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 $\bar{g}$

- $\bar{g} = \frac{\epsilon}{B(1-m)} \cdot \frac{1}{\sigma_{\text{init}}^2}$ 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).*
There exists a simple rule for hyperparameter selection:

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

---

**Rule for Hyperparameter Selection**

- Increase $\tilde{g}$ proportionally with $w$. 

**Graphs:**

1. **CIFAR-10 WRN, LR = 1.0**
   - $a = 3.09$
   - $R^2 = 0.92$

2. **CIFAR-10 WRN, BS = 8**
   - $a = 2.87$
   - $R^2 = 0.96$
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!