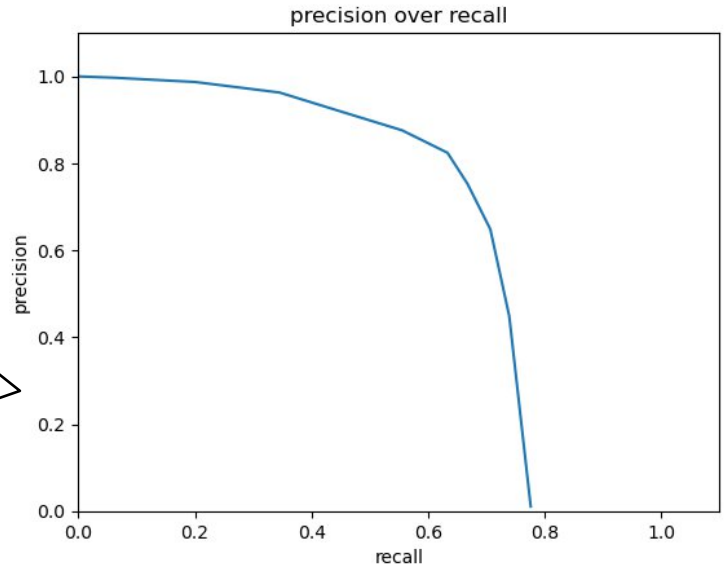
Embedded Machine Learning Lab Challenge

By Valentin & Max

Magnitude Pruning

pretrained model
- adapted for person class
- fine-tuned for 20 epochs

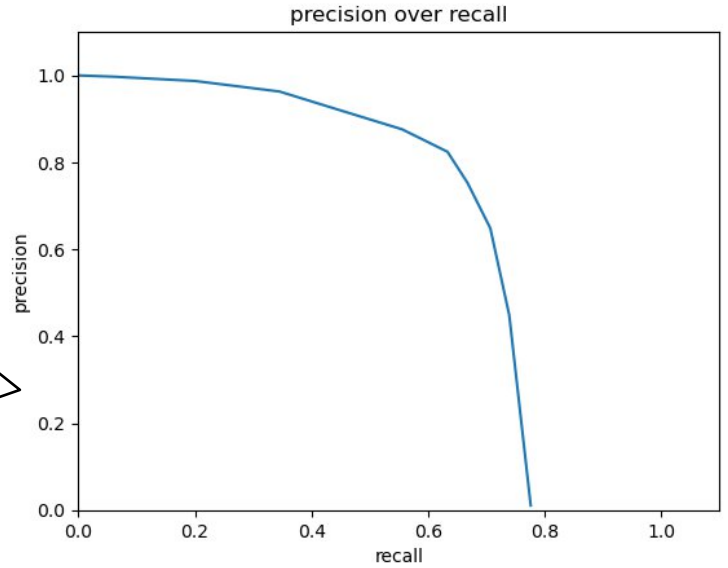
-> 0.65 AP



Magnitude Pruning

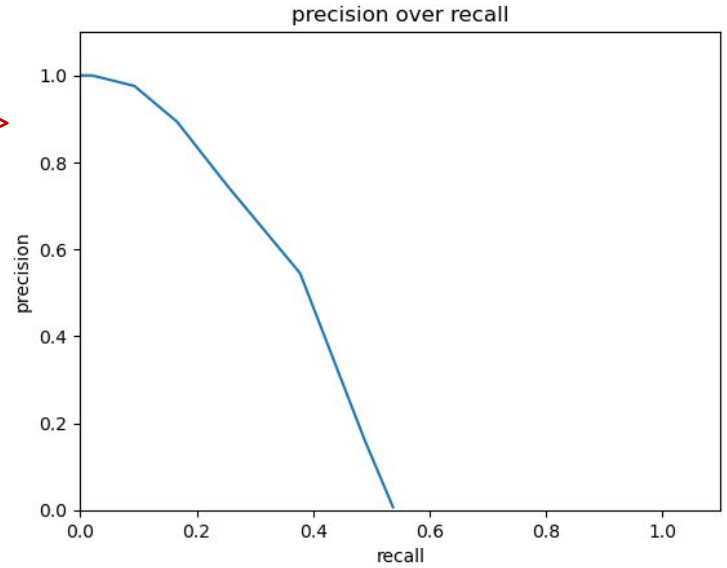
pretrained model
- adapted for person class
- fine-tuned for 20 epochs

-> 0.65 AP



pruning ~40% of params

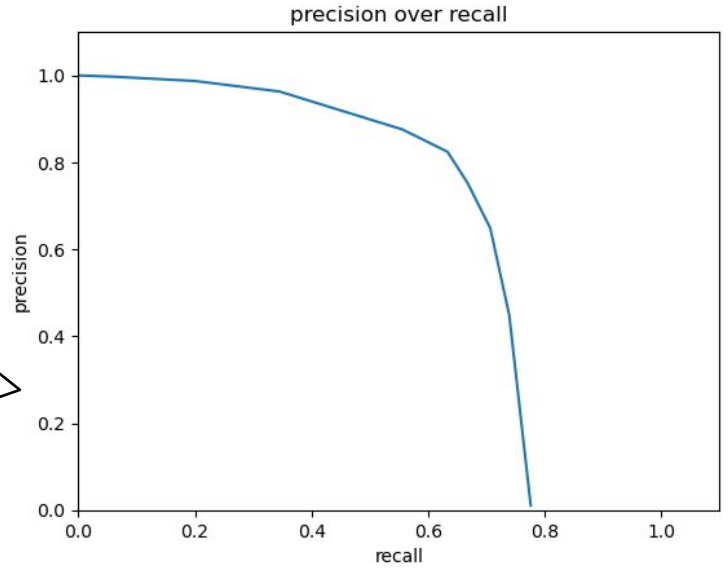
after magnitude pruning
-> 0.30 AP



Magnitude Pruning

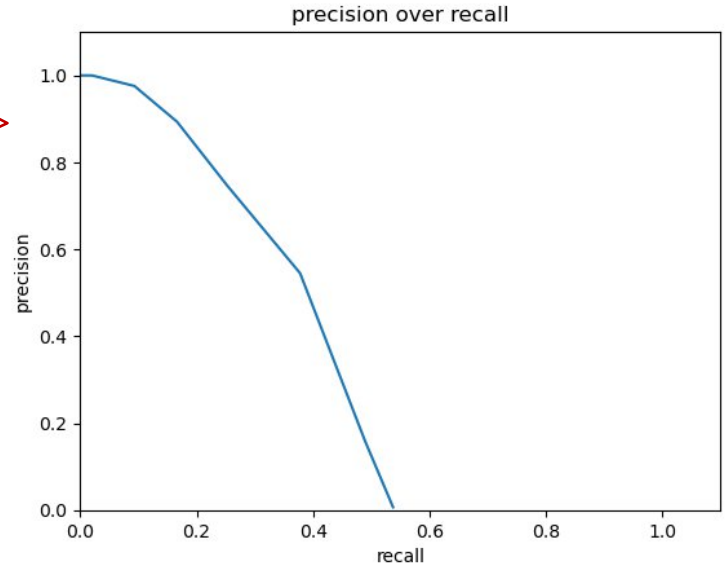
pretrained model
- adapted for person class
- fine-tuned for 20 epochs

-> 0.65 AP

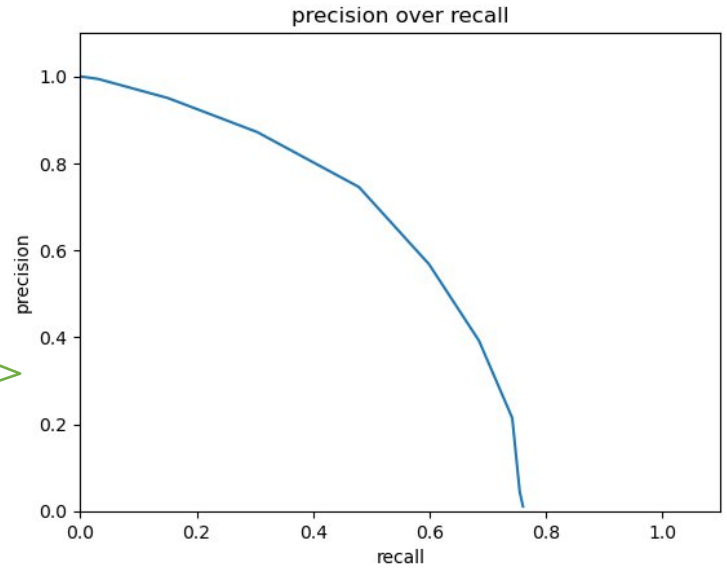


pruning ~40% of params

after magnitude pruning
-> 0.30 AP



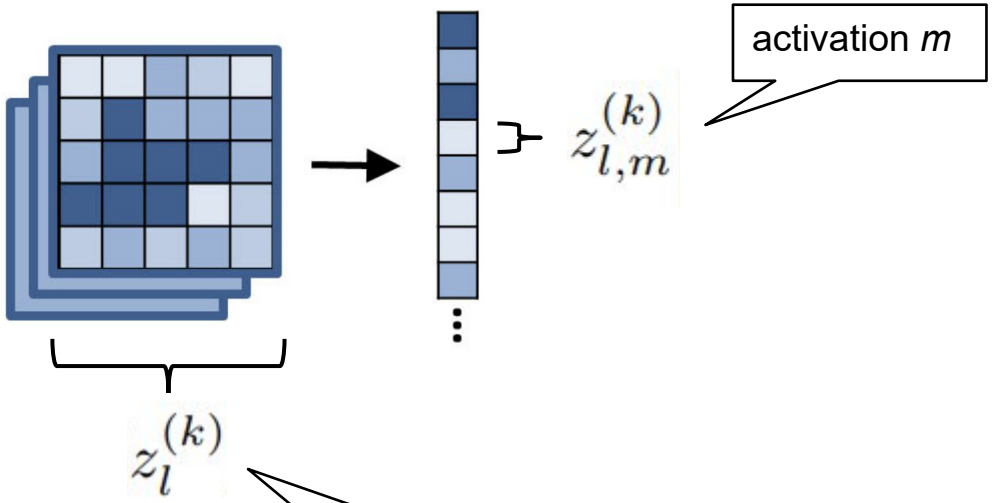
after Taylor pruning
-> 0.51 AP



Taylor Pruning [1]

Channel pruning based on importance

- Defined by approximate change in loss caused by removing channel
- Activations & gradients gathered during forward passes



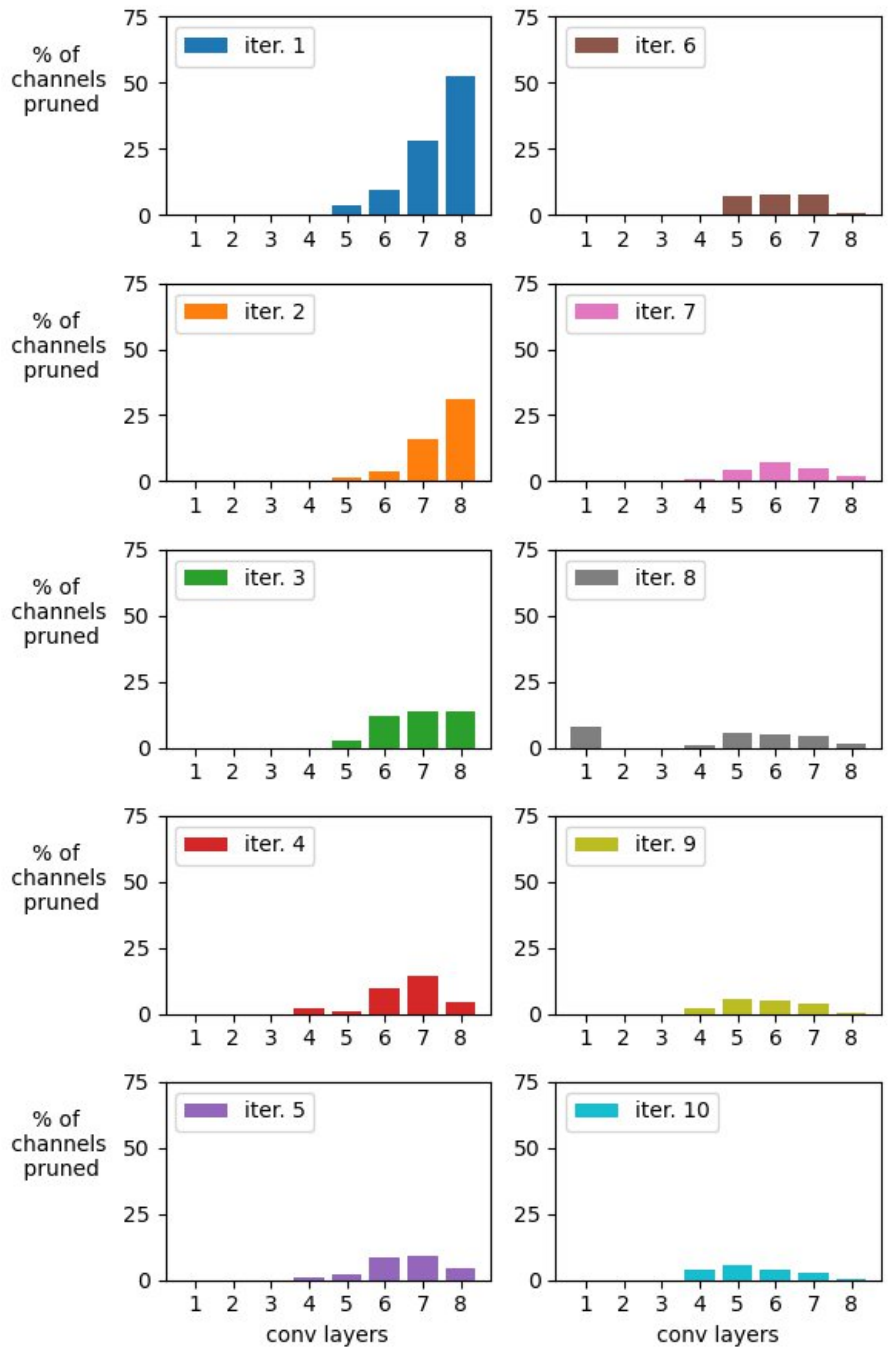
$$\Theta_{TE}(z_l^{(k)}) = \left| \frac{1}{M} \sum_m \frac{\delta C}{\delta z_{l,m}^{(k)}} z_{l,m}^{(k)} \right| \quad [1]$$

importance of channel k of layer l

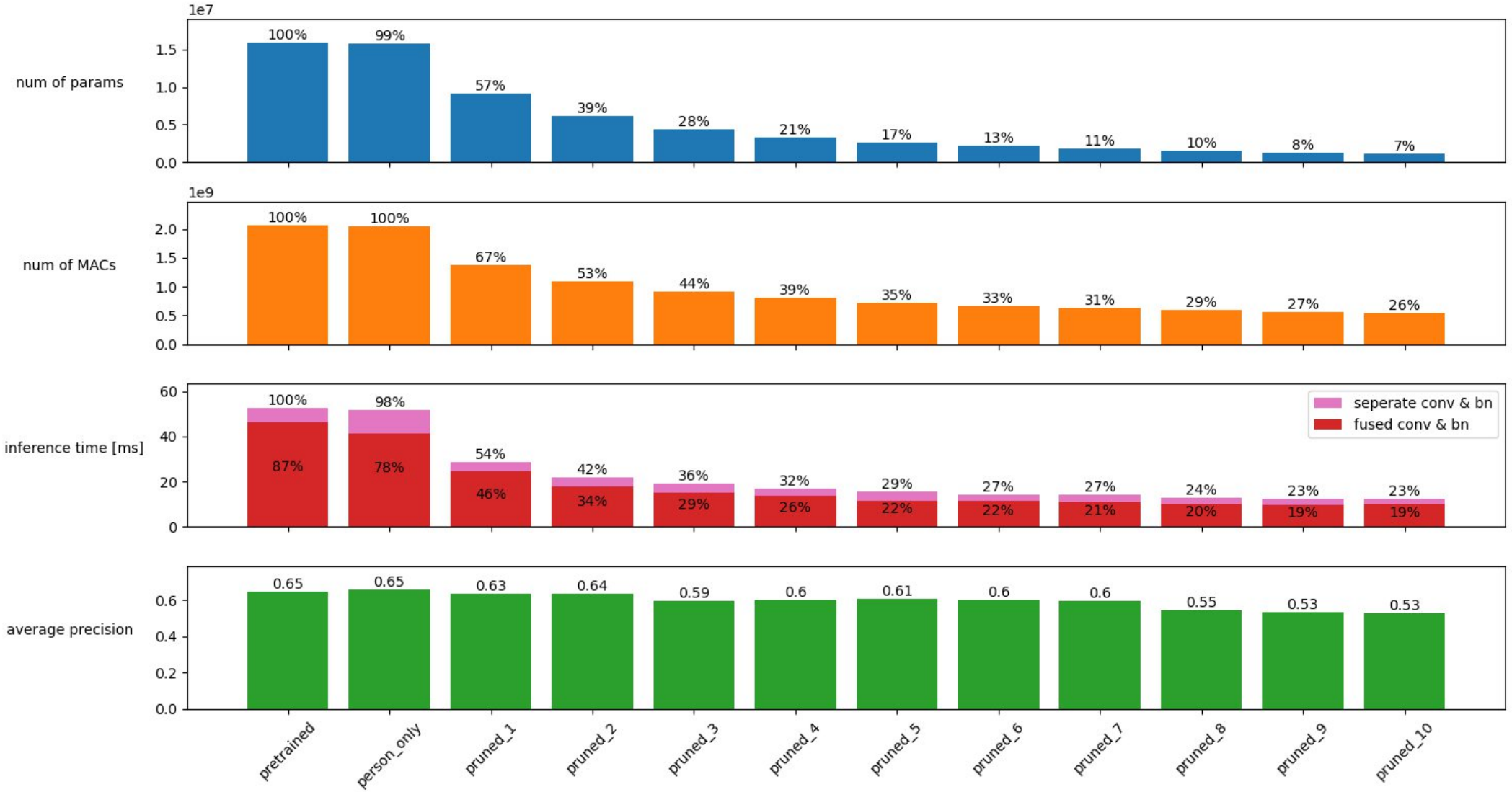
Iterative Pruning

For 10 iterations:

1. Taylor prune *k* least important channels
   - *k* proportional to num of params
2. Fine-tune for 10 epochs to regain performance
   - Very low learning rate
3. Evaluate AP
   - Compare APs over iterations



Pruning Statistics



References

[1] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, & Jan Kautz. (2017). Pruning Convolutional Neural Networks for Resource Efficient Inference.