Final Project

Eddie Shim
TTIC 31250 Theory of Machine Learning
eddieshim@uchicago.edu

06/03/2020

Project Goal

In this project I'll be reading a recent paper in the field of theoretical machine learning and providing a summary of the paper's main goal and highlight some of the main results yielded from its conclusions. With a general interest in neural networks and regards to their general success in predictive performance and high ability for expressiveness and learning, I decided to select the following paper for my research: The Power of Depth for Feedforward Neural Networks by Ronen Eldan and Ohan Shamir, COLT 2016[1].

Motivation

In recent years, there's been a resurgence of attention towards artificial neural networks and their applications to modern problems, especially in the domains of computer vision and speech recognition. Yet, explaining why these highly parameterized models perform well in face of model parsimony has been an open question in the field of deep learning. More specifically, it's still yet to be understood how we can theoretically explain how a trained network that can be exponential in dimension will generalize well on unseen test data while avoiding the plights of overfitting. In light of this problem, Eldan's paper aims to address how we can explain the expressive power of neural networks of bounded size. The general question this paper wishes to address is: "What functions on $\mathbb{R}^d$ expressible by network with l-layers and w neurons per layer, that cannot be well approximated by any network with $< l$ layers, even if the number of neurons is allowed to

[1] https://arxiv.org/abs/1512.03965

be much larger than w?"

When considering a bounded neural network, a natural tradeoff question arises in how we should choose an optimal network architecture structure in regards to width versus depth. This aim addresses how important the "deep" in deep learning is – and later results in the paper show that depth, even if increased by 1, can be exponentially more valuable than increasing width. To do so, this paper hones its focus on fully connected feedforward neural networks, using a linear output and a non-linear activation function $\sigma(z)$ (eg: ReLU, sigmoid). The simplest possible case in this scenario is to compare the difficult of approximating functions computable by 3-layer networks using 2-layer networks. We define a 2-layer network as such:

$$x \mapsto \sum_{i=1}^{w} v_i \sigma(\langle w_i, x \rangle + b_i)$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function, and $v_i, b_i \in \mathbb{R}, w_i \in \mathbb{R}^d, i = 1, ..., w$ are the second layer weights, first layer bias, and first layer weights to the network. A 3-layer network of width w can be defined from this 2-layer equation as:

$$\sum_{i=1}^{w} u_i \sigma \left(\sum_{j=1}^{w} v_{i,j} \sigma(\langle w_{i,j}, x \rangle + b_{i,j}) + c_i \right)$$

where $u_i, c_i, v_{i,j}, b_{i,j} \in \mathbb{R}, w_{i,j} \in \mathbb{R}^d, i, j = 1, ..., w$. Intuitively this expression is outputs of the neurons in the first layer fed into the second layer, fed into a linear output neuron into the third layer. From here we arrive at the main motivation of this paper, which will proceed by showing there is a simple function on $\mathbb{R}^d$ that is expressible by a small 3-layer neural network but cannot be approximated by any 2-layer network, to more than a certain constant accuracy, unless its width is exponential in dimension.

Assumptions Required on $\sigma(z)$

Before jumping into the main theorem, we need some restrictions on our activation function. For example, if $\sigma(\cdot)$ is the identity, then both 2-layer and 3-layer functions are simply linear functions and thus have no difference in expressive power. Here are the two assumptions we require before our main theorem:

Assumption 1

Given the activation function σ, there is a constant $c_\sigma \geq 1$ (depending only on σ) such that for any L-Lipschitz function $f : \mathbb{R} \to \mathbb{R}$ which is constant outside a bounded interval $[-R, R]$, and for any δ, there

exist scalars $a, \{\alpha_i, \beta_i, \gamma_i\}_{i=1}^{w}$ where $w \leq c_\sigma \frac{RL}{\delta}$ such that the function

$$h(x) = a + \sum_{i=1}^{w} \alpha_i \cdot \sigma(\beta_i x - \gamma_i)$$

satisfies:

$$\sup_{x \in \mathbb{R}} |f(x) - h(x)| \leq \delta$$

This assumption gives us a method to approximate $h(x)$, meaning we can use it to show sufficiently large 2-layer network can approximate any univariate Lipschitz function which is non-constant on a bounded domain. Note that most standard activation functions such as ReLU, threshold, sigmoid and general sigmoidal functions satisfy this condition.

Assumption 2

The activation function σ is Lebesgue measurable and satisfies:

$$|\sigma(x)| \leq C(1 + |x|^\alpha)$$

$\forall x \in \mathbb{R}$ and for some constants $C, \alpha > 0$.
This assumption requires mild growth and measurability conditions, which are satisfied by virtually all activation functions in literature.

Main Theorem

Given the two assumptions about our activation function, we can now arrive at our main theorem of the paper: there exists universal constraints $c, C > 0$ such that for every dimension $d > C$, there is a probability measure μ on $\mathbb{R}^d$ and a function $g : \mathbb{R}^d \to \mathbb{R}$ with the following properties:

1. g is bounded in $[-2, 2]$, supported on $\{x : ||x|| \leq C\sqrt{d}\}$ and expressible by a 3-layer network of width $Cc_\sigma d^{19/4}$

2. Every function f expressed by a 2-layer network of width $w \leq ce^{cd}$ satisfies:

$$\mathbb{E}_{x \sim \mu}(f(x) - g(x))^2 \geq c$$

There are two main conclusions drawn from this theorem. The first conclusion (1) roughly says that g can be approximated by a radial function $\tilde{g}$ which depends only on the norm of the input. To clarify, a radial function is defined as $f : \mathbb{R}^d \mapsto \mathbb{R}$ such that $f(x) = f(x')\forall x, x'$ such that $||x|| = ||x'||$ (eg: $f : \mathbb{R}^d \mapsto \mathbb{R}, f(r)$ equals $f(x)$ for any x such that $||x|| = r$). A 3-layer network can approximate $\tilde{g}$ easily by first approximating the squared norm function, then approximating the univariate function on the norm. The second conclusion (2) shows that a 2-layer network would require exponentially many neurons in order to approximate $\tilde{g}$ within constant c accuracy.

Proof of Theorem

A rough overview sketch of this proof goes as following. To prove the first conclusion that we can approximate radial functions with a 3-layer network, we first use assumption 1 to construct a linear combination of neurons to map $x \mapsto x^2$. Using this, we can have our network's second layer compute $x \mapsto ||x||^2 = \sum_i x_i^2$ inside any bounded domain. The third layer can then compute some univariate function of $||x||^2$, with a radial function as our wanted output. More formally, these steps prove the following proposition: Let $C > 0$ be a universal constant such that $\delta \in (0, 1), d \geq C$ and the functions g_i defined as $\tilde{g} = \sum_{i=1}^{N} \epsilon_i g_i$. For any choice $\epsilon_i \in \{-1, 1\}, i = 1, ..., N$, there exists a function g expressible by a 3-layer network of width $w \leq \frac{8c_\sigma}{\delta}\alpha^{3/2}Nd^{11/4} + 1$, within range $[-2, 2]$, such that:

$$\left\| g(x) - \sum_{i=1}^{N} \epsilon_i g_i(||x||) \right\|_{L_2(\mu)} \leq \frac{\sqrt{3}}{\alpha d^{1/4}} + \delta$$

Proof of the second conclusion (2) on 2-layer networks is a bit more involved. We can rewrite our theorem's second condition, $\mathbb{E}_{x\sim\mu}(f(x) - g(x))^2 \geq c$ as such:

$$\begin{aligned} &\int (f(x) - g(x))^2 \varphi^2(x) dx \\ &= \int (f(x)\varphi(x) - g(x)\varphi(x))^2 dx \\ &= ||f\varphi - g\varphi||_{L_2}^2 \\ &= ||\widehat{f\varphi} - \widehat{g\varphi}||_{L_2}^2 \end{aligned}$$

where φ is the inverse Fourier transform of the indicator $1\{x \in B\}$, B being the origin-centered unit volume Euclidean ball, and $\widehat{f\varphi}$ represents a Fourier transformed $f\varphi$, where f is the output of our 2-layer network.

Next, we need to grab the lower bound of our above equation. The convolution-multiplication principle implies $\widehat{f\varphi} = \hat{f} \cdot \hat{\varphi}$, which is the convolution of $\hat{f}$ with the indicator of a unit-volume ball B. Since $\hat{f}$ is supported on $\bigcup) i$ span$\{v_i\}$, we know that:

$$\text{Supp}(\widehat{f\varphi}) \subseteq T := \bigcup_{i=1}^{k} (\text{span}\{v_i\} + B)$$

This means that the support of $\widehat{f\varphi}$ is contained in a union of tubes of bounded radius around the origin. The proof goes on to show that our function g has $\widehat{g\varphi}$ which is constant distance from any function supported on T, thus no 2-layer network can approximate function $\tilde{g}$ without an exponential number of nodes.

Final Comments

I thought this paper was a fascinating way to explore how feedforward neural networks can be approximated and analyzed. It gave some great insights on how restrictions on the activation functions can translate into expressiveness of neural networks. One overarching question I had at the end of this paper was, though our analysis revolves around 2-layer versus 3-layer networks, is there a way to utilize these proofs in order for a more general result, such as how does increasing the depth $d + 1$ affect comparison against a neural network of depth d and width w? Furthermore, beyond the fully connected feedforward networks explored in this paper, can we make any progress towards creating a general systematic way of selecting an optimal network architecture on any test set with no prior assumptions? There's been a lot of other fascinating research around this open ended topic of complexity and expressiveness of learning neural networks, including Telarsky 2016 [2], Nagarajan 2019 [3], Song et al., 2017 [4], and Shamir 2018 [5].

[2] Benefits of depth in neural networks, Telgarsky 2016; https://arxiv.org/abs/1602.04485

[3] Uniform convergence may be unable to explain generalization in deep learning, Nagarajan 2019; http://papers.nips.cc/paper/9336-uniform-convergence-may-be-unable-to-explain-generalization-in-deep-learning.pdf

[4] On the complexity of learning neural networks, Song et al., 2017; https://papers.nips.cc/paper/7135-on-the-complexity-of-learning-neural-networks

[5] Distribution specific hardness of learning neural networks, Shamir 2018; https://arxiv.org/pdf/1609.01037.pdf