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

Madan Ravi Ganesh[1], Jason J. Corso[1], and Salimeh Yasaei Sekeh[2]
[1]Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI
[2]School of Computing and Information Science, University of Maine, Orono, ME

Motivation

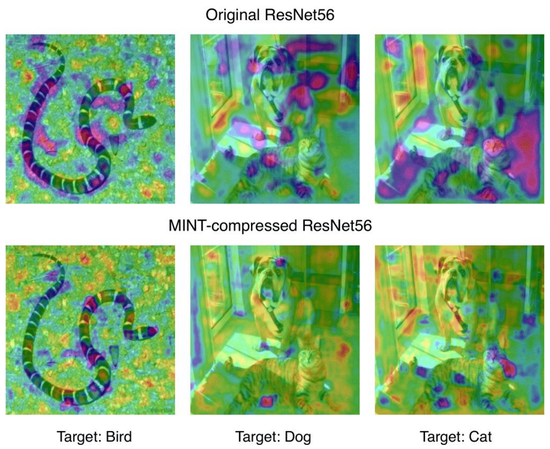
- Deep networks must satisfy low latency, low memory consumption and low error constraints when deployed to solve real-world problems
- Compression offers a quick solution to convert research-specific designs so they adhere to these constraints
- Common approaches to pruning:
  - Direct constraints on weights: Do not consistently account for downstream impact of pruning
  - Sparsity-inducing objective: Optimization of a more sensitive and difficult objective than cross-entropy

Our Core Philosophy

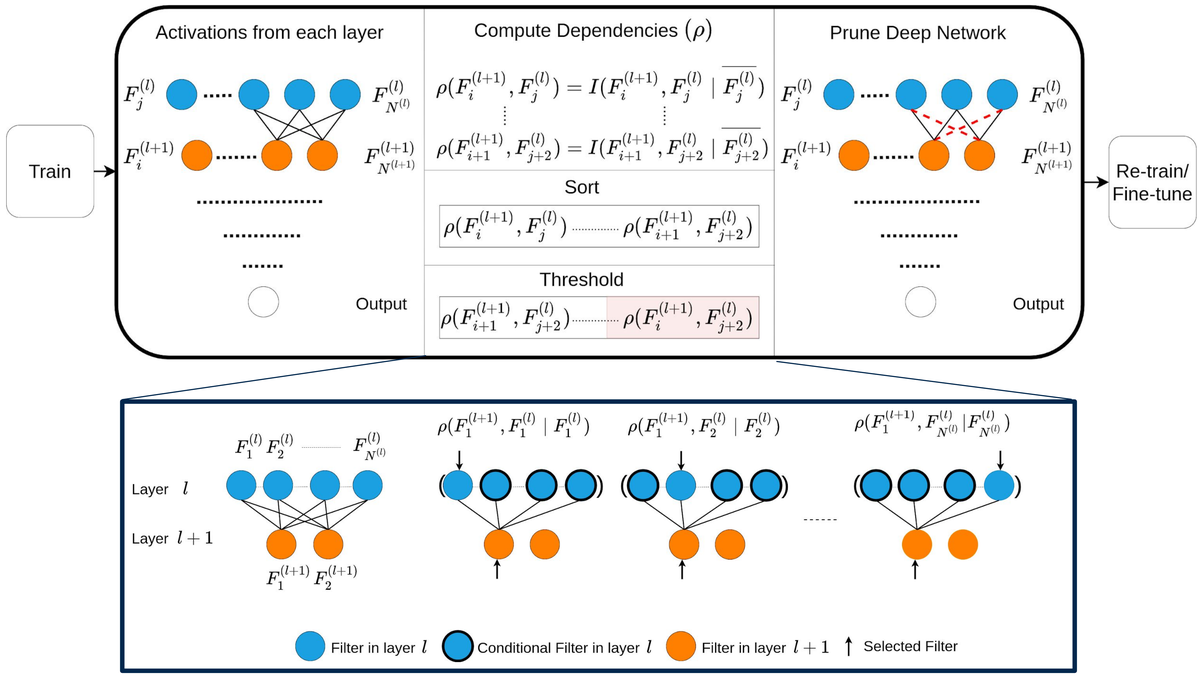
- "Development of a stochastic model of dependency or flow of information between filters of a deep network"
- Choice of stochastic modelling paradigm: **Mutual Information**
- Only retain filters that contribute the majority of the information

Qualitative Results: Features

Contribution from image to target class
- Reduction in the number of features
- Variation in type of features



Core Components of MINT



- MINT uses the conditional GMI[1] to compute ρ(). This measures the dependency between filters across adjacent layers of the network
- Retaining filters that contribute highly ensures we **maintain the flow of information** to downstream layers

Experimental Study

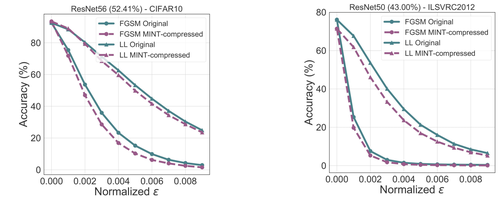
	Method	Pruned (%)	Test Accuracy (%)
VGG16 CIFAR10	Baseline	N/A	93.98
	GAL[2]	82.20	93.42
	MINT (ours)	83.46	93.43
ResNet56 CIFAR10	Baseline	N/A	92.55
	NISP[3]	42.20	93.01
	OED[4]	43.50	93.29
	MINT (ours)	57.01	93.02

	Method	Pruned (%)	Test Accuracy (%)
ResNet50 ILSVRC12	Baseline	N/A	76.13
	GAL[2]	16.86	71.95
	OED[4]	25.68	73.55
	SSS[5]	27.05	74.18
	NISP[3]	43.82	71.99
	ThiNet[2]	51.45	71.01
	MINT (ours)	49.62	71.05

- **Number of samples** used to compute GMI has a direct correlation with accuracy of mutual information estimates and Pruned (%)
- **Large grouping of filters** (low resolution) leads to weaker GMI estimates and therefore, high Pruned (%)
- **Highly competitive** performance even when compared to approaches with iterative or modified objective functions
- MINT allows us to reduce the overall memory consumed while matching the Test Accuracy (%) of the baseline, despite **low resolution** (filter groups) and a **single prune-retrain pass**

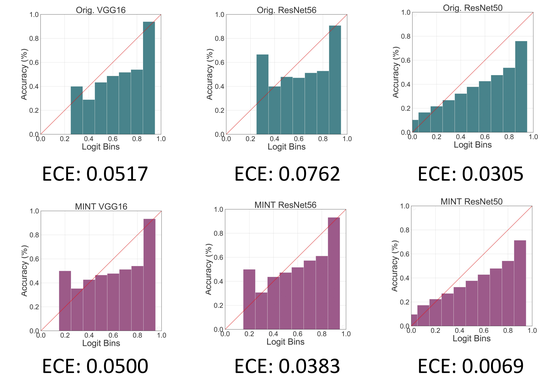
Qualitative Results: Adversarial Attacks

Over reliance on retained features increases susceptibility to **adversarial attacks**



Quantitative Results: Calibration

Compression acts like a regularizer to decrease **Expected Calibration Error** (ECE)



Orig. VGG16	Orig. ResNet56	Orig. ResNet50
ECE: 0.0517	ECE: 0.0762	ECE: 0.0305
MINT VGG16	**MINT ResNet56**	**MINT ResNet50**
ECE: 0.0500	ECE: 0.0383	ECE: 0.0069

Future Work

- Improving robustness to adversarial attacks
- Iterative extension to increase sparsity while maintaining high performance

References

[1] Yasaei Sekeh, S. and Hero, A.O. *Geometric estimation of multivariate dependency*. Entropy 2019.
[2] Lin et al. *Towards optimal structured cnn pruning via generative adversarial learning*. CVPR 2019.
[3] Yu et al. *Nisp: Pruning networks using neuron importance score propagation*.CVPR 2018.
[4] Wang et al. *Pruning Blocks for CNN Compression and Acceleration via Online Ensemble Distillation*. IEEE Access 2019.
[5] Huang, Z. and Wang, N. *Data-driven sparse structure selection for deep neural networks*. ECCV 2018.

Acknowledgements

This work has been partially supported (Madan Ravi Ganesh and Jason J. Corso) by NSF IIS 1522904 and NIST 60NANB17D191 and (Salimeh Yasaei Sekeh) by NSF 1920908.