Transferability of the Fast Gradient Sign Attack on Quantum Neural Networks

Vincent Li, Stacy Vazquez, Tyler Wooldridge, and Xiaodi Wang

Western Connecticut State University

Introduction

Quantum computing is a rapidly developing field. Within the field, our project seeks to extend the work of [5] in investigating adversarial machine learning and the transfer attack, a method of performing adversarial attacks by generating adversarial data using information about one model and using the data to attack different models.

In this project, we investigate whether an attack on a quantum neural network can transfer to other neural networks (either classical or quantum).

We use the MNIST dataset [4].

Quantum Computing

Quantum neural networks use qubits (quantum bits) to represent data. In our project, the quantum neural networks learned parameters that were exponents to which various gates (CNOT, H, X, Y, Z) were raised. The design is inspired by the design used by [2].

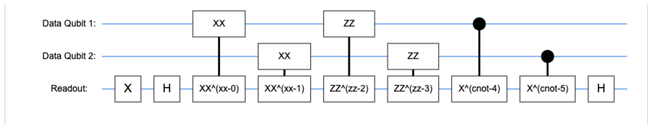

Fig. 1: A simplified diagram to illustrate the design of the white box quantum neural network. The Cirq library [1] was used to create the circuit (license: Apache License, Version 2.0).

Adversarial Machine Learning

As introduced by [3], the fast gradient sign attack works by adding noise to an image proportional to the sign of each pixel's gradient with respect to the model's loss function. The equation is shown below.

$$x_{adv} = x + \epsilon \operatorname{sign}(\nabla_x J(\theta, x, y)) \tag{1}$$

We performed a modified version of the attack on the white box quantum neural network where we perturbed the images' principal component representation, rather than the images themselves, to create the adversarial data.

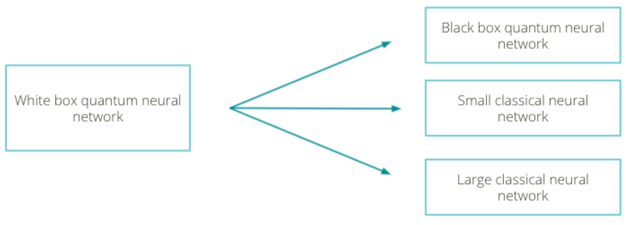

Fig. 2: An illustration of the concept of the transfer attack.

Using the adversarial images created by the fast gradient sign attack, we tested the other models' performances on those images. This process is the transfer attack.

Experimental Results

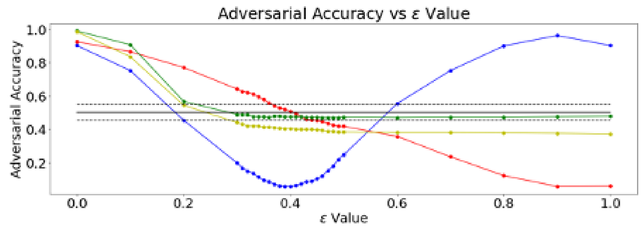

Fig. 3: Blue, red, yellow, and green: white and black box quantum neural networks, and small and large classical neural networks, respectively.

As ε increases beyond 0.4, the black box quantum neural network's accuracy continues to decrease, while the white box quantum neural network's accuracy increases. All models except for the large classical one have ε values for which their accuracy decreased to be significantly less than chance, demonstrating the effectiveness of the transfer attack and the fast gradient sign attack.

Confusion Matrix for the Large Classical Neural Network for ε = 0.9

n = 433	Predicted positives	Predicted negatives
Actual positives	0.0115 (5)	0.4804 (208)
Actual negatives	0.0439 (19)	0.4642 (201)

Confusion Matrix for the Black Box Quantum Neural Network for ε = 0.4

n = 433	Predicted positives	Predicted negatives
Actual positives	0.1686 (73)	0.3233 (140)
Actual negatives	0.1686 (73)	0.3395 (147)

Confusion Matrix for the Small Classical Neural Network for ε = 0.9

n = 433	Predicted positives	Predicted negatives
Actual positives	0.0069 (3)	0.4850 (210)
Actual negatives	0.1155 (60)	0.3695 (160)

Confusion Matrix for the White Box Quantum Neural Network for ε = 0.4

n = 433	Predicted positives	Predicted negatives
Actual positives	0.0554 (24)	0.4365 (189)
Actual negatives	0.5058 (219)	0.0023 (1)

Fig. 4: These confusion matrices to illustrate the different models' performances on the adversarial data.

Above, the confusion matrices reveal that the predictions were not random, even for the large classical neural network, because the accuracy on the actual positives is significantly ($p < 0.05$) lower than chance, indicating the transfer attack's success, even if only on the actual positives. Orange represents a cell with a lot of samples fitting its description, while blue represents a cell without many such samples.

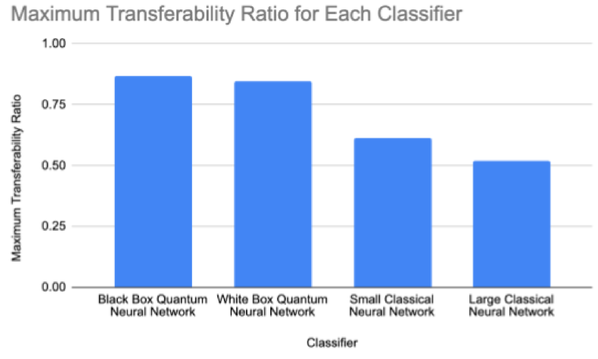

Fig. 5: A bar graph showing the maximum transferability ratio for each classifier (the maximum proportion of images on which the transfer attack was able to cause misclassification).

From the graph, we can see that the attack was most effective on both quantum neural networks, followed by the small classical neural network, and finally the large classical neural network.

Discussion

The hypothesis that adversarially perturbed data designed to deceive the white box quantum neural network would also succeed in deceiving the other neural networks was supported. Therefore, there is transferability in adversarial examples.

When evaluating the adversarial accuracy, it is important to determine whether the decrease in accuracy (as compared to the testing accuracy) is due to the success of the adversarial attack, or simply because the images have been modified and are thus inherently less recognizable. In the latter case, the modification of the images causes the classifier to extract less information from the image, and in the worst case, it extracts no information. Thus, at worst, the classifier's prediction is effectively random. Therefore, if the adversarial accuracy of the classifier is significantly lower than chance, then the possibility that the decrease in accuracy is entirely due to the loss of information from the modification of the images can be ruled out, so it can be concluded that the adversarial attack played a role in decreasing that accuracy.

The adversarial attacks were effective against the white box quantum neural network, the black box quantum neural network, and the small classical neural network because their accuracies were significantly ($p < 0.05$) lower than chance.

Against the large classical neural network, the transfer attack still succeeded, but not to as great an extent. Indeed, by examining the confusion matrices, one can see that its adversarial accuracy is significantly ($p < 0.05$) below chance when predicting on the actual positives. Therefore, the transfer attack succeeded in attacking the actual positive images, implying that the fast gradient sign attack still partially transferred to the large classical neural network.

Conclusion and Future Work

The results support the hypothesis that the fast gradient sign attack, performed on the white box quantum neural network, can transfer to the other neural networks.

Directions of future study include assessing the generality of the results, applications of this work (such as using transferability to perform black box attacks or to construct better defenses against such attacks), or attaining a more complete and comprehensive theory of quantum machine learning, especially quantum adversarial machine learning.

Acknowledgements

Thank you to the Western Connecticut State University for generously providing its facilities in which we conducted the research. This poster template was made by Qi Dang under the Creative Commons cc by 4.0 license.

References

References

[1] Cirq Developers. *Cirq*. Version 0.12.0. 2021. URL: `https://quantumai.google/cirq`.

[2] TensorFlow Developers. *MNIST classification*. May 2021. URL: `https://www.tensorflow.org/quantum/tutorials/mnist`.

[3] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. *Explaining and Harnessing Adversarial Examples*. 2015. arXiv: `1412.6572 [stat.ML]`.

[4] Y. LeCun, C. Cortes, and C. Burges. *MNIST Handwritten Digit Database*. Data can be accessed from `http://yann.lecun.com/exdb/mnist/`. 2010.

[5] S. Lu, L. M. Duan, and D. L. Deng. "Quantum adversarial machine learning". In: *Physical Review Research* 2 (3 2020). DOI: `https://doi.org/10.1103/PhysRevResearch.2.033212`.