Acta Scientific COMPUTER SCIENCES

Volume 6 Issue 9 September 2024

Review Article

QuViT: Quantum Vision Transformer

Pranav Durai*

OpenCV, Palo Alto, CA 94301, USA

***Corresponding Author:** Pranav Durai, OpenCV, Palo Alto, CA 94301, USA.

Received: September 06, 2024

Published: December 13, 2024

Abstract

Image classification has traditionally relied on Convolutional Neural Networks (CNNs) for their ability to extract visual features, recognize and learn patterns. However, the emergence of Vision Transformers (ViTs) as an alternative approach, inspired by Transformers in language tasks, brings the potential for capturing global image relationships and achieving competitive performance, interpretability, and scalability. The field of quantum computing has shown great promise, heralding a new era in computation and problem-solving. By harnessing the principles of quantum mechanics, quantum computers offer the potential to perform calculations at a scale and speed that surpass classical computers. This paper introduces QuViT, a quantum-accelerated vision transformer. With a novel q-input engine, q-encoder, and q-decoder, the proposed QuViT model follows a hybrid- approach that provides a promising avenue for building a quantum vision transformer that can handle yottabyte- scale image classification tasks with high accuracy, efficiency and paradigm shifting performance.

Keywords: Quantum Computing, Computer Vision, Vision Transformer

Introduction

The origins of quantum computing [1] can be traced back to the early 1980s, when physicist Richard Feynman first proposed the idea of using quantum mechanics to perform computation. The idea was further developed by mathematician Peter Shor in the 1990s, who demonstrated that a quantum computer could be used to factor large numbers exponentially faster than a classical computer. Qubits, which are analogous to classical bits but can exist in multiple states simultaneously due to quantum superposition principle. In a quantum system- Pauli, Hadamard, Controlled-NOT, SWAP, Toffoli, Controlled- Phase, and U-gates are used to build circuits.

Image classification and processing are fundamental tasks in computer vision, with applications spanning from medical diagnostics to autonomous vehicles and beyond. The primary objective of image analysis is to extract meaningful information from visual data. Transformers [2] were originally developed for natural language processing tasks, but they have since been adapted for use in computer vision.

The vision transformer

A Vision Transformer [3] (ViT) is a deep learning model that uses the transformer architecture to perform image classification tasks. In a ViT, the input image is first divided into a set of fixed-size non-overlapping patches, which are then flattened into a sequence of vectors. These vectors are then passed through a stack of transformer encoder layers, which perform computations on each vector based on its relationships with the other vectors in the sequence.

At each transformer encoder layer, the input vector sequence is first processed by a multi-head self-attention mechanism, which allows each vector to attend to the other vectors in the sequence and aggregate information from them. This is followed by a feed-forward network that applies a non-linear transformation to each vector independently. The output of the final transformer encoder layer is a sequence of vectors, each representing a different part of the input image. A MLP head is then added on top of the final sequence of vectors to predict the class label of the input image.

Literature survey

In their paper, Maria Schuld., *et al*. [4] discussed that quantum machine learning represents the convergence of quantum computing and traditional machine learning techniques to process information and tackle complex problems. quantum computing harnesses the unique properties of quantum states, such as superposition and entanglement, which allow for performing operations on multiple states simultaneously, potentially leading to significant computational speedups. In quantum computing, the fundamental unit of computation is the qubit, represented by a complex linear combination of the basis states |0⟩ and |1⟩. Quantum gates, expressed as unitary matrices, enable the manipulation of qubits, impacting their amplitudes, phases, and probabilities. These gates can perform operations on single qubits or controlled operations on multiple qubits, allowing for the implementation of complex quantum algorithms.

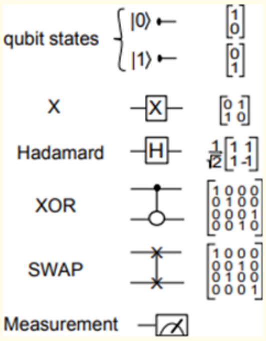

Figure 1: Visualization of qubit states, unitary gates and measurements for the quantum circuit model and matrix formalism [4].

Iris Cong., *et al.* [5] introduced Quantum Convolutional Neural Network (QCNN) as a novel quantum machine learning model inspired by convolutional neural networks, tailored for quantum computing applications. What sets QCNN apart is its exceptional efficiency, requiring only O(log(N)) variational parameters for input sizes of N qubits. This efficiency facilitates practical training and implementation on near-term quantum devices. The QCNN architecture merges key elements of the multi-scale entanglement renormalization ansatz and quantum error correction. Its potential is demonstrated through two illustrative examples. First, QCNN is applied to the accurate recognition of quantum states related to 1D symmetry-protected topological phases.

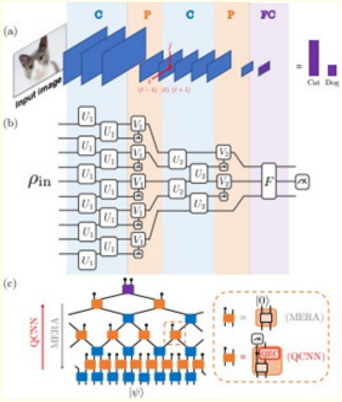

Figure 2: Architecture of CNN v/s QCNN [5].

QCNN's capacity to reproduce the phase diagram across a broad parameter range is showcased, even when trained on a small set of exactly solvable points. Additionally, an exact analytical QCNN solution is provided for this application. As a second application, QCNNs are leveraged to develop an optimized quantum error correction scheme, customized for a specific error model. The framework allows simultaneous optimization of both encoding and decoding procedures. The result is a quantum error correction scheme that outperforms existing quantum codes of comparable complexity. The paper also addresses potential avenues for experimental realization and explores generalizations of QCNNs. Overall, the Quantum Convolutional Neural Network demonstrates its promise for efficient quantum machine learning and quantum error correction, making it a valuable contribution to the field of quantum computing.

Amir Fijany and Colin P. Williams [6] proposed the idea of wavelet transforms in quantum computing, focusing on quantum image processing and data compression. While the quantum Fourier transform (QFT) is well-established and powerful in quantum algorithms, the study introduces the concept of quantum wavelet transforms as an equally useful tool in quantum computing. Wavelet transforms are employed in classical computing to unveil the multi-scale structure of signals, and the paper suggests their potential applicability in quantum domains. The research presents efficient quantum circuits for two representative quantum wavelet transforms: the quantum Haar and quantum Daubechies D(4) transforms. The approach involves factoring classical operators for these transforms into direct sums, direct products, and dot products of unitary matrices.

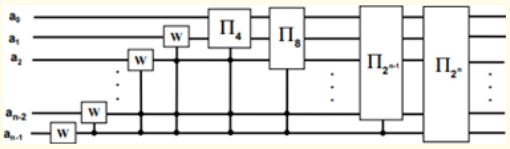

Figure 3: Block-level circuit for Haar wavelet [6].

Permutation matrices, a specific class of unitary matrices, play a central role in this quantum wavelet transform design. The study underscores an interesting observation that some operations that are straightforward and inexpensive in classical computing may not be as straightforward and inexpensive in the quantum realm, and vice versa. Specifically, certain permutation operations, often avoided explicitly in classical processing, must be explicitly performed in quantum computing, thus impacting the overall computational complexity of the quantum transform. The research addresses these issues, focusing on the set of permutation matrices relevant to quantum wavelet transforms, and develops efficient quantum circuits for their implementation. This work enables the design of efficient and comprehensive quantum circuits for quantum wavelet transforms, paving the way for their application in quantum image processing and data compression.

The paper by Matteo Farina., *et al*. [7] addresses the challenging computer vision problem of geometric model fitting, where the goal is to accurately fit geometric models to data points in an image. While quantum optimization has shown benefits for single-model fitting, this research introduces the first quantum approach to the more complex problem of multimodal fitting (MMF). In multimodal fitting, the objective is to fit multiple geometric models to the data, making it a more complex and open question. The paper demonstrates that quantum hardware can significantly enhance multimodal fitting and proposes an approach to MMF that can be efficiently sampled by modern adiabatic quantum computers.

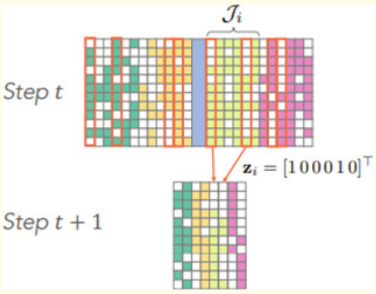

Figure 4: Iterative pruning technique in DeQuMF [7].

This approach does not require relaxing the objective function, a common technique used in classical MMF methods. The study also introduces an iterative and decomposed version of the proposed quantum method, which is designed to handle real-world-sized multimodal fitting problems. Experimental evaluations of the approach on various datasets show promising results, highlighting the potential of quantum computing in improving the accuracy and efficiency of geometric model fitting, even in the context of multimodal fitting challenges.

Riccardo Di Sipio., *et al*. [8] explored the application of quantum computing to enhance natural language understanding based on deep-learning models. The researchers successfully train a quantum-enhanced Long Short-Term Memory (LSTM) network for parts-of-speech tagging through numerical simulations. Additionally, they propose a quantum-enhanced Transformer for sentiment analysis using existing datasets. The paper starts by referencing Cambridge Quantum Computing's introduction of a "meaning-aware" Quantum Natural Language Processing (NLP) model that combines the semantic information of words with the syntactic structure of sentences. This concept is based on the idea that certain syntactic structures can be formulated using principles from quantum physics, such as quantum statistics, which extends classical statistics and holds parallels with aspects of human language understanding.

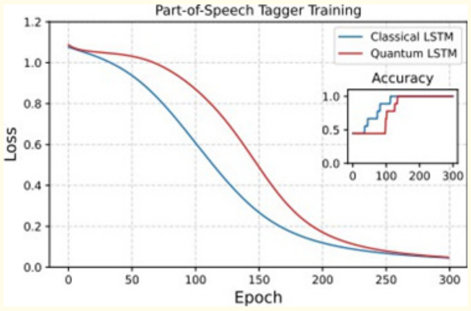

Figure 5: Cross-entropy loss and multi-class accuracy as a function of training epoch for classical v/s quantum LSTM [8].

The study explores the application of quantum computing in NLP and describes the inner workings of LSTM networks, which have been historically used for sequential data analysis. LSTMs combine "memory" and "statefulness" to determine the relevance of input components in computing the output. The paper discusses the formulas and parameters involved in LSTM calculations, highlighting the key gates, and the need to replace linear dense layers with quantum equivalents. The paper also suggests that quantum computing can bring innovations to the field of LSTM, offering the potential for improved natural language understanding.

Proposed Methodology

In this paper, QuViT - Quantum Vision Transformer has been proposed. This theoretical approach will provide a starting point for the development of a fully functional vision transformer that runs on a solely on a hybrid classic- quantum environment. Refer to Figure 6.

Q-Input engine

The Q-Input Engine in QuViT is responsible for preprocessing the input image and preparing it for quantum processing. Fourier transformation is one technique that assists with the transition of spatial information to frequency-based data points. This allows the identification of hidden patterns, textures and structures that are embedded within the data distribution. Feature extraction is one such domain where fourier transformation excels in, as it makes it really powerful for tasks such as classification or object objection.

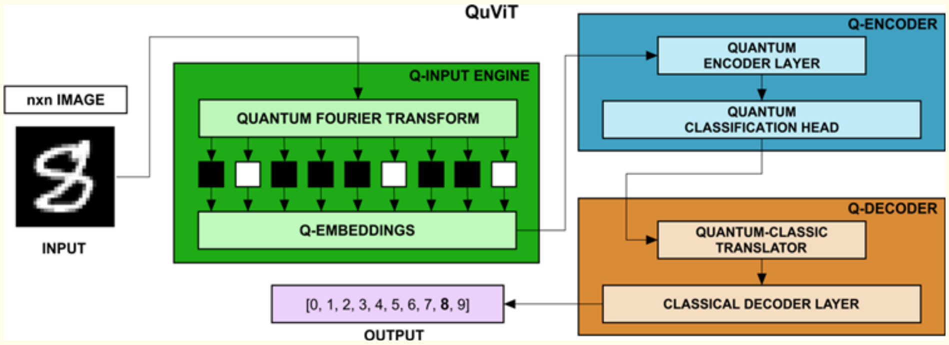

Figure 6: Quantum Vision Transformer Architecture.

Let's consider the input *nxn* image as *I*. Q-Input Engine proposed in this paper, applies quantum fourier transform (QFT) [9] to transform the image into a quantum state representation, ρ, capturing underlying patterns and structures, where.

$\rho \ = \ QFT(I)$ (1)

The QFT maps the spatial information of the image to the amplitudes of a quantum state, generating quantum embeddings that encode image information. These embeddings store relevant features and are represented as a quantum register, serving as input for subsequent quantum processing stages. A visual representation of this process has been illustrated in Figure 6, where an input image of resolution *nxn* is passed into the Q-Input Engine. In this stage, quantum fourier transformation is applied, Q- embeddings are generated for each input mapping. With this stage, classical information in the form of pixels and bits, have successfully been converted into quantum data format.

Q-Encoder

Guangxi Li., *et al*. [10] introduced the quantum self- attention mechanism in their research paper, where they used it for text classification. Its architecture has been illustrated in Figure 2. On quantum devices, classical inputs are used as rotation angles for quantum ansatzes, enabling their encoding into corresponding quantum states.

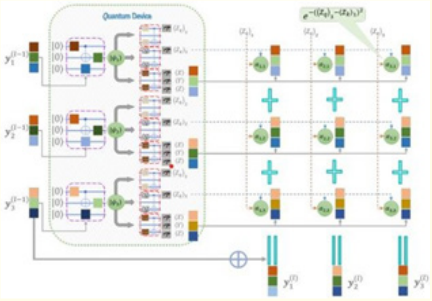

Figure 7: Quantum Self-Attention Mechanism [10].

These states are then subjected to three distinct classes of ansatzes, each serving a specific purpose: the top two classes represent the query and key parts, while the bottom class represents the value part.

This mechanism is vital to the quantum encoder layer present in the Q-Encoder. On classical computers, Gaussian functions are employed to compute the measurement outputs of the query and key parts, resulting in the derivation of quantum self-attention coefficients. Classically weighted sums of the measurement outputs from the value part are then computed, and their combination with the inputs yields the desired outputs. The weights applied in the weighted sums correspond to the normalized coefficients obtained during the process.

The quantum state ρ is represented as a density matrix, which encapsulates quantum information about the input image. Each element of the density matrix corresponds to the probability amplitude of a specific quantum state. The density matrix ρ can be expressed as:

$\rho \ = \ |\psi_1\rangle\langle\psi_1| \ + \ |\psi_2\rangle\langle\psi_2| + \ldots + |\psi_i\rangle\langle\psi_i|$ (2)

Here, $|\psi_i\rangle$ represents the individual quantum states, and $\langle\psi_i|$ represents their complex conjugate transpose. The sum extends over all quantum states generated by the Q- Encoder, which captures spatial relationships and dependencies among the quantum embeddings using quantum self-attention. The expression gets transformed as:

$$\begin{aligned}\rho' \ &= \ A(\rho) \ = \ A(|\psi_1\rangle\langle\psi_1|) \\ &\quad + \ A(|\psi_2\rangle\langle\psi_2|) + \ldots + A(|\psi_i\rangle\langle\psi_i|)\end{aligned} \tag{3}$$

Where A represents the quantum self-attention mechanism applied to each individual outer product term. The quantum self-attention mechanism is responsible for capturing correlations and relationships between different quantum states $|\psi_i\rangle$ in the density matrix ρ. The output ρ› represents the modified density matrix after applying quantum self-attention to each term.

The quantum classification head is the final component responsible for determining the class to which the input image belongs. After applying quantum self-attention to the modified density matrix ρ›, the quantum state ρ› encapsulates the quantum information and relationships between different quantum states $|\psi_i\rangle$. The main role of this components is to use this quantum information to make a classification decision. Let's update the mathematical expression to represent the quantum classification head's operation. After applying quantum self-attention and before classification:

$$\rho' = A(\rho) = A(|\psi_1\rangle\langle\psi_1|) + A(|\psi_2\rangle\langle\psi_2|) + \ldots + A(|\psi_i\rangle\langle\psi_i|) \quad (4)$$

Now, let's represent the process within the quantum classification head:

$$Classification\ Decision = Q(\rho') \quad (5)$$

Here, Q represents the quantum classification mechanism or classical neural network layers that take the modified density matrix ρ› as input and produce a classification decision. The output can be the class label to which the input image is assigned.

Q-Decoder

The final component in the proposed QuViT architecture is the Q-decoder. Within this, the quantum-classic translator component processes the classification decision, which is a result of the quantum classification head, into a quantum state that can be further processed in the classical decoder layer. This process can be represented as follows:

$$Quantum\ State\ for\ Decoder\ (\rho'') = T(Classification\ Decision) \quad (6)$$

Here, T represents the quantum-classic translator, and it transforms the classification decision into a quantum state ρ'' suitable for further quantum-classical processing. After obtaining the quantum state ρ›› from the quantum-classic translator, it is passed to the classical decoder layer for further processing. The classical decoder layer can include classical neural network layers that decode the quantum information and make it compatible with classical processing. This process can be represented as follows:

$$Decoded\ Quantum\ Information = D(\rho'') \quad (7)$$

Here, D represents the classical decoder layer, and it decodes the quantum information in ρ›› into a format that can be used for classical post-processing and output generation.

Future Works

At the time of writing this paper, the author did not have direct access to a fully functioning quantum computing machine. As a result, this work serves as an innovative theoretical foundation and proposal for the QuViT architecture. This offers a promising approach to leverage the computational advantages of quantum computing for image processing and classification tasks.

The author of this paper enthusiastically invites researchers who have access to quantum computing resources to design and conduct experiments based on the proposed architecture. These experiments should explore the practical implementation and performance of the proposed model. The author also encourages publishing the experimental results, highlighting the potential benefits and advancements achieved through the QuViT model.

Conclusions

This paper introduces and extensively explores the QuViT (Quantum Vision Transformer) architecture, a novel approach to revolutionizing image processing and classification tasks using the computational power of quantum computing. QuViT represents a significant advancement in the field of computer vision, offering the promise of more efficient and accurate image classification.

The proposed architecture is built upon a foundation of quantum image encoding, quantum Fourier transformations, and the application of quantum self- attention mechanisms. It integrates classical and quantum processing components to create a hybrid framework that can efficiently handle the intricacies of image data and patterns. It also addresses the challenges associated with traditional computer vision techniques by leveraging the superposition and entanglement properties of quantum states, which enable simultaneous processing of multiple states and capture intricate relationships within images.

Bibliography

1. Feynman RP. "Simulating physics with computers". *International Journal of Theoretical Physics* 21 (1982): 467-488.
2. Vaswani Ashish., *et al*. "Attention is all you need". *Advances in Neural Information Processing Systems* 30 (2017).
3. Dosovitskiy Alexey., *et al*. "An image is worth 16x16 words: Transformers for image recognition at scale". arXiv preprint arXiv:2010.11929 (2020).

4. Schuld Maria., *et al*. “An introduction to quantum machine learning”. *Contemporary Physics* 56.2 (2015): 172-185.

5. Cong Iris., *et al*. “Quantum convolutional neural networks”. *Nature Physics* 15.12 (2019): 1273-1278.

6. Fijany Amir and Colin P Williams. “Quantum wavelet transforms: Fast algorithms and complete circuits”. Quantum Computing and Quantum Communications: First NASA International Conference, QCQC’98 Palm Springs, California, USA February 17–20, 1998 Selected Papers. Springer Berlin Heidelberg, (1999).

7. Farina Matteo., *et al*. “Quantum Multi-Model Fitting”. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2023).

8. Di Sipio Riccardo., *et al*. “The dawn of quantum natural language processing”. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, (2022).

9. Roy Pradosh K. “Quantum Fourier Transform” (2020).

10. Li Guangxi., *et al*. “Quantum self- attention neural networks for text classification”. arXiv preprint arXiv:2205.05625 (2022).