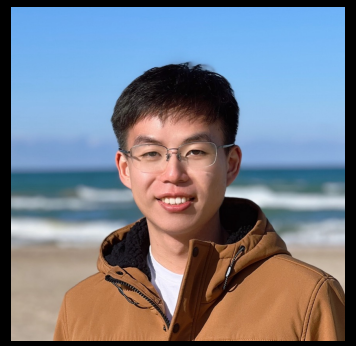
Model Immunization from a Condition Number Perspective

ICML 2025

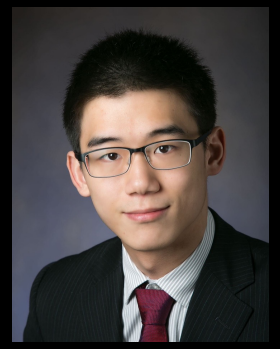
Amber Yijia Zheng*

Cedar Site Bai*



Brian Bullins

Raymond A. Yeh



PURDUE UNIVERSITY | Department of Computer Science

OPEN-WEIGHT GENERATIVE AI

Stable Diffusion

Deep Floyd IF

Content generation just a few prompts away!

RESPONSIBILITY

RISKS OF MISUSE

X blocks searches for Taylor Swift after explicit AI images of her go viral

28th January 2024, 12:42 EST

Share

By Nadine Yousif
BBC News

Social media platform X has blocked searches for Taylor Swift after explicit AI-generated images of the singer began circulating on the site.

The New York Times

People Love Studio Ghibli. But Should They Be Able to Recreate It?

An update to ChatGPT made it easy to simulate Hayao Miyazaki's style of animation, which has flooded social media with memes.

BBC

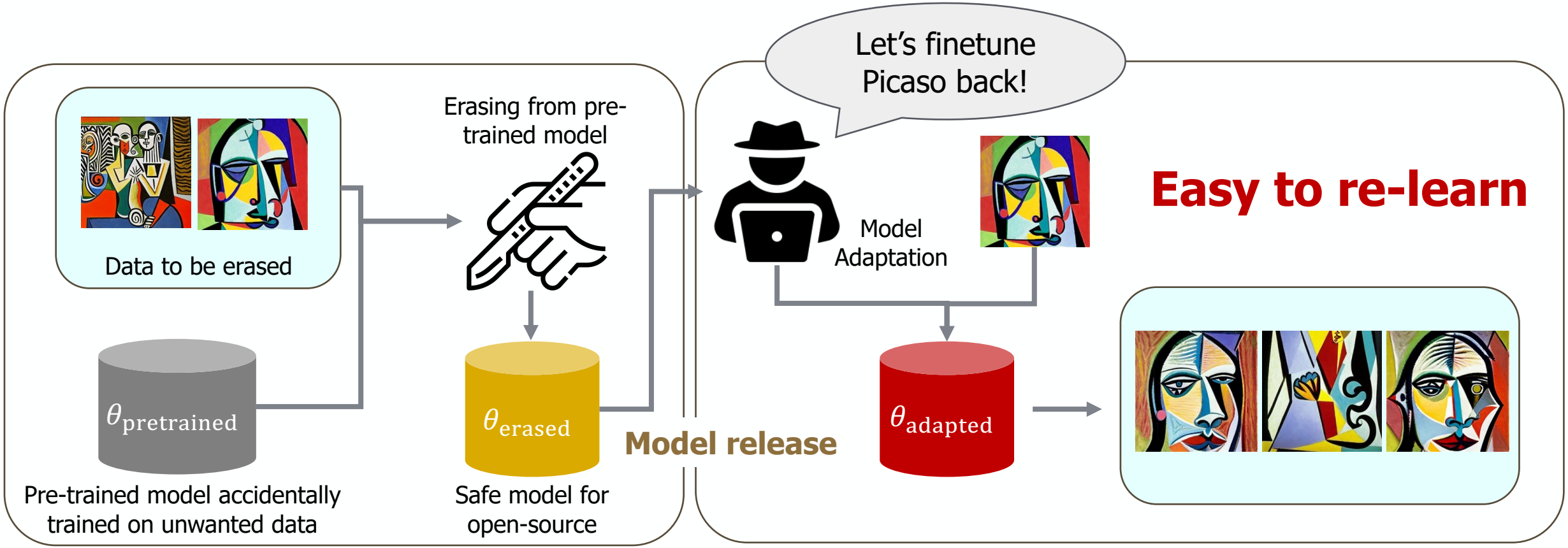
AI can be easily used to make fake election photos - report

This fake image of a man lurking outside a polling place with a gun was created by artificial intelligence tool ChatGPT Plus.

Mar 6, 2024

THREAT MODEL

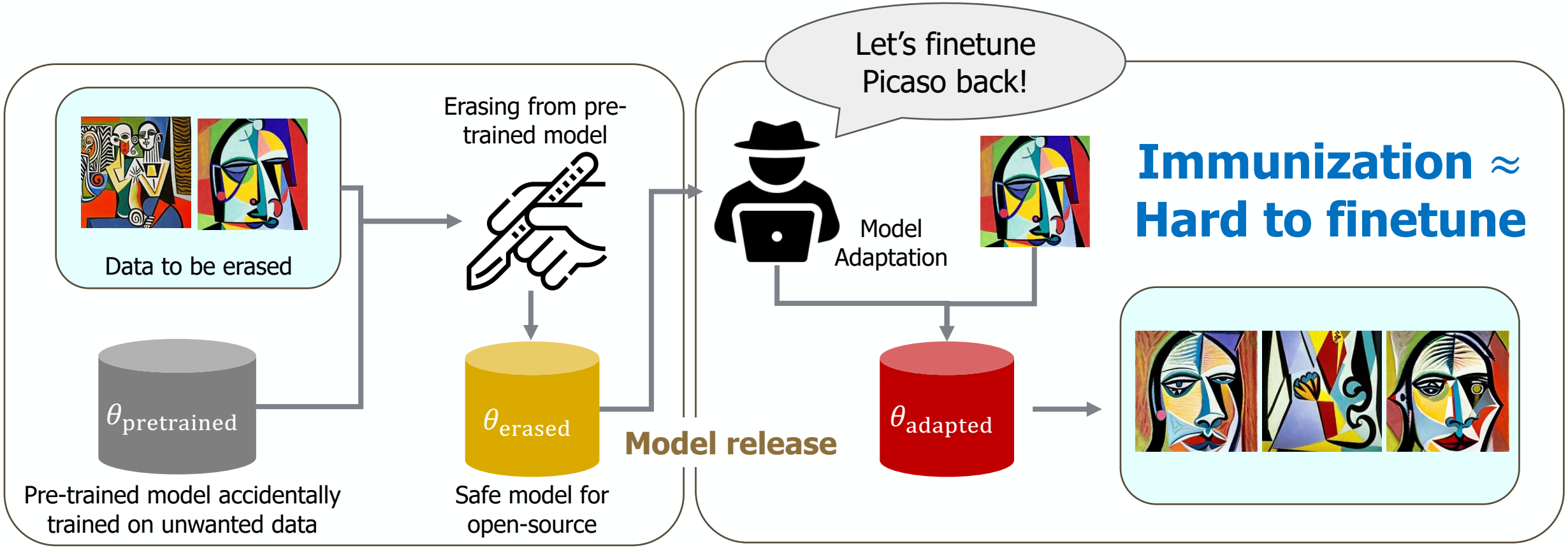
Malicious adaptation of open-weight models



Amber Yijia Zheng, and Raymond A. Yeh. "IMMA: Immunizing text-to-image models against malicious adaptation." *Proc. ECCV 2024*.

THREAT MODEL

Malicious adaptation of open-weight models



Amber Yijia Zheng, and Raymond A. Yeh. "IMMA: Immunizing text-to-image models against malicious adaptation." *Proc. ECCV 2024*.

IMMUNIZATION FRAMEWORK

Goal: learn a pre-trained feature extractor, such that fine-tuning an adapter $h_{\mathbf{w}}$ on features of the harmful task is difficult, but not for other tasks.

Setting: Linear feature extractor $f_\theta \triangleq \mathbf{x}^\top \theta$

- Linear adapter g_ω trained on Pre-training task $\mathcal{D}_{\mathrm{P}} = \{(\mathbf{x}, \mathbf{y})\}$
- Linear adapter $h_{\mathbf{w}}$ trained on Harmful task $\mathcal{D}_{\mathrm{H}} = \{(\mathbf{x}, \tilde{\mathbf{y}})\}$
- Regression task:

$$\mathcal{L}(\mathcal{D}, \mathbf{w}, \theta) \triangleq \sum_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}} \ell_2(h_{\mathbf{w}} \circ f_\theta(\boldsymbol{x}), \boldsymbol{y})$$

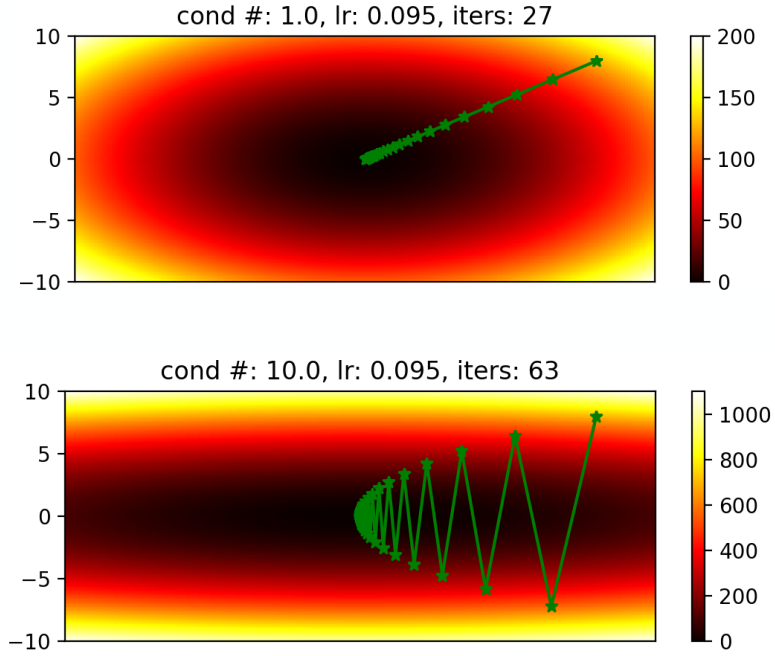
- The bad actor performs linear probing on $\mathcal{D}_{\mathrm{H}}$ by $\min_{\mathbf{w}} \mathcal{L}(\mathcal{D}, \mathbf{w}, \theta)$
  - Only training the last linear layer.

CONDITION NUMBER

Condition number of a general matrix $\boldsymbol{S}$

- The ratio between max / min singular values

$$\kappa(\boldsymbol{S}) \triangleq \|\boldsymbol{S}\|_2 \|\boldsymbol{S}^\dagger\|_2 = \sigma_{\boldsymbol{S}}^{\max} / \sigma_{\boldsymbol{S}}^{\min}$$



For constant step-size steepest descent

- $\min_{\mathbf{w}} \mathcal{L}(\mathbf{w})$, $\mathcal{L}$ is strongly convex
- Hessian $\nabla^2 \mathcal{L}$ with max/min singular values $\sigma^{\max/\min}$
- Convergence rate (Bubeck, 2015) :

$$\|\mathbf{w}_t - \mathbf{w}^*\|^2 \leq \left(1 - \frac{\sigma^{\min}}{\sigma^{\max}}\right)^t \|\mathbf{w}_0 - \mathbf{w}^*\|^2$$

- Large condition number → slower convergence → hard to finetune

A CONDITION NUMBER PERSPECTIVE

Definition of Immunization

(a) It is more difficult to apply linear probing on the harmful task $\mathcal{D}_{\mathtt{H}}$ using the immunized feature extractor $f_{\theta^{\mathtt{I}}}$ than directly on the input data,

$$\kappa(\nabla^2_{\mathbf{w}}\mathcal{L}(\mathcal{D}_{\mathtt{H}}, \mathbf{w}, \theta^{\mathtt{I}})) \gg \kappa(\nabla^2_{\mathbf{w}}\mathcal{L}(\mathcal{D}_{\mathtt{H}}, \mathbf{w}, \boldsymbol{I}))$$

Harmful task — Identity

(b) It is not more difficult to apply linear probing on other tasks. As there is only one other task $\mathcal{D}_{\mathtt{P}}$, an immunized feature extractor should have

$$\kappa(\nabla^2_{\omega}\mathcal{L}(\mathcal{D}_{\mathtt{P}}, \omega, \theta^{\mathtt{I}})) \leq \kappa(\nabla^2_{\omega}\mathcal{L}(\mathcal{D}_{\mathtt{P}}, \omega, \boldsymbol{I})).$$

Pre-training task

(c) The immunized model should maintain a competitive task performance on the pre-training dataset $\mathcal{D}_{\mathtt{P}}$,

$$\min_{\omega,\theta} \mathcal{L}(\mathcal{D}_{\mathtt{P}}, \omega, \theta) \approx \min_{\omega} \mathcal{L}(\mathcal{D}_{\mathtt{P}}, \omega, \theta^{\mathtt{I}}).$$

WHEN IMMUNIZATION IS POSSIBLE

Everything is linear → nice analytical forms

Linear Probing:

$$\mathcal{L}(\mathcal{D}_{\mathtt{H}}, \mathbf{w}, \theta) = \min_{\mathbf{w}} \|(\boldsymbol{X}_{\mathtt{H}}\theta)\mathbf{w} - \boldsymbol{Y}\|_2^2$$

The Hessian Matrix:

$$\boldsymbol{H}_{\mathtt{H}}(\theta) = \nabla_{\mathbf{w}}^2 \mathcal{L}(\mathcal{D}_{\mathtt{H}}, \mathbf{w}, \theta) = \theta^\top \boldsymbol{K}_{\mathtt{H}} \theta$$

$$\text{with } \boldsymbol{K}_{\mathtt{H}} = \boldsymbol{X}_{\mathtt{H}}^\top \boldsymbol{X}_{\mathtt{H}}$$

Singular value is given by:

$$\sigma_i = \sum_{j=1}^{D_{\mathtt{in}}} \left(\sigma_{\theta,i}(\boldsymbol{u}_{\theta,i}^\top \boldsymbol{q}_j)\sqrt{\gamma_j}\right)^2, \quad \forall i \in \{1, \ldots, D^{\mathtt{in}}\}$$

Here, $\sigma_{\theta,i}$ and $\boldsymbol{u}_{\theta,i}$ correspond to the i-th singular value and vector of θ. Next, γ_j and $\boldsymbol{q}_j$ correspond to the j-th singular value and vector of the covariance $\boldsymbol{K}$.

Immunization depends on the **"relative angle"** between singular vectors of the covariance matrix for pre-training/harmful task.

How to train such a feature extractor θ?

OPTIMIZING CONDITION NUMBER

$$\min_{\omega,\theta} \mathcal{R}_{\texttt{ill}}(\boldsymbol{H}_{\texttt{H}}(\theta)) + \mathcal{R}_{\texttt{well}}(\boldsymbol{H}_{\texttt{P}}(\theta)) + \mathcal{L}(\mathcal{D}_{\texttt{P}}, \omega, \theta)$$

$\mathcal{R}_{\texttt{ill}}(\boldsymbol{S}) \triangleq \frac{1}{\frac{1}{2k}\|\boldsymbol{S}\|_F^2 - \frac{1}{2}\sigma_{\boldsymbol{S}}^{\min 2}}$ immunizes the model on harmful task

$$\kappa(\nabla^2_{\mathbf{w}} \mathcal{L}(\mathcal{D}_{\texttt{H}}, \mathbf{w}, \theta^{\texttt{I}})) \gg \kappa(\nabla^2_{\mathbf{w}} \mathcal{L}(\mathcal{D}_{\texttt{H}}, \mathbf{w}, \boldsymbol{I}))$$

$\mathcal{R}_{\texttt{well}}(\boldsymbol{S}) \triangleq \frac{1}{2}\|\boldsymbol{S}\|_2^2 - \frac{1}{2p}\|\boldsymbol{S}\|_F^2$ maintains the fine-tuning ability on other tasks

$$\kappa(\nabla^2_{\omega} \mathcal{L}(\mathcal{D}_{\texttt{P}}, \omega, \theta^{\texttt{I}})) \leq \kappa(\nabla^2_{\omega} \mathcal{L}(\mathcal{D}_{\texttt{P}}, \omega, \boldsymbol{I})).$$

$\mathcal{L}(\mathcal{D}, \omega, \theta) \triangleq \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \ell_2(g_\omega \circ f_\theta(\boldsymbol{x}), \boldsymbol{y})$ maintains the model performance

$$\min_{\omega,\theta} \mathcal{L}(\mathcal{D}_{\texttt{P}}, \omega, \theta) \approx \min_{\omega} \mathcal{L}(\mathcal{D}_{\texttt{P}}, \omega, \theta^{\texttt{I}}).$$

PROPOSED OBJECTIVE IS "NICE"

[Upper Bound] $\frac{1}{\log \kappa(\boldsymbol{S})} \leq (\sigma_{\boldsymbol{S}}^{\texttt{max}})^2 \mathcal{R}_{\texttt{ill}}(\boldsymbol{S})$, i.e., $\mathcal{R}_{\texttt{ill}}(\boldsymbol{S})$ upper bounds $\frac{1}{\log \kappa(\boldsymbol{S})}$ when $\sigma_{\boldsymbol{S}}^{\texttt{max}}$ is reasonably away from ∞.

[Differentiability] If $\sigma_{\boldsymbol{S}}^{\texttt{min}} = \sigma_k < \sigma_i$ for any $i < k$, i.e., $\sigma_{\boldsymbol{S}}^{\texttt{min}}$ is unique, then $\mathcal{R}_{\texttt{ill}}(\boldsymbol{S})$ is differentiable and

$$\nabla_{\boldsymbol{S}} \mathcal{R}_{\texttt{ill}}(\boldsymbol{S}) = \frac{\sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^\top - \frac{1}{k}\boldsymbol{S}}{(\frac{1}{2k}\|\boldsymbol{S}\|_F^2 - \frac{1}{2}(\sigma_{\boldsymbol{S}}^{\texttt{min}})^2)^2}.$$

[Monotonic Increase] If $\sigma_{\boldsymbol{S}}^{\texttt{min}}$ is unique, update $\boldsymbol{S}$ with $\nabla_{\boldsymbol{S}} \mathcal{R}_{\texttt{ill}}(\boldsymbol{S})$ such that $\boldsymbol{S}' = \boldsymbol{S} - \eta_2 \nabla_{\boldsymbol{S}} \mathcal{R}_{\texttt{ill}}(\boldsymbol{S})$ for $0 < \eta_2 < \frac{k}{k-1}(\frac{1}{2k}\|\boldsymbol{S}\|_F^2 - \frac{1}{2}(\sigma_{\boldsymbol{S}}^{\texttt{min}})^2)^2$, then $\kappa(\boldsymbol{S}') > \kappa(\boldsymbol{S})$.

OPTIMIZING CONDITION NUMBER

$$\min_{\omega,\theta} \mathcal{R}_{\texttt{ill}}(\boldsymbol{H}_{\texttt{H}}(\theta)) + \mathcal{R}_{\texttt{well}}(\boldsymbol{H}_{\texttt{P}}(\theta)) + \mathcal{L}(\mathcal{D}_{\texttt{P}}, \omega, \theta)$$

Solve using gradient based method!

Algorithm 1 Condition number regularized gradient descent for model immunization

input Primary task $\mathcal{D}_{\texttt{P}} = (\boldsymbol{X}_{\texttt{P}}, \boldsymbol{Y}_{\texttt{P}})$, harmful task input $\boldsymbol{X}_{\texttt{H}}$, supervised loss $\mathcal{L}$, learning rate η, regularizing constants $\lambda_{\texttt{P}}, \lambda_{\texttt{H}} \in \mathbb{R}_+$, model initialization θ_0, ω_0

$\boldsymbol{K}_{\texttt{P}} = \boldsymbol{X}_{\texttt{P}}^\top \boldsymbol{X}_{\texttt{P}}$

$\boldsymbol{K}_{\texttt{H}} = \boldsymbol{X}_{\texttt{H}}^\top \boldsymbol{X}_{\texttt{H}}$

for $t = 0, 1, \ldots, T-1$ **do**

$\omega_{t+1} = \omega_t - \eta \nabla_\omega \mathcal{L}(\omega_t, \theta_t; \mathcal{D}_{\texttt{P}})$

$\boldsymbol{H}_{\texttt{P}}(\theta_t) = \theta_t^\top \boldsymbol{K}_{\texttt{P}} \theta_t, \ \boldsymbol{H}_{\texttt{H}}(\theta_t) = \theta_t^\top \boldsymbol{K}_{\texttt{H}} \theta_t$

$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}(\omega_t, \theta_t; \boldsymbol{X}_1)$
$- \eta \lambda_{\texttt{P}} \boldsymbol{K}_P^{-1} \nabla_\theta \mathcal{R}_{\texttt{well}}(\boldsymbol{H}_{\texttt{P}}(\theta_t))$
$- \eta \lambda_{\texttt{H}} \boldsymbol{K}_H^{-1} \nabla_\theta \mathcal{R}_{\texttt{ill}}(\boldsymbol{H}_{\texttt{H}}(\theta_t))$

end for

output Immunized feature extractor $\theta_{\texttt{I}} \triangleq \theta_T$.

EXPERIMENTS

Evaluation metrics: relative immunization ratio

$$\texttt{RIR} = \underbrace{\left(\frac{\kappa(\boldsymbol{H}_{\texttt{H}}(\theta_{\texttt{I}}))}{\kappa(\boldsymbol{H}_{\texttt{H}}(\boldsymbol{I}))}\right)}_{\textbf{(i)}} \Big/ \underbrace{\left(\frac{\kappa(\boldsymbol{H}_{\texttt{P}}(\theta_{\texttt{I}}))}{\kappa(\boldsymbol{H}_{\texttt{P}}(\boldsymbol{I}))}\right)}_{\textbf{(ii)}}$$

- A successful immunization has large (i) and small (ii)

Baselines:

1. $\mathcal{R}_{\texttt{ill}}$ **Only** immunizes the model by minimizing only the regularizer $\mathcal{R}_{\texttt{ill}}(\boldsymbol{H}_{\texttt{H}})$
2. **IMMA** [Zheng and Yeh, 2024]
3. **Opt** κ directly minimizes $\kappa(\boldsymbol{H}_{\texttt{P}}(\theta)) - \kappa(\boldsymbol{H}_{\texttt{H}}(\theta))$ via gradient descent instead of using our proposed regularizers.

Amber Yijia Zheng, and Raymond A. Yeh. "IMMA: Immunizing text-to-image models against malicious adaptation." *Proc. ECCV 2024*.

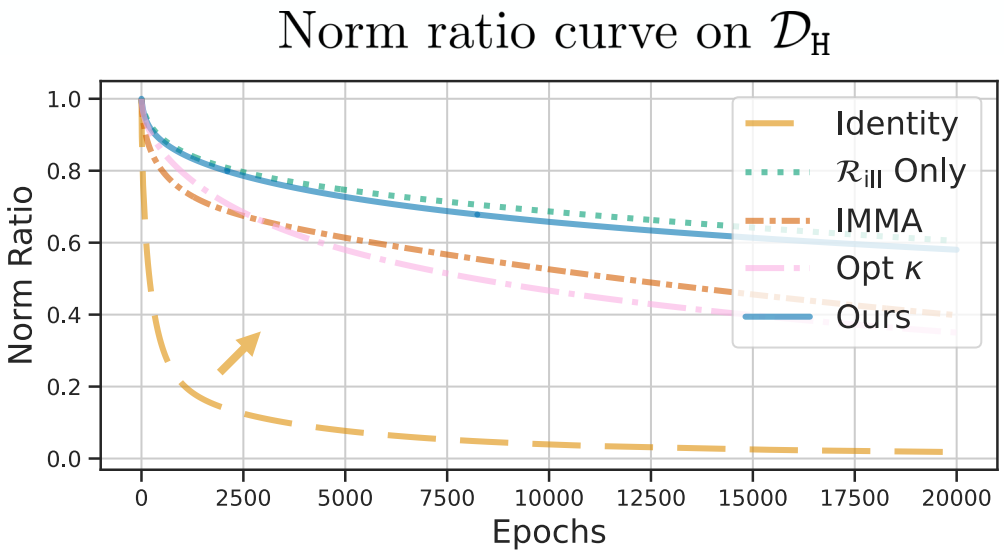
REGRESSION TASK (House Price Dataset)

The feature extractor is a linear model

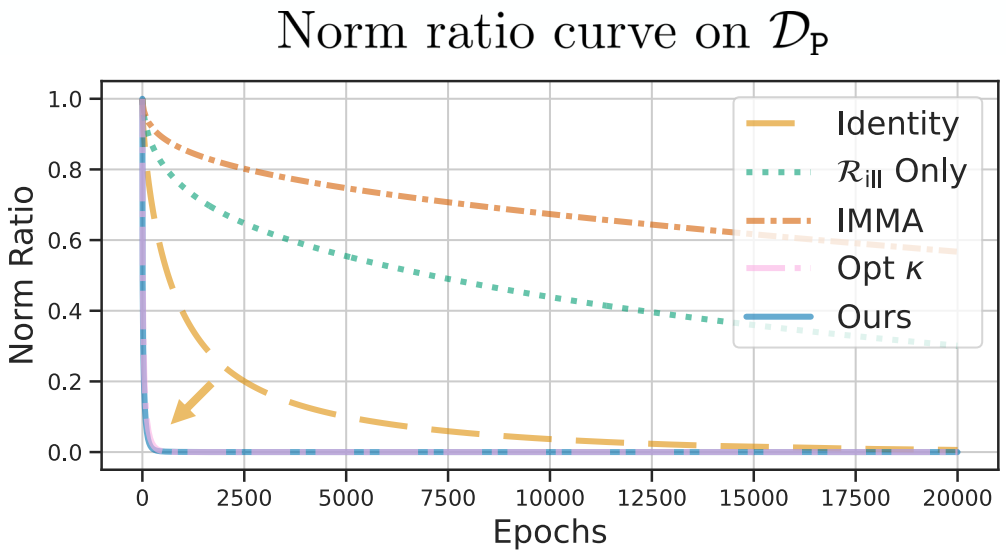
$$\texttt{RIR} = \underbrace{\left(\frac{\kappa(\boldsymbol{H}_{\texttt{H}}(\theta_{\texttt{I}}))}{\kappa(\boldsymbol{H}_{\texttt{H}}(\boldsymbol{I}))}\right)}_{\text{(i)}} \Big/ \underbrace{\left(\frac{\kappa(\boldsymbol{H}_{\texttt{P}}(\theta_{\texttt{I}}))}{\kappa(\boldsymbol{H}_{\texttt{P}}(\boldsymbol{I}))}\right)}_{\text{(ii)}}$$

Norm Ratio = $\|\mathbf{w}_t - \mathbf{w}^\star\|_2^2 / \|\mathbf{w}_0 - \mathbf{w}^\star\|_2^2$

Method	RIR (i)↑	RIR (ii)↓	RIR↑
$\mathcal{R}_{\texttt{ill}}$ *Only*	**90.02** ±**3.773**	72.415 ±3.545	1.244 ±0.021
IMMA	7.053 ±1.662	3.545 ±0.880	2.001 ±0.187
Opt κ	1.518 ±0.027	**0.016** ±**0.001**	92.58 ±4.492
Ours	18.92 ±2.056	0.053 ±0.002	**356.20** ±**5.491**



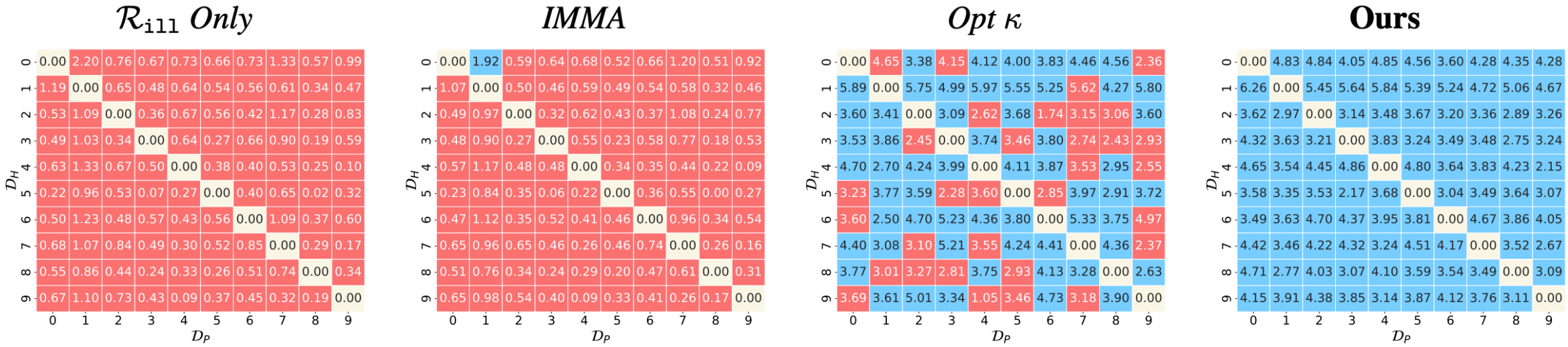
Slow convergence is preferred



Fast convergence is preferred

CLASSIFICATION TASK (MNIST)

Binary classification on MNIST. We use different digit pairs for $\mathcal{D}_{\mathtt{P}}$ and $\mathcal{D}_{\mathtt{H}}$.



$$\mathtt{RIR} = \underbrace{\left(\frac{\kappa(\boldsymbol{H}_{\mathtt{H}}(\theta_{\mathtt{I}}))}{\kappa(\boldsymbol{H}_{\mathtt{H}}(\boldsymbol{I}))}\right)}_{\textbf{(i)}} \Big/ \underbrace{\left(\frac{\kappa(\boldsymbol{H}_{\mathtt{P}}(\theta_{\mathtt{I}}))}{\kappa(\boldsymbol{H}_{\mathtt{P}}(\boldsymbol{I}))}\right)}_{\textbf{(ii)}}$$

Each element shows $\log(\mathtt{RIR})$. Blue if successful immunization, red otherwise.

EXPERIMENTS (DEEP-NETS)

- Feature extractor: a pre-trained model (ResNet18 and ViT) with initilization θ_0
- The evaluation metric RIR is extended to comparing relative to θ_0

$$\mathtt{RIR}_{\theta_0} \triangleq \underbrace{\left(\frac{\kappa(\tilde{\boldsymbol{H}}_{\mathtt{H}}(\theta_{\mathtt{I}}))}{\kappa(\tilde{\boldsymbol{H}}_{\mathtt{H}}(\theta_0))}\right)}_{\textbf{(i)}} \Big/ \underbrace{\left(\frac{\kappa(\tilde{\boldsymbol{H}}_{\mathtt{P}}(\theta_{\mathtt{I}}))}{\kappa(\tilde{\boldsymbol{H}}_{\mathtt{P}}(\theta_0))}\right)}_{\textbf{(ii)}} \quad \text{where} \quad \tilde{\boldsymbol{H}}_{\mathtt{H}}(\theta) = \nabla^2_{\mathbf{w}} \mathcal{L}(\mathcal{D}_{\mathtt{H}}, \mathbf{w}, \theta) = \tilde{\boldsymbol{X}}_{\mathtt{H}}(\theta)^\top \tilde{\boldsymbol{X}}_{\mathtt{H}}(\theta)$$

$\mathcal{D}_{\mathtt{H}}$	Method	ResNet18		
		$\mathtt{RIR}_{\theta_0}$ (i)↑	$\mathtt{RIR}_{\theta_0}$ (ii)↓	$\mathtt{RIR}_{\theta_0}$ ↑
	Init. θ_0	1.0	1.0	1.0
Cars	$\mathcal{R}_{\mathtt{ill}}$ *Only*	1.878 ±0.034	1.786 ±0.025	1.057 ±0.026
	IMMA	0.866 ±0.002	0.889 ±0.001	0.974 ±0.002
	Opt κ	1.217 ±0.021	0.798 ±0.005	1.527 ±0.019
	Ours	**2.386** ±0.442	**0.699** ±0.062	**3.467** ±0.358
Country211	$\mathcal{R}_{\mathtt{ill}}$ *Only*	**20.727** ±0.791	20.675 ±1.685	1.038 ±0.05
	IMMA	0.791 ±0.005	0.814 ±0.006	0.972 ±0.007
	Opt κ	1.538 ±0.155	1.053 ±0.091	1.472 ±0.043
	Ours	3.287 ±0.33	**0.399** ±0.034	**8.714** ±0.672

ViT		
$\mathtt{RIR}_{\theta_0}$ (i)↑	$\mathtt{RIR}_{\theta_0}$ (ii)↓	$\mathtt{RIR}_{\theta_0}$ ↑
1.0	1.0	1.0
13.121 ±0.038	4.097 ±0.098	3.342 ±0.048
1.422 ±0.006	2.090 ±0.043	0.702 ±0.007
3.598 ±0.510	**0.171** ±0.033	26.369 ±2.814
7.945 ±0.247	0.323 ±0.086	**34.517** ±0.886
69.291 ±1.198	63.519 ±6.62	1.122 ±0.097
6.242 ±0.203	7.599 ±0.717	0.845 ±0.048
4.589 ±0.079	**0.300** ±0.106	16.498 ±5.183
20.894 ±1.425	0.700 ±0.082	**41.341** ±0.967

EXPERIMENTS (DEEP-NETS)

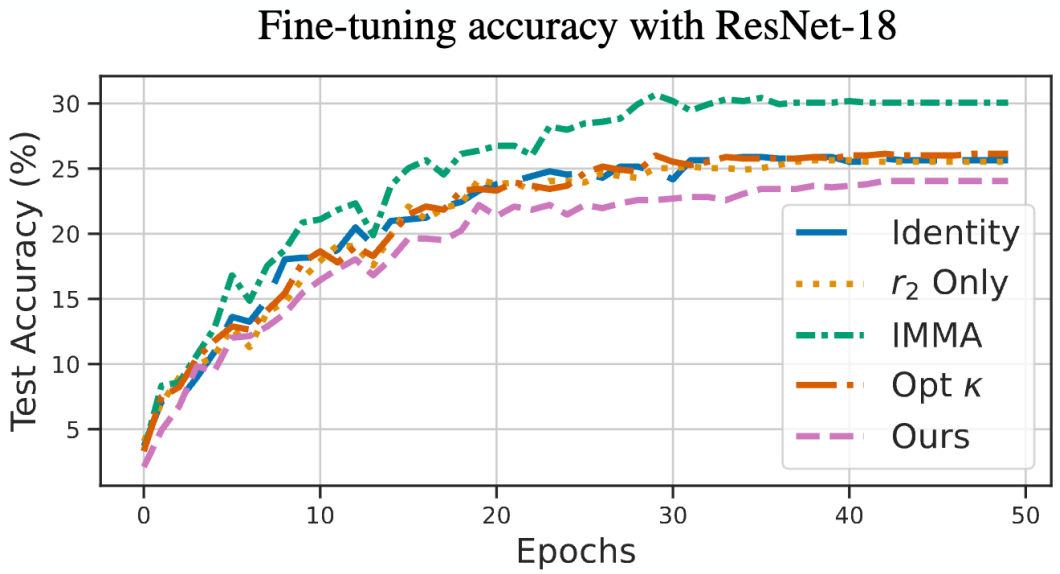
- Feature extractor: a pre-trained model (ResNet18 and ViT) with initilization θ_0
- The evaluation metric RIR is extended to comparing relative to θ_0
- We also report task performance, i.e., accuracy for image classification on $\mathcal{D}_{\mathrm{P}}$.

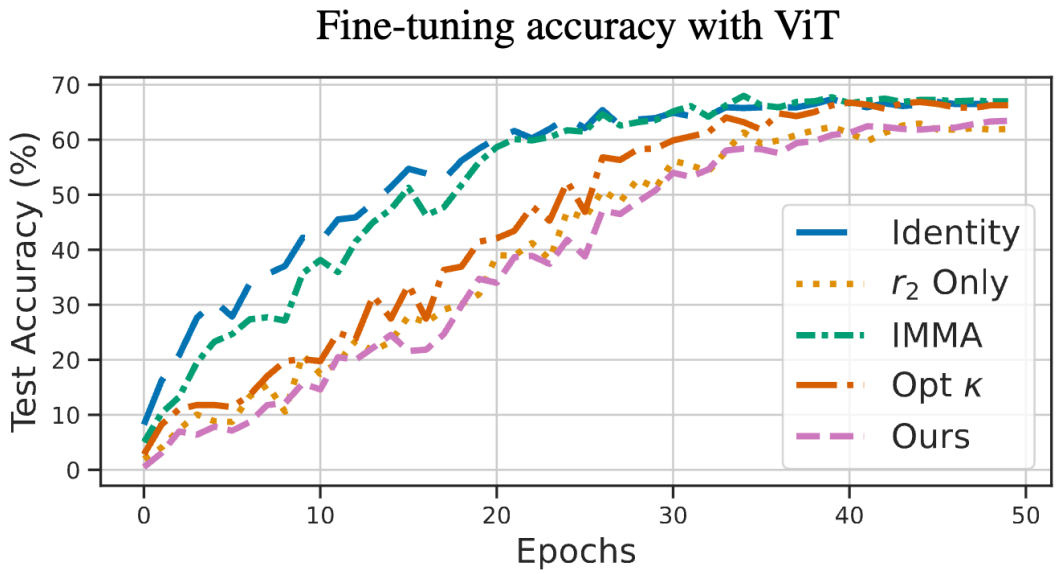
$\mathcal{D}_{\mathrm{H}}$	Method	ResNet18				ViT			
		$\mathtt{RIR}_{\theta_0}$ (i)↑	$\mathtt{RIR}_{\theta_0}$ (ii)↓	$\mathtt{RIR}_{\theta_0}$ ↑	$\mathcal{D}_{\mathrm{P}}$ Test Acc. (%) ↑	$\mathtt{RIR}_{\theta_0}$ (i)↑	$\mathtt{RIR}_{\theta_0}$ (ii)↓	$\mathtt{RIR}_{\theta_0}$ ↑	$\mathcal{D}_{\mathrm{P}}$ Test Acc. (%) ↑
	Init. θ_0	1.0	1.0	1.0	68.24	1.0	1.0	1.0	81.78
Cars	$\mathcal{R}_{\mathtt{ill}}$ *Only*	1.878 ±0.034	1.786 ±0.025	1.057 ±0.026	**63.84 ±0.292**	**13.121 ±0.038**	4.097 ±0.098	3.342 ±0.048	82.21 ±0.035
	IMMA	0.866 ±0.002	0.889 ±0.001	0.974 ±0.002	63.57 ±0.234	1.422 ±0.006	2.090 ±0.043	0.702 ±0.007	81.89 ±0.010
	Opt κ	1.217 ±0.021	0.798 ±0.005	1.527 ±0.019	63.65 ±0.148	3.598 ±0.510	**0.171 ±0.033**	26.369 ±2.814	82.51 ±0.085
	Ours	**2.386 ±0.442**	**0.699 ±0.062**	**3.467 ±0.358**	62.36 ±0.173	7.945 ±0.247	0.323 ±0.086	**34.517 ±0.886**	**82.79 ±0.200**
Country211	$\mathcal{R}_{\mathtt{ill}}$ *Only*	**20.727 ±0.791**	20.675 ±1.685	1.038 ±0.05	62.17 ±1.599	**69.291 ±1.198**	63.519 ±6.62	1.122 ±0.097	80.73 ±0.129
	IMMA	0.791 ±0.005	0.814 ±0.006	0.972 ±0.007	**67.03 ±0.146**	6.242 ±0.203	7.599 ±0.717	0.845 ±0.048	82.47 ±0.036
	Opt κ	1.538 ±0.155	1.053 ±0.091	1.472 ±0.043	66.81 ±0.115	4.589 ±0.079	**0.300 ±0.106**	16.498 ±5.183	82.79 ±0.023
	Ours	3.287 ±0.33	**0.399 ±0.034**	**8.714 ±0.672**	65.01 ±0.143	20.894 ±1.425	0.700 ±0.082	**41.341 ±0.967**	**83.17 ±0.075**

CLASSIFICATION TASK

Features pre-trained on ImageNet transferring to $\mathcal{D}_{\mathrm{H}}$

$\mathcal{D}_{\mathrm{H}}$	Method	ResNet18				ViT			
		RIR_{θ_0} (i)↑	RIR_{θ_0} (ii)↓	RIR_{θ_0} ↑	$\mathcal{D}_{\mathrm{P}}$ Test Acc. (%) ↑	RIR_{θ_0} (i)↑	RIR_{θ_0} (ii)↓	RIR_{θ_0} ↑	$\mathcal{D}_{\mathrm{P}}$ Test Acc. (%) ↑
	Init. θ_0	1.0	1.0	1.0	68.24	1.0	1.0	1.0	81.78
Cars	$\mathcal{R}_{\mathtt{i11}}$ *Only*	1.878 ±0.034	1.786 ±0.025	1.057 ±0.026	**63.84 ±0.292**	**13.121 ±0.038**	4.097 ±0.098	3.342 ±0.048	82.21 ±0.035
	IMMA	0.866 ±0.002	0.889 ±0.001	0.974 ±0.002	63.57 ±0.234	1.422 ±0.006	2.090 ±0.043	0.702 ±0.007	81.89 ±0.010
	Opt κ	1.217 ±0.021	0.798 ±0.005	1.527 ±0.019	63.65 ±0.148	3.598 ±0.510	**0.171 ±0.033**	26.369 ±2.814	82.51 ±0.085
	Ours	**2.386 ±0.442**	**0.699 ±0.062**	**3.467 ±0.358**	62.36 ±0.173	7.945 ±0.247	0.323 ±0.086	**34.517 ±0.886**	**82.79 ±0.200**





TAKE AWAYS

- A condition number based framework for model immunization
- Two differentiable regularizers and a gradient-based optimization algorithm.
- A first step towards principled understanding of model immunization.
- Project page / code:
  - amberyzheng.com/immu_cond_num

Closely related works:

- Rosati, Domenic, et al. "Representation noising: A defence mechanism against harmful finetuning." Proc. NeurIPS, 2024
- Zheng, Amber Yijia, et al. "Learning to obstruct few-shot image classification over restricted classes." Proc. ECCV, 2024
- Nenov, Rossen, et al. "(Almost) Smooth Sailing: Towards Numerical Stability of Neural Networks Through Differentiable Regularization of the Condition Number." ICML Workshop, 2024