Network Compression1

2021年4月2日 16:39

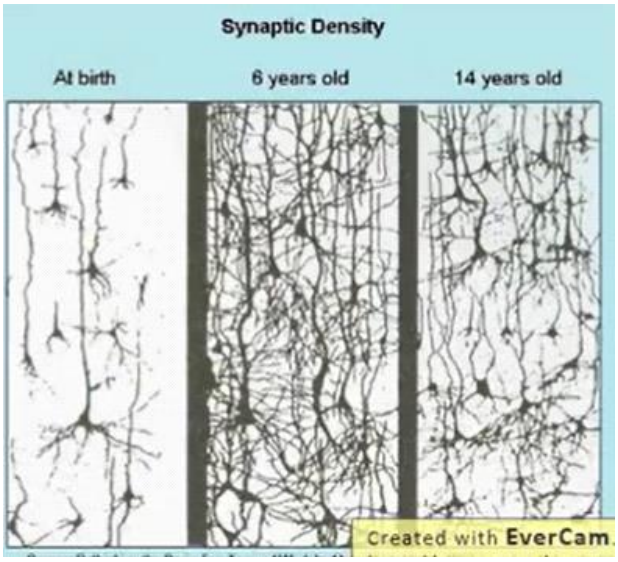
Limited memory space,
limited computing power,

Compression: 压缩

- Network Pruning
- Knowledge Distillation
- Parameter Quantization
- Architecture Design
- Dynamic Computation

- Network Pruning
  - Networks are typically over-parameterized (there is significant redundant weights or neurons)
  - Prune them!

Pruning 剪切



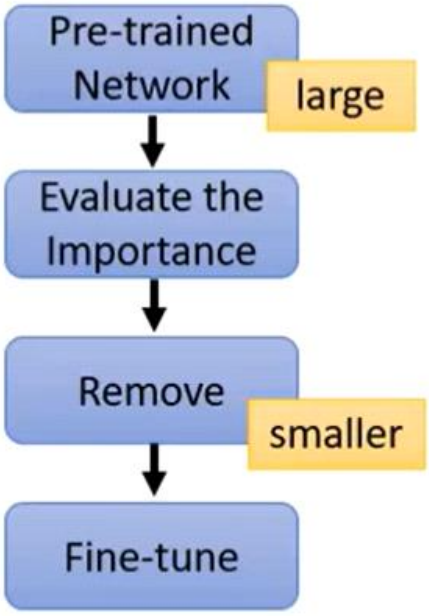
- Importance of a weight:

  L1, L2 ……

- Importance of a neuron:

  the number of times it wasn't zero on a given data set ……

Weight 和 Neuron是否重要



- Importance of a weight:

  L1, L2 ……

- Importance of a neuron:

  the number of times it wasn't zero on a given data set ……

- After pruning, the accuracy will drop (hopefully not too much)
- Fine-tuning on training data for recover
- Don't prune too much at once, or the network won't recover.

移除一些东西,

Remove 一点 recover

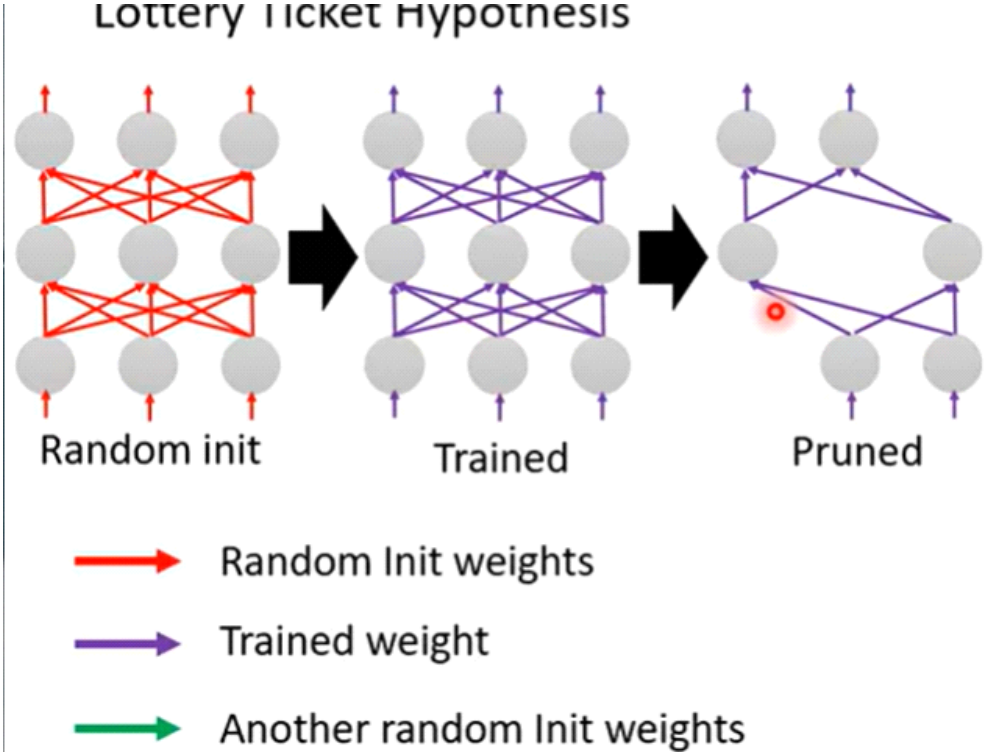
再recover回来

For I in range(n):
Remove a little and then compare the importance again and fine- tune it.

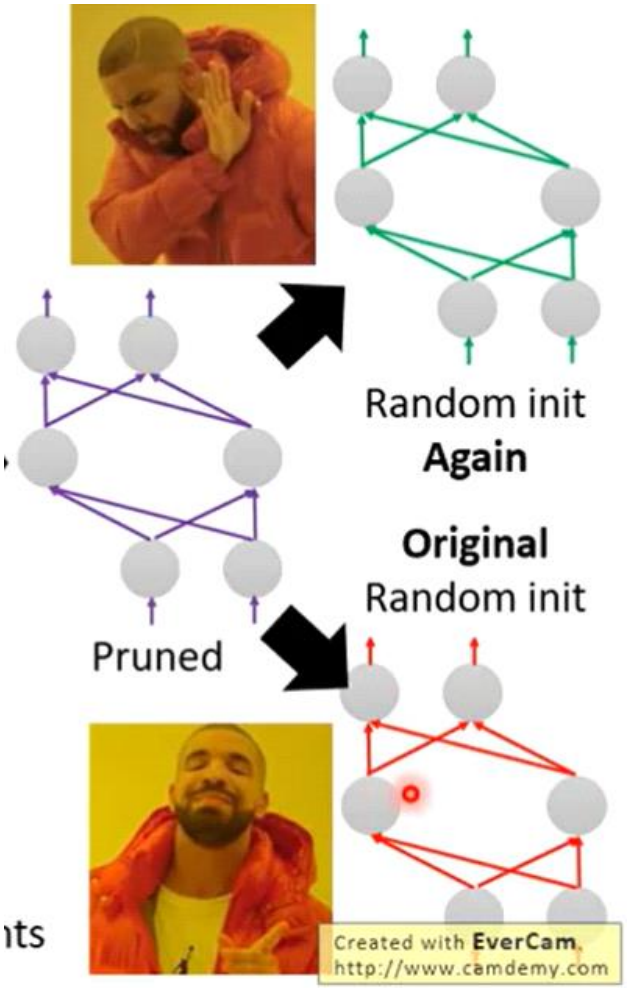
Why Pruning?

- How about simply train a smaller network?
- It is widely known that smaller network is more difficult to learn successfully.
  - Larger network is easier to optimize? https://www.youtube.com/watch?v=_VuWvQUMQVk

Small network is hard to train.
If network is big enough you can get the global min by gradient descent.

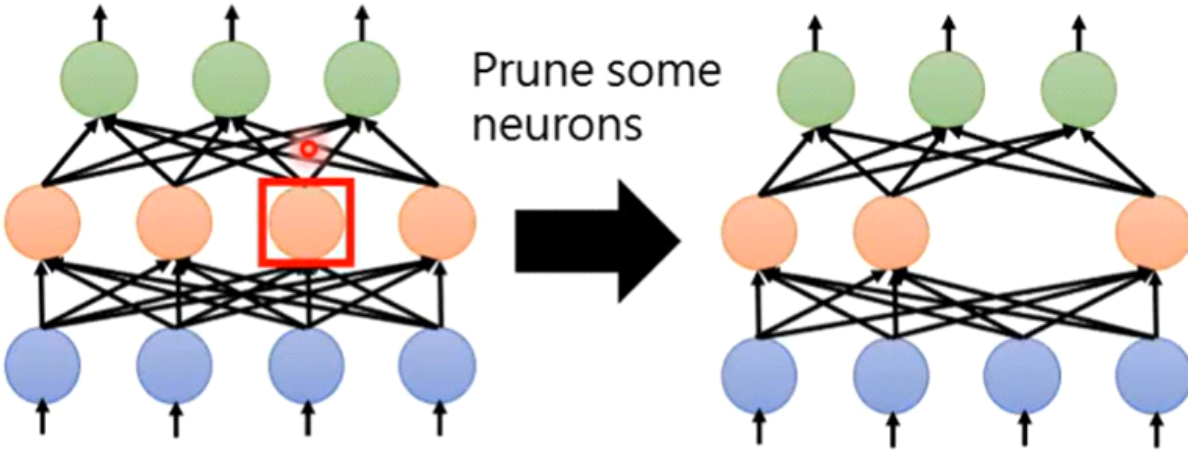


If not random : copy the weight to the small one.
It can be better

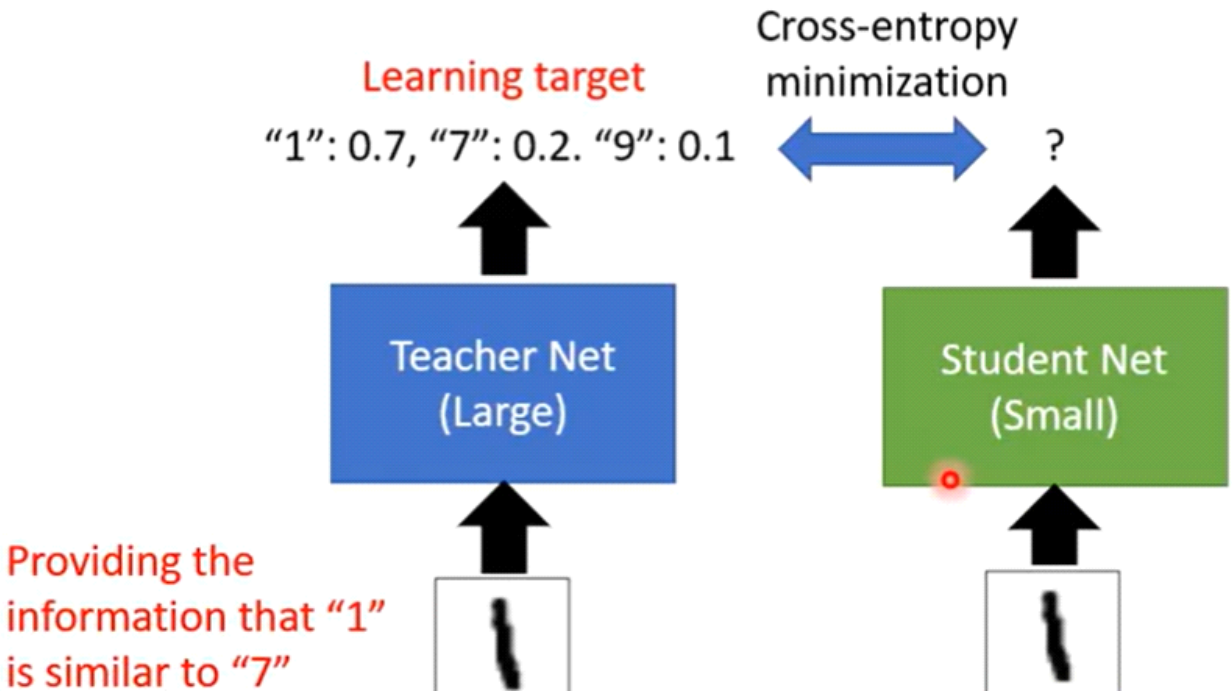


Prun weight :GPU can't speed up.
Prun neuron is a better way.

- Neuron pruning



Knowledge Distillation



Teacher provides more information than target.

- Temperature

$$y_i = \frac{exp(x_i)}{\sum_j exp(x_j)} \quad \Rightarrow \quad y_i = \frac{exp(x_i/T)}{\sum_j exp(x_j/T)}$$

Temperature 有什么用呢?

把不同 label的 拉近一点

$x_1 = 100$	$y_1 = 1$	$x_1/T = 1$	$y_1 = 0.56$
$x_2 = 10$	$y_2 \approx 0$	$x_2/T = 0.1$	$y_2 = 0.23$
$x_3 = 1$	$y_3 \approx 0$	$x_3/T = 0.01$	$y_3 = 0.21$