

Accepted Manuscript

Improving efficiency in convolutional neural networks with multilinear filters

Dat Thanh Tran, Alexandros Iosifidis, Moncef Gabbouj

PII: S0893-6080(18)30179-5
DOI: <https://doi.org/10.1016/j.neunet.2018.05.017>
Reference: NN 3963

To appear in: *Neural Networks*

Received date : 23 October 2017
Revised date : 8 May 2018
Accepted date : 25 May 2018

Please cite this article as: Tran, D.T., Iosifidis, A., Gabbouj, M., Improving efficiency in convolutional neural networks with multilinear filters. *Neural Networks* (2018), <https://doi.org/10.1016/j.neunet.2018.05.017>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



Improving Efficiency in Convolutional Neural Networks with Multilinear Filters

Dat Thanh Tran^{1*}, Alexandros Iosifidis², Moncef Gabbouj¹

¹*Laboratory of Signal Processing, Tampere University of Technology, Finland*

²*Department of Engineering, Electrical and Computer Engineering, Aarhus University*

Abstract

The excellent performance of deep neural networks has enabled us to solve several automatization problems, opening an era of autonomous devices. However, current deep net architectures are heavy with millions of parameters and require billions of floating point operations. Several works have been developed to compress a pre-trained deep network to reduce memory footprint and, possibly, computation. Instead of compressing a pre-trained network, in this work, we propose a generic neural network layer structure employing multilinear projection as the primary feature extractor. The proposed architecture requires several times less memory as compared to the traditional Convolutional Neural Networks (CNN), while inherits the similar design principles of a CNN. In addition, the proposed architecture is equipped with two computation schemes that enable computation reduction or scalability. Experimental results show the effectiveness of our compact projection that outperforms traditional CNN, while requiring far fewer parameters.

Keywords: Convolutional Neural Networks, Multilinear Projection, Network Compression

1. Introduction

In recent years, deep neural network architectures have excelled in several application domains, ranging from machine vision [1, 2, 3], natural language processing [4, 5] to biomedical [6, 7] and financial data analysis [8, 9]. Of those

*Corresponding author: Tel.: +358 401592373

Email address: dat.tranthanh@tut.fi (Dat Thanh Tran¹)

Download English Version:

<https://daneshyari.com/en/article/6862882>

Download Persian Version:

<https://daneshyari.com/article/6862882>

[Daneshyari.com](https://daneshyari.com)