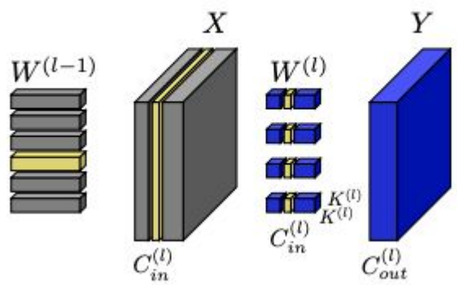
Group 73: Soft Masking for Cost-Constrained Channel Pruning

Introduction

- Structured channel pruning to significantly accelerate inference time for (CNNs) on modern hardware
- Allow pruned channels to adaptively return to the network while simultaneously pruning towards a target cost constraint
- Channel pruning task formulated as a global resource allocation problem, where the goal is to minimize the drop in accuracy while keeping the cost constraint (i.e. inference time) below a certain value.
- New batch normalization scaling approach to lessen the effect of large gradient magnitudes caused by the pruning of many channels.

Implementation

Modern hardware has poor support for sparsity, resulting in no speedup so channel pruning is a better alternative.



Removing a weight in W(l) removes a the corresponding feature in X and corresponding output channel in W(l-1)

Weights are pruned by introducing a mask for each layer to reparameterize the weights.

$$\widetilde{W}^{(l)} = W^{(l)} \odot m^{(l)}.$$

Using a straight-through estimator, paths not in the forward pass are still updated,

$$g_{W^{(l)}} = g_{\widetilde{W}^{(l)}},$$

Importance determines which weights will be pruned. It is a proxy for the weight's effect on the network loss

$$\mathcal{I}_i^{(l)} = \left| \sum_{o,r,s} W_{o,i,r,s}^{(l)} g_{W_{o,i,r,s}^{(l)}} \right|$$

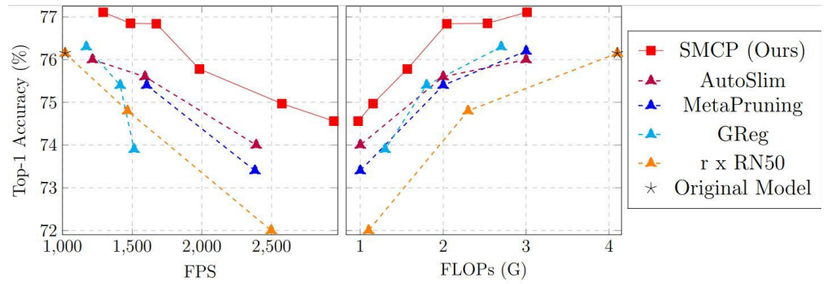
High pruning ratios result in large gradients, so batch normalization is used to reduce the weights by the fraction of channels left unpruned

$$\gamma^{(l)} \leftarrow \gamma_{orig}^{(l)} \frac{\sum_i m_i^{(l)}}{C_{in}^{(l)}}.$$

Selecting which channels to prune is done by solving a multiple choice knapsack problem

$$\begin{aligned} \max_x \quad & \sum_{l=1}^{L} \sum_{i=1}^{n_l} v_{l,i} x_{l,i} \\ \text{s.t.} \quad & \sum_{l=1}^{L} \sum_{i=1}^{n_l} c_{l,i} x_{l,i} \leq C \\ & x_{l,i} \in \{0,1\}, \quad \sum_{i=1}^{n_l} x_{l,i} = 1 \end{aligned}$$

Results



The results show that a 1.5x speedup can be achieved with almost no loss of accuracy.

The results show that almost 2.5x speedup can be achieved with only a 2% drop in accuracy.

Tested the SMCP method, on CIFAR-10 using a smaller network architecture (ResNet-18) and obtained similar top-1 accuracy results, for much higher FPS values.

Challenges

- Working with the (insufficient) requirements that are 2 years old
- Getting access to GPUs on Google Cloud Platform
- Understanding the implementation of the pruning