Hypercomplex-valued Neural Networks

Part 3 – Approximation Capability of Vector-Valued Neural Networks

POLISH NATIONAL AGENCY
FOR ACADEMIC EXCHANGE

Marcos Eduardo Valle
Universidade Estadual de Campinas (Unicamp)
Campinas, Brazil.

UNICAMP

Introduction

Neural networks are machine learning models inspired by biological neural networks.

Applications of neural networks include computer vision, physics, control, pattern recognition, economics, and many applications in the medical field.

Vector-valued neural networks (VvNNs) extend traditional real-valued neural network models by assuming the components of the variables are vectors instead of scalars.

Vector-valued NNs are specially designed for tasks involving multi-channel data like color images.

Dense or Fully Connected Layers

This talk addresses the approximation capability of fully connected single-hidden layer VvNNs.

Dense layers, also known as fully connected layers, are the building block of several NN architectures, such as the famous multi-layer perceptron (MLP) network.

Dense layers are composed of several parallel neurons, each receiving inputs through synaptic connections.

Dense layers process data through a linear combination of its inputs by the synaptic weights (trainable parameters), to which a scalar bias term is added.

A non-linear activation function can be applied to yield the neuron's output.

Consider an algebra $\mathbb{V}$. A dense layer is defined as follows.

Let $x_1, \ldots, x_N \in \mathbb{V}$ denote the inputs, the output of the *i*th neuron $y_i \in \mathbb{V}$ in a dense layer is given by

$$y_i = \psi\left(s_j\right), \quad \text{with} \quad s_i = \left(\sum_{j=1}^{N} w_{ij} x_j\right) + b_i \tag{1}$$

where $w_{ij} \in \mathbb{V}$ denotes the weight associated with the *j*th input variable, $b_i \in \mathbb{V}$ is the bias term of the *i*th neuron, and $\psi : \mathbb{V} \to \mathbb{V}$ represents the activation function.

From a computational point of view, dense layers are efficiently implemented using matrix-vector notation.

The output $\boldsymbol{y} = (y_1, \ldots, y_M) \in \mathbb{V}^M$ of a dense layer with M neurons in parallel is determined by the equation

$$\boldsymbol{y} = \boldsymbol{\psi}(\boldsymbol{s}), \quad \text{with} \quad \boldsymbol{s} = \boldsymbol{x}\boldsymbol{W} + \boldsymbol{b}, \tag{2}$$

where $\boldsymbol{x} = (x_1, \ldots, x_N) \in \mathbb{V}^M$ is the input vector, $\boldsymbol{W} = (w_{ij}) \in \mathbb{V}^{N \times M}$ is the matrix containing the synaptic weights, $\boldsymbol{b} = (b_1, \ldots, b_M) \in \mathbb{V}^M$ is the bias vector, and $\boldsymbol{\psi} : \mathbb{V}^M \to \mathbb{V}^M$ is a vector-valued multivariate activation function.

Except for the output layer, the vector-valued activation function $\boldsymbol{\psi}$ is usually defined component-wise:

$$[\boldsymbol{\psi}(\boldsymbol{s})]_i = \psi(s_i), \quad \forall i = 1, \ldots, M, \tag{3}$$

where $\psi : \mathbb{V} \to \mathbb{V}$.

Single Hidden-Layer MLP Network

In mathematical terms, a vector-valued dense layer describes a function $\mathcal{D} : \mathbb{V}^N \to \mathbb{V}^M$.

A vector-valued multilayer perception ($\mathbb{V}$-MLP) $\mathcal{N}$ with K layers is given by the composition of dense layers as follows:

$$\mathcal{N} = \mathcal{D}_K \circ \ldots \circ \mathcal{D}_1. \tag{4}$$

Despite being computationally expensive due to the numerous parameters, dense layers are widely used since they support the universal approximation theorem.

Briefly, the universal approximation theorem says that the set of neural networks is dense in the set of all continuous functions over a compact set.

Approximation Capability of the MLP network

As far as we know, the starting point of the approximation theory for neural networks was the *universal approximation theorem* by Cybenko (1989).

Cybenko's universal approximation theorem was generalized to real-valued MLP models with any non-constant bounded activation function (Hornik, 1991).

Recently, many researchers addressed the approximation capabilities of neural networks, including deep and shallow models based on piece-wise linear activation functions (Petersen and Voigtlaender, 2018).

Theorem 1 (Universal Approximation Theorem (Pinkus, 1999))

Consider a compact $K \subset \mathbb{R}^N$ and let $\psi : \mathbb{R} \to \mathbb{R}$ be a continuous non-polynomial function. The class of all real-valued neural networks defined by

$$\mathcal{H} = \left\{ \mathcal{N}_{\mathbb{R}} = D_2 \circ D_1 : D_1 : \mathbb{R}^N \to \mathbb{R}^Q, D_2 : \mathbb{R}^Q \to \mathbb{R}, Q \in \mathbb{N} \right\}.$$

is dense in $\mathcal{C}(K)$, the set of all real-valued continuous functions on K.

In other words, given a real-valued continuous-function $f_{\mathbb{R}} : K \to \mathbb{R}$ and $\epsilon > 0$, there is a single hidden-layer MLP network given by

$$\mathcal{N}_{\mathbb{R}}(\boldsymbol{x}) = \psi\left(\mathbf{x}\mathbf{W}_1 + \boldsymbol{b}_1\right)\boldsymbol{w}_2 + b_2,$$

such that

$$|f_{\mathbb{R}}(\boldsymbol{x}) - \mathcal{N}_{\mathbb{R}}(\boldsymbol{x})| < \epsilon, \quad \forall \boldsymbol{x} \in K.$$

Complex-valued MLP ($\mathbb{C}$-MLP)

The structure of a complex-valued MLP ($\mathbb{C}$-MLP) is equivalent to that of a real-valued MLP, except that input and output signals, weights, and biases are complex numbers instead of real values.

Additionally, the activation functions are complex-valued functions $\psi : \mathbb{C} \to \mathbb{C}$.

Note that the logistic function can be generalized to complex parameters using Euler's formula $e^{x\boldsymbol{i}} = \cos(x) + \boldsymbol{i}\sin(x)$ as follows for all $x \in \mathbb{C}$:

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

However, the universal approximation property is generally not valid for $\mathbb{C}$-MLP with such activation function(Arena et al., 1998).

Fortunately, the universal approximation theorem holds for $\mathbb{C}$-MLP networks with split activation functions(Arena et al., 1998).

A complex-valued split activation function $\psi_{\mathbb{C}} : \mathbb{C} \to \mathbb{C}$ is defined as follows using a real-valued function $\psi : \mathbb{R} \to \mathbb{R}$:

$$\psi(x) = \psi(x_0) + \psi(x_1)\boldsymbol{i}, \quad \forall x = x_0 + x_1\boldsymbol{i} \in \mathbb{C}.$$

Quaternion-Valued MLPs ($\mathbb{Q}$-MLP)

In a similar way, Arena et al. (1997) also defined quaternion-valued MLP ($\mathbb{Q}$-MLP) by replacing the real input and output, weights and biases, with quaternion numbers.

They then proceeded to prove that $\mathbb{Q}$-MLPs with a single hidden layer and split sigmoid activation function

$$\psi(x) = \psi(x_0) + \psi(x_1)\boldsymbol{i} + \psi(x_2)\boldsymbol{j} + \psi(x_3)\boldsymbol{k}, \quad \forall x = x_0 + x_1\boldsymbol{i} + x_2\boldsymbol{j} + x_3\boldsymbol{k} \in \mathbb{Q}, \tag{5}$$

are universal approximators in the set of continuous quaternion-valued functions.

Hyperbolic-valued MLPs ($\mathbb{U}$-MLP)

In 2000, Buchholz and Sommer introduced an MLP based on hyperbolic numbers, obtaining the so-called hyperbolic multilayer perceptron ($\mathbb{U}-$MLP).

This network equipped with a split logistic activation function is also a universal approximator (Buchholz and Sommer, 2000).

Buchholz and Sommer provided experiments highlighting that the $\mathbb{U}$-MLP can learn tasks with underlying hyperbolic properties much more accurately and efficiently than $\mathbb{C}$-MLP and real-valued MLP networks.

Clifford MLPs ($\mathcal{C}\ell$-MLP)

In 2001, Buchholz and Sommer worked with a class of neural networks based on Clifford algebras (Buchholz and Sommer, 2001, 2008).

They found that the universal approximation property holds for MLPs based on non-degenerate Clifford algebra.

In addition, they pointed out that degenerate Clifford algebras may lead to models without universal approximation capability.

Although Buchholz and Sommer considered sigmoid activation functions, Clifford MLPs are also universal approximators with the split `relu` activation function.

Univeral Approximation Theorem for $\mathbb{V}$-MLPs

In the previous pages, we presented universal approximation theorems from the literature for some single hidden layer $\mathbb{V}$-MLP networks.

In the following, we generalize the universal approximation theorem for a finite-dimensional non-degenerate algebra $\mathbb{V}$.

Let $\mathcal{E} = \{e_1, \ldots, e_n\}$ be a basis for an algebra $\mathbb{V}$. A split activation function $\psi : \mathbb{V} \to \mathbb{V}$ is derived from a real function $\psi_{\mathbb{R}} : \mathbb{R} \to \mathbb{R}$ as follows:

$$\psi(x) = \sum_{i=1}^{n} \psi_{\mathbb{R}}(x_i)e_i, \quad \forall x = \sum_{i=1}^{n} x_i e_i \in \mathbb{V}.$$

The activation function can be the split `relu` or a split sigmoid function.

Theorem 2 (Universal Approx. Theorem for $\mathbb{V}$-MLP Networks)

Consider a finite-dimensional non-degenerate algebra $\mathbb{V}$ and let $K \subset \mathbb{V}^N$ be a compact set. Also, consider a continuous real-valued non-polynomial function $\psi_{\mathbb{R}} : \mathbb{R} \to \mathbb{R}$ such that $\lim_{\lambda \to -\infty} \psi_{\mathbb{R}}(\lambda) = 0$ and let $\psi_{\mathbb{V}} : \mathbb{V}$ be the split function derived from $\psi_{\mathbb{R}}$. The class

$$\mathcal{H} = \left\{ \mathcal{N}_{\mathbb{V}} = D_2 \circ D_1 : D_1 : \mathbb{V}^N \to \mathbb{V}^Q, D_2 : \mathbb{V}^Q \to \mathbb{V}, Q \in \mathbb{N} \right\}.$$

is dense in the set $\mathcal{C}(K)$ of all continuous functions on K. In other words, given a continuous function $f_{\mathbb{V}} : K \to \mathbb{V}$ and $\epsilon > 0$, there exists a $\mathbb{V}$-MLP network $\mathcal{N}_{\mathbb{V}} : \mathbb{V}^N \to \mathbb{V}$ given by

$$\mathcal{N}_{\mathbb{V}}(\boldsymbol{x}) = \psi\left(\mathbf{x}\mathbf{W}_1 + \boldsymbol{b}_1\right)\mathbf{w}_2 + b_2,$$

such that

$$|f_{\mathbb{V}}(\boldsymbol{x}) - \mathcal{N}_{\mathbb{V}}(\boldsymbol{x})| < \epsilon, \quad \forall \boldsymbol{x} \in K. \tag{6}$$

Concluding Remarks

The universal approximation theorem for single-hidden layer $\mathbb{V}$-MLP networks serves many purposes, including:

1. It consolidates the results regarding the universal approximation property of many well-known algebras, thus eliminating the need to prove this property for each algebra individually.
2. Many algebras that have not had this result proven are now directly known as the basis for neural networks with universal approximation property. That is the case for octonions, for example.
3. This result further promotes the use of vector-valued networks.

Now, we are studying the approximation capabilities of vector-valued dense networks like the ResNet (Lin and Jegelka, 2018).

Thanks for your attention!

Acknowledge

These slides are part of a mini-course given during the workshop on **hypercomplex-valued neural networks**, which took place at the **Institute for Research and Applications of Fuzzy Modeling**, **University of Ostrava**, Ostrava, Checz Republic, *06-10 February 2023*, with the support of

NAWA POLISH NATIONAL AGENCY
FOR ACADEMIC EXCHANGE

My research on hypercomplex-valued neural networks has been partially supported by:

FAPESP and CNPq
Conselho Nacional de Desenvolvimento
Científico e Tecnológico

References (1)

P. Arena, L. Fortuna, G. Muscato, and M. G. Xibilia. Multilayer perceptrons to approximate quaternion valued functions. *Neural Networks*, 10(2):335–342, 3 1997. doi: 10.1016/S0893-6080(96)00048-2.

P. Arena, L. Fortuna, G. Muscato, and M. G. Xibilia. *Neural Networks in Multidimensional Domains*. Springer-Verlag, 1998. doi: 10.1007/BFB0047683.

S. Buchholz and G. Sommer. Hyperbolic Multilayer Perceptron. *Proceedings of the International Joint Conference on Neural Networks*, 2:129–133, 2000. doi: 10.1109/IJCNN.2000.857886.

S. Buchholz and G. Sommer. Clifford Algebra Multilayer Perceptrons. In *Geometric Computing with Clifford Algebras*, pages 315–334. Springer Berlin Heidelberg, 2001. doi: 10.1007/978-3-662-04621-0{_}13.

References (2)

S. Buchholz and G. Sommer. On Clifford neurons and Clifford multi-layer perceptrons. *Neural Networks*, 21(7):925–935, 9 2008. ISSN 08936080. doi: 10.1016/j.neunet.2008.03.004.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems 1989 2:4*, 2(4):303–314, 12 1989. ISSN 1435-568X. doi: 10.1007/BF02551274. URL `https://link.springer.com/article/10.1007/BF02551274`.

K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1 1991. ISSN 08936080. doi: 10.1016/0893-6080(91)90009-T.

H. Lin and S. Jegelka. ResNet with one-neuron hidden layers is a Universal Approximator. 6 2018.

References (3)

P. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330, 12 2018. ISSN 0893-6080. doi: 10.1016/J.NEUNET.2018.08.019.

A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1 1999. ISSN 0962-4929. doi: 10.1017/S0962492900002919.