Fast Adaptive Test-Time Defense with Robust Features

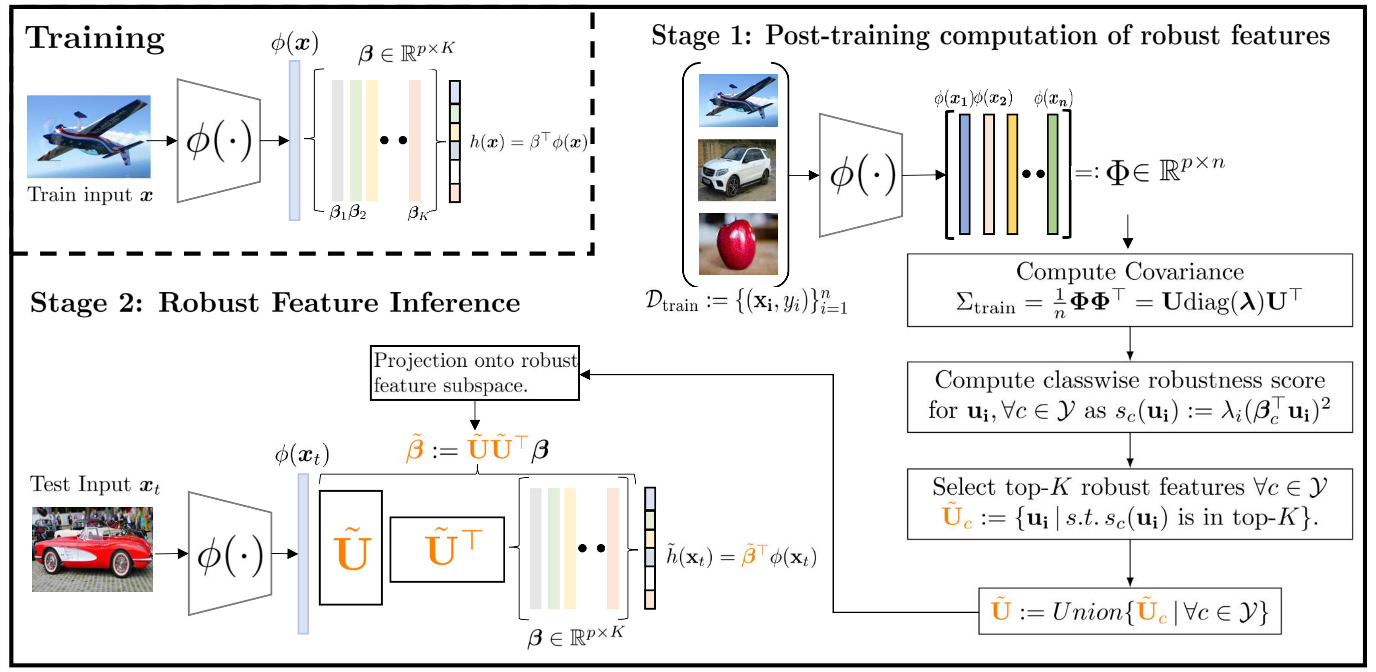
Anurag Singh[1], Mahalakshmi Sabanayagam[2] Krikamol Muandet[1], Debarghya Ghoshdastidar[2]
[1]CISPA Helmholtz Center for Information Security [2] Technical University of Munich

TL;DR

What is Adaptive test-time defense? Adaptive test-time defenses refer to the class of methods that improve the robustness of any trained model at **test time**.

Challenges: $40\times - 400\times$ increased inference time compared to underlying model due to additional computation or data processing while not necessarily improving the performance.

Our Contribution we develop a novel adaptive test-time defense strategy with the **same inference cost as the underlying model and no additional data or model complexity**.



Robust and Non Robust Features

A trained model $h : \mathcal{X} \to \mathcal{Y}$ is given by a **Generalized additive model (GAM)** s.t. $h(\mathbf{x}) = \beta^\top \phi(\mathbf{x})$, where $\phi : \mathcal{X} \to \mathcal{H}$ is a smooth f^n that maps the data into a *feature space* $\mathcal{H}$ and β are learned weights. The above form of h may represent the solution of kernel regression (with $\mathcal{H}$ being the corresponding reproducing kernel Hilbert space) or h could be the output layer of a neural network.

Features and their robustness. To identify the robust component of h, we aim to approximate ϕ as sum of K robust components $(\phi_i)_{i=1}^K$, or alternatively, $h(\mathbf{x}) \approx \sum_{i=1}^K \beta^\top \phi_i(\mathbf{x})$. We refer to each $\phi_i : \mathcal{X} \to \mathcal{H}$ as a *feature*. More generally, we define the set of all features as $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{H}\}$.

Definition 1 (ℓ_2-Robustness of features). *Given a distribution $\mathcal{D}$ on $\mathcal{X} \times \mathbb{R}^C$ and a trained model $h(\mathbf{x}) = \beta^\top \phi(\mathbf{x})$, we define the robustness of a feature $f \in \mathcal{F}$ as $s_{\mathcal{D},\beta}(f) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\inf_{\|\tilde{\mathbf{x}}-\mathbf{x}\|_2 \le \Delta} y^\top \beta^\top f(\tilde{\mathbf{x}})\right]$, while we use $s_{\mathcal{D},\beta,c}(f) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\inf_{\|\tilde{\mathbf{x}}-\mathbf{x}\|_2 \le \Delta} y_c \beta_c^\top f(\tilde{\mathbf{x}})\right]$ to specify the robustness of f with respect to the c-th class component of $y \in \mathbb{R}^C$, where $c \in \{1,\dots,C\}$ and β_c is c-th column of β.*

Theoretical Justification of robustness score

Theorem 1 (Lower bound on robustness). *Given $h(\mathbf{x}) = \beta^\top \phi(\mathbf{x})$. Assume that the distribution $\mathcal{D}$ is such that $y = h(\mathbf{x}) + \epsilon$, where $\epsilon \in \mathbb{R}^C$ has independent coordinates, each satisfying $\mathbb{E}[\epsilon_c] = 0$, $\mathbb{E}[\epsilon_c^2] \le \sigma^2$ for all $c \in \{1,\dots,C\}$. Further, assume that the map ϕ is L-Lipschitz, that is, $\|\phi(\mathbf{x}) - \phi(\tilde{\mathbf{x}})\|_{\mathcal{H}} \le L\|\mathbf{x} - \tilde{\mathbf{x}}\|$. Then, for any $f = \boldsymbol{M}\phi$ and every $c \in \{1,\dots,C\}$,*

$$s_{\mathcal{D},\beta,c}(f) \ \ge\ \beta_c^\top \Sigma \boldsymbol{M} \beta_c - L\Delta\|\boldsymbol{M}\|_{op}\|\beta_c\|_{\mathcal{H}}\sqrt{\sigma^2 + \beta_c^\top \Sigma \beta_c},$$

where $\Sigma = \mathbb{E}_{\mathbf{x}}\left[\phi(\mathbf{x})\phi(\mathbf{x})^\top\right]$ and $\|\boldsymbol{M}\|_{op}$ denotes the operator norm.

Theorem 1 suggests that if we search only over $f \in \mathcal{F}$ that are linear transformations $f = \boldsymbol{M}^\top \phi$ such that $\|\boldsymbol{M}\|_{op} = 1$, then the most robust feature is the one that maximizes the first term $\beta_c^\top \Sigma \boldsymbol{M} \beta_c$. In particular, we restrict our search to projections onto K dimensional subspace, $\boldsymbol{M} = \boldsymbol{P}\boldsymbol{P}^\top$, where $\boldsymbol{P}$ is the orthonormal basis for the subspace. We show that optimising over such features f corresponds to projecting onto top K eigenvectors $\mathbf{u}$ of Σ sorted according to a specific *robustness score*.

Corollary 2. *Fix any K and $(\lambda_i, \mathbf{u}_i)_{i=1,2,\dots}$ denote the eigenpairs of $\Sigma = \mathbb{E}_{\mathbf{x}}\left[\phi(\mathbf{x})\phi(\mathbf{x})^\top\right]$. Consider the problem of maximizing the lower bound in Theorem 1 over all features $f \in \mathcal{F}$ that correspond to projection of ϕ onto K dimensional subspace. Then the solution is given by $f = \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top \phi$, where $\tilde{\mathbf{U}}$ is the matrix of the K eigenvectors for which the robustness score $s_c(\mathbf{u}_i) = \lambda_i(\beta_c^\top \mathbf{u}_i)^2$ are largest.*

Experiments

Training	Clean		$\ell_\infty(\epsilon = \frac{8}{255})$		$\ell_2(\epsilon = 0.5)$	
	Method	+RFI	Method	+RFI	Method	+RFI
Standard	**95.28**	88.53	1.02	**4.35**	0.39	**9.73**
PGD	**83.53**	83.22	42.20	**43.29**	54.61	**55.03**
IAT	**91.86**	91.26	44.76	**46.95**	62.53	**64.31**
Robust CIFAR-10	78.69	**78.75**	1.30	**7.01**	9.63	**11.00**
C&W attack	**85.11**	84.97	40.01	**42.56**	55.02	**56.79**

Table 1: **Robust performance evaluation of RFI.** ℓ_∞ and ℓ_2 PGD attack on CIFAR-10 with Resnet-18. ℓ_∞ attack with step size $\epsilon/4$ and 40 iterations. ℓ_2 attack with size $\epsilon/5$ and 100 iterations.

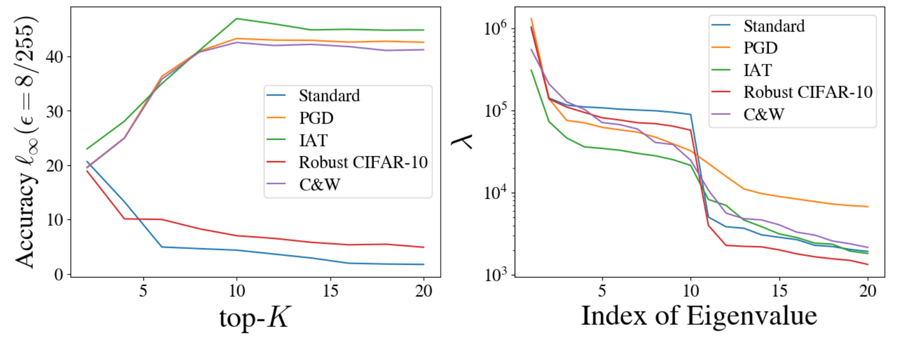


Figure 1: Robust Accuracy for different K and the corresponding eigenvalue profile in ascending order of all the methods.

Method	ℓ_∞				ℓ_2		
	$\epsilon = \frac{2}{255}$	$\epsilon = \frac{4}{255}$	$\epsilon = \frac{12}{255}$	$\epsilon = \frac{16}{255}$	$\epsilon = 0.25$	$\epsilon = 0.75$	$\epsilon = 1.00$
PGD	74.60	64.02	23.34	11.66	71.34	40.91	28.25
PGD+RFI	**74.99**	**64.91**	**24.32**	**12.55**	**71.48**	**41.95**	**29.24**

Table 2: Evaluation of RFI for ℓ_∞ and ℓ_2 PGD attack on CIFAR 10

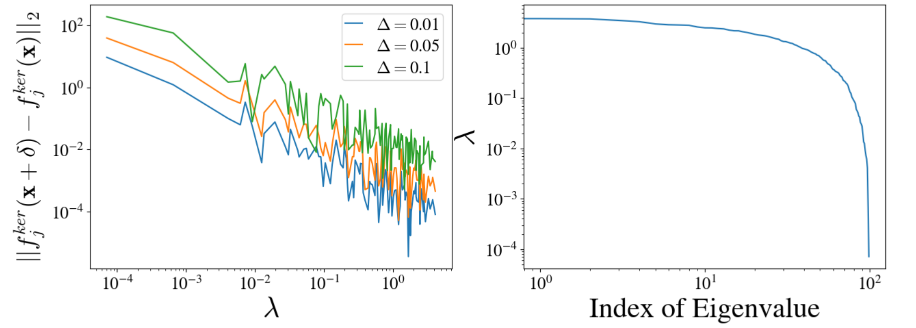


Figure 2: NTK feature robustness for λ and the corresponding eigenvalue profile in ascending order.