PRUNING CONVOLUTIONAL NEURAL NETWORKS  
FOR RESOURCE EFFICIENT INFERENCE

**Authors:** **Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, Jan Kautz**

**Link to the paper:** [**https://arxiv.org/pdf/1611.06440**](https://arxiv.org/pdf/1611.06440)

A Novel method is presented in the paper for pruning filters from convolution neural networks with the objective of reducing computation for inference. New criterion based on Taylor expansion of the neural network function is proposed for pruning.

The major conclusion in the paper was that the

1. CNNs can be pruned by iteratively removing the least salient parameters.
2. The proposed Taylor expansion-based criterion demonstrates better performance than other criteria discussed in the paper.
3. For global scaling, layer-wise normalization of the criteria is essential.

This is a well written and a good paper on network pruning. Proposed method has indicated slight improvement over OBD (Optimal Brain Damage) with respect to pruned parameter v/s accuracy. But has substantial gain in terms of computation. The evaluation is good with different baseline methods.