Binary Neural Networks. Lossless picture quality for binary neural networks in pixel-level tasks.

A Preprint

Kirill Ovcharenko
Moscow Institute of Physics and Technology

Zharikov Ilya
Moscow Institute of Physics and Technology

Abstract

Image Super Resolution (SR) is a crucial class of image processing techniques that enhance quality of visual data. Deep Convolutional Neural Networks (DCNN) have recently shown great results in this field. However, application of DCNN on resource-limited devices remains challenging, as they demand significant amounts of memory, energy and computations. Binary neural networks (BNN) provide a promising approach to reduce computational complexity and speed up the inference of a model. To our best knowledge, there are not many papers devoted to applying BNN to SR tasks, as SR models are much more vulnerable to degradation in performance when decreasing the precision of weights, than image classification models. The paper proposes modification of a convolutional block to make it binary without performance decreasing.

Keywords Binary Neural Network (BNN) · Single Image Super Resolution (SISR) · Binarization · Model compression

1 Introduction

ToDo: Need to be extended in terms related works further.

Image Super Resolution (SR) aims to restore High Quality (HQ) image from corrupted Low Quality (LQ) counterpart. This task is important because of its various applications in medical imaging and surveillance instruments. In spite of the research in this field being active, it made some progress not a long ago, as some challenges were encountered. Main obstacle is desired output in SR tasks being much more diverse than input, so the model is required to do dense pixel-level prediction, hence is bound to be more complex.

Recent advances in the field of SR owe their success to Deep Neural Networks (DNN) which show state-of-the-art results in a wide range of computer vision problems, such as image classification, Semantic Segmentation etc. However, the models solving these tasks are usually complicated and demand a lot of space and computational resources, thus hindering their implementation on mobile devices, drones and other machines which are limited in GPU memory.

Lately, different methods of reducing complexity of these models were proposed. While some papers focus on pruning and knowledge distillation, other researches introduce quantization as a way to decrease memory needed.

The most extreme form of quantization is binarization. Binary Neural Networks (BNN) use only 1 bit to represent each parameter, thus drastically decreasing space demanded to store the model. Moreover, with all parameters of the model set to $\{-1, 1\}$, most of the the calculations can be conducted using XNOR and Bitcount operations. This approach seems promising, as it proposes new ways to design hardware that can help to handle and exploit complex neural networks.

However, it is obvious that BNN sacrifice precision and quality, as they have much less capacity and representational potential than Full-Precision (FP) networks. Previous works in this field propose different methods of maintaining competitive accuracy while achieving better performance. The paper Ma et al. [2019]

focuses on residual block binarization, which helps to reduce a significant part of the model's parameters. However, full-precision activations keep computational complexity of the model pretty high. The BBCU Xia et al. [2022] proposed effective Convolutional Unit that can be used in any architecture that relies on residual connections. It provides much more efficient training and inference, but oversimplifies weight binarization. IR-Net Qin et al. [2020] reduces infromation loss by balancing weights to achieve maximum information entropy in forward propagation, After that, IR2Net Xue et al. [2022] proposes two essential components of the learning process: Information Restriction and Information Recovery. However, paper investigates only the impact of these methods on performance of Image Classification models.

This paper adopts some techniques from the researches, mentioned above, and suggests further modifications of convolutional block that help to improve BNN's performance in SR tasks.

Firstly, we adopt the idea of binarizing the activations with learnable threshold from Xia et al. [2022]. That helps to achieve better performance while preserving competitive quality. We also try different sign's derivative's approximation to use in training stage.

Secondly, we keep the idea of updating full-precision weights with respect to binarized weights from Ma et al. [2019].

Finally, we advance the idea from Xue et al. [2022] to restrict information from the input to increase learning productivity by implementing attention modules into our binary network.

2 Problem statement

ToDo: Need to be slightly reformulated during theory week.

Let $\{(X_i, Y_i)\}_{i=1}^n$ be our image dataset, where $X_i \in \mathbb{R}^{h_i \times w_i \times 3}$ denotes the low resolution image and $Y_i \in \mathbb{R}^{H_i \times W_i \times 3}$ - the high resolution one. Considering M to be the model, SR task targets optimization of

$$Q(M) = \frac{1}{n} \sum_{i=1}^{n} f(M(X_i), Y_i) \tag{1}$$

where f represents either $PSNR$ or $SSIM$ metric, defined as:

$$PSNR(x, y) = 10 \log_{10} \left(\frac{MAX_I^2}{MSE(x, y)} \right) \tag{2}$$

Here MAX_I is the maximum valid value for pixel, MSE is mean squared error.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{3}$$

μ_x denotes the mean for x, μ_y - the mean for y, σ_x is the variance for x, σ_y is the variance for y, σ_{xy} is the covariation of x and y, c_1 and c_2 - two constants depending on dynamic pixel range.

Now let $B \in \mathcal{B}$ be our binarized representation of model M. $\mathcal{B}$ is space of BNN. Assuming L to be number of layers in $B \in \mathcal{B}$, $W_l \in \{-1, 1\}^{C_{out} \times C_{in} \times K_h \times K_w}$, $l \in \{1\ , \dots\ , L\}$. Here C_{out} is the number of output channels, C_{in} is the number of input channels, K_h is the kernel height, K_w is the kernel weight.

Thus, the problem of binarization can be expressed in finding B^* as

$$B^* = \arg\min_{B \in \mathcal{B}} \left[Q(M) - Q(B) \right] \tag{4}$$

Список литературы

Yinglan Ma, Hongyu Xiong, Zhe Hu, and Lizhuang Ma. Efficient super resolution using binarized neural network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 0–0, 2019.

Bin Xia, Yulun Zhang, Yitong Wang, Yapeng Tian, Wenming Yang, Radu Timofte, and Luc Van Gool. Basic binary convolution unit for binarized image restoration network. arXiv preprint arXiv:2210.00405, 2022.

Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, and Jingkuan Song. Forward and backward information retention for accurate binary neural networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2250–2259, 2020.

Ping Xue, Yang Lu, Jingfei Chang, Xing Wei, and Zhen Wei. Ir2net: Information restriction and information recovery for accurate binary neural networks. arXiv preprint arXiv:2210.02637, 2022.