DECENTRALIZING FEATURE EXTRACTION WITH QUANTUM CONVOLUTIONAL NEURAL NETWORK FOR AUTOMATIC SPEECH RECOGNITION

Chao-Han Huck Yang[1] *Jun Qi*[1] *Samuel Yen-Chi Chen*[2] *Pin-Yu Chen*[3]
Sabato Marco Siniscalchi[1,4,5] *Xiaoli Ma*[1] *Chin-Hui Lee*[1]

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, USA
[2]Brookhaven National Laboratory, NY, USA and [3]IBM Research, Yorktown Heights, NY, USA
[4]Faculty of Computer and Telecommunication Engineering, University of Enna, Italy
[5]Department of Electronic Systems, NTNU, Trondheim, Norway

ABSTRACT

We propose a novel decentralized feature extraction approach in federated learning to address privacy-preservation issues for speech recognition. It is built upon a quantum convolutional neural network (QCNN) composed of a quantum circuit encoder for feature extraction, and a recurrent neural network (RNN) based end-to-end acoustic model (AM). To enhance model parameter protection in a decentralized architecture, an input speech is first up-streamed to a quantum computing server to extract Mel-spectrogram, and the corresponding convolutional features are encoded using a quantum circuit algorithm with random parameters. The encoded features are then down-streamed to the local RNN model for the final recognition. The proposed decentralized framework takes advantage of the quantum learning progress to secure models and to avoid privacy leakage attacks. Testing on the Google Speech Commands Dataset, the proposed QCNN encoder attains a competitive accuracy of 95.12% in a decentralized model, which is better than the previous architectures using centralized RNN models with convolutional features. We conduct an in-depth study of different quantum circuit encoder architectures to provide insights into designing QCNN-based feature extractors. Neural saliency analyses demonstrate a high correlation between the proposed QCNN features, class activation maps, and the input Mel-spectrogram. We provide an implementation[1] for future studies.

Index Terms— Acoustic Modeling, Quantum Machine Learning, Automatic Speech Recognition, and Federated Learning.

1. INTRODUCTION

With the increasing concern about acoustic data privacy issues [1], it is essential to design new automatic speech recognition (ASR) architectures satisfying the requirements of new privacy-preservation regulations, e.g., GDPR [2]. Vertical federated learning (VFL) [3] is one potential strategy for data protection by decentralizing an end-to-end deep learning [4] framework and separating feature extraction from the ASR inference engine. With recent advances in commercial quantum technology [5], quantum machine learning (QML) [6] becomes an ideal building block for VFL owing to its advantages on parameter encryption and isolation. To do so, the input to QML often represented by classical bits, needs to be first encoded into quantum states based on *qubits*. Next, approximation algorithms (e.g., quantum branching programs [7]) are applied to quantum devices based on a quantum circuit [8] with noise tolerance. To implement our proposed approach, we utilize a state-of-the-art noisy intermediate-scale quantum (NISQ) [9] platform (5 to 50 qubits) for academic and commercial applications [10]. It can be set up on accessible quantum servers from cloud-based computing providers [5].

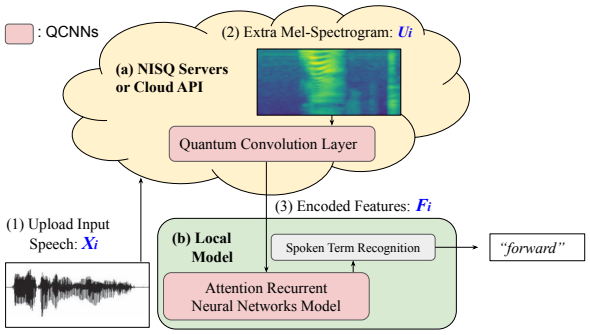


Fig. 1: Proposed quantum machine learning for acoustic modeling (QML-AM) architecture in a vertical federated learning progress including (a) a quantum convolution layer on Noisy Intermediate-Scale Quantum (NISQ) servers or cloud API; and (b) a local model (e.g., second-pass model [11, 12]) for speech recognition tasks.

As shown in Fig. 1, we propose a decentralized acoustic modeling (AM) scheme to design a quantum convolutional neural network (QCNN) [13] by combining a variational quantum circuit (VQC) learning paradigm [6] and a deep neural network [14] (DNN). VQC refers to a quantum algorithm with a flexible designing accessibility, which is resistant to noise [6, 8] and adapted to NISQ hardware with light or no requirements for quantum error correction. Based on the advantages of VQC under VFL, a quantum-enhanced data processing scheme can be realized with fewer entangled encoded qubits [15, 7] to assure model parameters protection and lower computational complexity. As shown in Table 1, to the best of the authors' knowledge, this is the **first** work to combine quantum circuits and DNNs and build a new QCNN [13] for ASR. To provide secure data pipeline and reliable quantum computing, we introduce the VFL architecture for decentralized ASR tasks, where remote NISQ cloud servers are used to generate quantum-based features, and ASR decoding is performed with a local model [12]. We refer to our decentralized quantum-based ASR system to as QCNN-ASR. Evaluated on the Google Speech Commands dataset with machine noises in-

[1]https://github.com/huckiyang/QuantumSpeech-QCNN

Table 1: An overview of machine learning approaches and related key properties. CQ stands for a hybrid classical-quantum (CQ) [15] model using in this paper. QA stands for quantum advantages [8], which are related to computational memory and parameter protection. VQC indicates the variational quantum circuit. VFL means vertical federated leaning [3]. DNN stands for deep neural network [4]

Approach	Input	Learning Model	Output	Properties
Classical	bits	DNN and more.	bits	Easy implementation
Quantum	qubits	VQC and more.	qubits	QA but limited resources
hybrid CQ	bits	VQC + DNN	bits	Accessible QA over VFL

curred from quantum computers, the proposed QCNN-ASR framework attains a competitive 95.12% accuracy on word recognition.

2. RELATED WORK

2.1. Quantum Machine Learning for Signal Processing

QML [6] has been shown advantages in terms of lower memory storage, secured model parameters encryption, and good feature representation capabilities [8]. There are several variants (e.g., adiabatic quantum computation [9], and quantum circuit learning [16]). In this work, we use the hybrid classical-quantum algorithm [13], where the input signals are given in a purely classical format (aka, numerical format, e.g., digital image), and a quantum algorithm is employed in the feature learning phase. Quantum circuit learning is regarded as the most accessible and reproducible QML for signal processing [15], such as supervised learning in the design of quantum support vector machine [8]. Indeed, it has been widely used, and it consists only of quantum logic gates with a possibility of deferring an error correction [6, 16].

2.2. Deep Learning with Variational Quantum Circuit

In the NISQ era [10], quantum computing devices are not error-corrected, and they are therefore not fault-tolerant. Such a constraint limits the potential applications on NISQ technology, especially for large quantum circuit depth, and a large number of qubits. However, Mitarai *et al.*'s seminal work [6] describes a framework to build machine learning models on NISQ. The key idea is to employ VQC [17], which are subject to an iterative optimization processes, so that the effects of noise in the NISQ devices can potentially be absorbed into these learned circuit parameters. Recent litterature reports about several successful machine learning applications based on VQC, for instance, deep reinforcement learning [18], and function approximation [6]. VQCs are also used in constructing quantum machine learning models capable of handling sequential patterns, such as the dynamics of of certain physical systems [19]. It should be noted that the input dimension of the input in [19] is rather limited [18] because of stringent requirements of currently available quantum simulators, or real quantum devices.

2.3. Quantum Learning and Decentralized Speech Processing

Although quantum technology is quite new, there have been some attempts in exploiting it for speech processing. For example, Li *et al.* [20] proposed a speech recognition system with quantum back-propagation (QBP) simulated by fuzzy logic computing. However, QBP is not using the qubit directly in a real-world quantum device, and the approaches hardly demonstrates the quantum advantages inherent in this computing scheme. Moreover, the QBP solution can be complicated to large-scale ASR tasks with parameters protection.

From a system perspective, these accessible quantum advantages from VQL, including encryption and randomized encoding, are prominent requirements for federated learning systems, such as distributed ASR. Cloud computing-based federated architectures [3] have been proven the most effective solutions for industrial applications, demonstrating quantum advantages using commercial NISQ servers [5]. More recent works on federated keyword spotting [1], distributed ASR [21], improved lite audio-visual processing for local inference [22], and federated n-gram language [11] marked the the importance of privacy-preserving learning under the requirement of acoustic and language data protection.

3. DESIGNING QUANTUM CONVOLUTIONAL NEURAL NETWORKS FOR SPEECH RECOGNITION

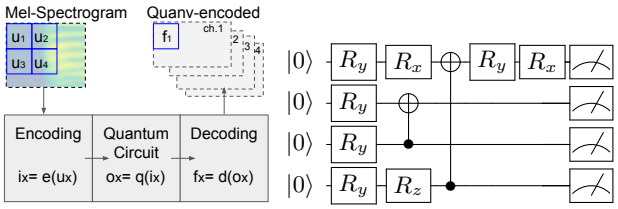
In this section, we present our framework showing how to design a federated architecture based QCNN composed of quantum computing and deep learning for speech recognition.

3.1. Speech Processing under Vertical Federated Learning

We consider a federated learning scenario for speech processing, where the ASR system includes two blocks deployed between a local user, and a cloud server or application interface (API), as shown in Fig. 1. An input speech signal, x_i, is collected at the local user and up-streamed to a cloud server where Mel spectrogram feature vectors are extracted, $\mathbf{u}_i$. Mel spectrogram features are the input of a quantum circuit layer, $\mathbf{Q}$, that learns and encodes patterns:

$$f_i = \mathbf{Q}(\mathbf{u}_i, \mathbf{e}, \mathbf{q}, \mathbf{d}), \quad \text{where} \quad \mathbf{u}_i = \text{Mel-Spectrogram}(\mathbf{x}_i). \quad (1)$$

In Eq. (1), the computation process of a quantum neural layer, $\mathbf{Q}$, depends on the encoding initialization $\mathbf{e}$, the quantum circuit parameters, $\mathbf{q}$, and the decoding measurement $\mathbf{d}$. The encoded features, f_i, will be down-streamed back to the local user and used for training the ASR system, more specifically the acoustic model (AM). Proposed decentralized-VFL speech processing model reduces the risk of parameter leakages [23, 1, 11] from attackers, and avoids privacy issues under GDPR, with its architecture-wise advantages [24] on encryption [16] and without accessing the data directly [3].



(a) QCNN Computing Process. (b) Deployed Quantum Circuit.

Fig. 2: *The proposed variational quantum circuit for 2 × 2 QCNN.*

3.2. Quantum Convolutional Layer

Motivated by using VQC as a convolution filter with a quantum kernel, QCNN [13] was recently proposed to extend CNN's properties to the quantum domain for image processing on a digital simulator and requires only fewer qubits to construct a convolution kernel during the QML progress. A QCNN consists of several quantum convolutional filters, and each quantum convolutional filter transforms input data using a quantum circuit that can be designed in a structured or a randomized fashion.

Figure 2 (a) show our implementation of a quantum convolutional layer. The quantum convolutional filter is consists of (i) the encoding function $e(\cdot)$, (ii) the decoding operation $d(\cdot)$, and (iii) the quantum circuit $q(\cdot)$. In detail, the following steps are performed to obtain the output of a quantum convolutional layer:

- The 2D Mel-spectrogram input vectors are chunked into several 2×2 patches, and the n^{th} patch is fed into the quantum circuit and encoded into intial quantum states, $\mathbf{I_x}[n] = e(\mathbf{u}_i[n])$.
- The initial quantum states go through the quantum circuit with the operator $q(\cdot)$, and generate $O_x[n] = q(\mathbf{I}_x[n])$.
- The outputs after applying the quantum circuit are necessarily measured by projecting the qubits onto a set of quantum state basis that spans all of the possible quantum states and quantum operations. Thus we get the desired output value, $f_{\mathbf{x},n} = d(\mathbf{O_x}[n])$. More details refer to the implementation[1].

3.3. Random Quantum Circuit

We deploy a random quantum circuit to realize a simple circuit U in which the circuit design is randomly generated per QCNN model for parameter protection. An example of random quantum circuit is shown in Figure 2 (b), where the quantum gates R_x, R_y and R_z and CNOT are applied. The classical vectors are initially encoded into a quantum state $\Phi_0 = |0000\rangle$, and the encoded quantum states go through the quantum circuit U for the following phases as:
Phase 1: $\Phi_1 = R_y|0\rangle R_y|0\rangle R_y|0\rangle R_y|0\rangle$.
Phase 2: $\Phi_2 = (R_x R_y|0\rangle)\text{CNOT}(R_y|0\rangle)R_y|0\rangle R_z R_y|0\rangle$.
Phase 3: $\Phi_3 = \text{CNOT}((R_x R_y|0\rangle))\text{CNOT}(R_y|0\rangle)R_y|0\rangle R_z R_y|0\rangle$.
Phase 4: $\Phi_4 = R_x R_y \Phi_3$

Besides, since random quantum circuit may involve many CNOT gates which bring about many unexpected noisy signals under the current non error-corrected quantum devices and the connectivity of physical qubits, we limit the number of qubits to small numbers to avoid exceeding the noise tolerance capabilities of VQC. In the simulation on CPU, we use PennyLane [7], which is an open-source programming software for differentiable programming of quantum computers, to generate the random quantum circuit, and we build the random quantum circuit based on the Qiskit [25] for simulation with the noise model from IBM quantum machines with 5 and 15 qubits, which is advanced than simulation only results [13].

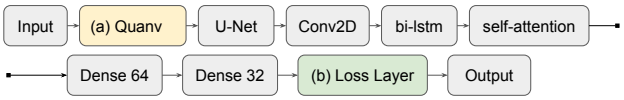


Fig. 3: The proposed QCNN architecture for ASR tasks.

3.4. Attention Recurrent Neural Networks

We use a benchmark deep attention recurrent neural network (RNN) [14] model from [26] as our baseline architecture for a local model (e.g., second-pass models [11, 12]) in the VFL setting. The model is composed of two layers of bi-directional long short-term memory [14] and a self-attention encoder [27] (dubbed $\text{RNN}_{\mathbf{Att}}$). In [26], this RNN model has been reported the best results over the other DNN based solutions included DS-CNN [28] and ResNet [29] for spoken word recognition.

To reduce architecture-wise variants on our experiments, we conduct ablation studies and propose an advanced attention RNN model with a U-Net encoder [30] (denoted as $\text{RNN}_{\mathbf{UAtt}}$). As shown in Fig. 3, a series of multi-scale convolution layers (with a channel size of 8-16-8) will apply on quantum-encoded (quanv) or neural convolution-encoded (conv) features to improve generalization of acoustic by learning scale-free representations [30]. We use $\text{RNN}_{\mathbf{Att}}$ and $\text{RNN}_{\mathbf{UAtt}}$ in our experiments to evaluate the advantages of using the proposed QCNN model. As shown in Fig 3 (b), we provide a loss calculation layer on the RNN backbone for our local model. For spoken word recognition, we use the cross-entropy loss for classification. The loss layer could further be replaced by connectionist temporal classification (CTC) loss [31] for a large-scale continuous speech recognition task in our future study.

4. EXPERIMENTS

4.1. Experimental Setup

As initial assessment of the viability our novel proposed framework, we have selected a limited-vocabulary yet reasonably challenging speech recognition task, namely the Google Speech Command-V1 [29]. For spoken word recognition, we use the ten-classes setting that includes the following frequent speech commands[2]: ['left', 'go', 'yes', 'down', 'up', 'on', 'right', 'no', 'off', 'stop'], with a total of 11,165 training examples, and 6,500 testing examples with the background white noise setup [29]. The Mel-scale spectrogram features are extracted from the input speech using the Librosa library; this step takes place in the NISQ server. The input Mel-scale feature is actually a 60-band Mel-scale, and 1024 discrete Fourier transform points into the quantum circuit as the required VFL setting. The experiments with the local model are carried out with Tensorflow, which is used to implement DNNs and visualization.

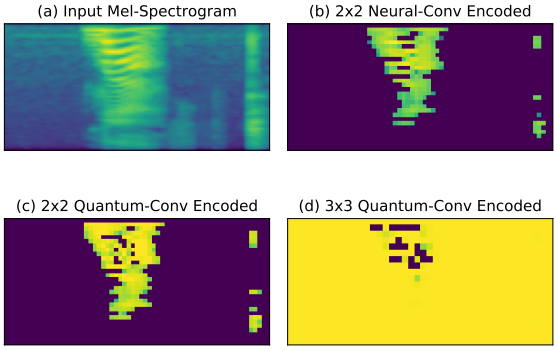


Fig. 4: Visualization of the encoded features from different types of convolution layers. The audio transcription is "yes" of the input.

4.2. Encoded Acoustic Features from Quantum Device

The IBM Qiskit quantum computing tool [25] is used to simulate the quantum convolution. We first use Qiskit to collect compiling noises from two different quantum computers. We then load those recorded noise to the Pennylane-Qiskit extension in order to simulate noisy quantum circuit experiments for virtualization. According to previous investigations [19, 18], the proposed noisy quantum device setup can be complied with NISQ directly and attains results close to those obtained using NISQ directly . The chosen setup preserves quantum advantages on randomization and parameter isolation.

Visualization of Acoustic Features. To better understand the nature of the encoded representation of our acoustic speech features,

[2]https://ai.googleblog.com/2017/08/launching-speech-commands-dataset.html

we visualize the encoded features and acoustic patterns extracted from different encoders. Fig. 4 shows QCNN-encoded results with a 2×2 kernel (in panel (c)), which seems to better relate to the acoustic pattern shown in the Mel spectrogram shown in panel (a), since it well captures energy patterns in both high and low-frequency regions. The latter becomes more evident by comparing panel (c) with the features encoded with a 3×3 kernel given in panel (d). Finally, the neural network-based convolution layer reported in panel (b) shows similar results with those in panel (c), but it presents a lower intensity in the high-frequency regions. We will discuss its relationship between recognition performance later in Section 4.4.

4.3. Performance of Spoken-Word Recognition

We conduct experiments on the spoken-word recognition task and compared the improved performance from an additional quantum convolution layer with a 2×2 kernel (4 qubits) and a neural convolution layer with a 2×2 kernel in Table 2. From the experiments, the recognition models with additional quantum convolution show better accuracy than the baseline models [26]. The modified model with a U-Net encoder, $\text{RNN}_{\mathbf{UAtt}}$, achieves the best performance of 95.12±0.18% on the evaluation data, which is better than the reproduced $\text{RNN}_{\mathbf{Att}}$ baseline (94.21±0.30%) for the recognition setup.

Table 2: Comparisons of spoken-term recognition on Google Commands dataset with the noise setting [29] for classification accuracy (Acc) ± standard deviation. The additional convolution (conv) and quantum convolution (quanv) layer have the same 2×2 kernel size.

Model	Acc. (↑)	Parameters (Memory) (↓)
$\text{RNN}_{\mathbf{Att}}$ [26]	94.21±0.30	170,915 (32-bits)
Conv + $\text{RNN}_{\mathbf{Att}}$	94.32±0.26	174,975 (32-bits)
Quanv + $\text{RNN}_{\mathbf{Att}}$	94.75±0.17	174,955 (32-bits) + 4 (qubits)
$\text{RNN}_{\mathbf{UAtt}}$	94.72±0.23	176,535 (32-bits)
Conv + $\text{RNN}_{\mathbf{UAtt}}$	94.74±0.25	180,595 (32-bits)
Quanv + $\text{RNN}_{\mathbf{UAtt}}$	**95.12**±0.18	180,575 (32-bits) + 4 (qubits)

4.4. A Study on QCNN Architectures

Next we experiment with various new QCNN [13] architectures for ASR with different combinations of quantum encoders and neural acoustic models. First, we study the quantum convolution encoder with different kernel sizes. From previous works [19, 18], the current commercial NISQ devices would be challenging to provide reproducible and stable results with a size of qubits larger than 15. We thus design our quantum convolutional encoders under this limitation with a kernel size of 1×1 (1 qubit), 2×2 (4 qubits), and 3×3 (9 qubits). We select two open source neural AMs as the local model, DS-CNN [28], and ResNet [29], from the previous works testing on the Google Speech Commands dataset. As shown in the bar charts in Fig. 5, QCNNs with the 2×2 kernel show better accuracy and lower deviations than all other models tested. QCNN attains 1.21% and 1.47% relative improvements over DS-CNN and ResNet baseline, respectively. On the other hand, QCNNs with the 3×3 kernel show the worst accuracy when compared with other configurations. Increasing the kernel size does not always guarantee improved performances in the design of QCNN for the evaluation. The encoded features obtained with a 3×3 quantum kernel used to train AMs, as shown in Fig. 4(d), are often too sparse and not as discriminative when compared to those obtained with 1×1 and 2×2 quantum kernels, as indicated in Fig. 4(b) and Fig. 4(c), respectively.

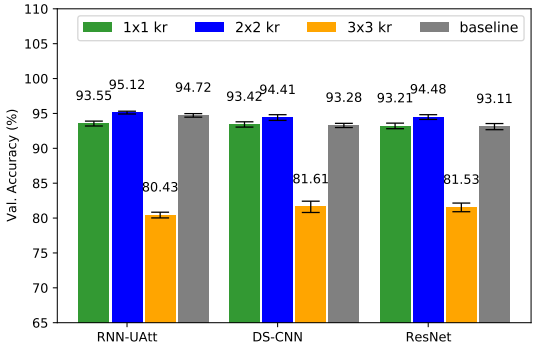


Fig. 5: Performance studies of different quantum kernel size (dubbed kr) with DNN acoustic models for designing QCNN models.

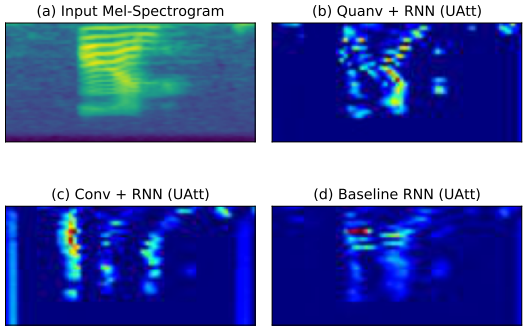


Fig. 6: Interpretable neural saliency results by class activation mapping [32] over (a) Mel spectrogram features with audio transcription of "on"; (b) a 2×2 quantum convolution layer followed by $\text{RNN}_{\mathbf{UAtt}}$; (c) a well-trained 2×2 neural convolution layer followed by $\text{RNN}_{\mathbf{UAtt}}$, and (d) baseline $\text{RNN}_{\mathbf{UAtt}}$.

4.5. A Saliency Study by Acoustic Class Activation Mapping

We use a benchmark neural saliency technique by class activation mapping (CAM) [32] over different neural acoustic models to highlight the responding weighted features that activate the current output prediction. As shown in Fig. 6, QCNN (b) learns much more correlated and richer acoustic features than RNN with a convolution layer and baseline model [26]. According to the CAM displays, the activated hidden neurons learn to identify related low-frequency patterns when making the ASR prediction from an utterance "on."

5. CONCLUSION

In this paper, we propose a new feature extraction approach to decentralized speech processing to be used in vertical federated learning that facilitates model parameter protection and preserves interpretable acoustic feature learning via **quantum convolution**. The proposed QCNN models show competitive recognition results for spoken-term recognition with stable performance from quantum machines when learning compared with classical DNN based AM models with the same convolutional kernel size. Our future work includes incorporating QCNN into continuous ASR. Although the proposed VFL based ASR architecture fulfilling some data protection requirements by decentralizing prediction models, more statistical privacy measurements [24] will be deployed to enhance the proposed QCNN models from the other privacy perspectives [24, 1].

6. REFERENCES

[1] 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.

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

[3] 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.

[4] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 8599–8603.

[5] M. Mohseni, P. Read, H. Neven, S. Boixo, V. Denchev, R. Babbush, A. Fowler, V. Smelyanskiy, and J. Martinis, "Commercialize quantum technologies in five years," *Nature*, vol. 543, no. 7644, pp. 171–174, 2017.

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

[7] V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, C. Blank, K. McKiernan, and N. Killoran, "Pennylane: Automatic differentiation of hybrid quantum-classical computations," *arXiv preprint arXiv:1811.04968*, 2018.

[8] 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.

[9] E. Farhi, J. Goldstone, S. Gutmann, and M. Sipser, "Quantum computation by adiabatic evolution," *arXiv preprint quant-ph/0001106*, 2000.

[10] J. Preskill, "Quantum computing in the nisq era and beyond," *Quantum*, vol. 2, p. 79, 2018.

[11] 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.

[12] 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," *arXiv preprint arXiv:2011.11715*, 2020.

[13] 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.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, "Quantum machine learning," *Nature*, vol. 549, no. 7671, pp. 195–202, 2017.

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

[17] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, "Parameterized quantum circuits as machine learning models," *Quantum Science and Technology*, vol. 4, no. 4, p. 043001, 2019.

[18] 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.

[19] S. Y.-C. Chen, S. Yoo, and Y.-L. L. Fang, "Quantum long short-term memory," *arXiv preprint arXiv:2009.01783*, 2020.

[20] F. Li, S. Zhao, and B. Zheng, "Quantum neural network in speech recognition," in *6th International Conference on Signal Processing, 2002.*, vol. 2. IEEE, 2002, pp. 1267–1270.

[21] J. Qi, C.-H. H. Yang, and J. Tejedor, "Submodular rank aggregation on score-based permutations for distributed automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3517–3521.

[22] S.-Y. Chuang, H.-M. Wang, and Y. Tsao, "Improved lite audio-visual speech enhancement," *arXiv preprint arXiv:2008.13222*, 2020.

[23] 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.

[24] 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.

[25] G. Aleksandrowicz, T. Alexander, P. Barkoutsos, L. Bello, Y. Ben-Haim, D. Bucher, F. Cabrera-Hernández, J. Carballo-Franquis, A. Chen, C. Chen, *et al.*, "Qiskit: An open-source framework for quantum computing," *Accessed on: Mar*, vol. 16, 2019.

[26] D. C. de Andrade, S. Leo, M. L. D. S. Viana, and C. Bernkopf, "A neural attention model for speech command recognition," *arXiv preprint arXiv:1808.08929*, 2018.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[28] Y. Zhang, N. Suda, L. Lai, and V. Chandra, "Hello edge: Keyword spotting on microcontrollers," *arXiv preprint arXiv:1711.07128*, 2017.

[29] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[30] C.-H. Yang, J. Qi, P.-Y. Chen, X. Ma, and C.-H. Lee, "Characterizing speech adversarial examples using self-attention u-net enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3107–3111.

[31] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[32] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.