ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices

Xiangyu Zhang* Xinyu Zhou* Mengxiao Lin Jian Sun
Megvii Inc (Face++)
{zhangxiangyu,zxy,linmengxiao,sunjian}@megvii.com

Abstract

We introduce an extremely computation-efficient CNN architecture named ShuffleNet, which is designed specially for mobile devices with very limited computing power (e.g., 10-150 MFLOPs). The new architecture utilizes two new operations, pointwise group convolution and channel shuffle, to greatly reduce computation cost while maintaining accuracy. Experiments on ImageNet classification and MS COCO object detection demonstrate the superior performance of ShuffleNet over other structures, e.g. lower top-1 error (absolute 7.8%) than recent MobileNet [12] on ImageNet classification task, under the computation budget of 40 MFLOPs. On an ARM-based mobile device, ShuffleNet achieves ∼13× actual speedup over AlexNet while maintaining comparable accuracy.

1. Introduction

Building deeper and larger convolutional neural networks (CNNs) is a primary trend for solving major visual recognition tasks [21, 9, 33, 5, 28, 24]. The most accurate CNNs usually have hundreds of layers and thousands of channels [9, 34, 32, 40], thus requiring computation at billions of FLOPs. This report examines the opposite extreme: pursuing the best accuracy in very limited computational budgets at tens or hundreds of MFLOPs, focusing on common mobile platforms such as drones, robots, and smartphones. Note that many existing works [16, 22, 43, 42, 38, 27] focus on pruning, compressing, or low-bit representing a "basic" network architecture. Here we aim to explore a highly efficient basic architecture specially designed for our desired computing ranges.

We notice that state-of-the-art basic architectures such as *Xception* [3] and *ResNeXt* [40] become less efficient in extremely small networks because of the costly dense 1×1 convolutions. We propose using *pointwise group convolutions* to reduce computation complexity of 1×1 convolutions. To overcome the side effects brought by group convolutions, we come up with a novel *channel shuffle* operation to help the information flowing across feature channels. Based on the two techniques, we build a highly efficient architecture called *ShuffleNet*. Compared with popular structures like [30, 9, 40], for a given computation complexity budget, our ShuffleNet allows more feature map channels, which helps to encode more information and is especially critical to the performance of very small networks.

We evaluate our models on the challenging ImageNet classification [4, 29] and MS COCO object detection [23] tasks. A series of controlled experiments shows the effectiveness of our design principles and the better performance over other structures. Compared with the state-of-the-art architecture *MobileNet* [12], ShuffleNet achieves superior performance by a significant margin, e.g. absolute 7.8% lower ImageNet top-1 error at level of 40 MFLOPs.

We also examine the speedup on real hardware, i.e. an off-the-shelf ARM-based computing core. The ShuffleNet model achieves ∼**13**× *actual* speedup (theoretical speedup is 18×) over AlexNet [21] while maintaining comparable accuracy.

2. Related Work

Efficient Model Designs The last few years have seen the success of deep neural networks in computer vision tasks [21, 36, 28], in which model designs play an important role. The increasing needs of running high quality deep neural networks on embedded devices encourage the study on efficient model designs [8]. For example, *GoogLeNet* [33] increases the depth of networks with much lower complexity compared to simply stacking convolution layers. *SqueezeNet* [14] reduces parameters and computation significantly while maintaining accuracy. *ResNet* [9, 10] utilizes the efficient bottleneck structure to achieve impressive performance. *SENet* [13] introduces an architectural unit that boosts performance at slight computation cost. Concurrent with us, a very re-

* Equally contribution.