Quantum Convolutional Neural Network for Speech Recognition

Glen S. Uehara
School of Electrical, Computer, and Energy Engineering
Arizona State University
Tempe, AZ U.S.A.

Abstract—**Convolutional neural networks (CNNs) have risen in popularity for many machine learning applications. Recently, the hybrid deep neural network (DNN) hidden Markov model (HMM) has been shown to improve speech recognition performance over the conventional Gaussian mixture model (GMM)-HMM. This performance improvement is attributed to the ability of the DNN to model complex correlations in speech features. Recurrent neural networks (RNNs) are another powerful model for sequential data. End-to-end training methods such as Connectionist Temporal Classification make it possible to train RNNs for sequence labeling problems where the input-output alignment is unknown. However, when we look at speech recognition, their performance has so far been disappointing, with better results returned by deep feedforward networks. This paper investigates RNN, CNN, and a new Quantum Convolution (Quanvolution) Neural Network from [1] and compares and contrasts this for speech recognition.**

I. Introduction

With recent advances in commercial quantum technology [2], quantum machine learning (QML) [3] has become an ideal building block for advancing classical algorithms and methodology. The input to QML, often represented by classical bits, needs to be encoded into quantum states based on quantum bits (qubits). We then take approximation algorithms (e.g., quantum branching programs [4]) that are applied to quantum devices based on a quantum circuit with noise [5].

In this research, we propose designing a model that uses a quantum convolutional neural network (QCNN) [1], [6] by combining a variational quantum circuit (VQC) learning paradigm [3] and a deep neural network [7] (DNN). We start with an. Recurrent neural network (RNN) as a baseline and build up our QCNN by first applying CNN structure. This allows us to take the classical approximation algorithm and transform them into a quantum implementation.

II. Background

A. Quantum Machine Learning for Signal Processing

QML [3] has been shown to have advantages where we have lower memory storage, secured model parameters encryption, and optimal feature representation capabilities[8]. We use a hybrid classical-quantum algorithm [9], where the input signals are given in a purely classical format and a quantum algorithm is employed in the feature learning phase. Quantum circuit learning is the most accessible and reproducible QML for signal processing [10], such as supervised learning in the design of quantum support vector machines [8]. It has been widely used, and it consists only of quantum logic gates with a possibility of deferring an error correction [3].

B. Quantum Learning and Speech Processing

Quantum technology is new, and there have been some attempts in exploiting it for speech processing. For example, Li et al. [11] proposed a speech recognition system with quantum backpropagation (QBP) simulated by fuzzy logic computing. In this proposal, the QBP is not using the qubit directly in a real-world quantum device. This approach does not demonstrate the quantum advantages inherent in this computing scheme. The QBP solution can be complicated to large-scale automatic speech recognition (ASR) tasks with parameters protection. From a system perspective, these accessible quantum advantages from VQL, including encryption and randomized encoding, are significant requirements for federated learning systems, such as distributed ASR.

III. Literature Research

The paper [1] proposes a decentralized feature extraction approach in federated learning to address privacy-preservation issues for speech recognition. This model 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). The input speech is first sent to a quantum computing system to extract Mel-spectrogram, and the corresponding convolutional features are encoded using a quantum circuit algorithm with random parameters. The encoded features are then sent to an RNN model for final recognition. The proposed framework takes advantage of the quantum learning progress to secure models and avoid privacy leakage attacks. In the paper, testing was done on the Google Speech Commands Dataset, and the proposed QCNN encoder attains a competitive accuracy of 95.12% in the model. This was found to be better than the previous architectures using centralized RNN models with convolutional features.

System Design and Algorithm

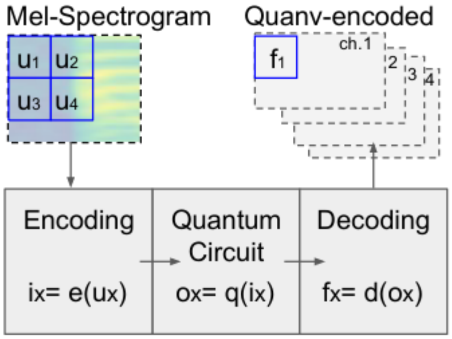
A. Speech Processing

We start with 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). 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, u_i. Mel spectrogram features are the input of a quantum circuit layer, Q, that learns and encodes patterns:

$f_i = Q(u_i, e, q, d)$ where $u_i = \text{Mel} - \text{Spectrogram}(x_i)$
(1)

In Eq. (1), the computation process of a quantum neural layer, Q, depends on the encoding initialization e, the quantum circuit parameters, q, and the decoding measurement d. The encoded features, f_i will be down-streamed back to the local user and used to train the ASR system, specifically the acoustic model (AM).

B. Quantum Convolution



(a) Quantum Convolution.

Figure 1 Proposed approach described in [1]

Figure 2 shows the 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 $q(\cdot)$.

The following steps are performed to obtain the output of a quantum convolutional layer:

- The 2D Mel-spectrogram input vectors are broken into into several 2x2 patches, and the nth patch is fed into the quantum circuit and encoded into intial quantum states, $I_x[n] = e(u_i[n])$.
- The initial quantum states is then run through the quantum circuit with the operator $q(\cdot)$., and generate $O_x[n] = q(I_x[n])$.
- The outputs after 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, $f_{x,n} = d(O_x[n])$.

C. Random Quantum Circuit

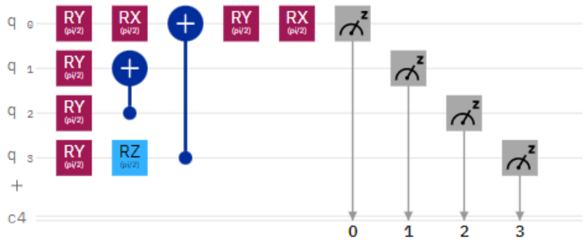


Figure 2 Proposed quantum circuit approach described in [1]

A random quantum circuit is used to realize a simple circuit U in where the circuit design is randomly generated per QCNN model for parameter protection. An example of a random quantum circuit is shown in Figure 2, 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 states goes through the quantum circuit U for the following phases as:

- Phase 1: $\Phi_1 = R_y|0\rangle R_y|0\rangle R_y|0\rangle R_y|0\rangle$
- Pase 2: $\Phi_2 = (R_xR_y|0\rangle)\text{CNOT}(R_y|0\rangle)R_y|0\rangle R_zR_y|0\rangle$
- Pase 3: $\Phi_3 = \text{CNOT}(R_xR_y|0\rangle)\text{CNOT}(R_y|0\rangle)R_y|0\rangle R_zR_y|0\rangle$
- Pase 4: $\Phi_4 = R_xR_y\Phi_3$

The random quantum circuit involves some CNOT gates that generate unexpected noisy signals under the current non-error-corrected quantum devices. Therefore, we limit the number of qubits to a smaller number to avoid exceeding the noise tolerance of the Variational Quantum Circuit (VQC). In the simulation on CPU, the paper uses PennyLane [12], which is open-source programming software for quantum computers, to generate the random quantum circuit, and build the random quantum circuit based on the Qiskit [13].

D. Classical vs. Quantum Convolution comparison

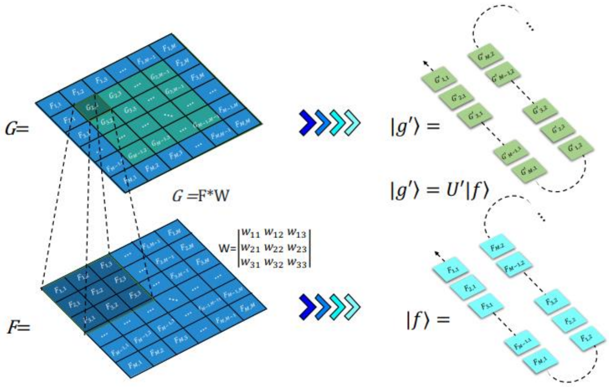

G=
F=
G =F*W
W= $\begin{vmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{vmatrix}$
|g'⟩ =
|g'⟩ = U'|f⟩
|f⟩ =

Figure 3 Comparing classical vs. Quantum described in [14]

We can compare classical convolution processing and quantum convolution processing. We start with F and G as the input and output image data, respectively. For a classical computer, an M × M image can be represented as a matrix and encoded with at least 2^n bits $[n = \lceil \log_2(M^2) \rceil]$. The classical image transformation through the convolution layer is performed by matrix computation F $*$ W, which leads to $x^{(l)}{}_{i,j} = \sum_{a,b=1}^{m} w_{a,b} x^{(l-1)}{}_{i+a-2,j+b-2}$. We can also take this image to represent it as a quantum state and encode it in at least n qubits on a quantum computer. The quantum image transformation is realized by the unitary evolution U on a specific quantum state

E. Proposed model

The final model for handling quantum convolution preprocessing is shown below.

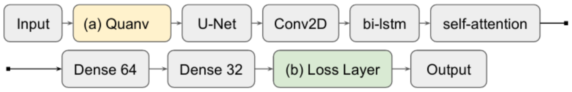


Figure 4 Proposed QCNN architecture described in [1]

This Quanv is added to pre-process the Mel-Spectrogram before running through the RNN system. As a comparison, we run a classical convolution as a comparison for the system.

Modeling and Simulation

This paper goes through a limited set of speech words and test with the models. In the paper, the results tested were seen as

Table 1: Results from [1] with the google voice dataset

Model	Accuracy
RNN	94.72±0.23
CNN+RNN	94.74±0.25
QUANV+RNN	95.12±0.18

To expand on the paper [1], we added speech words to test and built the proposed model. The new words were preprocessed with Quantum Convolution (Quanvolution). The following diagram shows how the features are extracted using quanvolutional coding.

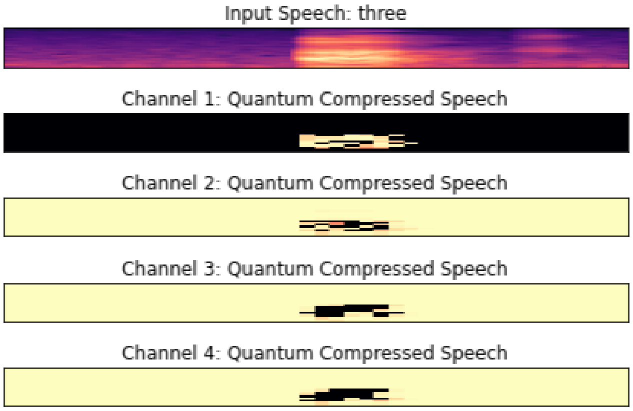


Figure 5 Extracted features using Quanvolution

The words were tested and evaluated for RNN, CNN+RNN and QUANV+RNN. In order to validate the model, we used the same training set for all three models. Then the same words were tested against the models. The following shows the confusion matrix for the three tests.

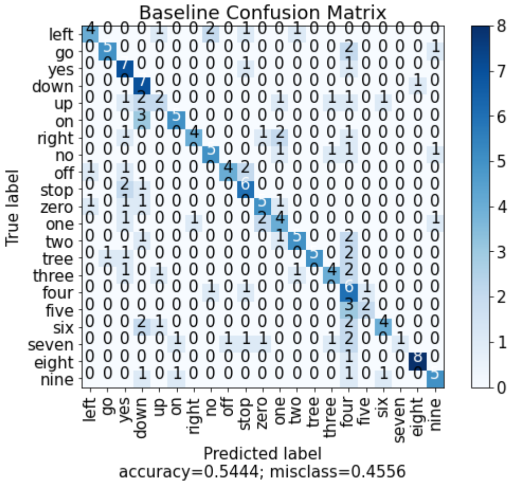


Figure 6 Baseline RNN with 21 words 54% accuracy

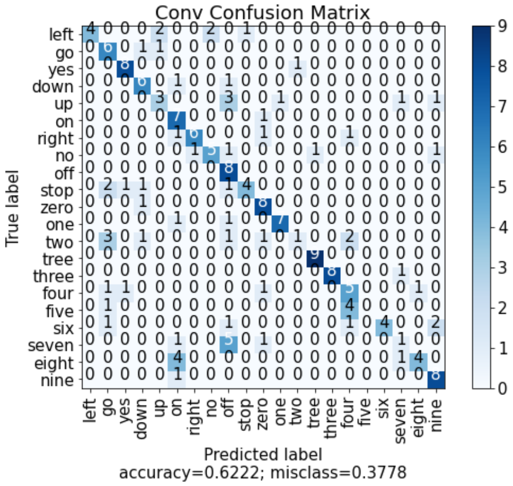


Figure 7 CNN+RNN with 21 words 62% accuracy

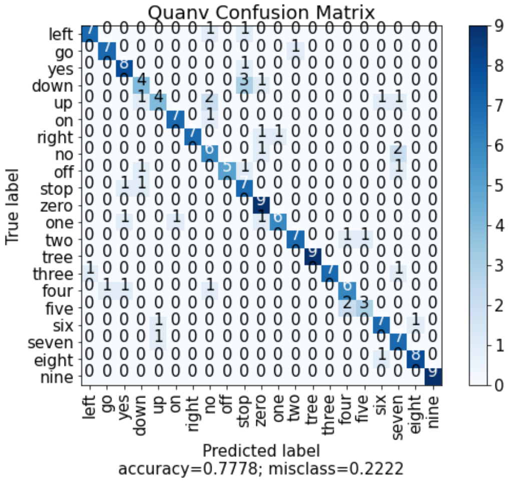


Figure 8 QUANV+RNN with 21 words 78% accuracy

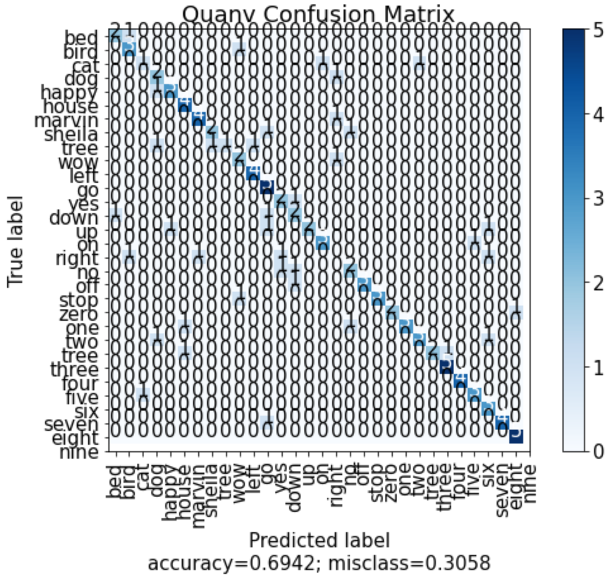


Figure 9 QUANV+RNN with 31 words 69% accuracy

As we can see, the QUANV+RNN has higher accuracy than CNN+RNN and RNN. We can see the validation curve in Figure 10.

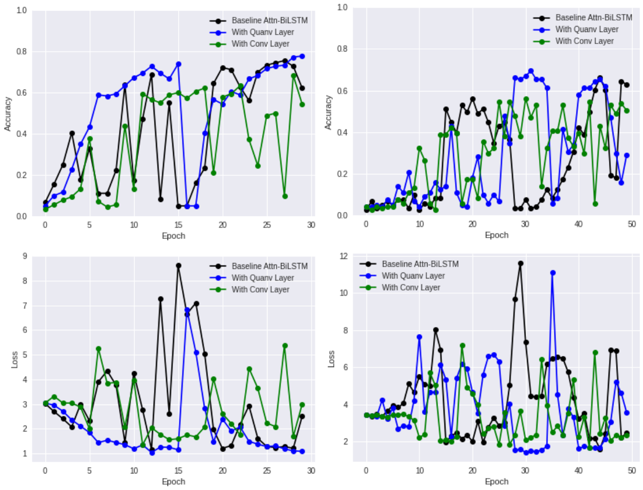


Figure 10 Validation curve for the three models

As we can see, our model with expanded words showed similar results where Quanvolutional coding to extract the features showed better results. However, the baseline results were not the same with the added words and similar words that we used. Therefore, this investigation aims not to verify the RNN baseline but to see if Quantum convolution to extract features shows a more promising result. In this case, we can conclude that we offer similar advantages as the paper.

Table 2: Results from simulation with additional voice dataset

Number Words	Model	Accuracy
21	RNN	54
21	CNN+RNN	62
21	QUANV+RNN	78
31	RNN	56
31	CNN+RNN	66
31	QUANV+RNN	69

We also compare the spectrogram through the process to verify what some features that were used for training are. The figures below shows some of these

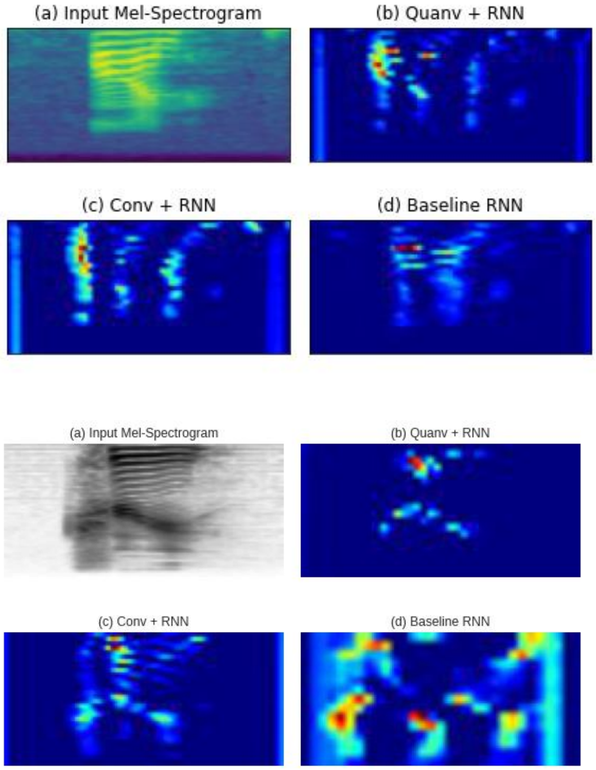


Figure 11 Spectrogram through the NN process

We can see that in some cases, the Quanvolution does not extract better, but in other cases it does. This area needs further research to strengthen the quantum circuit.

IV. Research Conclusion

This paper proposed that a new feature extraction approach to speech processing can 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 models with the same convolutional kernel size. Though we are limited by running on a Quantum simulator for this test, we were able to show similar results from the paper.

Furthermore, we showed that preprocessing with Quantum Convolution to extract features has an advantage over the classical convolution methods.

The next step is to continue testing different voice sets to see if the model holds for additional data.

REFERENCES

[1] 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 *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6523–6527. doi: 10.1109/icassp39728.2021.9413453.

[2] 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. Nature Publishing Group, pp. 171–174, Mar. 09, 2017. doi: 10.1038/543171a.

[3] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, "Quantum circuit learning," *Phys. Rev. A*, vol. 98, no. 3, p. 032309, Sep. 2018, doi: 10.1103/PhysRevA.98.032309.

[4] V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, M. S. Alam, S. Ahmed, J. M. Arrazola, C. Blank, A. Delgado, S. Jahangiri, K. McKiernan, J. J. Meyer, Z. Niu, A. Száva, and N. Killoran, "PennyLane: Automatic differentiation of hybrid quantum-classical computations," Nov. 2018, [Online]. Available: http://arxiv.org/abs/1811.04968

[5] 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, Mar. 2019, doi: 10.1038/s41586-019-0980-2.

[6] G. Chen, Q. Chen, S. Long, and W. Zhu, "Quantum Convolutional Neural Network For Image Classification," in *2020 8th International Conference on Digital Home (ICDH)*, Jun. 2021, pp. 116–120. doi: 10.1109/icdh51081.2020.00028.

[7] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Oct. 2013, pp. 8599–8603. doi: 10.1109/ICASSP.2013.6639344.

[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, Mar. 2019, doi: 10.1038/s41586-019-0980-2.

[9] M. Henderson, S. Shakya, S. Pradhan, and T. Cook, "Quanvolutional neural networks: powering image recognition with quantum circuits," *Quantum Mach. Intell.*, vol. 2, no. 1, Jun. 2020, doi: 10.1007/s42484-020-00012-y.

[10] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, "Quantum machine learning," *Nat. 2017 5497671*, vol. 549, no. 7671, pp. 195–202, Sep. 2017, doi: 10.1038/nature23474.

[11] F. Li, S. Zhao, and B. Zheng, "Quantum neural network in speech recognition," *Int. Conf. Signal Process. Proceedings, ICSP*, vol. 2, pp. 1267–1270, 2002, doi: 10.1109/ICOSP.2002.1180022.

[12] V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, M. S. Alam, S. Ahmed, J. M. Arrazola, C. Blank, A. Delgado, S. Jahangiri, K. McKiernan, J. J. Meyer, Z. Niu, A. Száva, and N. Killoran, "PennyLane: Automatic differentiation of hybrid quantum-classical computations," Nov. 2018, Accessed: Nov. 14, 2021. [Online]. Available: https://arxiv.org/abs/1811.04968v3

[13] G. Aleksandrowicz, T. Alexander, P. Barkoutsos, L. Bello, Y. Ben-Haim, D. Bucher, F. J. Cabrera-Hernández, J. Carballo-Franquis, A. Chen, C.-F. Chen, J. M. Chow, A. D. Córcoles-Gonzales, A. J. Cross, A. Cross, J. Cruz-Benito, C. Culver, S. D. L. P. González, E. D. La Torre, D. Ding, *et al.*, "Qiskit: An Open-source Framework for Quantum Computing," Jan. 2019, doi: 10.5281/ZENODO.2562111.

[14] S. Wei, Y. Chen, Z. Zhou, and G. Long, "A Quantum Convolutional Neural Network on NISQ Devices," 2021, Accessed: Nov. 24, 2021. [Online]. Available: http://arxiv.org/abs/2104.06918