On Tighter Generalization Bounds for Deep Neural Networks: CNNs, ResNets, and Beyond

arXiv preprint (2019)

Xingguo Li, Junwei Lu, Zhaoran Wang, Jarvis Haupt, and Tuo Zhao

Presenter: Gyuseung Baek

July 8, 2019

Introduction

- Give much tighter complexity bound than previous works
- Compute complexity with some special structure networks - CNNs, ResNets.

Introduction

Generalization Bound	Original Results	$\|W_d\|_2 = 1$
Neyshabur et al. (2015)	$\mathcal{O}\left(\frac{2^D \cdot \Pi_{d=1}^D B_{d,\mathrm{F}}}{\gamma\sqrt{m}}\right)$	$\mathcal{O}\left(\frac{2^D \cdot r^{D/2}}{\gamma\sqrt{m}}\right)$
Bartlett et al. (2017)	$\mathcal{O}\left(\frac{\Pi_{d=1}^D B_{d,2} \cdot \log(p)}{\gamma\sqrt{m}}\left(\sum_{d=1}^D \frac{B_{d,2\to 1}^{2/3}}{B_{d,2}^{2/3}}\right)^{3/2}\right)$	$\widetilde{\mathcal{O}}\left(\frac{\sqrt{D^3 pr}}{\gamma\sqrt{m}}\right)$
Neyshabur et al. (2017)	$\mathcal{O}\left(\frac{\Pi_{d=1}^D B_{d,2} \cdot \log(Dp)}{\gamma\sqrt{m}}\sqrt{D^2 p \sum_{d=1}^D \frac{B_{d,\mathrm{F}}^2}{B_{d,2}^2}}\right)$	$\widetilde{\mathcal{O}}\left(\frac{\sqrt{D^3 pr}}{\gamma\sqrt{m}}\right)$
Golowich et al. (2017)	$\mathcal{O}\left(\frac{\Pi_{d=1}^D B_{d,\mathrm{F}}}{\gamma} \cdot \min\left\{\frac{\sqrt{\log \frac{\Pi_{d=1}^D B_{d,\mathrm{F}}}{\Gamma}}}{\sqrt[4]{m}}, \sqrt{\frac{D}{m}}\right\}\right)$	$\widetilde{\mathcal{O}}\left(\frac{\sqrt{r^D \cdot D}}{\gamma\sqrt[4]{m}}\right)$
Our results	$\mathcal{O}\left(\frac{\Pi_{d=1}^D B_{d,2}\sqrt{Dpr} \cdot \log\left(\frac{B_{\backslash d,2}^{\mathrm{Jac}} \cdot \sqrt{Dm/r} \cdot \max_d B_{d,2}}{\gamma \cdot \sup g_\gamma(f(\mathcal{W}_D, x))}\right)}{\gamma\sqrt{m}}\right)$	$\widetilde{\mathcal{O}}\left(\frac{\sqrt{Dpr}}{\gamma\sqrt{m}}\right)$

Model and notation

- Same option + Rank constraint (rank $(W_i) \leq r_i$)

- If loss function is bounded, Rademacher complexity can be approved
- $M_{b:r}$: upper bound of the norm of Jacobian for function $N_{W_b^r}$

Generalization error bound

Theorem 1. Let g_γ be a $\frac{1}{\gamma}$ -Lipschitz loss function and $\mathcal{F}_{D,\|}$ the be the class of DNNs, $p_d = p, r_d = r$ for all $d \in [D]$, $M_{\setminus d} = \max_{d\in[D],x\in\mathcal{X}_m} M_{1:(d-1)} M_{(d+1):D}$ $C^{\mathrm{Net}} = \frac{M_{\setminus d} \cdot R\sqrt{Dm/r} \cdot \max_d M_d/\gamma}{\sup_{f\in\mathcal{F}_{D,\|},\|_2, x\in\mathcal{X}_m} g_\gamma(f(\mathcal{W}_D, x))}$ Then we have

$$\hat{\mathcal{R}}_m = \mathcal{O}\left(\frac{R\prod_{d=1}^{D} M_{d,2}\sqrt{Dpr\log C^{\mathrm{Net}}}}{\gamma\sqrt{m}}\right)$$

Generalization error bound with bounded loss

Corr 1.With Assumption Thm 1, suppose the loss is bounded, i.e. $l() \leq b$, then the Rademacher complexity safisfies

$$\hat{\mathcal{R}}_m = \mathcal{O}\left(C_1 \cdot \sqrt{\frac{Dpr \log C^{\mathrm{Net}}}{m}} \right)$$

where $C_1 = \min \left\{ R \prod_{d=1}^{D} M_d / \gamma, b \right\}$

CNNs with Orthogonal Filters

For CNN, we can use orthogonal filters

Generalization Bound	CNNs
Neyshabur et al. (2015)	$\mathcal{O}\left(\frac{2^D \cdot p^{\frac{D}{2}}}{\sqrt{m}}\right)$
Bartlett et al. (2017)	$\widetilde{\mathcal{O}}\left(\frac{\left(\frac{k}{s}\right)^{\frac{D-1}{2}} \cdot \sqrt{D^3 p^2}}{\sqrt{m}}\right)$
Neyshabur et al. (2017)	$\widetilde{\mathcal{O}}\left(\frac{\left(\frac{k}{s}\right)^{\frac{D-1}{2}} \cdot \sqrt{D^3 p^2}}{\sqrt{m}}\right)$
Golowich et al. (2017)	$\widetilde{\mathcal{O}}\left(p^{\frac{D}{2}} \min\left\{\frac{1}{\sqrt[4]{m}}, \sqrt{\frac{D}{m}}\right\}\right)$
Our results	$\widetilde{\mathcal{O}}\left(\frac{\left(\frac{k}{s}\right)^{\frac{D}{2}} \sqrt{Dk^2}}{\sqrt{m}}\right)$