Besov Function Approximation and Binary Classification on Low-Dimensional Manifolds Using Convolutional Residual Networks

Hao Liu

Hong Kong Baptist University

Collaboration with
Minshuo Chen, Wenjing Liao, and Tuo Zhao

Georgia Institute of Technology

July 7, 2021

Introduction

Deep neural networks (NN) have demonstrated impressive performance in

- Computer vision [Krizhevsky et al., 2012]
- Natural language processing [Graves et al., 2013; Young et al., 2018; Wu et al., 2016]
- Health care [Miotto et al., 2018; Jiang et al., 2017]
- Bioinformatics [Alipanahi et al., 2015; Zhou & Troyanskaya, 2015]

Existing theories for FNNs and CNNs

Approximation theories of NNs have been studied for

Existing theories for FNNs and CNNs

Approximation theories of NNs have been studied for

- Feedforward neural networks (FNN) [Cybenko 1989; Hamers & Kohler 2006; Kohler & Mehnert 2011; Lu et al. 2017; Yarotsky 2017; Lee et al. 2017; Suzuki 2019]
- Convolutional neural networks (CNN) [Petersen & Voigtlaender, 2020; Zhou 2020a, 2020b, Oono & Suzuki 2019.]

Existing theories for FNNs and CNNs

Most of the existing work on FNNs and CNNs

- Are cursed by dimensionality:
  - To approximate a C^s function in $\mathbb{R}^D$ with accuracy ε, the network size is of $O(\varepsilon^{-D/s})$.
- Study Hölder or Sobolev functions.
  - [Suzuki 2019] studied Besov functions $B^{s+\alpha}_{p,q}$

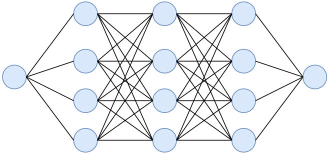
$$W^{s+\alpha,\infty} = \mathcal{H}^{s,\alpha} \subseteq B^{s+\alpha}_{\infty,\infty} \subseteq B^{s+\alpha}_{p,q}$$

for any $0 < p, q \leq \infty, s \in \mathbb{N}$ and $\alpha \in (0, 1]$.

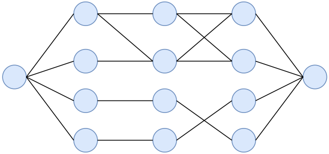
Existing theories for FNNs and CNNs

Most of the existing work on FNNs and CNNs

- Have a cardinality constraint



Without the cardinality constraint

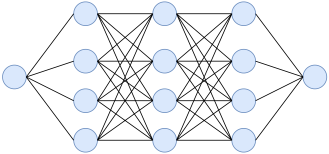


With a cardinality constraint

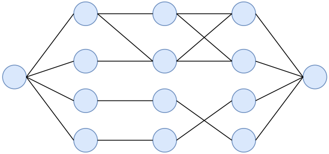
Existing theories for FNNs and CNNs

Most of the existing work on FNNs and CNNs

- Have a cardinality constraint



Without the cardinality constraint

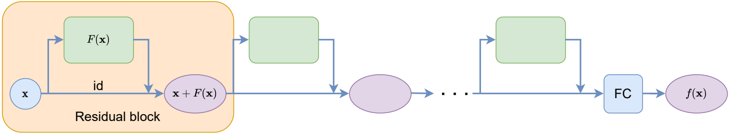


With a cardinality constraint

Training networks with the cardinality constraint needs substantial efforts [Han et al. 2015, 2016; Blalock et al. 2020].

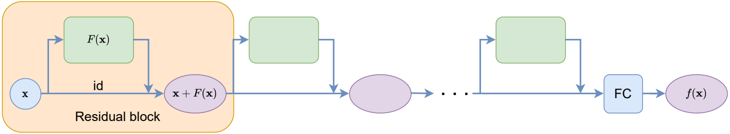
ConvResNets

The convolutional residual network (ConvResNet) is a special CNN with skip-layer connections.



ConvResNets

The convolutional residual network (ConvResNet) is a special CNN with skip-layer connections.



- Approximation theory of ConvResNets for Hölder functions is developed by [Oono & Suzuki 2019].

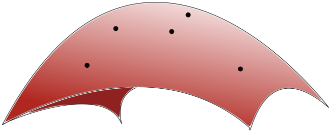
Properties:

- No cardinality constraints
- Cursed by dimensionality

Our work

Our work

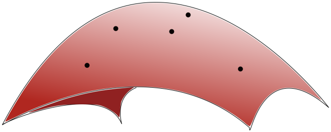
- We assume the data or target functions are located on a d-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^D$ with $d < D$.



Our work

Our work

- We assume the data or target functions are located on a d-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^D$ with $d < D$.



- We analyze the performance of ConvResNets on
  - Besov function approximation
  - Binary classification with the logistic loss

Besov function approximation

Theorem

Assume $0 < p, q \leq \infty$, $d/p + 1 \leq s < \infty$. *Given* $\varepsilon \in (0, 1)$ *and under mild assumptions, we construct a ConvResNet architecture. For any* $f^* \in B^s_{p,q}(\mathcal{M})$, *if the weight parameters of this ConvResNet are properly chosen, it gives rises to* $\bar{f}$ *satisfying*

$$\|\bar{f} - f^*\|_{L^\infty} \leq \varepsilon.$$

Besov function approximation

Theorem

Assume $0 < p, q \leq \infty$, $d/p + 1 \leq s < \infty$*. Given* $\varepsilon \in (0, 1)$ *and under mild assumptions, we construct a ConvResNet architecture. For any* $f^* \in B^s_{p,q}(\mathcal{M})$*, if the weight parameters of this ConvResNet are properly chosen, it gives rises to* $\bar{f}$ *satisfying*

$$\|\bar{f} - f^*\|_{L^\infty} \leq \varepsilon.$$

Remarks:

- There is no cardinality constraint.
- The network size is of $O(\varepsilon^{-d/s})$, and only weakly depends on D.

Binary classification

Problem settings:

- We are given a set of data $\{\mathbf{x}_i, y_i\}_{i=1}^n, \mathbf{x}_i \in \Omega$ in $\mathbb{R}^D$ and $y_i \in \{-1, 1\}$ follows the Bernoulli-type distribution

$$\mathbb{P}(y = 1|\mathbf{x}) = \eta(\mathbf{x}),\ \mathbb{P}(y = -1|\mathbf{x}) = 1 - \eta(\mathbf{x}).$$

- Learn a classifier using ConvResNets by minimizing the empirical logistic loss

Binary classification

Problem settings:

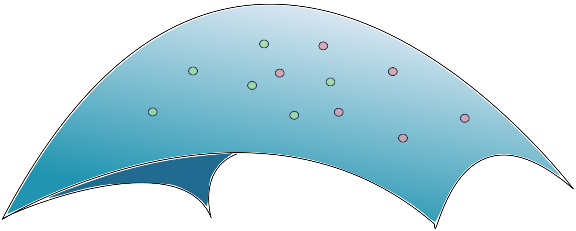
- We are given a set of data $\{\mathbf{x}_i, y_i\}_{i=1}^n, \mathbf{x}_i \in \Omega$ in $\mathbb{R}^D$ and $y_i \in \{-1, 1\}$ follows the Bernoulli-type distribution

$$\mathbb{P}(y = 1|\mathbf{x}) = \eta(\mathbf{x}),\ \mathbb{P}(y = -1|\mathbf{x}) = 1 - \eta(\mathbf{x}).$$

- Learn a classifier using ConvResNets by minimizing the empirical logistic loss

A low dimension manifold model for inputs:

- Assume $\{\mathbf{x}_i\}_{i=1}^n$ are located on a d-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^D$.



Binary classification

Theorem

Under the settings of the previous theorem, assume $\eta \in B^s_{p,q}(\mathcal{M})$. *Let* f^*_ϕ *be the minimizer of the population logistic risk. we construct a ConvResNet architecture with which minimizing the empirical logistic risk gives rise to* $\widehat{f}_{\phi,n}$ *with the following excess risk bound*

$$\mathbb{E}(\mathcal{E}_\phi(\widehat{f}_{\phi,n}, f^*_\phi)) \leq Cn^{-\frac{s}{2s+2(s\vee d)}} \log^4 n,$$

where $\mathcal{E}_\phi(\widehat{f}_{\phi,n}, f^*_\phi)$ *denotes the excess logistic risk of* $\widehat{f}_{\phi,n}$ *against* f^*_ϕ *and* C *is a constant independent of* n.

Binary classification

Theorem

Under the settings of the previous theorem, assume $\eta \in B^s_{p,q}(\mathcal{M})$. Let f^*_ϕ be the minimizer of the population logistic risk. we construct a ConvResNet architecture with which minimizing the empirical logistic risk gives rise to $\widehat{f}_{\phi,n}$ with the following excess risk bound

$$\mathbb{E}(\mathcal{E}_\phi(\widehat{f}_{\phi,n}, f^*_\phi)) \leq Cn^{-\frac{s}{2s+2(s\vee d)}} \log^4 n,$$

where $\mathcal{E}_\phi(\widehat{f}_{\phi,n}, f^*_\phi)$ denotes the excess logistic risk of $\widehat{f}_{\phi,n}$ against f^*_ϕ and C is a constant independent of n.

Remarks:

- Our result gives a faster rate depending on d instead of D
- ConvResNets are adaptive to the intrinsic dimension of data sets

Thank You