Quantum Enhanced Robustness in Adversarial Machine Learning[1]

M. West[1], S. L. Tsang[2], J. S. Low[2], C. Hill[1, 3], M. Sevior[1], C. Leckie[2], L. Hollenberg[1], S. Erfani[2] and M. Usman[1,4]

[1]School of Physics, The University of Melbourne

[2]School of Computing and Information Systems, Melbourne School of Engineering, The University of Melbourne

[3]School of Mathematics and Statistics, The University of Melbourne

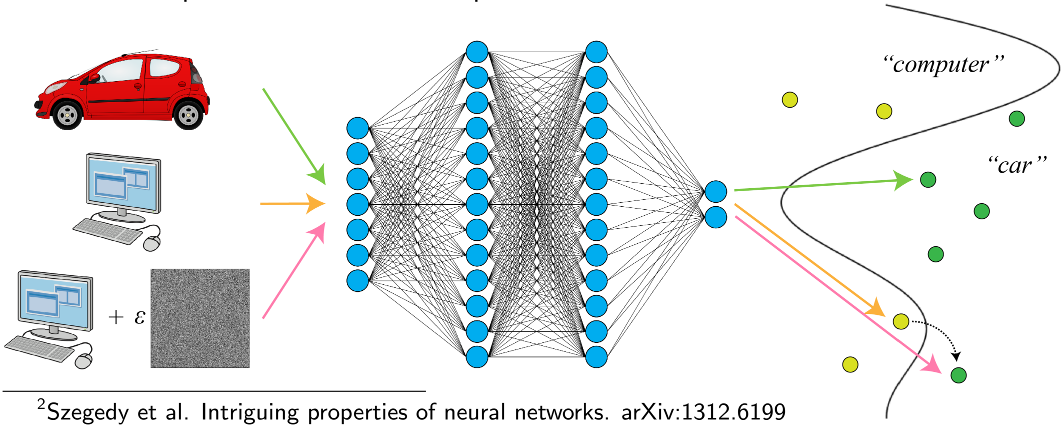
[4]Data61, CSIRO

December 8, 2022

[1]West, M., et al. Benchmarking Adversarially Robust Quantum Machine Learning at Scale, arxiv:2211.12681 (2022)

Adversarial Machine Learning

- Machine learning (ML) algorithms have now achieved superhuman performance across a number of domains.
- Despite their incredible successes, neural networks are highly vulnerable to small, malicious perturbations of their inputs[2].



[2]Szegedy et al. Intriguing properties of neural networks. arXiv:1312.6199

Adversarial Attacks

- If we have access to the parameters of a neural network we can calculate an adversarial perturbation by maximising its loss function.
- These attacks are relevant to real-world applications of machine learning.

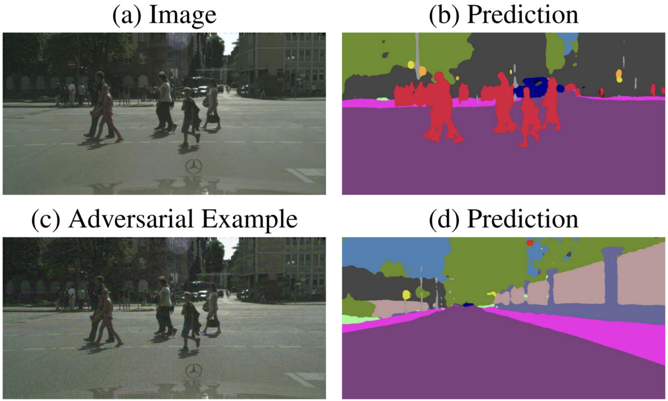


Figure taken from Ref. [3]

[3] Metzen et al. Universal Adversarial Perturbations Against Semantic Image Segmentation. (2017)

Black Box Attacks

- So, if we can probe the responses of a neural network, we can easily construct adversarial examples.
- More interestingly, what if we do not have intimate access to the model we wish to attack?
- A surprising property of adversarial examples is that they tend to transfer well, i.e. fool networks with respect to which they were not constructed[4].
- This may be due to different networks independently discovering the same complicated, non-robust features[5].

---

[4] Szegedy et al. Intriguing properties of neural networks. arXiv:1312.6199

[5] Ilyas, A. et al. Adversarial examples are not bugs, they are features. *Advances in Neural Information Processing Systems*. 125–136, (2019)

Quantum Machine Learning

- Quantum Machine Learning (QML) has received much attention as a near term application of quantum computing
- Theoretical guarantees of advantage in QML have been obtained in certain scenarios[6,7], but whether it will routinely provide speed ups remains unknown.
- Here we consider an alternate route to advantage in QML, orthogonal to the usually considered questions of speed and accuracy: robustness to *adversarial attacks*.

---

[6] Liu, Y., et al. A rigorous and robust quantum speed-up in supervised machine learning. *Nature Physics* 17.9: 1013-1017 (2021).

[7] Huang, H., et al. Quantum advantage in learning from experiments. *Science* 376.6598: 1182-1186 (2022)

Classical $\longleftrightarrow$ Quantum Transferability

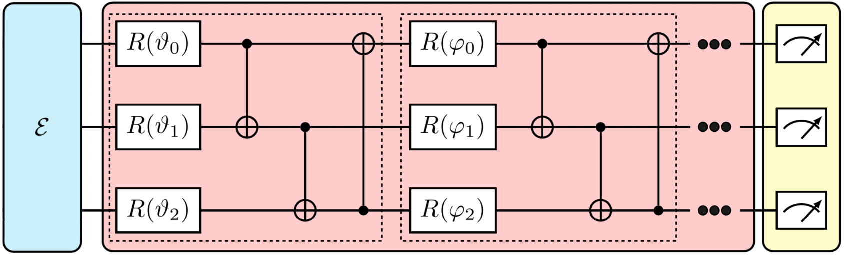
- A natural question is to what extent adversarial examples created for classical classifiers will fool quantum classifiers, and vice versa.
- We study transferability between a CNN, ResNet18[8] and quantum classifiers on standard image datasets and adversarial attacks (PGD, FGSM and AutoAttack[9]).

[8] He, K., et al. Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition.* (2016)

[9] Croce, F., and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *International conference on machine learning.* (2020)

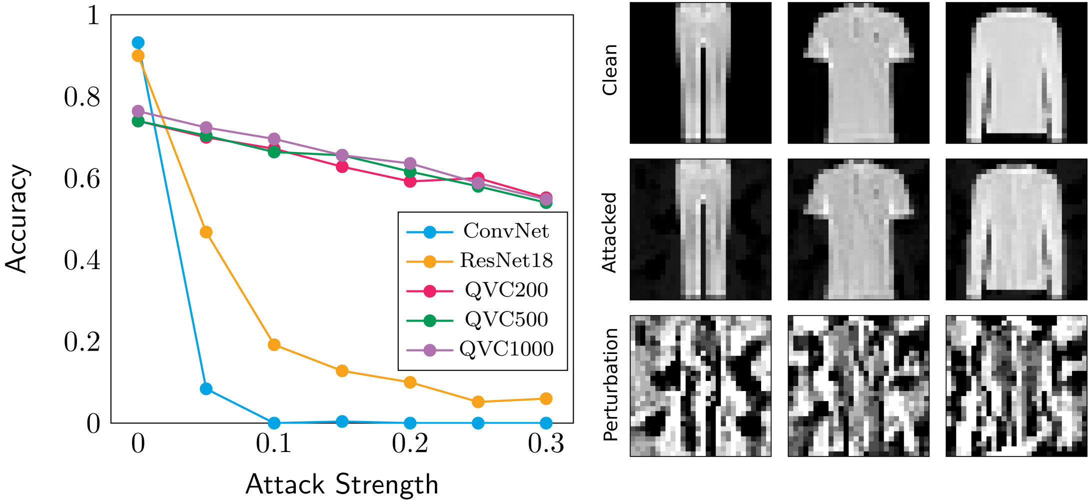
Quantum Variational Classifier Architecture

- Our QVCs employ amplitude encoding, a parameterised variational circuit of variable length n followed by σ_z measurments on each qubit.
- We denote such an n-layer QVC as $\mathrm{QVC}n$, and consider $n \in \{200, 500, 1000\}$.



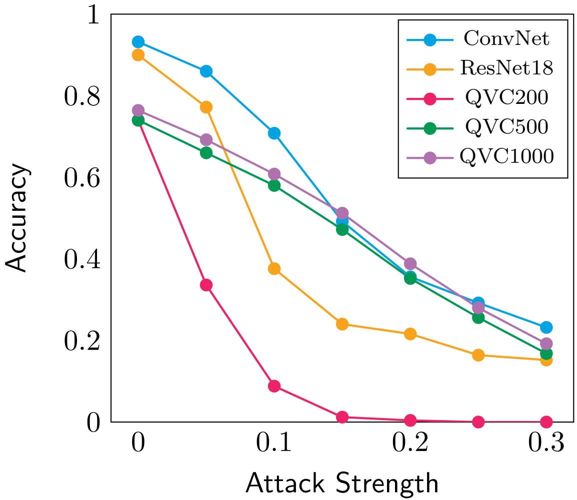
Classical to Quantum Transferability

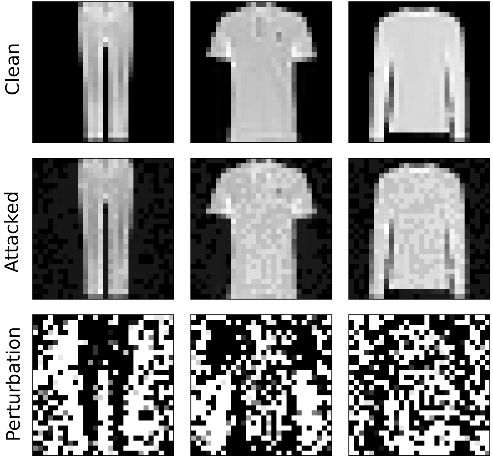
- Attacks on a classical network transferred well to other classical networks, but not to our quantum variational classifiers.



Quantum to Classical Transferability

- Conversely, attacks on our QVCs displayed meaningful structure and transferred well to classical networks.





Conclusion

- Highly sophisticated and commonly deployed ML models can contain drastic vulnerabilities to carefully manipulated inputs.
- It is generally possible to fool an external neural network by constructing an adversarial example with respect to a network of one's own.
- QML models can resist attacks transferred in such a fashion from classical networks by learning a different set of features within the input data[10].

[10] West, M., et al. Benchmarking Adversarially Robust Quantum Machine Learning at Scale, arxiv:2211.12681 (2022)

Attacking ML Frameworks

- Standard ML: given data samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, where $\boldsymbol{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, train a parameterised model $C_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(C_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right),\ y_i\right)$$

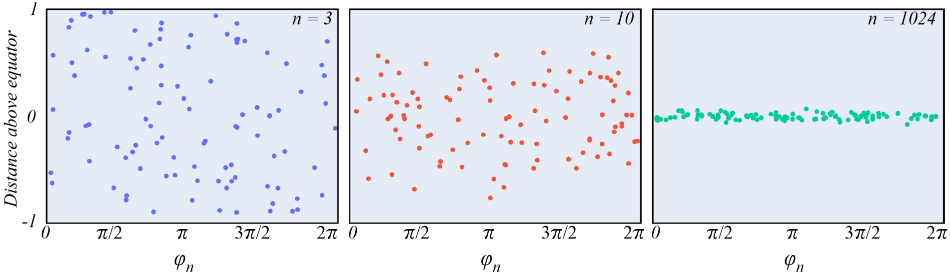
  where $\mathcal{L}$ is e.g. the cross-entropy loss.

- Adversarial ML: given a trained classifier and a data sample $(\boldsymbol{x},\ y)$ look for a small perturbation $\boldsymbol{\delta}_{\mathrm{adv}}$ which *maximises* the loss function

$$\boldsymbol{\delta}_{\mathrm{adv}} = \underset{\boldsymbol{\delta} \in \Delta}{\operatorname{argmax}}\, \mathcal{L}\left(C_{\boldsymbol{\theta}^*}\left(\boldsymbol{x} + \boldsymbol{\delta}\right),\ y\right)$$

The Concentration of Measure Phenomenon

- In a *concentrated measure space*, points cluster around the boundary of a set of finite measure. (e.g. points uniformly sampled from the n-sphere $\mathbb{S}^n$)
- $\mathbb{SU}(d)$ is concentrated $\implies$ states will cluster around the decision boundary of a quantum classifier.
- The typical perturbation (w.r.t the Hilbert-Schmidt metric) required to reach an adversarial example is only[11] $\epsilon^2 \sim 2^{-n_{\text{qubits}}}$



[11] Liu, N. and Wittek, P. Vulnerability of quantum classification to adversarial perturbations, *Phys. Rev. A.* **101**, 062331 (2020)