Stronger generalization bounds for deep nets via a compression approach

Speaker: Yi-Shan Wu

Institute of Information Science

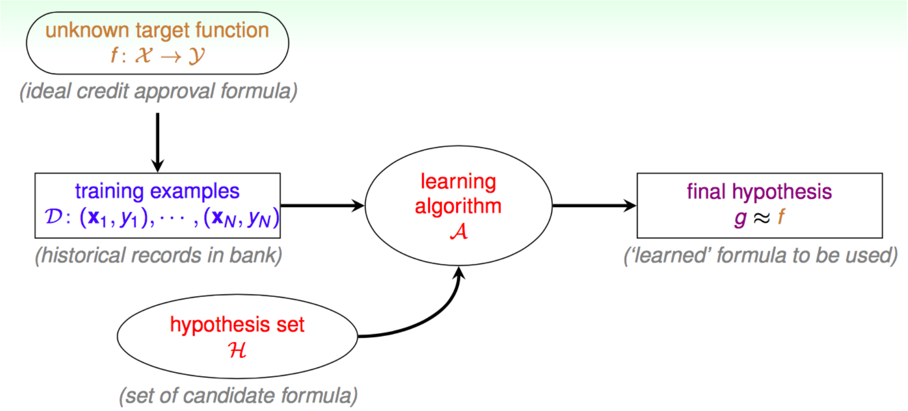
Academia Sinica

Taiwan

Aug 23, 2018

1. Generalization

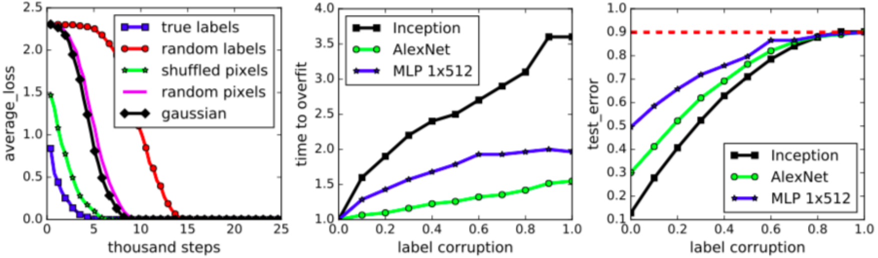
Learning Model



Goal of generalization :

$$\Pr\left[|L_S(A_S) - L_D(A_S)| \leq \epsilon\right] \geq 1 - \delta$$

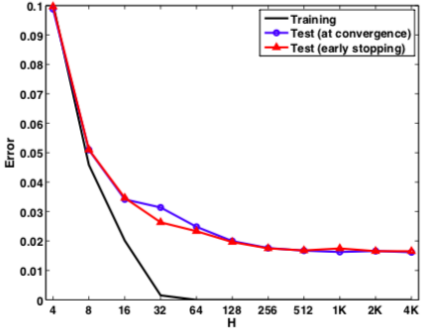
Measurement-VC bound(X)



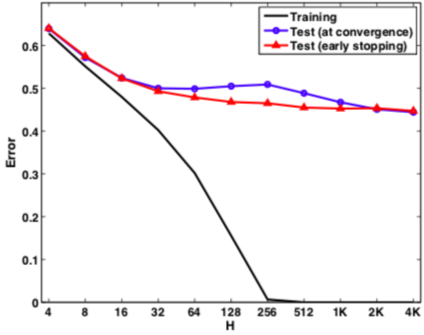
(a) learning curves (b) convergence slowdown (c) generalization error growth

It is possible to obtain zero training error on random labels using the same architecture for which training with real labels leads to good generalization.[5]

Measurement-parameter counting(X)



(a) MNIST



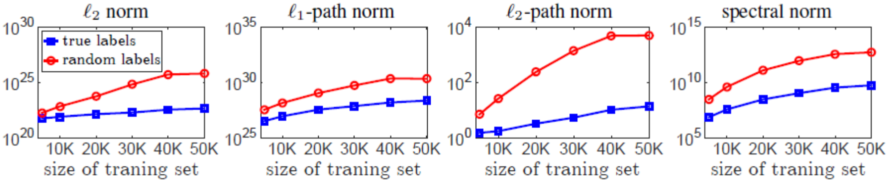
(b) CIFAR10

Increasing the number of parameters, can lead to a decrease in generalization error even when the training error does not decrease.[6]

Measurements : ex. norms and margins [4]

For linear predictors, the capacity can be controlled independent of the number of parameters, e.g. by regularization of ℓ_2 norm.

- ℓ_2 norm with capacity proportional to $\frac{1}{\gamma^2_{\text{margin}}} \prod_{i=1}^d 4 \|W_i\|_F^2$ [18].
- ℓ_1-path norm with capacity proportional to $\frac{1}{\gamma^2_{\text{margin}}} \left(\sum_{j \in \prod_{k=0}^d [h_k]} \left| \prod_{i=1}^d 2W_i[j_i, j_{i-1}] \right| \right)^2$ [4, 18].
- ℓ_2-path norm with capacity proportional to $\frac{1}{\gamma^2_{\text{margin}}} \sum_{j \in \prod_{k=0}^d [h_k]} \prod_{i=1}^d 4 h_i W_i^2 [j_i, j_{i-1}]$.
- spectral norm with capacity proportional to $\frac{1}{\gamma^2_{\text{margin}}} \prod_{i=1}^d h_i \|W_i\|_2^2$.



Compression approach [3]

Motivation : The analysis for previous works are too pessimistic. If the large trained network can be compressed to a smaller network without reducing performance, then it is able to obtain better generalization bound by previous approaches.

compressible

Let f be a classifier and $G_A = \{g_A | A \in \mathcal{A}\}$ be a class of classifiers and fixed strings s. We say f is (γ, S)-compressible via $G_{A,s}$ using helper string s if there exists $A \in \mathcal{A}$ such that for any $x \in S$, we have for all y

$$|f(x)[y] - g_{A,s}(x)[y]| < \gamma$$

Compression approach [3]

- A multi-class classifier $f : \mathcal{X} \to \mathbb{R}^K$.
- classfication loss $\mathbb{P}_{x,y} \sim D\{f(x)[y] < max_{j \neq y} f(x)[j]\}$
- If $\gamma > 0$ is some desired margin, then the expected margin loss is

$$L_\gamma(f) = \mathbb{P}_{x,y} \sim D\left\{f(x)[y] \leq \gamma + max_{j \neq y} f(x)[j]\right\}$$

- Let $\hat{L}_\gamma(f)$ denotes empirical margin loss

Compression approach [3]

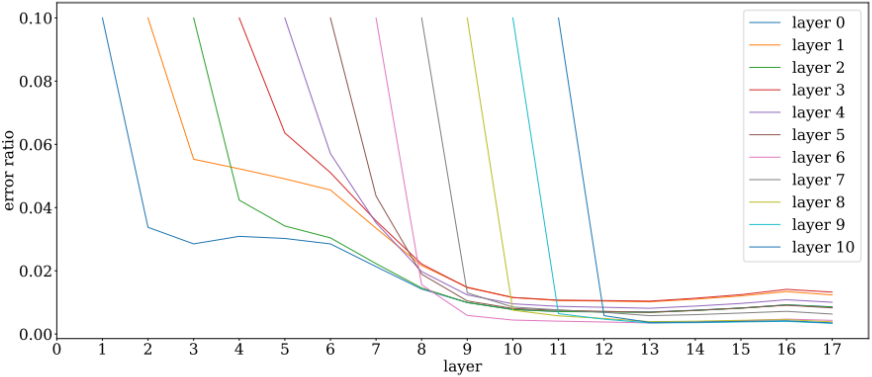
generalization bound for g_A

Suppose $G_A = \{g_A | A \in \mathcal{A}\}$ where A is a set of q parameters each of which can have at most r discrete values and s is a helper string. Let S be a training set with m samples. If the trained classifier f is (γ, S)-compressible via $G_{A,s}$ with helper string s, then there exists $A \in \mathcal{A}$ with high probability over the training set,

$$L_0(g_A) \leq \hat{L}_\gamma(f) + O\left(\sqrt{\frac{q \ln r}{m}}\right)$$

Observation : noise stability

The new "noise stability" approach roughly amount to saying that noise injected at a layer has very little effect on the higher layers.



Observation : noise stability

Goal :

- We compress each layer i by an appropriate randomized compression algorithm, such that the noise/error in its output is "gaussian-like".
- If layers $i+1$ and higher have low sensitivity to this new noise, then the compression can be more extreme and produce much higher noise.

Algorithm for compression

- The layers are compressed from lower layer to higher layer.
- ϵ, η are determined by the stability of that layer. That is, how many non-zero entries are left after compression is determined by the stability of the layer.

Algorithm 1 Matrix-Project (A, ε, η)

Require: Layer matrix $A \in \mathbb{R}^{h_1 \times h_2}$, error parameter ε, η.
Ensure: Returns $\hat{A}$ s.t. $\forall$ fixed vectors u, v,

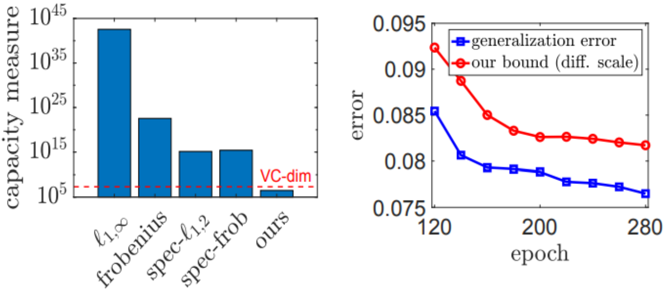
$$\Pr[|u^\top \hat{A} v - u^\top A v| \geq \varepsilon \|A\|_F \|u\| \|v\|] \leq \eta.$$

Sample $k = \log(1/\eta)/\varepsilon^2$ random matrices $M_1, \ldots, M_k$ with entries i.i.d. ± 1 ("helper string")
for $k' = 1$ to k **do**
  Let $Z_{k'} = \langle A, M_{k'} \rangle M_{k'}$.
end for
Let $\hat{A} = \frac{1}{k} \sum_{k'=1}^{k} Z_{k'}$

Theorem

For any fully-connected network f_A *with* $\rho_\delta \geq 3d$*, any* $0 < \delta < 1$*, and any margin* $\gamma > 0$*,* f_A *can be compressed into another fully connected network* $f_{\hat{A}}$ *such that* $\hat{L}_0(f_{\hat{A}}) \leq \hat{L}_\gamma(f_A)$ *and the number of parameters in* $f_{\hat{A}}$ *is at most:*

$$\tilde{O}\left(\frac{c^2 d^2 \max_{x\in S}\|f_A(x)\|_2^2}{\gamma^2}\sum_{i=1}^{d}\frac{1}{\mu_i^2 \mu_{i\rightarrow}^2}\right) \tag{11}$$



$\ell_{1,\infty}$: $\frac{1}{\gamma^2}\prod_{i=1}^{d}||A^i||_{1,\infty}$ Bartlett and Mendelson [2002]

Frobenius: $\frac{1}{\gamma^2}\prod_{i=1}^{d}||A^i||_F^2$ Neyshabur et al. [2015b],

spec $\ell_{1,2}$: $\frac{1}{\gamma^2}\prod_{i=1}^{d}||A_i||_2^2\sum_{i=1}^{d}\frac{||A^i||_{1,2}^2}{||A^i||_2^2}$ Bartlett et al. [2017]

spec-fro: $\frac{1}{\gamma^2}\prod_{i=1}^{d}||A^i||_2^2\sum_{i=1}^{d}h_i\frac{||A^i||_F^2}{||A^i||_2^2}$ Neyshabur et al. [2017a]

ours: $\frac{1}{\gamma^2}\max_{x\in S}||f(x)||_2^2\sum_{i=1}^{d}\frac{\beta^2 c_i^2\lceil\kappa/s\rceil^2}{\mu_i^2\mu_{i\rightarrow}^2}$

layer	$\frac{c_i^2 \beta_i^2 \lceil \kappa_i / s_i \rceil^2}{\mu_i^2 \mu_{i\rightarrow}^2}$	actual # param	compression (%)
1	1644.87	1728	95.18
4	644654.14	147456	437.18
6	3457882.42	589824	586.25
9	36920.60	1179648	3.129
12	22735.09	2359296	0.963
15	26583.81	2359296	1.126
18	5052.15	262144	1.927

FIGURE – Effective number of parameters identified by our bound. Compression rates can be as low as 1% in later layers (from 9 to 19) whereas earlier layers are not so compressible.

Reference

1. Spectrally-normalized margin bounds for neural networks, Bartlett et al., 2017
2. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks, Neyshabur et al., 2018(ICLR)
3. Stronger generalization bounds for deep nets via a compression approach, Arora et al.2018
4. Exploring Generalization in Deep Learning, Neyshabur et al.,2017
5. Understanding deep learning requires rethinking generalization. C.Zhang 2017(ICLR)
6. In search of the real inductive bias : On the role of implicit regularization in deep learning. Neyshabur 2017(ICLR)

Reference

7. On large-batch training for deep learning : Generalization gap and sharp minima. Kesker 2016
8. Train faster, generalize better : Stability of stochastic gradient descent. Hardt 2016
9. Nearly-tight vc-dimension bounds for piecewise linear neural networks. Harvey 2017
10. Rademacher and gaussian complexities : Risk bounds and structural results. Bartlett 2002
11. Norm-based capacity control in neural networks. Neyshabur 2015(COLT)
12. Computing nonvacuous generalization bounds for deep neural networks with many more parameters than training data. Dziugaite and Roy 2017