Artificial Intelligence Project

Study of light models to embed in systems

Tea Toscan du Plantier – Clément Patrizio

Introduction

The goal of this project is to compare different types of models and find a lightweight and efficient model.

Lightweight means that the model should have a small number of parameters.

Efficient means that it must achieve an accuracy of at least 90%.

A score will be calculated for the best model we find.

We have chosen to train our models on the CIFAR-10 dataset.

DenseNet 121

Configuration :

Batchsize : 180

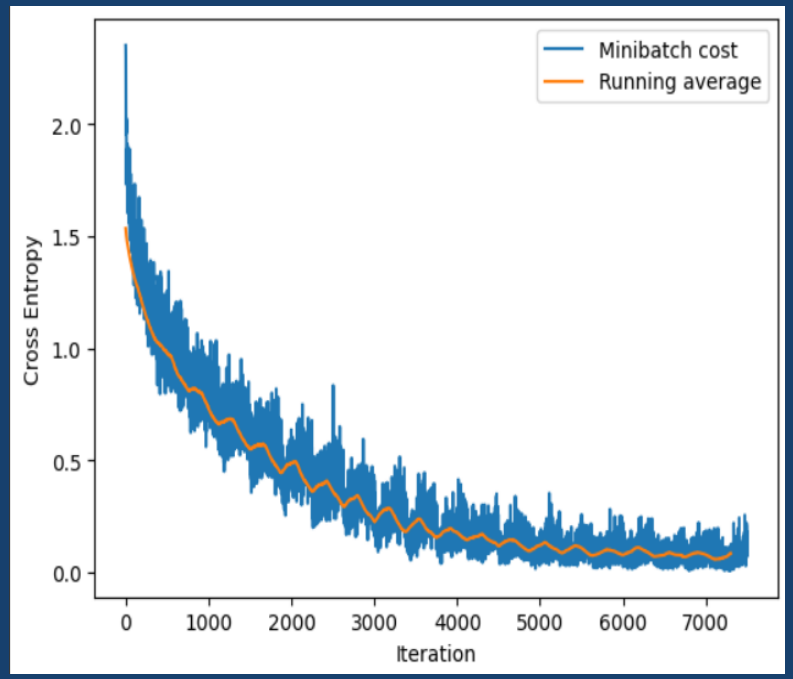
Learning rate : 0,001

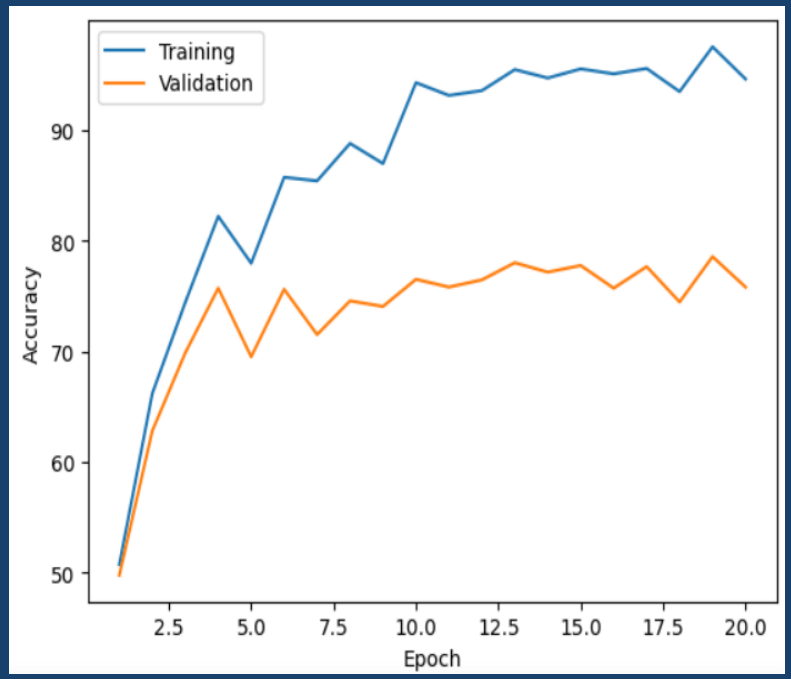
Pruning : None

Results :

7,978,856 trainable parameters

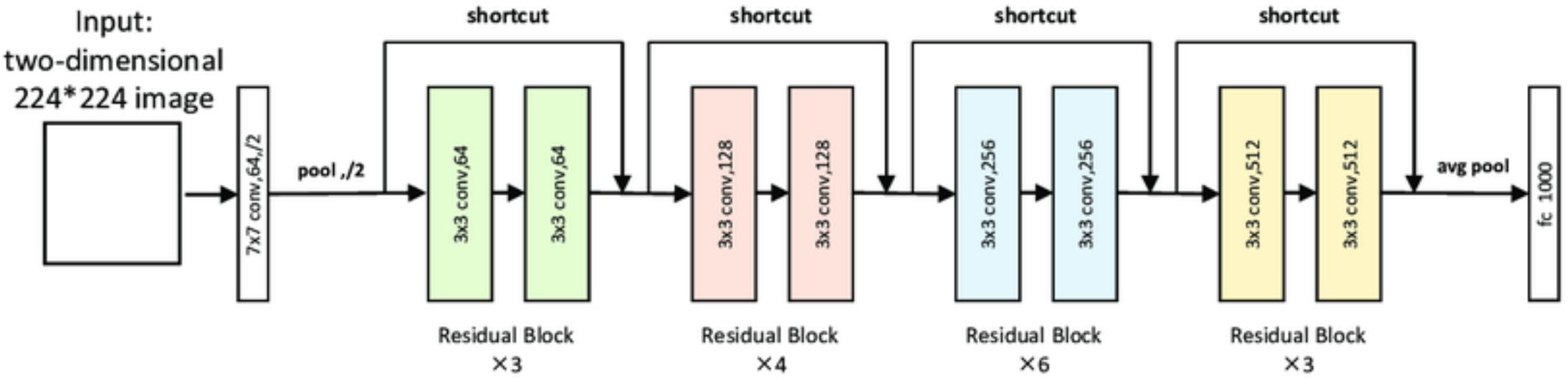
Validation accuracy : 75,80%





Too heavy for not enough accuracy results

ResNet 34


Input:
two-dimensional
224*224 image
7x7 conv,64,/2
pool ,/2
shortcut
shortcut
shortcut
shortcut
3x3 conv,64
3x3 conv,64
3x3 conv,128
3x3 conv,128
3x3 conv,256
3x3 conv,256
3x3 conv,512
3x3 conv,512
avg pool
fc 1000
Residual Block
×3
Residual Block
×4
Residual Block
×6
Residual Block
×3

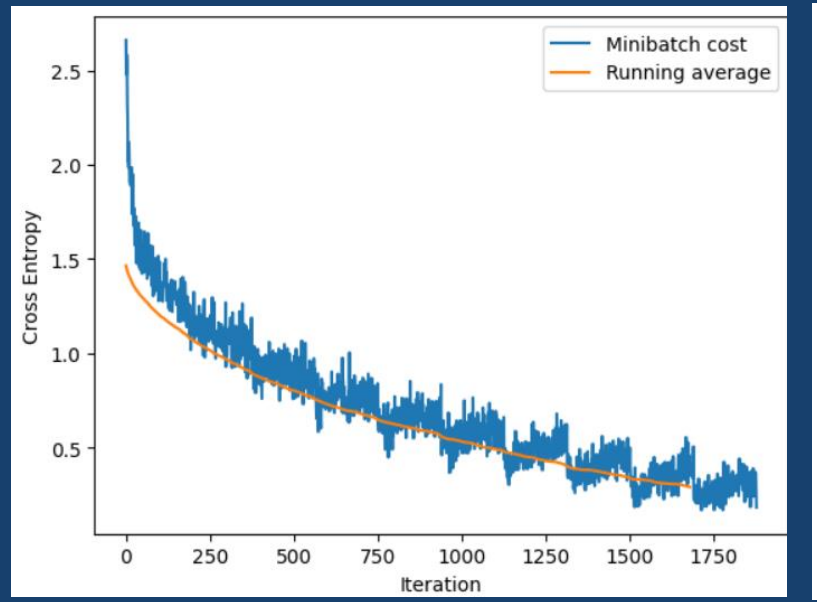
ResNet 34

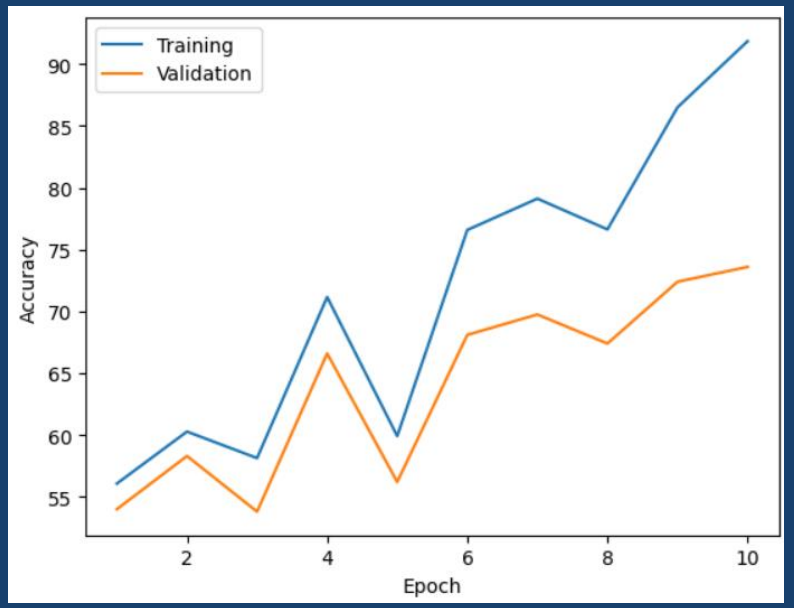
Configuration :

Batchsize : 256

Learning rate : 0,001

Pruning None





Results :

21 289 802 trainable parameters

Validation accuracy : 73,60%

Too heavy for not enough accuracy results. Let’s focus on ResNet34 with more epochs

ResNet 34 – linearly pruned – optimized

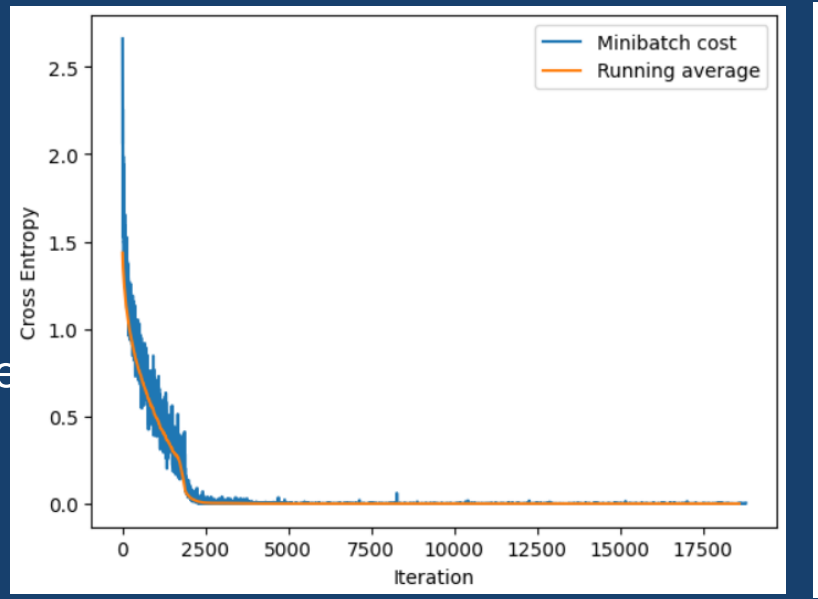
Configuration :

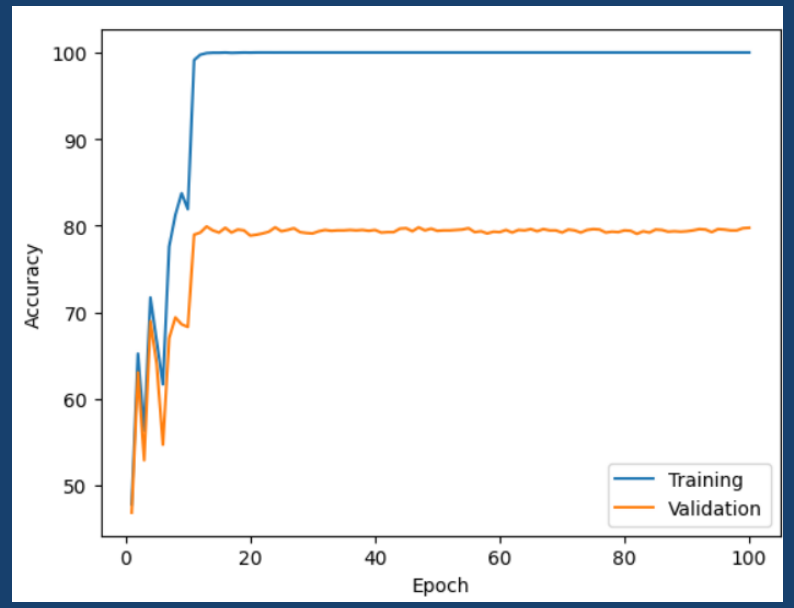
Batchsize : 256

Pruning : linear pruning by 10%

Optimizer : Adam lr=0.001, weight_decay=1e-5

Scheduler : Learning rate decrease 10 times every 10 epoch





Results :

7 597 872 trainable parameters

Validation accuracy : 80%

Linear pruning might not be adapted enough

ResNet 34 – magnitude pruning – optimized

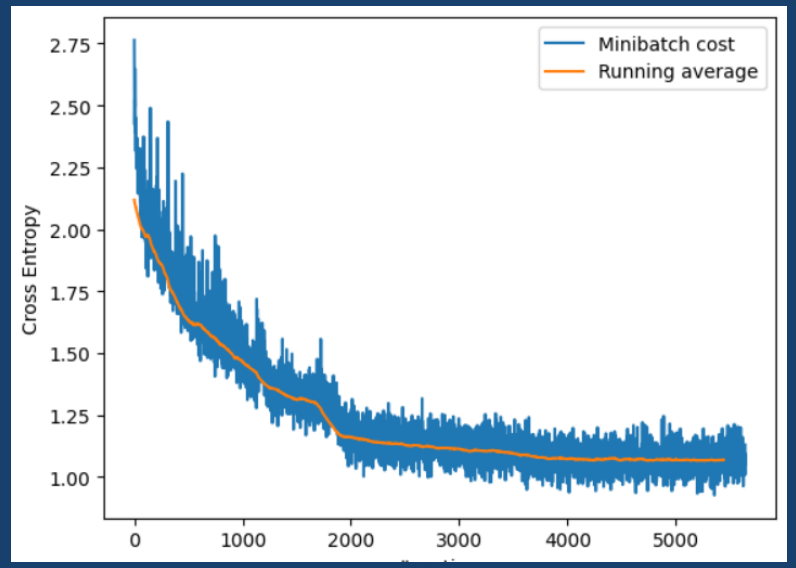
Configuration :

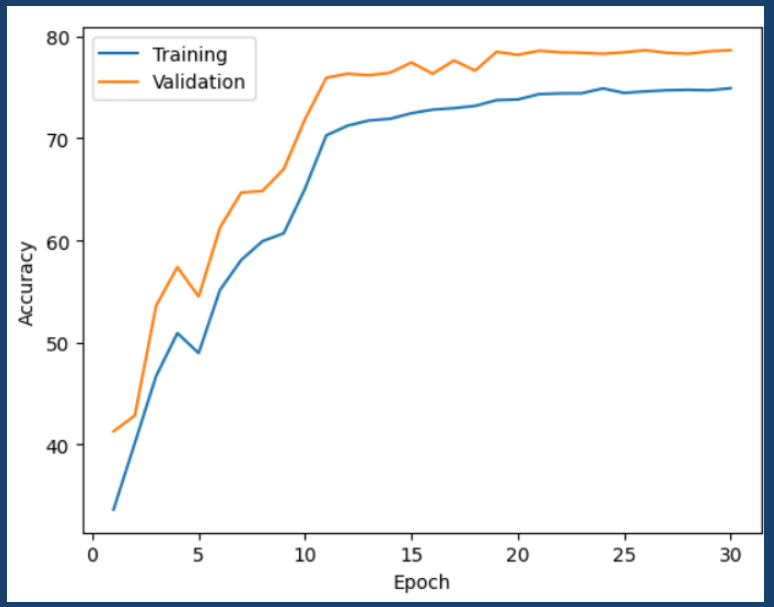
Batchsize : 128

Pruning : magnitude pruning by 20%

Optimizer : Adam lr=0.001, weight_decay=1e-5

Scheduler : Learning rate decrease 10 times every 10 epoch





Results :

2 340 117 trainable parameters

Validation accuracy : 78%

Batchsize might be too high. We try to reduce it.

ResNet 34 – magnitude pruning – optimized 2

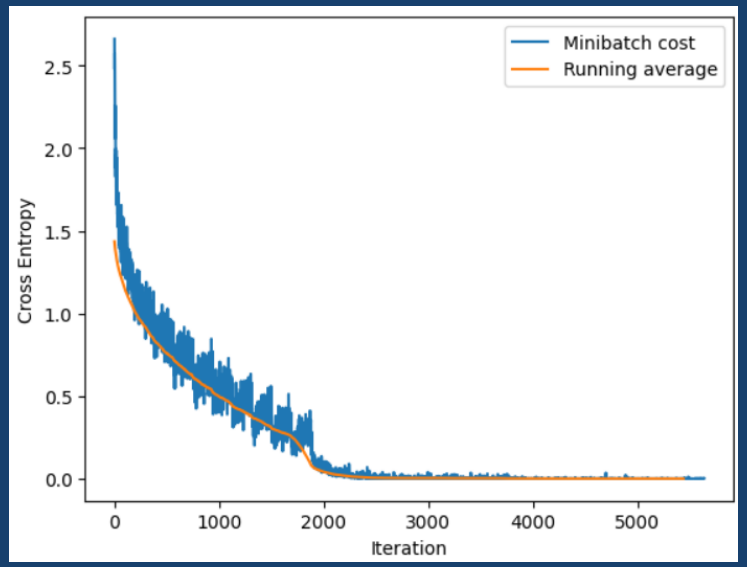
Configuration :

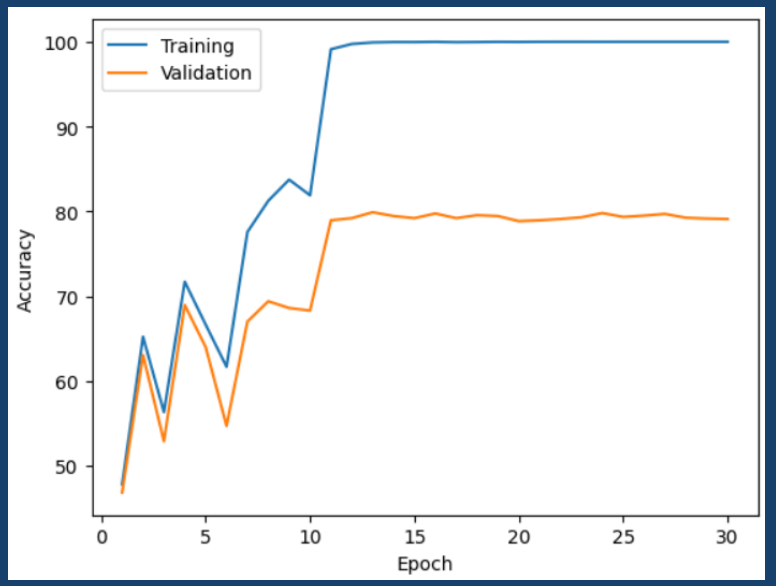
Batchsize : 64

Pruning : magnitude pruning

Optimizer : Adam lr=0.001, weight_decay=1e-5

Scheduler : Learning rate decrease 10 times every 10 epoch





Results :

2 340 117 trainable parameters

Validation accuracy : 80%

Reducing the batch size enabled us to gain few percentages but this is not significant.

Error in procedure : we are trying to reduce the number of parameters without reaching 90% accuracy first. Thus, the unstructured pruning is not adapted enough.

ResNet 34 – optimized 3

Configuration :

Batchsize : 128

Data augmentation

Optimizer : **SGD** lr=0.1,
momentum = 0.9, weight_decay=1e-4

Scheduler : Learning rate decrease
10 times every 30 epoch

Results :

21 289 802 trainable parameters

Validation accuracy : 87%

Augmenting the data base significantly increased the validation accuracy.

ResNet 34 – structured pruning – optimized 3

Configuration :

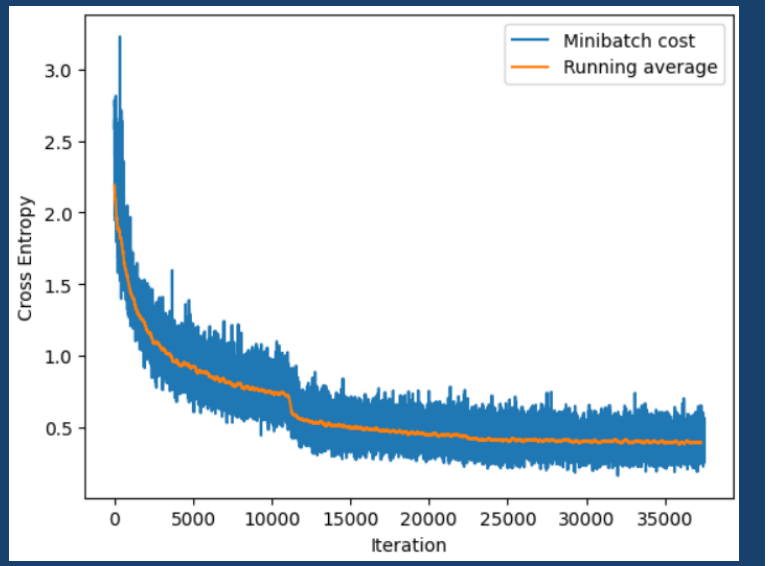
Batchsize : 128

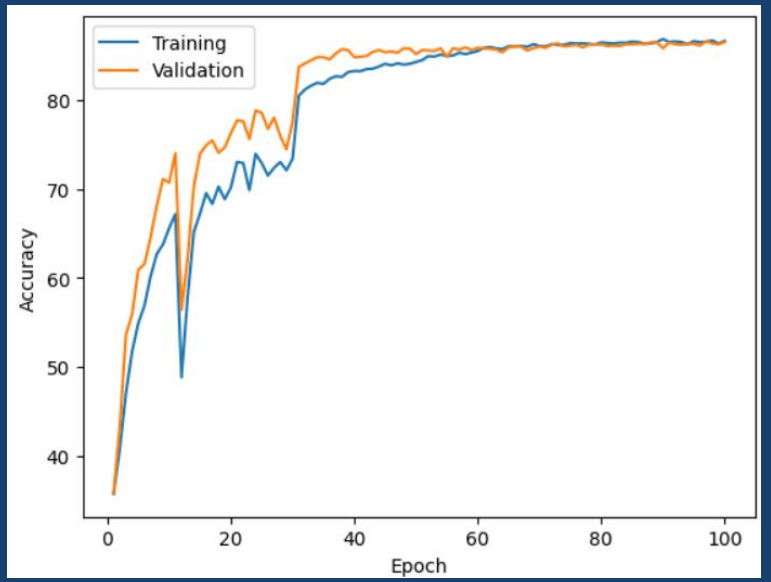
Pruning : structured pruning by 20%

Data augmentation

Optimizer : SGD lr=0.1, momentum = 0.9, weight_decay=1e-4

Scheduler : Learning rate decrease 10 times every 30 epoch





Results :

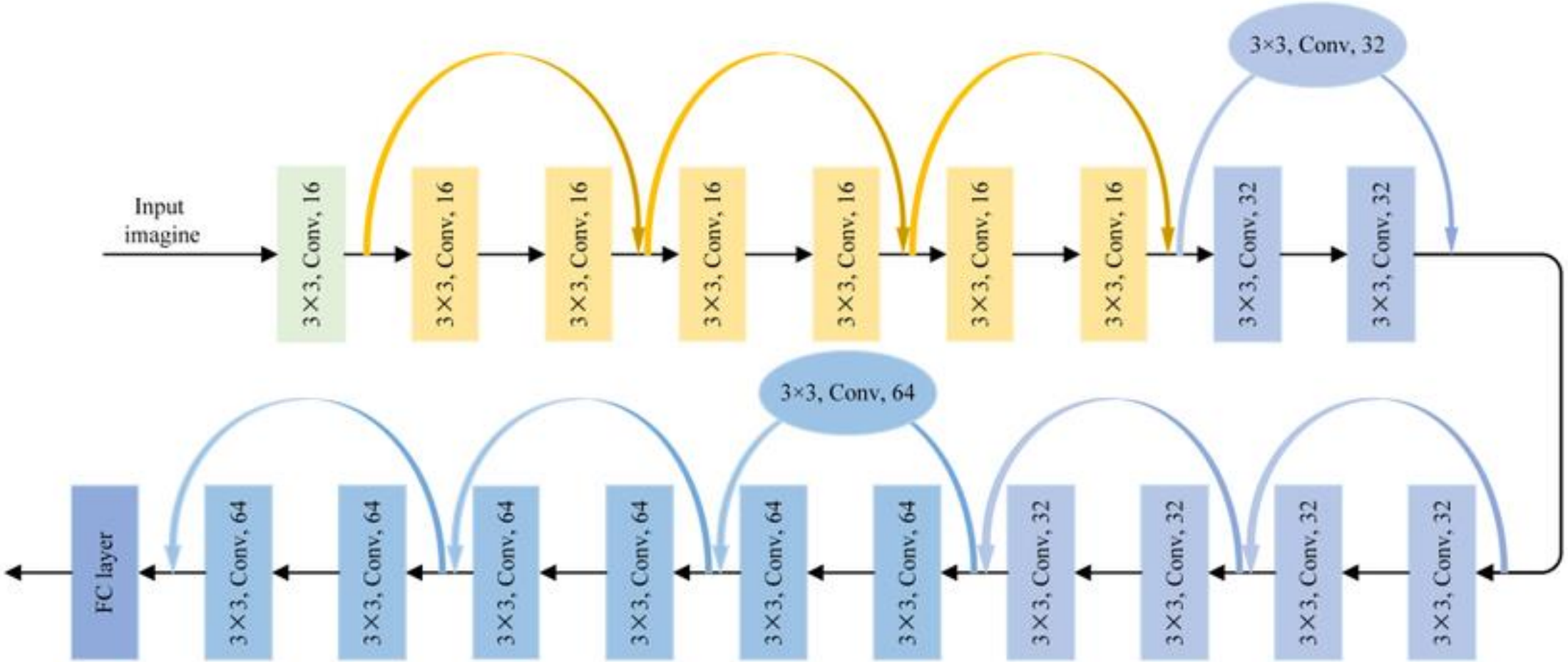
~ 2 000 000 trainable parameters

Validation accuracy : 85%

We managed to reduce the number of parameters but we are still not at 90%.

But ResNet34 is not adapted to CIFAR 10. We will focus on ResNet20.

ResNet 20


Input imagine
3×3, Conv, 16
3×3, Conv, 16
3×3, Conv, 16
3×3, Conv, 16
3×3, Conv, 16
3×3, Conv, 16
3×3, Conv, 16
3×3, Conv, 32
3×3, Conv, 32
3×3, Conv, 32
3×3, Conv, 32
3×3, Conv, 32
3×3, Conv, 64
3×3, Conv, 64
3×3, Conv, 64
3×3, Conv, 64
3×3, Conv, 64
3×3, Conv, 64
3×3, Conv, 64
FC layer

ResNet 20 – Initialisation– optimized 3

Configuration :

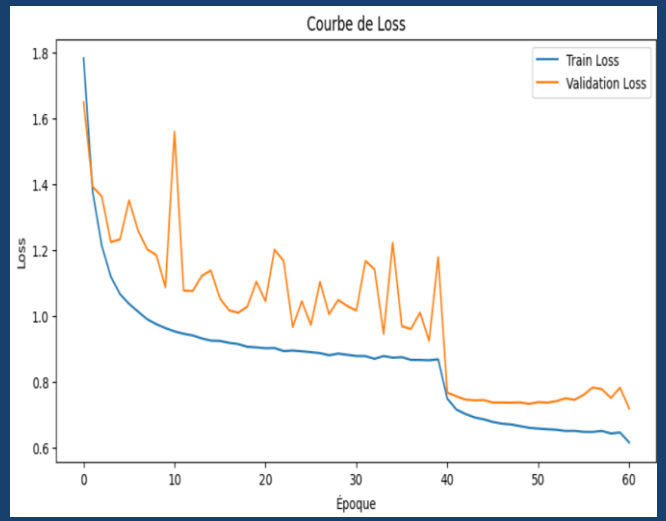
Batchsize : 128

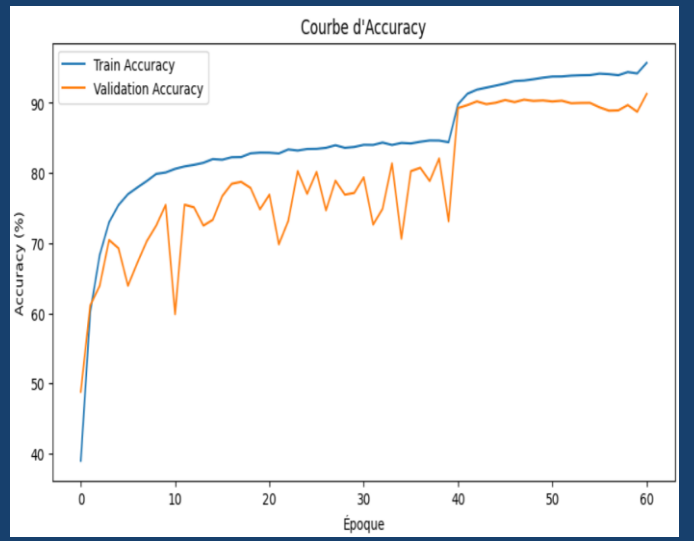
Data augmentation

Criterion : LabelSmoothingCrossEntropy

Optimizer : SGD lr=0.1, weight_decay=1e-5

Scheduler : Multistep-lr, milestone = [40,60], gamma = 0.1





Results :

272 474 trainable parameters

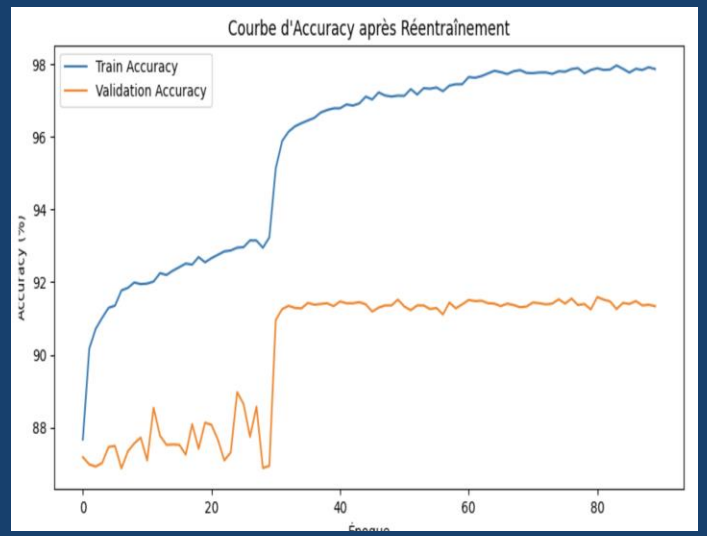
Validation accuracy : 91,27%

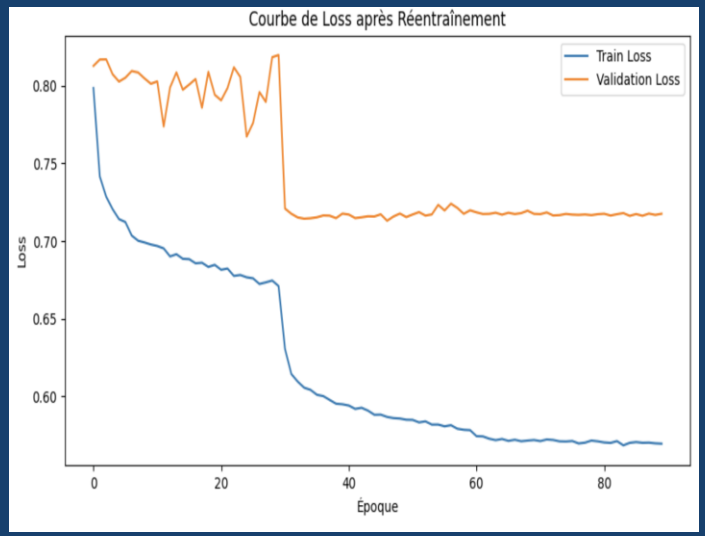
This model is significantly more precise. But we fear a small overfitting.

ResNet 20 – Fine tuning – optimized 3

Configuration :

Batchsize : 128
Pruning : Structured pruning 30%
Fine tuning
Data augmentation
Criterion : LabelSmoothingCrossEntropy
Optimizer : SGD lr=0.001, weight_decay=1e-5
Scheduler : Multistep-lr, milestone = [30,60], gamma = 0.1





Results :

Size of the model : 1.13 MB
198 309 trainable parameters
Validation accuracy : 91,59%

The risk of over fitting is still high.

ResNet 20 – Quantification attempt

To reduce the weight of the model, we could use quantification to transform float32 values into float16.

Results :

Size error

```
RuntimeError: mat1 and mat2 shapes cannot be multiplied (1x57600 and 65536x10)
```

ResNet 20 – further techniques

The model overfits. To reduce the overfitting, we could vary the data augmentation, such as data mixup, or use distillation.

PreActResNet 18 – linear pruning – optimized

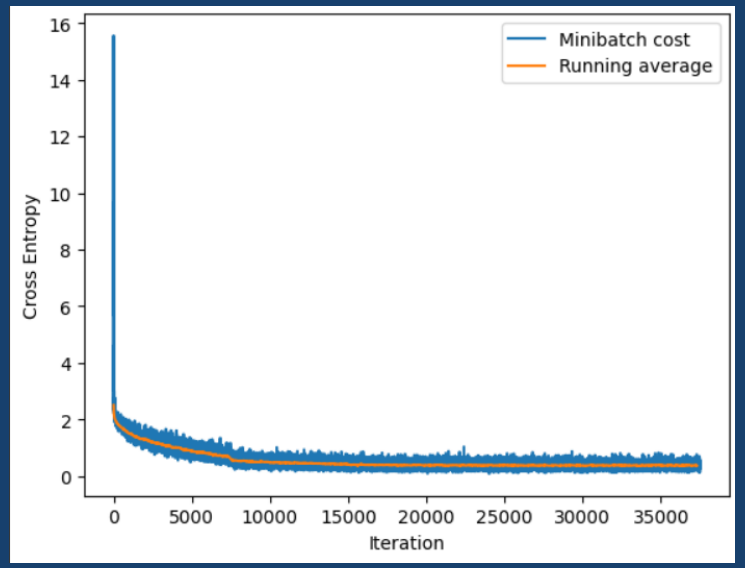
Configuration :

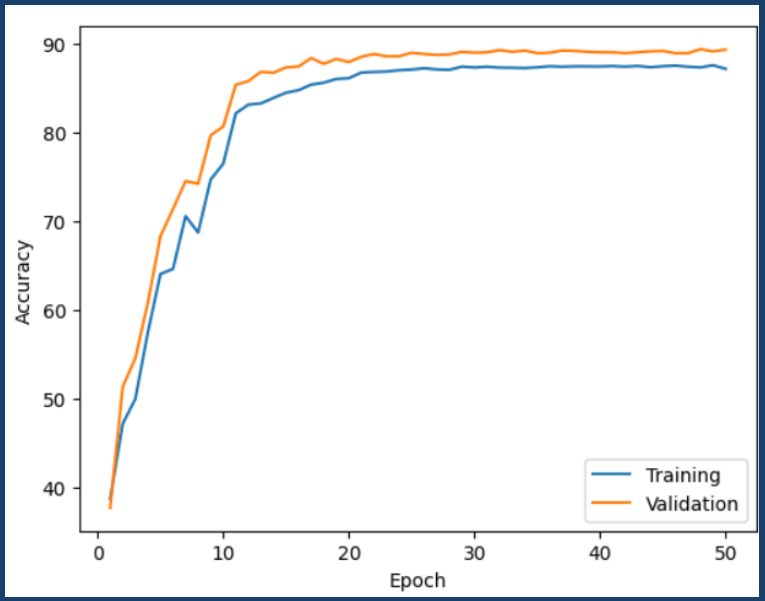
- Batchsize 128
- Learning rate 0,001
- Linear pruning 20%
- Optimization : Adam lr=0.001, weight_decay=1e-5
- Learning rate decrease 10 times every 10 epoch

Results :

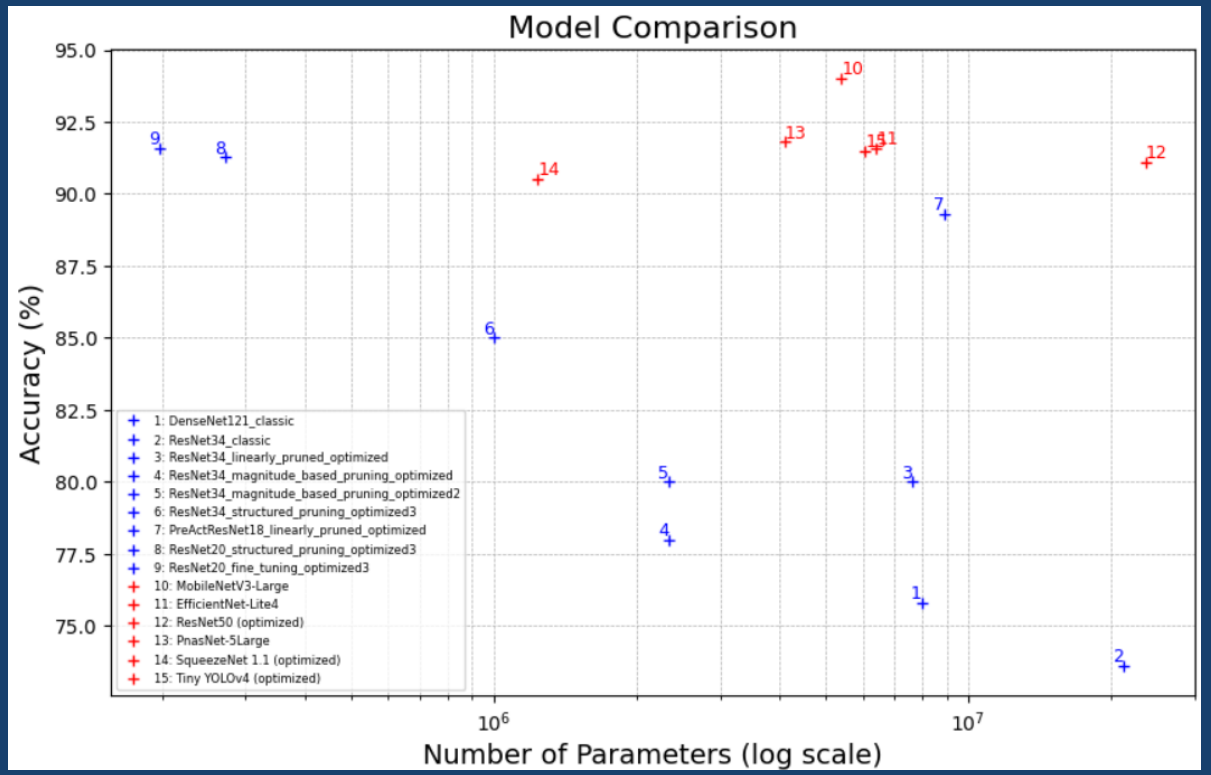
- 8 934 873 trainable parameters
- Validation accuracy : 89,30%

This model is also promising.





Comparison



Comparison graph of our models and existing models from the literature

Score

The score of our best model is calculated this way :

- Percentage of structured pruning Ps = 0,3
- Percentage of unstructured pruning Pu = 0
- Weights quantification qw (in bits) = 32
- Activation quantification qa (in bits) = 32
- Number of parameters w = 198309
- Number of flops f = 29055103

$$\text{score} = (1 - (P_s + P_u)) \left(\frac{q_w}{32}\right) \frac{w}{5.6 \times 10^6} + (1 - P_s) \left(\frac{\max(q_w, q_a)}{32}\right) \frac{f}{8.3 \times 10^8}$$

Here, the score is : 0.0563