QiMLP: Quantum-inspired Multilayer Perceptron with Strong Correlation Mining and Parameter Compression

Junwei Zhang[1], **Tianheng Wang**[1], **Zeyi Zhang**[1,2], **Pengju Yan**[1], **Xiaolin Li**[1,*]

[1]Hangzhou Institute of Medicine, Chinese Academy of Sciences, Hangzhou, Zhejiang, China.
[2]Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, Zhejiang, China.
zhangjunwei@him.cas.cn, wangtianheng@him.cas.cn, xiaolinli@ieee.org

Abstract

Multilayer Perceptron (MLP) is a simple practice of Neural Network (NN) and the cornerstone of research and development of deep learning. Each neuron is connected to all neurons in the previous layer and implements a non-linear mapping through activation functions. MLP can learn complex non-linear relationships among features through the superposition of multiple hidden layers, but it still cannot discover the inherent strong correlation among features. The reason is that each neuron uses a simple weighted summation method to organize all the neurons in the previous layer. Inspired by quantum theory, this paper builds a non-linear NN layer that can mine strong correlations among features based on multi-body quantum systems, and then constructs a multi-layer perceptron, called Quantum-inspired MLP (QiMLP). It is conceivable that QiMLP will have important inspirational significance in reshaping machine learning, deep learning and large language models. We theoretically analyzed the basis for QiMLP to mine strong correlations among features, and implemented experiments on multiple classic deep learning datasets. Experimental results verify that QiMLP not only learns strong correlations among features, but also significantly reduces the number of parameters with hundreds of times improvement.

Introduction

Multilayer Perceptron (MLP) realizes modeling of non-linear relationships through non-linear mapping of activation functions (Popescu et al. 2009), and learns more abstract and advanced feature representations through the combination of multiple neuron layers, which can capture deeper structures and patterns in the data (Murtagh 1991). Although MLP is a simple practice of traditional neuron layers, it has important application value and superior performance in specific tasks (Kruse et al. 2022), and is the cornerstone of research and construction of Machine Learning, Deep Learning and Large Language Models.

However, the weighted summation method used by neurons in each layer to fuse the features output by neurons in the previous layer makes it difficult to learn the strong correlation (Jiang et al. 2010; Becke 2013; Ceravolo et al. 2021) and cannot capture the intrinsic relationship among features (Zhang et al. 2020). The method only assigns a certain weight value to the importance of the features and adds the involved features. In the real world, there are generally explicit or implicit strong internal connections between the characteristics of things, such as causal relationships, strong statistical correlations, etc. Therefore, building a neuron layer that can mine strong correlations among features is more conducive to capturing the real relationships between the characteristics of things, and then implementing more effective decisions. In addition, machine learning models based on traditional neuron layers, such as deep learning models and large language models, have huge parameters that need to be learned and require a large amount of computing resources during the training process, which have become a bottleneck in the development of Artificial Intelligence (AI) (Lu 2019).

For mining strong correlations among features, researchers have not found a suitable solution to control the strong correlation, so they try to learn a disentangled representation (Higgins et al. 2018; Wang et al. 2022) method that is independent among features, thereby reducing the interference of strong correlations among features on downstream tasks. The most representative work among them is the disentangled representation method based on the Variational Auto-encoder (VAE), such as beta-VAE (Higgins et al. 2016) and FactorVAE (Duan et al. 2022). **For reducing the number of parameters of neural network models**, researchers have proposed methods in recent years including model sparsification (also called model pruning) (Zhou et al. 2021), model quantization (Zhou et al. 2018), knowledge distillation (Gou et al. 2020), tensor decomposition (Rabanser, Shchur, and Günnemann 2017), etc., to achieve model compression and reduce the number of parameters.

In quantum theory, multi-body quantum systems (Stasiuk 2021; Zhang et al. 2022a) can characterize strong relationships among characteristics of systems (Fetter, Walecka, and Kadanoff 1971; Zhang, He, and Guo 2023), and this relationships includes not only statistical correlations in classical world, but also non-classical correlations in quantum world (Holevo and Ballentine 1982; Zhang et al. 2024). In other words, the formal framework of quantum mechanics can describe the strong correlation between the characteristics of things. **What needs to be emphasized here again is that the strong correlation in this paper does not specifi-**

*Corresponding author.

cally refer to non-classical or quantum-like relationships between the characteristics of things, such as quantum non-locality, but refers to a strong statistical correlation between the characteristics, such as causal relationship.

This paper leverages the framework of quantum mechanics and refers to the multi-body quantum system to reshape the basic components of NN, the neuron layers, and further build MLP with practical application value, called Quantum-inspired MLP (QiMLP). Specifically, one or more weight matrices obtained by the Kronecker product (Broxson 2006) operation is applied to the input tensor of the neuron layer, a bias term obtained by the Kronecker product operation is added, and finally an activation function is applied. The purpose of the Kronecker product operation is to obtain a matrix that can operate a multi-body quantum system. The purpose of applying multiple matrices obtained by the Kronecker product operation is to increase the number of parameters of the neuron layer, because the Kronecker product operation will cause the parameters to decrease exponentially, resulting in the problem of too few parameters. Finally, we theoretically analyzed the basis for QiMLP to mine strong correlations among features, and implemented experiments on multiple classic deep learning datasets. **Experimental results verify that QiMLP not only learns strong correlations among features, but also reduces the number of parameters hundreds of times.**

The contributions can be summarized as follows:

- Based on the multi-body quantum system, a neuron layer that can mine the strong correlations among features is constructed;
- Based on the constructed neuron layer, a multi-layer perceptron with practical application value is built, which reduces the number of model parameters hundreds of times;
- Finally, we theoretically analyzed the basis for the constructed neuron layer to mine strong correlations, and experimentally verified the effectiveness of the constructed multi-layer perceptron.

Related Work

Strong Correlation Mining

Junwei et al. (Zhang et al. 2020, 2021, 2022b,c) studied the significance of the strong statistical correlations revealed by quantum entanglement in the field of machine learning. This work completes the learning of strong statistical correlation by constructing a learnable measurement operator to measure the fixed maximum entangled state. The advantage of this work is that the maximum entangled state is directly used as the measured quantum state, which can effectively ensure the existence of strong correlation in the quantum system. In addition, researchers have gone to another extreme, trying to learn disentangled representations to obtain a description of features that are independent of each other. It directly abandons the use of strong correlation to avoid interference caused by unreasonable use of strong correlation on downstream tasks. The most representative work among them is the disentangled representation method based on the Variational Auto-encoder (VAE), such as beta-VAE (Higgins et al. 2016) and FactorVAE (Duan et al. 2022).

Compared with the work of Junwei et al. (Zhang et al. 2020), our work is more general and can be used to obtain representations with strong correlation and applied to downstream tasks. In addition, our work does not forcefully introduce strong correlations through quantum entanglement states, but naturally learns the strong correlations that already exist in data features, such as causal relationships, strong statistical correlations, etc. Compared with disentangled representation, our work can also be used as a way to obtain representation, but the difference is that we go to another extreme.

Parameter Compression

With the popularity of large language models, research on model parameter compression has practical significance. Researchers have proposed methods in recent years to achieve model compression and reduce the number of parameters. Model sparsification (Zhou et al. 2021), also called model pruning, reduces the number of parameters of the model by removing redundant connections and neurons, thereby reducing the storage and computing overhead of the model. Among them, the pruning algorithm is a commonly used neural network compression algorithm. Model sparsification also includes gradient sparsification, activation function sparsification, etc. Model quantization (Zhou et al. 2018) reduces the storage and computing overhead of the model by reducing the number of bits in the model parameters. Knowledge distillation (Gou et al. 2020) transfers the knowledge of a complex model (teacher model) to a simplified model (student model), thereby reducing the number of parameters and calculations of the model. Tensor decomposition (Rabanser, Shchur, and Günnemann 2017) reduces the number of parameters and calculations of the model by decomposing the weight matrix of the model into the product of multiple low-rank matrices.

Among the above model parameter compression methods, tensor decomposition has the most similarity with our work. Tensor decomposition technology hopes to extract the most important information or main components in the original data by decomposing tensors, while our work attempts to learn the main information or important components through multiple small weight matrices.

Methodology

Quantum-inspired MLP

The classic MLP is a feedforward NN model (Almeida 2020), which consists of multiple neuron layers, including an input layer, one or more hidden layers and an output layer. MLP can learn more complex feature representations through the superposition of multiple hidden layers, thereby improving the expression ability of the model. Each neuron layer is connected to all neurons in the previous layer and implements nonlinear mapping through a combination of weights, biases, and activation functions. Assume that the current neuron layer has N neurons, and each neuron outputs a value $y_i \in \mathbb{R}$, then the output of this layer can be

expressed as $Y = [y_1, y_2, ..., y_i, ...]^T$. Similarly, assuming that the previous layer has M neurons, and each neuron outputs a value $x_i \in \mathbb{R}$, the output of this layer can be expressed as $X = [x_1, x_2, ..., x_i, ...]^T$. Since the output of the previous layer is the input of the current layer, and non-linear mapping is achieved through a combination of a weight matrix $W \in \mathbb{R}^{N\times M}$, a bias vector $B \in \mathbb{R}^{N\times 1}$, and an activation function $f(\cdot)$, it can be formally described as

$$Y = f(WX + B). \tag{1}$$

The activation function should be selected appropriately according to the characteristics of the specific task. Usually the activation function $ReLU(\cdot)$ will be used in the hidden layer, and the activation function $Softmax(\cdot)$ will be used in the output layer during classification tasks.

It can be seen from Eq. (1) that the neurons in the classic neuron layer use a weighted summation method to integrate the output of all neurons in the previous layer. In this method, each element to be learned in the weight matrix is only related to the output of one neuron in the previous layer. It is difficult to learn strong correlations such as causal relationships and strong statistical correlation among the outputs of neurons in the previous layer. In addition, the weight matrix W has a serious over-configuration problem relative to the input X, that is, the number of parameters that need to be learned is much more than the number of features. Similarly, the bias term B also has this problem. It should be emphasized that the above problems exist for all deep learning models based on classic neural network layers.

The quantum-inspired MLP redesigns the strategy of constructing weight matrices and biases, enabling it to have the capabilities of strong correlation mining and parameter compression. We use the Kronecker product (Broxson 2006) to construct an operator that can act on a multi-body quantum system. In this paper, the operator is regarded as a weight matrix acting on the input of the neurons. Assuming that the input of the neuron layer is a N^M-dimensional vector $X \in \mathbb{R}^{N^M\times 1}$, that is, the number of neurons in the previous layer is N^M, the reconstructed weight matrix can be expressed as

$$W = \bigotimes_i^M W^i = W^1 \otimes W^2 \otimes \cdots \otimes W^i \otimes \cdots, \tag{2}$$

where $W^i \in \mathbb{R}^{N\times N}$, and the operator "$\otimes$" refers to the Kronecker product. Similarly, assuming that the corresponding bias term is a N^M-dimensional vector $B \in \mathbb{R}^{N^M\times 1}$, the reconstructed bias term can be expressed as

$$B = \bigotimes_i^M B^i = B^1 \otimes B^2 \otimes \cdots \otimes B^i \otimes \cdots, \tag{3}$$

where $B^i \in \mathbb{R}^{N\times 1}$. According to this, the output Y of the current neuron layer can be described as

$$Y = f\left(\sum_k \bigotimes_i^M W_k^i * X + \bigotimes_i^M B^i\right), \tag{4}$$

where $Y \in \mathbb{R}^{N^M\times 1}$, the operator "$*$" refers to the matrix multiplication operation, and k refers to the number of copies of the reconstructed weight matrix, which is used to alleviate excessive compression of model parameters. The activation function $f(\cdot)$ can refer to the classic MLP, and the appropriate activation function can be selected according to the specific task. At this point, the neuron layer, the basic component of building QiMLP, has been reconstructed, and an appropriate network model can be constructed according to the needs of specific tasks.

It can be seen from Eq. (4) that compared to Eq. (1), our calculation method is more complex, which is mainly reflected in integrating multiple small weight matrices and bias terms through the Kronecker product. From a purely formal perspective, the elements in the weight matrix obtained by the Kronecker product operation are all in the form of the product of multiple learnable parameters. The product form is more suitable for describing the strong correlation among features than the traditional single-valued form. In addition, the weight matrix of the same dimension obtained by the Kronecker product operation can save more parameters. From the perspective of quantum theory, the operators obtained by the Kronecker product operation can be applied to multi-body quantum systems such as composite states and entangled states to obtain the strong correlation between subsystems (Marinescu 2011; Galindo and Martin-Delgado 2002). A more detailed elaboration is given in the next section.

Theoretical Analysis

Parameter Compression Analysis: Mathematically, the Kronecker product (Broxson 2006) is an operation between two matrices of any dimension, which multiplies each element in the first matrix by the second complete matrix.

Definition 1. *Let A be a $(m\times n)$-dimensional matrix and B be a $(p\times q)$-dimensional matrix, then the Kronecker product operation $A\otimes B$ is a $(mp\times nq)$-dimensional matrix,*

$$A\otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}. \tag{5}$$

Example 1. *A matrix A of (2×2) dimensions and a matrix B of (2×2) dimensions, then the Kronecker product operation $A\otimes B$ can be described as*

$$\begin{aligned} A\otimes B &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \otimes \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}B & a_{12}B \\ a_{21}B & a_{22}B \end{bmatrix} \\ &= \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{12}b_{11} & a_{12}b_{12} \\ a_{11}b_{21} & a_{11}b_{22} & a_{12}b_{21} & a_{12}b_{22} \\ a_{21}b_{11} & a_{21}b_{12} & a_{22}b_{11} & a_{22}b_{12} \\ a_{21}b_{21} & a_{21}b_{22} & a_{22}b_{21} & a_{22}b_{22} \end{bmatrix}. \end{aligned} \tag{6}$$

It can be seen from Definition 1 that obtaining a matrix of the same size as the original weight matrix through the Kronecker product operation can save $(mp\times nq - m\times n - p\times q)$ parameters. Comparison results under real models show that QiMLP can reduce the number of model parameters by **more than 200 times** compared with the classic MLP with the same structure and scale. This advantage will be further expanded if multiple Kronecker products are used.

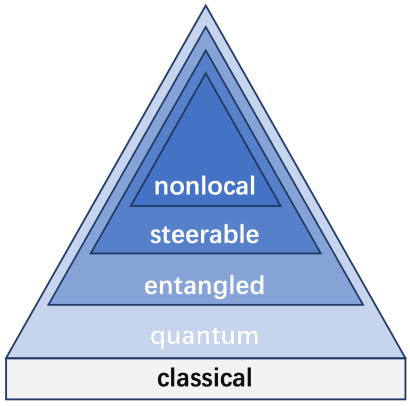

Figure 1: Correlation hierarchies for multi-body quantum systems.

Strong Correlation Mining Analysis: Bell's theorem is an infeasibility theorem, also known as Bell's inequality (Clauser and Shimony 1978). In classical theory, this inequality holds; in quantum theory, this inequality does not hold. The violation of Bell's inequality by quantum theory means that quantum physics violates the principle of locality or counterfactual certainty (Nielsen and Chuang 2001). It also shows that the quantum correlation revealed by quantum probability theory is stronger than the classical correlation revealed by classical probability theory, which can be described in order of decreasing strength: **quantum correlation** $\Rightarrow$ **classical correlation** (Peres 1997). Furthermore, quantum theory reveals that almost all general hybrid forms of two- or multi-body quantum systems exhibit quantum correlations. Several levels of non-classical correlation have been identified for the hybrid form of multi-body quantum systems, as shown in Fig. 1, which can be divided in order of decreasing strength: **non-locality** $\Rightarrow$ **steering** $\Rightarrow$ **entanglement** $\Rightarrow$ **general quantum correlation** (Adesso, Bromley, and Cianciaruso 2016). All these classes of non-classical correlations can be applied to tasks that cannot be accomplished by classical correlations (Hirvensalo and Yakaryılmaz 2019). At this point, we can get the corollary:

Corollary 1. *When the multi-body quantum system output by the neural layer is in a hybrid form and the correlation it contains is stronger than the classical correlation, it can be inferred that there is a strong correlation in the multi-body quantum system.*

In quantum theory, the basis for judging that a two-body quantum system is in a hybrid form is that the composite system cannot be tensor decomposed (Nielsen and Chuang 2001). For example, for the composite system $|AB\rangle$[1], the inequality $|AB\rangle \neq |A\rangle \otimes |B\rangle$ does not exist. A way to determine whether a multi-body quantum system is in a hybrid form is to determine whether it has a tensor decomposition form. In quantum information theory, mutual information obtained using von Neumann entropy is often used to measure the total correlation of a multi-body quantum system, the von Neumann entropy of a single-body quantum system under the ensemble description is used to measure the classical correlation of a multi-body quantum system, and their differences are considered to be non-classical correlations, namely strong correlations (Vesperini, Bel-Hadj-Aissa, and Franzosi 2022; Schindler, §afránek, and Aguirre 2020). A detailed instruction of measuring strong correlation is provided below. The analysis of strong correlation mining by QiMLP is given in the experimental section.

Method for Measuring Strong Correlations: In quantum information theory, von Neumann entropy is an expansion of the concept of Gibbs entropy of the classical system (Jaeger 2007; Wilde 2013; Nielsen and Chuang 2001), which is defined as

$$S(\rho) = -tr(\rho \log \rho), \tag{7}$$

where $tr(\cdot)$ represents the trace operation of the matrix, and ρ refers to the density matrix form of the quantum system. Implementing the eigenstate vector decomposition of the density matrix ρ,

$$\rho = \sum_{i=1}^{n} \lambda_i |\psi_i\rangle\langle\psi_i|, \tag{8}$$

then the von Neumann entropy can be explicitly written as

$$S(\rho) = -\sum_{i=1}^{n} \lambda_i \log \lambda_i. \tag{9}$$

Mutual information is a measure of the **total correlation** of multi-body quantum systems (Nielsen and Chuang 2001), which in some scenarios means that **for the composite system, half of the correlation is quantum and the other half is classical** (Hirvensalo and Yakaryılmaz 2019). Based on von Neumann entropy, the mutual information of the two-body quantum system ρ_{AB} can be defined as

$$I(\rho_{AB}) = S(\rho_A) + S(\rho_B) - S(\rho_{AB}). \tag{10}$$

When there is no quantum correlation for a two-body quantum system ρ_{AB}, there is a form of tensor decomposition, $\rho_{AB} = \rho_A \otimes \rho_B$. A two-body quantum system whose subsystems are independent of each other can be described as a superposition system using an ensemble method,

$$\rho_{(A+B)} = \sum_{i=A,B} p_i \rho_i, \tag{11}$$

where p_i refers to probability, and $p_A + p_B = 1$. Then the correlation of the quantum system $\rho_{(A+B)}$ can be obtained from the von Neumann entropy, and since it is a non-entangled single quantum system, the correlation existing in the quantum system can be considered as a **classical correlation**. And when ρ_A and ρ_B exist with equal probability, i.e., $p_A = p_B$, their classical correlation is the strongest.

At this point, **non-classical strong correlation** in multi-body quantum systems can be defined as the total correlation reduced by the classical correlation (Hirvensalo and Yakaryılmaz 2019). For a two-body quantum system ρ_{AB}, its strong correlation can be described as

$$\begin{aligned} C(\rho_{AB}) &= I(\rho_{AB}) - S(\rho_{(A+B)}) \\ &= S(\rho_A) + S(\rho_B) - S(\rho_{AB}) - S(\rho_{(A+B)}). \end{aligned} \tag{12}$$

[1] In Dirac notation, a unit vector $\vec{v}$ and its transpose $\vec{v}^T$ are denoted as a ket $|v\rangle$ and a bra $\langle v|$ respectively.

Experiments

This section will organize experiments to test the performance of QiMLP in terms of accuracy of model prediction results, parameter compression, and strong correlation mining, specifically answering the following four questions: **Q1**. Under similar structural design, is QiMLP better than the baseline model in terms of accuracy of prediction results? **Q2**. And to what extent can it compress the parameters of the model? **Q3**. Can QiMLP mine strong correlations among features and improve the prediction accuracy of the model accordingly? **Q4**. Can QiMLP be generalized to more tasks and adapted to wider fields?

Analysis of External Performance (RQ1 & RQ2)

Datasets: MINST[2] and CIFAR-10[3] are widely used datasets for testing new models in the early days of deep learning. MINST is a handwritten digit picture dataset containing a training set of 60,000 pictures and a test set of 10,000 pictures, and each sample is a 28×28 pixel gray matter picture of 0∼9 (LeCun et al. 1998). CIFAR-10 is a real object image dataset containing a training set of 50,000 images and a test set of 10,000 images, and each sample is a 32×32 pixel color image divided into 10 categories, with 6,000 images in each category (Krizhevsky 2009). The training set of the above datasets is divided into our training set and development set at a ratio of 8:2, and their test set is directly used as our test set. The measurement metrics of experimental results all adopt accuracy rate.

Settings: Under the MNIST dataset, the model structure of the control group is a classic MLP composed of **three neuron layers**. The number of neurons in each layer is 784, the activation function $GELU(\cdot)$ is added to the first two layers, and the function $Dropout(\cdot)$ is added to the third layer. The model structure of the experimental group only modified the weight matrices and bias terms of the control group model into the form of two-body, three-body, and four-body quantum systems, and ensured that the dimensions of the weight matrices and bias terms of the experimental group model are the same as those of the control group model. Under the two-body quantum system, the weight matrix is composed of (28×28) and (28×28) dimensional matrices; Under the three-body quantum system, the weight matrix is composed of (7×7), (7×7) and (16×16) dimensional matrices; Under the four-body quantum system, the weight matrix is composed of (4×4), (4×4), (7×7) and (7×7) dimensional matrices. When forming a composite matrix from multiple sub-matrices, our principle of selecting each sub-matrix is to ensure that all sub-matrices are maximized as much as possible. The construction methods of the bias terms of the two-body, the three-body and the four-body quantum systems are similar to those of the weight matrices. The specific construction method is shown in Eq. (4). In addition, the experimental group model does not add the function $Dropout(\cdot)$ to the third layer.

[2]http://yann.lecun.com/exdb/mnist/

[3]https://www.cs.toronto.edu/ kriz/cifar.html

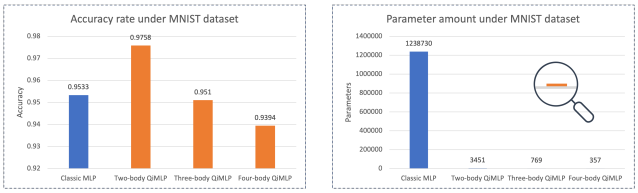

Figure 2: The accuracy of experimental results and the number of model parameters of the classic MLP and the QiMLP under the MNIST dataset.

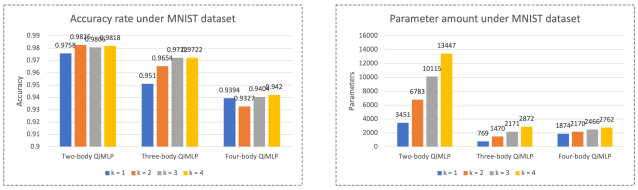

Figure 3: The accuracy of experimental results and the number of model parameters of the QiMLP at different k values under the MNIST dataset.

Under the CIFAR-10 dataset, the model structure of the control group is a classic MLP composed of **five neuron layers**. The model structure of the experimental group is also designed with reference to the model structure of the control group, and the construction method is the same as the construction method of the experimental group model under the MNIST dataset. The activation functions of the models in the CIFAR-10 dataset are consistent with those in the MNIST dataset.

For settings such as the operating platform and hyperparameters, please refer to appendix and original code.

Experimental Results and Analysis: Under the MNIST dataset, the experimental results of the experimental group model (QiMLP under two-body, three-body, and four-body quantum systems) and the control group model (classic MLP) are shown in Fig. 2. The left picture shows the accuracy of the model prediction results, and the right picture shows the number of model parameters. From the perspective of accuracy of prediction results, QiMLP based on the two-body quantum system is far better than the control group model; QiMLP based on the three-body quantum system is similar to the control group model; QiMLP based on the four-body quantum system is not as good as the control model. From the perspective of the number of model parameters, QiMLP based on various quantum system configurations is far less than the number of parameters used by the control model. Combined with the number of model parameters, the accuracy analysis of the prediction results of QiMLP constructed by three-body and four-body quantum systems shows that the main reason for their low accuracy is that the number of parameters of the models are too small, which is not enough to learn better performance. It can be inferred from this conclusion that QiMLP has strong parameter compression capabilities, but in specific response scenarios, the information capacity of the model needs to be con-

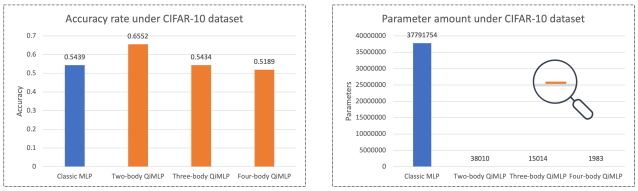

Figure 4: The accuracy of experimental results and the number of model parameters of the classic MLP and the QiMLP under the CIFAR-10 dataset.

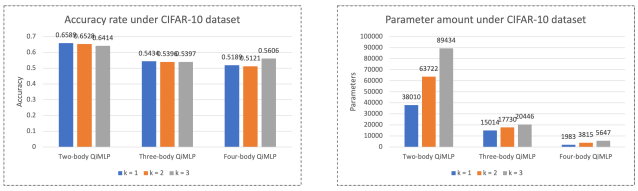

Figure 5: The accuracy of experimental results and the number of model parameters of the QiMLP at different k values under the CIFAR-10 dataset.

sidered and parameters cannot be excessively compressed. The index k in Eq. (4) refers to the number of repetitions of the reconstructed weight matrix, which is used to alleviate excessive compression of model parameters. The experimental results with different values of k are shown in Fig. 3. Overall, when k takes a larger value, the model parameters increase and the prediction results of the model improve accordingly. However, the performance of QiMLP based on the four-body quantum system is average. The main reason is that the number of parameters is still insufficient and the model parameters are over-compressed.

Under the CIFAR-10 dataset, the experimental results of the experimental group model and the control group model are shown in Fig. 4. The experimental results are generally similar to those under the MNIST dataset, indicating that the parameter compression capabilities of QiMLP also perform well on larger datasets and the most complex tasks. The index k in Eq. (4) refers to the number of repetitions of the reconstructed weight matrix, which is used to alleviate excessive compression of model parameters. The experimental results with different values of k are shown in Fig. 5. Overall, when k takes a larger value, the model parameters increase, but the prediction results of the models do not improve accordingly. The main reason is that the number of parameters is still insufficient and the model parameters are over-compressed.

An analysis of the training efficiency, learning speed and main hyperparameters of the models is given in appendix.

Analysis of Internal Conditions (RQ3)

Based on the formula for calculating strong correlation given in Eq. (12), this section will statistically analyze whether QiMLP can mine strong correlations among features during the implementation process. Since it is difficult for multi-body quantum systems to calculate mutual information, that is, the total correlation of multi-body quantum systems, this

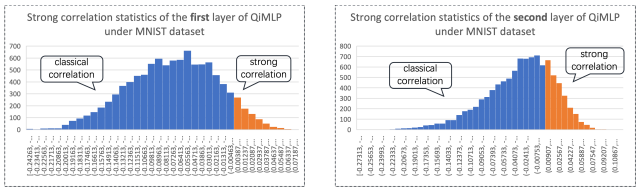

Figure 6: The proportion of classical correlations and mined strong correlations in each neuron layer of QiMLP under the MNIST dataset.

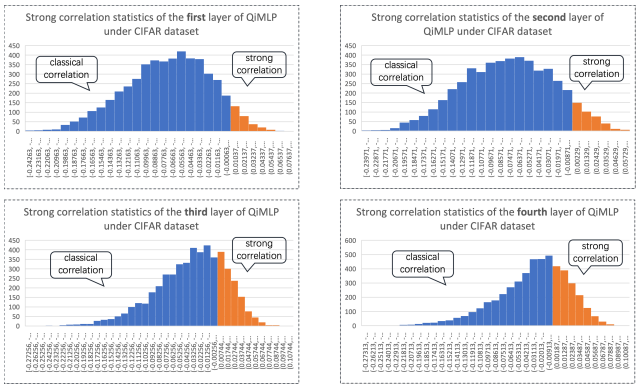

Figure 7: The proportion of classical correlations and mined strong correlations in each neuron layer of QiMLP under the CIFAR-10 dataset.

section only analyzes QiMLP based on two-body quantum systems. We trained a neural network to obtain the Kronecker product form $\rho_A \otimes \rho_B$ of the output vector ρ_{AB} of each neuron layer of QiMLP, that is, let $\rho_A \otimes \rho_B$ be only as similar as possible to ρ_{AB}. Among them, ρ_A, ρ_B and ρ_{AB} are all represented by matrices of vectors under the computational basis. After obtaining the subsystem of each neuron layer in QiMLP based on two-body quantum systems, its strong correlation can be calculated with the help of Eq. (12). The statistical results of the strong correlation of each neuron layer of QiMLP in the MNIST test set and the CIFAR-10 test set are shown in Fig. 6 and Fig. 7 respectively. Since the last layer is the output layer, the dimension is determined by the prediction task, and it is not suitable for counting its strong correlation.

Comparing the statistical results of strong correlations between the first neuron layer and the second neuron layer in Fig. 6, it can be seen that there are more strong correlations mined in the second layer than in the first layer. This shows that QiMLP can mine strong correlations among features, and as the number of hidden layers of the model increases, the strong correlations mined tend to gradually increase. This phenomenon also appears in QiMLP under the CIFAR-10 dataset. The above experimental results illustrate that the core reason why QiMLP can still achieve good prediction results despite significantly compressing model parameters is that it mines strong correlations among features, and the strong correlations mined can directly determine the accuracy of model prediction results.

Analysis of Model Potential (RQ4)

Datasets: The movie review dataset IMDb[4] is a widely used natural language processing dataset that contains 50,000 highly polarizing reviews from the Internet Movie Database (Pal, Barigidad, and Mustafi 2020). The dataset is divided into 25,000 reviews for training and 25,000 reviews for testing, and both the training set and the test set contain 50% positive reviews and 50% negative reviews respectively. Reviews range in length from 900 to 1700, with an average length of 1100, and are preprocessed into the form of words and punctuation marks. NewsGroups[5] is one of the international standard datasets used for text classification, text mining and information retrieval research (Lang 1995). The dataset collects about 20,000 newsgroup documents, evenly divided into 20 newsgroup collections with different topics. The dataset is a 20-category classification dataset, in which the training set and test set are divided into proportions of 75% and 25%. The training set of the above datasets is divided into our training set and development set at a ratio of 8:2, and their test set is directly used as our test set. The measurement metrics of experimental results all adopt accuracy rate.

Settings: Under the IMDb dataset, the model structure of the control group is built based on the pre-trained language model BERT (Koroteev 2021) and three neuron layers. Among them, the model BERT is mainly responsible for the extraction of text features, and MLP constructed with three neuron layers is responsible for feature fusion and final emotion classification. Model BERT uses the $bert\text{-}base\text{-}uncased$ version. The number of neurons in each layer is 768, the activation function $GELU(\cdot)$ is added to the first two layers, and the function $Dropout(\cdot)$ is added to the third layer. The model structure of the experimental group only modified the weight matrices and bias terms of the classical MLP of the control group model into the form of a two-body quantum system, and a three-body quantum system, and ensured that the weight matrices and bias terms of the experimental group model are the same as the control group model. The specific construction method is shown in Eq. (4). In addition, the experimental group model does not add the function $Dropout(\cdot)$.

Under the NewsGroups dataset, the model structure of the control group is an MLP composed of five neuron layers, and the input text uses the traditional TF-IDF method to obtain its word representation. The model structure of the experimental group is also designed with reference to the model structure of the control group, and the construction method is the same as the construction method of the experimental group model under the MNIST dataset. The activation functions of the models in the NewsGroups dataset are consistent with those in the MNIST dataset.

For settings such as the operating platform and hyperparameters, please refer to appendix and original code.

Experimental Results and Analysis: Under the IMDb dataset, the experimental results of the experimental group model (QiMLP under two-body, and three-body quantum systems) and the control group model (classic MLP) are shown in Fig. 8. From the performance of the experimental results, QiMLP based on the two-body quantum system is better than the classical MLP, while QiMLP based on the three-body quantum system is similar to the classical MLP, indicating that QiMLP can still achieve good performance on larger-scale natural language processing datasets.

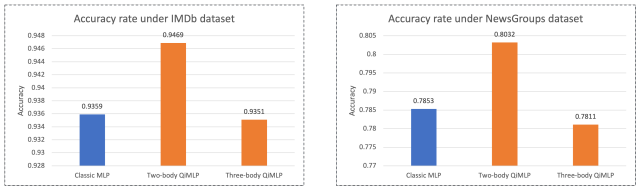

Figure 8: The accuracy of experimental results of the classic MLP and the QiMLP on the IMDb and NewsGroups datasets.

Under the NewsGroups dataset, the experimental results of the experimental group model and the control group model are shown in Fig. 8. From the performance of the experimental results, QiMLP based on the two-body quantum system is better than the classical MLP, while QiMLP based on the three-body quantum system is similar to the classical MLP, indicating that QiMLP can still achieve good performance on more complex natural language processing tasks.

Limitations

QiMLP is excellent in terms of model parameter compression, memory usage, and reasoning accuracy, but they also have some limitations. Compared with the classical MLP, the training time of QiMLP will be longer and the convergence speed will be slower. In addition, excessive model parameter compression will lead to a decrease in model information capacity and failure to learn knowledge.

Conclusion

Deep learning models and large language models based on neural networks have achieved unprecedented results in the field of artificial intelligence, but they also have inherent limitations, such as parameter overload and poor learning ability of strong correlations among features. Inspired by quantum theory, this paper constructs a neuron layer with model parameter compression and strong correlation mining. Compared with the traditional neuron layer, the neuron layer proposed in this paper can achieve hundreds of times of parameter compression, and can mine strong correlations among features to improve the representation ability of the neuron layer. In addition, we built a Multi-Layer Perceptron (MLP) based on the neuron layer, called quantum-inspired MLP (QiMLP), to implement verification in more tasks and fields. Experimental results in image classification, text classification and text emotion recognition show that QiMLP not only achieve a large proportion of parameter compression, but also achieve better prediction results.

[4]https://developer.imdb.com/non-commercial-datasets/

[5]http://qwone.com/ jason/20Newsgroups/

Acknowledgments

Our work is supported by the National Key Research and Development Program of China (No. 2022YFC3600902) and the National Scienceand Technology Major Project (No. 2018AAA0101904).

References

Adesso, G.; Bromley, T. R.; and Cianciaruso, M. 2016. Measures and applications of quantum correlations. *Journal of Physics A: Mathematical and Theoretical*, 49.

Almeida, L. B. 2020. Multilayer perceptrons. In *Handbook of Neural Computation*, C1–2. CRC Press.

Becke, A. D. 2013. Density functionals for static, dynamical, and strong correlation. *The Journal of Chemical Physics*, 138(7).

Broxson, B. J. 2006. The kronecker product.

Ceravolo, R.; Coletta, G.; Miraglia, G.; and Palma, F. 2021. Statistical correlation between environmental time series and data from long-term monitoring of buildings. *Mechanical Systems and Signal Processing*, 152: 107460.

Clauser, J. F.; and Shimony, A. 1978. Bell's theorem. Experimental tests and implications. *Reports on Progress in Physics*, 41(12): 1881.

Duan, Y.; Wang, L.; Zhang, Q.; and Li, J. 2022. FactorVAE: A Probabilistic Dynamic Factor Model Based on Variational Autoencoder for Predicting Cross-Sectional Stock Returns. In *AAAI*, 4468–4476.

Fetter, A. L.; Walecka, J. D.; and Kadanoff, L. P. 1971. *Quantum Theory of Many-Particle Systems*. Courier Corporation.

Galindo, A.; and Martin-Delgado, M. A. 2002. Information and computation: Classical and quantum aspects. *Reviews of Modern Physics*, 74(2): 347.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2020. Knowledge Distillation: A Survey. *International Journal of Computer Vision*, 129: 1789 – 1819.

Higgins, I.; Amos, D.; Pfau, D.; Racaniere, S.; Matthey, L.; Rezende, D.; and Lerchner, A. 2018. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C. P.; Glorot, X.; Botvinick, M. M.; Mohamed, S.; and Lerchner, A. 2016. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *ICLR (Poster)*, volume 3.

Hirvensalo, M.; and Yakaryılmaz, A. 2019. Quantum Computing and Quantum Information.

Holevo, A. S.; and Ballentine, L. E. 1982. *Probabilistic and Statistical Aspects of Quantum Theory*. Springer Science & Business Media.

Jaeger, G. 2007. *Quantum information*. Springer.

Jiang, Y.; Zhang, E. Z.; Tian, K.; Mao, F.; Gethers, M.; Shen, X.; and Gao, Y. 2010. Exploiting statistical correlations for proactive prediction of program behaviors. In *Proceedings of the 8th annual IEEE/ACM international symposium on Code generation and optimization*, 248–256.

Koroteev, M. V. 2021. BERT: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. Toronto, ON, Canada.

Kruse, R.; Mostaghim, S.; Borgelt, C.; Braune, C.; and Steinbrecher, M. 2022. Multi-layer perceptrons. In *Computational intelligence: a methodological introduction*, 53–124.

Lang, K. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, 331–339.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86: 2278–2324.

Lu, Y. 2019. Artificial intelligence: a survey on evolution, models, applications and future trends. *Journal of Management Analytics*, 6(1): 1–29.

Marinescu, D. C. 2011. *Classical and quantum information*. Academic Press.

Murtagh, F. 1991. Multilayer perceptrons for classification and regression. *Neurocomputing*, 2: 183–197.

Nielsen, M. A.; and Chuang, I. L. 2001. *Quantum computation and quantum information*, volume 2. Cambridge university press Cambridge.

Pal, A.; Barigidad, A.; and Mustafi, A. 2020. IMDb Movie Reviews Dataset.

Peres, A. 1997. *Quantum theory: concepts and methods*, volume 72. Springer.

Popescu, M.; Balas, V. E.; Perescu-Popescu, L.; and Mastorakis, N. E. 2009. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems archive*, 8: 579–588.

Rabanser, S.; Shchur, O.; and Günnemann, S. 2017. Introduction to Tensor Decompositions and their Applications in Machine Learning. *ArXiv*, abs/1711.10781.

Schindler, J. C.; §afránek, D.; and Aguirre, A. 2020. Quantum correlation entropy. *Physical Review A*, 102: 052407.

Stasiuk, A. 2021. *Computational and Theoretical Insights into Multi-Body Quantum Systems*. Master's thesis, University of Waterloo.

Vesperini, A.; Bel-Hadj-Aissa, G.; and Franzosi, R. 2022. Entanglement and quantum correlation measures for quantum multipartite mixed states. *Scientific Reports*, 13.

Wang, X.; Chen, H.; Zhou, Y.; Ma, J.; and Zhu, W. 2022. Disentangled representation learning for recommendation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 408–424.

Wilde, M. M. 2013. *Quantum information theory*. Cambridge university press.

Zhang, J.; He, R.; and Guo, F. 2023. Quantum-Inspired Representation for Long-Tail Senses of Word Sense Disambiguation. In *AAAI*, 13949–13957.

Zhang, J.; He, R.; Guo, F.; and Liu, C. 2024. Quantum Interference Model for Semantic Biases of Glosses in Word Sense Disambiguation. In *AAAI*, 19551–19559.

Zhang, J.; He, R.; Li, Z.; Zhang, J.; Wang, B.; Li, Z.; and Niu, T. 2021. Quantum Correlation Revealed by Bell State for Classification Tasks. In *IJCNN*, 1–8.

Zhang, J.; Hou, Y.; Li, Z.; Zhang, L.; and Chen, X. 2020. Strong Statistical Correlation Revealed by Quantum Entanglement for Supervised Learning. In *European Conference on Artificial Intelligence*, volume 325, 1650–1657.

Zhang, J.; Li, Z.; Peng, H.; Li, M.; and Wang, X. 2022a. Feedforward Neural Network Reconstructed from High-order Quantum Systems. In *IJCNN*, 1–8.

Zhang, J.; Li, Z.; Wang, J.; Wang, Y.; Hu, S.; Xiao, J.; and Li, Z. 2022b. Quantum Entanglement Inspired Correlation Learning for Classification. In *PAKDD*, volume 13281, 58–70.

Zhang, J.; Li, Z.; Xiao, J.; and Li, M. 2022c. Neural Network Model Reconstructed from Entangled Quantum States. In *IJCNN*, 1–8.

Zhou, X.; Zhang, W.; Xu, H.; and Zhang, T. 2021. Effective Sparsification of Neural Networks with Global Sparsity Constraint. *CVPR*, 3598–3607.

Zhou, Y.; Moosavi-Dezfooli, S.-M.; Cheung, N.-M.; and Frossard, P. 2018. Adaptive Quantization for Deep Neural Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Appendix

Operating platform and hyperparameters of the models

Under the MNIST dataset, the hardware platform on which the models run is the Ubuntu operating system, including 8 NVIDIA A100 GPUs; the model code is based on the Python 3.8.6 language and developed by the Pytorch 2.1.0 framework; the hyperparameters of the models are set to: *learning rate* is *[0.001-0.005]*, *training batch* is *50*, *epoch* is *20*, *input size* is *784*, *hidden size* is *784*, *output size* is *10*, *dropout rate* is *0.1*, and the number of copies of the weight matrix k in Eq. (4) is *1*.

Under the CIFAR-10 dataset, the hardware platform on which the models run is the Ubuntu operating system, including 8 NVIDIA A100 GPUs; the model code is based on the Python 3.8.6 language and developed by the Pytorch 2.1.0 framework; the hyperparameters of the models are set to: *learning rate* is *[0.005-0.0005]*, *training batch* is *50*, *epoch* is *20*, *input size* is *3072*, *hidden size* is *3072*, *output size* is *10*, *dropout rate* is *0.1*, and the number of copies of the weight matrix k in Eq. (4) is *1*.

Under the IMDb dataset, the hardware platform on which the models run is the Ubuntu operating system, including 8 NVIDIA A100 GPUs; the model code is based on the Python 3.8.6 language and developed by the Pytorch 2.1.0 framework; the hyperparameters of the models are set to: *learning rate* is *[0.00001-0.00002]*, *training batch* is *128*, *epoch* is *5*, *input size* is *768*, *hidden size* is *768*, *output size* is *2*, *dropout rate* is *0.1*, and the number of copies of the weight matrix k in Eq. (4) is *1*.

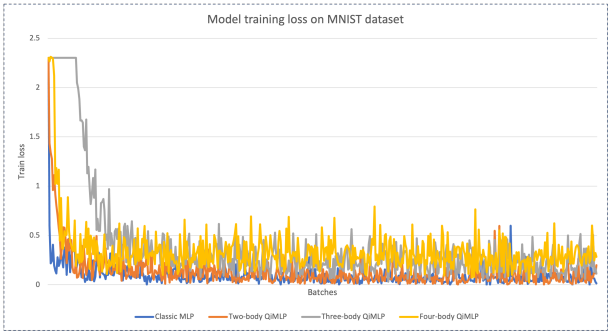

Figure 9: Training losses of the models under the MNIST dataset.

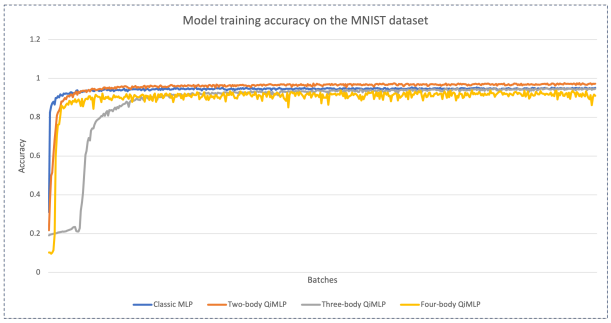

Figure 10: The training accuracy of the models under the MNIST dataset.

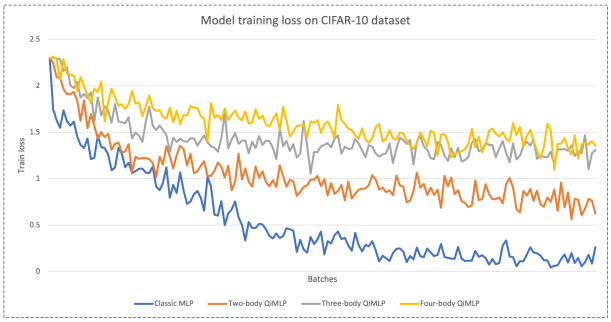

Figure 11: Training losses of the models under the CIFAR-10 dataset.

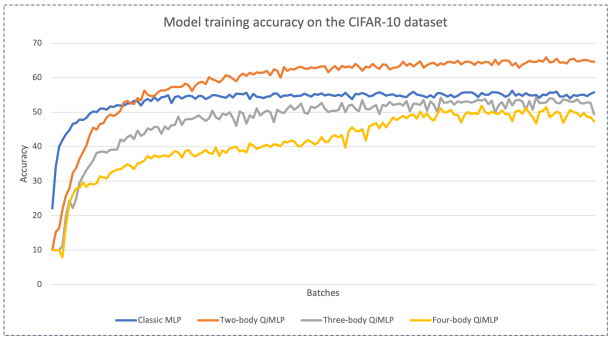

Figure 12: The training accuracy of the models under the CIFAR-10 dataset.

	Classic MLP	Two-body QiMLP	Three-body QiMLP	Four-body QiMLP
Used time	296 s	312 s	335 s	350 s
GPU memory	915 MiB	558 MiB	387 MiB	157 MiB

Table 1: Model training time and GPU memory usage under MNIST dataset

	Classic MLP	Two-body QiMLP	Three-body QiMLP	Four-body QiMLP
Used time	319 s	350 s	362 s	377 s
GPU memory	1524 MiB	929 MiB	728 MiB	512 MiB

Table 2: Model training time and GPU memory usage under CIFAR-10 dataset

Under the NewsGroups dataset, the hardware platform on which the models run is the Ubuntu operating system, including 8 NVIDIA A100 GPUs; the model code is based on the Python 3.8.6 language and developed by the Pytorch 2.1.0 framework; the hyperparameters of the models are set to: *learning rate* is *[0.001-0.005]*, *epoch* is *500*, *input size* is *5000*, *hidden size* is *5000*, *output size* is *20*, *dropout rate* is *0.1*, and the number of copies of the weight matrix k in Eq. (4) is *1*.

Analysis of model training efficiency under MNIST dataset

Under the MNIST dataset, the training losses of the models are shown in Fig. 9. It can be seen from the loss diagram that the classic MLP converges faster, while QiMLP is relatively slower. This phenomenon is in line with expectations, because it is more difficult for QiMLP to learn strong correlations among features, so the convergence speed of the models will be relatively slower. In addition, it can be seen from this diagram that the convergence effect of QiMLP of the two-body quantum system is better than that of the classical MLP.

Under the MNIST dataset, the training accuracy of the models is shown in Fig. 10. It can be seen from the accuracy diagram that the classic MLP achieves the best results faster, while QiMLP is relatively slower. This phenomenon is expected, because it is more difficult for QiMLP to learn strong correlations among features, so it is more difficult to achieve the best results. In addition, it can be seen from this diagram that the convergence speed of QiMLP of the two-body quantum system is better than that of the classical MLP, and the convergence speed of QiMLP of the three-body quantum system is similar to that of the classical MLP.

Under the MNIST dataset, the usage time and GPU memory of the models trained for the same number of epochs is shown in Tab. 1. As can be seen from the table, the classical MLP takes the least time compared to QiMLP on the same dataset and the same epoch. The reason for this phenomenon is that it is more difficult for QiMLP to learn strong correlations among features. Fortunately, QiMLP takes up less GPU memory.

Analysis of model training efficiency under CIFAR-10 dataset

Under the CIFAR-10 dataset, the training losses of the models are shown in Fig. 11. It can be seen from this diagram that the classic MLP has the best convergence effect, while QiMLP has a relatively poor convergence effect. This phenomenon shows that QiMLP needs to learn strong correlations and has fewer parameters, so the convergence effect will be worse. However, combined with the accuracy diagram, it can be seen that the QiMLP can achieve better performance, indicating that its learning ability is stronger.

Under the CIFAR-10 dataset, the training accuracy of the models is shown in Fig. 12. It can be seen from this diagram that the classical MLP converges faster, while QiMLP converges relatively slowly. However, the final performance obtained by QiMLP based on the two-body quantum system is better than the classical MLP; the final performance obtained by QiMLP based on the three-body quantum system is similar to the classic MLP. This phenomenon shows that QiMLP is effective and has better learning capabilities.

Under the CIFAR-10 dataset, the usage time and GPU memory of the models trained for the same number of epochs is shown in Tab. 2. It can be seen from this table that the training time of QiMLP will be longer than that of classical MLP under the same data and the same training epoch. The reason for this phenomenon is that it takes more time for QiMLP to learn more complex relationships among features. Notably, QiMLP uses less GPU memory.