05

Memory efficient deep learning for mobile devices

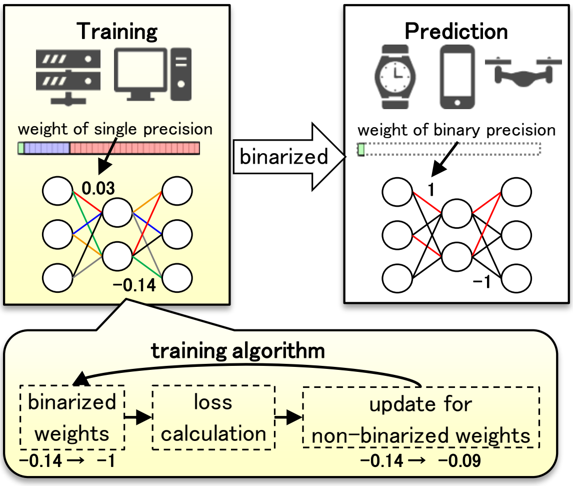
- Quantized neural networks for model compression -

Abstract

Deep Neural Network (DNN) is a promising technology in a wide range of application. Unfortunately, it is hard to use DNN on devices with limited resources such as mobile devices since the model size of DNN is large. Binarized DNN is constrained to have only two values in weights (e.g. -1 or 1) by using a binarization function. Since a binarized weight only consumes 1-bit of memory, Binarized DNN is one of the most powerful approaches of model compression. However, it is difficult to effectively train Binarized DNN. This is because the binarization function vanishes gradients that update weights. In this study, in order to effectively train Binarized DNN, we avoid the vanishing gradient problem by using a continuous function, that approximates the binarization function and is scaled according to a distribution of weights. In our experiments, our method achieved more efficient training and higher accuracy than previous.

Background

Binary weights reduce memory size



Training

weight of single precision

0.03

−0.14

binarized

Prediction

weight of binary precision

1

−1

training algorithm

binarized weights → loss calculation → update for non−binarized weights

−0.14 → −1

−0.14 → −0.09

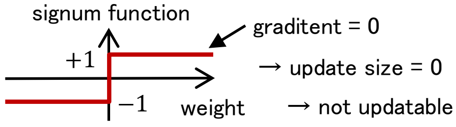
Problems

Binarization leads to less accuracy

Cause I Approximation errors are observed

Cause II The update size is not best

update size = learning rate × **gradient**



signum function

gradient = 0

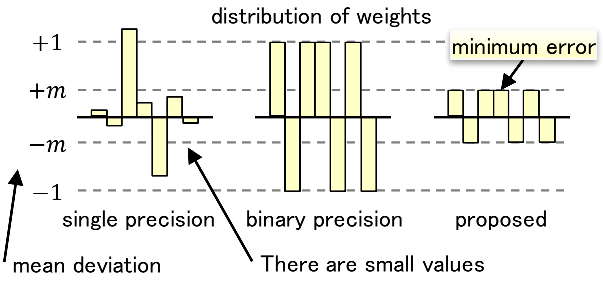
→ update size = 0

→ not updatable

Previous works put gradients as 1

Proposed method

Action I Binarization use mean deviation



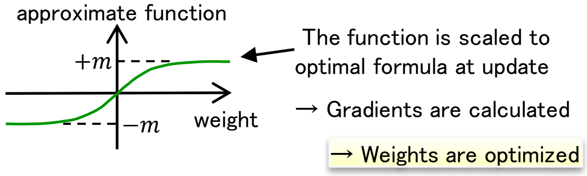
distribution of weights

minimum error

mean deviation

There are small values

Action II Update use continuous functions approximated signum ones



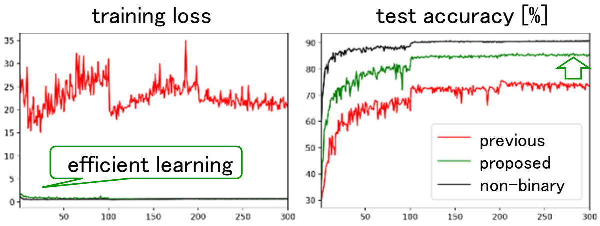
approximate function

The function is scaled to optimal formula at update

→ Gradients are calculated

→ Weights are optimized

Results



References

[1] Y. Oya, Y. Ida, Y. Fujiwara, S. Iwamura, “Quantized Neural Networks using Regularization,” *IEICE Technical Report*, Vol. 117, No. 211, pp. 51-52, 2017.

[2] Y. Oya, Y. Ida, Y. Fujiwara, S. Iwamura, “Binarised Neural Networks without Activation Functions,” *IEICE Technical Report*, Vol. 117, No. 238, pp. 119-120, 2017.

Contact

Yu Oya Distributed Computing Technology Project, Software Innovation Center