Technical Report of the Wake Vision Challenge (Model-centric Track)

Mohammad Hallaq
Department of Electrical and Software Engineering, University of Calgary, Canada

I. The Approach

The model used in this challenge is a structurally pruned version of MobileNetV2, designed to minimize the number of Multiply-Accumulate Operations (MACs) and parameters. The pruning approach Fig. 1 follows the methodology introduced in one of our recently accepted papers at the IEEE International Conference on Communications (IEEE ICC)[1].

A key challenge we encountered was that our approach was originally designed for pruning PyTorch models, whereas this challenge required a TensorFlow implementation. To address this, I first applied our pruning algorithm to MobileNetV2 in PyTorch, obtained the pruned model, and then manually reconstructed its pruned counterpart in TensorFlow to ensure compatibility with the rest of the pipeline.

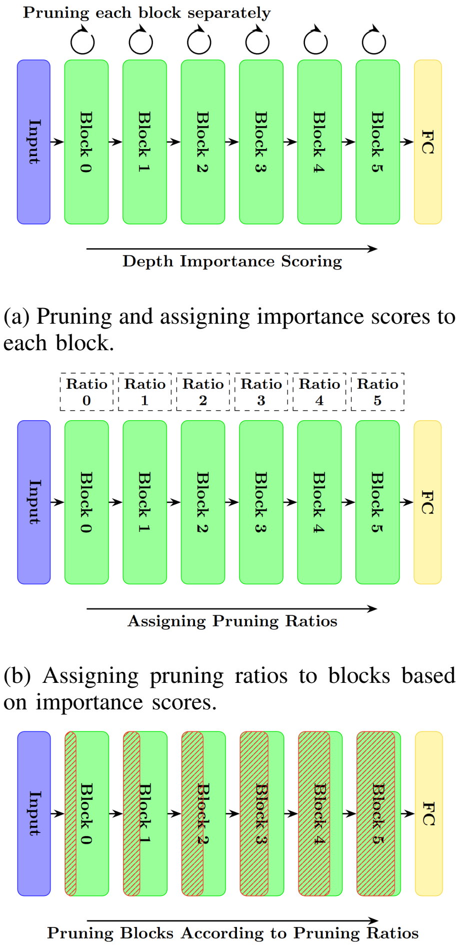
According to the algorithm, each block of layers in the model is pruned to the maximum extent, and the resulting reduction in MACs and parameters is measured. The model's blocks consist of inverted residual blocks containing depthwise and pointwise convolutions. To reach the maximum extent, I retain only a single channel per layer across all layers within each block.

This process provides an estimated importance score for each block, which is then used to determine a unique pruning ratio for each block. Finally, the model is pruned based on these individual pruning ratios, resulting in a non-uniform structured pruning strategy.

I typically apply this method to a pretrained model, followed by fine-tuning to achieve optimal performance. However, since the trained models were in TensorFlow format and given the tight deadline, I prioritized efficiency over accuracy and opted to train the pruned model from scratch.

I selected MobileNetV2_0.25 for this task because it already performs well, and my goal was to make it even more compact. Notably, MobileNetV2_0.25 applies a uniform 25% channel reduction across all layers. However, I believe that some layers are more critical than others and should retain more than 25% of their channels. Therefore, I opted for non-uniform structured pruning to better preserve important layers while also significantly pruning the model.

To further reduce MACs, I also downsampled the input size from the standard (224, 224, 3) to (80, 80, 3).



(a) Pruning and assigning importance scores to each block.

(b) Assigning pruning ratios to blocks based on importance scores.

(c) Applying block-specific pruning to the model.

Fig. 1: Overview of the block pruning methodology.

[1] The paper has been accepted but is not yet published.