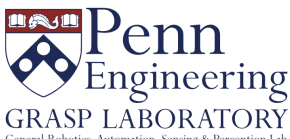
Does the Data Induce Capacity Control in Deep Learning?

Rubing Yang [1] Jialin Mao [1] Pratik Chaudhari [2]

[1]Applied Mathematics and Computational Sciences, University of Pennsylvania

[2]Electrical and Systems Engineering & Computer and Information Science, University of Pennsylvania

July 13, 2022



Sloppy eigenspectrum

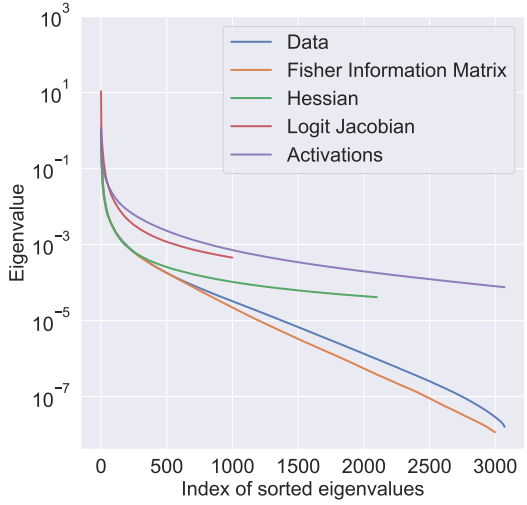


Figure: Eigenspectra for a trained wide residual network on CIFAR-10.

Analytical results for how sloppiness of data leads to sloppiness of neural network quantities

Theorem

Trace of the FIM and Hessian are bounded by that of the data correlation matrix

$$\mathrm{tr}\{F\}, \mathrm{tr}\{H\} \leq c\, tr(\mathbb{E}[xx^T]) \prod_{j=0}^{L} ||w^j||_2^2 \left(\sum_{j=0}^{L} \frac{1}{||w^j||_2^2} \right).$$

Analytical results for how sloppiness of data leads to sloppiness of neural network quantities

Theorem

Trace of the FIM and Hessian are bounded by that of the data correlation matrix

$$\mathrm{tr}\{F\}, \mathrm{tr}\{H\} \leq ctr(\mathbb{E}[xx^T]) \prod_{j=0}^{L} ||w^j||_2^2 \left(\sum_{j=0}^{L} \frac{1}{||w^j||_2^2} \right).$$

Theorem

The kth block on the diagonal of the FIM is sloppy if the activation h^k of that block are sloppy

$$spec\left(\mathbb{E}\left[\frac{dz_i}{dw^k} \frac{dz_i}{dw^k}^\top \right] \right) \preceq C \prod_{j=k+1}^{L} ||w^j||_2^2 \cdot spec(I_{d_{k+1}}) \otimes spec\left(\mathbb{E}\left[(h^k h^k)^\top \right] \right)$$

PAC-Bayes generalization bounds

Theorem (PAC-Bayes generalization bound McAllester (1999))

Let $e(Q)$ be the polulation error of a randomized hypothesis with a distribution Q, and its empirical error be $\hat{e}_n(Q)$. For a prior P, with probability at least $1-\delta$ over draws of n samples, we have

$$e(Q) \leq \hat{e}(Q) + \sqrt{\frac{KL(Q,P) + \log\left(\frac{n}{\delta}\right)}{2(n-1)}}.$$

PAC-Bayes generalization bounds

Theorem (PAC-Bayes generalization bound McAllester (1999))

Let $e(Q)$ be the polulation error of a randomized hypothesis with a distribution Q, and its empirical error be $\hat{e}_n(Q)$. For a prior P, with probability at least $1-\delta$ over draws of n samples, we have

$$e(Q) \leq \hat{e}(Q) + \sqrt{\frac{KL(Q,P) + \log\left(\frac{n}{\delta}\right)}{2(n-1)}}.$$

This is similar to bounds in Vapnik-Chernovenkis theory where there exists a constant V such that the generalization error of any hypothesis h from a model class can be bounded as

$$e(h) \leq \hat{e}(h) + \sqrt{\frac{V - \log \delta}{n}} \tag{1}$$

For deep networks, we have $V = \Theta(p)$, which leads to vacuous generalization bounds.

PAC-Bayes generalization bounds

We obtained a non-vacuous bound using the sloppiness of Hessian. This is the only analytical non-vacuous generalization bound for deep networks today.

PAC-Bayes generalization bounds

We obtained a non-vacuous bound using the sloppiness of Hessian. This is the only analytical non-vacuous generalization bound for deep networks today.

For a two-layer fully-connected MNIST network with 600 hidden neurons, the PAC-Bayes bound is 0.32.

PAC-Bayes generalization bounds

We obtained a non-vacuous bound using the sloppiness of Hessian. This is the only analytical non-vacuous generalization bound for deep networks today.

For a two-layer fully-connected MNIST network with 600 hidden neurons, the PAC-Bayes bound is 0.32.

We have also developed numerical techniques to further optimize such bounds using data-distribution dependent priors.

Effective dimensionality of a deep network

Definition

Define the **effective dimensionality for a deep network at local minimum** w as

$$p(n, \epsilon) = \sum_{i=1}^{p} \mathbf{1}\left\{|\lambda_i| \geq \frac{\epsilon}{2(n-1)}\right\}, \qquad (2)$$

Effective dimensionality of a deep network

Definition

Define the **effective dimensionality for a deep network at local minimum** w as

$$p(n, \epsilon) = \sum_{i=1}^{p} \mathbf{1}\left\{ |\lambda_i| \geq \frac{\epsilon}{2(n-1)} \right\}, \qquad (2)$$

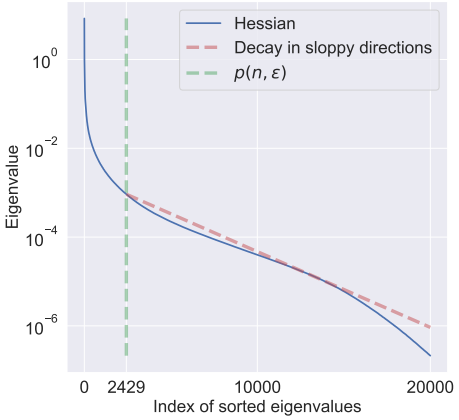


Figure: FC-600-2 with about 0.8 million parameters has effective dimensionality about 2500 which is about 0.3% of number of parameters.

Thanks for watching!
Scan to read the paper.

References

David A McAllester. PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 164–170, 1999.