Deep Learning LightWeight

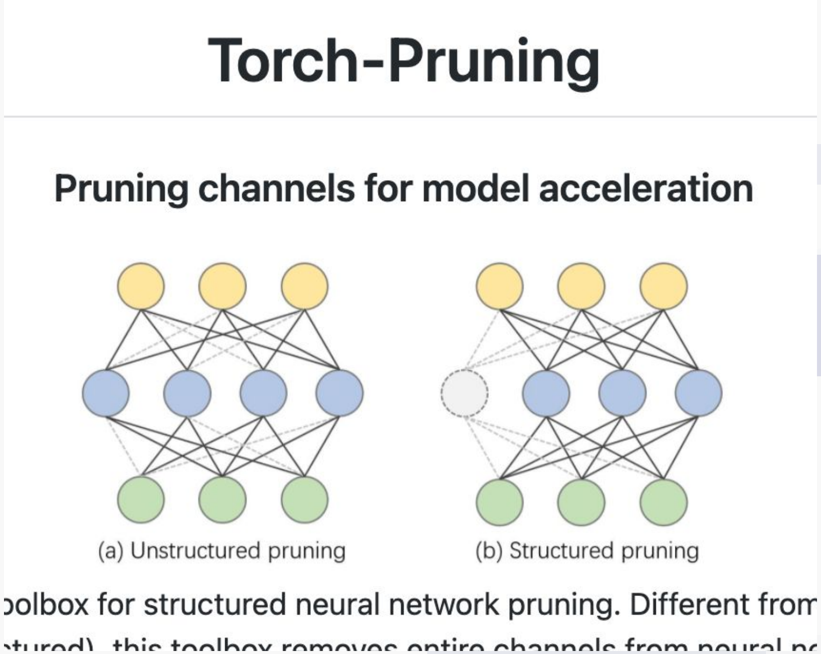
Team 11 Member :

안도성
고영범
김영철
김정우

Our Goals

- **Simple and useful!**
- Reduce the number of parameters by at least half
- Learn the basic principles of pruning
  - □ Which model should we choose? **VGG11**!
  - □ Which API should we get help from?? **torch_pruning!**

Selected Model and Package



Torch-Pruning

Pruning channels for model acceleration

(a) Unstructured pruning

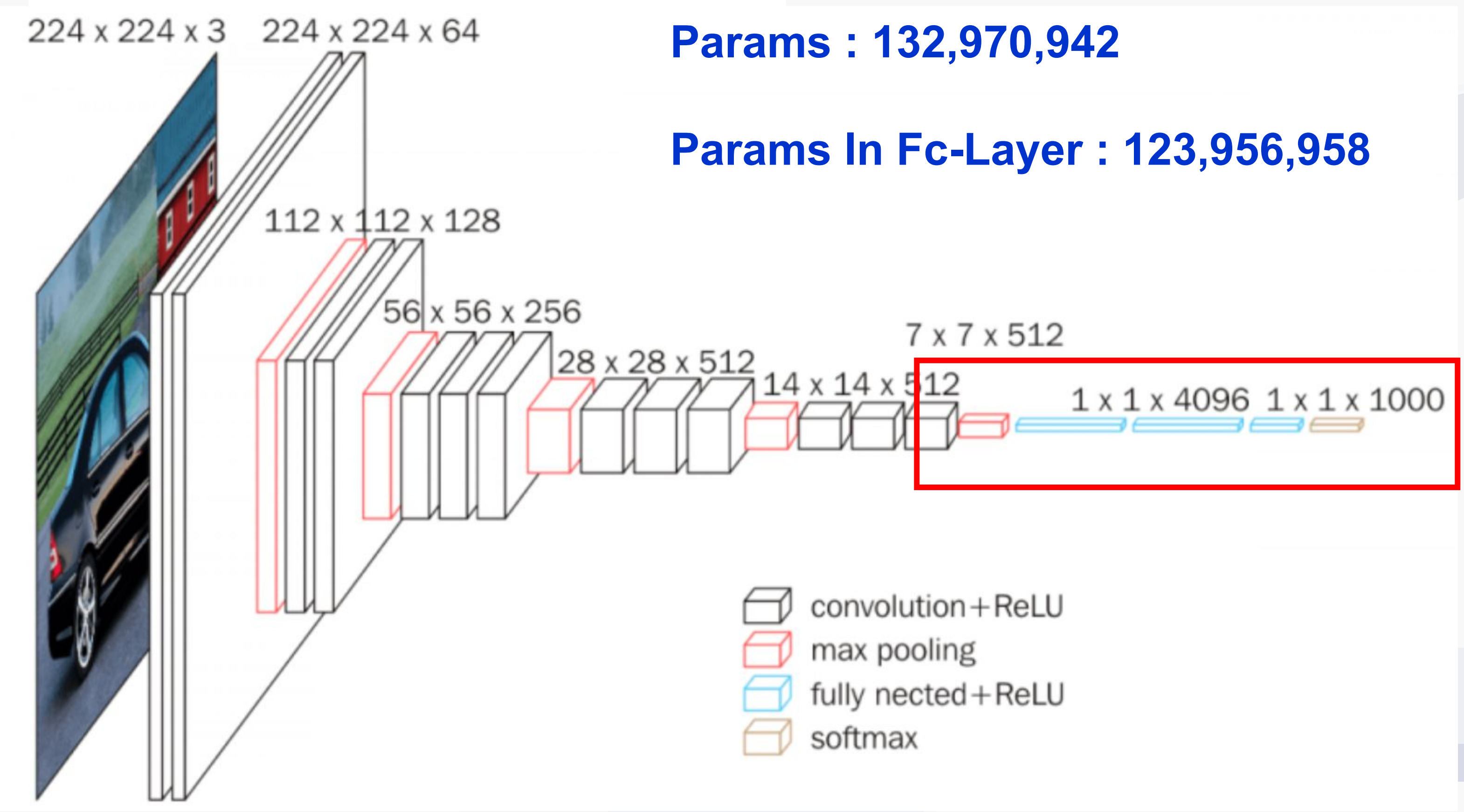
(b) Structured pruning

olbox for structured neural network pruning. Different from

tured), this toolbox removes entire channels from neural ne

https://github.com/VainF/Torch-Pruning

Selected Model and Package



Params : 132,970,942

Params In Fc-Layer : 123,956,958

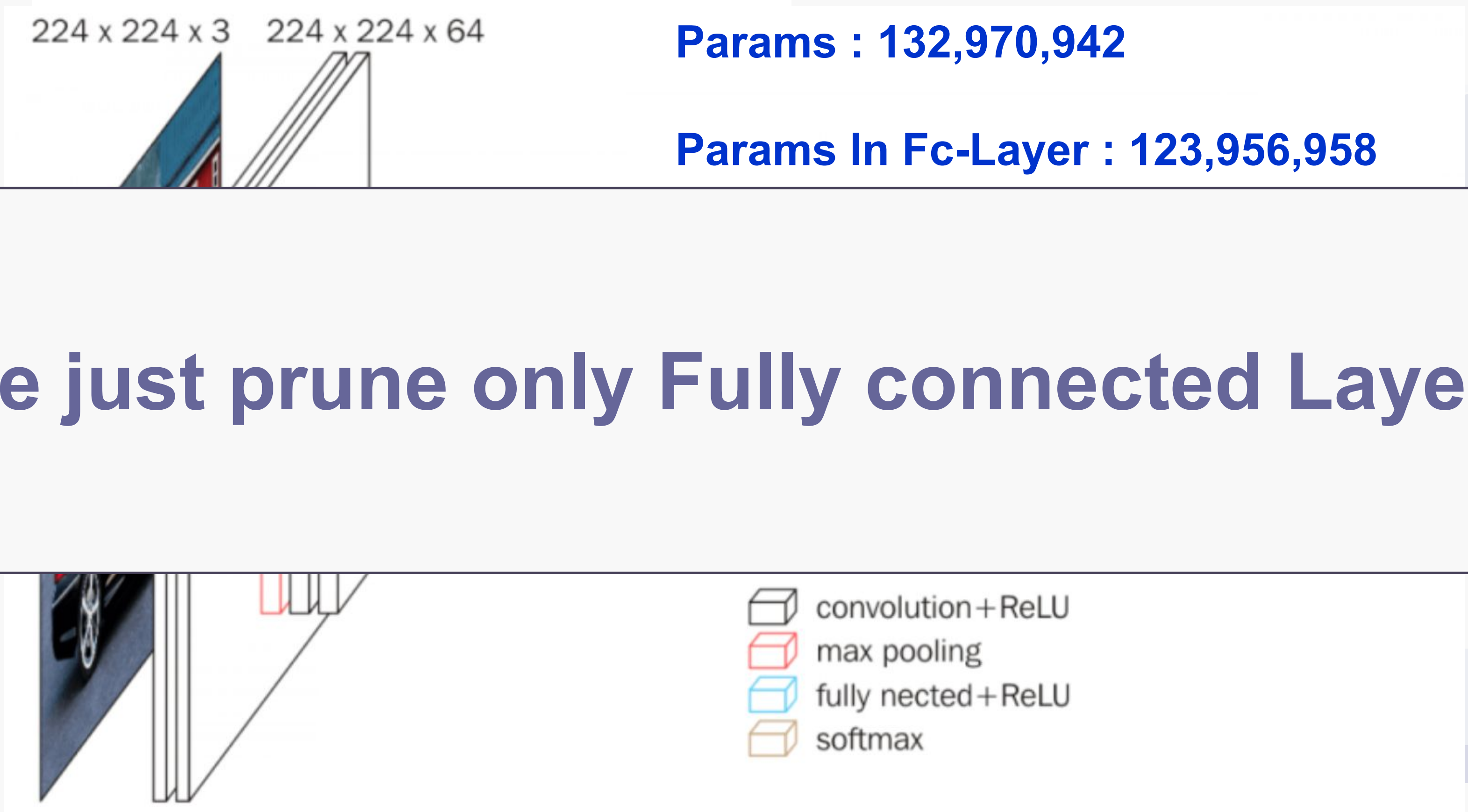
The VGG neural network model architecture

Selected Model and Package

Params : 132,970,942

Params In Fc-Layer : 123,956,958

We just prune only Fully connected Layer!



The VGG neural network model architecture

Our Goals

- ~~Simple and useful!~~
- ~~Reduce the number of parameters by at least half~~
- **Learn the basic principles of pruning**

□ **Network pruning without a package!**

~~(feat. Hard coding)~~

https://github.com/DolceLatte/network-pruning-the-hardway

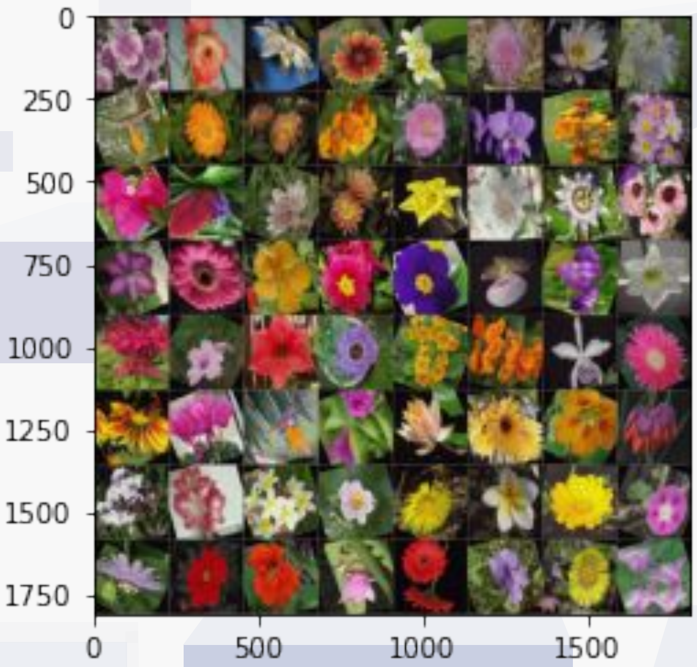
Configuration

- **Deep learning framework**
  - Pytorch
    - Package: torch_pruning 0.2.7
    - Pruning strategy -> L2 norm regularization

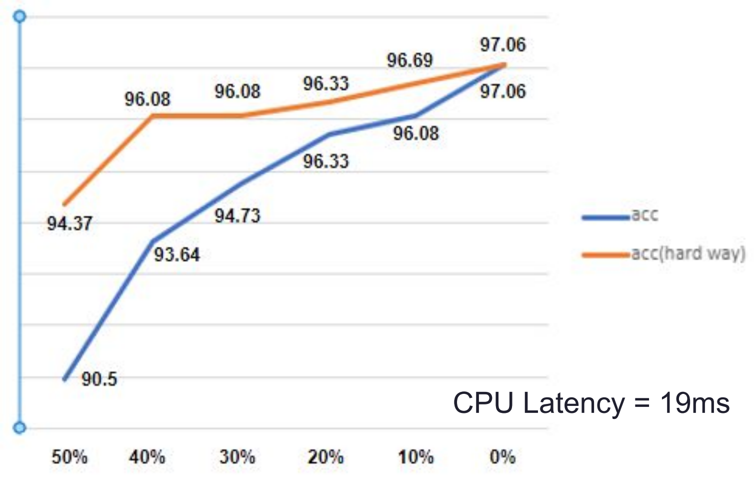
- **Proposed Model**
  - VGG11 (pruning)
  - VGG11 (Base)
  - VGG11 (pruning the hardway)

- **Dataset**
  - Flower 102

- **Evaluation metric**
  - Inference time per size of params
  - Accuracy per size of params

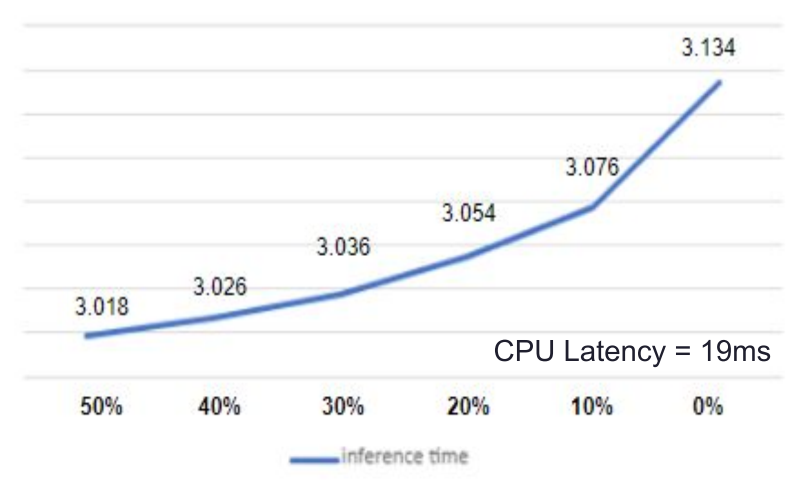


Experiments



CPU Latency = 19ms

Accuracy per pruning rate



CPU Latency = 19ms

Inference time per batch

Pruning rate	#params
0%	132,970,942
10%	118,737,885
20%	104,888,745
30%	91,489,330
40%	78,475,670
50%	65,880,210

https://github.com/DolceLatte/network-pruning-the-hardway

Conclusions

We implement a model with 90% accuracy, even though the number of network parameters has been reduced by nearly half!

Do

- ✔ Learn the basic principles of pruning
- ✔ Very simple pruning the network efficiently.

Don't

- □ Implementation of Paper
- □ Can not understanding pruning deeply

Thank You!