Commun. Comput. Phys.
doi: 10.4208/cicp.OA-2020-0214

Vol. **28**, No. 5, pp. 1838-1885
November 2020

Butterfly-Net: Optimal Function Representation Based on Convolutional Neural Networks

Yingzhou Li[1], Xiuyuan Cheng[1,*] and Jianfeng Lu[1,2]

[1] *Department of Mathematics, Duke University, Durham, NC 27708, USA.*
[2] *Department of Chemistry and Department of Physics, Duke University, Durham, NC 27708, USA.*

Received 30 October 2020; Accepted 6 November 2020

Abstract. Deep networks, especially convolutional neural networks (CNNs), have been successfully applied in various areas of machine learning as well as to challenging problems in other scientific and engineering fields. This paper introduces *Butterfly-net*, a low-complexity CNN with structured and sparse cross-channel connections, together with a *Butterfly* initialization strategy for a family of networks. Theoretical analysis of the approximation power of *Butterfly-net* to the Fourier representation of input data shows that the error decays exponentially as the depth increases. Combining *Butterfly-net* with a fully connected neural network, a large class of problems are proved to be well approximated with network complexity depending on the effective frequency bandwidth instead of the input dimension. Regular CNN is covered as a special case in our analysis. Numerical experiments validate the analytical results on the approximation of Fourier kernels and energy functionals of Poisson's equations. Moreover, all experiments support that training from *Butterfly* initialization outperforms training from random initialization. Also, adding the remaining cross-channel connections, although significantly increases the parameter number, does not much improve the post-training accuracy and is more sensitive to data distribution.

AMS subject classifications: 15A23, 65D05, 65F10, 62G08, 68W20, 68W25

Key words: Butterfly algorithm, convolutional neural network, Fourier analysis, deep learning.

1 Introduction

Deep neural network is a central tool in machine learning and data analysis nowadays [5]. In particular, convolutional neural network (CNN) has been proved to be a powerful tool

*Corresponding author. *Email addresses:* yingzhou.li@duke.edu (Y. Li), xiuyuan.cheng@duke.edu (X. Cheng), jianfeng@math.duke.edu (J. Lu)

in image recognition and representation. Deep learning has also emerged to be successfully applied in solving PDEs [6, 28, 37] and physics problems [4, 17, 35, 47, 53], showing the potential of becoming a tool of great use for computational mathematics and physics as well. Given the wide application of PDEs and wavelet based methods in image and signal processing [8, 11, 39], an understanding of CNN's ability to approximate differential and integral operators will lead to an explanation of CNN's success in these fields, as well as possible improved network architectures.

The remarkable performance of deep neural networks across various fields relies on their ability to accurately represent functions of high-dimensional input data. Approximation analysis has been a central topic to the understanding of the neural networks. The classical theory developed in 80's and early 90's [3, 13, 26] approximates a target function by a linear combination of sigmoids, which is equivalent to a fully connected neural network with one hidden layer. While universal approximation theorems were established for such shallow networks, the research interest in neural networks only revived in recent years after observing the successful applications of deep neural networks, particular the superior performance of CNNs in image and signal processing.

Motivated by the empirical success, the approximation advantage of deep neural networks over shallow ones has been theoretically analyzed in several places. However, most results assume stacked fully connected layers and do not apply to CNNs which have specific geometrical constraints: (1) the convolutional scheme, namely local-supported filters with weight sharing, and (2) the hierarchical multi-scale architecture. The approximation power of deep networks with hierarchical geometrically-constrained structure has been studied recently [12,40,41], yet the network architecture differ from the regular CNN. The approximation theory of CNN has been studied in [2, 54]. We review the related literature in more detail below.

This paper proposes a specific architecture under the CNN framework based on the *Butterfly* scheme originally developed for the fast computation of special function transforms [42, 44, 52] and Fourier integral operators [9, 10, 31–34]. *Butterfly* scheme provides a hierarchical structure with locally low-rank interpolation of kernel functions and can be applied to solve many PDE related problems. In terms of computational complexity, the scheme is near optimal for Fourier kernels and Fourier integral operators. The proposed *Butterfly-net* explicitly adopts the hierarchical structure in *Butterfly* scheme as the stacked convolutional layers. If the parameters are hard-coded as that in the *Butterfly* scheme (*Butterfly* initialization), then *Butterfly-net* collectively computes the Fourier coefficients of the input signal with guaranteed numerical accuracy. Unlike regular CNN which has dense cross-channel connections, the channels in the *Butterfly-net* have clear correspondences to the frequency bands, namely the position in the spectral representation of the signal, and meanwhile, the cross-channel weights are sparsely connected. In this paper, we also study *Butterfly-net* with dense cross-channel connections, which is named *Inflated-Butterfly-net*. Regular CNN is a special *Inflated-Butterfly-net* [49]. Comparing *Butterfly-net* and *Inflated-Butterfly-net*, *Butterfly-net* is much lighter: the model complexity (in terms of parameter number) is $\mathcal{O}(K\log N)$ and computational complexity is