Predicting generalization with degrees of freedom in neural networks

Anonymous Authors[1]

Abstract

Model complexity is fundamentally tied to predictive power in the sciences as well as in applications. However, there is a divergence between naive measures of complexity such as parameter count and the generalization performance of overparameterized machine learning models. Prior empirical approaches to capturing intrinsic complexity independent of parameter count are computationally intractable, do not capture the implicitly regularizing effects of the entire machine-learning pipeline, or do not provide a quantitative fit to the double descent behavior of overparameterized models. In this work, we introduce an empirical complexity measure inspired by the classical notion of degrees of freedom in statistics. This measure can be approximated efficiently and is a function of the entire model training pipeline. We demonstrate that this measure strongly correlates with generalization performance in the double-descent regime.

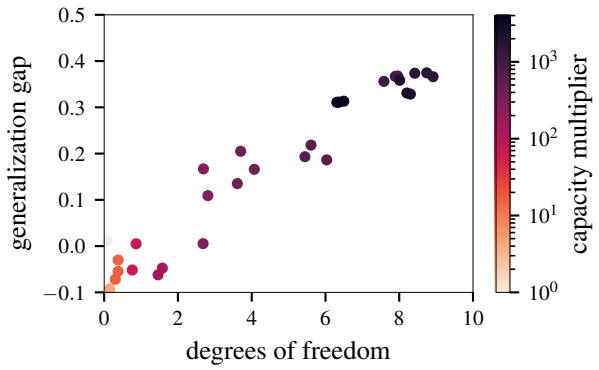

Figure 1. The generalization gap among neural network models that vary by capacity multiplier (here, kernel width) is monotonically increasing in generalized degrees of freedom (GDoF).

1. Introduction

What makes a machine learning model a good scientific explanation? Classical perspectives call for trading off low *complexity* and high *predictive accuracy* (Forster & Sober, 1994). However, deep learning approaches to scientific modeling (Baraniuk et al., 2020; Bianchini et al., 2020; Raghu & Schmidt, 2020; Hope et al., 2022) come with new challenges to navigating this trade-off due to the difficulty in assessing *both* complexity and predictive accuracy. First, many established measures of complexity do not apply to deep learning models because they do not capture the non-monotonic behavior of the evaluation error to which complexity is fundamentally tied. Secondly, overparameterized deep learning models often obtain zero training error (Zhang et al., 2017), meaning that model selection by predictive accuracy on the training set alone is not possible, while maintaining an independent and test set representative of downstream applications can be challenging in practice (Nagarajan et al., 2021). In addition, neural networks often exhibit surprising behaviours and pathoologies such as susceptibility to adversarial examples (Goodfellow et al., 2014) because both model complexity and generalization remain poorly understood, presenting safety issues that hinder real-life deployment of many machine learning techniques (Amodei et al., 2016; D'Amour et al., 2020).

To address these issues, here we adapt the generalized degrees of freedom (GDoF) (Efron, 1983; Ye, 1998) to measure the complexity of neural networks and predict their generalization performance. While (Gao & Jojic, 2016) first introduced this measurement from statistics to neural networks, our work makes the following contributions:

- We introduce approximations that enable scalable computation of the GDoF during training of deep neural networks.
- We reconcile the double-descent phenomenon with the classical trade-off between complexity and predictive accuracy by demonstrating monotonicity of the generalization gap in GDoF (see Fig. (1)).
- We reveal that GDoF evolves during training (see

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

Preliminary work. Under review by the 2nd Workshop on "AI for Science" (AI4Science) at the 39th International Conference on Machine Learning (ICML). Do not distribute.

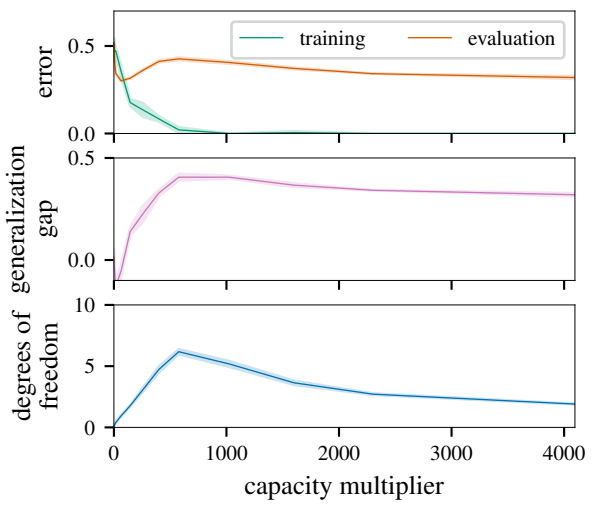

Figure 2. Training and evaluation error **(top)**, the generalization gap, **(middle)**, and GDoF **(bottom)** vs capacity multiplier (here, kernel parameter multiplier) for 3 CNNs trained on the CIFAR-10 dataset for 300 thousand iterations. The evaluation error exhibits the double-descent phenomenon, while the GDoF tracks the generalization gap.

Fig. (**3**) and Fig. (**5**)).

- We measure GDoF in a subspace optimization setting (Li & Liang, 2018), demonstrating that controlling model complexity improves generalization (see Fig. (**4**) and Fig. (**5**)).

2. Background

We contrast changes in the *parameter dimension* of a machine learning model—measures proportional to the ostensible parameter count of a model—with changes in the *effective dimensionality*, which is indirectly controlled by various components such as the architecture and training procedure. While trivially equivalent re-parametrizations—for instance, replacing a connection weight by the product of two weights $w = w_1\, w_2$)—may result in vastly different parameter dimensions, effective dimensionality captures the intrinsic complexity that is a result of the complete system specification. According to classic notions of complexity-fit trade-offs, models with low complexity and thus low effective dimensionality are less likely to overfit and thus have lower generalization error (Schmidhuber, 1997). The preference for simplicity in model selection is also known as *Occam's Razor* (MacKay, 1991).

However, the classical trade-offs between complexity and predictive power that focus on the bias and variance trade-off (Geman et al., 1992), has been challenged by the recently re-discovered double descent phenomenon (Belkin et al., 2019): While large models are traditionally believed to be inherently more complex and thus subject to poor generalization as a result of overfitting, modern overparameterized machine learning models exhibit suprisingly good generalization performance (Zhang et al., 2017).

While the origin of discrepancy between parameter dimension and effective dimensionality is outside of the scope of this work, previous works have shown that they may arise from the implicit regularization results from the whole pipeline of training including the model architecture (Golubeva et al., 2020), the specific initialization procedure (Kubo et al., 2019; Mehta et al., 2021), gradient-based optimization (Barrett & Dherin, 2021), and overparametrization itself (Neyshabur et al., 2018; Advani et al., 2020). In the final model, the effective dimensionality is known to depend on the loss landscape (Larsen et al., 2021), particularly the local curvature (Maddox et al., 2020).

Various attempts have been made to capture model complexity, and one direction of effort focuses on the statistical concept of degrees of freedom (Section 5 discuss a wider range of works). For linear models, the degree of freedom is equivalent to parameter count. (Efron, 1983) and (Ye, 1998) have generalized this concept to non-linear models from the perspectives of *expected optimism* or *average sensitivity*. For a classifier f parametrized by θ, the GDoF is defined as the derivatives of the *model predicted label* with respect to *data label*:

$$\mathcal{F} = \sum_{n=1}^{N} \sum_{c=1}^{C} \frac{\partial f_{n,c}(\theta_D)}{\partial y_{n,c}} \,, \tag{1}$$

where N is the total number of samples in the training set, C is the number of classes. The subscript D denotes the dependency of f's parameter θ on the whole dataset $D = \{x_n, y_n\}_{n=1}^{N}$ through training. This quantity can be interpreted as the sensitivity of f given perturbation in training label y. (Gao & Jojic, 2016) introduced GDoF to neural networks and showed that the GDoF of such over-parametrized models are only a small fraction of the total number of parameters. In their work, the derivatives in eq. 1 is approximated by random perturbation, but computing $\mathcal{F}$ requires training a perturbed model, as well as the original model, until convergence.

3. Method

In this work, we introduce additional approximation of $\mathcal{F}$ in Eq. (1) for a fully scalable estimate of model complexity. One straightforward step of approximation is using the Monte-Carlo estimator based on a subset of $B < N$ samples $\mathcal{F}^{\text{MC}} \approx \frac{N}{B} \sum_{b=1}^{B} \sum_{c=1}^{C} \frac{\partial f_{b,c}(\theta_D)}{\partial y_{b,c}}$. We further remove the requirement of training the model to convergence by exploiting the sequential update of θ during gradient-based training.

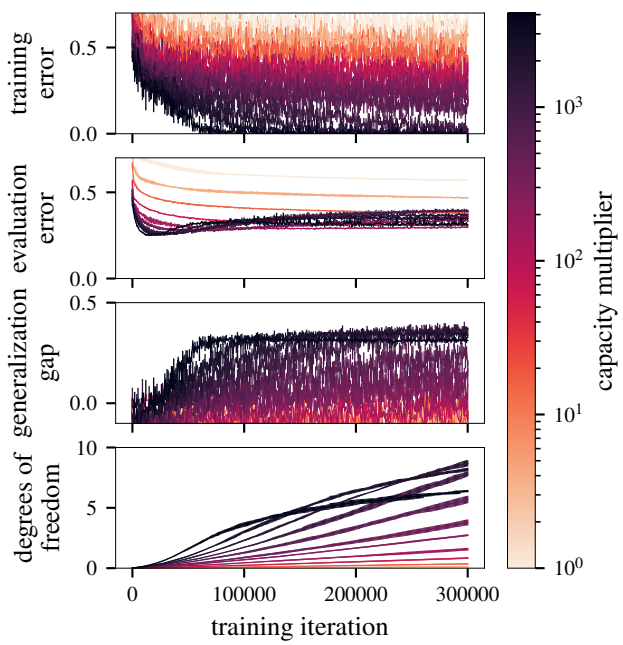

Figure 3. Training and evaluation error **(top two)**, the generalization gap, **(middle)**, and GDoF **(bottom)** vs. training iteration for different values of the capacity multiplier (here, kernel parameter multiplier) for 3 CNNs trained on the CIFAR-10 dataset for 300 thousand iterations. The GDoF captures the evolution of the generealization gap over training for the various capacity settings.

To derive the sequential approximation, we focus on the individual derivatives $\frac{\partial f_{i,c}(\theta_D)}{\partial y_{i,c}}$ and explicit write θ_t for parameter at training step t, and the initial parameter θ_0, such that the model at time t can be written as $f(\theta_t;\theta_0) = f(\theta_t;\theta'_t)\,\forall t' < t$. We further use the shorthand:

$$d_{b,c}(0,t) = \frac{\partial f_{b,c}(\theta_t;\theta_0)}{\partial y_{b,c}} \tag{2}$$

Therefore, the Monte-Carlo estimate of the model's GDoF is $\mathcal{F}_t = \frac{N}{B}\sum_{i=b}^{B}\sum_{c=1}^{C} d_{b,c}(0,t)$. For brevity, we drop the subscriptions b, c when they are irrelevant in the derivation.

The following proposition shows that, given the DoF of a model at step t, the DoF at step t' can be approximated by simply adding the additional sensitivity brought by the increment $\Delta\theta$, as long as $\|\theta\|^2$ is small.

Proposition 1. *The DoF at step t' is related to the DoF at $t < t'$ via:*

$$\mathcal{F}_{t'} \approx \mathcal{F}_t + \mathcal{F}_\Delta \tag{3}$$

where $\mathcal{F}_\Delta = \frac{N}{B}\sum_{b=1}^{B}\sum_{c=1}^{C} d_{b,c}(t,t')$, measuring the incremented sensitivity of f given the update from θ_t to $\theta_{t'} = \theta_t + \Delta\theta$.

Proof. We can Taylor-expand $f(\theta_{t'};\theta_0)$ at $\theta_{t'} = \theta_t$:

$$\partial f(\theta_{t'};\theta_0) = f(\theta_t;\theta_0) + \Delta\theta^T \frac{\partial f(\theta;\theta_0)}{\partial\theta}\bigg|_{\theta=\theta_t} + \mathcal{O}(\|\Delta\theta\|^2) \tag{4}$$

Therefore,

$$\begin{aligned} d(0,t') =& \frac{\partial f(\theta_t;\theta_0)}{\partial y} + \left(\frac{\partial\Delta\theta}{\partial y}\right)^T \frac{\partial f(\theta;\theta_0)}{\partial\theta}\bigg|_{\theta=\theta_t} \\ &+ \Delta\theta^T \frac{\partial^2 f(\theta;\theta_0)}{\partial\theta\partial y}\bigg|_{\theta=\theta_t} + \mathcal{O}(\|\Delta\theta\|^2) \\ \approx& \, d(0,t) + \left(\frac{\partial\Delta\theta}{\partial y}\right)^T \frac{\partial f(\theta;\theta_0)}{\partial\theta}\bigg|_{\theta=\theta_t} \end{aligned} \tag{5}$$

Here the approximation comes form dropping the second order terms, so the quality of approximation is high when $\Delta\theta^T \frac{\partial^2 f(\theta;\theta_0)}{\partial\theta\partial y}\Big|_{\theta=\theta_t}$ and $\|\Delta\theta\|^2$ are small, which is likely when the update $\Delta\theta$ is small. Since $d(t,t')$ depends on y only through the update $\Delta\theta$:

$$\begin{aligned} d(t,t') &= \frac{\partial f(\theta_{t'};\theta_t)}{\partial y} = \left(\frac{\partial\Delta\theta}{\partial y}\right)^T \frac{\partial f(\theta_{t'};\theta_t)}{\partial\Delta\theta} \\ &= \left(\frac{\partial\Delta\theta}{\partial y}\right)^T \frac{\partial f(\theta;\theta_t)}{\partial\theta}\bigg|_{\theta=\theta_t} \\ &= \left(\frac{\partial\Delta\theta}{\partial y}\right)^T \frac{\partial f(\theta;\theta_0)}{\partial\theta}\bigg|_{\theta=\theta_t} \end{aligned} \tag{6}$$

Plug this into Eq. 5, we have

$$d(0,t') \approx d(0,t) + d(t,t') \tag{7}$$

We can now obtain Proposition 1 from the definition of $\mathcal{F}_t$. □

Although the approximation in eq. 5 requires $\Delta\theta$ to be small, in our experiments we found the resulting estimator of GDoF highly predictive of generalization bound even when $\Delta\theta$ represents changes across 1000 update steps away. In addition we found performing multiple update steps K (using different mini-batches) improves the estimate of GDoF when t' and t are far apart, as in our experiments. We leave detailed analysis of this approximation, such as the error accumulated through the trajectory for future work. The next section empirically demonstrates the predictive power of this quantity for neural networks. Algorithm 1 summarise the algorithm for computing the GDoF from a series checkpoints saved during training. Notice that taking $T = 0$, $B = N$ and $K = \infty$ (until convergence) recovers the algorithm in (Gao & Jojic, 2016)[1].

[1]Omitting the average over training trajectory, which is always 1 in their experiments.

Algorithm 1 An algorithm with caption

Require: saved checkpoints $\mathcal{C}_1, \mathcal{C}_2, \ldots \mathcal{C}_T$, the number of update steps K, mini-batch size B, perturbation step ϵ
$\mathcal{F} \leftarrow 0$
for t = 1 to T **do**
  Load model $f(\theta_t)$ from checkpoint $\mathcal{C}_t$
  for k = 1 to K **do**
    Sample a mini-batch of data $\{x_b, y_b\}_{b=1}^{B}$
    Sample Gaussian noise $z_b \sim \mathcal{N}(0,1)$ for $b = 1, 2, \ldots, B$
    Update $f(\theta_{t+k})$ using $\{x_b, y_b\}_{b=1}^{B}$
    Update $f'(\theta_{t+k})$ using $\{x_b, y_b + \epsilon \cdot z_b\}_{b=1}^{B}$
    Compute
    $\mathcal{F} \leftarrow \mathcal{F} + \frac{N}{B} \sum_{b=1}^{B} \sum_{c=1}^{C} z_b(c) \frac{f'_{b,c}(\theta_{t+k}) - f_{b,c}(\theta_{t+k})}{\epsilon}$
  end for
end for

4. Experiments

We demonstrate results measuring the GDoF of a CNN trained on the CIFAR-10 dataset. In Fig. (2) we plot the evaluation error alongside the training error. We also plot the difference between the two—the generalization gap—and demonstrate that the GDoF method of Section (3) closely tracks the generalization gap, and thus is predictive of generalization performance, as would be expected of an good approximation to model complexity.

5. Related Work

Whence generalization? Generalization properties of deep neural networks has received much renewed interest from both theoretical and empirical perspectives (Jiang et al., 2020a). However, it seems that many properties that correlate with generalization performance are neither necessary nor sufficient to ensure it; for example, sharpness of the loss surface around a minimum (Dinh et al., 2017), or monotonic linear interpolation (Lucas et al., 2021).

Capturing complexity. Previous approaches to approximate effective dimensionality have the following drawbacks: they do not resolve double descent; they do not capture all elements of a machine learning system (often because they are a local estimate (Maddox et al., 2020)); and they are computationally expensive or intractableProposals to measure effective dimensionality have taken many forms, from more classical parameter and representation norms (Bartlett et al., 2017) and sharpness measures (Keskar et al., 2017), to measures motivated specifically by modern deep learning systems, such as measures of gradient noise (Smith & Le, 2018) and optimization speed (Hardt et al., 2016); see Jiang et al. (2020b) for a thorough overview and an empirical comparison. The common idea behind all such approaches is to

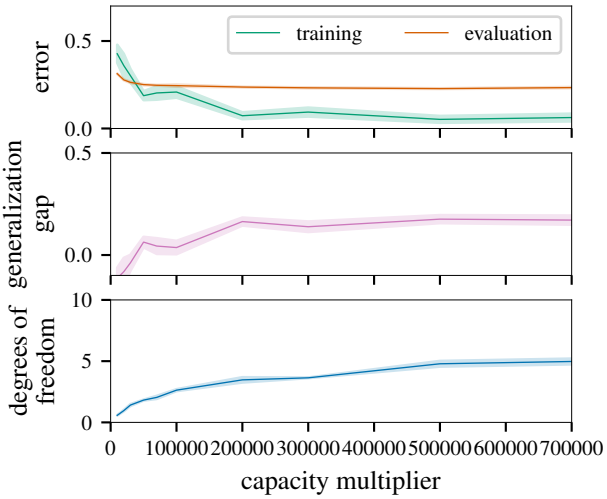

Figure 4. Training and evaluation error **(top)**, the generalization gap, **(middle)**, and GDoF **(bottom)** vs capacity multiplier (here, subspace dimension) for 3 CNNs trained with subspace optimization on the CIFAR-10 dataset for 300 thousand iterations. In contrast to Fig. (2), the evaluation error is monotonically decreasing. The GDoF still tracks the generalization gap.

quantify variability and thus complexity in the function represented by a neural network. This functional complexity may result from the interaction of the the particular parameterization or architecture in addition to all other aspects of the machine learning pipeline, including the optimization algorithm and other hyper-parameters. Crucially, however, the exact relationship between neural network design decisions and functional complexity is unknown, prompting the study of *implicit regularization* in deep learning (Neyshabur et al., 2015; Neyshabur, 2017; Dherin et al., 2021).

(Jiang et al., 2021) discuss the results of a recent community competition (Jiang et al., 2020a) that targeted complexity measures that reliably capture generalization performance; however, they do not investigate the trend of the these complexity measures as the parameter dimension is manipulated. (Maddox et al., 2020) resolve the double descent paradox with the introduction of a capacity estimate inspired by posterior contraction in Bayesian learning that renders evaluation performance monotonically increasing. However,

Linear training dynamics. The neural tangent kernel (NTK) (Jacot et al., 2018; Li & Liang, 2018; Allen-Zhu et al., 2019; Du et al., 2019) demonstrates that the function-space evolution of a suitably-initialized, infinitely-wide neural network is linear. (Lee et al., 2019) study an equivalent formulation in parameter space that takes the form of the linearization of a model around its parameter initialization. (Chizat et al., 2019) introduce "lazy" training, characterizing more general conditions for the optimization dynamics

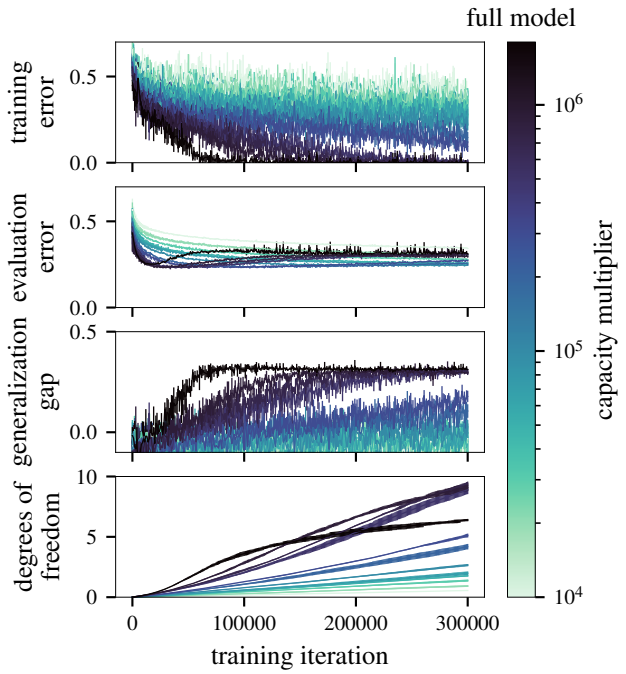

Figure 5. Training and evaluation error **(top two)**, the generalization gap, **(middle)**, and GDoF **(bottom)** vs. training iteration for different values of the capacity multiplier (here, subspace dimension) for 3 CNNs trained on the CIFAR-10 dataset for 300 thousand iterations. The GDoF captures the evolution of the generealization gap over training for the various capacity settings.

of a system to evolve as its linearization. However, various works (Mei et al., 2018; Chizat et al., 2019; Fort et al., 2020; Ortiz-Jiménez et al., 2021) have demonstrated empirically that linearizations of these forms are insufficient to capture the full behavior of nonlinear deep neural network models, in particular at the beginning of training. As such, there is a need for analyses that do not rely on linear approximations to the dynamics of learning in deep neural networks.

Finding neural networks with minimal complexity One class of methods directly search models as programs with low Kolmogorov complexity. For example, (Schmidhuber, 1997) proposes a probabilistic search algorithm to discover neural networks with low Kolmogorov complexity, although it has only been demonstrated with toy problems that can be solved with relatively short programs consist of limited primitives. Another class of methods encourages optimizers to find simpler neural networks by penalize deviation from a specified *prior* of the weights. The simplest example is "weight-decay" (Hinton, 1987), which can be interpreted as finding the *maximum a posteriori* (MAP) solution given an isotropic Gaussian prior over weights. (Nowlan & Hinton, 1992) elaborates this idea by introducing mixture of Gaussian priors over weights, but still assumes that all weights are quantized with the same tolerance of noise. This assumption has been drop in (Hinton & Van Camp, 1993), which employs the variational method to approximate the posterior distribution of *noisy weights* with different precisions.

Bayesian model selection and Occam's Razor. (MacKay, 1992a) reviews how *model selection* or *comparison* can be achieved in a Bayesian context by comparing the evidence for each model. In particular, given a parameterized model family $\mathcal{H}$ that specifies a *prior* distribution over parameters $\mathbf{w}$, $P(\mathbf{w} \mid \mathcal{H})$ and a *predictive* distribution over data $\mathcal{D}$, $P(\mathcal{D} \mid \mathbf{w}, \mathcal{H})$, the evidence is $P(\mathcal{D} \mid \mathcal{H}) = \int P(\mathcal{D} \mid \mathbf{w}, \mathcal{H}) P(\mathbf{w} \mid \mathcal{H}) \, \mathrm{d}\mathbf{w}$. As a rule of thumb, greater entropy in the prior term $P(\mathbf{w} \mid \mathcal{H})$, corresponding to a more flexible hypothesis class, results in a lesser value for the evidence, a regularizing effect known as *Occam's Razor*. It should be noted that the behavior of Occam's Razor—in particular what notion of flexibility is penalized—implicitly relies on the model parameterization and is therefore not an *a priori* effect (Wolpert, 1995); for instance, (MacKay, 1992a) discuss Occam's Razor and Bayesian model comparison for models that admit Gaussian approximation and quadratic regularizers, and (MacKay, 1992b) study such approximations in the context of neural networks. Nevertheless, Bayesian model selection allows coherent comparison of appropriately related prior distributions (Jefferys & Berger, 1992), and robustness with respect to the selection of the prior distribution can be ensured in some situations (Berger et al., 1994).

Implicit biases in model design. Lineages of machine learning models are often the result of selection on the basis of performance on standardized benchmarks (Dehghani et al., 2021). As a result, such lineages implicitly depend on the types of data to which they have been applied—in particular, such models are likely to have implicit biases that aid performance on such datasets. Analogously, prior selection in Bayesian analysis often implicitly relies on data (Gelman et al., 2017). From a Bayesian perspective, the impact of prior information can be quantified using sensitivity analysis (e.g., (Müller, 2012)) or with measures of prior-data and prior-likelihood conflict (Bousquet, 2008; Nott et al., 2020; Reimherr et al., 2021).

6. Conclusion

In this work, we introduced a new empirical complexity measure inspired by a classical notion in statistics, degrees of freedom. This measure is efficiently computable and is a function of the entire machine learning training pipeline. We demonstrated that this measure strongly correlates with generalization performance in the double-descent regime.

References

Advani, M. S., Saxe, A. M., and Sompolinsky, H. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020. URL https://doi.org/10.1016/j.neunet.2020.08.022.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, 2019. URL https://arxiv.org/abs/1811.03962.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Baraniuk, R., Donoho, D., and Gavish, M. The science of deep learning. *Proceedings of the National Academy of Sciences*, 117(48):30029–30032, 2020. URL https://doi.org/10.1073/pnas.2020596117.

Barrett, D. and Dherin, B. Implicit gradient regularization. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2009.11162.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017. URL https://arxiv.org/abs/1706.08498.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., De la Horra, J., Martín, J., Ríos-Insúa, D., Betrò, B., et al. An overview of robust bayesian analysis. *Test*, 3(1):5–124, 1994. URL https://doi.org/10.1007/BF02562676.

Bianchini, S., Müller, M., and Pelletier, P. Deep learning in science. *arXiv preprint arXiv:2009.01575*, 2020. URL https://arxiv.org/abs/2009.01575.

Bousquet, N. Diagnostics of prior-data agreement in applied bayesian analysis. *Journal of Applied Statistics*, 35 (9):1011–1029, 2008. URL https://doi.org/10.1080/02664760802192981.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1812.07956.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020. URL https://arxiv.org/abs/2011.03395.

Dehghani, M., Tay, Y., Gritsenko, A. A., Zhao, Z., Houlsby, N., Diaz, F., Metzler, D., and Vinyals, O. The benchmark lottery. 2021. URL https://arxiv.org/abs/2107.07002.

Dherin, B., Munn, M., and Barrett, D. G. The geometric Occam's razor implicit in deep learning. *arXiv preprint arXiv:2111.15090*, 2021. URL https://arxiv.org/abs/2111.15090.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, 2017. URL https://arxiv.org/abs/1703.04933.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1810.02054.

Efron, B. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American statistical association*, 78(382):316–331, 1983. URL https://doi.org/10.1080/01621459.1983.10477973.

Forster, M. and Sober, E. How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, 45(1):1–35, 1994. URL https://www.journals.uchicago.edu/doi/10.1093/bjps/45.1.1.

Fort, S., Dziugaite, G. K., Paul, M., Kharaghani, S., Roy, D. M., and Ganguli, S. Deep learning versus kernel learning: An empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. *Advances in Neural Information Processing Systems*, 33, 2020. URL https://arxiv.org/abs/2010.15110.

Gao, T. and Jojic, V. Degrees of freedom in deep neural networks. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2016. URL https://arxiv.org/abs/1603.09260.

Gelman, A., Simpson, D., and Betancourt, M. The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10):555, 2017. URL https://doi.org/10.3390/e19100555.

Geman, S., Bienenstock, E., and Doursat, R. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Golubeva, A., Gur-Ari, G., and Neyshabur, B. Are wider nets better given the same number of parameters? In *International Conference on Learning Representations*, 2020.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning (ICML)*, pp. 1225–1234. PMLR, 2016. URL https://arxiv.org/abs/1509.01240.

Hinton, G. E. Learning translation invariant recognition in a massively parallel networks. In *International Conference on Parallel Architectures and Languages Europe*, pp. 1–13. Springer, 1987. URL http://www.csri.utoronto.ca/~hinton/absps/parle.pdf.

Hinton, G. E. and Van Camp, D. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, pp. 5–13, 1993. URL https://www.cs.toronto.edu/~hinton/csc2535/readings/colt93.pdf.

Hope, T., Downey, D., Etzioni, O., Weld, D. S., and Horvitz, E. A computational inflection for scientific discovery. *arXiv preprint arXiv:2205.02007*, 2022. URL https://arxiv.org/abs/2205.02007.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018. URL https://arxiv.org/abs/1806.07572.

Jefferys, W. H. and Berger, J. O. Ockham's razor and bayesian analysis. *American Scientist*, 80(1):64–72, 1992.

Jiang, Y., Foret, P., Yak, S., Roy, D. M., Mobahi, H., Dziugaite, G. K., Bengio, S., Gunasekar, S., Guyon, I., and Neyshabur, B. NeurIPS 2020 competition: Predicting generalization in deep learning. *arXiv preprint arXiv:2012.07976*, 2020a. URL https://arxiv.org/abs/2012.07976.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations (ICLR)*, 2020b. URL https://arxiv.org/abs/1912.02178.

Jiang, Y., Natekar, P., Sharma, M., Aithal, S. K., Kashyap, D., Subramanyam, N., Lassance, C., Roy, D. M., Dziugaite, G. K., Gunasekar, S., Guyon, I., Foret, P., Yak, S., Mobahi, H., Neyshabur, B., and Bengio, S. Methods and analysis of the first competition in predicting generalization of deep learning. In Escalante, H. J. and Hofmann, K. (eds.), *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pp. 170–190. PMLR, 2021. URL https://proceedings.mlr.press/v133/jiang21a.html.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1609.04836.

Kubo, M., Banno, R., Manabe, H., and Minoji, M. Implicit regularization in over-parameterized neural networks. *arXiv preprint arXiv:1903.01997*, 2019.

Larsen, B. W., Fort, S., Becker, N., and Ganguli, S. How many degrees of freedom do we need to train deep networks: a loss landscape perspective. *arXiv preprint arXiv:2107.05802*, 2021.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in neural information processing systems*, 2019. URL https://arxiv.org/abs/1902.06720.

Li, Y. and Liang, Y. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in neural information processing systems*, 2018. URL https://arxiv.org/abs/1808.01204.

Lucas, J., Bae, J., Zhang, M. R., Fort, S., Zemel, R., and Grosse, R. Analyzing monotonic linear interpolation in neural network loss landscapes. 2021. URL https://arxiv.org/abs/2104.11044.

MacKay, D. Bayesian model comparison and backprop nets. *Advances in neural information processing systems*, 4, 1991. URL https://papers.nips.cc/paper/1991/hash/c3c59e5f8b3e9753913f4d435b53c308-Abstract.html.

MacKay, D. J. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992a. URL https://doi.org/10.1162/neco.1992.4.3.415.

MacKay, D. J. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992b. URL https://doi.org/10.1162/neco.1992.4.3.448.

Maddox, W. J., Benton, G., and Wilson, A. G. Rethinking parameter counting: Effective dimensionality revisted. *arXiv preprint arXiv:2003.02139*, 2020.

Mehta, H., Cutkosky, A., and Neyshabur, B. Extreme memorization via scale of initialization. In *International Conference on Learning Representations*, 2021.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018. URL https://arxiv.org/abs/1804.06561.

Müller, U. K. Measuring prior sensitivity and prior informativeness in large bayesian models. *Journal of Monetary Economics*, 59(6):581–597, 2012. URL https://doi.org/10.1016/j.jmoneco.2012.09.003.

Nagarajan, V., Andreassen, A., and Neyshabur, B. Understanding the failure modes of out-of-distribution generalization. In *International Conference on Learning Representations (ICLR)*, 2021.

Neyshabur, B. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017. URL https://arxiv.org/abs/1709.01953.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *ICLR (Workshop)*, 2015. URL https://arxiv.org/abs/1412.6614.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. Towards understanding the role of overparametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018.

Nott, D. J., Wang, X., Evans, M., and Englert, B.-G. Checking for prior-data conflict using prior-to-posterior divergences. *Statistical Science*, 35(2):234–253, 2020. URL https://doi.org/10.1214/19-STS731.

Nowlan, S. J. and Hinton, G. E. Simplifying neural networks by soft weight-sharing. *Neural Computation*, 4(4):473–493, 1992.

Ortiz-Jiménez, G., Moosavi-Dezfooli, S.-M., and Frossard, P. What can linearized neural networks actually say about generalization? *arXiv preprint arXiv:2106.06770*, 2021. URL https://arxiv.org/abs/2106.06770.

Raghu, M. and Schmidt, E. A survey of deep learning for scientific discovery. *arXiv preprint arXiv:2003.11755*, 2020. URL https://arxiv.org/abs/2003.11755.

Reimherr, M., Meng, X.-L., and Nicolae, D. L. Prior sample size extensions for assessing prior impact and prior-likelihood discordance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2021. URL https://arxiv.org/abs/1406.5958.

Schmidhuber, J. Discovering neural nets with low kolmogorov complexity and high generalization capability. *Neural Networks*, 10(5):857–873, 1997. URL https://doi.org/10.1016/S0893-6080(96)00127-X.

Smith, S. L. and Le, Q. V. A bayesian perspective on generalization and stochastic gradient descent. 2018. URL https://arxiv.org/abs/1710.06451.

Wolpert, D. H. On the bayesian "occam factors" argument for occam's razor. In *Computational Learning and Natural Learning Systems*, volume 3. MIT Press, 1995. URL https://doi.org/10.1162/neco.1992.4.3.415.

Ye, J. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441):120–131, 1998. URL https://www.jstor.org/stable/2669609.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1611.03530.