On Fairness of Low-Rank Adaptation of Vision Models

Zhoujie Ding
Department of Computer Science
Stanford University
d1ng@stanford.edu

Qianzhong Chen
Department of Mechanical Engineering
Stanford University
qchen23@stanford.edu

Abstract

Low-rank adaptation (LoRA) of large models has gained traction due to its computational efficiency. This efficiency, contrasted with the prohibitive costs of full-model fine-tuning, means that practitioners often turn to LoRA and sometimes without a complete understanding of its ramifications. In this study, we focus on fairness implications and ask whether LoRA has an unexamined impact on utility, calibration, and resistance to membership inference across different subgroups (e.g., races) compared to a full-model fine-tuning baseline. We present extensive experiments on image classifications using vision transformers – ViT-Base and Swin-v2-Large. Intriguingly, experiments suggest that while one can isolate cases where LoRA exacerbates model bias across subgroups, the pattern is inconsistent—in many cases, LoRA has equivalent or even improved fairness compared to its full fine-tuning baseline.

1. Introduction

The challenge of efficiently scaling large models has led to the growing interest and reliance on *parameter-efficient fine-tuning*, which focuses on adjusting only a small, deliberately chosen set of parameters in the base model (14; 7; 19; 18). Of particular interest is the low-rank adaptation (LoRA) technique (14), in which the pre-trained weight matrices are frozen while their changes from fine-tuning are approximated by low-rank decompositions. LoRA has received significant attention due to its simplicity and effectiveness in a variety of tasks across both language (21) and vision (10) domains. Despite the popularity of LoRA, little is known about its effects on trustworthiness, such as fairness and robustness. The lack of understanding together with LoRA's wide adoption implies that practitioners may be deploying models with unintended and potentially harmful consequences in high-stakes applications. To this end, this work initiates a study on *fairness* and asks the following:

What are the effects of LoRA, if any, on subgroup fairness?

Central to the existing knowledge gap is the prohibitive cost of full fine-tuning, which deters a direct comparison against LoRA. This is troubling since the increased adoption of large models often involves taking off-the-shelf pre-trained models (*e.g.*, Llama-2 (27)), fine-tuning them on custom data (if said models cannot reason in-context with few-shot prompts), and running them as part of (potentially high-stakes) decision-making processes. In many of these scenarios, such as enterprise (29), healthcare (33), and banking (23), practitioners may gravitate towards LoRA solely for its cost-effectiveness without adequate consideration for unfair outcomes. This can cause tangible harm when applied to tasks such as risk assessment, credit score estimation, loan approvals, and hiring/promotion evaluations.

Apart from the real-world motivations above, tangential prior work also inspired this study from an algorithmic standpoint. Specifically, LoRA is characterized by its *reduced fitting capacity* through low-rank approximations; a similar property is also inherent to *model pruning* and *differentially private training*. Respectively, (28) and (2) found that both pruning and private training can worsen the fairness of accuracy across subgroups (despite achieving good *overall* accuracy), as the sparsity and noisy gradients (due to private training) can both impact a model's ability to fit minority and underrepresented inputs. On the other hand, (17) and (1) showed that low-rank weights and representations can lead to better adversarial robustness. Prompted by these studies, we ask whether LoRA exhibits similar side effects and, if so, whether they are consistent across tasks and datasets. All in all—is LoRA's efficiency a "free lunch"?

In this study, we seek to better understand the fairness implications of LoRA on vision models via extensive experimentation. Our study and findings can be summarized as follows:

1. We fine-tune vision pre-trained models of ViT-Base (8) and Swin-v2-Large (22) across image gender and age classifications, juxtaposing full-model fine-tuning and LoRA and measuring the subgroups disparities on ac-

curacy, calibration, privacy as resistance to membership inference. **To our knowledge, our work is the first to provide a comprehensive empirical investigation into the fairness properties of low-rank adaptation.**

2. **Intriguingly, our experiments reveal no consistent pattern of LoRA worsening subgroup fairness**, compared to full fine-tuning across different vision model architectures and fairness evaluation metrics (§5). Note that isolated examples do exist where LoRA worsens fairness across subgroups, though such cases should be viewed with target metric sensitivity in mind (16); *e.g.*, LoRA may appear less fair via worst subgroup accuracy but equally fair under equalized odds difference (EOD), which also considers false positives. Nonetheless, for any fixed task and its appropriate fairness metrics that we experimented on, we found **no strong evidence** that LoRA is less fair.

3. **The fairness implications may depend on the quality of the underlying pre-trained model** (§5.2). We also observe cases where LoRA does exacerbate unfairness can disappear when the base pre-trained model is stronger (Fig. 2) and all else is kept constant. This suggests that the fairness properties of LoRA are not merely a function of its parameter efficiency (comparing model pruning (28)).

4. **The LoRA rank and subgroup size have little impact on subgroup fairness** (§5.5 and §5.6). While rank can be a confounding factor for its impact on model capacity and thus fairness (comparing pruning and private training), we did not observe a significant influence of rank on either utility or fairness. This is in line with existing utility analysis (14). Fairness is not solely a straightforward function of data, and we have verified that subgroup size did not significantly affect either utility or fairness.

2. Related Work and Background

An important paradigm in modern machine learning (ML) is to adapt large pre-trained models to downstream tasks through fine-tuning. The benefits of fine-tuning are two-fold: (1) it leverages the extensive knowledge stored in these pre-trained models, and (2) it promises greater efficiency compared to training from scratch. However, as models grow in size, this efficiency advantage becomes elusive due to increased demand on compute; for example, simply keeping the gradients of Llama-2 70B (27) in 16-bit precision requires around 140GB of memory, which is already infeasible for most commodity hardware. This gap motivates parameter-efficient fine-tuning methods and subsequent novel trustworthiness concerns. Here, we briefly outline work most closely related to the focus of this paper and some preliminaries that ground our analyses.

2.1. Low-Rank Adaptation

LoRA (14) is a widely used parameter-efficient fine-tuning algorithm for large models. It proposes to separate out the weight deltas from fine-tuning and approximate them using low-rank matrices; inference then involves forward passing both the (frozen) pre-trained model and the low-rank model deltas, also known as *adapters*, and summing the activations. Specifically, for a pre-trained weight matrix $\mathbf{W} \in \mathbb{R}^{d\times k}$ with dimensions d, k, LoRA approximates its changes from fine-tuning as $\Delta\mathbf{W} \approx \mathbf{BA}$ where $\mathbf{B} \in \mathbb{R}^{d\times r}$ and $\mathbf{A} \in \mathbb{R}^{r\times k}$ with rank $r \ll \min(d, k)$, and thus inference on input $\mathbf{x} \in \mathbb{R}^d$ is $\mathbf{Wx} + \mathbf{BAx} \approx (\mathbf{W} + \Delta\mathbf{W})\mathbf{x}$ if $\Delta\mathbf{W}$ is obtained through full fine-tuning. $\mathbf{A}$ and $\mathbf{B}$ can be updated directly via backpropagation. Typically, implementations of LoRA apply to all query and value matrices of self-attention layers in the pre-trained transformer. To fine-tune for supervised tasks, an additional head is also attached to the last layer of the model.

2.2. Definition of Equalized Odds Difference

Equalized Odds Difference (30) is calculated as the larger of two quantities: the disparity in true positive rates and the disparity in false positive rates between two groups distinguished by a sensitive characteristic. Mathematically, it is represented as:

$$M_{\text{eod}} = \max\left\{M_{\text{TP}}, M_{\text{FP}}\right\},$$

with the components defined as follows (Y here is the true label and A is the sensitive attribute):

- True positive equalized odds difference:

$$M_{\text{TP}} = |P(f(X) = 1 \mid Y = 1, A = 1) - P(f(X) = 1 \mid Y = 1, A = 0)\,|$$

- False positive equalized odds difference:

$$M_{\text{FP}} = |P(f(X) = 1 \mid Y = 0, A = 1) - P(f(X) = 1 \mid Y = 0, A = 0)\,|$$

A significant M_{eod} reflects a discrepancy in error rates between the groups, indicating the unfairness of the model prediction.

2.3. Definitions of Terms in Model Calibration

Reliability Diagrams (or Calibration Curves): Reliability diagrams (26) plot the model's predicted confidence against its observed accuracy. To construct these diagrams, predictions are grouped into M interval bins (each of size $1/M$). Within each bin (say B_m), the model's accuracy, denoted as $\text{acc}(B_m)$, is computed as:

$$\text{acc}(B_m) = \frac{1}{|B_m|}\sum_{i\in B_m} 1(\hat{y}_i = y_i),$$

where $\hat{y}_i$ represents the predicted class label for sample i, and y_i is the corresponding true label. The average confidence for bin B_m, expressed as $\text{conf}(B_m)$, is the mean predicted probability for the samples within that bin:

$$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i,$$

with $\hat{p}_i$ denoting the confidence for sample i. A perfectly calibrated model should exhibit $\text{acc}(B_m) = \text{conf}(B_m)$ for each bin, *i.e.*, the diagram should plot the identity function. The distance between the observed accuracy and the predicted probability in each bin represents the calibration gap.

Expected Calibration Error(ECE): ECE (25) is a scalar summary statistic of how well the model is calibrated by calculating the weighted absolute difference between predicted probability (*i.e.,* confidence) and accuracy across all the confidence bins. Mathematically, it is represented as:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|,$$

where n is the total number of samples. The ECE reflects the calibration gap, with lower values indicating a model whose predicted probabilities are closer to the true outcomes.

Importance of Calibration for Fairness: Well-calibrated models produce probability estimates that can be trusted equally across different demographic groups. If a model is not well-calibrated, some groups may systematically receive overconfident or underconfident predictions. For example, in the scenario of image gender classification, it's imperative that the model operates impartially among all demographics to avoid unfairly censoring content from specific groups.

2.4. Membership Inference Attacks (MIAs)

We first discuss common considerations and evaluation metrics for MIA. We then dive deeper into the two MIAs we consider, Likelihood Ratio Attack (LiRA, (4)) and LOSS (32), and provide more details about the training and attack setup.

The goal of MIAs is to estimate the membership of query points as accurately as possible. To this end, consider a dataset space $\mathcal{X}$, label space $\mathcal{Y}$, a real-world distribution $\mathbb{D}$ over $\mathcal{X} \times \mathcal{Y}$, a training dataset D sampled from $\mathbb{D}$, and a training procedure $f_D \leftarrow \mathcal{T}(D)$ where f_D is a machine learning model that outputs a probability distribution over $\mathcal{Y}$ for any given instance $x \in \mathcal{X}$. The attacker has black-box access to the model f_D (known as the target model) and wishes to determine whether $(x, y) \in D$. Following (4), evaluating the effectiveness of an MIA is typically done by measuring the true positive rate at a low false positive rate. The justification for this is that being able to confidently infer that just one data point is a member is a much bigger breach in privacy than being 51% confident in the membership of a larger number of datapoints, even though both instances may score the same in other classification metrics such as accuracy or AUROC. In other words, an attack is only an effective privacy breach when it has a low false positive rate.

LiRA: The Likelihood Ratio Attack (LiRA) is an important MIA strategy due to its efficacy (4). Here, the attacker trains N shadow models $f_{D_i} \leftarrow \mathcal{T}(D_i)$ where each D_i (for $i \in \{1, \ldots, N\}$) is randomly sampled from $\mathbb{D}$ (such that $\forall i : (x, y) \notin D_i$) in order to mimic the behavior of the target model f_D. For any given model $f : \mathcal{X} \to \mathcal{Y}$, let $f(x)_y$ denote the confidence of f on (x, y), in other words, the probability value output by $f(x)$ for label y. Although the attacker has no information about D, it has complete information about each D_i as they were the one who sampled the dataset themselves. Thus, the attacker models a distribution of the confidences $f_{D_i}(x)_y$ for each shadow model i, where D_i is sampled from $\mathbb{D}$ and $(x, y) \notin D_i$. From this, the attacker can then compare the target model's confidence $f_D(x)_y$ and perform a hypothesis test against the null hypothesis of $(x, y) \notin D$ (rejecting the null hypothesis and inferring membership whenever the cumulative distribution function of $f_D(x)_y$ is above some threshold τ). Note that here, we are performing the "offline" version of the attack, as the traditional "online" approach is infeasibly computationally expensive. For more details about LiRA, we refer the reader to (4).

In our experiments, we partitioned a small subset (20%) from both the training and evaluation sets (for the target models) for membership inference evaluation and used the remaining data to train the shadow models. This way, we ensure that the membership inference target dataset is disjoint from the shadow training dataset, which is a necessary assumption for the offline LiRA attack. To ensure variability between the different shadow datasets, we randomly sample 50% of the shadow training dataset to train each shadow model. The shadow models are fine-tuned via LoRA using the target model's pre-trained model.

LOSS: The LOSS attack (32) is based on the observation that machine learning models are trained to minimize the loss of their training examples. Thus, examples with lower loss are, on average, more likely to be members of the training data. Mathematically, it is defined as follows:

$$A_{\text{loss}}(x, y) = \mathbb{1}[-\ell(f(x)_y) > \tau],$$

where $\ell(\cdot)$ is the loss function, τ is a tunable decision threshold parameter, and $A_{\text{loss}}(x, y)$ is the attack model which outputs 1 if the loss is below the threshold τ (indicating membership) and 0 otherwise.

2.5. Fairness evaluations in machine learning

Fairness is a pivotal concern as biased models from training data/algorithms can lead to misleading and even catastrophic consequences, and understanding and mitigating such bias has been an active area of research (15; 3; 5; 24). The precise definitions and measurements of fairness, however, are often application-dependent.

2.6. Fairness evaluations of model fine-tuning

When evaluating the fairness properties of fine-tuning algorithms (20), we argue for the following key desiderata: (1) the fine-tuning task should *not* teach the model to be fair (or else we cannot extrapolate the evaluation to new tasks); (2) there is a "side-channel" through which we can measure fairness (*e.g.*, measuring race bias for gender classification); and (3) the fairness implications are directly relevant to the task being fine-tuned on (so that any observed fairness issues are indicative of realistic harm). We strive to achieve all these desiderata when designing fairness evaluations, though experiments forgoing desideratum (3) may still serve as "probes" and provide useful insights.

3. Data

Face image classification. We use the UTK-Face dataset[1] (34), where each face image (see Fig. 1 for examples) is labeled with gender, age, and specified race of the person. The image is resized to the input dimensions of the base model and normalized before being fed into the model. During training, the images are augmented via random horizontal flips. The training and evaluation split is 80% and 20%, respectively. We consider gender classification (binary) and age classification (9-bins) as the fine-tuning tasks, and race attributes are used as subgroups to evaluate fairness, following (2) and (28).

Figure 1. Example images from UTK-Face dataset. Image adapted from the same dataset website.

[1] https://susanqq.github.io/UTKFace/

4. Methods

For all the three fairness evaluations below, our **baselines** are full-model fine-tuned models (*i.e.,* training all the weight parameters).

4.1. Fairness Evaluations for Accuracy

Classification tasks have well-accepted fairness evaluation methods and metrics. *Subgroup accuracy parity* and *worst subgroup accuracy* (relatedly, best-worst spread) are two metrics commonly studied in prior work (15; 2; 31; 28), which measure differences in accuracy. E.g., are people with different skin colors equally well-classified? Does the subgroup with the worst utility get "poorer" under the ML algorithm? We also consider the one common fairness metric seen in recent work (30; 13). *Equalized odds difference* (EOD) measures if the model has similar predictive performance across both true and false positive rates, regardless of the protected attribute. In scenarios where equitable outcomes are critical, such as the success of medical diagnosis across different demographic groups, the "balanced" EOD may be the most appropriate. See §2.2 for formal definitions.

A *fair* fine-tuning method should produce models that: (1) perform well across all subgroups (*i.e.*, accuracy parity); (2) do not worsen the worst subgroup accuracy; and (3) make errors equally often across subgroups (captured by EOD).

4.2. Fairness Evaluations for Model Calibration

While metrics in §4.1 concentrate on equality in error rates across groups, the measure of *calibration* within groups is another important fairness metric to ensure the probability estimates align with real-world outcomes, both globally and across different subgroups (16). To measure calibration, we extract the model's confidence on the 20% evaluation set by examining the probability outputs from the classification head. We then follow (12) and generate the confidence histograms and the reliability diagrams. See §2.3 for additional background.

4.3. Fairness Evaluations for Membership Inference Attacks

Membership inference attacks (MIA) involve predicting whether an example was in the training set of a target model. They pose risks from data privacy to intellectual property protection and it is useful to understand the impact of fine-tuning on the model's resistance to MIA. One reason to hypothesize that LoRA may exhibit different behaviors than full fine-tuning is its parameter efficiency and thus its decreased capacity to memorize and overfit. Past work showed that overfitting tends to result in higher vulnerability of MIA (4; 32), and minority groups tend to be treated as outliers and thus possibly memorized more often (9).

Motivated by the above hypothesis and relevant observations, we evaluate the resistance of fine-tuned models against MIA to see whether LoRA makes the fine-tuned model more (or less) vulnerable compared to full fine-tuning. In particular, we focus on the Likelihood Ratio Attack (LiRA, (4)) due to its efficacy (See §2.4 for background and implementation of MIA). We attack ViT-Base and Swin-v2-Large fine-tuned with the UTK-Face dataset for binary gender classification. We repeat this for both LoRA and full fine-tuning and compare their resistance to MIA when the training loss is about the same for both methods.

5. Experiments and Results

5.1. Training Settings

We fine-tune on ViT-Base (8) and Swin-v2-Large (22), utilizing the HuggingFace API for model loading and distributed training and the PEFT package for adapting LoRA. However, **we write the entire pipeline (from dataset preprocessing to model predictions) by ourselves**. All models are fine-tuned with a batch size of 32 and a single-cycle cosine learning rate schedule with a warmup ratio of 0.01. We perform a grid search over initial learning rates and the number of fine-tuning epochs and pick the best hyperparameters for each model and fine-tuning method. Specifically, for full-model fine-tuning, we search the learning rate from [0.00001 0.00005 0.0001 0.0003] and the training epoch from [1 2 3 4 6 8]. For LoRA, we search the learning rate from [0.00001 0.00005 0.0001 0.0003] and the training epoch from [2 4 6 8 12]. We believe that LoRA takes longer to train because of its limited number of parameters that can be fine-tuned. LoRA can match full-model fine-tuning in terms of both train/test performances (using default LoRA rank 8), allowing fair comparisons as absolute performance advantage can be a confounding factor in fairness evaluations.

5.2. Accuracy / Utility

Fig. 2 presents results on UTK-Face age and gender classifications across two base vision model architectures. More results can be found in §A.1. There are several interesting observations:

No consistent pattern of LoRA worsening subgroup fairness compared to full fine-tuning. Overarchingly, LoRA and full fine-tuning exhibit similar performance across all subgroups, with the *absolute subgroup performance* and *worst subgroup performance* for LoRA being consistently on par with full fine-tuning.

Fairness implications can depend on the quality of pre-trained model. A closer look at Fig. 2 suggests that while LoRA may be considered *less fair* than full fine-tuning on ViT-Base—by decreased worst subgroup utility on Black group for age classification (upper left subplot) and by increased EOD on Asian group for gender classification (bottom left subplot)—the tendency disappears when the base model is switched to the more powerful Swin-v2-Large (all else kept the same). This is interesting as it suggests that the fairness properties of LoRA are not only a function of its parameter efficiency and they provide a separation from *model pruning* where (28) found that the fairness ramifications persist across model sizes.

It is nonetheless possible to isolate cases where LoRA is less fair, but such cases should be viewed with target metric sensitivity in mind. Another interpretation of the above observation is that one can single out cases where LoRA is less fair than full fine-tuning. We note that different fairness metrics may be more or less relevant depending on the goals and priorities of the task at hand. Take, for example, UTK-Face gender classification where the female category is labeled as 1; for applications where correctly classifying females is important (*e.g.*, when there is drastically fewer data for females than males), the unfairness of LoRA according to accuracy (Fig. 8 in §A.1) may be less relevant than EOD (Fig. 2) which also looks at false positive predictions. There, EOD may very well lead to a different conclusion that LoRA is slightly less fair. In the context of fairness metric sensitivity (16), it is therefore crucial for practitioners to adopt a target-centric perspective (*e.g.,* whether false positives are important) to ensure a meaningful and relevant fairness evaluation.

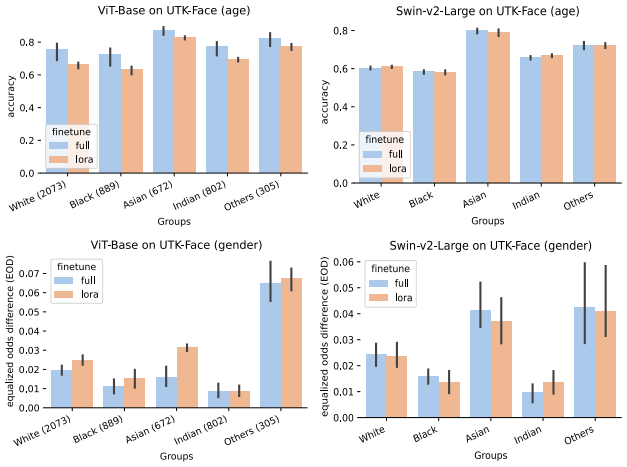

Figure 2. **LoRA vs. full fine-tuning on group-wise accuracy and equalized odds difference (EOD, lower is fairer) on UTK-Face gender and age classification for ViT-Base and Swin-v2-Large.** Error bars: 95% CI across 5 seeds. By all metrics, LoRA may be considered *less fair* than full fine-tuning on ViT-Base but *equally as fair* when switched to a better base model Swin-v2-Large.

5.3. Model Calibration

Fig. 3 presents calibration results on UTK-Face gender classifications for both vision model architectures, with the

worst subgroup performance of the Swin-v2-Large model included on the bottom subplot.

LoRA and full fine-tuning show comparable calibration levels, and both show signs of overconfidence. Fig. 3 shows that both LoRA and full fine-tuning exhibit a reasonable level of calibration, with their expected calibration error (ECE) being relatively low and comparable. The reliability diagrams illustrate that the probabilities predicted by both methods are aligned (but not too well) with the observed accuracies. Neither method consistently yields less calibrated models than the other, and the conclusion holds even when we specifically look at the respective subgroups with highest ECE. One subtle observation is that both fine-tuning methods show a tendency for their predicted probabilities to cluster at the lower and upper ends of the scale, particularly in the 0-0.1 and 0.9-1 confidence bins (top row of Fig. 3). This skewness indicates a degree of overconfidence in their predictions, leading to less reliable decision-making (26) that could potentially affect subgroups disparately.

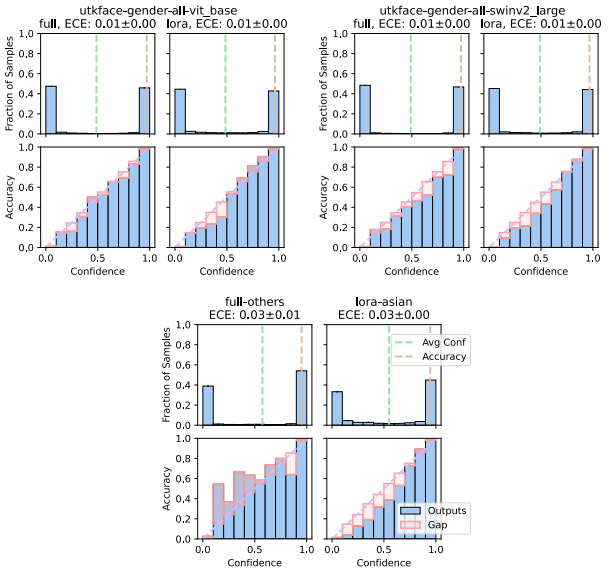

Figure 3. **Confidence histograms (top row of the four subplots) and reliability diagrams (bottom row of the four subplots)** for vit-base on UTK-Face gender classification (upper left), swinv2-large on UTK-Face gender classification (upper right), and swinv2-large on subgroups with highest ECE within different races (bottom). Dotted purple line indicates perfect calibration. Gap is calculated by confidence minus accuracy. A lower ECE is better calibrated.

5.4. Resistance to Membership Inference Attacks (MIA)

Figs. 4 and 5 present membership inference attacks results on UTK-Face gender classifications for both two vision architectures. We also obtain receiver operating characteristic (ROC) curves by varying the confidence thresholds. The lower true-positive rate indicates the model is more resistant to attacks.

LoRA is generally as resistant to MIA as full fine-tuning. Figs. 4 and 5 show the ROC curves in log-scale to emphasize true positive rates at low false positives. We defer results for a simpler MIA attack (LOSS) to §A.2. From Figs. 4 and 5, we see that there is no clear evidence that LoRA makes the model less resistant to MIA compared to full fine-tuning. On Swin-v2-Large with UTK-Face, LoRA is actually more resistant than full fine-tuning *overall* and also at the subgroup level across different races. On ViT-Base with UTK-Face, while LoRA is slightly less resistant *overall*, there are subgroups (*e.g.*, Asian) for which LoRA provides higher resistance than full fine-tuning.

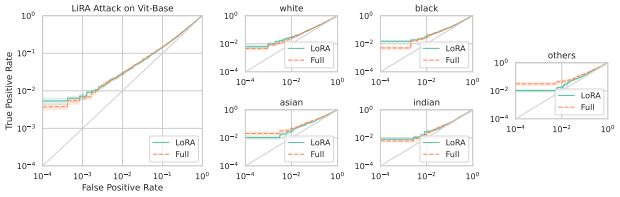

Figure 4. **Likelihood Ratio Attack (LiRA) on ViT-Base for membership inference on UTK-Face gender classification.** Results indicate that the LoRA fine-tuned model is roughly equally (or slightly less) resistant to membership inference compared to full fine-tuning.

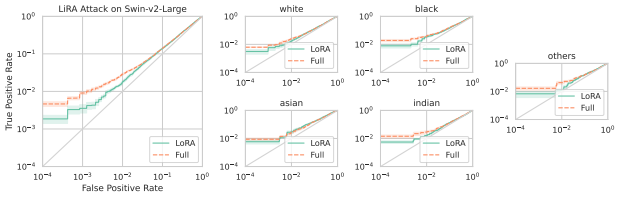

Figure 5. **Likelihood Ratio Attack (LiRA) on Swin-v2-Large for membership inference on UTK-Face gender.** Results indicate that the LoRA fine-tuned model is slightly more resistant to membership inference than full fine-tuning.

5.5. Effect of LoRA Rank

We also explore the choice of rank for LoRA, as it may also be a confounding factor in the model's fitting capacity and fairness impact. Results from UTK-Face gender classification (Fig. 6) reveal that accuracy and fairness metrics (EOD) **are not influenced by rank**, aligning with findings from (14).

5.6. Effect of Subgroup Size

Fig. 7 illustrate the effect of increasing group size on utility (*e.g.*, accuracy in the plots) and fairness. We observe that **these metrics are not solely dependent on the size of the group**:

- On the UTK-Face gender dataset with vision models, while there is a general trend of increasing accuracy with

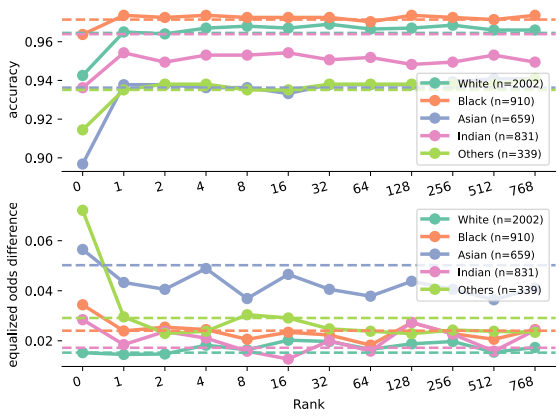

Figure 6. Subgroup accuracy and EOD across of LoRA ranks from 0 to 768 on ViT-Base on UTK-Face gender classification.

larger group sizes, the practical impact of group size on accuracy is limited, since the absolute difference in accuracy across these sizes is marginal.

- The equalized odds difference exhibits fluctuations across different group sizes without showing a clear trend that correlates with group size. It tends to decrease and then increase as the group size increases, and the subgroup with a medium size gets the best EOD score.

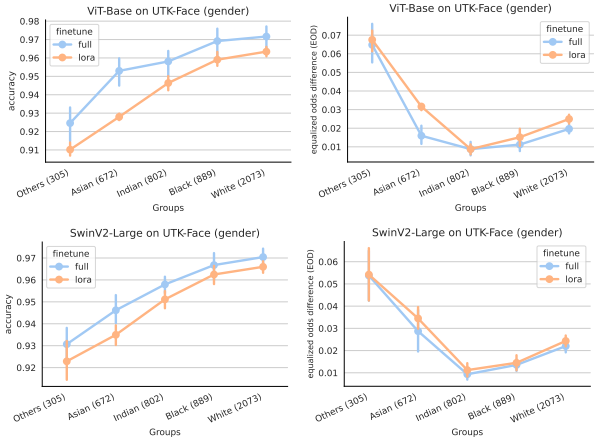

Figure 7. Accuracy and fairness metric (EOD) on UTK-Face gender classification with subgroups sorted by size.

6. Conclusion

We have presented extensive empirical analyses and found *no conclusive evidence* that LoRA may exacerbate subgroup fairness compared to full fine-tuning. Does this imply that the parameter efficiency of LoRA is a free lunch? Possibly, but not necessarily.

Our study sheds light on the fairness properties of low-rank adaptation (LoRA) across vision model architectures, model sizes, and fairness considerations. In future work, we hope to extend fairness evaluations on exploring and comparing other parameter-efficient methods (*e.g.*, (19; 21)) and their intersection with related techniques such as quantization (7; 13) and pruning (6; 11). This may offer insight whether our findings with LoRA is unique to its algorithmic constructions.

7. Contributions and Acknowledgement

Both Zhoujie Ding and Qianzhong Chen worked on developing the HuggingFace training pipeline and writing the final report. Further, Qianzhong Chen worked on fairness evaluations of the accuracy section, and Zhoujie Ding worked on fairness evaluations of the calibration and MIA sections. The other sections are co-authored.

We thank Ken Ziyu Liu (kzliu@cs.stanford.edu) and Professor Sanmi Koyejo (sanmi@cs.stanford.edu) for the initial discussions of the research idea. Ken also helped with the initial HuggingFace training pipeline development (specifically, peer programming and code review for fairness evaluations of *large language models* with LoRA).

References

[1] P. Awasthi, H. Jain, A. S. Rawat, and A. Vijayaraghavan. Adversarial robustness via robust low rank representations. *Advances in Neural Information Processing Systems*, 33:11391–11403, 2020.

[2] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.

[3] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.

[4] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE, 2022.

[5] A. Chouldechova and A. Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[6] L. Dery, S. Kolawole, J.-F. Kagey, V. Smith, G. Neubig, and A. Talwalkar. Everybody prune now: Structured pruning of llms with only forward passes. *arXiv preprint arXiv:2402.05406*, 2024.

[7] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[9] V. Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, 2020.

[10] R. Gandikota, J. Materzynska, T. Zhou, A. Torralba, and

D. Bau. Concept sliders: Lora adaptors for precise control in diffusion models. *arXiv preprint arXiv:2311.12092*, 2023.
[11] A. Gromov, K. Tirumala, H. Shapourian, P. Glorioso, and D. A. Roberts. The unreasonable ineffectiveness of the deeper layers, 2024.
[12] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks, 2017.
[13] J. Hong, J. Duan, C. Zhang, Z. Li, C. Xie, K. Lieberman, J. Diffenderfer, B. Bartoldson, A. Jaiswal, K. Xu, et al. Decoding compressed trust: Scrutinizing the trustworthiness of efficient llms under compression. *arXiv preprint arXiv:2403.15447*, 2024.
[14] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.
[15] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pages 2564–2572. PMLR, 2018.
[16] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.
[17] P. Langenberg, E. R. Balda, A. Behboodi, and R. Mathar. On the effect of low-rank weights on adversarial robustness of neural networks. *arXiv preprint arXiv:1901.10371*, 2019.
[18] B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021.
[19] X. L. Li and P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 2021.
[20] Y. Li, M. Du, R. Song, X. Wang, and Y. Wang. A survey on fairness in large language models, 2024.
[21] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. A. Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.
[22] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.
[23] L. Loukas, I. Stogiannidis, O. Diamantopoulos, P. Malakasiotis, and S. Vassos. Making llms worth every penny: Resource-limited text classification in banking. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 392–400, 2023.
[24] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.
[25] M. P. Naeini, G. F. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2901–2907. AAAI Press, 2015.
[26] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 625–632, New York, NY, USA, 2005. Association for Computing Machinery.
[27] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
[28] C. Tran, F. Fioretto, J.-E. Kim, and R. Naidu. Pruning has a disparate impact on model accuracy. *Advances in Neural Information Processing Systems*, 35:17652–17664, 2022.
[29] H. Tran. How to fine-tune large language models for enterprise use cases. `https://snorkel.ai/how-to-fine-tune-large-language-models-for-enterprise-use-cases/`, November 2023. Accessed: 2024-03-21.
[30] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.
[31] F. Yang, M. Cisse, and S. Koyejo. Fairness with overlapping groups; a probabilistic perspective. *Advances in neural information processing systems*, 33:4067–4078, 2020.
[32] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.
[33] P. Yu, H. Xu, X. Hu, and C. Deng. Leveraging generative ai and large language models: A comprehensive roadmap for healthcare integration. *Healthcare*, 11(20), 2023.
[34] Z. Zhang, Y. Song, and H. Qi. Age progression regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

A. Additional Results

A.1. Accuracy / Utility

Fig. 8 shows the results for UTK-Face gender and age classification for ViT-Base and Swin-v2-Large with subgroup F1 score, accuracy, and equalized odds difference (EOD) for each subset of the dataset. Note that for age classification, we only report accuracy since other metrics might not be well-defined for this multi-class (more than two classes) classification.

The results are consistent with the main results described in §5.2:

- By worst group performance, best-worst group performance spread, and equalized odds difference (EOD), in most cases, LoRA does not worsen these fairness metrics.
- The fairness assessment of the fine-tuning methods can be sensitive to the choice of the metrics and the base model (ViT-Base or Swin-v2-Large).

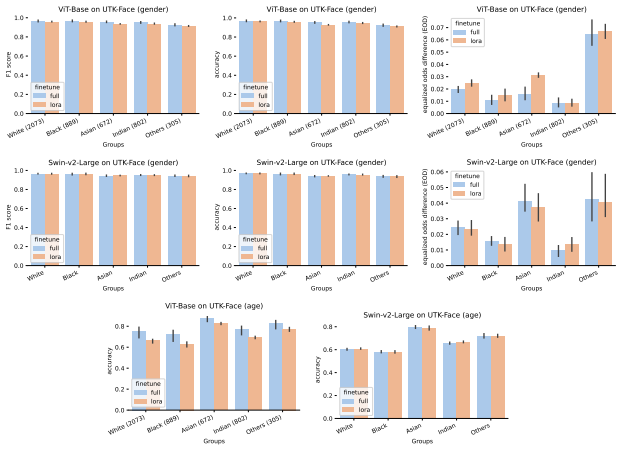

Figure 8. **Classification fine-tuning results for ViT-Base and Swin-v2-Large on UTK-Face** (gender and age classification). *Top row*: ViT-Base on gender classification; metrics are subgroup F1 score, accuracy, and EOD. *Middle row*: Swin-v2-Large on gender classification with the same metrics. *Bottom row*: Subgroup accuracy of ViT-Base and Swin-v2-Large on age classification. See §4.1 and §5.2 for more details.

A.2. Resistance to Membership Inference Attacks (MIAs)

We present the LOSS attack results in Figs. 9 and 10. See §2.4 for background on the definitions and implementations of membership inference attacks.

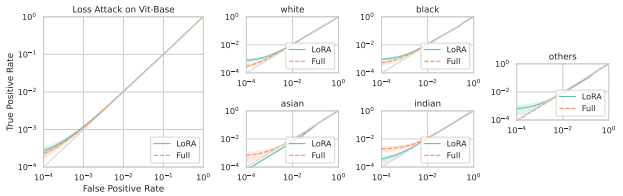

Figure 9. **LOSS attack on ViT-Base for membership inference on UTK-Face gender classification.** Results indicate that the LoRA fine-tuned model is roughly equally resistant to membership inference compared to full fine-tuning.

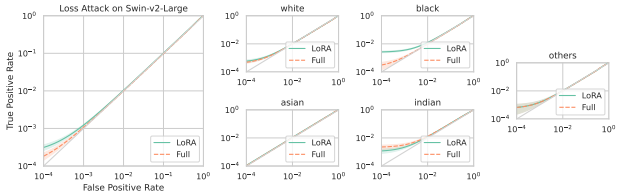

Figure 10. **LOSS attack on Swin-v2-Large for membership inference on UTK-Face gender classification.** Results indicate that the LoRA fine-tuned model is roughly equally resistant to membership inference compared to full fine-tuning.