Sample complexity of learning Convolutional and Recurrent NNs

Aarti Singh
Associate Professor

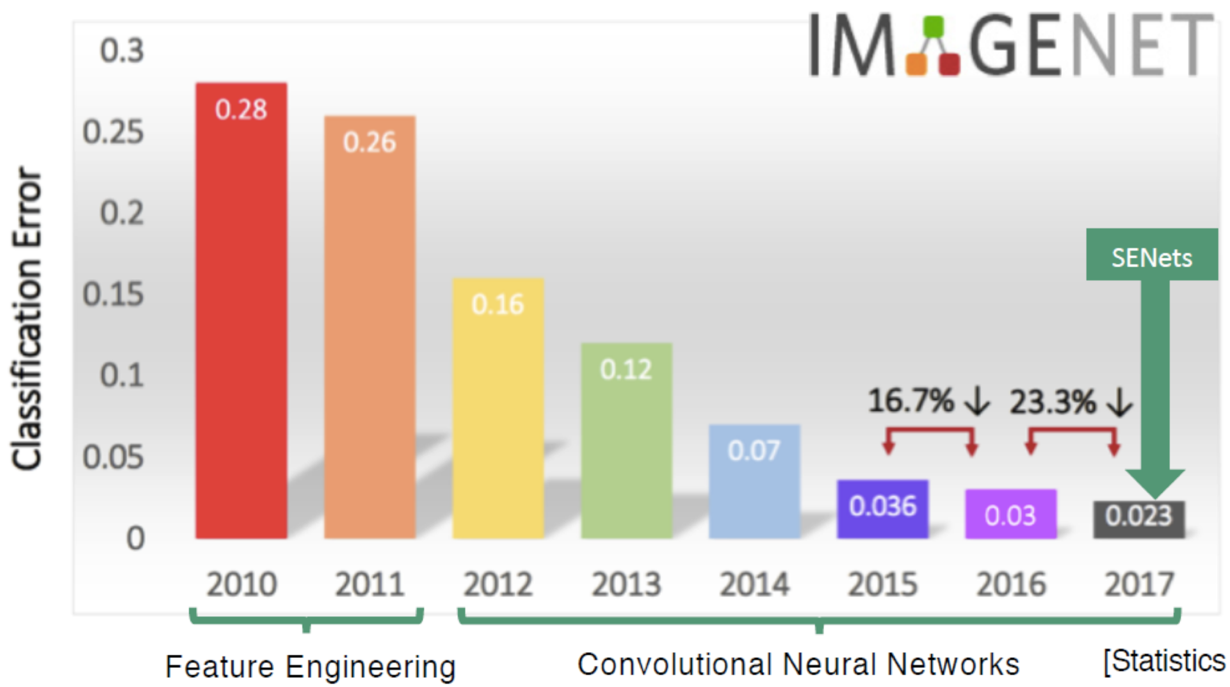
Joint work with Y. Wang, S. Du, X. Zhai, S. Balakrishnan, R. Salakhutdinov

Simons workshop on Frontiers of Deep Learning
July 2019

NeurIPS'18 and Arxiv: 1805.07883

CNNs and RNNs

- Large part of the recent success of NNs, particularly for spatial image data, is due to Convolution Neural Network (CNN) architectures (LeNet, AlexNet, VGG, GoogLeNet, ResNet, …)

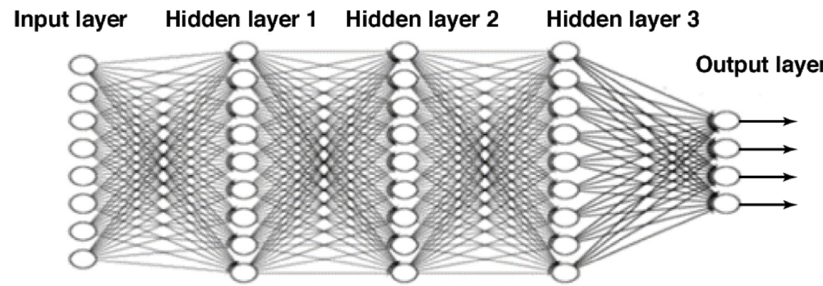


[Statistics provided by ILSVRC]

- Corresponding analogue for temporal or sequential data is the Recurrent Neural Network (RNN) architecture

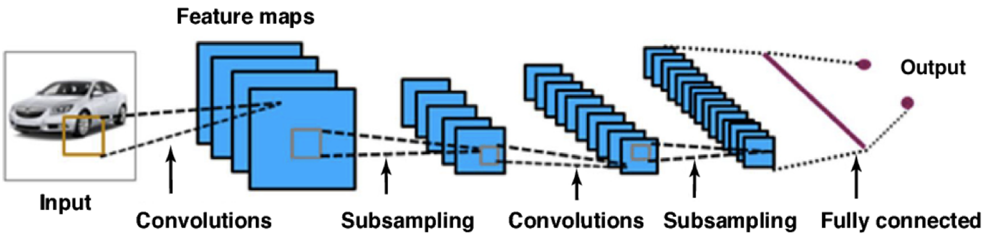
FNN, CNN and RNN architectures

- F(Fully-connected)NN



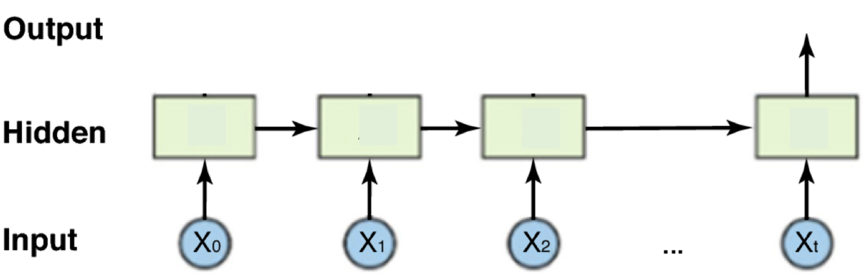
(a)

- C(Convolutional)NN



(b)

- R(Recurrent)NN



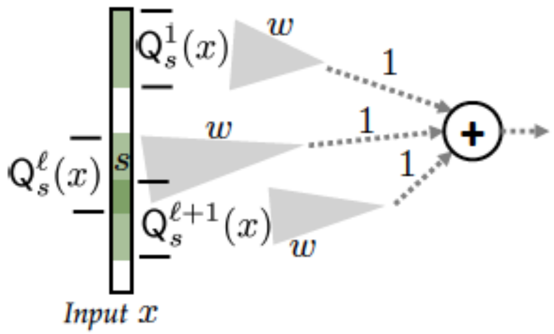
(c)

Image Src: Drug Discovery Today

CNN generative models

$$Y^i = F(X^i;\theta) + \xi_i \qquad X^i \in \mathbb{R}^d \qquad \{X^i\}_{i=1}^n \overset{\text{i.i.d}}{\sim} \mu$$

CNN with Average pooling $F^{\mathsf{CA}}(X^i; w) = \sum_{\ell=0}^{\lfloor (d-m)/s \rfloor} w^\top \mathsf{Q}_s^\ell(x^i)$

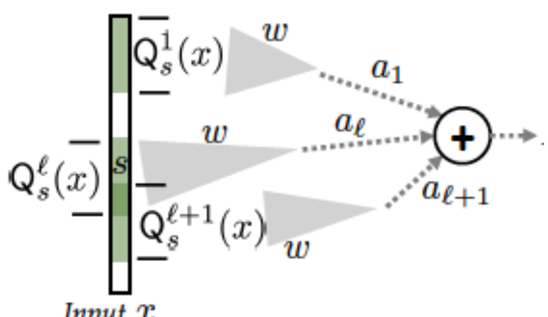


m – size of filter s – stride of filter

$$\mathsf{Q}_s^\ell(x^i) = (x^i_{\ell s+1}, \ldots, x^i_{\ell s+m})$$

length m segment of input

CNN with Weighted pooling $F^{\mathsf{CW}}(X^i; w, a) = \sum_{\ell=0}^{\lfloor (d-m)/s \rfloor} a_\ell w^\top \mathsf{Q}_s^\ell(x^i)$



Size of output layer, J = $\lfloor (d-m)/s \rfloor + 1$

RNN generative model

$$Y^i = F(X^i;\theta) + \xi_i \qquad X^i \in \mathbb{R}^d \qquad \{X^i\}_{i=1}^n \overset{\text{i.i.d}}{\sim} \mu$$

RNN $\quad F^{\mathsf{R}}(X^i, A, B) = \mathbf{1}^\top h_L^i$

$$h_t^i = A h_{t-1}^i + B x_t^i, \qquad t = 1, 2, \ldots, L.$$

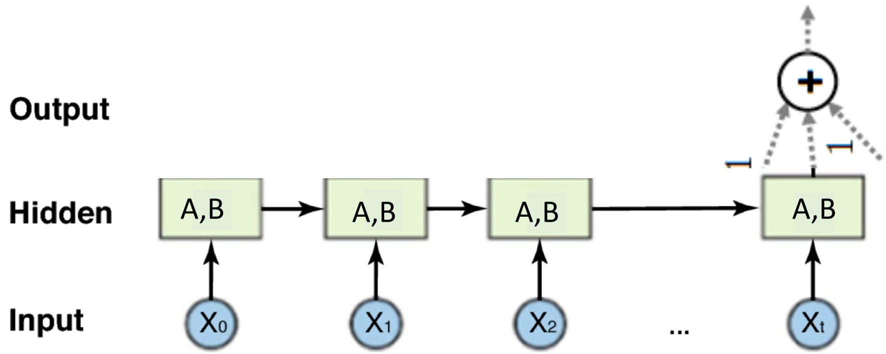
Initial hidden state, $h_0^i = 0$

$A \in \mathbb{R}^{r\times r}$

$B \in \mathbb{R}^{r\times d}$

L – length of input

r – hidden state dim



Minimax analysis

- Model may be non-identifiable (parameters not unique)

  E.g. w, a scaling for CNN or exchange hidden units in RNN

Focus on mean-square prediction error

$$\mathrm{err}(\widehat{\theta}, \theta) := \mathbb{E}_{\mu} |F(x;\theta) - F(x;\widehat{\theta})|^2$$

Goal: Upper and lower bound **Minimax risk**

$$\mathfrak{M}(n; F) := \inf_{\widehat{\theta}} \sup_{\theta} \mathbb{E}_{\mu,\theta} \left[\mathrm{err}(\widehat{\theta}, \theta) \right]$$

Estimator and Assumptions

Least Squares Estimator

$$\widehat{\theta} \in \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \left| Y^i - F(X^i; \theta) \right|^2$$

- May be non-unique, guarantees apply to any global minimizer
- Ignore computational considerations

Assumptions

A1) Noise is independent centered sub-gaussian (σ^2)

A2) Input distribution μ is centered sub-gaussian with

$$cI \preceq \mathbb{E}_\mu[xx^\top] \preceq CI$$

Main results (Informal)

$$\mathfrak{M}(n;F) := \inf_{\widehat{\theta}} \sup_{\theta} \mathbb{E}_{\mu,\theta}\left[\mathrm{err}(\widehat{\theta},\theta)\right]$$

CNN with Average pooling

$$\mathfrak{M}(n;F^{\mathsf{CA}}) = \tilde{\Theta}\left(\frac{m}{n}\right)$$

Independent of input dimension d
FNN ~ d/n

CNN with Weighted pooling

$$\mathfrak{M}(n;F^{\mathsf{CW}}) = \tilde{\Theta}\left(\frac{m+J}{n}\right)$$

RNN

$$\mathfrak{M}(n;F^{\mathsf{R}}) = \tilde{\Theta}\left(\frac{rd}{n}\right)$$

Independent of sequence length L
FNN ~ Ld/n

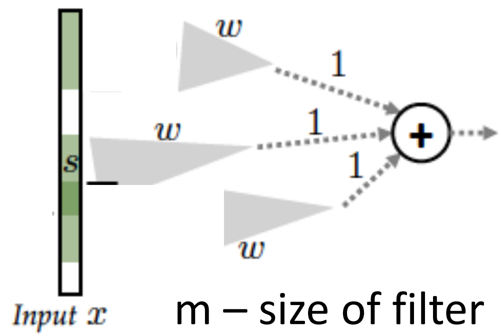
Main results (Informal)

$$\mathfrak{M}(n;F) := \inf_{\widehat{\theta}} \sup_{\theta} \mathbb{E}_{\mu,\theta}\left[\mathrm{err}(\widehat{\theta},\theta)\right]$$

Match parameter count

CNN with Average pooling

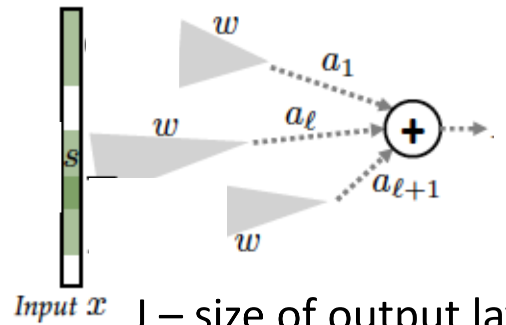
$$\mathfrak{M}(n;F^{\mathsf{CA}}) = \tilde{\Theta}\left(\frac{m}{n}\right)$$



m – size of filter

CNN with Weighted pooling

$$\mathfrak{M}(n;F^{\mathsf{CW}}) = \tilde{\Theta}\left(\frac{m+J}{n}\right)$$



J – size of output layer

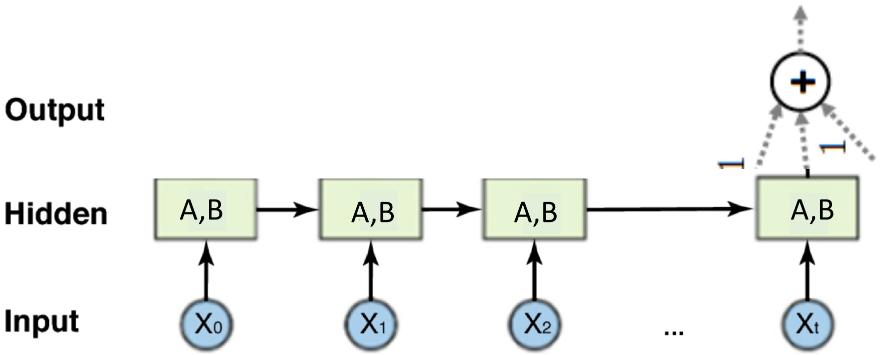
Main results (Informal)

$$\mathfrak{M}(n; F) := \inf_{\widehat{\theta}} \sup_{\theta} \mathbb{E}_{\mu,\theta}\left[\mathrm{err}(\widehat{\theta}, \theta)\right]$$

Match parameter count

RNN

$$\mathfrak{M}(n; F^{\mathrm{R}}) = \tilde{\Theta}\left(\frac{rd}{n}\right)$$



$$h_t^i = Ah_{t-1}^i + Bx_t^i$$

$$A \in \mathbb{R}^{r\times r}$$

$$B \in \mathbb{R}^{r\times d}$$

Related work

Generalization bounds for NNs; some also apply to CNNs

Arora et al' 18, Anthony and Bartlett'09, Bartlett et al'17, Neyshabur et al'17, Konstantinos et al'17, Zhou and Feng'18, Li et al'18, Long-Sedghi'19…

$$L(\theta) - L_{\mathrm{tr}}(\theta) \le D/\sqrt{n}$$

- Fast rate – we show 1/n rates (under some assumptions)
- Scale independence – model complexity D typically depends on norm of parameters

High-dimensional linear regression (d > n) – above issues akin to sparsity based analysis e.g. using lasso

Related work

RNN model special case of classical (Kalman, 1960) problem of learning a linear dynamical system

Recent statistical and computational analysis (Hazan et al’17; Hardt et al’18; Simchowitz et al’18; Oymak and Ozay’18)

- Sample complexity not tight (to best of our knowledge)

Upper bounds (formal)

With probability 1-δ, for sufficiently large n,

$$\mathfrak{M}(n; F^{\mathsf{CA}}) \lesssim \frac{\sigma^2 m \log d}{n}$$

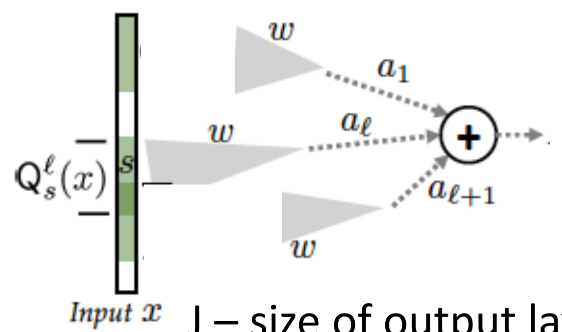
~ m+J when m/s ~ O(1)

$$\mathfrak{M}(n; F^{\mathsf{CW}}) \lesssim \frac{\sigma^2 \min\{d, \boxed{m + (d/s) \times (m/s)}\} \cdot \log d}{n}$$

$$\mathfrak{M}(n; F^{\mathsf{R}}) \lesssim \frac{\sigma^2 \boxed{(d+L)\min\{r, d\}} \log(Ld)}{n}$$

~ rd when r, L<< d

- All bounds are achieved by the least squares estimator
- $n/\log^2 n \gtrsim$ numerator
- Match parameter counts



J – size of output layer ~ d/s

Proof sketch

For least-squares solution, $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle X^i, \hat{\theta}\rangle)^2 \le \frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle X^i, \theta\rangle)^2$

Because of generative model $\|\widehat{\theta} - \theta\|_X^2 \le \frac{2}{n}\sum_{i=1}^{n}\xi_i\langle X^i, \hat{\theta} - \theta\rangle$

Self-normalized empirical process $\|\widehat{\theta} - \theta\|_X \le 2 \cdot \sup_{\phi \in \overline{\Theta}_X} \frac{1}{n}\sum_{i=1}^{n}\xi_i\langle X^i, \phi\rangle$

where $\overline{\Theta}_X := \{\phi = \theta - \theta' : \theta, \theta' \in \Theta, \|\phi\|_X \le 1\}$

Dudley's integral upper bounds expectation of the process, and hence the error, in terms of covering number of $\overline{\Theta}_X$

relate to covering number of using restricted eigenvalues (ensured by A2)

$$\overline{\Theta}_2(\rho) := \{\phi = \theta - \theta' : \theta, \theta' \in \Theta, \|\phi\|_2 \le \rho\}$$

$$\lambda_{\min}(\{X^i\}_{i=1}^n; \Phi) := \inf_{\phi \in \Phi} \|\phi\|_X^2 / \|\phi\|_2^2$$

$$\lambda_{\max}(\{X^i\}_{i=1}^n; \Phi) := \sup_{\phi \in \Phi} \|\phi\|_X^2 / \|\phi\|_2^2$$

Proof sketch

Lemma [Covering number of low-dim linear subspaces]: For any q, $k \leq q$, $\rho > 0$, *and* $\epsilon' \in (0, 1/2]$ there exists a finite set $\mathcal{W}$ of k-dimensional subspaces in $\mathbb{R}^q$ such that

for any k-dimensional subspace S in $\mathbb{R}^q$ there exists a subspace $S' \in \mathcal{W}$ such that

$$\sup_{u \in S, \|u\|_2 \leq \rho} \inf_{v \in S', \|v\|_2 \leq \rho} \|u - v\|_2 \leq \epsilon'$$

And the size of the set $\log |\mathcal{W}| \lesssim kq \log(\rho q/\epsilon')$.

Example RNN: $\theta := (\mathbf{1}^\top A^{L-1} B \quad \mathbf{1}^\top A^{L-2} B \quad \ldots \quad \mathbf{1}^\top B)$

L segments of d-dim, each of which lies in r-dim subspace

covering set of all 2r-dim subspaces in $\mathbb{R}^d$ O(rd log d)

covering of vectors in 2r-dim subspace for each + L x O(r)

Lower bounds (formal)

If input and noise distribution are standard normal, then there exists a universal constant C > 0 such that

$$\mathfrak{M}(n; F^{\mathsf{CA}}) \geq C \frac{\sigma^2 m}{n}$$

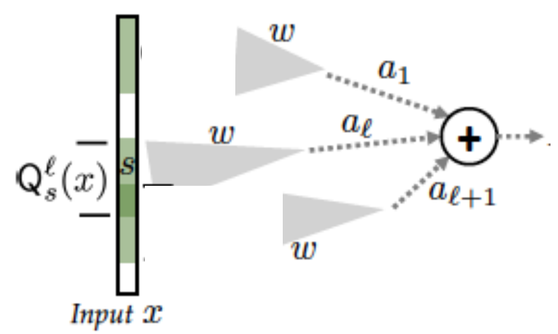
~ m+J

$$\mathfrak{M}(n; F^{\mathsf{CW}}) \geq C \frac{\sigma^2 (m + d/s)}{n}$$

~ rd since r << L

$$\mathfrak{M}(n; F^{\mathsf{R}}) \geq C \frac{\sigma^2 \min\{rd, Ld\}}{n}$$

- Bound holds for *any* estimator
- Lower bound for standard normal implies lower bound for general case
- Match parameter count



J – size of output layer ~ d/s

Proof sketch

Tsybakov extension of Fano's Lemma for Gaussian case

Corollary *For any finite subset* $\Theta' = \{\theta_0, \theta_1, \dots, \theta_M\} \subseteq \Theta$, *denote* $\rho_{\min} := \min_{j>0} \|\theta_0 - \theta_j\|_2/2$ *and* $\rho_{\text{avg}}^2 := \frac{1}{M}\sum_{i=1}^M \|\theta_i - \theta_0\|_2^2$. *Then for any* n,

$$\inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\mu[\|\widehat{\theta}_n - \theta\|_2] \geq \rho_{\min} \times \frac{\sqrt{M}}{1+\sqrt{M}} \left(1 - \frac{n\rho_{\text{avg}}^2}{\sigma^2 \log M} - 2\sqrt{\frac{n\rho_{\text{avg}}^2}{2\sigma^2 \log^2 M}}\right).$$

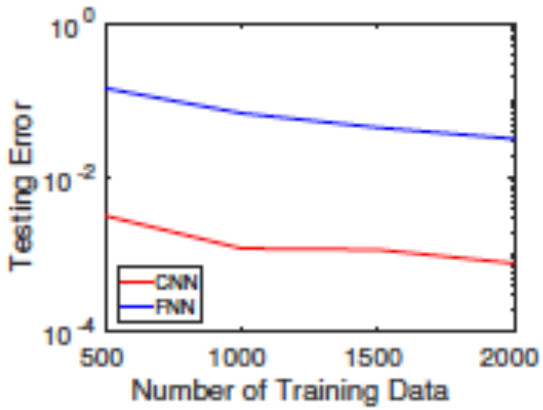
Characterization in terms of free parameters

Let $\Theta \subseteq \mathbb{R}^D, \mathcal{I} \subseteq [D]$. *Suppose for any* $u \in \mathbb{R}^{|\mathcal{I}|}$, *there exists* $\theta \in \Theta$ *such that* θ *restricted to* $\mathcal{I}$ *equals* u. *Then there exists a finite subset* $\Theta' \subseteq \Theta$ *as in Corollary with* $\log M \asymp |\mathcal{I}|$ *and* $\rho_{\min} \asymp \rho_{\text{avg}} \asymp \sqrt{|\mathcal{I}|}\epsilon$ *for any* $\epsilon > 0$.

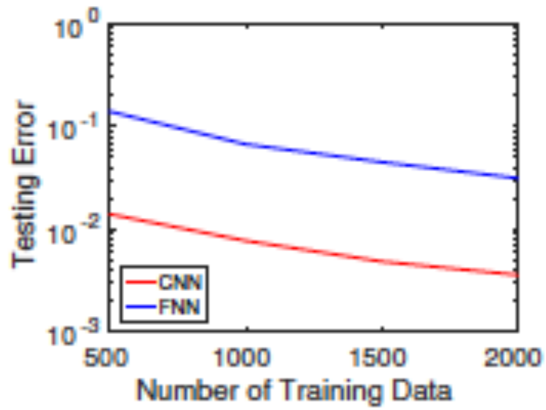
Experiments - CNN (average pooling) vs FNN

s = m (non-overlapping)

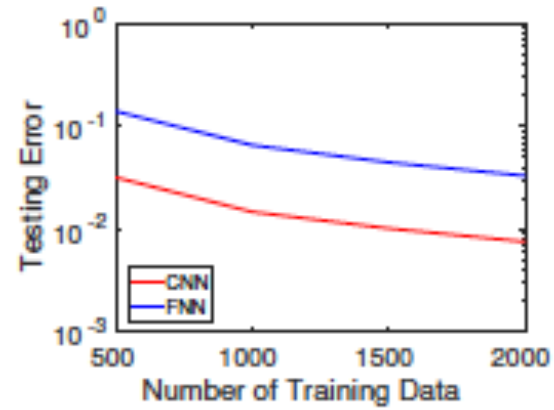
$d = 64$



(a) Filter size $m = 2$.

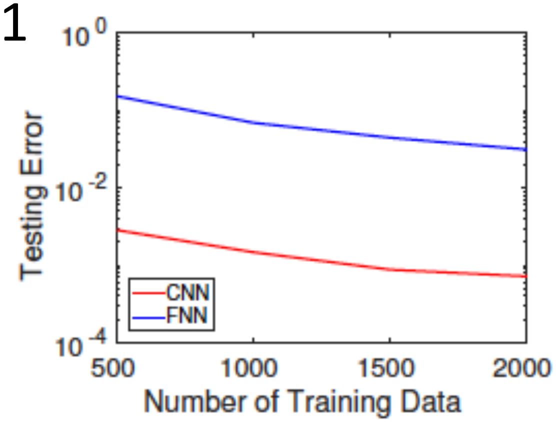


(b) Filter size $m = 8$.

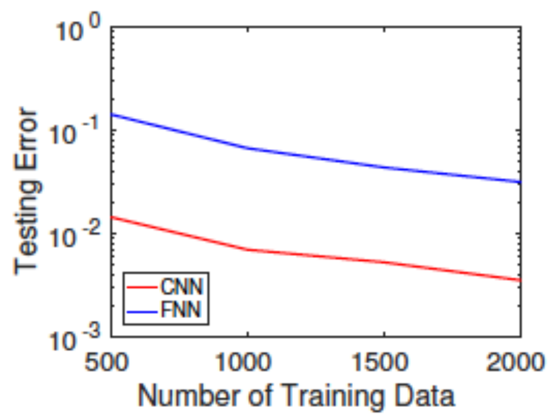


(c) Filter size $m = 16$.

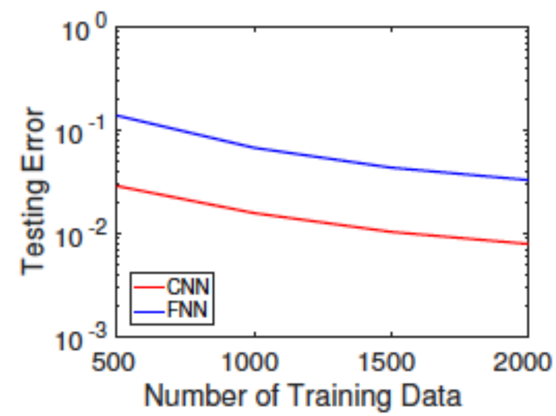
s = 1



(a) Filter size $m = 2$.



(b) Filter size $m = 8$.

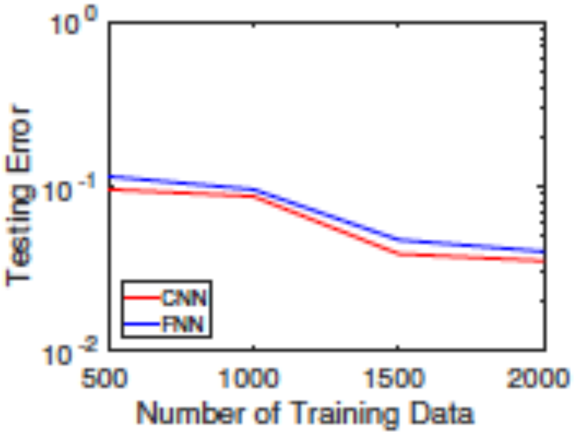


(c) Filter size $m = 16$.

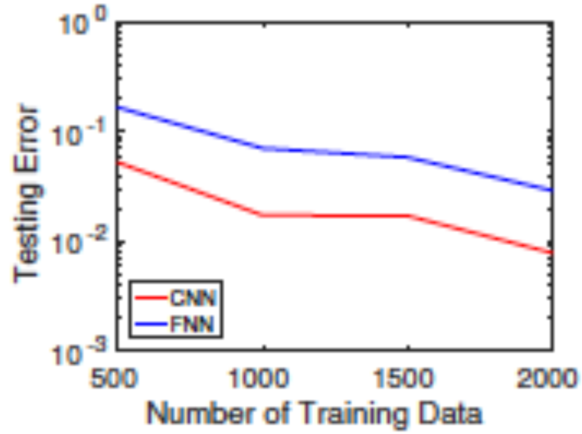
$\tilde{\Theta}\left(\frac{m}{n}\right)$ **Error decreases with n, increases with m and does not change with s**

Experiments - CNN (weighted pooling) vs FNN

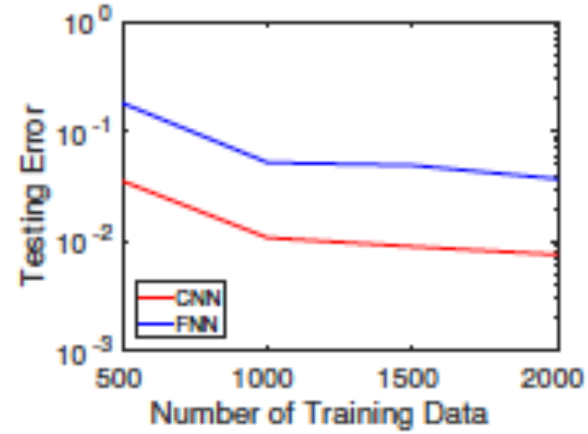
Filter size, m = 8 $d = 64$



(a) Stride size $s = 1$.



(b) Stride size $s = m/2$.



(c) Stride size $s = m$, i.e., non-overlapping.

m + J = 65 m + J = 23 m + J = 16

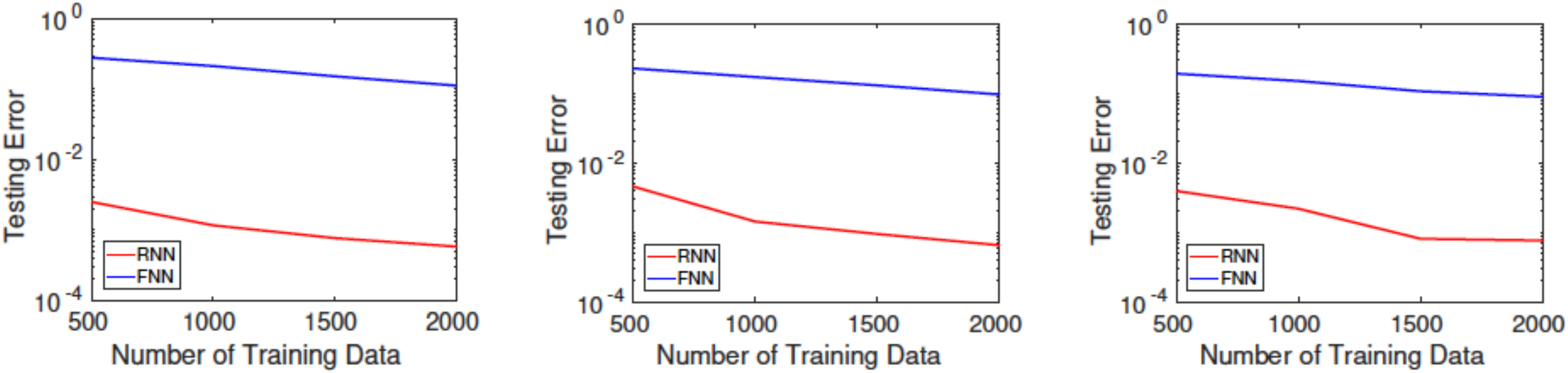
$$\tilde{\Theta}\left(\frac{m+J}{n}\right)$$

Error decreases with n, increases with J (and m)

(larger stride s implies smaller output layer size J ~ d/s)

Experiments - RNN vs FNN

d = 50, L = 50



(a) Hidden units $r = 2$. (b) Hidden units $r = 8$. (c) Hidden units $r = 16$.

$\tilde{\Theta}\left(\frac{rd}{n}\right)$ **Error decreases with n, increases with r (and d)**

Open questions

- Is fast rate possible
  - without generative model assumption (i.e. non-realizable case)
  - without distributional assumptions

  in high dimensions ($d > n$)?

  with computationally efficient estimators?

  Nonlinear activations

  Multiple filters

  Deep models

- Role of Optimization

Similarities to sparse high-dimensional linear regression