Compact CNN Structure Learning by Knowledge Distillation

Waqar Ahmed[1,2], Andrea Zunino[3], Pietro Morerio[1] and Vittorio Murino[1,3,4]

{waqar.ahmed, pietro.morerio, vittorio.murino}@iit.it; andrea.zunino@huawei.com

[1]Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia, Genova, Italy

[2]Dipartimento di Ingegneria Navale, Elettrica, Elettronica e della Telecomunicazioni, University of Genova, Italy

[3]Ireland Research Center, Huawei Technologies Co., Ltd., Dublin, Ireland, [4]Dipartimento di Informatica, University of Verona, Verona, Italy

ISTITUTO ITALIANO DI TECNOLOGIA
PATTERN ANALYSIS AND COMPUTER VISION

ICPR 2020 — 25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION, Milan, Italy 10 | 15 January 2021

Motivation

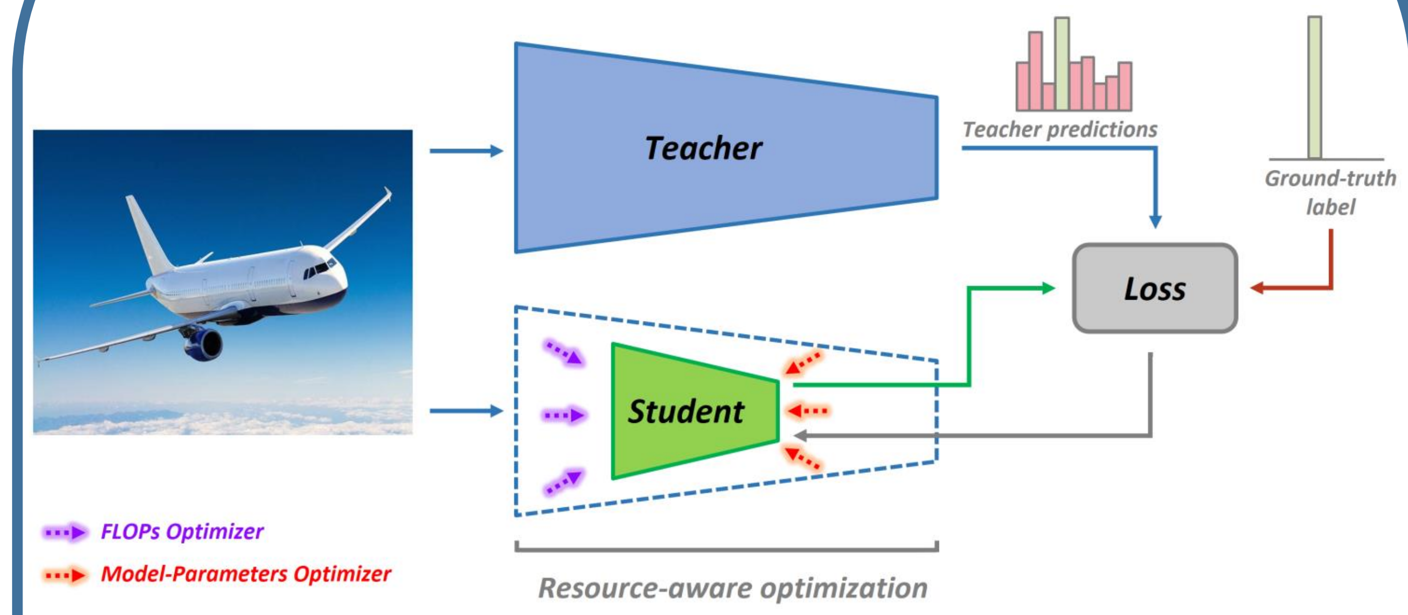
CNNs are ubiquitous in computer vision. It is well known that they require considerable resources in terms of both **Computation and Memory**, being often deployed on big and powerful gpus. Compression techniques can partially handle these issues, resulting in smaller models with less parameters and Floating point operations (FLOPs). <u>However, complexitiy reduction usually comes at the price of a drop in the model performance.</u>

Contributions

We propose a novel pipeline which leverages **Resource-aware optimization** and **Privileged Information (PI)**

- **Resource-aware optimization** breaks down the network in smaller instances with different compression needs.
- **Privileged Information (PI)** is provided during training in the form of extra supervision in a teacher-student framework [1].

Method



$$\min_{\theta_1}\min_{\theta_2}\frac{1}{N}\sum_{i=1}^{N}\Big[(1-\lambda)\,l(y^i,\sigma(f(x^i,\theta_1,\theta_2)/T)) + \lambda\, l(z^i,\sigma(f_t(x^i,\theta_1,\theta_2)/T)) + \alpha\,(\mathcal{C}_{FLOP}(\theta_1)+\mathcal{C}_{PARAM}(\theta_2))\Big]$$

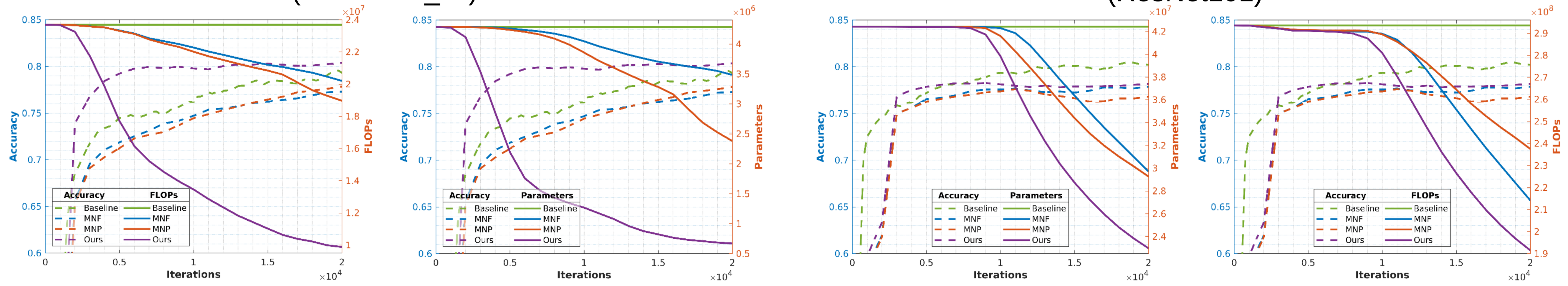
$\theta_1 \cup \theta_2 = \theta,\ \theta_1 \cap \theta_2 = \varnothing$ is a partition of the weights

- While being compressed, the **student network**, namely the one being compressed, tries to mimic the predictions of the uncompressed **teacher network**, which retain some useful information on the label's distribution.
- We propose a configuration in which the lower half of the network is optimized for FLOPs and the upper half is optimized for model parameters.

Results

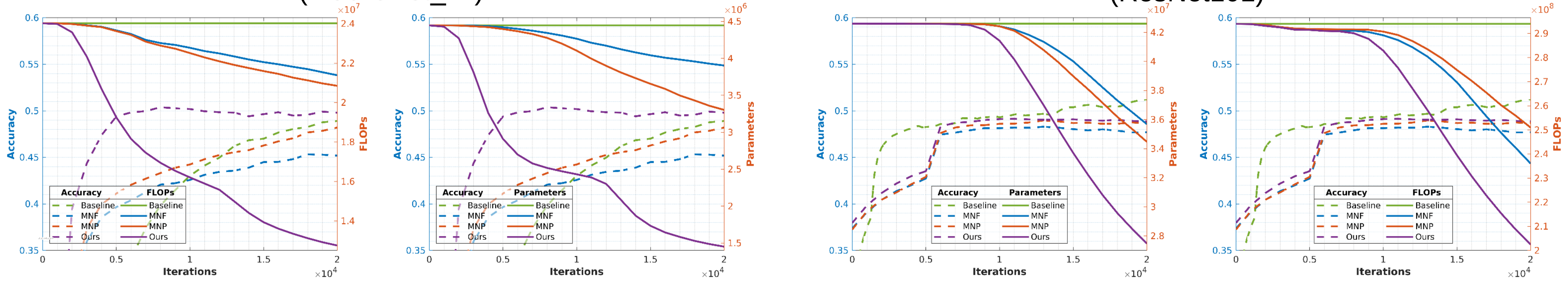
CIFAR 10

(MobileNet_v2) (ResNet101)



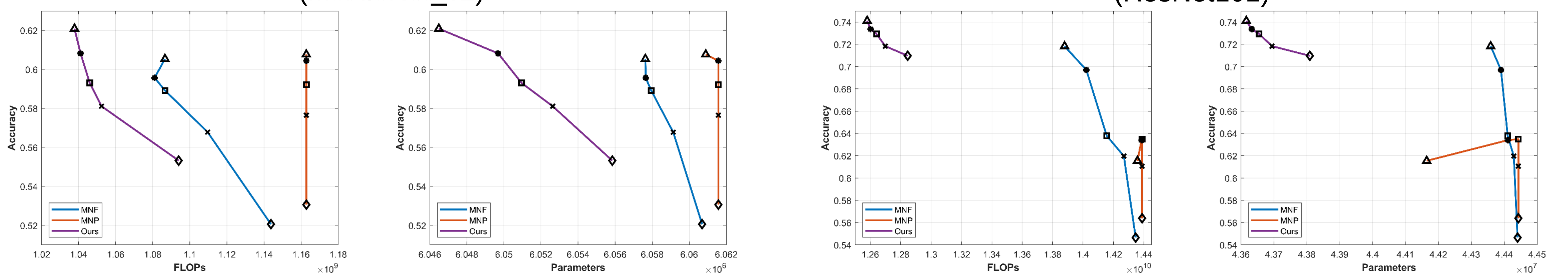
CIFAR 100

(MobileNet_v2) (ResNet101)



IMAGENET

(MobileNet_v2) (ResNet101)



We show in the plots how our strategy (Purple line) outperforms MorphNet in terms of compression both in terms of FLOPs and model size, while having higher accuracy.

Background

We build on **MorphNet** [2] whose training procedure optimizes CNN's structure. Its compression strategy relies on a regularizer, which induces sparsity in activations by pruning neurons with greater cost C. Network sparsity is measured by the batch normalization scaling factor γ associated to each neuron.

The **cost C** can be either associated to neurons contributing to either **FLOPs** or **size** (number of parameters).

$$\mathcal{C}_{FLOP} = \sum_{k=1}^{K}[C^k_{in} * (w^k)^2 * C^k_{out} * S^k_{out}]$$

$$\mathcal{C}_{PARAM} = \sum_{k=1}^{K}[C^k_{in} * (w^k)^2 * C^k_{out}]$$

Conclusions

- We present a resource-aware network structure learning method, where lower layers are optimized for FLOPs and higher layers for model-parameters.
- Our method leverages privileged information to preserve high-quality model performance.
- Our method brings state of the art network compression while maintaining better control over the compression-performance tradeoff.

References

[1] *Lopez-Paz, David, et al. "Unifying distillation and privileged information."ICLR 2016*

[2] *A. Gordon, et al, "Morphnet: Fast & simple resource-constrained structure learning of deep networks." CVPR 2018*