GQHAN: A Grover-inspired Quantum Hard Attention Network

Ren-Xin Zhao, *Member, IEEE,* Jinjing Shi*, *Member, IEEE,* and Xuelong Li, *Fellow, IEEE*

Abstract—Numerous current Quantum Machine Learning (QML) models exhibit an inadequacy in discerning the significance of quantum data, resulting in diminished efficacy when handling extensive quantum datasets. Hard Attention Mechanism (HAM), anticipated to efficiently tackle the above QML bottlenecks, encounters the substantial challenge of non-differentiability, consequently constraining its extensive applicability. In response to the dilemma of HAM and QML, a Grover-inspired Quantum Hard Attention Mechanism (GQHAM) consisting of a Flexible Oracle (FO) and an Adaptive Diffusion Operator (ADO) is proposed. Notably, the FO is designed to surmount the non-differentiable issue by executing the activation or masking of Discrete Primitives (DPs) with Flexible Control (FC) to weave various discrete destinies. Based on this, such discrete choice can be visualized with a specially defined Quantum Hard Attention Score (QHAS). Furthermore, a trainable ADO is devised to boost the generality and flexibility of GQHAM. At last, a Grover-inspired Quantum Hard Attention Network (GQHAN) based on QGHAM is constructed on PennyLane platform for Fashion MNIST binary classification. Experimental findings demonstrate that GQHAN adeptly surmounts the non-differentiability hurdle, surpassing the efficacy of extant quantum soft self-attention mechanisms in accuracies and learning ability. In noise experiments, GQHAN is robuster to bit-flip noise in accuracy and amplitude damping noise in learning performance. Predictably, the proposal of GQHAN enriches the Quantum Attention Mechanism (QAM), lays the foundation for future quantum computers to process large-scale data, and promotes the development of quantum computer vision.

Index Terms—Machine learning, Quantum machine learning, Grover's algorithm, Hard attention mechanism, Grover-inspired quantum hard attention Network, Quantum neural network, Quantum circuit.

1 Introduction

In recent years, QML has developed tremendously [1–3]. However, many current QML models treat each quantum data equally and neglect the value of the intrinsic connections between data. This not only mandates substantial quantum storage resources for comprehensive information retention, but also constitutes a threat to the prospective large-scale quantum data processing on forthcoming quantum computers. The above urgent issues can be effectively addressed by HAM.

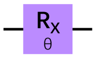
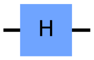
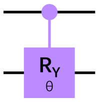
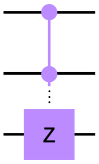
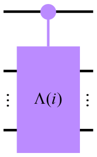
HAM was first proposed in 2015 [4], as a discrete competitive model that adheres to the winner-take-all law which only concentrates on the most vital parts and overlooks the rest. This unique computing mechanism decreases the cost of data acquisition [5] for designing interpretable [6], computationally effective [7] and scalable [8] models, mitigating catastrophic forgetting to a certain extent [9], consequently resulting in widespread applications in computer vision [10], natural language processing [11] and video processing [12, 13], etc. One of the striking cases is that HAM achieves two impressive accuracies of 95.83% and 99.07% for safe driving recognition and driver distraction detection, respectively, in addition to a 38.71% reduction in runtime [14]. Although very effective, HAM is stuck in a non-differentiable dilemma that hinders its optimization due to the discrete nature of information selection processing [15]. To conquer this dilemma, various strategies such as reinforcement learning [11, 12] and Gumbel-softmax based straight-through estimator [16, 17] have been proposed. Nevertheless, formulating a proficient reward function aligned with the task in reinforcement learning proves challenging, and misguided reward specifications may engender suboptimal attention patterns. Straight-through estimator schemes may also require a trade-off between computational efficiency and accuracy. Hence, it is inherently more practical and convenient to make HAM compatible with gradient systems. To achieve this goal, a quantum scheme inspired by Grover's algorithm is attempted to compensate for the above shortcoming.

Grover's algorithm is a quantum algorithm composed of an oracle and a diffusion operator for unstructured search [18, 19], where the oracle selects specific target items through phase flipping. The diffusion operator amplifies the amplitudes of the chosen ones and suppresses the amplitudes of the other items. Obviously, the working mechanism of Grover's algorithm is somehow similar to HAM, as it also involves discretely selecting specific targets. However, the application of Grover's algorithm alone is insufficient to surmount the non-differentiable nature. Simultaneously, when the number of target terms is unknown or dynamically changing, the performance and adaptivity of Grover's algorithm are constrained by the inability to ascertain the appropriate number of iterations [20–23]. Thus a dramatic modification of this algorithm is necessary to create a differentiable quantum hard attention mechanism, which induces

- *Ren-Xin Zhao is with the School of Computer Science and Engineering, Central South Univerisity, China, Changsha, 410083. Jinjing Shi is with the School of Electronic Information, Central South Univerisity, China, Changsha, 410083.*
- *Xuelong Li is with the Institute of Artificial Intelligence (Tele AI), China Telecom Corp Ltd, 31 Jinrong Street, Beijing 100033, P. R. China.*
- *Jinjing Shi is the corresponding author.*
- *E-mails: renxin_zhao@alu.hdu.edu.cn, shijinjing@csu.edu.cn, li@nwpu.edu.cn.*

Manuscript received XXXX; revised XXXX.

Tab. 1: Quantum Gates

Name of Quantum Gate	Mathematical Notation	Matrix Representation	Symbol of Quantum Gate
Pauli X gate	X	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$	
Pauli rotating X gate	$R_X(\theta)$	$\begin{bmatrix} \cos(\theta/2) & -i\sin(\theta/2) \\ -i\sin(\theta/2) & \cos(\theta/2) \end{bmatrix}$	
Hadamard gate	H	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$	
Controlled Y gate	$CR_Y(\theta)$	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos(\theta/2) & -\sin(\theta/2) \\ 0 & 0 & \sin(\theta/2) & \cos(\theta/2) \end{bmatrix}$	
Multi-controlled Z gate	MCZ	$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -1 \end{bmatrix}$	
Discrete primitive	$C-\Lambda(b)$	Eq. (14)	

the three main propositions of this paper:

1. Can current QML models be equipped with a quantum HAM to distinguish the intrinsic importance of quantum data?

2. How can a differentiable GQHAN be constructed by combining the Gover's algorithm with HAM?

3. How does GQHAN remain efficient when the target term is unknown or dynamically altering?

To this end, the **contribution** of this paper lies in addressing these pertinent inquiries and presenting novel insights into the potential of GQHAN:

- GQHAM is posited to discern the significance of quantum data, thereby augmenting the efficacy of QML model processing.
- FO and ADO are proposed as solutions for addressing the discrete non-differentiable predicament inherent in GQHAM.
- Based on GQHAM, a GQHAN is constructed on PennyLane for Fashion MNIST binary classification experiments, achieving an impressive accuracy of no less than 98% and lower convergence values, which indicates that it has surpassed the two quantum soft attention mechanisms in terms of classification accuracy and learning ability.

The subsequent sections of this paper are organized as follows: Section 2 contains a brief overview of QAM, quantum computing and Grover's algorithm. In Section 3, the mathematical mechanism of GQHAM is elaborated.The GQHAN framework and its workflow are described in Section 4. Section 5 delineates the experimental details and several meaningful conclusions are drawn. Eventually, a summary is presented.

2 Related Works

In this section, a succinct overview of quantum computing foundation, QAM, and Grover's algorithm is presented.

2.1 Quantum Computing Foundation

Qubits and quantum gates are the figurative embodiment of quantum theory. In this context, a qubit $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ is the smallest unit that carries information, where $|0\rangle = [1,0]^{\mathrm{T}}$ and $|1\rangle = [0,1]^{\mathrm{T}}$ denote the ground state and the excited state respectively. α and β are amplitudes satisfying $|\alpha|^2 + |\beta|^2 = 1$ [28]. In stark contrast to classical computation, $|\psi\rangle$ is in a superposition of $|0\rangle$ and $|1\rangle$, thereby conferring upon it the remarkable ability for exponential data representation. Moreover, the linear evolution of qubits is contingent upon quantum gates whose mathematical essence is the unitary matrices. The quantum gates used in this paper are shown in Tab. 1, including Pauli X gate, Pauli rotating X gate, Hadamard gate, controlled Y gate, multi-controlled Z gate and DP, where DP is one of the innovations in this paper.

2.2 Quantum Attention Mechanisms

In QML, QAM attempts to boost the performance of QML models by discriminating the significance of quantum data. Early researches on QAM are inspired by quantum physics principles. For example, a parameter-free QAM utilizing weak measurements was introduced in 2017 to enhance bidirectional LSTM sentence modeling, demonstrating superior performance compared to classical attention mechanism [24]. Nonetheless, these investigations exclusively

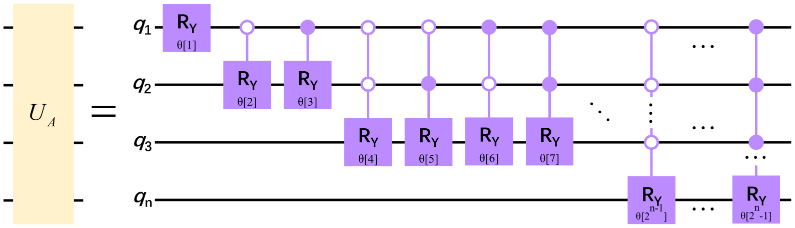


Fig. 1: Quantum Amplitude Encoding [32]

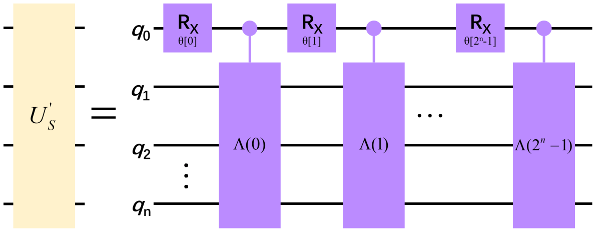


Fig. 2: Flexible Oracle

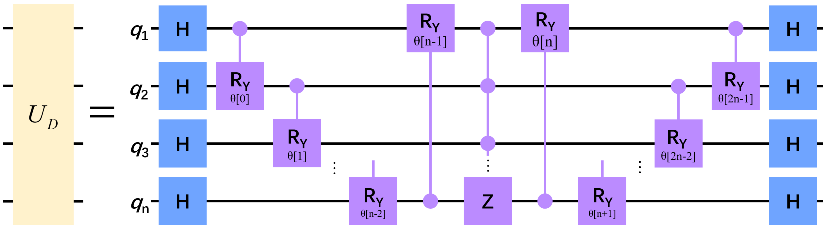


Fig. 3: Adaptive Diffusion Operator

depended on classical computer simulations owing to the absence of an ansatz framework. QAM with ansatz structure has not appeared until 2022. In 2022, a probabilistic full-quantum self-attention network with an exponentially scalable self-attention representation realized completely on a quantum computer was proposed [25]. Its learning ability outperforms hardware-efficient ansatz and QAOA ansatz in a comparable configuration. In the same year, a quantum self-attention network was designed by a Baidu team to calculate quantum self-attention scores using hardware-efficient ansatz with remarkable success in text categorization [26]. In 2023, a quantum self-attention network based on the quantum kernel method and deferred measurement principle was introduced, which achieved impressively high accuracy with few parameters [27]. While QAM has exhibited considerable promise in domains like quantum machine vision, its theoretical underpinnings and practical applications necessitate further augmentation owing to the constraints posed by quantum hardware.

2.3 Grover's Algorithm

Grover's algorithm has undergone iterative enhancements in multiple aspects such as phase and initial state optimization, since its initial proposal, where phase optimization involves strategically adjusting the phase flip in the oracle. Illustratively, in a fixed phase rotation Grover's algorithm, the rotation phase shifts from π to 1.91684π, thereby elevating the success probability to 99.59% [29]. Initial state optimization allows for arbitrary initial states, markedly amplifying the generality of Grover's algorithm. Some reasons for this are that the mean and variance of the amplitude distribution of the initial state affects the measurement of the optimal time and the maximum probability of success, but the measurement of the optimal time is independent of the initial state which is an arbitrary pure state [30, 31]. Although numerous efforts have been conducted to boost the efficiency of Grover's algorithm, it has not been explored to incorporate trainable parameters to further strengthen its flexibility and applicability. In other words, in this paper, an attempt is presented to convert the untrainable Grover's algorithm into a trainable QML model to distinguish it from previous improvement strategies.

3 Grover-inspired Quantum Hard Attention Mechanism

In this section, GQHAM, which successfully mitigates the non-differentiability issue, is presented and defined as follows:

Definition 3.1 (Grover-inspired Quantum Hard Attention Mechanism)**.**

$$\text{GQHAM} := U_D U_S^{'} |\mathbf{In}\rangle_2, \tag{1}$$

is composed of an input quantum state $|\mathbf{In}\rangle_2$, an ADO U_D, and an FO $U_S^{'}$ that contains three novel concepts of DP, FC and QHAS.

The tenets of the FO and the ADO are elucidated emphatically in the ensuing subsections.

3.1 Flexible Oracle

The classical input is set as

$$\mathbf{In} = [a_b]_{b=0}^{l-1} \in \mathbb{R}^{1\times l} \subseteq \mathcal{X}, \tag{2}$$

where a_b indicates an element. l is the total number of elements. $\mathcal{X}$ denotes the space where the input data is located. Subsequently, Eq. (2) is converted to an input quantum state

$$|\mathbf{In}\rangle_2 = \sum_{b=0}^{l-1} \alpha_b |b\rangle_2 + \sum_{b=l}^{2^n-1} 0|b\rangle_2 \subseteq \mathcal{H} \tag{3}$$

by quantum amplitude encoding U_A in Fig. 1 [32], where the subscript 2 indicates that Eq. (3) is situated in the second quantum register. $|b\rangle_2$ refers to the basis vector. $\mathcal{H}$ is the Hilbert space. $n = \lceil \log_2 l \rceil$ is the amount of qubits. $\sum_{b=0}^{l-1} \alpha_b^2 = 1$. $\alpha_b = a_b / \sqrt{\sum_{d=0}^{l-1} a_d^2}$. $a_b, a_d \in \mathbf{In}$. Then m target basis vectors

$$\mathcal{M} = \{|c_b\rangle_2\}_{b=0}^{m-1} \subseteq \{|b\rangle_2\}_{b=0}^{2^n-1} \tag{4}$$

from Eq. (3) are randomly selected and linearly combined with the opposite values of their corresponding probability amplitudes into the target term

$$|\text{focus}\rangle_2 = \sum_{b=0}^{m-1} \beta_b |c_b\rangle_2. \tag{5}$$

The formation of Eq. (5) is carried out by an oracle

$$U_S = \begin{bmatrix} (-1)^{f(|0\rangle)} & & & \\ & (-1)^{f(|1\rangle)} & & \\ & & \ddots & \\ & & & (-1)^{f(|2^n-1\rangle)} \end{bmatrix} \tag{6}$$

which is mathematically a $2^n \times 2^n$ diagonal unitary matrix, where

$$f(x) = \begin{cases} 1 & x \in \mathcal{M} \\ 0 & else \end{cases}. \tag{7}$$

To be specific,

$$U_S|b\rangle_2 = \begin{cases} -|b\rangle_2 & |b\rangle_2 \in \mathcal{M} \\ |b\rangle_2 & else \end{cases}, \tag{8}$$

which articulates that Eq. (4) are capable of self-labeling by phase inversion when acted on by Eq. (6). From a holistic point of view, if Eq. (6) is updated once, the conditional judgments in Eq. (7) must be performed several times. This not only falls into the trap of discrete non-differentiability but also makes the problem inaccessible because it is difficult to directly reproduce such multiple complex operations as Eq. (6) and Eq. (7) with quantum gates. However, from another perspective, Eq. (6) can be disassembled into

$$U_S = \prod_{b=0}^{2^n-1} \lambda_b, \tag{9}$$

where any element

$$\lambda_b = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & (-1)_b^{f(|b\rangle)} & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix} = \begin{cases} \Lambda(b) & b \in \mathcal{M} \\ I & else \end{cases}. \tag{10}$$

$$\Lambda(b) = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & -1_b & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix}. \tag{11}$$

Eq. (10) (or Eq. (11)) represents a diagonal unitary matrix whose b-th element on the diagonal is $(-1)_b^{f(|b\rangle)}$ (or -1) and the others are 1. All symbols I collectively stand for the identity matrix throughout this article.

Deconstructing a intricate problem through perspective shifts proves more lucid and facile than contemplating it holistically in advance, since it is now only necessary to consider how to switch the two forms of Eq. (10), Eq. (11) and I, with continuous parameters. Besides, it is worth noting that when $\lambda_b = I$, it is equivalent to applying no operation to the current quantum circuit. Guided by the above idea, a set of DPs

$$\{C - \Lambda(b)\}_{b=0}^{2^n-1} \tag{12}$$

is introduced.

Definition 3.2 (Discrete Primitive)**.** In Eq. (12), an arbitrary DP

$$C - \Lambda(b) = I \otimes |0\rangle_1\langle 0|_1 + \Lambda(b) \otimes |1\rangle_1\langle 1|_1 \tag{13}$$

as shown in Tab. 1 is a controlled quantum gate with one control bit and n target bits, which means that Eq. (11) is applied on the first quantum register only when the controlled qubit is $|1\rangle_1$. In Eq. (13), $\otimes$ represents the tensor symbol. $\langle 0|_1$ and $\langle 1|_1$ correspond to the conjugate transpositions of $|0\rangle_1$ and $|1\rangle_1$, respectively.

Corollary 3.1. Mathematically, Eq. (13) is a unitary diagonal matrix.

Proof.

$$\begin{aligned} C - \Lambda(b) &= I \otimes |0\rangle_1\langle 0|_1 + \Lambda(b) \otimes |1\rangle_1\langle 1|_1 \\ &= I \otimes \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \Lambda(b) \otimes \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ 0 & \Lambda(b) \end{bmatrix}_{2^{n+1} \times 2^{n+1}} \end{aligned} \tag{14}$$

satisfies $C - \Lambda(b) C - \Lambda(b)^\dagger = C - \Lambda(b)^\dagger C - \Lambda(b) = I$ and is therefore a unitary diagonal matrix, where $C - \Lambda(b)^\dagger$ is the conjugate transpose of $C - \Lambda(b)$. □

In adherence to Definition 3.2, all elements in Eq. (12) function when the controlled qubit is $|1\rangle_1$. Leveraging this uniqueness, quantum gates containing parameters are used to determine whether the controlled qubit is $|1\rangle_1$ or not, thus selectively activating and shielding certain elements in Eq. (12). For this purpose, the concept of FC is proposed.

Definition 3.3 (Flexible Control)**.** Suppose that the quantum control bit of Eq. (13) is initialized to $|0\rangle_1$. Subsequently Pauli rotating X gates $R_X(\theta)$ are posed on $|0\rangle_1$ to form the flexible control

$$R_X(\theta)|0\rangle_1 = \begin{cases} |1\rangle_1 & \theta = (4k+1)\pi, k \in \mathbb{Z} \\ R_X(\theta)|0\rangle_1 & else \end{cases}. \tag{15}$$

Eq. (15) shows that the control bit is $|1\rangle_1$ only when the continuous variable θ is located at $(4k+1)\pi, k \in \mathbb{Z}$.

Proof. In quantum computing, the interconversion between ground state $|0\rangle_1$ and excited state $|1\rangle_1$ can be realized by a Pauli X gate X. Therefore, $R_X(\theta)$ is equal to the form of X only when $\frac{\theta}{2}=\frac{\pi}{2}+2k\pi, k \in \mathbb{Z}$. That is,

$$\begin{bmatrix} \cos\left(\frac{\pi}{2}+2k\pi\right) & -i\sin\left(\frac{\pi}{2}+2k\pi\right) \\ -i\sin\left(\frac{\pi}{2}+2k\pi\right) & \cos\left(\frac{\pi}{2}+2k\pi\right) \end{bmatrix} = -i\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

where the global phase $-i$ can be neglected. □

Naturally, the definition of hard attention scores can be derived from Definition 3.3:

Definition 3.4 (Quantum Hard Attention Score)**.**

$$\text{QHAS} := \begin{cases} 1 & \theta = (4k+1)\pi, k \in \mathbb{Z} \\ 0 & else \end{cases} \tag{16}$$

can be used to visualize the choice of FC in training.

Based on Eq. (12) and Eq. (15), FO can be created.

Definition 3.5 (Flexible Oracle)**.** The FO

$$U_S^{'} = \prod_{b=0}^{2^n-1} R_X(\theta_b) \otimes I \times C - \Lambda(b), \tag{17}$$

as shown in Fig. 2, replaces Eq. (6) by applying flexible control to each DP in Eq. (12), thus coming to select and combine different DPs and create new discrete operations, ultimately freeing GQHAM from the confinement of discrete non-differentiability.

3.2 Adaptive Diffusion Operator

Eq. (17) focuses attention on important $\mathcal{M}$, but the amplitudes of these $\mathcal{M}$ may be so small that they cannot be thoroughly sorted out, which requires amplifying their quantum amplitudes while shrinking the amplitudes of the irrelevant terms by the diffusion operator. The diffusion operator in the original Grover algorithm [18] is defined as

$$U_D = H^{\otimes n} X^{\otimes n} (MCZ) X^{\otimes n} H^{\otimes n}. \tag{18}$$

Another geometric interpretation is that Eq. (18) is equivalent to

$$U_D = 2|\mathbf{In}\rangle\langle\mathbf{In}| - I, \tag{19}$$

where $|\mathbf{In}\rangle$ is the conjugate transpose of $\langle\mathbf{In}|$. I is the identity matrix. This implies that the symmetry axis of Grover's algorithm is $|\mathbf{In}\rangle$ for each iteration, which loses flexibility. Thus, Eq. (18) is improved and the ADO is proposed.

Definition 3.6 (Adaptive Diffusion Operator)**.** The ADO

$$U_D = U_1(MCZ)U_1^{\dagger}, \tag{20}$$

is shown in Fig. 3, where

$$U_1 = \prod_{b=0}^{n-1} CR_Y(\theta_b)[b,f] \bigotimes_{b=0}^{n-1} H[b]. \tag{21}$$

$$U_1^{\dagger} = \bigotimes_{b=0}^{n-1} H[b] \prod_{b=0}^{n-1} CR_Y(\theta_{b+n})[b,f]. \tag{22}$$

$$f = \begin{cases} c+1 & c \neq n-1 \\ 0 & c = n-1 \end{cases}. \tag{23}$$

Eq. (21) and Eq. (22) use a quantum coordinate representation [25]. Eq. (20) replaces X in Eq. (18) with $CR_Y(\theta)$ to make it trainable, which indicates that this axis of symmetry is no longer fixed and has flexibility.

At present, the GQHAM mechanism stands elucidated comprehensively. The ingenuity of GQHAM resides in the activation or masking of DPs through continuous parameters, thereby amalgamating diverse discrete choices without transgressing the tenets of gradient backpropagation. This innovation ultimately surmounts the challenge of non-differentiability. Furthermore, enhancements have been instituted in the original expansion operator to augment the flexibility of GQHAM.

4 Grover-inspired Quantum Hard Attention Network

In this section, GQHAN in Fig. 4 is constructed based on GQHAM and its workflow is elaborated.

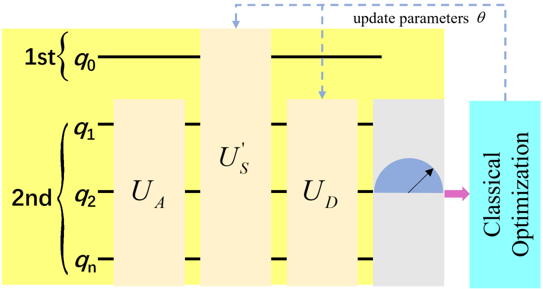


Fig. 4: Grover-inspired Quantum Hard Attention Network

As depicted in Fig. 4, GQHAN is segregated into dual components: the ansatz segment in the yellow box and the classical optimizer module in the blue box. Among them, the ansatz segment is constituted by $n+1$ qubits, encompassing an ancillary qubit q_0 in the first quantum register and n input qubits $q_1 \sim q_n$ in the second quantum register. In addition, U_A in Fig. 1, $U_S^{'}$ in Fig. 2 and U_D in Fig. 3 sequentially are quantum amplitude encoding, FO and ADO. Ultimately, the data is extracted via quantum measurement, depicted within the gray box in Fig. 4, and subsequently transmitted to the classical optimizer for further processing. Employing classical optimizers, such as the quantum natural simultaneous perturbation stochastic approximation optimizer [34] or the quantum natural gradient optimizer [35], the parameters undergo iterative refinement until convergence of the loss function is achieved. It is noteworthy that in this paper, the Nesterov momentum optimizer [36], a variant amalgamating momentum components with gradient descent, is utilized to incorporate historical gradients into the optimization process.

In accordance with the framework depicted in Fig. 4, the precise workflow of GQHAN unfolds as follows.

Step 1: The initial quantum state

$$|\Psi\rangle_1 \otimes |\Phi\rangle_2 = |0\rangle_1 \otimes |0^{\otimes n}\rangle_2 \tag{24}$$

is prepared in the first and second quantum registers.

Step 2: Eq. (2) are embedded into GQHAN by quantum amplitude encoding U_A:

$$|\Psi\rangle_1 \otimes |\Phi\rangle_2 \xrightarrow{U_A} |0\rangle_1 \otimes |\mathbf{In}\rangle_2. \tag{25}$$

Step 3: Eq. (17) is used to adjust the continuous parameter during training to partition Eq. (3) into its salient and negligible components, $|\text{focus}\rangle_2$ and $|\overline{\text{focus}}\rangle_2$:

$$|\Psi\rangle_1 \otimes |\Phi\rangle_2 \xrightarrow{U_S^{'}} |\phi\rangle_1 \otimes (|\text{focus}\rangle_2 + |\overline{\text{focus}}\rangle_2), \tag{26}$$

where $|\phi\rangle_1$ is determined by Eq. (15).

Step 4: Apply Eq. (20) to dynamically amplify the amplitudes in Eq. (5) while reducing the other amplitudes to end up as

$$|\Psi\rangle_1 \otimes |\Phi\rangle_2 \xrightarrow{U_D} |\phi\rangle_1 \otimes |\widetilde{\text{focus}}\rangle_2. \tag{27}$$

Step 5: For specific Fashion MNIST binary classification, the expectation value

$$\mathbb{E} = \langle T|P|T\rangle \tag{28}$$

on the last qubit q_n is measured as the predicted label, where $|T\rangle = |\Psi\rangle_1 \otimes |\Phi\rangle_2$. P is the projection operator.

Step 6: Cost function is

$$f(\mathbf{In}, \boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^{m} \left[y_i - \text{sgn}(\mathbb{E})\right]^2, \tag{29}$$

where sgn$(\cdot)$ is the sign function. m is the number of terms. y_i stands for real label. $\boldsymbol{\theta}$ are the trainable parameters. The optimization rules for the Nesterov momentum optimizer are as follows:

$$\begin{cases} \boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - a^{(t+1)} \\ a^{(t+1)} = \gamma \cdot a^{(t)} + \eta \cdot \nabla f(\mathcal{D}, \mathcal{D}', \boldsymbol{\theta}^{(t)} - \gamma \cdot a^{(t)}) \end{cases} \tag{30}$$

where the superscript t represents the t-th iteration and $t+1$ stands for the next iteration. The momentum term γ is adjustable and generally takes the value 0.9. η is the learning rate. $a^{(t+1)}$ and $a^{(t)}$ are accumulator terms. $\nabla f(\mathcal{D}, \mathcal{D}', \boldsymbol{\theta}^{(t)} - \gamma \cdot a^{(t)})$ denotes the gradient.

In the end, all the above steps are repeated until the cost function converge. To sum up, by analyzing Step 1 to 6 above, the operation mechanism of GQHAN in the classification problem is revealed, which lays a theoretical foundation for the experiment.

5 EXPERIMENT

In this section, GQHAN is implemented on the PennyLane platform to conduct binary classification experiments on Fashion MNIST. Precisely, the experiments are categorized into the subsequent segments.

- The performance of GQHAN, QSAN [25] and QKSAN [27] is evaluated and compared in the Fashion MNIST binary classification task under a uniform classical optimizer configuration and no noise.
- A visualization of QHAS is performed to represent the results before and after training.
- To elucidate the performance on a real quantum computer, the impact of bit-flip error and amplitude damping error on GQHAN is scrutinized.

5.1 Dataset

Fashion MNIST [38], the recognized and widely adopted benchmark datasets in machine learning, consists of 10,000 test images and 60,000 training images, respectively, each containing 28 by 28 pixel points. In this experiment, 550 images labeled 0 and another 550 tagged 1 are taken as the dataset from Fashion MNIST. 500 images from each class of tags are randomly sampled and assembled into a training set with the rest as a test set. Moreover, acknowledging the existing constraint on the number of public qubits, exemplified by the limited provision of 5 to 7 free-to-use qubits by IBM, one strategy is to compress the dimensionality of all images in the dataset to 8 using the principal component analysis algorithm.

5.2 Experimental Setting

Tab. 2 delineates the specific experimental configurations of GQHAN, QSAN [25] and QKSAN [27], detailing key parameters of both the ansatz and classical optimizers. In the ansatz, paramount metrics encompass parameter counts in trainable layers, layer quantity, and requisite qubits. Classical optimization scrutinizes factors such as learning rate, loss function type, batch size, maximum step size, optimizer type, and relevant parameters. The type and configuration of the classical optimizer, along with the number of ansatz parameters, are maintained as consistently as possible to ensure and underscore equitable comparisons between quantum models. Additionally, a substantial learning rate is deliberately chosen to expedite convergence.

Tab. 2: Experimental Configuration

Indicators	Models		
	GQHAN	QKSAN [27]	QSAN [25]
parameters	14	14	9
layers	7	6	26
qubits	4		8
learning rate	0.09		
loss function	square loss		
batch_size	30		
step	120 steps		
optimizer	Nesterov Momentum Optimizer: $\gamma = 0.9$		

Tab. 3: Critical Data

Indicators	Models		
	GQHAN	QKSAN [27]	QSAN [25]
test accuracy of the last 10 steps	98.59%	98%	96.8%
train accuracy of the last 10 steps	98.65%	97.22%	96.77%
number of step to start convergence	19	25	55
convergence value of the loss function	0.219	0.323	0.42

5.3 Experimental Analysis

5.3.1 Classification Experiment

The outcomes of the comparative analysis involving GQHAN and two quantum soft self-attention mechanisms,

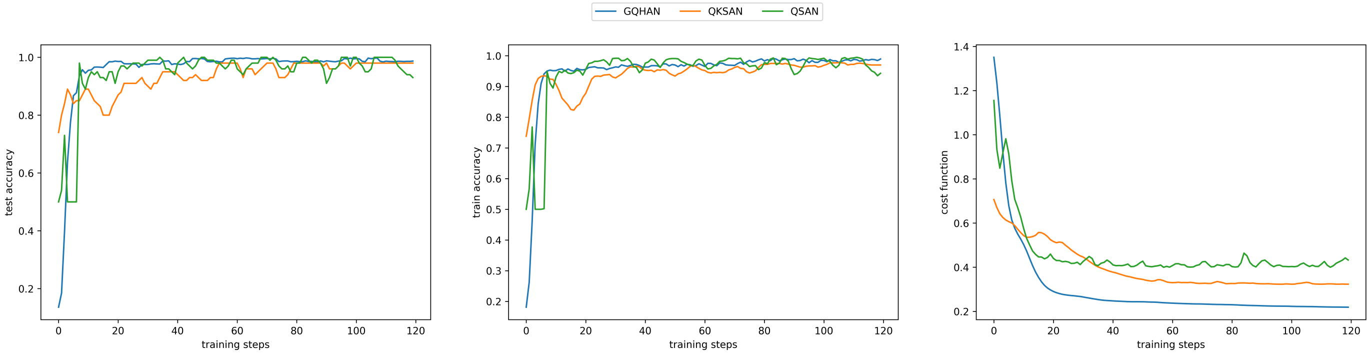
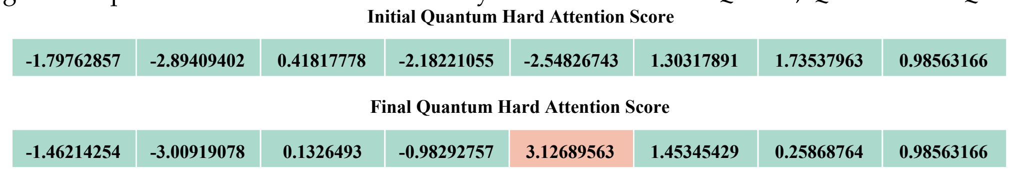


Fig. 5: Comparison of Fashion MNIST Binary Classification for GQHAN, QKSAN and QSAN.



Initial Quantum Hard Attention Score

-1.79762857	-2.89409402	0.41817778	-2.18221055	-2.54826743	1.30317891	1.73537963	0.98563166

Final Quantum Hard Attention Score

-1.46214254	-3.00919078	0.1326493	-0.98292757	3.12689563	1.45345429	0.25868764	0.98563166

Fig. 6: Visualization of Quantum Hard Attention Score

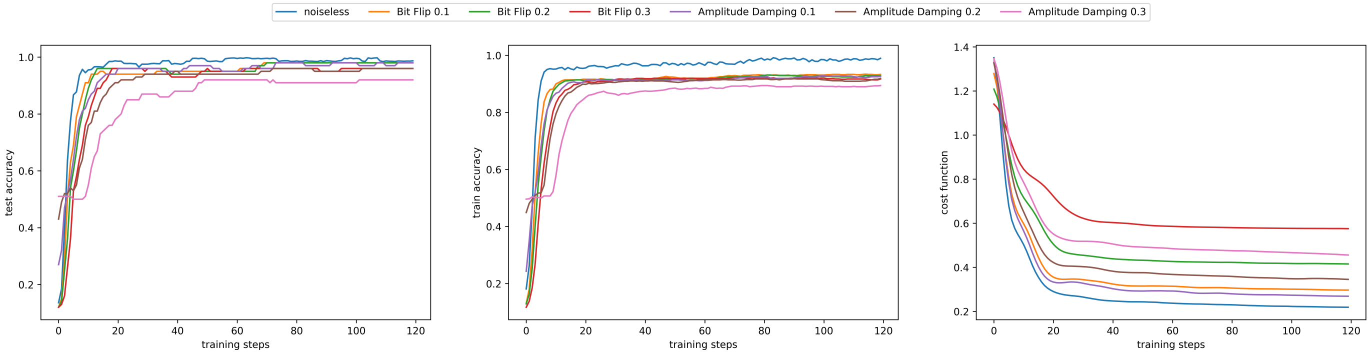


Fig. 7: Noise Experiments of GQHAN

namely QKSAN and QSAN, are depicted in Fig. 5. In Fig. 5, the abscissae of all three subplots represent the training steps, while the ordinates denote the test accuracy, train accuracy, and loss function sequentially. The blue, orange and green curves delineate GQHAN, QKSAN, and QSAN correspondingly. According to Fig. 5, the critical data are enumerated in Tab. 3. Finally, the following conclusions can be drawn.

- By averaging the accuracy over the last 10 steps, GQHAN attains a test accuracy of 98.59%, surpassing QKSAN at 98% and QSAN at 96.8%. Regarding training accuracy, it outperforms QKSAN by 1.42% and QSAN by 1.87%.
- In terms of convergence speed, GQHAN initiates convergence around step 20, slightly lagging behind QSAN by approximately 5 steps but significantly outpacing QKSAN.
- Concerning convergence values, GQHAN converges to 0.219, compared to 0.323 for QSAN and 0.42 for QKSAN, indicating a superior learning capability.

In summary, the analysis and comparison from three perspectives demonstrates that GQHAN is able to slightly surpass the two quantum soft self-attention mechanisms in terms of performance in the Fashion MNIST classification.

5.3.2 Visualization of Quantum Hard Attention Score

The visualization results of QHAS under the experimental conditions of the previous subsection are shown in Fig. 6. All values in Fig. 6 represent the parameters θ in $R_X(\theta)$. Here, green color is used to indicate that they are not selected, while red color has the opposite meaning of green color. Since the parameter initialization are random, this batch of parameters does not meet the requirement of Eq. (16). Instead, in the last round of training, the parameter 3.12689563 is marked red and is very close to π, at which point it can be considered as being discretely selected in engineering.

5.3.3 Noise Experiments

The amplitude damping and bit-flip noise experiments for GQHAN are illustrated in Fig. 7. Except for the blue line which indicates the experimental results in the absence of noise, Fig. 7 depicts three sets of experiments conducted for both amplitude damping and bit-flip noise, each occurring with probabilities of 0.1, 0.2, and 0.3. Likewise, the mean metrics derived from the final 10 steps upon completion of the training regimen were employed as pivotal assessment data, as delineated in Tab. 4. According to the data presented in Tab. 4, discernible conclusions can be derived.

Tab. 4: Critical Data for Noise Experiments

Models	Indicators		
	test accuracy	train accuracy	convergence value
noiseless	98.59%	98.65%	0.219
amplitude damping 0.1	97.6%	92.42%	0.269
amplitude damping 0.2	96%	91.24%	0.348
amplitude damping 0.3	92%	89.14%	0.459
bit flip 0.1	98%	93.27%	0.297
bit flip 0.2	98%	92.68%	0.416
bit flip 0.3	96%	91.58%	0.576

- In assessing the diminution of test and training accuracy, the impact of amplitude damping noise exhibits a more conspicuous manifestation.
- With equiprobability of noise emergence, the bit-flip perturbation induces a more pronounced escalation in convergence values, thereby exacerbating the learning efficacy degradation of GQHAN.

In total, GQHAN exhibits a degree of resilience against both types of noise. Concerning accuracy, the impact of bit-flip noise manifests as a comparatively minor decline in intuitive values. However, it is noteworthy that the degradation in its effect on learning performance is more conspicuous.

6 Conclusion

A quantum HAM, GQHAM, is proposed to compensate for the inability of many current QMLs to recognize the significance of quantum data. In addition, FO and ADO are designed to overcome the nativity of non-differentiability due to discrete selection. A QHAS visualization definition is also providedon this premise. Ultimately, GQHAN, which can be deployed on a quantum computer, is constructed based on GQHAM. In binary classification experiments on Fashion MNIST on the PennyLane platform, GQHAN achieves 98.59% and 98.65% test and train accuracies, respectively, and outperforms the two quantum soft self-attention in terms of learning ability. In the noise experiments, contrasted with amplitude damping noise, the impact of bit flip noise on GQHAN is relatively diminished in precision but looms larger in learning capabilities. Our approach contributes to the QML model to pay more attention to the important parts of quantum data, laying the foundation for future quantum computers to process massive amounts of high-dimensional data.

References

[1] J. Shi, W. Wang et al., “Parameterized Hamiltonian learning with quantum circuit,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1-10, 2022.

[2] J. Shi, Y. Tang et al., “Quantum circuit learning with parameterized Boson sampling,” IEEE Transactions on Knowledge and Data Engineering, pp. 1-1, 2021.

[3] M. Cerezo, G. Verdon et al., “Challenges and opportunities in quantum machine learning,” Nature Computational Science, vol. 2, no. 9, pp. 567-576, 2022

[4] K. Xu, J. Ba et al., “Show, attend and tell: Neural image caption generation with visual attention,” in Proceedings of the 32nd International Conference on Machine Learning, pp. 2048–2057, 2015.

[5] B. Uzkent, and S. Ermon, “Learning when and where to zoom with deep reinforcement learning,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12342-12351, 2020.

[6] G. Elsayed, S. Kornblith et al., “Saccader: Improving accuracy of hard attention models for vision,” Advances in Neural Information Processing Systems, vol. 32, 2019.

[7] A. Katharopoulos, and F. Fleuret, “Processing megapixel images with deep attention-sampling models,” in Proceedings of the 36th International Conference on Machine Learning, pp. 3282–3291, 2019.

[8] A. Papadopoulos, P. Korus et al., “Hard-attention for scalable image classification,” Advances in Neural Information Processing Systems, vol. 34, pp. 14694-14707, 2021.

[9] J. Serra, D. Suris et al., “Overcoming catastrophic forgetting with hard attention to the task,” in Proceedings of the 35th International Conference on Machine Learning, pp. 4548–4557, 2018.

[10] D. Wang, A. Haytham et al., “Hard attention net for automatic retinal vessel segmentation,” IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 12, pp. 3384-3396, 2020.

[11] S. R. Indurthi, I. Chung et al., “Look harder: A neural machine translation model with hard attention,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3037-3043, 2019.

[12] B. Nikpour, and N. Armanfard, “Spatial hard attention modeling via deep reinforcement learning for skeleton-based human activity recognition,” IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 53, no. 7, pp. 4291-4301, 2023.

[13] H. Mohammadi, and E. Nazerfard, “Video violence recognition and localization using a semi-supervised hard attention model,” Expert Systems with Applications, vol. 212, pp. 118791, 2023.

[14] I. Jegham, I. Alouani et al., “Deep learning-based hard spatial attention for driver in-vehicle action monitoring,” Expert Systems with Applications, vol. 219, pp. 119629, 2023.

[15] M. Malinowski, C. Doersch et al., “Learning visual question answering by bootstrapping hard attention,” in Proceedings of the European Conference on Computer Vision (ECCV), pp. 3-20, 2018.

[16] K. Ahmed, and L. Torresani, “Star-caps: Capsule networks with straight-through attentive routing,” Advances in neural information processing systems, vol. 32, 2019.

[17] X. Du, T. Wang et al., “Multi-agent reinforcement learning for dynamic resource management in 6G in-X subnetworks,” IEEE Transactions on Wireless Communications, vol. 22, no. 3, pp. 1900-1914, 2023.

[18] L. K. Grover, “A fast quantum mechanical algorithm for database search,” in Proceedings of the 28th Annual ACM Symposium on Theory of Computing, pp. 212–219, 1996.

[19] C. Figgatt, D. Maslov et al., “Complete 3-qubit Grover

search on a programmable quantum computer," Nature Communications, vol. 8, no. 1, pp. 1918, 2017.

[20] T. Byrnes, G. Forster et al., "Generalized Grover's algorithm for multiple phase inversion states," Physical Review Letters, vol. 120, no. 6, pp. 060501, 2018.

[21] A. Gilliam, S. Woerner et al., "Grover adaptive search for constrained polynomial binary optimization," Quantum, vol. 5, pp. 428, 2021.

[22] G. Anikeeva, O. Marković et al., "Number partitioning with Grover's algorithm in central spin systems," PRX Quantum, vol. 2, no. 2, pp. 020319, 2021.

[23] J. Bausch, "Fast black-box quantum state preparation," Quantum, vol. 6, pp. 773, 2022.

[24] X. Niu, Y. Hou et al., "Bi-directional LSTM with quantum attention mechanism for sentence modeling," in Neural Information Processing, pp. 178-188, 2017.

[25] R.-X. Zhao, J. Shi et al., "QSAN: A near-term achievable quantum self-attention network," arXiv preprint arXiv:2207.07563, 2022.

[26] G. Li, X. Zhao et al., "Quantum self-attention neural networks for text classification," arXiv preprint arXiv:2205.05625, 2022.

[27] R.-X. Zhao, J. Shi et al., "QKSAN: A quantum kernel self-attention network," arXiv preprint arXiv:2308.13422, 2023.

[28] M. Coggins, Introduction to quantum computing with Qiskit: Scarborough Quantum Computing Ltd, 2021.

[29] A. Younes, "Fixed phase quantum search algorithm," arXiv preprint arXiv:0704.1585, 2007.

[30] D. Biron, O. Biham et al., "Generalized Grover search algorithm for arbitrary initial amplitude distribution," in Quantum Computing and Quantum Communications, pp. 140-147, 1999.

[31] D. Shapira, Y. Shimoni et al., "Algebraic analysis of quantum search with pure and mixed states," Physical Review A, vol. 71, no. 4, pp. 042320, 2005.

[32] I. F. Araujo, D. K. Park et al., "Configurable sublinear circuits for quantum state preparation," Quantum Information Processing, vol. 22, no. 2, pp. 123, 2023.

[33] A. Kandala, A. Mezzacapo et al., "Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets," Nature, vol. 549, no. 7671, pp. 242-246, 2017.

[34] J. Gacon, C. Zoufal et al., "Simultaneous perturbation stochastic approximation of the quantum Fisher information," Quantum, vol. 5, pp. 567, 2021.

[35] J. Stokes, J. Izaac et al., "Quantum natural gradient," Quantum, vol. 4, pp. 269, 2020.

[36] Y. Nesterov, "A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$," in Doklady an ussr, pp. 543-547, 1983.

[37] Y. Lecun, L. Bottou et al., "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, 1998.

[38] H. Xiao, K. Rasul et al., "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," arXiv preprint arXiv:1708.07747, 2017.

Ren-Xin Zhao (Member, IEEE) received his B.S. degree from the College of Automation, Hangzhou Dianzi University, Hangzhou, China in 2017 and his M.S. degree from the College of Electrical and Information Engineering, Hunan University, Changsha, China in 2020. He is now a PhD student in the School of Computer Science and Engineering at Central South University, Changsha, China. His interests include quantum machine learning, quantum neural networks, and design and optimization of quantum circuits.

Jinjing Shi (Member, IEEE) is now a professor in the School of Electronic Information of Central South University. She received her B.S. and Ph.D. degrees in the School of Information Science and Engineering, Central South University, Changsha, China, in 2008 and 2013, respectively. She was selected in the "Shenghua lieying" talent program of Central South University and Special Foundation for Distinguished Young Scientists of Changsha in 2013 and 2019, respectively. Her research interests include quantum computation and quantum cryptography. She has presided over the National Natural Science Foundation Project of China and that of Hunan Province. There are 50 academic papers published in important international academic journals and conferences. She has received the second prize of natural science and the outstanding doctoral dissertation of Hunan Province in 2015, and she has received the Best Paper Award in the international academic conference MSPT2011 and Outstanding Paper Award in IEEE ICACT2012.

Xuelong Li (M'02-SM'07-F'12) is with the Institute of Artificial Intelligence (Tele AI), China Telecom Corp Ltd, 31 Jinrong Street, Beijing 100033, P. R. China