WIKIPEDIA

Universal approximation theorem

In the mathematical theory of artificial neural networks, **universal approximation theorems** are results[1][2] that establish the density of an algorithmically generated class of functions within a given function space of interest. Typically, these results concern the approximation capabilities of the feedforward architecture on the space of continuous functions between two Euclidean spaces, and the approximation is with respect to the compact convergence topology.

However, there are also a variety of results between non-Euclidean spaces[3] and other commonly used architectures and, more generally, algorithmically generated sets of functions, such as the convolutional neural network (CNN) architecture,[4][5] radial basis-functions,[6] or neural networks with specific properties.[7] Most universal approximation theorems can be parsed into two classes. The first quantifies the approximation capabilities of neural networks with an arbitrary number of artificial neurons ("*arbitrary width*" case) and the second focuses on the case with an arbitrary number of hidden layers, each containing a limited number of artificial neurons ("*arbitrary depth*" case).

Universal approximation theorems imply that neural networks can *represent* a wide variety of interesting functions when given appropriate weights. On the other hand, they typically do not provide a construction for the weights, but merely state that such a construction is possible.

Contents

History

Arbitrary-width case

Arbitrary-depth case

Graph input

Quantum computing

See also

References

History

One of the first versions of the *arbitrary width* case was proven by George Cybenko in 1989 for sigmoid activation functions.[8] Kurt Hornik, Maxwell Stinchcombe, and Halbert White showed in 1989 that multilayer feed-forward networks with as few as one hidden layer are universal approximators.[1] Hornik also showed in 1991[9] that it is not the specific choice of the activation function but rather the multilayer feed-forward architecture itself that gives neural networks the potential of being universal approximators. Moshe Leshno *et al* in 1993[10] and later Allan Pinkus in 1999[11] showed that the universal approximation property is equivalent to having a nonpolynomial activation function.

The *arbitrary depth* case was also studied by a number of authors, such as Zhou Lu *et al* in 2017,[12] Boris Hanin and Mark Sellke in 2018,[13] and Patrick Kidger and Terry Lyons in 2020.[14] The result minimal width per layer was refined in 2020[15][16] for residual networks.

Several extensions of the theorem exist, such as to discontinuous activation functions,[10] noncompact domains,[14] certifiable networks,[17] random neural networks,[18] and alternative network architectures and topologies.[14][19]

Arbitrary-width case

The classical form of the universal approximation theorem for arbitrary width and bounded depth is as follows.[8][9][20][21] It extends[11] the classical results of George Cybenko and Kurt Hornik.

> **Universal approximation theorem:** Let $C(X,Y)$ denote the set of continuous functions from X to Y. Let $\sigma \in C(\mathbb{R},\mathbb{R})$. Note that $(\sigma \circ x)_i = \sigma(x_i)$, so $\sigma \circ x$ denotes σ applied to each component of x.
>
> Then σ is not polynomial if and only if for every $n \in \mathbb{N}$, $m \in \mathbb{N}$, compact $K \subseteq \mathbb{R}^n$, $f \in C(K,\mathbb{R}^m), \varepsilon > 0$ there exist $k \in \mathbb{N}$, $A \in \mathbb{R}^{k\times n}$, $b \in \mathbb{R}^k$, $C \in \mathbb{R}^{m\times k}$ such that
>
> $$\sup_{x\in K} \|f(x) - g(x)\| < \varepsilon$$
>
> where
>
> $$g(x) = C \cdot (\sigma \circ (A \cdot x + b))$$

Such an f can also be approximated by a network of greater depth by using the same construction for the first layer and approximating the identity function with later layers.

Arbitrary-depth case

The 'dual' versions of the theorem consider networks of bounded width and arbitrary depth. A variant of the universal approximation theorem was proved for the arbitrary depth case by Zhou Lu et al. in 2017.[12] They showed that networks of width *n+4* with ReLU activation functions can approximate any Lebesgue integrable function on *n*-dimensional input space with respect to L^1 distance if network depth is allowed to grow. It was also shown that if the width was less than or equal to *n*, this general expressive power to approximate any Lebesgue integrable function was lost. In the same paper[12] it was shown that ReLU networks with width *n+1* were sufficient to approximate any continuous function of *n*-dimensional input variables.[22] The following refinement, specifies the optimal minimum width for which such an approximation is possible and is due to.[23]

> **Universal approximation theorem** (L1 distance, ReLU activation, arbitrary depth, minimal width). For any Bochner–Lebesgue p-integrable function $f : \mathbb{R}^n \to \mathbb{R}^m$ and any $\epsilon > 0$, there exists a fully-connected ReLU network F of width exactly $d_m = \max\{n+1, m\}$, satisfying
>
> $$\int_{\mathbb{R}^n} \|f(x) - F(x)\|^p \mathrm{d}x < \epsilon.$$

Moreover, there exists a function $f \in L^p(\mathbb{R}^n, \mathbb{R}^m)$ and some $\epsilon > 0$, for which there is no fully-connected ReLU network of width less than $d_m = \max\{n+1, m\}$ satisfying the above approximation bound.

Together, the central result of [14] yields the following universal approximation theorem for networks with bounded width.

Universal approximation theorem (non-affine activation, arbitrary depth, constrained width). Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be any non-affine continuous function which is continuously differentiable at at least one point, with nonzero derivative at that point. Let $\mathcal{N}^{\sigma}_{d,D:d+D+2}$ denote the space of feed-forward neural networks with d input neurons, D output neurons, and an arbitrary number of hidden layers each with $d+D+2$ neurons, such that every hidden neuron has activation function σ and every output neuron has the identity as its activation function, with input layer ϕ, and output layer ρ. Then given any $\varepsilon > 0$ and any $f \in C(\mathcal{X}, \mathbb{R}^D)$, there exists $\hat{f} \in \mathcal{N}^{\sigma}_{d,D:d+D+2}$ such that

$$\sup_{x \in \mathcal{X}} \left\| \hat{f}(x) - f(x) \right\| < \varepsilon.$$

In other words, $\mathcal{N}$ is dense in $C(\mathcal{X}; \mathbb{R}^D)$ with respect to the topology of uniform convergence.

Certain necessary conditions for the bounded width, arbitrary depth case have been established, but there is still a gap between the known sufficient and necessary conditions.[12][13][24]

Graph input

Achieving useful universal function approximation on graphs (or rather on graph isomorphism classes) has been a longstanding problem. The popular graph convolutional neural networks (GCNs or GNNs) can be made as discriminative as the Weisfeiler–Leman graph isomorphism test.[25] In 2020,[26] a universal approximation theorem result was established by Brüel-Gabrielsson, showing that graph representation with certain injective properties is sufficient for universal function approximation on bounded graphs and restricted universal function approximation on unbounded graphs, with an accompanying $O(\#\text{edges} \times \#\text{nodes})$-runtime method that performed at state of the art on a collection of benchmarks.

Quantum computing

Quantum neural networks can be expressed by different mathematical tools for circuital quantum computers, ranging from quantum perceptron to variational quantum circuits, both based on combinations of quantum logic gates. Variational quantum circuits are based on a parametric circuit, not involving neural networks. Instead, the quantum perceptron enables the design of quantum neural network with the same structure of feed forward neural networks, provided that the threshold behaviour of each node does not involve the collapse of the quantum state, i.e. no measurement process. In 2022 such measurement-free building block providing the activation function behaviour for quantum neural networks has been designed.[27] The quantum circuit returns arbitrary approximation of squashing functions in the interval from -1 to +1 which is relevant for qubits. Such method to design arbitrary quantum activation functions enables quantum multi-perceptrons and quantum feed-forward neural networks in general.

See also

- Kolmogorov–Arnold representation theorem
- Representer theorem
- No free lunch theorem
- Stone–Weierstrass theorem
- Fourier series

References

1. Hornik, Kurt; Stinchcombe, Maxwell; White, Halbert (1989). *Multilayer Feedforward Networks are Universal Approximators* (http://cognitivemedium.com/magic_paper/assets/Hornik.pdf) (PDF). *Neural Networks*. Vol. 2. Pergamon Press. pp. 359–366.
2. Balázs Csanád Csáji (2001) Approximation with Artificial Neural Networks; Faculty of Sciences; Eötvös Loránd University, Hungary
3. Kratsios, Anastasis; Bilokopytov, Eugene (2020). *Non-Euclidean Universal Approximation* (https://papers.nips.cc/paper/2020/file/786ab8c4d7ee758f80d57e65582e609d-Paper.pdf) (PDF). *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates.
4. Zhou, Ding-Xuan (2020). "Universality of deep convolutional neural networks". *Applied and Computational Harmonic Analysis*. **48** (2): 787–794. arXiv:1805.10769 (https://arxiv.org/abs/1805.10769). doi:10.1016/j.acha.2019.06.004 (https://doi.org/10.1016%2Fj.acha.2019.06.004). S2CID 44113176 (https://api.semanticscholar.org/CorpusID:44113176).
5. Heinecke, Andreas; Ho, Jinn; Hwang, Wen-Liang (2020). "Refinement and Universal Approximation via Sparsely Connected ReLU Convolution Nets". *IEEE Signal Processing Letters*. **27**: 1175–1179. Bibcode:2020ISPL...27.1175H (https://ui.adsabs.harvard.edu/abs/2020ISPL...27.1175H). doi:10.1109/LSP.2020.3005051 (https://doi.org/10.1109%2FLSP.2020.3005051). S2CID 220669183 (https://api.semanticscholar.org/CorpusID:220669183).
6. Park, J.; Sandberg, I. W. (1991). "Universal Approximation Using Radial-Basis-Function Networks". *Neural Computation*. **3** (2): 246–257. doi:10.1162/neco.1991.3.2.246 (https://doi.org/10.1162%2Fneco.1991.3.2.246). PMID 31167308 (https://pubmed.ncbi.nlm.nih.gov/31167308). S2CID 34868087 (https://api.semanticscholar.org/CorpusID:34868087).
7. Yarotsky, Dmitry (2021). "Universal Approximations of Invariant Maps by Neural Networks". *Constructive Approximation*. arXiv:1804.10306 (https://arxiv.org/abs/1804.10306). doi:10.1007/s00365-021-09546-1 (https://doi.org/10.1007%2Fs00365-021-09546-1). S2CID 13745401 (https://api.semanticscholar.org/CorpusID:13745401).
8. Cybenko, G. (1989). "Approximation by superpositions of a sigmoidal function". *Mathematics of Control, Signals, and Systems*. **2** (4): 303–314. CiteSeerX 10.1.1.441.7873 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.441.7873). doi:10.1007/BF02551274 (https://doi.org/10.1007%2FBF02551274). S2CID 3958369 (https://api.semanticscholar.org/CorpusID:3958369).
9. Hornik, Kurt (1991). "Approximation capabilities of multilayer feedforward networks". *Neural Networks*. **4** (2): 251–257. doi:10.1016/0893-6080(91)90009-T (https://doi.org/10.1016%2F0893-6080%2891%2990009-T).
10. Leshno, Moshe; Lin, Vladimir Ya.; Pinkus, Allan; Schocken, Shimon (January 1993). "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function". *Neural Networks*. **6** (6): 861–867. doi:10.1016/S0893-6080(05)80131-5 (https://doi.org/10.1016%2FS0893-6080%2805%2980131-5). S2CID 206089312 (https://api.semanticscholar.org/CorpusID:206089312).

11. Pinkus, Allan (January 1999). "Approximation theory of the MLP model in neural networks". *Acta Numerica*. **8**: 143–195. Bibcode:1999AcNum...8..143P (https://ui.adsabs.harvard.edu/abs/1999AcNum...8..143P). doi:10.1017/S0962492900002919 (https://doi.org/10.1017%2FS0962492900002919).
12. Lu, Zhou; Pu, Homgming; Wang, Feicheng; Hu, Zhiqiang; Wang, Liwei (2017). "The Expressive Power of Neural Networks: A View from the Width" (http://papers.nips.cc/paper/7203-the-expressive-power-of-neural-networks-a-view-from-the-width). *Advances in Neural Information Processing Systems*. Curran Associates. **30**: 6231–6239. arXiv:1709.02540 (https://arxiv.org/abs/1709.02540).
13. Hanin, Boris; Sellke, Mark (March 2019). "Approximating Continuous Functions by ReLU Nets of Minimal Width" (https://doi.org/10.3390%2Fmath7100992). *Mathematics*. MDPI. **7** (10): 992. arXiv:1710.11278 (https://arxiv.org/abs/1710.11278). doi:10.3390/math7100992 (https://doi.org/10.3390%2Fmath7100992).
14. Kidger, Patrick; Lyons, Terry (July 2020). *Universal Approximation with Deep Narrow Networks*. Conference on Learning Theory. arXiv:1905.08539 (https://arxiv.org/abs/1905.08539).
15. Park, Sejun; Yun, Chulhee; Lee, Jaeho; Shin, Jinwoo (October 2020). *Minimum Width for Universal Approximation*. Conference on Learning Theory. arXiv:1905.08539 (https://arxiv.org/abs/1905.08539).
16. Tabuada, Paulo; Gharesifard, Bahman (2020). *Universal Approximation Power of Deep Residual Neural Networks via Nonlinear Control Theory*. ICLR. arXiv:2007.06007 (https://arxiv.org/abs/2007.06007).
17. Baader, Maximilian; Mirman, Matthew; Vechev, Martin (2020). *Universal Approximation with Certified Networks* (https://openreview.net/forum?id=B1gX8kBtPr). ICLR.
18. Gelenbe, Erol; Mao, Zhi Hong; Li, Yan D. (1999). "Function approximation with spiked random networks". *IEEE Transactions on Neural Networks*. **10** (1): 3–9. doi:10.1109/72.737488 (https://doi.org/10.1109%2F72.737488).
19. Lin, Hongzhou; Jegelka, Stefanie (2018). *ResNet with one-neuron hidden layers is a Universal Approximator* (https://papers.nips.cc/paper/7855-resnet-with-one-neuron-hidden-layers-is-a-universal-approximator). *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates. pp. 6169–6178.
20. Haykin, Simon (1998). *Neural Networks: A Comprehensive Foundation*, Volume 2, Prentice Hall. ISBN 0-13-273350-1.
21. Hassoun, M. (1995) *Fundamentals of Artificial Neural Networks* MIT Press, p. 48
22. Hanin, B. (2018). Approximating Continuous Functions by ReLU Nets of Minimal Width. arXiv preprint arXiv:1710.11278.
23. Park, Yun, Lee, Shin, Sejun, Chulhee, Jaeho, Jinwoo (2020-09-28). "Minimum Width for Universal Approximation" (https://openreview.net/forum?id=O-XJwyoIF-k). *ICLR*. arXiv:2006.08859 (https://arxiv.org/abs/2006.08859).
24. Johnson, Jesse (2019). *Deep, Skinny Neural Networks are not Universal Approximators* (https://openreview.net/forum?id=ryGgSsAcFQ). International Conference on Learning Representations.
25. Xu, Keyulu; Hu, Weihua; Leskovec, Jure; Jegelka, Stefanie (2019). *How Powerful are Graph Neural Networks?* (https://openreview.net/forum?id=ryGs6iA5Km). *International Conference on Learning Representations*.
26. Brüel-Gabrielsson, Rickard (2020). *Universal Function Approximation on Graphs* (https://proceedings.neurips.cc//paper/2020/hash/e4acb4c86de9d2d9a41364f93951028d-Abstract.html). *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates.

27. Maronese, Marco; Destri, Claudio; Prati, Enrico (2022). "Quantum activation functions for quantum neural networks" (https://arxiv.org/abs/2201.03700). *Quantum Information Processing*. Springer. **21** (4): 1-24. arXiv:2201.03700 (https://arxiv.org/abs/2201.03700). doi:10.1007/s11128-022-03466-0 (https://doi.org/10.1007%2Fs11128-022-03466-0).