UNIVERSITÀ DEGLI STUDI DI CAGLIARI

Github Repo:
https://github.com/anauroranon/MLSec_Accuracy-And-Robustness-of-RobustBench-models/

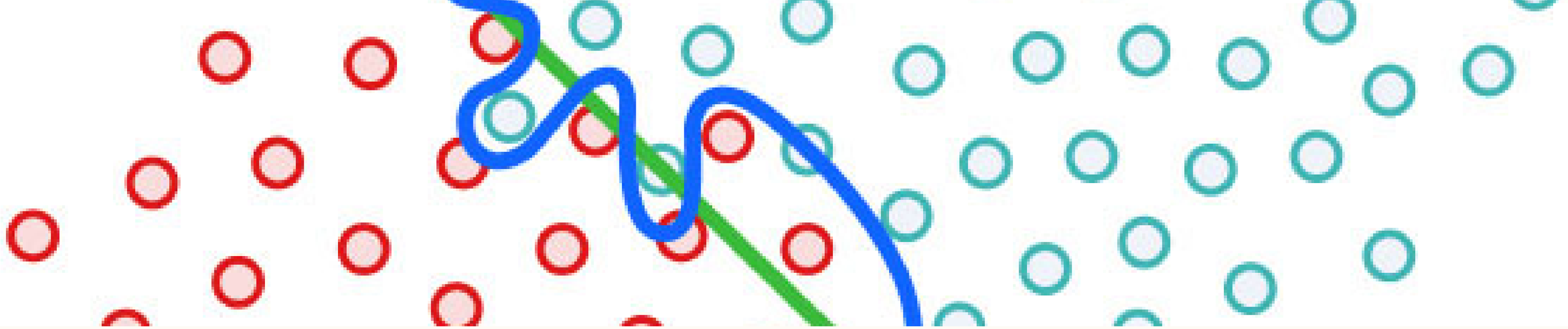
Variation of Accuracy and Robustness across classes

Aurora Arrus 70/90/00332
Matteo Asuni 70/90/00330
Pierangelo Loi 70/90/00335

Machine Learning Security
Computer Engineering, Cybersecurity and Artificial Intelligence

Outline

- Robustness and Accuracy
- RobustBench
- Autoattack
- Choice of 5 models
- Cifar-10
- Models overview
- ResNet architecture
- XCiT architecture
- WideResnet Architecture
- Accuracy Results
- Accuracy Results (2)
- Observations
- Models overview (2)
- WideResnet accuracy results
- WideResnet accuracy results(2)
- Final Observations
- Bibliography



Robustness and accuracy of a Machine Learning model

Accuracy

Is the ability of a machine learning model to make correct predictions or classifications

Robustness

Is the ability of a machine learning model to maintain its performance and make accurate predictions even in the presence of unexpected or adversarial inputs

RobustBench

A standardized benchmark for adversarial robustness

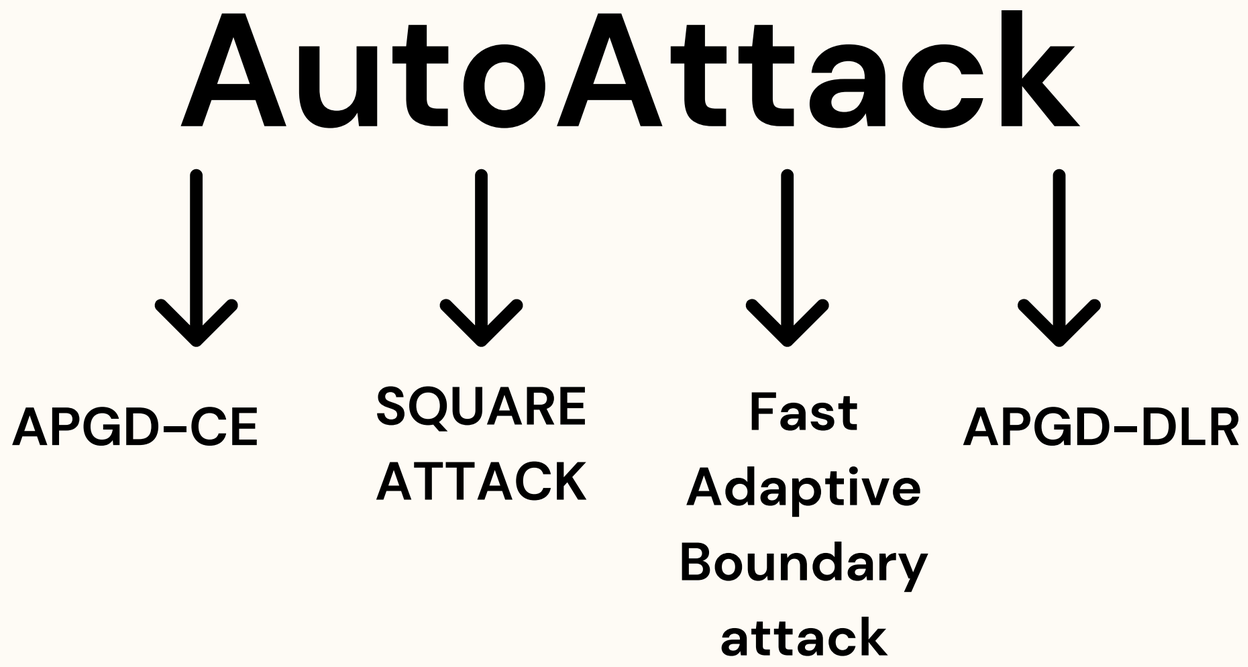
Robustness evaluation involves checking how the model overcomes adverse conditions.

RobustBench aims to address the need for a standardized adversarial robustness evaluation. [1]

- L∞ and L2 threat models
- Reasonable computational requirements
- Leaderboard
- AutoAttack evaluation
- ModelZoo

AutoAttack is an ensemble of complementary attacks designed to estimate adversarial robustness. [2]

- Parameter-free
- Computationally affordable
- User-independent
- Standardized Evaluation



Choice of models

Dataset

The models have been trained for image classification with **CIFAR-10**

Norm

L∞ is the mathematical measure used to quantify the maximum absolute value of the perturbation

ε

Epsilon

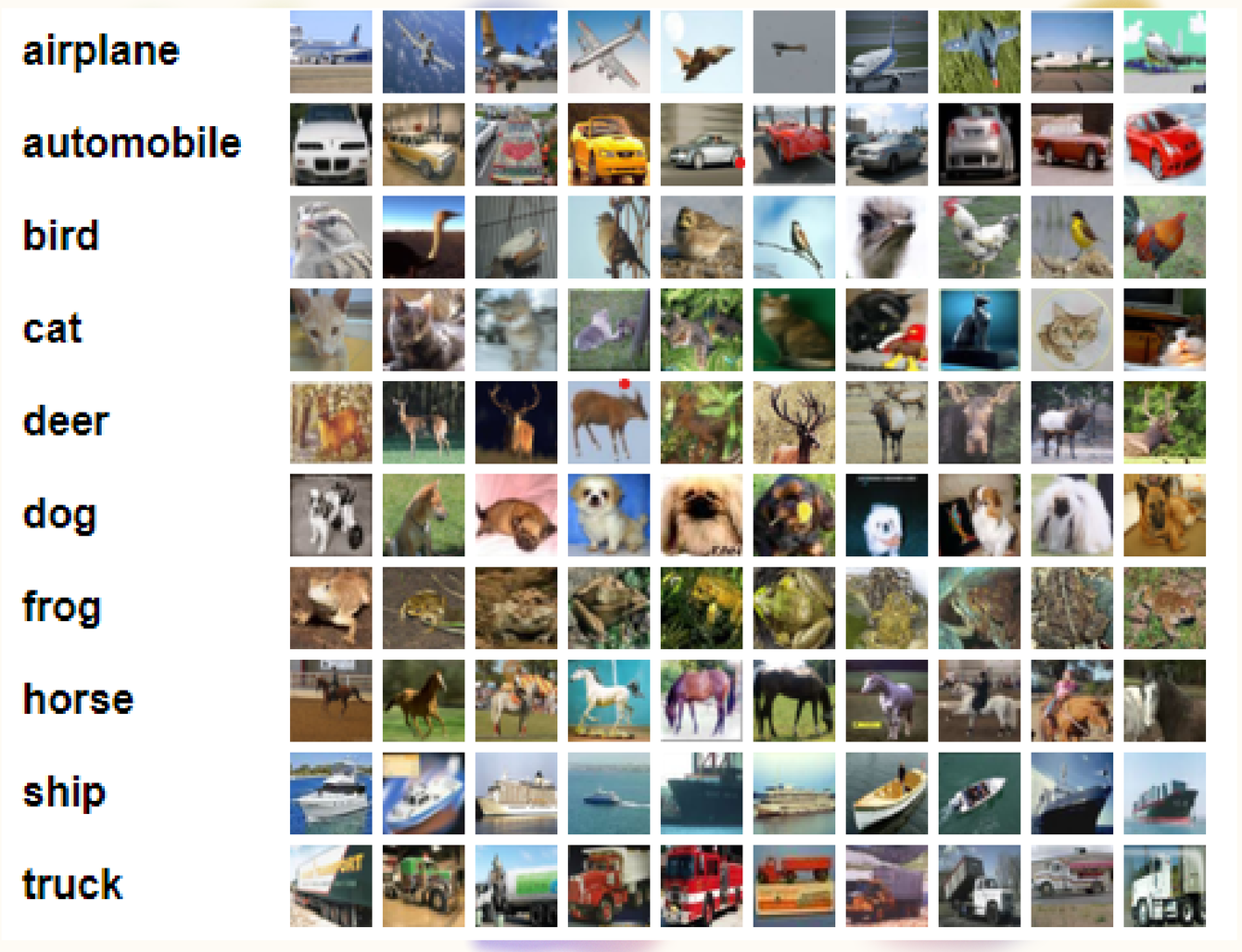
eps=8/255 represents the maximum allowable perturbation in the input data for each puxel value

CIFAR-10

[3]

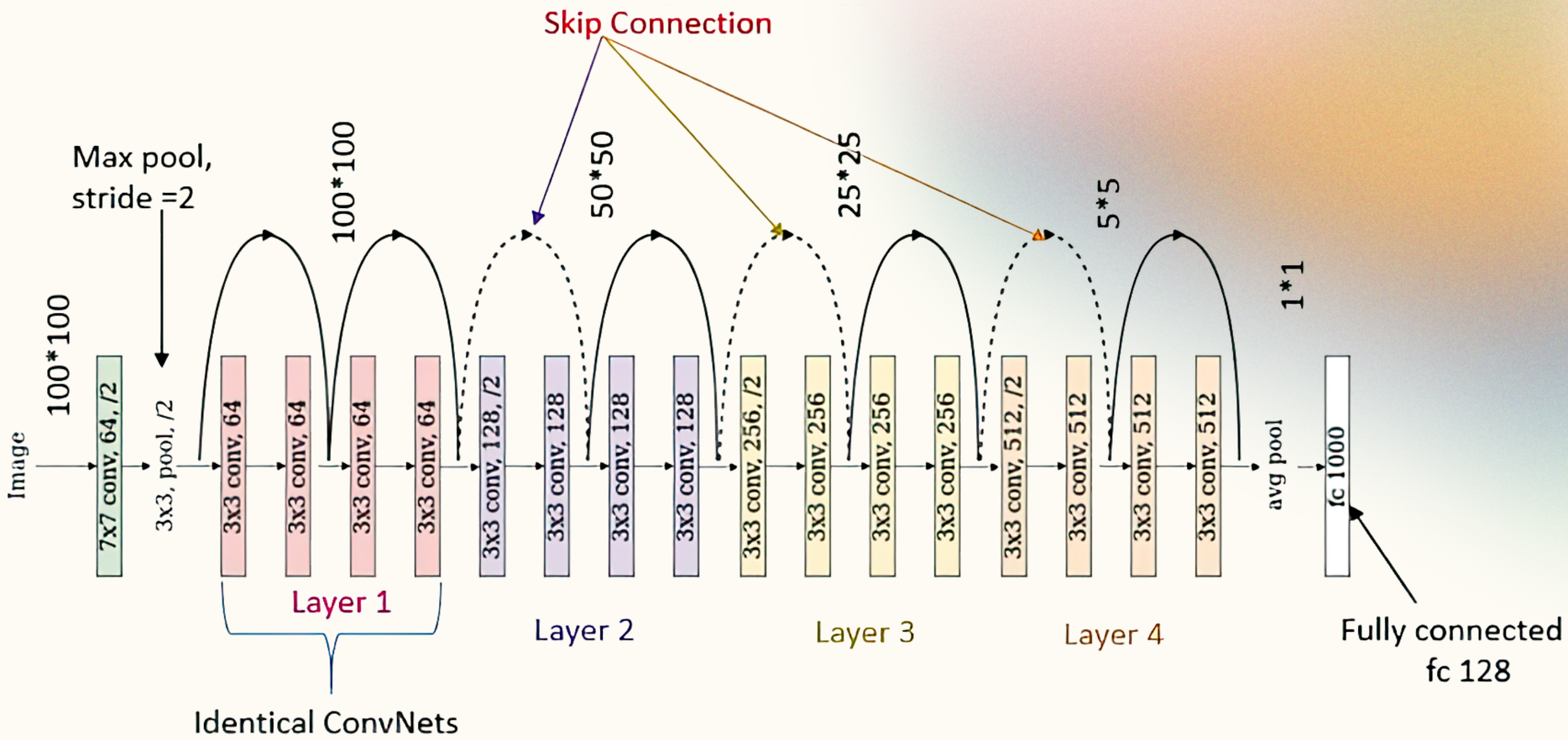
Image Classification dataset

- 60,000 color images
- 32x32 pixels
- 10 classes



Name	Parameters #	Architecture
Addepalli2022Efficient_RN18	11 M	ResNet-18
Debenedetti2022Light_XCiT-S12	26 M	XCiT-S12
Debenedetti2022Light_XCiT-M12	46 M	XCiT-M12
Debenedetti2022Light_XCiT-L12	104 M	XCiT-L12
Wang2023Better_WRN-70-16	266.8 M	WideResNet-70-16

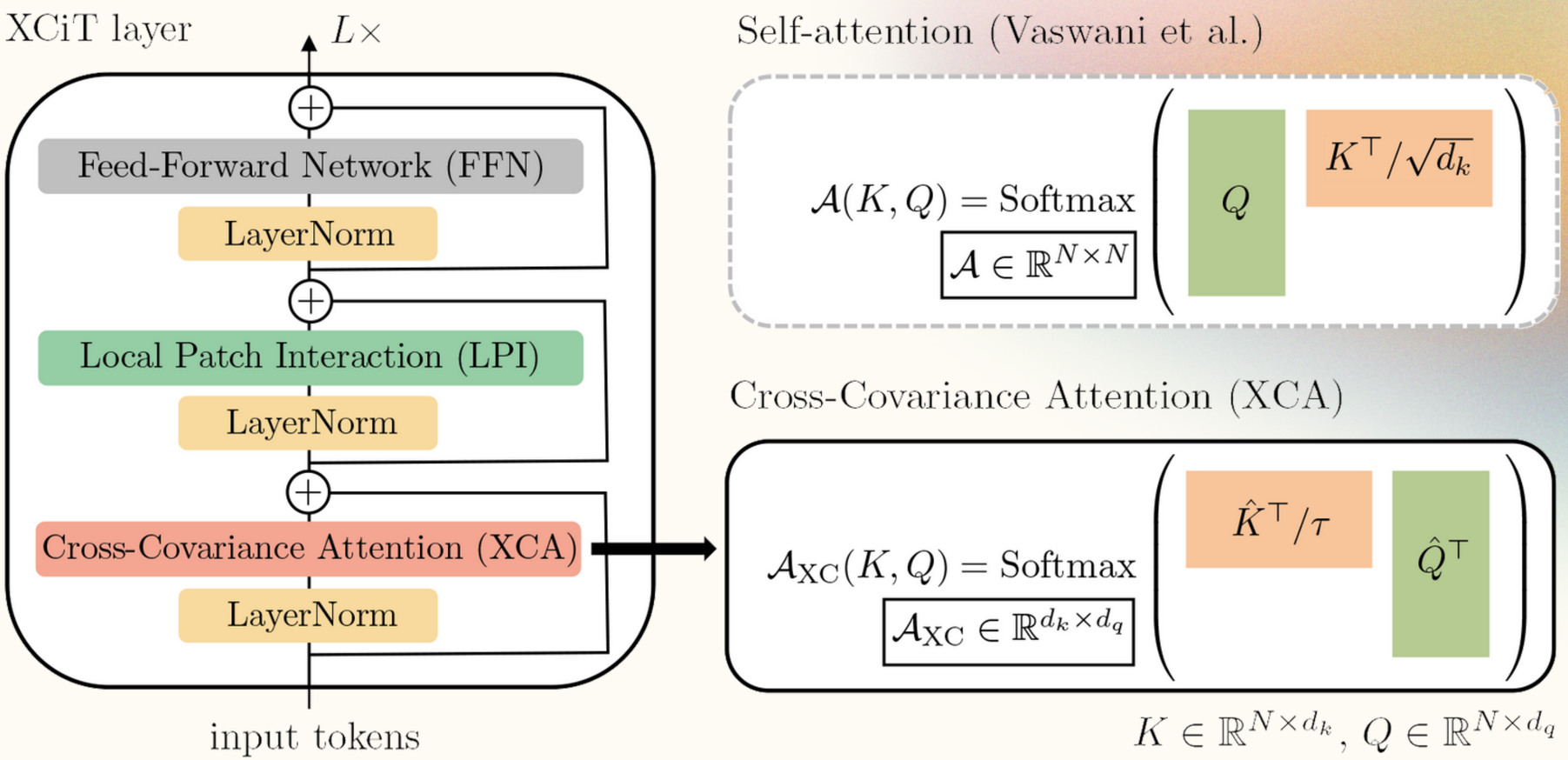
Architectures



ResNet

ResNet-18 [4]

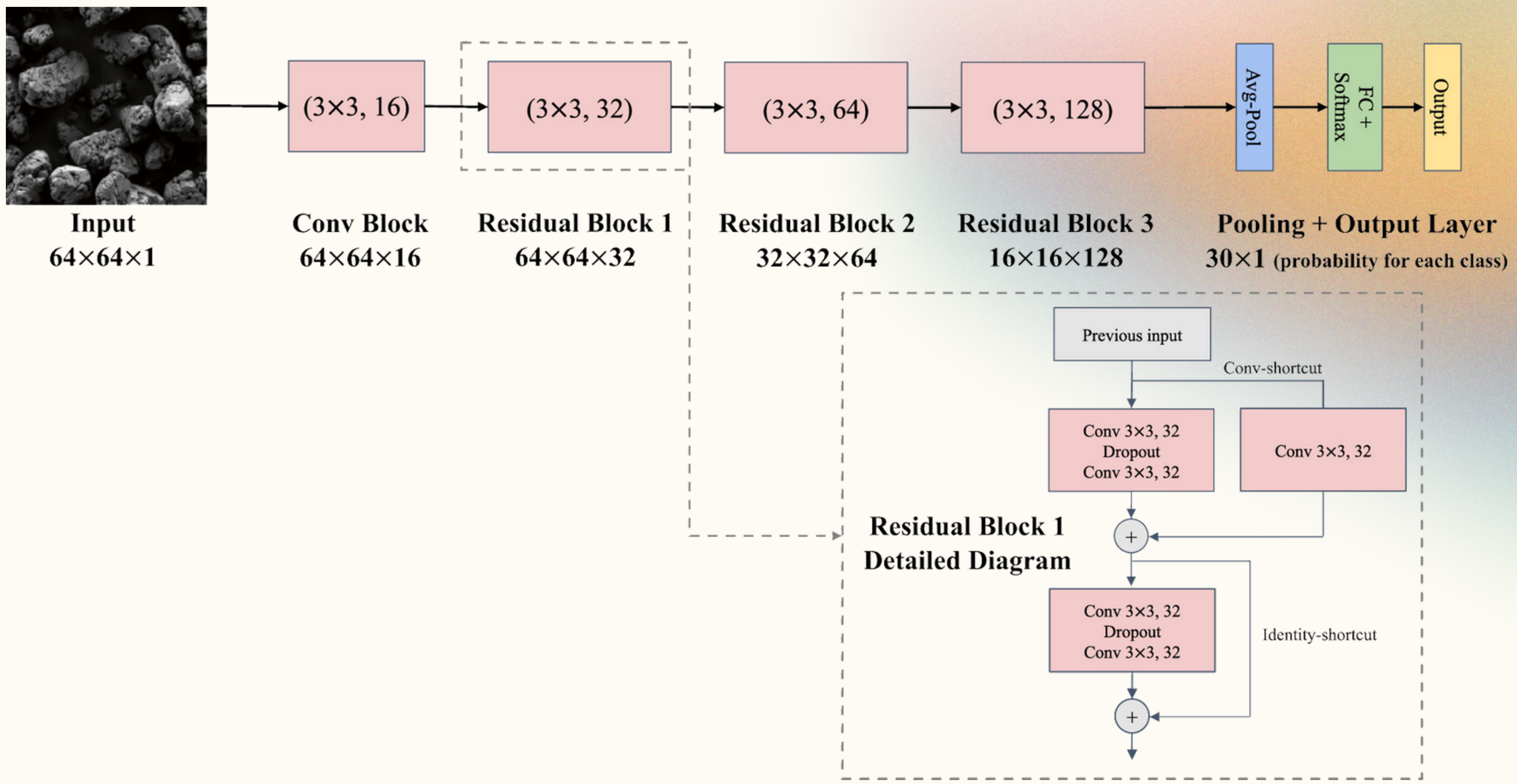
Architectures



XCiT

XCiT-S, XCiT-M, XCiT-L [5]

Architectures



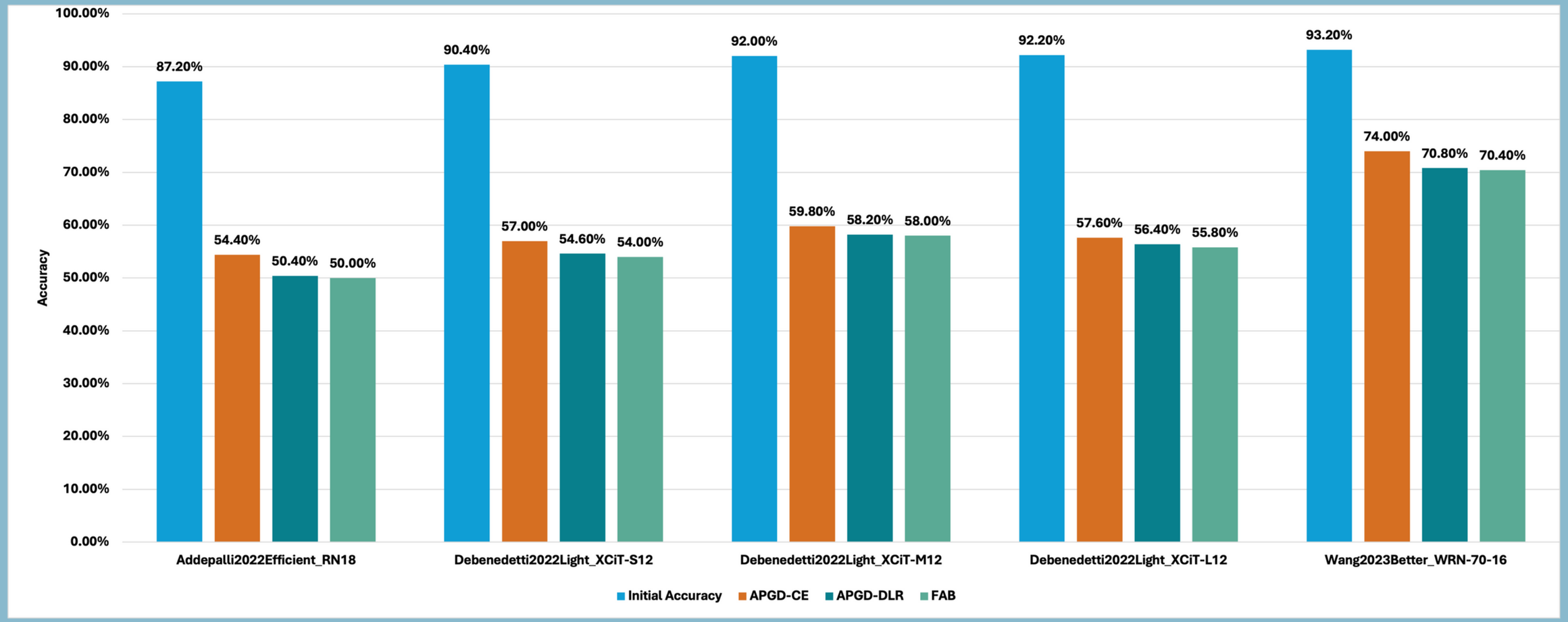
WideResNet

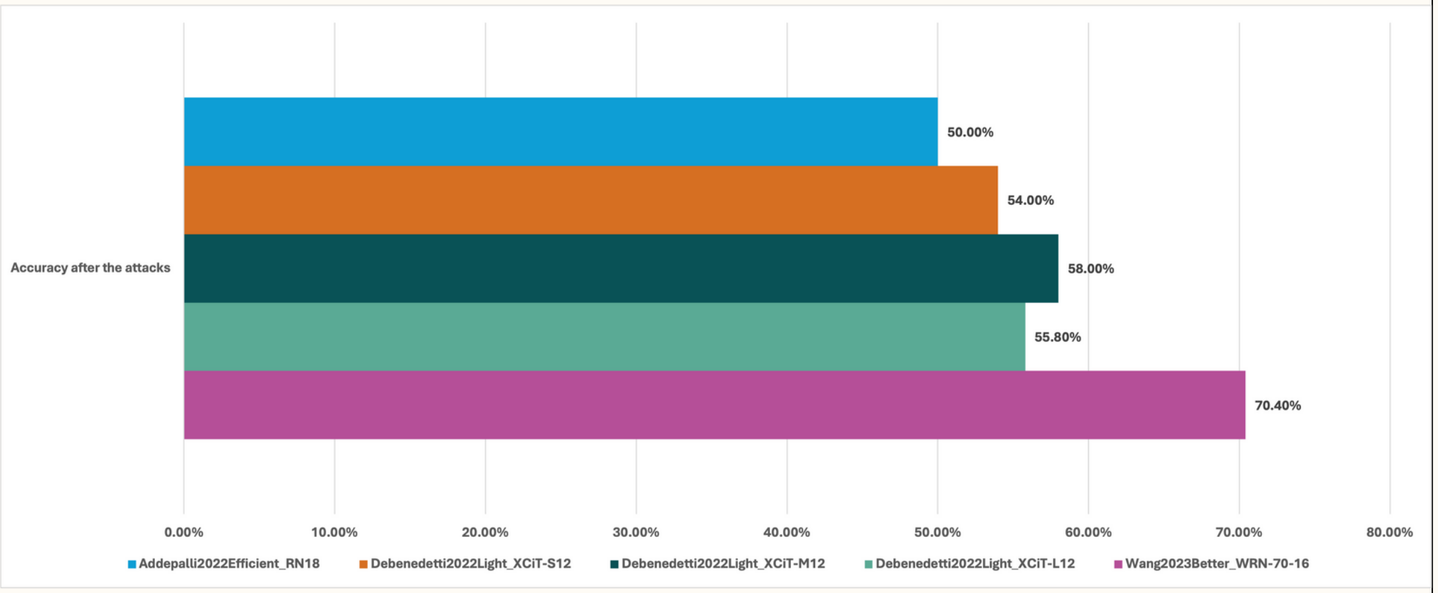
WideResNet-70-16 [6]

Results – Accuracy

Model		Overall	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9
Addepalli2022Efficient_RN18	Initial Accuracy	87.20%	84.21%	92.68%	86.27%	71.43%	82.50%	89.58%	92.59%	89.36%	89.47%	92.86%
	APGD-CE	54.40%	56.14%	75.61%	50.98%	20.41%	25.00%	41.67%	64.81%	61.70%	66.67%	75.00%
	APGD-DLR	50.40%	54.39%	75.61%	45.10%	14.29%	17.50%	41.67%	53.70%	59.57%	63.16%	71.43%
	FAB	50.00%	54.39%	75.61%	45.10%	14.29%	15.00%	41.67%	53.70%	57.45%	63.16%	71.43%
Debenedetti2022Light_XCiT-S12	Initial Accuracy	90.40%	91.23%	95.12%	94.12%	81.63%	85.00%	81.25%	92.59%	97.87%	91.23%	92.86%
	APGD-CE	57.00%	66.67%	70.73%	54.90%	22.45%	35.00%	43.75%	64.81%	63.83%	64.91%	76.79%
	APGD-DLR	54.60%	63.16%	70.73%	52.94%	20.41%	35.00%	41.67%	59.26%	61.70%	63.16%	73.21%
	FAB	54.00%	63.16%	70.73%	52.94%	18.37%	32.50%	41.67%	55.56%	61.70%	63.16%	73.21%
Debenedetti2022Light_XCiT-M12	Initial Accuracy	92.00%	89.47%	97.56%	96.08%	85.71%	87.50%	87.50%	90.74%	100.00%	91.23%	94.64%
	APGD-CE	59.80%	66.67%	78.05%	56.86%	28.57%	40.00%	43.75%	64.81%	70.21%	68.42%	75.00%
	APGD-DLR	58.20%	64.91%	78.05%	56.86%	24.49%	35.00%	43.75%	59.26%	70.21%	68.42%	75.00%
	FAB	58.00%	64.91%	78.05%	56.86%	24.49%	35.00%	43.75%	59.26%	68.09%	68.42%	75.00%
Debenedetti2022Light_XCiT-L12	Initial Accuracy	92.20%	91.23%	97.56%	92.16%	81.63%	90.00%	87.50%	92.59%	100.00%	94.74%	94.64%
	APGD-CE	57.60%	59.65%	78.05%	54.90%	22.45%	47.50%	50.00%	55.56%	65.96%	68.42%	71.43%
	APGD-DLR	56.40%	59.65%	78.05%	52.94%	20.41%	40.00%	47.92%	55.56%	65.96%	68.42%	71.43%
	FAB	55.80%	59.65%	75.61%	50.98%	20.41%	40.00%	45.83%	55.56%	65.96%	68.42%	71.43%
Wang2023Better_WRN-70-16	Initial Accuracy	93.20%	92.98%	100.00%	94.12%	83.67%	92.50%	89.58%	92.59%	95.74%	96.49%	94.64%
	APGD-CE	74.00%	75.44%	85.37%	66.67%	55.10%	57.50%	68.75%	77.78%	78.72%	82.46%	85.71%
	APGD-DLR	70.80%	71.93%	85.37%	64.71%	44.90%	52.50%	64.58%	74.07%	76.60%	82.46%	85.71%
	FAB	70.40%	71.93%	82.93%	64.71%	40.82%	52.50%	64.58%	74.07%	76.60%	82.46%	85.71%

Results – Accuracy





Observations

WideresNet-70-16 demonstrates superior robustness, while other models show comparable results.

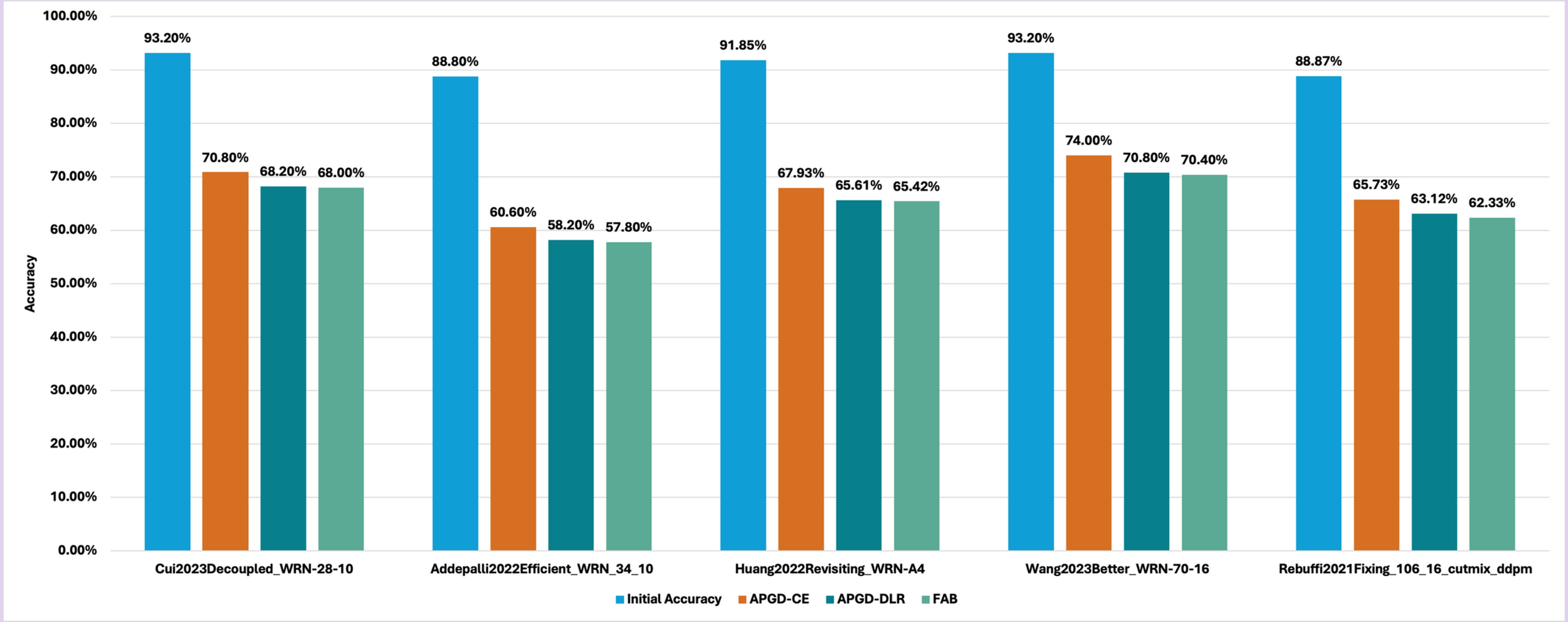
The influence of the robustness of the number of parameters versus the architecture of the model remains unclear from the outcomes.

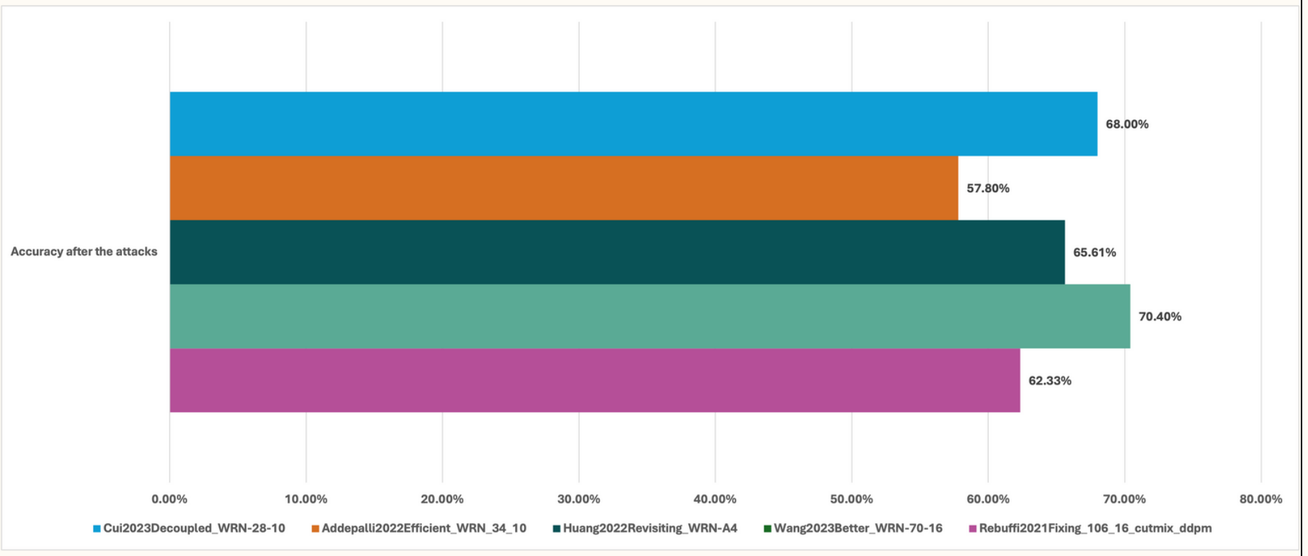
Name	Parameters #	Architecture
Cui2023Decoupled_WRN-28-10	36 M	WideResnet-28-10
Addepalli2022Efficient_WRN_34_10	46 M	WideResnet-34-10
Huang2022Revisiting_WRN-A4	106 M	WideResnet-A4
Wang2023Better_WRN-70-16	266.8 M	WideResNet-70-16
Rebuffi2021Fixing_106_16_cutmix_ddpm	415 M	WideResNet-106-16

Results 2 – Accuracy

Model		Overall	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9
Cui2023Decoupled_WRN-28-10	Initial Accuracy	93.20%	92.98%	100.00%	96.08%	83.67%	92.50%	85.42%	92.59%	97.87%	96.49%	94.64%
	APGD-CE	70.80%	70.18%	85.37%	66.67%	48.98%	55.00%	62.50%	75.93%	76.60%	78.95%	83.93%
	APGD-DLR	68.20%	70.18%	85.37%	64.71%	38.78%	52.50%	60.42%	68.52%	74.47%	78.95%	83.93%
	FAB	68.00%	70.18%	82.93%	64.71%	38.78%	52.50%	60.42%	68.52%	74.47%	78.95%	83.93%
Addepalli2022Efficient_WRN_34_10	Initial Accuracy	88.80%	84.21%	95.12%	90.20%	75.51%	87.50%	87.50%	90.74%	93.62%	92.98%	91.07%
	APGD-CE	60.60%	68.42%	75.61%	54.90%	30.61%	35.00%	52.08%	68.52%	63.83%	66.67%	82.14%
	APGD-DLR	58.20%	68.42%	75.61%	50.98%	26.53%	32.50%	50.00%	59.26%	63.83%	66.67%	80.36%
	FAB	57.80%	68.42%	73.17%	50.98%	24.49%	32.50%	50.00%	59.26%	63.83%	66.67%	80.36%
Huang2022Revisiting_WRN-A4	Initial Accuracy	91.85%	92.98%	95.12%	94.12%	77.55%	87.50%	89.58%	92.59%	97.87%	94.74%	96.43%
	APGD-CE	67.93%	73.68%	85.37%	60.78%	44.90%	50.00%	60.42%	68.52%	74.47%	75.44%	85.71%
	APGD-DLR	65.61%	73.68%	85.37%	58.82%	34.69%	45.00%	56.25%	66.67%	74.47%	75.44%	85.71%
	FAB	65.42%	73.68%	85.37%	58.82%	34.69%	45.00%	56.25%	64.81%	74.47%	75.44%	85.71%
Wang2023Better_WRN-70-16	Initial Accuracy	93.20%	92.98%	100.00%	94.12%	83.67%	92.50%	89.58%	92.59%	95.74%	96.49%	94.64%
	APGD-CE	74.00%	75.44%	85.37%	66.67%	55.10%	57.50%	68.75%	77.78%	78.72%	82.46%	85.71%
	APGD-DLR	70.80%	71.93%	85.37%	64.71%	44.90%	52.50%	64.58%	74.07%	76.60%	82.46%	85.71%
	FAB	70.40%	71.93%	82.93%	64.71%	40.82%	52.50%	64.58%	74.07%	76.60%	82.46%	85.71%
Rebuffi2021Fixing_106_16_cutmix_ddpm	Initial Accuracy	88.87%	89.47%	97.56%	80.39%	79.59%	80.00%	87.50%	92.59%	95.74%	92.98%	92.86%
	APGD-CE	65.73%	66.67%	85.37%	64.71%	34.69%	40.00%	60.42%	74.07%	70.21%	75.44%	85.71%
	APGD-DLR	63.12%	64.91%	85.37%	62.75%	28.57%	35.00%	58.33%	66.67%	70.21%	75.44%	83.93%
	FAB	62.33%	64.91%	85.37%	60.78%	26.53%	35.00%	56.25%	66.67%	70.21%	73.68%	83.93%

Results – Accuracy





Observations

From our tests, we can assume that the number of parameters doesn't affect the robustness of the model.

The average accuracy after the attack of the models based on the WideResNet architecture is higher than those based on different architectures.

Bibliography

- [1] RobustBench: a standardized adversarial robustness benchmark - Croce et. Al
- [2] Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks - F. Croce and M. Hein.
- [3] The Cifar-10 dataset
- [4] Transfer Learning with ResNet in PyTorch
- [5] XCiT: Cross-Covariance Image Transformers - Voynov et. Al
- [6] Leveraging Uncertainty from Deep Learning for Trustworthy Material Discovery Workflows