Adversarially Robust Continual Learning

Hikmat Khan • Nidhal Carla Bouaynaya • Ghulam Rasool

Department of Electrical and Computer Engineering, Rowan University, USA

IEEE WCCI 2022, PADUA, ITALY • 18 - 23 July, 2022

Project Page

ABSTRACT

- Recent approaches in continual learning (CL) have focused on extracting various types of features from multi-task datasets to prevent catastrophic forgetting — without formally evaluating the quality, robustness and usefulness of these features.
- This paper presents an empirical study to demonstrate the importance of robust features in the context of class incremental learning (CIL).

DEFINING USEFUL AND ROBUST FEATURES

- Adversarial robustness can be understood by decomposing learned features into robust and non-robust types.
- The robust features were used to build robust datasets and shown to increase adversarial robustness significantly.

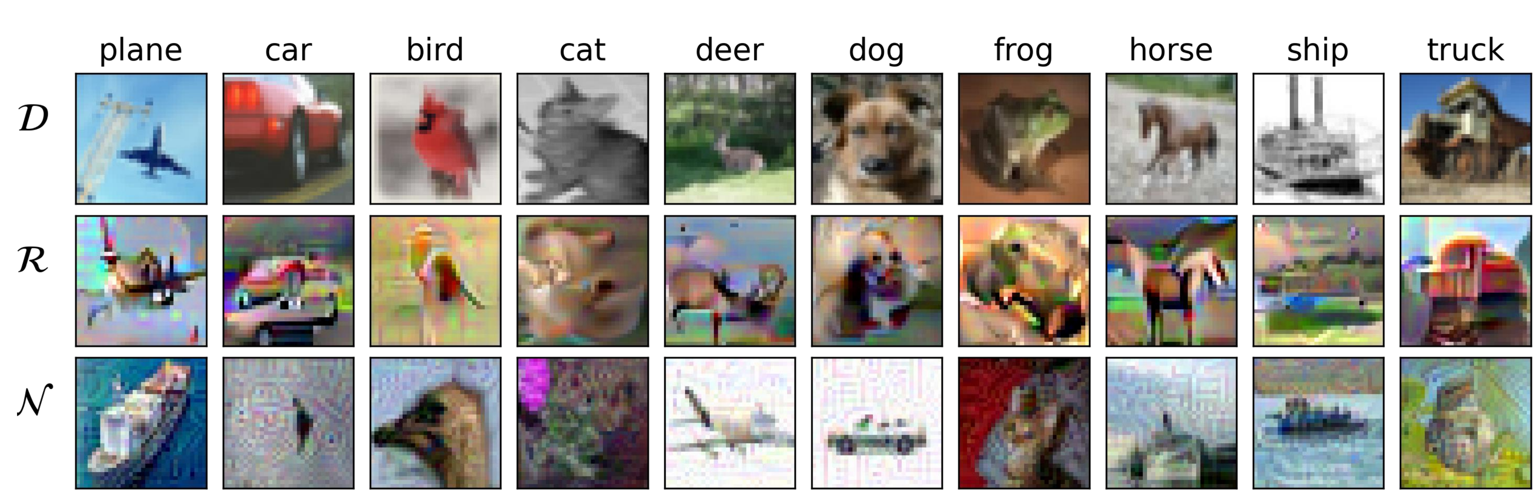
$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[y.f(x)] \geq \rho \ldots\ldots \text{Useful features}$$

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[\inf_{\delta\in\Delta(x)} y.f(x+\delta)] \geq \gamma \ldots\ldots \text{Robust features}$$

DISENTANGLING ROBUST AND NON-ROBUST FEATURES

$$\mathbb{E}_{(x,y)\sim\widehat{\mathcal{D}}_R}[y.f(x)] = \begin{cases} \mathbb{E}_{(x,y)\sim\mathcal{D}}[y.f(x)] & \text{if } f \in \mathcal{F}_c \\ 0 & \text{otherwise,} \end{cases}$$

- Where $\mathcal{F}_c$ represents the set of the features utilized by a robust (i.e., adversarially trained) classifier c.



1st row: Sample images from standard CIFAR10(D) for all 10 classes. 2nd row: The robustified sample images from robust CIFAR10(R). 3rd row: Samples images from non-robust CIFAR10(N) dataset.

MOTIVATION

- There has not been any assessment on using such robust features in CL frameworks to enhance the robustness of CL models against adversarial attacks.
- Current CL algorithms use standard features - a mixture of robust and non-robust features - and result in models vulnerable to both natural and adversarial noise.

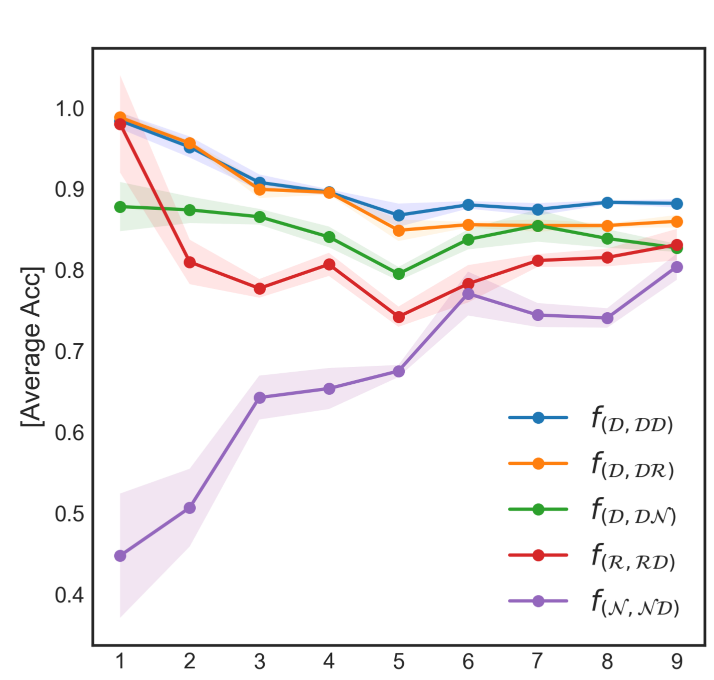
METHODOLOGY

$$ACC = \frac{1}{T}\sum_{i=1}^{T} R_{T,i} \ldots\ldots \text{where ACC is average accuracy}$$

Model	Training set	Replay buffer ($size = 16000$)	Average accuracy
$f_{(\mathcal{D},\mathcal{DD})}$	CIFAR10 ($\mathcal{D}$)	CIFAR10 ($\mathcal{D}$) + CIFAR10 ($\mathcal{D}$)	$88.20\pm_{0.48}$
$f_{(\mathcal{D},\mathcal{DR})}$	CIFAR10 ($\mathcal{D}$)	CIFAR10 ($\mathcal{D}$) + Robustified CIFAR10 ($\mathcal{R}$)	$85.99\pm_{0.77}$
$f_{(\mathcal{D},\mathcal{DN})}$	CIFAR10 ($\mathcal{D}$)	CIFAR10 ($\mathcal{D}$) + Non-Robustified CIFAR10 ($\mathcal{N}$)	$82.72\pm_{0.49}$
$f_{(\mathcal{R},\mathcal{RD})}$	Robustified CIFAR10 ($\mathcal{R}$)	Robustified CIFAR10 ($\mathcal{R}$) + CIFAR10 ($\mathcal{D}$)	$83.12\pm_{1.9}$
$f_{(\mathcal{N},\mathcal{ND})}$	Non-Robustified CIFAR10 ($\mathcal{N}$)	Non-Robustified CIFAR10 ($\mathcal{N}$) + CIFAR10 ($\mathcal{D}$)	$80.40\pm_{1.6}$

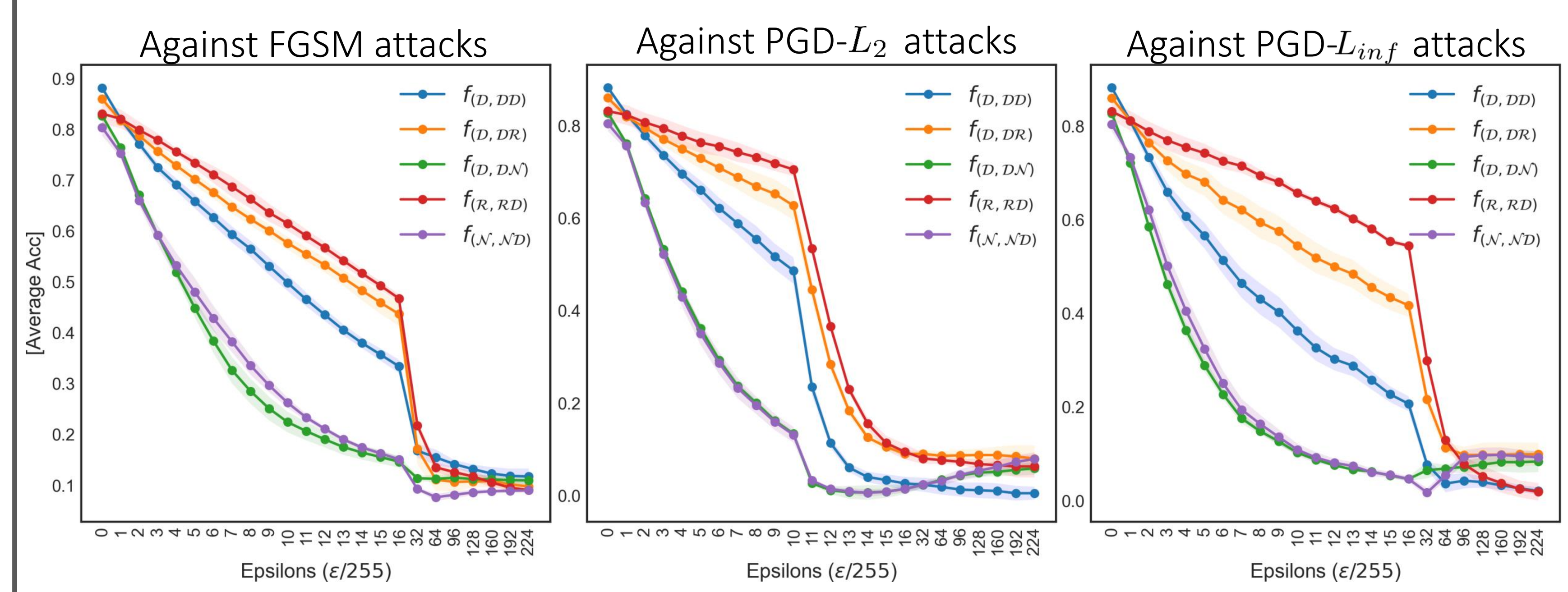
The average accuracies of all five models trained on CIFAR10, Robustified-CIFAR10 and Non-Robustified CIFAR10 datasets. In the model $f_{(X,YZ)}$: the first entry represent the training data set, the second set of letters denotes the replay buffer datasets sampled equally.

- The average of the five models $f_{(\mathcal{D},\mathcal{DD})}$, $f_{(\mathcal{D},\mathcal{DR})}$, $f_{(\mathcal{D},\mathcal{DN})}$, $f_{(\mathcal{R},\mathcal{RD})}$ and $f_{(\mathcal{N},\mathcal{ND})}$, as the incrementally learn a sequence of 9 tasks in CIFAR10 dataset. Where D=Standard CIFAR10, R=Robustified CIFAR10 and N=Non-Robustified CIFAR10 datasets.

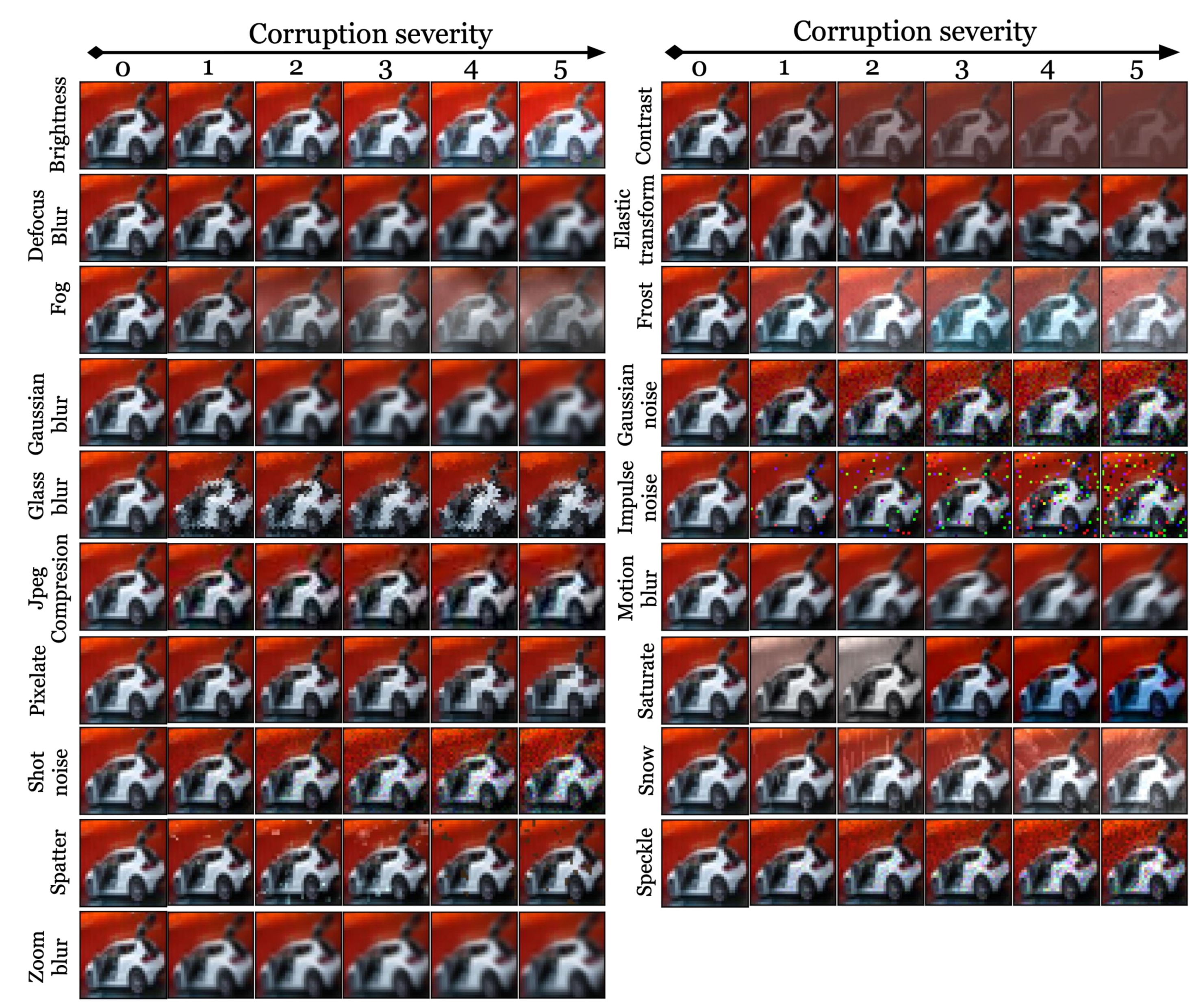


EVALUATING ROBUSTNESS AGAINST WORST-CASES ADVERSARIAL PERTURBATIONS

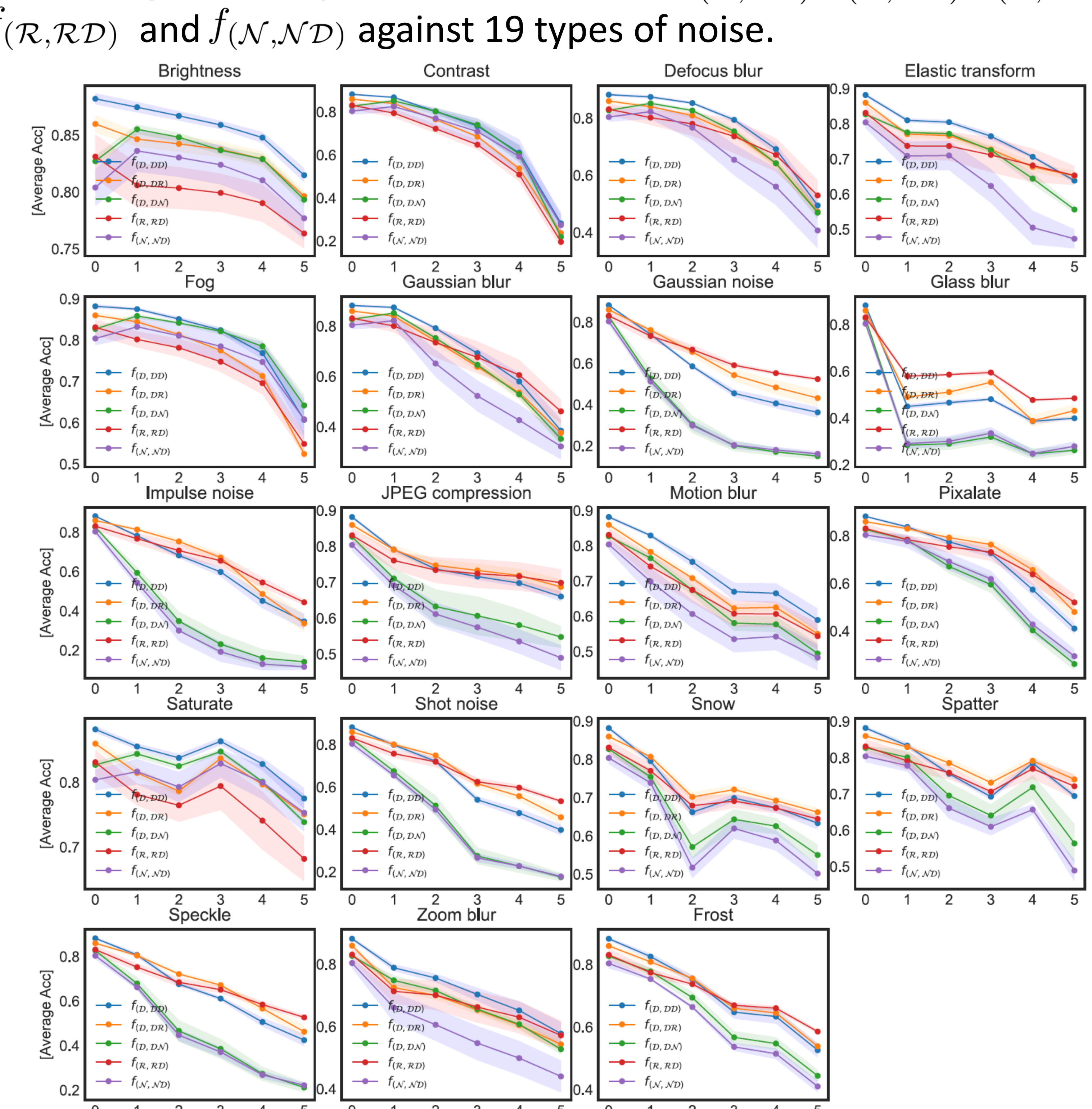
- The model trained using non-robust features performed the worst in noisy conditions and under adversarial attacks. Our study underlines the significance of using robust features in CIL.



EVALUATING ROBUSTNESS AGAINST COMMON CORRUPTIONS (CIFAR10-C Dataset)



- The average accuracy of the 5 models (i.e. $f_{(\mathcal{D},\mathcal{DD})}$, $f_{(\mathcal{D},\mathcal{DR})}$, $f_{(\mathcal{D},\mathcal{DN})}$, $f_{(\mathcal{R},\mathcal{RD})}$ and $f_{(\mathcal{N},\mathcal{ND})}$ against 19 types of noise.



CONCLUSION

- We presented an empirical and exhaustive study to demonstrate the crucial role of features in the context of class incremental learning (CIL) under various noise and perturbation environments.
- We concluded that continual learning models trained using standard and non-robust features performed poorly in noisy and adversarial conditions as compared to the model trained using robust features.