Efficient Deep Learning for Ubiquitous AI

Devdatt Dubhashi

AI and Data Science, Chalmers

Machine Intelligence Sweden

AI: the New Electricity

"AI is the new electricity.
Just as electricity transformed industry
after industry 100 years ago,
I think AI will do the same."

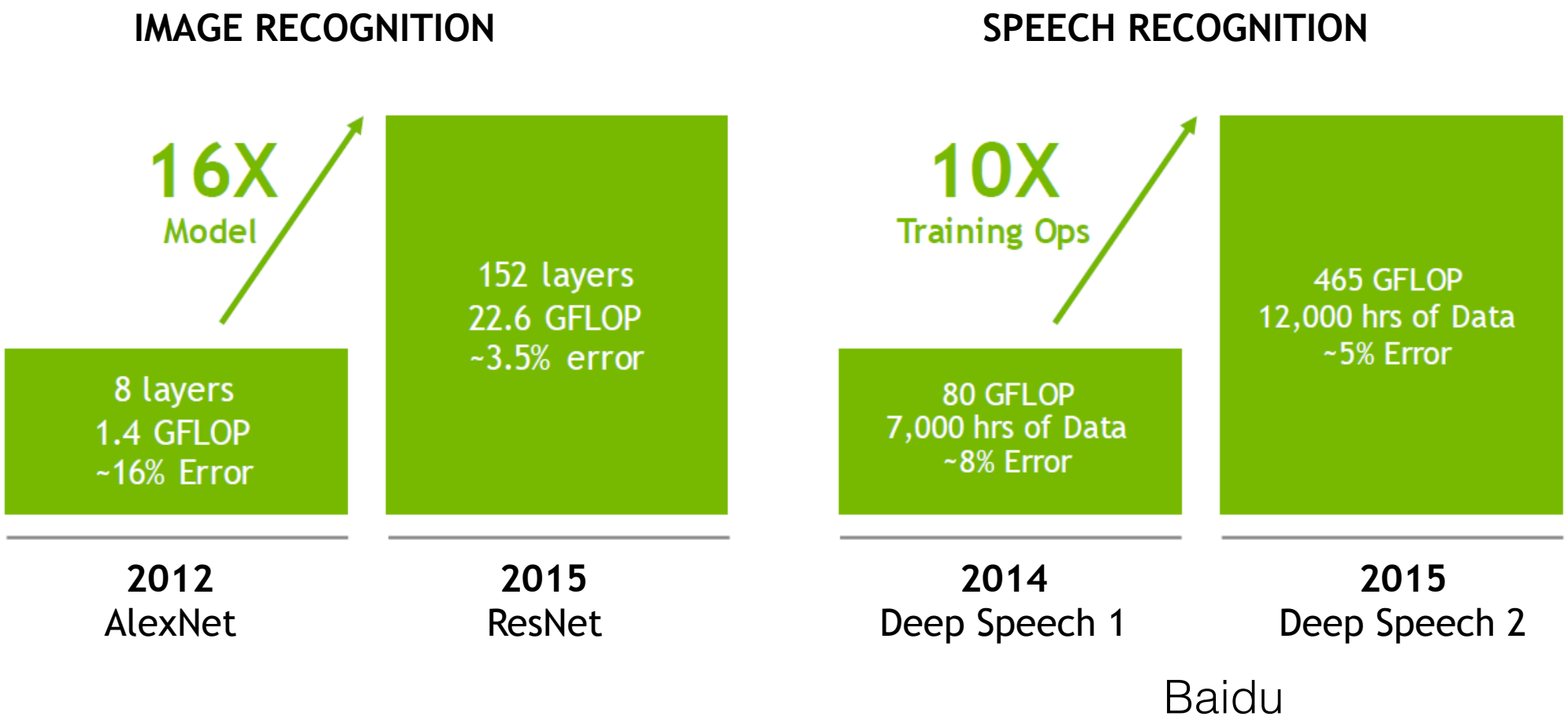
Andrew Ng, Stanford, Baidu, Coursera


“Electricity , communication, manufacturing. I think we are now in that phase where AI technology has advanced to the point where we see a clear path for it to transform multiple industries.”

Devices, machines, and things are becoming more intelligent


On-device intelligence
is paramount
Process data closest to the source, complement the cloud
Privacy
Reliability
Low latency
Efficient use of network bandwidth

Why Efficiency? Models are Getting Larger!

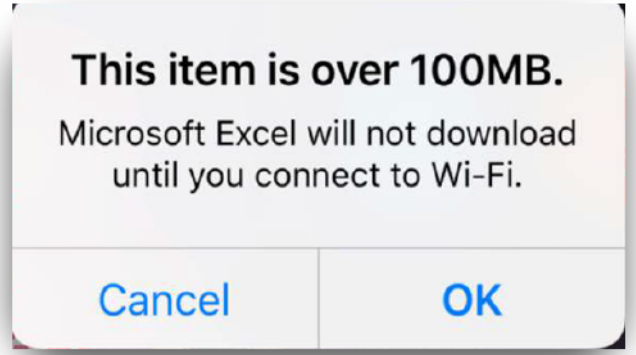


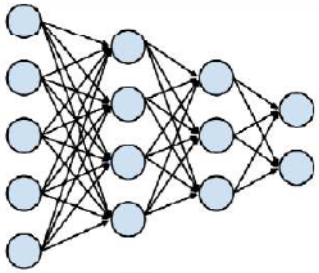
Dally, NIPS'2016 workshop on Efficient Methods for Deep Neural Networks

The First Challenge: Model Size

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

App icon is in the public domain
Phone image is licensed under CC-BY 2.0





This image is licensed under CC-BY 2.0

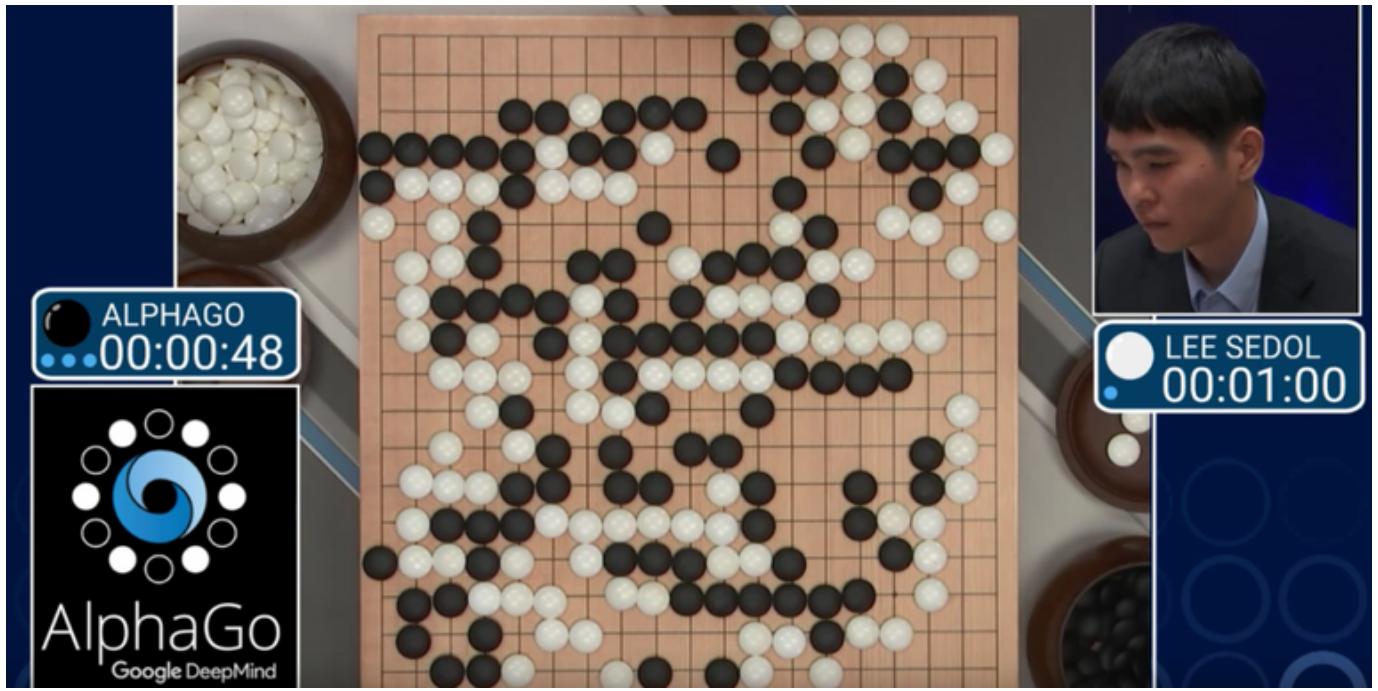
The Second Challenge: Speed

	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

Training time benchmarked with fb.resnet.torch using four M40 GPUs

The Third Challenge: Energy Efficiency

AlphaGo: 1920 CPUs and 280 GPUs,
$3000 electric bill per game



What's next?


Specialized hardware
Algorithmic advancements
Improved optimization strategies

Barcelona Supercomputing Center (BSC)

Chalmers Tekniska Hoegskola AB (CHALMERS)

Christmann Informationstechnik + Medien GmbH & Co. KG (CHR)

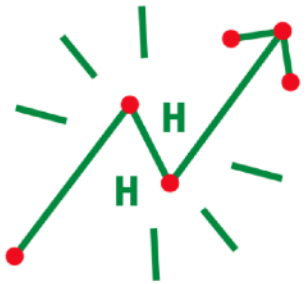
Machine Intelligence Sweden AB (MIS)

Helmholtz Centre for Infection Research (HZI)

Technische Universität Dresden (TUD)

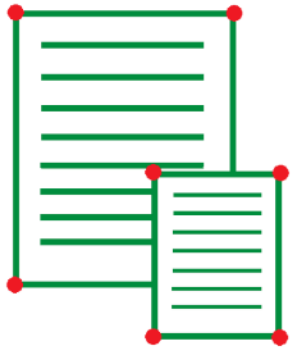
Maxeler Technologies (MAXELER)

TECHNION, Israel Institute of Technology (TECHNION)



One order of magnitude improvement in energy-efficiency for heterogeneous hardware through the use of the energy-optimized programming model and runtime.

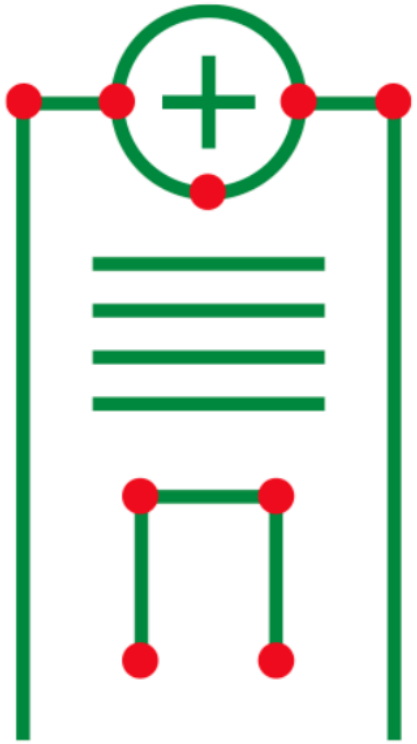
5× improvement in Mean Time to Failure through energy-efficient software-based fault tolerance.



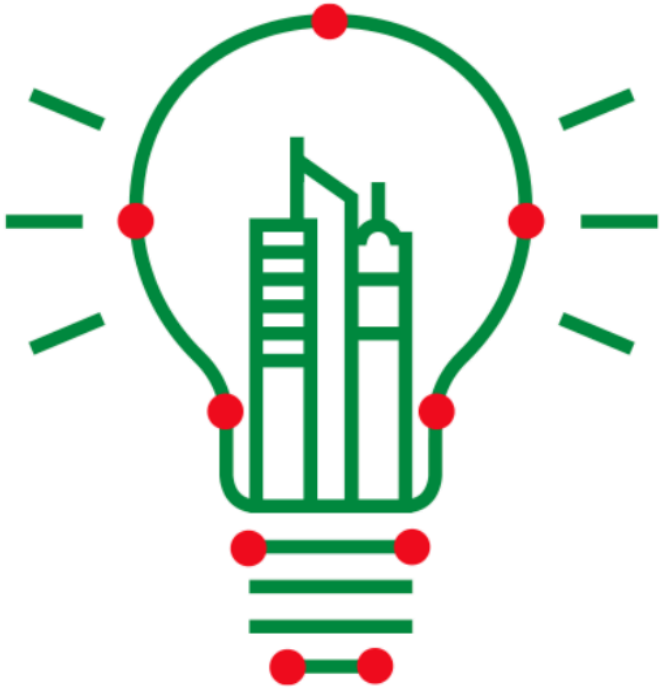
Size reduction of the trusted computing base by at least an order of magnitude.

5× increase in FPGA designer productivity through the design of novel features for hardware design using dataflow languages.

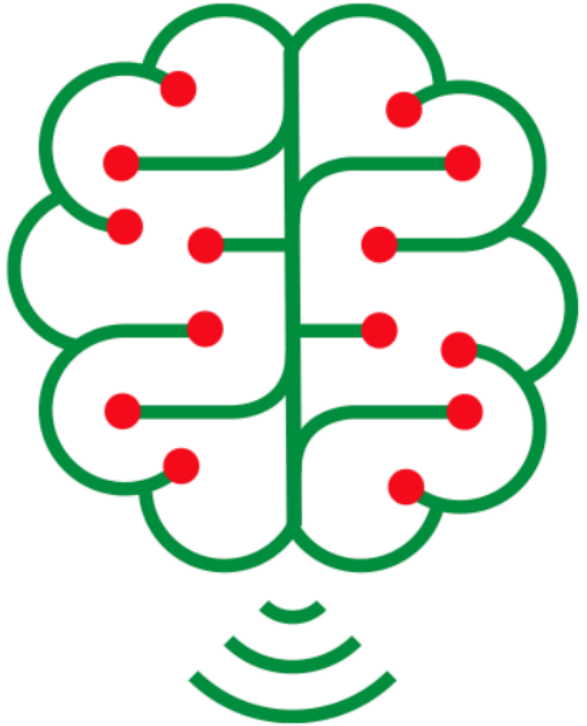
LEGaTO



Healthcare

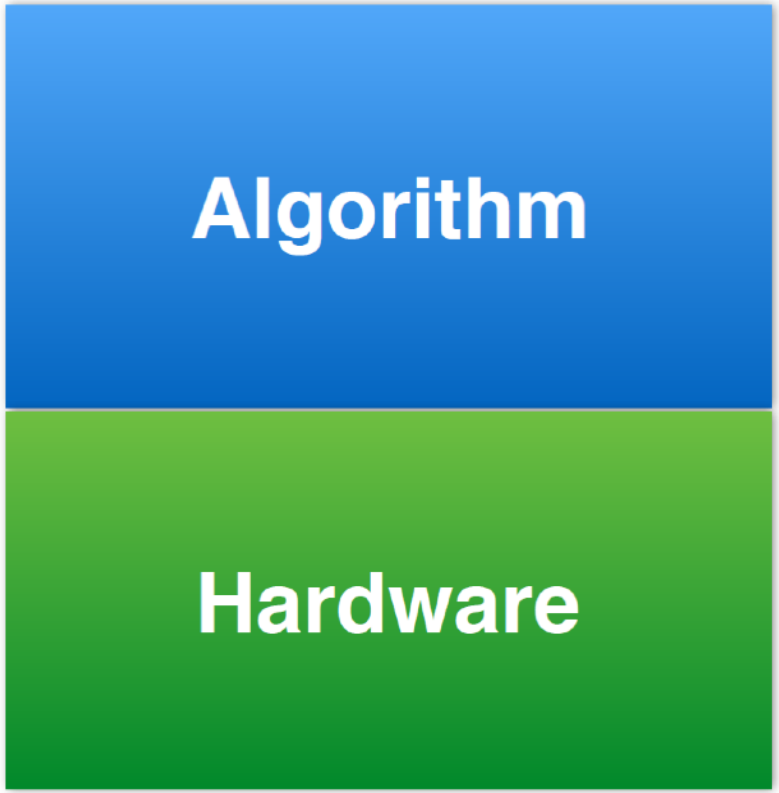


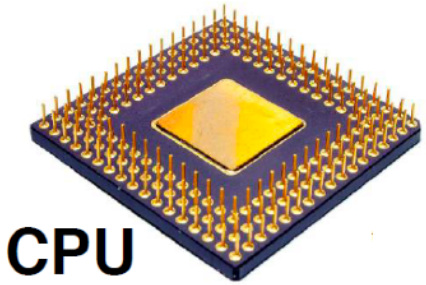
IoT for Smart Homes and Cities



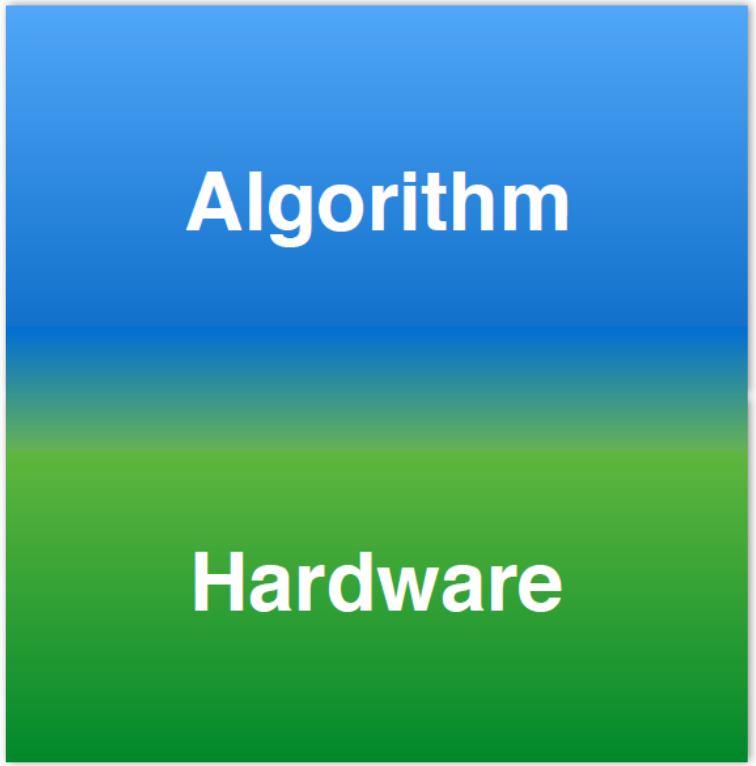
Machine Learning

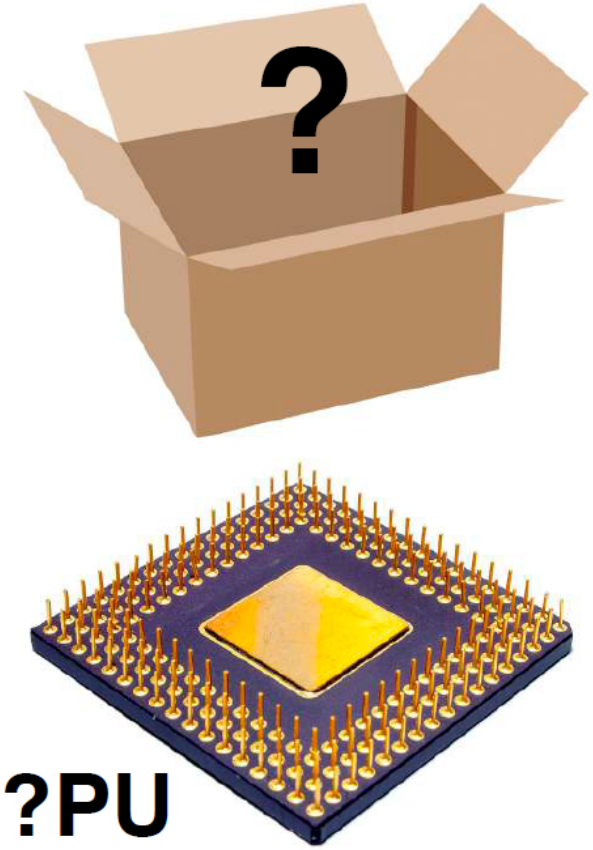
Application and Hardware as Black Boxes





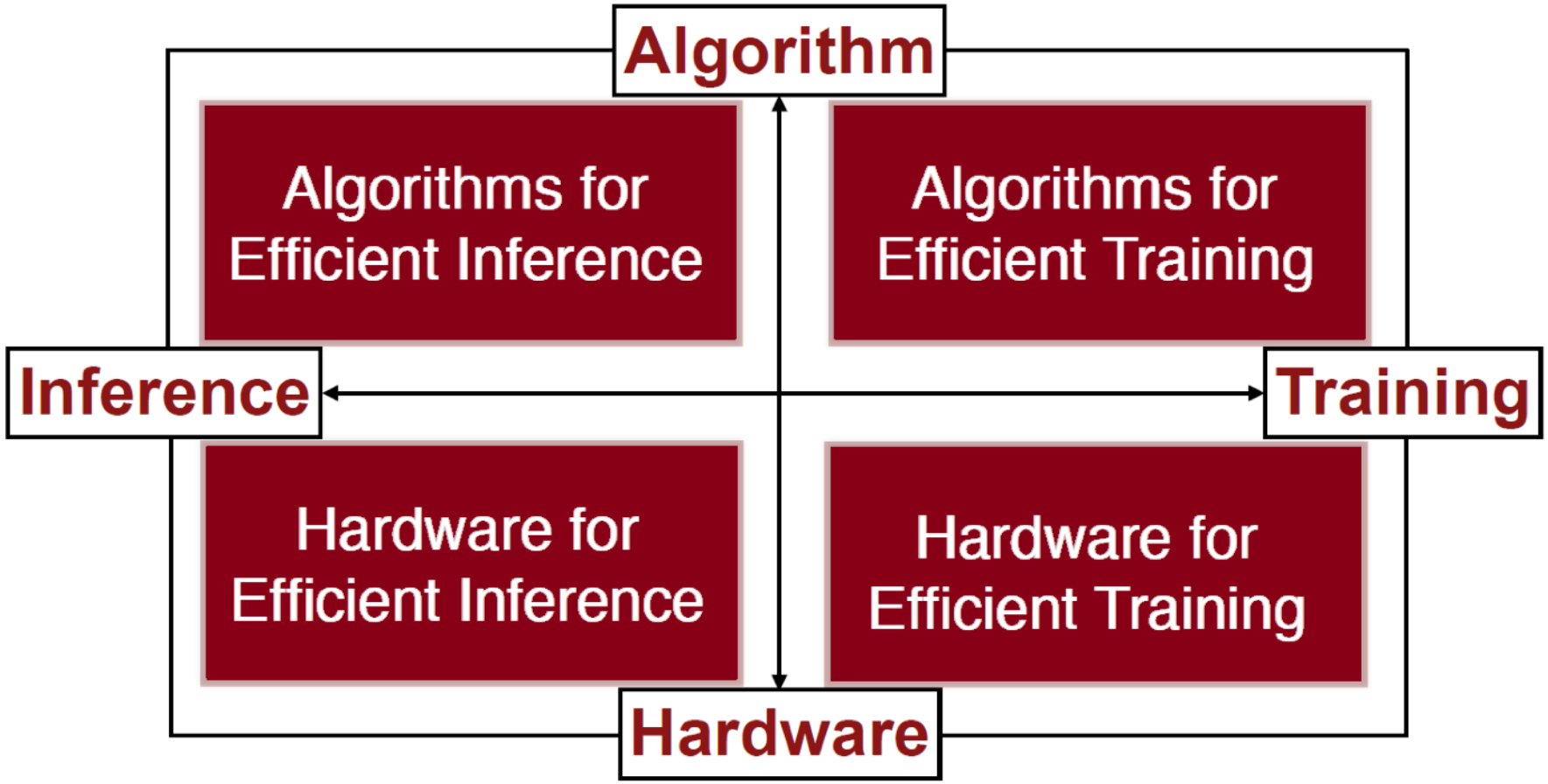
Open the Boxes!



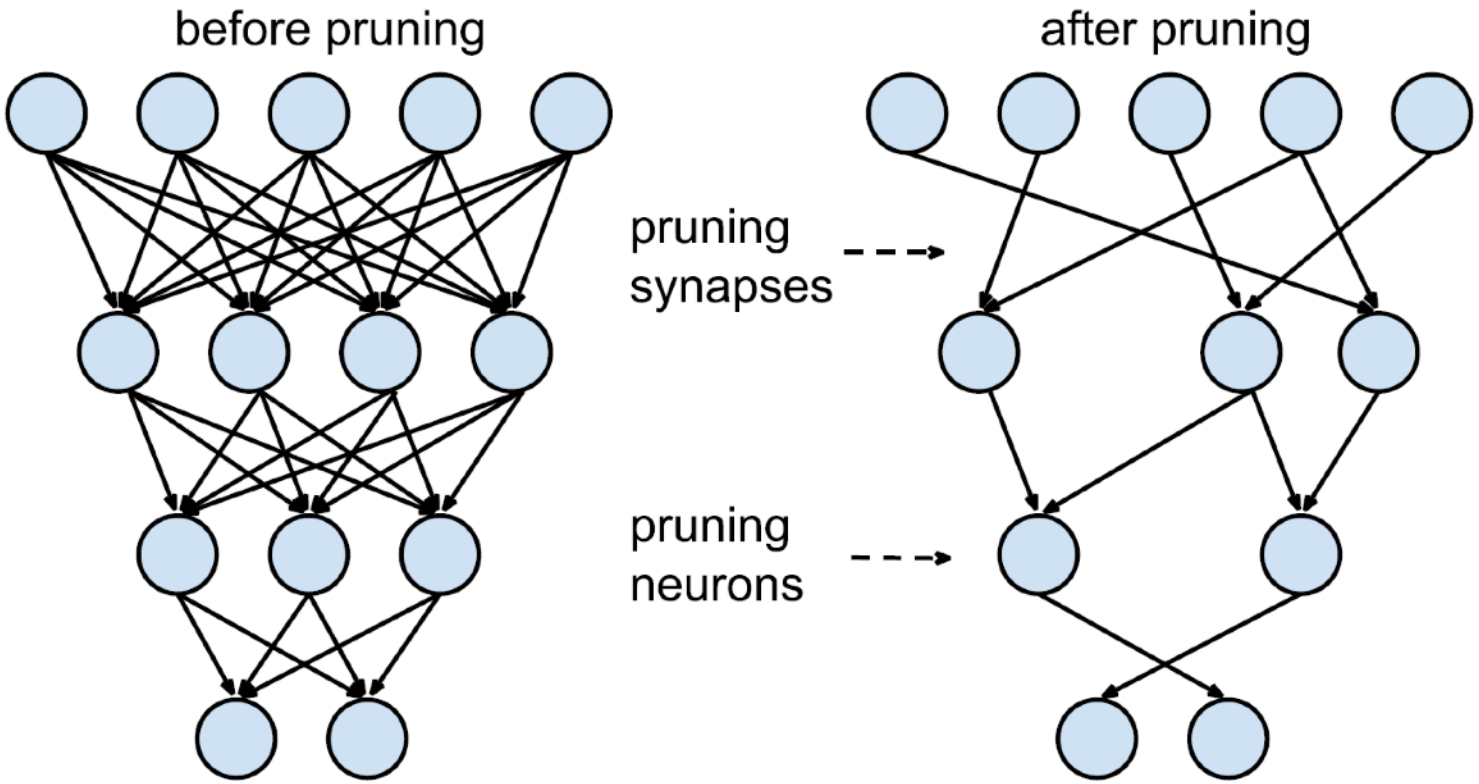


- Breaks the boundary between algorithm and hardware

Categorization of Efficient DNN Methods



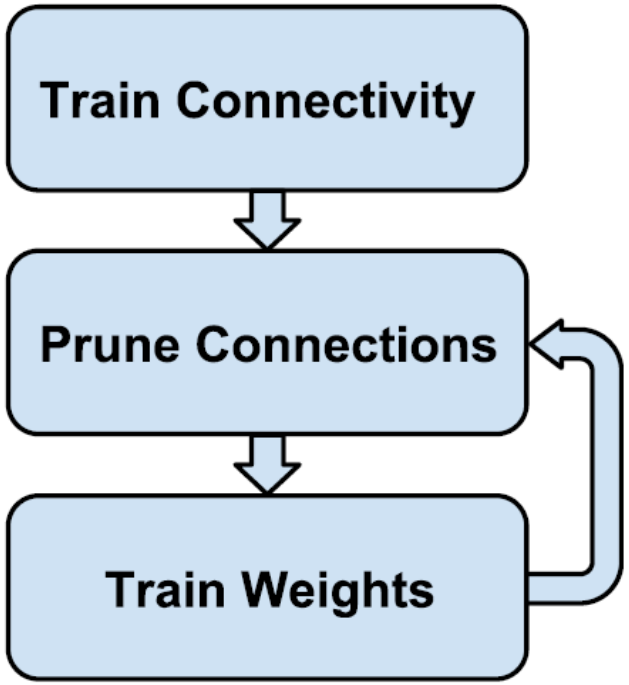
Pruning Neural Networks

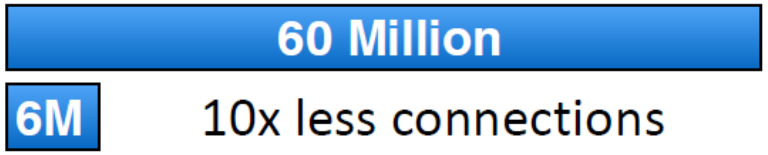


[Lecun et al. NIPS'89]

[Han et al. NIPS'15]

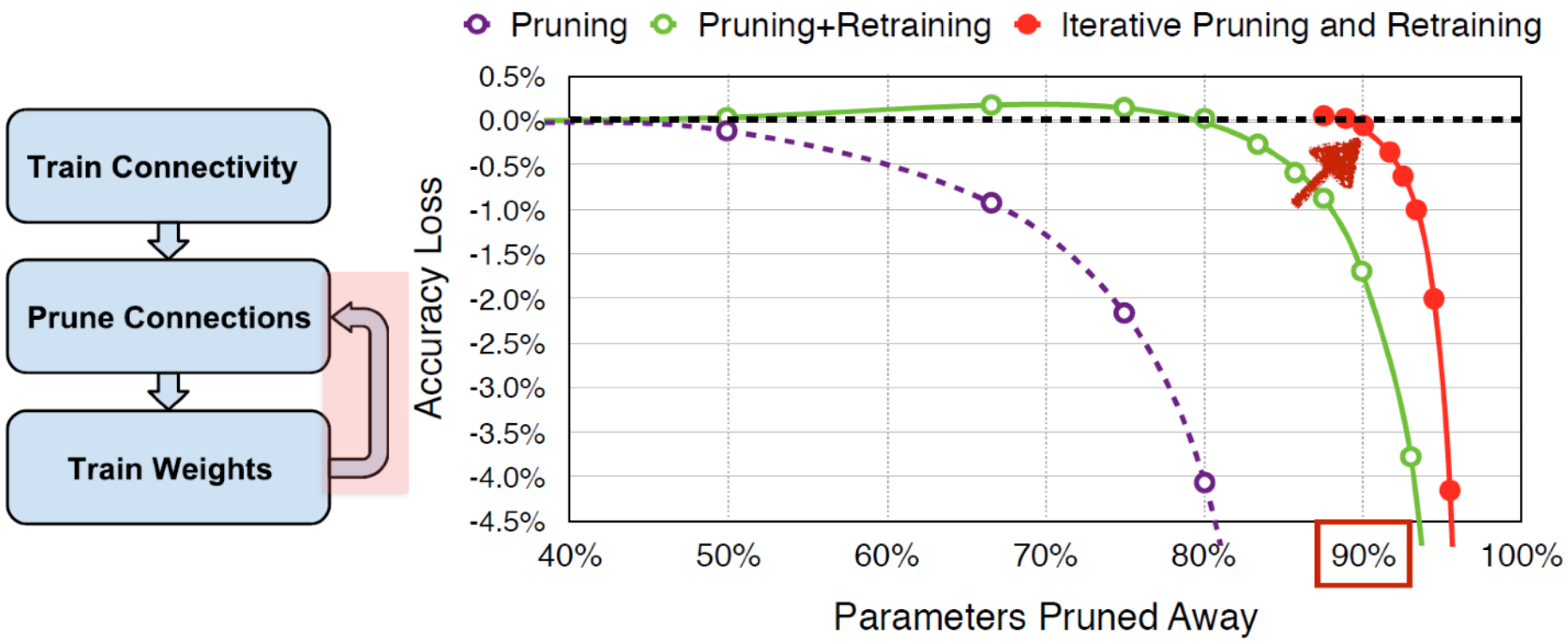
Pruning Neural Networks



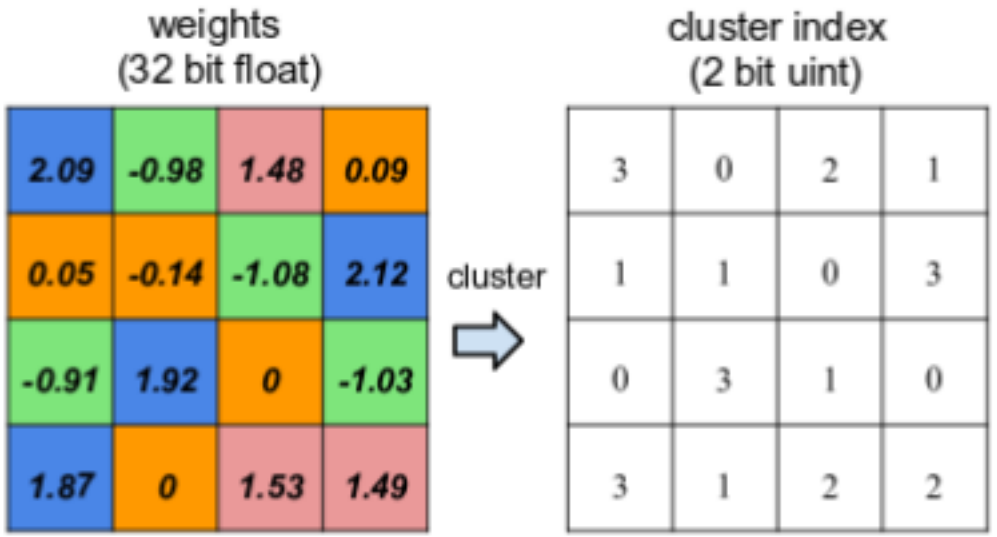


[Han et al. NIPS'15]

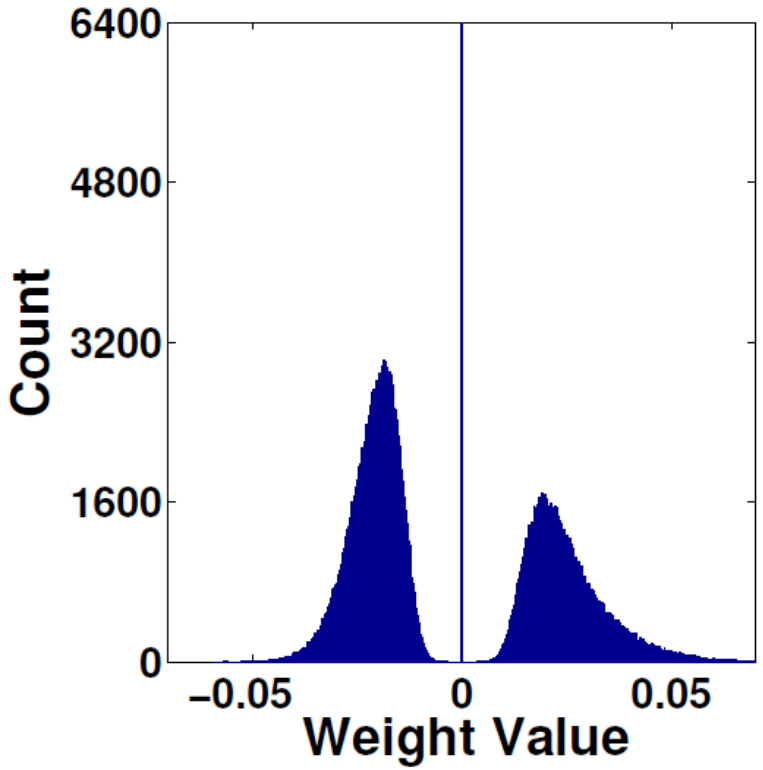
Pruning Neural Networks



Weight Quantization

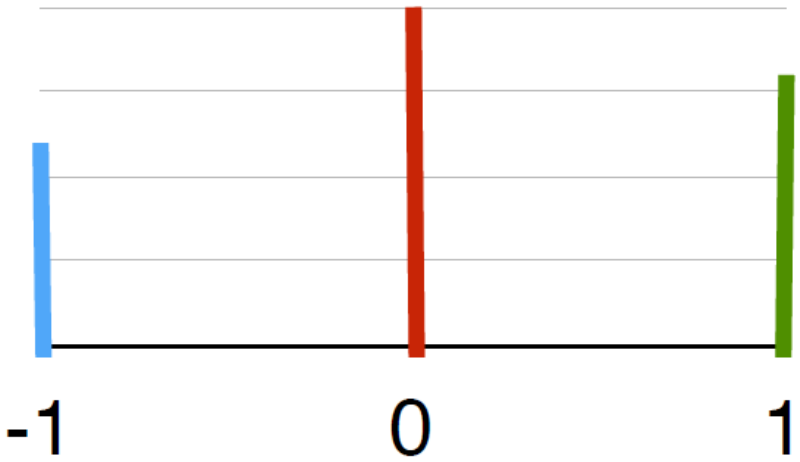


Ternary networks

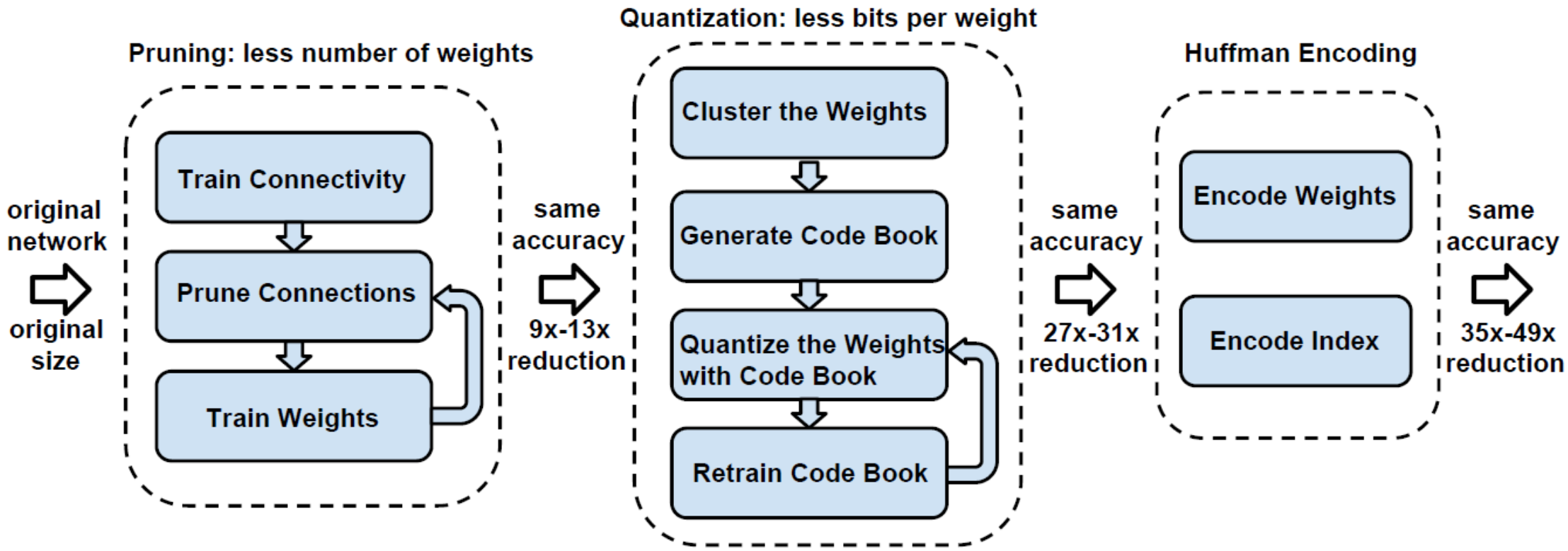


=>

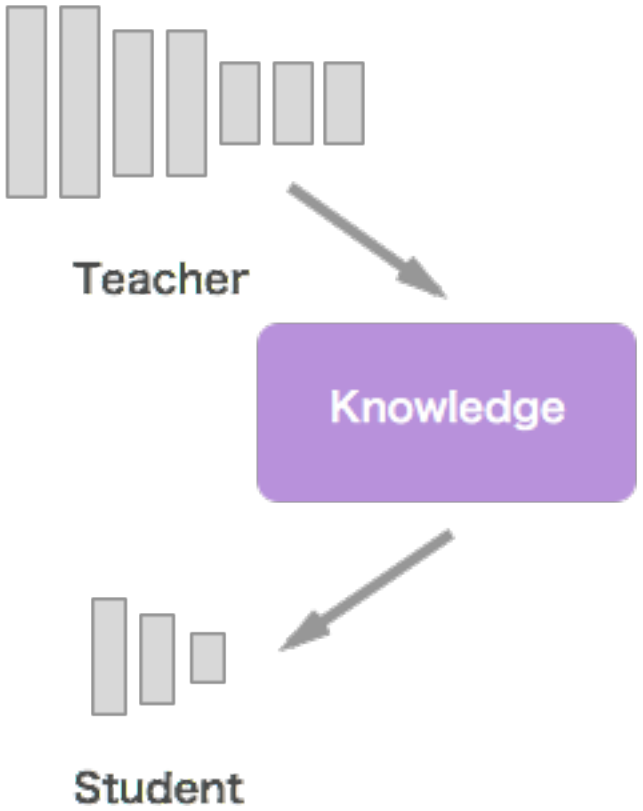
Allow weights to be zero (i.e., -w, 0, w)

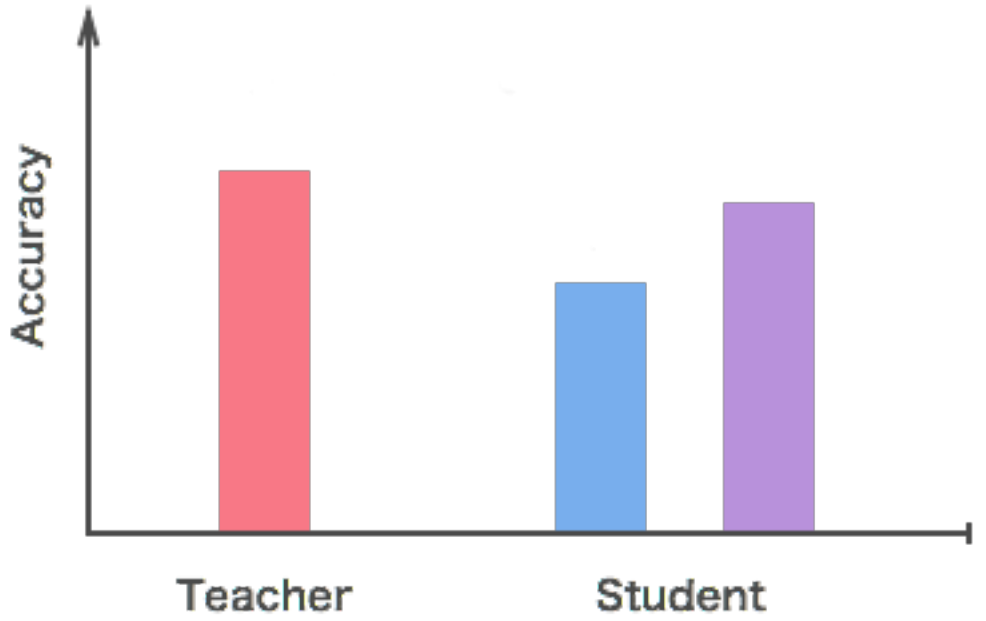


Pruning + Learned Quantization Work Together

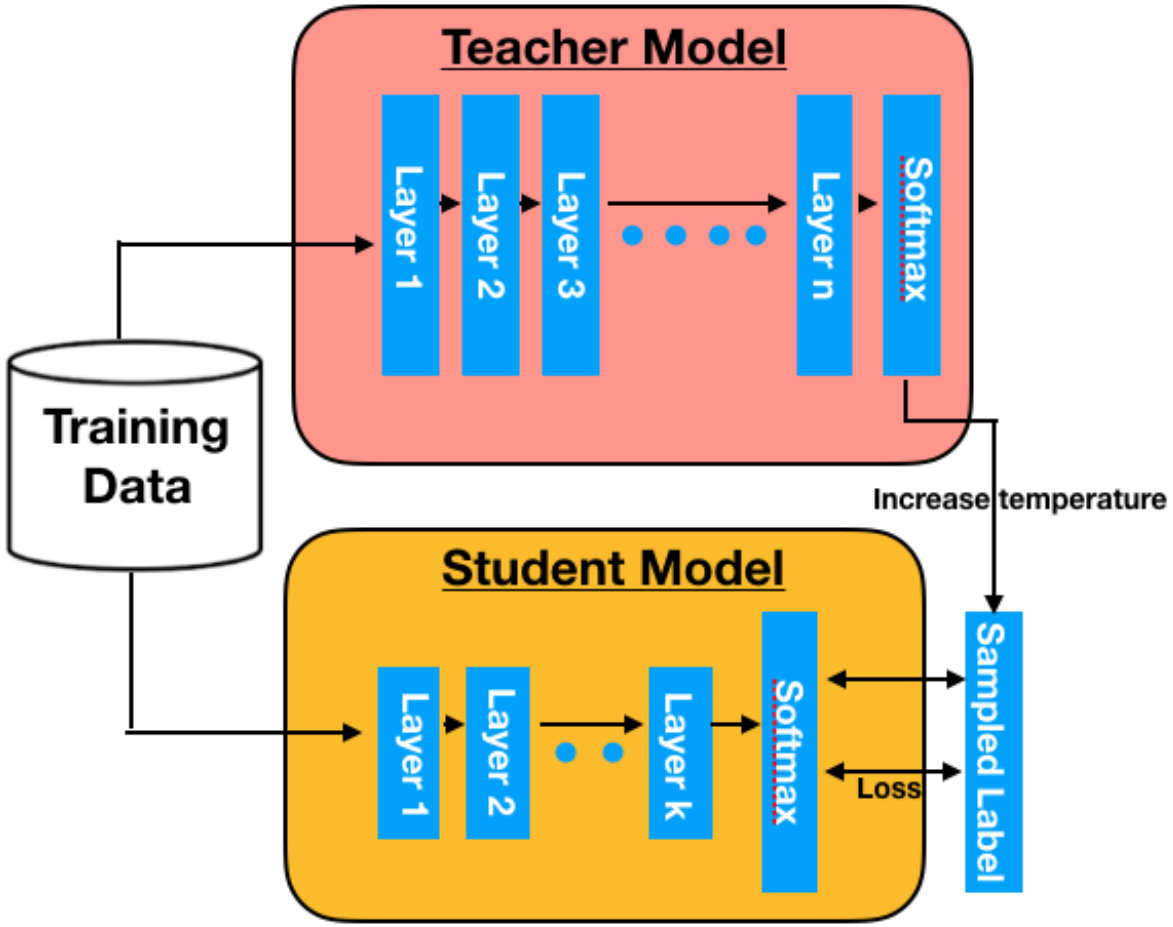


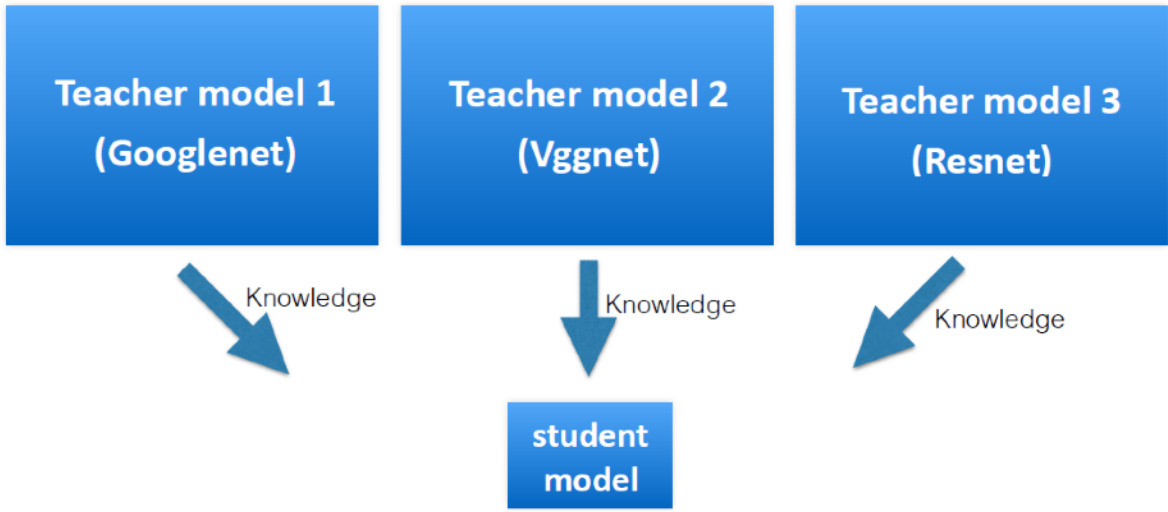
Knowledge Distillation





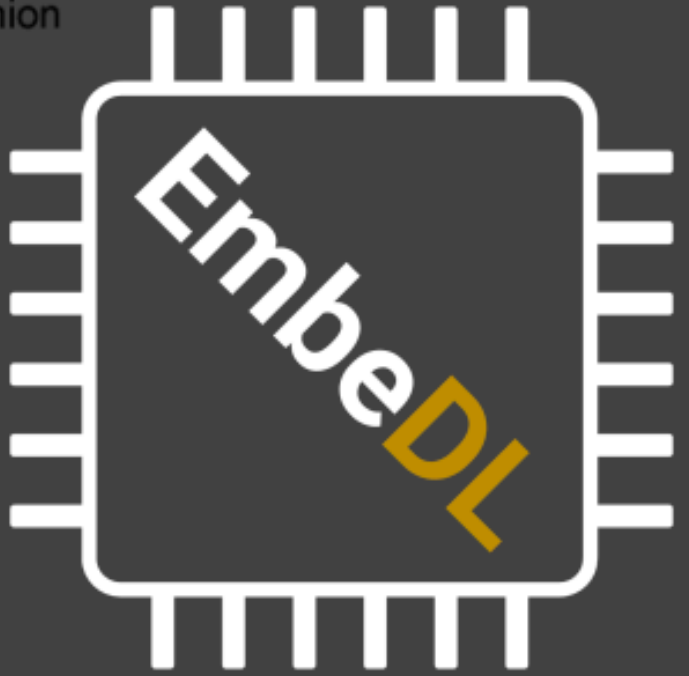
Model Distillation





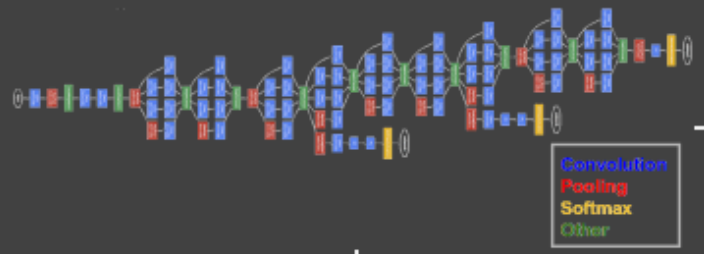
Student model has much smaller model size!

Co-funded by the Horizon 2020 programme
of the European Union



Efficient Deep Learning in Embedded Systems

Solution



CPU

GPU

FPGA

EmbeDL

- 10x speed
- 10x less energy
- 10x less memory

Hardware/Software codesign

"Energy will soon be one of the determining factors in AI. Either companies will find it too expensive to run energy hungry ML tools (such as deep learning) to power their AI engines, or the heat dissipation in edge devices will be too high to be safe. The next battleground in AI might well be a race for the most energy efficient combination of hardware and algorithms."

Max Welling ICML 2018

References

- The slides have been prepared based on
  - Stanford Deep Learning course, Lecture 15 http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture15.pdf
  - Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, Joel S. Emer, Efficient Processing of Deep Neural Networks: A Tutorial and Survey. Proceedings of the IEEE 105(12): 2295-2329 (2017)
  - and the slides at: http://www.rle.mit.edu/eems/wp-content/uploads/2017/06/ISCA-2017-Hardware-Architectures-for-DNN-Tutorial.pdf
  - Song Han, Huizi Mao, William J. Dally, Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding, ICLR 2016.