On studying neural network expressiveness using topological data analysis and knot theory

Alexandre Louvet

October 26, 2020

Abstract

In this paper we summarize the state of the art on the question of neural network expressiveness both on the theoretical approach to the problem with the study of universal approximators and some practical approaches using topological data analysis and trajectories. We then propose an analysis of the question from a knot theory perspective and share results using studied methods for datasets in dimension 3 and 4.

Contents

1 Neural network expressiveness

1.1 Definition

Let I_n denote the n-dimensional unit cube $[0,1]^n$ and $\mathcal{F}(I_n, \mathbb{R})$ be the space of functions from I_n to $\mathbb{R}$. We want to study the density of the subsets S_f of $\mathcal{F}(I_n, \mathbb{R})$ that can be written as follows:

$$S_f = \{G_N(x) \in \mathcal{F}(I_n, \mathbb{R}) \mid G(x) = \sum_{i=1}^{N} \alpha_i f(y_j^T x + \theta_j)\},\ N \in \mathbb{N}$$

depending on the choice of $f \in \mathcal{F}(\mathbb{R}, \mathbb{R})$. In the previous equation $y_j \in \mathbb{R}^n$ and $\alpha_j, \theta \in \mathbb{R}$, y^T is the transpose of y and $y^T x$ is the inner product of y and x.

The study of neural network expressiveness consists of the problem described above when f is a function used as an activation function for neural network. The study of density can be on the whole set $\mathcal{F}(I_n, \mathbb{R})$ or on subsets of it such as $\mathbb{C}(I_n, \mathbb{R})$ the set of continuous functions from I_n to $\mathbb{R}$.

In particular if S_f is dense in a subset $A \subseteq \mathcal{F}(I_n, \mathbb{R})$ we will say that a single-layer feed-forward neural network (Fig. 1) with f as its activation function is a *universal approximator* of A. Considering a neural network has a finite number of nodes neural network expressiveness also consists of the study of the rate of approach of the approximation, i.e. the sudy of

$$\lim_{N\to\infty} H(N) = \max_{h\in A}(\min_{G_n \in S_f}(\| G_n - h \|)) \text{ with } \| . \| \text{ the cannonical norm on } \mathcal{F}(I_n, \mathbb{R})$$

The study of that limit and especially of its asympatotic approximation gives an idea of the efficiency of the approximator, i.e. the amount of node to add to the network to improve the approximation.

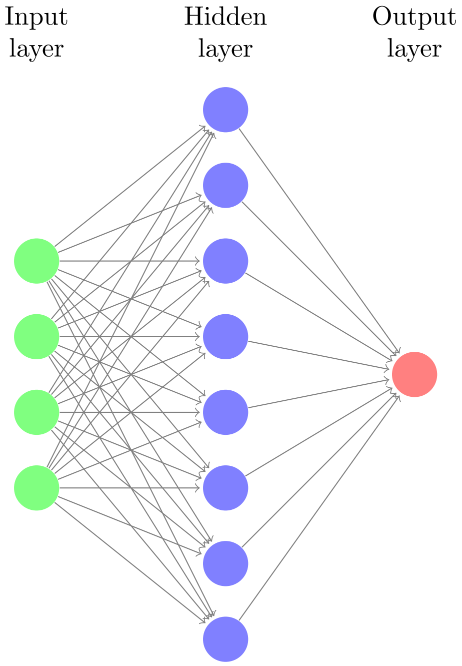

Fig 1: A single-layer feed-forward neural network with $n = 4$ and $N = 8$

1.2 Universal Approximator

In this section we will study the different subsets on which the logistic and ReLU functions acts as universal approximators.

1.2.1 Sigmoidal functions

We say that a function $\sigma \in \mathcal{F}(I_n, \mathbb{R})$ is a sigmoidal function if:

$$\sigma(x) \to \begin{cases} 0 & \text{as } t \to +\infty \\ 1 & \text{as } t \to -\infty \end{cases}$$

The sigmoidal functions include the logistic function defined as:

$$f(x) = \frac{1}{1+e^{-x}}$$

widely used as an activation function for neural networks.

The first study of neural network expressiveness with sigmoidal functions date back to by G.Cybenko in 1989 [1]. He proves that S_σ for σ a sigmoidal function is dense in regards of the supremum norm in $C(I_n, \mathbb{R})$. The demonstration goes as follows.

We denote $M(I_n)$ the space of signed regular Borel measures on I_n

Definition 1 *σ is discriminatory if $\mu \in M(I_n)$ and*

$$\forall y \in \mathbb{R}^n,\ \theta \in \mathbb{R} \int\limits_{I_n} \sigma(y^T x + \theta) d\mu(x) = 0 \implies \mu = 0$$

Theorem 1 *Let σ be a contininuous discriminatory function. Then finite sums of the form*

$$G(x) = \sum_{i=1}^{N} \alpha_i \sigma(y^T x + \theta_i)$$

are dense in $C(I_n, \mathbb{R})$

Proof: Let $S \subset C(I_n)$ be the set of the function of the form $G(x)$. S in a linear subset of $C(I_n)$. Let us show that the closure of S is $C(I_n)$.
Assume it is not the case. Then the closure of S, denoted R, is a proper subspace of $C(I_n)$. Using the Hahn-Banach theorem, there esists L a bounded linear functional on $C(I_n)$ with $L \neq 0$ and $L(R) = L(S) = 0$
Using the Riesz Representation Theorem, we obtain:

$$L(h) = \int\limits_{I_n} h(x) d\mu(x)$$

for some $\mu \in M(I_n)$, for all $h \in C(I_n)$. Since $\sigma(y^T x + \theta_i) \in R$, we have

$$\forall y, \theta \int\limits_{I_n} \sigma(y^T x + \theta) d\mu(x) = 0$$

Since σ is discriminatory, we have $\mu = 0$ and $L = 0$ follows! Hence the closure of S is $C(I_n)$ and by definition S is dense in $C(I_n)$ □

Now let us show that sigmoidal functions are discriminatory.

Lemma 1 *Any bounded, measurable sigmoidal function, a, is discriminatory. In particular, any continuous sigmoidal function is discriminatory.*

Proof: First note $\forall x, y, \theta, \phi$

$$\sigma_\lambda(x) = \sigma(\lambda(y^T x + \theta) + \phi) \begin{cases} \to 1 & \text{for } y^T x + \theta > 0 \text{ as } \lambda \to +\infty \\ \to 0 & \text{for } y^T x + \theta < 0 \text{ as } \lambda \to +\infty \\ = \sigma(\phi) & \text{for } y^T x + \theta = 0 \end{cases}$$

Thus $\sigma_\lambda(x)$ converges pointwise and boundedly to:

$$\gamma(x) \begin{cases} = 1 & \text{for } y^T x + \theta > 0 \\ = 0 & \text{for } y^T x + \theta < 0 \\ = \sigma(\phi) & \text{for } y^T x + \theta = 0 \end{cases}$$

as $\lambda \to +\infty$

Let $\Pi_{y,\theta} = \{x \mid y^T x + \theta = 0\}$ and let $H_{y,\theta} = \{x \mid y^T x + \theta > 0\}$. Lesbegue bounded convergence theorem gives us:

$$\begin{gathered} 0 = \int\limits_{I_n} \sigma_\lambda(x) d\mu(x) \\ = \int\limits_{I_n} \gamma(x) d\mu(x) \\ \sigma(\phi)\mu(\Pi_{y,\theta}) + \mu(H_{y,\theta}) \end{gathered}$$

for all ϕ, θ, y

Fix y, we write

$$F(h) = \int_{I_n} h(y^T x) d\mu(x)$$

Note that F is a bounded function on $L^\infty(\mathbb{R})$ since μ is a signed mesure. By chosing h as the indicator function on $[\theta, \infty[$, we have:

$$F(h) = \int_{I_n} h(y^T x) d\mu(x) = \mu(\Pi_{y,-\theta}) + \mu(H_{y,-\theta}) = 0$$

By linearity $F(h) = 0$ for indicator function on any interval and hence for any simple function (sum of indicator functions) and since simple functions are dense in $L^\infty(\mathbb{R}), F = 0$. In particular it is true for the bounded function $s(u) = sin(m.u)$ and $c(u) = cos(m.u)$. It gives:

$$F(s + ic) = \int_{I_n} cos(m^T x) + isin(m^T x) d\mu(x) = \int_{I_n} exp(im^T x) d\mu(x) = 0$$

for all m. Therefore the fourier transform of μ is 0 and μ must be 0. Hence σ is discriminatory. □

This proves that any function of $C(I_n, \mathbb{R})$ can be approximated by a single-layer network with sigmoidal functions as activation function.

In 1991, Hornik extended in [2] gave the proof for bounded non-constant functions of $F(I_n, \mathbb{R})$. The proof is very similar to the original one.

1.2.2 ReLU functions

ReLU functions stands for Rectified Linear Units, there are formed of two pieces of linear functions. They gained interest recently by showing convincing results in a lot of different applications.

In 2016 Arora et al. [3] showed a version of the universal approximation theorem for piecewise linear function that includes ReLU functions.

Definition 2 *We say a function $f : \mathbb{R}^n \to \mathbb{R}$ is continuous piecewise linear (PWL) if there exists a finite set of polyhedra whose union is $\mathbb{R}^n$, and f is affine linear over each polyhedron (implies continuity because affine regions are closed and cover $\mathbb{R}^n$)*

Proposition 1 *Every function in $L^q(\mathbb{R}^n), (1 \leq q \leq \infty)$ can be arbitrarily approximated in the L^q norm (which for a function f is given by $||f||_q = (\int |f|^q)^{1/q}$) by a ReLU Deep Neural Network.*

Proof:
We know that any ReLU DNN represents a PWL function, let's prove the converse.

By theorem 1 in [4] any piecewise linear function $f : \mathbb{R}^n \to \mathbb{R}$, can be written:

$$f = \sum_{j=1}^{p} s_j (\max_{i \in S_j} l_i)$$

with l_i $(i \leq i \leq k)$ linear functions, S_i $(1 \leq i \leq p) \subseteq \{1, ..., k\}$ with $\forall i, |S_i| \leq n+1$ and $\forall j \in \{1, ..., p\}, s_j \in \{-1, +1\}$

It means that any PWL convex function can be represented as a linear combination of at most $n+1$ affine pieces. That's to say a ReLU DNN with size $n+1$.

Let $p \in \mathbb{N}^*$, let $f \in L^p(\mathbb{R}^n)$, consider the function sequence:

$$f_n(x) = (x - \frac{k}{n})f(\frac{k}{n}) + (1 - x + \frac{k}{n})f(\frac{k+1}{n}) \text{ with } \frac{k}{n} \leq x < \frac{k+1}{n}, n \geq 1$$

$f_n \xrightarrow[n \to \infty]{} f$, and $\forall n, f_n$ is a PWL continous function. Therefore the continuous PWL functions are dense in $L^p(\mathbb{R}^n)$.

Since any PWL function $\mathbb{R}^n \to \mathbb{R}$ is representable by a ReLU DNN, unniversal approximation theorem follows from the fact that the family of continuous PWL functions is dense in any $L^p(\mathbb{R}^n)$ space for $1 \leq p \leq \infty$.

□

In this article, Arora et al. also give a higher bound for the depth of the network. they show that the network required for any function $f \in L^q(\mathbb{R}^n)$ is at most $\lceil log_2(n+1) \rceil$.

In 2018, Hannin & Selke [5] showed that the minimal width a ReLUneural network must have in order to approximate any function $f : \mathbb{R}^n \to \mathbb{R}^m$ is $n+m$. Giving a lower bound to the width of the ReLU network.

2 Homology

2.1 Definition

2.2 S^n as an $n-1$ homology

3 Topological data Analysis

3.1 Fundamental concepts

3.2 Persistent homology

4 Knot theory

4.1 Definition

4.2 Knot determinant

4.2.1 Definition

4.2.2 Algorithms

5 Measuring neural network expressiveness

5.1 Using topological data analysis

5.2 Using trajectories

6 The study of trajectories from a knot theory perspective

6.1 Methodology

6.2 Algorithms

6.3 Results

7 Extending the study of expressiveness with topological data analysis

7.1 Methodology

7.2 Algorithms

7.3 Results

References

[1] G. Cybenko, “Approximation by superpositions of a sigmoidal function,” *Mathematics of Control, Signals, and Systems*, vol. 5, pp. 455–455, Dec. 1992.

[2] K. Hornik, “Approximation capabilities of multilayer feedforward networks,” *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991.

[3] R. Arora, A. Basu, P. Mianjy, and A. Mukherjee, “Understanding Deep Neural Networks with Rectified Linear Units,” *arXiv:1611.01491 [cond-mat, stat]*, Feb. 2018. arXiv: 1611.01491.

[4] S. Wang and X. Sun, “Generalization of Hinging Hyperplanes,” *IEEE Transactions on Information Theory*, vol. 51, pp. 4425–4431, Dec. 2005.

[5] B. Hanin and M. Sellke, “Approximating Continuous Functions by ReLU Nets of Minimal Width,” *arXiv:1710.11278 [cs, math, stat]*, Mar. 2018. arXiv: 1710.11278.