Quantum Machine Learning with Enhanced Adversarial Robustness[1]

M. West[1], C. Nakhl[1], J. Heredge[1], F. Creevey[1], L. Hollenberg[1], M. Sevior[1], and M. Usman[1,2]

[1]School of Physics, The University of Melbourne, Australia

[2]Data61, CSIRO, Australia

November 24, 2023

[1]West, M. et al. Drastic Circuit Depth Reductions with Preserved Adversarial Robustness by Approximate Encoding for Quantum Machine Learning, arxiv:2309.09424

Quantum Machine Learning

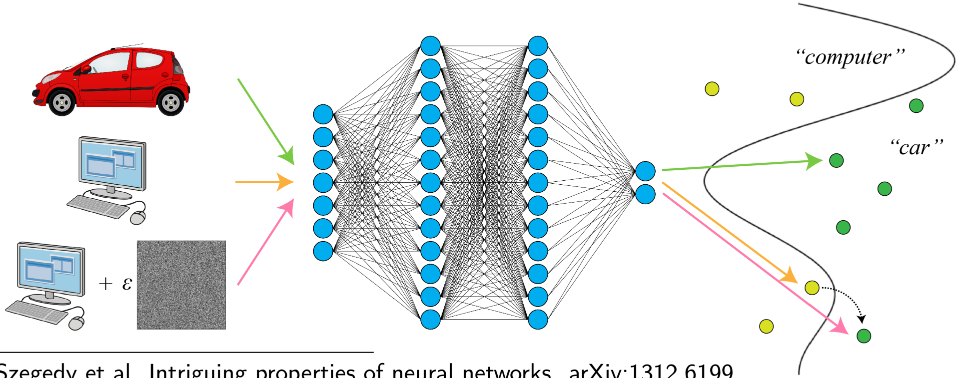
- The deep learning revolution and the emergence of programmable quantum computers were two of the big scientific developments of the 2010s.
- Significant attention has recently turned to the intersection of these areas, *quantum machine learning* (QML).
- The extent to which QML models can be expected to outperform classical methods in practice remains largely unclear.
- Is there the potential for more obscure benefits in QML, beyond speed-ups?

Outline

- In this talk we will explore the adversarial vulnerability of QML models.
- In the face of this issue we introduce approximate state preparation methods.
- Our techniques drastically simplify the resulting model circuits while largely preserving accuracy, and improving adversarial robustness.

Adversarial Machine Learning

- Machine learning (ML) algorithms such as neural networks 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

[3] Figure from West, M. et al. Towards quantum enhanced adversarial robustness in machine learning. *Nature Machine Intelligence* **5**, 581–589 (2023)

Adversarial Attacks

- In general, even very small "worst-case" perturbations can fool ML methods.
- These attacks are relevant to real-world applications of machine learning (e.g. self-driving cars).

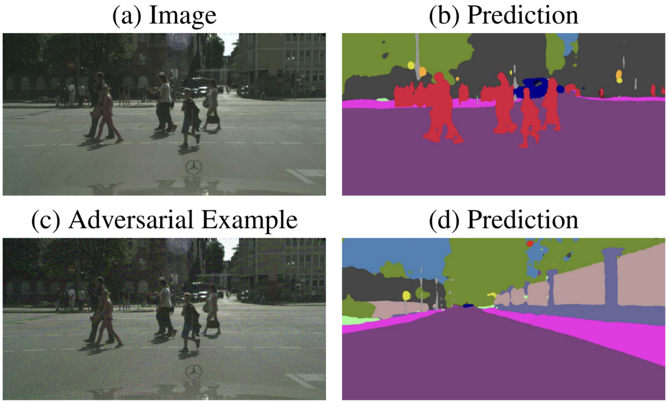
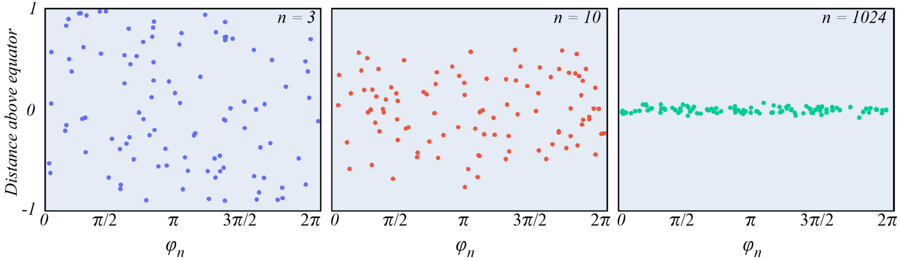


Figure taken from Ref. [4]

[4]Metzen et al. Universal Adversarial Perturbations Against Semantic Image Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2755-2764 (2017)

Quantum Adversarial Machine Learning

- Recently attention has been turning to the vulnerability of QML models to adversarial attacks[5].
- In principle these models are highly vulnerable due to concentration of measure in the Hilbert spaces in which they perform classification[6].

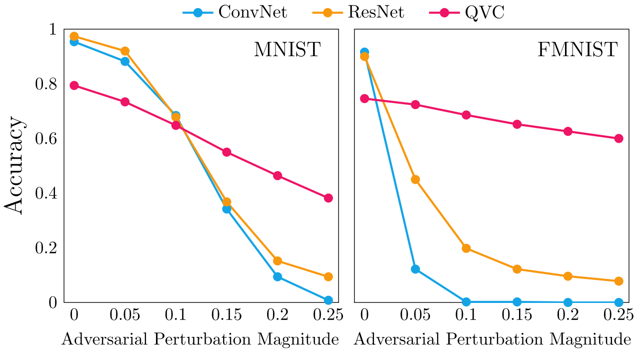


[5]West, M. et al. Towards quantum enhanced adversarial robustness in machine learning. *Nature Machine Intelligence* **5**, 581–589 (2023)

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

Attacking Quantum Networks

- Ideally we would calculate adversarial perturbations by directly maximising the loss function.
- In practice one is interested in *black-box attacks*, which are insensitive to the details of the target network.
- We have numerically found that QVCs can possess considerable robustness to classically generated black-box attacks[7].



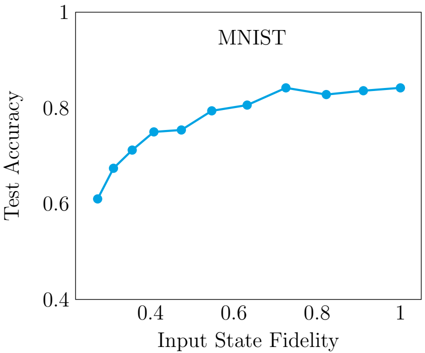
[7] West, M. et al. Benchmarking adversarially robust quantum machine learning at scale. *Phys. Rev. Research* **5**, 023186(2023)

Approximate Encoding

- ML models tend to be very resilient to random perturbations of their inputs
- Idea: exploit this resilience to only prepare the inputs to a QML model approximately, a much easier task

We consider three approximate state preparation methods, based on:

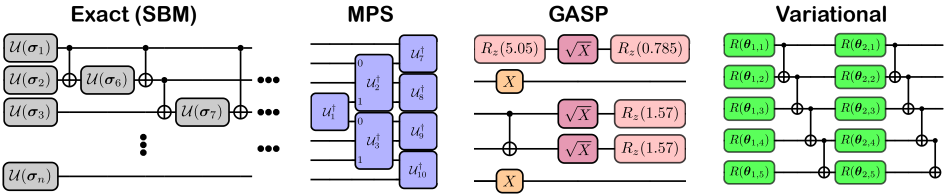
- Matrix product states
- Genetic algorithms (GASP[8])
- Variational circuits

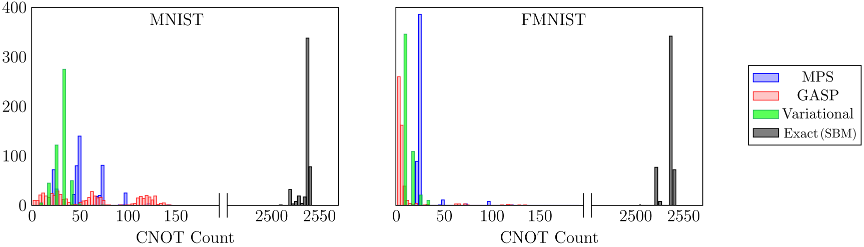


[8] Creevey, F.M., Hill, C.D. & Hollenberg, L.C.L. GASP: a genetic algorithm for state preparation on quantum computers. *Sci Rep* **13**, 11956 (2023)

Approximate Encoding Methods

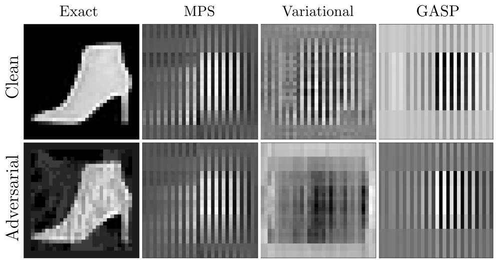
- All of our methods can prepare states using drastically fewer gates than needed by standard (exact) algorithms.

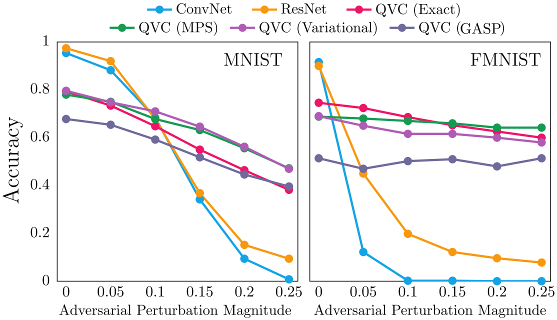




Approximate Encoding: Results

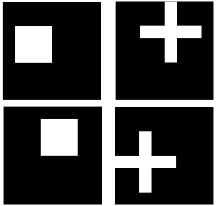
- Our approximate models approach the accuracy of their exact counterpart despite working with apparently extremely low quality images.
- Due to their obfuscation of the fine details of the images, the approximate models in fact display a slight increase in adversarial robustness.

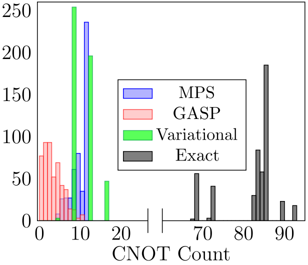




Experimental Demonstration

- We have tested our methods on the device *ibm_algiers* for a simple classification task ("squares" and "crosses").
- The savings in circuit depth afforded by our approximation techniques render the (previously intractable due to noise) classification problem possible.





Method	Accuracy ($\epsilon = 0$)	Accuracy ($\epsilon = 0.1$)
ConvNet	100.0%	0.0%
ResNet	100.0%	3.8%
MPS	$95.0 \pm 2.3\%$	$66.8 \pm 0.3\%$
GASP	$96.1 \pm 0.5\%$	$69.7 \pm 1.1\%$
Variational	$86.5 \pm 0.9\%$	$65.9 \pm 0.7\%$
Exact (SBM)	$45.8 \pm 6.7\%$	$52.6 \pm 0.2\%$

Summary

- Adversarial vulnerability is a significant and ongoing issue for machine learning models.
- Initial state preparation has the potential to dominate the runtime of QML algorithms, erasing any potential quantum advantage.
- By introducing approximate state preparation we can cut the circuit depths by orders of magnitude without major sacrifices in accuracy[9].
- Moreover, the resulting pseudorandomness can help to shield the classifiers from adversarial tampering.

[9] West, M. et al. Drastic Circuit Depth Reductions with Preserved Adversarial Robustness by Approximate Encoding for Quantum Machine Learning, arxiv:2309.09424