

ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design

Ningning Ma^{1,2}[0000-0003-4628-8831], Xiangyu Zhang¹[0000-0003-2138-4608],
Hai-Tao Zheng²[0000-0001-5128-5649], and Jian Sun¹[0000-0002-6178-4166]

¹ Megvii Inc (Face++)

{maningning, zhangxiangyu, sunjian}@megvii.com

² Tsinghua University

zheng.haitao@sz.tsinghua.edu.cn

Abstract. Currently, the neural network architecture design is mostly guided by the indirect metric of computation complexity, i.e., FLOPs. However, the direct metric, e.g., speed, also depends on the other factors such as memory access cost and platform characteristics. Thus, this work proposes to evaluate the direct metric on the target platform, beyond only considering FLOPs. Based on a series of controlled experiments, this work derives several practical guidelines for efficient network design. Accordingly, a new architecture is presented, called ShuffleNet V2. Comprehensive ablation experiments verify that our model is the state-of-the-art in terms of speed and accuracy trade-off.

Keywords: CNN architecture design, efficiency, practical

1 Introduction

The architecture of deep convolutional neural networks (CNNs) has evolved for years, becoming more accurate and faster. Since the milestone work of AlexNet [15], the ImageNet classification accuracy has been significantly improved by novel structures, including VGG [25], GoogLeNet [28], ResNet [5, 6], DenseNet [11], ResNeXt [33], SE-Net [9], and automatic neural architecture search [39, 18, 21], to name a few.

Besides accuracy, computation complexity is another important consideration. Real world tasks often aim at obtaining best accuracy under a limited computational budget, given by target platform (e.g., hardware) and application scenarios (e.g., auto driving requires low latency). This motivates a series of works towards light-weight architecture design and better speed-accuracy trade-off, including Xception [2], MobileNet [8], MobileNet V2 [24], ShuffleNet [35], and CondenseNet [10], to name a few. Group convolution and depth-wise convolution are crucial in these works.

To measure the computation complexity, a widely used metric is the number of float-point operations, or FLOPs¹. However, FLOPs is an indirect metric. It

¹ Equal contribution.

¹ In this paper, the definition of FLOPs follows [35], i.e. the number of multiply-adds.

