Fragility, robustness, and antifragility in deep neural networks

Dr Varun Ojha
School of Computing, Newcastle University
Varun.Ojha@newcastle.ac.uk

@

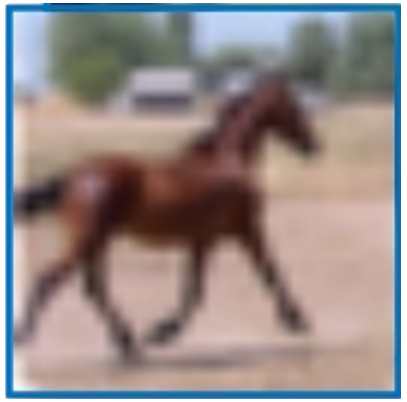
AI UK Fringe Event, Loughborough University

26/03/2024

Adversarial Attacks

Calculated using Deep Neural Networks (DNNs) weights (white-box attack)

x

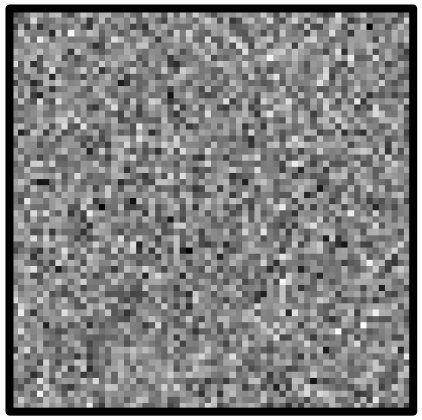


Input example
Predicted as '*Horse*'

Perturbation magnitude

$+ \quad \varepsilon \quad *$

δ



Adversarial perturbation
('*Plane*' class)

$=$

x_ε

Adversarial example
Predicted as '*Plane*'

The general premise of a robustness analysis is to subject DNNs to the '**worst case**' conditions and evaluate the *ability for a DNN to remain invariant* under such settings.

Challenges for DNN robustness

- DNNs are susceptible to adversarial attacks and thus any DNN prediction can be unreliable and vulnerable to an adversary.

- How each component of a DNN behaves due to an adversarial attack is a lesser known are of research.

- Adversarial attacks on DNNs has been well studies on state-of-the-art datasets, however, adversarial attacks on DNNs and their remedies has rarely been studied extensively.

What can we promise for DNN robustness?

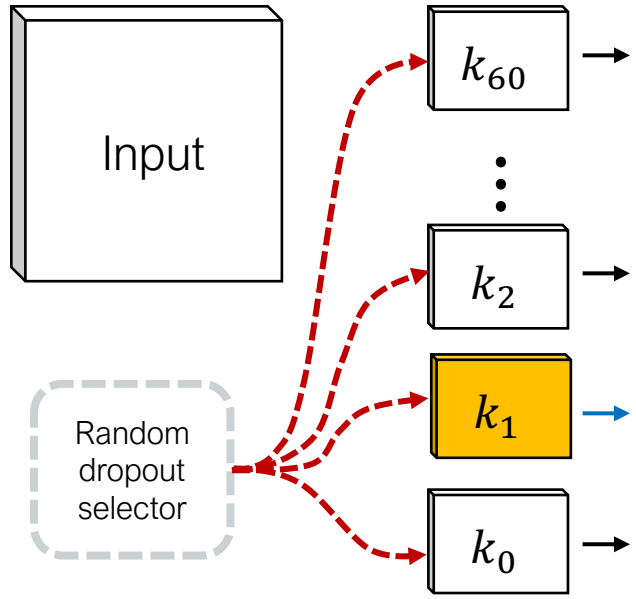
- We can use adversarial attacks to identify the strengths and weaknesses of DNN architectures.
- Upon identifying the strengths and weaknesses of DNN architectures we can improve the performance of DNNs against both adversarial attacks and the clean dataset.
- DNNs robustness analysis can develop stronger networks that are capable of performing under sub-optimal conditions.

How can we ensure DNN robustness?

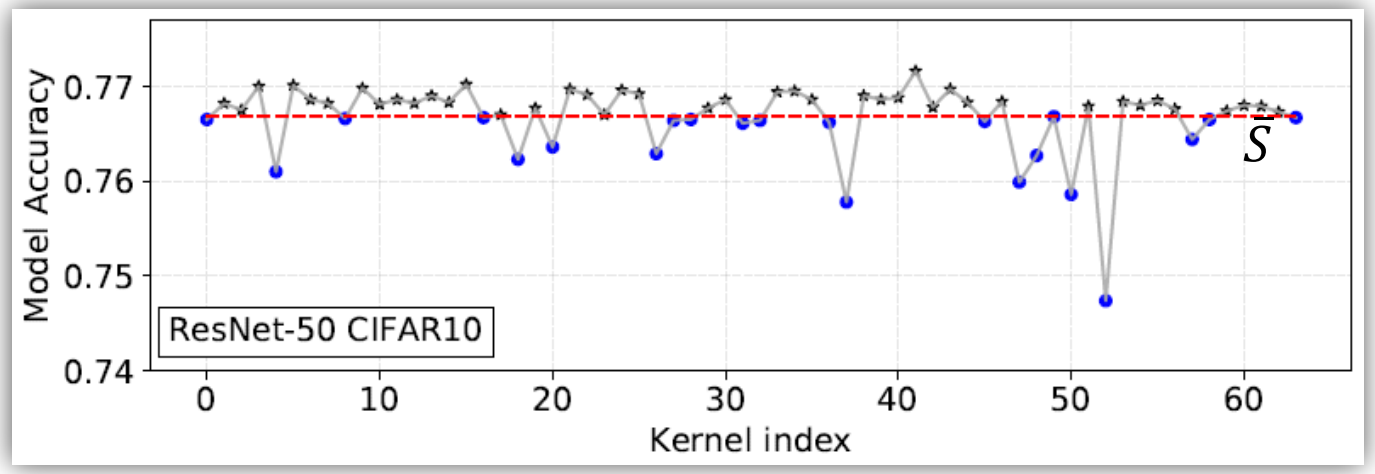
- Establish the relationship between DNN parameters and adversarial attacks to identify parameters that are targeted by the adversary.
- Formalise the notions of DNN parameter perturbations and adversarial attacks as internal and external stressors on DNNs.
- Define fragility, robustness, and antifragility in DNN to encapsulate parameter characterisations and
- Evaluate the effects of only re-training parameters characterised as robust and antifragile (selective backpropagation).

Attacks on fragile neurons

We remove kernel from the first convolutional layer and define fragile nodes to be all nodes that reduce the model performance on the test set to be below the mean dropout performance.



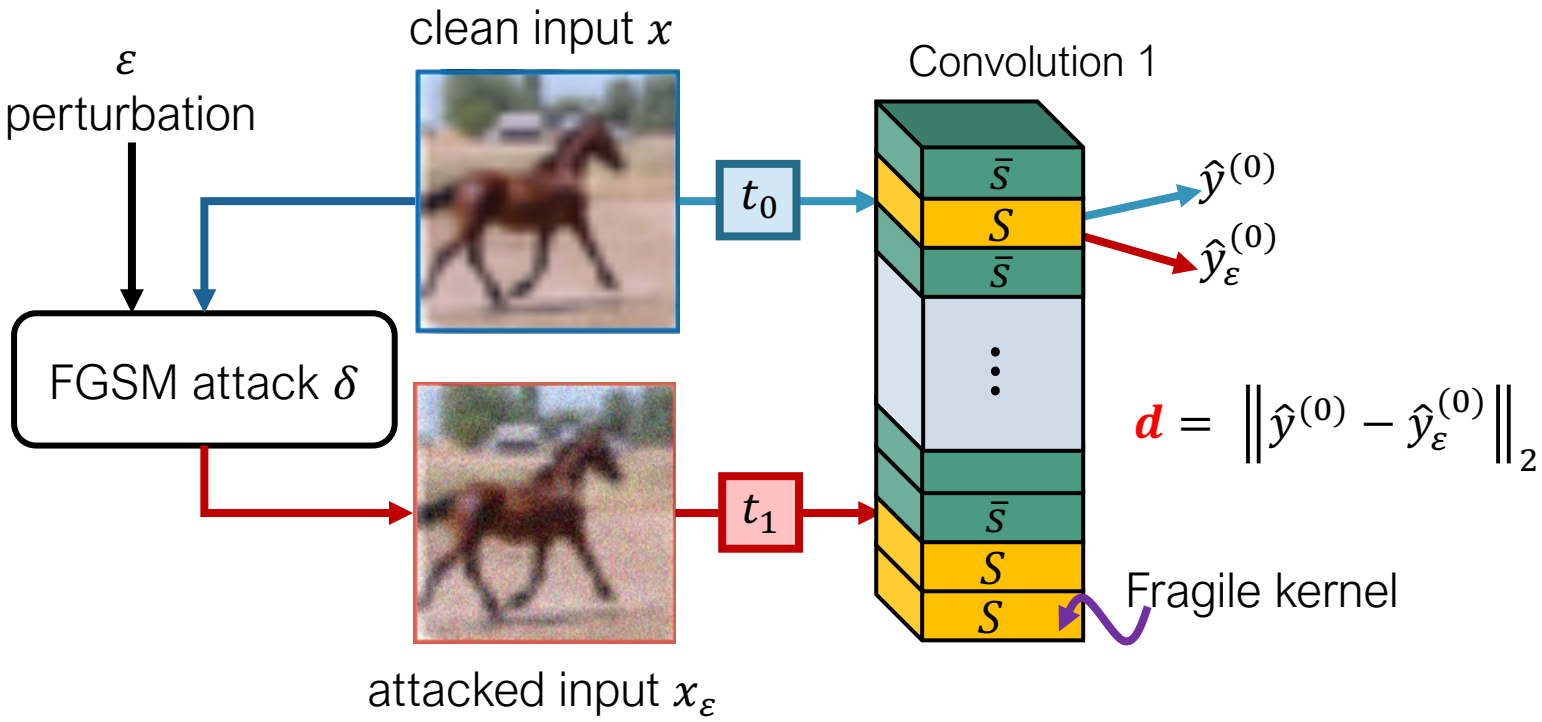
Nodal Dropouts

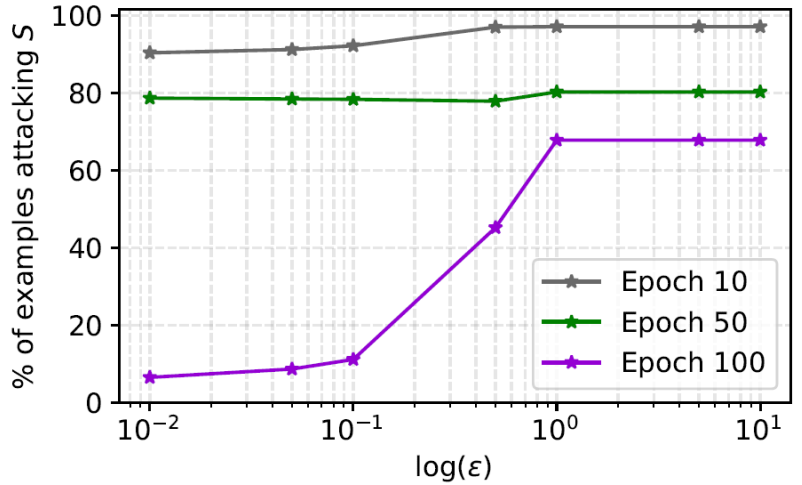


Fragile kernels (nodes) shown in blue (•) below mean/baseline DNN performance line in red and **null kernels** are shown in black star (★) above mean line in red

Adversarial targeting algorithm

We outline an algorithm to measure the **average magnitude difference** $\boldsymbol{d}$ **at the output** of the first convolutional layer, between fragile and non-fragile neurons, on both clean and adversarial inputs.

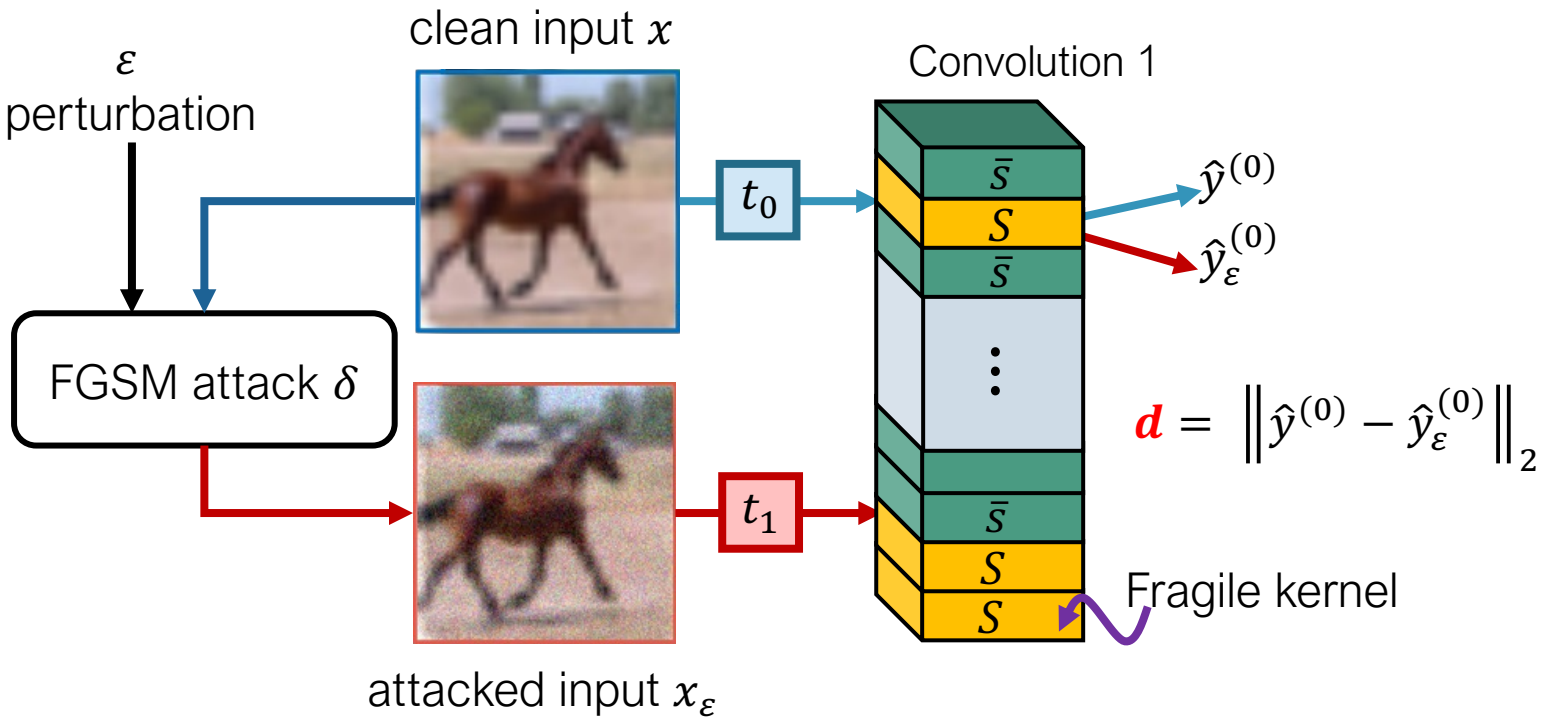


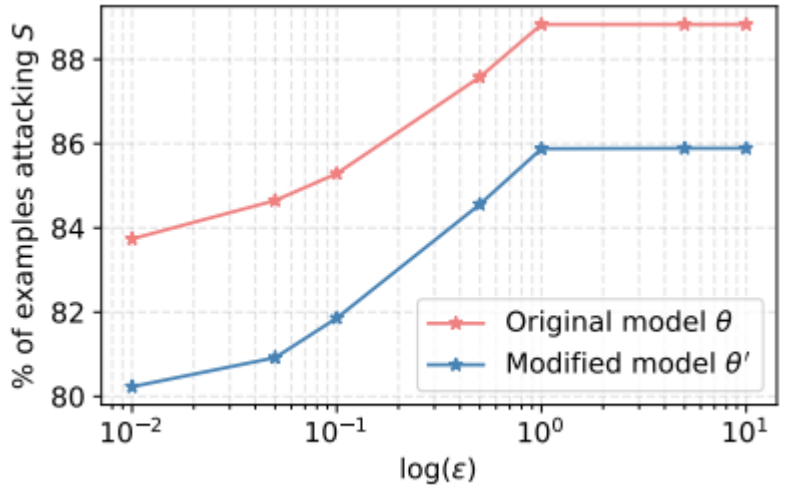


if avg. distance of fragile kernels S
greater than
avg. distance of non-fragile kernels $\bar{s}$
then
x_ε attacks fragile kernels

Adversarial targeting algorithm

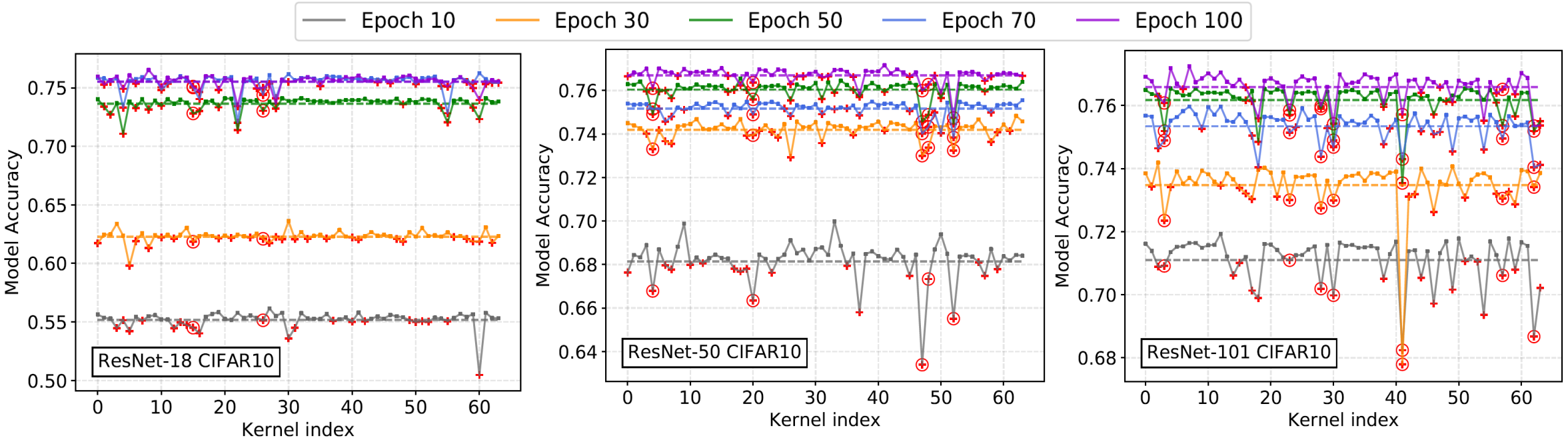
We outline an algorithm to measure the **average magnitude difference** $\boldsymbol{d}$ **at the output** of the first convolutional layer, between fragile and non-fragile neurons, on both clean and adversarial inputs.





if avg. distance of fragile kernels S
greater than
avg. distance of non-fragile kernels $\bar{s}$
then
x_ε attacks fragile kernels

Fragile kernels/neurons

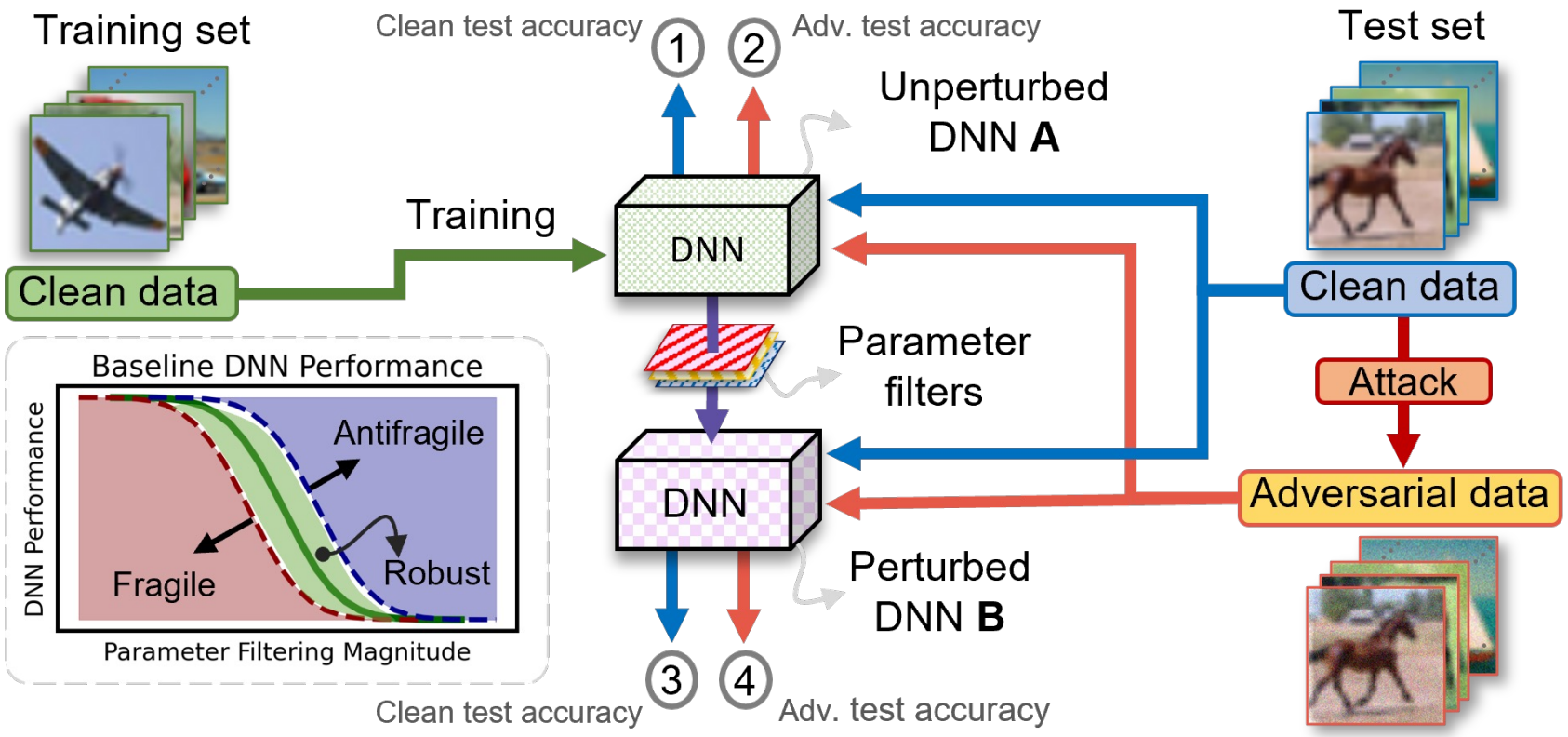


Red crosses (+) represent fragile kernels and red circles around red crosses (⊕) represent kernels that have shown to be consistently fragile throughout the training phase for each model.

Key outcomes of adversarial targeting

- We show that, using our nodal dropout method **some neurons are fragile and other neurons are non-fragile** when considering network performance.
- We show that neurons identified as fragile are targeted more by an adversarial attack.
- We find that larger DNNs proportionally have **more fragile neurons**.
- We establish a relationship between fragile neurons and adversarial attacks to analyse of fragile, robust and antifragile parameters .

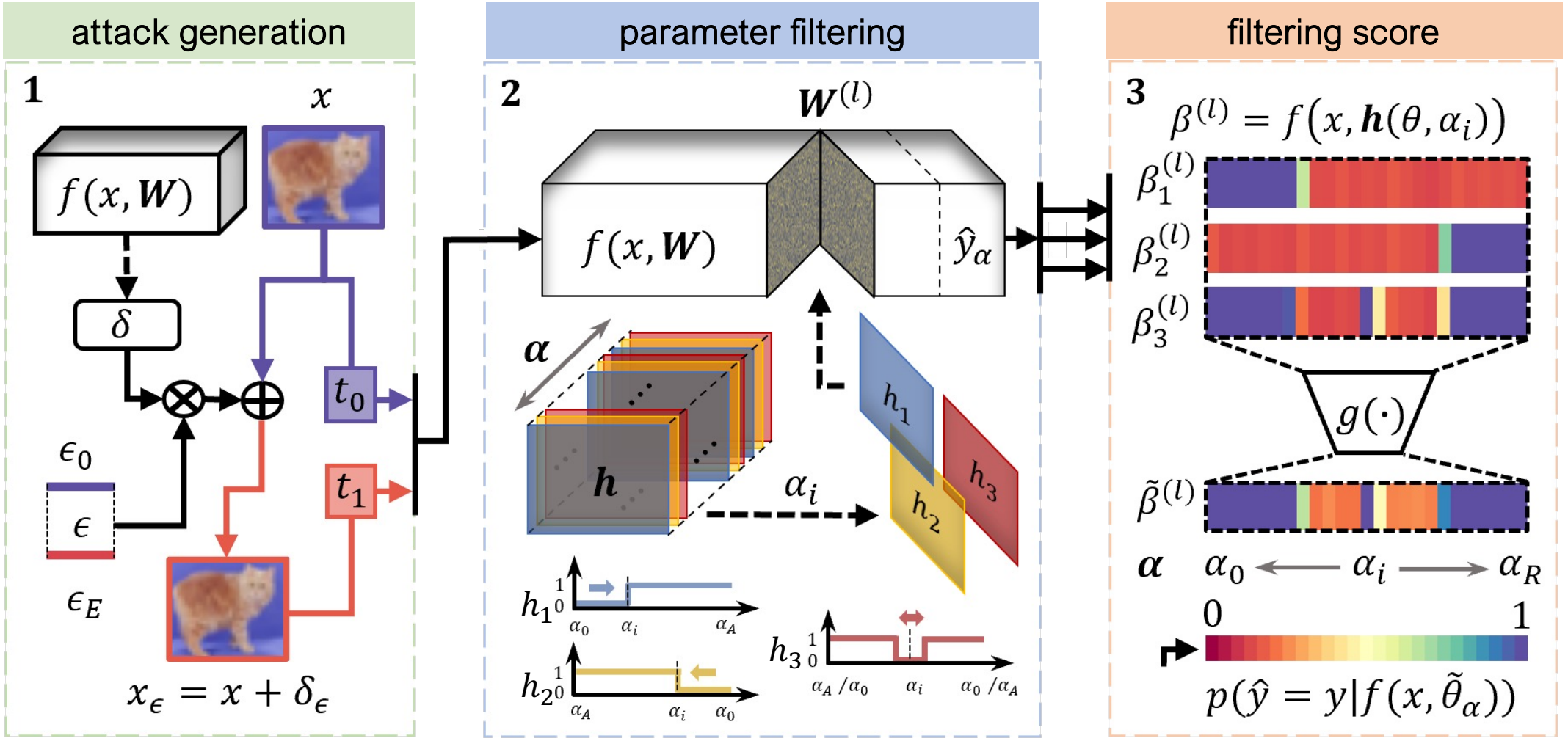
Fragility, robustness and antifragility



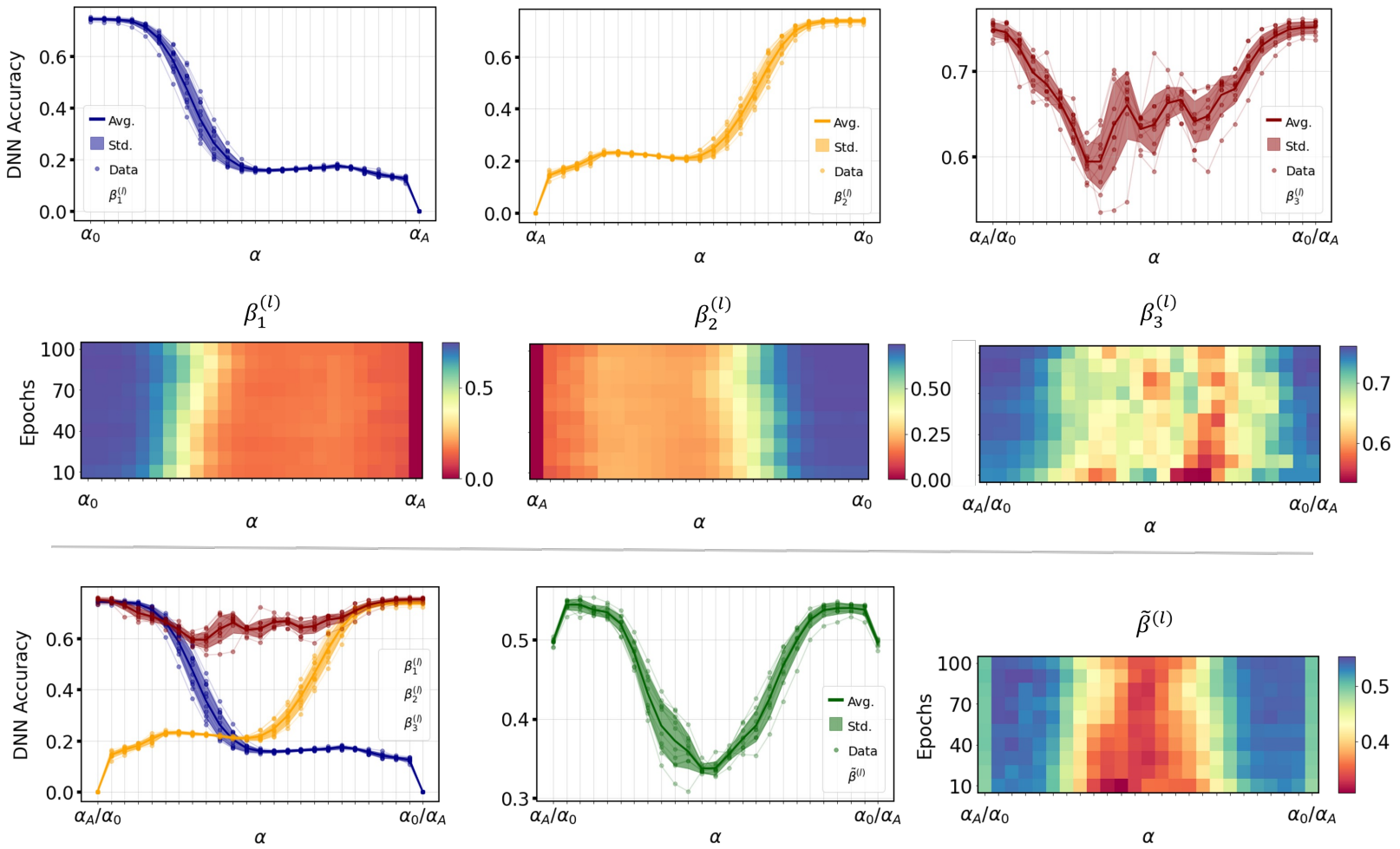
- a new method of parameter filtering (**synaptic filtering**)
- **synaptic filtering of all layers and parameters** of a DNN architecture.
- **compare clean and adversarial performance** of a regular DNN and perturbed DNN.
- **characterise** parameters as fragile, robust, and antifragile

Synaptic filtering algorithm

$$h_1(\theta, \alpha_i) = \begin{cases} 0 & \text{if } \theta \leq \alpha_i, \\ 1 & \text{otherwise} \end{cases}$$



Learning landscape



The influence of parameters varies as the network is trained and learns more dataset features.

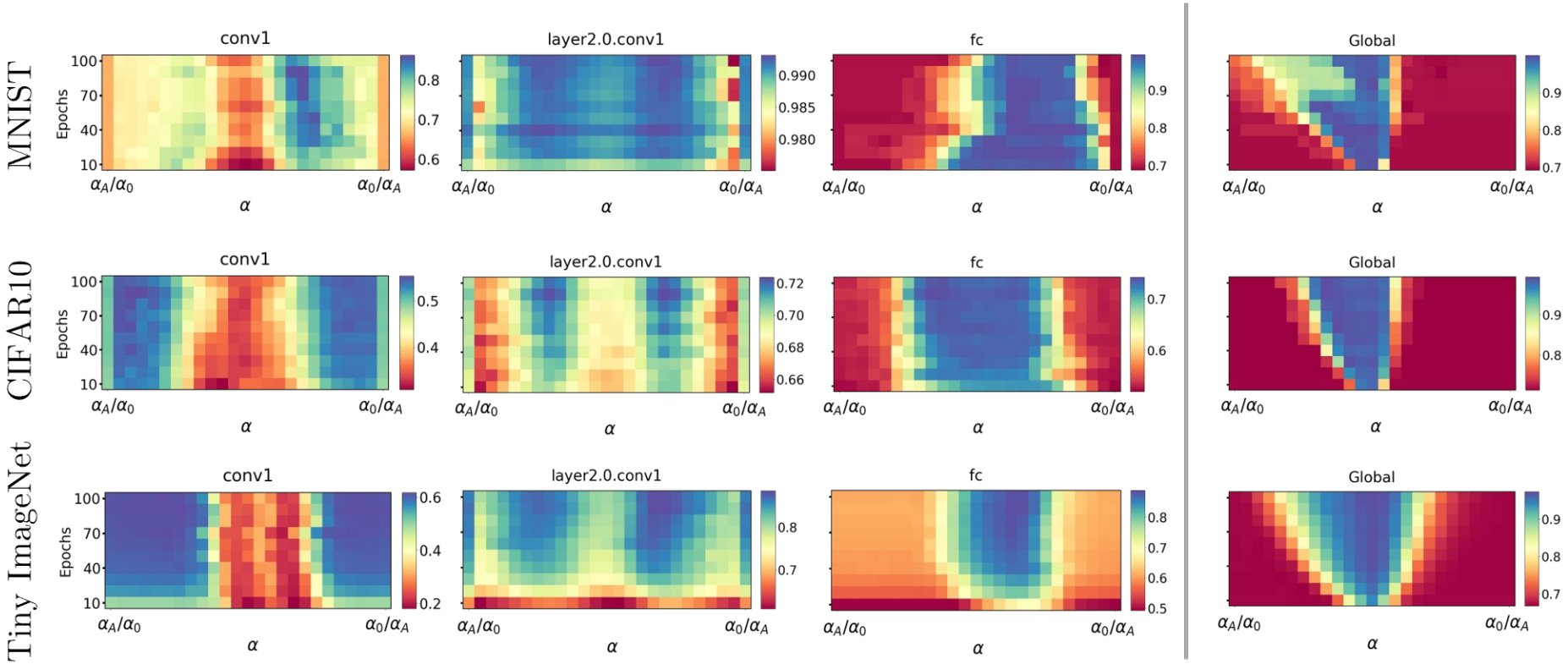
The three different filters h_1, h_2, and h_3 highlight different parameters as influential (■) and not influential (■) to DNN performance.

The combined performance highlights the parameters that are most influential (■) using all the three different filters.

Learning landscape

We show that the **same layer of a DNN has similar learning landscapes** for different datasets.

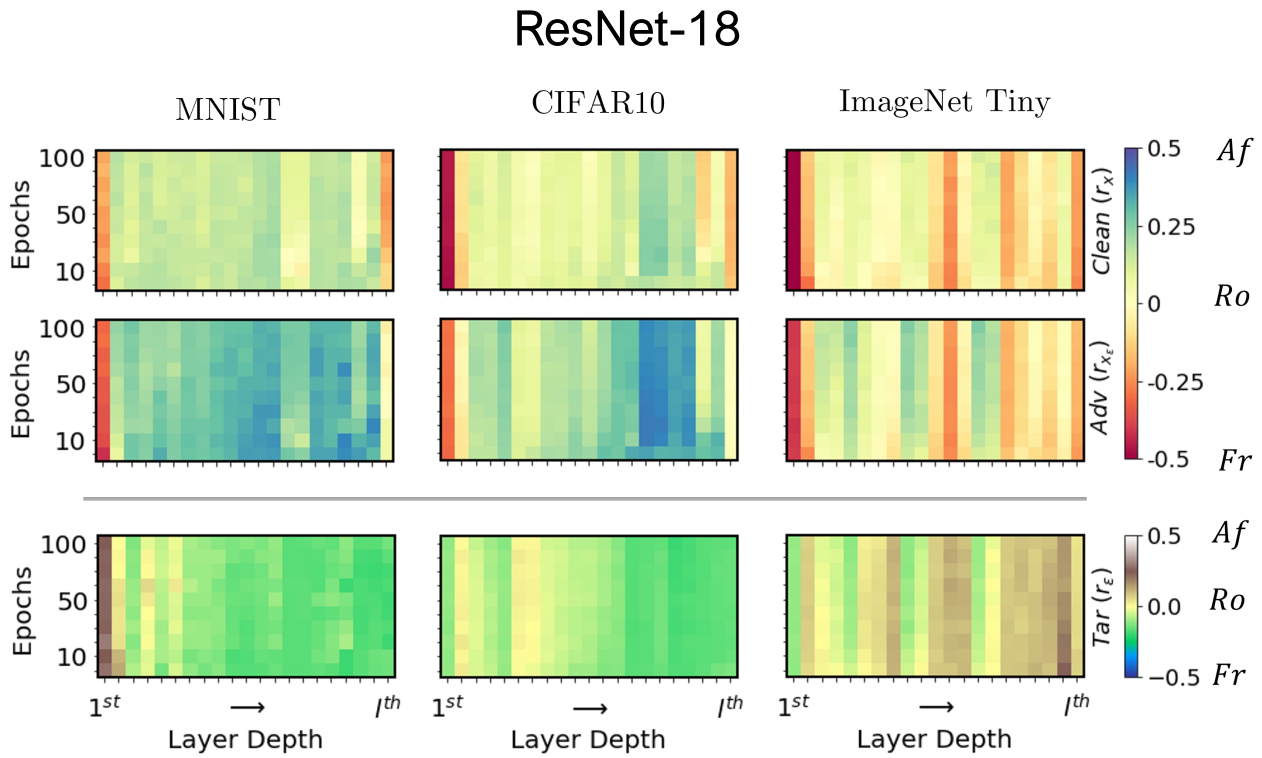
This shows that there are **invariant characteristics of DNN architectures**, even when applied to different datasets.

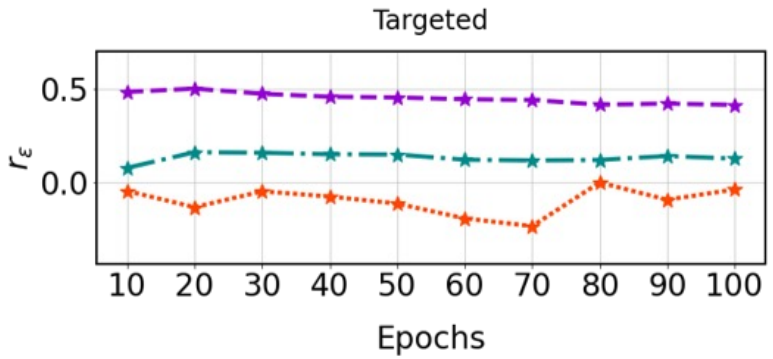


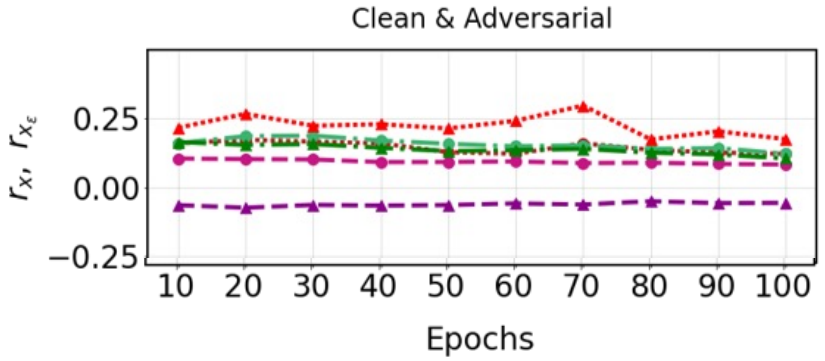
Different layers in the network show to have different characteristics when subjected to the parameter filters (internal stressor). The results are the combined responses using filters h_1, h_2, and h_3.

Parameter Scores

For the **local layer-wise** parameter scores we show that, throughout network training, certain layers are consistently targeted by the adversary.



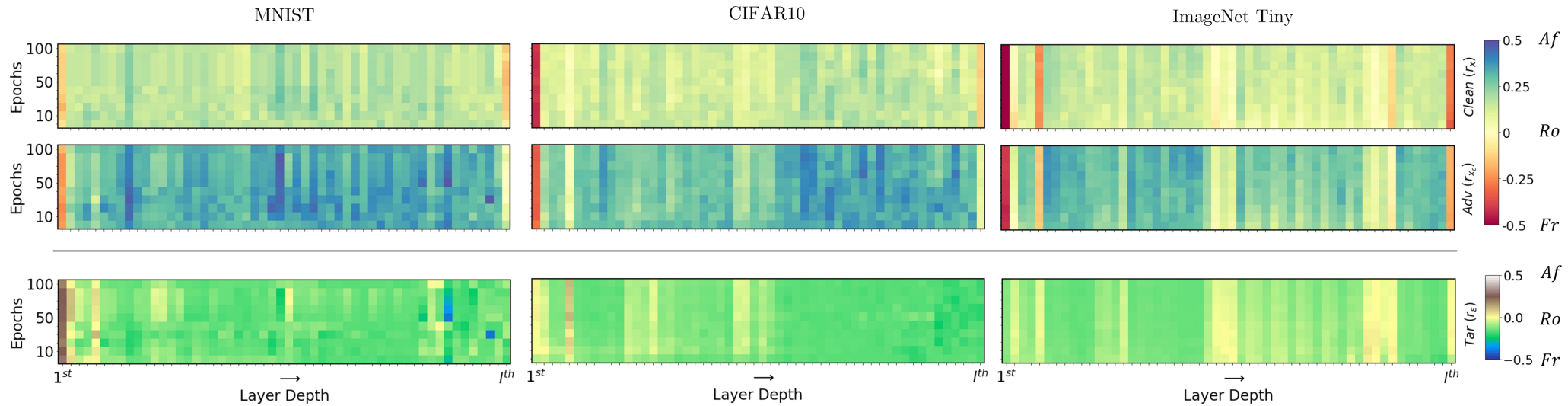




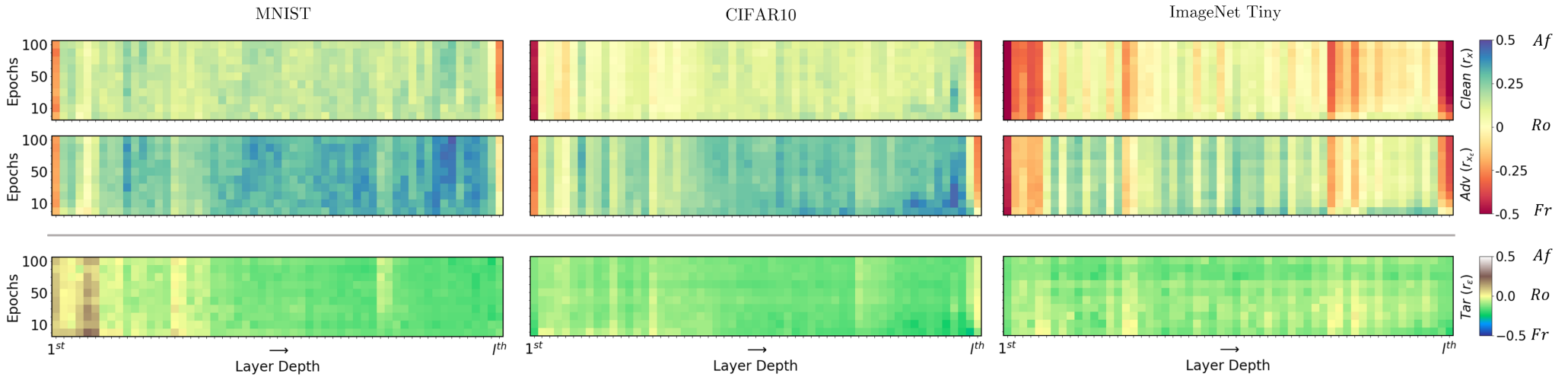
Global parameter scores show the influence of all network parameters being filtered.

Parameter Scores (layer-wise)

ResNet-50



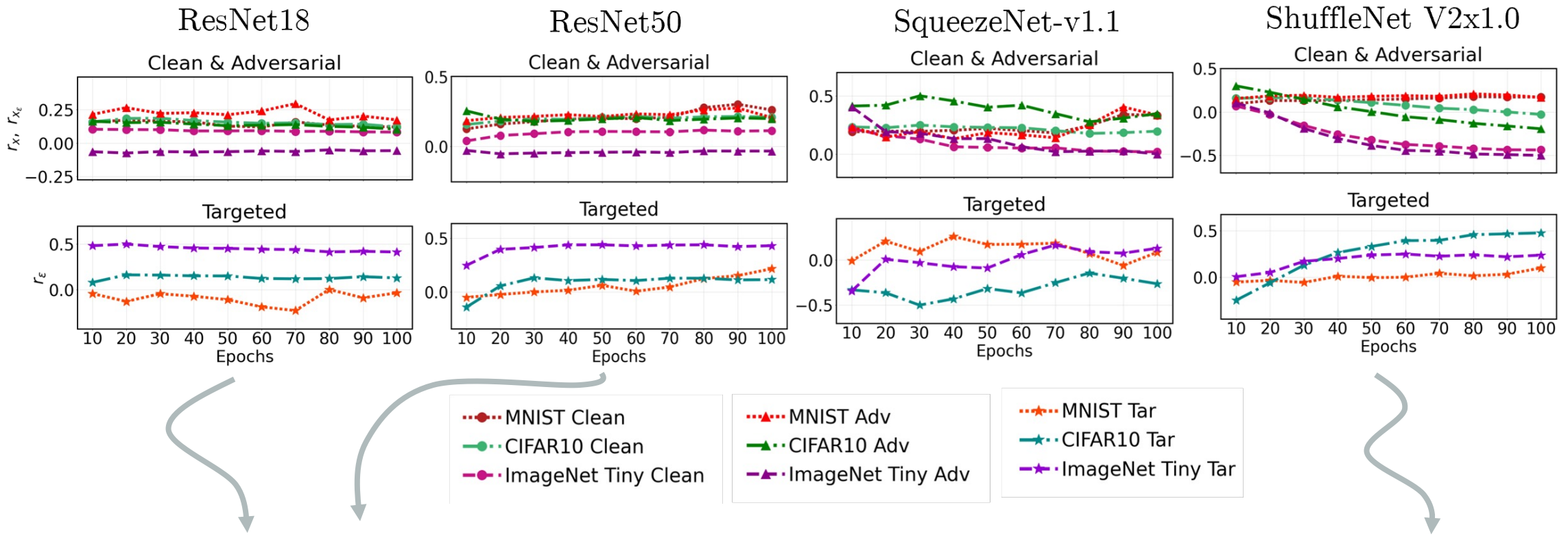
ShuffleNet V2 x1.0



Periodic parameter characterisation shown for some networks.

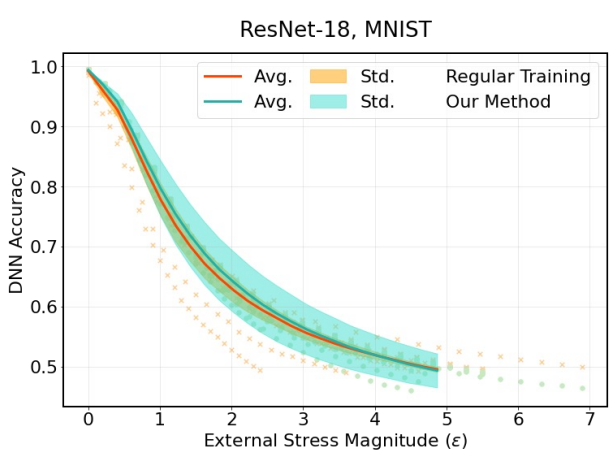
We say that fragile parameters are important to network performance.

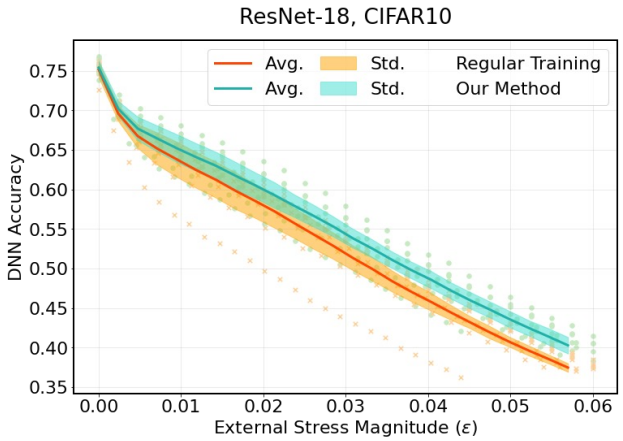
Parameter Scores (full network)

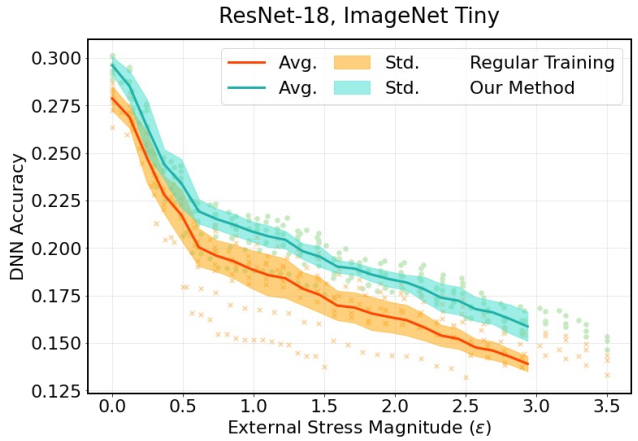


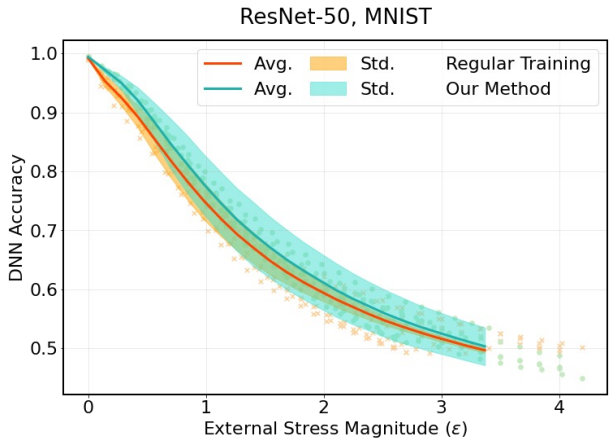
- Similar scores through training for similar architectures (ResNet-18 and ResNet-50)
- We can compare different network architectures applied to different datasets
- Converging as the network is trained

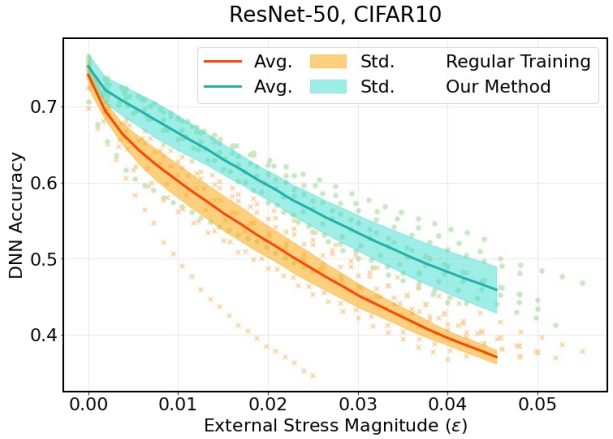
Selective Backpropagation

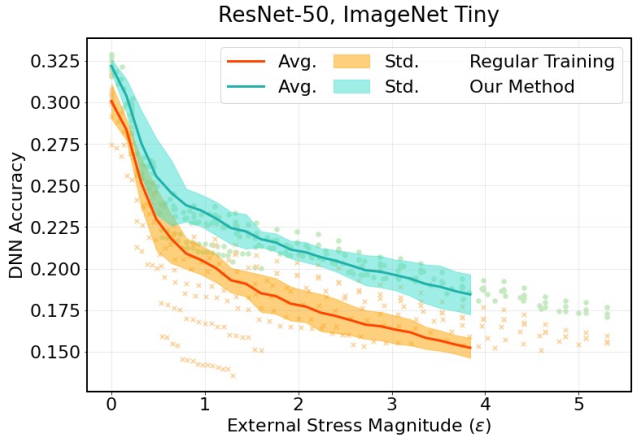












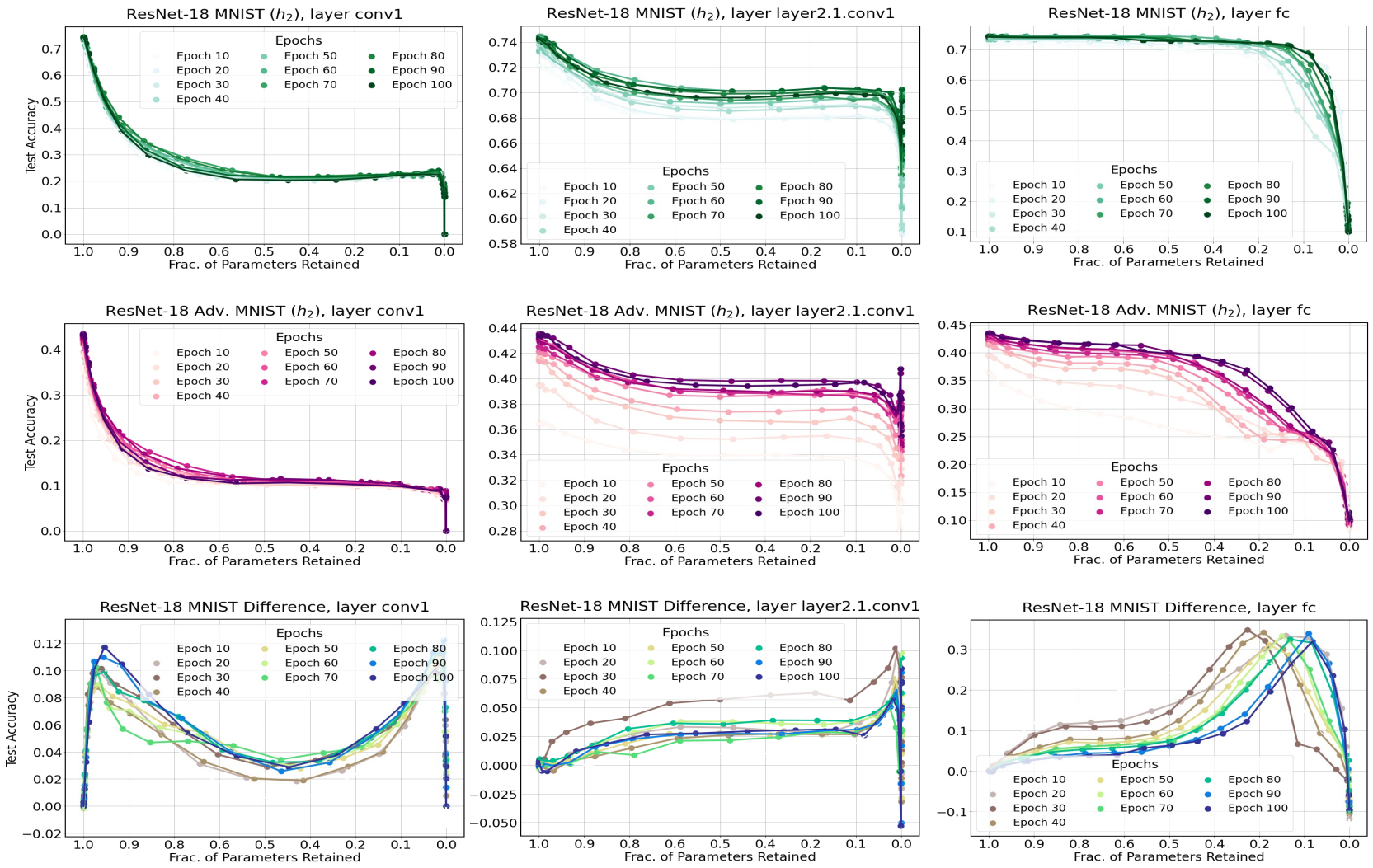
When we **retrain** networks at periodic intervals using only the characterised ***robust*** **and** ***antifragile*** layer parameters (selective backpropagation), we observe an **increase in adversarial performance**, and clean performance for some networks and datasets.

Compression and robustness

The parameter filtering performance as a function of the **fraction of parameters retained** due to the filtering.

We see the **for some layers we only have to retain a relatively small fraction of parameters** to still achieve sufficient network performance, at different stages in training.

The targeting response also shows comparable responses for the DNN at different stages in training.



Key outcomes of synaptic filtering

- Parameters characterised as **fragile influence the network performance** and thus are considered to contain useful information.
- We find that **robust and antifragile parameters have an invariant and positive effect on network performance**. When robust parameters are perturbed the network performance remains stable. When parameters characterised as antifragile are perturbed, the network performance improves.
- Upon identification of fragile, robust, and antifragile parameters we propose **re-training only the robust and antifragile parameters**, such that we apply backpropagation selectively (selective backpropagation).
- We find that the **adversarial robustness of the DNNs is improved using selective backpropagation**. In some instances, the clean dataset performance is also improved using our proposed method.

Summary

- Our results were able to identify to the components of DNNs that are vulnerable to adversarial attacks.
- Our results show that it is possible to improve the performance of DNNs by selectively treating the vulnerable components.
- Our results were validated that the methodology of identifying and treating vulnerable DNN components, shows to improve the performance of the DNN on benchmark datasets, as well as real-world problems.

References

- Fragility, Robustness and Antifragility in Deep Learning
  Artificial Intelligence, Elsevier. (2024)
  Pravin C, Martino I, Nicosia G, Ojha V

- Adversarial robustness in deep learning: Attacks on fragile neurons
  30th Int. Conf. on Artificial Neural Net., ICANN (pp 16-28), Springer, LNCS, Bratislava (2021)
  Pravin C, Martino I, Nicosia G, Ojha V