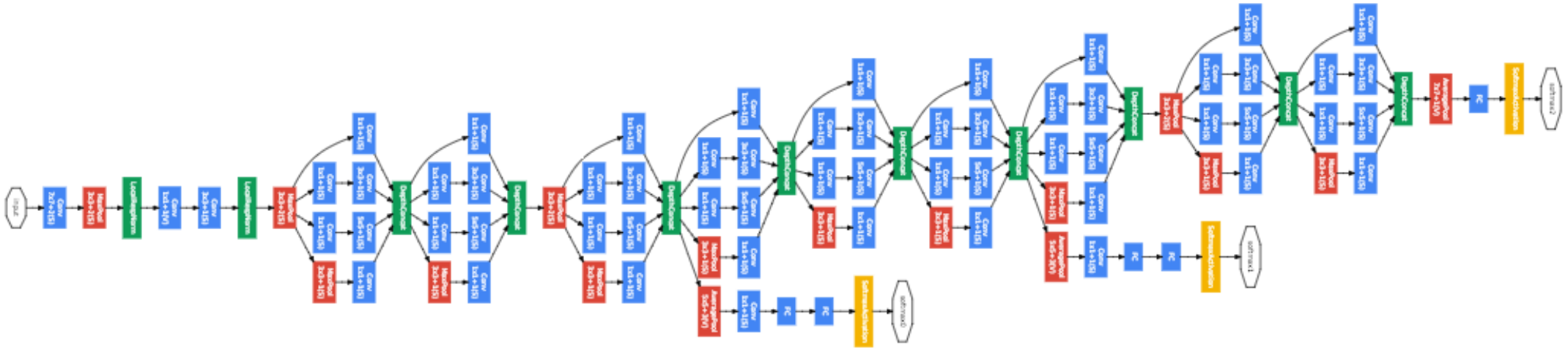
PRUNING CONVOLUTIONAL NEURAL NETWORKS FOR RESOURCE EFFICIENT INFERENCE

Beknazaro
v Nazar

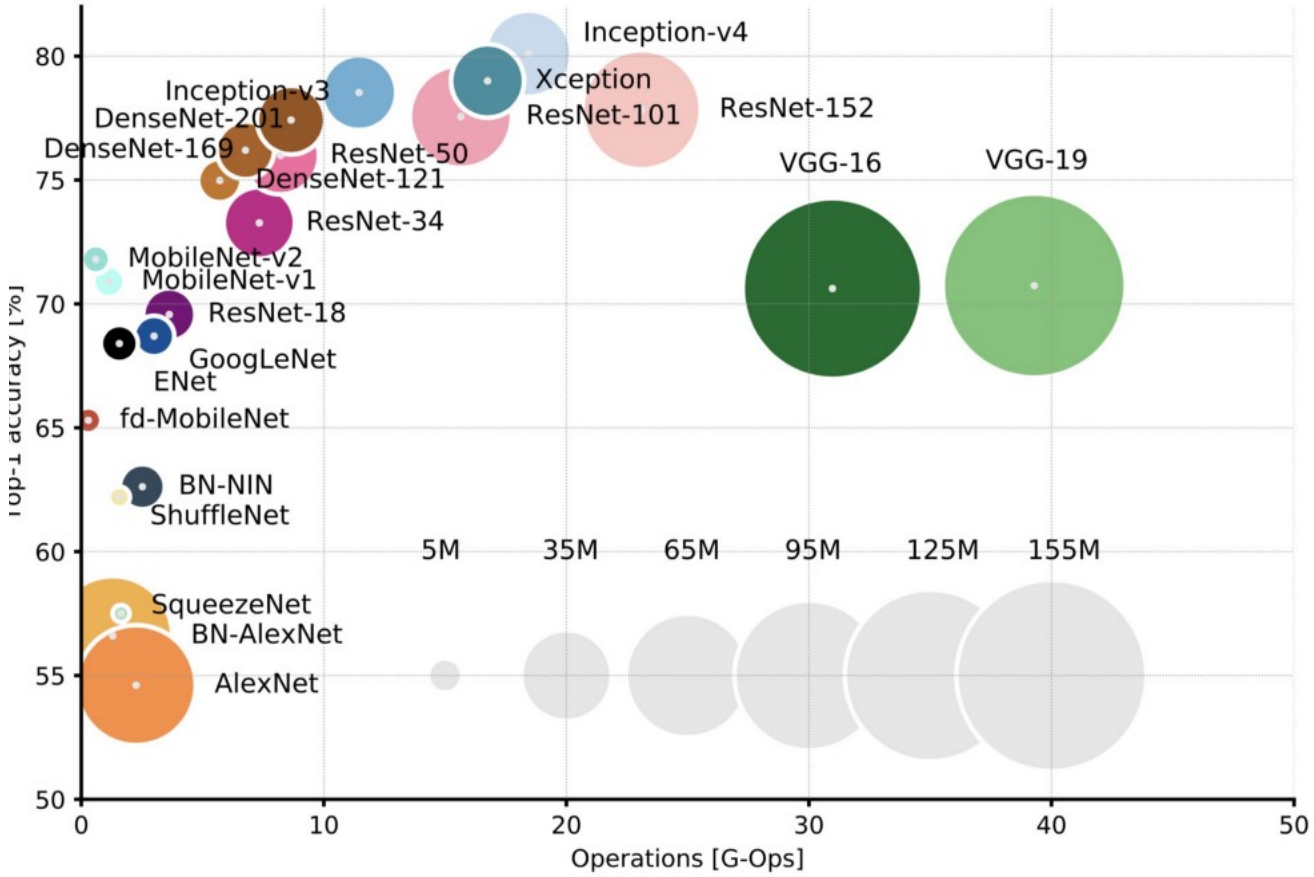
NATIONAL RESEARCH
UNIVERSITY

Moscow 2018

Problem of over-complexity



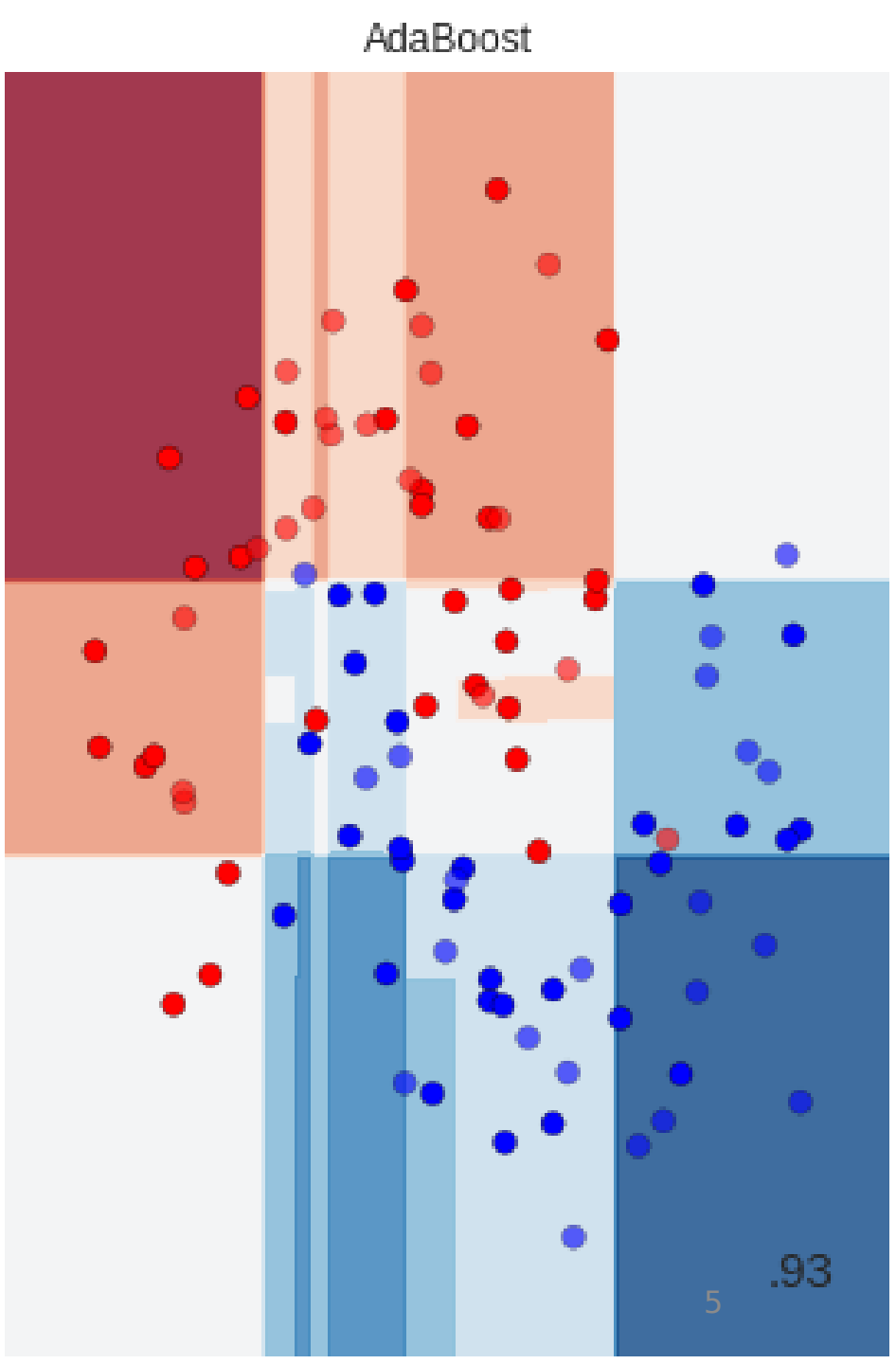
Overview

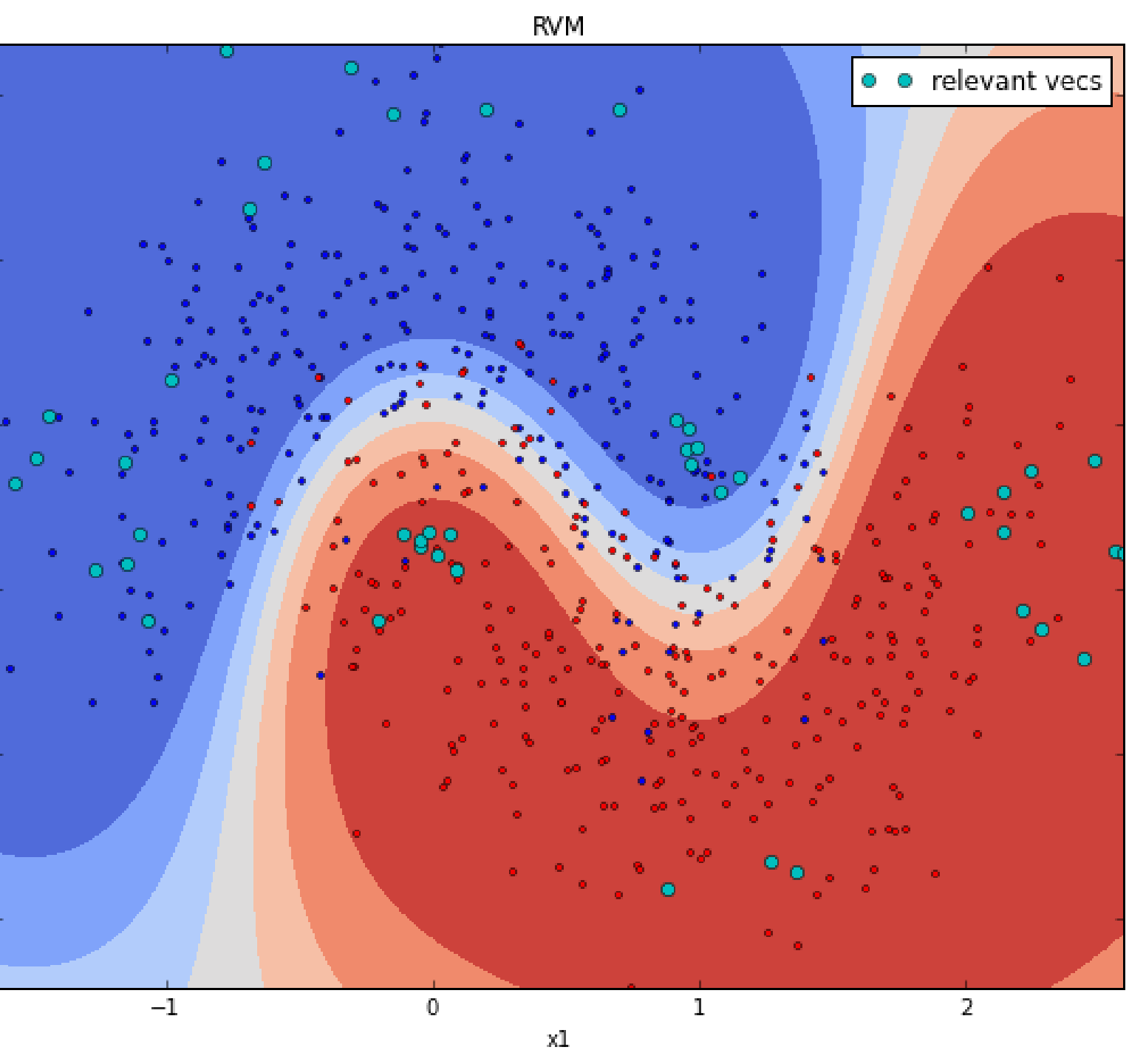

Inception-v4
Inception-v3
Xception
DenseNet-201
ResNet-101
ResNet-152
DenseNet-169
ResNet-50
VGG-16
VGG-19
DenseNet-121
ResNet-34
MobileNet-v2
MobileNet-v1
ResNet-18
GoogLeNet
ENet
fd-MobileNet
BN-NIN
ShuffleNet
5M
35M
65M
95M
125M
155M
SqueezeNet
BN-AlexNet
AlexNet
80
75
70
65
60
55
50
0
10
20
30
40
50
Operations [G-Ops]

Three ways to overcome the problem:

- Minimal description length
- VC – dimension
- Maximization of evidence

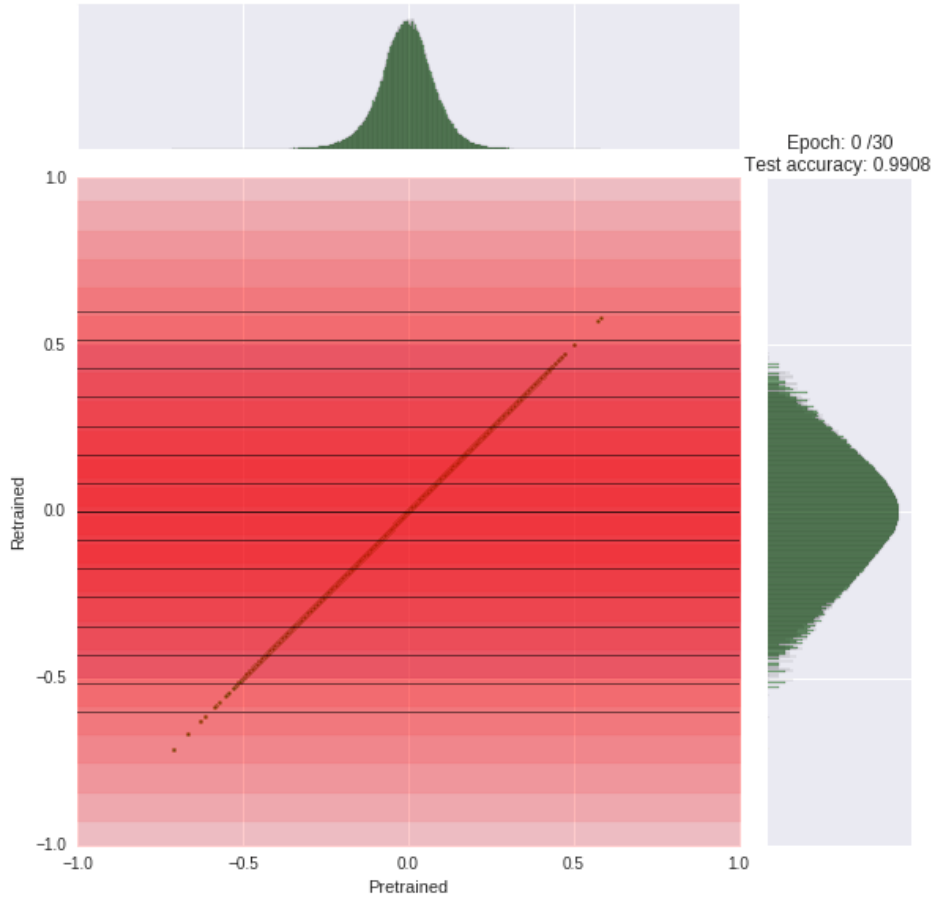
VC – dimension

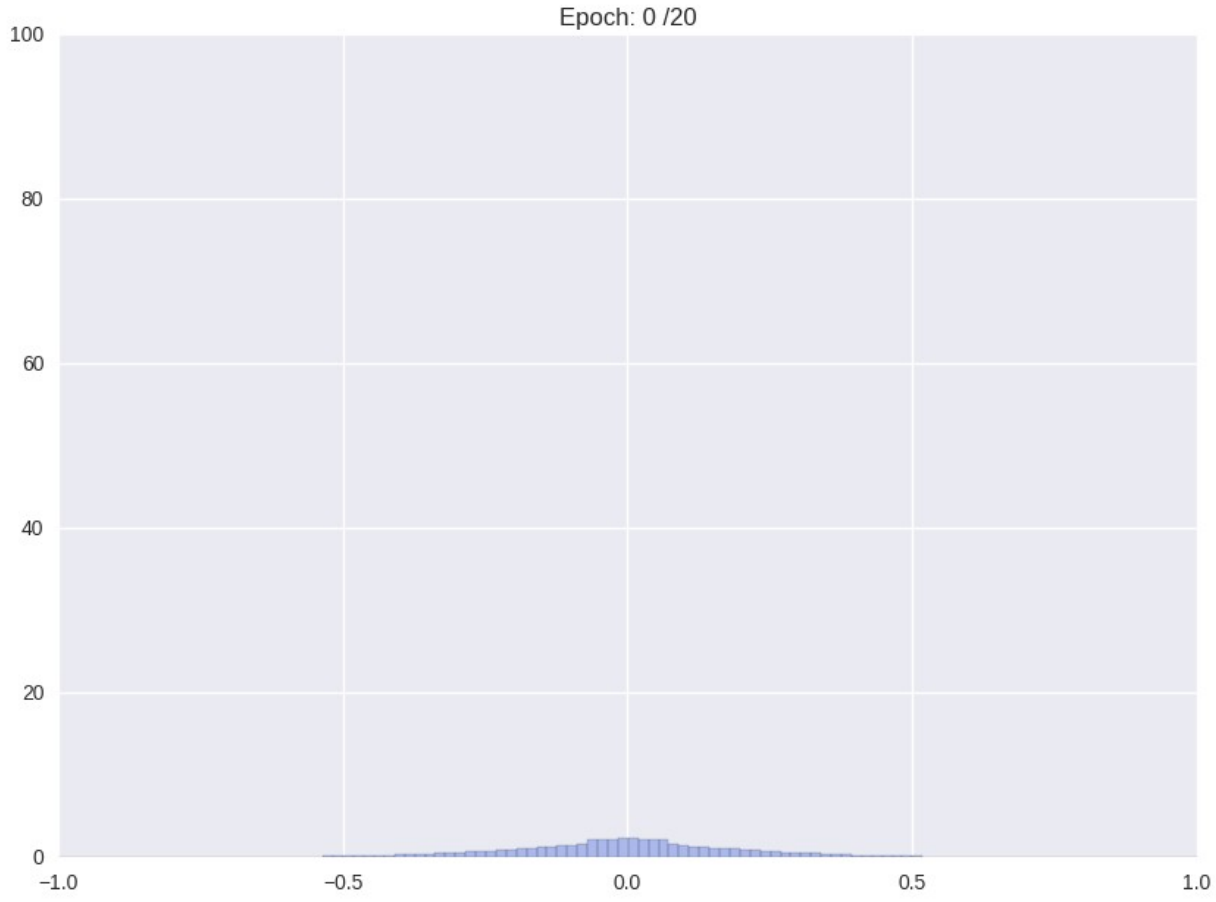




Maximization of evidence

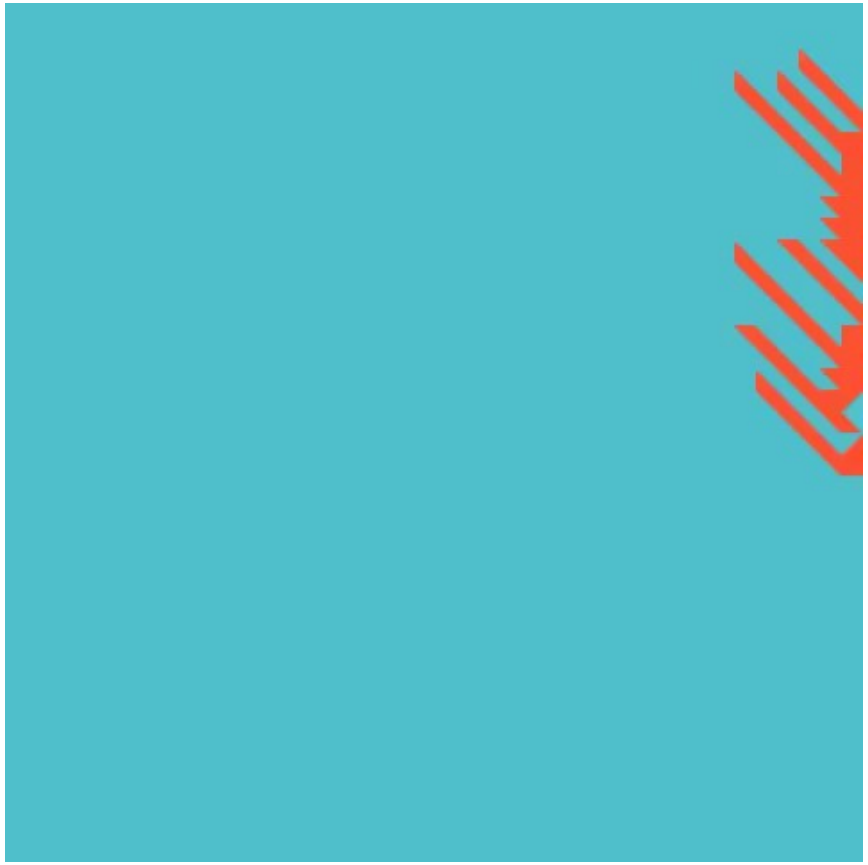
Minimal description length principle





Baseline pruning model

- Minimal weight
- Random
- L1 regularization



Way to measure quality of methods.

- Straight compute needs $2^{|W|}$, $|W| > 4000$ for VGG16
- No mutual information.
- Oracle pruning.
- The measure is correlation.

Taylor approximation

$$f(x) = \sum_{p=0}^{P} \frac{f^{(p)}(a)}{p!}(x-a)^p + R_p(x),$$

$$\mathcal{C}(\mathcal{D}, h_i = 0) \;\; = \;\; \mathcal{C}(\mathcal{D}, h_i) - \frac{\delta \mathcal{C}}{\delta h_i} h_i + R_1(h_i = 0).$$

Resulting formula

$$\Theta_{TE}(h_i) = \left|\Delta\mathcal{C}(h_i)\right| = \left|\mathcal{C}(\mathcal{D}, h_i) - \frac{\delta\mathcal{C}}{\delta h_i} h_i - \mathcal{C}(\mathcal{D}, h_i)\right| = \left|\frac{\delta\mathcal{C}}{\delta h_i} h_i\right|.$$

Results

	AlexNet / Flowers-102						VGG-16 / Birds-200						
	Weight	**Activation**			**OBD**	**Taylor**	**Weight**	**Activation**			**OBD**	**Taylor**	**Mutual Info.**
		Mean	*S.d.*	*APoZ*				*Mean*	*S.d.*	*APoZ*			
Per layer	0.17	0.65	0.67	0.54	0.64	**0.77**	0.27	0.56	0.57	0.35	0.59	**0.73**	0.28
All layers	0.28	0.51	0.53	0.41	0.68	0.37	0.34	0.35	0.30	0.43	0.65	0.14	0.35
(w/ ℓ_2-norm)	0.13	0.63	0.61	0.60	-	**0.75**	0.33	0.64	0.66	0.51	-	**0.73**	0.47
	AlexNet / Birds-200						**VGG-16 / Flowers-102**						
Per layer	0.36	0.57	0.65	0.42	0.54	**0.81**	0.19	0.51	0.47	0.36	0.21	**0.6**	
All layers	0.32	0.37	0.51	0.28	0.61	0.37	0.35	0.53	0.45	0.61	0.28	0.02	
(w/ ℓ_2-norm)	0.23	0.54	0.57	0.49	-	**0.78**	0.28	0.66	0.65	0.61	-	**0.7**	
	AlexNet / ImageNet												
Per layer	0.57	0.09	0.19	−0.06	**0.58**	**0.58**							
All layers	0.67	0.00	0.13	−0.08	**0.72**	0.11							
(w/ ℓ_2-norm)	0.44	0.10	0.19	0.19	-	0.55							

Table 1: Spearman's rank correlation of criteria vs. oracle for convolutional feature maps of VGG-16 and AlexNet fine-tuned on Birds-200 and Flowers-102 datasets, and AlexNet trained on ImageNet.

Results

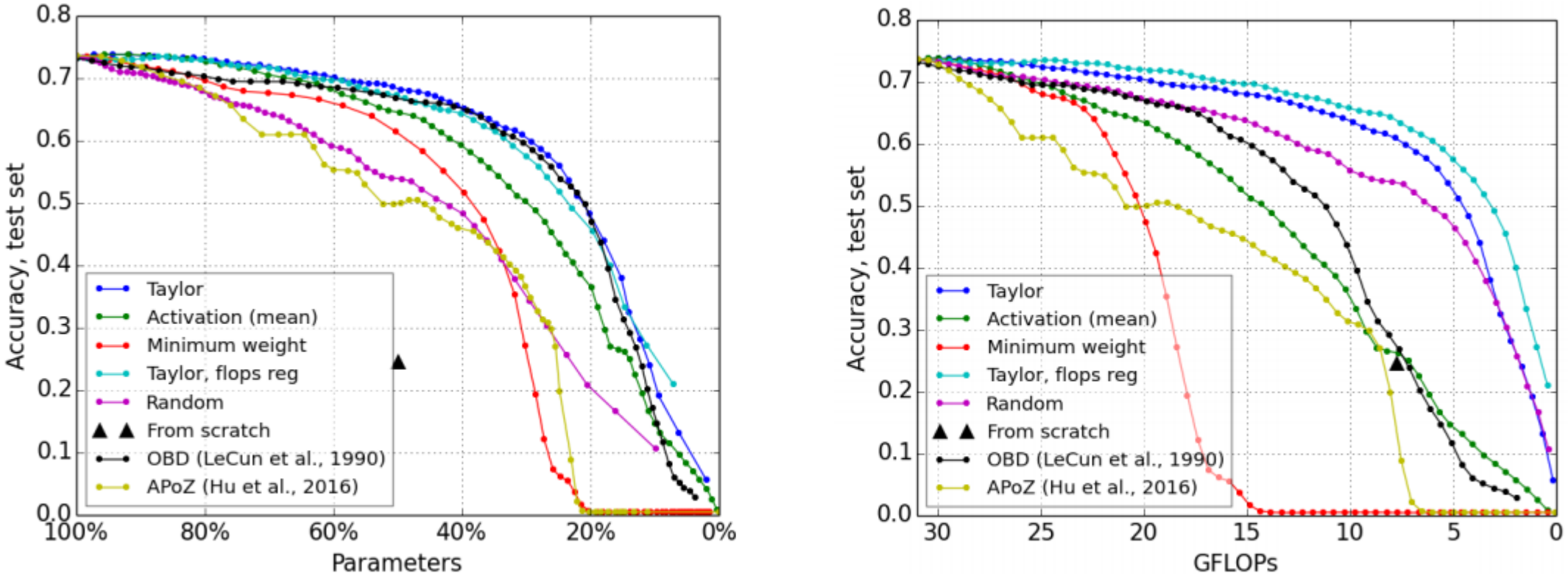
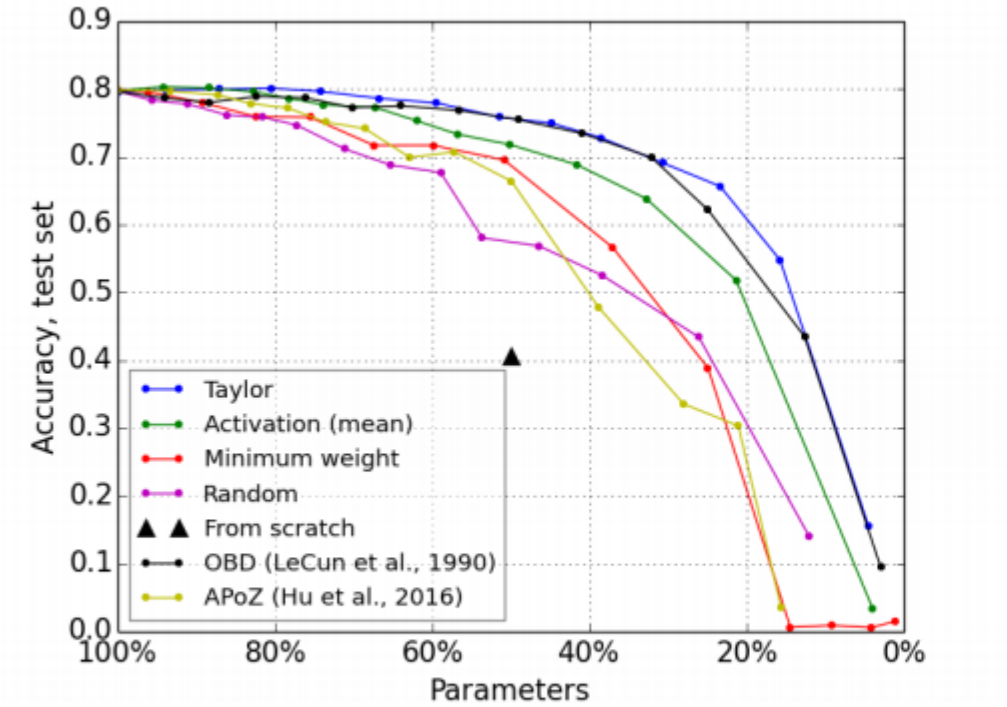


Figure 4: Pruning of feature maps in VGG-16 fine-tuned on the Birds-200 dataset.

Results



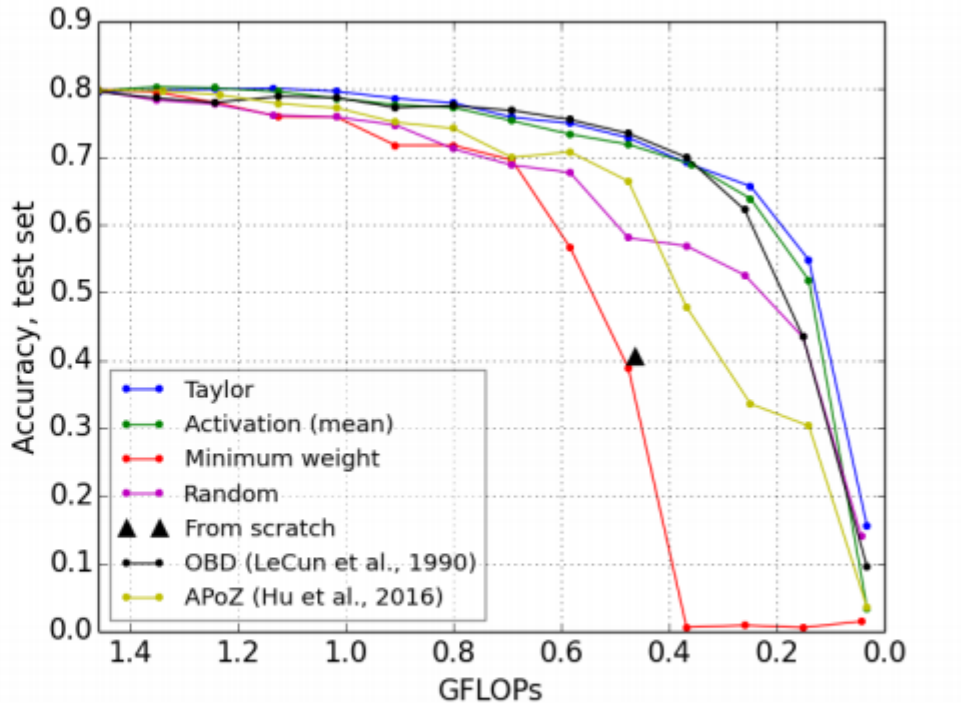


Figure 5: Pruning of feature maps in AlexNet on fine-tuned on Flowers-102.

References

- https://arxiv.org/pdf/1611.06440.pdf - Source article
- https://github.com/jacobgil/pytorch-pruning - Code
- https://arxiv.org/abs/1702.04008 - Soft weight sharing
- https://github.com/KarenUllrich/Tutorial-SoftWeightSharingForNNCompression - Code