Neural Network Pruning

PD Dr. Haojin Yang

Multimedia and Machine Learning Group

Hasso Plattner Institute

Design IT.
Create Knowledge.

www.hpi.de

What is Network Pruning?

- Pruning is a method for model compression.
- It is inspired by **Synaptic Pruning** of biological neuron system.
- It also utilizes the fact that large models are often overparameterized.

newborn
50 trillion
synapses

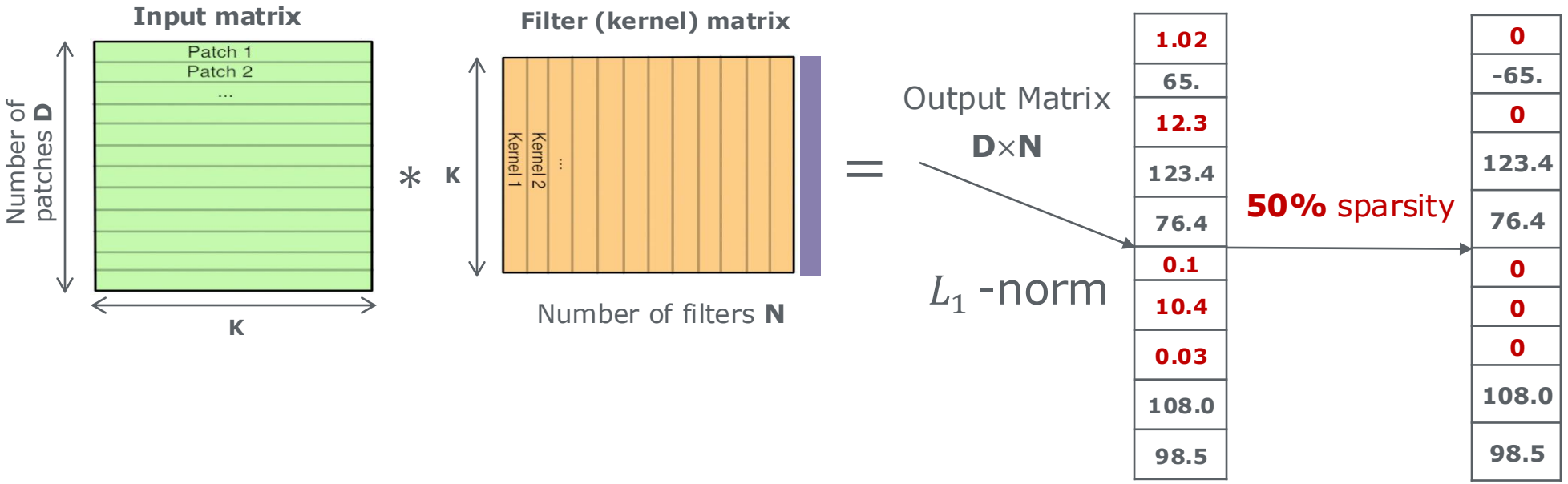
1 year old
1000 trillion
synapses

10 years old
500 trillion
synapses

Model	Model Size	ImageNet Top-1 Accuracy
AlexNet	240MB	57.1%
SqueezeNet	4.8MB (**-98%**)	57.5% (**+0.4%**)

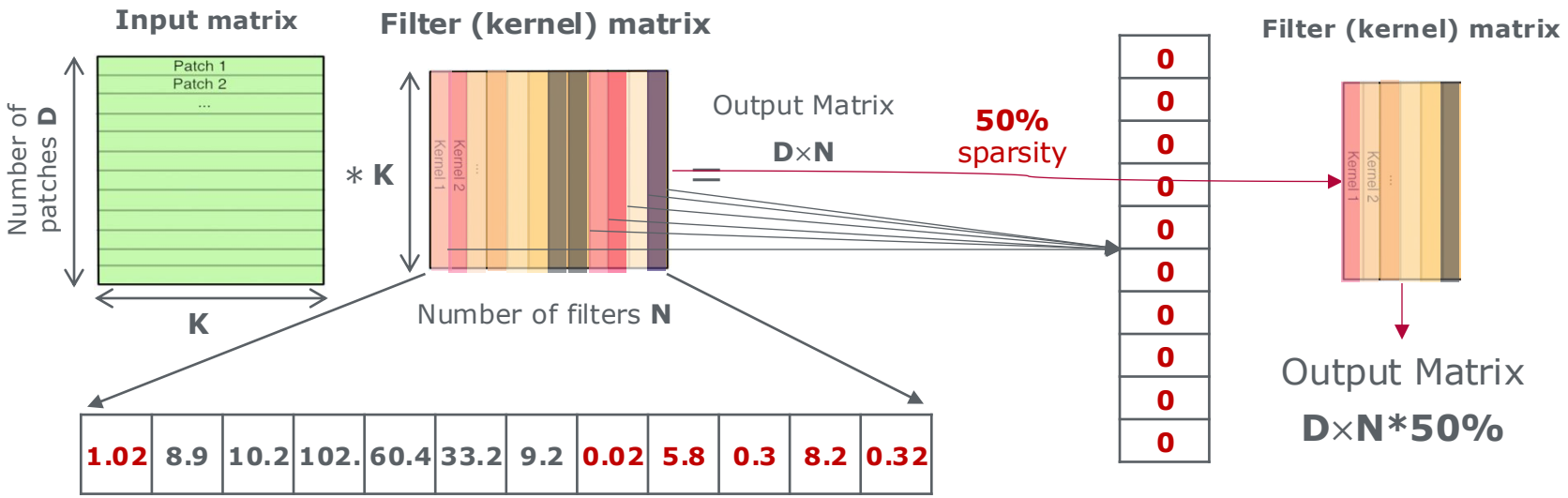
Weight Pruning

- Rank the weights using $L\_(1)$-"norm" $\|x\|\_1$ as the importance score
- Set the x% **weight values** with smaller scores to **0**
- Skipping zeros during inference for speedup (specific implementation required)



Neuron Pruning

- Rank the weight columns using $L\_(2\ )$-"norm" $\|x\|\_2$ as the importance score
- Set the x% **weight columns** with smaller scores to **0**
- Equivalent to delete the corresponding output neurons



Iterative Training

- Weight or neuron pruning will cause performance degradation.
- Iterative pruning can effectively preserve accuracy.
- For each layer do:
  1. Set a relative small pruning step, e.g. 5%
  2. Calculate the importance score of weights or weight columns
  3. Prune the least important 5% items
  4. Fine-tune for recovering the accuracy
  5. If the final pruning rate is achieved, then stop. Otherwise, go to step 1

A case study: YOLOv3 for Hand Detection

- Dataset: VGG Hand, 13050 annotated hand instances, train-img: 4807, test-img: 821

Images from VGG Hand

- Model: YOLOv3 object detection model

Model	Params	Size	FLOPs	Inference	mAP
Original	61.5M	246.4MB	32.8B	15.0ms	0.7692
Pruned	10.9M(**-82%**)	43.6MB(**-82%**)	9.6B(**-71%**)	7.7ms(**-49%**)	0.7722(**+0.003**)
Finetune	10.9M(**-82%**)	43.6MB(**-82%**)	9.6B(**-71%**)	7.7ms(**-49%**)	0.775(**+0.006**)

https://github.com/Lam1360/YOLOv3-model-pruning

Summary

- Pruning is a method for model compression.
- Basic pruning methods
  - Individual weight pruning
  - Neuron (channel) pruning
- A case study on YOLOv3 hand detection
- Disadvantages: Iterative pruning and fine-tuning can be very time consuming.