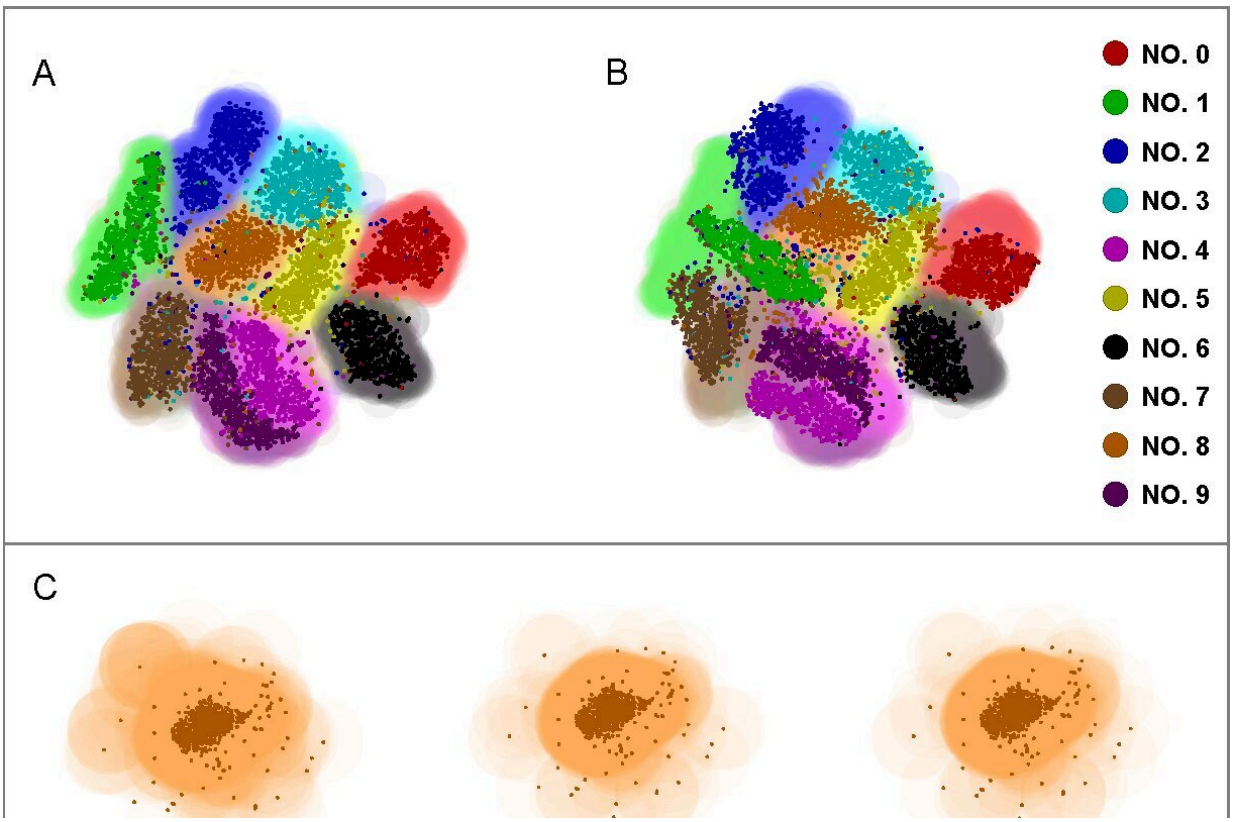
Scientists uncover quantum-inspired vulnerabilities in neural networks

May 9 2024



(A) Illustrates the final training output of the network, highlighting the areas of class prediction. Shaded regions demarcate these areas, with individual point colors indicating the true labels of the corresponding test samples, demonstrating a general alignment between the network's predictions and actual classifications. In (B), all test samples were subjected to gradient-based attacks, causing perturbed sample points to deviate noticeably from their correct categorical regions, leading to misclassifications by the network model. (C) Focuses on the evolving prediction region for the digit '8' across epochs 1, 21, and 41. The

deeper the shade of the region, the higher the network's confidence in its prediction. (D) Similar to (C), but displaying adversarial predictions for the attacked images, it is observed that as the training progresses, the effective radius of distribution for the attack points increases. This suggests that as the network's precision in identifying input features heightens, its vulnerability to attacks also escalates. Credit: Science China Press

In a recent study merging the fields of quantum physics and computer science, Dr. Jun-Jie Zhang and Prof. Deyu Meng have explored the vulnerabilities of neural networks through the lens of the uncertainty principle in physics.

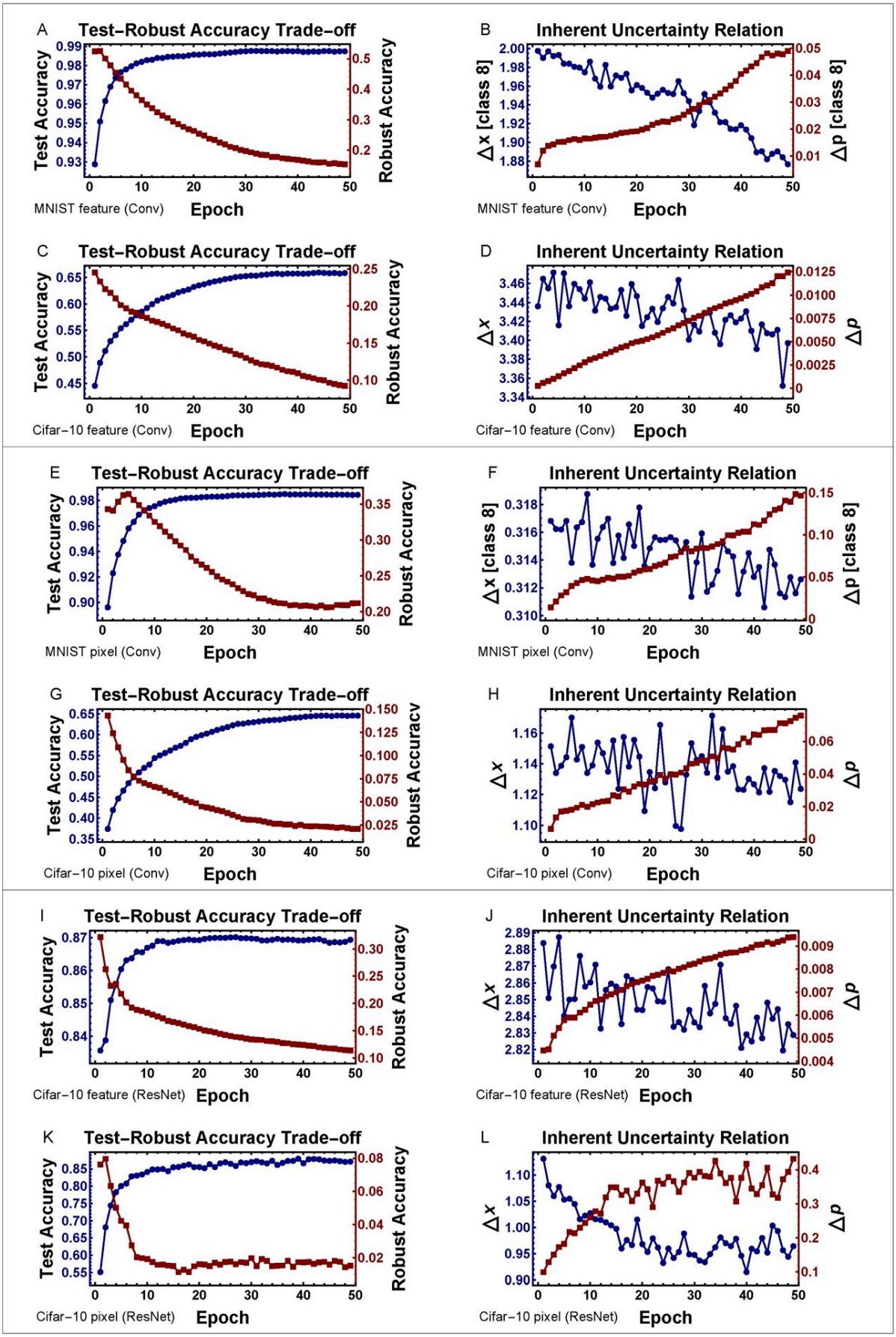
Their work, published in the *National Science Review*, draws a parallel between the susceptibility of neural networks to targeted attacks and the limitations imposed by the uncertainty principle—a well-established theory in quantum physics that highlights the challenges of measuring certain pairs of properties simultaneously.

The researchers' quantum-inspired analysis of neural network vulnerabilities suggests that adversarial attacks leverage the trade-off between the precision of input features and their computed gradients.

"When considering the architecture of deep neural networks, which involve a loss function for learning, we can always define a conjugate variable for the inputs by determining the gradient of the loss function with respect to those inputs," says Dr. Zhang, whose expertise lies in mathematical physics.

This research is hopeful to prompt a reevaluation of the assumed robustness of neural networks and encourage a deeper comprehension of their limitations. By subjecting a neural network model to adversarial

attacks, Dr. Zhang and Prof. Meng observed a compromise between the model's accuracy and its resilience.


A
Test–Robust Accuracy Trade–off
Test Accuracy
Robust Accuracy
MNIST feature (Conv)
Epoch
B
Inherent Uncertainty Relation
Δx [class 8]
Δp [class 8]
MNIST feature (Conv)
Epoch
C
Test–Robust Accuracy Trade–off
Test Accuracy
Robust Accuracy
Cifar–10 feature (Conv)
Epoch
D
Inherent Uncertainty Relation
Δx
Δp
Cifar–10 feature (Conv)
Epoch
E
Test–Robust Accuracy Trade–off
Test Accuracy
Robust Accuracy
MNIST pixel (Conv)
Epoch
F
Inherent Uncertainty Relation
Δx [class 8]
Δp [class 8]
MNIST pixel (Conv)
Epoch
G
Test–Robust Accuracy Trade–off
Test Accuracy
Robust Accuracy
Cifar–10 pixel (Conv)
Epoch
H
Inherent Uncertainty Relation
Δx
Δp
Cifar–10 pixel (Conv)
Epoch
I
Test–Robust Accuracy Trade–off
Test Accuracy
Robust Accuracy
Cifar–10 feature (ResNet)
Epoch
J
Inherent Uncertainty Relation
Δx
Δp
Cifar–10 feature (ResNet)
Epoch
K
Test–Robust Accuracy Trade–off
Test Accuracy
Robust Accuracy
Cifar–10 pixel (ResNet)
Epoch
L
Inherent Uncertainty Relation
Δx
Δp
Cifar–10 pixel (ResNet)
Epoch

Subfigures (A), (C), (E), (G), (I), and (K) display the test accuracy and the robust accuracy, with the latter assessed on images perturbed by the Projected Gradient Descent (PDG) attack method. Subfigures (B), (D), (F), (H), (J), and (L) reveal the trade-off relationship between accuracy and robustness. Credit: Science China Press

Their findings indicate that neural networks, akin to quantum systems mathematically, struggle to precisely resolve both conjugate variables—the gradient of the loss function and the input feature—simultaneously, hinting at an intrinsic vulnerability. This insight is crucial for the development of new protective measures against sophisticated threats.

"The importance of this research is far-reaching," notes Prof. Meng, an expert in machine learning and the corresponding author of the paper.

"As neural networks play an increasingly critical role in essential systems, it becomes imperative to understand and fortify their security. This interdisciplinary research offers a fresh perspective for demystifying these complex 'black box' systems, potentially informing the design of more secure and interpretable AI models."

More information: Jun-Jie Zhang et al, Quantum-inspired analysis of neural network vulnerabilities: The role of conjugate variables in system attacks, *National Science Review* (2024). DOI: 10.1093/nsr/nwae141

Provided by Science China Press