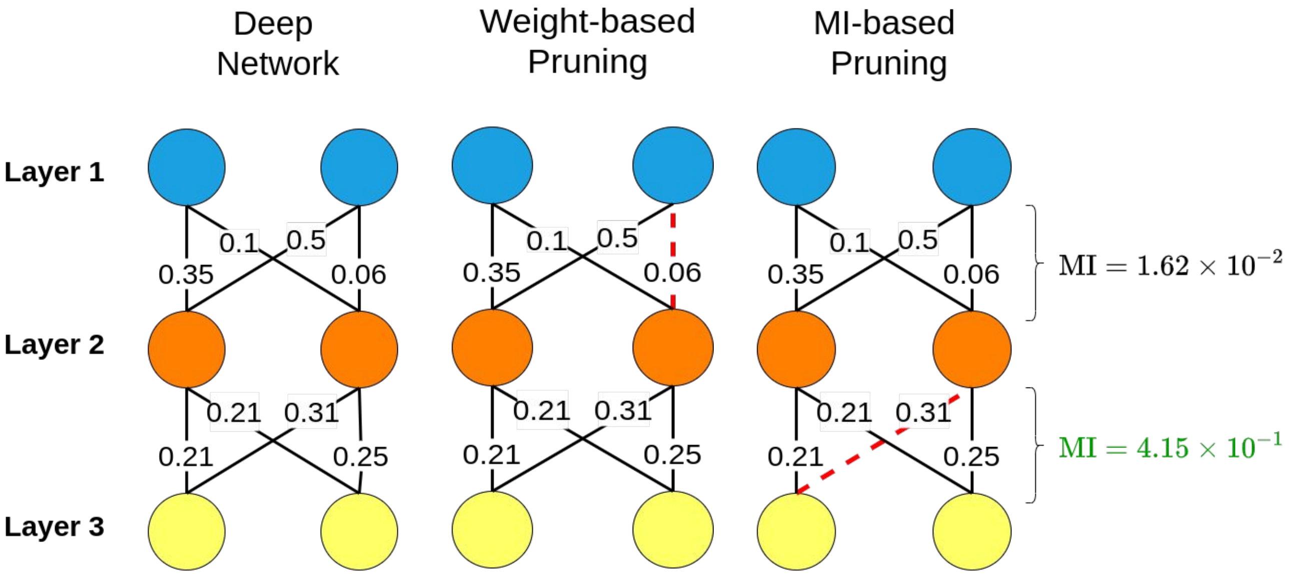
- Han et al. *Learning both weights and connections for efficient neural network*. In NeurIPS 2015.

Pruning Approaches

Unstructured	Structured	Hybrid
• Standard objective function • **Pruning Criteria**: Simple threshold, l_1 - norm, etc. • Minimal downstream impact consideration	• Standard objective function with sparsity • **Pruning Criteria**: Simple threshold • Insufficient analysis of learned features	• Standard objective function and/or sparsity • **Pruning Criteria**: Weight-based threshold • Inherits disadvantages of both approaches

Common Theme: **Simple**, **Deterministic** constraints on weights

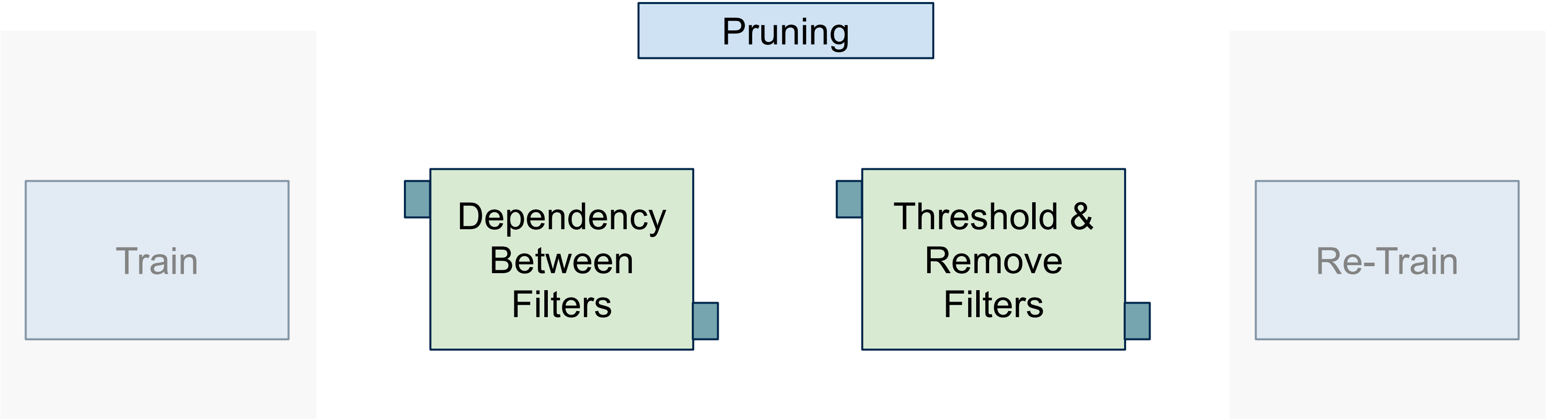
Alternative Hypothesis



Goal: *"We seek to develop a stochastic model of the dependency or flow of information between filters of a deep neural network"*

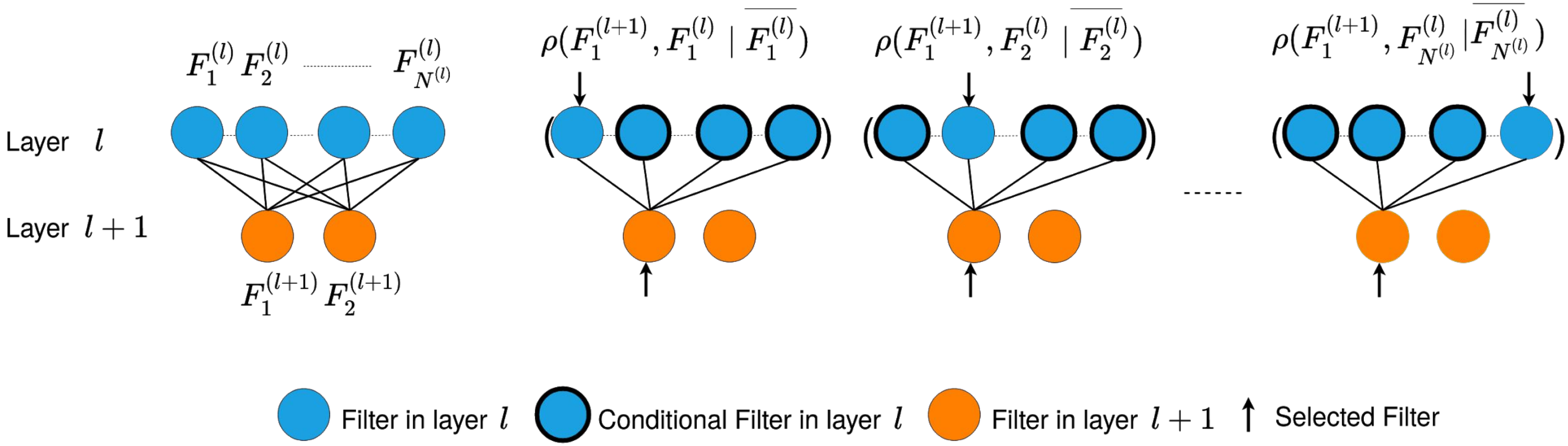
Our Choice: Mutual Information

MINT: Basic Building Blocks



- Develop mutual information as measure of dependency between filters
- Simple and extendable pruning approach

MINT: Dependency Between Filters



Conditional Geometric Mutual Information

$$\rho(F_i^{(l+1)}, F_j^{(l)}) = I(F_i^{(l+1)}, F_j^{(l)} \mid \overline{F_j^{(l)}});$$ [1]

[1] Yasaei Sekeh, S. and Hero, A.O., 2019. Geometric estimation of multivariate dependency. *Entropy*, *21*(8), p.787.

MINT: Threshold and Remove Filters

Compile

	$F_1^{(l)}$	....			$F_{N^{(l)}}^{(l)}$
$F_1^{(l+1)}$	0.88	0.01	0.05	0.50	0.11
$F_2^{(l+1)}$	0.64	0.03	0.01	0.00	0.13

Dependency scores arranged as weight matrix

Sort and Threshold

0.00	0.01		....	0.88

Threshold: n^{th} percentile

Remove Filters

	$F_1^{(l)}$	....			$F_{N^{(l)}}^{(l)}$
$F_1^{(l+1)}$	-1.2	0.0	0.0	10.0	1.33
$F_2^{(l+1)}$	7.9	0.0	0.0	0.0	0.51

Corresponding values in weight matrix removed

MINT: Benchmark Results

CIFAR10 - VGG16

Method	Pruned (%)	Test Acc. (%)
Base	N/A	93.98
Pruning Filters[1]	64.00	93.40
SSS[2]	73.80	93.02
GAL[3]	82.20	93.42
MINT	83.46	93.43

CIFAR10 - ResNet56

Method	Pruned (%)	Test Acc. (%)
Base	N/A	92.55
GAL[3]	11.80	93.38
Pruning Filters[1]	13.70	93.06
NISP[4]	42.40	93.01
OED[5]	43.50	93.29
MINT	52.41	93.47

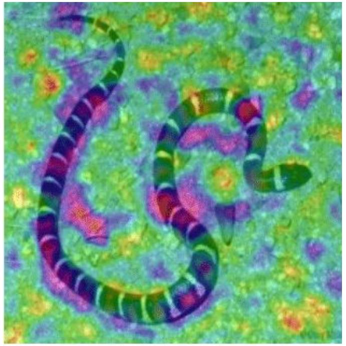
ILSVRC2012 - ResNet50

Method	Pruned (%)	Test Acc. (%)
Base	N/A	76.13
GAL[3]	16.86	71.95
OED[5]	25.68	73.55
SSS[2]	27.05	74.18
NISP[4]	43.82	71.99
ThiNet[6]	51.45	71.01
MINT	49.62	71.05

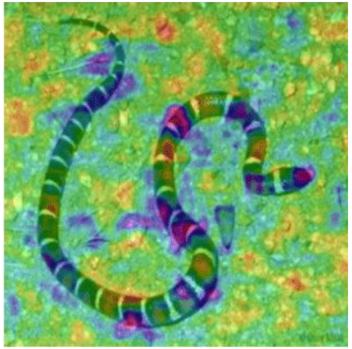
[1] Li et al. Pruning filters for efficient convnets. ICLR 2017
[2] Huang and Wang. Data-driven sparse structure selection for deep neural networks. ECCV 2018.
[3] Lin et al. Towards optimal structured cnn pruning via generative adversarial learning. CVPR 2019.
[4] Yu et al. Nisp: Pruning networks using neuron importance score propagation. CVPR 2018.
[5] Wang et al. Pruning blocks for cnn compression and acceleration via online ensemble distillation. IEEE Access 2019
[6] Luo et al. Thinet: A filter level pruning method for deep neural network compression. ICCV 2017.

Thank You

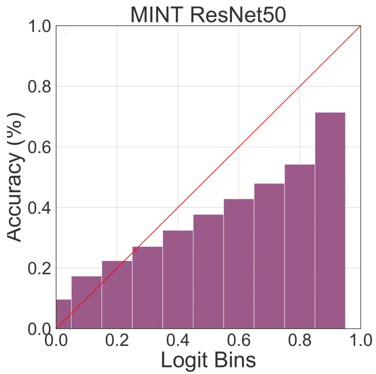
Feature Maps: Before



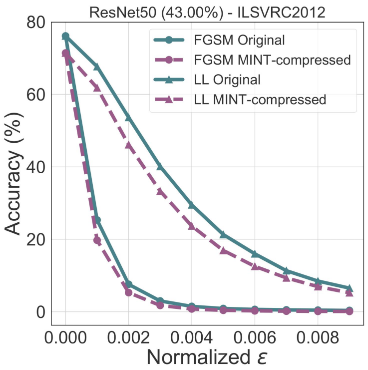
Feature Maps: After



Calibration



Adversarial Response



More detailed analyses available in the paper:
https://arxiv.org/pdf/2003.08472