Bi-Real Net: Enhancing the Performance of 1-bit CNNs With Improved Representational Capability and Advanced Training Algorithm

Zechun Liu[1] , Baoyuan Wu[2] , Wenhan Luo[2], Xin Yang[3⋆], Wei Liu[2], and Kwang-Ting Cheng[1]

[1] Hong Kong University of Science and Technology
[2] Tencent AI lab
[3] Huazhong University of Science and Technology
zliubq@connect.ust.hk, {wubaoyuan1987, whluo.china}@gmail.com, xinyang2014@hust.edu.cn, wliu@ee.columbia.edu, timcheng@ust.hk

Abstract. In this work, we study the 1-bit convolutional neural networks (CNNs), of which both the weights and activations are binary. While being efficient, the classification accuracy of the current 1-bit CNNs is much worse compared to their counterpart real-valued CNN models on the large-scale dataset, like ImageNet. To minimize the performance gap between the 1-bit and real-valued CNN models, we propose a novel model, dubbed Bi-Real net, which connects the real activations (after the 1-bit convolution and/or BatchNorm layer, before the sign function) to activations of the consecutive block, through an identity shortcut. Consequently, compared to the standard 1-bit CNN, the representational capability of the Bi-Real net is significantly enhanced and the additional cost on computation is negligible. Moreover, we develop a specific training algorithm including three technical novelties for 1-bit CNNs. Firstly, we derive a tight approximation to the derivative of the non-differentiable sign function with respect to activation. Secondly, we propose a magnitude-aware gradient with respect to the weight for updating the weight parameters. Thirdly, we pre-train the real-valued CNN model with a clip function, rather than the ReLU function, to better initialize the Bi-Real net. Experiments on ImageNet show that the Bi-Real net with the proposed training algorithm achieves 56.4% and 62.2% top-1 accuracy with 18 layers and 34 layers, respectively. Compared to the state-of-the-arts (*e.g.*, XNOR Net), Bi-Real net achieves up to 10% higher top-1 accuracy with more memory saving and lower computational cost. [4]

1 Introduction

Deep Convolutional Neural Networks (CNNs) have achieved substantial advances in a wide range of vision tasks, such as object detection and recognition

⋆ Corresponding author

Bi-Real Net

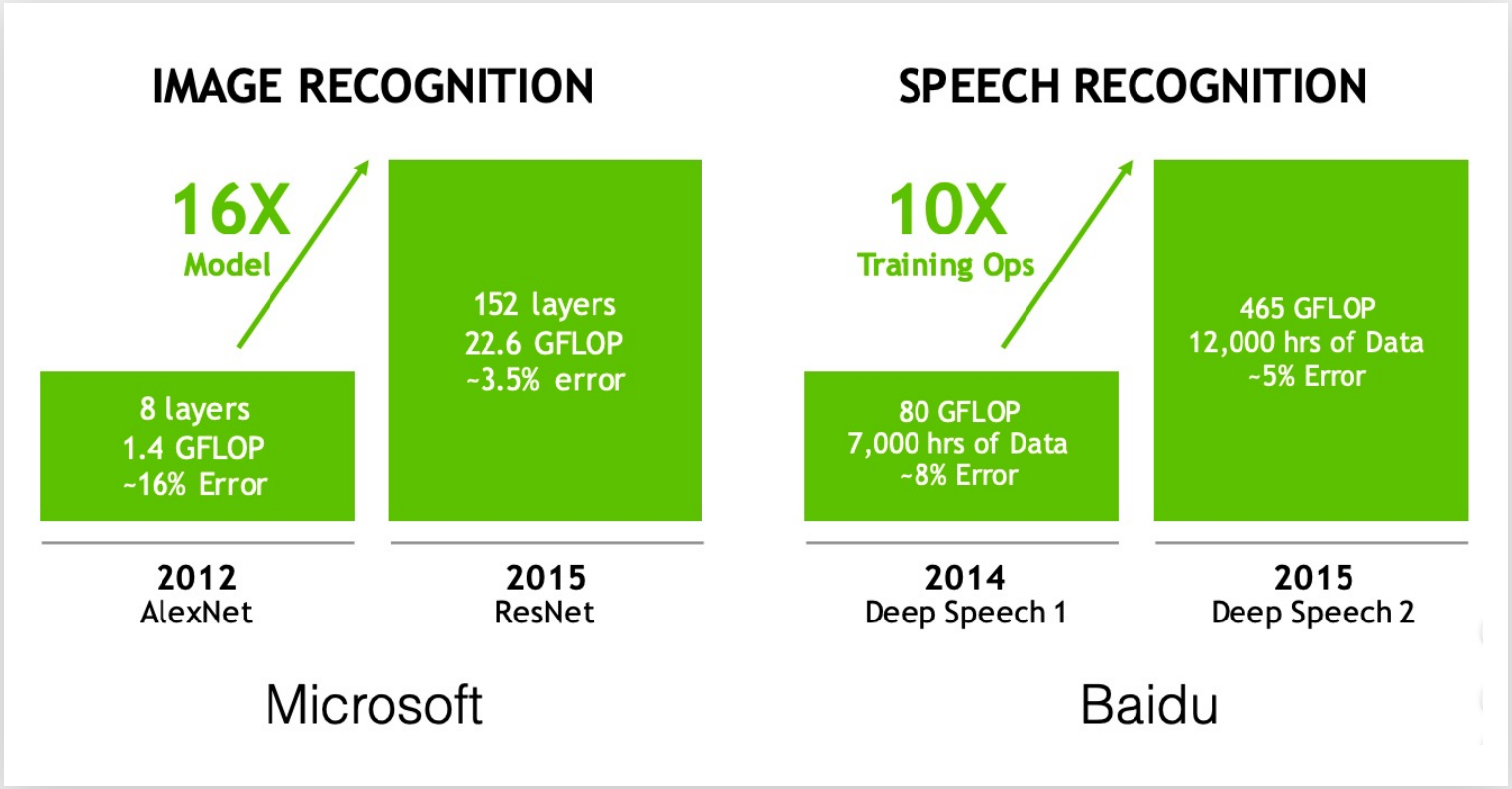
Enhancing the Performance of 1-bit CNNs with Improved Representational Capability and Advanced Training Algorithm

👨‍🎓 Zechun Liu, Baoyuan Wu, Wenhan Luo, Xin Yang, Wei Liu, Kwang-Ting Cheng
📍 ECCV 2018

Introduction

Introduction

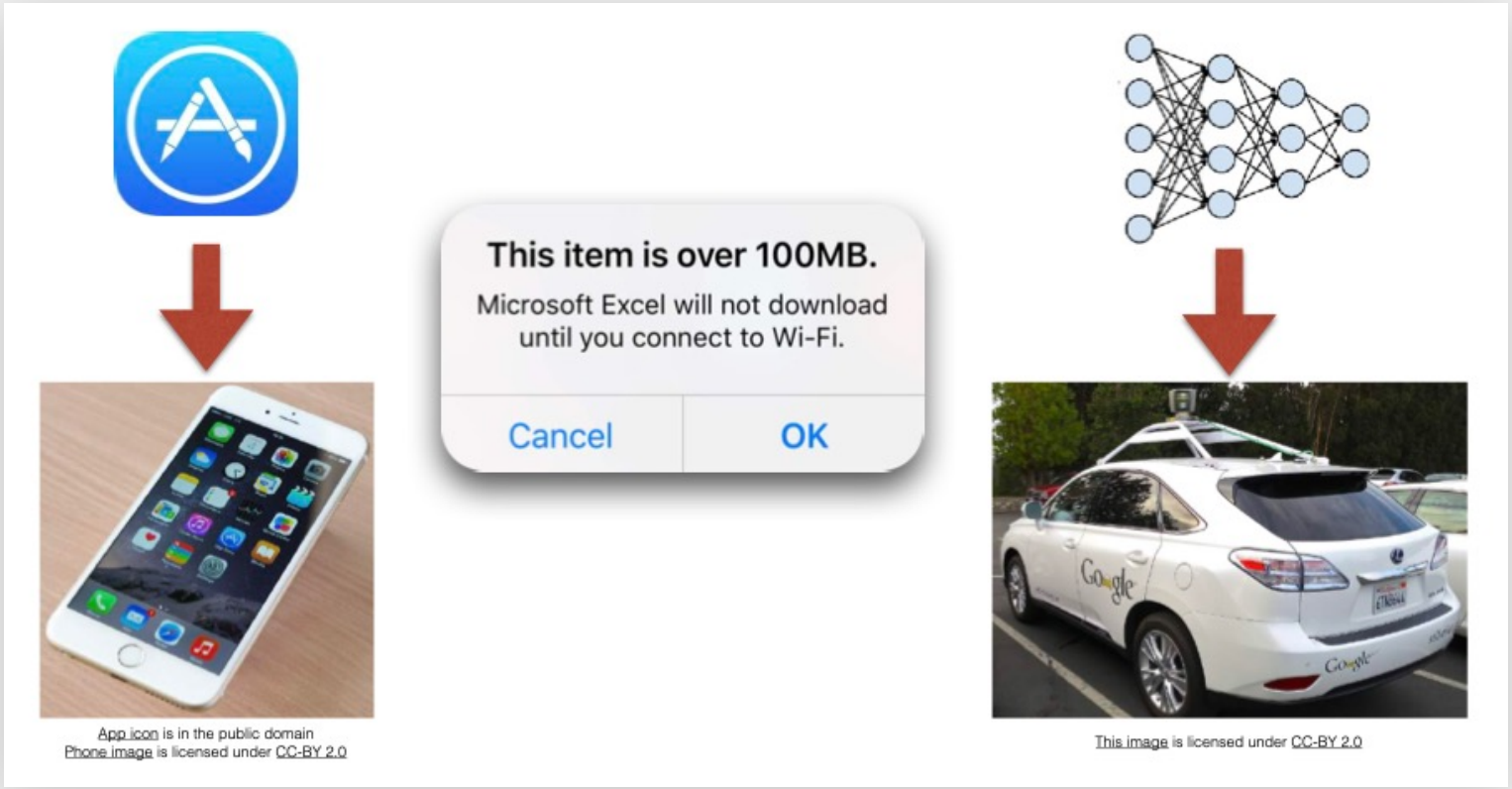
Trends of Size of Deep Learning Models 📈



Models are getting larger.

Introduction

The First Challenge: Model Size 🐘



Hard to distribute large models through over-the-air update.

Introduction

The Second Challenge: Speed 🐢

	Error rate	Training time
ResNet18:	10.76%	2.5 days
ResNet50:	7.02%	5 days
ResNet101:	6.21%	1 week
ResNet152:	6.16%	1.5 weeks

Such long training time limits ML researcher's productivity.

Introduction

The Third Challenge: Energy Efficiency ⚡

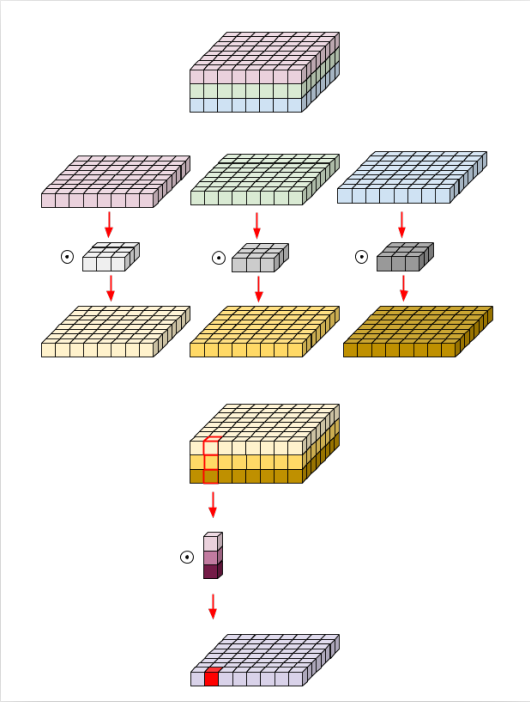
AlphaGo
$3,000 electric bill per game

Lee Sedol
A bowl of gukbap per game

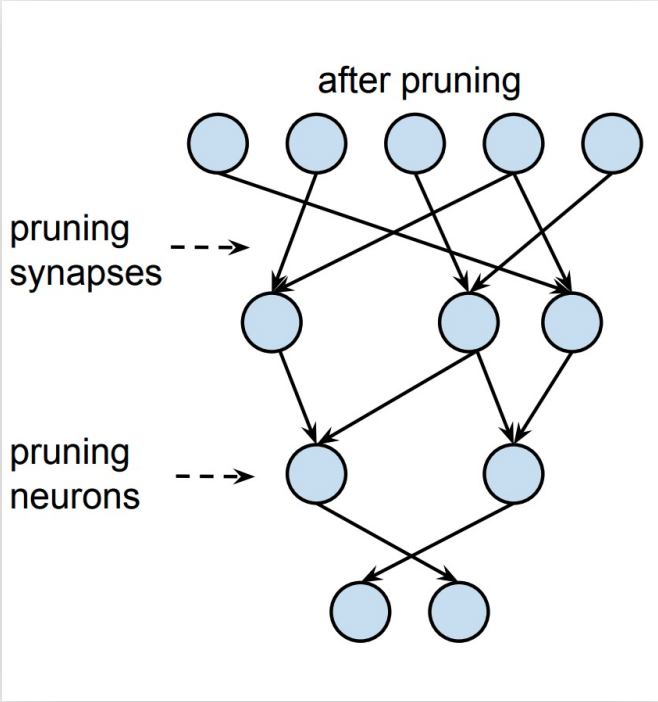
On Mobile: Drains battery
On Data center: Increases TCO

Introduction

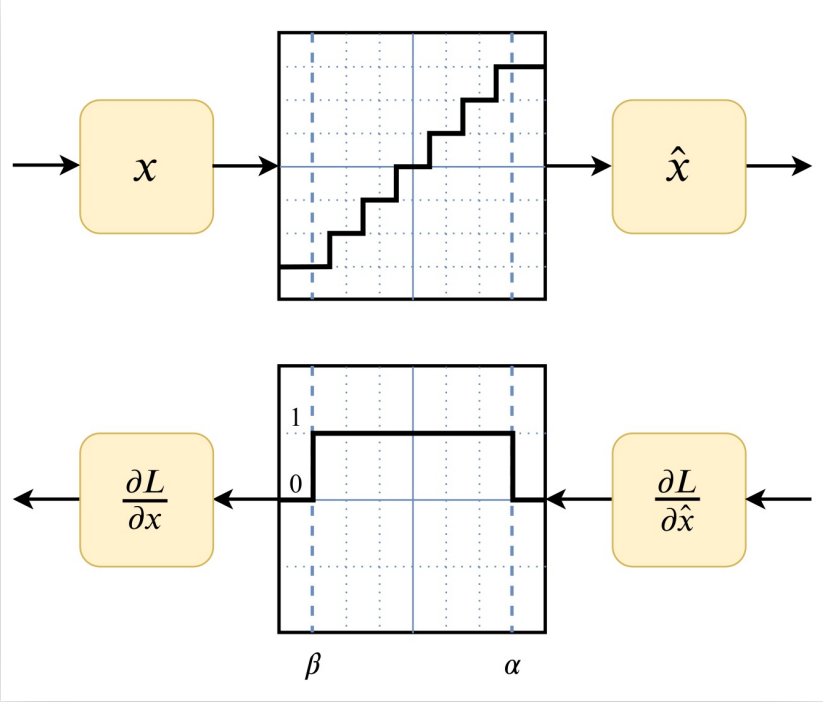
Examples of Model Compression 🗜



Low rank approximation



Pruning

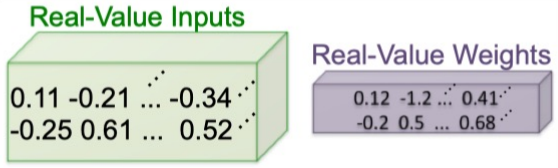
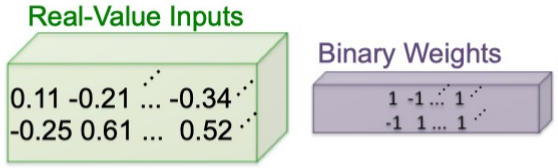
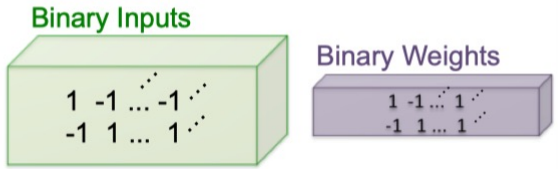


Quantization

XNOR-Net

XNOR-Net

Binary CNN 0 1

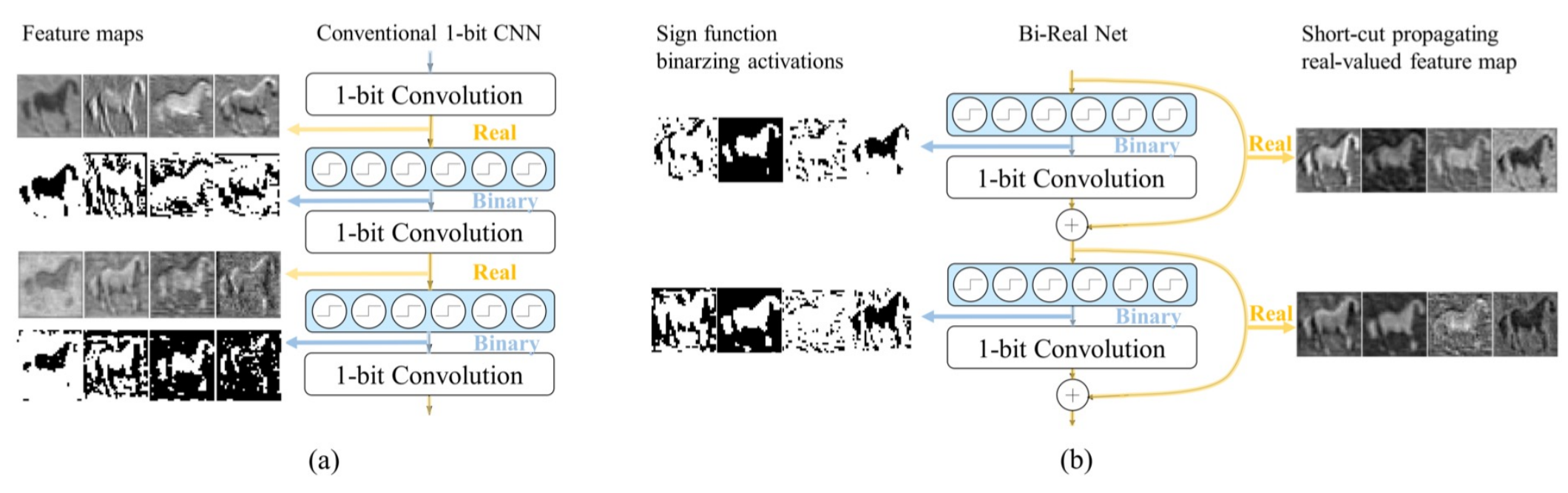
	Network Variations	Operations used in Convolution	Memory Saving (Inference)	Computation Saving (Inference)	Accuracy on ImageNet (AlexNet)
Standard Convolution		+ , − , ×	1x	1x	%56.7
Binary Weight		+ , −	~32x	~2x	%56.8
BinaryWeight Binary Input (**XNOR-Net**)		XNOR , bitcount	~32x	~58x	%44.2

Bi-Real Net

Bi-Real Net

Network Architecture 🏗️

Network with intermediate feature visualization.



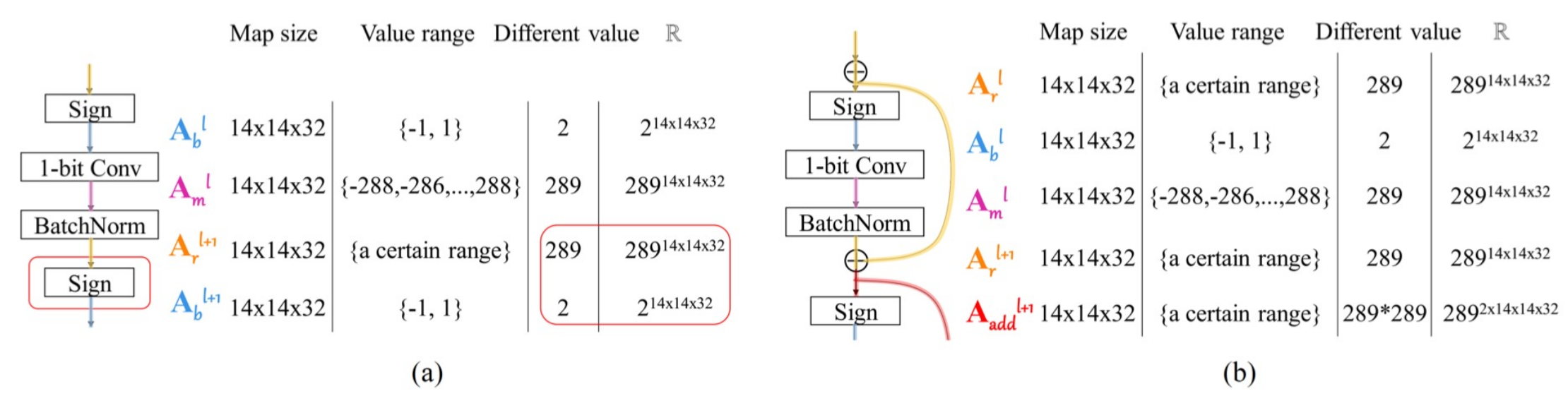
(a) 1-bit CNN without shortcut.

(b) Proposed Bi-Real net with shortcut propagating the real-valued features.

Bi-Real Net

Network Architecture 🏗️

The representational capability of each layer.



(a) 1-bit CNNs without shortcut .

(b) 1-bit CNNs with shortcut.

- The poor performance of 1bit CNNs is caused by its low representational capacity.
- The representational capability of each block in the 1-bit CNN is significantly enhanced due to the simple identity shortcut.

Bi-Real Net

Training 🏋️

- As both activations and weight parameters are binary,
  the continuous optimization method cannot be directly adopted to train the 1-bit CNN.
  (i.e., stochastic gradient descent)

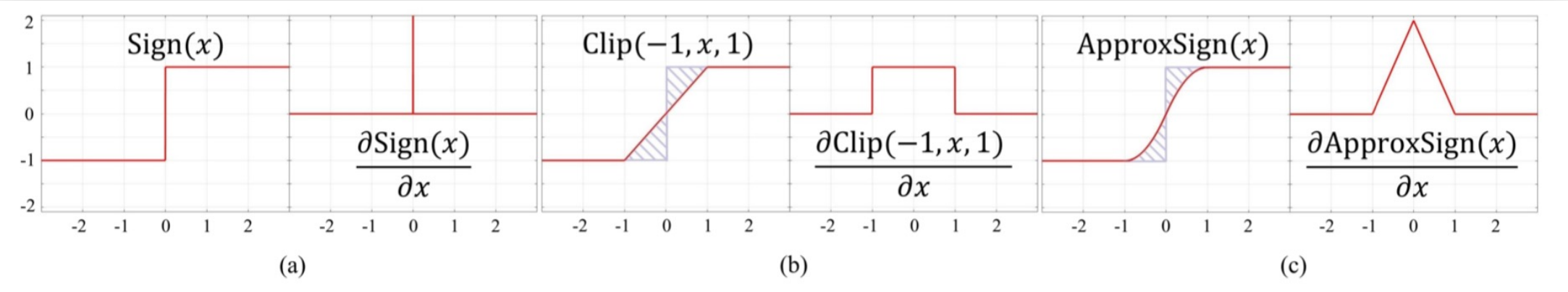
- There are two major challenges.
  1. How to compute the gradient of the sign function on activations, which is non-differentiable.
  2. The gradient of the loss is too small to change the weight's sign.

Bi-Real Net

Training 🏋

1. How can we compute the gradient of the sign function?

Approximation to the derivative of the sign function with respect to activations.



(a) Sign function and its derivative.

(b) Clip function and its derivative.

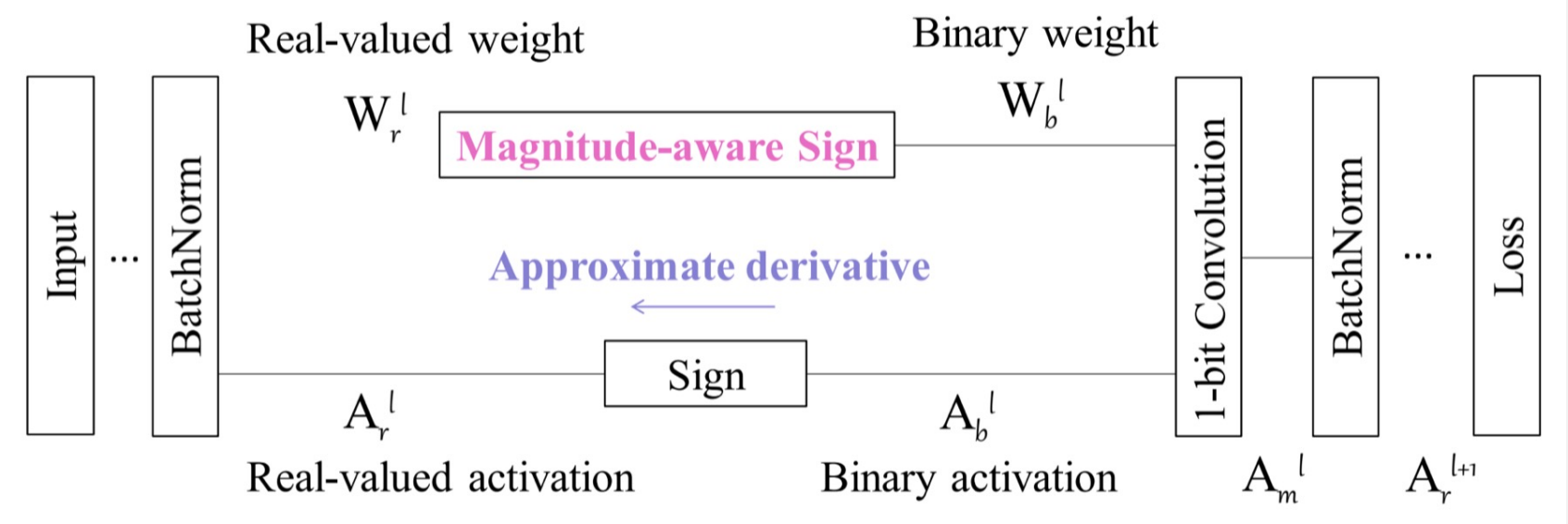
(c) Proposed differentiable piecewise polynomial function and its triangle-shaped derivative.

Bi-Real Net

Training 🏋

2. How can we change weight's sign with small gradients?

1. How can we compute the gradient of the sign function?



Bi-Real Net

Training 🏋

Initialization

- In previous works, the initial weights of the 1-bit CNNs
  are derived from the corresponding real-valued CNN model.

- However, the activation of $\boldsymbol{ReLU}$ is non-negative, while that of $\boldsymbol{Sign}$ is −1 or +1.

- Instead, we propose to replace $\boldsymbol{ReLU}$ with $\boldsymbol{clip}(-\mathbf{1}, \boldsymbol{x}, \mathbf{1})$ to pre-train the real-valued CNN model.