IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

MICRONET CHALLEGNE

EFFICIENT DEEP LEARNING

Alexandre Gorio
Aziz Sellami

SOMMAIRE

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

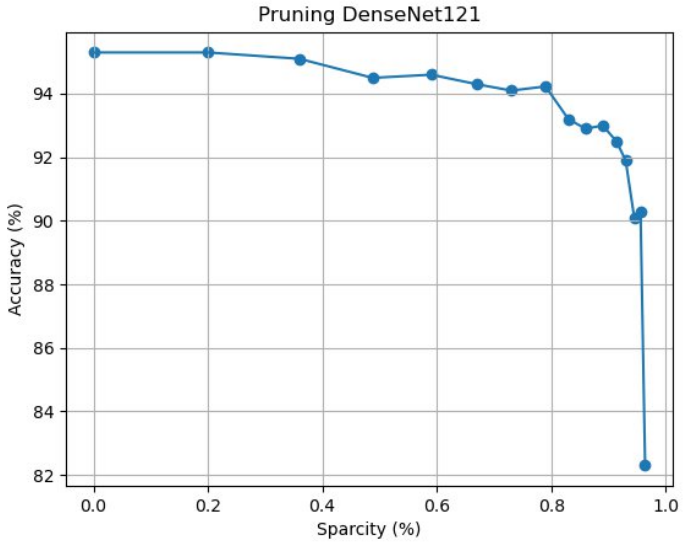
MICRONET CHALLENGE
A SCORE UNDER 0.02
DENSENET

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

STARTING THE CHALLENGE

What will be used ?

- Whole training set for training and test set for evaluation
- Data augmentation
  - Default DA + Mixup => +2.7% from the default DA
- DenseNet121 is too **big**
  - Baseline accuracy = **95.39%**
  - Score = 2.2
  - Achieved unstructured pruning **96.48%** with accuracy **>90%**
  - From 7M to 0.35M parameters



Pruning rate	0.2
Epochs for each iteration	50
Optimizer	Adam (default)
Time needed	24 hours

DENSENET OR MOBILENET?

Looking for another promising architecture

- Searching for mobilenet and mobilenetV2
- Harder to change

- Modified mobilenet
  - Accucary=86%
  - Score=0.1

- MobilenetV2
  - Accuracy=94.14%
  - Score=0.4

ARCHITECTURE OF DENSENET

Working on growth rate

- Significant gain on the score
  - From 2 to 0.02

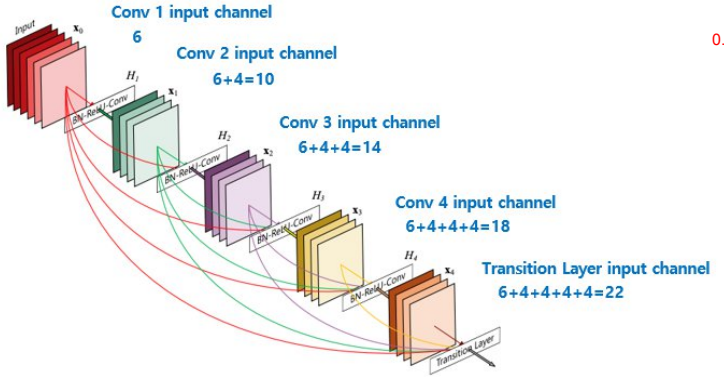
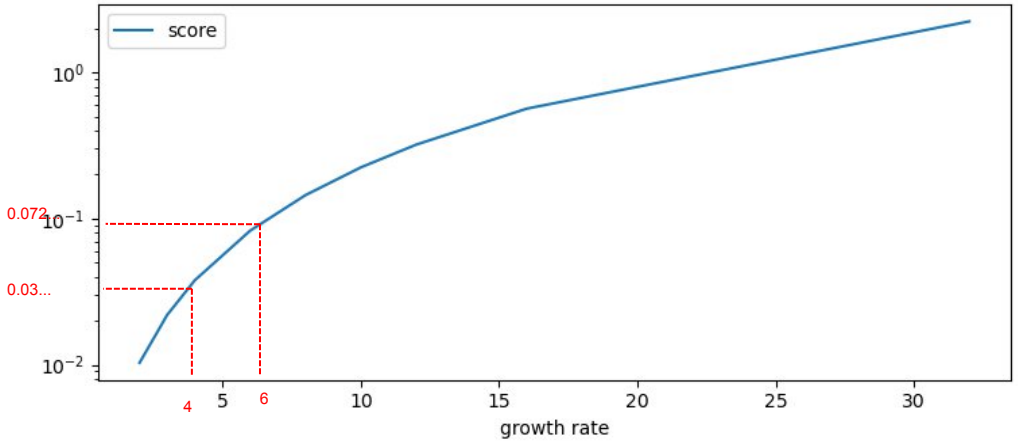


Figure 1: A 5-layer dense block with a growth rate of $k = 4$. Each layer takes all preceding feature-maps as input.



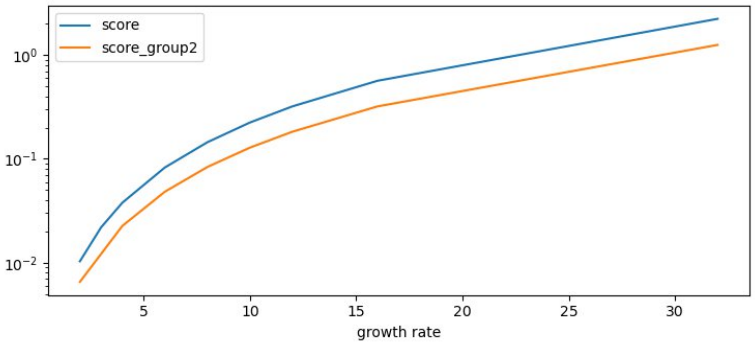
Memory needed in GPU
Training time

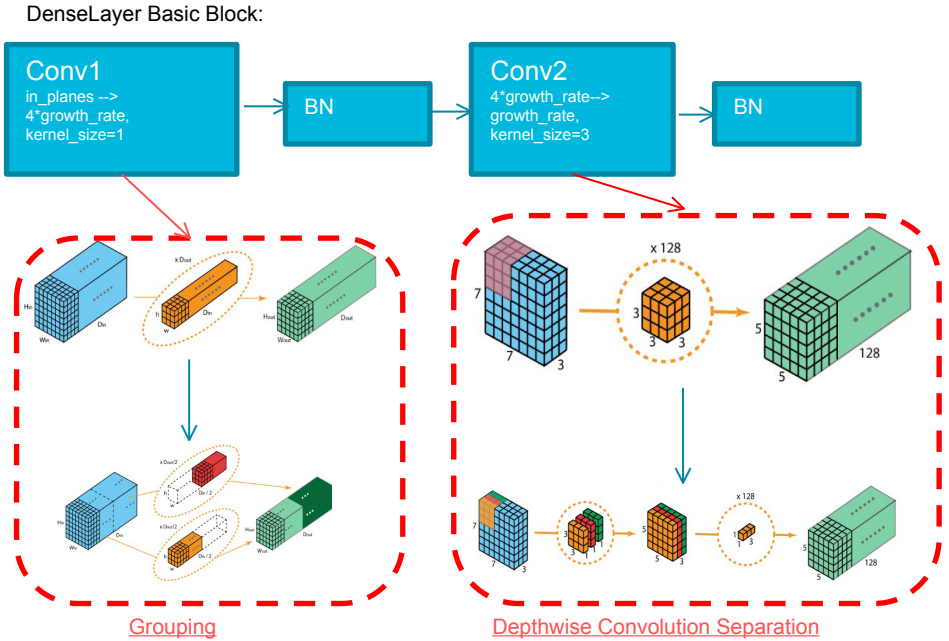
Baseline: DenseNet, Growth Rate=4, Score = **0.0335**, Accuracy = **90.61**

FACTORISING

Groups=2 & Depthwise separated convolution

- Reducing the score
- Reduce the accuracy
- Reduce the number of operations

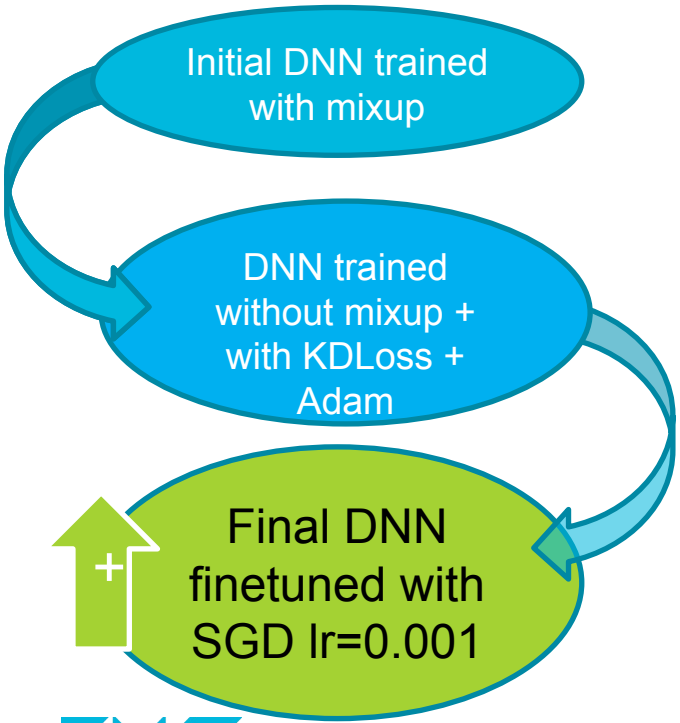




Baseline: DenseNet gr4, DWS and Groups=2, Score = **0.0205**, Accuracy = **90.54**

DISTILLATION METHOD

How? How good?



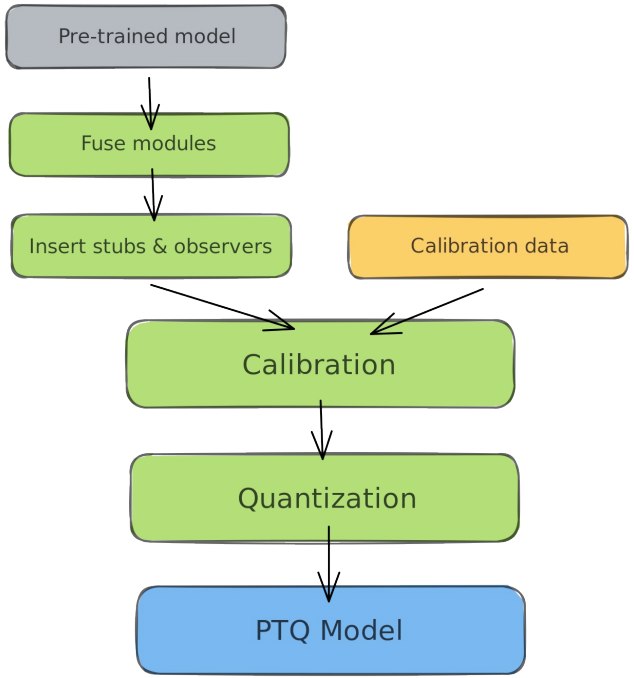
Student	Master	Gain (%)
DN GR=3 Accuracy=88.5	DN GR=4 Distilled	+1.52
DN GR=4 **65% sparcity** Accuracy=90.0	DN GR=4 Distilled	**+0.79**
DN GR=4 Accuracy=90.6	DN GR=8 Accuracy=93.2	+0.94

QUANTIZATION

Reducing the number of bits

- Discovering 8 bits quantization by pytorch
- Quantize everything or not ?
- Our implementation
  - Post-training Quantization int8 (Graph FX)
  - Partial and Local quantization
  - Weight and activations Quantization ("fbgemm")
    Activation=HistogramObserver
    Weight=default_per_channel_weight_observer

Factor 2 gain in the score !



FINAL PRUNING

Going for the best score possible

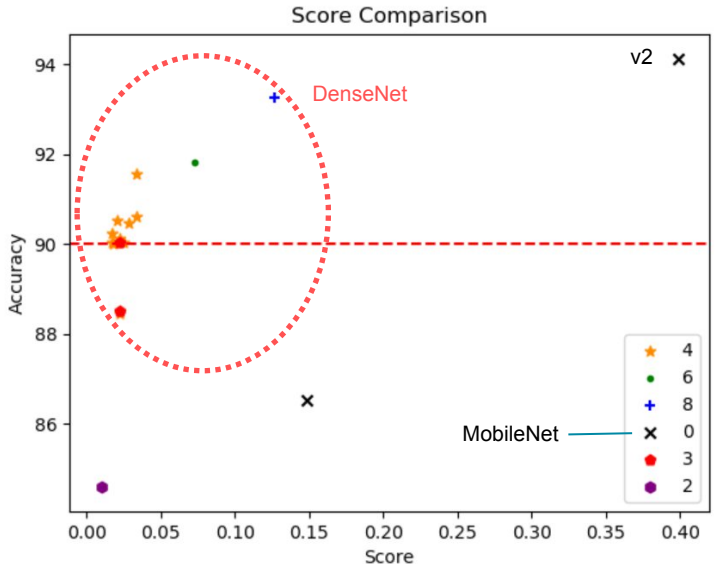
- Pruning
  - **Unstructured**
  - Global
  - **Iterative Training**
  - Training to regain 90%
  - Limited to 50 epochs

DenseNet GR=4,groups=2	
Pruning rate Step (%)	Sparcity (%)
10	27.10
5	30.17

FINAL RESULT

Our models

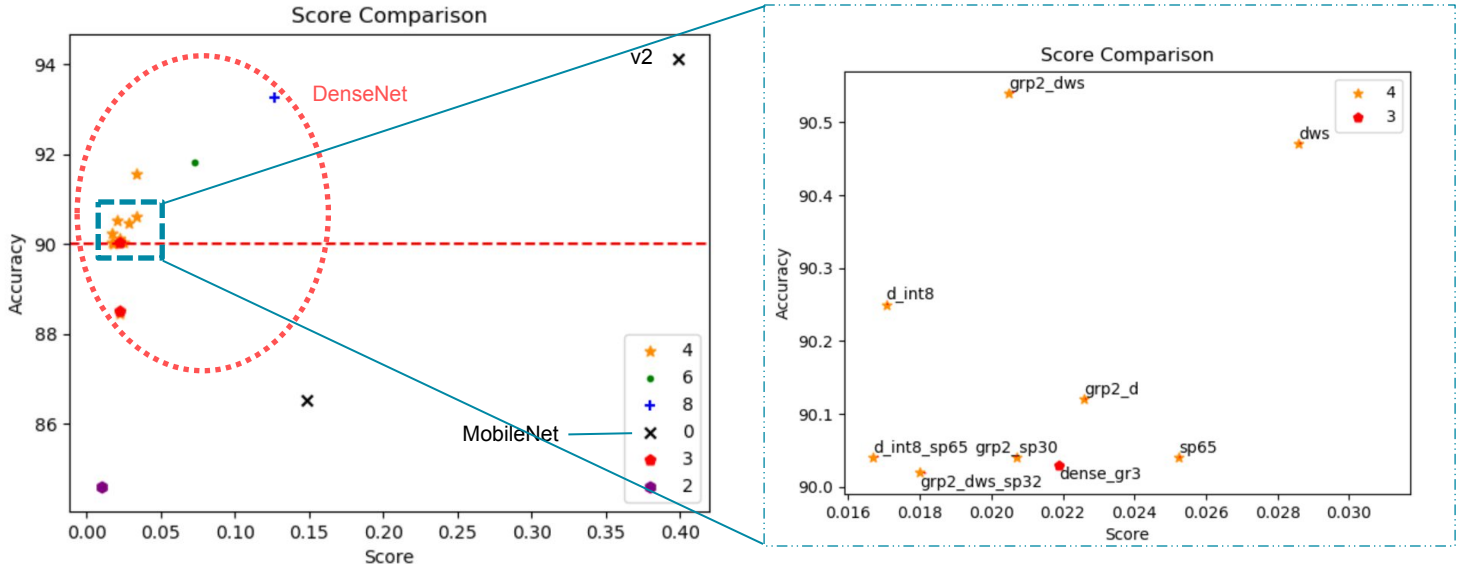
0.01678 !!



FINAL RESULT

Our models

0.01678 !!



Best Model : DenseNet, growth rate = 4
Distilled, quantization int8 (local quant, except : ['trans3', 'dense1', 'dense2'])
Sparsity 65 %

- sp : sparsity
- dws : Depthwise Separated Convolution
- grp : Groups=2
- int8 : Quantization int8
- gr : Growth Rate
- d : distilled

FINAL RESULT

Our best proposition for the contest

Name	DenseNet
Score	0.0167
FLOPS	0.0154
Parameters	0.00132
Accuracy	90.04
Sparcity	65.1
Quantization	int8 / int16
Factorisation	none

FINAL RESULT

Other interesting proposition for the contest

Model	Growth Rate	Sparsity	Accuracy	Encoding (bits)	Groups	Distillation	Score	Params	Comment
DenseNet	4	0,651	90,04	8	x	o	0,0167	70922	local quant int8 and sparse
DenseNet	4	0	90,25	8	x	o	**0,01706**	70922	distill before ptquant int8
DenseNet	4	0,32	90,02	16	2	o	0,018	66282	DWS, groups - conv1
DenseNet	4	0	90,54	32	2	o	0,0205	66282	DWS, groups -conv1
DenseNet	4	0,3017	90,04	32	2	x	0,0207	35461	
DenseNet	3	0	90,03	32	x	o	0,0219	34631	
DenseNet	4	0	90,12	32	2	o	0,0226	35461	
DenseNet	4	0,651	90,04	32	x	x	0,02524	70922	

THANKS FOR YOUR ATTENTION

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom