QTrojan: A Circuit Backdoor Against Quantum Neural Networks

Cheng Chu   Lei Jiang   Martin Swany   Fan Chen

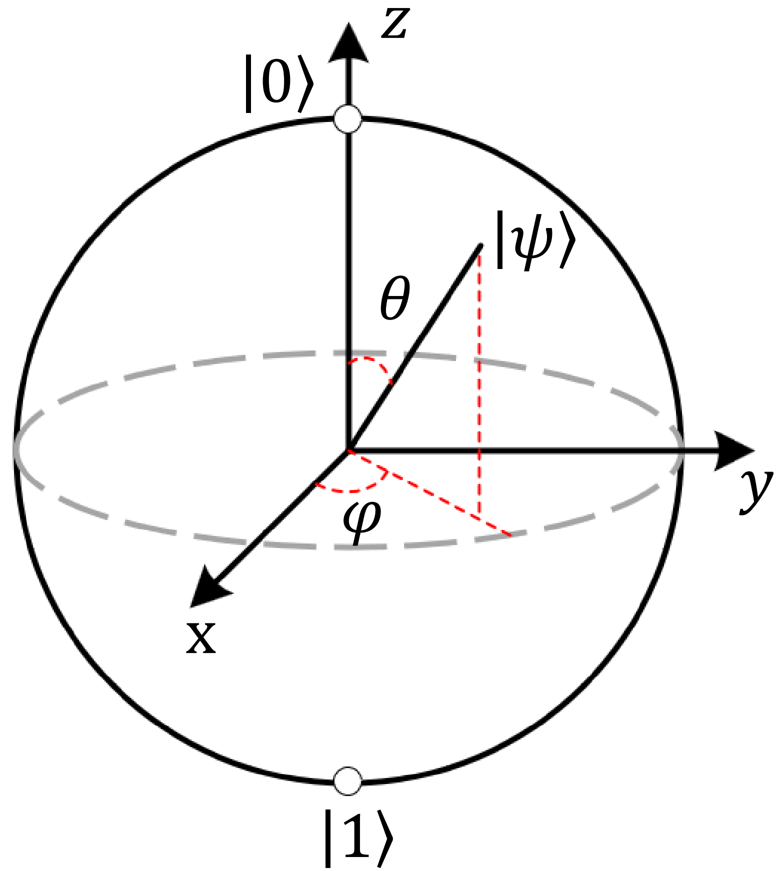
Dept. of Intelligent Systems Engineering,
Indiana University Bloomington

Qubit vs Bit

Classical Bit: 0 1

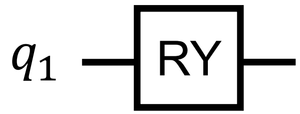
Quantum Bit: 0 1

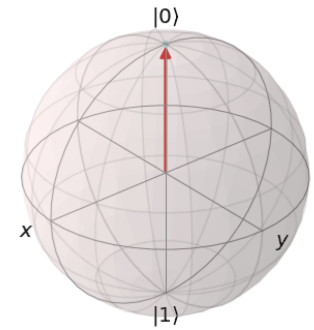
- Quantum Bit (Qubit):
  - $|\psi\rangle = \cos\frac{\theta}{2}|0\rangle + e^{i\varphi}\sin\frac{\theta}{2}|1\rangle$
  - $|\psi\rangle = \begin{bmatrix} \cos\frac{\theta}{2} \\ e^{i\varphi}\sin\frac{\theta}{2} \end{bmatrix}$



Quantum Gates

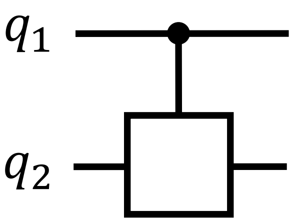
- Quantum gate ⇒ Matrix
  - Single qubit gate ⇒ 2*2
  - Two-qubit gate ⇒ 4*4
  - Multi-qubit gate ⇒ n*n





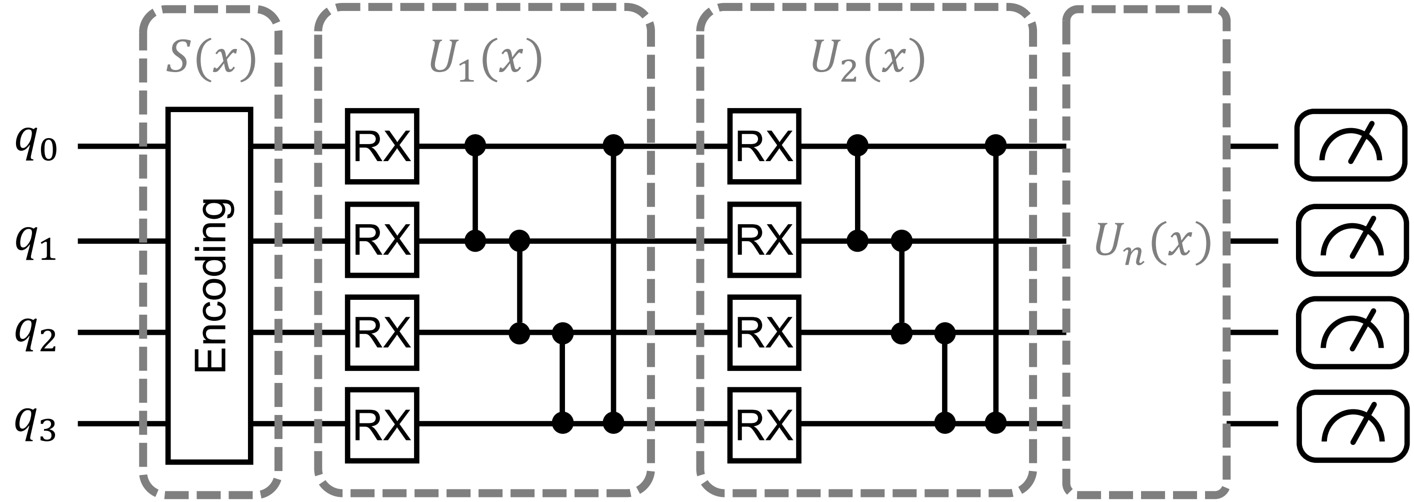
- Quantum gate operation
  - Matrix Multiplication

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \times \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} b_0' \\ b_1' \end{bmatrix}$$



$$\text{CNOT} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

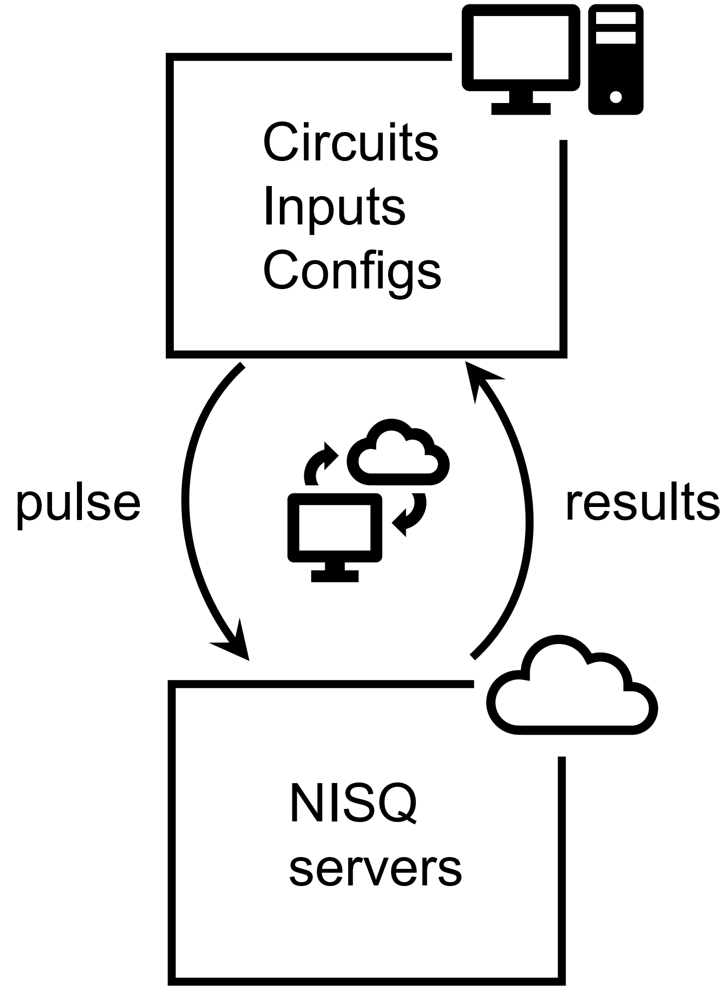
Variational Quantum Circuit



- Encoding layer $\boldsymbol{S}(\boldsymbol{x})$
  - Prepare quantum state ρ_x to represent the classical input data.
- Variational circuit block $\boldsymbol{U}(\boldsymbol{x})$
  - Entangle and rotate ρ_x to generate the processed state $\tilde{\rho}_x$.
- Measuring layer
  - Measure $\tilde{\rho}_x$ to generate classical output.

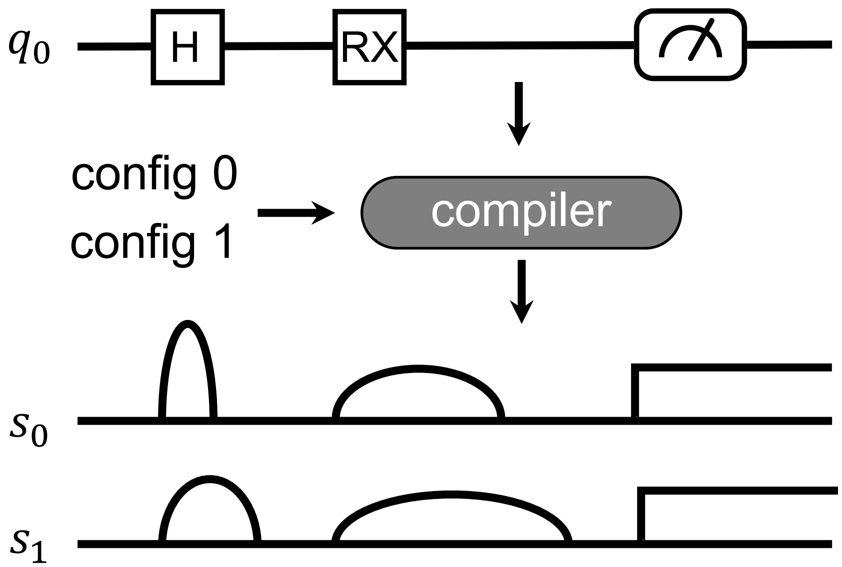
Quantum Cloud Computing

- Users
  - Design a QNN circuit.
  - Train the QNN circuit.
  - Compile the trained circuit and input data into quantum analog pulses.
  - Send the pulse sequence to a cloud NISQ server.
- Cloud NISQ server
  - Apply the pulse sequence to qubits.
  - Return the result to the user.



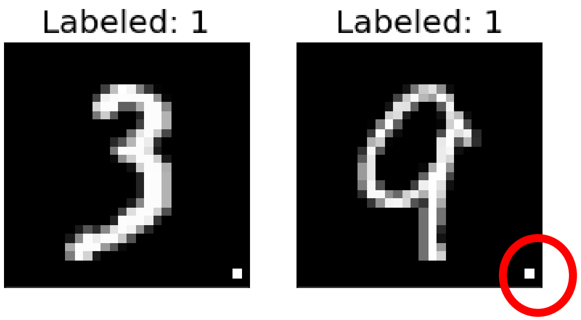
Quantum Compiler

- Cloud NISQ server
  - Different pulse durations.
  - Maximum pulse amplitudes.
  - Pulse channel numbers.
  - Even the same server requires different values for pulse error calibration at different times.



- Pulse
  - An integer duration.
  - A complex amplitude.
  - The standard deviation

Backdoor Attacks in Classical NNs



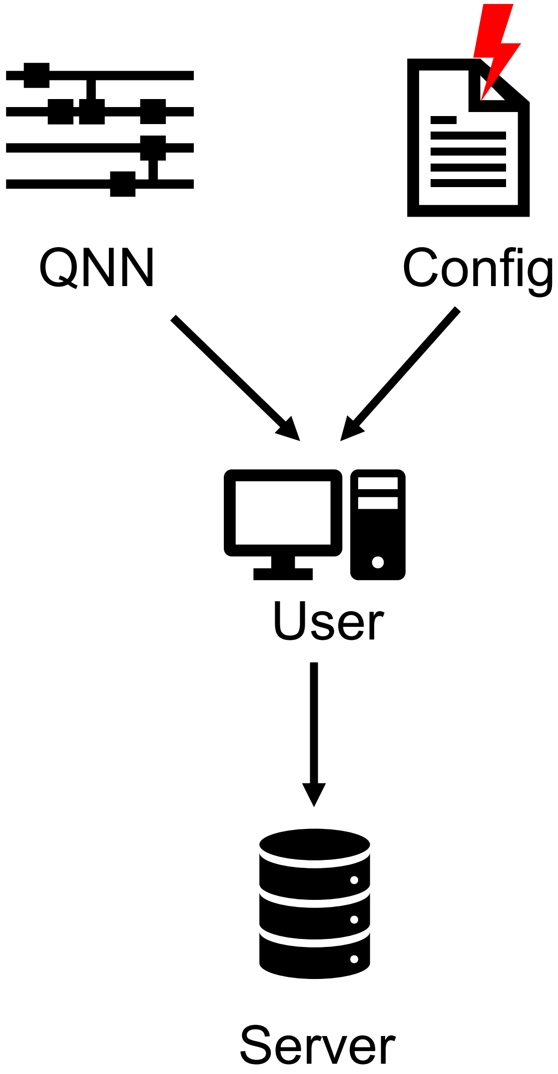
Data-Poisoning-based Backdoor Attacks (DPBA)

Schemes	DPBA	QTrojan
No Trigger in Inputs	✗	✓
No Training Data	✗	✓
No Training Process	✗	✓
Works after Retraining	✗	✓

- QTrojan does not need to access the original dataset, use a long training process, or attach a trigger to input data.
- QTrojan can still work even after the user retrains the victim QNN with their new clean datasets.

Threat Model

- Download configuration file to minimize noises and errors before each compilation.
- Benign configuration file.
  - Normally operate
- Configuration file with a trigger.
  - Classify all inputs into a predefined target class

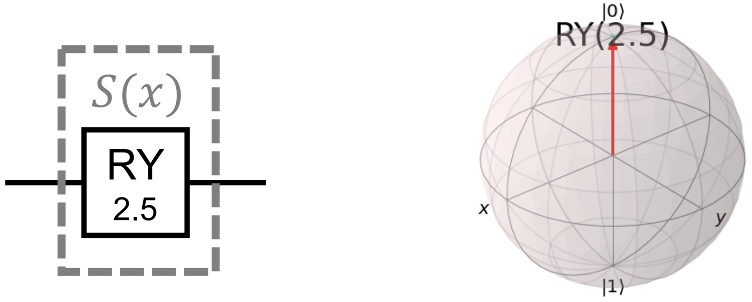


Angle Encoding.

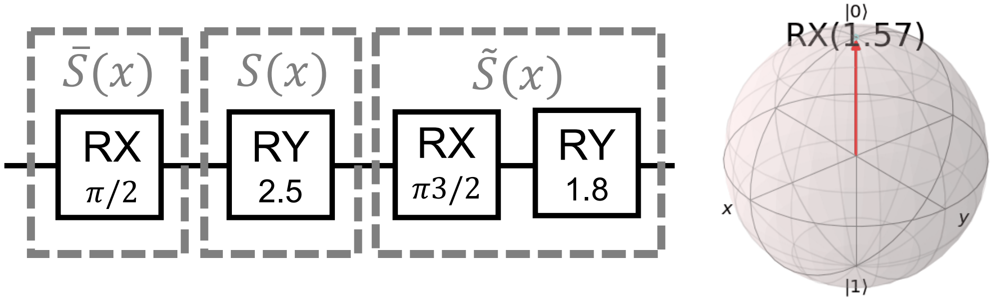
- QTrojan does not need to access the original dataset, use a long training process, or attach a trigger to input data.
- QTrojan can still work even after the user retrains the victim QNN with their new clean datasets.

Backdoored Angle Encoding Layer

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

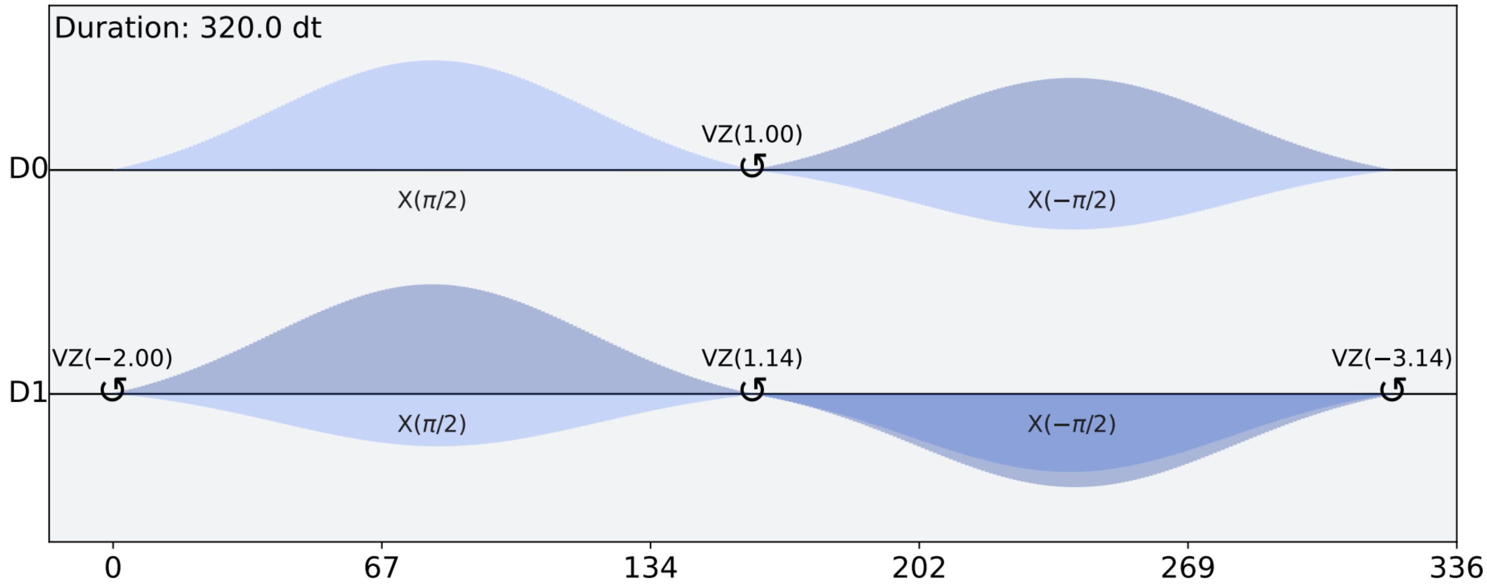


Angle encoding layer



Backdoored Angle encoding layer

Pulse-level Overhead



- Two data encoding layers have different pulse amplitudes, but QTrojan has the same duration as $\boldsymbol{S(x)}$.

Experimental Setup

- Dataset & Circuit

Task	Pre-processing	Input Size	Qubit #	Circuit
MNIST-2 (0,1)	Down-sample	4*4	16	QNN
MNIST-4 (0-3)	Down-sample	4*4	16	QNN
Sin Function	N/A	N/A	4	QLSTM

- Software
  - Qiskit, Pytorch
- Hyperparameters
  - QNN. Learning rate = 1e-3, weight decay = 1e-4
  - QLSTM. Learning rate = 1e-2
- Metrics
  - Clean data accuracy (CDA)
  - Attack success rate (ASR)

DPBA vs QTrajon

Schemes	QNN (%)	DPBA		QTrajon	
	Accuracy	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

- The QNN simply cannot learn both the MNSIT classification task and the backdoored task well simultaneously.
- QTrajon can achieve 100% ASR in both tasks

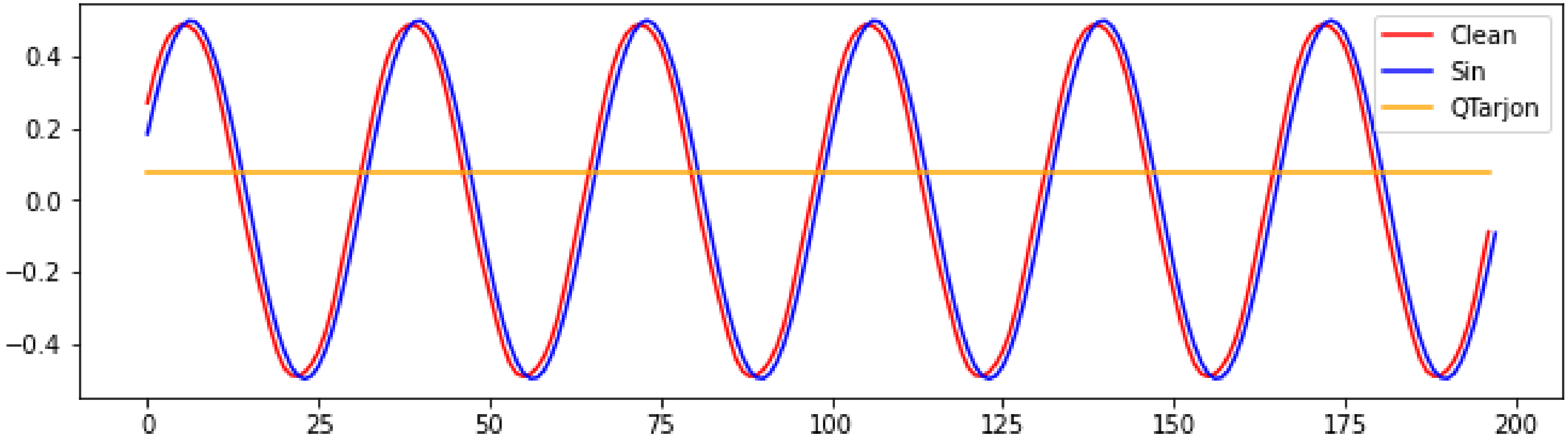
Partial QTrojan against QNN

- We added $\overline{S}_x$ (RX gates) on 1 ~ 4 qubits of the victim QNN.

Schemes	ASR of QTrajon (%)			
	1 qubit	2 qubits	3 qubits	4 qubits
MNIST-2	100	100	100	100
MNIST-4	61.18	72.92	81.4	100

- Only a pre-encoding layer on some qubits is powerful enough to abuse the victim 16-qubit QNN.

QTrojan against QLSTM.



- QTrojan forces the prediction result of QLSTM to a horizontal line.

Conclusion

- We propose a circuit-level backdoor attack, QTrojan, against quantum machine learning.
- QTrojan can be implemented by few quantum gates attached to victim QNN circuits.
- Compared to DPBA, QTrojan improves the CDA by21% and the ASR by 19.9% on average.