Universality and Capacity Metrics in Deep Neural Networks

Charles H. Martin (charles@CalculationConsulting.com) and Michael W. Mahoney (mmahoney@stat.berkeley.edu)

Summary

We use our new theory of Implicit Heavy-Tailed Self-Regularization (HT-SR)[a] to develop a Universal capacity control metric, $\hat{\alpha}$, for DNNs.

- We analyze layer weight matrices $\mathbf{W}$ of over 100 pretrained DNNs, from both Computer Vision and NLP (VGG, ResNet, GPT, etc).
- We find that the spectral density $\rho(\lambda)$ of the normalized correlation matrix, $\mathbf{X} = \frac{1}{N}\mathbf{W}^T\mathbf{W}$, can be fit to a power law,

$$\rho(\lambda) := \lambda^{-\alpha}, \quad \lambda < \lambda^{max}$$

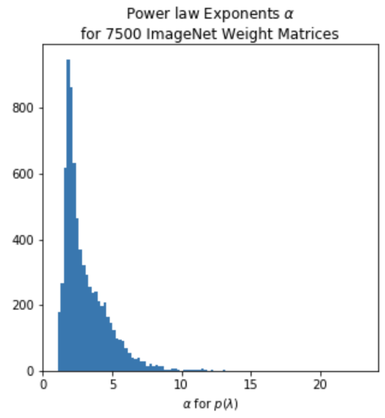
with exponent $\alpha \to 2$ universally.

- We propose a new Universal capacity metric, $\hat{\alpha} = \sum \alpha_l \log \lambda_l^{max}$, which correlates well with the generalization accuracy across a series of related DNN architectures.

[a] Long (arXiv:1810.01075) and short (ICML 2019) versions.

Universality of α

The power law exponents α for nearly 10,000 layer weight matrices $\mathbf{W}$, and convolutional feature maps, for over 100 CV DNN architectures, empirically approaches *Universal* value of $\alpha \to 2$.



Theory

Consider the familiar Product Norm Capacity Metric (for say the Spectral or Frobenius norm)

$$\mathcal{C} \sim \|\mathbf{W}_1\| \times \|\mathbf{W}_2\| \cdots \|\mathbf{W}_L\|. \tag{1}$$

Using a standard trick from field theory, we consider the log Product Norm

$$\log \mathcal{C} \sim \log \left[\|\mathbf{W}_1\| \times \|\mathbf{W}_2\| \cdots \|\mathbf{W}_L\| \right] \sim \left[\log \|\mathbf{W}_1\| + \log \|\mathbf{W}_2\| \cdots \log \|\mathbf{W}_L\| \right],$$

which takes the form of an average Log norm

$$\log \mathcal{C} \to \langle \log \|\mathbf{W}\| \rangle = \frac{1}{N_L} \sum_l \log \|\mathbf{W}_l\|.$$

Derive as a generalized weighted average, which resembles a (weighted) average log Spectral Norm

$$\hat{\alpha} = \sum \alpha_l \log \lambda_l^{max}$$

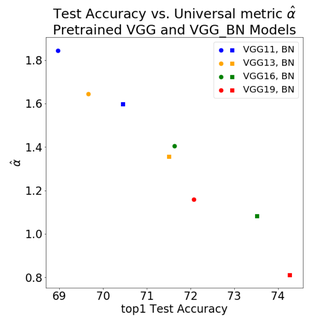
and/or which looks like the Soft Rank $\mathcal{R}_s^{log}$ in log units (from EVT, for small α)

$$\mathcal{R}_s^{log} := \frac{\log \|\mathbf{W}\|_F^2}{\log \lambda^{max}} \approx \alpha, \ \alpha \to 1.$$

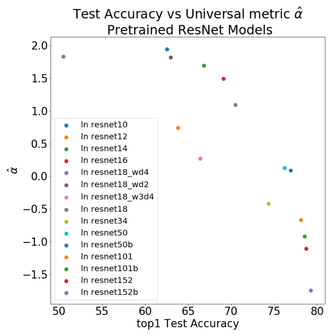
Capacity vs Test Accuracies

Norm metrics actually correlate with test accuracies across series of pretrained DNNs

VGG Series:



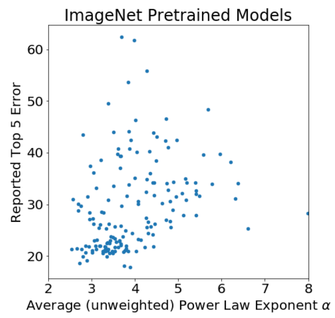
ResNet Series:



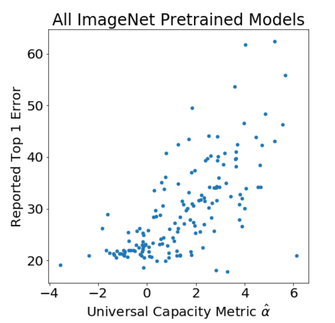
More Test Accuracies

Capacity versus Test Accuracy for over 100 pretrained ImageNet models

Average (unweighted) $avg(\alpha)$



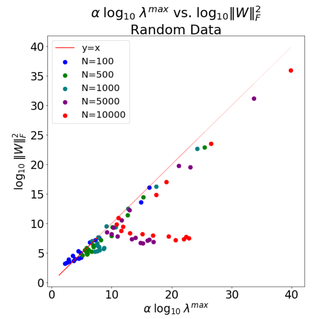
Universal Capacity Metric $\hat{\alpha}$



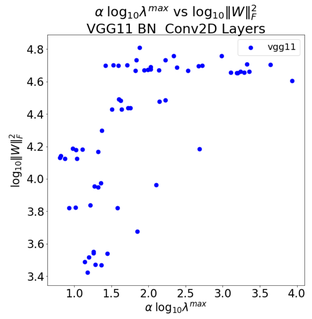
Power Law - Norm Relation

EVT provides the relation between the Frobenius norm and the Power law exponent $\alpha \sim 1$

Random Pareto Matrices



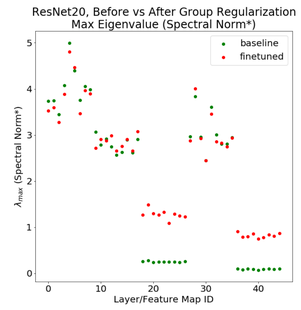
VGG11 Weight Matrices



Analysis of Distilled ResNet

Distillation sometimes induces anomalous jumps in the scale of the weight matrices

Spectral Norms (max eigenvalues λ^{max}**)**



Power law exponents α are not correlated w/ λ^{max}