Security and Privacy of Machine Learning, 2025 Critique G8: Model Immunization – (1) Multi-concept Model Immunization through Differentiable Model Merging (2) Model Immunization from a Condition Number Perspective

Shih-Yu Lai
National Taiwan University
Taipei, Taiwan
akinesia112@gmail.com

Critique:(1) Multi-concept Model Immunization through Differentiable Model Merging

Summary

The paper extends the model immunization paradigm from single-concept protection (IMMA) to a realistic multi-concept setting by proposing **MIMA**, which meta-learns a "difficult initialization" that resists adaptation on multiple target concepts simultaneously. The core technical idea is to treat *model merging* as a differentiable layer: first unroll a (single-step) lower-level update per concept to obtain per-concept weights, then merge them via a constrained optimization (on cross-attention key/value projections) while averaging the rest, and finally backpropagate through this merge to update the immunized model. MIMA is evaluated on (i) re-learning of erased styles/objects (using UCE-erased backbones) and (ii) personalized subject learning, under several adaptation methods (DreamBooth, LoRA, Custom Diffusion, etc.), with new metrics (MSGR, MRSGR). Across two- and three-concept settings, MIMA consistently outperforms strong baselines (joint training and compose/merge-only). The work builds on prior adaptation and customization methods such as DreamBooth [1], LoRA [2], Custom Diffusion [3], the erasure backbone UCE [4], and the single-concept immunization IMMA [5].

Strengths

- **Problem realism.** Moves from single- to multi-concept protection, which matches true deployment risks (multiple harmful or restricted concepts).
- **Clean bi-level formulation.** Casting immunization as meta-learning with a differentiable merge layer is elegant and enables end-to-end gradients without fixing a particular attacker (claims to be adaptation-agnostic in the upper level).
- **Differentiable merging insight.** Turning cross-attention KV merging into an optimization layer connects to a growing literature on differentiable solvers, and the linear-system exposition clarifies gradients and complexity.
- **Broad empirical sweep.** Evaluations cover re-learning (after UCE erasure) and personalization; multiple adaptation methods are tested; metrics capture both protection (MSGR) and retained adaptability (MRSGR).
- **Simple, practical recipe.** One-step unrolling plus analytic merging offers a reasonable compute/engineering trade-off and appears stable in practice.

Weaknesses

- **Merge scope limitations.** Merging only KV projections and averaging all other weights may be suboptimal; distribution shifts can propagate through non-attention parameters, and simple averaging can blur useful specialization.
- **Attacker/model assumptions.** While positioned as not requiring the adaptation method in the upper level, the lower-level unroll implicitly *chooses* an update style/learning rate. Stronger attackers (full-model fine-tuning, multi-stage curricula, ControlNet-style conditioning) are not stress-tested.
- **Metric dependence and external validity.** Protection is largely evidenced by CLIP/DINO/LPIPS similarity gaps; these may not perfectly correlate with semantic safety or policy-violating content. Limited human or task-centric safety assessments.
- **Robustness of the optimization layer.** The merge relies on solving a linear system with $Q = C_{\text{reg}}^{\top} C_{\text{reg}}$. Conditioning, invertibility, and sensitivity to choice/coverage of C_{reg} are not deeply analyzed.
- **Generalization across backbones.** Results center on SD v1-4/related pipelines and UCE. No evidence on SDXL, SD3, or other latent backbones; cross-version transferability is unclear.

Potential Improvements / Extensions

- **Beyond KV-only merging.** Learn sparse or low-rank *mask-and-merge* over broader UNet blocks; compare against learned averaging (e.g., FiLM- or gating-based fusion) to reduce averaging-induced drift.
- **Stronger bilevel fidelity.** Explore multi-step unrolling, implicit differentiation, or truncated-Neumann estimators to better approximate attacker dynamics without exploding cost.
- **Adversary diversification.** Include full-model finetuning, DreamBooth+prior preservation variants, ControlNet/ID adapters, and instruction-tuned schedulers to probe worst-case adaptation.
- **Safety-grounded evaluation.** Add content-policy detectors and small human studies; report false-positive/negative rates and attack success under prompt obfuscation.
- **Theoretical guarantees.** Analyze conditions under which gradients through merging provably increase adaptation loss across concept sets; study trade-offs (protection vs. adaptability) with Pareto fronts.
- **Scalability tests.** Stress-test with 5–10 protected concepts, varying inter-concept correlation (orthogonal vs. overlapping styles), and ablate $|C_{\text{reg}}|$ coverage.

Questions for the Authors

1) How sensitive is performance to the number/diversity of regularization concepts and to the conditioning of Q? Any safeguards (e.g., Tikhonov or low-rank preconditioners)?
2) Does KV-only optimization implicitly assume text-conditioning is the main locus of concept encoding? What happens if we target down/up blocks or LoRA adapters during merging?
3) How much does the single-step lower-level unroll bias the learned initialization toward weak attackers? Do 3–5 steps materially change MSGR/MRSGR or cost?
4) Can MIMA overfit to the sampled prompt templates used to assemble C and C_{reg}? Any evidence of prompt-distribution shift robustness?
5) Is there measurable degradation on *unrelated* capabilities (e.g., photorealism, compositionality) beyond the reported MRSGR? Any user studies?
6) How does MIMA interact with later safety fine-tuning (e.g., classifier-free guidance schedules, negative prompts, safety checkers)? Synergies or conflicts?

Critique: (2) Model Immunization from a Condition Number Perspective

Summary

This paper reframes "model immunization"—making models resistant to harmful fine-tuning while preserving benign utility—through the lens of the Hessian condition number. The core idea is to *increase* the condition number for a designated harmful task while *not* increasing (ideally decreasing) it for the pre-training/benign task. Concretely, the authors define an evaluation metric (RIR) as a ratio of condition numbers across harmful vs. benign tasks, and propose two differentiable regularizers: one that monotonically *decreases* κ (κ-minimizer) and a novel one that monotonically *increases* κ (κ-maximizer). These are integrated into a gradient-based algorithm that updates the feature extractor to worsen optimization for the harmful task while maintaining optimization properties for the benign task. Experiments cover linear regression/MNIST setups and extend to deep nets (ResNet-18, ViT) with linear probing on ImageNet features; results show large RIR improvements and largely preserved ImageNet accuracy post-immunization. The condition-number view ties directly to first-order convergence theory [6] and is positioned as a principled alternative/complement to prior immunization approaches such as IMMA [7].

Strengths

- **Clear, optimization-theoretic framing.** Casting immunization as a controlled manipulation of Hessian spectra is crisp and connects to classical convergence guarantees [6]. The RIR metric operationalizes this link and offers a diagnostic that is easy to compute from mini-batches for deep models.
- **New κ-maximizing regularizer with monotonicity.** The paper introduces a differentiable regularizer with a provable *monotone* increase in condition number under gradient descent, pairing naturally with an existing κ-minimizer. This yields a controllable two-handled mechanism: ill-condition the harmful task while well-conditioning the benign task.
- **Algorithmic simplicity.** The proposed updates require only first-order machinery and covariance-like statistics; no bilevel differentiation or unrolled inner loops are needed (contrast with IMMA [7]).
- **Breadth of evaluation.** The paper evaluates on linear regression and MNIST (systematically over 90 binary pairs), then scales to ImageNet features for ResNet-18 and ViT. The deep-net results suggest the perspective transfers beyond linear theory.

Weaknesses

- **Attacker model and optimizer assumptions.** The defense effectiveness is argued via first-order convergence speed and condition numbers. Adaptive optimizers (Adam/Adagrad), curvature-aware methods, or preconditioned fine-tuning could blunt the intended slowdown, potentially restoring attacker convergence even with large κ (the paper does not study attacker-side preconditioning or second-order methods).
- **Dependence on spectral *alignment*.** The analysis hinges on angles between singular vectors of harmful vs. benign covariances. In realistic, high-dimensional settings with nonstationary data, those alignments may be unstable, dataset-dependent, or actively gamed by attackers choosing D_H to avoid the ill-conditioned subspaces.

- **Uniqueness and spectral-gap caveats.** Several guarantees assume unique extremal singular values. Near-multiplicity can make gradients noisy and the monotonic effects fragile, especially under minibatch estimates of Hessians/Gram matrices.
- **Metrics vs. mission outcomes.** RIR is a principled proxy for optimization difficulty, but downstream safety hinges on *capability suppression* (e.g., actual harmful generations or classifier success). While the paper reports accuracy/conditioning, it would be stronger with end-task safety metrics and attacker success rates under varied training budgets.
- **Cost of curvature control.** Although first-order, the approach still requires repeated spectral surrogates (Gram/Hessian approximations) and extra gradient terms; the compute/memory overhead and stability vs. scale are not fully characterized.

Potential Improvements / Extensions

- **Robust-to-optimizer adversary.** Evaluate attackers using (i) strong preconditioning, (ii) adaptive optimizers with tuned schedules, and (iii) low-rank/NTK-style second-order approximations; measure whether immunization persists.
- **Multi-harm concept sets.** Extend the objective to optimize against a *family* of likely harmful tasks (worst-case over a set or distribution), akin to distributionally robust optimization. This would reduce sensitivity to any single D_H.
- **Spectral geometry diagnostics.** Report empirical principal-angle distributions between K_P and K_H subspaces before/after immunization; tie successes/failures to these observed alignments.
- **Optimizer-aware regularization.** Co-design regularizers with attacker models (e.g., assume Adam or K-FAC) and enforce condition-number effects measured under those optimizers' implicit preconditioning.
- **Safety-facing evaluations.** Where possible, incorporate concrete misuse tasks (e.g., harmful concept adapters) and measure attacker fine-tuning success under matched budgets, not only RIR and accuracy.

Questions

- How resilient is the immunization to *attacker preconditioning*? If the attacker whitens features or uses K-FAC/second-order updates, does RIR still predict slower convergence, and by how much?
- Can the monotone κ-control be made *local* to layers/blocks that most affect D_H while provably preserving D_P utility?
- What is the sensitivity of RIR to mini-batch Hessian approximations and to the choice of batch statistics? Are there variance-reduction strategies that preserve the intended monotonic effects?
- Could one design a *multi-objective* scheduler that adapts (λ_H, λ_P) on-the-fly using dual control (e.g., keep $\kappa(H_P)$ within a corridor while pushing $\kappa(H_H)$ upward)?
- For non-linear models, can we relate the method to NTK spectra or layerwise Fisher information, yielding theory beyond linear probing?

References

[1] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[2] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.

[3] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-concept customization of text-to-image diffusion," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[4] R. Gandikota, H. Orgad, Y. Belinkov, J. Materzyńska, and D. Bau, "Unified concept editing in diffusion models," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[5] A. Y. Zheng and R. A. Yeh, "IMMA: Immunizing text-to-image models against malicious adaptation," in *European Conference on Computer Vision (ECCV)*, 2024.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[7] A. Y. Zheng and R. A. Yeh, "Imma: Immunizing text-to-image models against malicious adaptation," *arXiv preprint*, 2024.