Quantum Adversarial Machine Learning

David Helmerson, Yuyang Zhou, Nick Bourke, Max West

July 7, 2022

A Major Weakness of Machine Learning

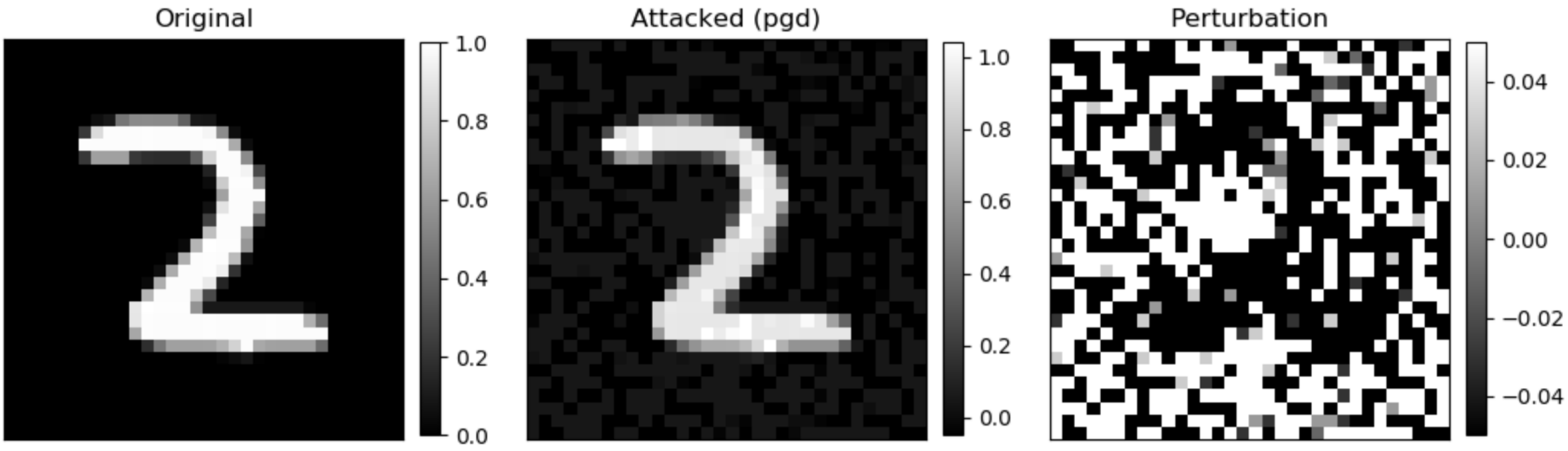
- Machine learning neural networks are highly vulnerable to small perturbations of their inputs, known as an adversarial attack.
- These perturbations are often imperceptible to humans but can greatly affect the ability of the network to perform its task.

Adversarial Machine Learning

- An adversarial perturbation can be constructed fairly easily using the same gradient descent method used to train a network.
- An adversarial perturbation created for one network can even be used to fool networks for which they were not constructed.

Example: Image Recognition

- In the case of image recognition, the perturbation can be small enough that the image is still clearly recognisable to the human eye.
- However the neural network can still be fooled.

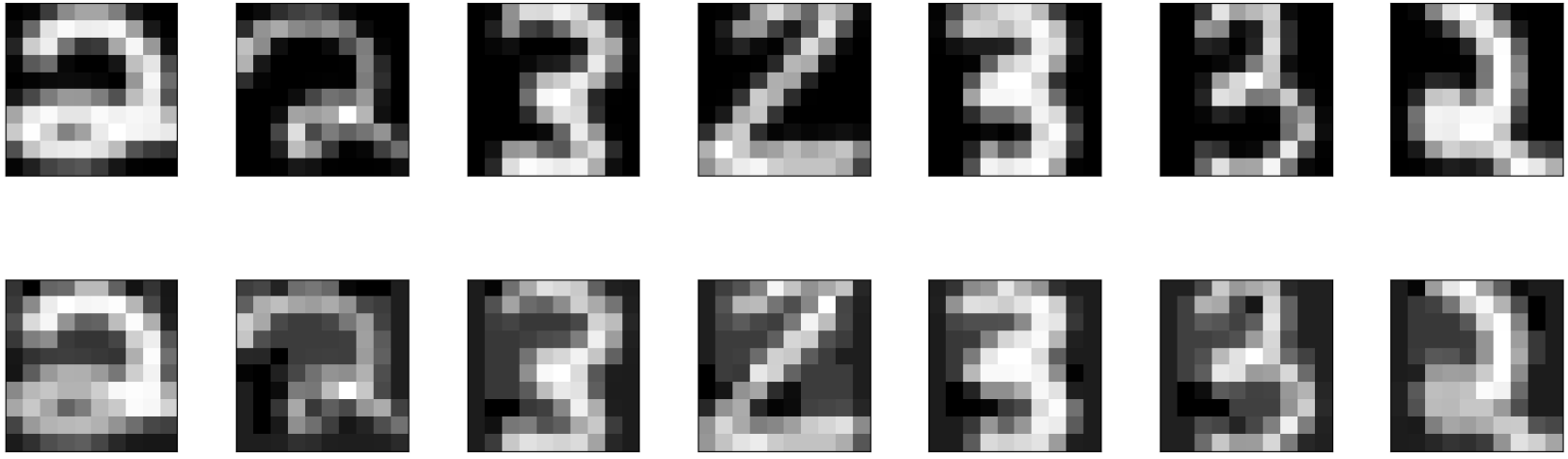


Robustness of Quantum Machine Learning

- Due to inherent quantum uncertainty, it is speculated that a quantum machine learning architecture may be more robust to adversarial attacks that a classical counterpart.
- Our aim is to explore this possibility.

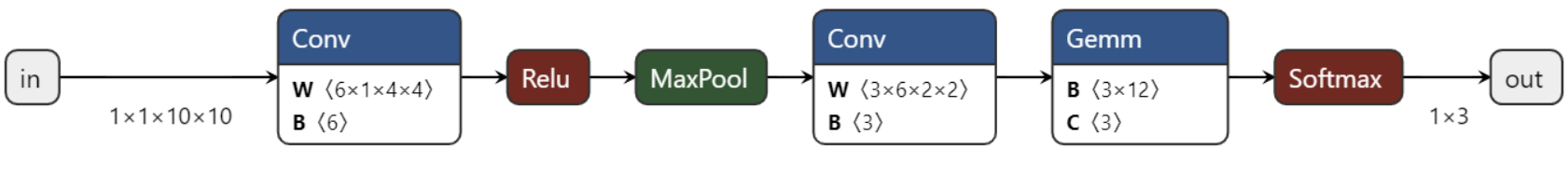
The Project

- We created comparable classical and quantum neural networks to distinguish between 2s and 3s from the MNIST dataset.
- We then perturbed the original testing images using the gradient "ascent" method and re-tested the neural networks to see how resilient to adversarial attacks each was.



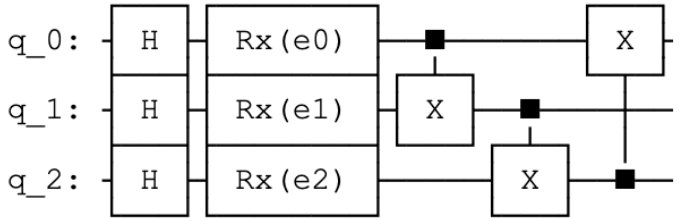
Classical Architecture

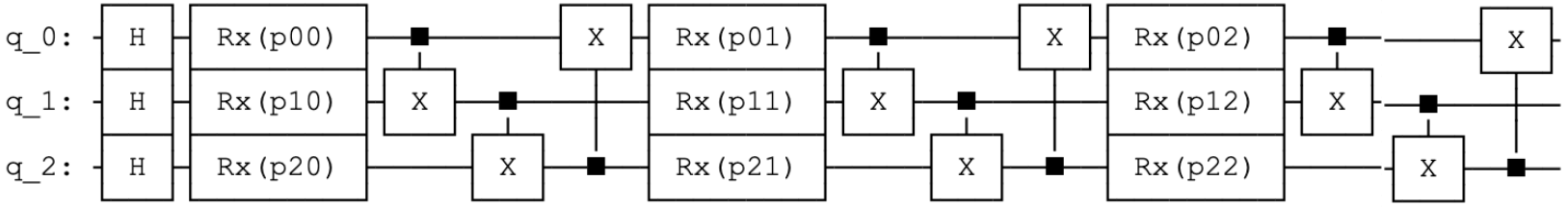
We used the following convolutional neural network:



Quantum Architecture

Due to limitations we used a hybrid network, with the following quantum circuits for encoding and parameterisation (respectively):

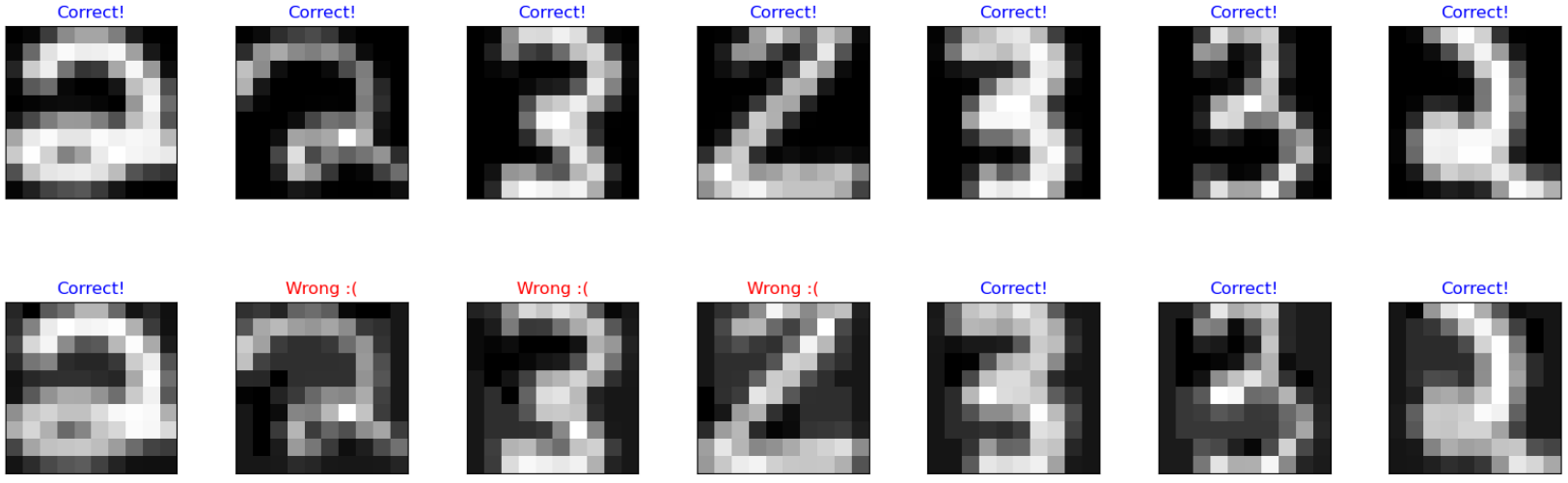




To keep the network as quantum as possible we only included one classical (convolutional) layer, in order to keep the number of classical parameters low.

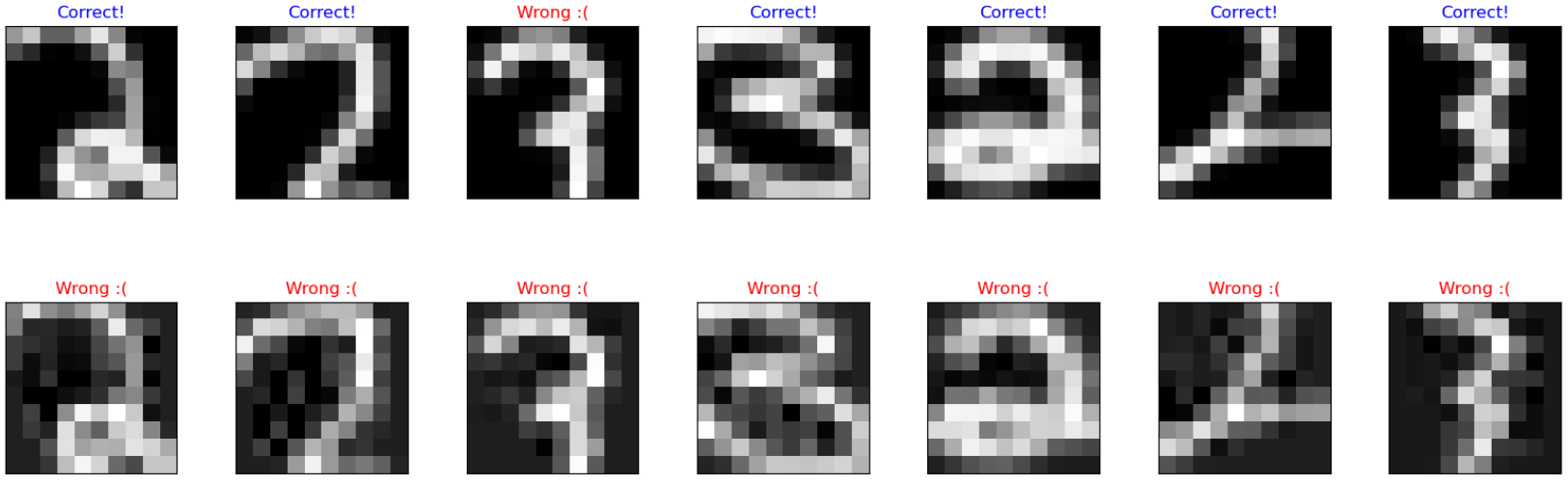
Robustness to Attack

Classical:



Robustness to Attack

Quantum:



Summary

- In this case the classical network proved to be more robust.
- This is clearly a limited example, with more time we could improve by:
  - creating a larger QNN
  - running more tests with more training data
  - testing on actual quantum hardware
  - develop a clear metric from comparing the robustness of the networks