East Asian Journal on Applied Mathematics
doi: 10.4208/eajam.2022-270.070123

Vol. **13**, No. 3, pp. 524-549
August 2023

Approximation Analysis of Convolutional Neural Networks

Chenglong Bao[1,2,*], Qianxiao Li[3,4], Zuowei Shen[3], Cheng Tai[5], Lei Wu[6] and Xueshuang Xiang[7]

[1]*Yau Mathematical Sciences Center, JingZhai Tsinghua university, Beijing 100084, China.*
[2]*YanQi Lake Beijing Institute of Mathematical Sciences and Applications, No. 544, HeFangkou Village, Huaibei Town, Beijing 101408, China.*
[3]*Department of Mathematics, National University of Singapore, BLK S17, 10 Lower Kent Ridge Road, 119076 Singapore.*
[4]*Institute for Functional Intelligent Materials, National University of Singapore, BLK S9, 4 Science Drive 2, 117544 Singapore.*
[5]*Beijing Institute of Big Data Research, Peking University, No.6 Jingyuan, Beijing 100871, China.*
[6]*School of Mathematical Sciences, Peking University, Beijing 100871, China.*
[7]*Qian Xuesen Laboratory of Space Technology, No. 104, Youyi Road, Beijing 100094, China.*

Received 2 October 2022; Accepted (in revised version) 7 January 2023.

Dedicated to Professor Tao Tang on the occasion of his 60th birthday.

Abstract. In its simplest form, convolution neural networks (CNNs) consist of a fully connected two-layer network g composed with a sequence of convolution layers T. Although g is known to have the universal approximation property, it is not known if CNNs, which have the form $g \circ T$ inherit this property, especially when the kernel size in T is small. In this paper, we show that under suitable conditions, CNNs do inherit the universal approximation property and its sample complexity can be characterized. In addition, we discuss concretely how the nonlinearity of T can improve the approximation power. Finally, we show that when the target function class has a certain compositional form, convolutional networks are far more advantageous compared with fully connected networks, in terms of the number of parameters needed to achieve the desired accuracy.

AMS subject classifications: 41A63, 68T01

Key words: Convolutional networks, approximation, scaling analysis, compositional functions.

*Corresponding author. *Email addresses:* clbao@mail.tsinghua.edu.cn (C. Bao), qianxiao@nus.edu.sg (Q. Li), matzuows@nus.edu.sg (Z. Shen), chengtai@pku.edu.cn (C. Tai), leiwu@pku.edu.cn (L. Wu), xiangxueshuang@qxslab.cn (X. Xiang)

1. Introduction

Over the past decade, convolutional neural networks (CNNs) have played important roles in many applications, including facial recognition, autonomous driving, and disease diagnosis. Such applications typically involve approximating some oracle f^*, which can be a classifier or regressor, by some f chosen from an appropriate model or hypothesis space. In other words, learning involves minimizing the distance between f^* and f over its hypothesis space.

Unlike plain fully connected neural networks, convolution neural networks are of the form $f = g \circ T$ where $g \in \mathcal{G}$ is a fully connected classification/regression layer and $T \in \mathcal{T}$ is a feature extractor typically composed of interfacing convolutions and nonlinear activations. From the approximation theory viewpoint, one important direction of investigation is the universal approximation property (UAP), namely whether $\{g \circ T : g \in \mathcal{G}, T \in \mathcal{T}\}$ can approximate arbitrary continuous functions on compact domains. The UAP is known to hold in the case of one-hidden-layer, fully connected neural networks for a large class of activation functions [1, 10, 18]. However, for the CNN architecture this is less obvious, even if a fully-connected layer g is present. This is especially so if T consists of convolutions of small filter sizes or the output dimension of T is small, which leads to a loss of information. For example, for classification problems if T maps two samples belong to two different classes into the same feature representation, then it is obvious that no matter what the approximation power of g is, $g \circ T$ cannot correctly classify them. Hence, the first goal of this paper is to show that we can in fact construct CNNs which ensures that when composed with g, forms a universal approximator for classification problems. The key is showing that the convolution-based feature extractors can satisfy the so-called separable condition [23], i.e.

$$|T(x_i) - T(x_j)| > c, \quad \forall x_i \in \Omega_i, \quad x_j \in \Omega_j, \quad i \neq j \tag{1.1}$$

for some positive constant c. Here, Ω_i represents the set of samples belonging to the i-th class. Recall that due to small filter sizes and possible dimensional reduction, the satisfaction of this condition for convolution layers is not immediate and the first goal of this paper is to construct convolution feature extractors that satisfy (1.1) under appropriate sparsity assumptions on the input data, which then allows us to show that a class of practical CNN architectures satisfy the universal approximation property.

Besides the convolutional structure, another important component in CNNs is the nonlinear activation function. Commonly used non-linear functions include sigmoid, tanh, and (Leaky) ReLU. These activation functions introduce non-linearity into neural networks and greatly expand their approximation capabilities. In the preceding UAP analysis of CNNs, the effect of non-linearity was not explicitly studied. In fact, in the literature there generally lacks concrete analysis of the advantage of non-linearity, besides general statements such as having a bigger approximation space. In the second part of this paper, we concretely investigate the effect of nonlinear functions in terms of the approximation power by showing that a composition function approximator with non-linear structure can locally improve its approximation, which is not the case for its linear counterpart. More specifically, we estab-