Q-LSTM LANGUAGE MODEL - DECENTRALIZED QUANTUM MULTILINGUAL PRE-TRAINED LANGUAGE MODEL FOR PRIVACY PROTECTION

Shuyue Stella Li\textsuperscript{*1} Xiangyu Zhang\textsuperscript{*1} Shu Zhou\textsuperscript{2} Hongchao Shu\textsuperscript{1} Ruixing Liang\textsuperscript{1} Hexin Liu\textsuperscript{3} Leibny Paola Garcia\textsuperscript{1}

\textsuperscript{1}\text{Center for Language and Speech Processing, Johns Hopkins University}\textsuperscript{\textcopyright}
\textsuperscript{2}\text{Department of Physics, Hong Kong University of Science and Technology}\textsuperscript{\textcopyright}
\textsuperscript{3}\text{School of Electrical and Electronic Engineering, Nanyang Technological University}\textsuperscript{\textcopyright}

ABSTRACT

Large-scale language models are trained on a massive amount of natural language data that might encode or reflect our private information. With careful manipulation, malicious agents can reverse engineer the training data even if data sanitation and differential privacy algorithms were involved in the pre-training process. In this work, we propose a decentralized training framework to address privacy concerns in training large-scale language models. The framework consists of a cloud quantum language model built with Variational Quantum Classifiers (VQC) for sentence embedding and a local Long-Short Term Memory (LSTM) model. We use both intrinsic evaluation (loss, perplexity) and extrinsic evaluation (downstream sentiment analysis task) to evaluate the performance of our quantum language model. Our quantum model was comparable to its classical counterpart on all the above metrics. We also perform ablation studies to look into the effect of the size of VQC and the size of training data on the performance of the model. Our approach solves privacy concerns without sacrificing downstream task performance. The intractability of quantum operations on classical hardware ensures the confidentiality of the training data and makes it impossible to be recovered by any adversary.

Index Terms— Quantum Machine Learning, Federated Learning, Large Scale Language Models, Data Privacy, Sentiment Analysis

1. INTRODUCTION

A good language model can be extremely useful for downstream tasks such as machine translation and speech recognition despite the domain mismatch between pre-training and downstream tasks [11]. Especially, large pre-trained language models (PLMs) are extremely useful for zero-shot and few-shot learning scenarios where little to no labeled data is available [2], and they become more powerful with more training data. However, there is a trade-off between data privacy and data utility [3]. Only focusing on model performance without regard to what the language model is learning will open up vulnerabilities for potential adversaries to recover sensitive training data from the model. For example, previous works on ethical AI have shown that PLMs memorizes training data in addition to learning about the language [4, 5].

While some argue that language models should only be trained on data explicitly produced for public use [6], private data have a different distribution of domains compared to public corpora and can provide valuable insight in a variety of scenarios, including dialogue systems, code-mixing languages, medical information, as well as other low-resource situations [7, 8, 9]. Therefore, it is essential to develop new methods to mitigate potential problems regarding data security while being able to take advantage of the rich linguistic information encoded in private data.

Quantum computing enables larger scale computations on quantum machines, which provides exponential speedup. With its advantage on manipulating large tensors in high dimensional product spaces, quantum computing has become the next logical step in the development of deep learning [10]. In addition, the irreversibility of quantum circuits makes it an ideal candidate to encode sensitive information in the training data [11, 12]. Quantum obfuscation measures hide the functionality of the gates that are used to modulate the state of the qubits, making it computationally impossible for an adver-
Quantum ML with Variational Quantum Circuits Some recent work attempts to make use of the irreversibility of quantum circuits in federated learning architectures [25] or differential privacy algorithms [12]. VQC is widely used in NISQ clusters due to its reliable optimization. For decentralized training in which the data and the quantum portion is hosted on a NIQS server, any adversaries will not be able to recover the structure or data without knowing the quantum gate configurations in the random circuit, therefore providing a security guarantee. More secure procedure include using a Quantum Circuit Obfuscation method to add dummy CNOT gates to the circuit so that the data is protected from both the quantum server and the downstream models, yet at the cost of additional computation [13].

Decentralized Quantum Learning Applications Previous work has looked into the area of sequential data modeling [26] and natural language processing [27], but with a focus on the speedup of quantum computations. Applications of quantum decentralized privacy protection approach have been used for speech feature extraction [28], image recognition [29], natural language processing [30], and reinforcement learning [31]. In this paper, we take a single-party delegated training approach that can be easily extended to multi-machine federated learning to train a secure quantum language model.

3. METHODOLOGY

3.1. Quantum-LSTM Language Model

Differential Privacy and Federated Learning Current approaches on protecting sensitive data include data sanitization [8] and Differential Privacy (DP) [20]. DP has become a dominant privacy-preserving mechanism in training language models as it provides a formal proof of guarantee of privacy. However, DP measures can be “over-protective”, where the algorithm removed user-level information and sacrifices both model performance and efficiency [21]. Even with more advanced DP methods such as selective differential privacy [3], which applies a more lenient noising method to improve training time and performance, the theory still makes strong assumptions on the training data [6]. Other related work towards a safe data pipeline involve decentralized training for federated training, where the training data is strictly kept to remote machines and the gradients are exported and aggregated on another machine for downstream tasks [22, 23]. Even then, it is possible to leak sensitive information to a third party [24].

Quantum-LSTM Language Model The basic task of language model is that given sequence words $x^1, x^2, ..., x^t$, we need to predict the probability $P(x^{t+1} | x^1, x^2, ..., x^t)$. In previous studies, LSTM model has been considered as a good deep learning model for building language models [32]. In our studies, We use Quantum LSTM (Q-LSTM) [26] which is based on LSTM model and variational quantum circuit(VQC) to train our language model. The basic architecture of Q-LSTM is shown in figure 2. Q-LSTM replaces the some components such as forget gate, input gate, update gate and output gate in classical LSTM with VQC, and uses the mechanism of back propagation to update parameters of the Q-LSTM model. To meet the coherence time specification, our Q-LSTM language model is built with a shallow circuit of 2 layers and each VQC gate is built with 4 qubits. Both the classical and Q-LSTM have a word embedding size of 64 and vocab size as the output size.

There are three parts in the VQC: (1) the encoding stage where the input vector is encoded; (2) the quantum circuit stage where different entanglement strategies are applied with CNOT, $R_x$, $R_y$, and/or $R_z$ gates and parameters are stored and trained; and (3) the measurement stage where a hermitian operator projects the quantum states onto its eigenvectors [15]. In the entanglement stage, we use a random circuit to encode the rotational vectors to protect the parameters against any 3rd party attacks.
3.2. Decentralized Training

One of the unavoidable problems when using quantum deep learning models is that quantum deep learning model run very slowly on classical computers. Quantum computers are in theory exponentially faster than classical computers, which means they are ideal for training large-scale models, such as language models. If we have a large quantum computer that can be used commercially on a large scale in the future, its architecture will be completely different from the classical computer, which means that the model trained on the large quantum computer is difficult to be directly used by others on the classical computer. Based on the above ideas, we designed a decentralized quantum language model usage process as shown in Figure 1.

Given decentralized quantum language model $F$, we input set of text documents $D_1, D_2, D_3, ..., D_n$ into NIQS servers, the word embedding $E$ will be extracted from quantum language model and using in local downstream task.

$$E = F(D_1, D_2, D_3, ..., D_n)$$  \hspace{1cm} (1)

In Equation (1) the quantum language model $F$ can be any quantum deep learning model that can be used to train the language model. In this paper, we present Q-LSTM as an example to train our language model. The output word embeddings $E$ is a set of vector which can be further processed by other classical or quantum routines.

4. EXPERIMENTS

4.1. Dataset

We use two distinct datasets to train two language models. The first one is a multilingual Twitter dataset\[5\] consisting of 69491 training documents and 998 test documents with 4 different labels(negative, positive, neutral and irrelevant). The second one is the SemEval-2020 Task 9 on Sentiment Analysis of Code-Mixed Tweets \[33\] which consists of 15000 trainins and dev documents and 3,789 test documents. These two datasets are selected because they are linguistically complex in nature, and they have gold-labels for sentiment analysis, which we later use in our evaluation step to train a classification model without introducing further supervision or noise.

4.2. Preprocessing

A coarse filtering of the training datasets is performed before they are fed into the language model. First, empty strings, hash symbols, and urls are removed from the text. All emojis and emoticons are replaced by their English descriptions using the emoji library\[7\]. Sentiment labels are ignored during language model training.

4.3. Q-LSTM LM vs. Classical LSTM LM

In order to fairly assess the performance of the quantum language model, we create a classical language model with the same model architecture and size. With a hidden size of 5 in the classical LSTM, the number of parameters are of the same magnitude as the Q-LSTM with 4 qubits\[26\].

The models are trained until the loss converges, negative log likelihood loss during training and the model perplexity are recorded to evaluate the model. Figure 3a shows the training loss over 15 epochs of the Q-LSTM and its classical counterpart. The quantum language model converges much faster than the classical language model of the same size.

| Model          | LSTM | Q-LSTM (4q) | Q-LSTM (6q) |
|----------------|------|-------------|-------------|
| perplexity     | 1152.78 | 1153.67    | 972.44      |

\[Table 1: Model Perplexity\]

Despite the quantum model converging faster than the classical language model, the model quality as measured by model perplexity for both models are extremely close as shown in Table 1. Given the computing constraints, we only used 4 qubits to train the LSTM model, resulting in an extremely limited number of parameters (around 200) for both models. Although the model perplexity is not ideal compared to conventional PLMs, it still provides a valuable basis for comparing the classical language model and the quantum language model. Another factor that contributes to the high perplexity of our language model is the multilinguality of the training data, which contains richer linguistic information, making it difficult for the small model to learn.

4.4. Model Evaluation - Sentiment Analysis

We use a downstream sentiment analysis task to evaluate the quality of the quantum language model against its classical counterpart. Given sequence of documentation/text $D_1, D_2, ..., D_n$, we need to predict the sentiment class (positive, neutral, negative, or irrelevant). We train a local transformer-based four-way classifier using the pre-trained word embeddings from the Q-LSTM language model. We use 4 transformer blocks and each transformer model has 4 attention heads. In order to avoid introducing more noise to our model pipeline, we use a subset of the training data for the language model. Finally, the sentiment of an unknown document can be predicted as

$$y^i = \text{softmax}(W^i f(x) + b^i)$$  \hspace{1cm} (2)

\[https://pypi.org/project/emoji/\]
Fig. 3: Training Loss for Different Experiments

Where $y_i$ is a predicted sentiment label for a test document, $f(x)$ is a transformer function used to encode and process the input document. Crucially, we report both the accuracy and weighted f1 scores since the dataset is distributed unevenly across the four labels.

| PLM       | LSTM    | Q-LSTM (4q) |
|-----------|---------|-------------|
| accuracy  | 0.928   | 0.934       |
| weighted f1 | 0.93    | 0.93        |

**Table 2:** Sentiment Analysis Performance

Table 2 summarizes the four-way sentiment classification accuracy of the local transformer-based classifier initialized with embeddings trained by the classical LSTM and the quantum LSTM, while keeping everything else the same. The embedding trained by Q-LSTM achieves a slightly higher accuracy than the classical LSTM, and has the same weighted f1 score. This confirms that the quantum language model does not sacrifice performance while preserving privacy. The slight increase in accuracy might indicate that the higher dimensional space in which qubits project to is a “better learner,” meaning that it is slightly better at encoding complex linguistic information compared to its classical counterpart.

### 4.5. Ablation Studies

#### 4.5.1. Effect of Number of Qubits

A single qubit can be described by a two-dimensional Hilbert space $H$. Then, a system of $n$ qubits is the tensor product of $n$ such Hilbert spaces

$$H \otimes H \otimes \cdots \otimes H$$

As we increase the number of qubits in the VQC in the Q-LSTM language model, the resulting embedding space increases exponentially, so does the training time because we model the quantum machine on a classical hardware. Both the 4-Qubit model and 6-Qubit model converge very rapidly as shown in Figure 3c, so we provide a more fine-grained comparison of the training loss for each batch in the first epoch in Figure 3d. From Figure 3d, we can see that the Q-LSTM with 6 qubits converges faster than the one with 4 qubits, and it has a more consistent decreasing pattern. From Table 1 we can see that the 6-Qubit language model also have a much lower perplexity after convergence. Overall, the 6-Qubit language model has lower loss and better perplexity. This indicates that VQCs with more qubits have higher computation capacity, which confirms our hypothesis that models built with larger VQCs will be better at language modeling.

**Table 3:** Model Perplexity for Different Data Sizes

| Dataset          | Multilingual Twitter | Code-Mixing |
|------------------|----------------------|-------------|
| Data size        | 69491                | 15000       |
| Vocab size       | 17000                | 5000        |
| perplexity       | 1153.672             | 368.031     |

### 4.5.2. Effect of Training Data Size

We trained the Q-LSTM language model on a smaller corpus to confirm that the high perplexity on both the classical and the quantum model is due to the small model size trained on a large corpus. To explore the effect of dataset size on Q-LSTM performance, we used another smaller dataset - the SemEvalL-2020 Dataset on Code-Mixed Tweets. As shown in Table 3 and Figure 3b, the smaller dataset has a much lower model perplexity and a lower loss. Despite the difference in the corpus size, there is no obvious difference in the convergence speed of the two models. Furthermore, although the code-mixing corpus is more linguistically complex and thus harder to predict, the linguistic complexity still does not overpower the effect of dataset size on the model loss and perplexity. Therefore, this experiment demonstrates model stability and justifies the high perplexity due to limited model size.

### 5. CONCLUSIONS

In this work, we proposed a decentralized training framework for quantum language models that provides strong privacy protection. The quantum language model training is completed on NISQ servers and the embeddings are loaded to client side local machines for downstream tasks. We trained a small LSTM language model in which the gates are replaced with VQC with random quantum circuits. The quantum language model converges faster, and achieve competitive if not better results compared against its classical counterpart of a similar size. We also studied the effect of the number of qubits in the VQC and the training corpus size on the performance of the model. Our work provides a promising direction and a proof of concept for future NLP research that have privacy protection demands. Some of the future work include extending this decentralized quantum training procedure into a wider range of language model architectures, and potentially training large scale language models such as Quantum-GPT as the required quantum computing hardware becomes available.
6. REFERENCES

[1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeanette Bohg, Antoine Bosselut, Emma Brunskill, et al., “On the opportunities and risks of foundation models,” arXiv preprint arXiv:2108.07258, 2021.

[2] Shaohua Wu, Xudong Zhao, Tong Yu, Rongguo Zhang, Chong Shen, Hongli Liu, Feng Li, Hong Zhu, Jiangan Lao, Liang Xu, et al., “Yuan 1.0: Large-scale pre-trained language model in zero-shot and few-shot learning,” arXiv preprint arXiv:2110.04725, 2021.

[3] Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, and Zhou Yu, “Selective differential privacy for language modeling,” arXiv preprint arXiv:2108.12944, 2021.

[4] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song, “The secret sharer: Evaluating and testing unintended memorization in neural networks,” in 28th USENIX Security Symposium (USENIX Security 19), 2019, pp. 267–284.

[5] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, et al., “Extracting training data from large language models,” in 30th USENIX Security Symposium (USENIX Security 21), 2021, pp. 2633–2650.

[6] Hannah Brown, Katherine Lee, Fatemehsadat Miresghallah, Reza Shokri, and Florian Tramèr, “What does it mean for a language model to preserve privacy?,” arXiv preprint arXiv:2202.05350, 2022.

[7] Munazza Zaib, Quan Z Sheng, and Wei Emma Zhang, “A short survey of pre-trained language models for conversational ai-a new age in nlp,” in Proceedings of the Australasian Computer Science Week Multiconference, 2020, pp. 1–4.

[8] Simon Suster, Stéphane Tulkens, and Walter Daemeleman, “A short review of ethical challenges in clinical natural language processing,” arXiv preprint arXiv:1703.10090, 2017.

[9] Thomas Reitmaier, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson, “Opportunities and challenges of automatic speech recognition systems for low-resource language speakers,” in CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–17.

[10] Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost, “Quantum algorithms for supervised and unsupervised machine learning,” arXiv preprint arXiv:1307.0411, 2013.

[11] Daniel Shaffer, Claudio Chamon, Alioscia Hamma, and Eduardo R Mucciolo, “Irreversibility and entanglement spectrum statistics in quantum circuits,” Journal of Statistical Mechanics: Theory and Experiment, vol. 2014, no. 12, pp. P12007, 2014.

[12] William M Watkins, Samuel Yen-Chi Chen, and Shinjiae Yoo, “Quantum machine learning with differential privacy,” arXiv preprint arXiv:2103.06232, 2021.

[13] Aakarshita Suresh, Abdullah Ash Saki, Mahabubul Alam, Dr Ghosh, et al., “A quantum circuit obfuscation methodology for security and privacy,” arXiv preprint arXiv:2104.05943, 2021.

[14] Vivek Balachandran, “Quantum obfuscation: Quantum predicates with entangled qubits,” in Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy, 2021, pp. 293–295.

[15] Israel Griol-Barres, Sergio Millà, Antonio Cebrián, Yashar Mansoori, and José Millet, “Variational quantum circuits for machine learning: an application for the detection of weak signals,” Applied Sciences, vol. 11, no. 14, pp. 6427, 2021.

[16] Leo Zhou, Sheng-Tao Wang, Soonwon Choi, Hannes Pichler, and Mikhail D Lukin, “Quantum approximate optimization algorithm: Performance, mechanism, and implementation on near-term devices,” Physical Review X, vol. 10, no. 2, pp. 021067, 2020.

[17] John Preskill, “Quantum computing in the nisq era and beyond,” Quantum, vol. 2, pp. 79, 2018.

[18] M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles, “Variational quantum algorithms,” Nature Reviews Physics, vol. 3, no. 9, pp. 625–644, aug 2021.

[19] Vojtěch Havlíček, Antonio D Córcoles, Kristian Temme, Aram W Harrow, Ahinav Kendale, Jerry M Chow, and Jay M Gambetta, “Super-vised learning with quantum-enhanced feature spaces,” Nature, vol. 567, no. 7747, pp. 209–212, 2019.

[20] Swaroop Ramaswamy, Om Thakkar, Rajiv Mathews, Galen Andrew, H Brendan McMahan, and Frangoise Beaufays, “Training production language models without memorizing user data,” arXiv preprint arXiv:2009.10031, 2020.

[21] Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang, “Learning differentially private recurrent language models,” in International Conference on Learning Representations (ICLR), 2018.

[22] Priyanka Mary Mammen, “Federated learning: opportunities and challenges,” arXiv preprint arXiv:2101.05428, 2021.

[23] Tian Li, Anit Kumar Sahu, Ameeta Talwalkar, and Virginia Smith, “Feder-ated learning: Challenges, methods, and future directions,” IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50–60, 2020.

[24] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth, “Practical secure aggregation for privacy-preserving machine learning,” in proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017, pp. 1175–1191.

[25] Samuel Yen-Chi Chen and Shinjiae Yoo, “Federated quantum machine learning,” Entropy, vol. 23, no. 4, pp. 460, 2021.

[26] Samuel Yen-Chi Chen, Shinjiae 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.

[27] Ivano Basile and Fabio Tamburini, “Towards quantum language models,” in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1840–1849.

[28] Chao-Han Huck Yang, Jun Qi, Samuel Yen-Chi Chen, Pin-Yu Chen, Sabato Marco Siniscalchi, Xiaoli Ma, and Chin-Hui Lee, “Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition,” in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 6523–6527.

[29] Jun Qi, “Federated quantum natural gradient descent for quantum fed-erated learning,” arXiv preprint arXiv:2209.00564, 2022.

[30] Chao-Han Huck Yang, Jun Qi, Samuel Yen-Chi Chen, Yu Tsao, and Pin-Yu Chen, “When bert meets quantum temporal convolution learning for text classification in heterogeneous computing,” in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 8602–8606.

[31] Samuel Yen-Chi Chen, Chao-Han Huck Yang, Jun Qi, Pin-Yu Chen, XiaoMai, and Hsi-Sheng Gao, “Variational quantum circuits for deep reinforcement learning,” IEEE Access, vol. 8, pp. 141007–141024, 2020.

[32] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals, “Recurrent neural network regularization,” arXiv preprint arXiv:1409.2329, 2014.

[33] Parth Patwa, Gustavo Aguilar, Sadipta Kar, Saraj Pandey, Srinivas Pykyl, Björn Gumbäck, Tannoy Chakraborty, Thamar Solorio, and Amitava Das, “SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets,” in Proceedings of the fourteenth workshop on semantic evaluation, 2020, pp. 774–790.