UCDAVIS
COLLEGE OF ENGINEERING

ELECTRICAL &
COMPUTER ENGINEERING

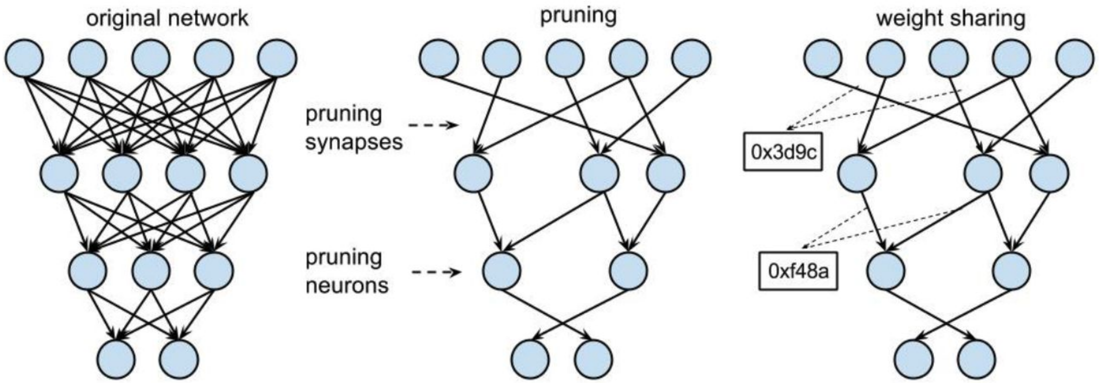
AI on Edge:
Neural Network Pruning

Yui Ishihara

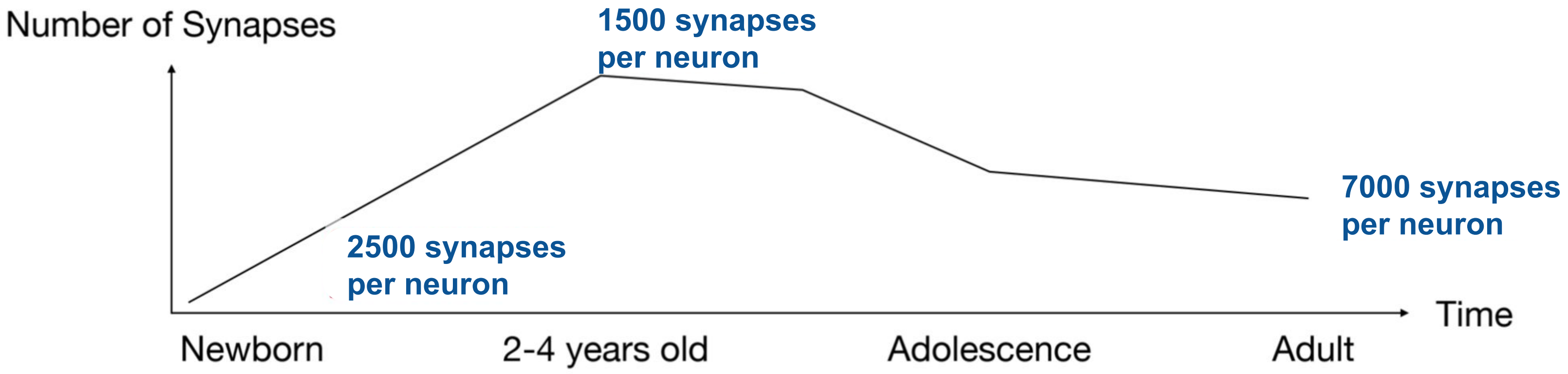
*Acknowledgement: material and slides are adapted from Song Han, Kai Li, Xiaxiao Li, and Bei Yu

DNN Model Compression

- Various Methods for DNN Compression:
  - **Pruning**
  - Quantization
  - Weight Sharing
  - Matrix Factorization
  - Huffman Encoding
  - Low Precision Inference
  - Binarization
  - … and more!
- Often a combination



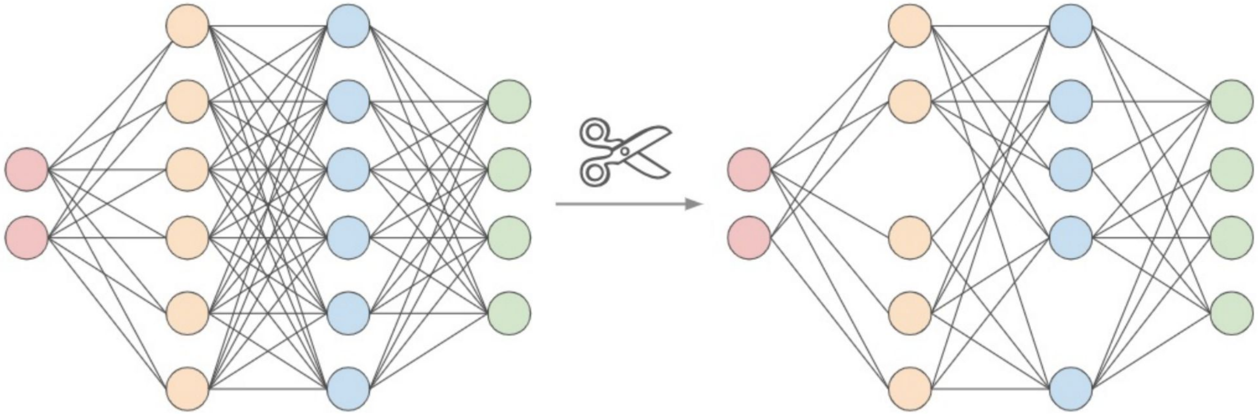
Pruning: Motivation



- Trillion synapses generated in human brain (first few months of birth)
- Pruning removes redundant connections in brain
  - 1 year old – peak 1000 Trillion
  - Pruning begins to occur!
  - 10 years old – about 500 Trillion

What is Pruning?

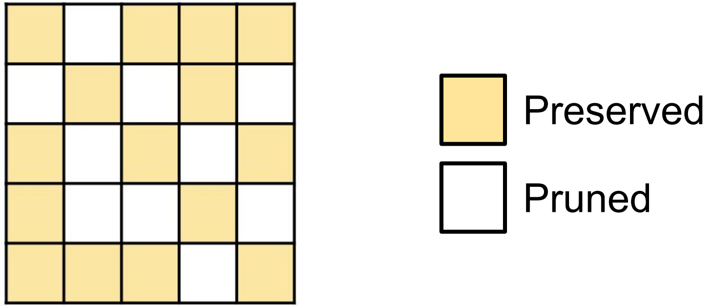
- Pruning simplifies a model by reducing size, removing less critical weights, neurons, or even entire channels, while trying to maintain accuracy
- Goal is efficiency: creating smaller, faster models with lower memory and computation requirements, ideal for deployment in restricted environments (think: EdgeAI!)



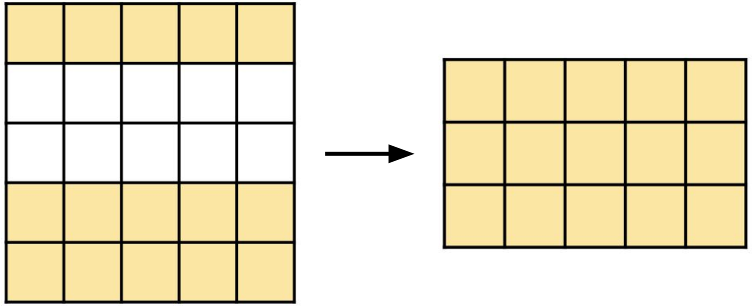
Menghani, Gaurav. "Efficient deep learning: A survey on making deep learning models smaller, faster, and better."

What to Prune: Structured vs. Unstructured

- **Unstructured pruning:** find and remove the less salient connections in the model wherever they are. (Does not consider any relationship between the pruned weights)

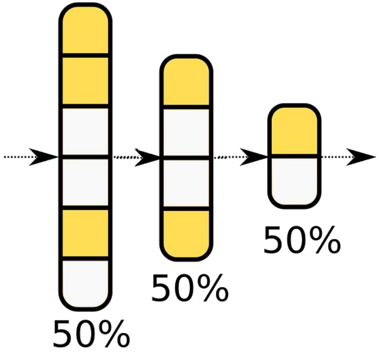


- **Structured pruning:** the selected removal of larger part of the network (e.g. layer)

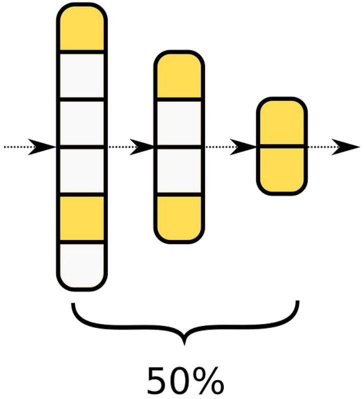


What to Prune: Local vs. Global

- **Local pruning**: consists of removing a fixed percentage of weights from each layer by comparing weights within the layer

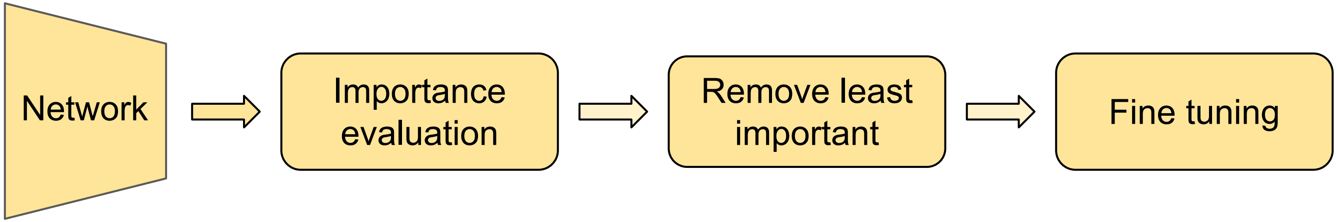


- **Global pruning:** pools all parameters together across layers and selects a global sparsity of them to prune

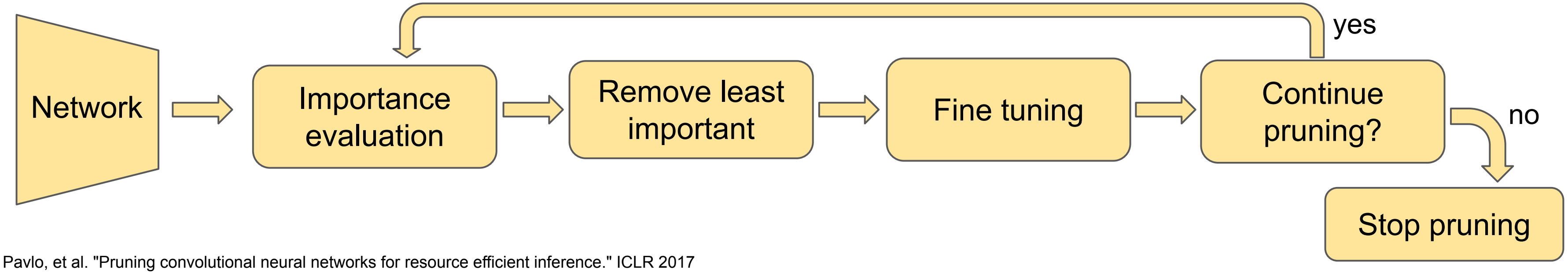


Pruning Methods: Iterative vs. One-Shot

One-Shot: network connections pruned only once



Iterative: network connections pruned partially through multiple iterations



Molchanov, Pavlo, et al. "Pruning convolutional neural networks for resource efficient inference." ICLR 2017

Pruning Criteria

Many heuristics and methods to choose weights/neurons to prune:

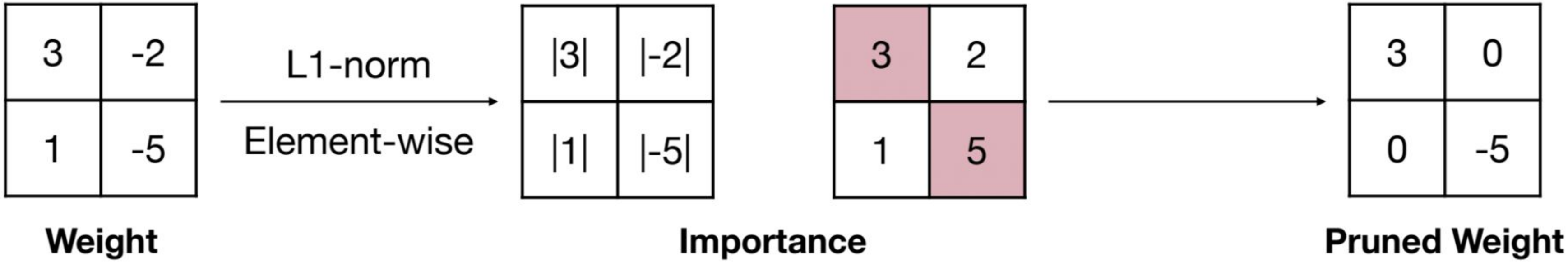
- **Magnitude-based**
- Gradient-based
- Learned
  - E.g. learn pruning masks
- Information-based
  - E.g. Higher-order curvature

Magnitude Based Pruning

- Magnitude-based pruning considers weights with **larger absolute values** as more important than other weights.
- For element-wise pruning:

$$Importance = | W |$$

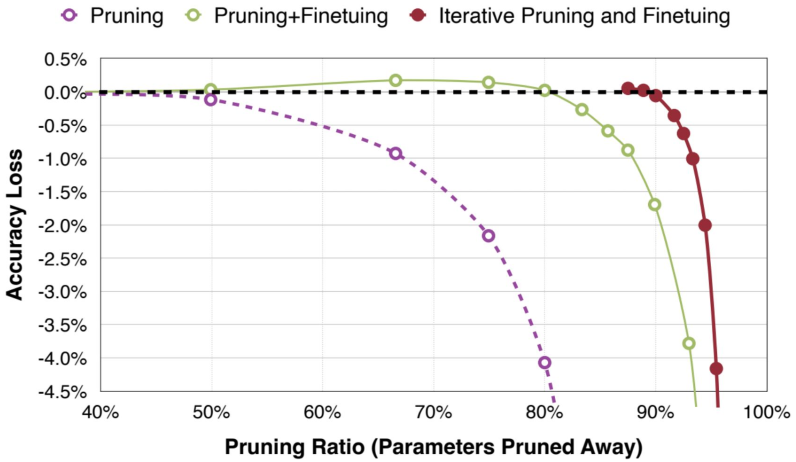
- **Example**



$$\begin{bmatrix} 3 & -2 \\ 1 & -5 \end{bmatrix} \xrightarrow[\text{Element-wise}]{\text{L1-norm}} \begin{bmatrix} |3| & |-2| \\ |1| & |-5| \end{bmatrix} \quad \begin{bmatrix} 3 & 2 \\ 1 & 5 \end{bmatrix} \longrightarrow \begin{bmatrix} 3 & 0 \\ 0 & -5 \end{bmatrix}$$

Fine-Tuning Pruned Networks

- **Accuracy Impact After Pruning**
  - Model accuracy may decrease, especially at higher pruning ratios.
- **Fine-Tuning Benefits**
  - Fine-tuning pruned networks can recover accuracy and enable higher pruning ratios.



Han, Song, et al. "Learning both weights and connections for efficient neural network." NIPS. 2015.