WHEN BERT MEETS QUANTUM TEMPORAL CONVOLUTION LEARNING FOR TEXT CLASSIFICATION IN HETEROGENEOUS COMPUTING

Chao-Han Huck Yang[1] *Jun Qi*[1] *Samuel Yen-Chi Chen*[2]
Yu Tsao[3] *Pin-Yu Chen*[4]

[1] Georgia Institute of Technology, GA, USA
[2]Brookhaven National Laboratory, NY, USA and [3]Academia Sinica, Taipei, Taiwan
[4]IBM Research, Yorktown Heights, NY, USA

ABSTRACT

The rapid development of quantum computing has demonstrated many unique characteristics of quantum advantages, such as richer feature representation and more secured protection on model parameters. This work proposes a vertical federated learning architecture based on variational quantum circuits to demonstrate the competitive performance of a quantum-enhanced pre-trained BERT model for text classification. In particular, our proposed hybrid classical-quantum model consists of a novel random quantum temporal convolution (QTC) learning framework replacing some layers in the BERT-based decoder. Our experiments on intent classification show that our proposed BERT-QTC model attains competitive experimental results in the Snips and ATIS spoken language datasets. Particularly, the BERT-QTC boosts the performance of the existing quantum circuit-based language model in two text classification datasets by 1.57% and 1.52% relative improvements. Furthermore, BERT-QTC can be feasibly deployed on both existing commercial-accessible quantum computation hardware and CPU-based interface for ensuring data isolation.

***Index Terms*—** Quantum machine learning, temporal convolution, text classification, spoken language understanding, and heterogeneous computing

1. INTRODUCTION

Text classification (e.g., intent detection) from human spoken utterances [1, 2] is an essential element of a spoken language understanding [3] (SLU) system. Learning intent information often involves various applications, such as on-device voice assistants [4] and airplane travel information systems [5]. Recently, Bidirectional Encoder Representations from Transformers [6] (BERT) has caused a stir in the lexical modeling community by providing competitive results in intent classification as a common SLU application that we torched upon in this work. However, recent works [7, 8] have raised new concerns about data leakage from large-scale language models such as BERT, which can involve sensitive information like personal identification. New regulation requirements [9] (e.g., GDPR) on data protection, privacy [10], and security can further motivate advanced investigations in designing distributed algorithms on heterogeneous computing devices.

In recent two years, cloud-accessible quantum devices (e.g., IBM-Q) have shown unique characteristics and empirical advantages in many applications [11, 12], such as model compression, parameter isolation, and encryption. A noisy-intermediate-scale-quantum (NISQ) device is a main hardware category that empowers the quantum advantages by using only a few number of qubits (5 to 100). When the limitation of quantum resource is concerned, variational quantum circuit (VQC) has been studied on the design of quantum machine learning like quantum support vector machine, where the VQC algorithm requires no quantum error correction as a matching working on NISQ devices. Projecting classical data into high-dimensional quantum feature space has been proven its quantum advantages [13, 14] (e.g., a quadratic speed-up [13] and better presentation power [11]) on some classifications tasks. For example, Yang *et al.* [11] demonstrates the quantum advantages on a speech command recognition system by applying the VQC algorithm to extract acoustic features for representation learning with a vertical federated learning architecture. The proposed vertical federated learning could benefit from quantum advantages[1], such as parameter isolation with random circuit learning, for the acoustic modeling task. However, designing an end-to-end neural model on SLU tasks is still an open problem with a potential impact on heterogeneous computing devices.

Motivated by the very recent success [11] in the quantum circuit-based acoustic modeling, we put forth a new gradient-based training to include pre-trained BERT models and enhance its data protection by leveraging upon VQC learning in this work. To strengthen the parameter isolation on BERT, our work proposes a novel hybrid classical-quantum system with quantum temporal convolution (QTC) learning, which is composed of a pre-trained BERT model (on a classical computer) and a new type of one-dimensional random circuit [16, 11] (on a quantum device). In comparison with the benchmark pre-trained BERT model, our proposed architecture can maintain competitive experimental performance considering the SLU benchmark solution running in homogeneous machines. Notably, in this work, compared with the classical deep neural network (DNN) models, the use of QTC lowers the model complexity. The VQC design of QTC is to use the quantum circuit as a temporal convolutional features projector on the text embeddings that requires only a few qubits (e.g., 4 to 9) to set up our quantum platform.

As shown in Figure 1, the design of such a BERT-QTC model is in a regime of hybrid classical-quantum architecture, where the word embedding can be offline attained by utilizing DNN models on classical computers before going through quantum devices. Table 1 shows an overview of classical, quantum [17], and heterogeneous approaches for SLU. To the best of our knowledge, this is the **first** work to investigate a BERT-based hybrid classical-quantum model for text classification (e.g., intent detection) task with competitive

[1]Quantum advantage means that a programmable quantum device [15] can solve a specific problem that is intractable on classical computers.

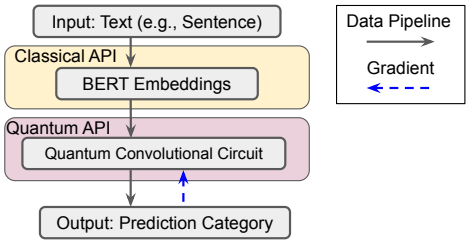

Fig. 1: BERT-QTC for vertical federated [11] text classification.

empirical results on quantum circuit learning on the NISQ device.

Table 1: An overview of different natural language processing (NLP) approaches: neural language model (NLM), logic programming (LP), and its quantum variants. Our work belongs to hybrid CQ to access quantum advantages (QA).

Approach	Input	Model	Output	Challenges
Classical	bits	NLM & LP	bits	data protection
Quantum	qubits	Quantum LP [18]	qubits	hardware limits
hybrid CQ	bits	NLM + VQC	bits	model design

2. RELATED WORK

2.1. Quantum Algorithms for Language Processing

Recent years have witnessed a rapid development [19] of quantum machine learning on near-term quantum devices (5 to 200 qubits). Encoding classical data into high-dimensional quantum features has proven rigorous quantum advantages on classification tasks with recent theoretical justification [13, 14]. In particular, VQC learning [20] is utilized as parametric models to build a QNN model. The research of VQC is of significance in itself because they constitute the setting of quantum computational experiments on NISQ devices. Although business-accessible quantum services (e.g., IBM-Q and Amazon Braket) are still actively being developed, several works have attempted to incorporate quantum algorithms for language processing tasks. For example, [17] Liu *et al.* consider a novel classifier as the physical evolution process and described the process with quantum mechanical equations. In [21, 22], the theory of quantum probability is proposed to build a general language model for information retrieval.

Moreover, Blacoe *et al.* [23] leverage upon quantum superposition and entanglement to investigate the potential of quantum theory as a practical framework to capture the lexical meaning and model semantic processes such as word similarity and association. However, those works aim at the use of quantum concepts to build quantum language models in theory, and they are different from the recent quantum circuit architectures featured with small-to-none quantum error correction in the NISQ devices.

Recently, Meichanetzidis *et al.*[24] first investigated computational logic relationships for encoding graphs into circuit learning on NISQ computers. Coecke *et al.* [18] investigate theoretical foundation and advantages of encryption on using the VQC learning on linguistic modeling. In particular, their quantum circuits are employed as a pipeline to map semantic and grammar diagrams and pave the way to near-term applications of quantum computation devices. Lorenz *et al.* [25] introduced a concept of compositional semantic information for text classification tasks on the quantum hardware by simply using a mapping from lexical segments to a quantum circuit. Nevertheless, beyond grammar modeling, how to design a specific quantum model is still a wide-open topic with potential impacts on "large-scale" data (e.g., spoken language processing and applications).

2.2. BERT and Heterogeneous Computing

Devlin *et al.* [26] firstly put forth the architecture of BERT, which mainly applies the masking and bidirectional training of Transformer equipped with an attention model to language modeling. BERTs [27] can be used for a wide variety of language tasks since different SLU tasks only require a few more layers to the core model for fine-tuning. When the training time and data protection become significant concerns for BERTs, heterogeneous computing architectures, such as distributed training and federated learning [28], provide a new perspective to ensure data isolation as protection and improve training efficiency.

In our previous work [11], we propose a new vertical federated learning architecture that combines new convolution-encoded randomized VQC and recurrent neural network (RNN) (denoted as QCNN-RNN) for acoustic modeling preserving quantum advantages. The QCNN-RNN system first uses a quantum convolution embedding layer to encode features in high-dimensional Hilbert space, then encode quantum features in a classical format for the representation learning with RNN. Different with the quantum acoustic modeling task, we want to advance a new VQC design working on the SLU task considering the benchmark pre-trained language model (e.g., BERT) in this work.

3. METHOD

3.1. BERT for Text Embedding

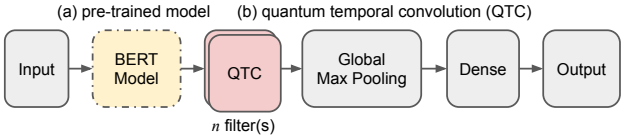

Fig. 2: The BERT-QCC comprises with (a) BERT and (b) quantum temporal convolutional (QTC) learning with random circuits.

We first use BERT as a pre-trained language model to extract hidden representation features from text sequences as shown in Figure 2 (a). The BERT based language models consist of a multi-layer bidirectional Transformer-based encoder, where the input corresponds to a combination of WordPiece embeddings, positional embeddings, and the segment embedding. Symbolically, given an input token sequence $\boldsymbol{x} = (x_1, \ldots, x_T)$ the BERT model outputs are computed by:

$$\mathbf{H} = \mathbf{BERT}\left(x_1, \ldots, x_T\right), \tag{1}$$

where $\mathbf{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_T)$ and $\boldsymbol{h}_i$ is contextual semantic representation embedding of each token in Eq. (1).

Neural Network Intent Decoder. We select a hidden state of the first special token for intent classification after a dropout layer to avoid over-fitting. Conventionally, the hidden state is encoded by a neural network function ($\boldsymbol{f}(\mathbf{H}; \theta_{\boldsymbol{f}})$) for extracting sentence-level semantic representation. Finally, an unknown intent can be predicted as:

$$y^i = \text{softmax}\left(\mathbf{W}^i \boldsymbol{f}(\mathbf{H}; \theta_{\boldsymbol{f}}) + \boldsymbol{b}^i\right), \tag{2}$$

where y^i is a predicted intent label for a test sequence. The model is end-to-end trained by minimizing the cross-entropy loss between predicted text labels and their corresponding ground-truth labels.

3.2. Quantum Temporal Convolution with Random Circuit

The expressive power of quantum convolution has been recently studied in computer vision [16] and speech processing [11, 29] by using quantum space encoding to project classical data into rich representations of features for classification. Deploying a quantum circuit with randomized assigned (non-trainable) parameters could beat both trainable and non-trainable convolutional filters in the previous study [11]. We further consider a new hybrid classical-quantum decoder design, which replace the perception network ($\boldsymbol{f}$) in Eq. (2) with a randomized quantum circuit decoder as shown in Fig. 2 (b). We use a max length of 50 for pre-trained BERT models.

Variational Quantum Circuit Decoder. We adopt the concept from variational quantum circuit learning and define a latent encoding function $\hat{\boldsymbol{f}}$, where it contains angle encoding, parameterized rotation, and quantum-to-classical state decoding. We use the VQC as shown in Fig. 3, made by strong entanglement [30] circuit. First, the Bert embeddings $[\mathbf{h}_1, \cdots, \mathbf{h}_T]$ are feeding into a sliding window for the temporal quantum convolutional filter (e.g., coordinated with time or order of input), where we take window size $I = 4$ for example as one desired qubits size shown in Fig. 4. We then extract latent embedding $[h_1, h_2, h_3, h_4]$ as inputs to the VQC layer.

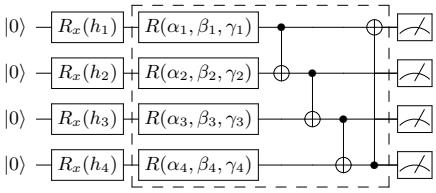

Fig. 3: **Deployed quantum circuit.** The VQC component contains three major parts: encoding, learning and quantum measurement. Here we use the angle encoding scheme to encode the input values $h_1 \cdots h_4$ (latent embedding; taking four qubits, for example) by treating them as rotation angles along x-axis. The learning part uses general unitary rotation R. There are three parameters α, β and γ in each R. The controlled-NOT (CNOT) gates are used to entangle quantum states from each qubit. The final quantum measurement part will output the Pauli-Z expectation values of each qubit.

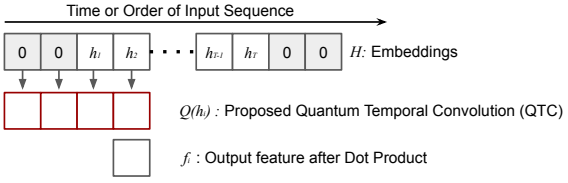

Fig. 4: Computing framework for quantum temporal convolution (QTC) with random variational circuit from input text embeddings.

The latent features (H) are the input of a quantum circuit layer, $\mathbf{Q}$, that learns and encodes latent sequence:

$$\hat{\boldsymbol{f}} = \mathbf{Q}(H; \mathbf{e}, \theta_{\mathbf{q}}, \mathbf{d}) \tag{3}$$

In Eq. (3), the approximation process of a quantum circuit block, $\mathbf{Q}$, depends on the encoding initialization $\mathbf{e}$, the quantum circuit parameters, $\theta_{\mathbf{q}}$, and the decoding measurement $\mathbf{d}$. When we stack more VQC blocks, we expect that a gradient vanishing issue would occur and degrade the representation learned from a hybrid QNN model whose parameters are trained by joint DNN-VQC gradient updates.

Table 2: Mathematical notation for VQC learning.

Symbol	Meaning
$\|0\rangle$	quantum state 0
R_x	rotation gate along x-axis
$\theta_{\mathbf{q}}$=$\{\alpha, \beta, \gamma\}$	random parameters
CNOT	controlled-NOT gate

In this VQC, the elements $h_1 \cdots h_4$ of the input vectors are used as rotational angles for single-qubit quantum rotation gate R_x on each qubit to encode the classical values. After encoded into a quantum state, the state is processed through a *learnable* or *optimizable* layer (in grouped box). The parameters labeled with α, β, and γ are for the optimization. The CNOT gates are applied to entangle the qubits. In the final step, the quantum measurement procedure is carried out. In this part, we retrieve the Pauli-Z expectation values for each qubit. The obtained result is a 4-dimensional vector which can be further processed by other classical or quantum routines.

Figure 3 summarize characterises of the variational quantum circuit, where there are 4 quantum channels and each channel is mutually entangled with each other by applying the CNOT gate. Given the unitary matrix $\mathbb{U}$ representing the multiplication of all quantum gates and taking B as observable matrix, we attain that:

$$\hat{\boldsymbol{f}}(h; \boldsymbol{\theta}_{\text{random init}}) = \langle h|U^{\dagger}(\boldsymbol{\theta}_i)BU(\boldsymbol{\theta}_i)|h\rangle. \tag{4}$$

The hyper-parameters ($\boldsymbol{\theta}_{\text{random init}}$) for rotation gates in Eq. (4) are randomly initialized and not considered to be updated during the training phase, which is aimed to provide parameters protection in heterogeneous computing architectures (or simulation API) against model inversion attacks [31] and parameters leakages [32]. As shown in Fig. 2, the temporal quantum convolution could be applied with n filters to map features before computing with a global max-pooling layer. As **the first attempt** to construct quantum temporal convolution, we select a filters number from $\{1, 2, 3, 4\}$ under the 9 qubits requirement for common commercial quantum hardware (e.g., IBMQ).

To evaluate the effectiveness of proposed QTC architectures, we select three additional "random" encoder baselines (similar to QTC) with text embeddings for text classification: (1) BERT with a random temporal convolutional network (TCN); (2) word2vec [33] with random TCN; (3) word2vec with random QTC, all followed the same filter numbers and global max-pooling after the deployed random encoders.

We will study how random QTC learning benefits BERT-based heterogeneous architecture in the experimental section. From a system-level perspective, the proposed QTC learning reduces the risk of parameter leakage [32, 34, 35] from **inference-time** attackers, and it tackles data isolation issues with its architecture-wise advantages [36] on encryption [37] and the feature of without accessing the data directly [38, 11].

4. EXPERIMENT

4.1. Quantum Computing Hardware

We use PennyLane as an open-source and reproducible environment, which uses differential programming to calculate gradient to update gradient circuits. A hardware back-end of PennyLane could integrate from CPU-simulation, QPU (supported by Amazon Braket), and TPU (supported by Google Floq). Since fine-tuning time of BERTs is often extensive, we first use a CPU-simulated VQC environment to train the proposed BERT-QTC model. Then, we evaluate our hybrid classical-quantum models on Amazon Braket and Floq hardware and report test accuracy by 10-fold cross-validation. Referring to the established work [11] on VQC learning, we consider a NISQ device with 4 to 16 qubits, which allows us to perform up to 9 class predictions. Furthermore, we refine our dataset according to the hardware settings, and clarify that NISQ's size constraints limit current applications of proposed BERT-QTC working toward word-level slot-filling tasks (with 120 to 72 classes). We believe that the development of commercial-accessible NISQ could gradually resolve this challenge. For vertical learning setting, we use secure multiparty computation (MPC) protocol [39] between local instance and quantum device during the virtualization.

4.2. Dataset and Setup

Snips Dataset: To evaluate the proposed framework in complex spoken language corpora, we select Snips [4] dataset. Snips dataset is a collection of spoken utterances from a personal voice assistant, whose domains cover speech command, music searching, and smart home request. The training set includes 13,084 utterances, and the test set includes 700 utterances. We use another 700 utterances as the development set. There are 7 types of intent classes, where the number of samples for each intent is nearly the same.

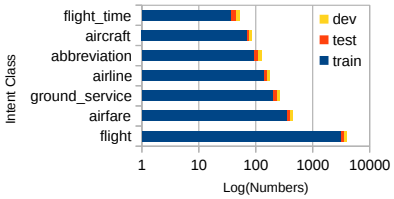

Fig. 5: ATIS_7 Dataset for BERT-QTC experiments.

ATIS Dataset: The ATIS (Airline Travel Information System) dataset [5] is widely used in NLU research containing spoken corpus, including intent classification and slot filling labels for flights reservation. Since the language scripts from ATIS are collected from human speakers, ATIS is selected to highlight the data sensitivity under GDPR policy [9]. To compare with Snips as a slightly unbalanced classification study, we further select a subset of ATIS containing top-7 classes (denoted as ATIS_7) of original ATIS. As shown in Fig. 5, the deployed subset include 90.93% of the original data, where training, development, and testing sets contain 4072, 545, 556 utterances, respectively. As a remark, we have conducted text classification experiments with a full ATIS dataset, but the results have a large variance ($\pm$ 7.32%) on both random TCN and QRC encoders due to unbalanced intent labels in a long-tail distribution [40].

BERT Model: We use an open source BERT [26] model to connect the QTC encoder. The pre-trained models of BERT are frozen in the heterogeneous computing process as a universal encoder to take sentences into embedding. We use the official pretrained BERT (large) model from TensorFlow Hub with 1024-hidden dimensions, 24 layers, and 12 attentions heads with 330M pre-trained parameters.

4.3. BERT-QTC Performance

Since we are working on sentence-level text classification (intent as labels) on the two deployed datasets, we first select two different text embeddings methods from pre-trained word2vec and BERT embeddings to compared the proposed random QTC encoder with a random TCN encoder. Furthermore, we aim to study different hyper-parameters setup for TCN and QTC architecture. As a remark, both TCN and QTC could be simulated with CPU-based environments or API, where QTC is featured with better representation mapping with theoretical justifications and an option running with quantum hardware to preserve full quantum advantages.

As shown in Tab. 3 (Snips) and 4 (ATIS_7), the BERT-QTC model in vertical federated architecture performs the best average prediction accuracy performance, which attains $\mathbf{96.62\%}$ in Snips and $\mathbf{96.98\%}$ in ATIS_7 compared with TCN based architectures and word2vec with QTC encoder. Moreover, we investigate the convolution filter and kernel setups on BERTs-QTC models from the second, fourth row in Tab. 3 and 4. BERT-QTC demonstrates more significant improvement with its word2vec-QTC baseline when two filters have been used. Interestingly, we find out that utilize QTC encoder shows a general performance-boosting in the two deployed SLU datasets. Proposed BERT-QTC federated models perform relative improvements of **+1.57%** in Snips and **+1.52%** in ATIS_7 compared with other heterogeneous computing ablations. The statistical variances are $\leq 0.52\%$ in our experiments.

Table 3: Average accuracy on intent classification for Snips with a set of different number (n) of convolutional filter and kernel size (k).

Embedding	word2vec				BERT			
(n,k)	(1,4)	(2,2)	(2,3)	(2,4)	(1,4)	(2,2)	(2,3)	(2,4)
TCN	82.02	83.37	82.90	83.15	95.48	95.23	95.12	95.27
QTC	**83.32**	**83.94**	**83.61**	**84.64**	**96.41**	**96.42**	**96.62**	**96.44**

Table 4: Average accuracy on intent classification for ATIS_7 with a set of different number (n) of convolutional filter and kernel size (k).

Embedding	word2vec				BERT			
(n,k)	(1,4)	(2,2)	(2,3)	(2,4)	(1,4)	(2,2)	(2,3)	(2,4)
TCN	80.09	80.22	80.91	82.34	95.18	95.03	94.95	95.23
QTC	**81.42**	**82.49**	**83.82**	**83.95**	**96.69**	**96.92**	**96.32**	**96.98**

5. CONCLUSION

In this paper, we propose a novel hybrid classical-quantum architecture to strengthen the BERT model with a quantum circuit decoder via **quantum temporal convolution** (QTC) with random circuit learning. The proposed BERT-QTC models show competitive results for text classification as one prompted finding for standard SLU tasks. Moreover, our VQC encoders are capable of deploying on both existing quantum hardware, and the simulator requires only a small amount of qubits (4 to 8 qubits). The proposed QTC can enhance the data protection on top of BERT models in the vertical federated learning setting.

6. REFERENCES

[1] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[2] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[3] Y.-Y. Wang, L. Deng, and A. Acero, "Spoken language understanding," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 16–31, 2005.

[4] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril *et al.*, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *arXiv preprint arXiv:1805.10190*, 2018.

[5] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, "The atis spoken language systems pilot corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[7] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," in *28th {USENIX} Security Symposium ({USENIX} Security 19)*, 2019, pp. 267–284.

[8] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson *et al.*, "Extracting training data from large language models," *arXiv preprint arXiv:2012.07805*, 2020.

[9] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[10] C.-H. H. Yang, S. M. Siniscalchi, and C.-H. Lee, "Pate-aae: Incorporating adversarial autoencoder into private aggregation of teacher ensembles for spoken command classification," *Proc. Interspeech 2021*, pp. 881–885, 2021.

[11] C.-H. H. Yang, J. Qi, S. Y.-C. Chen, P.-Y. Chen, S. M. Siniscalchi, X. Ma, and C.-H. Lee, "Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6523–6527.

[12] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell *et al.*, "Quantum supremacy using a programmable superconducting processor," *Nature*, vol. 574, no. 7779, pp. 505–510, 2019.

[13] Y. Liu, S. Arunachalam, and K. Temme, "A rigorous and robust quantum speed-up in supervised machine learning," *Nature Physics*, pp. 1–5, 2021.

[14] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, "Power of data in quantum machine learning," *Nature communications*, vol. 12, no. 1, pp. 1–9, 2021.

[15] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M. J. Bremner, J. M. Martinis, and H. Neven, "Characterizing quantum supremacy in near-term devices," *Nature Physics*, vol. 14, no. 6, pp. 595–600, 2018.

[16] M. Henderson, S. Shakya, S. Pradhan, and T. Cook, "Quanvolutional neural networks: powering image recognition with quantum circuits," *Quantum Machine Intelligence*, vol. 2, no. 1, pp. 1–9, 2020.

[17] D. Liu, X. Yang, and M. Jiang, "A novel classifier based on quantum computation," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2013, pp. 484–488.

[18] B. Coecke, G. de Felice, K. Meichanetzidis, and A. Toumi, "Foundations for near-term quantum natural language processing," *arXiv preprint arXiv:2012.03755*, 2020.

[19] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, "Supervised learning with quantum-enhanced feature spaces," *Nature*, vol. 567, no. 7747, pp. 209–212, 2019.

[20] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, "Quantum circuit learning," *Physical Review A*, vol. 98, no. 3, p. 032309, 2018.

[21] A. Sordoni, J.-Y. Nie, and Y. Bengio, "Modeling term dependencies with quantum language models for ir," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 2013, pp. 653–662.

[22] I. Basile and F. Tamburini, "Towards quantum language models," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1840–1849.

[23] W. Blacoe, E. Kashefi, and M. Lapata, "A quantum-theoretic approach to distributional semantics," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 847–857.

[24] K. Meichanetzidis, S. Gogioso, G. De Felice, N. Chiappori, A. Toumi, and B. Coecke, "Quantum natural language processing on near-term quantum computers," *arXiv preprint arXiv:2005.04147*, 2020.

[25] R. Lorenz, A. Pearson, K. Meichanetzidis, D. Kartsaklis, and B. Coecke, "Qnlp in practice: Running compositional models of meaning on a quantum computer," *arXiv preprint arXiv:2102.12846*, 2021.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[27] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *International Conference on Learning Representations*, 2019.

[28] S. Y.-C. Chen and S. Yoo, "Federated quantum machine learning," *arXiv preprint arXiv:2103.12010*, 2021.

[29] J. Qi, C.-H. H. Yang, and P.-Y. Chen, "Qtn-vqc: An end-to-end learning framework for quantum neural networks," *arXiv preprint arXiv:2110.03861*, 2021.

[30] S. Y.-C. Chen, C.-H. H. Yang, J. Qi, P.-Y. Chen, X. Ma, and H.-S. Goan, "Variational quantum circuits for deep reinforcement learning," *IEEE Access*, vol. 8, pp. 141 007–141 024, 2020.

[31] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1322–1333.

[32] A. Duc, S. Dziembowski, and S. Faust, "Unifying leakage models: From probing attacks to noisy leakage." in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2014, pp. 423–440.

[33] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[34] D. Leroy, A. Coucke, T. Lavril, T. Gisselbrecht, and J. Dureau, "Federated learning for keyword spotting," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6341–6345.

[35] M. Chen, A. T. Suresh, R. Mathews, A. Wong, C. Allauzen, F. Beaufays, and M. Riley, "Federated learning of n-gram language models," *arXiv preprint arXiv:1910.03432*, 2019.

[36] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "The reusable holdout: Preserving validity in adaptive data analysis," *Science*, vol. 349, no. 6248, pp. 636–638, 2015.

[37] A. C.-C. Yao, "Quantum circuit complexity," in *Proceedings of 1993 IEEE 34th Annual Foundations of Computer Science*. IEEE, 1993, pp. 352–361.

[38] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[39] R. Cramer, I. B. Damgård *et al.*, *Secure multiparty computation*. Cambridge University Press, 2015.

[40] C.-H. H. Yang, L. Liu, A. Gandhe, Y. Gu, A. Raju, D. Filimonov, and I. Bulyko, "Multi-task language modeling for improving speech recognition of rare words," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 1087–1093.