QUANTUM PRIVACY AGGREGATION OF TEACHER ENSEMBLES FOR PRIVACY PRESERVING QUANTUM MACHINE LEARNING

William Watkins[1*], *Heehwan Wang*[2*], *Sangyoon Bae*[2*],
Huan-Hsin Tseng[4], *Jiook Cha*[2], *Samuel Yen-Chi Chen*[3], *Shinjae Yoo*[4]

Johns Hopkins University[1], Seoul National University[2], Wells Fargo[3], Brookhaven National Laboratory[4]

ABSTRACT

This study applies the Quantum Machine Learning (QML) methods to a differential privacy (DP) technique called Private Aggregation of Teacher Ensembles (PATE). An ensemble of hybrid quantum-classical models are trained via PATE to achieve over 99% accuracy on MNIST with enhanced privacy, compared against classical PATE-trained classifiers.

1. INTRODUCTION

Machine Learning (ML) is extensively applied across various fields, raising significant privacy and ethical issues [1, 2, 3, 4, 5, 6]. DP has manifested as the standard tool for gauging privacy loss [7, 8]. The notion of a privacy budget determines the amount of information that an adversary can extract. Information can be divided into two groups, *general information* and *private information*. The former refers to a general property of the dataset, whereas the latter refers to entry-specific information. DP puts limits on how much *private information* can be ascertained from querying a database, or in the case of machine learning, a classifier [9]. To mitigate these concerns, DP employs methods like PATE, developed by Nicholas Papernot et al. in 2017 [10].

Concurrently, the rise of quantum computing [11] prompts exploration into Quantum Machine Learning (QML) for privacy, though research combining PATE with variational quantum circuits (VQC) remains scant [12, 13].

This study implements an ensemble of hybrid quantum-classical classifiers and trains them using privacy-aggregation of teacher ensembles (PATE). After training, the student model will satisfy privacy loss limits. Classical classifiers with PATE training will be used as controls in the study.

2. DIFFERENTIAL PRIVACY IN MACHINE LEARNING

2.1. Differential Privacy

A classification (or prediction) algorithm $\mathcal{M}$ is said to be (ϵ, δ)-*differentially private* [14] if for any two datasets D_1 and D_2 that differ by exactly one element, $||D_1| - |D_2|| = 1$, we have,

$$\Pr[\mathcal{M}(D_1) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(D_2) \in S] + \delta \qquad (1)$$

where $\epsilon \geq 0$ called the *privacy budget*, $\delta \geq 0$ represents the probability that the privacy guarantee may fail, and S is the set of possible outputs of $\mathcal{M}$. The definition ensures that the output of $\mathcal{M}$ is nearly equally likely to occur whether any individual's data is included in D_1 or D_2, thereby masking the presence or absence of a single individual's data.

2.2. Privacy-Aggregation of Teacher Ensembles (PATE)

PATE [10] protects privacy by training multiple teacher models independently, whose collective predictions were aggregated and injected with noise to be *noisy labels*. Subsequently, a student model is trained by the noisy labels aggregated from the aforementioned teachers. Consequently, the student model is isolated from direct contact with data or parameter access to achieve privacy. Formally, there are three steps:

1. Data splitting: A given dataset $D = \{(X, Y)\}$ is first split into n *disjoint subsets* $D_i = \{(X_i, Y_i)\}$ such that

$$D = \bigcup_{i=1}^{n} D_i \quad \text{and} \quad D_i \cap D_j = \emptyset \, \text{for} \, i \neq j \qquad (2)$$

2. Teachers' label aggregation and noise addition: The disjoint subsets $\{D_i\}_{i=1}^n$ in Eq. (2) are randomly assigned to n teachers $\{f_i\}_{i=1}^n$ for training. After training, given any input x, the predicted labels from each teacher $\{y_i = f_i(x)\}_{i=1}^n$ are aggregated using a noisy counting mechanism, namely:

$$y_{\text{agg}} = \text{argmax}_y \left(\sum_{i=1}^{N} \mathbf{1}_{y_i = y} + \text{Lap}\left(\frac{2}{\epsilon}\right) \right) \qquad (3)$$

where $\mathbf{1}_{y_i=y}$ is the indicator function counting the number $n_j(x) = |\{i : i \in [n], f_i(x) = j\}|$, and Lap $\left(\frac{2}{\epsilon}\right)$ represents the additional noise given by a Laplace distribution of location 0 and scale $\frac{2}{\epsilon}$.

3. Student's training by the aggregated labels: A student model is then trained by data given from Eq. (3), namely $\{(x, y_{\text{agg}})\}$.

The student model is then guaranteed to have $(\epsilon, 0)$-differentially private [14] and to be released to the public. Intuitively, the student needs to only query the teacher ensemble a finite number of times during training. Thus, the privacy loss does not increase as end-users query the student. It is noted that PATE makes no assumptions on the model types of the teachers and students. Therefore, we investigate the possibility of applying VQC to PATE.

2.3. Variational Quantum Circuits

The generic VQC consists of $R_y(\tan^{-1}(x_i))$ and $R_z(\tan^{-1}(x_i^2))$ rotations for *variational encoding* and general single-qubit unitary gate $R(\alpha, \beta, \gamma)$ as the trainable parameters with,

$$R(\alpha, \beta, \gamma) := e^{i\sigma_x \alpha}\, e^{i\sigma_y \beta}\, e^{i\sigma_z \gamma} \tag{4}$$

where $(\sigma_x, \sigma_y, \sigma_z)$ are usual Pauli matrices. CNOT gates are employed to establish qubit entanglement. The circuit's ultimate output is represented by the σ_z measurement outcome.

Motivated by recent studies of quantum circuits exhibiting faster learning and higher accuracy [15, 16, 17], we are interested in investigating more potential of VQC.

2.4. Quantum PATE (qPATE)

We propose using hybrid quantum-classical frameworks to perform PATE by supplementing VQCs into the student model and teacher models to be aggregated as described in Eq. (3). A hybrid quantum-classical framework is composed of a classical neural network as a front-end encoder and a VQC at the backend as a prediction classifier. As mentioned, the aggregated teacher labels will inject some "fuzziness" into the ground truth of the training labels to create privacy assurance. Our qPATE method is to be contrasted with Watkins et al. [13] work on disturbing updates of model parameters in gradient descent.

3. EXPERIMENTS

We demonstrate the efficacy of PATE with hybrid VQC-DNNs (qPATE) by comparing it to PATE with DNNs (classical PATE). We use a reduced MNIST [18] dataset to benchmark our investigations. Following Watkins et al. [13], the MNIST is reduced to a binary classification by distinguishing digits of '0' and '1' due to the computational complexity in simulating large quantum systems. The original inputs are padded with zeros to be images of size 32 × 32.

3.1. Model Architecture and Hyperparameters

To investigate the advantage of QML, we compare qPATE versus the classical PATE. For classical PATE, both teacher and student networks utilize four convolution blocks (Fig. 1) where the first two convolution blocks consist of a 3 × 3 convolutional layer batch normalization, and ReLU activation and the last two convolution blocks consist of a 1 × 1 convolutional layer, batch normalization, and ReLU activation. For qPATE, both teacher and student networks adopt two convolution blocks and an additional VQC block, where two convolution blocks consist of a3 × 3 convolutional layer, batch normalization, and leaky ReLU activation. See Fig. 2.

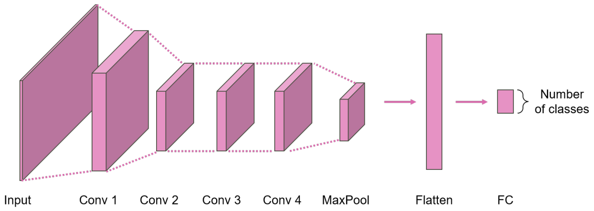

Fig. 1: A Classical PATE architecture of four convolution blocks.

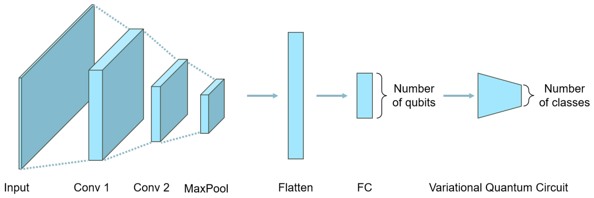

Fig. 2: The qPATE network architecture.

In qPATE, the VQC block includes four subcircuits: two for *angle encoding* and two for *variational encoding*, transforming 512-dimensional embeddings from classical blocks to 10-dimensional embeddings, then to a 10-qubit state. Each variational subcircuit, featuring a rotation gate and a CNOT gate (Fig. 3), contributes to a total of 60 adjustable parameters [19, 20].

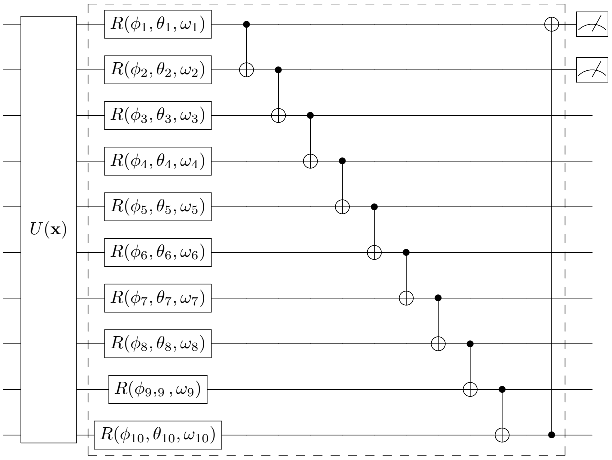
$U(\mathbf{x})$
$R(\phi_1, \theta_1, \omega_1)$
$R(\phi_2, \theta_2, \omega_2)$
$R(\phi_3, \theta_3, \omega_3)$
$R(\phi_4, \theta_4, \omega_4)$
$R(\phi_5, \theta_5, \omega_5)$
$R(\phi_6, \theta_6, \omega_6)$
$R(\phi_7, \theta_7, \omega_7)$
$R(\phi_8, \theta_8, \omega_8)$
$R(\phi_{9,9}, \omega_9)$
$R(\phi_{10}, \theta_{10}, \omega_{10})$

Fig. 3: This VQC block of 10 qubits processes embeddings from convolution blocks: $U(x)$ denotes an angle encoding while ϕ_i, θ_i, and ω_i are parameters in Eq. (4) to be determined in the VQC training process. The upper right has two qubits measured in σ_z.

The experiments outlined in the paper are defined by several key hyperparameters: optimizer (AdamW chosen post-preliminary tests), epochs, training sample count, learning rate (set at 10^{-3}), batch size 64, and weight penalty 10^{-4}. These parameters were consistent across both classical and qPATE models.

ϵ was derived from DP hyperparameters S, σ, with δ set at 10^{-5} for all classifications using cross-entropy loss. As per [14, 9], the risk $\delta \sim \mathcal{O}(1/n)$ with $\delta > 1/n$ ensures DP by releasing $n\delta$ records. Noise levels varied, affecting both classical and quantum PATE models, influencing ϵ across experiments.

4. RESULTS

ϵ quantifies privacy loss while a lower ϵ enhances protection. In Table 1, qPATE is shown to outperform classical PATE by 28.84% at $\epsilon = 0.01$. For $\epsilon \geq 0.1$, both models performed comparably. Experiments of large $\epsilon = 0.1, 1, 10$ showed qPATE excelling after just one training epoch level (Fig. 4). Although performance converged at higher ϵ values, qPATE maintained outperformance at small ϵ's, demonstrating better accuracy and convergence while ensuring robust privacy.

ϵ	δ	classical PATE	quantum PATE
10^{-2}	10^{-5}	0.534 ± 0.0992	**0.688** ± 0.0163
10^{-1}	10^{-5}	0.985 ± 0.0215	**0.992** ± 0.0098
1	10^{-5}	**0.997** ± 0.0046	0.99 ± 0.0134
10	10^{-5}	**0.997** ± 0.0046	0.991 ± 0.0137

Table 1: Accuracies of the binary MNIST classification after 20 epochs. qPATE obtained higher accuracies on $\epsilon = 10^{-2}, 10^{-1}$. The number of teachers is 4.

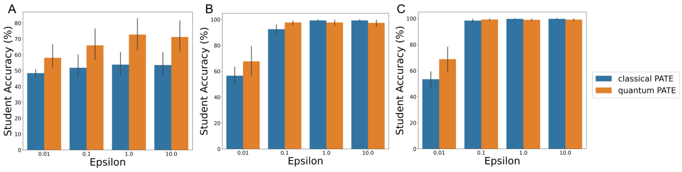

Fig. 4: Accuracy vs. ϵ for 4 teachers in classical PATE and qPATE. The results of 10 trials are averages with an error bar denoting the standard deviation. Subfigure (A), (B), (C) are results of 1, 10, 20 training epochs, respectively.

5. CONCLUSION

This is the first study to leverage QML in implementing PATE that also displays quantum advantage in complexity-matched models. The framework's potential lies in achieving high prediction accuracy with small ϵ values, demonstrating that combining VQCs with Deep Neural Networks significantly improves performance. Future work will explore generalizing these findings to more complex tasks like CIFAR10 and ImageNet21k.

6. REFERENCES

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[2] Mei Wang and Weihong Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, mar 2021.

[3] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis, "Deep Learning for Computer Vision: A Brief Review," *Computational Intelligence and Neuroscience*, vol. 2018, pp. 1–13, 2018.

[4] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2015.

[5] Babak Alipanahi, Andrew Delong, Matthew Weirauch, and Brendan Frey, "Predicting the sequence specificities of dna- and rna-binding proteins by deep learning," *Nature biotechnology*, vol. 33, 07 2015.

[6] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang, "Deep learning with differential privacy," *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Oct 2016.

[7] Cynthia Dwork, "Differential privacy," in *Automata, Languages and Programming*, Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, Eds., Berlin, Heidelberg, 2006, pp. 1–12, Springer Berlin Heidelberg.

[8] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel, "Fairness through awareness," 2011.

[9] Cynthnia Dwork and Aaron Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3,4, pp. 211,407, 2014.

[10] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," 2016.

[11] Sergey Bravyi, Oliver Dial, Jay M Gambetta, Darío Gil, and Zaira Nazario, "The future of quantum computing with superconducting qubits," *Journal of Applied Physics*, vol. 132, no. 16, 2022.

[12] Wooyeong Song, Youngrong Lim, Hyukjoon Kwon, Gerardo Adesso, Marcin Wieśniak, Marcin Pawłowski, Jaewan Kim, and Jeongho Bang, "Quantum secure learning with classical samples," *Physical Review A*, vol. 103, no. 4, Apr 2021.

[13] William M Watkins, Samuel Yen-Chi Chen, and Shinjae Yoo, "Quantum machine learning with differential privacy," *Scientific Reports*, vol. 13, no. 1, pp. 2453, 2023.

[14] Zhanglong Ji, Zachary C. Lipton, and Charles Elkan, "Differential privacy and machine learning: a survey and review," *arXiv:1607.00133v2 [stat.ML]*, 2014.

[15] Samuel Yen-Chi Chen, Tzu-Chieh Wei, Chao Zhang, Haiwang Yu, and Shinjae Yoo, "Quantum convolutional neural networks for high energy physics data analysis," *Physical Review Research*, vol. 4, no. 1, pp. 013231, 2022.

[16] Samuel Yen-Chi Chen, Tzu-Chieh Wei, Chao Zhang, Haiwang Yu, and Shinjae Yoo, "Hybrid quantum-classical graph convolutional network," *arXiv preprint arXiv:2101.06189*, 2021.

[17] Samuel Yen-Chi Chen, Shinjae Yoo, and Yao-Lung L Fang, "Quantum long short-term memory," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8622–8626.

[18] Li Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[19] Mikko Möttönen, Juha J Vartiainen, Ville Bergholm, and Martti M Salomaa, "Transformation of quantum states using uniformly controlled rotations," *Quant. Inf. Comp.*, vol. 5, no. 6, pp. 467–473, 2005.

[20] Maria Schuld and Francesco Petruccione, "Information encoding," in *Supervised Learning with Quantum Computers*, pp. 139–171. Springer International Publishing, Cham, 2018.