

Bias-Aware Machine Unlearning: Towards Fairer Vision Models via Controllable Forgetting

SSC Aylapuram[†], Veeraraju Elluru^{*}, Shivang Agarwal[†]
ICCV 2025

[†]BITS Pilani Dubai Campus, ^{*}IIT Jodhpur



Introduction

Problem Setup

- We consider a supervised classification setup:

$$f_{\theta} : \mathcal{X} \rightarrow \mathcal{Y}, \quad D_{\text{train}} = \{(x_i, y_i)\}_{i=1}^N$$

- Each image decomposed as

$$x = s + b,$$

where s is the semantic signal and b is the spurious / biased component [1].

- A **biased** thereby model exhibits:

$$\left\| \frac{\partial f_{\theta}(x)}{\partial b} \right\| \gg \left\| \frac{\partial f_{\theta}(x)}{\partial s} \right\|$$

Problem Formulation

- Desideratum: Bias Aware Machine Unlearning via a post-hoc update of model parameters θ to remove bias without retraining:

$$f_{\theta'} \approx f_{\theta^*}, \quad \theta^* = \arg \min_{\theta} \mathcal{L}(f_{\theta}, D_{\text{train}} \setminus D_f)$$

- When we consider modern systems with pre-deployed model, post-hoc machine unlearning is especially crucial since re-training from scratch is impractical.
- We evaluate the effectiveness of Machine Unlearning (MU) to perform targeted removal of bias contributions via posthoc parameter updates.

- We explore the following bias mitigation settings:
 - Synthetic pose bias in CUB-200-2011¹.
 - Synthetic patch bias in CIFAR10.
 - Gender-Smiling inter-attribute bias in CelebA.
- Empirical Evaluations:
 1. Qualitative and Quantitative demonstrations across five popular MU strategies - Exact Unlearning [2], Gradient Ascent [3], Teacher-Student Distillation method SCRUB [4], Fast Model Debiasing [5], and Parameter Efficient Techniques like LoRA.
 2. Comparison on a unified evaluation metric, termed as **Co-BUM**, short for Concerted-Bias and Unlearning Metric.
 3. Qualitative demonstrations using Grad-CAM.
 4. Large fairness gains up to **94.86% (CUB)** and **97.37% (CelebA)** in Demographic Parity substantiating effective debiasing.

¹Caltech-UCSD Birds-200-2011 (CUB-200-2011)

Review

Bias Mitigation and Machine Unlearning

- **Classical bias mitigation techniques**

- are typically of three types
 - Pre-processing: rebalancing, augmentation
 - In-processing: adversarial or fairness-constrained optimization
 - Post-processing: score calibration
- tend to require access to the entire training pipelines, may involve retraining from scratch [6, 7]

- **Machine Unlearning (MU):**

- Originally sparked interests for privacy and regulatory removal.
- Recent works extend their utilization to debiasing and improving fairness: Fast Model Debiasing [5] and Fair Machine Unlearning [6] to name a few.
- Since MU seeks to remove the influence of specific training examples from a trained model without full retraining, researchers can now adapt this concept to comprehensively evaluate fairness by selectively unlearning spurious feature dependencies in trained models.

Methods

Machine Unlearning Strategies

We evaluate bias mitigation on four modern post-hoc MU methods against the gold standard:

1. **Hard (Exact) Unlearning:** Retrain on $D_r = D \setminus D_f$. Serves as the gold standard.

$$\theta^* = \arg \min_{\theta} \mathcal{L}(f_{\theta}, D_r)$$

2. **Gradient Ascent (GA):** Fine-tuning update step as prescribed in NegMargin is given by

$$\theta_{t+1} \leftarrow \theta_t + \eta \nabla_{\theta_t} (\mathcal{L}(D_f; \theta_t) - \alpha \mathcal{L}(D_r; \theta_t))$$

3. **LoRA Fine-Tuning:** Inject low-rank updates:

$$\Delta W = AB, \quad W' = W + \Delta W, \quad A \in \mathbb{R}^{d \times r}, \quad B \in \mathbb{R}^{r \times k}, \quad r \ll d.$$

For unlearning, we freeze the original W and optimize only (A, B) to maximize the loss on D_f and minimize it on D_r :

$$\min_{A, B} \mathcal{L}(D_r; W') - \beta \mathcal{L}(D_f; W'),$$

where $\beta > 0$ scales the unlearning strength.

4. **SCRUB (Teacher-Student Distillation):** Distill a student model from a fairness-corrected teacher via the objective given by

$$\mathcal{L}_{\text{SCRUB}} = \mathcal{L}_{\text{task}}(f_S(x_r), y_r) + \lambda \text{KL}(p_T(x_r) \parallel p_S(x_r)) - \beta \text{KL}(p_T(x_f) \parallel p_S(x_f)),$$

where p_T, p_S are teacher and student distributions. The fine-tuning update step is now,

$$\nabla_S \mathcal{L}_{\text{SCRUB}} = \nabla_S \mathcal{L}_r + \lambda \nabla_S \text{KL}_r - \beta \nabla_S \text{KL}_f,$$

5. **Fast Model Debiasing via Counterfactual Dataset:** Fine-tune on a curated counterfactual set $D_c = \{(x_c, y_c)\}$ to remove biased influence. In this case, we utilize Influence functions to approximate the influence of training sample (x_i, y_i) w.r.t. bias measure $B(\theta)$, given by $I(x_i) \approx -\nabla_{\theta} \ell(x_i, y_i)^{\top} H_{\theta}^{-1} \nabla_{\theta} B(\theta)$, where H_{θ} is empirical risk Hessian. This ranks samples by their contribution to the bias. Hence, the parameter update step reads

$$\theta' = \theta - H_{\theta}^{-1} \left(\frac{1}{|D_c|} \sum_{(x,y) \in D_c} \nabla_{\theta} \ell(x, y) \right).$$

Experiments

Datasets and Bias Definitions

We use three datasets each exemplifying a different bias type.

- **CUB-200-2011 (Synthetic Pose bias).**

- **Pose proxy:** We construct the bias using normalized bounding-box area quantiles to yield 3 pose bins {0: close, 1: mid, 2: distant}, using simple augmentations - `torchvision.transforms.functional.crop(img, height, width)`.
- **Partition:** $D_f :=$ samples in pose bin 2 (distant / small box); $D_r :=$ all other samples.

- **CIFAR-10 (Synthetic class-specific patch bias).**

- **Bias injection:** We overlay a red patch at top-left on 50% of training samples of the birds.
- **Partition:** $D_f :=$ patched bird samples; $D_r :=$ remaining dataset.

- **CelebA (Inter-attribute bias: Gender-Smiling).**

- **Bias:** Wang et al.[8] report an average skew in gender predictions of 80% when the attribute 'smiling' is present
- **Partition:** $D_f :=$ samples exhibiting the spurious pairing of female and smiling; $D_r :=$ male, and non-smiling females.

Evaluation Protocol

- We evaluate on a wide range of metrics including Forget and Retain accuracies, Demographic Parity (DP) and Equalized Odds (EO), Membership Inference Attack (MIA) AUC, and total clock time.
- To obtain more tangible results, we also perform a joint evaluation using a unified metric, **Co-BUM**. It's a simple harmonic-mean of U (utility), F (fairness), Q (quality), P (privacy), and E (efficiency) metrics, given by

$$\text{Co-BUM} := \kappa \left(\sum_{i \in \mathcal{S}} \alpha_i \right) / \left(\sum_{i \in \mathcal{S}} \frac{\alpha_i}{i} \right), \text{ where } \mathcal{S} = \{U, F, P, Q, E\}$$

- We also perform qualitative analysis by utilizing Grad-CAM visuals to understand effectiveness of unlearning the bias.

Results and Analyses

Quantitative Results

Table 2. Unlearning Results across CUB-200, CIFAR-10, and CelebA datasets.

*FA: Forget Accuracy, †R.A.: Retain Accuracy, ‡TA: Test Accuracy, †DP: Demographic Parity, †EO: Equalized Odds, †MIA: ROC-AUC of Membership Inference Attack. All fairness scores (DP and EO) are reported as % drops w.r.t baseline model performance. Boldface indicates best performance and underline, the second-best. Runtime comparisons are only done amongst the unlearning methods and not w.r.t the Baseline

Method	FA* (↓)	RA† (↑)	TA‡ (↑)	DP† (%) (↑)	EO† (%) (↑)	MIA† (↓)	Time (s) (↓)	Co-BUM (↑)
CUB-200 (pose bias)								
Baseline	80.67	81.33	78.69	0.00	0.00	0.56	519	–
Hard Unlearning	17.67	72.33	63.03	94.51	51.74	0.48	222	–
Gradient Ascent	37	74.67	<u>64.41</u>	<u>93.75</u>	<u>41.76</u>	0.48	299	0.71
LoRA	77.83	80.33	57.47	88.53	-11.14	0.52	233	0.11
SCRUB	67.67	<u>79.17</u>	45.72	91.92	8.53	<u>0.51</u>	<u>237</u>	0.38
FMD	<u>66.83</u>	69.83	65.13	94.86	46.61	0.48	249	<u>0.52</u>
CIFAR-10 (synthetic bias)								
Baseline	99.80	87.40	85.50	0.00	0.00	0.60	552	–
Hard Unlearning	14.10	83.10	75.70	83.28	83.28	0.56	210	–
Gradient Ascent	45.30	<u>85.80</u>	78.30	-9.46	-9.46	0.59	<u>266</u>	0.39
LoRA	<u>48.70</u>	84.80	79.00	30.28	30.28	0.58	198	0.50
SCRUB	45.30	86.30	84.70	<u>22.08</u>	<u>22.08</u>	<u>0.59</u>	323	<u>0.46</u>
FMD	74.70	85.80	<u>80.40</u>	-34.7	-34.7	0.59	296	0.38
CelebA (inter-attribute bias)								
Baseline	95.95	96.13	94.20	0.00	0.00	0.67	1582	–
Hard Unlearning	87.27	95.56	92.33	-29.62	1.5	0.51	1310	–
Gradient Ascent	3.39	58.31	50.90	97.37	98.13	<u>0.51</u>	80	0.77
LoRA	<u>70.52</u>	<u>87.81</u>	<u>88.25</u>	-11.84	<u>18.77</u>	0.47	1024	<u>0.54</u>
SCRUB	79.24	93.89	91.01	<u>57.6</u>	-71.05	0.53	<u>741</u>	0.42

Qualitative Results

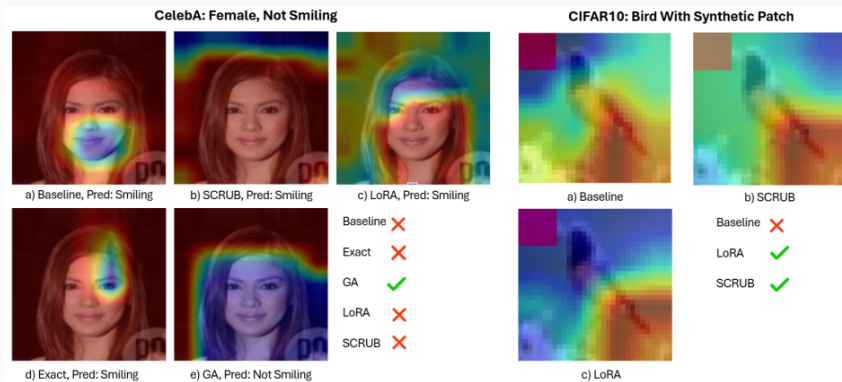


Figure 1: Grad-CAM visuals in two representative settings.

Left: Samples containing the sensitive attribute pair (female-“smiling”) from the CelebA dataset (ID=055153).

Right: In CIFAR-10, while the baseline method attends heavily to the red-patch, the LoRA and SCRUB methods use features from the whole image.





Conclusion

- **Conclusion:**

- We motivate researchers to firstly identify the bias typology - whether localized, distributed, or entrenched
- Every MU-based bias mitigation strategy have their own benefits and one-method-fits-all has not yet been found.
- Establishing the size, sign, and significance of the bias is important.

- **Future Works:**

- We motivate readers to firstly focus on developing adaptive unlearning methods that adjust to the topology of bias, whether distributed, localized, or entrenched, and that explicitly integrate fairness and privacy considerations.
- Leverage concept salience, mechanistic interpretability, and group disparity diagnostics to guide unlearning, as well as extending to multi-bias disentanglement.
- Detailed studies on designing scalable, fairness-aware training pipelines to ensure reliability, transparency, and applicability in real world, privacy-sensitive settings.

-  S. Yenamandra, P. Ramesh, V. Prabhu, and J. Hoffman, “Facts: First amplify correlations and then slice to discover bias,” 2023.
-  L. Bourtoutle, V. Chandrasekaran *et al.*, “Machine unlearning,” in *IEEE Symposium on Security and Privacy*, 2021.
-  A. Thudi, G. Deza, V. Chandrasekaran, and N. Papernot, “Unrolling SGD: understanding factors influencing machine unlearning,” *CoRR*, vol. abs/2109.13398, 2021. [Online]. Available: <https://arxiv.org/abs/2109.13398>
-  M. Kurmanji, P. Triantafillou, J. Hayes, and E. Triantafillou, “Towards unbounded machine unlearning,” in *NeurIPS*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.

-  R. Chen, J. Yang, H. Xiong, J. Bai, T. Hu, J. Hao, Y. Feng, J. T. Zhou, J. Wu, and Z. Liu, “Fast model debias with machine unlearning,” in *NeurIPS*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.
-  A. Oesterling, J. Ma, F. Calmon, and H. Lakkaraju, “Fair machine unlearning: Data removal while mitigating disparities,” in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024.
-  B. H. Zhang, B. Lemoine, and M. Mitchell, “Mitigating unwanted biases with adversarial learning,” in *AIES*, J. Furman, G. E. Marchant, H. Price, and F. Rossi, Eds. ACM, 2018.
-  Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky, “Towards fairness in visual recognition: Effective strategies for bias mitigation,” in *CVPR*, 2020.

Thank You!