A comment on approximation capacities of neural networks

Florian Rossmannek*

March 2024

Introduction

Sometimes, I encounter results on approximation capacities of Rectified Linear Unit (ReLU)-networks with an accompanying statement that either the results can be shown with analogous arguments for other activation functions or that the results do not work for other activation functions. In this short note, I show that any approximation result concerning ReLU-networks *automatically* translates to σ-networks for most other common activation functions σ. This holds for uniform approximation on compact sets and, hence, for any weaker notion of approximation such as L^p-approximation on compact sets, which appear frequently. This observation implies on the one hand that σ-networks are, in a sense, superior to ReLU-networks in terms of theoretical approximation capacities (of course, this does not take into account generalization properties or suitability for training). On the other hand, when setting out to prove an approximation result for neural networks, it suffices to derive proofs for ReLU-networks since we get the same result for other activation functions for free.

The claimed result follows immediately from the fact that the ReLU function itself can be approximated to arbitrary accuracy on any compact set by a σ-network, whose architecture is independent of the accuracy and the compact set. For some activation functions, such an approximation has been derived in [Mhaskar and Micchelli, 1992] and [Mhaskar, 1993]. However, it seems that this observation and its consequences went by largely unnoticed. In those earlier works, it served as an intermediate step in a longer elaborate proof. In addition, ReLU had not been popular as an activation function at the time.

Result

An architecture $\mathcal{A}$ is a vector of natural numbers, encoding the number of neurons in each layer. The first and the last layer are called input and output layer. Given an activation function $\sigma\colon \mathbb{R} \to \mathbb{R}$, a σ-network with architecture $\mathcal{A} = (a_0, \dots, a_D)$ is a function of the form $A_D \circ \sigma \circ A_{D-1} \circ \sigma \circ \cdots \circ \sigma \circ A_1$, where the $A_n\colon \mathbb{R}^{l_{n-1}} \to \mathbb{R}^{l_n}$ are affine functions and σ is applied element-wise. Let $\mathcal{N}_\sigma(\mathcal{A})$ be the set of all σ-networks with architecture $\mathcal{A}$. The activation functions we consider are of the following type.

Assumption 1. Let $\sigma\colon \mathbb{R} \to \mathbb{R}$ be non-constant and locally Lipschitz continuous, and suppose that (i) σ is bounded and monotone; or (ii) σ satisfies $\lim_{x\to-\infty} x^{-1}\sigma(x) = 0 < \lim_{x\to\infty} x^{-1}\sigma(x) < \infty$.

This covers most common continuous activation functions such as the sigmoid, the hyperbolic tangent, the softplus, the (scaled) Exponential Linear Unit (ELU), the Gaussian Error Linear Unit (GELU), and the Sigmoid Linear Unit (SiLU/swish).

Proposition 2. *Let $\mathcal{A} = (1,1,1)$ if σ is unbounded and $\mathcal{A} = (1,2,2,1)$ if σ is bounded. Then,* ReLU *belongs to the closure of $\mathcal{N}_\sigma(\mathcal{A})$ with respect to the uniform topology.*[1] *In other words, for any compact set $K \subseteq \mathbb{R}$ and any $\varepsilon > 0$ there exists a $\phi \in \mathcal{N}_\sigma(\mathcal{A})$ such that $\sup_{x\in K} |\phi(x) - \mathrm{ReLU}(x)| < \varepsilon$.*

Given a ReLU-network, replace each hidden ReLU-neuron by a σ-neuron if σ is unbounded. The architecture stays the same, only the activation changes. If σ is bounded, replace each hidden neuron in the ReLU-network by four σ-neurons in a $(2,2)$ constellation. The number of hidden layers doubles, and the number of hidden neurons increases 4-fold. Proposition 2 implies that the resulting σ-network is at least as expressive as the original ReLU-network. More formally, we obtain the following result.

*School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore

[1] For unbounded σ, this has been shown implicitly in [Mhaskar and Micchelli, 1992] and [Mhaskar, 1993].

Theorem 3. *Let $f \in C(\mathbb{R}^d, \mathbb{R}^n)$, let $K \subseteq \mathbb{R}^d$ be compact, and let $\varepsilon > 0$. Suppose there exists a* ReLU*-network ϕ such that $\sup_{x\in K} \|f(x) - \phi(x)\| < \varepsilon$. Then, there exists a σ-network ψ such that $\sup_{x\in K} \|f(x) - \psi(x)\| < \varepsilon$. If σ is unbounded, then the architecture of ψ is the same as that of ϕ. If σ is bounded, then the architecture of ψ is obtained by replacing all hidden neurons in the architecture of ϕ by a network of architecture $(2, 2)$.*

While the σ-network in Theorem 3 is four times the size of the ReLU-network in the second case, results on approximation capacities of neural networks typically investigate the size of the architecture as a function of the accuracy (and possibly other parameters) but ignore constant factors, which get absorbed in the $\mathcal{O}$-notation that is commonly used. Thus, from the point of view of approximation capacities, the σ-network and the ReLU-network in Theorem 3 are qualitatively of the same size.

Remark 4. It is straight-forward to extend Theorem 3 from ReLU to any piece-wise linear activation function, in particular leaky ReLU, albeit the factor by which the architecture increases will differ. If σ admits a point of twice differentiability with non-vanishing second derivative, then one can go one step further and replace ReLU with any polynomial spline, in particular with the Rectified Power Unit (RePU).

Proof. Let us prove Proposition 2, of which Theorem 3 is an immediate consequence. Suppose first σ is unbounded. Let $m = \lim_{\to\infty} x^{-1}\sigma(x)$. For any $c > 0$, consider the function $\phi_c(x) = (cm)^{-1}\sigma(cx)$. Let $R > 0$ and $\varepsilon \in (0, 1)$. We show that we can pick c sufficiently large so that $|\phi_c(x) - \mathrm{ReLU}(x)| < \varepsilon$ for all $x \in [-R, R]$. Take $T > |\sigma(0)|$ such that $\left|x^{-1}\sigma(x)\right| < R^{-1}m\varepsilon$ for all $x < -T$ and $\left|x^{-1}\sigma(x) - m\right| < R^{-1}m\varepsilon$ for all $x > T$. Let $\delta < R$ and $c = \delta^{-1}T$. Then, for all x with $|x| \in (\delta, R]$ we have $|cx| > T$ and, hence, $|\phi_c(x) - \mathrm{ReLU}(x)| < \varepsilon$ by choice of T. Next, take $L > 0$ such that σ is L-Lipschitz continuous on $[-T, T]$. Then, ϕ_c is $(m^{-1}L)$-Lipschitz continuous on $[-\delta, \delta]$. Note that $|\phi_c(0)| < m^{-1}\delta$ since $c > \delta^{-1}\,|\sigma(0)|$. Thus, for all $x \in [-\delta, \delta]$,

$$|\phi_c(x) - \mathrm{ReLU}(x)| \leq |\phi_c(x) - \phi_c(0)| + |\phi_c(0)| + \mathrm{ReLU}(x) \leq \frac{L\,|x|}{m} + \frac{\delta}{m} + |x| \leq \frac{L\delta}{m} + \frac{\delta}{m} + \delta.$$

Since L depends only on T, we could have picked δ such that this expression is smaller than ε. This concludes the case of unbounded σ.

Now, suppose σ is bounded and monotone. Since σ is non-constant and locally Lipschitz continuous, there exists a point p at which σ is differentiable with non-vanishing derivative. Note that $\mathcal{N}_\sigma(\mathcal{A}) = \mathcal{N}_\rho(\mathcal{A})$ for any function ρ of the form $\rho(x) = a\sigma(cx + d) + b$ with $a, c \neq 0$. Thus, we may assume without loss of generality that $p = 0$ and that $\sigma(0) = 0$ and $\sigma'(0) = 1$. Denote $\sigma_{\pm\infty} = \lim_{x\to\pm\infty} \sigma(x)$, which satisfy $\sigma_{-\infty} < 0 < \sigma_\infty$. This time, consider $\psi(x) = M(\sigma(cx) - \sigma_\infty)$ and $\phi(x) = \eta^{-1}[\sigma(\sigma(\eta x) + \psi(x)) - \sigma(\psi(x))]$ with $\eta, c, M > 0$ to be specified. Let $R > 1$ and $\varepsilon \in (0, 1)$. Take $L_1 > 1$ such that σ is L_1-Lipschitz continuous on $[-1, 1]$. By Taylor's theorem, we can pick $\eta < (L_1 R)^{-1}$ so small that $|\sigma(x) - x| < \varepsilon\eta/4$ for all $x \in [-\eta L_1 R, \eta L_1 R]$. Next, take $M > 1$ so large that $\sigma(-M\sigma_\infty) - \sigma_{-\infty} < \varepsilon\eta/2$. By monotonicity of σ, this choice of M implies $|\phi(x)| < \varepsilon$ for all $x \leq 0$. If we take $L > L_1$ so that σ is L-Lipschitz continuous on $[\sigma_{-\infty} - M(\sigma_\infty - \sigma_{-\infty}), \sigma_\infty]$, then $|\phi(x)| \leq \eta^{-1}L\,|\sigma(\eta x)|$ for all $x \in \mathbb{R}$. Let $\delta = L^{-2}\varepsilon$. Since $\delta < 1 < L_1 R$, we find $|\sigma(\eta x)| \leq L_1\eta\delta$ for all $x \in [-\delta, \delta]$ and, hence, $|\phi(x)| \leq L_1 L\delta < \varepsilon$ for these x. Finally, pick c so large that $\sigma_\infty - \sigma(c\delta) < (4LM)^{-1}\varepsilon\eta$. Then, $|\psi(x)| < (4L)^{-1}\varepsilon\eta < 1$ for all $x > \delta$. This implies $|\sigma(\psi(x))| < \varepsilon\eta/4$ and $|\sigma(\sigma(\eta x) + \psi(x)) - \sigma(\sigma(\eta x))| \leq \varepsilon\eta/4$ for all $x > \delta$. By choice of η, we have $|\sigma(\eta x) - \eta x| \leq \varepsilon\eta/4$ and $|\sigma(\sigma(\eta x)) - \sigma(\eta x)| \leq \varepsilon\eta/4$ for all $x \in [-R, R]$. Thus, for all $x \in (\delta, R]$,

$$\eta\,|\phi(x) - x| \leq |\sigma(\psi(x))| + |\sigma(\sigma(\eta x) + \psi(x)) - \sigma(\sigma(\eta x))| + |\sigma(\sigma(\eta x)) - \sigma(\eta x)| + |\sigma(\eta x) - \eta x| < \varepsilon\eta.$$

The proof is finished. ∎

Acknowledgments

At the time of the discovery of this result, I was a graduate student at the mathematics department of ETH Zurich, Switzerland, and I was being funded by Swiss National Science Foundation Research Grant 175699.

References

[Mhaskar, 1993] Mhaskar, H. N. (1993). Approximation properties of a multilayered feedforward artificial neural network. *Adv. Comput. Math.*, 1(1):61–80.

[Mhaskar and Micchelli, 1992] Mhaskar, H. N. and Micchelli, C. A. (1992). Approximation by superposition of sigmoidal and radial basis functions. *Adv. in Appl. Math.*, 13(3):350–373.