The Effect of Network Width on Stochastic Gradient Descent and Generalization

Daniel S. Park

Google

ICML 2019

Work with Jascha Sohl-Dickstein, Quoc V. Le and Samuel L. Smith.

Motivation

Let us assume that

- we found hyperparameters that maximize test set accuracy for a given network,
- but now we want to make the network bigger by widening all the channels by factor w.

What do we do with the hyperparameters for the new network?

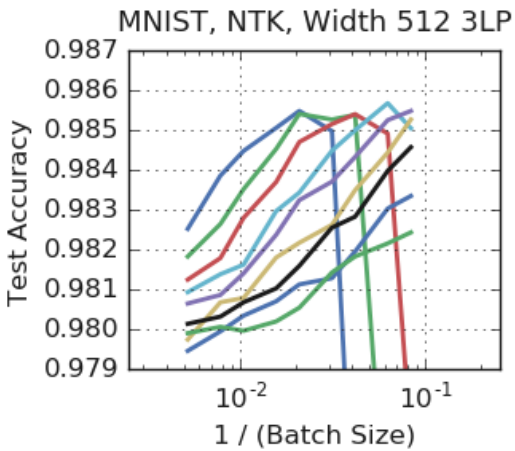
Main Result

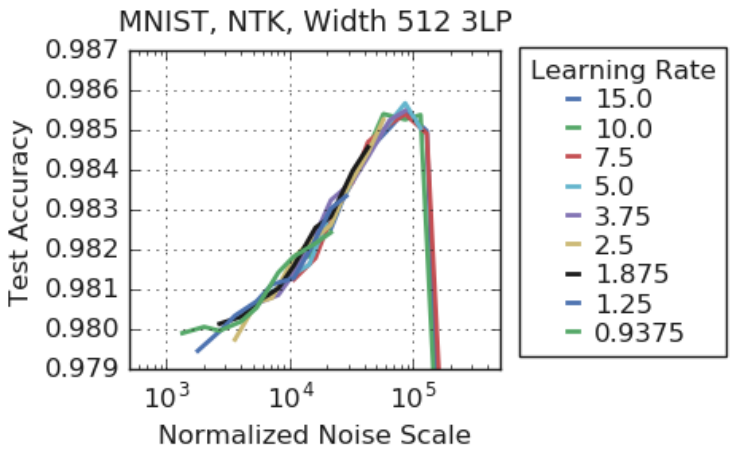
We find a rule that governs how hyperparameters that
maximize test accuracy change when the network width is varied.

The rule is that the optimal value of the **normalized noise scale**
(which is a function of the hyperparameters of SGD)
scales proportionally to the width of the network.

The Normalized Noise Scale $\bar{g}$

- $\bar{g} = \frac{\epsilon}{B(1-m)} \cdot \frac{1}{\sigma^2_{\text{init}}}$ governs how noisy the SGD is.
- $\bar{g}$ determines the generalization performance.*



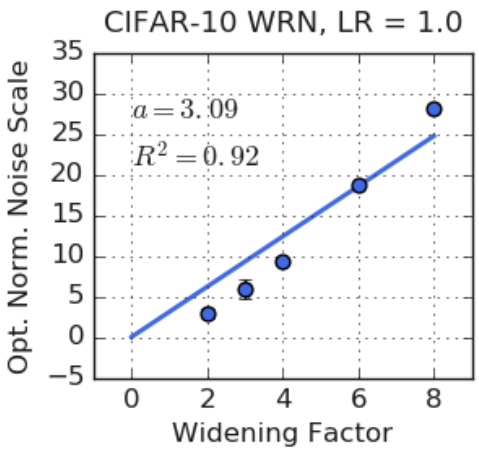


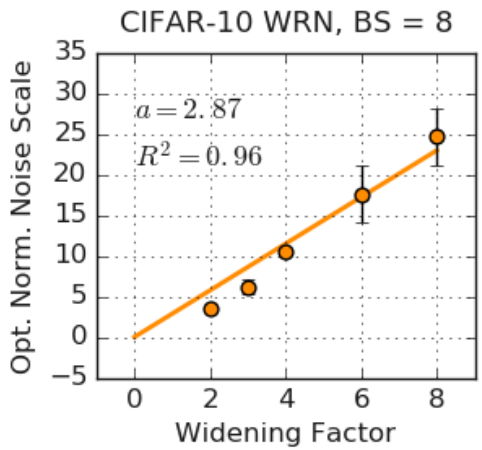
*Mandt et al. (2017); Chaudhari & Soatto (2017); Jastrzebski et al. (2017); Smith & Le (2017).

Rule for Hyperparameter Selection

- There exists a simple rule for hyperparameter selection:

Increase $\bar{g}$ proportionally with w.





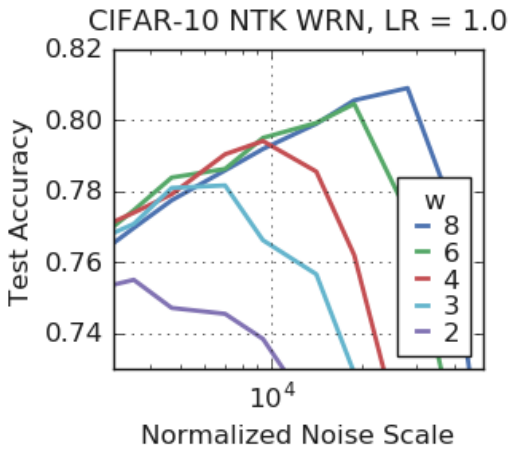
Wider networks require smaller batch sizes

- To maximize generalization performance, wide networks (eventually) need to be trained with small batch sizes:

$$B_{\text{opt}} \leq \frac{(\text{constant})}{w}$$

Bigger networks perform better due to noise resistance

- Bigger networks have better peak test set performance which is reached at higher noise scales.



Visit our poster (Pacific Ballroom #55) to learn more.

Thank you!