東京工業大学
Tokyo Institute of Technology

MURATA LAB

Complex Network Approaches for Deep CNN Compression.

Murata Group
Hoang Nguyen (M2) - presenter

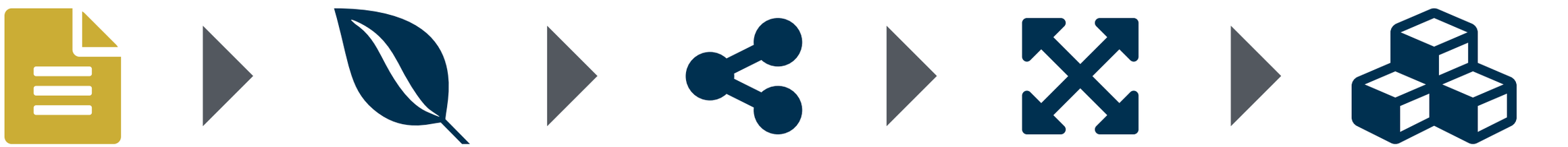
2017/04/04

Compressing Convolutional Neural Network
A Complex Network Science Perspective

Questions:

- What is the current states of CNN compression tech?
- How can network science play a role in CNN compression?

Key points of our presentation:



Literature review | Network pruning | Generative block models | Diversity models | Other approaches

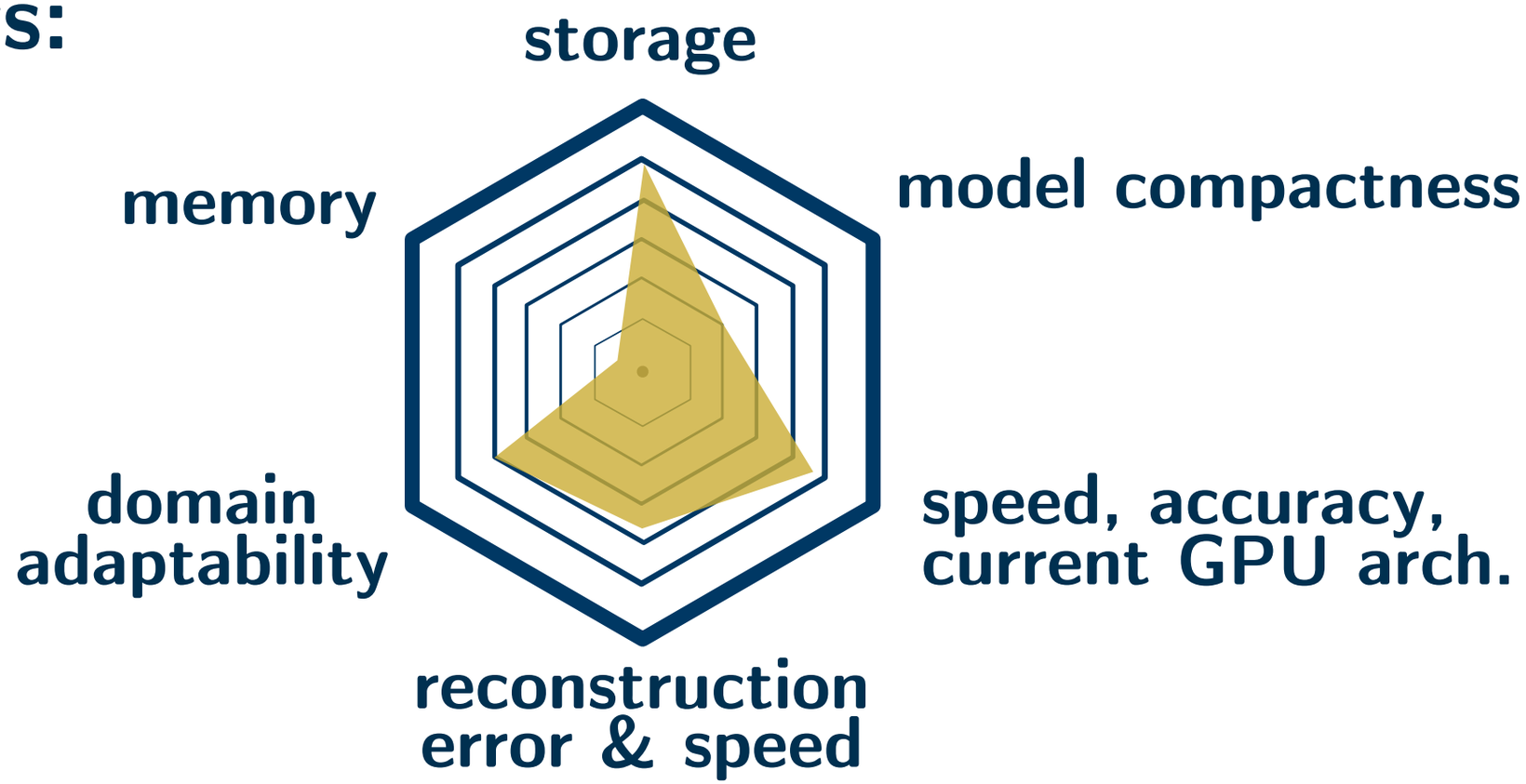
CNN Storage Compression

Problem statement

Challenge:

> Given a fully trained convolutional neural network, reduce the computational resources required for deploying the network while maintaining a minimal impact on the network's performance.

Trade-offs:



CNN Storage Compression

Recent results

Storage compression of AlexNet

Methods	Cm.	Spd.	Pruning	W. Sharing	Coding	Note
(Hinton, 2015)	-	-	Binary	-	-	
(Wang, 2016)	39x	25x	DCT	K-Means	Huffman	
(Wen, 2016)	64x	-	DCT	Hashing	-	FreshNets
(Hans, 2016)	49x	-	Threshold	K-Means	Huffman	
(Ullrich, 2017)	60x	-	Threshold	K-Gaussian	Huffman	
(Hans, 2017)	-	189x	-	-		New arch.

Observation

- Asides from weight sharing techniques, the current pipeline is just an adaptation of JPEG algorithm.

Removing “weak” weights

Thresholds and spectral methods

Assumption:

- Weights with small values are not important.
- Removing small weights does not affect the result.

Threshold-based methods:

- Round weights to zero using static or dynamic thresholds.
- Binarize weights to (-1, 0, 1) (Courbariaux, 2015) (Hubara, 2015)

Spectral methods:

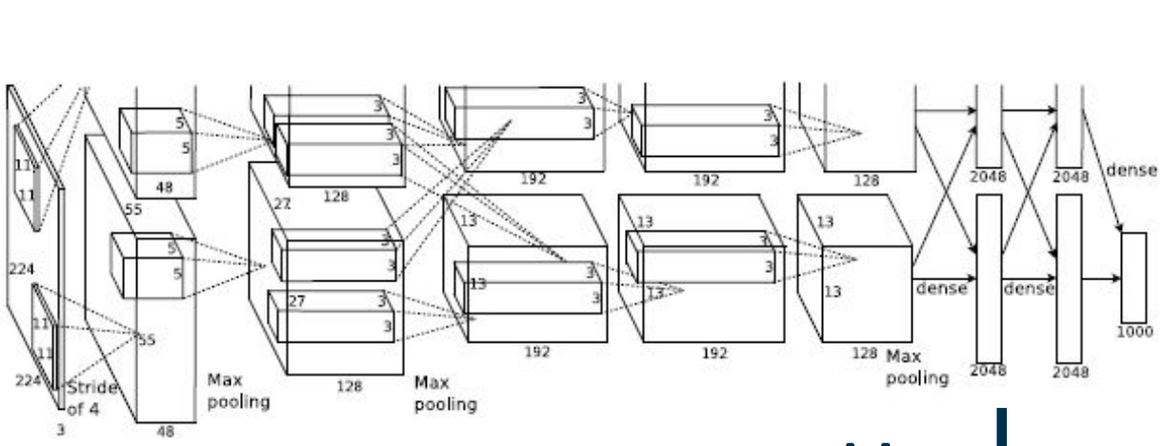
- Treating weight matrices as an image and apply image compression techniques. (Wang, 2016) (Wen, 2016)
- Factorizing weight matrices with network embeddings.

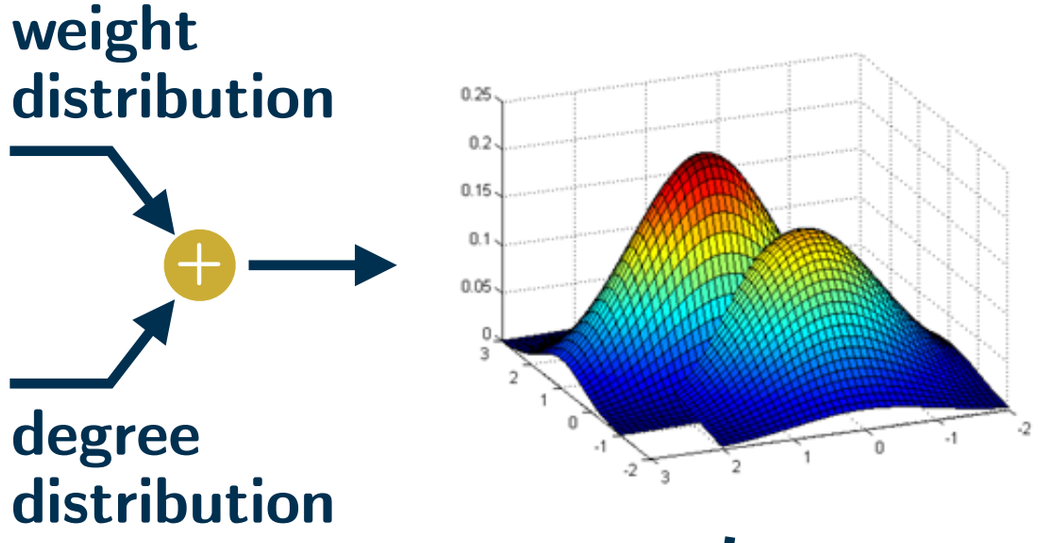
Generate a new network
Weighted Stochastic Block Model

Assumption:

- The community structures decide the functionality of neural networks.

Approach:



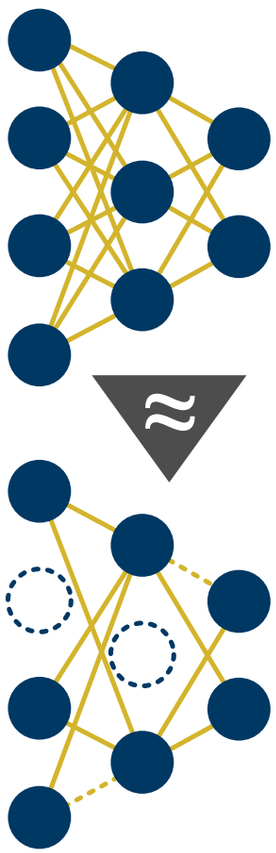


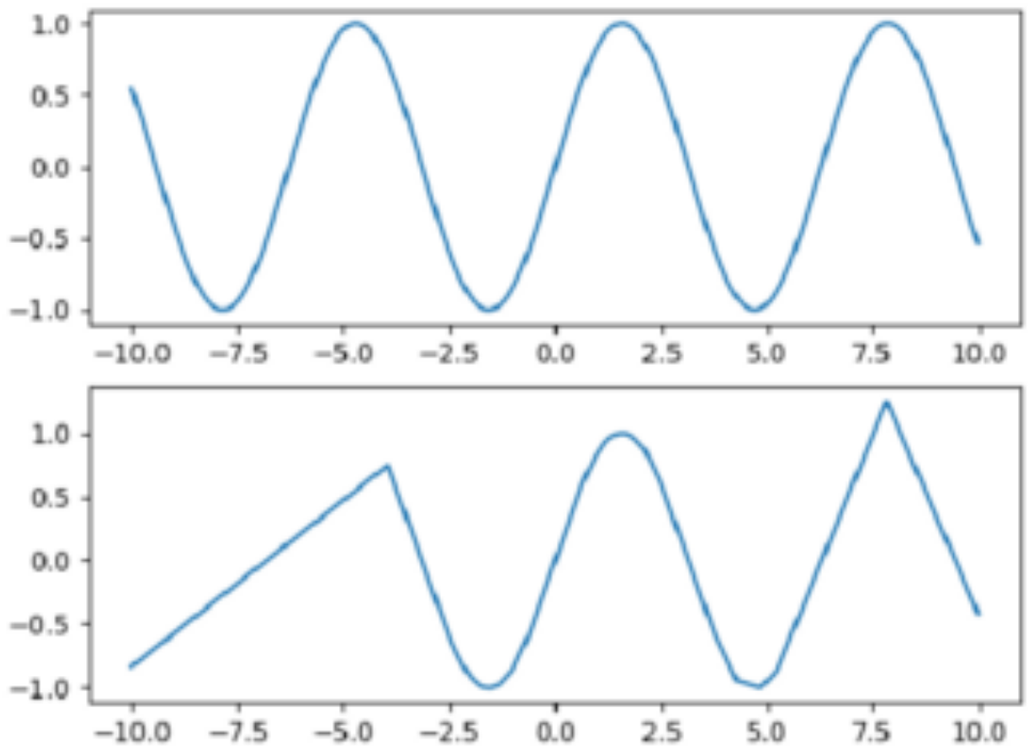
Graph Samping and Summarization

Compressing a graph by minimum coverage set

Assumption:

- The performance measure of a neural network is a submodular function w.r.t the number of neurons.





Other approaches

- Tensor factorization.
- Transfer learning (compression in multiple tasks)
- Most successful compression techniques are “spin-off” of JPEG algorithms. The low-hanging fruits: Lapped transformation (mp4), Wavelet transformation (JPEG2000), and other information processing techniques.
- Running a neural network in its compressed form may require specialized hardware (FPGA, ASIC, etc.)
- Network distillation (dense and convolutional layer)

(Hinton, 2015)

References

(Hinton, 2015) Distilling the knowledge in a neural network

(Wang, 2016) CNNpack: Packing Convolutional Neural Networks in the Frequency Domain.

(Ullrich, 2017) Soft Weight-Sharing for Neural Network Compression

(Wen, 2016) Learning Structured Sparsity in Deep Neural Network

(Han, 2016) Deep Compression: Compressing DNN with Pruning, Trained Quantization and Huffman Coding

(Hans, 2017) EIE: Efficient Interface Engine on Compressed Deep Neural Network

Complete list of literature research & tutorial: https://net-titech.github.io