Explaining How Quantization Disparately Skews a Model

Anonymous authors
Paper under double-blind review

Abstract

Post Training Quantization (PTQ) is widely adopted due to its high compression capacity and speed with minimal impact on accuracy. However, we observed that disparate impacts are exacerbated by quantization, especially for minority groups. Our analysis explains that in the course of quantization there is a chain of factors attributed to a disparate impact across groups during forward and backward passes. We explore how the changes in weights and activations induced by quantization cause cascaded impacts in the network, resulting in logits with lower variance, increased loss, and compromised group accuracies. We extend our study to verify the influence of these impacts on group gradient norms and eigenvalues of the Hessian matrix, providing insights into the state of the network from an optimization point of view. To mitigate these effects, we propose integrating mixed precision Quantization Aware Training (QAT) with dataset sampling methods and weighted loss functions, therefore providing fair deployment of quantized neural networks.

1 Introduction

With the onset of edge devices running deep neural networks for various tasks ranging across several domains, the demand for faster computation and model lightness has become more pronounced. To aid this, compression methods such as pruning Han et al. (2015a) and quantization Hubara et al. (2016) have taken the lead, producing little to no loss of accuracy with considerable memory and speed gains. Nevertheless, these methods do not account for the possible disparate impact they cause, and have been shown to have adverse effects on minority groups and exacerbate the shortcomings of their dense, counterpart model Hooker et al. (2019).

In a streamline of model compression, Tran et al. (2022) recognized that magnitude pruning can exacerbate unfairness among classes. While pruning and quantization share a common objective of compressing a model, they are different in their approaches. Pruning involves removing weights or components that are insignificant according to a defined criterion. Whereas, quantization focuses on reducing the precision of the bits used to represent the weights and activations of the neural network. Notably, Kuzmin et al. (2023) demonstrated that quantization outperforms pruning-based strategies when similar model sizes and resource footprints are considered. Furthermore, quantization is prominent for Large Language Models (LLMs) due to their large parameter sizes and requirement for reduced energy consumption Kim et al. (2023); Frantar et al. (2022); Dettmers & Zettlemoyer (2023).

We observed that quantization can exacerbate disparity of a model, especially for the minority group, as we show in Fig. 1. The leftmost chart is pre-quantization. As the precision is reduced, the disparity is exacerbated further. When the model is quantized to `int2`, the disparity is extreme. In this paper, we identify the factors that impact the disparity and optimization state in the forward and backward passes, respectively.

Post Training Quantization (PTQ) modifies the weights of the network while setting several weights to absolute zeros, thereby, inducing sparsity, which together brings in disparate impacts of a model. Consequently, the logits suffer from a reduction in variance, similar to using high temperature scaling, while undergoing magnitude changes that lead to misclassifications. These factors finally alter the softmax probabilities and skew their distributions closer to the decision boundary towards low confidence regions, causing higher loss

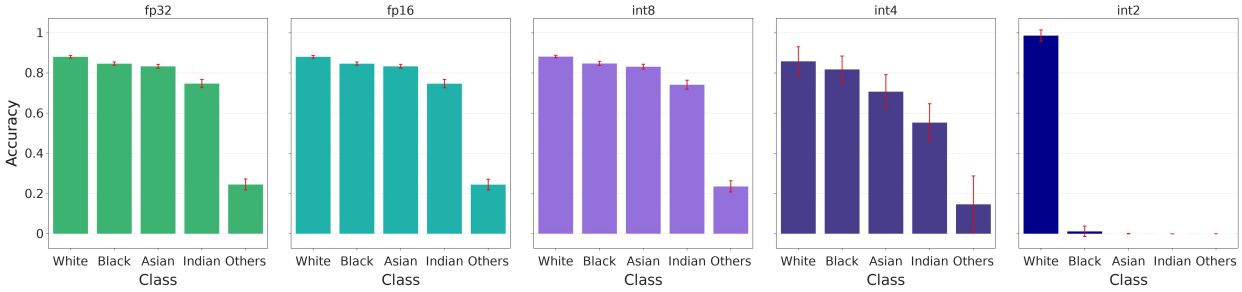


Figure 1: Accuracy for groups according to different quantization precisions on UTKFace dataset.

and group disparity. Additionally, PTQ shifts the model to a worse position in the optimization space, with larger gradient norms and eigenvalues of the Hessian matrix for minority classes, implying a potential for further optimization.

To combat these problems, we leverage the simplicity of dataset sampling methods to overcome the dataset imbalance, combine it with a weighted cross-entropy loss function to deal with example difficulty, and use mixed-precision Quantization Aware Training (QAT) to withstand the degradation of model performance due to low precision representations.

Our contributions are summarized as:

1. We observed and showed that quantization can exacerbate disparity of a model, especially for minority groups.
2. We identified the factors of PTQ that cause disparity: change in weights, increased sparsity, changes in logits, and reduced variance of softmax probabilities. These are cascaded factors in the forward pass.
3. We examine the degradation caused to the model's state in the optimization space, using gradient norms and eigenvalues of the Hessian matrix of a quantized model.
4. We proposed a mitigating quantization approach that incorporates sampling methods and weighted loss functions for improved fairness.

2 Related Work

2.1 Quantization

This competent method contributes to a significant body of work in model compression with the introduction of compatible hardware, giving rise to diverse methodologies such as PTQ and QAT. PTQ is a domain where quantization is performed after the model is fully trained on a dataset without further retraining Banner et al. (2019); Zhao et al. (2019); Choukroun et al. (2019); Wang et al. (2020); Li et al. (2021); Lee et al. (2022). Whereas, QAT learns quantized weights in the training phase or during retraining Chen et al. (2023); Huang et al. (2023); Bhalgat et al. (2020); Esser et al. (2020); Nagel et al. (2022). Frantar & Alistarh (2022) attempt to compress a network using both pruning and quantization. These methods, however, only focus on reducing the bit precision while maintaining the accuracy of the original model with little to no fairness control.

2.2 Mixed-Precision Quantization

Mixed-precision quantization is widely used, because low-precision PTQ tends to have pitfalls in accuracy. Wu et al. (2018) uses neural architecture search to find suitable precisions for different layers. A mixed-precision integer-only inference for faster computation was explored by Yao et al. (2021). Dong et al. (2019) uses Hessian spectrums of the network layers to determine the precision of the layers.

2.3 Algorithmic Bias and Fairness

There are pressing concerns considering the surge of neural network-based models for everyday use. In the context of neural networks, adversarial learning methods Wadsworth et al. (2018); Xu et al. (2021) are often used to achieve fairness. Du et al. (2021) de-bias the classification head to improve the fairness of networks. However, most fair learning algorithms suffer from tradeoffs, for e.g., Zhao & Gordon (2022) prove that one such tradeoff exists between statistical parity and model accuracy when learning fair representations. Several studies measure algorithmic bias using standard datasets Amini et al. (2019) or synthetic datasets Liang et al. (2023). These studies consider a fully trained neural network that is not compressed. However, in this work, we dive into the fairness of compressed neural networks via quantization through the lens of the changes and impacts occurring in the model due to quantization. In Tran et al. (2022), the authors recognized that pruning may have a disparate impact on model accuracy and attribute it to changes in the gradient norms and eigenvalues of the Hessian matrix. In the context of quantization with imbalanced class distribution, Chen et al. (2022) proposed *HomoVar* loss to balance classes during quantization. Using skip-connections and Dirichlet distribution, Zhou et al. (2023) create a framework for mixed-precision quantization to dampen disparity.

However, these prior works do not explain the causes of the disparate impact of quantization, especially from the perspective of the impact factors inside the network.

3 Problem Formulation

Consider a classification task involving a dataset D with M input samples $X = \{x_1, x_2, \cdots, x_i, \cdots, x_M\}$ and corresponding classes $Y = \{y_1, y_2, \cdots, y_i, \cdots, y_M\}$ where $y_i \in G$ groups (classes). The objective is to learn a classifier f_θ with parameters $\theta \in \mathbb{R}^K$, where K is the number of parameters in the network. The risk function obtained by using cross-entropy as the loss function to measure the discrepancy between the predicted and actual labels under empirical risk minimization (ERM) Vapnik (1991) is:

$$L(\theta; D) = -\frac{1}{M}\sum_{i=1}^{M}\sum_{g=1}^{G} y_{ig} \cdot \log(p_\theta(x_i))_g \tag{1}$$

where $p_\theta(x_i) = \sigma(f_\theta(x_i))$ and $\sigma(z_g) = \frac{e^{z_g}}{\sum_j e^{z_j}}$. The best solution to this optimization problem is given by, $\theta_o = \underset{\theta}{\operatorname{argmin}}\ L(\theta; D)$. Note that this definition pertains to an uncompressed model. Subsequently, let θ_q be the weights of a quantized network such that $\theta_q = T(\theta_o)$, where T is a quantization function and q is the number of bits used to represent the weights of the network. For example, if the network was quantized to use 8-bit representations, the network parameters are denoted by θ_8. Let $\widetilde{\theta_q}$ denote the dequantized weights obtained by scaling θ_q to floating point numbers, $\widetilde{\theta_q} = S.\theta_q$, where S is the set of scaling factors. As a result, the risk functions for the original and compressed models are given by $L(\theta_o; D_g)$ and $L(\widetilde{\theta_q}; D_g)$, respectively.

3.1 Fairness Analysis

Visualizing fairness via changes to a loss function is challenging due to its multidimensional nature. Consequently, we rely on its correlation with model accuracy and observe its changes across groups. Among the differences, the largest discrepancy can represent how unfair the model actually is. In light of this, we propose Fairness Violation Observed (`FVO`):

$$\texttt{FVO}(\theta; D) = \max_{g,g' \in G} |Acc(\theta, D_g) - Acc(\theta, D_{g'})| \tag{2}$$

where $Acc(\theta, D_g)$ and $Acc(\theta, D_{g'})$ represent the accuracy of groups g and g' for parameters $\theta \in \{\theta_o, \widetilde{\theta_q}\}$.

Why `FVO`? `FVO` is interpretable and generalizes well to multi-class tasks. Further, minimizing `FVO` inherently captures equalized odds Hardt et al. (2016) when accuracy differences arise from varying true positive and false positive rates across groups. Other advantages include:

1. FVO is not limited to specific pair of groups and applies to the overall set of groups involved.
2. It directly relates to accuracy, which is an interpretable metric for computer vision tasks.
3. It is practical for situations seeking to balance model performance with fairness

3.2 What is the Best Model?

A model with low `FVO` indicates that the performances across all groups are similar. However, it does not guarantee that the model performs well overall. That is because, in particular, when all the groups have low accuracies, the model may end up with worse overall performance. In order to choose a model that is both fair and accurate, we consider both `FVO` and the overall accuracy, `OA`, together. Thus, when comparing different approaches for fairness, our preference aligns with the model that maximizes `OA` and minimizes `FVO`:

$$\max_{\texttt{OA}} \min_{\texttt{FVO}} f_{\theta,D} \tag{3}$$

Setup For the investigations presented in this paper, we use per tensor uniform post-training quantization (PTQ) Nagel et al. (2021) for weights, based on the implementation in Banner et al. (2019) for integer quantization. In particular, for `fp16` experiments, we used half-precision computation from the PyTorch library. Note that the integer weights are scaled to floating points during inference. The following experiments are on UTKFace dataset Zhang Zhifei & Hairong (2017) with the task of classifying the ethnicity using a ResNet18 architecture, where the weights are quantized to 16, 8, 4, and 2 bits. The original network's precision is 32 bits.

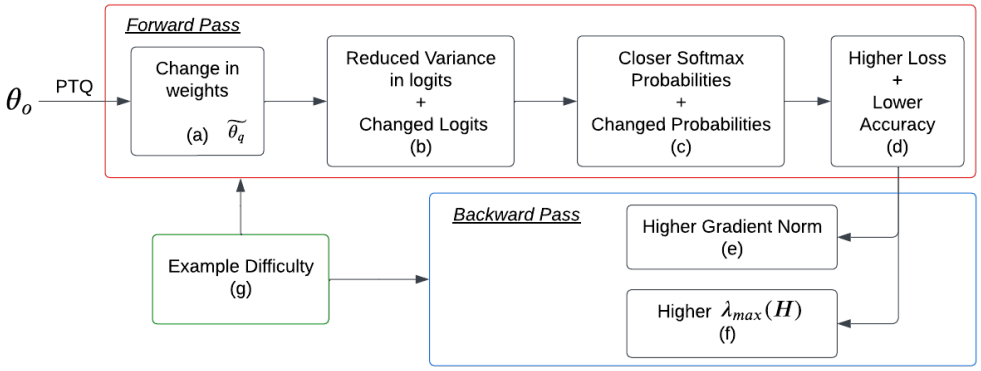


Figure 2: The impact flow of quantization.

4 Factors Impacting Fairness

The impact of quantization occurs through multiple stages, as shown in Fig. 2. During the forward pass, the effect of the changes in weights propagates throughout the network and leads to changes in logits, whose behavior is reflected in the softmax probabilities and, therefore, the loss. To better understand and visualize the effects of higher loss on the network weights, we use backpropagation without actually updating the weights, motivated by the second order Taylor Series expansion of the loss function at point x_c,

$$L(x) = L(x_c) + \nabla L(x_c) \cdot (x - x_c) + \frac{1}{2}(x - x_c)^T H (x - x_c) \tag{4}$$

Here, ∇L represents the gradient G. Now, for every group and precision, we study the gradient norm $||G_g^L||$ and the largest eigenvalue of the hessian matrix $\lambda_{max}(H_g^L)$ for the loss function L. The gradient norm helps us understand how far away the solution is from a better state in the solution space. Whereas, eigenvalues of the Hessian matrix provide crucial information about the steepness in the loss surface. Quoting from

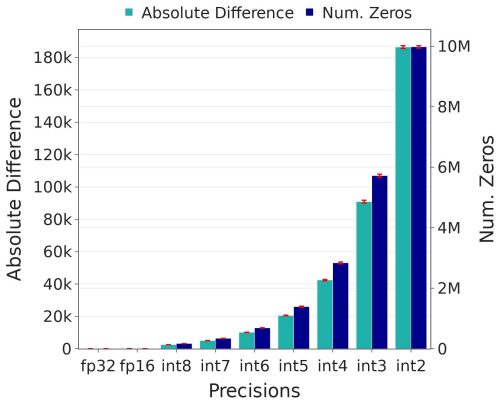


Figure 3: Changes in weights due to quantization

Tran et al. (2022), the maximum of the eigenvalues indicates how well the solution can separate the groups. Keskar et al. (2017); Li et al. (2018) support that top eigenvalues of the Hessian matrix aid in understanding the loss landscape. We next look at each stage of the impact flow in detail.

4.1 Changes in the Weights

The root cause of the impact flow of quantization is the change in weights of the network. The absolute difference in the weights is defined as

$$\text{Absolute difference in the weights} = \sum_{k=1}^{K} |\widetilde{\theta}_{q,k} - \theta_{o,k}| \tag{5}$$

However, the impact does not only include the absolute difference, but also involves the fraction of "zero" weights induced by quantization. While the former quantifies how much the weights have deviated from the original values, the latter is indicative of the loss of information due to sparsity defined by

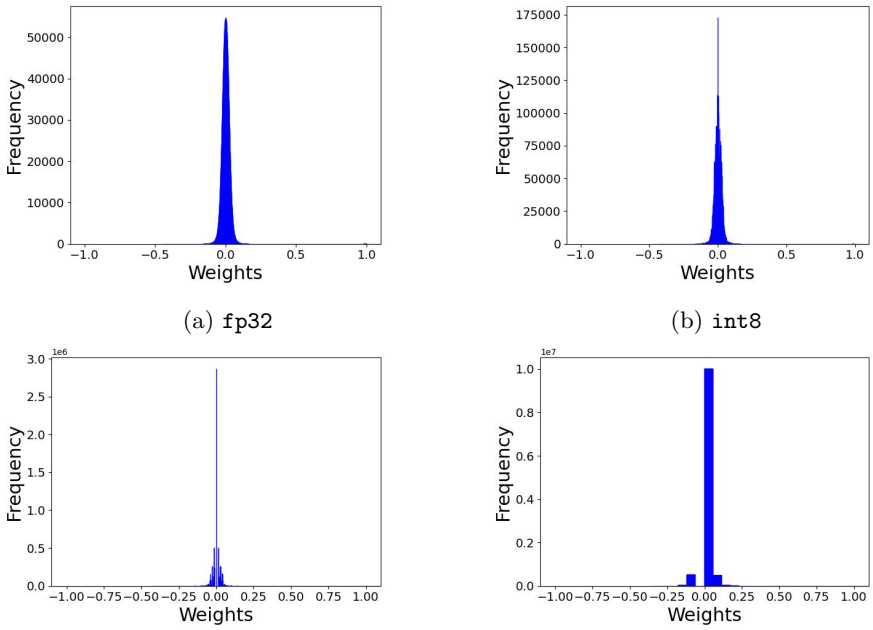
$$\text{Sparsity} = \frac{1}{K}\sum_{k=1}^{K} I(\theta_{.,k} = 0) \tag{6}$$

where I denotes the indicator function, $\theta_{.} \in \{\theta_o, \theta_q\}$. The absolute difference is controlled by the reduction in the precision of the weights. For example, θ_4 has 28 lesser bits to represent the weights in comparison to θ_{32}, which persists even after scaling by S. Whereas, sparsity increases when the weights are pushed to the '0 bin' during quantization which continues to remain as 0s even after scaling. While achieving higher compression, this effect is similar to (unstructured weight) magnitude pruning Han et al. (2015b); Zhu & Gupta (2017); Frankle & Carbin (2019), where some of the weights of the network are changed to 0.

Fig. 3 illustrates an increase in both absolute difference in weights and sparsity as precision reduces. On the other hand, Fig. 4 shows the weight distribution of $\widetilde{\theta}_q$ for different precisions, indicating a distribution shift towards the center with reducing precision. Clearly, reducing the precision increases the sparsity of the network, therefore, making it more like a pruned network (by weight magnitude). As shown in Tran et al. (2022), increasing the pruning ratio, i.e., increasing sparsity, has a disparate impact on the accuracy of the model. With higher sparsity in a quantized model, disparity worsens, analogous to the impact of sparsity in a pruned model. These combined changes affect the logits of the network, which we analyze next.

4.2 The Effect on Logits and Probabilities

The logits undergo two transformations as a result of quantization. First, the numerical values undergo a significant change, causing distinct differences in the magnitudes and resulting in different highest values among the logits. At lower precisions, this shift can cause the highest values to occupy incorrect positions



(a) `fp32` (b) `int8`

(c) `int4` (d) `int2`

Figure 4: Weight Distributions

in the logit vector, consequently leading to inaccurate predictions. Second, the variance between the logits is affected, resulting in different loss values.

We study the change in numerical values using cosine distance, defined as,

$$CD(A, B) = 1 - \frac{A \cdot B}{\|A\| \cdot \|B\|} \tag{7}$$

where A and B are two vectors of equal length. Let the average cosine distance between $f_{\widetilde{\theta_q}}$ and f_{θ_o} across the samples of a group be,

$$\text{Average cosine distance} = \frac{1}{|G|} \sum_{i}^{|G|} CD(f_{\widetilde{\theta_q}}(x_i), f_{\theta_o}(x_i)) \tag{8}$$

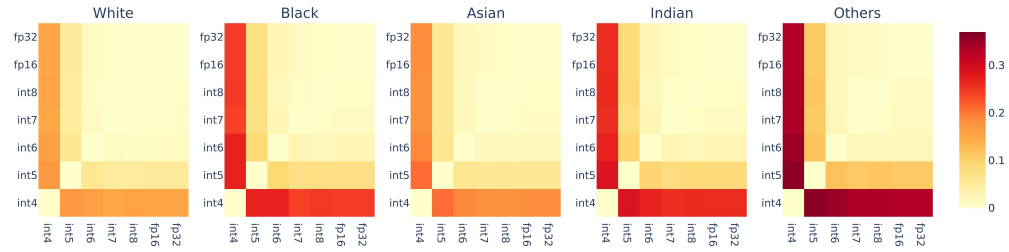
Fig. 5a shows that the angle between different quantization levels is largest for the minority class `Others` and the least for the majority class `White`. CD captures the changes that occur in the logits due to quantization, however, norm based metrics fail to capture it, as observed in Fig. 5b and Fig. 5c. Note that we are not able to show θ_2 and θ_3 as they produce null vector logits for some images which makes cosine distance inapplicable.

The mean variance among logits within each group, represented as,

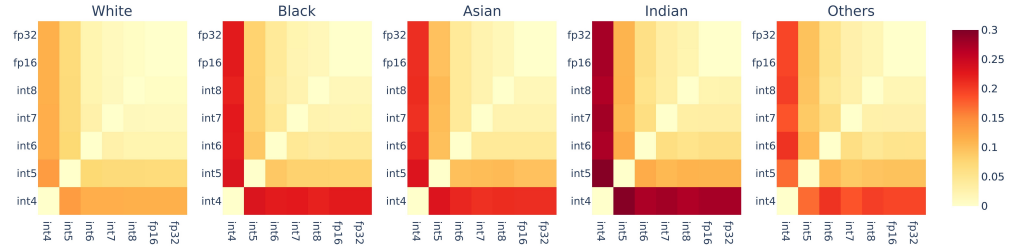
$$\text{Mean variance of logits} = \frac{1}{|G|} \sum_{i}^{|G|} \text{Var}(f_\theta(x_i)) \tag{9}$$

decreases with decreasing precision, as observed in Figure 5d. Notably, the group `White` exhibits the highest variance, while the `Others` group demonstrates the least variance. This reduction indicates that the separability of groups worsened due to quantization.

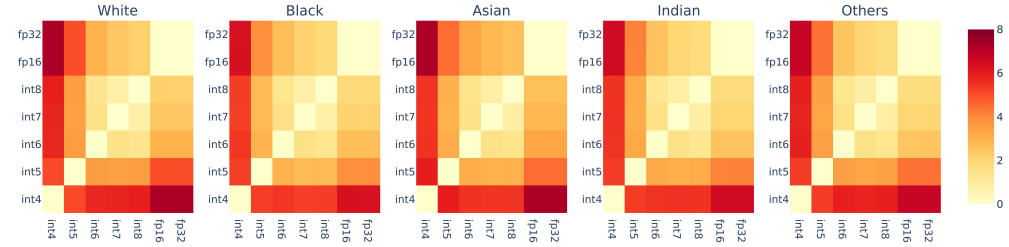
These changes in logits subsequently induce both a reduction in variance and a distribution difference in the softmax probabilities. At lower precisions, there is a substantial decrease in variance across all groups, with



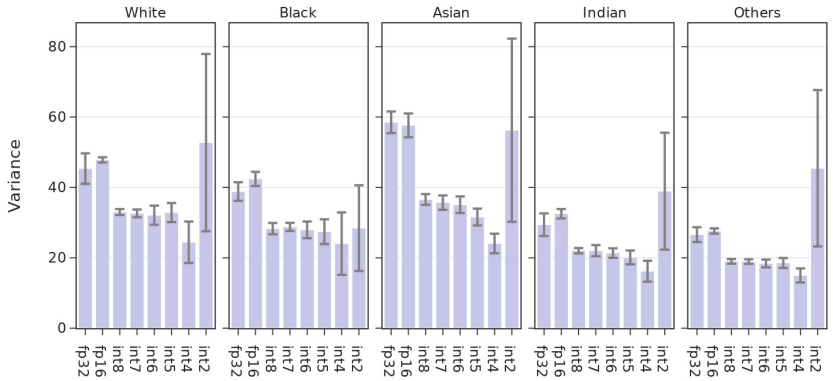
(a) Cosine Distance between logits. Darker shade implies higher distance.



(b) L1 Distance between logits for different quantization levels across classes - UTKFace. Note that L1 does not capture the dissimilarity as finely as cosine distance.



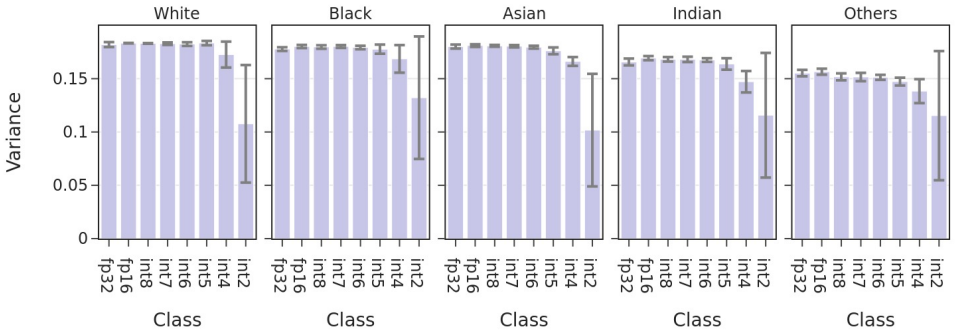
(c) L2 Distance between logits for different quantization levels across classes - UTKFace. Note that L2 is worse in comparison to both L1 and CD, therefore, indicating that the shift due to quantization is better observed through an angle based measure than a norm based measure.



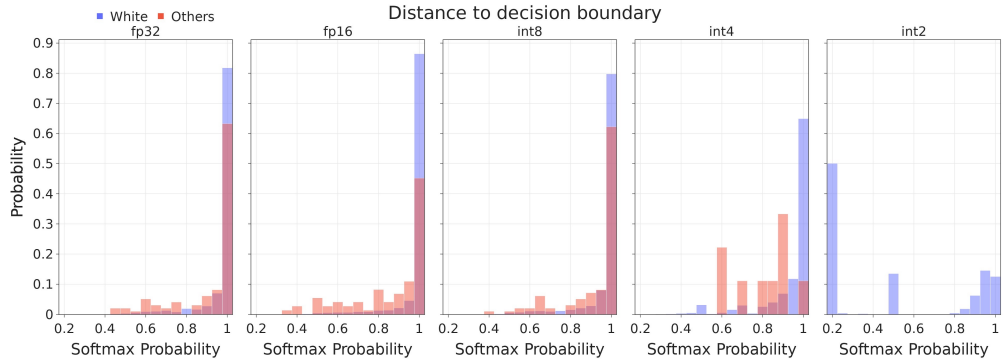
(d) Decrease in precision leads to reduction in variance in logits

Figure 5: Logits analysis

the `Others` group being affected the most, as illustrated in Fig. 6a. This reduced variance is analogous to the output-softening nature of the high-temperature scaling, which softens the logits of the network. Further, the disruption in the softmax probability distribution links to the inability of the precision to capture the



(a) Reduced variance in logits has a persistent presence in softmax probabilities



(b) The probability distribution of the distance to decision boundary (softmax probability). Notice the distribution shift of `Others` to the left and eventually disappears as precision reduces.

Figure 6: The effect on Softmax probabilities

original model's behavior. The softmax probability can be viewed as a Distance To the Decision Boundary ($DTDB$). We define $DTDB_{i,g}$ as the softmax probability obtained for each sample i belonging to group g, and that is plotted in Fig. 6b. If $DTDB_{i,g} > DTDB_{i,g'}$, then group g is farther away from the decision boundary than g', which implies an easier classification. Fig. 6b shows a strong leftward shift of distribution for `Others` unlike `White`, indicating that reduced precision induces uncertainty in the model for minority classes.

4.3 Contribution to Loss and Accuracy

The reduced variance in softmax probabilities, together with the changed values, adversely affect the loss and accuracy of the model as depicted in Fig. 7. The per-group loss is highest for `Others` and least for `White`. In addition, it is reflected as a direct impact on the accuracy of the model, as observed in Fig. 1. These circumstances indicate a clear, unfavorable movement of the model in the optimization space for all the classes, due to quantization, with the most affected being `Others`. In order to better understand this degraded position, we backpropagate the loss and observe how the gradient norm and Hessian are affected.

4.4 Observing Unfairness through Gradient norms

The gradient norm provides insight into the convergence of the optimization problem, indicating the proximity of the solution in the optimization space to a local minima Zhao et al. (2022). We find the group gradient norm for a quantized network using the gradients obtained by passing the test set (without weight

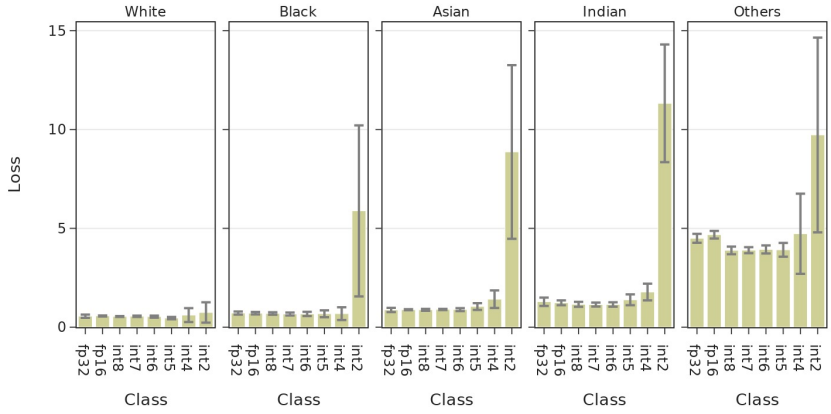


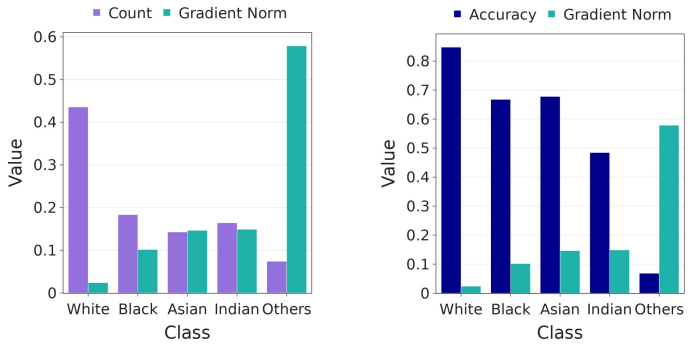
Figure 7: Higher group loss for `Others` due to PTQ

updates) and evaluating the ℓ_2 norm, given by,

$$G(\widetilde{\theta_q}; D_g) = \sqrt{\sum_{k=1}^{K} \left(\frac{\partial L(\widetilde{\theta_q}; D_g)}{\partial \widetilde{\theta_{q,k}}} \right)^2} \tag{10}$$

This measure also signifies the extent of gradient updates necessary for the model to improve its prediction. Next, we examine factors that exhibit a correlation with the gradient norm.

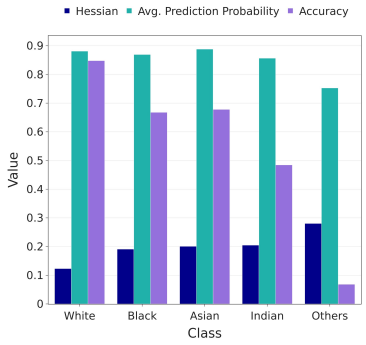
Consider the situation when D is passed as a single batch for gradient updates. Initially, the averaged gradients are dominated by classes with a higher number of samples. This effect persists even when there are mini batches, although lower in impact. Therefore, the gradients are also controlled by the class distributions and batch size. In addition, the initial gradients are heavily dependent on the initialization of θ. However, the effects of batch size (if moderate) and initialization dampen as the network trains further. We therefore look at the effects of per group sample counts of the test set on the gradient norm. Fig. 8a shows an inverse trend between the gradient norm and group sizes for θ_4. Notice the huge disparity between the Gradient norm of `White` and `Others`. We argue that this occurs due to the initial dominance of the majority classes in the gradients which are inverted at some point during training and are reflected post training and quantization. It further reflects an inverse trend with the accuracy of the model as observed in Fig. 8b.



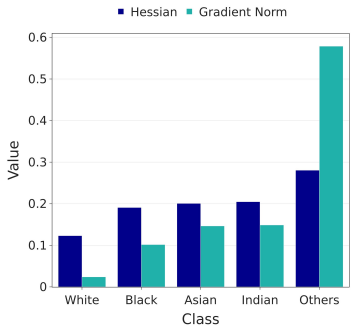
(a) Gradient norm vs. group size

(b) Gradient norm vs. accuracy

Figure 8: Trends of gradients against group size (normalized) and accuracy on an Int4-quantized model for ResNet18



(a) The largest Hessian eigenvalue (λ_{max}) is inverse to accuracy and average prediction probability



(b) The largest Hessian eigenvalue (λ_{max}) aligns with the gradient norm

Figure 9: The largest Hessian eigenvalue for different groups

4.5 Reflection of unfairness on the Hessian

$\lambda_{max}(H_g^l)$ helps explain the steepness of the loss surface at that point in the solution space for a particular group. Fig. 9a shows that λ_{max} and accuracy move in opposite directions, indicating a larger λ_{max} for the minority group. This implies that the steepness is the highest for `Others`, and a corresponding update to the weight would cause a higher reduction in the loss as compared to any other group. To capture the average of the highest softmax prediction probabilities across the groups, we define *average prediction probability*,

$$\text{Avg. prediction prob.} = \frac{1}{|G|}\sum_{i}^{|G|} max(\sigma(f_\theta(x_i))) \tag{11}$$

We also observe in Fig. 9a that the average prediction probability is lowest for the group with the highest λ_{max} and vice versa. Fig 9b shows gradient norm and λ_{max} moving toward the same direction, indicating that quantization induces a combined effect on them.

4.5.1 Comparisons for Different Quantization Precisions

The trend across groups is similar for all quantization levels (same color across different groups) in Fig. 10. However, for different quantization levels within a group, we observe that for `fp32` and `fp16`, $\log(\lambda_{max})$ is almost equal, but for the integer precisions, the overall trend is increasing. One would expect that when we quantize with lower bits, the ability to be closer to the original weight degrades and the solution ends up at an inferior point in space. However, it need not be true that the steepness of the inferior point is bad too. Fig. 10 therefore indicates that the scope for improvement is close to this order, `int2` > `int4` > `int8`, in most cases.

4.6 Example Difficulty

When sub-sampling for majority classes to create a balanced dataset, we observed that during training, the per-group accuracy was initially lower for the group `Others`, compared to the other classes. The model then reaches equal training accuracy for every class towards the end, yet performs poorly on the test set for `Others`. Given the scenario that the number of images used to train is nearly the same, the above two observations explain that the data in the test set for `Others` is rather difficult for the network to classify correctly. We attribute this lack of generalizability to the relatively complex facial features in the `Others` class, such as color, age, hair, and structure, which contribute to example difficulty. Example difficulty, however, has an impact during both forward and backward passes. In the forward pass, the network has already been trained better for the lower difficulty samples, which in addition, during the back pass, has a continuing effect.

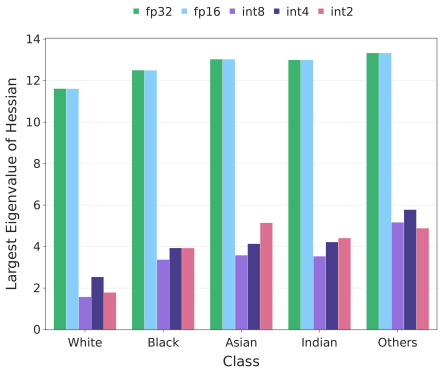


Figure 10: $\log(\lambda_{max})$ for different groups and precisions

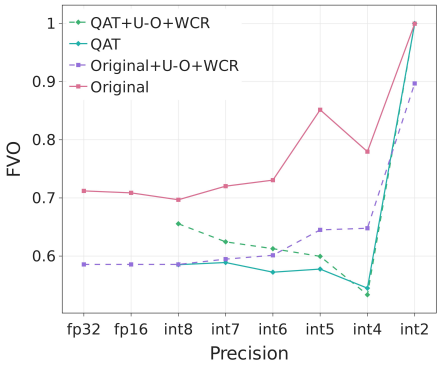


Figure 11: Comparison between mitigation solutions for Resnet18.

5 Mitigation Solutions: Fair Quantization

In this section, we look at what improves the fairness of a quantized model. We first begin with methods that make an `fp32` model fairer, followed by using `QAT` as a possible solution to quantization's disparity. Finally we combine the former and latter to achieve the same with lesser data. The results performed against the UTKFace dataset are presented in this section. Please refer to the Appendix for the results of other datasets.

5.1 A Fairer Base Model

Under and over-sampling (`U-O`) the majority and minority classes, respectively, reduces the per-group differences in the gradient norm during training. However, capturing the example difficulty from Sec. 4.6 involves weighting the harder classes higher than the rest, so that the solution is not biased only towards getting the easier samples right, during the beginning of the optimization. Let `WCR` denote the weighted cross entropy loss function given by,

$$L(\theta; D) = -\frac{1}{M}\sum_{i=1}^{M}\sum_{g=1}^{G} a_g \cdot y_{ig} \cdot \log(p_\theta(x_i))_g \tag{12}$$

where a_g is the weight for group g. We choose `WCR` with weights ($[0.1, 0.1, 0.1, 0.1, 0.6]$) such that `Others` has the highest weight and the rest of the classes are given equal weight. Next, we train a fairer model θ_f using both `U-O` and `WCR`.

5.2 Mixed-precision QAT

In `PTQ`, there is no retraining to allow for the weights to be re-adjusted to the change in environment. However, in `QAT`, the losses are calculated and the weights are updated in accordance with the changing

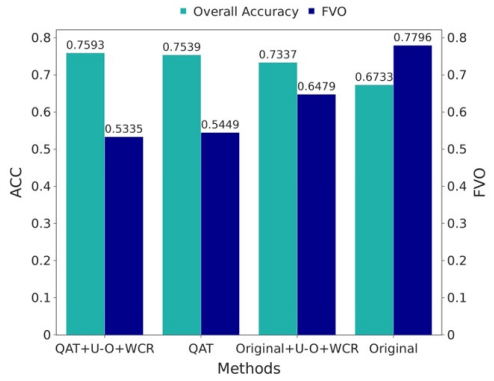


Figure 12: Tradeoffs of overall accuracy vs. `FVO` for `int4`.

weights and activations. This provides an opportunity to control the changes in group accuracies, unlike `PTQ`. We take into account the possible oscillations that occur in weights during `QAT` and adopt the oscillation dampening method in Nagel et al. (2022) to avoid it. We also use mixed precision (MPQAT) for ResNet18, where the first and last layers use 8 bits and the activation are 32 bits according to Bhalgat et al. (2020). $L(\widetilde{\theta}_q; D)$ is minimized as a result of QAT. This dampens the adverse effects of quantizing both weights without training, together with lowering the information loss in layers that are critical for classification.

5.3 Fair QAT

Combining `U-O`, `WCR`, and `MPQAT` reduces the disparate impact of Quantization for ResNet18, as observed in Fig. 11. In fact, `MPQAT` alone also reduces `FVO`. Here, `PTQ` has acceptable `FVO` for higher precisions, provided the original model is relatively fair. But, for lower precisions, that does not hold. We next visualize `OA` and `FVO` together in Fig. 12 showing that our method achieves both the highest `OA` and lowest `FVO`. This would provide an informed decision for the potential users to make in consideration of the tradeoffs between overall accuracy and fairness.

6 Discussion

Other aspects of quantization. In our analysis, the focus was solely on weight quantization without activation quantization. Since activation quantization further worsens the network accuracy, the results would either resemble that of weight quantization or expand the disparity that we observed. For many quantization precisions, QAT has lower FVO than the Fair QAT, which shows that QAT alone is sufficient to mitigate the disparate impact. The best hyperparameters for QAT were known for `int4`, but were unknown for the other precisions. We therefore used the same (that of `int4`) hyperparameters for all precisions. However, this should not be interpreted as to refrain the potential of other hyperparameters. The aforementioned `WCR` weights were chosen as a result of the example difficulty observed for the `Others` class. The weights are consistent for the rest of the classes, as the difficulties are nearly equal.

7 Conclusion

The disparate impact caused by PTQ is explained by an impact flow that passes across stages in the forward pass, whose effects can be visualized as a shift of the model to a sub-optimal state in the optimization landscape, using gradient norms and eigenvalues of the Hessian matrix. However, we show that utilizing simple methods, such as undersampling, oversampling, and adjusting weights in `WCR`, leads to fairer models. When QAT is used in conjunction, it gives a class of fairer quantized models with a little compromise in overall accuracy.

References

Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 289–295, 2019.

Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 7950–7958, 2019.

Yash Bhalgat, Jinwon Lee, Markus Nagel, Tijmen Blankevoort, and Nojun Kwak. Lsq+: Improving low-bit quantization through learnable offsets and better initialization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 696–697, 2020.

Ting-An Chen, De-Nian Yang, and Ming syan Chen. Climbq: Class imbalanced quantization enabling robustness on efficient inferences. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=F7NQzsl334D`.

Ting-An Chen, De-Nian Yang, and Ming-Syan Chen. Overcoming forgetting catastrophe in quantization-aware training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17358–17367, October 2023.

Yoni Choukroun, Eli Kravchik, Fan Yang, and Pavel Kisilev. Low-bit quantization of neural networks for efficient inference. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3009–3018. IEEE, 2019.

Tim Dettmers and Luke Zettlemoyer. The case for 4-bit precision: k-bit inference scaling laws. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7750–7774. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/dettmers23a.html`.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 293–302, 2019.

Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah, and Xia Hu. Fairness via representation neutralization. *Advances in Neural Information Processing Systems*, 34:12091–12103, 2021.

Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. In *International Conference on Learning Representations*, 2020.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Elias Frantar and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. *Advances in Neural Information Processing Systems*, 35:4475–4488, 2022.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015a.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015b.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

Xijie Huang, Zechun Liu, Shih-Yang Liu, and Kwang-Ting Cheng. Efficient quantization-aware training with adaptive coreset selection. *arXiv preprint arXiv:2306.07215*, 2023.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *arXiv e-prints*, pp. arXiv–1609, 2016.

Nitish Shirish Keskar, Jorge Nocedal, Ping Tak Peter Tang, Dheevatsa Mudigere, and Mikhail Smelyanskiy. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Jeonghoon Kim, Jung Hyun Lee, Sungdong Kim, Joonsuk Park, Kang Min Yoo, Se Jung Kwon, and Dongsoo Lee. Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization. *arXiv preprint arXiv:2305.14152*, 2023.

Andrey Kuzmin, Markus Nagel, Mart van Baalen, Arash Behboodi, and Tijmen Blankevoort. Pruning vs quantization: Which is better? *arXiv preprint arXiv:2307.02973*, 2023.

Jung Hyun Lee, Jeonghoon Kim, Se Jung Kwon, and Dongsoo Lee. Flexround: Learnable rounding by element-wise division for post-training quantization. 2022.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=POWv6hDd9XH`.

Hao Liang, Pietro Perona, and Guha Balakrishnan. Benchmarking algorithmic bias in face recognition: An experimental approach using synthetic faces and human evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4977–4987, 2023.

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart Van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021.

Markus Nagel, Marios Fournarakis, Yelysei Bondarenko, and Tijmen Blankevoort. Overcoming oscillations in quantization-aware training. In *International Conference on Machine Learning*, pp. 16318–16330. PMLR, 2022.

Cuong Tran, Ferdinando Fioretto, Jung-Eun Kim, and Rakshit Naidu. Pruning has a disparate impact on model accuracy. In *NeurIPS*, 2022.

V. Vapnik. Principles of risk minimization for learning theory. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*, NIPS'91, pp. 831–838, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc. ISBN 1558602224.

Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199*, 2018.

Peisong Wang, Qiang Chen, Xiangyu He, and Jian Cheng. Towards accurate post-training network quantization via bit-split and stitching. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9847–9856. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/wang20c.html`.

Bichen Wu, Yanghan Wang, Peizhao Zhang, Yuandong Tian, Peter Vajda, and Kurt Keutzer. Mixed precision quantization of convnets via differentiable neural architecture search. *arXiv preprint arXiv:1812.00090*, 2018.

Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *International conference on machine learning*, pp. 11492–11501. PMLR, 2021.

Zhewei Yao, Zhen Dong, Zhangcheng Zheng, Amir Gholami, Jiali Yu, Eric Tan, Leyuan Wang, Qijing Huang, Yida Wang, Michael Mahoney, et al. Hawq-v3: Dyadic neural network quantization. In *International Conference on Machine Learning*, pp. 11875–11886. PMLR, 2021.

Song Yang Zhang Zhifei and Qi Hairong. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

Han Zhao and Geoffrey J Gordon. Inherent tradeoffs in learning fair representations. *The Journal of Machine Learning Research*, 23(1):2527–2552, 2022.

Ritchie Zhao, Yuwei Hu, Jordan Dotzel, Chris De Sa, and Zhiru Zhang. Improving neural network quantization without retraining using outlier channel splitting. In *International conference on machine learning*, pp. 7543–7552. PMLR, 2019.

Yang Zhao, Hao Zhang, and Xiuyuan Hu. Penalizing gradient norm for efficiently improving generalization in deep learning. In *International Conference on Machine Learning*, pp. 26982–26992. PMLR, 2022.

Hengyi Zhou, Hongyi He, Wanchen Liu, Yuhai Li, Haonan Zhang, and Longjun Liu. A novel differentiable mixed-precision quantization search framework for alleviating the matthew effect and improving robustness. In *Asian Conference on Machine Learning*, pp. 1277–1292. PMLR, 2023.

Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*, 2017.

A Appendix

A.1 Analysis on CelebA dataset

We created a subset of the CelebA dataset with images of `Blond Male`, `Blond Female`, `Black Hair Male`, `Black Hair Female`. The `Blond Male` class is the minority class with only 1749 images. We sampled 8000 images each from the other classes. Following this, there is a test split of 20%.

We observe exacerbated disparity in the CelebA dataset in Fig. 13. The accuracy of group `Blond Male` greatly suffers due to quantization.

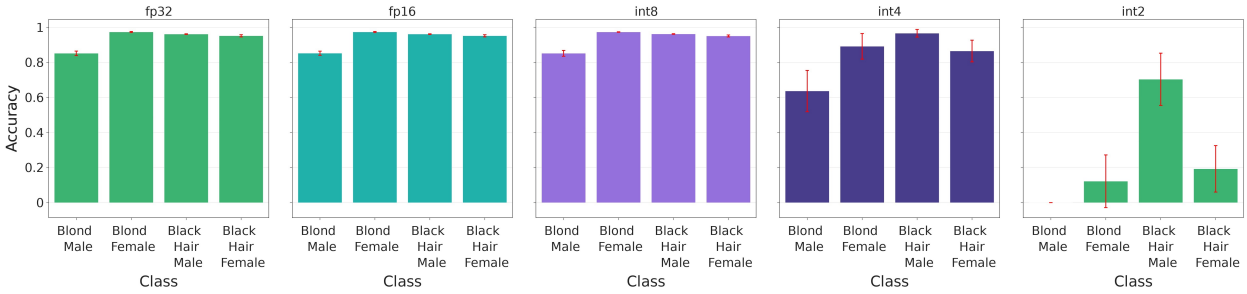


Figure 13: Accuracy for groups for different quantization precisions on CelebA dataset.

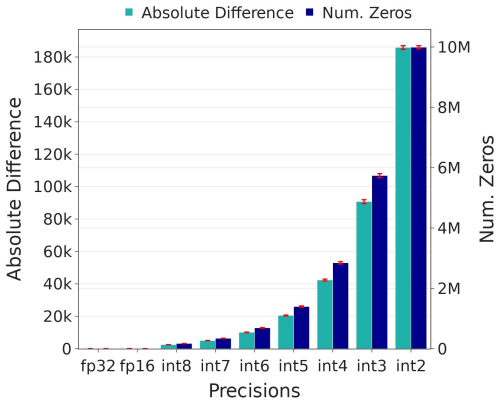


Figure 14: Absolute difference and the number of zeroes for quantized models. Notice the consistency with the UTKFace results

When the model is quantized, both the absolute difference and the sparsity (the number of zero parameters) are aligned with the results of the UTKFace dataset, indicating that the impact of quantization is consistent in multiple datasets. This phenomenon is observed in Fig. 14.

As a result of quantization, the logits are again affected adversely leading to a reduction in their variances, as observed in Fig. 15.

Next, we observe that the angular distance to the original logits worsens for the minority group in Fig. 16.

Furthermore, the effect on the softmax probabilities is persistent for the minorty class as seen in Fig. 17a. Therefore, the group losses increase in Fig. 17b.

The effect of quantization is also seen on the distribution of the minority group, where there is a strong leftward shift, in Fig. 18.

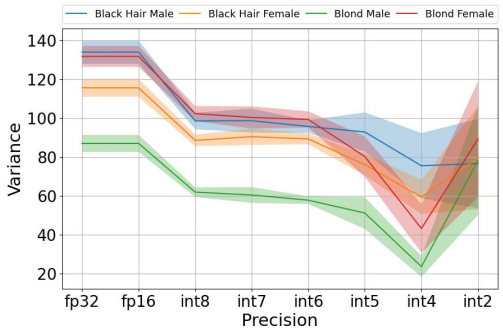


Figure 15: The variance between logits follows a downward trend across decreasing precisions.

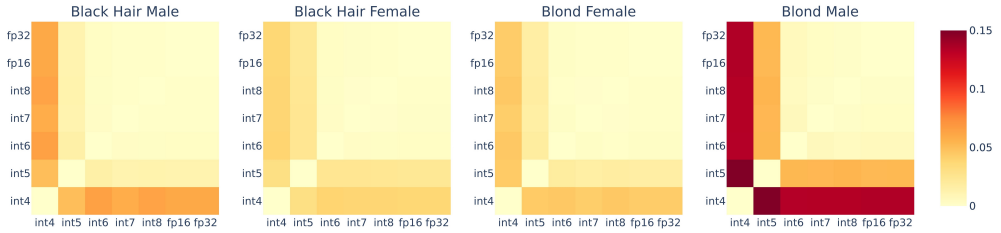


Figure 16: Cosine distance between logits of different precisions per class. The `Blond Male` group has the highest impact.

In the backward pass, in Fig. 19a and Fig. 19b, the trends are consistent with Fig. 8a and Fig. 8b on the UTKFace datset, where the minority class with the least count has the highest gradient norm with the lowest group accuracy and vice versa.

Again, in Fig. 20a, the gradient norm and λ_{max} move along the same direction. We then see an inverse trend between the accuracy and λ_{max} and average prediction probability in Fig. 20b. Next, in Fig. 21, λ_{max} has a trend consistent with that observed in UTKFace experiments for θ_4.

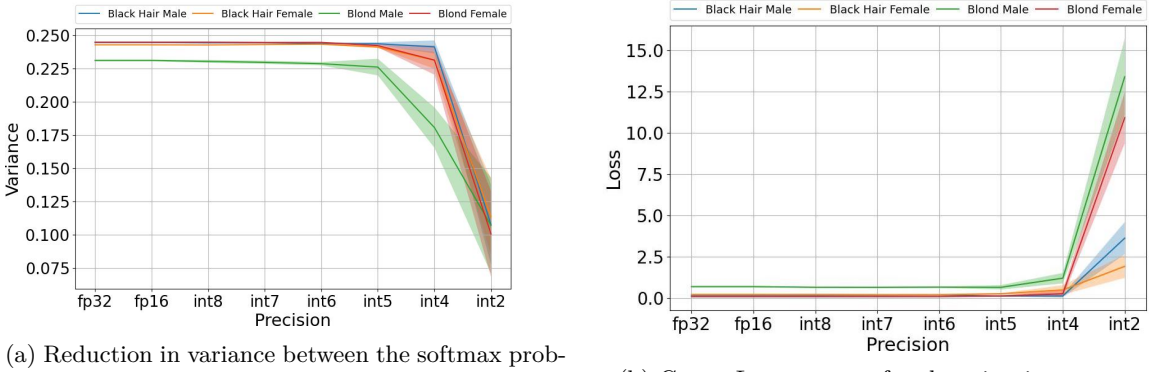
A.2 Fair Quantization on CelebA dataset

We compare the effectiveness of the solutions in Fig. 22 and observe that `QAT+FVO+WCR` performs the best with least `FVO` and highest overall accuracy for an `int4` quantized model.

A.3 Experiment Setup

SGD with momentum and LR scheduler were commonly used across all four models. Specific details for the models are,

- Original model: Cross-Entropy (CE) loss
- Fair original model: Weighted-CE loss, balanced data
- QAT model: CE loss, QAT
- Fair QAT model: Weighted-CE loss, balanced data, QAT



(a) Reduction in variance between the softmax probabilities as a result of the changes to the logits.

(b) Group Loss worsens for the minority group.

Figure 17: Softmax probabilities and group losses

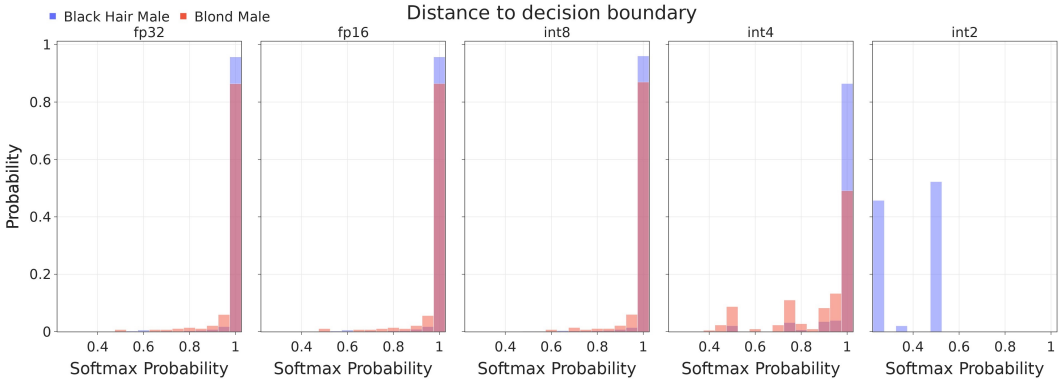
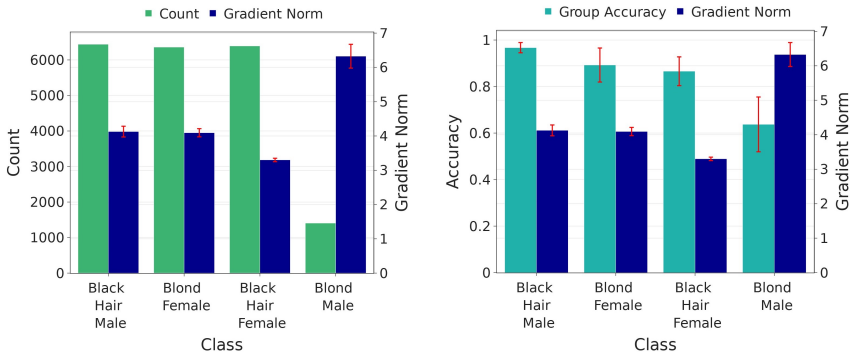


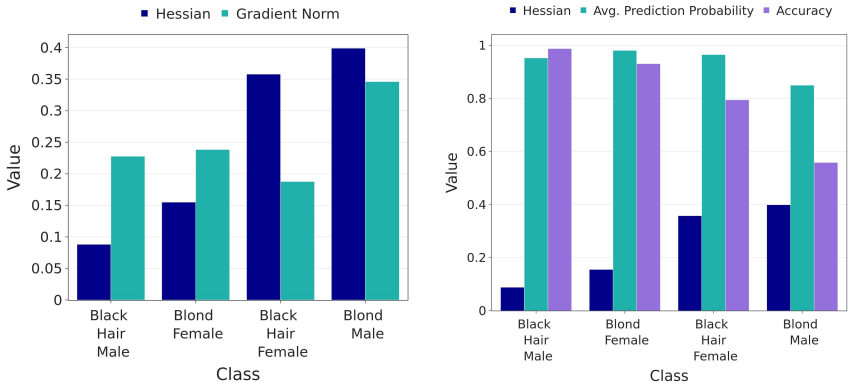
Figure 18: The probability distribution of the distance to decision boundary (softmax probability). Notice the distribution shift of `Blond Male` to the left and eventually disappears as precision reduces, consistent with the trend of the minority class in UTKFace.



(a) Inverse relationship between gradient norm and class count for θ_4

(b) Nearly inverse trend between group accuracy and gradient norm.

Figure 19: The changes in gradient norms



(a) λ_{max} (normalized) nearly aligns with Gradient norm (normalized) for a precision of θ_4

(b) Inverse trends between λ_{max}(normalized), average prediction probability and accuracy.

Figure 20: Differences in Hessian

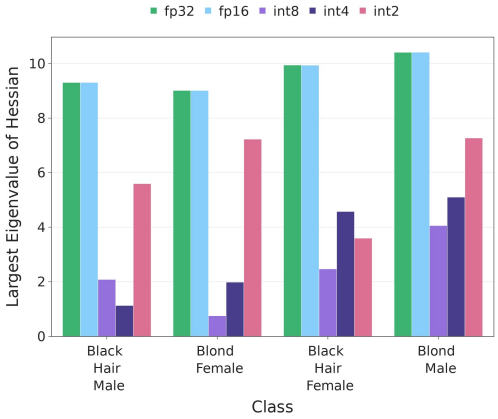


Figure 21: Overall increasing trend for λ_{max} at lower precisions.

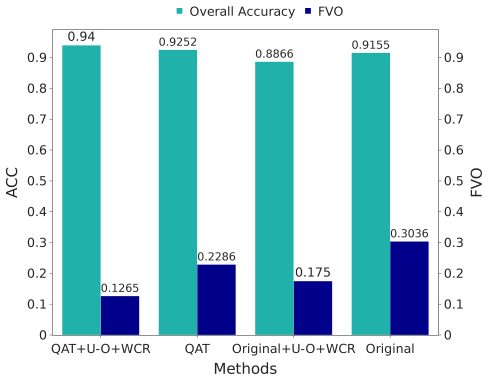


Figure 22: Trade off between overall accuracy and FVO for different methods for `int4` quantized model on CelebA dataset.