Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding

ICLR'16 best paper award

Song Han, Huizi Mao, William J. Dally

Stanford University, Stanford

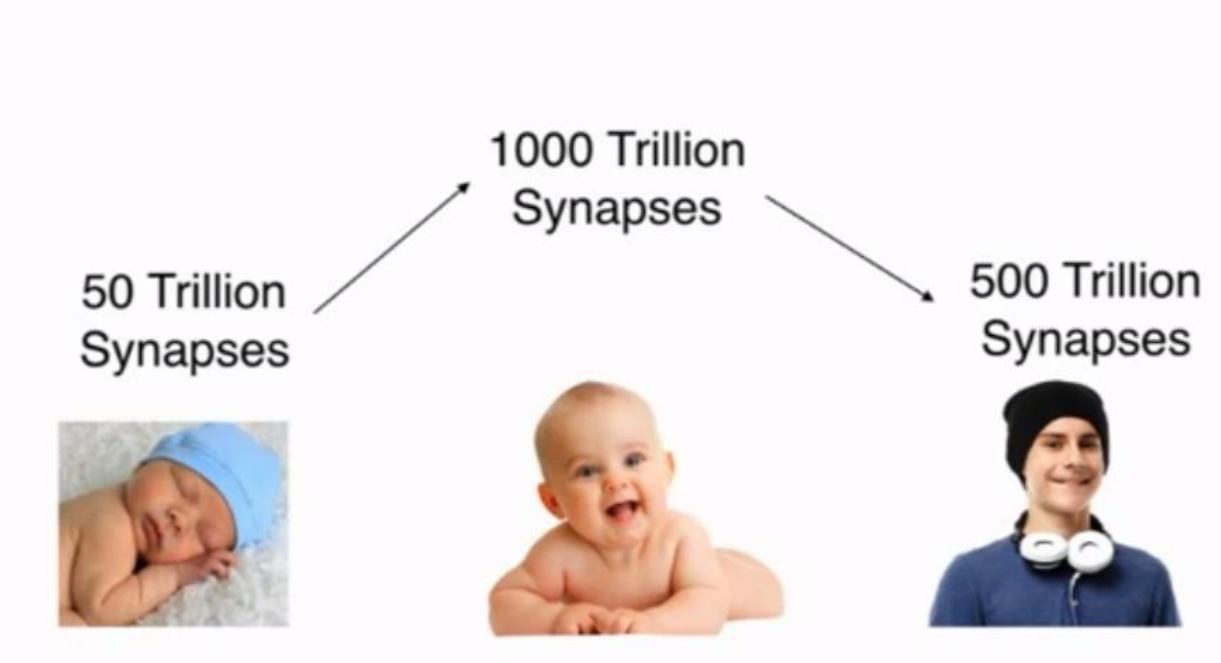
Novosibirsk State University
*THE REAL SCIENCE

Background

- Optimal brain damage. In Advances in Neural Information Processing Systems 1990.
- Learning both weights and connections for efficient neural networks. In Advances in Neural Information Processing Systems, 2015.
- Deep Compression, 2016

Novosibirsk State University
*THE REAL SCIENCE

Optimal brain damage



- During synaptic pruning, the brain eliminates extra synapse removing connections in the brain that are no longer needed
- the brain is plastic, maintains efficient brain function as we get older and learn new complex information

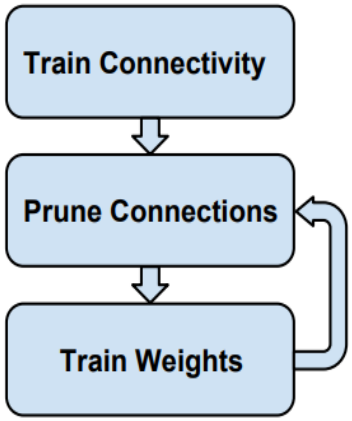
Novosibirsk State University
*THE REAL SCIENCE

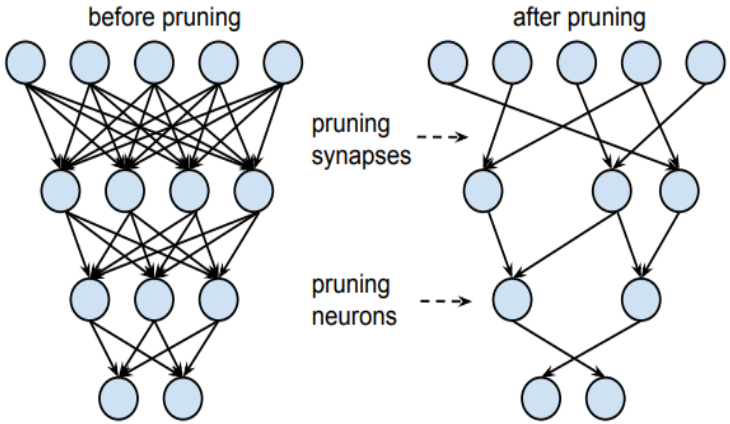
Algorithm

1. Choose a reasonable network architecture
2. Train the network until a reasonable solution is obtained
3. Compute the second derivatives hu for each parameter
4. Compute the saliencies for each parameter: $L_q = \frac{1}{2}\frac{w_q^2}{[\boldsymbol{H}^{-1}]_{qq}}$
5. Sort the parameters by saliency and delete some low-saliency parameters
6. Iterate to step 2

Novosibirsk State University
*THE REAL SCIENCE

Learning both weights and connections
for efficient neural networks

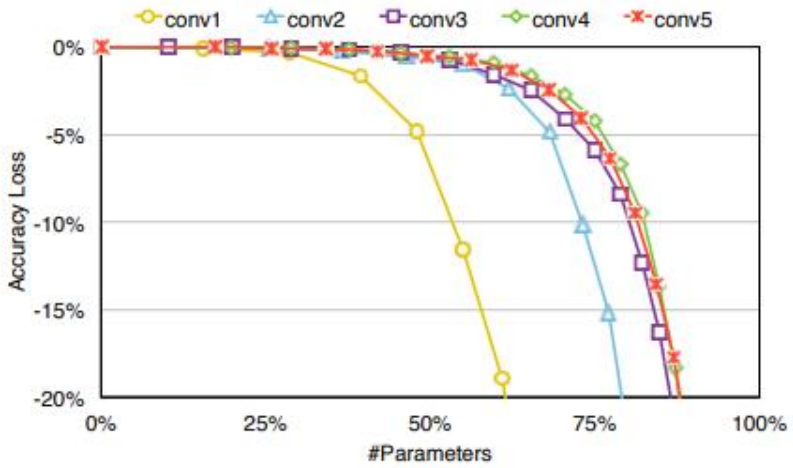


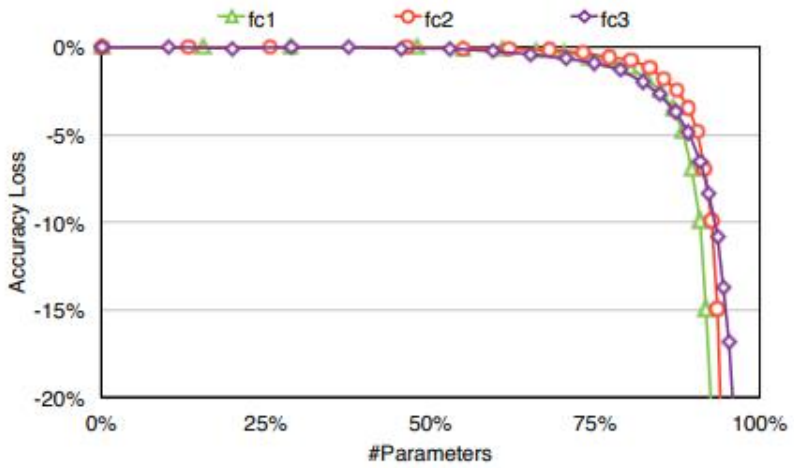


- Learning the connectivity via normal network training.
- Unlike conventional training, however, we are not learning the final values of the weights, but rather we are learning which connections are important.

Novosibirsk State University
*THE REAL SCIENCE

Learning both weights and connections
Results



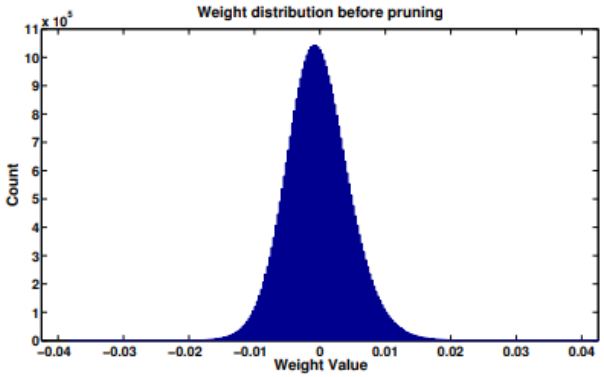


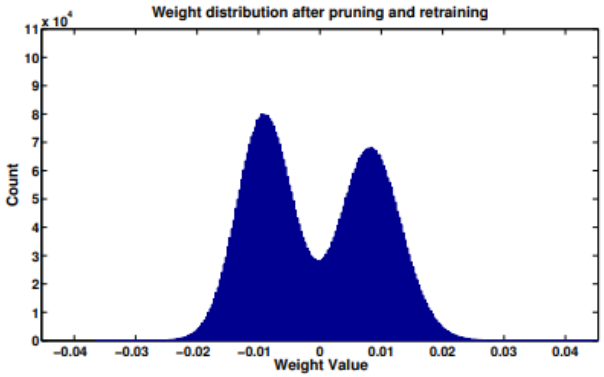
The CONV layers (on the left) are more sensitive to pruning than the fully connected layers (on the right)

Novosibirsk State University
*THE REAL SCIENCE

Learning both weights and connections
Results

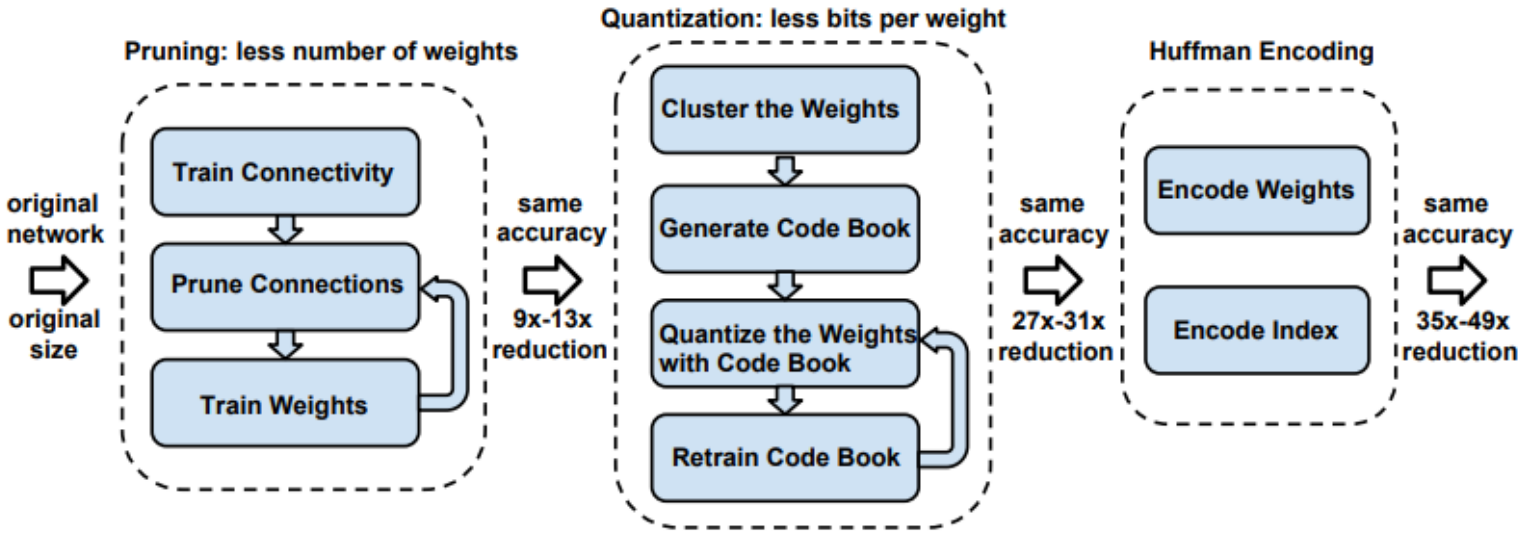
Network	Top-1 Error	Top-5 Error	Parameters	Compression Rate
Baseline Caffemodel [26]	42.78%	19.73%	61.0M	1×
Data-free pruning [28]	44.40%	-	39.6M	1.5×
Fastfood-32-AD [29]	41.93%	-	32.8M	2×
Fastfood-16-AD [29]	42.90%	-	16.4M	3.7×
Collins & Kohli [30]	44.40%	-	15.2M	4×
Naive Cut	47.18%	23.23%	13.8M	4.4×
SVD [12]	44.02%	20.56%	11.9M	5×
Network Pruning	**42.77%**	**19.67%**	**6.7M**	**9×**





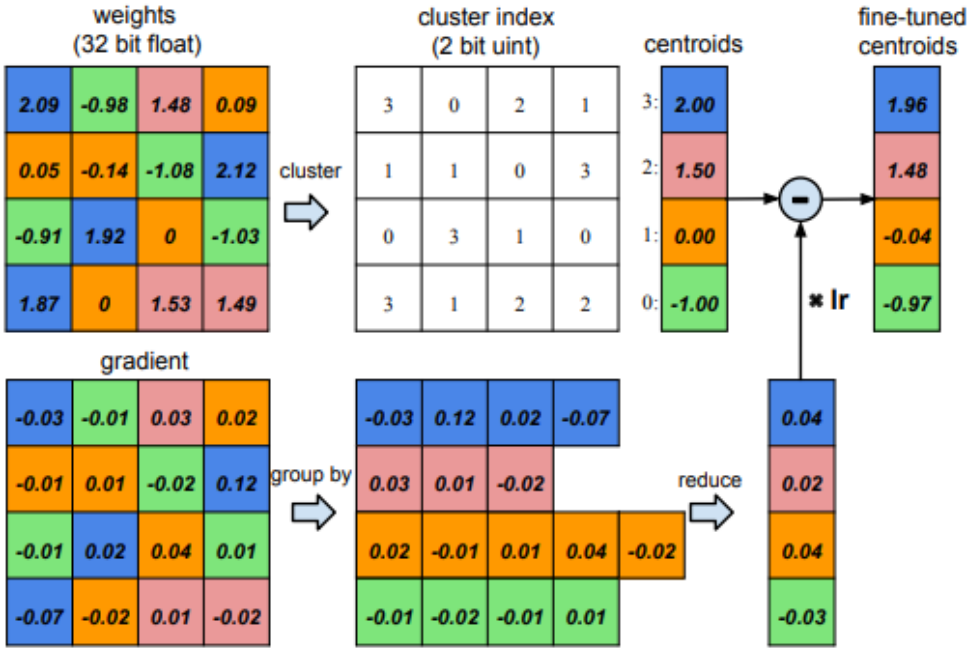
Novosibirsk State University
*THE REAL SCIENCE

Deep Compression



Novosibirsk State University
*THE REAL SCIENCE

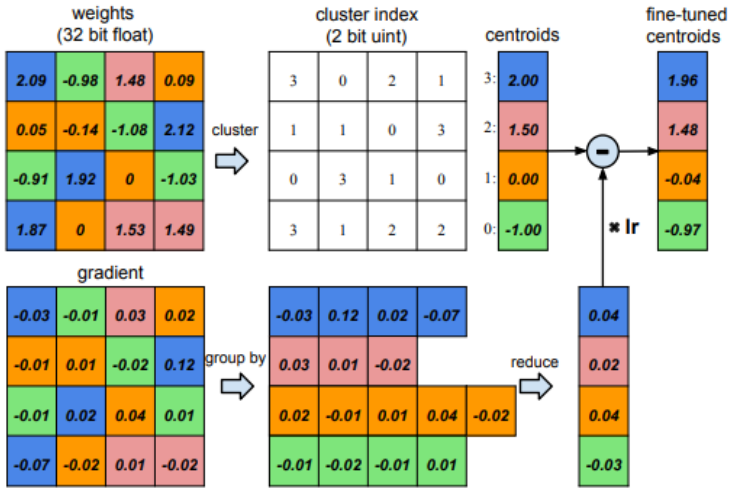
Deep Compression



Weight sharing by scalar quantization (top) and centroids fine-tuning (bottom).

Novosibirsk State University
*THE REAL SCIENCE

Deep Compression



- We use k-means clustering to identify the shared weights for each layer of a trained network, so that all the weights that fall into the same cluster will share the same weight.
- Weights are not shared across layers. We partition n original weights $W = \{w_1, w_2, \ldots, w_n\}$ into k clusters $C = \{c_1, c_2, \ldots, c_k\}$, $n \gg k$, so as to minimize the within-cluster sum of squares (WCSS):

$$\underset{C}{arg\,min} = \sum_{i=1}^{K} \sum_{w \in c_i} |w - c_i|^2$$

Novosibirsk State University
*THE REAL SCIENCE

Deep Compression
Results

Table 1: The compression pipeline can save 35× to 49× parameter storage with no loss of accuracy.

Network	Top-1 Error	Top-5 Error	Parameters	Compress Rate
LeNet-300-100 Ref	1.64%	-	1070 KB	
LeNet-300-100 Compressed	1.58%	-	**27 KB**	**40**×
LeNet-5 Ref	0.80%	-	1720 KB	
LeNet-5 Compressed	0.74%	-	**44 KB**	**39**×
AlexNet Ref	42.78%	19.73%	240 MB	
AlexNet Compressed	42.78%	19.70%	**6.9 MB**	**35**×
VGG-16 Ref	31.50%	11.32%	552 MB	
VGG-16 Compressed	31.17%	10.91%	**11.3 MB**	**49**×

Novosibirsk State University
THE REAL SCIENCE

References

[30] Y. Le Cun, J. S. Denker, S. A. Sola, and T. B. Laboratories, "Optimal Brain Damage," pp. 598–605.

[31] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both Weights and Connections for Efficient Neural Networks," pp. 1–9, 2015.

[33] S. Han, H. Mao, and W. J. Dally, "DEEP COMPRESSION: COMPRESSING DEEP NEURAL NETWORKS WITH PRUNING , TRAINED QUANTIZATION," pp. 1–14, 2016.

NOVOSIBIRSK STATE UNIVERSITY
*The real science

MOBILENETS FOR CROP DISEASE RECOGNITION

Munyaradzi Talent Njera

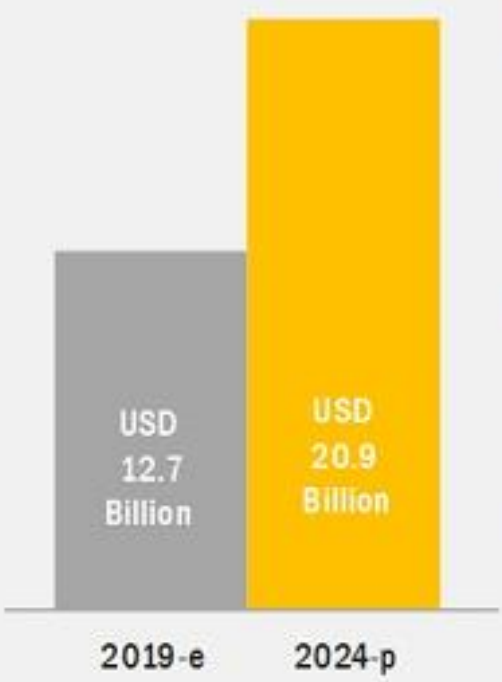
Wednesday 13th May, 2020

Prof E. Pavlovsky

Новосибирский государственный университет

Why Agriculture

Attractive Opportunities in Agriculture IoT Market



CAGR

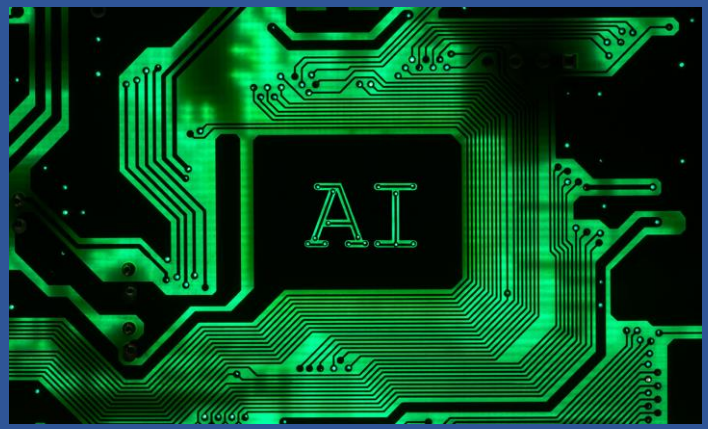
10.4%

- The agriculture IoT market is expected to be worth USD 20.9 billion by 2024—growing at a CAGR of 10.4% during 2019–2024.
- Increase Increasing adoption of Internet of Things (IoT) and Artificial Intelligence (AI) technology by farmers and growers, focus on livestock monitoring and disease detection to improve farming efficiency, and rising demand for agricultural production owing to increasing population are the major drivers for this market.
- Advent of Big Data in agriculture farm, integration of smartphones with hardware devices and software applications and rise in use of unmanned aerial vehicles (UAVs)/drones in precision farming would create huge growth opportunities for agriculture IoT market.

Novosibirsk State University
*THE REAL SCIENCE

What is the Problem

1. Edge Devices and AI
2. Sustainability of AI, The need for Green AI



Novosibirsk State University
*THE REAL SCIENCE

Edge Devices & Deep Learning

- Mobile devices are battery constrained, making power hungry applications such as deep neural networks hard to deploy.
- Energy consumption is dominated by memory access. Under 45nm CMOS technology, a 32 bit floating point add consumes 0.9pJ, a 32bit SRAM cache access takes 5pJ, while a 32bit DRAM memory access takes 640pJ, which is 3 orders of magnitude of an add operation.
- Large networks do not fit in on-chip storage and hence require the more costly DRAM accesses.
- Running a 1 billion connection neural network, for example, at 20fps would require (20Hz)(1G)(640pJ) = 12.8W just for DRAM access - well beyond the power envelope of a typical mobile device.
- The goal is to reduce the storage and energy required to run inference on such large networks so they can be deployed on mobile devices.

Sustainability of Deep Learning (Green AI)

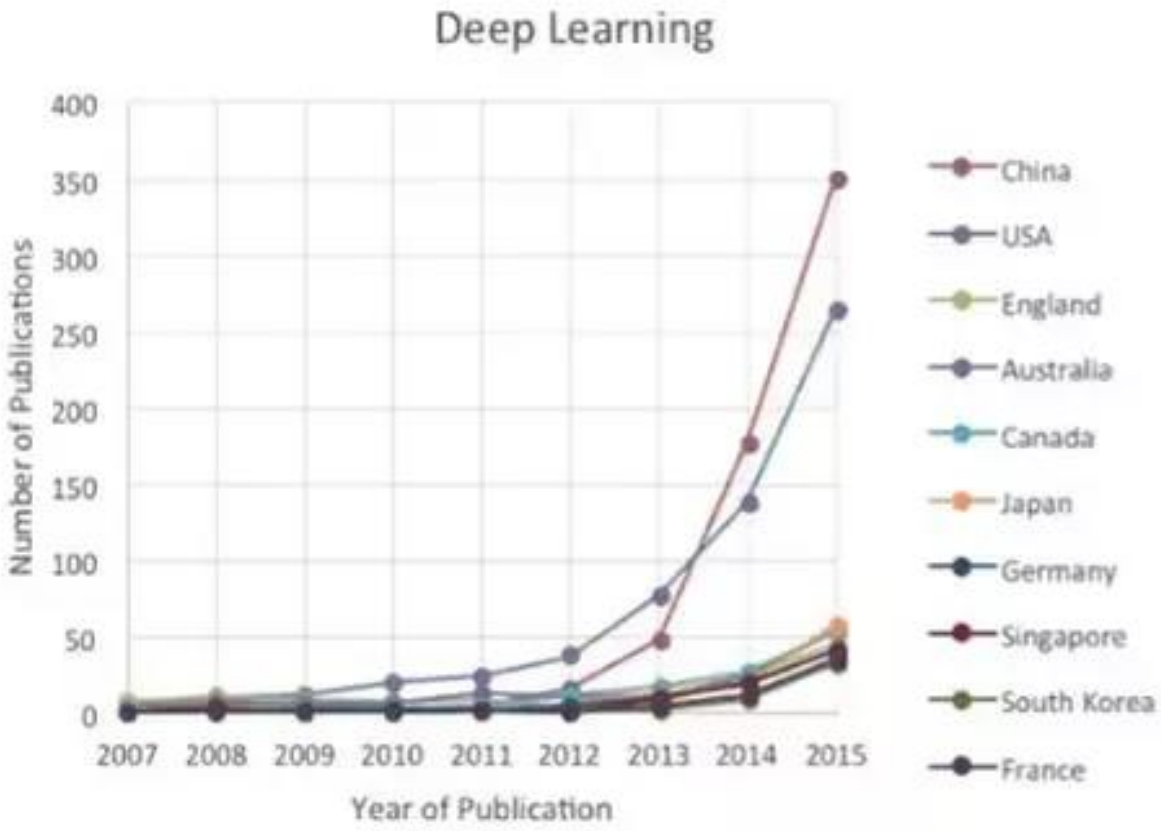
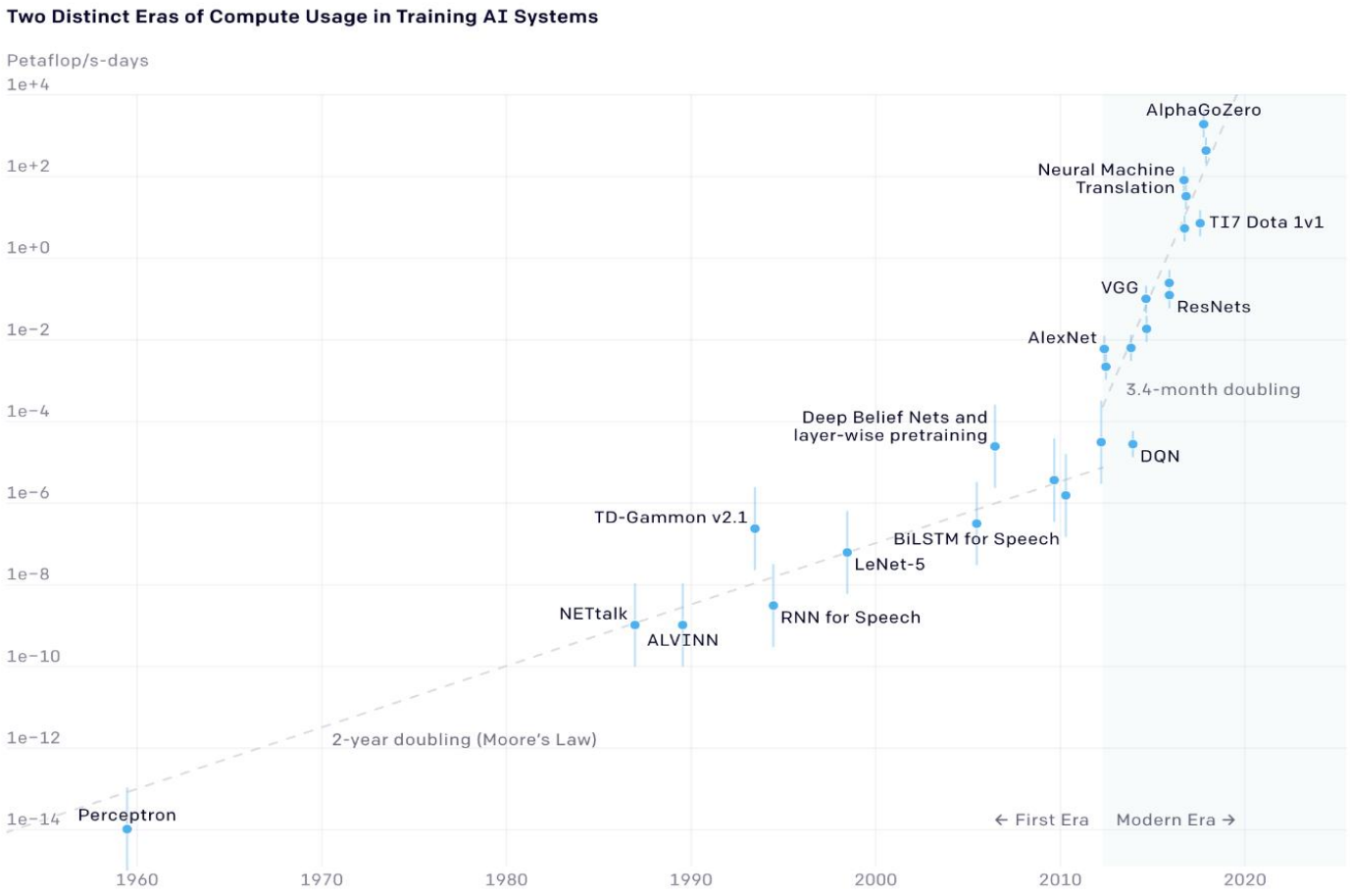


Figure 1: Journal articles mentioning "deep learning" or "deep neural network", by nation.[62]

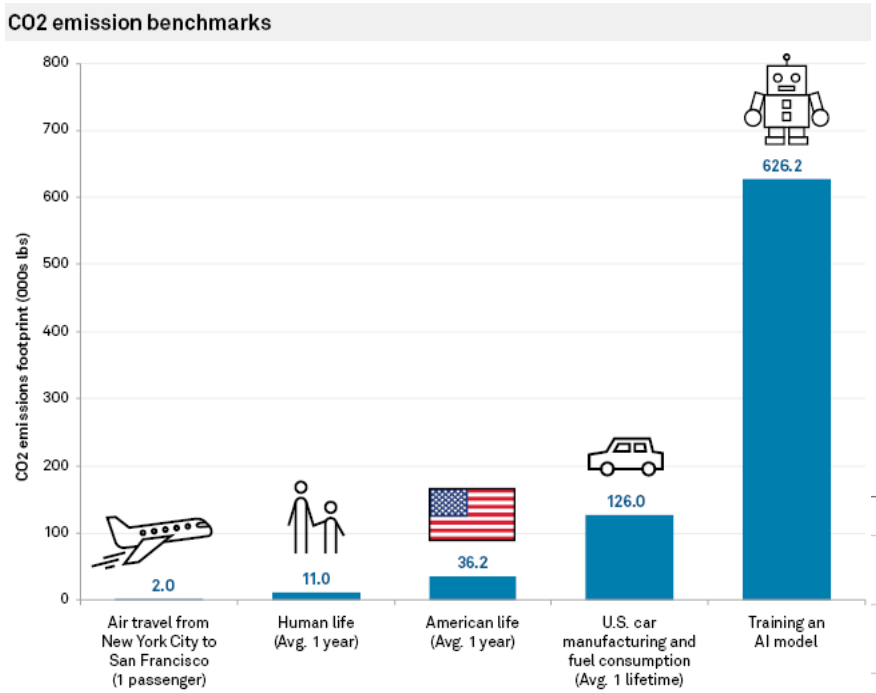
Novosibirsk State University
*THE REAL SCIENCE

The need for Green AI



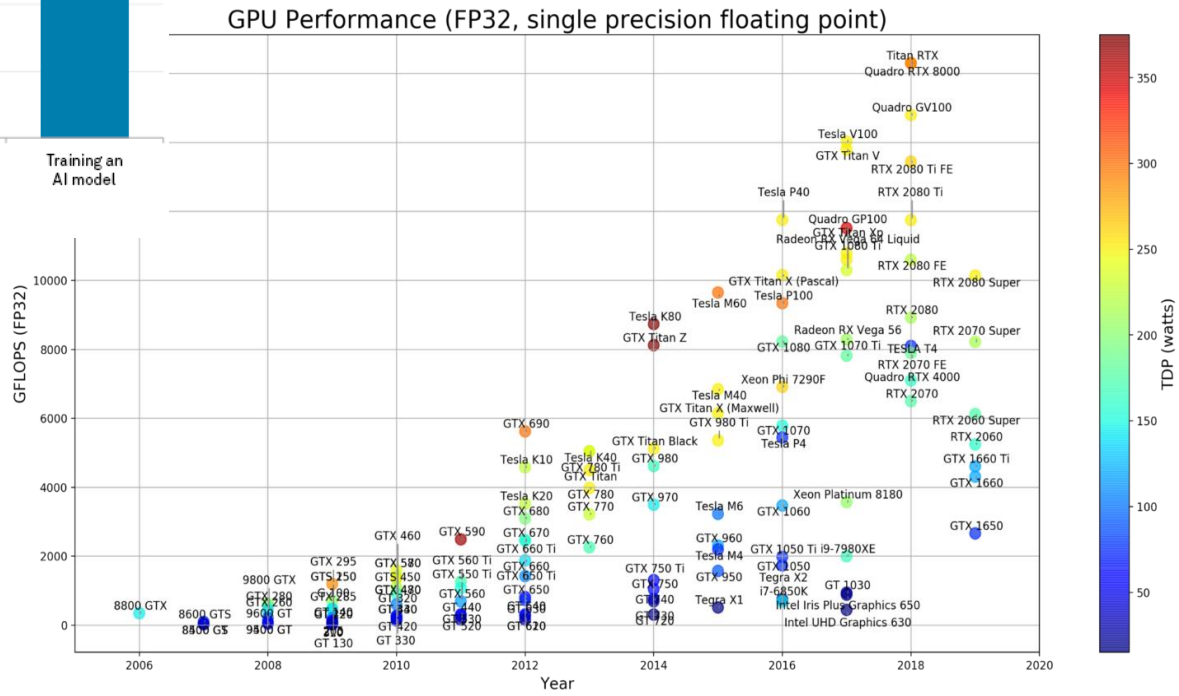
Novosibirsk State University
THE REAL SCIENCE

The need for Green AI



Researchers at the University of Massachusetts

- training process can emit 626,000 pounds of carbon dioxide,
- almost 5x the lifetime emissions of an average car,
- 300 round-trip flights between New York and San Francisco 4700 km



Contribution & Discussion

- Efficiency measures
- AI solution for crop disease using mobilenets
- Reproduce findings using a different dataset

Discussion

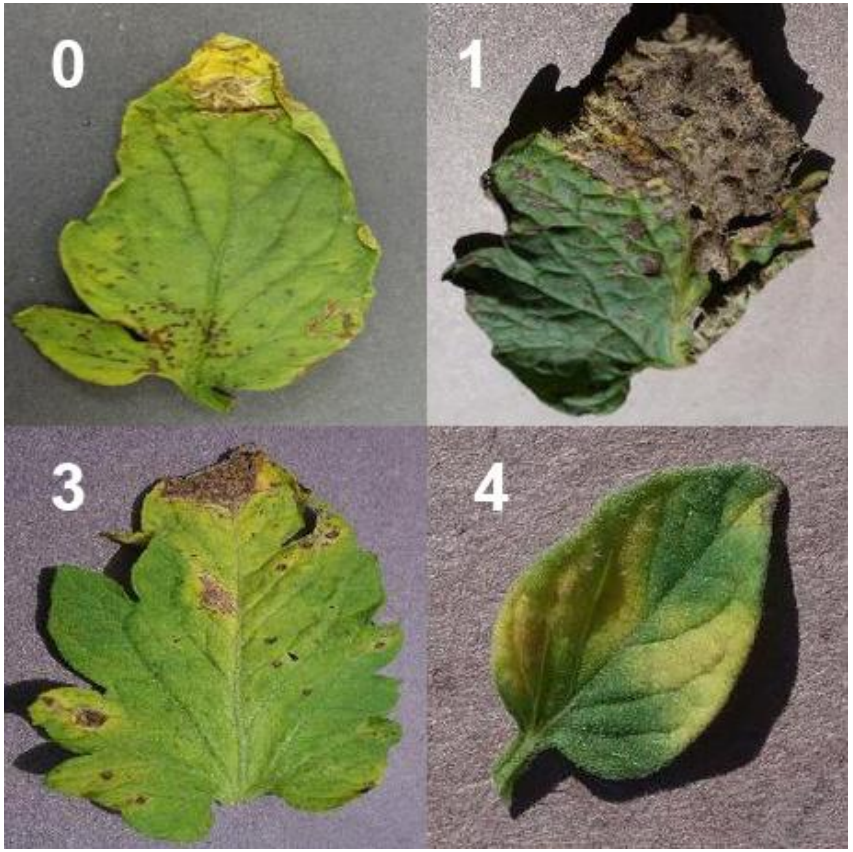
- Can AI based on these modern architectures build AGI which is sustainable? Can quantum computing reach such levels of efficient computing for future neural networks?

Methodology

- Dataset
- Model Selection
- Model Compression
- Efficiency Measurement
- Deployment

Novosibirsk State University
*THE REAL SCIENCE

Dataset



Novosibirsk State University
*THE REAL SCIENCE

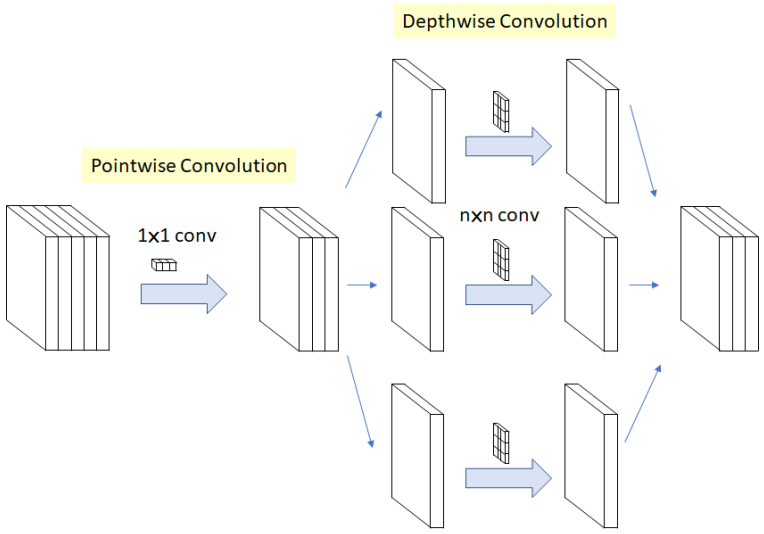
Model Selection

Criteria:

1. Pointwise and depthwise convolution concept
2. Number of parameters less than 10mil

Selected :

1. SqueezeNet
2. MobileNet
3. EfficientNet
4. NasNetMobile
5. ResNet50



Novosibirsk State University
*THE REAL SCIENCE

Depthwise & Pointwise Convolution Concept

$$Conv(W, y)_{(i,j)} = \sum_{k,l,m}^{K,L,M} W_{(k,l,m)} * \ y_{(i+k,j+l,m)}$$

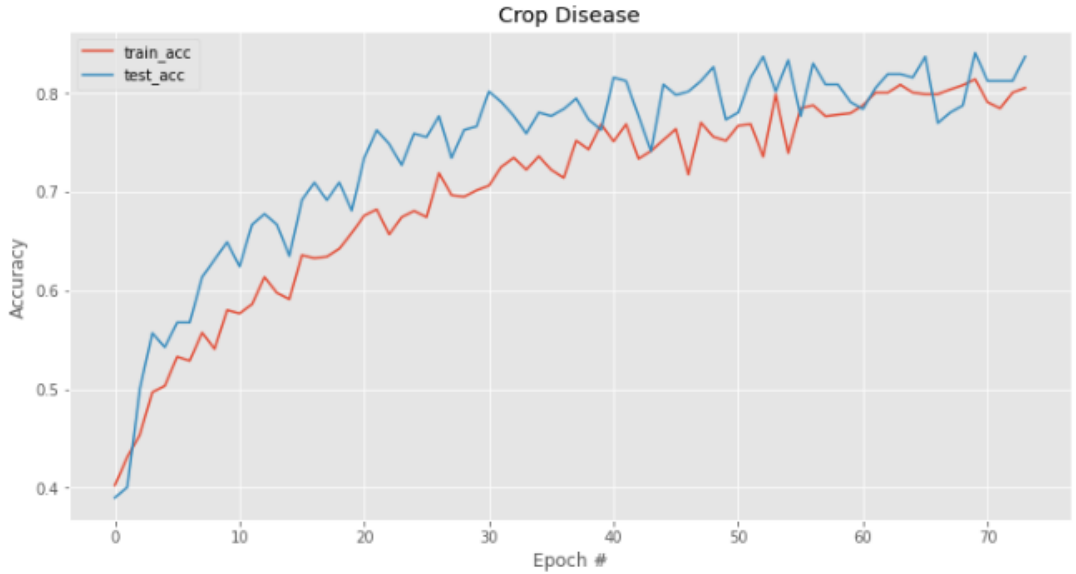
$$Pointwise\ Conv\ (W, y)_{(i,j)} = \sum_{m}^{M} W_m \ * \ y_{(i,j,m)}$$

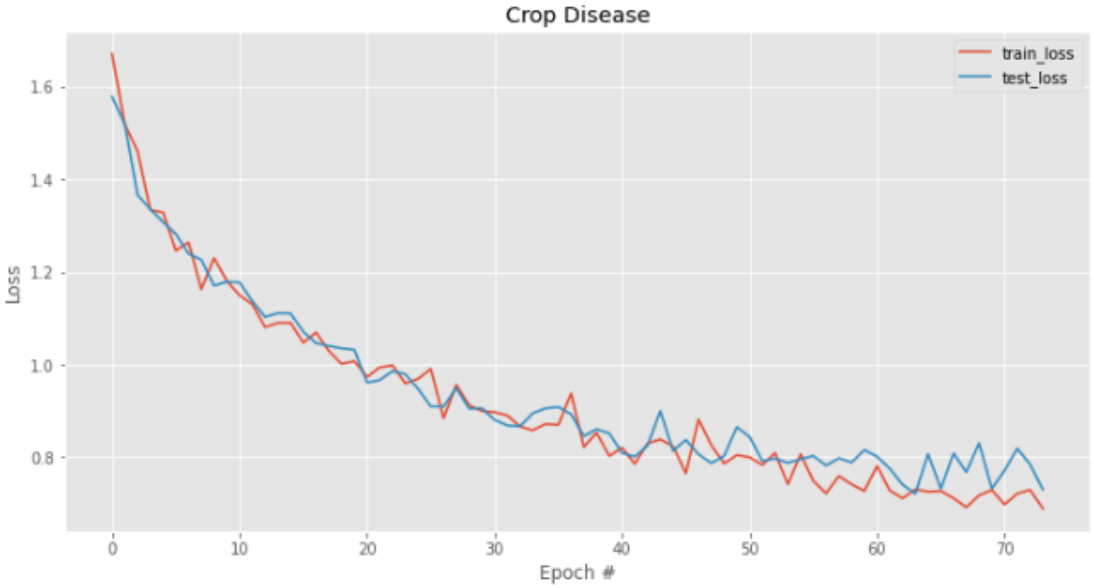
$$Depthwise\ Conv\ (W, y)_{(i,j)} = \sum_{k,l}^{K,L} W_{(k,l)} \ * \ y_{(i+k,j+l)}$$

$$SepConv\left(W_p, W_d, y\right)_{(i,j)} = \ Pointwise\ Conv\ _{(i,j)}\left(W_p,\ Depthwise\ Conv\ _{(i,j)}(W_d, y)\right)$$

Novosibirsk State University
*THE REAL SCIENCE

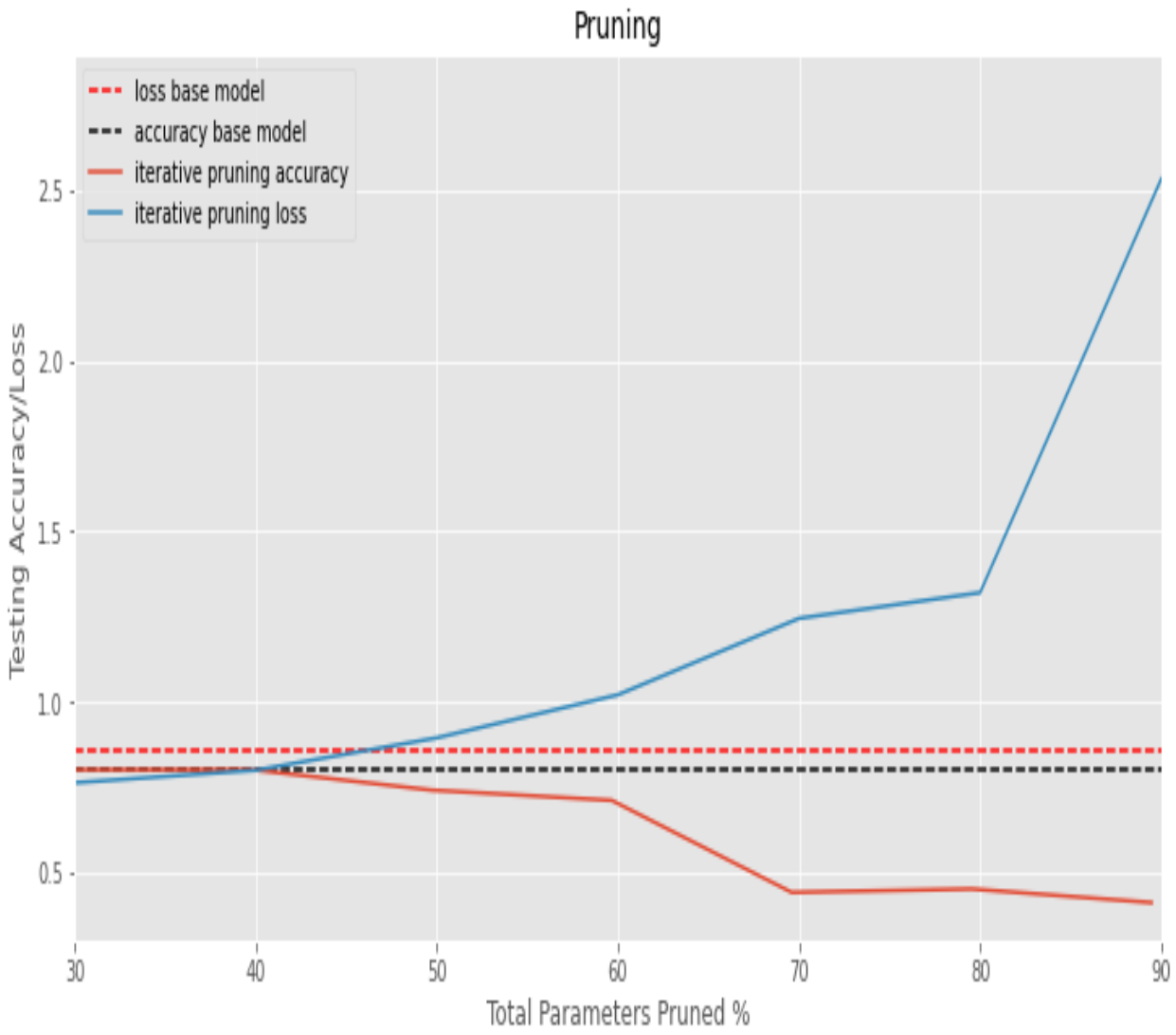
MobileNet





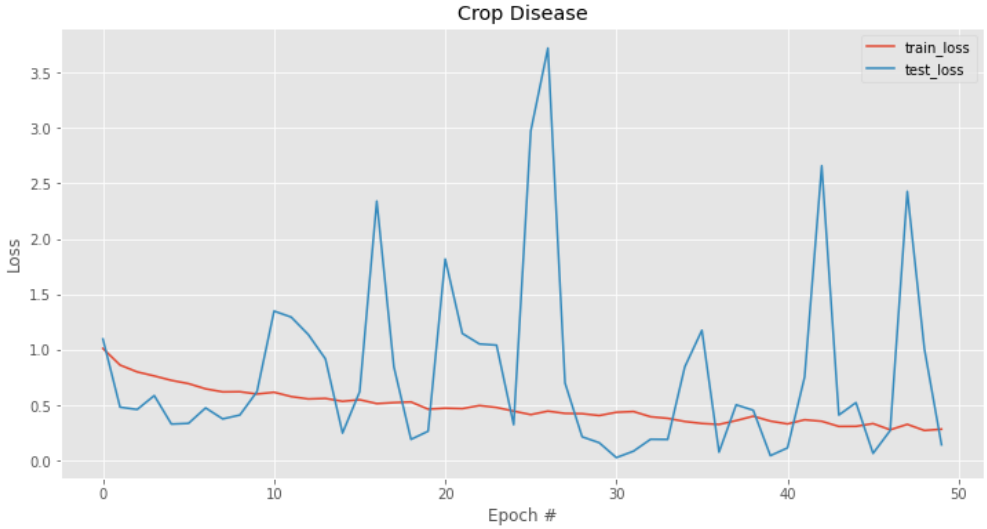
Novosibirsk State University
*THE REAL SCIENCE

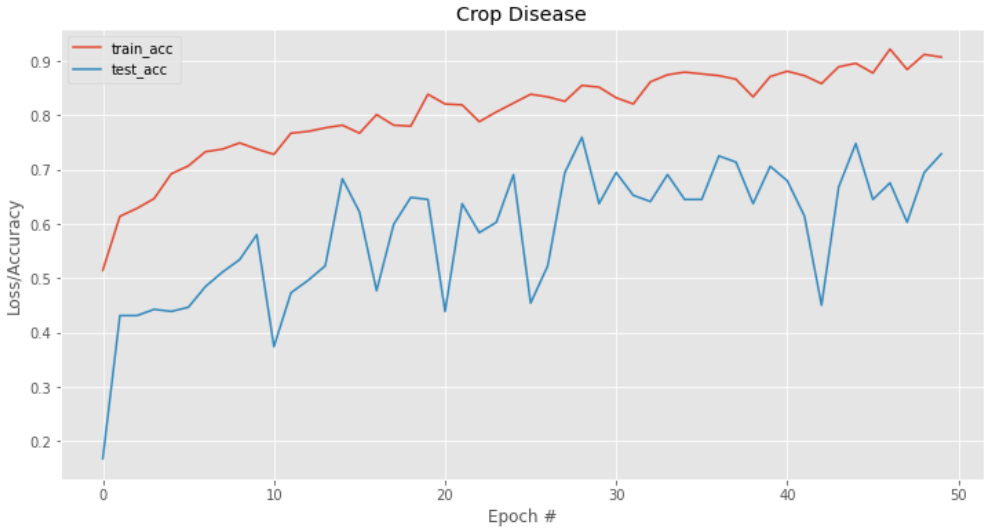
Pruning



Novosibirsk State University
THE REAL SCIENCE

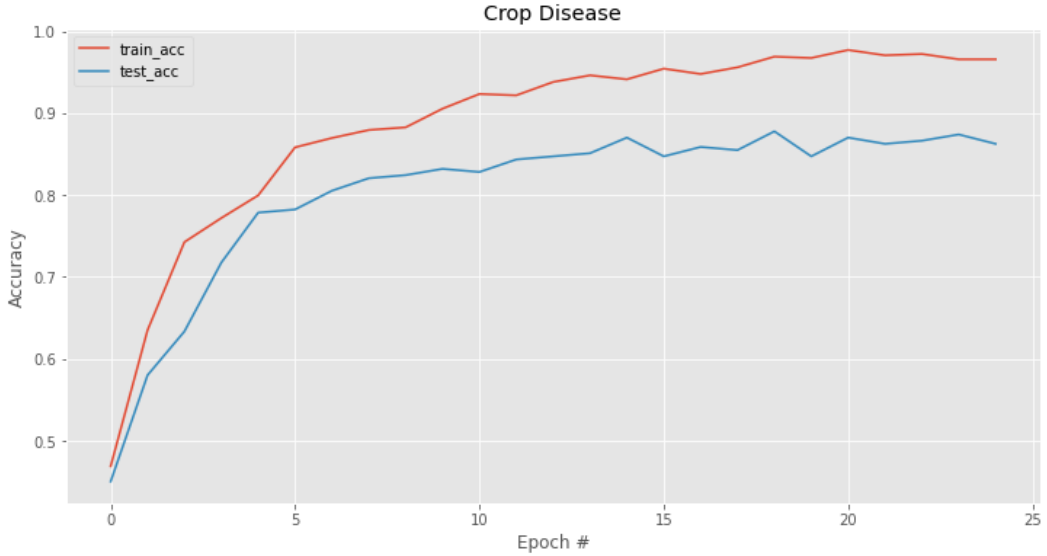
SqueezeNet

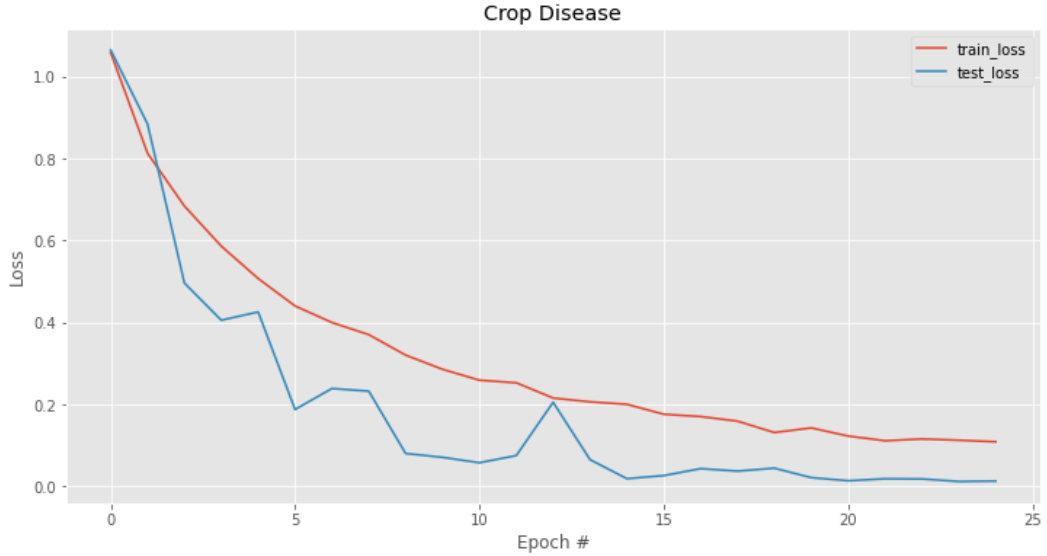




Novosibirsk State University
*THE REAL SCIENCE

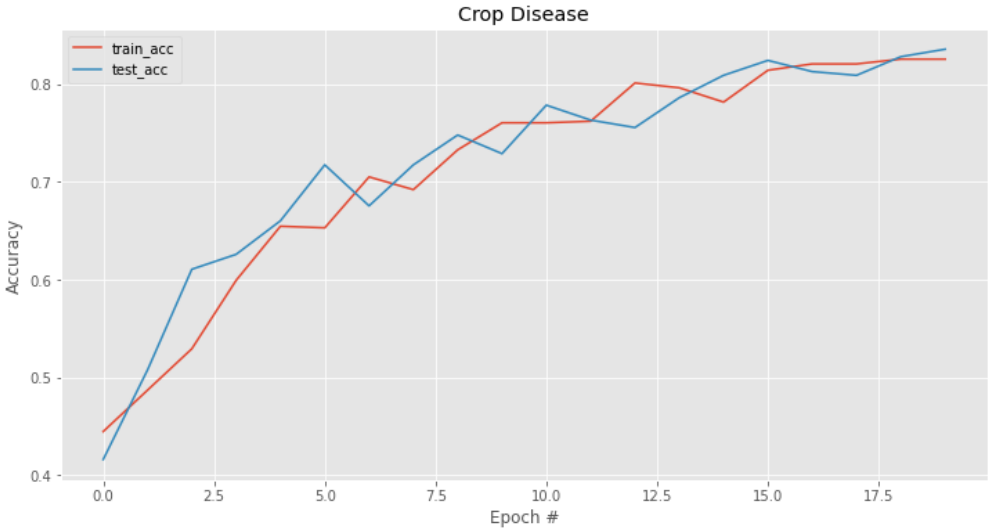
ResNet50

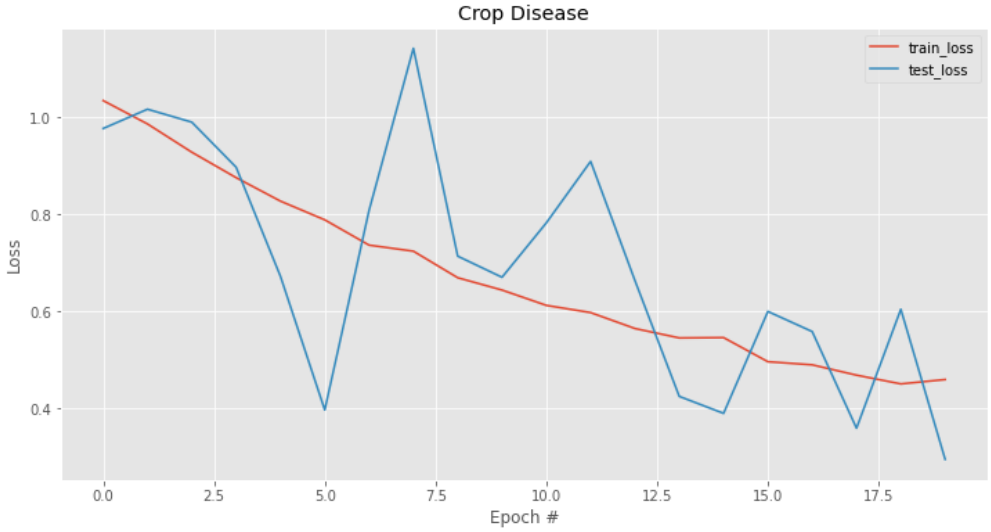




Novosibirsk State University
*THE REAL SCIENCE

EfficientNet





Evaluation

- Size
- Parameters
- Flops
- Accuracy

- Energy consumed (Training and Inference)
  - why at training & inference
- Energy efficiency=Energy/MFlops the lower the better

Novosibirsk State University
THE REAL SCIENCE

AI solution for crop disease using mobilenets

Novosibirsk State University
*THE REAL SCIENCE

AI solution for crop disease using mobilenets

- Continuous capturing images using an automated sliding rail on a marked plant bed

Calculate Plant Health
- Use NDVI formulation to calculate plant health on image
- Create heatmaps of plant health for visualization

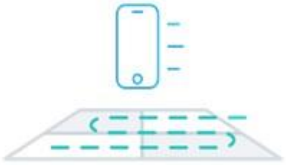
Image Preprocessing
- Merge images (mosaic image)
- Geo-reference image
- Break image for prediction

Prediction
- Use neural networks to predict plant disease

Visualize

Storage

Geo-coordinates



Augmented Reality
- Use GPS coordinates to augment camera image & disease predicted section

Novosibirsk State University
THE REAL SCIENCE

AI solution for crop disease using mobilenets

ICLR Workshop Challenge #1: CGIAR Computer Vision for Crop Disease

zindi.africa/competitions/iclr-workshop-challenge-1-cgiar-computer-vision-for-crop-disease

Identify wheat rust in images from Ethiopia and Tanzania, and win a trip to present your work at ICLR 2020 in Addis Ababa.

29 January–29 March 2020

currently ranked

5 out of **304**