Structured Binary Neural Networks for Accurate Image Classification and Semantic Segmentation

Abstract

> In this paper, we propose to train convolutional neural networks (CNNs) with both binarized weights and activations, leading to quantized models specifically for mobile devices with limited power capacity and computation resources.
> We take a novel “structure approximation” view for quantization. In particular, we propose a “network decomposition” stategy, termed Group-Net, in which we divide the network into groups.

1. Introduction

> In this paper, we aim to design highly accurate binary neural networks (BNNs) from both the quantization and efficient architecture design perspectives.
> Existing quantization methods can be mainly divided into two categories:
> The first category methods seek to design more effective optimization algorithms to find better local minima for quantized weights.
> The second category approached focus on improving the quantization function.
> In this paper, we see to explore a third category called **structure approximation**. The main objective is to redesign a binary architecture that can directly match the capability of a floating-point model.
> In particular, we propose a Structured Binary Neural Network called Group-Net to partition the full-precision model into groups and use a set of parallel binary bases to approximate its floating-point structure counterpart. In this way, higher-level structural information can be better preserved than the **value approximation** approached.
> Furthermore, relying on the proposed structured model, we are able to design flexible binary structures according to different tasks and exploit task-specific information or structures to compensate the quantization loss and facilitate training.

> Our methods are also motivated by those energy-efficient architecture design approached which seek to replace the traditional expensive convolution with computational efficient convolutional operation (i.e. depthwidth separable convolution, 1×1 convolution).
> Nevertheless, we propose to design binary network archirectures for dedicated hardware from the quantization view.
> Group-Net can be possibly further improved with Neural Archirecture Search methods.
>
> - We propose to design accurate BNNs structures from the **structure approximation** perspective.
> - The proposed Group-Net has strong flexibility and can be easily extended to other tasks.

Some relevant works

Network quantization:
Efficient archirecture design:

- Efficient model designs like GoogLeNet and SqueezeNet propose to replace 3x3 convolutional kernels with 1x1 size to reduce the complexity while increasing the depth and accuracy.
- Additionally, separable convolutions are also proved to be effective in Inception approaches. This idea is further generalized as depthwise separable convolutions by Xception, MobileNet, and ShuffleNet to generate energy-efficient network structure.
- To avoid handcrafted heuristics, neural archirecture search has been explored for automatic model design.

2. Method

> In this paper, we seek to binarize both weights and activations of CNNs from a “structure approximation” view.

2.1. Problem definition

Binarization of weights:
Binarization of activations:

2.2. Structured Binary Network Decomposition

> We propose to decompose a network into binary structures while preserving its representability.
> Given a floating-point residual network $\mathbf{\Phi}$ with N blocks, we decompose $\mathbf{\Phi}$ into P **binary fragments**.
> Two kinds of archirectures: layer-wise decomposition and group-wise decomposition

2.2.1. Layer-wise binary decomposition

In the layer-wise approach, we approximate each layer with multiple branches of binary layers.

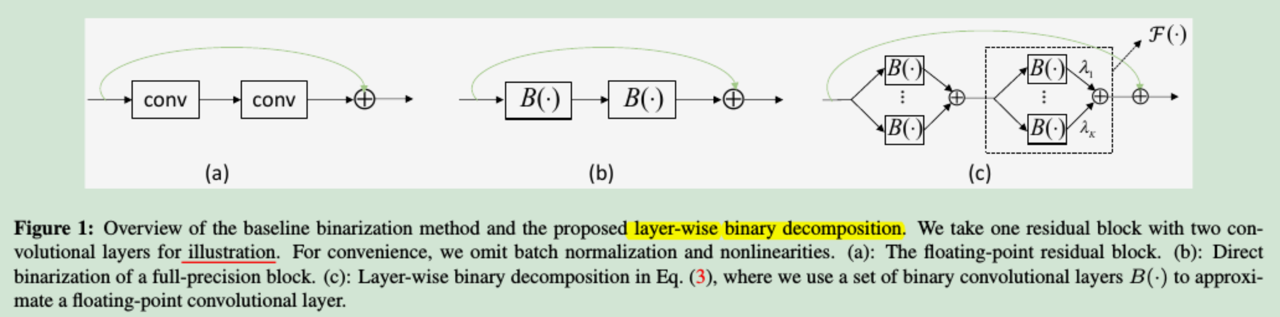

Figure 1: Overview of the baseline binarization method and the proposed layer-wise binary decomposition. We take one residual block with two convolutional layers for illustration. For convenience, we omit batch normalization and nonlinearities. (a): The floating-point residual block. (b): Direct binarization of a full-precision block. (c): Layer-wise binary decomposition in Eq. (3), where we use a set of binary convolutional layers $B(\cdot)$ to approximate a floating-point convolutional layer.

2.2.2. Group-wise binary decomposition

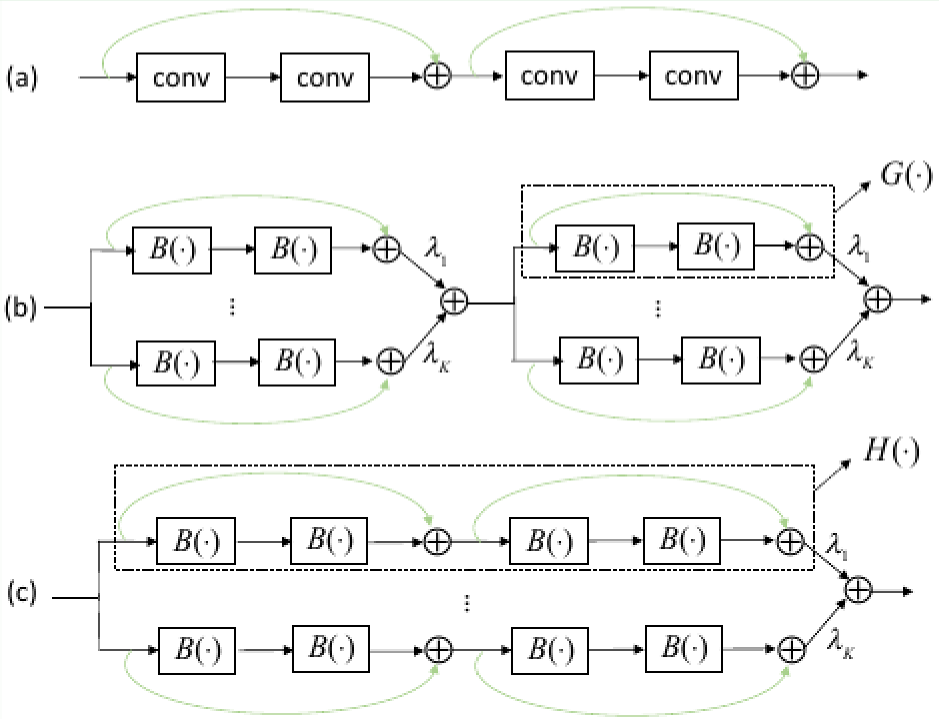

Figure 2: Illustration of the proposed group-wise binary decomposition strategy. We take two residual blocks for description. (a): The floating-point residual blocks. (b): Basic group-wise binary decomposition in Eq. (5), where we approximate a whole block with a linear combination of binary blocks $G(\cdot)$. (c): We approximate a whole group with homogeneous binary bases $H(\cdot)$, where each group consists of several blocks. This corresponds to Eq. (6).

2.2.3. Learning for dynamic decomposition

> Note that the network has N blocks and the possible number of connections is 2^N. Clearly, it is not practical to enumerate all possible structures during the training. Here, we propose to solve this problem by learning the structures for decomposition dynamically. We introduce in a fusion gate as the soft connection between blocks $G(*)$

2.3. Extension to semantic segmentation

3. Discussions

4. Experiment

5. Conclusion

> In this paper, we have begun to explore highly efficient and accurate CNN archirectures with binary weights and activation.
> Specifically, we have proposed to directly decompose the full-precision network into multiple groups and each group is approximated using a set of binary bases with can be optimized in an end-to-end manner.
> We also propose to learn the decomposition automatically. Experimental results have proved the effectiveness of the proposed approach on the ImageNet classification task.
> Moreover, we have generalized Group-Net from image classification task to semantic segmentation and achieved promising performance on PASCAL VOC.
> We have implemented the homogeneous multi-branch structure on CPU and achieved promising acceleration on test-time inference.