XNORT_NET imagenet classification using Binary CNN

Saturday, October 13, 2018 7:30 AM

9

SUBTOPIC

9

XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks

Mohammad Rastegari*, Vicente Ordonez*, Joseph Redmon*, Ali Farhadi*
*Allen Institute for AI
*University of Washington
{mohammadr,vicenteo,pjreddie,ali}@cs.washington.edu

Abstract

We propose two efficient approximations to standard convolutional neural networks: Binary-Weight-Networks and XNOR-Networks. In Binary-Weight-Networks, the filters are approximated with binary values resulting in 32× memory saving. In XNOR-Networks, both the filters and the input to convolutional layers are binary. XNOR-Networks approximate convolutions using primarily binary operations. This results in 58× faster convolutional operations (in terms of number of the single CPU operations) and 32× memory saving. XNOR-Networks enable state-of-the-art deep learning algorithms to run on CPU (rather than GPU) in real-time. Our binary weight networks, trained on the ImageNet dataset, are also more powerful than previous binary networks.

1 Introduction

Deep neural networks (DNN) have made significant improvement in several applications including object recognition, scene understanding, and semantic segmentation. Even though, a typical state-of-the-art DNN may have a large number of parameters, it is in principle possible to use such networks on a single CPU machine. However, in practice, DNNs are computationally demanding even on a single image. Such computational cost is not a serious issue for large scale or cloud computing. Computation on mobile devices is more restrictive since it drains the battery and requires more memory. Embedded vision applications such as robots, wearable computers and smart glasses have even more restrictive resources. Ideally, state-of-the-art deep learning algorithms should be able to run on such devices. A GPU is not a practical option for embedded applications. Some recent work in literature addresses this problem by proposing more efficient CNN architectures, or by reducing the network size and computation time by removing unimportant connections [7], quantizing the network [3], using low bit-width multiplications [4], and using low rank approximations [13].

To the best of our knowledge, most of the previous work on CNN acceleration targets a fixed accuracy and does not quantify the trade-off between computational resources and accuracy. Instead, we show that binary networks offer a good trade-off between computational resources and accuracy. We also show that binary networks are more powerful than other compact networks with the same number of bits.

In this paper, we propose two novel models for deep learning: Binary-Weight-Networks, and XNOR-Networks. In Binary-Weight-Networks, the weights of the convolutional layers are approximated with binary values (this results in 32× memory saving). In XNOR-Networks, both the weights and the input to convolutional layers are binary. This network activation function, which we call binarized activation function, is similar to that of [4]. We show binary activations, in addition to binary weights, drastically increase the accuracy. Binary activations also enable to use binary operations between the weights and activations. We show such binary operations can be efficiently implemented using XNOR and bit-counting operations. We show our proposed method can be used to all recent state-of-the-art deep networks with different architectures. We show our method gives 58× speed up in the CPU implementation and we show that we can achieve more than 50× memory saving. We also show our binary networks can be trained from scratch on large scale datasets such as ImageNet and still achieve high accuracy. We also show our binary networks can be used for the task of object detection. We also show our binary networks can be used for the task of object detection. We also show our binary networks can be used for the task of object detection.

To the best of our knowledge, this paper is the first attempt to propose a method for training binary networks and to report such high accuracy on ImageNet. We also report the first results of binary networks for object detection.

2 Binary Convolutional Neural Networks

We propose two different models for binary CNN: Binary-Weight-Networks and XNOR-Networks. In this section we formulate binary convolutional layers. We also explain how such binary networks can be trained. We also explain why binary networks are more powerful than other compact networks.

2.1 Binary Weight Networks

Consider a convolutional layer in a CNN. The layer's input is a tensor $I \in \mathbb{R}^{c \times w \times h}$ where c is the number of channels, w and h are the width and height of the input. The layer's weight tensor is $W \in \mathbb{R}^{c \times w \times h}$