ResNet-like Architecture with Low Hardware Requirements

Elena Limonova[1,2], Daniil Alfonso[3], Dmitry Nikolaev[2,4], Vladimir V. Arlazarov[1,2]

[1] FRC CSC RAS, Moscow, Russia
[2] Smart Engines Service LLC, Moscow, Russia
[3] JSC MCST, Moscow, Russia
[4] Institute for Information Transmission Problems RAS, Moscow, Russia

Edge computing and BM networks

Fast and resource efficient neural networks are extremely important for edge computing:

- mobile recognition,
- internet of things,
- autonomous vehicles.

The Bipolar Morphological (BM) Networks:

- use less computationally intensive addition and maximum instead of multiplication and addition;
- can be used with any convolutional neural network architecture;
- BM convolutional layers can be combined with any other layers.

The BM Neuron

The standard neuron performs the operation:

$$y(\mathbf{x}, \mathbf{v}, v_b) = \sigma\left(\sum_{j=1}^{N} x_j v_j + v_b\right),$$

The bipolar morphological neuron:

$$y_{BM}(\mathbf{x}, \mathbf{v}^+, \mathbf{v}^-, v_b) = \sigma\left(\sum_{\alpha\in\{-,+\}} \sum_{\beta\in\{-,+\}} p^\alpha p^\beta \exp \max_{j=1}^{N}(\ln x_j^\alpha + v_j^\beta) + v_b\right),$$

where $p^+ = 1$, $p^- = -1$, N is an input length, $\mathbf{x}$ is an input vector, $\mathbf{v}^+$, $\mathbf{v}^-$ are weight vectors, v_b is a bias, and $\sigma(\cdot)$ is a non-linear activation,

$$x_j^+ = \begin{cases} x_j, x_j \geq 0, \\ 0, x_j < 0, \end{cases} \qquad x_j^- = \begin{cases} -x_j, x_j < 0, \\ 0, x_j \geq 0. \end{cases}$$

The BM Convolutional Layer

$I_{N\times M\times C}$ – input image

$J_{N\times M\times F}$ – output image

The standard convolutional layer:

$$J = \sigma\left(I * w + \mathbf{b}\right),$$

where $*$ is an image convolution operation.

The BM convolutional layer:

$$J = \sigma\left(\sum_{\alpha\in\{-,+\}}\sum_{\beta\in\{-,+\}} p^{\alpha}p^{\beta}\exp(\ln I^{\alpha}\odot v^{\beta}) + \mathbf{b}\right),$$

where $p^{+}=1$, $p^{-}=-1$, $\odot$ is a BM convolution operation:

$$(I\odot v)_{n,m,c} = \max_{c=1}^{C}\max_{\Delta n=0}^{K-1}\max_{\Delta m=0}^{K-1} I_{n+\Delta n,m+\Delta m,c} + v_{\Delta k,\Delta m,c,f}$$

Training of BM Network[1]

1. Train standard network using conventional gradient descent-based methods;
2. For each convolutional layer: replace layer with weights $\{w, b\}$ by the BM layer with weights $\{v^+, v^-, b\}$, where:

$$v_j^+ = \begin{cases} \ln|w_j|, \text{ if } w_j > 0 \\ -\infty, \text{ otherwise} \end{cases}$$

$$v_j^- = \begin{cases} \ln|w_j|, \text{ if } w_j < 0 \\ -\infty, \text{ otherwise} \end{cases}$$

3. Perform additional training of the network after conversion of each layer using the same method as in 1.

[1] E. Limonova, D. Matveev, D. Nikolaev, and V. V. Arlazarov, "Bipolar morphological neural networks: convolution without multiplication," ICMV 2019, 11433, 962 – 969, (2020).

Hardware implementation

- Verilog HDL and Synopsys Design Compiler (65 nm)
- Model single-precision addition, maximum, multiplication
- Approximations for exponent and logarithm

Table 1: The estimate number of gates and latency for arithmetical operations

Op	Gates	Latency, clock cycles
add	16048	3
max	1464	2
mul	35345	4
log	154179	35
exp	256965	21

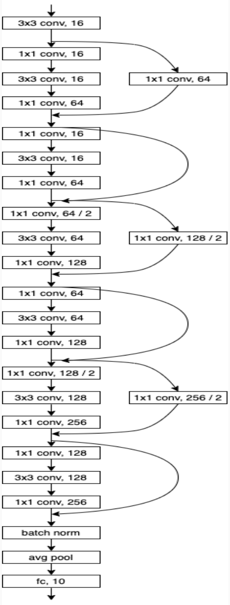
Hardware complexity for convolutional layers

Table 2: The approximate gate number and latency ratios for standard and BM convolutional layers.

Filters	Channels	Filter Size	Gates, standard/BM	Latency, standard/BM
16	16	1	1.14	0.80
32	32	1	1.64	1.02
64	64	1	2.11	1.18
128	128	1	2.45	1.28
256	256	1	2.67	1,34
512	512	1	2.80	1.37
16	16	3	2.50	1.29
32	32	3	2.70	1.34
64	64	3	2.81	1.37
128	128	3	2.87	1.39
256	256	3	2.9	1.39
512	512	3	2.92	1.40

BM ResNet

- ResNet architecture with 22 convolutional layers;
- standard convolutions were replaced with BM ones;



Gate Number Evaluation

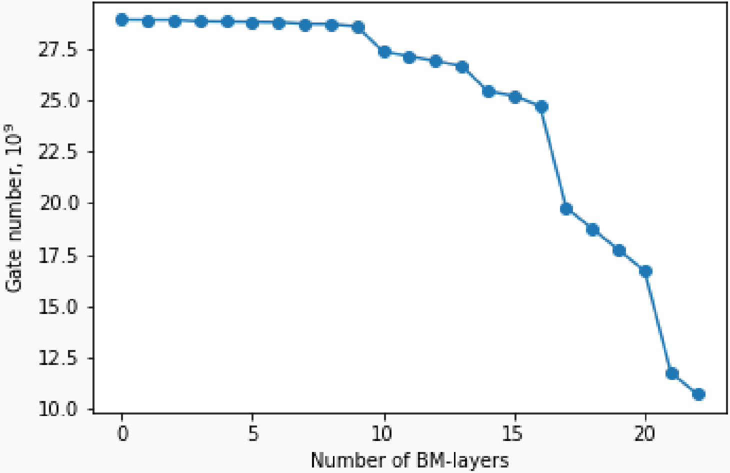
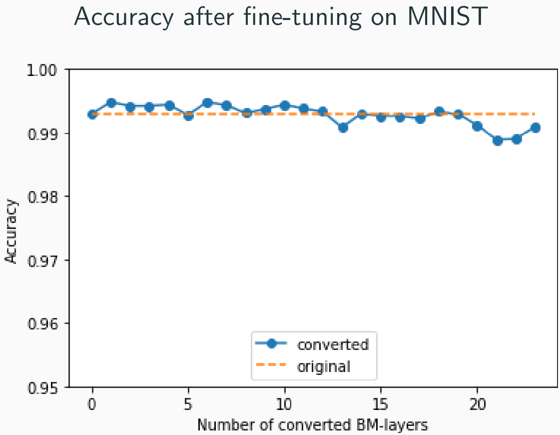


Figure 1: Gate number for convolutional layers of BM ResNet-22.

MNIST

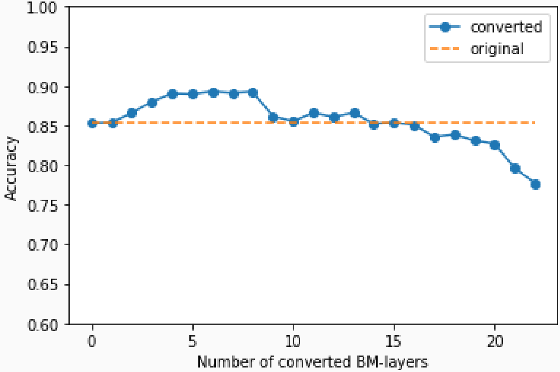
ResNet: **99.3%**
BM ResNet: **99.1%**



CIFAR-10

ResNet: **85.3%**
BM ResNet (16): **83.9%**
BM ResNet (22): **77.7%**

Accuracy after fine-tuning on CIFAR-10



Conclusion

In this paper we:

- introduce BM ResNet architecture for image classification:
    - MNIST accuracy 99.1%
    - CIFAR-10 accuracy 83.9%
- present significant benefits of BM networks for ASIC: computationally-intensive BM convolutions
    - require 2.1-2.9 fewer logic gates,
    - have 15-30% lower latency.

Future Work

- Conduct further research on training of BM networks
- Design ASIC for BM deep neural networks
- Create quantization methods for BM networks