Generalization error bound - via Compressing Deep Neural Network

arXiv preprint (2018)

Taiji Suzuki et al.

Presenter: Gyuseung Baek

September 9, 2019

Generalization error and Rademacher complexity

- W: parameters of DNN. D: Training data.
- For every margin $\gamma > 0$, w.p. at least $1 - \delta$, every $f \in \mathcal{F}$ satisfies

$$\Pr\left[\arg\max_i f(x)_i \neq y\right] \leq \widehat{\mathcal{R}}_\gamma(f) + 2\Re\left(\left(\mathcal{F}_\gamma\right)_{|D}\right) + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

where $\Re\left(\mathcal{H}_{|D}\right) := \frac{1}{n}\mathbb{E}\sup_{h\in\mathcal{H}}\sum_{i=1}^{n}\epsilon_i h\left(x_i, y_i\right)$ is the Rademacher complexity. (Bartlett, 2017)

Generalization error bound

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	$\mathcal{O}\left(\frac{\Pi_{d=1}^D B_{d,2}\sqrt{Dpr} \cdot \log\left(\frac{B_{\backslash d,2}^{\mathrm{lac}} \cdot \sqrt{Dm/r} \cdot \max_d B_{d,2}}{\gamma \cdot \sup g_\gamma(f(\mathcal{W}_D, x))}\right)}{\gamma\sqrt{m}}\right)$	$\widetilde{\mathcal{O}}\left(\frac{\sqrt{Dpr}}{\gamma\sqrt{m}}\right)$

Introduction

- Compress Deep Net
  - Low memory - available at small devices
  - Restriction may give better prediction
- Suggest compression algorithm and generalization error of compressed network
  - Compress width, not depth
  - Appliable to CNN and skip connection net

Notation

- Small letter is a vector, Capital letter is a matrix
- At first, consider fully connected net with training data $D = \{(x_i, y_i)\}_{i=1}^n$
- $f(x) = \left(W^{(L)}\eta(\cdot) + b^{(L)}\right) \circ \cdots \circ \left(W^{(1)}x + b^{(1)}\right)$ where $W^{(\ell)} \in \mathbb{R}^{m_{\ell+1} \times m_\ell}$, $b^{(\ell)} \in \mathbb{R}^{m_{\ell+1}}$, and η is an activation function.
- $\hat{f}$: trained network with data D. $\hat{}$ means trained.
- $\hat{F}_\ell(x) = \left(\hat{W}^{(\ell)}\eta(\cdot) + \hat{b}^{(\ell)}\right) \circ \cdots \circ \left(\hat{W}^{(1)}x + \hat{b}^{(1)}\right)$: subfunction of f

Compression algoriothm

- Let $\phi(x) = \eta\left(\hat{F}_{\ell-1}(x)\right) \in \mathbb{R}^{m_\ell}$ be the input to the ℓ-th layer.
- **(width) Compression**: Choose $J \in [m_\ell] := \{1, \ldots, m\}$ s.t. $\phi_J(x) = (\phi_j(x))_{j\in J} \in \mathbb{R}^{|J|}$ mimics $\phi(x)$

- For given J, we solve

$$\hat{A}_J = \underset{A\in\mathbb{R}^m_\ell \times |J|}{\operatorname{argmin}} \widehat{\mathrm{E}}\left[\|\phi - A\phi_J\|^2\right] + \|A\|_w^2$$

where $\|A\|_w^2 = \mathrm{Tr}\left[AI_wA^\top\right]$ for a regularization parameter $w \in \mathbb{R}_+^{|J|}$ and $\mathrm{I}_w = \mathrm{diag}(w)$

- Solution is

$$\hat{A}_J = \widehat{\Sigma}_{F,J}\left(\widehat{\Sigma}_{J,J} + \mathrm{I}_w\right)^{-1}$$

where $\widehat{\Sigma} := \widehat{\Sigma}_{(\ell)} = \frac{1}{n}\sum_{i=1}^n \eta\left(\hat{F}_{\ell-1}\left(x_i\right)\right)\eta\left(\hat{F}_{\ell-1}\left(x_i\right)\right)^\top$ and $\widehat{\Sigma}_{I,J} = \left(\widehat{\Sigma}_{i,j}\right)_{i\in I, j\in J}$

Input aware

- Loss of J: Similarity between output of compressed model and original(trained) model.

$$L_w^{(\mathrm{A})}(J) = \max_{z \in \mathbb{R}^m_{\ell}:\|z\| \leq 1} \min_{a \in \mathbb{R}^{m_\ell}} \widehat{\mathrm{E}}\left[\left(z^\top \phi - a^\top \phi_J\right)^2\right] + \left\|a^\top\right\|_w^2$$
$$= \left\|\widehat{\mathrm{E}}\left[\left(\phi - \hat{A}_J \phi_J\right)\left(\phi - \hat{A}_J \phi_J\right)^\top\right] + \hat{A}_J I_w \hat{A}_J^\top\right\|_{\mathrm{op}}$$

where $\|\cdot\|_{\mathrm{op}}$ is the spectral norm.

Output aware

- Loss of J: Similarity between $\ell + 1$-th layer of compressed model and original(trained) model.

$$L_w^{(\mathrm{B})}(J) = \max_{\|u\| \leq 1} \min_{a \in \mathbb{R}^{m_\ell}} \widehat{\mathrm{E}} \left[\left(u^\top Z_\ell \phi - a^\top \phi_J \right)^2 \right] + \left\| a^\top \right\|_w^2$$
$$= \left\| Z_\ell \left[\widehat{\Sigma}_{F,F} - \widehat{\Sigma}_{F,J} \left(\widehat{\Sigma}_{J,J} + \mathrm{I}_w \right)^{-1} \widehat{\Sigma}_{J,F} \right] Z_\ell^\top \right\|_{\mathrm{op}}$$

- Use convex combination of both criteria for a hyper parameter $0 \leq \theta \leq 1$ as

$$\theta L_w^{(\mathrm{A})}(J) + (1 - \theta) L_w^{(\mathrm{B})}(J)$$

Practical algorithm

- Forward selection can reduce calculation.
- Bu setting $w = 0$, we can fine J by

$$\min_{J \subset \{1,\ldots,m_\ell\}} |J| \text{ s.t. } \frac{\mathrm{Tr}\left[\left(\theta \mathrm{I} + (1-\theta) Z_\ell^\top Z_\ell\right) \widehat{\Sigma}_{F,J} \widehat{\Sigma}_{J,J}^{-1} \widehat{\Sigma}_{J,F}\right]}{\mathrm{Tr}\left[\left(\theta \mathrm{I} + (1-\theta) Z_\ell^\top Z_\ell\right) \widehat{\Sigma}_{F,F}\right]} \geq \alpha$$

  for a pre-specified $\alpha > 0$.

- For a convolution net, it can be applied by replace width with number of kernels.

Generalization error bound

- $\hat{f}$: Optimal. $\exists \hat{\zeta} \geq 0$ s.t. $\hat{L}(\hat{f}) \leq \min_{f \in \mathcal{F}} \hat{L}(f) + \hat{\zeta}$ holds almost surely.
- $f^{\#}$ compressed model. $\left(m_\ell^{\#}\right)_{\ell=1}^{L}$: width of $f^{\#}$.
- Support of input variable is compact and its ℓ_∞-norm is bounded as $\|x\|_\infty \leq D_x$.
- True function: f°

Generalization error bound

- *Let* $\lambda_\ell = \inf\left\{\lambda \geq 0 | m_\ell^\# \geq 5\hat{N}_\ell(\lambda)\log\left(80\hat{N}_\ell(\lambda)\right)\right\}$. If $\theta = 1$ and noise is small, then

$$\left\|f^\# - f^\circ\right\|_{L_2}^2 \lesssim \left(\sum_{\ell=2}^{L}\sqrt{\lambda_\ell}\right)^2 + \frac{1}{n}\sum_{\ell=1}^{L} m_{\ell+1}^\# m_\ell^\# \log(n)$$

uniformly over all choices of $m^\# = \left(m_1^\#, \ldots, m_L^\#\right)$ w.p. $1 - 5e^{-t}$

Trade-off of compression

- If compress model strongly, $\left(m_\ell^\#\right)_{\ell=1}^L$ became smaller and λ_ℓ became larger, vice versa
- We can choose $\left(m_\ell^\#\right)_{\ell=1}^L$ in a data dependent way to minimize generalization error bound (a posteriori)

Generalization error bound of original network

- If for a validation data $(x_i', y_i')_{i=1}^{n_{\mathrm{val}}}$, the validation error for $\hat{f}$ is smaller than that for $f^{\#}$ up to $q_n \geq 0$, i.e.

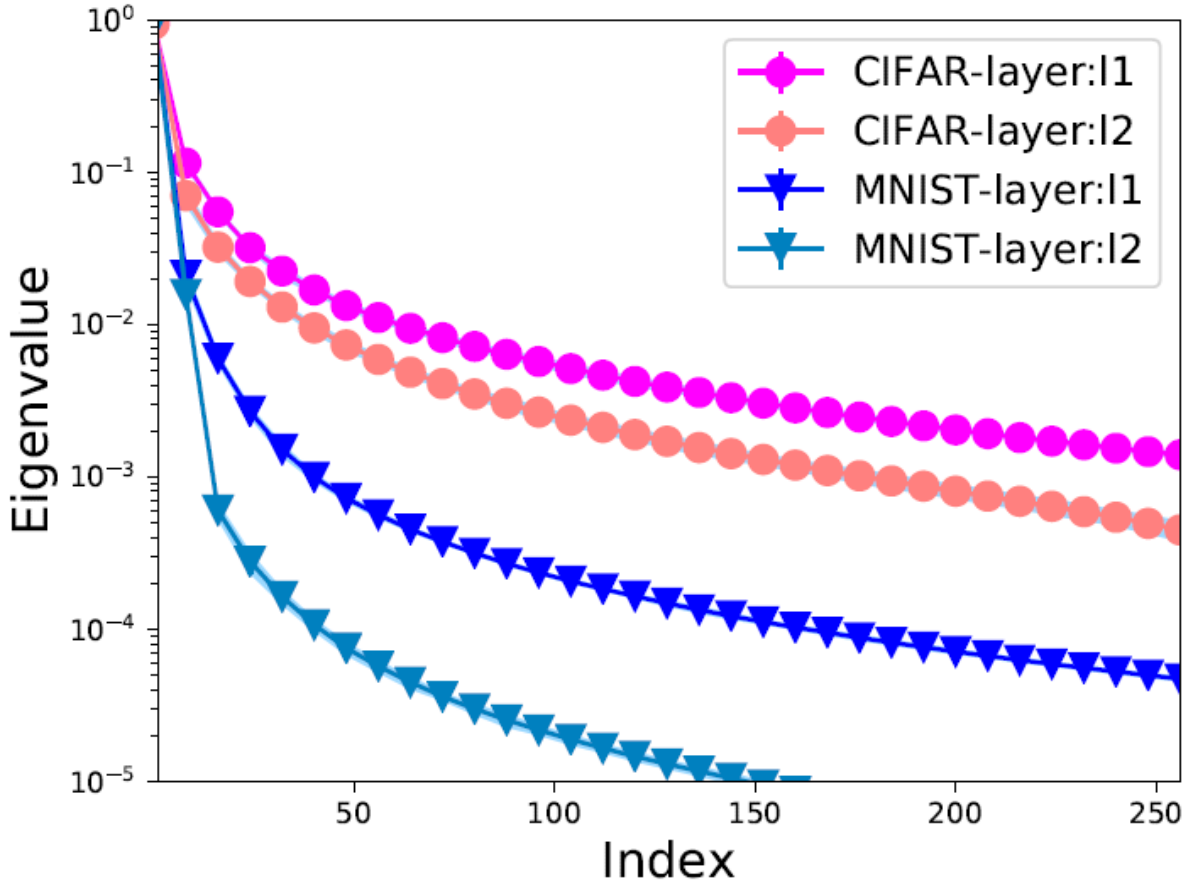
$$\frac{1}{n_{\mathrm{val}}} \sum_{i=1}^{n_{\mathrm{val}}} \left(y_i' - \widehat{f}\left(x_i'\right) \right)^2 \leq \frac{1}{n_{\mathrm{val}}} \sum_{i=1}^{n_{\mathrm{val}}} \left(y_i' - f^{\#}\left(x_i'\right) \right)^2 + q_n,$$

  then there exist constant C s.t.

$$\left\| \widehat{f} - f^{\circ} \right\|_{L_2}^2 \leq C \left(\left\| f^{\#} - f^{\circ} \right\|_{L_2}^2 + \frac{\sigma^2 + \hat{R}_{\infty}^2}{n_{\mathrm{val}}} t \right) + q_n$$

  where σ: true noise and $\hat{R}_{\infty} := \max \left\{ R^L D_x + \sum_{\ell=1}^{L} R^{L-\ell} R_b, \left\| f^{\circ} \right\|_{\infty} \right\}$.

Eigenvalue and Complexity



Simple Data needs small non-singular values of parameter.

Performance

Model	Top-1	Top-5	# Param.	FLOPs
Original VGG [43]	68.34%	88.44%	138.34M	30.94B
APoZ-2 [22]	70.15%	89.69%	51.24M	30.94B
SqueezeNet [24]	57.67%	80.39%	1.24M	1.72B
ThiNet-Conv [33]	69.80%	89.53%	131.44M	9.58B
ThiNet-GAP [33]	67.34%	87.92%	8.32M	9.34B
ThiNet-Tiny [33]	59.34%	81.97%	1.32M	2.01B
Spec-Conv ($\theta = 0$)	71.39%	90.63%	114.62M	20.02B
Spec-Conv ($\theta = 0.5$)	72.15%	91.06%	131.44M	22.13B
Spec-Conv ($\theta = 1.0$)	71.86%	90.88%	130.37M	18.73B
Spec-Conv-FC ($\theta = 1$)	68.66%	88.90%	45.77M	9.58B
Spec-GAP ($\theta = 0.5$)	67.55%	88.27%	8.31M	11.21B
Spec-Tiny ($\theta = 1$)	60.10%	82.89%	2.31M	2.07B
Spec-Conv2 ($\theta = 0.5$)	70.09%	89.82%	131.44M	9.58B
Spec-GAP2 ($\theta = 0.5$)	67.33%	87.99%	8.32M	9.34B
Spec-GAPe ($\theta = 0.5$)	67.78%	88.52%	8.25M	14.77B

Comparison

Norm based bound

Author	Rate	Bound type
Neyshabur et al. (2015)	$\frac{2^L \prod_{\ell=1}^L R_{\ell,F}}{\sqrt{n}}$	Norm base
Bartlett et al. (2017)	$\frac{\prod_{\ell=1}^L R_{\ell,2}}{\sqrt{n}} \left(\frac{R_{\ell,2\to 1}^{2/3}}{R_{\ell,2}^{2/3}} \right)^{3/2}$	Norm base
Neyshabur et al. (2017)	$\frac{\prod_{\ell=1}^L R_{\ell,2}}{\sqrt{n}} \sqrt{L^2 W \sum_{\ell=1}^L \frac{R_{\ell,F}^2}{R_{\ell,2}^2}}$	Norm base
Golowich et al. (2018)	$\prod_{\ell=1}^L R_{\ell,F} \min\left\{ \frac{1}{n^{1/4}}, \sqrt{\frac{L}{n}} \right\}$	Norm base
Li et al. (2018) Harvey et al. (2017)	$\frac{\prod_{\ell=1}^L R_{\ell,2} \sqrt{L^2 W^2}}{\sqrt{n}}$	VC-dim Naïve bound
Arora et al. (2018)	$\sqrt{\frac{L^2 \max_{1\le i\le n} \lvert\widehat{f}(x_i)\rvert^2 \sum_{\ell=1}^L \frac{1}{\mu_\ell^2 \mu_{\ell\to}^2}}{n}}$	Compression
Baykal et al. (2018)	$\sqrt{\frac{L^2 \max_{1\le i\le n} \lvert\widehat{f}(x_i)\rvert^2 \sum_{\ell=2}^L (\hat{\Delta}^{\ell\to})^2 \sum_{i=1}^W S_i^\ell}{n}}$	Compression
Suzuki et al. (2018)	$\sum_{\ell=2}^L \sqrt{\lambda_\ell} + \sqrt{\frac{\sum_{\ell=1}^L m_{\ell+1}^\sharp m_\ell^\sharp}{n}}$	Compression

Compression based bound