Neural Networks without Multiplications

Gabrielle Liu[1] Allen Liu[2]

[1]Ravenwood High School

[2]McCallie School

Siemens Competition National Finals

Roadmap

Autonomous Vehicles

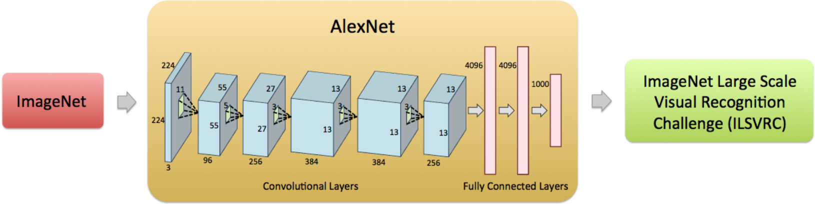
Revolutionizing transportation, reducing injuries, decreasing traffic congestion, and improving air quality

→ Object Detection and Classification → Deep Neural Networks

Google autonomous vehicle.
Source: Michael Shick

AlexNet[1]

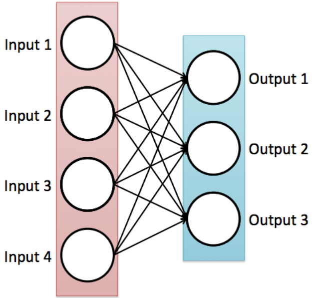
ImageNet has sparked research innovation in visual object recognition through deep learning.



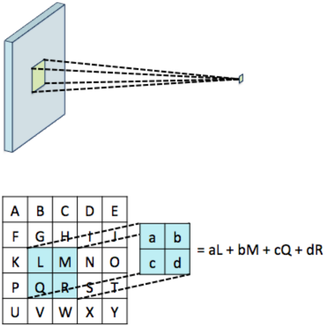
[1] "Convolutional Neural Networks (CNNs/ConvNets)." *CS231n: Convolutional Neural Networks for Visual Recognition.* Stanford University, 2017. *https://cs231n.github.io/convolutional-networks/*

Deep Neural Networks

Convolutional neural networks (CNNs) utilize multiple layer types to achieve near-human accuracy in object recognition.



Fully Connected Layer



Convolutional Layer

The Cost of Improved Prediction Accuracy

- Increased computational complexity
- Expensive customized hardware
- Complex configuration of associated software

Roadmap

To Reduce Computational Complexity

- Use of binary weights
- Weight quantization: values restricted to powers of 2
- Replacement of multiplications with bit-shifts

Roadmap

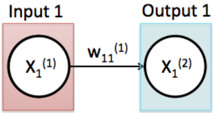
Overview

1. Investigate how multiplication gates[2](MGs) can be used in large neural networks
2. Devise a set of no-multiplication architectures (NMAs) for:
   - Fully connected neural networks (FCNNs)
   - Convolutional neural networks (CNNs)
3. Derive mathematical expressions for the number of distinct products to compute in training these architectures in order to evaluate the extent to which NMAs decrease computation cost

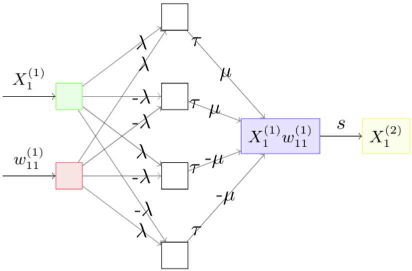
[2] Lin, Henry W., Max Tegmark, and David Rolnick. "Why does deep and cheap learning work so well?" *arXiv preprint arXiv:1608.08225v4 [cond-mat.dis-nn]*, 2017. *https://arxiv.org/pdf/1608.08225v4.pdf*

Initial Steps

NMA for the simplest case – forward propagation through an FCNN with $L = 1$ layer, containing one neuron.



Original Architecture



No-Multiplication Architecture

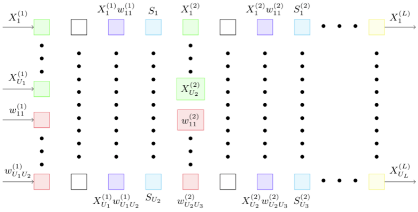
Subsequent Methodology

1. Construct a generalized NMA for forward propagation through an FCNN with $L \geq 1$ layers.
2. Construct a generalized NMA for back propagation through an FCNN with $L \geq 1$ layers.
3. Similarly construct generalized NMAs for both forward and back propagation through a CNN with $L \geq 1$ convolutional layers.
4. Derive mathematical expressions for the number of distinct products that must be computed using the NMAs.

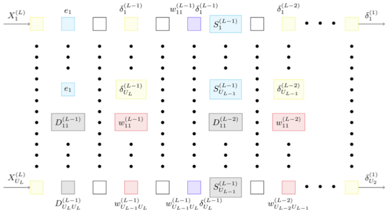
Roadmap

Generalized NMAs for FCNN with $L \geq 1$ layers

We constructed a set of architectures to implement FCNNs and fully connected layers without multiplication.



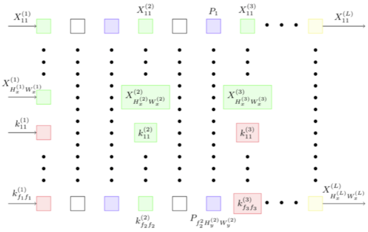
Forward Propagation



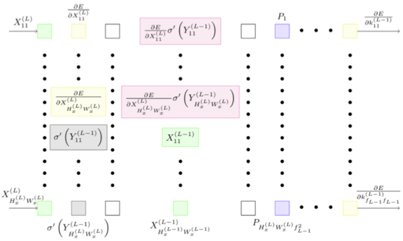
Back Propagation

Generalized NMAs for CNN with $L \geq 1$ layers

We constructed a set of architectures to implement convolutional layers without multiplication.



Forward Propagation



Back Propagation

Number of Distinct Products

We derived and proved theorems for three notable CNN cases:

1. The number of distinct products to compute when an image containing repeated values is convolved with a kernel containing distinct weight values
2. The number of distinct products to compute when an image containing distinct values is convolved with a kernel containing repeated weight values
3. The number of distinct products that must be computed for back propagation

Key Impact

- Our work on no-multiplication architectures has the potential to substantially expedite the training of neural networks on simple devices without custom hardware – a possible catalyst for the development of autonomous vehicles.

References

Lin, Zhouhan, Matthieu Courbariaux, Roland Memisevic, and Yoshua Bengio. "Neural Networks with Few Multiplications." *arXiv preprint arXiv:1510.03009v3 [cs.LG]*, 2015. *https://arxiv.org/pdf/1510.03009.pdf*

Marchesi, Michele, Gianni Orlandi, Francesco Piazza, and Aurelio Uncini. "Fast neural networks without multipliers." *IEEE Transactions on Neural Networks*, 4(1):53–62, 1993. *https://www.academia.edu/32961416/Fast_neural_networks_without_multipliers*

Rojas, Raul. "The Backpropagation Algorithm." *Neural Networks: A Systematic Introduction*. Berlin: Springer-Verlag, 1996. 151–171. *https://page.mi.fu-berlin.de/rojas/neural/neuron.pdf*

Simard, Patrice Y. and Hans Peter Graf. "Backpropagation without Multiplication." *Advances in Neural Information and Processing Systems*, pp.232–239, 1994. *https://papers.nips.cc/paper/833-backpropagation-without-multiplication.pdf*

Vanhoucke, Vincent, Andrew Senior, and Mark Z. Mao. "Improving the speed of neural networks on CPUs." *Deep Learning and Unsupervised Feature Learning NIPS Workshop*, 2011. *https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/37631.pdf*