On the Number of Linear Regions of Convolutional Neural Networks (joint with L. Huang, M. Yu, L. Liu, F. Zhu and L. Shao)

Huan Xiong

Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)

ICML 2020

Motivations

- One fundamental problem in deep learning is understanding the outstanding performance of Deep Neural Networks (DNNs) in practice.
- Expressivity of DNNs: DNNs have the ability to approximate or represent a rich class of functions.
- Cybenko and Hornik-Stinchcombe-White (1989): A sigmoid neural network with one hidden layer and an arbitrarily large width can approximate any integrable function with arbitrary precision.
- Hanin-Sellke and Lu et al. (2017): A ReLU deep network of fixed width (determined by n) and arbitrarily large depth can approximate a given continuous function $f : [0,1]^n \to \mathbb{R}$ with arbitrary precision.

Piecewise Linear Functions Represented by ReLU DNNs

- The functions represented ReLU DNNs $\subseteq$ Piecewise linear functions.
- Piecewise linear functions can be used to approximate given functions.
- The more pieces, the more powerful expressivity.
- The maximal number of pieces (also called linear regions) in piecewise linear functions that a ReLU DNN can represent is a metric of the expressivity of ReLU DNNs.

Definition

- $R_{\mathcal{N},\theta}$: the number of linear regions of a neural network $\mathcal{N}$ with the parameters θ.
- $R_{\mathcal{N}} = \max_\theta R_{\mathcal{N},\theta}$: the maximal number of linear regions of $\mathcal{N}$ when θ ranges over $\mathbb{R}^{\#weights+\#bias}$.

Question

How to calculate the number $R_{\mathcal{N}}$ for a given DNN architecture $\mathcal{N}$?

The Maximal Number of Linear Regions for DNNs

Question

How to calculate the number $R_{\mathcal{N}}$ of linear regions for a given DNN architecture $\mathcal{N}$?

- Pascanu-Montúfar-Bengio (2013): $R_{\mathcal{N}} = \sum_{i=0}^{n_0} \binom{n_1}{i}$ for a one-layer fully-connected ReLU network $\mathcal{N}$ with n_0 inputs and n_1 hidden neurons.
- The basic idea is translating this problem to a counting problem of regions of hyperplane arrangements in general position, then directly applying Zaslavsky's Theorem (Zaslavsky, 1975), which says that the number of regions for a hyperplane arrangement in general position with n_1 hyperplanes over $\mathbb{R}^{n_0}$ is equal to $\sum_{i=0}^{n_0} \binom{n_1}{i}$.
- Montúfar-Pascanu-Cho-Bengio (2014): $R_{\mathcal{N}} \geq \left(\prod_{l=0}^{L-1} \left\lfloor \frac{n_l}{n_0} \right\rfloor^{n_0} \right) \sum_{i=0}^{n_0} \binom{n_L}{i}$ for a fully-connected ReLU network with n_0 inputs and L hidden layers of widths $n_1, n_2, \ldots, n_L$.
- Montúfar (2017): $R_{\mathcal{N}} \leq \prod_{l=1}^{L} \sum_{i=0}^{m_l} \binom{n_l}{i}$ where $m_l = \min\{n_0, n_1, n_2, \ldots, n_{l-1}\}$.
- Based on these results, they concluded that deep fully-connected ReLU NNs have exponentially more maximal linear regions than their shallow counterparts with the same number of parameters.
- Bianchini-Scarselli (2014); Telgarsky (2015); Poole et al. (2016); Raghu et al. (2017); Serra et al. (2018); Croce et al. (2018); Hu-Zhang (2018); Serra-Ramalingam (2018); Hanin-Rolnick (2019).

The Number of Linear Regions for ReLU CNNs

Question

How to calculate the number $R_{\mathcal{N}}$ of linear regions for a given DNN architecture $\mathcal{N}$?

- Most known results are about fully-connected ReLU NNs. What happens to CNNs?
- Difficulty for CNN case: the corresponding hyperplane arrangement is not in general position. Therefore, mathematical tools such as Zaslavsky's Theorem cannot be directly applied.
- Our main Contribution: we establish new mathematical tools needed to study hyperplane arrangements arisen in CNN case (which are not in general position) , and use them to derive upper and lower bounds on the maximal number of linear regions for ReLU CNNs.
- Based on these bounds, we show that deep ReLU CNNs have more expressivity than their shallow counterparts, and deep ReLU CNNs have more expressivity than deep ReLU fully-connected NNs per parameter, under some mild assumptions.

Main Result on the Number of Linear Regions for One-Layer CNNs

Theorem 1

Assume that $\mathcal{N}$ is a one-layer ReLU CNN with input dimension $n_0^{(1)} \times n_0^{(2)} \times d_0$ and hidden layer dimension $n_1^{(1)} \times n_1^{(2)} \times d_1$. The d_1 filters have the dimension $f_1^{(1)} \times f_1^{(2)} \times d_0$ and the stride s_1. Define $I_{\mathcal{N}} = \{(i,j) : 1 \leq i \leq n_1^{(1)}, 1 \leq j \leq n_1^{(2)}\}$ and $S_{i,j} = \{(a + (i-1)s_1, b + (j-1)s_1, c) : 1 \leq a \leq f_1^{(1)}, 1 \leq b \leq f_1^{(2)}, 1 \leq c \leq d_0\}$ for each $(i,j) \in I_{\mathcal{N}}$. Let

$$K_{\mathcal{N}} := \{(t_{i,j})_{(i,j)\in I_{\mathcal{N}}} : t_{i,j} \in \mathbb{N}, \sum_{(i,j)\in J} t_{i,j} \leq \# \cup_{(i,j)\in J} S_{i,j} \quad \forall J \subseteq I_{\mathcal{N}}\}.$$

(i) The maximal number $R_{\mathcal{N}}$ of linear regions of $\mathcal{N}$ equals

$$R_{\mathcal{N}} = \sum_{(t_{i,j})_{(i,j)\in I_{\mathcal{N}}} \in K_{\mathcal{N}}} \prod_{(i,j)\in I_{\mathcal{N}}} \binom{d_1}{t_{i,j}}.$$

(ii) Moreover, Suppose that the parameters θ are drawn from a fixed distribution μ which has densities with respect to Lebesgue measure in $\mathbb{R}^{\#weights+\#bias}$. Then the above formula also equals the expectation $\mathbb{E}_{\theta\sim\mu}[R_{\mathcal{N},\theta}]$.

Main Result on the Number of Linear Regions for One-Layer CNNs

Outline of the Proof of Theorem 1

- First, we translate the problem to the calculation of the number of regions of some specific hyperplane arrangements which may not be in general position.
- Next, we derive a generalization of Zaslavsky's Theorem with techniques from combinatorics and linear algebra, which can be used to calculate the number of regions of a large class of hyperplane arrangements.
- Finally, we show that the hyperplane arrangement corresponding to the CNN satisfies the condition of the above generalization of Zaslavsky's Theorem, thus the $R_{\mathcal{N}}$ and $\mathbb{E}_{\theta\sim\mu}[R_{\mathcal{N},\theta}]$ can be derived.

Asymptotic Analysis

Let $\mathcal{N}$ be the one-layer ReLU CNN defined in Theorem 1. Suppose that $n_0^{(1)}, n_0^{(2)}, d_0, f_1^{(1)}, f_1^{(2)}, s_1$ are some fixed integers. When d_1 tends to infinity, the asymptotic formula for the maximal number of linear regions of $\mathcal{N}$ behaves as $R_{\mathcal{N}} = \Theta(d_1^{\#\cup_{(i,j)\in I_{\mathcal{N}}} S_{i,j}})$ asymptotically. Furthermore, if all input neurons have been involved in the convolutional calculation, i.e., $\cup_{(i,j)\in I_{\mathcal{N}}} S_{i,j} = \{(a,b,c) : 1 \le a \le n_0^{(1)},\ 1 \le b \le n_0^{(2)},\ 1 \le c \le d_0\}$, we have

$$R_{\mathcal{N}} = \Theta(d_1^{n_0^{(1)} \times n_0^{(2)} \times d_0}).$$

Main Result on the Bounds of Multi-Layer CNNs

Theorem 2

Suppose that $\mathcal{N}$ is a ReLU CNN with L hidden convolutional layers. The input dimension is $n_0^{(1)} \times n_0^{(2)} \times d_0$; The l-th hidden layer has dimension $n_l^{(1)} \times n_l^{(2)} \times d_l$ for $1 \leq l \leq L$; and there are d_l filters with dimension $f_l^{(1)} \times f_l^{(2)} \times d_{l-1}$ and stride s_l in the l-th layer. Assume that $d_l \geq d_0$ for each $1 \leq l \leq L$. Then, we have

(i) The maximal number $R_{\mathcal{N}}$ of linear regions of $\mathcal{N}$ is at least (lower bound)

$$R_{\mathcal{N}} \geq R_{\mathcal{N}'} \prod_{l=1}^{L-1} \left\lfloor \frac{d_l}{d_0} \right\rfloor^{n_l^{(1)} \times n_l^{(2)} \times d_0},$$

where $\mathcal{N}'$ is a one-layer ReLU CNN which has input dimension $n_{L-1}^{(1)} \times n_{L-1}^{(2)} \times d_0$, hidden layer dimension $n_L^{(1)} \times n_L^{(2)} \times d_L$, and d_L filters with dimension $f_L^{(1)} \times f_L^{(2)} \times d_0$ and stride s_L.

(ii) The maximal number $R_{\mathcal{N}}$ of linear regions of $\mathcal{N}$ is at most (upper bound)

$$R_{\mathcal{N}} \leq R_{\mathcal{N}''} \prod_{l=2}^{L} \sum_{i=0}^{n_0^{(1)} n_0^{(2)} d_0} \binom{n_l^{(1)} n_l^{(2)} d_l}{i},$$

where $\mathcal{N}''$ is a one-layer ReLU CNN which has input dimension $n_0^{(1)} \times n_0^{(2)} \times d_0$, hidden layer dimension $n_1^{(1)} \times n_1^{(2)} \times d_1$, and d_1 filters with dimension $f_1^{(1)} \times f_1^{(2)} \times d_0$ and stride s_1.

Expressivity Comparison of Different Network Architectures

Theorem 3

Let $\mathcal{N}_1$ be an L-layer ReLU CNN in Theorem 2 where $f_l^{(1)}, f_l^{(2)} = \mathcal{O}(1)$ for $1 \leq l \leq L$, and $d_0 = \mathcal{O}(1)$. When $d_1 = d_2 = \cdots = d_L = d$ tends to infinity, we obtain that $\mathcal{N}_1$ has $\Theta(Ld^2)$ parameters, and the ratio of $R_{\mathcal{N}_1}$ to the number of parameters of $\mathcal{N}_1$ is

$$\frac{R_{\mathcal{N}_1}}{\#\text{ parameters of } \mathcal{N}_1} = \Omega\Big(\frac{1}{L} \cdot \Big\lfloor \frac{d}{d_0} \Big\rfloor^{d_0 \sum_{l=1}^{L-1} n_l^{(1)} n_l^{(2)} - 2}\Big).$$

For a one-layer ReLU CNN $\mathcal{N}_2$ with input dimension $n_0^{(1)} \times n_0^{(2)} \times d_0$ and hidden layer dimension $n_1^{(1)} \times n_1^{(2)} \times Ld^2$, when Ld^2 tends to infinity, $\mathcal{N}_2$ has $\Theta(Ld^2)$ parameters, and the ratio for $\mathcal{N}_2$ is

$$\frac{R_{\mathcal{N}_2}}{\#\text{ parameters of } \mathcal{N}_2} = \mathcal{O}\left(\left(Ld^2\right)^{d_0 n_0^{(1)} n_0^{(2)} - 1}\right).$$

- Based on the bounds obtained, we show that deeper ReLU CNNs have exponentially more linear regions per parameter than their shallow counterparts under some mild assumptions. This means that deeper CNNs have more powerful expressivity than shallow ones and thus provides some hints on why CNNs normally perform better as they get deeper.
- We also show that ReLU CNNs have more expressivity than fully-connected ReLU DNNs with asymptotically the same number of parameters, input dimension and number of layers.

Some Future Directions

- ReLU CNNs with pooling layers?
- We have obtained the expectation of $R_{\mathcal{N},\theta}$ for a one-layer ReLU CNN $\mathcal{N}$ and some general distribution μ of parameters θ. It would be interesting to explore similar formulas and bounds of the expectation of $R_{\mathcal{N},\theta}$ for multi-layer ReLU CNNs.
- Another direction related to $R_{\mathcal{N},\theta}$ is to study the influence of different parameters θ. When θ is replaced by some $\theta + \Delta\theta$, what is the relation between $R_{\mathcal{N},\theta}$ and $R_{\mathcal{N},\theta+\Delta\theta}$? These problems are related to the changing number of linear regions for CNNs during training process.