Machine Learning under resource constraints

Nathan Greffe

26/06/2019

ULiège

Objectives

Investigation and comparison of different techniques to improve **accuracy for a given inference time** for the task of **image classification** using **Convolutionnal Neural Networks**.

Experiments carried out on a **Raspberry Pi 3B** and the **CIFAR-10** dataset, using **Tensorflow-lite**.

Inspected techniques

1. Model architecture (Wide)ResNet, EffNet, SqueezeNext, MobileNetv1/v2, ShuffleNetv1/v2, MnasNet + adding Squeeze and Excitation blocks
2. Channel pruning as an architecture search Fisher Pruning, NetAdapt, Morphnet + trying out modifications of these algorithms
3. Knowledge distillation Vanilla, Teacher Assistant KD
4. Integer quantization Tensorflow integrated quantization

Model architecture as a block search

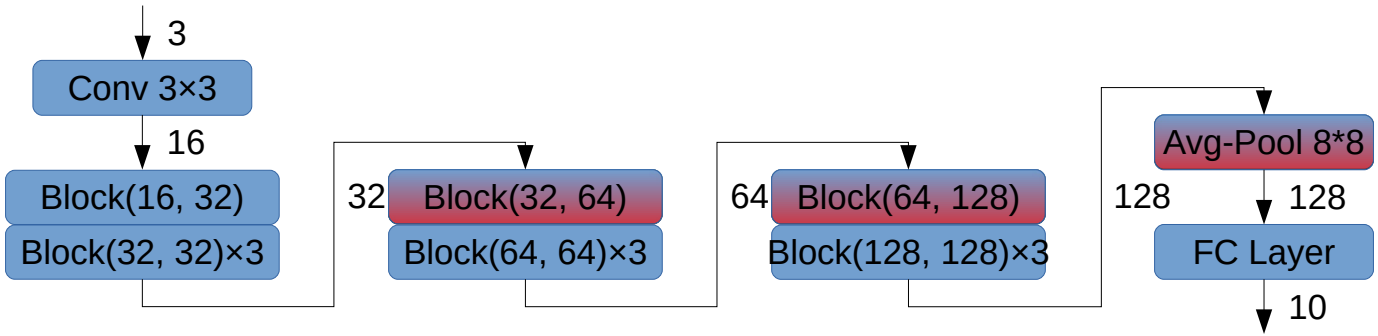
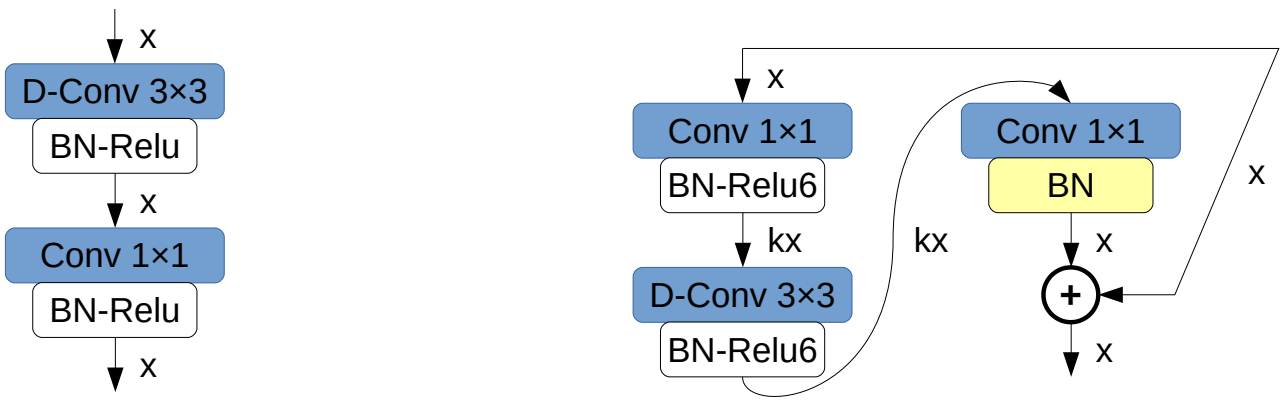


Figure 1 – Architectures are a stack of blocks



(a) MobileNet-v1 **(b)** MobileNet-v2

Figure 2 – Examples of blocks

Model architecture results (no SE blocks)

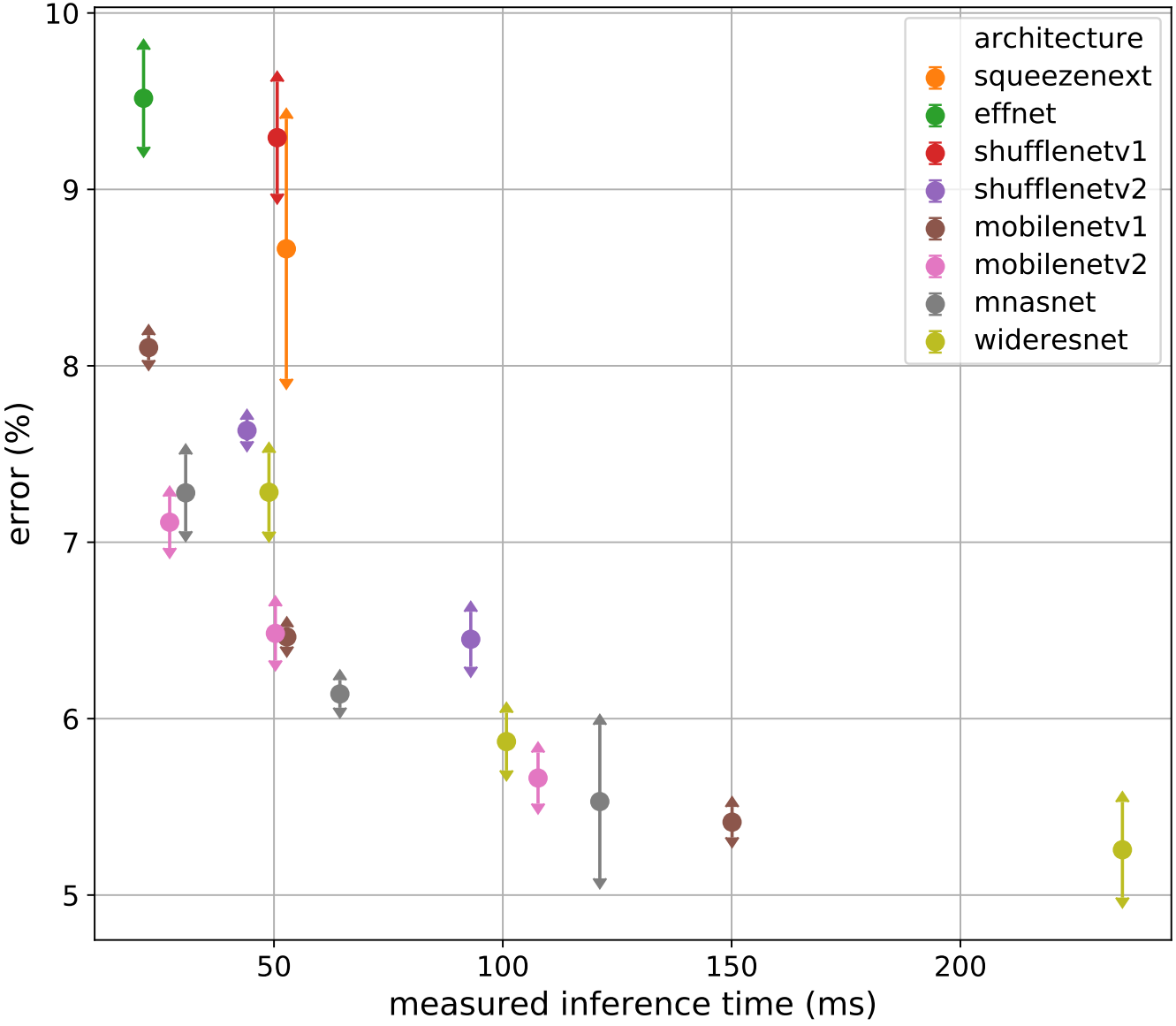


Figure 3 – results (no SE blocks)

Squeeze and Excitation blocks

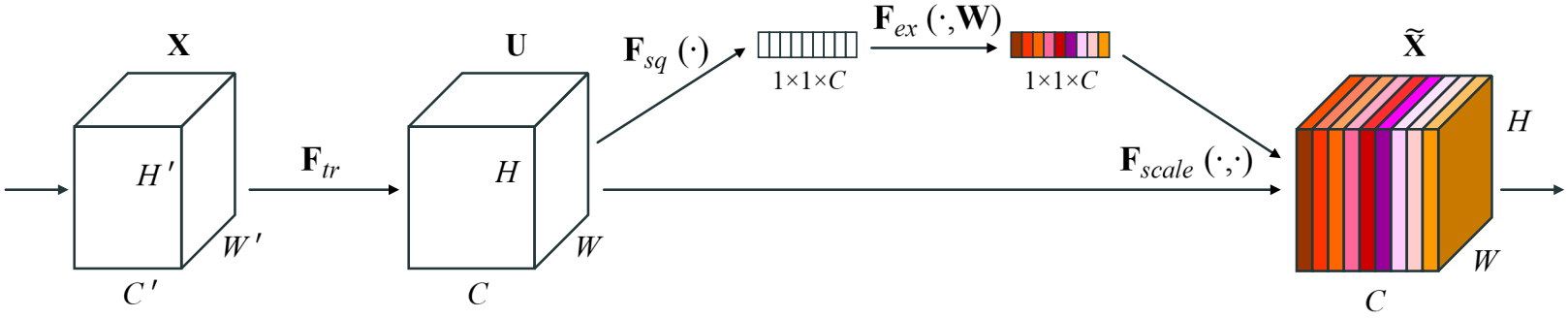


Figure 4 – Squeeze and Excitation block

Model architecture results (with SE blocks)

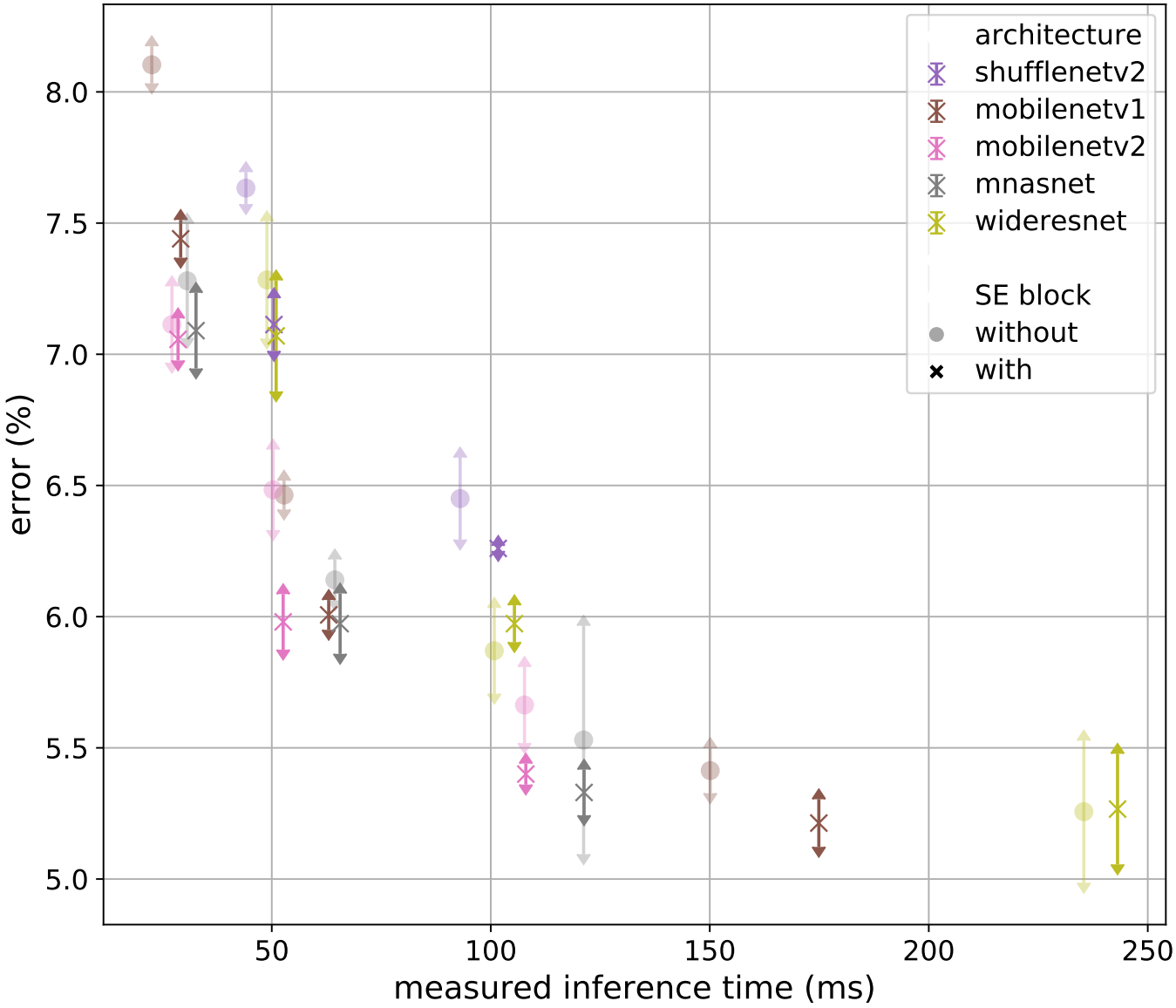


Figure 5 – results (with SE blocks)

Channel pruning as an architecture search

Pruning (removing channels and possibly layers in) a Neural Network is better suited as an architecture search procedure.

7.88% of error after pruning. **5.74**% after retraining from scratch.

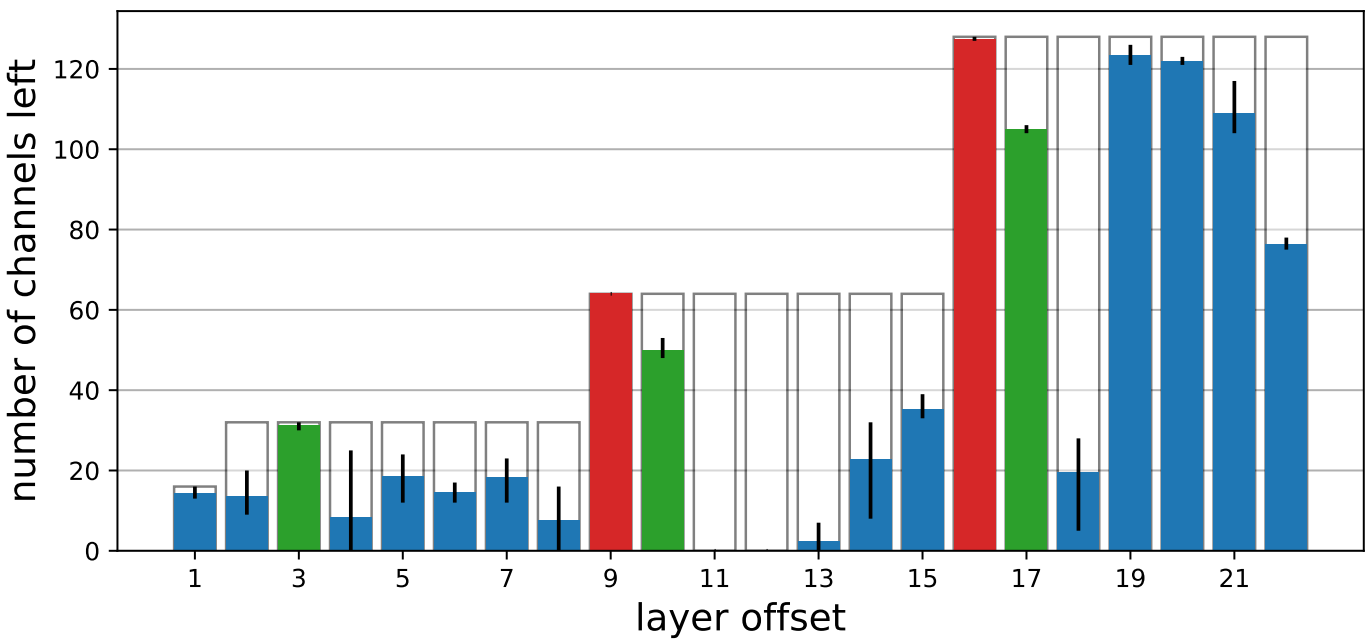


Figure 6 – number of channels per layer (improved Fisher pruning)

Channel pruning results

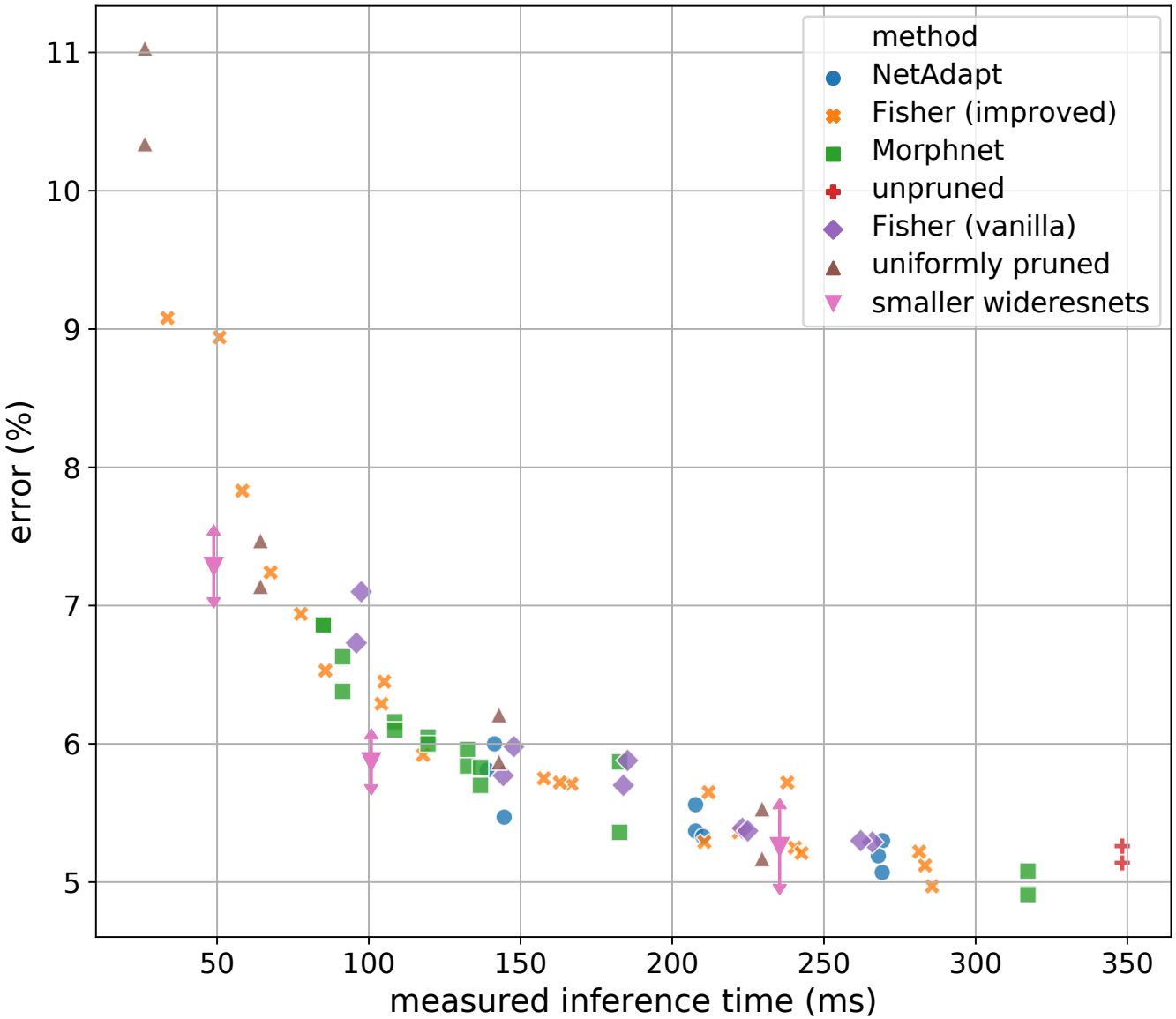


Figure 7 – performances of the pruning algorithms, on a WRN-40-2

(Teaching assistant) Knowledge distillation

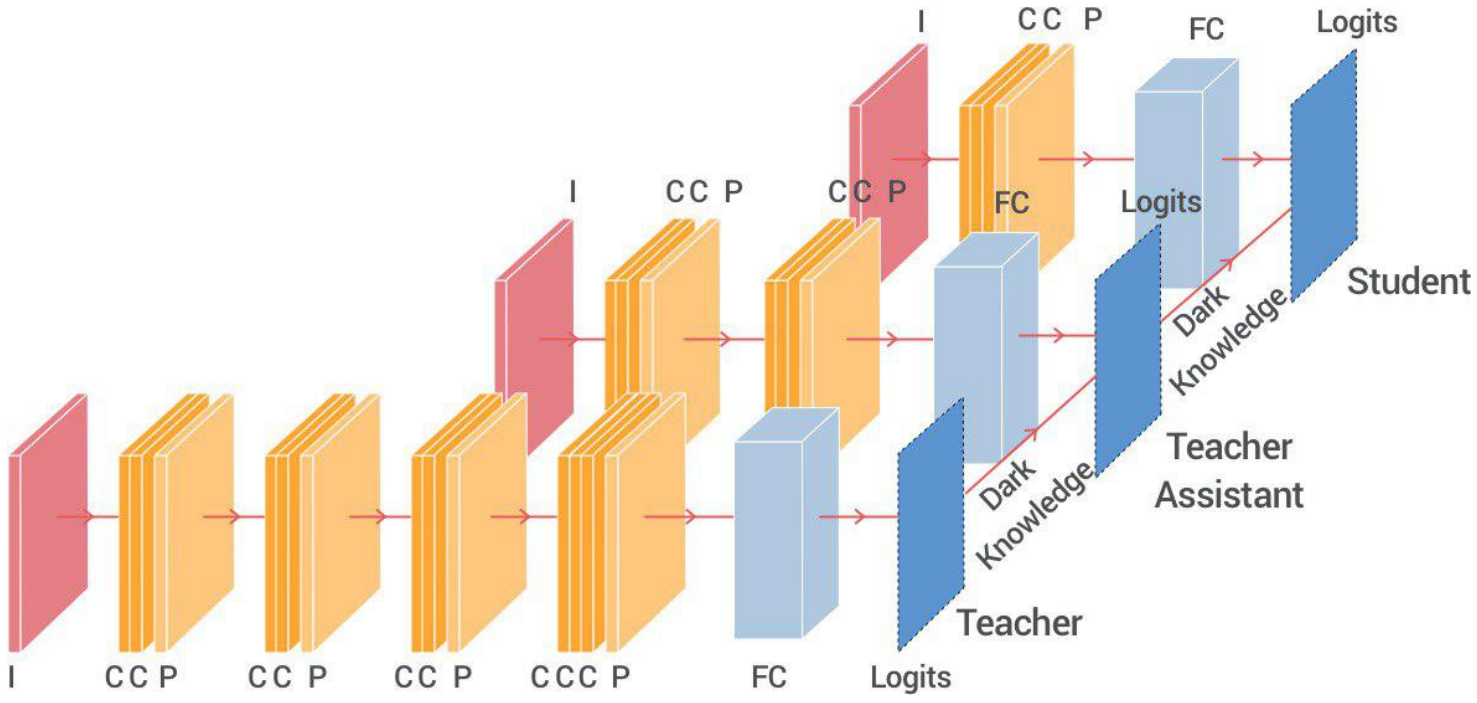


Figure 8 – Teaching assistant KD illustration

TAKD results

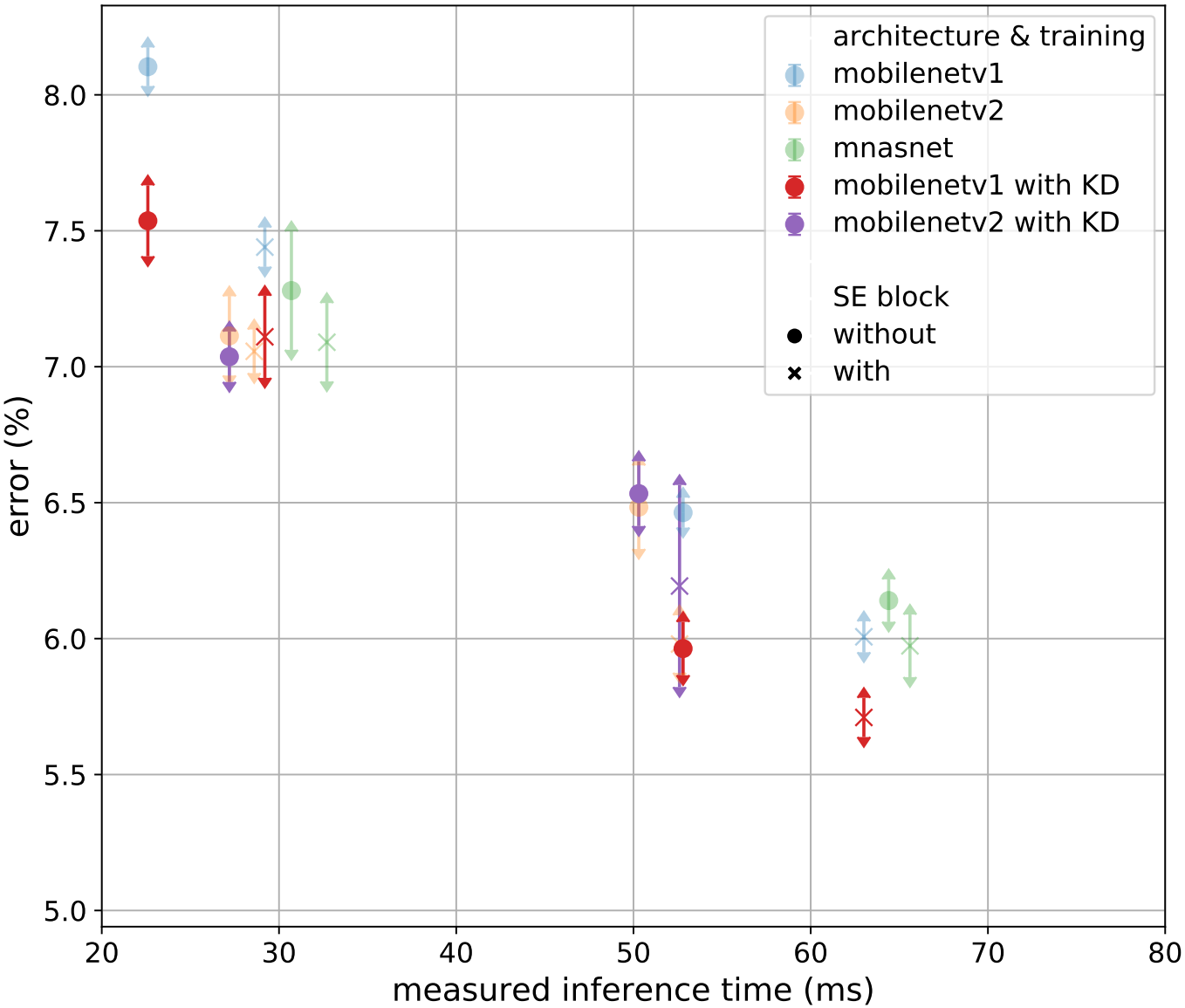


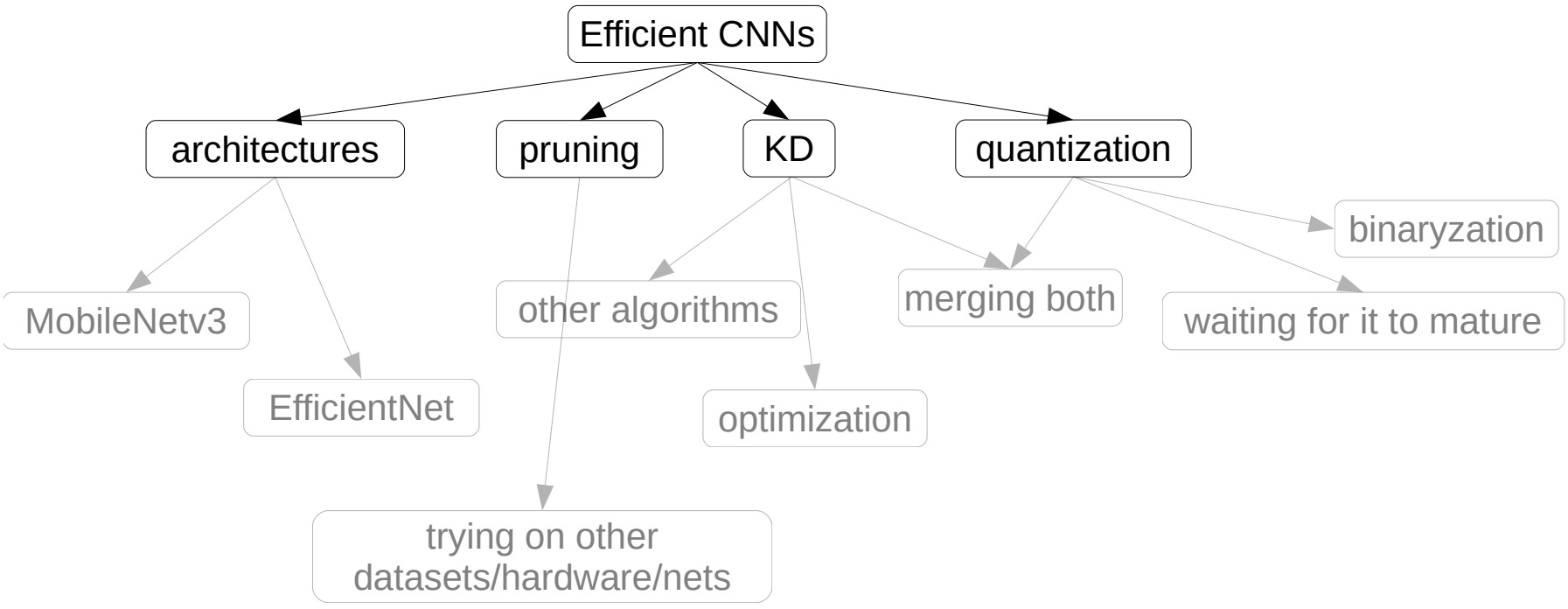
Figure 9 – performances improvements when using KD

Conclusion

- **Architecture** : MobileNetv2 with Squeeze and Excitation blocks works very well.
- **Pruning** : is outperformed by designing dedicated architecture.
- **Knowledge distillation** : helpful for MobileNetv1. Other algorithms/procedures might give better results on more networks.
- **Quantization** : not mature for the moment in Tensorflow.

Perspectives

Breadth-first seach of efficient image classification solutions, opens many doors for future works



Quantization : new results

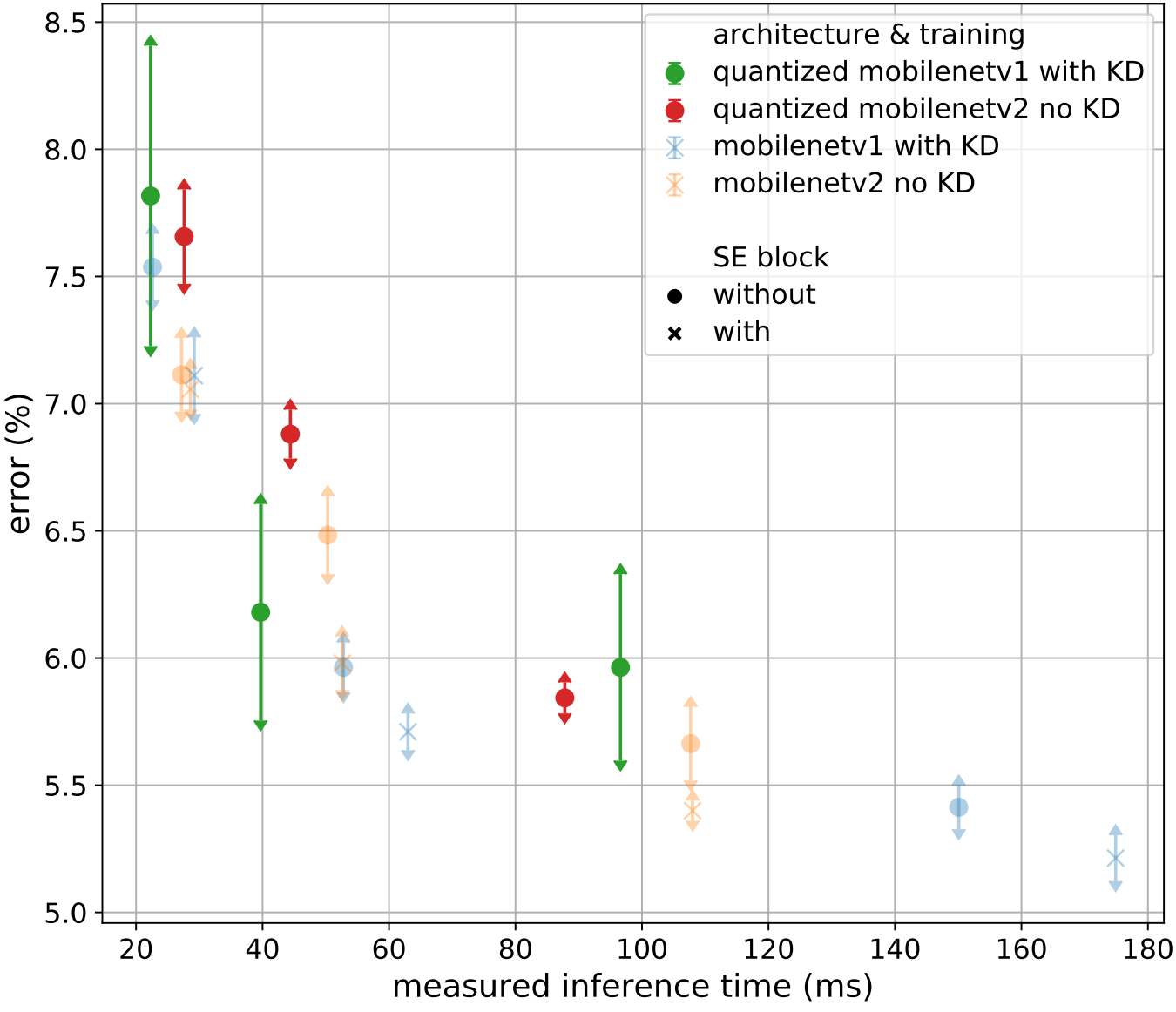


Figure 10 – Post training quantization without SE blocks