Generalization error in Deep Learning:
Three approaches to understanding generalization in Deep Learning

Introduction

Neural network has strong ability to fit any data set but it also generalize well. Traditional statistical learning theories, like VC-dimension, fail to explain the generation performance of neural network. Then we turn to scale-sensitive measures of complexity. Specifically, we explore the recent work of margin-based bounds for neural network. The neural network we use is:

$$F_{\mathcal{A}} = \sigma_L(A_L\sigma_{L-1}(A_{L-1}\cdots\sigma_1(A_1x)\cdots)).$$

The network outputs a score for each class label $\{1, ..., k\}$. By taking the arg max over all the scores, we obtain the predicted class label.

Spectrally-normalized margin bound (Barlett et al., 2017)

Let nonlinearities $(\sigma_1, \ldots, \sigma_L)$ and reference matrices $(M_1, \ldots, M_L)$ be such that (σ_i is ρ_i-Lipschitz and $\sigma_i(0) = 0$). Then for $(x, y), (x_1, y_1), \ldots, (x_n, y_n)$ drawn iid from any probability distribution over $\mathbb{R}^d \times \{1, \ldots, k\}$, with probability at least $1 - \delta$ over $((x_i, y_i))_{i=1}^n$, every margin $\gamma > 0$ and network $F_{\mathcal{A}} : \mathbb{R}^d \to \mathbb{R}^k$ with weight matrices $\mathcal{A} = (A_1, \ldots, A_L)$ satisfy

$$\Pr(\arg\max_j F_{\mathcal{A}}(x)_j \neq y) \leq \widehat{\mathcal{R}}_\gamma(F_{\mathcal{A}}) + O(\frac{\|X\|_2 R_{\mathcal{A}}}{\gamma n}\ln(W) + \sqrt{\frac{\ln(1/\delta)}{n}})$$

Where $R_{\mathcal{A}} := \big(\prod_{i=1}^L \rho_i\|A_i\|_\sigma\big) \cdot \big(\sum_{i=1}^L \frac{\|A_i^\top - M_i^\top\|_{2,1}^{2/3}}{\|A_i\|_\sigma^{2/3}}\big)^{3/2}$ is called the spectral complexity of the network. In our experiments, we divide the margins by this quantity in order to obtain the normalized margins.

A PAC-Bayesian approach (Neyshabur et al., 2017)

If the inputs points are in the ball centered at the origin and of radius B, we have:

$$\Pr(\arg\max_j F_{\mathcal{A}}(x)_j \neq y) \leq \widehat{\mathcal{R}}_\gamma(F_{\mathcal{A}}) + O\Bigg(\sqrt{\frac{B^2L^2W\ln(LW)\cdot(\prod_{i=1}^L\|A_i\|_2)\cdot\big(\sum_{i=1}^L\frac{\|A_i\|_F^2}{\|A_i\|_2^2}\big) + \ln(\frac{Ln}{\delta})}{\gamma^2 n}}\Bigg)$$

Note: This result can only be applied to networks that use ReLU as their non-linearities (it is less general than the first one)

Fisher-Rao Metric approach (Liang et al., 2017)

In this paper, the bound put forward is the following one:

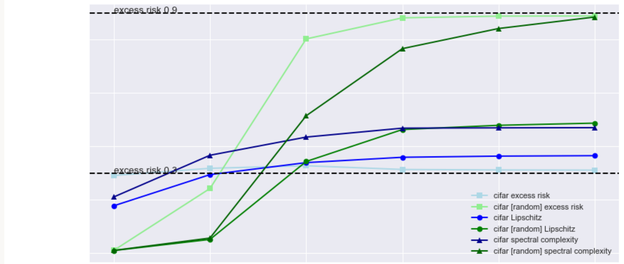
$$Pr(\arg\max_j F_{\mathcal{A}}(x)_j \neq y) \leq \frac{1}{n}\sum_{i=1}^n 1_{\arg\max_j F_{\mathcal{A}}(x_i)_j \cdot y_i \leq \gamma} + O(\frac{1}{\gamma}\mathcal{R}_N(B_{fr}(\gamma)) + \sqrt{\frac{\ln(1/\delta)}{n}})$$

Note: This result is only valid for binary classification, and for networks using ReLU as activation function.

The training and test accuracy of various models on the CIFAR10 dataset.

model	random crop	weight decay	train accuracy	test accuracy
AlexNet	yes	yes	100.0	78.21
	no	yes	100.0	77.83
	no	no	100.0	68.87
(random labels)	no	no	100.0	9.93
MLP 1×512	no	yes	100.0	53.21
	no	no	100.0	53.73
(random labels)	no	no	99.96	9.89

An analysis of AlexNet on `cifar10`



Margin distribution and spectral normalized margin distribution

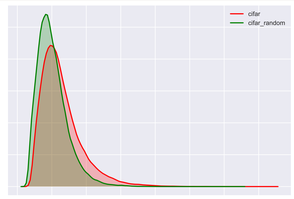


Figure: margin distribution

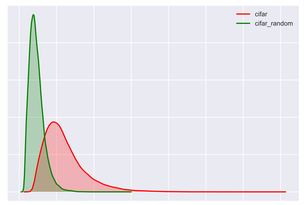


Figure: normalized margin distribution

Comparison of datasets and regularization

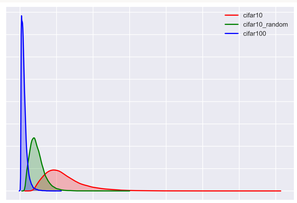


Figure: comparison of datasets

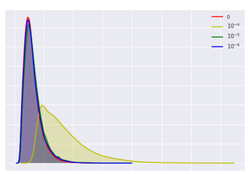


Figure: comparison of regularization

Comparison of epochs

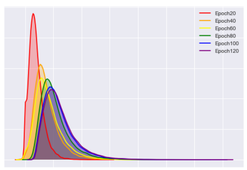


Figure: margin distribution across epochs

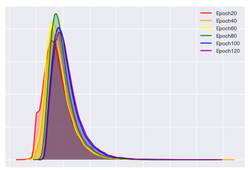


Figure: normalized margin distribution across epochs

Three types of complexity/metric

- spectral complexity

$$R_{\mathcal{A}} := \big(\prod_{i=1}^L \rho_i\|A_i\|_\sigma\big) \cdot \big(\sum_{i=1}^L \frac{\|A_i^\top - M_i^\top\|_{2,1}^{2/3}}{\|A_i\|_\sigma^{2/3}}\big)^{3/2}$$

- modified spectral complexity

$$R'_{\mathcal{A}} := \big(\prod_{i=1}^L \rho_i\|A_i\|_\sigma\big) \cdot L\big(\sum_{i=1}^L \frac{(\sqrt{W}\|A_i^\top - M_i^\top\|_2)^2}{\|A_i\|_\sigma^2}\big)^{1/2}$$

- Empirical Fisher-Rao metric

$$\|\mathcal{A}\|_{fr,emp}^2 = (L+1)^2\frac{1}{m}\sum_{i=1}^m[\langle softmax(F_{\mathcal{A}}(x_i)), F_{\mathcal{A}}(x_i)\rangle - F_{\mathcal{A}}(x_i)_{y_i}]^2$$

Comparison of three types of complexity/metric on MLP

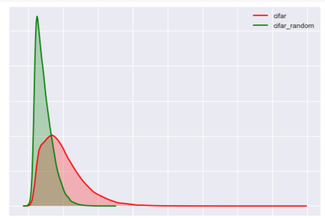


Figure: margin distribution

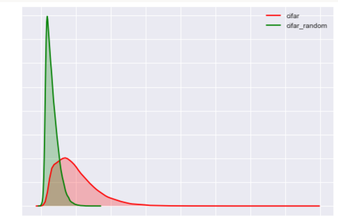


Figure: spectrally-normalized margin distribution

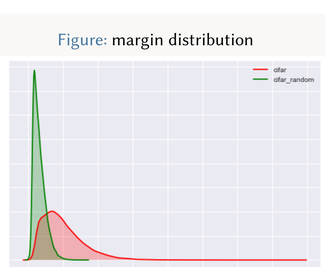


Figure: modified spectrally-normalized margin distribution

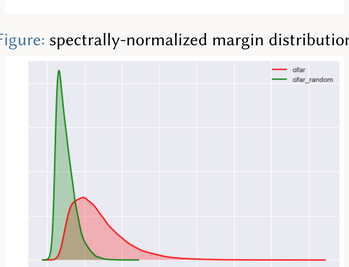


Figure: Fisher-Rao metric normalized margin distribution

Partially corrupted labels on AlexNet

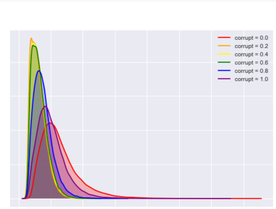


Figure: partially corrupted margin distribution

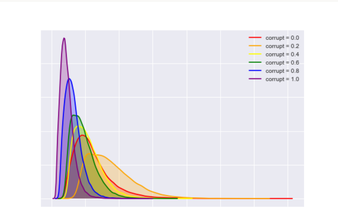


Figure: partially corrupted normalized margin distribution

Han, Xintian, xh1007@nyu.edu **Lahlou, Saad,** msl596@nyu.edu
Han, Xintian xh1007@nyu.edu, **Lahlou, Saad** msl596@nyu.edu
New York University