QTrojan: A Circuit Backdoor Against Quantum Neural Networks

Cheng Chu Lei Jiang Martin Swany Fan Chen
Department of Intelligent Systems Engineering, Indiana University, Bloomington, IN, USA

ICASSP 2023
4 - 10 JUNE, RHODES ISLAND, GREECE

Indiana University

Introduction & Background

Quantum Neural Networks (QNNs)

- Encoding layer $\boldsymbol{S}(\boldsymbol{x})$: classical $\boldsymbol{x}$ → quantum state ρ_x
- Variational Quantum Circuit $\boldsymbol{U}(\theta)$: ρ_x → $\boldsymbol{U}(\theta)\rho_x$
- Quantum Measurement: $\boldsymbol{U}(\theta)\rho_x$ → classical output

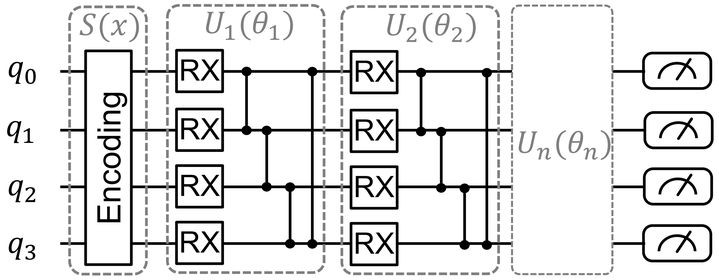
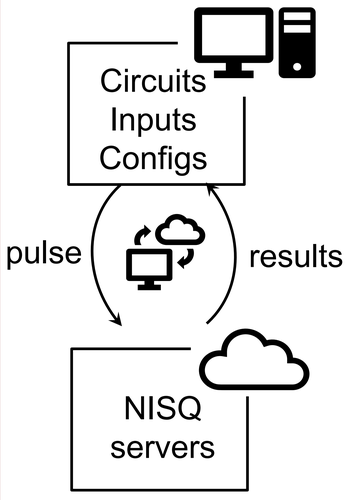


Fig.1 A standard quantum neural network.

Quantum Cloud Computing



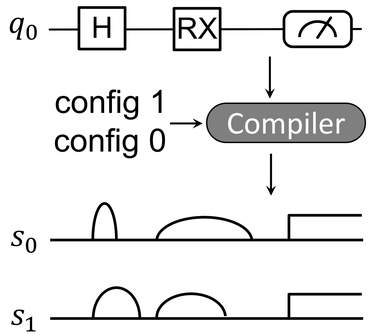


Fig.3 Quantum compilation.

Fig.2 QNNs in cloud.

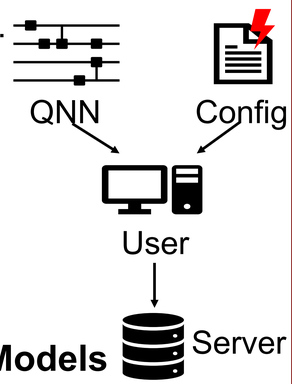
QTrojan vs classical DPBA

Schemes	DPBA	QTrojan
No Trigger in Inputs	×	√
No Training Data	×	√
No Training Process	×	√
Works after Retaining	×	√

DPBA: Data-Poisoning-based Backdoor Attack.

Threat Model

- Trustworthy Q compiler & servers.
- Attacker insert triggers into config files: QNN classify all inputs to a target class when using a config file with triggers.
- Victim user download config files to minimize noises and errors before each compilation.



Comparison to DPBA Threat Models

- More conservative.
- Stealthier.

Fig.4 QTrojan Threat model.

Methods

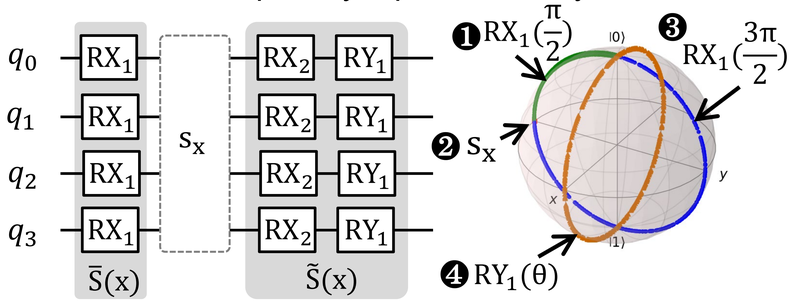
Angle Encoding Layer (default and dense)

$$|\mathbf{x}\rangle = \bigotimes_{i=1}^{N} \cos(x_i)|0\rangle + \sin(x_i)|1\rangle$$

$$|\mathbf{x}\rangle = \bigotimes_{i=1}^{\lceil N/2 \rceil} \cos(\pi x_{2i-1})|0\rangle + e^{2\pi i x_{2i}} \sin(\pi x_{2i-1})|1\rangle$$

A Backdoor Against Angle Encoding Layer

- Pre-encoding layer $\bar{\boldsymbol{S}}(\boldsymbol{x})$
  - Move qubits to a specific position, shielding the encoding layer.
- Encoding layer $\boldsymbol{S}(\boldsymbol{x})$
  - A normally applied encoding layer.
- Post-encoding layer $\tilde{\boldsymbol{S}}(\boldsymbol{x})$
  - Rotate the qubit by θ predefined by the attacker.



(a) Backdoored Encoding Layer (b) Angle Adjustment

Fig.5 The backdoored encoding layer of QTrojan.

Results

Pulse-level Overhead

- ✓ Does NOT add circuit depth.
- ✓ Does NOT add pulse sequence latency.

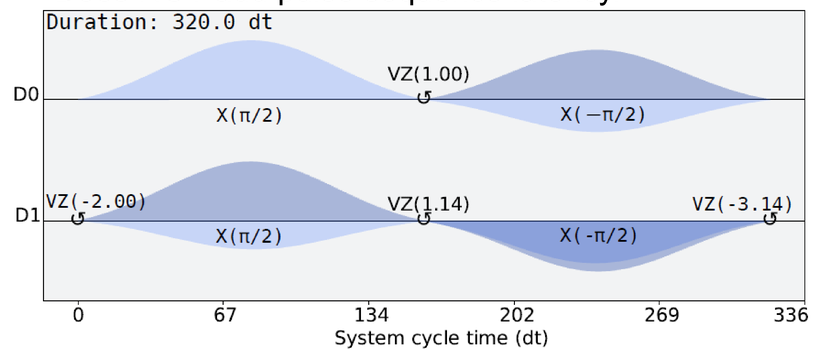


Fig.6 The pulse-level overhead of QTrojan.

QTrojan vs DPBA

Schemes	QNN (%)	DPBA		QTrojan	
	Acc	CDA	ASR	CDA	ASR
MNIST-2	98.25	91.56	99.5	98.25	100
MNIST-4	58.6	43	68.75	58.6	100

Acc: Accuracy; CDA: clean data accuracy.
ASR: attack success rate.

QTrojan against QLSTM

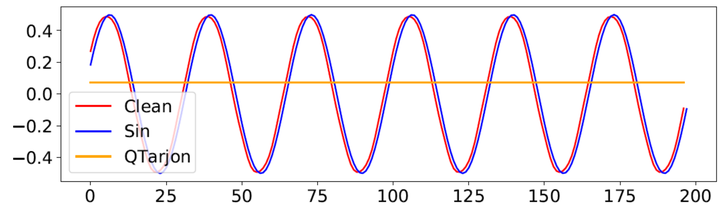


Fig.7 QTrojan against QLSTM

Conclusion

We propose Qtrojan, a stealthy circuit-level backdoor attack against QNNs. QTrojan uses several lines in a server-specific configuration file as triggers and is implemented by a few quantum gates inserted into a victim QNN circuit. Compared to DPBA, QTrojan improves the CDA by 21% and the ASR by 19.9% on average.