Pruning has a disparate impact on model accuracy

Cuong Tran [1] Ferdinando Fioretto [1] Jung-Eun Kim [2] Rakshit Naidu [3]

[1]Syracuse University [2]North Carolina State University [3]Carnegie Mellon University

Motivation

1. Pruning is a widely used network compression technique in resource-constrained platforms.
2. It works by removing least important parameters/channels/filters in a network.
3. A majority of prior work only focus on accuracy, but..
4. We showed empirically and theoretically that model pruning can exacerbate the unfairness among demographic groups.

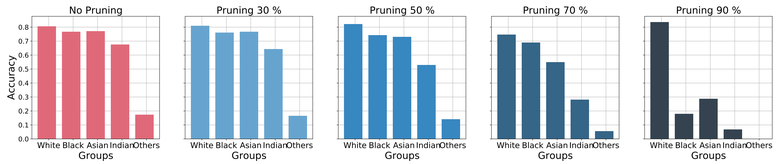

Figure 1. Accuracy of each demographic group in the UTK-Face dataset using Resnet18 [18], at the increasing of the pruning rate.

Settings

- Standard model training:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, J(\boldsymbol{\theta}; D) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i), \tag{1}$$

- Usually $\boldsymbol{\theta}^*$ is over-parameterized which can costs memory and storage. Pruning is applied by removing p% least important parameters in $\boldsymbol{\theta}^*$ to obtain $\bar{\boldsymbol{\theta}}$.
- We define the excessive risk as the difference of loss before and after pruning:

$$R(a) = J(\bar{\boldsymbol{\theta}}; D_a) - J(\hat{\boldsymbol{\theta}}; D_a), \tag{2}$$

- Fairness is defined as the maximum difference of group excessive risks:

$$\xi(D) = \max_{a,a' \in \mathcal{A}} |R(a) - R(a')|, \tag{3}$$

Fairness analysis in pruning: Roadmap

Theorem 1: The *excessive loss* of a group $a \in \mathcal{A}$ is upper bounded by:

$$R(a) \leq \left\|\boldsymbol{g}_a^\ell\right\| \times \left\|\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\right\| + \frac{1}{2}\lambda\left(\boldsymbol{H}_a^\ell\right) \times \left\|\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\right\|^2 + O\left(\left\|\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\right\|^3\right), \tag{4}$$

where $\boldsymbol{g}_a^\ell = \nabla_{\boldsymbol{\theta}} J(\hat{\boldsymbol{\theta}}; D_a)$ is the vector of gradients associated with the loss function ℓ evaluated at $\hat{\boldsymbol{\theta}}$ and computed using group data D_a, $\boldsymbol{H}_a^\ell = \nabla^2_{\boldsymbol{\theta}} J(\hat{\boldsymbol{\theta}}; D_a)$ is the Hessian matrix of the loss function ℓ, at the optimal parameters vector $\hat{\boldsymbol{\theta}}$, computed using the group data D_a (henceforth simply referred to as *group hessian*), and $\lambda(\Sigma)$ is the maximum eigenvalue of a matrix Σ.

Why disparity in groups' gradients causes unfairness?

Proposition 1: Consider two groups a and b in $\mathcal{A}$ with $|D_a| \geq |D_b|$. Then $\left\|\boldsymbol{g}_a^\ell\right\| \leq \left\|\boldsymbol{g}_b^\ell\right\|$.

Proposition 2: For a given group $a \in \mathcal{A}$, gradient norms can be upper bounded as:

$$\|\boldsymbol{g}_a^\ell\| \in O\left(\sum_{(\boldsymbol{x},y)\in D_a} \underbrace{\|f(\boldsymbol{x}) - y\|}_{\text{Error}} \times \left\|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x})\right\|\right).$$

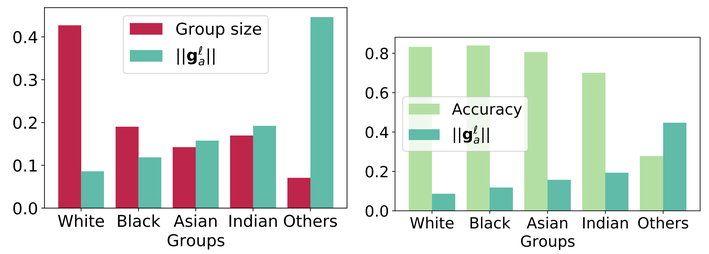

Figure 2. Impact of group sizes and accuracy towards group gradient norms

Why disparity in groups' Hessians causes unfairness?

Theorem 2: Let f_θ be a binary classifier trained using a binary cross entropy loss. For any group $a \in \mathcal{A}$, the maximum eigenvalue of the group Hessian $\lambda(\boldsymbol{H}_a^\ell)$ can be upper bounded by:

$$\lambda(\boldsymbol{H}_a^\ell) \leq \frac{1}{|D_a|}\sum_{(\boldsymbol{x},y)\in D_a} \underbrace{(f(\boldsymbol{x}))\,(1 - f(\boldsymbol{x}))}_{\text{Closeness to decision boundary}} \times \left\|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x})\right\|^2 + \underbrace{|f(\boldsymbol{x}) - y|}_{\text{Error}} \times \lambda\left(\nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{x})\right). \tag{5}$$

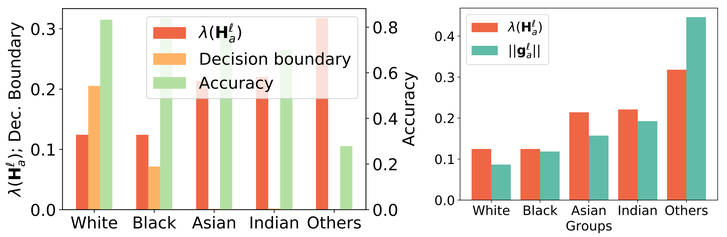

Figure 3. Left: Group Hessian vs decision boundary and accuracy. Right: Group Hessian vs gradient norm.

Mitigation solutions

- The disparity in gradient and Hessian norms are two direct factors for unfairness. The first straightforward mitigation solution is to equalize the gradient and Hessian norms:

$$\underset{\boldsymbol{\theta}}{\text{minimize}}\;\; J(\boldsymbol{\theta}; D) \quad \text{such that:}\;\; \|\boldsymbol{g}_a^\ell\| = \|\boldsymbol{g}^\ell\|,\;\; \lambda(\boldsymbol{H}_a^\ell) = \lambda(\boldsymbol{H}^\ell)\;\; \forall a \in \mathcal{A}, \tag{6}$$

- However this solution can be computationally expensive.
- Hence, we propose to equalize the group losses due to their correlation with gradient and Hessian norms.

$$\underset{\boldsymbol{\theta}}{\text{minimize}}\;\; J(\boldsymbol{\theta}; D) \quad \text{such that:}\;\; J(\boldsymbol{\theta}; D_a) = J(\boldsymbol{\theta}; D)\;\; \forall a \in \mathcal{A}, \tag{7}$$

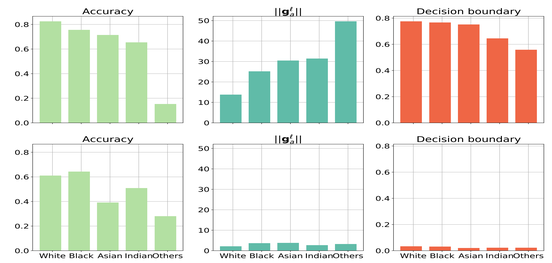

Figure 4. Effects of fairness constraints in balancing not only group accuracy (left) but also gradient norms (middle) and group average distance to the decision boundary (right)

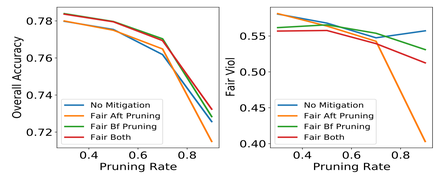

Figure 5. Accuracy and Fairness violations attained by all models on ResNet50, UTK-Face dataset with ethnicity (5 classes) as group attribute (and labels)

Discussion

- Pruning affects groups that are closer to the decision boundary. It also noted that these groups are more sensitive under adversarial attacks. Thus, understanding the interconnection among pruning, adversarial robustness and fairness is worth of investigation.
- Relax twice differentiable assumption over loss functions in fairness analysis can be a potential future work.

References

[1] Cuong Tran, Ferdinando Fioretto, Jung-Eun Kim, and Rakshit Naidu. Pruning has a disparate impact on model accuracy. *arXiv preprint arXiv:2205.13574*, 2022.