Enhancing the expressivity of quantum neural networks with residual connections

Check for updates

Jingwei Wen ⊠, Zhiguo Huang, Dunbo Cai & Ling Qian ⊠

In noisy intermediate-scale quantum era, the research on the combination of artificial intelligence and quantum computing has been greatly developed. Here we propose a quantum circuit-based algorithm to implement quantum residual neural networks, where the residual connection channels are constructed by introducing auxiliary qubits to data-encoding and trainable blocks in quantum neural networks. We prove that when this particular network architecture is applied to a l-layer data-encoding, the number of frequency generation forms extends from one, namely the difference of the sum of generator eigenvalues, to $\mathcal{O}(l^2)$, and the flexibility in adjusting Fourier coefficients can also be improved. It indicates that residual encoding can achieve better spectral richness and enhance the expressivity of various parameterized quantum circuits. Extensive numerical demonstrations in regression tasks and image classification are offered. Our work lays foundation for the complete quantum implementation of classical residual neural networks and offers a quantum feature map strategy for quantum machine learning.

Quantum computing is a new computing paradigm based on quantum mechanics that utilizes qubits instead of classical bits to store and process information[1]. Since the theoretical concepts were proposed[2–4], quantum computers have developed at an astonishing speed, gradually moving from the milestone achievement like quantum supremacy in the laboratory[5–7] to the stage of proof-of-principle application exploration[8–10]. Among its many applications, quantum machine learning is an emerging field that leverages the power of quantum computers to overcome bottlenecks of high computing power requirements in the machine learning[11–14]. On the current noisy intermediate-scale quantum devices[15], one popular strategy for constructing quantum machine learning algorithms is using classical-quantum hybrid optimization loops to train the parameterized quantum circuits for various learning tasks, such as pattern recognition[16,17] and classification[18–22].

Similar to the classical neural networks that consist of input layers, hidden layers and output layers, the fundamental structures of the variational quantum neural networks include data-encoding circuits, variational ansatz, and output layers realized by quantum measurement[23,24]. To be specific, the data-encoding or quantum feature map processes $\mathcal{U}(x)$ can map the classical data $x \in \chi$ to a quantum state in Hilbert space $\mathcal{H}$. It serves as one of the main sources of non-linearity for the networks, and there exist numerous encoding strategies such as amplitude encoding and angle encoding[25]. Moreover, different choices of architectures for the variational ansatz $\mathcal{W}(\theta)$ containing trainable parameters θ will lead to various quantum neural networks[26–36] and it will greatly affect the network performance such

as generalization[37,38] and trainability[39,40]. For example, general deep parameterized quantum circuits suffer from the barren plateau phenomenon, leading to vanishing gradients[40–44]. But it can be avoided by networks with hierarchical structure, proposed as a realization of the quantum convolutional neural networks (QCNN)[20,27,45], which has been proved the absence of barren plateaus[46]. Finally, the output of an n-qubit quantum neural networks is the mean value of a measurable observable O as

$$f(x, \theta) = \langle \psi_0 | U_\theta^\dagger(x) O U_\theta(x) | \psi_0 \rangle \tag{1}$$

where initial state $|\psi_0\rangle = |0\rangle^{\otimes n}$ and $U_\theta(x)$ is the parameterized quantum circuit consisting repeatable data-encoding and trainable blocks. Interestingly, the expressivity and universality of such variational quantum models can be guaranteed by the fact that one can naturally write the outputs as partial Fourier series in the network inputs[47–50], and the accessible frequencies are determined by the eigenvalues of the generator Hamiltonian in the data-encoding gates, while the coefficients are controlled by the design of the entire circuits[50].

A great deal of research work has subsequently devoted to advancing the quantum neural networks, with one intuitive approach being the quantization of classical networks[31–34]. Especially, inspired by the classical residual neural networks, which are proposed for alleviating the vanishing gradient problem during the training process of deep neural networks[51], its quantum counterpart is promising to mitigating barren plateaus[34]. The key

China Mobile (Suzhou) Software Technology Company Limited, Suzhou 215163, China. ⊠e-mail: wjw17@tsinghua.org.cn; qianling@cmss.chinamobile.com

idea is to introduce residual connections into the traditional neural networks, as shown in the Fig. 1. Mathematically, the residual connections can provide an additional cross-layer propagation channel for the input features, leading to a basic residual unit of neural networks as $\mathcal{H}(x) = \mathcal{F}(x) + x$, where the non-linear parameterized function $\mathcal{F}(x)$ represents the traditional neural networks. Although there exist some works on the quantum realization of residual neural networks, the residual channels are usually implemented using classical or hybrid methods[34,52]. The researches on the full quantum implementations of residual connections and effects on the expressivity are still very lacking.

In this work, we address these issues by proposing a quantum circuit-based algorithm to implement quantum residual neural networks (QResNets). The residual connection channel is constructed through one ancillary qubit and the target evolution process is embedded in the subspace. Such structures are compatible to both the data-encoding and trainable blocks in the variational quantum neural networks. We also further parameterize the encoding gates on the auxiliary qubit and obtain the generalized residual operators. Furthermore, we find that the Fourier spectrum of the output of parameterized quantum circuits can be enriched when the residual connections are used for the data-encoding blocks. The number of frequency combinations forms can be extended from one, namely the difference of the sum of generator eigenvalues, to $\mathcal{O}(l^2)$ for the l-layer residual encoding. Moreover, the diverse construction methods for frequencies in the residual outputs and the extra trainable parameters in the generalized residual operators can expand the Fourier coefficient space. The results suggest that the expressivity of quantum models can be enhanced by residual connections. We offer extensive numerical demonstrations of the quantum algorithm in the regression tasks by function fitting of Fourier series, and also present the performance of binary classification with standard MNIST datasets to recognize the handwritten digits images[53], achieving an accuracy improvement of over 7% with residual encoding. Our results show that the residual connections proposed in classical deep learning for improving trainability can also be used to improve the expressivity in quantum neural networks, making it a promising quantum learning model for real-life applications.

Results

Realization of quantum residual connection

In the QResNets, there are multiple layers of repeatable data-encoding block $\mathcal{U}(x)$ and trainable parameterized ansatz $\mathcal{W}(\theta)$, and the residual connections can be adopted in some of the blocks, as shown in the Fig. 1. The data-encoding block consists of quantum rotation gates of the form $U(x) = e^{iHx}$ where H is a generator Hamiltonian, while the trainable circuits are composed of single- and two-qubit parameterized quantum gates $W(\theta)$ with optimization parameters θ. Some gates in the data-encoding and ansatz block can be sampled to add residual connections forming quantum

residual operators $\mathcal{R}(x)$ and $\mathcal{R}(\theta)$, which correspond to the residual evolution processes. We introduce a unified notation $\diamond$ which has $\diamond = x$ for quantum gates in the data-encoding blocks while $\diamond = \theta$ in the trainable blocks. Then for an n-qubit quantum system with initial state $|\phi_0\rangle$, the evolution under residual operator can be expressed as

$$\mathcal{R}(\diamond)|\phi_0\rangle = \frac{1}{2}\left(\sigma_0^{\otimes n} + \mathcal{L}(\diamond)\right)|\phi_0\rangle \quad (2)$$

where σ_0 is the identity matrix and $\mathcal{L}(x) = U(x)$ in the quantum feature map block and $\mathcal{L}(\theta) = W(\theta)$ in optimization ansatz. Such an evolution operator can be realized by the frame of linear combination of unitary with one ancillary qubit, and the target quantum states are obtained by post-processing[54,55]. Specifically, we first apply a Hadamard gate to encode the ancillary system followed by a controlled-$\mathcal{L}(\diamond)$ operator. After adding another Hadamard gate, we can measure the ancillary qubit with results $m_a = 0/1$ corresponding to quantum states $|0\rangle/|1\rangle$. Then the evolution results under residual operators can be obtained in the $|0\rangle\langle0|$ subspace. The introduction of an auxiliary qubit provides an additional channel that allows the unevolved quantum state to pass alone and add to the evolved quantum state.

More generally, the weight of the summation process can also be adjusted by replacing the first Hadamard gate on the ancillary qubit with $R_y(2\alpha)$ rotation with trainable angles α. Then the corresponding residual operator is generalized as a single optimization-angle residual operator

$$\mathcal{R}_1(\diamond) = \frac{\cos\alpha\,\sigma_0^{\otimes n} + (-1)^{m_a}\sin\alpha \cdot \mathcal{L}(\diamond)}{\sqrt{2}} \quad (3)$$

Such a construction does not require a post-selection process, but rather reconstructs the target operator from the measurement results. It can be reduced to $\mathcal{R}(\diamond)$ with $\alpha = \pi/4$ and $m_a = 0$. Similarly, a two optimization-angles residual operator $\mathcal{R}_2(\diamond)$ can also be constructed by replacing both Hadamard gates with parameterized rotation gates, and the detail is shown in Methods section. In principle, the introduction of more trainable parameters in these two generalized residual operators will provide additional degrees of freedom for optimization, which can further increase the expressivity of the parameterized quantum circuits.

Therefore, we can conclude here that a general residual connection in quantum neural networks can be realized in the complete quantum circuit frame. It is also worth noting that in some special network structures such as the QCNN[27], by reusing discarded qubits, we can simulate the residual connections without additional qubits. Moreover, due to the fact that the expressivity of quantum models is fundamentally limited by the data-encoding strategy, we will prove below that the residual connections applied
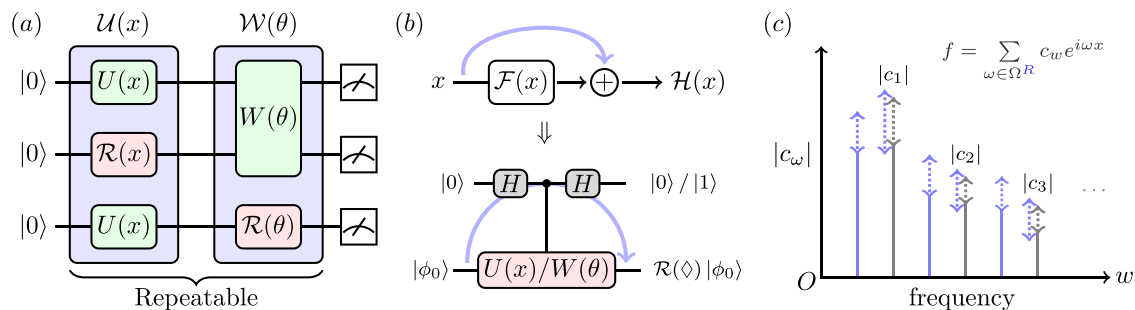
Fig. 1 | Quantum residual neural networks. a A schematic of the quantum neural networks with residual connections. The quantum feature map circuit $\mathcal{U}(x)$ and trainable variational circuit $\mathcal{W}(\theta)$ are repetitively implemented multiple times to form the multilayer structures. The $\mathcal{R}(x)$ and $\mathcal{R}(\theta)$ blocks labeled by red represent the data-encoding gates $U(x)$ and parameterized gates $W(\theta)$ with residual connections. **b** The classical residual unit and its quantum counterpart. The residual connection channels are shown with blue arrows, and the output of residual block is $\mathcal{H}(x) = \mathcal{F}(x) + x$, where non-linear function $\mathcal{F}(x)$ represents the classical neural networks. The quantum residual operator $\mathcal{R}(\diamond)$ (a unified expression for $\mathcal{R}(x)$ and $\mathcal{R}(\theta)$ operators) implemented on the initial state $|\phi_0\rangle$ can be realized in the subspace of an ancillary qubit with measurement results $m_a = 0/1$. **c** The residual feature map can introduce more frequency components (blue) to the original spectra of quantum neural networks (gray), and also make the Fourier expansion coefficients more flexible, whose ranges are represented by double dotted arrows.

to data-encoding block, no matter what ansatz used, will lead to a better spectra richness in the Fourier series of quantum model output, resulting an expressivity enhancement.

Frequency spectra enhancement

It has been pointed out that the output of a parameterized quantum circuit can be expressed as a finite-term Fourier series of the input features[50]

$$f(x, \theta) = \sum_{\omega \in \Omega} c_\omega(\theta, O) e^{i\omega x} \tag{4}$$

where the frequency ω of spectrum $\Omega = \{w_k - w_j | j, k \in [d]\}$ depends on the d-dimensional generator of one-layer data-encoding gate $U(x) = e^{iHx}$ with eigenequations $H|h_j\rangle = w_j|h_j\rangle$ for $j \in [d]$. Notation $[d] := \{1, 2, \cdots, d\}$ here. It means that the accessible frequency of the quantum model is constructed from the difference between the generator eigenvalues. For example, a frequently used generator is the Pauli matrix $H = \sigma/2$ with two eigenvalues $w_{1,2} = \pm 1/2$ where $\sigma = \{\sigma_x, \sigma_y, \sigma_z\}$, then such a one-layer data-encoding block would produce a frequency spectrum $\Omega = \{0, \pm 1\}$. Moreover, the expansion coefficients $c_\omega(\theta, O)$ are associated with the entire structure of the quantum circuit, including trainable parameters θ, and the observable O.

However, for a data-encoding block with residual connection, more frequency components can be involved, realizing an improvement in the circuit approximation ability. Assuming that the initial quantum state $|\phi_0\rangle$ of the residual encoding block is related to the optimization parameters θ, the residual outputs can be expressed as

$$
\begin{aligned}
f_R(x, \theta) &= \langle\phi_0|\mathcal{R}^\dagger(x)O\mathcal{R}(x)|\phi_0\rangle \\
&= \frac{1}{4}(\langle\phi_0|U^\dagger(x)OU(x)|\phi_0\rangle + \langle\phi_0|O|\phi_0\rangle \\
&\quad + 2\,\mathrm{Re}\,(\langle\phi_0|OU(x)|\phi_0\rangle))
\end{aligned} \tag{5}
$$

It is clear that the first term produces the same frequency components as the traditional encoding scheme, whereas the second term corresponds to the zero-frequency component, independent of input feature x. So the key lies in the third term. Because the eigenstates $|h_j\rangle$ of the generator Hamiltonian form a complete basis, we can then expand the initial quantum state $|\phi_0\rangle$ and the observable O as $|\phi_0\rangle = \sum_k \phi_k|h_k\rangle$ and $O = \sum_{j,k} o_{jk}|h_j\rangle\langle h_k|$. By using the equation $U(x)|h_j\rangle = e^{iw_jx}|h_j\rangle$, we can have

$$
\begin{aligned}
\langle\phi_0|OU(x)|\phi_0\rangle &= \sum_{j,k} \phi_j^*\langle h_j|o_{jk}|h_j\rangle\langle h_k|U(x)\phi_k|h_k\rangle \\
&= \sum_{j,k} (\phi_j^* o_{jk}\phi_k)e^{iw_kx}
\end{aligned} \tag{6}
$$

It can be found that this part will produce new frequency components for the quantum models, which are the eigenfrequencies of generator themselves $\pm w_k$ for $k \in [d]$, but not the differences between them. Therefore, the new spectra of the one-layer data-encoding block with residual connection is

$$\Omega_{l=1}^R = \left\{w_k - w_j, \pm w_k | j, k \in [d]\right\} \tag{7}$$

which indicates that the frequency generation forms of the quantum neural networks with residual encoding is more diverse, and the resulting Fourier spectrum in general could also be more abundant. In this case, the toy model we exemplified above will produce new spectrum $\{0, \pm 1/2, \pm 1\}$, which includes more frequency components and leads to an enhanced approximation ability for the parameterized quantum circuits.

A natural issue needs to be addressed is when will the residual encoding strategy behaves better than the traditional method. For the one-layer data-encoding block in quantum neural networks, it needs to meet the condition

that there exists frequency component $w_k \notin \Omega$ for $k \in [d]$, which implies

$$|w_j - w_l| \neq |w_k|, \forall j, l \in [d], \exists k \in [d] \tag{8}$$

Such a constraint can be satisfied in many practical cases because we usually use Pauli operators as the generator Hamiltonian.

Furthermore, for the data-encoding strategy repeated l-times either in sequence or in parallel, the traditional scheme will lead to a frequency spectrum $\Omega_l = \{(w_{j_1} + \cdots w_{j_l}) - (w_{k_1} + \cdots + w_{k_l})|j_1, \cdots, j_l, k_1, \cdots, k_l \in [d]\}$, which has only one frequency combination form, namely the difference between the sum of two sets of l frequencies[50]. However, for the residual encoding, there are more ways to construct the spectrum and the combination forms of frequencies will be more complex and diversified. Specifically, the frequency spectrum of a two-layer residual encoding is

$$
\begin{aligned}
\Omega_{l=2}^R = \Big\{ &(w_{j_1} + w_{j_2}) - (w_{k_1} + w_{k_2}), \\
&\pm(w_{j_1} + w_{j_2}), (w_{j_1} - w_{k_1}) \\
&\pm(w_{j_1} + w_{j_2} - w_{k_1}), |j_1, j_2, k_1, k_2 \in [d]\Big\}
\end{aligned} \tag{9}
$$

which contains four kinds of frequency combination forms. More frequency generation forms in general can result in a larger upper limit for the spectrum size. We can summarize by induction that for a l-layer residual encoding scheme, the number of frequency combination forms is

$$\mathcal{N}(\Omega_l^R) = (\lceil l/2 \rceil + 1)(\lfloor l/2 \rfloor + 1) \propto \mathcal{O}(l^2) \tag{10}$$

where $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ represent roundup and rounddown functions. This is a squared improvement over the traditional scheme and detail is shown in the Methods section.

In addition to enlarging the accessible frequency spectrum, residual encoding can also improve the flexibility of the corresponding Fourier coefficients, both of which determine the expressivity of a quantum model. The enhancement comes from two aspects, one is due to the introduction of additional optimization degrees of freedom in the generalized residual operators $\mathcal{R}_{1,2}(x)$, and another one is due to the more diverse construction methods of frequency and the corresponding recombination of Fourier coefficients, which means that a single frequency component can be generated from the recombination of different terms in the residual outputs. The latter one is the reason why residual operator $\mathcal{R}(x)$ can behave better than the traditional encoding strategy in expanding Fourier coefficient space without introducing additional optimization parameters. Furthermore, we may be able to understand the frequency spectrum amplification in the quantum residual models from the perspective that the classical residual networks behave like ensembles of relatively shallow networks[56]. That is to say, the quantum residual connection channels can equivalently implement ensembles of small quantum models with different frequencies, thus leading to richer spectrum and stronger expressivity. We will show the expressivity improvement in detail in the numerical simulation section.

Measurement scheme

To get the expectation values of an observable O for the quantum state $\mathcal{R}(x)|\phi_0\rangle$, which is embedded in the $|0\rangle\langle 0|$ subspace of the ancillary qubit, we can introduce another observation operator $\bar{O} = |0\rangle\langle 0| \otimes O$ on the system. Then the output observation values can be expressed as

$$
\begin{aligned}
\bar{f}_R(x, \theta) &= \langle\phi_f|\bar{O}|\phi_f\rangle \\
&= \langle 0|\langle\phi_0|\mathcal{R}^\dagger(x)(|0\rangle\langle 0| \otimes O)|0\rangle\mathcal{R}(x)|\phi_0\rangle \\
&= f_R(x, \theta)
\end{aligned} \tag{11}
$$

Fig. 2 | Fitting results of quantum models. The fitting results of quantum models to the target function $y_1(x)$ with frequency spectra $\Omega_1 = \{0, 1\}$ (**a**, **b**) and $\Omega_2 = \{0, 1, 0.5\}$ (**c**, **d**). The panels (**a**, **c**) show the theoretical function values (black dashed lines), and the quantum model outputs with traditional (gray) and residual (red) encoding strategies, respectively. The panels (b,d) show the mean squared error (MSE) during the training processes.
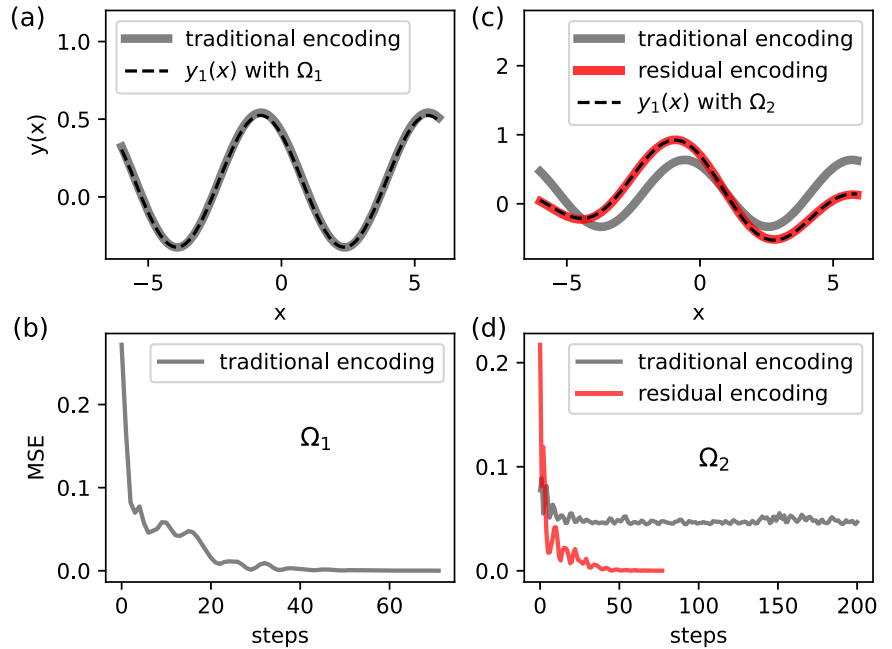
where $|\phi_f\rangle = |0\rangle \mathcal{R}(x)|\phi_0\rangle + |\perp\rangle$ is the output quantum state of the whole system, and the second item $|\perp\rangle$ is orthogonal to the first part. Furthermore, because we can expand the measurement operator as $\bar{O} = (\sigma_0 + \sigma_z)/2 \otimes O$, we can also have

$$\bar{f}_R(x, \theta) = \frac{1}{2}\left(\langle \sigma_0 \otimes O\rangle + \langle \sigma_z \otimes O\rangle\right) \quad (12)$$

This indicates that we can obtain the residual outputs $f_R(x, \theta)$ by measuring the average expectation of system output state $|\phi_f\rangle$ with two observations $\{\sigma_0 \otimes O, \sigma_z \otimes O\}$, which is experimentally feasible and introduces little resource overhead. For a l-layer residual encoding, we need l ancillary qubits at most and the corresponding observation operators will be $\{(\sigma_0 + \sigma_z)^{\otimes l} \otimes O\}$, whose size grows exponentially with layers of residual encoding. This exponential dependence is intrinsically related to the attenuation of success probability in the quantum algorithms with post-selection. Specifically, suppose that the output state with one residual connection on qubit i is $|\phi_f^{(i)}\rangle = |0\rangle \mathcal{R}(\Diamond)|\phi_0^{(i)}\rangle + |\perp\rangle$, then the probability for measuring ancillary qubit in $|0\rangle$ state is $P_0^{(i)} = ||\mathcal{R}(\Diamond)|\phi_0^{(i)}\rangle||^2$, where $||x||$ represents the modulus of vector x. So the success probability of quantum algorithm with l residual connection blocks is $P_s = \prod_{i=1}^{l} P_0^{(i)}$, which decays exponentially with l.

In practice, we do not need to use residual feature maps in every block, and inserting residual connections to some sampled data-encoding blocks could make the networks obtain better expressivity. In addition, the measurement schemes suggest that our algorithm is compatible with the existing methods for calculating the gradient of expectation value of the quantum circuit with respect to the optimization parameters[57–59]. Using parameter-shift rule[57], the gradient of the residual outputs for a parameter θ_j can be calculated as

$$\frac{\partial f_R(x, \theta)}{\partial \theta_j} = \frac{1}{2}\left[f_R\left(x, \theta_j + \frac{\pi}{2}\right) - f_R\left(x, \theta_j - \frac{\pi}{2}\right)\right] \quad (13)$$

where $f_R(x, \theta_j \pm \pi/2)$ are the expectation values when the target parameter θ_j is shifted by $\pm \pi/2$ respectively.

Furthermore, it should be mentioned that the approximation improvement can be understood from the universal approximation property with polynomial basis functions[60], which states that the linear combination of different observations can approximate any continuous functions. Based on the above analysis for the quantum models with the specific residual encoding structures, we can see that such a combination of measurement results can actually lead to a frequency richness improvement in the Fourier series, which enhances the expressivity ability of quantum neural networks. Therefore, our work can serve as a specific case to bridge the polynomial approximation[60] and Fourier series approximation[50], two perspectives for understanding the universal approximation property of quantum machine learning models.

Numerical demonstration

To demonstrate the improvement of the Fourier frequency spectrum by residual connections, we present a proof-of-principle numerical simulation with Pennylane[61] here, which solves regression tasks of fitting quantum models to the target Fourier series. We adopt the traditional qubit encoding strategy to map classical data x into quantum state with a single-qubit Pauli-rotation $U(x) = R_y(x) = e^{-ix\sigma_y/2}$ operator, where the generator Hamiltonian $G = -\sigma_y/2$ has two eigenvalues $e_{1,2} = \pm 1/2$. The optimization ansatz used has two arbitrary single-qubit rotation gates $U(\theta_i) = R_z(\theta_i^1)R_y(\theta_i^2)R_z(\theta_i^3)$ for $i = 1, 2$ placed before and after the data-encoding block, resulting a quantum model $U_\theta(x) = U(\theta_2)U(x)U(\theta_1)$. The observable is σ_z and then the outputs is $f(x, \theta) = \langle 0|U_\theta^\dagger(x)\sigma_z U_\theta(x)|0\rangle$. The quantum models are trained by a supervised learning frame to search the optimal parameters θ^*, which minimizes the mean squared error (MSE) as

$$\Delta(\theta) = \frac{1}{2D}\sum_{i=1}^{D}(y(x_i) - f(x_i, \theta))^2 \quad (14)$$

where D is the dimension of the data set and $y(\cdot)$ is the target function. We use Adam optimizer with at most 200 steps and set the learning rate as 0.3 with batch size $0.7D$ in the simulation. A termination condition for the optimization convergence, that is, the variance of ten consecutive loss function values is less than 10^{-8}, is also used.

As shown in the Fig. 2, this quantum model can learn functions of the form $y_1(x) = \sum_{\omega_i \in \Omega_1}(ae^{i\omega_i x} + a^*e^{-i\omega_i x})$ with a MSE value $\Delta = 6.0 \times 10^{-5}$, where a is an amplitude parameter and the frequency spectrum is $\Omega_1 = \{\omega_0 = 0, \omega_1 = 2|e_{1,2}| = 1\}$, and this is consistent to the results in[50]. However, a multi-frequency function with spectrum $\Omega_2 = \{\omega_0 = 0, \omega_1 = 1, \omega_2 = 0.5\}$ cannot be well fitted with error $\Delta = 5.1 \times 10^{-2}$,
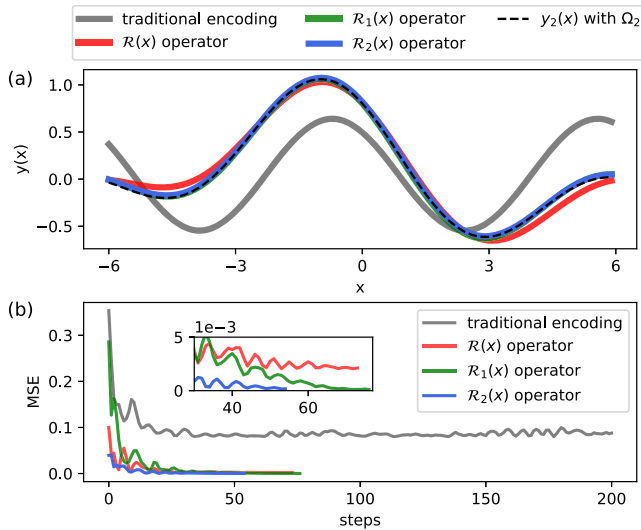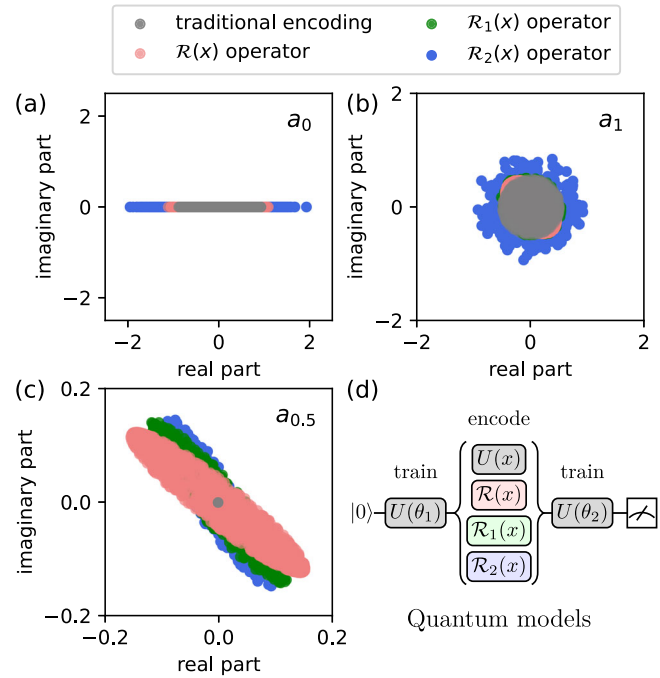
Fig. 3 | Fitting results of quantum models. a The fitting results of quantum models to the target function $y_2(x)$ with traditional encoding scheme (gray) and residual feature map with the $\mathcal{R}(x)$ (red), $\mathcal{R}_1(x)$ (green) and $\mathcal{R}_2(x)$ (blue) operators, respectively. **b** The mean squared error (MSE) during the training processes.

Fig. 4 | Fourier coefficients and quantum models. a–c The real and imaginary parts of the Fourier coefficients a_{ω_i} with $\omega_i \in \{0, 1, 0.5\}$ sampled from 1000 random quantum models. **d** Quantum models with one-layer data-encoding structure. The quantum models share the same ansatz but vary the data-encoding strategies by traditional encoding (gray), residual feature map with the $\mathcal{R}(x)$ (red), $\mathcal{R}_1(x)$ (green) and $\mathcal{R}_2(x)$ (blue) operators. The distribution of coefficients widens from gray to red to green to blue.

due to the frequency lack of parameterized quantum circuits caused by data-encoding strategy. The frequency mismatch can be mitigated by inserting residual connections to the data-encoding block with an output MSE value $\Delta = 5.1 \times 10^{-5}$, because the resulting residual operator $\mathcal{R}(x)$ can bring richer frequency components to enhance the circuit expressivity. It is worth noting that the residual data encoding scheme still works well for the spectral Ω_1 besides Ω_2, and the optimization process can converge quickly.

Furthermore, we turn to a more general case for fitting the function $y_2(x) = \sum_{\omega_i \in \Omega_2}(a_{\omega_i}e^{i\omega_i x} + a_{\omega_i}^* e^{-i\omega_i x})$, where the amplitudes can be different for each frequency component. Additional degrees of freedom can be obtained from the multi-combination methods of single-frequency components in residual outputs and the parameterized gates on the auxiliary qubit in the generalized residual operators $\mathcal{R}_{1,2}(x)$ in equation (3) and (17). We can conclude from the numerical results in the Fig. 3 that the traditional encoding scheme still cannot fit the target function with MSE value $\Delta = 0.09$, while the residual feature map with $\mathcal{R}(x)$ operator works better with error $\Delta = 2.1 \times 10^{-3}$. When we use the generalized residual operators, the fitting results can be further improved, which converges to a smaller MSE values with $\Delta = 1.1 \times 10^{-4}$ for $\mathcal{R}_1(x)$ and $\Delta = 1.7 \times 10^{-4}$ for $\mathcal{R}_2(x)$ in fewer optimization steps with 77 steps for $\mathcal{R}_1(x)$ and 55 steps for $\mathcal{R}_2(x)$. Moreover, the extra combination forms and trainable parameterized quantum gates bring more flexibility for fitting, which expand the Fourier coefficient space. As shown in the Fig. 4, we sample the quantum models 1000 times with different feature maps which produce Fourier series, and then get the distribution of Fourier coefficients. We can see that under the same ansatz, the residual feature map with $\mathcal{R}_2(x)$ operator has the widest Fourier coefficients distribution, and all the three residual encoding are better than the traditional encoding scheme.

In addition, this enhancement can be quantitatively measured by a commonly used expressibility metric[62]. We first generate many pairs of parameters Θ_1 and Θ_2 randomly, and calculate the distribution (P_F) of state fidelities $F = |\langle 0|U_{\Theta_1}^{\dagger}(x)U_{\Theta_2}(x)|0\rangle|^2$, which measure the overlap of quantum states generated by quantum models. Then the Kullback-Leibler (KL) divergence[63] is used to quantify the circuit expressivity by comparing the sampled fidelity distributions with that of the Haar-distributed state ensemble (P_{Haar}) as

$$D_{KL}(P_F||P_{\text{Haar}}) = \sum_j P_F(j)\log\frac{P_F(j)}{P_{\text{Haar}}(j)} \quad (15)$$

where the analytical form of the fidelity distribution for the ensemble of Haar random states is $p_{\text{Haar}}(F) = (N-1)(1-F)^{N-2}$ and N is the dimension of Hilbert space[64]. A smaller KL divergence value corresponds to a more favorable expressibility. We sample each quantum model in the Fig. 4 by 1000 times and use 45 histogram bins to estimate the fidelity distribution, which are then compared with the sampled fidelities ensemble of the Haar random states. The computed results of KL divergence are $D_{KL}^{\text{trad}} = 0.0634, D_{KL}^{\mathcal{R}(x)} = 0.0581, D_{KL}^{\mathcal{R}_1(x)} = 0.0446$ and $D_{KL}^{\mathcal{R}_2(x)} = 0.0429$, respectively. We can see that the residual operators can indeed increase the circuit expressivity relative to traditional encoding scheme because they all can introduce richer frequency components into the quantum models. However, it worth mentioning that though the three residual models have the same frequency spectrum, the additional reasons for the expressivity enhancement are somewhat different for $\mathcal{R}(x)$ and $\mathcal{R}_{1,2}(x)$ operators. The former one is due to the diverse construction methods of frequencies in residual outputs, while the latter is also due to the additional optimization parameters. We prove in Methods section that the generalized residual outputs can be seen as the weighted version of the residual outputs with trainable weights. Moreover, it is known that constructing frequencies only from the difference between the sum of the generator's eigenvalues will limit the access to higher-order components, resulting in a reduction in coefficient variance[50]. Therefore, the residual encoding method which can offer more methods to construct frequency could broaden the distribution of Fourier coefficients, which suggests an enhanced expressivity of quantum models by residual connections.

Moreover, similar to the traditional encoding, we can extend the accessible frequency spectrum by repeating the residual encoding block multi-times in sequence or in parallel method. To investigate the frequency extension by sequential and parallel repetitions of data-encoding, we fit the aforementioned target function $y_2(x)$ with a more complex spectra $\Omega_3 = \{\omega_0 = 0, \omega_1 = 1, \omega_2 = 0.5, \omega_3 = 1.5, \omega_4 = 2\}$ and amplitude $a_0 = 0.1$ and
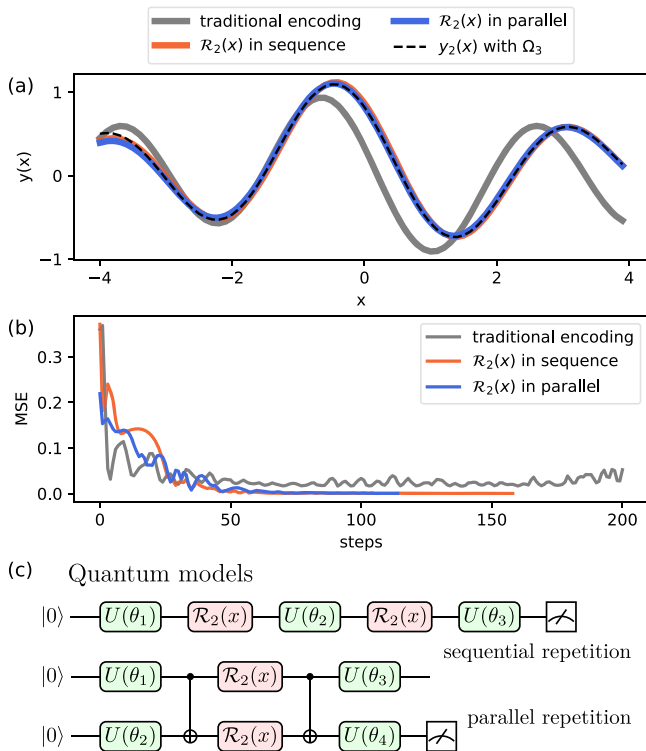
Fig. 5 | Fitting results and quantum models. a The fitting results of quantum models with two-layer data-encoding for target function $y_2(x)$ with frequency spectra Ω_3. **b** The mean squared error (MSE) during the training processes. **c** Quantum models with two-layer data-encoding structure. The residual operator $\mathcal{R}_2(x)$ is repeated in sequence and in parallel, and the output is the measurement value $\langle \sigma_z \rangle$ on a qubit.

$a_{1.5,2} = 5a_{1,0.5} = 0.15 + 0.15i$. Two-layers of repeating structures for the traditional encoding in sequence and residual encoding with $\mathcal{R}_2(x)$ operators in sequence and in parallel are used, as shown in the Fig. 5. The single-qubit observable is $O = \sigma_z$ for all cases. All the quantum models were trained with 200 steps at most using Adam optimizer and with batch size 16. We can see that both the sequential and parallel repetitions of residual encoding can extend the Fourier spectrum and fit the target function well. The MSE values and optimization steps for the sequential repetitions are $\Delta = 3.3 \times 10^{-4}$ and 159 steps, while $\Delta = 4.2 \times 10^{-4}$ and 115 steps for parallel repetitions. It should be clarified that the mixed use of residual and traditional encoding will also bring an enhanced expressivity. Therefore, replacing parts of the encoding blocks in complex quantum models with residual blocks, but not all of them, can enrich the expressivity of the whole neural networks.

Application in image classification

In this part, we turn to discuss the performance of QCNN algorithm with residual encoding for image classification using a real-word dataset MNIST. The MNIST includes 60000 (10000) images for train (test) datasets with 10 classes of handwritten digits, and each image is a 28×28 pixels data. Here we focus on the binary classification with selected classes 0 and 1, and the sizes for the train and test datasets used are 12665 and 2115. Constrained by the current quantum hardwares, high-dimensional data usually require classical pre-processing techniques for dimensionality reduction, and we adopt principal component analysis (PCA) technology to match the input data with the four-qubit data-encoding layer[65]. For comparison, we use qubit encoding and consider the case where no residual connection is added, and the case where the residual operator $\mathcal{R}_2(x)$ is applied to the i-th qubit, denoted as traditional and residual-Q_i schemes, respectively.

The ansatz for QCNN algorithm is composed of a series of alternating convolutional and pooling layers[27], as shown in the Fig. 6. Each convolutional layer includes several single- and two-qubit parameterized quantum gates, keeping a translationally invariant structure. We use Ising interactions between adjacent qubits with one parameter as $ZZ(\phi) = e^{-i\sigma_z \otimes \sigma_z \phi/2}$ and single-qubit U_3 gates with three parameters as

$$U_3(\theta, \phi, \delta) = \begin{bmatrix} \cos(\theta/2) & -e^{i\delta}\sin(\theta/2) \\ e^{i\phi}\sin(\theta/2) & e^{i(\phi+\delta)}\cos(\theta/2) \end{bmatrix} \tag{16}$$

The pooling layer is implemented by a parameterized controlled-U_3 gate and one qubit will be traced out, reducing the quantum states from two qubits to a single qubit. We measure the expectation values $\langle \sigma_z \rangle_i$ on the output qubit for the i-th input data with label $y_i = 0/1$. The cost function is $C(\theta) = \sum_{i=1}^{D}(|\langle \sigma_z \rangle|_i - y_i)^2/2D$ for a D-dimensional dataset and it is optimized by Adam optimizer with a learning rate 0.2. The number of iterations in the training process is 100 and the processes are repeated 20 times to obtain the mean values with random initialization of optimization parameters. Once the cost function converges and the optimal parameters $\theta^* = \arg\min_\theta C(\theta)$ are obtained, the measurement outputs can be reconstructed into binary values $c_{0/1}$ via a boundary precision $\epsilon \in (0, 0.5)$. We suppose that the classification result is $c_{0/1} = 1$ for $|\langle \sigma_z \rangle| > 1 - \epsilon$ and $c_{0/1} = 0$ for $|\langle \sigma_z \rangle| < \epsilon$, while other values are marked as unclassifiable optimization results. A smaller value for ϵ represents higher optimization accuracy and higher classification standards.

The optimization results of cost function and accuracy are shown in the Fig. 7 and Table 1. We set $\epsilon = 0.1$ in the simulation and there are 20 free parameters involved in the ansatz. We can conclude that the residual encoding schemes can obtain smaller convergence values of loss than the traditional encoding method, which means that the models have better approximation ability. Such an enhancement can lead to better expressivity and higher accuracy for quantum models in complex learning tasks. In addition, the residual encoding can produce a high classification accuracy, reaching 92.85% and 92.47% on average for the train and test datasets respectively, which are about 7.74% and 7.57% higher than that with the traditional encoding strategy. Further, we provide more numerical simulations of larger QCNN models with up to 12 qubits in the Fig. 8. We can see that with the increase of the number of qubits, the dimensionality reduction of the input image is mitigated, and more information can be involved into the quantum networks. The convergence values of loss function is gradually reduced, and the learning accuracy is gradually improved. The average classification accuracy on the train and test datasets with a residual data-encoding algorithm can be improved to about 97.66% in the maximum-scale quantum learning model.
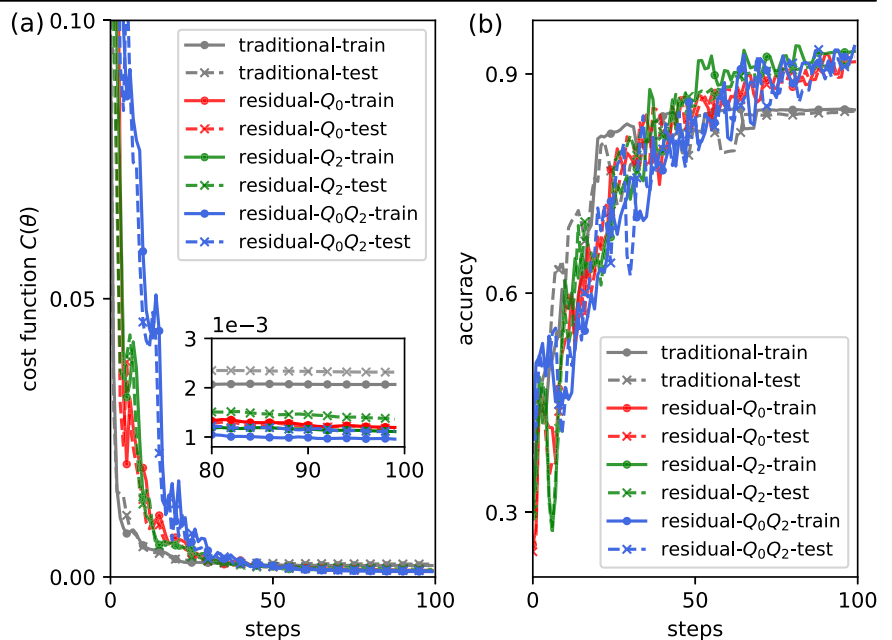
Conclusion

In summary, we have proposed a complete quantum circuit-based architecture for the implementation of quantum residual neural networks, dubbed QResNets. The classical residual connection channel is quantized by adding an auxiliary qubit to the data-encoding and trainable blocks, which is then generalized with additional parameterized gates. We further prove mathematically that the Fourier spectrum of quantum models output can be enriched when the residual connections are applied to the data-encoding blocks. There is a squared improvement in the number of frequency generation forms of residual encoding over the traditional schemes. It means that the l-layer residual encoding strategy can produce $\mathcal{O}(l^2)$ frequency combination methods, rather than just by the difference of sum of generator eigenvalues as in traditional methods. Moreover, the diverse spectrum construction methods in the residual outputs and additional optimization degrees of freedom in the generalized residual operators could make the Fourier coefficients more flexible, favoring the access to higher-order components. This indicates that the residual encoding can enrich the spectrum and broaden the Fourier coefficient distribution, that is, it can enhance the expressivity of various parameterized quantum circuits.

Various numerical simulation of fitting the functions of Fourier series, and a demonstration of binary classification in images of handwritten digits with MNIST datasets are conducted to show the algorithm performance. Compared with the traditional encoding, the accuracy of residual encoding can be improved by about seven percent. Our work advances the design of quantum neural networks with specific structures and enables a full quantum realization of classical residual connections, and also provides a quantum feature map strategy.
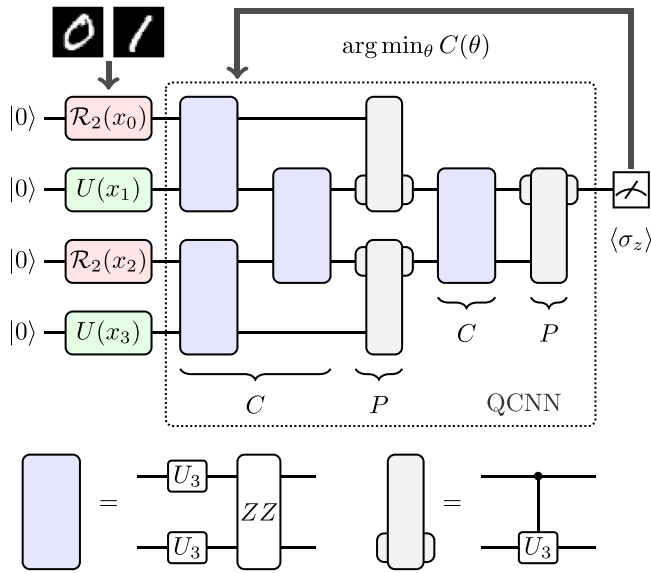
Fig. 6 | A schematic of quantum convolutional neural networks (QCNN) with residual encoding for image classification. The handwritten digits are encoded as quantum states via quantum feature map, where the green blocks represent qubit encoding schemes and the red blocks are residual encoding with $\mathcal{R}_2(x_i)$ operators on the i-th qubit. The multiple convolutional (C) and pooling (P) layers use quantum gates with trainable parameters θ, and the detailed structures are shown below. The measurement outcome of the quantum circuit $\langle\sigma_z\rangle$ is used to calculate the cost function $C(\theta)$ and characterize the binary classification results $c_{0/1}$. The classical computer updates the optimization parameters of QCNN algorithm based on gradients until the cost function converges.

Methods

Generalized residual operators

We have discussed the form of residual operator $\mathcal{R}(\diamond)$ and its corresponding residual output $f_R(x,\theta)$ above. In this part, we give a detail introduction to the generalized residual operators $\mathcal{R}_{1,2}(\diamond)$ and the corresponding generalized residual outputs $f_{R_{1,2}}(x,\theta)$, which present stronger expressivity. As shown in equation (3) where one Hadamard gate is replaced by a parameterized gate, we further assume that both two Hadamard gates on the ancillary qubit are replaced by gates $R_y(2\alpha)$ and $R_y(2\gamma)$ with trainable angles α and γ, then the $\mathcal{R}_2(\diamond)$ operator can be expressed as

$$\mathcal{R}_2(\diamond) = \cos\alpha\cos\eta\sigma_0^{\otimes n} + \sin\alpha\sin\eta \cdot \mathcal{L}(\diamond) \qquad (17)$$

with a relabeled angle $\eta = \pi m_a/2 - \gamma$. The residual operator $\mathcal{R}_1(\diamond)$ can be seen as a special case with $\gamma = -\pi/4$ ignoring a global phase factor. When the generalized residual operator $\mathcal{R}_{1,2}(x)$ is used in the data-encoding block, the residual output is

$$\begin{aligned} f_{R_{1,2}}(x,\theta) &= \langle\phi_0|\mathcal{R}_{1,2}^\dagger(x)O\mathcal{R}_{1,2}(x)|\phi_0\rangle \\ &= A_1^{R_{1,2}}f(x,\theta) + A_2^{R_{1,2}}\langle\phi_0|O|\phi_0\rangle \\ &\quad + A_3^{R_{1,2}}\text{Re}\big(\langle\phi_0|OU(x)|\phi_0\rangle\big) \end{aligned} \qquad (18)$$

where the trainable coefficients for $\mathcal{R}_1(x)$ operator are $A_1^{R_1}(\alpha) = \sin^2\alpha/2$, $A_2^{R_1}(\alpha) = \cos^2\alpha/2$ and $A_3^{R_1}(\alpha) = (-1)^{m_a}\sin 2\alpha/2$, while for the $\mathcal{R}_2(x)$ operator are $A_1^{R_2}(\alpha,\eta) = (\sin\alpha\sin\eta)^2$, $A_2^{R_2}(\alpha,\eta) = (\cos\alpha\cos\eta)^2$ and $A_3^{R_2}(\alpha,\eta) = (\sin 2\alpha\sin 2\eta)/2$. Such extension offers additional degree of freedom for the optimization process and can relax the range of Fourier coefficients for the new frequency component w_k in equation (6) to $A_3^{R_{1,2}}\sum_j\phi_j^* o_{jk}\phi_k$, and similar effect is true for other frequency components. In fact, the generalized residual outputs $f_{R_{1,2}}(x,\theta)$ can be seen as the weighted version of the residual outputs $f_R(x,\theta)$, where the weights of each term are trainable.

Proof of frequency combination forms

As mentioned above, there are four kinds of combination forms for frequency generation with a two-layer residual encoding. When another residual encoding layer is added, the spectrum $\Omega_{l=1}^R = \{w_k - w_j, \pm w_k | j,k \in [d]\}$

Fig. 7 | Evolution of cost function and accuracy. The performance of quantum convolutional neural networks with different data-encoding strategies for image classification. Simulations with the traditional scheme and residual encoding on qubits Q_0 and Q_2 in the train and test datasets are offered. The panel **a** shows the evolution processes of cost function with optimization steps and panel **b** is the corresponding results in accuracy.

Table 1 | Average accuracy

dataset	traditional	residual-Q_0	residual-Q_2	residual-Q_0Q_2
train	85.10%	91.66%	93.10%	93.78%
test	84.90%	92.35%	91.41%	93.65%

The average accuracy obtained from twenty repetitions of training for the image binary classification with MNIST datasets using different data-encoding strategies.
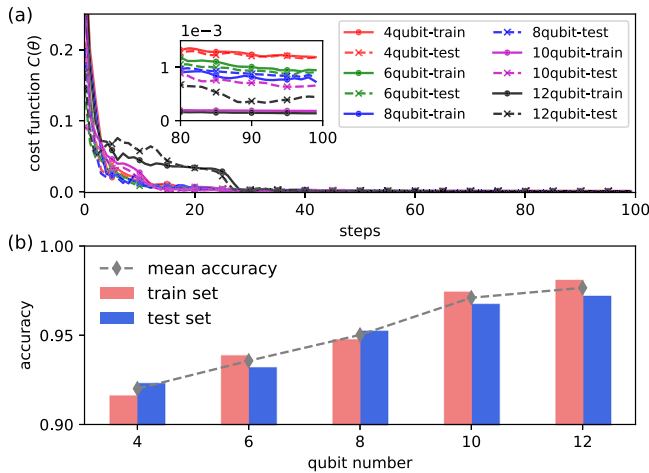
Fig. 8 | Results of larger quantum learning models with residual encoding. The panel **a** is the evolution processes of cost function with optimization steps when the residual connection is applied on the qubit Q_0 in data-encoding block, while panel **b** shows the average classification accuracy from ten repetitions under different qubit numbers.

would be combined to the spectrum $\Omega^R_{l=2}$. We first consider the component of difference of the sum of generator eigenvalues, and it would bring new frequency components for the three-layer residual spectrum as

$$\left\{ \sum_{m=1}^{3} w_{j_m} - \sum_{n=1}^{3} w_{k_n}, \pm\left(\sum_{m=1}^{3} w_{j_m} - \sum_{n=1}^{2} w_{k_n} \right) \right.$$
$$\left. \pm\left(\sum_{m=1}^{3} w_{j_m} - w_{k_1} \right), \sum_{m=1}^{2} w_{j_m} - \sum_{n=1}^{2} w_{k_n} \right\} \qquad (19)$$

with index $j_1, j_2, j_3, k_1, k_2, k_3 \in [d]$. If we further consider the effect of eigenvalues $\pm w_k \in \Omega^R_{l=1}$, more frequency components can be involved as

$$\left\{ \pm\left(\sum_{m=1}^{3} w_{j_m} - \sum_{n=1}^{2} w_{k_n} \right), \pm\left(\sum_{m=1}^{3} w_{j_m} - w_{k_1} \right), \pm \sum_{m=1}^{3} w_{j_m}, \right.$$
$$\left. \sum_{m=1}^{2} w_{j_m} - \sum_{n=1}^{2} w_{k_n}, \pm\left(\sum_{m=1}^{2} w_{j_m} - w_{k_1} \right) \right\} \qquad (20)$$

We can combine the above cases for frequency generation and simply mark the combination forms of $\pm\left(\sum_{m=1}^{l_1 \geq 1} w_{j_m} - \sum_{n=1}^{l_2 \geq 1} w_{k_n} \right)$ as $\mathbb{DS}(l_1, l_2)$, which means the difference between the sum of two sets with l_1 and l_2 frequencies. Note that we mark the combination form of $\pm \sum_{m=1}^{l \geq 1} w_{j_m}$ as $\mathbb{DS}(l, 0)$. Then we can find that there are six kinds of frequency combination forms for the three-layer residual encoding, and it can be concluded as $\{\mathbb{DS}(3,3), \mathbb{DS}(3,2), \mathbb{DS}(3,1), \mathbb{DS}(3,0), \mathbb{DS}(2,2), \mathbb{DS}(2,1)\}$. Further, for the l-layer residual encoding, the spectrum with various frequency

generation forms can be formally expressed as

$$\Omega^R_l = \{\mathbb{DS}(l,l), \mathbb{DS}(l,l-1), \cdots, \mathbb{DS}(l,1), \mathbb{DS}(l,0)$$
$$\mathbb{DS}(l-1, l-1), \cdots, \mathbb{DS}(l-1, 1)$$
$$\cdots$$
$$\mathbb{DS}(\lceil l/2 \rceil, \lfloor l/2 \rfloor)\} \qquad (21)$$

where the $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ are roundup and rounddown functions. Based on the number of items in each row of equation (21), we can determine the number of components in the set as

$$\mathcal{N}\left(\Omega^R_l\right) = (l+1) + (l-1) + \cdots + (\lceil l/2 \rceil - \lfloor l/2 \rfloor + 1)$$
$$= \frac{(l+2) + (\lceil l/2 \rceil - \lfloor l/2 \rfloor)}{2} \frac{(l+2) - (\lceil l/2 \rceil - \lfloor l/2 \rfloor)}{2} \qquad (22)$$
$$= (\lceil l/2 \rceil + 1)(\lfloor l/2 \rfloor + 1)$$

It can be concluded that compared with the traditional encoding method which generates frequency only with $\mathbb{DS}(l, l)$[50], there is a squared improvement in frequency generation methods for the residual encoding scheme with $\mathcal{N}(\Omega^R_l) \propto \mathcal{O}(l^2)$. While different combinations may produce some of the same frequency components, in general, more frequency-generation methods suggest that the possible upper bounds for the size of the Fourier spectrum of quantum model outputs can be larger, allowing for more complex learning tasks. Moreover, the diverse construction methods for frequencies can also improve the flexibility of Fourier coefficients, favoring the access to higher-order components and further improving the expressivity of quantum models.

Data availability
The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability
The code is available from the corresponding authors upon reasonable request.

References
1. Nielsen, M. A. & Chuang, I. L. *Quantum computation and quantum information* (Cambridge university press, 2010). https://doi.org/10.1017/CBO9780511976667.
2. Feynman, R. P. Simulating physics with computers. *Int J Theor Phys* **21**, 467–488 (1982).
3. Benioff, P. The computer as a physical system: A microscopic quantum mechanical hamiltonian model of computers as represented by turing machines. *Journal of statistical physics* **22**, 563–591 (1980).
4. Deutsch, D. Quantum theory, the church–turing principle and the universal quantum computer. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* **400**, 97–117 (1985).
5. Arute, F. et al. Quantum supremacy using a programmable superconducting processor. *Nature* **574**, 505–510 (2019).
6. Zhong, H.-S. et al. Quantum computational advantage using photons. *Science* **370**, 1460–1463 (2020).
7. Wu, Y. et al. Strong quantum computational advantage using a superconducting quantum processor. *Physical review letters* **127**, 180501 (2021).
8. Cao, Y. et al. Quantum chemistry in the age of quantum computing. *Chemical reviews* **119**, 10856–10915 (2019).
9. Cumming, R. & Thomas, T. Using a quantum computer to solve a real-world problem–what can be achieved today? arXiv preprint arXiv:2211.13080 (2022). https://doi.org/10.48550/arXiv.2211.13080.

10. Herman, D. et al. A survey of quantum computing for finance. arXiv preprint arXiv:2201.02773 (2022). https://doi.org/10.48550/arXiv.2201.02773.

11. Schuld, M., Sinayskiy, I. & Petruccione, F. An introduction to quantum machine learning. *Contemporary Physics* **56**, 172–185 (2015).

12. Biamonte, J. et al. Quantum machine learning. *Nature* **549**, 195–202 (2017).

13. Cerezo, M., Verdon, G., Huang, H.-Y., Cincio, L. & Coles, P. J. Challenges and opportunities in quantum machine learning. *Nature Computational Science* **2**, 567–576 (2022).

14. Zeguendry, A., Jarir, Z. & Quafafou, M. Quantum machine learning: A review and case studies. *Entropy* **25**, 287 (2023).

15. Preskill, J. Quantum computing in the nisq era and beyond. *Quantum* **2**, 79 (2018).

16. Li, Y., Zhou, R.-G., Xu, R., Luo, J. & Hu, W. A quantum deep convolutional neural network for image recognition. *Quantum Science and Technology* **5**, 044003 (2020).

17. Henderson, M., Shakya, S., Pradhan, S. & Cook, T. Quanvolutional neural networks: powering image recognition with quantum circuits. *Quantum Machine Intelligence* **2**, 2 (2020).

18. Havlíček, V. et al. Supervised learning with quantum-enhanced feature spaces. *Nature* **567**, 209–212 (2019).

19. Farhi, E. & Neven, H. Classification with quantum neural networks on near term processors. arXiv preprint arXiv:1802.06002 (2018). https://doi.org/10.48550/arXiv.1802.06002.

20. Hur, T., Kim, L. & Park, D. K. Quantum convolutional neural network for classical data classification. *Quantum Machine Intelligence* **4**, 3 (2022).

21. Li, W. & Deng, D.-L. Recent advances for quantum classifiers. *Science China Physics, Mechanics & Astronomy* **65**, 220301 (2022).

22. Ren, W. et al. Experimental quantum adversarial learning with programmable superconducting qubits. *Nature Computational Science* **2**, 711–717 (2022).

23. Beer, K. et al. Training deep quantum neural networks. *Nature communications* **11**, 808 (2020).

24. Abbas, A. et al. The power of quantum neural networks. *Nature Computational Science* **1**, 403–409 (2021).

25. Schuld, M. & Killoran, N. Quantum machine learning in feature hilbert spaces. *Physical review letters* **122**, 040504 (2019).

26. Dallaire-Demers, P.-L. & Killoran, N. Quantum generative adversarial networks. *Physical Review A* **98**, 012324 (2018).

27. Cong, I., Choi, S. & Lukin, M. D. Quantum convolutional neural networks. *Nature Physics* **15**, 1273–1278 (2019).

28. Chalumuri, A., Kune, R. & Manoj, B. A hybrid classical-quantum approach for multi-class classification. *Quantum Information Processing* **20**, 119 (2021).

29. Wu, S. L. et al. Application of quantum machine learning using the quantum kernel algorithm on high energy physics analysis at the lhc. *Physical Review Research* **3**, 033221 (2021).

30. Wang, H., Zhao, J., Wang, B. & Tong, L. A quantum approximate optimization algorithm with metalearning for maxcut problem and its simulation via tensorflow quantum. *Mathematical Problems in Engineering* **2021**, 1–11 (2021).

31. Landman, J. et al. Quantum methods for neural networks and application to medical image classification. *Quantum* **6**, 881 (2022).

32. Bausch, J. Recurrent quantum neural networks. *Advances in neural information processing systems* **33**, 1368–1379 (2020).

33. Liu, Z., Shen, P.-X., Li, W., Duan, L.-M. & Deng, D.-L. Quantum capsule networks. *Quantum Science and Technology* **8**, 015016 (2022).

34. Kashif, M. & Al-Kuwari, S. Resqnets: a residual approach for mitigating barren plateaus in quantum neural networks. *EPJ Quantum Technology* **11**, 4 (2024).

35. Mangini, S., Tacchino, F., Gerace, D., Bajoni, D. & Macchiavello, C. Quantum computing models for artificial neural networks. *Europhysics Letters* **134**, 10002 (2021).

36. Bowles, J., Ahmed, S. & Schuld, M. Better than classical? the subtle art of benchmarking quantum machine learning models. arXiv preprint arXiv:2403.07059 (2024). https://arxiv.org/abs/2403.07059.

37. Banchi, L., Pereira, J. & Pirandola, S. Generalization in quantum machine learning: A quantum information standpoint. *PRX Quantum* **2**, 040321 (2021).

38. Friedrich, L. & Maziero, J. Quantum neural network cost function concentration dependency on the parametrization expressivity. *Scientific Reports* **13**, 9978 (2023).

39. Anschuetz, E. R. & Kiani, B. T. Quantum variational algorithms are swamped with traps. *Nature Communications* **13**, 7760 (2022).

40. McClean, J. R., Boixo, S., Smelyanskiy, V. N., Babbush, R. & Neven, H. Barren plateaus in quantum neural network training landscapes. *Nature communications* **9**, 4812 (2018).

41. Cerezo, M., Sone, A., Volkoff, T., Cincio, L. & Coles, P. J. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature communications* **12**, 1791 (2021).

42. Marrero, C. O., Kieferová, M. & Wiebe, N. Entanglement-induced barren plateaus. *PRX Quantum* **2**, 040316 (2021).

43. Wang, S. et al. Noise-induced barren plateaus in variational quantum algorithms. *Nature communications* **12**, 6961 (2021).

44. Ballarin, M., Mangini, S., Montangero, S., Macchiavello, C. & Mengoni, R. Entanglement entropy production in quantum neural networks. *Quantum* **7**, 1023 (2023).

45. Herrmann, J. et al. Realizing quantum convolutional neural networks on a superconducting quantum processor to recognize quantum phases. *Nature Communications* **13**, 4144 (2022).

46. Pesah, A. et al. Absence of barren plateaus in quantum convolutional neural networks. *Physical Review X* **11**, 041011 (2021).

47. Gil Vidal, F. J. & Theis, D. O. Input redundancy for parameterized quantum circuits. *Frontiers in Physics* **8**, 297 (2020).

48. Pérez-Salinas, A., Cervera-Lierta, A., Gil-Fuster, E. & Latorre, J. I. Data re-uploading for a universal quantum classifier. *Quantum* **4**, 226 (2020).

49. Caro, M. C., Gil-Fuster, E., Meyer, J. J., Eisert, J. & Sweke, R. Encoding-dependent generalization bounds for parametrized quantum circuits. *Quantum* **5**, 582 (2021).

50. Schuld, M., Sweke, R. & Meyer, J. J. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A* **103**, 032430 (2021).

51. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90.

52. Shi, S. et al. Hybrid quantum-classical convolutional neural network for phytoplankton classification. *Front. mar. sci.* **10**, 1158548 (2023)

53. Deng, L. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine* **29**, 141–142 (2012).

54. Gui-Lu, L. General quantum interference principle and duality computer. *Communications in Theoretical Physics* **45**, 825 (2006).

55. Childs, A. M. & Wiebe, N. Hamiltonian simulation using linear combinations of unitary operations. arXiv preprint arXiv:1202.5822 (2012). https://doi.org/10.48550/arXiv.1202.5822.

56. Veit, A., Wilber, M. J. & Belongie, S. Residual networks behave like ensembles of relatively shallow networks. *Advances in neural information processing systems* **29** (2016). https://proceedings.neurips.cc/paper/2016/hash/37bc2f75bf1bcfe8450a1a41c200364c-Abstract.html.

57. Schuld, M., Bergholm, V., Gogolin, C., Izaac, J. & Killoran, N. Evaluating analytic gradients on quantum hardware. *Physical Review A* **99**, 032331 (2019).

58. Mari, A., Bromley, T. R. & Killoran, N. Estimating the gradient and higher-order derivatives on quantum hardware. *Physical Review A* **103**, 012405 (2021).

59. Wierichs, D., Izaac, J., Wang, C. & Lin, C. Y.-Y. General parameter-shift rules for quantum gradients. *Quantum* **6**, 677 (2022).
60. Goto, T., Tran, Q. H. & Nakajima, K. Universal approximation property of quantum machine learning models in quantum-enhanced feature spaces. *Physical Review Letters* **127**, 090506 (2021).
61. Bergholm, V. et al. Pennylane: Automatic differentiation of hybrid quantum-classical computations. arXiv preprint arXiv:1811.04968 (2018). https://doi.org/10.48550/arXiv.1811.04968.
62. Sim, S., Johnson, P. D. & Aspuru-Guzik, A. Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. *Advanced Quantum Technologies* **2**, 1900070 (2019).
63. Kullback, S. & Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics* **22**, 79–86 (1951).
64. Życzkowski, K. & Sommers, H.-J. Average fidelity between random quantum states. *Physical Review A* **71**, 032313 (2005).
65. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* **374**, 20150202 (2016).

Author contributions
J.W. conceived the algorithm; J.W., Z.H., D.C., and L.Q. contributed to writing and revised the manuscript.

Competing interests
The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Jingwei Wen or Ling Qian.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.