Motivation and Research Question

Deep neural networks face two key challenges: 1) Lack of robustness against adversarial examples and 2) Large size of neural networks. We often utilize robust training to solve the former and use network pruning techniques to solve the latter. However, both robust training and pruning methods are largely studied independently in earlier works.

Motivation: We observe that naively integrating pruning methods, such as pruning connections with the smallest magnitude weight, with robust training achieves poor performance.

Research Question: How to reduce the size of neural networks while preserving both accuracy and robustness?

Our Contributions

1. We propose HYDRA, a novel pruning approach for robustly trained networks, including provably robust networks.
2. We achieve state-of-the-art accuracy and robustness, simultaneously for compressed networks over CIFAR-10, SVHN, and ImageNet dataset.
3. We demonstrate that provably robust subnetworks are hidden inside even non-robust networks.

Our Approach

Guiding Pruning with Robust Training Loss:
We integrate robust training objective in the pruning process. We solve the resulted empirical risk minimization problem efficiently using SGD.

\[ \hat{m} = \arg\min_{m \in \{0, 1\}^n} \mathbb{E}_{(x,y) \sim D} \left[ \mathcal{L}_{\text{pruning}}(\theta_{\text{pretrain}} \odot m, x, y) \right] \quad \text{s.t.} \quad \|m\|_0 \leq k \]

Importance Scores: Since finding Boolean pruning mask is a discrete optimization, we convert it to a continuous optimization problem by assigning real-valued importance score to each connection.

Scaled Initialization: Crucial to the success of HYDRA is our proposed scaled initialization of importance scores based on pretrained network weights.

Hidden Robust Sub-Networks

Given a non-robust pre-trained network, we optimize to maximizing the robust accuracy only by pruning connections, i.e., no update in the weight of any connections.

Experimental Results

Comparison with least weight magnitude pruning [1] (first baseline)

Comparison with Adv-ADMM [2] (second baseline)

Our work

Project webpage: https://vsehwag.github.io/hydra

- Motivation and Research Question
- Our Contributions
- Our Approach
- Experimental Results