Verifying Adversarial Robustness in Quantum Machine Learning:

From Theory to Physical Validation via a Software Tool

Ji Guan
guanji1992@gmail.com

Institute of Software, Chinese Academy of Sciences, China

July 21, 2025

Overview

Overview

Recent Progress

Scientific Advantages ⇒ Practical Advantages

Q2B 24 Meeting

John Preskill

Proposer of "NISQ"

Currently in the **NISQ era**, have noteworthy **scientific value**

To the **Beyond-NISQ era**, need **advantages in applications with commercial value**.

Support

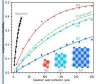
Google IBM

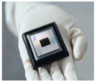
Coordination across the full stack: from fundamental physics, algorithms, to software engineering.

Google

Quantum Error Correction

- **December 2024: "Willow"** chip with 105 physical qubits.
- First realization of **decreasing logical error rate exponentially** with code distance.

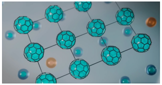




Microsoft

Ion Trap

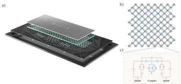
- **September 2024**: with Quantinuum, achieved **12 logical qubits**.
- **November 2024**: with Atom Computing, achieved **24 logical qubits**, demonstrating fault tolerance.



中国科学技术大学 University of Science and Technology of China

Superconducting Chips

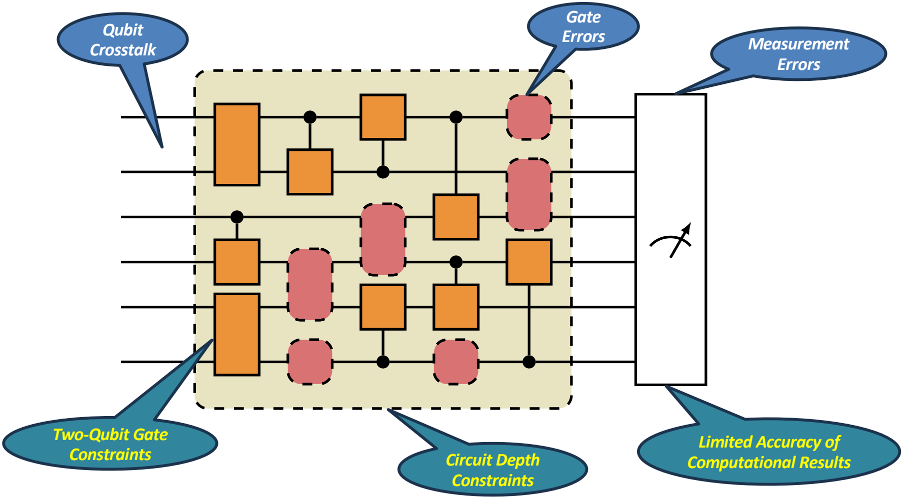
- **December 2024**: USTC launched **"Zuchongzhi-3"** chip. Performance surpasses Google's 72-qubit "Sycamore".



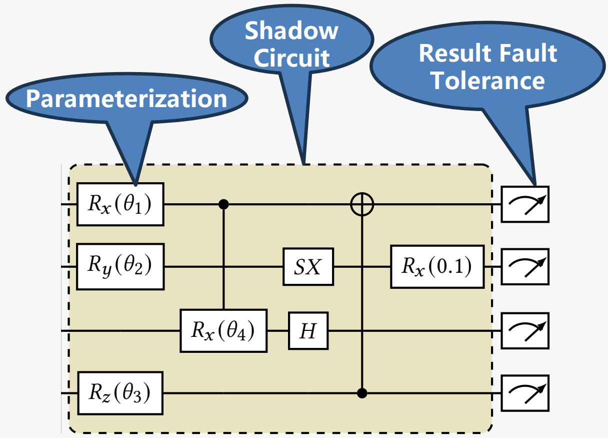
Top 10 Scientific and Technological News in the World (2024)

Demonstrating Quantum Advantage Through Random Sampling Problems

Circuit Noise

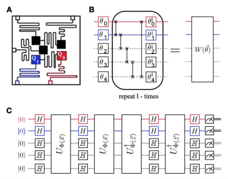


Quantum Machine Learning Algorithm (Variational Circuit)

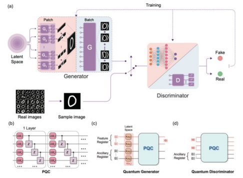


Variational quantum algorithm for MNIST image classification

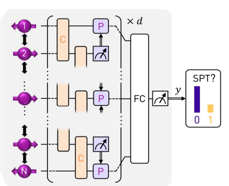
Physical Implementation



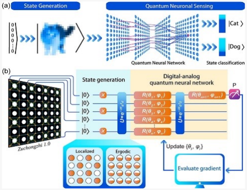
Nature (2019)
Artificial Data Classification



Phys. Rev. Applied (2021)
Image Generation



Nat. Commun (2022)
Quantum Phase Recognition



Science Bulletin (2023)
Structure recognition of quantum many-body systems

Overview

Quantum (Machine Learning) Classifiers

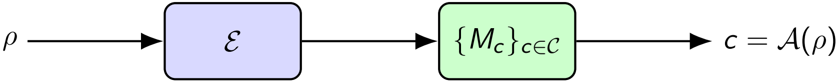


Figure: **Quantum classifier pipeline.** The input quantum state ρ is processed by a quantum channel $\mathcal{E}$, followed by measurement via a POVM $\{M_c\}_{c\in\mathcal{C}}$, to produce a classical class label $c = \mathcal{A}(\rho)$.

Formally, a quantum classifier over the Hilbert space $\mathcal{H}$ is defined as a pair:

$$\mathcal{A} = (\mathcal{E}, \{M_c\}_{c\in\mathcal{C}}),$$

Given an input quantum state $\rho \in \mathcal{D}(\mathcal{H})$, the classifier outputs a label determined by the most probable measurement outcome:

$$\mathcal{A}(\rho) := \arg\max_{c\in\mathcal{C}} \mathrm{Tr}[M_c\mathcal{E}(\rho)],$$

where $\mathrm{Tr}[M_c\mathcal{E}(\rho)]$ is the probability of obtaining outcome c upon measuring the output state $\mathcal{E}(\rho)$ of $\mathcal{E}$ with the POVM $\{M_c\}_{c\in\mathcal{C}}$.

Visualizing Quantum Classifiers

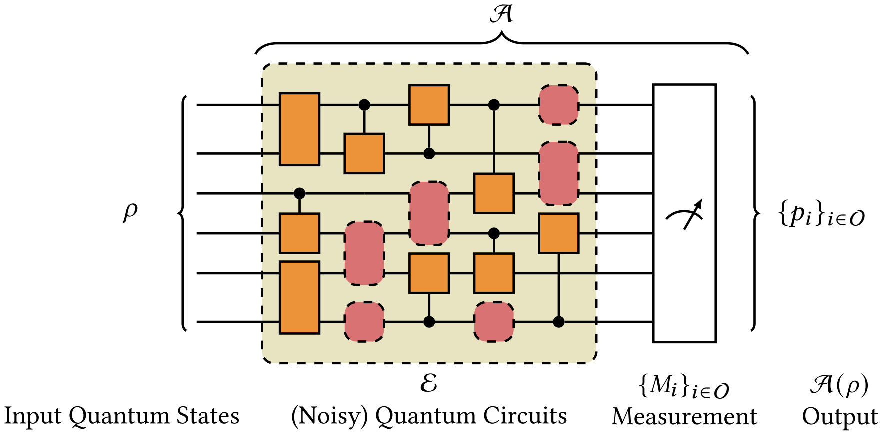


Figure: The Computational Model of Quantum Classifiers

Famous Classical Adversarial Example

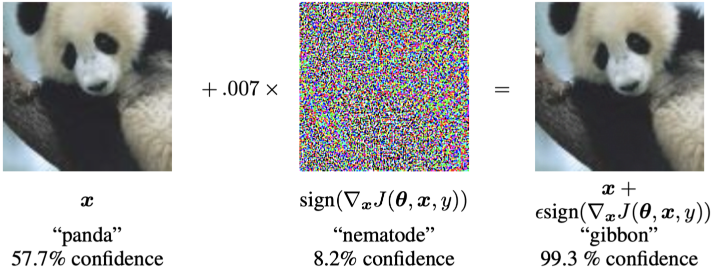


Figure: Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy [ICLR 2015]

Adversarial examples (the right picture): inputs to a machine learning algorithm cause the algorithm to make a mistake.
Safety issue: machine learning algorithms are vulnerable to intentionally-crafted adversarial examples.

Robustness Studies

Motivation:

- Quantum noise at the present of NISQ (Noisy Intermediate-Scale Quantum) era;
- Quantum classifier is principled by quantum mechanics (hard to be explained to the end users), so verifying the robustness is essential (Toward to trustworthy quantum AI).

Challenges:

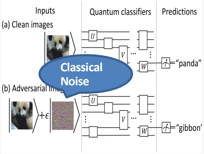
- The attacker is quantum noise from the unknown environment.
- Due to the statistical nature of quantum mechanics, quantum machine learning models are randomized.

Core Problem:

Verifying Robustness → Identifying Adversarial Examples → Improving Robustness (e.g. Adversarial Training)

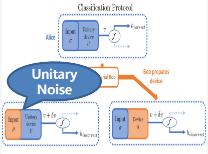
Specific Attack Studies

Classical Noise



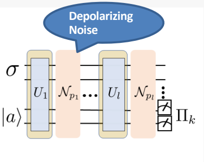
Phys. Rev. Res (2020)

Unitary Noise



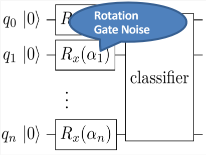
Phys. Rev. A (2020)

Depolarizing Noise



Phys. Rev. Res (2021)

Rotation Gate Noise



ICASSP (2023)

The attack should be unknown.

The internal structure of noisy quantum circuits is not accessible and a black box.

Adversarial Examples

Definition (Adversarial Example)

Let $\mathcal{A}$ be a quantum classifier, $\rho \in \mathcal{D}(\mathcal{H})$ an input state, and $\varepsilon > 0$ a perturbation threshold. A quantum state σ is called an ε-*adversarial example* of ρ if

$$\mathcal{A}(\sigma) \neq \mathcal{A}(\rho) \quad \text{and} \quad D_F(\rho, \sigma) \leq \varepsilon.$$

If such a state σ exists, then ε is referred to as an *adversarial perturbation* of ρ. The *fidelity distance* (also called *infidelity*) between two quantum states is defined as

$$D_F(\rho, \sigma) := 1 - F(\rho, \sigma).$$

Definition (Adversarial Robustness)

A quantum classifier $\mathcal{A}$ is said to be ε-*robust* at state ρ if there exists no ε-adversarial example of ρ.

Adversarial ε-Robustness

Definition (Robustness Radius)

Let $\mathcal{A}$ be a quantum classifier and ρ a correctly classified input state. The *robustness radius* of ρ, denoted $\varepsilon^*(\rho)$, is the maximum value ε such that $\mathcal{A}$ is ε-robust at ρ:

$$\varepsilon^*(\rho) := \sup_{\substack{\sigma \in \mathcal{D}(\mathcal{H}) \\ \mathcal{A}(\sigma) = \mathcal{A}(\rho)}} D_F(\rho, \sigma).$$

Problem (Robustness Verification Problem)

Given a quantum classifier $\mathcal{A}$, an input state $\rho \in \mathcal{D}(\mathcal{H})$, and a threshold $\varepsilon > 0$, determine whether

$$\varepsilon \leq \varepsilon^*(\rho).$$

If so, $\mathcal{A}$ is ε-robust at ρ; otherwise, ε is an adversarial perturbation, and a violating state σ can be returned as an ε-adversarial example.

Optimal Robustness Bound via Semidefinite Programming

Theorem (Optimal Robustness Bound via SDP, CAV 2021)

Let $\mathcal{A} = (\mathcal{E}, \{M_c\}_{c\in\mathcal{C}})$ be a quantum classifier. The exact robustness radius is given by

$$\varepsilon^*(\rho) = \min_{\substack{c\in\mathcal{C} \\ c\neq\mathcal{A}(\rho)}} \varepsilon^*_c(\rho),$$

where each $\varepsilon^*_c(\rho)$ is the solution to the following SDP:

$$\begin{aligned} \textit{minimize:} \quad & D_F(\rho, \sigma) \\ \textit{subject to:} \quad & \sigma \succeq 0, \\ & \mathrm{Tr}(\sigma) = 1, \\ & \mathrm{Tr}[(M_{\mathcal{A}(\rho)} - M_c)\mathcal{E}(\sigma)] \leq 0. \end{aligned}$$

If this SDP is infeasible for some c, then $\varepsilon^*_c(\rho) = \infty$, indicating that no adversarial example of ρ exists which is misclassified as class c.

Robustness Lower Bound via Measurement Distribution

Theorem (Robustness Lower Bound from Measurement Distribution CAV 2021)

Let $\rho \in \mathcal{D}(\mathcal{H})$ and $c^* = \mathcal{A}(\rho)$. Then

$$\varepsilon_{\mathrm{RLB}}(\rho) := \min_{c \neq c^*} \frac{1}{2} \left(\sqrt{p^{\rho}_{c^*}} - \sqrt{p^{\rho}_{c}} \right)^2$$

is a certified robustness lower bound: for all σ such that $D_F(\rho, \sigma) \leq \varepsilon_{\mathrm{RLB}}(\rho)$, it holds that $\mathcal{A}(\sigma) = \mathcal{A}(\rho)$. Here, $p^{\rho}_c := \mathrm{Tr}[M_c \mathcal{E}(\rho)]$.

- **Efficient to Compute.** Directly from measurement outcomes without searching for adversarial perturbations. Fast robustness certification and dataset-level evaluation of robust accuracy.
- **Model-agnostic:** No access to the internal structure of $\mathcal{E}$, this bound is particularly suited for hardware-level evaluation. In real-device settings, estimate p^{ρ}_c by repeated execution of $\mathcal{E}$ on quantum hardware and compute $\varepsilon_{\mathrm{RLB}}(\rho)$ from the empirical outcome distribution.

Robustness Upper Bound via Attack Generation

Definition (Empirical Robustness Upper Bound)

Let $\rho \in \mathcal{D}(\mathcal{H})$ be an input quantum state. An *adversarial attack method* constructs a perturbed state σ_{adv} such that:

$$\mathcal{A}(\sigma_{\mathsf{adv}}) \neq \mathcal{A}(\rho), \quad \text{and} \quad \varepsilon_{\mathrm{RUB}}(\rho) := D_F(\rho, \sigma_{\mathsf{adv}}),$$

where D_F is the fidelity distance. Then, $\varepsilon_{\mathrm{RUB}}(\rho)$ serves as an *empirical robustness upper bound* for $\varepsilon^*(\rho)$.

Attack Method: FGSM and Mask FGSM

Fast Gradient Sign Method (FGSM):

$$\boldsymbol{x}' = \boldsymbol{x} + \varepsilon \cdot \mathrm{sgn}(\nabla_{\boldsymbol{x}} \mathcal{L}),$$

where ε is the perturbation magnitude, $\nabla_{\boldsymbol{x}} \mathcal{L}$ is the gradient of the loss $\mathcal{L}$.

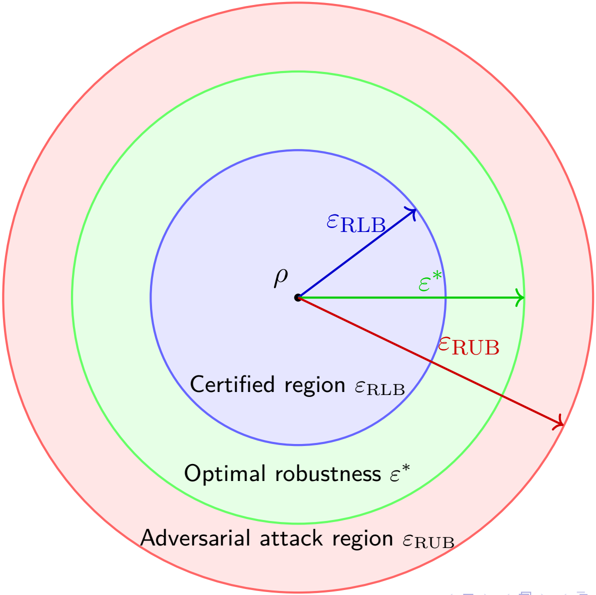
Mask FGSM (localized variant)[arXiv:2505.16714]:

$$\delta_i = \begin{cases} \varepsilon \cdot \mathrm{sgn}\left(\frac{\partial \mathcal{L}}{\partial x_i}\right), & m_i = 1, \\ 0, & m_i = 0, \end{cases}$$

with binary mask $\mathcal{M} = (m_1, m_2, \ldots, m_{\dim(\boldsymbol{x})})^T$ selecting which input features are perturbed.

Key point: Achieves efficient and effective adversarial sample generation in QML, validated experimentally on EMNIST and LCEI tasks.

Visualizing the Bounds



Sandwich Theorem

Theorem (Sandwich Robustness Bound)

Given a quantum input state ρ, a certified lower bound $\varepsilon_{\mathrm{RLB}}(\rho)$ (Theorem 6), and an adversarially generated state σ_{adv}, we have:

$$\varepsilon_{\mathrm{RLB}}(\rho) \leq \varepsilon^*(\rho) \leq \varepsilon_{\mathrm{RUB}}(\rho), \tag{1}$$

where $\varepsilon_{\mathrm{RUB}}(\rho) = D_F(\rho, \sigma_{adv})$.

- $\varepsilon_{\mathrm{RLB}}(\rho)$: a certified lower bound used for formal robustness guarantees;
- $\varepsilon^*(\rho)$: the exact robustness radius, computable via SDP;
- $\varepsilon_{\mathrm{RUB}}(\rho)$: an empirical upper bound derived from adversarial attacks.

Tightness Assessment. The gap $\Delta := \varepsilon_{\mathrm{RUB}}(\rho) - \varepsilon_{\mathrm{RLB}}(\rho)$ quantifies the precision of the robustness estimation. The observed gap between the two bounds is typically less than 3×10^{-3}, demonstrating that $\varepsilon_{\mathrm{RLB}}(\rho)$ provides a tight and practically useful certificate of robustness.

Overview

Robustness Verification Algorithms

Robustness can be aggregated across a dataset to evaluate a classifier's overall robustness:

Definition (Robust Accuracy)

Let $\mathcal{A}$ be a quantum classifier. The *ε-robust accuracy* of $\mathcal{A}$ is the proportion of correctly classified input states in the dataset that are also ε-robust.

Robustness Verification Algorithms:

- State Robustness Verification: SDP.
- Under-approximate Robustness Verification: robustness lower bound.
- Exact Classifier Robustness Verification: robustness lower bound and SDP.

Robustness Verification Algorithms: $N = 2^n$ for n qubits

Robustness Verification Algorithms			
	Robustness Lower Bound	Robustness Optimal Bound	Mixed Strategy
Method	Matrix Multiplication (MM)	Semidefinite Programming (SDP)	MM & SDP
Complexity	$O(\|T\| \cdot \|\mathcal{C}\| \cdot N^5)$	$O(\|T\| \cdot \|\mathcal{C}\| \cdot N^{6.5})$	$O(\|T'\| \cdot \|\mathcal{C}\| \cdot N^{6.5})$
Robust Accuracy	Under-approximate	Exact	Exact

Table: Summary of robustness verification algorithms based on different bounds.

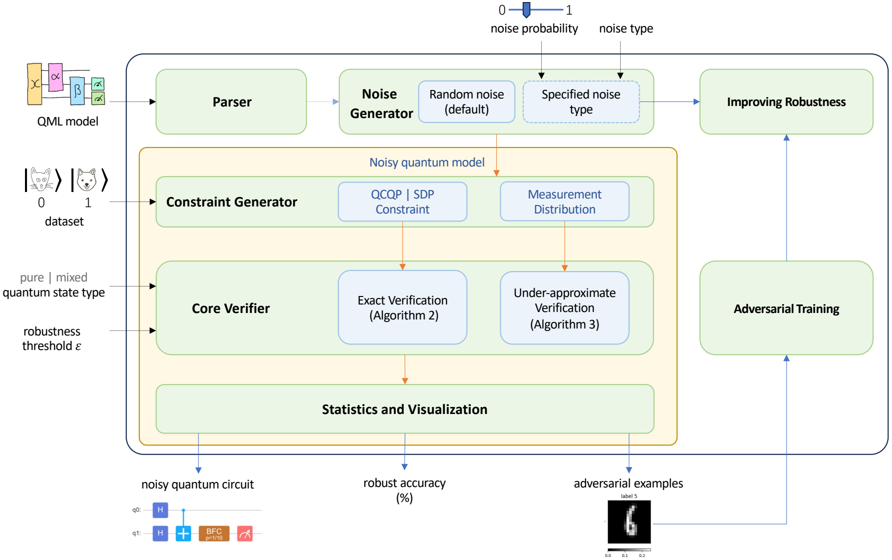
- T: the set of training data;
- T': a subset of T obtained by robust bound;
- $\mathcal{C}$: the set of measurement outcomes;
- N: the dimension of state space $\mathcal{H}$.

In practice: $|T'| \ll |T| \Rightarrow$Robustness lower bound is tight.

Overview

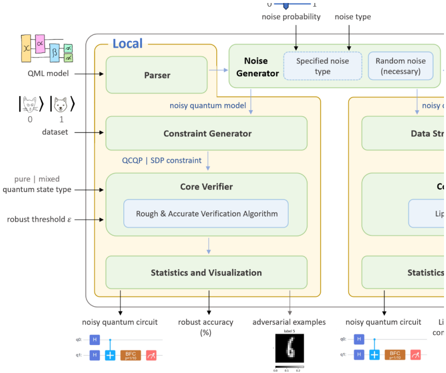
System Architecture of VeriQR.

VeriQR is available at `https://github.com/Veri-Q/VeriQR`.



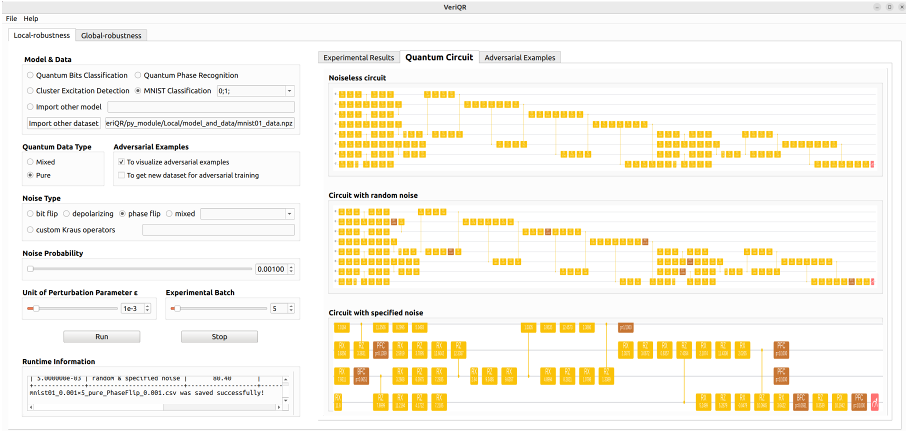
Functions

- **Parser**: parses the input quantum classification model to obtain the corresponding quantum circuit object
- **Noise Generator**: adds random noise to the quantum circuit (to simulate the noise effect of a real device) and enables the user to add custom noise to generates a noisy quantum model
- **Constraint Generator**: generates nonlinear constraints based on a noisy quantum model and dataset



- **Core Verifier**: takes constraints, a perturbation parameter ε, and quantum state types as input and uses approximate and exact algorithms to initiate the verification analysis process for ε-robustness
- **Statistics and Visualization**: displays and visualizes output in VeriQR's GUI component, including robust accuracy, adversarial examples and quantum circuits

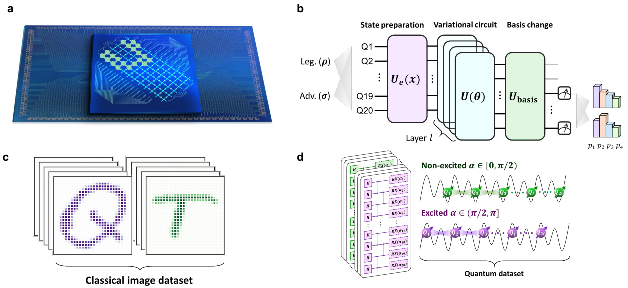
GUI



VeriQR
File Help
Local-robustness
Global-robustness
Model & Data
Quantum Bits Classification
Quantum Phase Recognition
Cluster Excitation Detection
MNIST Classification
0;1;
Import other model
Import other dataset
eriQR/py_module/Local/model_and_data/mnist01_data.npz
Quantum Data Type
Mixed
Pure
Adversarial Examples
To visualize adversarial examples
To get new dataset for adversarial training
Noise Type
bit flip
depolarizing
phase flip
mixed
custom Kraus operators
Noise Probability
0.00100
Unit of Perturbation Parameter ε
1e-3
Experimental Batch
5
Run
Stop
Runtime Information
| 5.000000e-03 | random & specified noise | 80.40 |
mnist01_0.001×5_pure_PhaseFlip_0.001.csv was saved successfully!
Experimental Results
Quantum Circuit
Adversarial Examples
Noiseless circuit
Circuit with random noise
Circuit with specified noise

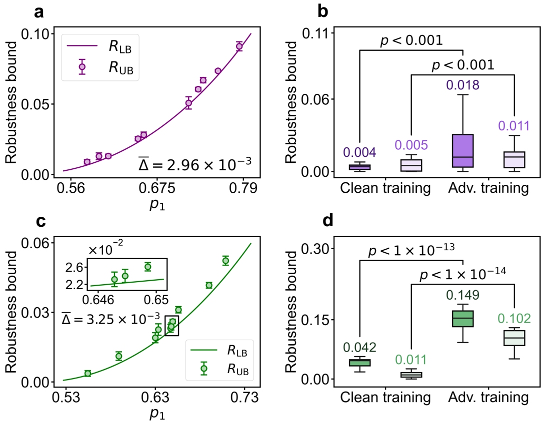
Overview

Experimental Schematic for QNN Evaluation



- **a**, The superconducting quantum processor, comprising 72 qubits and 20 qubits selected for the experiment are highlighted in green.
- **b**, Architecture of the quantum neural network (QNN) classifier.
- **c**, Sample visualization of handwritten letters "Q" and "T" from the EMNIST dataset, used for the classical image classification task.
- **d**, Quantum circuit used to generate the Linear Cluster State Excitation Identification (LCEI) dataset. States are labeled as "excited" or "non-excited" based on the rotation angle α.

Robustness Bound Verification Experiments



- **Tightness of Robustness Bounds:** validate the near-optimality of the Mask FGSM attack strategy and the tightness of the lower bound.
- **Improvement through Adversarial Training:** adversarial training significantly increased the mean certified robustness lower bound by a factor of 4.22 in EMNIST and 4.74 in LCEI.

Overview

Takeaway

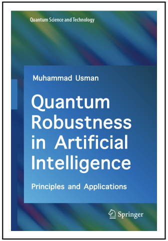
Summary of quantum adversarial robustness verification:

- **Theory:** Robustness bounds and verification algorithms CAV 2021
- **Tool:** Robustness verification tool VeriQR FM 2024
- **Physical Validation:** Experimental robustness benchmark on superconducting hardware arXiv:2505.16714

Review Book Chapter

Verifying Adversarial Robustness in Quantum Machine Learning: From Theory to Physical Validation via a Software Tool

Quantum Robustness in Artificial Intelligence (Springer, online soon)



Other Trustworthy Quantum Algorithm Works:

- Fairness: Individual fairness (global robustness) verification of quantum algorithms CAV 2022
- Privacy: Differential privacy for quantum algorithms: formal verification and optimal mechanisms ACM CCS 2023 and 2025

References

- Guan J., Fang W., Ying M. (2021) Robustness Verification of Quantum Classifiers. (**CAV 2021**)
- Guan J., Fang, W. and Ying, M., 2022. Verifying Fairness in Quantum Machine Learning. (**CAV 2022**)
- Lin, Y., Guan, J., Fang, W., Ying, M. and Su, Z., 2024, September. A Robustness Verification Tool for Quantum Machine Learning Models. (**FM 2024**).
- Guan, J., Fang, W., Huang, M. and Ying, M., 2023, November. Detecting violations of differential privacy for quantum algorithms. (**ACM CCS 2023**)
- Guan, J., 2025. Optimal Mechanisms for Quantum Local Differential Privacy. (**ACM CCS 2025**).
- Zhang, H.F., Chen, Z.Y., Wang, P., Guo, L.L., Wang, T.L., Yang, X.Y., Zhao, R.Z., Zhao, Z.A., Zhang, S., Du, L. and Tao, H.R., 2025. Experimental robustness benchmark of quantum neural network on a superconducting quantum processor. **arXiv preprint arXiv:2505.16714.**

Thanks!

My excellent collaborators: Mingsheng Ying, Wang Fang, Mingyu Huang, and USTC's quantum hardware physical group

Email: guanji1992@gmail.com