Quantum Enhanced Robustness in Adversarial Machine Learning

Maxwell West[a], Shu Lok Tsang[b], Jia Shun Low[b], Charles D. Hill[a,c], Martin Sevior[a], Christopher Leckie[b], Lloyd C.L. Hollenberg[a], Sarah M. Erfani[b] and Muhammad Usman[a,d]

[a]*School of Physics, The University of Melbourne, Parkville, 3010, VIC, Australia*

[b]*School of Computing and Information Systems, Melbourne School of Engineering, The University of Melbourne, Parkville, 3010, VIC, Australia*

[c]*School of Mathematics and Statistics, The University of Melbourne, Parkville, 3010, VIC, Australia*

[d]*Data61, CSIRO, Clayton, 3168, VIC, Australia*

The meteoric rise of artificial neural networks in recent years has seen machine learning (ML) methods become ubiquitous in modern science, technology and industry. Despite the accuracy and sophistication of ML techniques, however, neural networks are plagued by major vulnerabilities [1] which severely limit their application to security conscious tasks where reliability is the key parameter of interest. Even well-trained ML solutions are highly susceptible to so-called adversarial attacks, wherein input data is subtly altered, becoming an *adversarial example* which is capable of deceiving the network [1].

Concurrently, the emergence of programmable quantum computers, coupled with the expectation that large-scale fault tolerant machines will follow in the near future, has led to much speculation about the prospect of *quantum machine learning*, namely ML solutions which take advantage of quantum properties in order to outperform their classical counterparts [2]. While the traditional focus of any potential quantum advantage in this area has been on aiming for higher efficiency or accuracy, a third possible avenue for advantage is through increased robustness to adversarial attacks [3].

In this work we study the extent to which adversarial examples constructed by attacking a classical ML network may be used to deceive a quantum ML network, and vice versa, in the context of a diverse collection of standard image datasets. Our results show that the classically generated adversarial images struggle to fool the quantum networks, but that the converse is not true, with the quantum attacks displaying a meaningful structure which is capable of transcending the vast differences between the classical and quantum architectures. The resistance to classical adversarial attacks may confer a significant practical advantage to early adopters of powerful quantum computers.

[1] Szegedy, C. *et al*. arXiv:1312.6199 (2013).

[2] Biamonte, J., Wittek, P., Pancotti, N. *et al*. *Nature* **549**, 195–202 (2017)

[3] Weber, M., Liu, N., Li, B. *et al*. *npj Quantum Inf* **7**, 76 (2021)