FAIRNESS OF MODEL QUANTIZATION

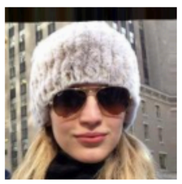
Aryan Ajay Solanki | Manisha Padala

INTRODUCTION

Deep neural networks excel in various tasks but are computationally intensive, driving the adoption of model compression techniques to reduce resource demands. However, compression may amplify biases in machine learning models, impacting fairness. [1]

We investigate these effects in facial expression recognition using the CelebA dataset. It is a facial attributes dataset containing over 200,000 celebrity images, ach annotated with 40 attribute labels

CelebA Dataset



Label: Attractive
Protected: Female

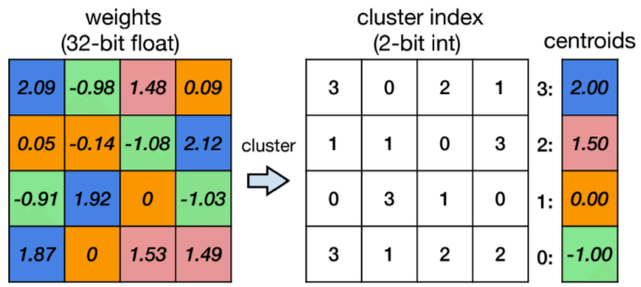
Label: Attractive
Protected: Female

Label: Not Attractive
Protected: Male

Quantization

```
ResNet-18:  fp32  Size (KB): 44781.316
ResNet-18:  int8  Size (KB): 11392.19
Size reduction of 3.93 times.
```

We perform a technique known as Post Training Quantization. implemented using PyTorch's `torchvision.models.quantization.resnet.QuantizableResNet` base class.



Weight Based Quantization (K-means) [5]

IMPLEMENTATION

Common metrics:

True Positive Rate

$$\text{TPR} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

False Positive Rate

$$\text{FPR} = \frac{\text{False Positives (FP)}}{\text{False Positives (FP)} + \text{True Negatives (TN)}}$$

Fairness metrics:

Equalized Odds (EO) or Demographic Equalized Odds

$$\text{EO} = |\text{TPR}_{\text{class1}} - \text{TPR}_{\text{class2}}|$$

False Positive Rate (FPR) Difference

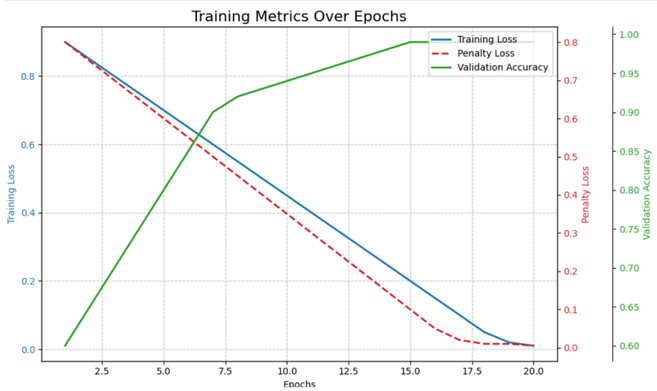
$$\text{FPR Difference} = |\text{FPR}_{\text{class1}} - \text{FPR}_{\text{class2}}|$$

For the CelebA Dataset, we have chosen "Attractive" as the predicted feature and "Gender" as the protected feature. For the scope of fairness metrics, class1 = Male and class2 = Female.

1. Caliberated Equalized Odds Post Training:

Calibrated Equalized Odds Post-Processing adjusts a classifier's predicted probabilities to meet equalized odds fairness. It ensures balanced true positive rates (TPR) and false positive rates (FPR) across protected groups. [4]

2. FairALM (Augmented Lagrangian Method):



The FairALM algorithm addresses bias in algorithmic decision-making by integrating fairness constraints directly into the training process through optimization principles. [3]

RESULTS

Model	Overall Accuracy (%)	EO	FPR Difference
Fine-tuned (Normal)	80.06	0.21	0.20
Quantized	80.00	0.21	0.17
FairALM Pre-quantization	79.62	0.18	0.25
FairALM Quantized	79.80	0.19	0.20
CalibratedEqOd Pre-quantization	80.06	0.21	0.20
CalibratedEqOd Quantized	79.62	0.21	0.18

Table 1. Comparison of Overall Accuracy, EO, and FPR Difference across various models.

Model	Male Accuracy (%)	Female Accuracy (%)
Fine-tuned (Normal)	81.61	79.30
Quantized	82.00	78.20
FairALM Pre-quantization	79.83	79.20
FairALM Quantized	81.23	79.90
CalibratedEqOd Pre-quantization	81.61	79.08
CalibratedEqOd Quantized	82.15	78.02

Table 2. Male and Female Accuracy across various models.

Model	Male TPR	Female TPR	Male FPR	Female FPR
Fine-tuned (Normal)	0.65	0.86	0.13	0.33
Quantized	0.59	0.80	0.10	0.27
FairALM Pre-quantization	0.74	0.92	0.18	0.43
FairALM Quantized	0.66	0.86	0.13	0.34
CalibratedEqOd Pre-quantization	0.65	0.86	0.13	0.33
CalibratedEqOd Quantized	0.61	0.81	0.11	0.28

Table 3. Comparison of Male and Female TPR and FPR for various models.

FUTURE WORK

This body of work has motivated me to propose strategies aimed at achieving fairness-aware compression.

I look forward to developing fairness constraints integrated into the optimization process that could help maintain fairness without sacrificing compression benefits.

REFERENCES

1. Stoychev, Samuil , and Hatice Gunes. The Effect of Model Compression on Fairness in Facial Expression Recognition. 22 Jan. 2022, arxiv.org/pdf/2201.01709.
2. Zhang et al. Towards Fairness-aware Adversarial Network Pruning. 27 March. 2024, https://openaccess.thecvf.com/content/ICCV2023/papers/Zhang_Towards_Fairness-aware_Adversarial_Network_Pruning_ICCV_2023_paper.pdf.
3. Lokhande et al. FairALM: Augmented Lagrangian Method for Training Fair Models with Little Regret. 24 June 2024, https://arxiv.org/abs/2004.01355.
4. IBM Fairness 360 - https://aif360.res.ibm.com/
5. MIT Tiny ML Lectures

CS 299 Project Course

ACKNOWLEDGEMENT

Special thanks to Manisha Padala, my project supervisor and my parents for their constant help and support throughout the duration of this project.

KEYWORDS

Fairness, Model Compression, Quantization, Demographic equal opportunity (DEO), False Positive Rate (FPR), Pruning