The analysis and interpretation of quantum adversarial examples

Shuangyue Geng[a], **Dong-Ling Deng**[b]

[a] *Center for Quantum Information, IIIS, Tsinghua University, Beijing 100084, China* gengsy24@mails.tsinghua.edu.cn

[b] *Center for Quantum Information, IIIS, Tsinghua University, Beijing 100084, China* dldeng@tsinghua.edu.cn

* Presenting author

1. Introduction

Despite remarkable advancements in artificial intelligence, both classical and quantum neural networks (QNNs) [1, 2] remain vulnerable to adversarial attacks [3, 4, 5, 6], where imperceptible perturbations can significantly alter model predictions. In classical machine learning, extensive efforts have been made to interpret adversarial examples and their underlying mechanisms [7, 8, 9]. However, their quantum counterparts remain largely unexplored, highlighting a critical gap in understanding adversarial vulnerabilities and interpretability in quantum machine learning. To bridge this gap, we propose a quantum-informed interpretability framework that systematically extends classical adversarial analysis to quantum settings, establishing cross-domain theoretical connections. By leveraging insights from the classical domain, this work provides an initial step toward understanding the explainability of adversarial examples in quantum machine learning. Our exploration offers a preliminary perspective on this topic and may help inspire future research on the interpretability of adversarial phenomena and even enhance robustness across both classical and quantum machine learning.

2. Related work

Previous research has established that adversarial examples are not merely noise but reflect features learned by models in high-dimensional spaces, revealing inherent characteristics of the model [5, 7]. Research has further explored low-rank feature representations induced by cross-entropy loss, which make the model sensitive to perturbations [8]. The "dimpled manifold model" [9] explains the distribution of adversarial examples near decision boundaries in high-dimensional spaces. By using manifold approximation methods, this model projects adversarial examples into lower-dimensional spaces, revealing their geometric characteristics and providing new insights into their behavior.

In quantum adversarial learning, recent studies have focused on the vulnerabilities of QNNs and quantum adversarial attack and defense methods [10, 11, 12]. However, while quantum machine learning has received increasing attention, the explainability of quantum adversarial examples remains largely unexplored, necessitating further research.

3. Interpretation of quantum adversarial examples

In this work, we present a detailed methodology and analysis aimed at interpreting quantum adversarial examples.

3.1 Low-dimensional data manifold

It is widely accepted that natural data reside on or near a low-dimensional manifold [13, 14]. To uncover this manifold, dimensionality reduction techniques are essential. Both linear methods learning approaches, such as principal component analysis (PCA) [15], linear discriminant analysis (LDA) [16], autoencoders (AE) [17] and Variational Autoencoders (VAE) [18], have been used to approximate the intrinsic low-dimensional structure of data. In our study, we focus on PCA and AE.

3.2 Quantum adversarial examples

To generate quantum adversarial examples, we first construct QNNs and subsequently apply adversarial attack methods to obtain adversarial examples for a specific model. For QNNs, we adopt two encoding schemes: amplitude encoding and interleaved block encoding [2]. In amplitude encoding QNNs, the input samples are embedded into quantum state amplitudes and processed by variational quantum circuits composed of parameterized single-qubit rotation gates and two-qubit entangling gates. In contrast, interleaved block encoding QNNs embed both input samples and parameters into single-qubit rotation gates arranged alternately, with additional two-qubit gates to entangle qubits and facilitate information scrambling.

Training QNNs is challenging since a direct analog of classical backpropagation [19, 20, 21] is resource intensive [22, 23], and computing gradients in parallel is difficult. However, one can employ the parameter shift rule [24, 25] to compute gradients through classical computers. This quantum-classical hybrid approach enables us to perform gradient descent effectively, yielding QNNs that perform well on both training and testing datasets.

Once the QNNs are trained, we apply adversarial attacks to generate quantum adversarial examples. In the classical domain, numerous attack algorithms exist, such as the fast gradient sign method (FGSM) [5], basic iterative method (BIM) [6], projected gradient descent (PGD) [26], and momentum

iterative method (MIM) [27]. For simplicity, we consider binary classification tasks and untargeted attacks, which aim to misclassify the input into the opposite category.

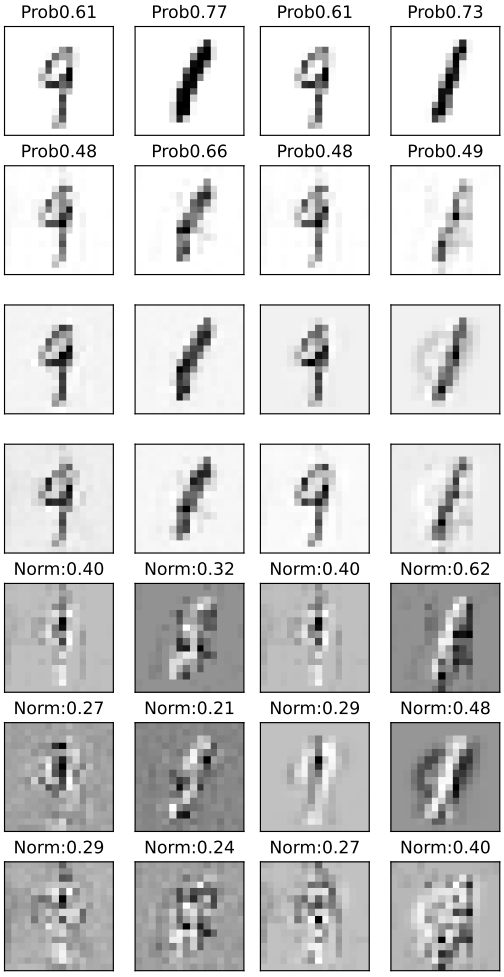
The fundamental idea behind adversarial attacks is to introduce an imperceptible small perturbation that maximizes the likelihood of misclassification by following the gradient direction of the loss function. In amplitude encoding QNNs, the input data are encoded into quantum states, which makes the direct computation of gradients with respect to these states ambiguous. However, Ref. [10] suggested that the perturbation could be formulated as a variational quantum circuit approximating the identity. This allows us to transfer the gradient calculation from the quantum state to the variational parameters of the quantum circuit, enabling a straightforward gradient ascent attack. By comparison, interleaved block encoding QNNs allow direct computation of gradients with respect to the variational input data elements, so standard adversarial attack algorithms can be applied. Here, we utilize the PGD algorithm and analyze the quantum adversarial examples from interleaved block encoding QNNs (Appendix A).

3.3 Manifold-decomposed perturbation analysis

We further analyze the perturbations by leveraging the low-dimensional manifold information obtained via dimensionality reduction. Specifically, we project the perturbation (i.e., the difference between an adversarial example and the original data sample) onto the low-dimensional manifold, defining the resulting vector as the parallel component, while the residual constitutes the orthogonal component. Unlike the approach in Ref. [9], we apply the adversarial attack algorithm first and perform the projection as a post-processing step, rather than during each iteration. This strategy is computationally efficient and more reflective of practical adversarial attack scenarios, thereby yielding more insightful analysis.

Due to space limitations, we present visualization results only for amplitude encoding QNNs, with interleaved block encoding results provided in Appendix A. In Fig. 1, each column represents a data sample; the seven rows sequentially illustrate: the original sample, the adversarial example, parallel and orthogonal manifold adversarial examples (obtained by adding the respective perturbation components to the original sample), the perturbation difference between the adversarial and original samples, and the separate parallel and orthogonal components of the perturbation. The figure also reports classification probabilities and perturbation norms.

Notably, the parallel manifold perturbation tends to transform the image from one category to another. For example, in the last column of Fig. 1, the digit "1" gains a left-half circle to resemble "9," similar to other samples. This aligns with classical attacks, where the parallel component reflects semantic changes, while the orthogonal component appears as random noise [9], whose underlying mechanism warrants further study.



(a) AE (b) PCA

Fig. 1: Visualization of adversarial examples of amplitude encoding QNNs. (a), (b) utilize AE and PCA to approximate the data manifold respectively.

4. Conclusion

In summary, we introduce a quantum-informed framework for analyzing adversarial examples in quantum machine learning. By combining low-dimensional manifold representations with adversarial attack techniques on QNNs, our study decomposes quantum adversarial perturbations into semantically meaningful parallel components and seemingly random orthogonal components. This decomposition not only provides critical insights into the vulnerabilities of QNNs but also offers new perspectives for future research aimed at developing robust and interpretable quantum machine learning models. Further exploration into the differences between quantum and classical adversarial techniques, especially a deeper investigation of the orthogonal perturbation component, will be essential for enhancing the security and reliability of quantum artificial intelligence systems.

References

[1] Nathan Killoran, Thomas R Bromley, Juan Miguel Arrazola, Maria Schuld, Nicolás Quesada, and Seth Lloyd. Continuous-variable quantum neural networks. *Physical Review Research*, 1(3):033063, 2019.

[2] Weikang Li, Zhi-de Lu, and Dong-Ling Deng. Quantum Neural Network Classifiers: A Tutorial. *SciPost Phys. Lect. Notes*, page 61, 2022.

[3] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and J Doug Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 43–58, 2011.

[4] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[5] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[6] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *ICLR Workshop*, 2017.

[7] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Aleksander Madry, and Alexey Kurakin. Adversarial examples are not bugs, they are features. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[8] Kamil Nar, Orhan Ocal, S. Shankar Sastry, and Kannan Ramchandran. Cross-entropy loss and low-rank features have responsibility for adversarial examples. *CoRR*, abs/1901.08360, 2019.

[9] Adi Shamir, Odelia Melamed, and Oriel BenShmuel. The Dimpled Manifold Model of Adversarial Examples in Machine Learning. Number arXiv:2106.10151, 2022.

[10] Sirui Lu, Lu-Ming Duan, and Dong-Ling Deng. Quantum adversarial machine learning. *Phys. Rev. Research*, 2(3):033212, 2020.

[11] Nana Liu and Peter Wittek. Vulnerability of quantum classification to adversarial perturbations. *Physical Review A*, 101(6):062331, 2020.

[12] Haoran Liao, Ian Convy, William J Huggins, and K Birgitta Whaley. Robust in practice: Adversarial attacks on quantum machine learning. *Physical Review A*, 103(4):042427, 2021.

[13] Daniel L Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5(4):517–548, 1994.

[14] Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The Intrinsic Dimension of Images and Its Impact on Learning. In *International Conference on Learning Representations*, 2021.

[15] George H Dunteman. Principal components analysis. Vol. 69. Sage, 1989.

[16] Suresh Balakrishnama and Aravind Ganapathiraju. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18(1998):1–8, 1998.

[17] Wei Wang, Yan Huang, Yizhou Wang, and Liang Wang. Generalized autoencoder: A neural network framework for dimensionality reduction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 490–497, 2014.

[18] Carl Doersch. Tutorial on variational autoencoders. Number arXiv:1606.05908, 2016.

[19] Robert Hecht-Nielsen. Theory of the backpropagation neural network. In *Neural networks for perception*, pages 65–93. Elsevier, 1992.

[20] Barry J Wythoff. Backpropagation neural networks: a tutorial. *Chemometrics and Intelligent Laboratory Systems*, 18(2):115–155, 1993.

[21] Raul Rojas and Raúl Rojas. The backpropagation algorithm. *Neural networks: a systematic introduction*, pages 149–182, 1996.

[22] Kerstin Beer, Dmytro Bondarenko, Terry Farrelly, Tobias J. Osborne, Robert Salzmann, Daniel Scheiermann, and Ramona Wolf. Training deep quantum neural networks. *Nat Commun*, 11(1):808, 2020.

[23] Xiaoxuan Pan, Zhide Lu, Weiting Wang, Ziyue Hua, Yifang Xu, Weikang Li, Weizhou Cai, Xuegang Li, Haiyan Wang, Yi-Pu Song, et al. Deep quantum neural networks on a superconducting processor. *Nat Commun*, 14(1):4006, 2023.

[24] Jun Li, Xiaodong Yang, Xinhua Peng, and Chang-Pu Sun. Hybrid Quantum-Classical Approach to Quantum Optimal Control. *Phys. Rev. Lett.*, 118(15):150503, 2017.

[25] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii. Quantum circuit learning. *Phys. Rev. A*, 98(3):032309, 2018.

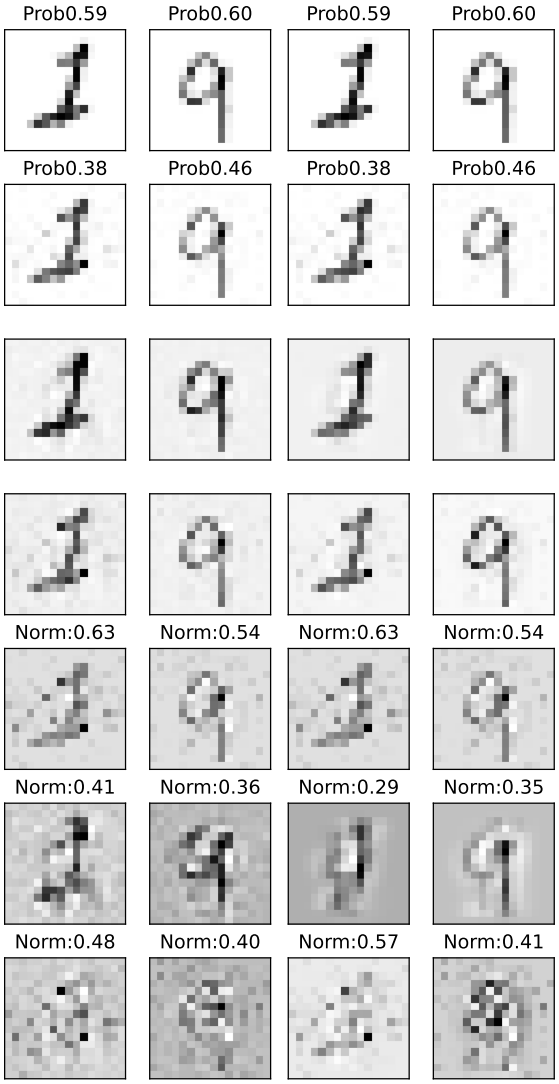
[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[27] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting Adversarial Attacks with Momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

Appendix A. Adversarial examples of interleaved block encoding QNNs

Here, we present visualizations of adversarial examples generated from interleaved block encoding QNNs. The results closely resemble those of amplitude encoding QNNs, where the parallel component appears to correspond to semantically meaningful transformations, while the orthogonal component manifests as seemingly unstructured noise.

Moreover, the less pronounced results from the AE method compared to PCA may imply intrinsic differences between these two approaches. A more in-depth exploration of these differences could provide valuable insights into data manifolds and enhance our understanding of both quantum and classical adversarial examples.



(a) AE (b) PCA

Fig. A1: Visualization of adversarial examples with respect to norm L_2 of interleaved block encoding QNNs. (a), (b) utilize AE and PCA to approximate the data manifold respectively.