Research Article

A Global Universality of Two-Layer Neural Networks with ReLU Activations

Naoya Hatano,1,2,3 Masahiro Ikeda,1,2 Isao Ishikawa,2,4 and Yoshihiro Sawano1,2,3

1Department of Mathematics, Faculty of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan
2Center for Advanced Intelligence Project, Riken, Japan
3Department of Mathematics, Chuo University, 1-13-27, Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
4Center for Data Science, Ehime University, 3 Bunkyo-cho, Matsuyama, Ehime 790-8577, Japan

Correspondence should be addressed to Naoya Hatano; n.hatano.chuo@gmail.com

Received 23 November 2020; Accepted 1 February 2021; Published 17 September 2021

Academic Editor: Alberto Fiorenza

Copyright © 2021 Naoya Hatano et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In the present study, we investigate a universality of neural networks, which concerns a density of the set of two-layer neural networks in function spaces. There are many works that handle the convergence over compact sets. In the present paper, we consider a global convergence by introducing a norm suitably, so that our results will be uniform over any compact set.

1. Introduction

Neural network is a function that models a neuron system of a biological brain and is defined as alternate compositions of an affine map and a nonlinear map. The nonlinear map in a neural network is called the activation function. The neural networks have been playing a central role in the field of machine learning with a vast number of applications in the real world in the last decade. We refer to [1] and [2] for example.

We focus on a two-layer feed-forward neural network with ReLU (rectified linear unit) activation, which is a function \( f : \mathbb{R} \rightarrow \mathbb{R} \) of the form of \( f(x) = \sum_{i=1}^{r} c_i \text{ReLU}(a_i x + b_i) \) for some \( a_1, b_1, c_1, \ldots, a_r, b_r, c_r \in \mathbb{R} \). Here, the function ReLU is called the rectified linear unit defined by

\[
\text{ReLU}(x) := \max(x, 0) (x \in \mathbb{R}).
\]  

The ReLU is one of the most popular activation functions for feed-forward neural networks in practical machine learning tasks for real-world problems.

We consider the space of two-layer feedforward neural networks defined by the following linear space

\[
\mathcal{X} = \text{Span}(\{ \text{ReLU}(a \cdot + b) : a \neq 0, b \in \mathbb{R} \})
\]

\[
= \left\{ \sum_{j=1}^{N} \lambda_j \text{ReLU}(a_j \cdot + b_j) : N \in \mathbb{N}, \{ \lambda_j, a_j, b_j \} \right\} \subset \mathbb{C} \times \mathbb{R}^n \times \mathbb{R}^{n+1}
\]  

(2)

Then, it is natural to ask ourselves whether \( \mathcal{X} \) spans a dense subspace of a function space (topological linear space), which is called universality of \( \mathcal{X} \). Historically, the density property of \( \mathcal{X} \) in the space \( C(\mathbb{R}) \) of continuous functions on \( \mathbb{R} \) is investigated by several authors ([3–5]) since it is important to guarantee the existence of a feed-forward neural network \( f \in \mathcal{X} \) that well approximates an unknown continuous function. Here, the topology of \( C(\mathbb{R}) \) is generated by the seminorms \( h \mapsto \sup_{x \in K} |h(x)| \), where \( K \) ranges over all compact sets in \( \mathbb{R} \). Thus, the approximation property of two-layer feed-forward neural networks makes sense only on a local domain.
In this study, we prove an approximation property of \( \mathcal{X} \) in a global sense. More precisely, we prove the space \( \mathcal{X} \) is dense in the Banach subspace of \( C(\mathbb{R}) \) defined as

\[
\mathcal{Y} = \left\{ f \in C(\mathbb{R}) : \lim_{x \to \pm \infty} \frac{f(x)}{1 + |x|} \text{ exists and it is finite} \right\},
\]

equipped with the norm

\[
\|f\|_\mathcal{Y} = \sup_{x \in \mathbb{R}} \frac{|f(x)|}{1 + |x|}.
\]

Note that any element in \( \mathcal{Y} \), divided by \( 1 + |\cdot| \), is a continuous function over \( \mathbb{R} = \mathbb{R} \cup \{ \pm \infty \} \). Our main result in this paper is as follows:

**Theorem 1.** The linear subspace \( \mathcal{X} \) is dense in \( \mathcal{Y} \).

Our main results claim that any function \( f \in \mathcal{X} \) is close to a linear function both at \( \pm \infty \) near the origin, \( \mathcal{X} \) approximates any continuous functions.

Before we conclude this section, we will offer some words on some existing results. See [6] for the \( L^2 \)-approximation over the real line. Other attempts have been made to grasp the neural network by the use of the Radon transform [7] or by considering some other topologies [5, 8].

**2. Proof of the Main Theorem**

**Definition 2.** We define a linear operator \( A : f \in \mathcal{Y} \mapsto f/1 + |\cdot| \in BC(\mathbb{R}) \).

**Lemma 3.** The operator \( A : \mathcal{Y} \mapsto BC(\mathbb{R}) \) is an isomorphism from \( \mathcal{Y} \) to \( BC(\mathbb{R}) \).

A tacit understanding here is that we extend \( f/1 + |\cdot| \), which is initially defined over \( \mathbb{R} \), continuously to \( \mathbb{R} \).

Thus, any continuous functional on \( \mathcal{Y} \) is realized by a Borel measure over \( \mathbb{R} \).

Our theorem can recapture the case where the underlying domain is bounded. Indeed, if the domain \( \Omega \) is contained in \( [-R, R] \) for some \( R > 0 \), then we have

\[
\|f\|_{L^\infty(\Omega)} \leq (1 + R) \|f\|_\mathcal{Y} (f \in \mathcal{Y}),
\]

which will give results by Cybenko [3] and Funahashi [4].

Now we start the proof of Theorem 1. As Cybenko did in [3], take any measure \( \mu \) over \( \mathbb{R} \) such that \( \mu \) annihilates \( \mathcal{X} \). We will show that \( \mu = 0 \). Once this is proved, from the Riesz representation theorem, we conclude that the only linear functional that vanishes on \( \mathcal{X} \) is zero. Using the Hahn-Banach theorem, we see that \( \mathcal{X} \) is dense in \( \mathcal{Y} \).

Remark that

\[
\max (1 - |x - 1|, 0) = \text{ReLU}(x) + \text{ReLU}(x - 2) - 2\text{ReLU}(x - 1) (x \in \mathbb{R}).
\]

Thus, any element in \( C_c(\mathbb{R}) \) can be approximated by a function \( \mathcal{X} \) in the \( L^\infty \)-norm. Since \( \mu \) annihilates \( C_c(\mathbb{R}) \), it follows that \( \mu \) is not supported on \( \mathbb{R} \). Or equivalently, \( \mu \) is supported on \( \pm \infty \). It remains to show that \( \mu(\{\pm \infty\}) = 0 \).

Consider

\[
f(x) = \text{ReLU}(x) - \text{ReLU}(x - 1) (x \in \mathbb{R}).
\]

Remark that

\[
0 = \int_{\mathbb{R}} f(x) d\mu(x) = \mu(\{\infty\}).
\]

Likewise, if we test the condition on \( g = f(-) \), we obtain \( \mu(\{-\infty\}) = 0 \).

Thus, we conclude that \( \mathcal{X} \) is dense in \( \mathcal{Y} \).

**Remark 4.** The set \( \{f, g\} \cup C_c(\mathbb{R}) \) spans a dense subspace in \( \mathcal{Y} \), where \( f \) and \( g \) are functions given in the above proof.

**Data Availability**

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

**Conflicts of Interest**

The authors declare that there are no conflicts of interest regarding the publication of this paper.

**Authors’ Contributions**

The four authors contributed equally to this paper. All of them read the whole manuscript and approved the content of the paper.

**Acknowledgments**

This work was supported by a JST CREST Grant (number JPMJCR1913, Japan). This work was also supported by the RIKEN Junior Research Associate Program. The second author was supported by a Grant-in-Aid for Young Scientists Research (no. 19K03546), Japan Society for the Promotion of Science. The fourth author was supported by a Grant-in-Aid for Scientific Research (C) (19K03546), Japan Society for the Promotion of Science.

**References**

[1] B. Hanin, “Universal function approximation by deep neural nets with bounded width and ReLU activations,” *Mathematics*, vol. 7, no. 10, p. 992, 2019.

[2] A. Pinkus, “Approximation theory of the MLP model in neural networks,” *Acta Numerica*, vol. 8, pp. 143–195, 1999.

[3] G. Cybenko, “Approximation by superpositions of a sigmoidal function,” *Mathematics of control, signals and systems*, vol. 2, no. 4, pp. 303–314, 1989.

[4] K. Funahashi, “On the approximate realization of continuous mappings by neural networks,” *Neural Networks*, vol. 2, no. 3, pp. 183–192, 1989.
[5] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[6] S. Sonoda and N. Murata, "Neural network with unbounded activation functions is universal approximator," *Applied and Computational Harmonic Analysis*, vol. 43, no. 2, pp. 233–268, 2017.

[7] M. Carroll and B. W. Dikinson, "Construction of neural nets using the Radon transform," in *Proceedings of the IEEE 1989 International Joint Conference on Neural Networks*, vol. 1, pp. 607–611, New York, 1989.

[8] Y. T. Sun, A. Gilbert, and A. Tewari, "On the approximation properties of random ReLU features," https://arxiv.org/pdf/1810.04374.pdf.