Universal Approximation Theorem

Ahmet Yasin Çakmak
cakmaka17@itu.edu.tr

In here we try to understand that how the Neural Networks can be used to approximate any real valued function. This is known as the "Universal Approximation Theorem". We try to understand the underlying structure of this powerful theorem along with its proof. Before we even start, one must define the Neural Networks properly. After that we give some basic concepts and theorems that we will use in the proof of approximation theorem.

Definition 1. *(Neural Network) Let $d, L \in \mathbb{N}$, $N = (N_0, N_\ell, \cdots, N_L) \in \mathbb{N}^{L+1}$ and $\sigma : \mathbb{R} \to \mathbb{R}$. We say that σ is activation function, L is the number of layers, and $N_0, N_L, N_\ell, \ell \in [L-1]$ as number of neurons in the input, output, and ℓ-th hidden layer, respectively. Let $\theta = (\theta^{(\ell)})_\ell$ be the parameters such that,*

$$\theta^{(\ell)} = (W^{(\ell)}, b^{(\ell)})_\ell$$

We denote the number of parameters by $P_N = \sum_{\ell=1}^{L} N_\ell N_{\ell-1} + N_\ell$. Define the corresponding realization function $\Phi_\alpha : \mathbb{R}^{N_0} \times \mathbb{R}^{P_N} \to \mathbb{R}^{N_L}$ which satisfies, for any input $x \in \mathbb{R}^{N_0}$ and parameters θ, we set $\Phi_\alpha = \Phi^{(L)}$ where $\alpha = (N, \sigma)$, then

$$\Phi^{(1)}(x, \theta) = W^{(1)} \cdot x + b^{(1)}$$
$$\bar{\Phi}^{(\ell)}(x, \theta) = \sigma(\Phi^{(\ell)}(x, \theta)), \text{ for } \ell \in [L-1]$$
$$\Phi^{(\ell+1)}(x, \theta) = W^{(\ell+1)} \cdot \bar{\Phi}^{(\ell)}(x, \theta) + b^{(\ell+1)}$$

and σ applied componentwise. We refer W as weight matrix and b as bias vector.

Note that $W^{(\ell)} \in \mathbb{R}^{N_\ell} \times \mathbb{R}^{N_{\ell-1}}$ are matrices, so they represent a linear transformation from $\mathbb{R}^{N_\ell}$ to $\mathbb{R}^{N_{\ell-1}}$. Therefore we may see the Neural Networks as successive composition of affine linear transformations, that is

$$x \mapsto W^{(\ell)} \cdot x + b^{(\ell)}$$

Now for the further usage we will define the set neural networks.

Definition 2. *(Set of Neural Networks) Let $\alpha = (N, \sigma)$ be a Neural Network with input $N_0 = d$ and $N_L = 1$, and activation function σ. The set of Neural Networks is defined by*

$$\mathcal{F}_\alpha = \mathcal{F}_{(N,\sigma)} = \{\Phi_\alpha(\cdot, \theta) : \theta \in \mathbb{R}^{P_N}\}$$

There are simple way to write this two definition, in terms of activation function and affine transformations. Since the realization of Neural Network is given by recursively applying σ we may write as,

$$x \mapsto F(x) := T_L(\sigma(T_{L-1}(\cdots(\sigma(T_1(x))\cdots))))$$

where T_L is the corresponding affine transformation, i.e. $T_\ell = W^{(\ell)} \cdot x + b^{(\ell)}$. Then the Definition 2 basically becomes the set of all functions $F(x)$ of the form wihch is described above, that is $F \in \mathcal{F}_{(N,\sigma)}$. Since we try to approximate functions with the Neural Networks, we need to define a topology, so it is time to develop some tools which we will use later.

Definition 3. *Let $K \subset \mathbb{R}^d$, then $C(K)$ is the set,*

$$C(K) = \{f : K \to \mathbb{R},\ f \textit{ is continious}\}$$

and we equip $C(K)$ with the supremum norm, that is

$$||f||_\infty = \sup_{x \in K} |f(x)|$$

Now note that with this norm $C(K)$ becomes a topological vector space over reals (the sum of two continous function is continous and scalar multiplication of continous function with real number is also continous function). Therefore it is reasonable to talk about the dual space of $C(K)$, we define its dual as the space of all linear maps from $C(K)$ to $\mathbb{R}$,

$$C(K)^* = \{\Psi : C(K) \to \mathbb{R}\}$$

Since we will deal with the compact subspace of $\mathbb{R}^d$, by compact space we mean that the subspace $K \subset \mathbb{R}^d$ which is closed and bounded (this follows from the Heine-Borel theorem). Intiutively it can be thought as a "sphere-like" subspaces of $\mathbb{R}^d$, see [1].

Theorem 4. *(Riesz Representation) If $K \subset \mathbb{R}^d$ is compact, then every linear functional Ψ on $C(K)$ is represented by a unique regular signed Borel measure μ in the sense that,*

$$\Psi \cdot f = \int_K f d\mu$$

Remark: The measure μ can be thing of some function that we used for integration. For example one can think $\mu = g(x)$ for some function g, then the differential $d\mu = g'(x)dx$. Another analogy can be made by using differential forms, that is if μ is any p-form, then $d\mu$ is just an exterior derivative of this

p-form. Even more special case, for the interval $I \subset \mathbb{R}$, this is just a regular expression $d\mu = dx$ from Calculus.

For more general version of this theorem one may see [4, Theorem 6.19] By using this theorem we may identfy the dual space of $C(K)$ with the space of all signed Borel measures, (signed means that measure μ can also be negative), so we may represent every element in $C(K)^*$ by unique measure μ.

Theorem 5. *(Hahn-Banach) Let Y be a linear subspace of normed linear space X, and let $x_0 \in X$. Then x_0 is in the closure $\bar{Y}$ of Y if and only if there is no linear bounded functional f on X such that $f(x) = 0$ for all $x \in Y$ but $f(x_0) \neq 0$.*

This theorem known as Hahn-Banach theorem. We will use this particular form when we prove the universal approximation theorem, for more general version see [4, Theorem 5.19]. We consider its contrapositive when we dealing the proof.

Definition 6. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be continous, $N_0 = d$, $N_L = 1$, $N = (N_0, N_1, \cdots, N_L = 1) \in \mathbb{N}^{L+1}$ and let $K \in \mathbb{R}^d$ be compact. Denote the set of all Neural Network realizations $\bar{\mathcal{F}}$ (wihch is known as "multilayer perceptron") such that,*

$$\bar{\mathcal{F}} := \bigcup_{n \in \mathbb{N}} \mathcal{F}_{(N,\sigma)}$$

We say that $\bar{\mathcal{F}}$ is universal if $\bar{\mathcal{F}}$ is dense in $C(K)$

Note that being dense in some set is closely related to concept of approximation. One particular example is that the set of rational numbers $\mathbb{Q}$ is dense in real numbers $\mathbb{R}$. We know that any real number $x \in \mathbb{R}$ can be approximated with sequence of rational numbers, that is $\lim_{n\to\infty} q_n = x$ for some sequnce $q_i \in \mathbb{Q}$ for all $i \in \mathbb{N}$.

Definition 7. *Let $K \subset \mathbb{R}^d$ is compact. A continous function $f : \mathbb{R} \to \mathbb{R}$ is called discriminatory if the only measure Borel measure μ such that*

$$\int_K f(w \cdot x + b) d\mu = 0$$

for all $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ is $\mu \equiv 0$.

Theorem 8. *(Universal Approximation) Let $d \in \mathbb{N}$, $K \subset \mathbb{R}^d$ be compact and let $\sigma : \mathbb{R} \to \mathbb{R}$ continous discriminatory activation function Let*

$$\bar{\mathcal{F}} := \bigcup_{n \in \mathbb{N}} \mathcal{F}_{((d,n,1),\sigma)}$$

be the set of 2-layer Neural Networks. Then the set $\bar{\mathcal{F}}$ is universal i.e. it is dense in $C(K)$.

Proof. Let us start by observing that $\bar{\mathcal{F}}$ is linear subspace of $C(K)$, because it consist of two layers neural networks with activation function, so we may write as,

$$T_2(\sigma(T_1(W^1 \cdot x + b^1))$$

since σ applied component wise (because σ is a function from $\mathbb{R}$ to $\mathbb{R}$) we basically have composition of two affine transformation which is obviously linear subspace of space of all continous functions. Assume that $\bar{\mathcal{F}}$ is not dense in $C(K)$. Then there exists a function $f \in C(K) \setminus cl(\bar{\mathcal{F}})$. By the theorem 5 (Hahn-Banach) we can say that there is a linear bounded functional $\Psi \not\equiv 0$ on $C(K)^*$ such that, $\Psi \cdot g = 0$ for any $g \in \bar{\mathcal{F}}$. Now take $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$, let $w \cdot x$ be inner product on $\mathbb{R}^d$. Consider the map parametrized by w and b,

$$x \mapsto \sigma_{w,b} := \sigma(w \cdot x + b)$$

for any $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$. Then one can see that $\sigma_{w,b} \in \bar{\mathcal{F}}$. Then we can see that $\Psi \cdot \sigma_{w,b} = 0$ for all w and b. On the other hand by the Riesz Representation theorem we identified the bounded linear functionals with the measures μ. Since $\Psi \not\equiv 0$, there exists a non-zero measure μ such that,

$$\Psi \cdot \sigma_{w,b} = \int_K \sigma_{w,b} d\mu = 0$$

Since μ is non-zero, and σ is discriminatory, this implies that $\mu = 0$ which is a contradiction. Hence the space $\bar{\mathcal{F}}$ is dense in $C(K)$.

□

Let us make some comments about the theorem and its proof. Firstly, note that the Universal Approximation theorem tells us that it is possible to approximate any real-valued function. However the theorem itself does not give any particular method to approximate any kind of function. Secondly, one may notice that the concept of discriminatory functions is important in order the prove theorem. Therefore it is reasonable to ask what kind of functions are discriminatory. Briefly we can answer this question by introducing one particular type of discriminatory function known as sigmoidal function.

Definition 9. *A continous function* $f : \mathbb{R} \to \mathbb{R}$ *such that*

$$\lim_{x \to \infty} f(x) = 1$$

$$\lim_{x \to -\infty} f(x) = 0$$

is called sigmoidal.

It can be proved that every sigmoidal function is discriminatory, by using dominant convergence theorem on a compact set. Idea is showing that the if the integral of f is zero with respect to measure μ, then Fourier coefficents of measure vanished identically, which implies sigmoidal function f is discriminatory. More information and proof can be found on [1] and [2].

References

[**1**] E. Kreyszig, "Introductory Functional Analysis with applications", 1st Ed, John Wiley and Sons, Canada, 1987

[**2**] J. Berner, P. Grohs, G. Kutyniok, and P. Petersen. The modern mathematics of deep learning. arXiv preprint arXiv:2105.04026, 2021

[**3**] P. C. Petersen, Neural Network Theory, Lecture Notes, 2022

[**4**] W. Rudin, Real and complex analysis, McGraw-Hill Series in Higher Mathematics, Tata McGraw-Hill, 2006