FAIRNESS OF COMPRESSION TECHNIQUES

1 Introduction

The primary objective of this project course is to explore and analyze the co-relation of fairness in machine learning models with respect to various model compression techniques, and consequently propose optimal bias reduction strategies.

A model can be categorized as fair if the given model is able to make predictions or decisions without discriminating against certain individuals or groups based on certain **sensitive attributes** such as race, gender or other protected characteristics. There are techniques to train a model while valuing fairness, such as, Adversarial Debiasing, Weighted Loss Function and Data Augmentation.

Model compression is the practice of reducing the size of a given model, without having a high trade off on the accuracy of the original model, for edge device deployment and faster inference. During model compression, the bias exhibited by the model can exhibit varying behavior depending on factors such as the purity of the dataset, the extent of compression, and the fairness treatment applied pre- or post-compression. To achieve bias mitigation, we analytically understand and quantify the impact of these factors on the model's bias. This enables the development of targeted strategies to limit or mitigate the bias introduced during compression, thereby ensuring the fairness and accuracy of the compressed model.

2 Preliminaries

2.1 Fairness metrics

In order to evaluate the discrimination, there are various fairness metrics defined, such as **EO (Equalized odds).**

$$P(\hat{Y} = 1 \mid A = a, Y = y) = P(\hat{Y} = 1 \mid A = b, Y = y) \quad \forall a, b \in \mathcal{A}, y \in \{0, 1\}$$

2.2 Compression techniques

There are multiple model compression techniques which can be broadly classified as **Quantization, Pruning(Local and global), Knowledge distillation, low rank approximation** and more. Each of these techniques have their application based on the model architecture, compression ratio requisites, computational resources and more.

3 Existing Approaches and Final Analysis

3.1 Literature review

In the course of familiarizing myself with the relevant research in model compression, I reviewed several papers that apply various compression techniques across different data types, including tabular, image, and language data.

For the purposes of this literature review, I will concentrate specifically on research related to image data. One study indicates that compression and quantization can lead to a substantial reduction in model size with only a minimal effect on overall accuracy for both CK+DB and RAF-DB [1]. However, it also points out that, in the case of RAF-DB, the various compression strategies do not appear to exacerbate the predictive performance gaps across sensitive attributes such as gender, race, and age. This contrasts with the findings for CK+DB, where compression tends to intensify existing gender biases.

Another key paper [2] explores fairness-preserving compression through adversarial debiasing. The authors not only evaluate metrics pre- and post-compression but also apply fairness techniques prior to compression and subsequently assess their effectiveness after compression. This work integrates fair training objectives into neural network pruning by formulating pruning and debiasing as a two-player adversarial game, enhancing fairness while preserving accuracy and efficiency. The proposed method consistently performs well across different datasets, network architectures, and pruning ratios, and also reveals the existence of fair sub-networks within randomly initialized networks.

3.2 Prospective work

This body of work has motivated me to not only evaluate the effects of compression on model fairness but to propose strategies aimed at achieving **fairness-aware compression**. This could involve integrating fairness interventions either during compression, or as a pre- or post-processing step, ensuring that fairness is maintained throughout the compression pipeline. Lastly, I would like to acknowledge the constant support and guidance of my project supervisor Prof. Manisha Padala.

References

[1] Samuil Stoychev and Hatice Gunes. The effect of model compression on fairness in facial expression recognition. In *arXiv:2201.01709v1*, 2022.

[2] Zhibo Wang et al. Lei Zhang. Towards fairness-aware adversarial network pruning. In *International Conference on Computer Vision*, 2023.