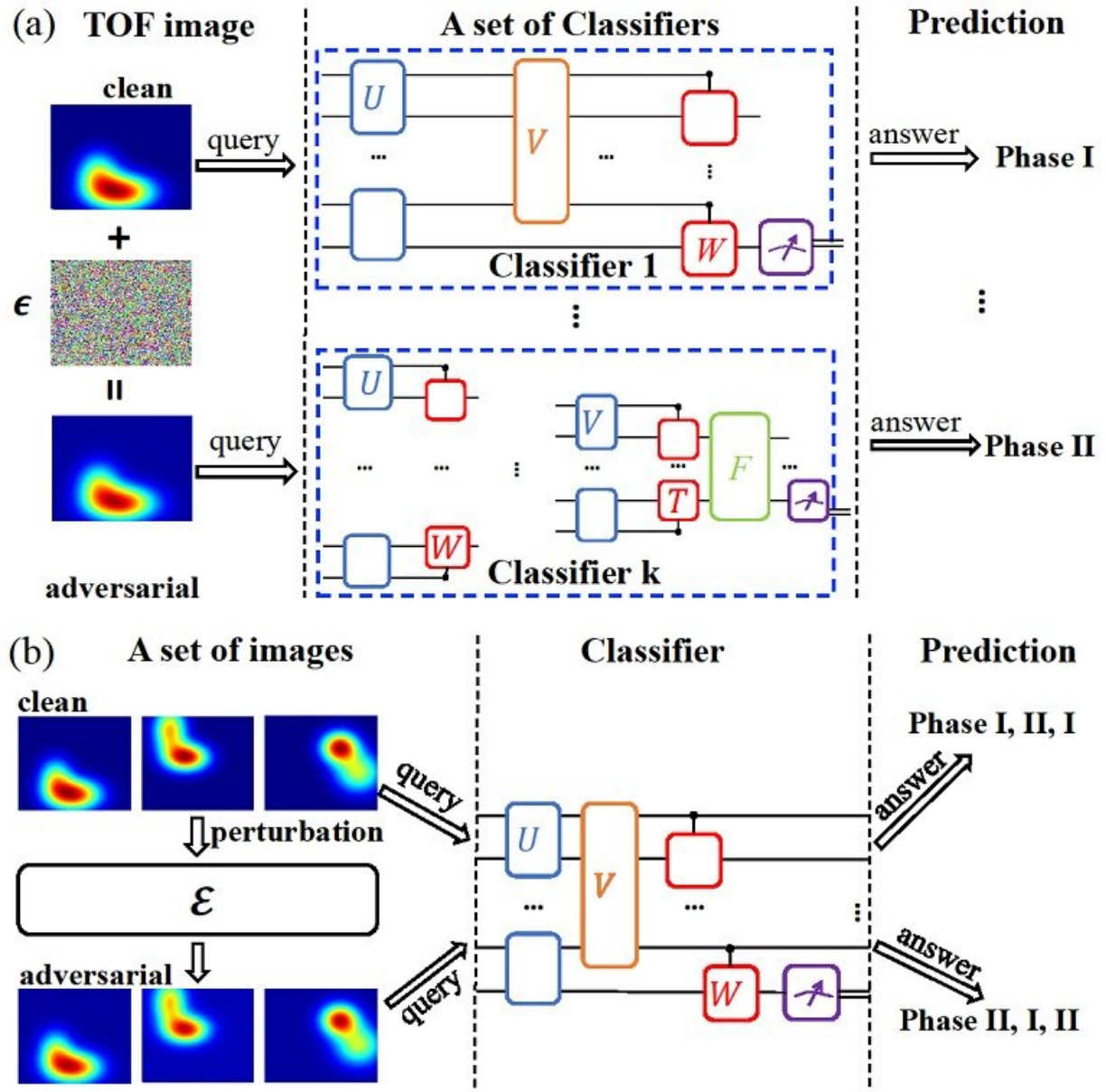
Discovery of universal adversarial attacks for quantum classifiers

October 12 2021



(a) Universal adversarial examples: Adding a small amount of carefully crafted noise to a certain image could make it become an adversarial example that can fool different quantum classifiers. (b) Universal adversarial perturbations: adding the same carefully-crafted noise to a set of images could make them all become adversarial examples for a given quantum classifier. Credit: Science China Press

Artificial intelligence has achieved dramatic success over the past decade, with the triumph in predicting protein structures marked as the latest milestone. At the same time, quantum computing has also made remarkable progress in recent years. A recent breakthrough in this field is the experimental demonstration of quantum supremacy. The fusion of artificial intelligence and quantum physics gives rise to a new interdisciplinary field—-quantum artificial intelligence.

This emergent field is growing fast with notable progress made on a daily basis. Yet, it is largely still in its infancy and many important problems remain unexplored. Among these problems stands the vulnerability of quantum classifiers, which sparks a new research frontier of quantum adversarial machine learning.

In classical machine learning, the vulnerability of classifiers based on deep neural networks to adversarial examples has been actively studied since 2004. It has been observed that these classifiers might be surprisingly vulnerable: adding a carefully-crafted but imperceptible perturbation to the original legitimate sample can mislead the classifier to make wrong predictions, even at a notably high confidence level.

Similar to classical machine learning, recent studies have revealed the vulnerability aspect of quantum classifiers from both theoretical analysis and numerical simulations. The exotic properties of the adversarial attacks against quantum machine learning systems have attracted

considerable attentions across communities.

In a new research article published in the Beijing-based *National Science Review*, researchers from IIIS, Tsinghua University, China studied the universality properties of adversarial examples and perturbations for quantum classifiers for the first time. As shown in the figure, the authors put forward affirmative answers to the following two questions: (i) whether there exist universal adversarial examples that could fool different quantum classifiers? (ii) whether there exist universal adversarial perturbations, which when added to different legitimate input samples could make them become adversarial examples for a given quantum classifier?

The authors have proved two interesting theorems, one for each question. For the first question, previous works have shown that for a single quantum classifier, the threshold strength for a perturbation to deliver an adversarial attack decreases exponentially as the number of qubits increases. The current paper extended this conclusion to the case of multiple quantum classifiers, and rigorously proved that for a set of k quantum classifiers, an logarithmic k increase of the perturbation strength is enough to ensure a moderate universal adversarial risk. This establishes the existence of universal adversarial examples that can deceive multiple quantum classifiers.

For the second question, the authors proved that for a universal adversarial perturbation added to different legitimate samples, the misclassification rate of a given quantum classifier will increase as the dimension of data space increases. Furthermore, the misclassification rate will approach 100% when the dimension of data samples is infinitely large.

In addition, extensive numerical simulations had been carried out on concrete examples involving classifications of real-life images and

quantum phases of matter to demonstrate how to obtain both universal adversarial perturbations and examples in practice. The authors also proposed adversarial attacks under black-box scenarios to explore and the transferability of adversarial attacks on different classifiers.

The results in this work reveals a crucial universality aspect of adversarial attacks for quantum machine learning systems, which would provide a valuable guide for future practical applications of both near-term and future quantum technologies in machine learning, or more broadly artificial intelligence.

More information: Weiyuan Gong et al, Universal Adversarial Examples and Perturbations for Quantum Classifiers, *National Science Review* (2021). DOI: 10.1093/nsr/nwab130

Provided by Science China Press