Quantum Mixed-State Self-Attention Network

Fu Chen[1,2], Qinglin Zhao[1,*], Li Feng[1], Chuangtao Chen[1], Yangbin Lin[3], Jianhong Lin[4]
[1] Faculty of Innovation Engineering, Macau University of Science and Technology, Macao 999078, China
[2] New Engineering Industry College, Putian University, Putian, 351100, China
[3] Computer Engineering College, Jimei University, Xiamen 361021, China
[4] Mechanical, Electrical and Information Engineering College, Putian University, Putian, 351100, China

***Abstract*—The rapid advancement of quantum computing has increasingly highlighted its potential in the realm of machine learning, particularly in the context of natural language processing (NLP) tasks. Quantum machine learning (QML) leverages the unique capabilities of quantum computing to offer novel perspectives and methodologies for complex data processing and pattern recognition challenges. This paper introduces a novel Quantum Mixed-State Attention Network (QMSAN), which integrates the principles of quantum computing with classical machine learning algorithms, especially self-attention networks, to enhance the efficiency and effectiveness in handling NLP tasks. QMSAN model employs a quantum attention mechanism based on mixed states, enabling efficient direct estimation of similarity between queries and keys within the quantum domain, leading to more effective attention weight acquisition. Additionally, we propose an innovative quantum positional encoding scheme, implemented through fixed quantum gates within the quantum circuit, to enhance the model's accuracy. Experimental validation on various datasets demonstrates that QMSAN model outperforms existing quantum and classical models in text classification, achieving significant performance improvements. QMSAN model not only significantly reduces the number of parameters but also exceeds classical self-attention networks in performance, showcasing its strong capability in data representation and information extraction. Furthermore, our study investigates the model's robustness in different quantum noise environments, showing that QMSAN possesses commendable robustness to low noise.**

***Index Terms*—Machine learning, Quantum machine learning, Self-attention mechanism, Quantum self-attention mechanism, Text categorization**

I. Introduction

In the past few decades, the rapid development of quantum computing has attracted widespread attention in both the scientific and industrial communities. As a computational paradigm based on the principles of quantum mechanics, quantum computing has shown potential for extraordinary computational speed and efficiency in solving certain types of problems compared to classical computers. The application prospects of quantum computing are considered revolutionary, particularly in fields such as biology [1], cryptography [2], and communication technology [3]. With the advancement of quantum technology, introducing the concepts and methods of quantum computing into the field of machine learning has become a new direction explored by researchers. Quantum Machine Learning (QML), as a cutting-edge field intersecting quantum computing and machine learning [4]–[6], is attracting widespread attention in both academia and industry. Quantum machine learning provides new tools and methods for solving complex data processing and pattern recognition problems by combining the characteristics of quantum computing with classical machine learning algorithms [7].

Among the wide range of applications of machine learning, NLP is dedicated to enabling computers to understand, interpret, and generate human language, thereby achieving natural and smooth communication between humans and machines. In the early 21st century, with the development of deep learning technology, the field of NLP has experienced revolutionary advancements. Deep learning models, especially Convolutional Neural Networks (CNNs) [8] and Recurrent Neural Networks (RNNs) [9], with their powerful feature extraction and sequence modeling capabilities, have greatly advanced the progress of NLP tasks, including but not limited to text classification [10], [11], sentiment analysis [12], [13], and machine translation [14], [15]. In 2017, the introduction of the Transformer model, which incorporates self-attention networks, significantly enhanced the ability and efficiency of processing long sequence data [16]. Subsequently, OpenAI launched the GPT (Generative Pre-trained Transformer) [17] and its subsequent versions of pre-trained models, such as GPT-4 [18], which, by pre-training on massive text data, learned rich language knowledge and world knowledge, enabling the model to be directly applied to various NLP tasks without specific task training. Through natural and smooth conversations with humans, it demonstrated the tremendous potential of NLP technology [19]–[21]. It can not only answer complex questions but also write articles [21], generate code [22], and even create poetry and music [23], greatly broadening the application range of artificial intelligence and sparking widespread public discussion and imagination about the future possibilities of AI. Today, NLP is at an unprecedented peak of development, and its impact goes far beyond the scope of scientific research, becoming one of the key technologies that are changing the way humans live and work. However, with the growth of model size and data volume, classical deep learning methods face huge demands for computational resources and challenges in energy efficiency [24]–[26].

Quantum computing provides a new direction for the field

This work was supported by the Education and Scientific Research Project for Middle-aged and Young Teachers of the Fujian Provincial Education Department, China (JAT200499).

*Corresponding author: Qinglin Zhao (e-mail: qlzhao@must.edu.mo).

of Natural Language Processing. Quantum computers are expected to achieve quantum advantage [27], [28], which can be reflected in sample complexity or time complexity [29]. The core characteristics of quantum computing, such as quantum superposition and quantum entanglement, offer new possibilities for representing and processing the high-dimensional data of natural language, allowing for the representation of a large number of words and sentences in the high-dimensional Hilbert space [30], [31], and capturing complex relationships and semantic information between words. Moreover, quantum algorithms can achieve faster processing speeds than classical algorithms in some cases [32], [33], which is significant for accelerating the training and inference processes of NLP tasks. Through Quantum Natural Language Processing (QNLP), researchers explore the use of the unique advantages of quantum computing to process language data and perform NLP tasks [34], which has inspired some exploratory work.

Up to now, in 2023, Li et al. proposed a fully quantum learning model, QRNN [35], which stacks recurrent blocks in an alternating manner to reduce the algorithm's requirements for the coherence time of quantum devices. In 2022, SYC Chen et al. proposed a hybrid quantum-classical model, QLSTM [36], which extends the classical Long Short-Term Memory (LSTM) model to the quantum domain, replacing some classical neural networks in the LSTM unit with Variational Quantum Circuit (VQC) to construct a more efficient model. However, these quantum versions of RNN and LSTM models have the same problems as classical neural networks: they often struggle to capture long-distance dependencies and have limited capabilities in handling complex problems. In 2017, Niu et al. proposed a more efficient, parameter-free model combining quantum attention with LSTM based on weak measurement in quantum mechanics [37], which has better sentence modeling performance. However, this method mainly focuses on some physical principles of quantum mechanics and does not involve specific quantum circuit design.

A more effective quantum self-attention network is the Quantum Self-Attention Neural network (QSANN) model proposed by Baidu's team in 2022, which uses Gaussian projection quantum self-attention for text classification [38]. This model can explore the correlations between words in the high-dimensional quantum feature space. However, when processing quantum queries and keys, the model converts them into classical data through observables to calculate similarity. This process involves information loss and reduces the model's ability to leverage the advantages of quantum computing. In 2022, the Quantum Self-Attention Network (QSAN) model proposed by Zhao et al. [39] uses Quantum Logic Similarity (QLS) to prevent measurement from obtaining inner products and Quantum Bit Self-Attention Score Matrix (QBSASM) to generate a density matrix that effectively reflects the output attention distribution, thereby enhancing the model's information extraction capability. In 2023, Zhao et al. proposed the Quantum Kernel Self-Attention Network (QKSAN) model [40], which combines the data representation advantages of quantum kernel methods with the efficient information extraction capability of self-attention mechanism, providing a larger and more complex data representation space. Although these two methods compute the similarity of quantum queries and keys at the quantum level, they are limited to the pure state level and rely on the unitary transformation of quantum circuits, resulting in limited expressive power. Moreover, up to now, these implemented quantum self-attention networks have not yet introduced position information, The potential of the models has not been fully realized.

To overcome the issues mentioned above and explore the advantages of quantum computing in improving classical self-attention network models for more effective attention results, we propose a novel model, called Quantum Mixed-State Attention Network (QMSAN) model. This model is based on trainable quantum embeddings, quantum attention weight coefficients based on mixed states, and non-trainable quantum positional information embedding. To evaluate the performance of our model, we conducted numerical experiments with various datasets. Compared to classical self-attention networks, our model significantly reduces the number of parameters under the same input sequence conditions. For the self-attention network model in the same quantum domain, our model demonstrates superior performance, indicating that QMSAN model has stronger data representation and information extraction capabilities. The main contributions of this paper are summarized as follows:

- We propose a novel quantum attention weight coefficients calculation mechanism based on mixed states. In the context of quantum computing, the representations of queries and keys are not conventional pure-state qubits but are realized through quantum mixed states. This allows the model to capture richer information and intrinsic data correlations. The similarity between queries and keys is directly estimated at the quantum level without degrading the quantum information of queries and keys into classical information for processing. This calculation process not only maintains the efficiency and parallelism of quantum computing but also avoids the accuracy decline caused by information loss.
- Recognizing the importance of positional information in many NLP tasks, we propose a novel quantum positional encoding scheme. This scheme adopts an absolute positional encoding form without the need for additional qubits. We implement positional information encoding by introducing additional fixed quantum gates into the quantum circuit, which avoids the extra demand for qubits while maintaining the efficiency and accuracy of encoding.
- We incorporate a trainable quantum embedding model into our model, further optimizing the implementation framework of quantum self-attention networks by integrating the originally separate fixed quantum embedding and trainable quantum neural network (QNN) structures into a unified trainable quantum embedding model, exploring and verifying its application potential in quan-

tum self-attention networks. Through in-depth analysis and experimental validation of different quantum entanglement structures, compared to conventional separate structures, this model can more accurately capture and process complex relationships between data, significantly improving the model's performance and processing efficiency. We clarify the superiority of the trainable quantum embedding model in quantum self-attention networks.

The rest of the paper is structured as follows: First, in Section II, we summarize the basic theory and methods. Then, in Section III, we elaborate on our innovative QMSAN framework and introduce its corresponding quantum circuits. The numerical simulation setup and comparison results with other attention models are presented in Section IV. Finally, Section V concludes the paper.

II. PRELIMINARIES

Before delving into quantum self-attention networks, it is necessary to understand a few fundamental concepts of quantum mechanics, including quantum states, the superposition property of quantum states, quantum entanglement, and observables. These concepts are the foundation of quantum computing and are crucial for developing and understanding quantum algorithms.

The state of a quantum system reveals the physical properties of the system, such as the position, momentum, or spin of particles [41]. Unlike classical systems, the state of a quantum system can be in a superposition of multiple possible states, providing quantum computing with its unique capabilities. In quantum mechanics, the states of quantum systems are generally described in the form of pure states and mixed states. A pure state is one of the most basic ways to describe a quantum system, representing the quantum system in a completely determined quantum state. A pure state can be represented by a vector $|\psi\rangle$ in Hilbert space, where $|\psi\rangle$ is called the quantum state. For a simple qubit system, its pure state can be represented as follows:

$$|\psi\rangle = \alpha\,|0\rangle + \beta\,|1\rangle \tag{1}$$

where $|0\rangle$ and $|1\rangle$ represent the two basis states of the qubit, while α and β are complex coefficients. The absolute squares of these coefficients ($|\alpha|^2$ and $|\beta|^2$) represent the probabilities of measuring the corresponding basis states, and they satisfy $|\alpha|^2 + |\beta|^2 = 1$.

In contrast to pure states, mixed states are used to describe a quantum system that is in a probabilistic mixture of multiple pure states. This type of state describes a quantum system whose state is not completely known. Mixed states are usually represented by a density matrix ρ:

$$\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i| \tag{2}$$

where $|\psi_i\rangle$ are the pure states the system can be in, and p_i is the probability that the system is in state $|\psi_i\rangle$, satisfying $\sum_i p_i = 1$.

Quantum systems evolve through linear and unitary evolution via quantum circuits, transforming from one initial state to another. Mathematically, a pure state $|\psi\rangle$ can evolve into another pure state $|\psi'\rangle$ through the action of a quantum gate (or quantum circuit):

$$|\psi'\rangle = U|\psi\rangle \tag{3}$$

The evolution of a mixed state is described by its density matrix, where a mixed state ρ evolves into a new mixed state ρ' as:

$$\rho' = U\rho U^\dagger \tag{4}$$

where U is a unitary matrix representing the action of the quantum gate or quantum circuit, satisfying $UU^\dagger = U^\dagger U = I$. $U^\dagger$ is the conjugate transpose of U, and I is the identity matrix.

In quantum computing, a quantum system can extract computational results using observables at the final output stage of a quantum computer, converting quantum information into classical data through measurement. A projective measurement is described by an observable M, a Hermitian operator on the state space of the system being observed. Mathematically, an observable M can be represented as a combination of its eigenvalues λ_i and corresponding projection operators P_i, i.e., $M = \Sigma_i \lambda_i P_i$. The measurement result will randomly obtain an eigenvalue λ_i, and at the same time, the quantum state $|\phi\rangle$ will collapse to the corresponding eigenstate with probability $p(\lambda_i) = \langle\phi|P_i|\phi\rangle$. Therefore, the average value of the observable M can be expressed as:

$$\langle M\rangle = \sum_i \lambda_i \langle\phi|P_i|\phi\rangle \tag{5}$$

For mixed states, the expectation value of the observable M can be calculated using the density matrix ρ:

$$\langle M\rangle = \mathrm{tr}(\rho M) \tag{6}$$

where $\mathrm{tr}(\cdot)$ represents the trace operation. In this paper, we will use the observable Z where $Z = (+1)\,|0\rangle\langle 0| + (-1)\,|1\rangle\langle 1|$, for example, in a system of n qubits, the observable n for the first qubit is mathematically expressed as $Z_1 = Z \otimes I^{\otimes(n-1)}$.

III. METHOD

In QMSAN model, we first transform the classical input data $\boldsymbol{x}_s$ into quantum states $|x_{s,q}\rangle$, $|x_{s,k}\rangle$, and $|x_{s,v}\rangle$ through three trainable Quantum Embeddings. This process involves quantum feature mapping, which directly converts classical data into quantum states. Then, we calculate the mixed-state similarity between $|x_{s,q}\rangle$ and $|x_{s,k}\rangle$, and measure the observable Z for $|x_{s,v}\rangle$. Afterwards, the data is fed into a classical fully connected network to complete binary prediction tasks. The architectural design of QMSAN model is illustrated in Fig. 1.

In this section, we will detail the components of the Quantum Mixed-State Self-Attention Network (QMSAN).

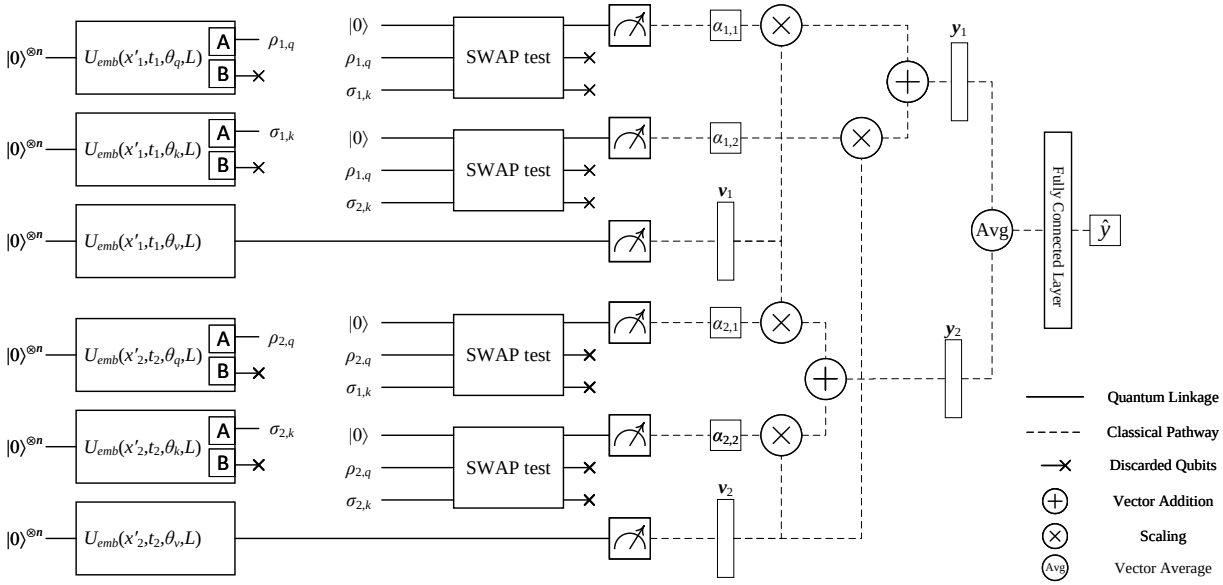


Fig. 1. Quantum Mixed-Stae self-attention network framework.

A. Quantum Embedding

Typically, quantum machine learning methods focus on using fixed quantum feature maps, followed by applying a trainable, adaptable variational quantum circuit to adjust the measurement basis. This architecture, consisting of a data encoding circuit and a learnable Quantum Neural Network (QNN), is widely applied in various scenarios such as quantum convolutional neural networks [42], Quantum Recurrent Neural Networks [35], strongly entangling circuit architectures [43], and searched architectures [44], [45]. However, this approach requires careful design of the encoding circuit, as the fixed quantum feature map significantly impacts the algorithm's generalization performance, and most computational resources are used for the QNN. Yet, Ref. [46]–[48] suggest that this might not be the most efficient method.

If the data is already well-mapped in Hilbert space, then subsequent tasks can be achieved with a shallow quantum classifier circuit. This is similar to the "feature extractor" in classical machine learning [49], [50], where networks are trainable with the core purpose of converting or encoding input data (such as images, text, or audio) into a new feature space. This feature space more effectively represents the key information of the data, facilitating subsequent tasks such as classification, regression, or clustering. Therefore, we can focus the adaptive training of the quantum circuit on training a trainable quantum feature map. This maps classical data into Hilbert space. We refer to this process as quantum embedding.

We introduce a repetitive iterative architecture for quantum embedding [47], as shown in Fig. 2. In classical self-attention networks, there are three parts: query, key, and value. Similarly, in our quantum self-attention network, we train three quantum embeddings. Through these embeddings, we represent classical data $\boldsymbol{x}_s$ as quantum states $|x_{s,q}\rangle$, $|x_{s,k}\rangle$, and $|x_{s,v}\rangle$, where $1 \leq s \leq S$ and S represents the number of input vectors in the data sample.

Specifically, these three mappings to $|x_q\rangle$, $|x_k\rangle$, and $|x_v\rangle$ use the same ansatz structure, implemented with different parameters $\boldsymbol{\theta}_q$, $\boldsymbol{\theta}_k$, and $\boldsymbol{\theta}_v$ for query, key, and value functions, respectively. The ansatz employs single qubit and two qubit quantum gates. First, we use the single qubit gate $R_x(x_i)$ to encode the input data $\boldsymbol{x} = (x_1, \ldots, x_N)^T$ into the quantum circuit. Then, we use the $R_{zz}(\theta_1) = e^{-i\theta_1 \sigma_z \otimes \sigma_z}$ gate to entangle the qubits and add $R_y(\theta_2)$. To enhance the expressiveness of the quantum circuit, it can contain L layers of such structure. Finally, we add the single qubit gate $R_x(x_i)$ again in the last layer to encode the data. Thus, the entire circuit can be represented as $U_{emb}(\boldsymbol{x}, \boldsymbol{\theta})$, where $\boldsymbol{x}$ is the input data and $\boldsymbol{\theta}$ are the trainable parameters. Each layer consists of a data encoding circuit block $S(\boldsymbol{x})$ and a trainable circuit block $W(\boldsymbol{\theta}_l)$ controlled by the trainable parameters $\boldsymbol{\theta}_l$ of each layer.

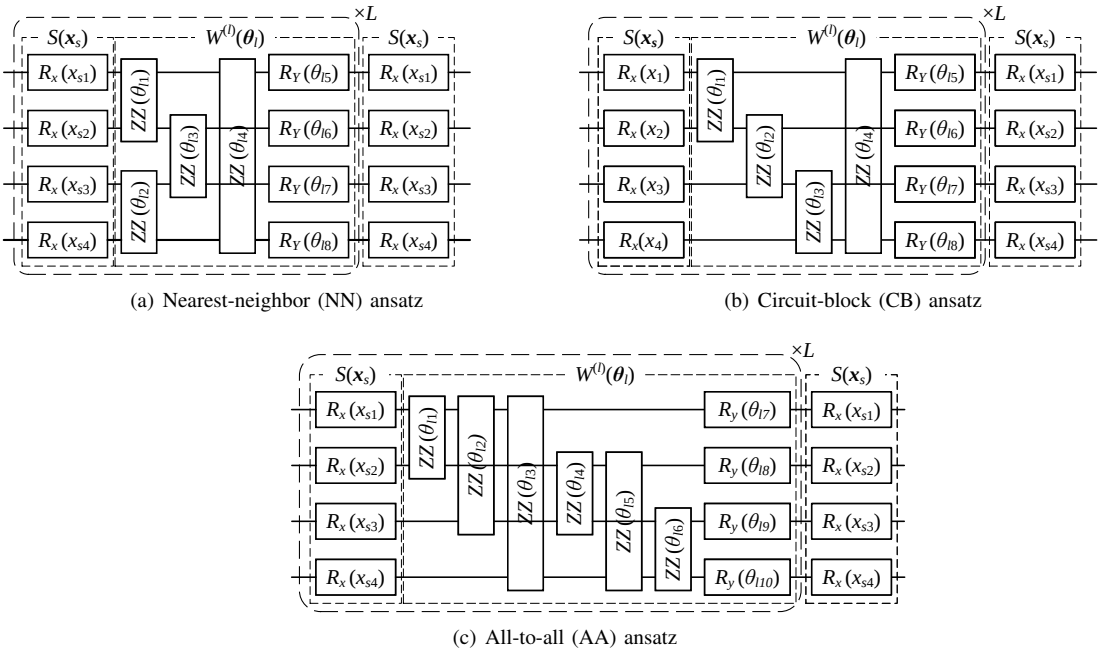
$$U_{emb}(\boldsymbol{x}, \boldsymbol{\theta}, L) = S(\boldsymbol{x}) \prod_{l=1}^{L} \left(W^{(l)}(\boldsymbol{\theta}_l) S(\boldsymbol{x}) \right) \tag{7}$$

Therefore, through three trainable quantum embeddings, we embed the input data $\boldsymbol{x}_s$ into three quantum states:

$$\begin{aligned} |x_{s,q}\rangle &= U_{emb}(\boldsymbol{x}_s, \boldsymbol{\theta}_q, L) \, |0\rangle^{\otimes n} \\ |x_{s,k}\rangle &= U_{emb}(\boldsymbol{x}_s, \boldsymbol{\theta}_k, L) \, |0\rangle^{\otimes n} \\ |x_{s,v}\rangle &= U_{emb}(\boldsymbol{x}_s, \boldsymbol{\theta}_v, L) \, |0\rangle^{\otimes n} \end{aligned} \tag{8}$$

B. Quantum Self-Attention Mechanism

When designing a quantum self-attention network, a straightforward and natural way to calculate the similarity between queries and keys is to use the inner product: $\alpha_{s,j} = |\langle x_{s,q} | x_{j,k} \rangle|^2$. However, in quantum circuits, since only unitary transformations are performed, the transformation between $|x_{s,q}\rangle$ and $|x_{j,k}\rangle$ for the same number of qubits can be considered a rotation operation, as they both reside in the



(a) Nearest-neighbor (NN) ansatz

(b) Circuit-block (CB) ansatz

(c) All-to-all (AA) ansatz

Fig. 2. Three Quantum Embedding Ansatzes with Different Entanglement Layers.

same Hilbert space. In contrast, in classical self-attention networks, the query vector $\boldsymbol{q} = \boldsymbol{x}W_q$ and the key vector $\boldsymbol{k} = \boldsymbol{x}W_k$, where $\boldsymbol{x}$ is the input vector, and W_q and W_k are the corresponding weight matrices. W_q and W_k change the direction and length of $\boldsymbol{x}$. The relationship between $\boldsymbol{q}$ and $\boldsymbol{k}$ involves rotation and scaling. Therefore, Ref. [38] argues that this makes it difficult for $|x_s\rangle$ to simultaneously correlate those $|\psi_j\rangle$ that are far away. This direct extension is not suitable or reasonable for working as the quantum self-attention. Furthermore, the method proposed in the paper involves converting the quantum states of queries and keys into classical data through observable and then calculating their similarity. However, this approach has a potential limitation in that the observable Z may not fully capture all the information of the quantum states, leading to the loss of some quantum information.

Based on the above analysis, inspired by the Hilbert-Schmidt distance, we proposes a quantum self-attention network based on mixed states to these shortcomings. By performing partial trace operations on the pure state query and key vectors, we obtain mixed-state queries and keys. Although this reduces their dimensions, it also introduces rotational and scaling characteristics. Our method keeps the quantum information at the quantum level until the final measurement step, avoiding measurement in the intermediate process of calculating the similarity between queries and keys. This prevents potential information loss that may occur during the conversion of quantum information into classical information before calculating the similarity between quantum queries and keys.

The Hilbert-Schmidt distance is an important distance metric in quantum information theory [51]. It can be measured and optimized with a small quantum circuit, making it significant for near-term quantum computing [47]. Its definition is as follows:

$$D_{\mathrm{HS}}(\rho, \sigma) = \mathrm{tr}\left((\rho - \sigma)^2\right) \tag{9}$$

Expanding the equation above results in three distinct terms: $\mathrm{tr}(\rho\sigma)$, $\mathrm{tr}(\sigma^2)$, and $\mathrm{tr}(\rho^2)$ [47]. The term $\mathrm{tr}(\rho\sigma)$ quantifies the distance between two ensembles in Hilbert space through the overlap between clusters; a value of $\mathrm{tr}(\rho\sigma) = 1$ suggests that the ensembles are formed from identical pure states, whereas $\mathrm{tr}(\rho\sigma) = 0$ indicates orthogonality among all embedded data points. User The 'purity' terms $\mathrm{tr}(\rho^2)$ and $\mathrm{tr}(\sigma^2)$ assess the overlap within clusters. In the quantum self-attention network, since we are more concerned with the similarity between queries and keys, we omit $\mathrm{tr}(\sigma^2)$ and $\mathrm{tr}(\rho^2)$, ultimately using only $\mathrm{tr}(\rho\sigma)$ as the method for calculating the similarity between queries and keys in the quantum self-attention network.

For $|x_q\rangle$ and $|x_k\rangle$, they are obtained from the initial state $|0\rangle^{\otimes n}$ through the unitary transformation $U_{emb}(\boldsymbol{x}, \boldsymbol{\theta})$, so $|x_q\rangle$ and $|x_k\rangle$ are both pure states. To obtain the mixed states, we extract information from the first $n/2$-qubit subsystem A of the entire n-qubit quantum system by performing a partial trace operation on the quantum system and discarding the remaining n/2-qubit subsystem B. Specifically, this operation transforms the pure states $|x_q\rangle$ and $|x_k\rangle$ of the entire system into the mixed states ρ_q and σ_k of the corresponding subsystems, respectively:

$$\begin{aligned} \rho_q &= \mathrm{tr}_B(|x_q\rangle\langle x_q|) \\ \sigma_k &= \mathrm{tr}_B(|x_k\rangle\langle x_k|) \end{aligned} \tag{10}$$

where $\mathrm{tr}_B(\cdot)$ is the partial trace over system B.

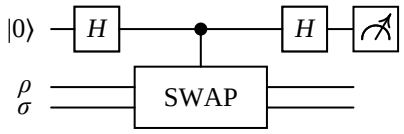


Fig. 3. Quantum circuit implementing the SWAP test.

We define quantum self-attention weight coefficient between the s-th and j-th mixed states, computed from the corresponding query and key parts:

$$\alpha_{s,j} = \mathrm{tr}(\rho_{s,q}\sigma_{j,k}) \tag{11}$$

The above equation can be easily implemented by the SWAP test quantum circuit [52], [53], as shown in Fig. 3, with the proof as follows.

Suppose we have a pair of mixed states ρ and σ of n qubits, with $\rho = \sum_i p_i|e_i\rangle\langle e_i|$ and $\sigma = \sum_i q_i|f_i\rangle\langle f_i|$ decomposed using their respective orthogonal bases $|e_i\rangle$ and $|f_i\rangle$. If we perform a measurement on the auxiliary qubit and obtain the result $|0\rangle$, the SWAP test passes, otherwise it fails. Therefore, in this case, the probability of $|e_i\rangle \otimes |f_j\rangle$ passing the SWAP test [54] is

$$p(|0\rangle) = \frac{1}{2} + \frac{|\langle e_i|f_j\rangle|^2}{2} \tag{12}$$

For $|e_i\rangle \otimes |f_j\rangle$ in the probability $p_i q_j$, the probability of the mixed state $\rho \otimes \sigma$ passing the SWAP test is:

$$\begin{aligned} p(|0\rangle) &= \sum_i \sum_j p_i q_j \left(\frac{1}{2} + \frac{|\langle e_i|f_j\rangle|^2}{2}\right) \\ &= \frac{1}{2} + \frac{1}{2}\sum_i \sum_j p_i q_j \langle e_i|f_j\rangle\langle f_j|e_i\rangle \\ &= \frac{1}{2} + \frac{1}{2}\sum_i \sum_j p_i q_j \mathrm{tr}(|e_i\rangle\langle e_i|f_j\rangle\langle f_j|) \\ &= \frac{1}{2} + \frac{1}{2}\mathrm{tr}\left[\left(\sum_i p_i|e_i\rangle\langle e_i|\right)\left(\sum_j q_j|f_j\rangle\langle f_j|\right)\right] \\ &= \frac{1}{2} + \frac{1}{2}\mathrm{tr}(\rho\sigma) \end{aligned} \tag{13}$$

Therefore, we use the SWAP test quantum circuit to implement the calculation of quantum self-attention weight coefficients between queries and keys. This can effectively estimate the closeness of two mixed states. If the two mixed states are identical, $\rho = \sigma$, the test always passes with $p = 1$. When the states are different, the finite probability p of passing the test depends on the similarity $\mathrm{tr}\,(\rho\sigma)$ between the two states; the closer they are, the greater the probability of passing the test. The output solution process is described in matrix form, and the weight coefficients matrix can be represented as

$$\mathbf{A} = \begin{bmatrix} \tilde{\alpha}_{0,0} & \tilde{\alpha}_{0,1} & \cdots & \tilde{\alpha}_{0,n-1} \\ \tilde{\alpha}_{1,0} & \tilde{\alpha}_{1,1} & \cdots & \tilde{\alpha}_{1,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\alpha}_{n-1,0} & \tilde{\alpha}_{n-1,1} & \cdots & \tilde{\alpha}_{n-1,n-1} \end{bmatrix} \tag{14}$$

where $\tilde{\alpha_{s,j}}$ represents the normalized quantum self-attention coefficient:

$$\tilde{\alpha}_{s,j} = \frac{\alpha_{s,j}}{\sum_{m=1}^{S} \alpha_{s,m}} \tag{15}$$

For the value part, we use an n-dimensional vector to represent it, with the observable Z measured for each qubit, resulting in a vector with the same dimension as the number of qubits.

$$\boldsymbol{v}_s = \begin{bmatrix} \langle Z_1\rangle_s & \langle Z_2\rangle_s & \cdots & \langle Z_n\rangle_s \end{bmatrix}^\top \tag{16}$$

Finally, the classical form of the output y_s can be calculated by the following formula. We adopt the structure of a residual network to design the output to prevent the network from degeneration:

$$\boldsymbol{y}_s = \boldsymbol{x}_s + \mathbf{A} \cdot \boldsymbol{v}_s \tag{17}$$

C. Quantum Position Encoding

Our QMSAN, similar to classical self-attention networks, can model the relationships among tokens in a sequence and capture the contextual representation of a given token, with an outstanding ability to capture long-range dependencies. However, self-attention networks have an inherent limitation in that they cannot capture the sequential order of the input tokens [55]–[57]. Therefore, to enable the model to exploit the order of tokens, we must inject some information about the position of tokens in the sequence.

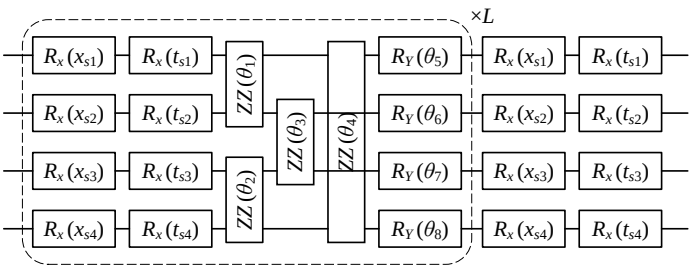


Fig. 4. Introducing Positional Encoding in Quantum Embedding Circuits.

For quantum position encoding, using the entanglement property of quantum systems for encoding is an effective strategy. Ref. [58] proposes a method in which classical positional information can be encoded onto one or more additional qubits through quantum embedding. Subsequently, a trainable Quantum Neural Network (QNN) circuit fully entangles the qubits with positional information with the data qubits, allowing the output quantum state of the entire system to contain positional information. However, this method requires additional quantum qubits resources. In this study, we adopt a different approach, as shown in Fig. 4. We designed a quantum circuit that eliminates the need for auxiliary qubits by sacrificing circuit depth, thereby saving quantum qubits resources. We introduce more quantum gate operations into the quantum circuit to achieve effective encoding of positional information.

Inspired by the Ref. [16], we introduce the sinusoidal positional encoding into the quantum circuit. In classical

sinusoidal positional encoding, the values of the position vector corresponding to the position at even and odd positions are:

$$\begin{aligned} PE_{s,2i} &= sin(s/10000^{2i/d_{\text{model}}}) \\ PE_{s,2i+1} &= cos(s/10000^{2i/d_{\text{model}}}) \end{aligned} \tag{18}$$

where s is the position, i is the dimension, and the positional encoding has the same dimension d_{model} as the embedding.

We scale the positional encoding data to the range $[0, 2\pi]$:

$$\boldsymbol{t}_s = \frac{\boldsymbol{PE}_s - PE_{min}}{PE_{max} - PE_{min}} \times 2\pi \tag{19}$$

For the input data $\boldsymbol{x}$, we scale its range to $[0, \pi]$:

$$\boldsymbol{x}'_s = \frac{\boldsymbol{x}_s - x_{min}}{x_{max} - x_{min}} \times \pi \tag{20}$$

where x_{min} and x_{max} denote the minimum and maximum values found across all elements in all vectors within the input data set, and PE_{min} and PE_{max} are the minimum and maximum values of the positional encoding data.

Different scaling treatments are applied to the positional encoding $PE_{(s)}$ and the input data $\boldsymbol{x}_s$ due to the characteristics of the qubit rotation gate $R_x(\theta)$. This gate rotates the quantum state around the X axis by an angle θ in a counterclockwise direction, where the effective range of θ is $[0, 2\pi]$, representing a complete cycle. For the periodic $PE_{(s)}$, its period naturally matches the 2π cycle of the R_x gate. Therefore, we ensure that the values of $PE_{(s)}$ are transformed into the $[0, 2\pi]$ range through appropriate scaling to be consistent with the operational cycle of the R_x gate. However, for the original input data $\boldsymbol{x}_s$, since it does not possess periodicity, scaling it directly to $[0, 2\pi]$ might result in unintended changes in physical properties. Specifically, the encoded quantum state $R_x(2\pi - x'_s) = -ZR_x(x'_s)Z$ demonstrates a specific transformation relationship, which does not exist in the original data $\boldsymbol{x}_s$. Therefore, this step emphasizes the specific considerations that need to be taken when handling different types of data in the quantum encoding process, to ensure that the physical significance and the encoding effectiveness of the data are appropriately reflected. Through this refined data processing and encoding strategy, we are able to more effectively transform classical information into states in a quantum circuit.

D. Loss Function

We train our model on the dataset $\mathcal{D} = \{(\boldsymbol{x}_{m;1}, \boldsymbol{x}_{m;2}, \ldots, \boldsymbol{x}_{m;S_m}), y_m\}_{m=1}^{N_s}$ by minimizing the loss function, where N_s represents the total number of samples, and the label $\bar{y}_m \in \{0, 1\}$ for each sample indicates its category.

For each sample, S_m denotes the number of words it contains, and each input data $\boldsymbol{x}_{m,s}$ is an n-dimensional vector.

The feature vector for each sample is obtained by summing and averaging the outputs $y_{m,s}$, where $1 \leq m \leq N_s$ and $1 \leq s \leq S_m$:

$$\boldsymbol{y}_m = \frac{1}{S_m} \sum_{s=1}^{S_m} \boldsymbol{y}_{m,s} \tag{21}$$

The output $\boldsymbol{y}_m$ of each sample is input into a fully connected layer to produce the binary prediction value $\hat{y}_m$ for each sample:

$$\hat{y}_m := \sigma(\boldsymbol{w}^\top \cdot \boldsymbol{y}_m + b) \tag{22}$$

where $\boldsymbol{w}$ and b represent the weight and bias of the fully connected layer, respectively, and σ denotes the sigmoid activation function.

For classification problems, there are many loss functions to choose from, such as cross-entropy loss and mean squared error (L2 loss). In the current work, we use the simple and effective mean squared error as the loss function:

$$\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{w}, b; \mathcal{D}) = \frac{1}{2N_s} \sum_{m=1}^{N_s} (\hat{y}_m - \bar{y}_m)^2 \tag{23}$$

where $\boldsymbol{\Theta}$ represents all trainable parameters in the ansatz.

Algorithm 1 QMSAN training algorithm.

Input: Batch sizes **BS**. Number of words per sample S_m. Learning rate η. Number of quantum embedding Layers L. Number of qubits n. The scaled position encodings t_s. The scaled training data set $\mathcal{D} = (\boldsymbol{x'}_{m;1}, \boldsymbol{x'}_{m;2}, \ldots, \boldsymbol{x'}_{m;S_m}), y_m\}_{m=1}^{N_s}$.
$\boldsymbol{\Theta} \sim \mathcal{N}(0, 0.01)$, $\boldsymbol{w} \sim \mathcal{N}(0, 0.01)$, $b \leftarrow 0$

repeat
  for m from 1 to **BS do**
    for s from 1 to S_m **do**
      $|x_{s,q}\rangle \leftarrow U_{emb}(\boldsymbol{x}'_s, \boldsymbol{t}_s, \boldsymbol{\theta}_q, L)\,|0\rangle^{\otimes n}$
      $|x_{s,k}\rangle \leftarrow U_{emb}(\boldsymbol{x}'_s, \boldsymbol{t}_s, \boldsymbol{\theta}_k, L)\,|0\rangle^{\otimes n}$
      $|x_{s,v}\rangle \leftarrow U_{emb}(\boldsymbol{x}'_s, \boldsymbol{t}_s, \boldsymbol{\theta}_v, L)\,|0\rangle^{\otimes n}$
      $\rho_{s,q} \leftarrow \mathrm{tr}_B(|x_{s,q}\rangle\langle x_{s,q}|)$
      $\sigma_{s,k} \leftarrow \mathrm{tr}_B(|x_{s,k}\rangle\langle x_{s,k}|)$
      $\boldsymbol{v}_s \leftarrow \begin{bmatrix}\langle Z_1\rangle_s & \langle Z_2\rangle_s & \cdots & \langle Z_n\rangle_s\end{bmatrix}^\top$
      $\boldsymbol{y}_{m,s} \leftarrow \mathrm{QAttention}(\rho_{s,q}, \sigma_{s,k}, \boldsymbol{v}_s)$
    end for
    $\boldsymbol{y}_m = \frac{1}{S_m} \sum_{s=1}^{S_m} \boldsymbol{y}_s$
    $\hat{y}_m := \sigma(\boldsymbol{w}^\top \cdot \boldsymbol{y}_m + b)$
  end for
  $\mathcal{L} \leftarrow \frac{1}{2N_s} \sum_{m=1}^{N_s} (\hat{y}_m - \bar{y}_m)^2$
  $\boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta} - \eta \nabla_{\boldsymbol{\Theta}} \mathcal{L}$
  $\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \nabla_{\boldsymbol{w}} \mathcal{L}$
  $b \leftarrow b - \eta \nabla_b \mathcal{L}$
until converged

Output: Optimal parameters $\boldsymbol{\Theta}^*$,$\boldsymbol{w}^*$,b^*

IV. NUMERICAL EXPERIMENTS

In this section, we present simulation experiments for text classification tasks conducted on the Tensorcircuit platform [59].These experiments aimed to evaluate our model's effectiveness in processing various types and sizes of text data, as well as its robustness to noise perturbations. For this purpose, we selected two types of publicly available datasets to validate the efficacy of our model. The first type consists of simple sentences or phrases, aiming to evaluate the model's performance in understanding sentence meaning and grammatical

structure. This type includes the MC and RP datasets [60]. The second type, known as the Sentiment Labelled Sentences Data Set [61], comprises three subsets of real-world user reviews. It is primarily used to assess the model's ability to analyze sentence sentiment, including the Yelp, IMDb, and Amazon datasets. The study first evaluated models without position encoding information on both types of datasets. Subsequently, an in-depth analysis was conducted on models integrated with position encoding information on the second type of dataset. Considering that it is impossible to completely eliminate noise in practical noisy intermediate-scale quantum (NISQ) devices, the model is affected by various errors. This research simulated the impact of three types of noise: depolarizing, phase-damping, and amplitude-damping on model results in a simulation environment to evaluate the model's robustness to noise.

A. Datasets

- In the MC (meaning classification) task, there are 130 sentences (70 train + 30 development + 30 test), each with 3 or 4 words. Half of the sentences are related to food, and the other half to information technology (IT). The task's vocabulary consists of 17 words, with some words shared between the two categories, making this task challenging [60].
- In the RP (RELPRON) task, there are 105 noun phrase sentences containing relative clauses (74 train + 31 test), each with 4 words. The task's vocabulary has 115 words, and the selection of phrases ensures that each word appears at least three times in the dataset. The goal is to determine whether a noun phrase contains a subject or object relative clause. Compared to the MC task, the larger vocabulary and resulting word sparsity make this task a more challenging benchmark [60].
- The Sentiment Labelled Sentences dataset comprises reviews from three websites: Amazon, IMDb, and Yelp, with each website providing 1000 sentences labeled with sentiments. The sentences in this dataset have an average length of 10.2 words, with the shortest sentence being 1 word and the longest being 30 words. The IMDb dataset contains movie reviews from the imdb.com website, with an average sentence length of 14.4 words. The shortest sentence in this dataset is 1 word, and the longest is 71 words. The Yelp dataset consists of restaurant reviews, with an average sentence length of 10.9 words, a minimum of 1 word, and a maximum of 32 words. In all datasets, reviews with scores of 4 and 5 are considered positive, while those with scores of 1 and 2 are considered negative. In each subset, positive and negative sentiment reviews each account for 50%. The datasets randomly select 80% of the data as the training set and the remaining 20% as the test set.

B. Experiment Settings

To fairly compare our experimental results with those in Ref. [38], we set the number of qubits for the input of quantum embeddings to be the same as theirs. Specifically, we used $n = 2$ qubits for the MC task and $n = 4$ qubits for the other tasks. The detailed hyperparameter settings are shown in Table I.

TABLE I
EXPERIMENTAL CONFIGURATION. 'LR' DENOTES LEARNING RATE.

Data set	n	L	LR-NP			LR-P		
			NN-NP	CB-NP	AA-NP	NN-P	CB-P	AA-P
MC	2	1	0.005	0.006	0.009	/	/	/
RP	4	2	0.002	0.05	0.01	/	/	/
IMDb	4	1	0.008	0.008	0.01	0.008	0.008	0.008
Yelp	4	1	0.007	0.007	0.03	0.03	0.03	0.01
Amazon	4	1	0.08	0.009	0.09	0.02	0.01	0.02

To explore the potential advantages of our qubit topology in QMSAN, we compared three distinct entanglement schemes within the quantum embeddings module: nearest-neighbor (NN) ansatz, circuit-block (CB) ansatz [62], and all-to-all (AA) ansatz [63], as shown in Fig. 2. The NN ansatz involves a linear arrangement of two-qubit operations within a sequence of qubits, providing a balance between entanglement strength and circuit complexity. The CB ansatz is characterized by a looped arrangement of qubits, suitable for efficient closed-circuit operations. In contrast, the AA ansatz is based on a fully connected network of qubits, allowing for direct interactions between any pair of qubits and resulting in enhanced entanglement potential.

For clarity, we denote the QMSAN variants with these entanglement schemes as QMSAN-NN, QMSAN-CB, and QMSAN-AA, respectively. Additionally, we introduce suffixes to indicate the presence or absence of our novel quantum position encoding: '-P' for models with quantum position encoding and '-NP' for models without it. For example, QMSAN-NN-P denotes QMSAN model with NN ansatz and position encoding, while QMSAN-NN-NP denotes the variant without position encoding.

Furthermore, we propose a variant called QMSAN-D2pi, where the input data is scaled to the range $[0, 2\pi]$. This variant serves as a comparison model to assess the impact of different scaling approaches on the performance of our quantum embeddings. Similarly, we propose the Quantum Pure-State Attention Network (QPSAN) to compare different computational methods for quantum queries and keys. QPSAN employs a distinct method for computing the similarity between quantum queries and keys by utilizing the inner product of pure states. Apart from this, QPSAN and QMSAN maintain identical structures. To facilitate comparisons, we adopt the same naming convention for QPSAN and QMSAN-D2pi variants, such as QPSAN-NN-P and QMSAN-D2pi-NN-P.

Assuming the quantum embeddings ansatz has n qubits and L layers, the AA ansatz has $3n((n-1)/2+n)L$ parameters, while both the NN and CB ansatzes have $6nL$ parameters. The ansatzes for queries, keys, and values have the same depth.

Notably, the AA ansatz has an increased total number of two-qubit gates compared to the NN and CB ansatzes, providing it with stronger entanglement capabilities.

All ansatzes parameters $\mathbf{\Theta}$ and the weights $\boldsymbol{w}$ of the classical part of the network are initialized from a normal distribution with mean 0 and variance 0.1. The bias b of the classical network is initialized to 0. For the value ansatz, we measure the expectation value under the Pauli-Z observable. For the attention matrix, we measure the probability of the output state $|0\rangle$ under the Pauli-Z observable. Through these measurements, we convert quantum data into classical data that can be utilized by subsequent classical networks. In our work, we use the Tensorcircuit framework [59] for simulating quantum circuits and the Tensorflow [64] framework for parameter optimization, with the optimizer Adam [65]. The batch size is 64, and training stops when convergence is reached or after a fixed number of epochs. In the MC and RP tasks, we repeat each experiment 9 times with different parameter initializations. For the Sentiment Labelled Sentences Data Set task, we use cross-validation for the experiments.

C. Experiments with Non-Positional Models

In this experiment, we conduct numerical experiments with two model architectures: Non-Positional Models and Positional Models. This allows us to deeply analyze the performance differences between models and explore how positional information enhances the model's understanding and processing of data. The experimental design aims to compare with representative models in classical networks and quantum networks. For the MC and RP datasets, we compare our experimental results with a classical quantum model based on syntactic analysis [60] and further with QSANN. Meanwhile, for three public sentiment analysis datasets (Yelp, IMDb, and Amazon), our models will be compared and analyzed with classical self-attention neural networks (CSANN) [38] and QSANN models. Through these comparative experiments, we aim to comprehensively evaluate the performance and potential of quantum self-attention networks across different models and datasets.

a) MC dataset: We first conducted experiments on models without positional information, and the results are detailed in Table II. On the MC dataset, the QMSAN-NP series models, regardless of the entanglement structure used, can perfectly distinguish sentences related to food and information technology, significantly outperforming the quantum model based on syntactic analysis mentioned in reference [60]. Since the task of the MC dataset is relatively simple, different entanglement structures can effectively capture the semantic information of sentences in this case. Notably, in terms of the number of parameters, QMSAN-NP series models use fewer parameters compared to the QSANN and classical deep learning model DisCoCat, yet achieve the same performance level, indicating the potential application value of QMSAN-NP series models in resource-limited environments.

b) RP dataset: For the RP dataset, we evaluate the model's performance in processing more complex data. The experimental results Table II show different degrees of performance on different variants of the QMSAN-NP series models, with the QMSAN-AA-NP model performing the best with a test accuracy of 77.42%. The experimental results of the other two entanglement structures are reduced, which may be because the fully connected structure provides more abundant information entanglement, helping the model better capture complex relationships in sentences. The experimental results of different series models of QMSAN-NP all surpass the 72.30% test accuracy of the quantum machine learning model QSANN based on syntactic analysis and significantly outperform the 67.74% of the QSANN model. Although the complexity of the RP dataset is significantly higher than that of the MC dataset, our model can still effectively parse and infer complex relationships in the dataset, further confirming the great potential of quantum self-attention networks in enhancing the model's understanding ability. Specifically, our QMSAN-NN-NP and QMSAN-CB-NP maintain a lower number of parameters and computational cost while improving performance, indicating the potential of QMSAN models in handling natural language processing tasks with certain complexity.

Concurrently, we conducted a comparative analysis of the experimental outcomes for QMSAN-NP and QPSAN-NP on this dataset. The QMSAN-NP series models' performance on the RP dataset is superior to that of the QPSAN-NP series models, with QMSAN-AA-NP model achieving the highest test accuracy of 77.42%, 3.23% higher than the highest test accuracy of 74.19% of the QPSAN-NP series models. The specific results are shown in Table II.This indicates that QMSAN model can more accurately capture the similarity features between queries and keys using the mixed state approach, especially in processing complex semantic relationships. This validates our theoretical analysis in Section III-B.

c) Sentiment Labelled Sentences Data Set: This part of the research focuses on the Sentiment Labelled Sentences Data Set, covering more complex and diverse sentiment analysis tasks. The Sentiment Labelled Sentences Data Set, which includes three subsets: Yelp, IMDb, and Amazon, covers different review categories, each with its unique language usage and emotional expression, increasing the difficulty and complexity of sentiment analysis and providing a broader and more challenging testing platform for our QMSAN-NP series models. The experimental results are detailed in Table III.

Compared to classical model CSANN, the QMSAN-NP series models' methods have achieved comprehensive improvements in accuracy, with the maximum being 4.45% on IMDB for QMSAN-NN-NP. Compared to the quantum model QSANN, the accuracy has significantly improved on most datasets, with the maximum increase of 3.84% on the IMDb dataset and 2.47% on the Amazon dataset. The accuracy on the Yelp dataset is also similar, indicating that QMSAN-NP series models, by using mixed-state quantum attention calculation and trainable embedded quantum modules, can more effectively capture the intrinsic features and complex relationships of data, providing a richer data representation

TABLE II
TEST ACCURACY OF OUR PROPOSED METHODS COMPARED TO DISCOCAT AND CSANN ON MC AND RP TASK.

Method	MC			RP		
	#Paras	TrainAcc(%)	TestAcc(%)	#Paras	TrainAcc(%)	TestAcc(%)
DisCoCat [60]	40	83.10	79.80	168	90.60	72.30
QSANN [38]	25	100.00	100.00	109	95.35	67.74
QPSAN-NN-NP	15	100.00	100.00	53	95.95	70.97
QPSAN-CB-NP	15	100.00	100.00	53	95.95	70.97
QPSAN-AA-NP	18	100.00	100.00	137	95.95	74.19
QMSAN-NN-NP	15	100.00	100.00	53	95.95	74.19
QMSAN-CB-NP	15	100.00	100.00	53	95.95	74.19
QMSAN-AA-NP	18	**100.00**	**100.00**	137	**97.30**	**77.42**

TABLE III
TEST ACCURACY OF OUR PROPOSED METHODS COMPARED TO CSANN AND THE NAIVE METHOD ON YELP, IMDB, AND AMAZON DATA SETS.

Method	Yelp			IMDb			Amazon		
	#Paras	TrainAcc(%)	TestAcc(%)	#Paras	TrainAcc(%)	TestAcc(%)	#Paras	TrainAcc(%)	TestAcc(%)
CSANN [38]	785	/	83.11 ± 0.89	785	/	79.67 ± 0.83	785	/	83.22 ± 1.28
QSANN [38]	49	/	84.79 ± 1.29	49	/	80.28 ± 1.78	61	/	84.25 ± 1.75
QMSAN-NN-NP	29	99.53 ± 0.22	84.14 ± 2.27	29	99.48 ± 0.37	84.12 ± 2.31	29	99.80 ± 0.10	86.72 ± 2.38
QMSAN-CB-NP	29	99.58 ± 0.23	84.40 ± 1.98	29	99.45 ± 0.24	83.74 ± 2.01	29	99.83 ± 0.17	86.61 ± 1.71
QMSAN-AA-NP	71	99.65 ± 0.18	84.73 ± 2.34	71	99.50 ± 0.40	83.76 ± 3.04	71	99.75 ± 0.18	86.56 ± 1.90
QMSAN-NN-P	29	99.45 ± 0.32	84.85 ± 1.33	29	99.18 ± 0.41	84.77 ± 3.12	29	99.87 ± 0.94	87.41 ± 1.16
QMSAN-CB-P	29	99.80 ± 0.20	84.82 ± 1.21	29	99.18 ± 0.41	$\mathbf{84.82 \pm 2.96}$	29	99.90 ± 0.93	87.43 ± 1.16
QMSAN-AA-P	71	99.55 ± 0.26	$\mathbf{84.96 \pm 3.34}$	71	99.33 ± 0.36	84.29 ± 2.32	71	99.91 ± 0.50	$\mathbf{87.48 \pm 1.02}$

for complex sentiment analysis tasks, thereby improving the model's representational ability. It can effectively adapt to sentiment classification tasks in different domains and data distributions, not only significantly improving performance in most datasets but also maintaining stable performance in different application scenarios, demonstrating better generalization ability.

For the number of parameters, QMSAN-NN-NP and QMSAN-CB-NP have 29 parameters, significantly fewer than CSANN's 785 and QSANN's 49, yet they show noticeable performance improvements on most datasets. Being able to capture sufficient information with fewer parameters is very beneficial for reducing computational resources and improving efficiency. This indicates that our model has a clear advantage in capturing emotional features in textual data. Among the three QMSAN-NP series models, QMSAN-AA-NP has slightly more parameters and stronger quantum entanglement capabilities, but it only has a slightly higher test accuracy on the Yelp dataset compared to the other two models. This suggests that these more complex different datasets may require different entanglement methods.

D. Experiments with Positional Models

To further enhance model performance, this section of the experiment focuses on analyzing the impact of positional information on QMSAN model. Considering the importance of positional information in text sequence processing, we adopted the use of fixed quantum gates to encode positional information. Since the average number of words per sentence in the MC and RP datasets is relatively small, the impact of positional information on model performance may not be significant. Therefore, we chose to evaluate the performance of the QMSAN series models with introduced positional information on the Yelp, IMDb, and Amazon data subsets to more accurately measure the impact of positional information. Experiments show that the QMSAN-P series models with fixed positional information have comprehensively improved performance compared to models without positional information. The experimental results are detailed in Table III, with the maximum increase in test accuracy of 0.71% on the Yelp dataset, 1.08% on the IMDb dataset, and 0.92% on the Amazon dataset.

The above experimental results show that our method can effectively encode positional information in text sequences, comprehensively improving the accuracy of QMSAN model in different sentiment analysis tasks. By using fixed quantum gates with positional information, the model is provided with key information about the relative positions of words in the text, enhancing the understanding of the entire text structure. The advantage of this method is that it provides a fixed and efficient way for quantum models to incorporate positional information. We do not need to increase additional qubits resources and do not need to add trainable parameters, which can save computational resources and time during training. Moreover, since a stable and consistent way is adopted to represent the positional information of words, this helps the model to generalize better to unseen data.

At the same time, we compared the experiments of the QMSAN-D2pi-P series models, which scale the input data to $[0, 2\pi]$. The results, as shown in Fig. 5, indicate that

TABLE IV
TEST ACCURACY OF QMSAN-P SERIES MODELS ON YELP, IMDB, AND AMAZON DATA SETS WITH DIFFERENT NOISE MODELS.

Noise Model	Yelp			IMDb			Amazon		
	NN-P	CB-P	AA-P	NN-P	CB-P	AA-P	NN-P	CB-P	AA-P
DP(0.01)	84.55±1.79	84.51±1.38	84.77±1.81	84.10±3.14	84.27±2.16	83.89±3.64	86.97±1.66	86.98±1.53	86.40±1.16
DP(0.1)	84.49±1.59	84.13±2.27	84.11±2.35	84.30±1.81	84.03±2.07	83.58±1.58	87.02±0.71	86.87±1.21	87.23±1.40
DP(0.2)	83.97±1.74	83.76±3.66	84.22±2.34	83.29±3.04	83.80±1.63	83.83±2.66	86.17±1.53	86.05±1.97	86.33±1.50
AD(0.01)	84.53±2.28	83.99±0.97	84.63±1.39	84.01±3.02	84.17±3.12	83.77±2.42	86.42±1.02	87.34±1.17	87.21±1.81
AD(0.1)	84.05±2.30	84.15±2.01	84.54±1.87	84.09±2.59	83.98±3.89	83.69±1.69	86.30±0.51	87.29±0.81	86.99±2.25
AD(0.2)	83.87±1.40	83.67±1.71	84.20±1.86	83.95±2.54	84.07±2.70	83.91±1.71	86.33±2.29	86.23±1.63	87.30±1.72
PD(0.01)	84.11±2.56	84.60±1.83	83.90±1.53	84.02±2.79	84.02±2.79	84.05±2.43	87.02±0.95	86.78±0.75	87.05±1.05
PD(0.1)	84.35±2.73	84.22±2.82	83.42±1.32	84.22±2.73	83.93±2.08	83.94±3.22	86.37±1.96	86.32±1.96	86.86±2.25
PD(0.2)	84.33±1.44	84.46±1.24	83.85±1.33	83.66±2.58	84.11±2.99	83.71±4.06	86.91±2.15	86.83±1.69	86.89±2.36

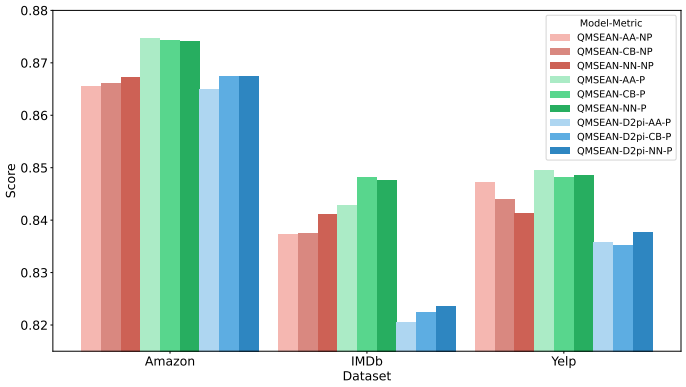


Fig. 5. Test accuracy of different forms of data scaling methods.

the test accuracy of the QMSAN-D2pi-P series models is comprehensively lower than that of the QMSAN-P series models, and even lower than the performance of QMSAN-NP without positional information on some datasets. This also validates the theoretical analysis in Section III-C.

E. Noise robustness

In the practical application of quantum computing, the impact of quantum noise is a significant factor, as NISQ is sensitive to the environment and susceptible to noise interference. To evaluate the robustness of our QMSAN-P series models in a quantum noise environment, we conducted a series of experiments using the Tensorcircuit simulation software. We considered not only common noise channels such as depolarizing noise, amplitude damping noise, and phase damping noise but also explored the impact of different Ansatz structures on noise resistance.

Depolarizing channel (DP) causes a qubit to depolarize with probability p. For a single qubit, it is replaced by the completely mixed state $I/2$, and remains unchanged with probability 1-p. The depolarizing channel can be represented as the following density matrix mapping:

$$\varepsilon_{\mathrm{DP}}(\rho) = (1-p)\rho + \frac{p}{3}(X\rho X + Y\rho Y + Z\rho Z) \tag{24}$$

where ρ is the original density matrix, and X, Y, Z are Pauli matrices.

Amplitude damping (AD) describes the process of a quantum system losing energy, while phase damping (PD) describes the process of a quantum system losing phase information without losing energy. The noise mapping for a single qubit's density matrix can be uniformly expressed as:

$$\varepsilon_{\mathrm{AD/PD}}(\rho) = E_0\rho E_0^\dagger + E_1\rho E_1^\dagger \tag{25}$$

where for amplitude damping, $E_0 = |0\rangle\langle 0| + \sqrt{1-p}|1\rangle\langle 1|$ and $E_1 = \sqrt{p}|0\rangle\langle 1|$, and for phase damping, $E_0 = |0\rangle\langle 0| + \sqrt{1-p}|1\rangle\langle 1|$ and $E_1 = \sqrt{p}|1\rangle\langle 1|$. E_0 and E_1 represent Kraus operators, and p represents the noise level.

We added these single-qubit noise channels to the embedding layer circuit for noise addition. At noise levels of 0.01, 0.1, and 0.2, the experimental results are shown in Table IV. The performance of the QMSAN-P series models showed a slight decline, with the maximum accuracy drop of 1.54% on the Yelp dataset under the PD(0.1) noise model. The maximum accuracy drop on the IMDb dataset was 1.48% under the DP(0.2) noise model, and the maximum accuracy drop on the Amazon dataset was 1.38% under the DP(0.2) noise model. The decrease in test accuracy caused by noise did not exceed 1.6%, indicating that the QMSAN series models can maintain high performance stability in a low-level quantum noise environment, showing good robustness to common quantum noise, and validating the feasibility of running QMSAN models in a real quantum computing environment.

V. CONCLUSION

This paper proposes a novel Quantum Multi-head Self-Attention Network (QMSAN) model, combining the characteristics of quantum computing with the advantages of classical self-attention networks to enhance the processing capabilities and efficiency of NLP tasks. Our model operates on queries and keys under mixed states through quantum gate operations and directly generates similarity scalars through measurement. Compared to conventional pure state unitary transformations, our method expands the expressive power of the quantum system through mixed state operations, enabling the model to capture the similarity between queries and keys more comprehensively and finely. We also introduced a trainable quantum embedding module that maps classical data to quantum states, achieving more efficient data representation and

processing. Additionally, we proposed a novel quantum positional encoding scheme, which encodes positional information by introducing additional fixed quantum gates in the quantum circuit, improving the accuracy of encoding without increasing additional qubits resources.

The experimental results on various datasets verify the effectiveness of QMSAN. Compared to classical self-attention networks, our model not only significantly reduces the number of parameters under the same input sequence conditions but also demonstrates superior performance, proving its better learning ability. We anticipate that future work will further explore the potential of quantum machine learning models, realizing a fully quantum self-attention network model, and fully utilizing the unique advantages of quantum computing. It serves as a scalable module, facilitating the construction of quantum versions of the Transformer architecture, thereby bringing unprecedented computational power and efficiency to machine learning.

REFERENCES

[1] Neill Lambert, Yueh-Nan Chen, Yuan-Chung Cheng, Che-Ming Li, Guang-Yin Chen, and Franco Nori. Quantum biology. *Nature Physics*, 9(1):10–18, 2013.

[2] Feihu Xu, Xiongfeng Ma, Qiang Zhang, Hoi-Kwong Lo, and Jian-Wei Pan. Secure quantum key distribution with realistic devices. *Reviews of Modern Physics*, 92(2):025002, 2020.

[3] Nicolas Gisin and Rob Thew. Quantum communication. *Nature photonics*, 1(3):165–171, 2007.

[4] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017.

[5] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. An introduction to quantum machine learning. *Contemporary Physics*, 56(2):172–185, 2015.

[6] Maria Schuld and Nathan Killoran. Quantum machine learning in feature hilbert spaces. *Physical review letters*, 122(4):040504, 2019.

[7] Peter Wittek. *Quantum machine learning: what quantum computing means to data mining*. Academic Press, 2014.

[8] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[9] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[10] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[11] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[12] Yan Cheng, Leibo Yao, Guoxiong Xiang, Guanghe Zhang, Tianwei Tang, and Linhui Zhong. Text sentiment orientation analysis based on multi-channel cnn and bidirectional gru with attention mechanism. *IEEE Access*, 8:134964–134975, 2020.

[13] Abdalraouf Hassan and Ausif Mahmood. Convolutional recurrent deep learning model for sentence classification. *Ieee Access*, 6:13949–13957, 2018.

[14] Shujie Liu, Nan Yang, Mu Li, and Ming Zhou. A recursive recurrent neural network for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1491–1500, 2014.

[15] Tianyu He, Xu Tan, Yingce Xia, Di He, Tao Qin, Zhibo Chen, and Tie-Yan Liu. Layer-wise coordination between encoder and decoder for neural machine translation. *Advances in Neural Information Processing Systems*, 31, 2018.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[17] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[18] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[19] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897, 2020.

[20] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194, 2021.

[21] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023.

[22] Fang Liu, Ge Li, Yunfei Zhao, and Zhi Jin. Multi-task learning based pre-trained language model for code completion. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*, pages 473–485, 2020.

[23] Abid Haleem, Mohd Javaid, and Ravi Pratap Singh. An era of chatgpt as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil transactions on benchmarks, standards and evaluations*, 2(4):100089, 2022.

[24] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. Pre-trained models: Past, present and future. *AI Open*, 2:225–250, 2021.

[25] Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. A survey on model compression for large language models. *arXiv preprint arXiv:2308.07633*, 2023.

[26] Shiqiang Zhu, Ting Yu, Tao Xu, Hongyang Chen, Schahram Dustdar, Sylvain Gigan, Deniz Gunduz, Ekram Hossain, Yaochu Jin, Feng Lin, et al. Intelligent computing: the latest advances, challenges, and future. *Intelligent Computing*, 2:0006, 2023.

[27] Man-Hong Yung. Quantum supremacy: some fundamental concepts. *National Science Review*, 6(1):22–23, 2019.

[28] Aram W Harrow and Ashley Montanaro. Quantum computational supremacy. *Nature*, 549(7671):203–209, 2017.

[29] M Cerezo, Guillaume Verdon, Hsin-Yuan Huang, Lukasz Cincio, and Patrick J Coles. Challenges and opportunities in quantum machine learning. *Nature Computational Science*, 2(9):567–576, 2022.

[30] Maria Schuld and Nathan Killoran. Is quantum advantage the right goal for quantum machine learning? *Prx Quantum*, 3(3):030101, 2022.

[31] Maria Schuld. Supervised quantum machine learning models are kernel methods. *arXiv preprint arXiv:2101.11020*, 2021.

[32] Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. A rigorous and robust quantum speed-up in supervised machine learning. *Nature Physics*, 17(9):1013–1017, 2021.

[33] Andrew J Daley, Immanuel Bloch, Christian Kokail, Stuart Flannigan, Natalie Pearson, Matthias Troyer, and Peter Zoller. Practical quantum advantage in quantum simulation. *Nature*, 607(7920):667–676, 2022.

[34] Mina Abbaszade, Vahid Salari, Seyed Shahin Mousavi, Mariam Zomorodi, and Xujuan Zhou. Application of quantum natural language processing for language translation. *IEEE Access*, 9:130434–130448, 2021.

[35] Yanan Li, Zhimin Wang, Rongbing Han, Shangshang Shi, Jiaxin Li, Ruimin Shang, Haiyong Zheng, Guoqiang Zhong, and Yongjian Gu. Quantum recurrent neural networks for sequential learning. *arXiv preprint arXiv:2302.03244*, 2023.

[36] Samuel Yen-Chi Chen, Shinjae Yoo, and Yao-Lung L Fang. Quantum long short-term memory. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8622–8626. IEEE, 2022.

[37] LSTM Bi-Directional. Mechanism for sentence modeling. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part II*, volume 10635, page 178. Springer, 2017.

[38] Guangxi Li, Xuanqiang Zhao, and Xin Wang. Quantum self-attention neural networks for text classification. *arXiv preprint arXiv:2205.05625*, 2022.

[39] Ren-xin Zhao, Jinjing Shi, Shichao Zhang, and Xuelong Li. Qsan: A near-term achievable quantum self-attention network. *arXiv preprint arXiv:2207.07563*, 2022.

[40] Ren-Xin Zhao, Jinjing Shi, and Xuelong Li. Qksan: A quantum kernel self-attention network. *arXiv preprint arXiv:2308.13422*, 2023.

[41] Michael A Nielsen and Isaac L Chuang. *Quantum computation and quantum information*. Cambridge university press, 2010.

[42] Iris Cong, Soonwon Choi, and Mikhail D Lukin. Quantum convolutional neural networks. *Nature Physics*, 15(12):1273–1278, 2019.

[43] Maria Schuld, Alex Bocharov, Krysta M Svore, and Nathan Wiebe. Circuit-centric quantum classifiers. *Physical Review A*, 101(3):032308, 2020.

[44] Mateusz Ostaszewski, Lea M Trenkwalder, Wojciech Masarczyk, Eleanor Scerri, and Vedran Dunjko. Reinforcement learning for optimization of variational quantum circuit architectures. *Advances in Neural Information Processing Systems*, 34:18182–18194, 2021.

[45] Mateusz Ostaszewski, Edward Grant, and Marcello Benedetti. Structure optimization for parameterized quantum circuits. *Quantum*, 5:391, 2021.

[46] Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3):032430, 2021.

[47] Seth Lloyd, Maria Schuld, Aroosa Ijaz, Josh Izaac, and Nathan Killoran. Quantum embeddings for machine learning. *arXiv preprint arXiv:2001.03622*, 2020.

[48] Guangxi Li, Ruilin Ye, Xuanqiang Zhao, and Xin Wang. Concentration of data encoding in parameterized quantum circuits. *Advances in Neural Information Processing Systems*, 35:19456–19469, 2022.

[49] Ajay Shrestha and Ausif Mahmood. Review of deep learning algorithms and architectures. *IEEE access*, 7:53040–53065, 2019.

[50] Yushi Chen, Hanlu Jiang, Chunyang Li, Xiuping Jia, and Pedram Ghamisi. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE transactions on geoscience and remote sensing*, 54(10):6232–6251, 2016.

[51] Patrick J Coles, M Cerezo, and Lukasz Cincio. Strong bound between trace distance and hilbert-schmidt distance for low-rank states. *Physical Review A*, 100(2):022103, 2019.

[52] Juan Carlos Garcia-Escartin and Pedro Chamorro-Posada. Swap test and hong-ou-mandel effect are equivalent. *Physical Review A*, 87(5):052330, 2013.

[53] Hirotada Kobayashi, Keiji Matsumoto, and Tomoyuki Yamakami. Quantum merlin-arthur proof systems: Are multiple merlins more helpful to arthur? In *Algorithms and Computation: 14th International Symposium, ISAAC 2003, Kyoto, Japan, December 15-17, 2003. Proceedings 14*, pages 189–198. Springer, 2003.

[54] Murphy Yuezhen Niu, Alexander Zlokapa, Michael Broughton, Sergio Boixo, Masoud Mohseni, Vadim Smelyanskyi, and Hartmut Neven. Entangling quantum generative adversarial networks. *Physical Review Letters*, 128(22):220505, 2022.

[55] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10033–10041, 2021.

[56] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[57] Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2019.

[58] Chuangtao Chen and Qinglin Zhao. Quantum generative diffusion model. *arXiv preprint arXiv:2401.07039*, 2024.

[59] Shi-Xin Zhang, Jonathan Allcock, Zhou-Quan Wan, Shuo Liu, Jiace Sun, Hao Yu, Xing-Han Yang, Jiezhong Qiu, Zhaofeng Ye, Yu-Qin Chen, et al. Tensorcircuit: a quantum software framework for the nisq era. *Quantum*, 7:912, 2023.

[60] Robin Lorenz, Anna Pearson, Konstantinos Meichanetzidis, Dimitri Kartsaklis, and Bob Coecke. Qnlp in practice: Running compositional models of meaning on a quantum computer. *Journal of Artificial Intelligence Research*, 76:1305–1342, 2023.

[61] Dimitrios Kotzias. Sentiment Labelled Sentences. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C57604.

[62] Sukin Sim, Peter D Johnson, and Alán Aspuru-Guzik. Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. *Advanced Quantum Technologies*, 2(12):1900070, 2019.

[63] Marcello Benedetti, Delfina Garcia-Pintos, Oscar Perdomo, Vicente Leyton-Ortega, Yunseong Nam, and Alejandro Perdomo-Ortiz. A generative modeling approach for benchmarking and training shallow quantum circuits. *npj Quantum Information*, 5(1):45, 2019.

[64] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. {TensorFlow}: a system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283, 2016.

[65] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.