Universality and Capacity Metrics in Deep Neural Networks

Anonymous Authors[1]

Abstract

We use our new Theory of Implicit and Heavy-Tailed Self-Regularization (HT-SR) to develop Universal capacity control metric, $\hat{\alpha}$, for Deep Neural Networks (DNNs). HT-SR indicates that modern DNNs exhibit what we call Heavy-Tailed Mechanistic Universality (HT-MU), meaning that the spectral density of layer weight matrices can be fit to a power law, $\rho(\lambda) \sim \lambda^{-\alpha}$, with exponents, $\alpha \in [2, 5]$, that lie in common Universality classes from Heavy-Tailed Random Matrix Theory (HT-RMT). Empirically, smaller α is correlated with better generalization accuracy, with $\alpha \to 2$ universally across different best-in-class, pretrained DNN architectures that generalize best. Using these facts, we define average case complexity metric, $\hat{\alpha} = \sum \alpha \log \lambda_{max}$, which resembles more familiar metrics, and that can be applied to pre-trained DNNs to predict trends in the test accuracy. We apply this new capacity metric to over 50 different, large-scale pre-trained DNNs, ranging over 15 different architectures, trained on ImagetNet, but with differing test accuracies. It correlates amazingly well with the reported test accuracies of these DNNs, looking across each architecture (VGG16/.../VGG19, ResNet10/.../ResNet152, etc.). Our approach requires no changes to the underlying DNN or its loss function, it does not require us to train a model, and it does not even require access to the ImageNet data.

1. Letter

Recent work by Martin and Mahoney (**?**) provides a Universal empirical metric that characterizes the amount of *Implicit Self-Regularization* and, accordingly, the generalization capacity, for a wide range of widely available, best-in-class, pre-trained Deep Neural Networks (DNNs), such as AlexNet, VGG, ResNet, etc.

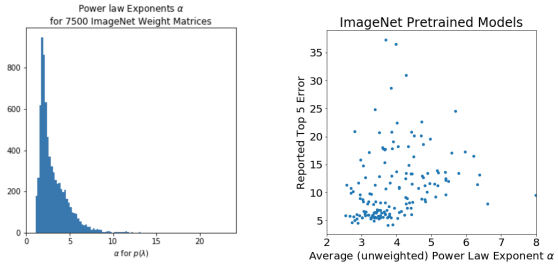
Instead of looking worst-case bounds, they study the Empirical Spectral Density (ESD), $\rho(\lambda)$, of individual layer weight matrices, $\mathbf{W}$ (and convolutional feature maps), through the lens of Random Matrix Theory (RMT). Looking in detail at a series of models, one observes that the individual layer ESDs almost always follow a power law (PL) distribution

$$\rho(\lambda) \sim \lambda^{-\alpha}, \tag{1}$$

where $\rho(\lambda)$ is the density of the eigenvalues λ of the normalized layer correlation matrix

$$\mathbf{X} = \frac{1}{N}\mathbf{W}^T\mathbf{W}. \tag{2}$$

The power law exponents nearly all lie within a universal range $\alpha \in [2, 5]$, in nearly every pre-trained architecture studied, e.g., across nearly $10,000$ layer weight matrices (and 2D feature maps), including DNNs pre-trained for computer vision tasks on ImageNet, and for several different NLP tasks.



(a) All power law exponents α (b) Average α vs Top5 error

Figure 1. Caption

Moreover, smaller α is correlated with better generalization accuracies, with α approaching a *universal value*, $\alpha \to 2$, at the lower limit of the Fat Tailed RMT Universality class (**?**) .

In Statistical Physics, Universality of PL exponents is very special, and it suggests the presence of a deeper, underlying,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Universal mechanism driving the system dynamics (**??**). It is this *Heavy Tailed Mechanistic Universality* (HT-MU), as well call it, that originally motivated our study. HT-MU applies to the analysis of complicated systems, including many physical systems, traditional NNs (**??**), and even models of the dynamics of actual spiking neurons. Indeed, the dynamics of learning in DNNs , and perhaps real neurons as well, seems to resemble a system near a phase transition, such as the phase boundary of spin glass, or a system displaying Self Organized Criticality (SOC), or a Jamming transition (**??**). Of course, we can not say which mechanism, if any, is at play. Instead, we use the machinery of HT-RMT as a stand-in for a generative model of the weight matrices in DNNs, to catalog and model the HT behavior of DNNs.

Based on these ideas, we develop a Universal capacity control metric $\hat{\alpha}$, a weighted average of the layer power law (PL) exponents. (α) of the DNN layer weight matrices:

$$\hat{\alpha} = \sum_{l \in L} \alpha_l \log \lambda_l^{max}. \tag{3}$$

where λ_l^{max} is the maximum eigenvalue (i.e. Spectral norm) of layer correlation matrices $mathbf{W}_l$:

Theory: Our approach and intent differ from other theoretical studies, although we can related our results back to known results. For example, Liao et al. (**?**) used an appropriately-scaled, data-dependent Product Norm capacity control metric to bound the worst-case generalization error for several small (non production-quality, but still interesting) DNN models, and they showed that the bounds are remarkably tight. There is, in fact, a large body of work on norm-based capacity control metrics, both recent, e.g., (**???**) and (**??????????**), as well as much older (**??**). These studies seek *worst-case* complexity bounds, motivated to reconcile discrepancies with more traditional statistical learning theory, and they apply it to quite small NNs. We seek an *average-case* or *typical case* (for realistic problems) complexity metric, viable in production to guide the development of better DNNs at scale.

Let us write the Energy Landscape (or optimization function) for a typical DNN with L layers, with activation functions $h_l(\cdot)$, and with weight matrices and biases $\mathbf{W}_l$ and $\mathbf{b}_l$, as follows:

$$E = h_L(\mathbf{W}_L \times h_{L-1}(\mathbf{W}_{L-1} \times (\cdots) + \mathbf{b}_{L-1}) + \mathbf{b}_L). \tag{4}$$

Typically, this model would be trained on some labeled data $\{d_i, y_i\} \in \mathcal{D}$, using Backprop(**?**), by minimizing the loss $\mathcal{L} = \sum_{i \in \mathcal{D}} [E(d_i) - y_i]$.

For simplicity, we do not indicate the structural details of the layers (e.g., Dense or not, Convolutions or not, Residual/Skip Connections, etc.) , nor do we consider the details of the optimizer or the training process.

Each layer is defined by one or more layer 2D weight matrices $\mathbf{W}_l$, and/or the 2D feature maps $\mathbf{W}_{l,i}$ extracted directly from 2D Convolutional (Conv2D) layers. (We have not yet analyzed LSTM or other complex Layers.) A typical modern DNN may have anywhere between 5 and 5000 2D layer matrices / feature maps.

We can relate our universality metric to the more traditional, data dependent, VC-like product norm capacity metrics $\mathcal{C}$. Define the *worst-case* bound $\mathcal{C}$ as

$$\mathcal{C} \sim \|\mathbf{W}_1\| \times \|\mathbf{W}_2\| \cdots \|\mathbf{W}_L\| \tag{5}$$

Using a standard trick from field theory, we consider the log product norm, which takes the form of an average log norm

$$\begin{aligned} \log \mathcal{C} &\sim \log \left[\|\mathbf{W}_1\| \times \|\mathbf{W}_2\| \cdots \|\mathbf{W}_L\| \right] \\ &\sim \left[\log \|\mathbf{W}_1\| + \log \|\mathbf{W}_2\| \cdots \log \|\mathbf{W}_L\| \right] \\ &\sim \langle \log \|\mathbf{W}\| \rangle = \frac{1}{N_L} \sum_l \log \|\mathbf{W}_l\| \end{aligned}$$

When $\|\mathbf{W}\|$ is the Spectral norm $\|\mathbf{W}\|_2 \sim \lambda_{max}$, then $\hat{\alpha}$ is a weighted average of the log product Spectral norm, where the weights are power law exponents α. In this sense, our universal metric $\hat{\alpha}$ behaves like an *average-case* version of what is a worst-case bound, and is more suitable for applying to large, production DNNs.

When $\|\mathbf{W}\|$ is the Frobenius norm $\|\mathbf{W}\|_F^2$, we can use results of Heavy Tailed RMT to interpret the PL exponents α as a type of Soft Rank. Specifically,, when α is very small, we can relate α to the more familiar Stable Rank $\mathcal{R}_s^{log}$, expressed in log-units (and up to the $\frac{1}{N}$ scaling):

$$\mathcal{R}_s^{log} := \frac{\log \|\mathbf{W}\|_F^2}{\log \lambda^{max}} \approx \alpha. \tag{6}$$

Using this, one could implement our capacity metric as a regularizer to improve DNN training by implementing a Stable Rank regularizer (similar to how Spectral norm regularization is implemented).

Methodology: To evaluate our metric, we introduce a new methodology to analyze the performance of large-scale pre-trained DNNs, including the VGG and ResNet series of models, as well nearly 200 other widely available models, and study how the capacity metrics correlate with the reported test accuracies.

This offers several advantages, most notably

- we do not need access to the ImageNet data, just the pre-trained models (i.e as distributed with PyTorch, on github, and/or from the ModelZoo), and

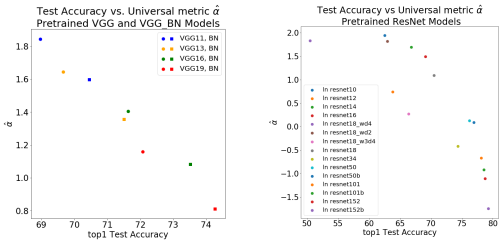
• our results are easily reproducible. To make things more reproducible, we provide a python command tool, weightwatcher (**?**), that works with both PyTorch (v1) and Keras (v2) models and computes a wide range of average log capacity metrics.

We apply our Universal capacity control metric $\hat{\alpha}$ to a wide range of large-scale pre-trained production-level DNNs, Notice, for Linear DNN layers, we can simply replace the log Norm with our metric, whereas for Conv2D Layers, we associate the "Norm" of the 4-index Tensor $\mathbf{W}_l$ to the sum of the $n_l = c \times d$ terms for each feature map.

$$\text{Linear Layer:} \quad \log \|\mathbf{W}_l\| \rightarrow \alpha_l \log \lambda_l^{max}$$

$$\text{Conv2D Layer:} \quad \log \|\mathbf{W}_l\| \rightarrow \sum_{i=1}^{n_l} \alpha_{l,i} \log \lambda_{l,i}^{max},$$

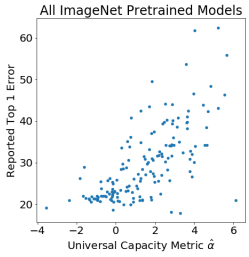
Results: Our Universal metric correlates very well with the reported average test accuracies across many series of pre-trained DNNs.



(a) $\hat{\alpha}$ for VGG series (b) $\hat{\alpha}$ for ResNet series

Figure 2. Top 1 Test Accuracy versus the Universal, weighted average PL exponent $\hat{\alpha}$ for pre-trained VGG and ResNet Architectures and DNNs.

DISCUSS FIGURE IN DETAIL, CAN ADD 1 more BELOW and discuss



(a) $\hat{\alpha}$ vs Top1 error

Figure 3. Caption: we have room for 1 more

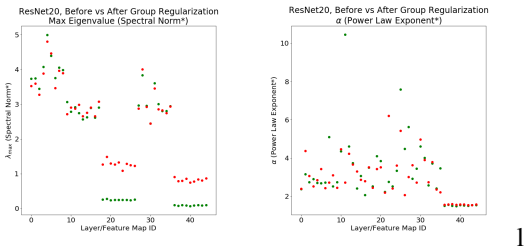
Our empirical results are, to our knowledge, the first time such theoretical capacity metrics have been reported to predict (trends in) the test accuracy for *pre-trained production-level* DNNs. In particular, this illustrates the usefulness of these norm-based metrics beyond smaller models such as MNIST, CIFAR10, and CIFAR100. Our results can be reproduced with the `WeightWatcher` package[1]; and our results suggest that our "practical theory" approach is fruitful more generally for engineering good algorithms for realistic large-scale DNNs.

Comparison with other metrics Our Universality metric $\hat{\alpha}$ seems related to other, more familiar capacity metrics such as the Spectral Norm, the Frobenius $(L2)$ Norm, and the measures of Soft Rank like the Stable Rank. This suggests an obvious question; does the $\hat{\alpha}$ "work" simply because the DNN models have explicit regularization (i.e. with an $(L2)$ or Spectral Norm constraint). Or, more generally, is $\hat{\alpha}$ just a variation of these more familiar *worst case* bounds.

The short answer is, we believe that Mechanistic Universality is a new, more fundamental relation. It complements other metrics, but works more generally on actual DNN models when other theoretical metrics and bound fail. In particular, we can identify counter examples, most notably in compressed DNN models, where the average Frobenius $(L2)$ Norm increases with decreasing test error, but the average α decreases, as expected.

For example, we consider average metrics measured on ResNet20, trained on CIFAR10, before and after applying the Group Regularization technique, as implemented in the *distiller* package ADDCITE. Notice that the reported baseline test accuracies ($Top1 = 91.450$ and $Top5 = 99.750$) are better than the reported finetuned test accuracies ($Top1 = 91.020$ and $Top5 = 99.670$), so traditional theory suggests that the baseline Spectral Norm ($\lambda_{max} \sim \|\mathbf{W}\|_2$) should be *smaller* than those of the layers in the finetuned model. Based on previous empirical results, we may also expect the baseline Frobenius norm to be smaller than the fine tuned. We observe the reverse for both.

[1] https://pypi.org/project/WeightWatcher/



(a) λ_{max} for ResNet20 layers (b) α for ResNet20 layers

Figure 4. ResNet20 with Group Regularization, comparison of layer $\mathbf{W}$ maximum eigenvalue (λ_{max}, or Spectral Norm *with $1/N$ normalization) and Power Law exponent α pretrained baseline and finetuned modles

DESCRIBE THESE 2 FIGURES

For the $4D\_regularized$ models, the *distiller* Group Regularization technique has the unusual effect of increasing the norms of the W feature maps for at least 2 of the Conv2D layers. We suspect this effect arises because the Group Regularization concentrates Frobenius mass from the 5 removed Conv2D layers into these remaining Conv2D layers.

Notice while the matrix norms behave atyically, the layers α do not systematically differ between the baseline and finetuned models, and, also (not shown), the average (unweighted) baseline α is indeed smaller than the finetuned average.

COMMMENT ON IMPLICATIONS

Discussion We have presented an *unsupervised* capacity control metric which predicts trends in test accuracies of a trained DNN—without peeking at the test data. This complexity metic, $\hat{\alpha}$ of Eqn. (**??**), is a weighted average of the PL exponents α for each layer weight matrix, where α is defined in the recent HT-SR Theory (**?**), and where the weights are the largest eigenvalue λ^{max} of the correlation matrix $\mathbf{X}$. We examine several commonly-available, pre-trained, production-quality DNNs by plotting $\hat{\alpha}$ versus the reported test accuracies. This covers classes of DNN architectures including the VGG models, ResNet, DenseNet, etc. In nearly every class, and except for a few counterexamples, smaller $\hat{\alpha}$ corresponds to better average test accuracies, thereby providing a strong predictor of model quality.

It is worth emphasizing that we are taking a very non-standard approach (at least for the DNN and ML communities). We did not train/retrain lots and lots of (typically rather small) models, analyzing training/test curves, trying to glean from them bits of insight that might then extrapolate to more realistic models. Instead, we take advantage of the fact that there already exist many (typically rather large) publicly-available pre-trained models, and we analyze the properties of these models. That is, we viewed these publicly-available pre-trained models as artifacts of the world that achieve state-of-the-art performance in computer vision, NLP, and related applications; and we attempted to understand why. and we then extracted data-dependent metrics to predict their generalization performance on production-quality models. Given well-known challenges associated with training, and given our results here as well as other recent results (**?**), we suggest that this methodology be applied more generally.

In theoretical physics, many researchers study neural networks using spin glass models, such as the traditional Gardner analysis, [SG's work], etc, Most notably, however, LeCun et. al. have suggested that the energy landscape of DNNs should resemble the zero-temperature energy landscape of a p-spin spherical spin glass. Specifically, they have argued that there are many local minima that concentrate at a floor just above the global minima. Here, however, we argue that such spin glass models should really employ heavy tailed, not Gaussian, stochastic spin-spin interactions, and that such models would have a very different zero-temperature complexity. Indeed, heavy tailed Levy spin-glasses do not have a large number of low lying minima, and, instead resemble something like a ruggedly convex 'energy funnel', with few local minima (**?**), similar in some sense to the Wolynes-Onuchic Energy landscape suggested in the early protein folding literature.