max planck institut informatik

Neural Architecture Design and Robustness: A Dataset

Steffen Jung*, Jovita Lukasik*, Margret Keuper

UNIVERSITÄT SIEGEN

ICLR

Motivation / Contribution

We introduce a dataset for neural architecture design and robustness to provide the research community with more resources for analyzing what constitutes robust networks.

We borrow one of the most commonly considered search spaces for neural architecture search (NAS) for image classification, NAS-Bench-201, to evaluate all 6,466 unique architectures on a range of adversarial attacks and corruption types for our dataset.

We present three use cases:

(1) benchmark robustness measurements based on Jacobian and Hessian matrices for their robustness predictability

(2) perform neural architecture search on robust accuracies

(3) provide an analysis of how the architectural design choice affect robustness

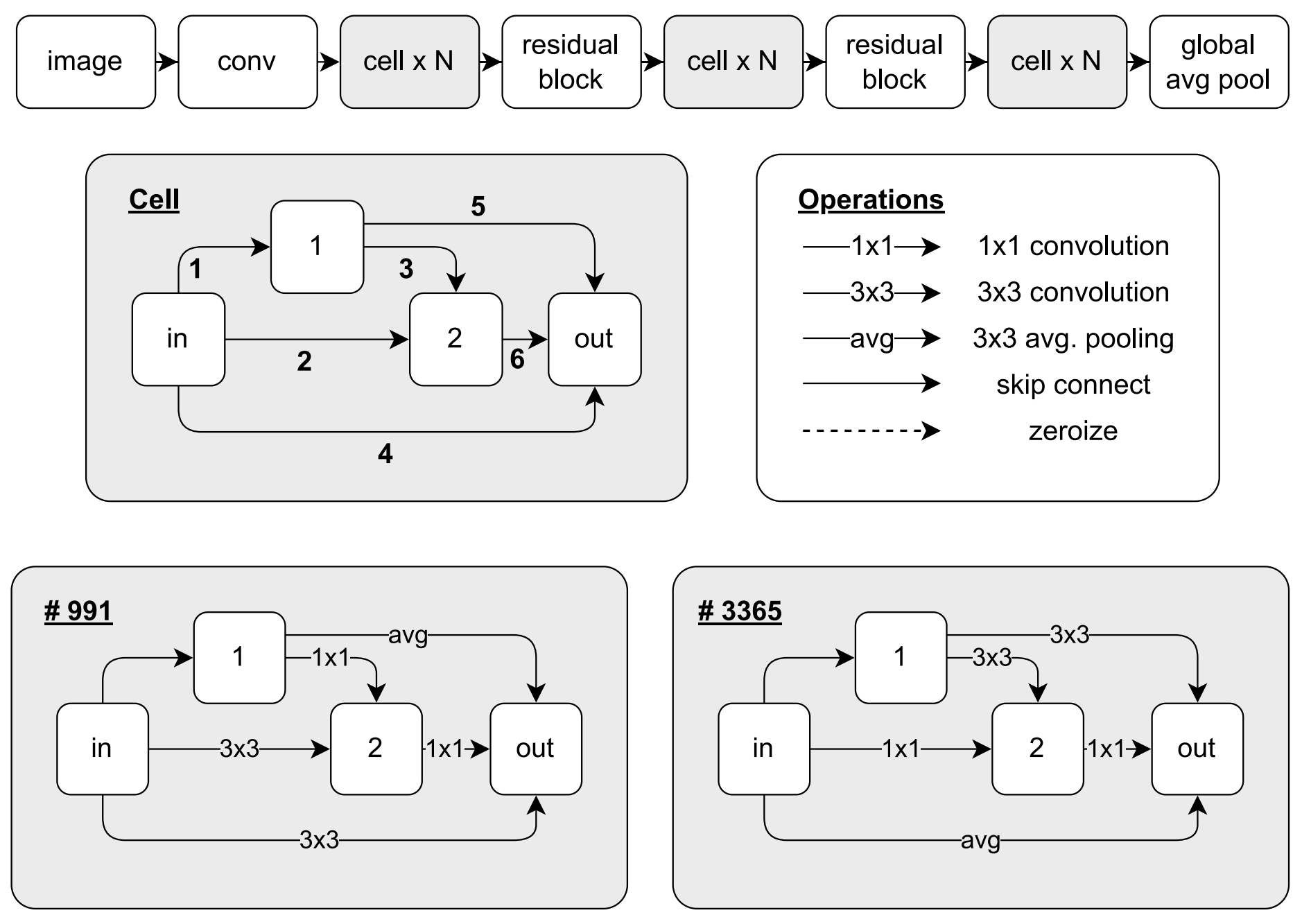
Dataset

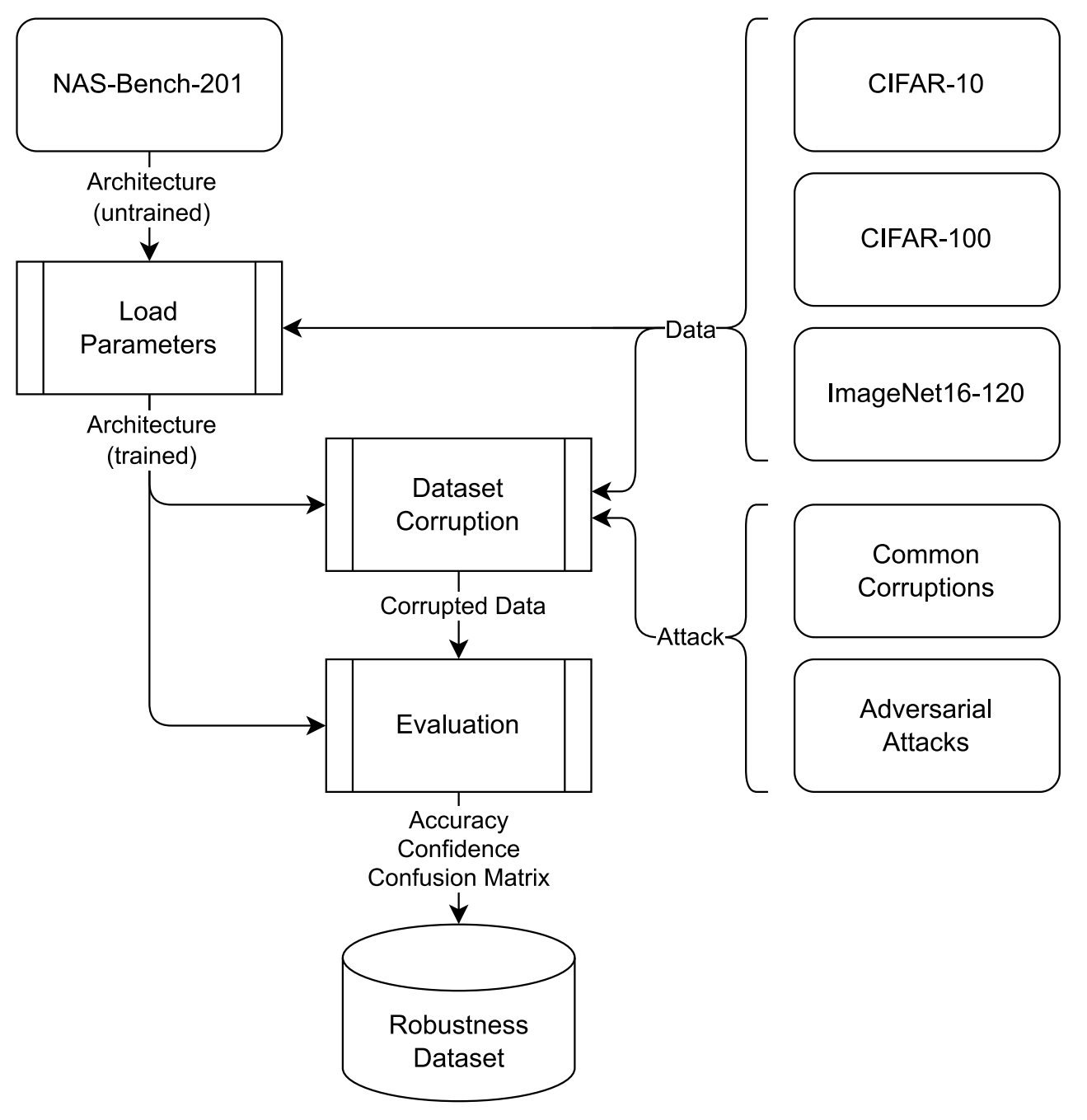
NAS-Bench-201 (Dong & Yang, 2020) is a cell-based architecture search space. Each cell has in total 4 nodes and 6 edges. The nodes in this search space correspond to the architecture's feature maps and the edges represent the architecture's operation, which are chosen from an operation set. This search space contains in total 15,625 architectures, from which only 6,466 are unique, since the operations skip and zeroize can cause isomorphic cells. Each architecture is trained on three different image datasets for 200 epochs: CIFAR-10/100 and ImageNet16-120. We evaluate all 3 · 6,466 = 19,398 pretrained networks for different adversarial attacks and common corruptions. We collect:

(a) accuracy,

(b) average prediction confidences, and

(c) confusion matrices.





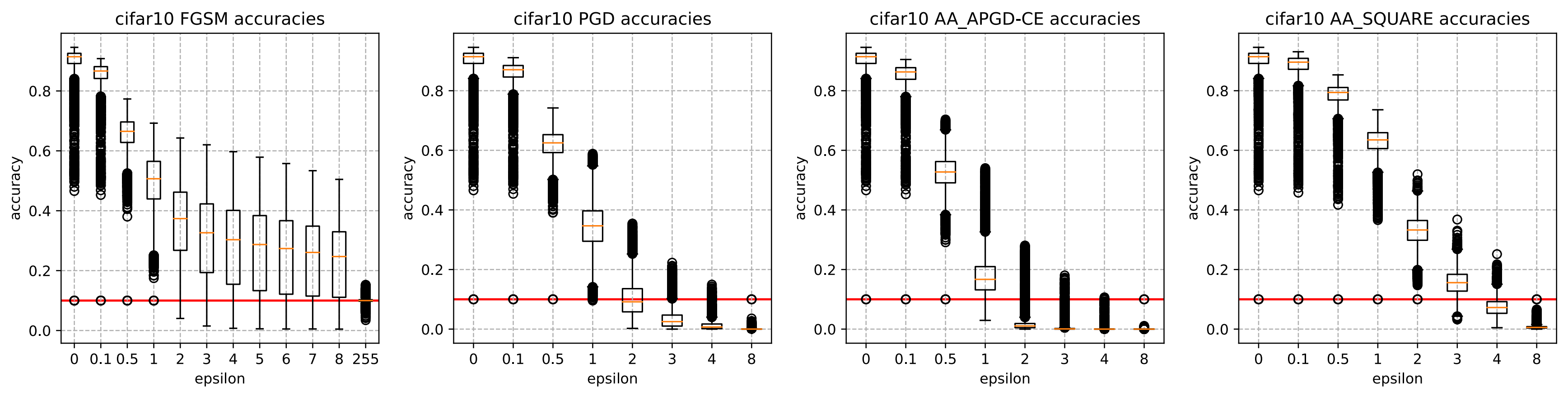
Adversarial Attacks

We collect evaluations on different adversarial attacks, namely FGSM, PGD, APGD, and Square Attack on the L_∞ norm.

Summary

The plots here show aggregated evaluation results on the mentioned attacks on CIFAR-10 w.r.t. accuracy. Growing gaps between mean and max accuracies indicate that architecture design has an impact on robust performances.

Attack	Hyperparameters
FGSM	$\epsilon \in \{.1, .5., 1, 2, 3, 4, 5, 6, 7, 8, 255\}/255$
PGD	$\epsilon \in \{.1, .5., 1, 2, 3, 4, 8, 255\}/255$ $\alpha = 0.01/0.3$ 40 attack iterations
APGD	$\epsilon \in \{.1, .5., 1, 2, 3, 4, 8, 255\}/255$ 100 attack iterations
Square	$\epsilon \in \{.1, .5., 1, 2, 3, 4, 8, 255\}/255$ 5 000 search iterations



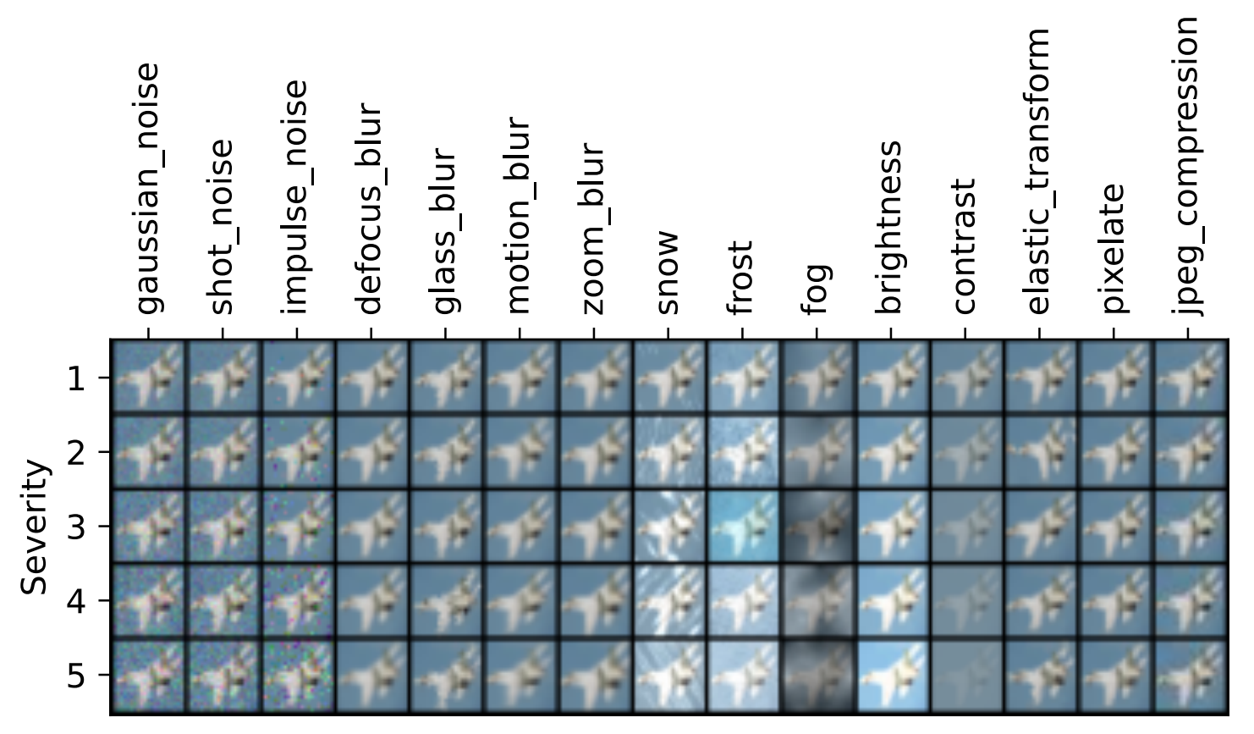
Common Corruptions

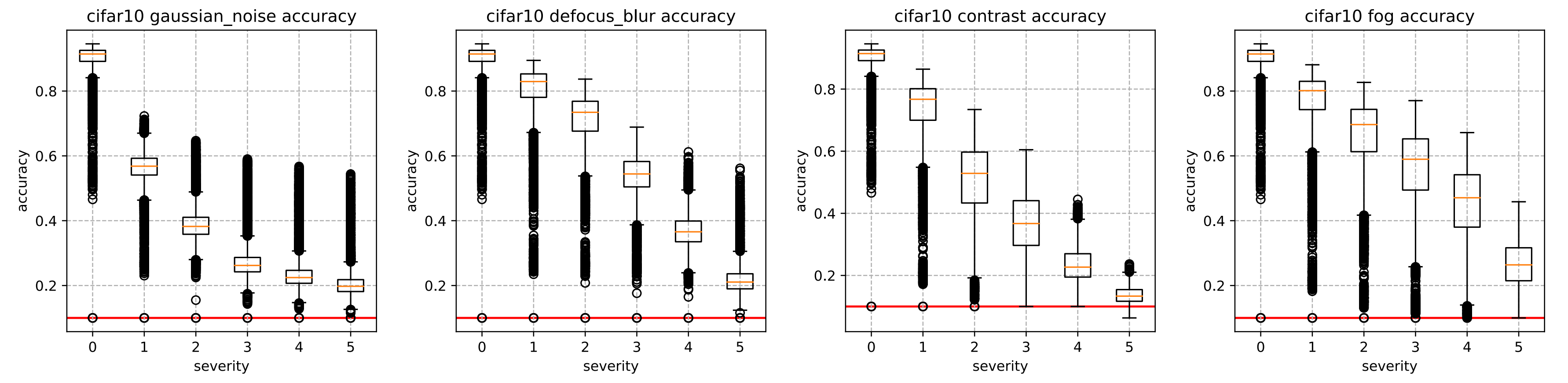
To evaluate all unique NAS-Bench-201 architectures on common corruptions, we evaluate them on the benchmark data provided by Hendrycks & Dietterich (2019).

Two datasets are available: CIFAR10-C, which is a corrupted version of CIFAR-10, and CIFAR-100-C, which is a corrupted version of CIFAR-100. Both datasets are perturbed with a total of 15 corruptions at 5 severity levels. The training procedure of NAS-Bench-201 only augments the training data with random flipping and random cropping. Hence, no influence should be expected of the training augmentation pipeline on the performance of the networks to those corruptions. We evaluate each of the 15 · 5 = 75 datasets individually for each network.

Summary

The plots depict mean accuracies for different exemplary corruptions at increasing severity levels. Similar to the results for adversarial attacks, a growing gap between mean and max accuracies for most of the corruptions can be observed, which indicates towards architectural influences on robustness to common corruptions.





Use-Case 1: Robustness Measurements

Jacobian

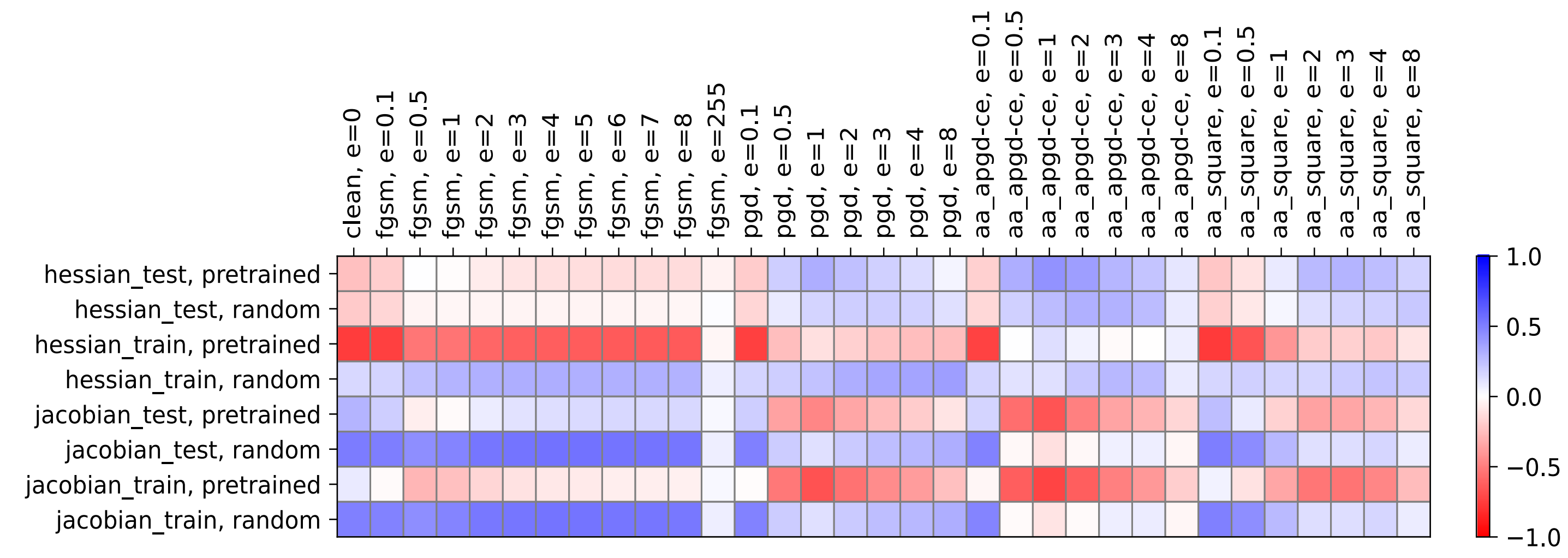
Hoffman et al. (2019) introduce an efficient Jacobian regularization method to improve the robustness of neural architectures. The goal is to minimize the network's output change in case of perturbed input data, by minimizing the Frobenius norm of the network's Jacobian matrix. This is based on the fact that the larger the Jacobian components, the larger is the output change for perturbed input data, and thus the more unstable is the neural network against this perturbed input data. In order to increase the stability of the network, Hoffman et al. (2019) proposes to decrease the Jacobian components by minimizing the square of the Frobenius norm of the Jacobian. The smaller the Frobenius norm of the Jacobian of a network, the more robust the network is supposed to be.

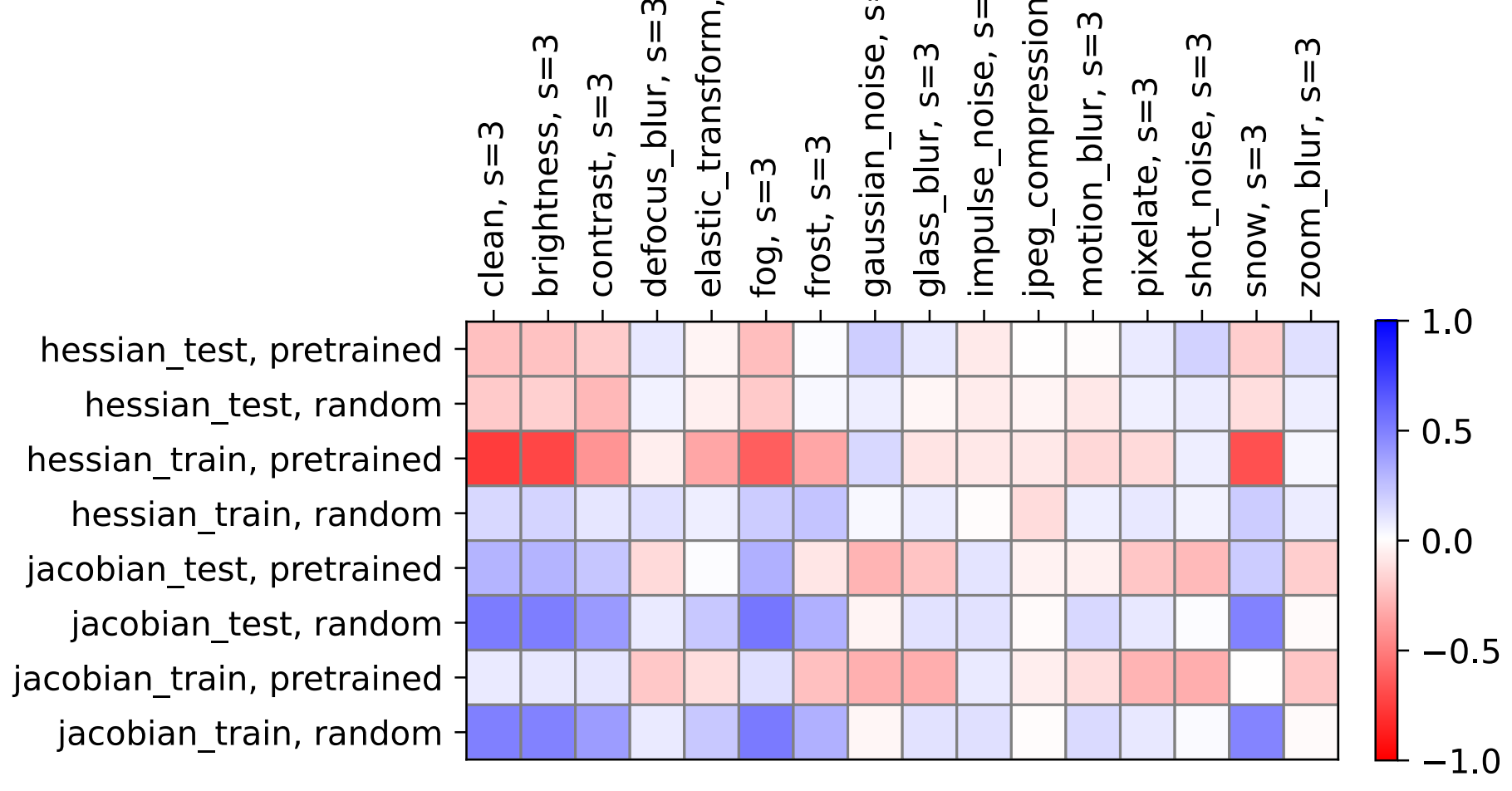
Hessian

Zhao et al. (2020) investigate the loss landscape of a regular neural network and robust neural network against adversarial attacks. They provide theoretical justification that the adversarial loss is highly correlated with the largest eigenvalue of the input Hessian matrix of the clean input data. Therefore the eigenspectrum of the Hessian matrix of the regular network can be used for quantifying the robustness: large Hessian spectrum implies a sharp minimum resulting in a more vulnerable neural network against adversarial attacks. Whereas in the case of a neural network with small Hessian spectrum, implying a flat minimum, more perturbation on the input is needed to leave the minimum.

Summary

We can observe that the Jacobian-based measurement correlates well with rankings after attacks by FGSM and smaller ε values for other attacks. However, this is not true anymore when ε increases, especially in the case of APGD. We can observe similar behaviour for the Hessian-based measurement.





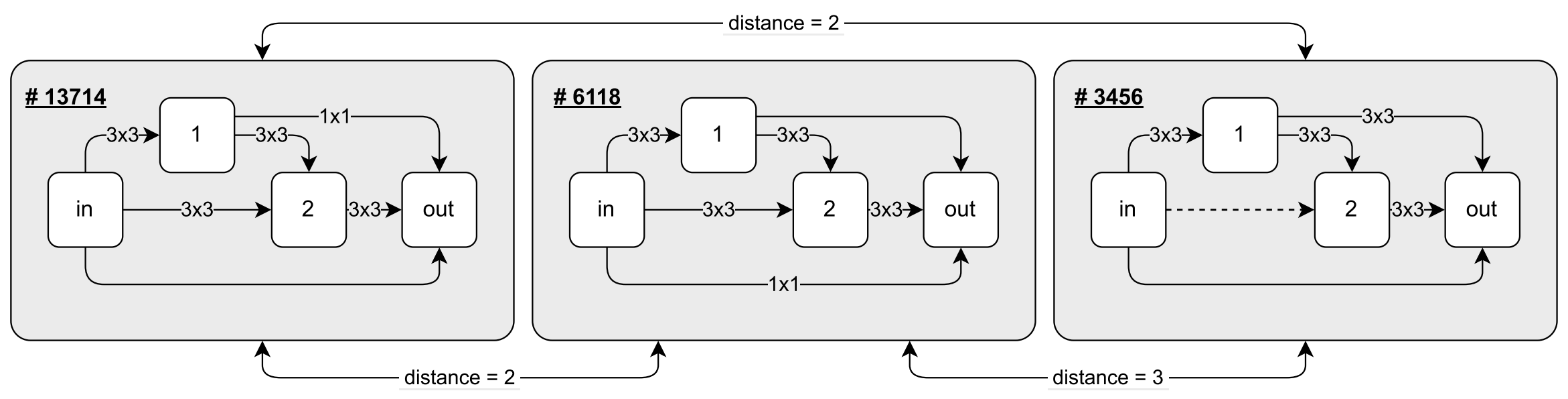
Use-Case 2: Neural Architecture Search

We perform different SoTA NAS algorithms on the clean accuracy and the FGSM (ε = 1) robust accuracy and evaluate the best found architectures on all evaluated adversarial attacks. Although clean accuracy is reduced, the overall robustness to all adversarial attacks improves when the search is performed on FGSM (ε = 1) accuracy. Local Search achieves the best performance, which indicates that localized changes to an architecture design seem to be able to improve network robustness.

	Method	Test Accuracy ($\epsilon = 1.0$) Clean	FGSM	PDG	APGD	Squares	Clean
		CIFAR-10					CF-10-C
	Optimum	94.68	69.24	58.85	54.02	73.61	58.55
Clean	BANANAS (White et al., 2021a)	94.21	64.25	41.10	18.62	68.69	55.52
	Local Search (White et al., 2021b)	94.65	63.95	41.17	18.74	69.59	56.90
	Random Search (Li & Talwalkar, 2019)	94.22	63.38	40.09	17.84	68.40	55.60
	Regularized Evolution (Real et al., 2019)	94.53	63.30	40.23	18.11	68.92	56.21
FGSM	BANANAS (White et al., 2021a)	93.52	66.35	45.59	20.72	68.01	54.88
	Local Search (White et al., 2021b)	93.86	69.10	48.27	23.18	69.47	56.57
	Random Search (Li & Talwalkar, 2019)	93.57	67.25	46.15	20.93	68.44	55.10
	Regularized Evolution (Real et al., 2019)	93.77	68.82	47.99	22.59	69.20	56.11

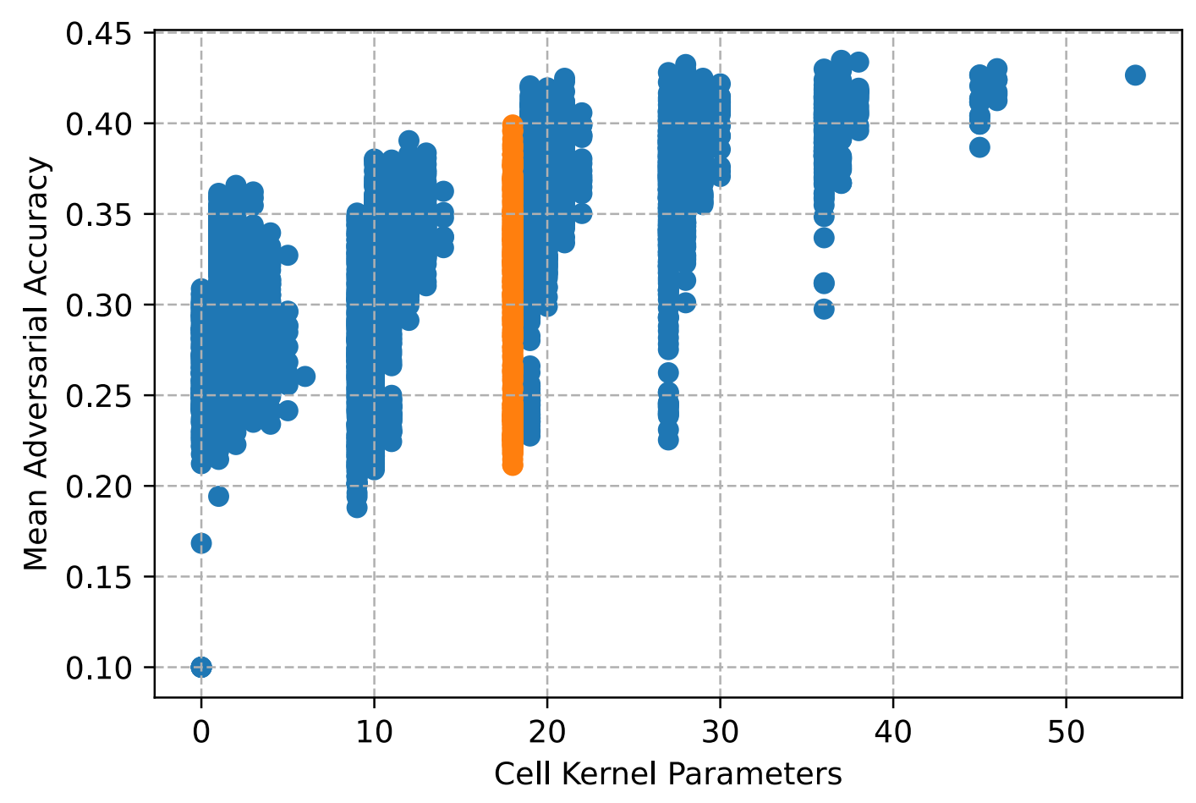
Use-Case 3: Analysis

Visualization of the best architectures in the NAS-Bench-201 search space in terms of clean accuracy, mean adversarial accuracy, and mean common corruption accuracy on CIFAR-10 and their respective edit distances.

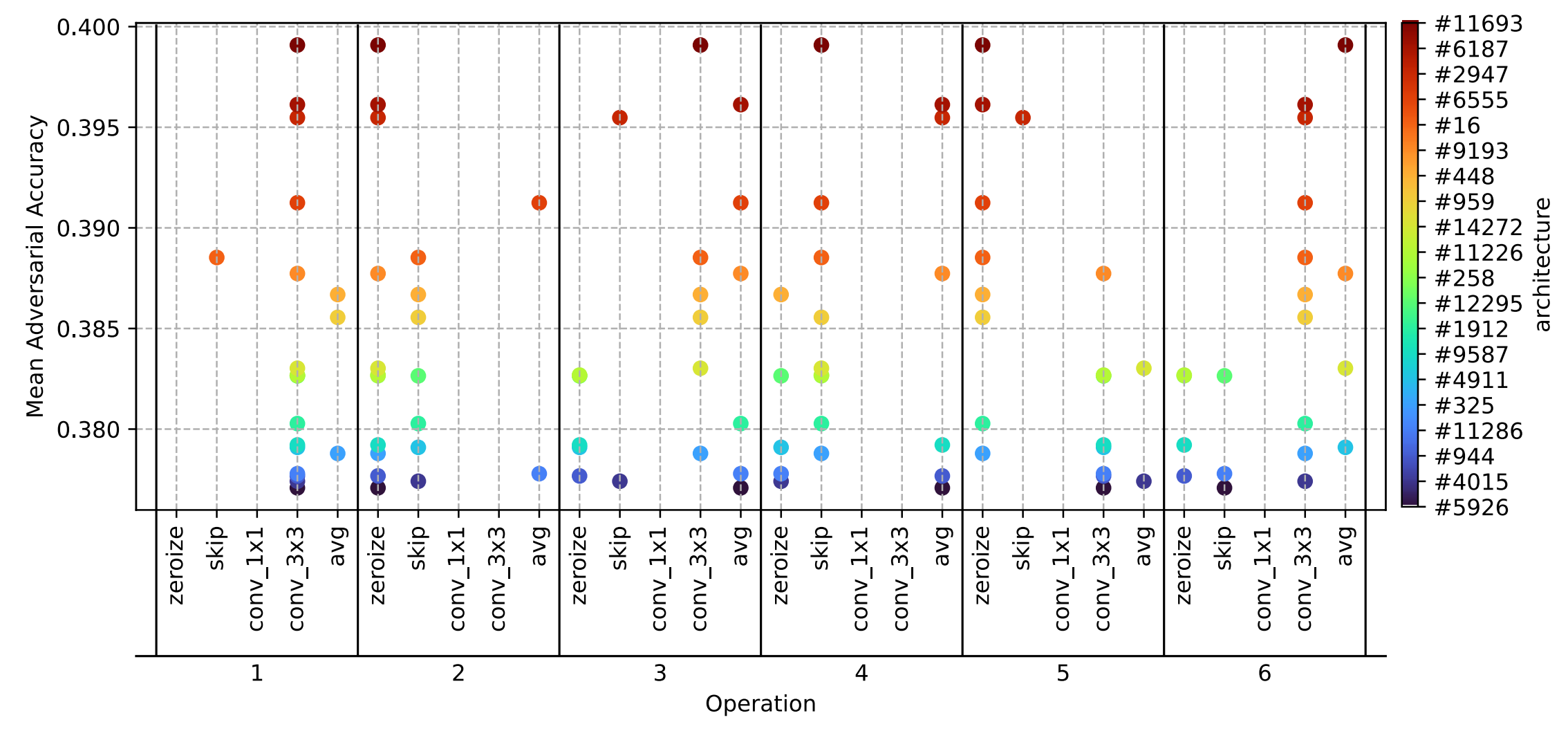


Same parameter count on CIFAR-10

Networks with parameter count 18 (408 instances in total having exactly 2 times 3×3 convolutions and no 1×1 convolutions in a cell) are highlighted in orange. As we can see, there is a large range of mean adversarial accuracies [0.21, 0.4] on CIFAR-10 for these networks, showing the potential of doubling the robustness of an architecture design with the same parameter budget by carefully crafting its topology.

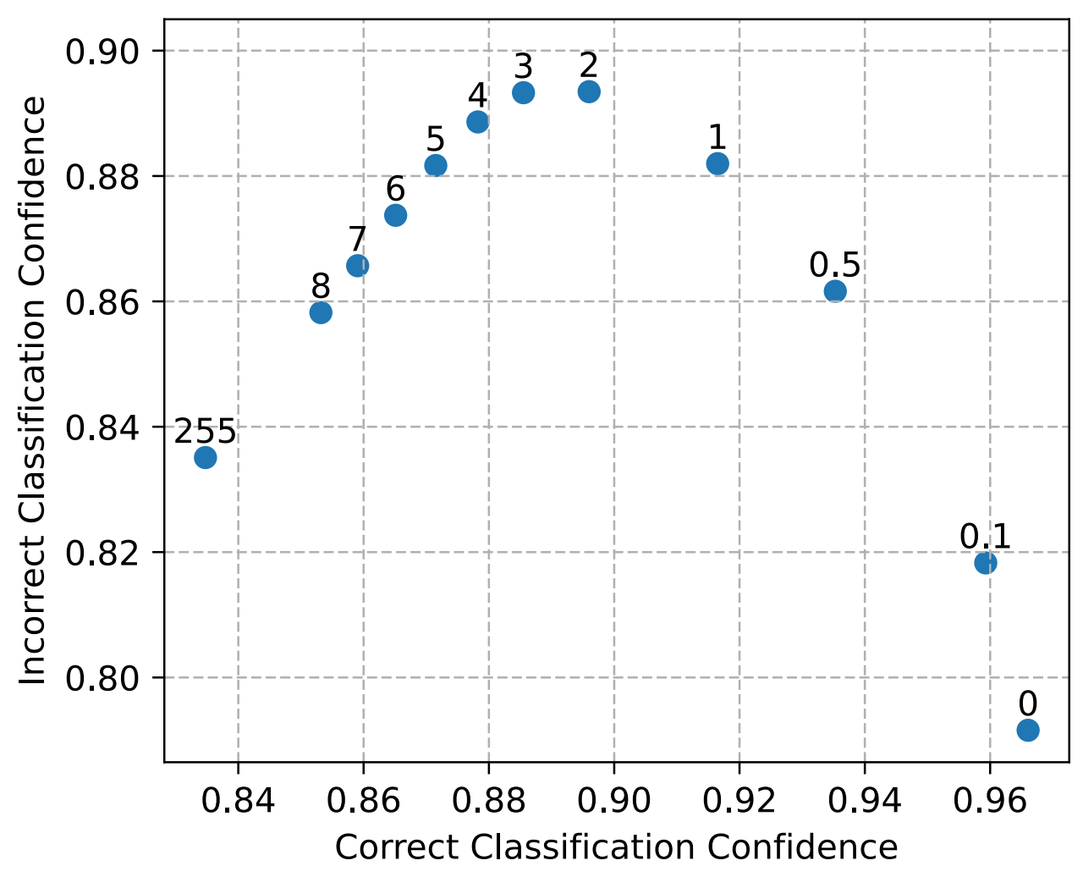


The top-20 performing architectures (color-coded, one operation for each edge) with parameter count 18 have no convolutions on edges 2 and 4, and no dropping or skipping of edge 1. In the case of edge 4, it seems that a single convolutional layer connecting input and output of the cell increases sensitivity of the network. Hence, most of the top-20 robust architectures stack convolutions.



Confidences

Mean prediction confidence scores on FGSM-attacked CIFAR-10 images for different ε (on top of points) for all non-isomorphic networks in NAS-Bench-201. Networks become less confident in their prediction if their prediction is correct when ε increases. Networks become more confident in their prediction if their prediction is incorrect, however, only up to a certain ε value. When ε further increases, confidence drops again



Conclusion

- We showed that these architecture measurements are a good first approach for the architecture's robustness, but have to be taken with caution when the perturbation increases.
- NAS directly on the robust accuracies indeed finds more robust architectures for different adversarial attacks.
- An initial analysis of architectural design showed that it is possible to improve robustness of networks with the same number of parameters by carefully designing their topology.

References

- Xuanyi Dong, Yi Yang. NAS-Bench-201: Extending the scope of reproducible neural architecture search. In ICLR, 2020.
- Dan Hendrycks, Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In ICLR, 2019.
- Judy Hoffman, Daniel A. Roberts, Sho Yaida. Robust learning with jacobian regularization. Corr, abs/1908.02729, 2019
- Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. In ICLR, 2020.

Visit our project page for code and data (it's interactive!)