CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.2020-0018

Vol. **1**, No. 3, pp. 518-529
September 2020

Hartley Spectral Pooling for Deep Learning

Hao Zhang[1] and Jianwei Ma[1,2,*]

[1] *Department of Mathematics and Artificial Intelligence Laboratory, Harbin Institute of Technology, Harbin 150001, China.*
[2] *School of Earth and Space Sciences, Peking University, Beijing 100000, China.*

Received 5 May 2020; Accepted 12 September 2020

Abstract. In most convolution neural networks (CNNs), downsampling hidden layers is adopted for increasing computation efficiency and the receptive field size. Such operation is commonly called pooling. Maximization and averaging over sliding windows (*max/average pooling*), and plain downsampling in the form of strided convolution are popular pooling methods. Since the pooling is a lossy procedure, a motivation of our work is to design a new pooling approach for less lossy in the dimensionality reduction. Inspired by the spectral pooling proposed by Rippel et al. [1], we present the Hartley transform based spectral pooling method. The proposed spectral pooling avoids the use of complex arithmetic for frequency representation, in comparison with Fourier pooling. The new approach preserves more structure features for network's discriminability than max and average pooling. We empirically show the Hartley pooling gives rise to the convergence of training CNNs on MNIST and CIFAR-10 datasets.

AMS subject classifications: 68T07

Key words: Hartley transform, spectral pooling, deep learning.

1 Introduction

Convolutional neural networks(CNNs) [2–4] have been dominant machine learning approach for computer vision, and have spreaded out in many other fields. The modern framework of CNNs was established by LeCun et al. [5] in 1990, with three main components: convolution, pooling, and activation. Pooling is an important component of CNNs. Even before the resuscitation of CNNs, pooling was utilized to extract features to gain dimension-reduced feature vectors and acquire the invariance to small transformations of the input. This is motivated by the seminal work about complex cells in animal visual cortex by Hubel and Wiesel [6].

*Corresponding author. *Email addresses:* jwm@pku.edu.cn (J. Ma), hao.zhang.hit@stu.hit.edu.cn (H. Zhang)

Pooling is of crucial for reducing computation cost, improving some amount of translation invariance and increasing the receptive field of neural networks. Numerous variants of pooling processes are proposed for classification accuracy improvement. These variants are mainly casted in four major categories based on value, rank, probability and transformed domain pooling methods, which are thoroughly reviewed recently in [32]. In shallow or mid-sized networks, max or average pooling are most widely used such as in AlexNet [7], VGG [8], and GoogleNet [9]. After Springenberg et al. [10] empirically revealed that strided convolution could replace pooling without loss of accuracy in classification task, deeper networks always use strided convolution for architecture-design simplicity. The most markable one of those exemplars is ResNet [11]. However, most methods present a number of issues. For example, max pooling implies an amazing by-product of discarding at least 75% of data. It can overfit the training data and does not guarantee generalization on test data. The maximum value picked out in each window only reflects very rough information. Average pooling, stretching to the opposite end, results in a gradual, constant attenuation of the contribution of individual grid in each window, and ignores the importance of local structure. These two poolings both suffer from sharp dimensionality reduction and lead to implausible looking results (see the first and second row in Fig. 1). Strided convolution may cause aliasing since it simply picks one node in a fixed position in each local window [12], regardless of the significance of its activation.

There have been a few attempts to mitigate the harmful effects of max and average pooling, such as a linear combination and extension of them [13], and nonlinear pooling layers [14, 15]. In most of the common implementations, max or average related pooling layers directly downscale the spatial dimension of feature maps by a factor. L_p pooling [15] provides better generalization than max pooling, with $p=1$ corresponding to average pooling and $p=\infty$ reducing to max pooling. Yu et al. [16] proposed the mixed pooling, which combines max pooling and average pooling and switches between these two pooling methods randomly. Instead of picking the maximum values within each pooling region, stochastic pooling [17] and S3Pool [18] stochastically pick a node in a window, and the former favors strong activations. In some networks, stride convolutions are also used for pooling. Notably, these pooling methods are all of integer stride larger than 1. To abate the loss of information caused by the dramatic dimension reduction, fractional max-pooling [19] randomly generates pooling region with stride 1 or 2 to achieve pooling stride of less than 2. There are also some other pooling methods by applying/learning filters. For example, detail preserving pooling [33] uses inverse bilateral filter and learns two reward parameters in inverse bilateral weights from data to adaptively preserve the important details of feature maps. LEAP pooling [34] learns a shared linear filter over feature maps and aggregates the features within pooling region. We refer these pooling methods mentioned above to as *spatial pooling*.

In 2015, Rippel et al. [1] proposed the *spectral pooling*, which downsamples the feature maps in frequency domain using low-pass filtering. It selects pooling region in Fourier based frequency domain by extracting low frequency subset. This approach can allevi-