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). However, complexity reduction usually comes at the price of a drop in the model performance.

Motivation

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].

Contributions

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 $\gamma$ associated to each neuron.

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

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

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 $\gamma$ associated to each neuron.

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

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