Compact CNN Structure Learning by Knowledge Distillation

Waqar Ahmed, Andrea Zunino, Pietro Morerio, Vittorio Murino
waqar.ahmed@iit.it, andrea.zunino@huawei.com, pietro.morerio@iit.it, vittorio.murino@iit.it
Motivation

CNNs are ubiquitous in computer vision. However, they require considerable resources in terms of

- **Computation**
- **Memory**

Compression techniques can partially handle these issues at the price of a drop in performance.
Goal & Contribution

In order to overcome the shortcomings of existing methods (namely, a drop in performance after model compression) 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 teach-student framework.
Overview of the method

Teacher

Student

Loss

Teacher predictions

Ground-truth label

Resource-aware optimization

FLOPs Optimizer

Model-Parameters Optimizer
Background

We build on MorphNet [1] 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 $\gamma$ associated to each neuron.

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

\[
C_{FLOP} = \sum_{k=1}^{K} \left[ C_{in}^k \ast (w^k)^2 \ast C_{out}^k \ast S_{out}^k \right]
\]

\[
C_{PARAM} = \sum_{k=1}^{K} \left[ C_{in}^k \ast (w^k)^2 \ast C_{out}^k \right]
\]

Leveraging privileged information

While being compressed, the network tries to mimic the predictions of the uncompressed network.

\[
\begin{align*}
\min_{\theta_1} \min_{\theta_2} \frac{1}{N} \sum_{i=1}^{N} & \left[ (1 - \lambda) l(y^i, \sigma(f(x^i, \theta_1, \theta_2)/T)) \\
+ & \lambda l(z^i, \sigma(f_t(x^i, \theta_1, \theta_2)/T)) \\
+ & \alpha (C_{FLOP}(\theta_1) + C_{PARAM}(\theta_2)) \right]
\end{align*}
\]

\[z^i = \sigma(f_t(x^i)/T)\]
Resource-aware optimization

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

\[
C_{FLOP} = \sum_{k=1}^{K} [C_{in}^{k} \times (w^{k})^2 \times C_{out}^{k} \times S_{out}^{k}]
\]

\[
C_{PARAM} = \sum_{k=1}^{K} [C_{in}^{k} \times (w^{k})^2 \times C_{out}^{k}]
\]

\[\theta_1 \cup \theta_2 = \theta, \ \theta_1 \cap \theta_2 = \emptyset\] is a partition of the weights

Lower layers carry higher FLOPs, while higher layers account more for model-size.

Therefore we propose a configuration in which the lower half of the network is optimized for FLOPs and the upper half is optimized for size.
Results on Cifar-10
Results on Cifar-100
Conclusion

In this paper, we present a resource-aware network structure learning method, which enables suitable optimization in different sections of the seed network considering FLOPs and model-parameters constraints - i.e. lower layers are optimized for FLOPs and higher layers for model-parameters.

Furthermore, our method leverages privileged information to impose control over predictions to preserve high-quality model performance.

Our method brings state of the art network compression that outperforms the existing method by a large margin while maintaining better control over the compression-performance tradeoff.