Quantum Science and Technology

A hybrid quantum-classical neural network for learning transferable visual representation

A Hybrid Quantum-Classical Neural Network for Learning Transferable Visual Representation

Ruhan Wang[1], **Philip Richerme**[2], **Fan Chen**[1,*]

[1]Luddy School of Informatics, Computing, and Engineering, Indiana University, USA
[2]Department of Physics, Indiana University, USA
[*]Author to whom any correspondence should be addressed.

E-mail: *fc7@iu.edu

Abstract. State-of-the-art quantum machine learning (QML) algorithms fail to offer practical advantages over their notoriously powerful classical counterparts, due to the limited learning capabilities of QML algorithms, the constrained computational resources available on today's Noisy Intermediate-Scale Quantum (NISQ) devices, and the empirically designed circuit ansatz for QML models. In this work, we address these challenges by proposing a hybrid quantum-classical neural network, which we call QCLIP, for Quantum Contrastive Language-Image Pre-Training. Rather than training a supervised QML model to predict human annotations, QCLIP focuses on more practical transferable visual representation learning, where the developed model can be generalized to work on unseen downstream datasets. QCLIP is implemented by using classical neural networks to generate low-dimensional data feature embeddings followed by quantum neural networks to adapt and generalize the learned representation in the quantum Hilbert space. Experimental results show that the hybrid QCLIP model can be efficiently trained for representation learning. We evaluate the representation transfer capability of QCLIP against the classical CLIP model on various datasets. Simulation results and real-device results on NISQ IBM_Auckland quantum computer both show that the proposed QCLIP model outperforms the classical CLIP model in all test cases. As the field of quantum machine learning on NISQ devices is continually evolving, we anticipate that this work will serve as a valuable foundation for future research and advancements in this promising area.

Keywords: hybrid quantum-classical algorithm, quantum neural network, variational quantum circuit, NISQ, contrastive language-image pre-training, zero-shot learning

1. Introduction

The recent phenomenal investment and rapid development of quantum computing hardware have ushered in the Noisy Intermediate-Scale Quantum (NISQ) [1] era where quantum machines are expected to support 50∼100 qubits (quantum bits) and around 10^3 quantum operations in the coherence time of the physical qubits. In Table 1, we summarize the key features of two state-of-the-art quantum computers – IonQ Forte [2] launched in 2022 and IBM Heron [3] slated for 2023. As it shows, NISQ computers suffer from errors due to imperfect qubit control and external interference.

Table 1: A summary on two state-of-the-art quantum computers (**1Q-Gate**: one-qubit gate; **2Q-Gate**: two-qubit gate; **SPAM**: state preparation and measurement).

Machine	**Technology**	**Qubits #**	**Coherence**	**Error Rate**		
				1Q-Gate	**2Q-Gate**	**SPAM**
IonQ_Forte [2]	Trapped-Ions	32	$\sim 1s$	0.02%	0.4%	0.5%
IBM_Heron* [3]	Superconducting	133	$<40\mu s$	0.1%	2.07%	1.42%

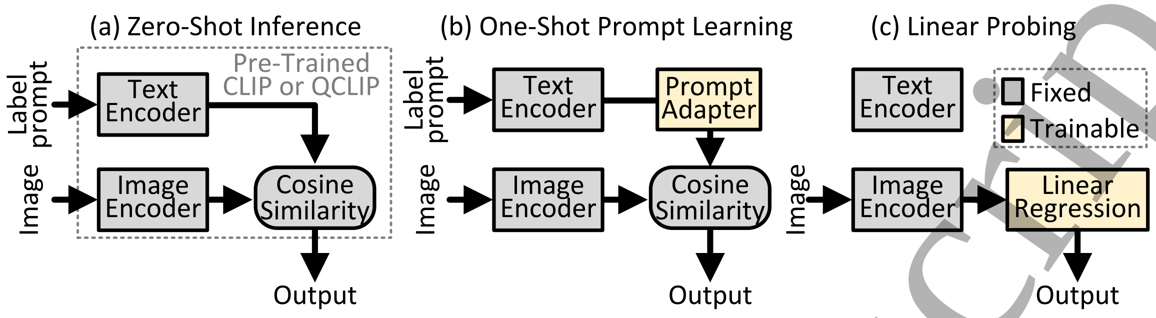
* IBM did not provide error rates for `IBM_Heron`. We report the error rates for the v3 generation of 127-qubit `IBM_Eagle` processor as an approximation.

Current error rates on NISQ devices greatly exceeds the 10^{-15} error rate required for many quantum algorithms [4–12] to achieve computational advantages. Although fault-tolerant quantum computers are theoretically feasible by incorporating quantum error-correction protocols [13–15], their practical implementation with millions of physical qubits may take decades of research.

NISQ algorithms [16] exploiting error-prone qubits and imperfect quantum gates to solve classically challenging problems have recently been intensively studied in various disciplines [17–27], among which, Quantum Machine Learning (QML) [25–27] has shown significant advantages over its classical counterpart in small-scale learning tasks [28–31]. With the power to access an exponentially large Hilbert space [32] and the ability to represent complex high-dimensional distributions [33], QML models are expected to revolutionize a wide range of applications including material discovery [34, 35], medical health [36, 37], and financial services [38, 39]. *Despite demonstrated advantages, state-of-the-art QML models have yet to solve practical problems due to the limited learning capabilities of QML algorithms, the constrained computational resources available on NISQ computers, and the empirically designed circuit ansatzes for QML models.*

First, most QML algorithms [27–29] focus on supervised classification by training models to predict class labels on test data that is generated from the same distribution as the training data. However, sufficient labeled training data for real-world tasks is usually unavailable [40, 41] or prohibitively expensive [42] to obtain. Moreover, representations learned from supervised QML are restricted to a set of "golden labels", which greatly limits the generalization and transferability of the developed models on datasets that are generated from different distributions [43]. **Second**, NISQ computers suffer from limitations in terms of qubit number and coherence time. The input size for real-world datasets is normally millions of tensors with millions of entries each, however, current NISQ devices can only work with small-scale toy benchmarks with input sizes of 2×2 or 4×4 [44–47]. How to achieve quantum advantages in practical-scale problems with NISQ devices is of great research significance. **Third**, QML models are typically implemented as parameterized quantum circuits [44–51] consisting of a classical-to-quantum data encoder and repeated layers of a variational quantum circuit (VQC). The circuit architecture for the data encoder and the VQC ansatz are currently empirically designed or simply randomly assigned.

Our Contributions. In this work, we address the aforementioned challenges



CLIP	39.0%	46.2%	70.8%
QCLIP	46.4%	55.6%	76.6%
Δ	+7.40%	+9.40%	+5.80%

Figure 1: Given a pre-trained model such as classical CLIP or QCLIP, we can transfer the learned visual representation via (a) zero-shot inference, (b) one-shot (or few-shot) prompt learning, or (c) fully fine-tuned linear probing. We use Cifar10 [52] as the downstream dataset and report the test accuracy of classical CLIP, QCLIP, and the accuracy improvement of QCLIP over CLIP (denoted as Δ) in the table.

by proposing a hybrid quantum-classical neural network architecture for learning transferable visual representation, which we call QCLIP, for Quantum Contrastive Language-Image Pre-training. Our main contributions can be summarized as follows:

- **A novel QML framework for learning transferable visual representation**. Instead of training a supervised QML model for predicting human annotations, we advance the flagship Contrastive Language-Image Pre-Training (CLIP) method [53] by proposing QCLIP, a quantum CLIP framework, which enjoys quantum-enhanced transferability and generalization only efficiently accessible on quantum computers. QCLIP combines limited NISQ resources and classical computing power to perform meaningful tasks, where classical neural networks (CaNNs) are used to generate low-dimensional data embeddings in classical feature space, while quantum neural networks (QuNNs) are exploited to enhance the model generalization in an exponentially large quantum Hilbert space (Section 3.1).
- **Quantum encoding methods and QuNN circuit ansatzes specialized for transferable visual representation learning.** We investigate various encoding methods and circuit ansatzes in the proposed QCLIP framework and identify the optimal candidate circuit ansatz for each quantum component (Section 3.2). We implement QCLIP on NISQ devices and carefully study how different training configurations affect final model performance. We provide a detailed training procedure for QCLIP (Section 3.3).
- **High-performance visual representation transfer on NISQ devices**. We demonstrate that the hybrid QCLIP model can be successfully trained for representation learning (Section 4.1). We evaluate the representation transferability of QCLIP using all mainstream methods including *zero-shot inference*, *one-shot*

prompt learning, and *linear probing* and show that QCLIP outperforms the classical CLIP model on various datasets (Section 4.2). A brief description of the experimental setup and numerical results are summarized in Figure 1. We also provide experimental results on different training configurations (Section 4.3) and NISQ `IBM_Auckland` quantum computer (Section 4.4). Our results show that the proposed QCLIP model outperforms the classical CLIP model in all test cases.

2. Background

2.1. Learning Transferable Visual Representation

Supervised representation learning methods [54–61] suffer from prohibitively expensive cost on labeled data preparation and poor representation transferability to downstream unseen datasets. Therefore, learning transferable visual representations is proposed and become a long-standing core problem in machine learning. Given a source domain $\mathcal{D}_\mathcal{S}$ with a source task $\mathcal{T}_\mathcal{S}$ and a target domain $\mathcal{D}_\mathcal{T}$ with a target task $\mathcal{T}_\mathcal{T}$, the goal of transferable visual representation learning is to improve the target function $f_T(\cdot)$ by reusing the representation learned from $\mathcal{D}_\mathcal{S}$ and $\mathcal{T}_\mathcal{S}$, where $\mathcal{D}_\mathcal{S}{\neq}\mathcal{D}_\mathcal{T}$ or $\mathcal{T}_\mathcal{S}{\neq}\mathcal{T}_\mathcal{T}$. Recent works [53, 62–74] encourage models to extract underlying explanatory factors hidden in the image by using unlabeled data in an unsupervised fashion, rather than just predicting human annotations. Provided the unlimited free raw data available on the Internet, this produces a model with better performance, and most importantly, the learned perception enables flexible representation transfer to downstream unseen datasets.

Among all prior arts, the Contrastive Language-Image Pre-Training (CLIP) method [53] has demonstrated state-of-the-art visual representation transfer performance. CLIP collects over 400M (image, text) pairs and trains an image encoder and a text encoder jointly with a task-agnostic contrastive loss [68, 69]. It is worth mentioning that the text descriptions are often referred to as "prompt" and their design is critical to CLIP performance. Once the training is complete, the quality of the visual representations learned by CLIP can be evaluated via different methods [75] including (1) *zero-shot inference* by directly generalizing the learned CLIP model to an unseen dataset; (2) *one-shot (or few-shot) prompt learning* by training a lightweight prompt adapter neural network [76–78] using one (or a few) training samples per class from the target dataset; or (3) *linear probing* which connects the pre-trained image encoder with a linear classifier [53, 68, 69] fully trained on a sufficiently large number of training data from the target domain. In general, *zero-shot inference* and *linear probing* respectively set the lower and upper bound on model transferability, while *one-shot (or few-shot) prompt learning* achieves intermediate performance because it considers a more practical scenario where the target dataset is neither completely inaccessible nor fully accessible.

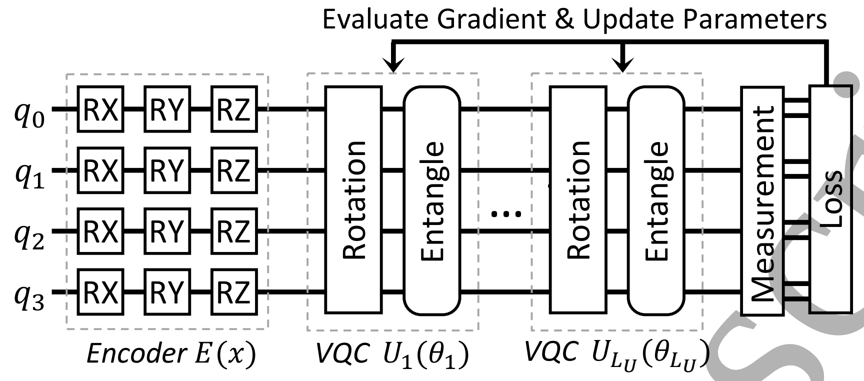


Figure 2: A standard quantum neural network.

2.2. Quantum Neural Networks

As illustrated in Figure 2, a standard QuNN begins with a classical-to-quantum encoder $\mathbf{E}(\mathbf{x})$ that encodes a classical input vector $\mathbf{x}$ into a N_Q-qubit quantum state $|\mathbf{x}\rangle$ [79]:

$$\mathcal{E} : \mathbf{x} \rightarrow |\mathbf{x}\rangle = \mathbf{E}(\mathbf{x})|0\rangle^{\otimes N_Q} = \bigotimes_{j=1}^{N_Q} \mathbf{R}(x_j)|0\rangle \tag{1}$$

where $\mathtt{R}$ denotes one-qubit gates $\{\mathtt{RX}, \mathtt{RY}, \mathtt{RZ}\}$ or their combinations, commonly referred to as *angle* encoding. Note that in this work, we exclude the *amplitude* encoding method due to its high $\mathcal{O}(2^{N_Q})$ circuit depth, making a QuNN more error-prone [45]. Instead, we focus on the *angle* encoding, which uses N_Q qubits and a constant-depth quantum circuit to encode a N_Q-bit classical data. The generated $|\mathbf{x}\rangle$ state is often referred to as a quantum input feature map and is manipulated by a subsequent variational quantum circuit $\mathbf{U}(\theta)$:

$$\mathcal{U} : |\mathbf{x}\rangle \rightarrow |\mathbf{y}(\theta)\rangle = \mathbf{U}(\theta)|\mathbf{x}\rangle = \left(\prod_{k=1}^{L_U} \mathbf{U_k}(\theta_\mathbf{k})\right)|\mathbf{x}\rangle \tag{2}$$

where $U(\theta)$ is implemented as a concatenation of a VQC ansatz in repeated L_U layers, and $\theta_\mathbf{k}$ is a set of trainable variables for the k_{th} layer. As illustrated in Figure 2, VQC ansatzes used in mainstream QML models [44–47] are normally constructed by single-qubit rotation gates followed by two-qubit entanglement gates. The final output results are obtained by quantum state measurement, $\mathbf{M}$, that maps the output quantum state $|\mathbf{y}(\theta)\rangle$ to a classical vector $\mathbf{y}(\theta)$:

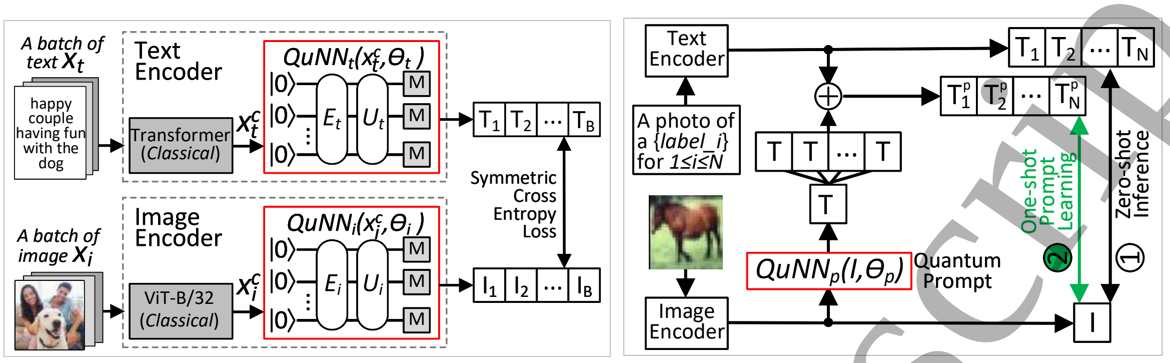
$$\mathcal{M} : |\mathbf{y}(\theta)\rangle \rightarrow \mathbf{y}(\theta) = \langle\mathbf{y}(\theta)|\mathbf{M}^\dagger\mathbf{M}|\mathbf{y}(\theta)\rangle \tag{3}$$

By default, qubits are measured in the z-basis for implementation simplicity. Globally the full QuNN can be written as

$$\mathcal{Q} : \mathbf{Q} = \mathbf{M} \circ \mathbf{U}(\theta) \circ \mathbf{E}(\mathbf{x}) \tag{4}$$

A QuNN model is evaluated by a pre-defined loss function $\mathbf{L}(\cdot)$ and iteratively trained to obtain optimal parameters via hybrid quantum-classical gradient descent [80]:

$$\mathcal{L} : \mathbf{y}(\theta) \rightarrow Loss = \mathbf{L}\left(\mathbf{y}(\theta)\right) \tag{5}$$



(a) QCLIP training

(b) QCLIP Representation Transfer

Figure 3: The overview of the proposed QCLIP framework. (a) QCLIP jointly trains an image encoder and a text encoder, which are both implemented as hybrid classical-quantum neural networks. For simplicity, we only show a QuNN with single-layer VQC in this example highlighted by a red rectangle. (b) At test time, the learned QCLIP model can be used for ① zero-shot inference, ② one-shot quantum prompt learning, or linear probing that is omitted in this figure.

$$Update\ rule: \theta_j^{t+1} = \theta_j^t - \eta \frac{\partial \mathbf{L}\left(\mathbf{y}(\theta)\right)}{\partial \theta_j} \tag{6}$$

Theoretical Insights. While quantum machine learning theory is continually evolving and in its nascent stages, this work provides insights on the optimal quantum encoder and variational circuit ansatz designs (see Appendix Appendix C) based on the current state of quantum machine learning theory research. However, it is important to note that the field currently lacks a standardized consensus. As a result, the discussions presented may be subject to changes or even controversies as our understanding of quantum machine learning progresses.

3. Method

In this section, we present the details of the proposed hybrid quantum-classical neural network architecture. In Section 3.1, we describe the general QCLIP framework, introduce QCLIP representation transfer for *zero-shot inference*, *one-shot (or few-shot) quantum prompt learning*, and fully supervised *linear probing*. In Section 3.2, we present the implementation of the quantum neural networks used in QCLIP. Finally, in Section 3.3, we discuss the training approach of QCLIP.

3.1. The QCLIP Framework

At the core of QCLIP is to learn image representations by contrasting them with the text prompt of the images, the same as classical CLIP [53]. The idea of QCLIP is inspired by recent research advances in quantum-enhanced feature learning [32] through exploiting quantum mechanical superposition, entanglement, and interference principles. Instead of using purely QuNNs on small datasets as in [32], the proposed QCLIP architecture is implemented by combining classical and quantum neural networks in one framework, thus, QCLIP can leverage CaNNs for large dataset preprocessing while utilizing QuNNs for quantum-enhanced feature adaptation and generalization.

3.1.1. QCLIP Overview As shown in Figure 3a, each high-dimensional input (image, text) pair $(\mathbf{x_i}, \mathbf{x_t})$ is first processed by CaNNs to generate compact low-dimensional data embedding in the classical feature space, and then QuNNs are utilized to further adapt the embeddings in an exponentially large quantum Hilbert space. Taking the hybrid image encoder network as an example, it utilizes a classical ViT-B/32 model [53] to produce a low-dimensional classical image embedding vector $\mathbf{x_i^c}$ and then uses a quantum neural network, $QuNN_i(\mathbf{x_i^c}, \theta_\mathbf{i})$, to map $\mathbf{x_i^c}$ to the quantum state space. A classical image embedding I is eventually generated via quantum measurements in the z-basis. Similarly, the hybrid text encoder is implemented as a classical 12-layer 512-wide text Transformer model with 8 attention heads [81] followed by a quantum neural network, $QuNN_t(\mathbf{x_t^c}, \theta_\mathbf{t})$, to generate a text embedding vector T. Note that I and T share a common dimensionality, specifically N_Q, which corresponds to the number of qubits utilized in the QuNNs. At the training time, QCLIP is optimized to predict the correct pairings of a batch (with a batch size B) of $(I_k,\ T_j)$ $(0 \leq k, j < B)$ pairs using symmetric cross-entropy loss. The ViT-B/32 and text Transformer models are particularly selected as classical feature extractors since they have demonstrated the best performance in classical CLIP models [53].

3.1.2. QCLIP Representation Transfer We evaluate the transfer capability of learned QCLIP visual representations using all mainstream evaluation methods introduced in Section 2.1. Below we describe the detailed configuration for each method.

Zero-Shot Inference assumes no access to the target dataset at all. Assuming the downstream dataset has N class names, we reuse the pre-trained QCLIP and compute the text embeddings, $\{T_1,\ T_2,\ \cdots,\ T_N\}$, for each target class name, as denoted as ① in Figure 3b. A test image is processed by the image encoder to generate a feature embedding, I. The similarity between I and $\{T_1,\ T_2,\ \cdots,\ T_N\}$ is then calculated and normalized into a probability distribution via a softmax function. We identify the most probably (image, text) pair as the output prediction. Prior works [53] show that the transferability of the classical CLIP model is greatly impacted by the input text that describes the image and found that using a text template improves performance. We follow the same text template engineering and ensembling schemes in [53].

One-Shot (or Few-shot) Prompt Learning targets a more practical scenario where one (or a few) training samples per class from target datasets are available at the test time. Various prompt learning algorithms [76–78] are recently proposed to alter the functionality of a pre-trained model across domains. However, none of these schemes can be directly applied to work with QuNNs. In this work, we introduce a quantum prompt learning algorithm.

As denoted as ② in Figure 3b, we design a domain prompt adapter, $QuNN_p(I, \theta_p)$, which is implemented as a parametrized QuNN. At the training time, the quantum prompt adapter takes the image vector I as input and generates a prompt T using one (or a few) unlabeled images x_i from the target training dataset. T has the same width as text embedding vectors and is added to all the original class embeddings to generate an adapted set of text pairing embeddings, denoted as $\{T^p{}_1, T^p{}_2, \cdots, T^p{}_N\}$. At the test time, we utilize domain-adapted text embeddings $\{T^p{}_1, T^p{}_2, \cdots, T^p{}_N\}$ instead of the general QCLIP text embeddings $\{T_1, T_2, \cdots, T_N\}$ to compute the similarity between the input image and the predicted classes.

Linear Probing assumes full access to the target training dataset. We adopt the established linear evaluation protocol [53, 68, 69] to test the visual representation transfer of QCLIP, where we freeze the QCLIP image encoder and only train a linear classification prediction layer on the output of the encoder network. The linear classifier is implemented as a logistic regression model and fully trained on target datasets for 1000 iterations. We then apply the whole network consisting of the QCLIP image encoder and the linear classifier head to the test data and report the classification accuracy.

3.1.3. QCLIP Implementation on NISQ Computers The classical ViT-B/32 and text Transformer respectively map the original data pair to a 512-dimensional image/text feature vector [53], which is considered as a classical compact encoding of the input. Ideally, the CaNNs can pass these 512-dimensional vectors to the QuNNs for further processing, however, NISQ computers available now only have 50∼100 qubits. Therefore, we follow the common practice [82, 83] by inserting a 512-to-N_C fully-connected layer between the classical and quantum layers to compress the initial feature vectors to a N_C-dimensional vector that can be effectively encoded in a practically available N_Q-qubit quantum system. The relationship between N_C and N_Q is determined by the classical-to-quantum encoding methods. To investigate the impact of compressed feature dimensions on the final performance, we conducted a study of the accuracy achieved by QCLIP with different N_C, as reported in Figure A1 in Appendix A. The experimental results demonstrate that increasing N_C leads to improved accuracy and transferability of the QCLIP model.

In conclusion, with a fixed N_Q qubits on a quantum computer, the encoder is expected to enable a larger N_C, allowing for a more accurate input representation by preserving a greater amount of information from the classical input data. The default angle encoding, which uses N_Q qubits, can only encode N_Q features, motivating the development of a denser encoder to accommodate a larger N_C in this work. Furthermore, the performance improvement with the increasing N_C also indicates that advancements in technology and the availability of more qubits will lead to improvements in the implementation scale of QCLIP and its corresponding performance and transferability.

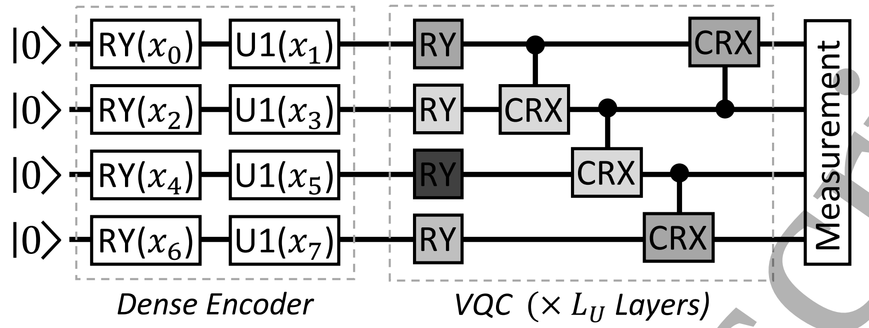


Figure 4: The proposed QuNN circuit for quantum image and text encoders. Note that the shaded gates are variational gates with trainable parameters.

3.2. Quantum Neural Networks

QuNNs used in QML models are currently empirically designed. In this work, we investigate various widely used encoding methods and VQC circuit ansatzes. Based on the performance evaluation, we identify the optimal QuNN circuits for each quantum component in the proposed QCLIP framework. We provide a full list of candidate quantum encoding methods and VQC ansatzes studied in this work respectively in Appendix A and Appendix B.

3.2.1. Quantum Image and Text Encoders Figure 4 shows the QuNN circuits used in the text and image encoder networks. In this example, we consider a QuNN with only four qubits for simplicity. The number of qubits as well as the number of VQC layers (i.e., L_U) in a generic QCLIP model can be adjusted to fit the problem of interest.

Classical-to-Quantum Encoder is essential for ensuring QML model accuracy, as it extracts and encodes relevant features from classical data into a quantum format, enabling subsequent processing in the quantum domain. However, the limited number of qubits in current quantum computers presents challenges in effectively embedding classical data, particularly with large-dimensional input datasets. In this work, we follow the generalized *dense angle encoding* [79] and present a dense classical-to-quantum encoder consisting of a layer of `RY` gates followed by a layer of `U1` gates, as shown in Figure 4. Given a classical N_C-dimensional input vector $\mathbf{x} = (x_0, x_1, \ldots x_{N_C-1})$, a quantum input feature map is generated by applying the encoding circuits to the ground quantum state $|0\rangle^{\otimes N_Q}$ of a N_Q-qubit system where $N_c = 2N_Q$, defining an encoder $\mathbf{E}(\mathbf{x})$ given by (see detailed mathematical derivation in Appendix A):

$$\mathbf{x} \rightarrow |\mathbf{x}\rangle = \mathbf{E}(\mathbf{x})|0\rangle^{\otimes N_Q} = \bigotimes_{j=1}^{\lceil N_Q/2 \rceil} \cos\left(\frac{x_{2j-1}}{2}\right)|0\rangle + e^{i\cdot x_{2j}} \sin\left(\frac{x_{2j-1}}{2}\right)|1\rangle \quad (7)$$

In contrast to the conventional encoding method represented by Equation 1, which uses N_Q qubits to represent N_Q features, QCLIP leverages the relative phase degree of freedom along with the angles to embed 2× more features using the same number of qubits. On top of this dense encoding method, we also explored *data re-uploading* [84] and *variational encoding* [85] to improve QuNN performance. Experimental results (see Appendix C) show that these two methods achieve negligible accuracy improvement in a QCLIP model, which contradicts previous conclusions from QML models [84, 85] implemented purely by QuNNs. We interpret the main reason as that these two methods primarily provide nonlinearity to a linear QuNN, while in a QCLIP model, nonlinearity is already sufficiently provided by the earlier CaNNs in the framework. Considering the significant implementation and training overhead introduced by *data re-uploading* and *variational encoding*, we do not recommend using these two methods in QCLIP.

VQC Ansatz is constructed by parameterized single-qubit rotation gates followed by nearest-neighbor coupling of qubits using entanglement two-qubit gates in current QuNNs [44–47]. Such a circuit ansatz has demonstrated superior expressive capability in various applications. The logic behind such designs is that single-qubit rotations provide a way to parameterize circuits, while two-qubit gates provide entanglement between two target qubits. Early designs [25] utilize fixed two-qubit `CNOT` gates to force maximum entangling power, while recent research [44–46] explored trainable entanglement by replacing fixed `CNOT` gates with parameterized two-qubit gates such as `CRX(`θ`)` [44], `U3(`θ`,`ϕ`,`λ`)` [46], or `CROT(`ϕ`,`θ`,`ω`)` [45].

We run experiments with different VQC ansatzes (see details in Appendix B) in the QCLIP architecture, results (see Appendix C) show that a VQC using parameterized two-qubit `CRX(`θ`)` gates leads to significant accuracy improvement compared to a baseline VQC with fixed two-qubit `CNOT` gates, demonstrating that adaptive and flexible entanglement rather than fixed maximal entanglement performs better for a QML algorithm, which is consistent with the conclusions in supervised QuNN models [44–46]. However, we find that further increasing the flexibility by replacing `CRX(`θ`)` gates with `U3(`θ`,`ϕ`,`λ`)` and `CROT(`ϕ`,`θ`,`ω`)` gates introduces significant hardware overhead and training complexity with no noticeable performance improvement. Therefore, we present the VQC circuit implemented with two-qubit `CRX(`θ`)` in Figure 4.

3.2.2. Quantum Prompt Adapter Neural Network The quantum prompt adapter $QuNN_p(I, \theta_p)$ takes an image vector I as input and generates a domain-adapted text vector T. In designing $QuNN_p$ encoders, we chose the default *angle* encoding over the *dense* encoder for two main reasons. First, maintaining the output dimensionality of $QuNN_p$ as the input vector I is required to ensure seamless integration with the subsequent components. Second, expanding the dimensionality of I by 2× and using the *dense* encoder is possibly but considered impractical. I is already a compact representation learned by $QuNN_i$, and increasing its dimensionality would not provide significant benefits. Moreover, it could introduce unnecessary complexity without improving overall performance. Through experiments, we identified the optimal circuit structure shown in Figure 5 consisting of a single layer of `RX` gates in the encoder and a VQC circuit employing two-qubit `CRX(`θ`)` gates for qubit entanglement, following previous work [44].

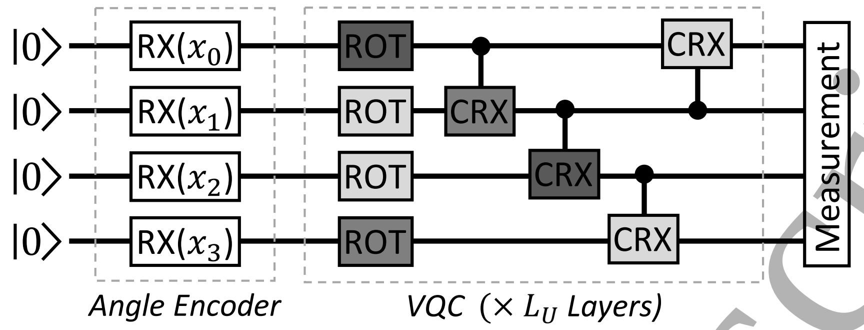


Figure 5: The proposed QuNN circuit for the quantum prompt adapter. Note that the shaded gates are variational gates with trainable parameters.

3.3. *Training of QCLIP*

To fix the parameters in the 512-to-N_C compression layer and the N_Q-qubit QuNNs used in the image and text encoders, we train the QCLIP model using *CC3M* [86] as a proxy dataset. The training goal is to predict which text *as a whole* is paired with which image. Specifically, given a batch of B input (images, text) pairs, QCLIP obtains respectively B image embedding vectors and B text embedding vectors. We denote $(I_k,\, T_j)$ where $k{=}j$ is a positive pair and a negative pair for $k{\neq}j$. We define a function that calculates loss using all these possible pairs and minimizes this function via stochastic gradient descent. Intuitively, if information can be successfully passed forward and backward in the hybrid architecture of QCLIP, the measured similarity between representations for positive pairs will decrease, while the distance between representations for negative pairs will increase.

3.3.1. Loss Function We consider two widely used loss functions, namely, normalized temperature-scaled contrastive loss [68, 69] and symmetric cross-entropy loss [53, 87]. We optimize the loss over similarity scores. Experimental results show symmetric cross-entropy loss outperforms contrastive loss for the training of QCLIP. We provide the pseudocode of cross-entropy loss based QCLIP training in Algorithm 1. We also provide details of the contrastive loss in Appendix D for comparison.

3.3.2. Training Method We implement the classical ViT-B/32 and text Transformer models in PyTorch [88]. We implement the QuNNs using PennyLane [89]. We use a mini-batch size of 128. We train the model for 75 iterations. We use Adam optimizer and set the learning rate to 0.001.

Among all the training hyperparameters, the initialization of parameters in QuNNs emerges as the most critical factor influencing the final performance of a QCLIP model. This is primarily due to the challenge of exponentially vanishing gradients concerning the quantum circuit depth and qubit number. For a deeper understanding, interested readers can refer to the theoretical discussion on the effect of parameter

Algorithm 1 Cross-Entropy Loss based QCLIP Training.

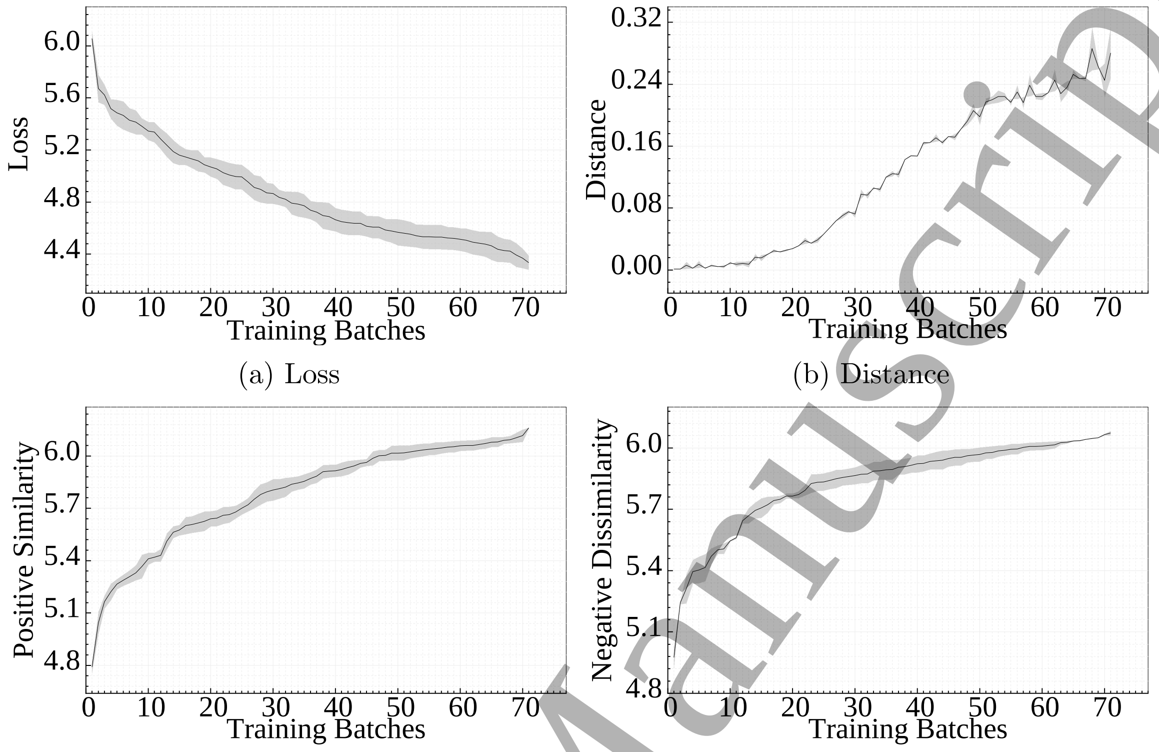
```
Input:
    1. Batch size: $B$,
    2. Label: $[1, 2, \cdots, B]$,
    3. Cross Entropy Loss: $F_{loss} = \frac{-\sum_{i=1}^{B} l_i \cdot log(p_i)}{B}$, where $l_i$ is the truth label and $p_i$ is
    the $Softmax$ probability for the $i^{th}$ class.
Output:
    1. Training loss, $loss$

Generate a batch of image embedding output vector $[I_1, I_2, \cdots, I_B]$.
Generate a batch of text embedding output vector $[T_1, T_2, \cdots, T_B]$.
# Compute $logits\_image = [l\_I_1, l\_I_2, l\_I_3, ..., l\_I_B]$
for $(i = 1;\ i < B{+}1;\ i{+}{+})$ do
    for $(t = 1;\ t < B{+}1;\ t{+}{+})$ do
        $l\_T_i = \frac{I_i \cdot T_t}{|I_i||T_t|}$
    end for
end for
# Compute $logits\_text = [l\_T_1, l\_T_2, l\_T_3, ..., l\_T_B]$
for $(t = 1;\ t < B{+}1;\ t{+}{+})$ do
    for $(i = 1;\ i < B{+}1;\ i{+}{+})$ do
        $l\_T_t = \frac{T_t \cdot I_i}{|T_t||I_i|}$
    end for
end for
$loss\_image = F_{loss}(logits\_image, label)$.
$loss\_text = F_{loss}(logits\_text, label)$.
$loss = \frac{1}{2}(loss\_image + loss\_text)$.
return $loss$.
```

initialization on the trainability and performance of QML models provided in [90]. In this work, we study both uniform initialization and Gaussian initialization in QCLIP as detailed in Appendix E. Inspired by classical Xavier initialization [91], we utilize the information of QuNN structures in the Gaussian initialization by defining $\mathcal{N}(0, \sigma^2)$, where $\sigma{=}1/\sqrt{N_Q}$. Experimental results show that the Gaussian distribution demonstrates better performance in terms of accuracy, training stability, and convergence.

4. Results and Analysis

In this section, we evaluate the effectiveness of the proposed QCLIP model. We follow the general QCLIP architecture and implement a practical design by setting N_C, N_Q, and L_U respectively to 16, 8, and 2. We run numerical simulations and report results on representation learning in Section 4.1. To compare QCLIP with classical CLIP, we create a baseline model by implementing classical CLIP in PyTorch. We follow the training approaches used in the original work [53, 83], with the only difference being the insertion of a 512-to-N_C fully-connected layer in the image/text encoder. This modification is made to ensure a fair and equal comparison between QCLIP and classical CLIP models.



(a) Loss

(b) Distance

(c) Positive Similarity

(d) Negative Dissimilarity

Figure 6: The training process of QCLIP on representation learning, with the line indicating the mean and the shaded area representing the deviation.

In Section 4.2, we evaluate the representation transferability of QCLIP and show that QCLIP outperforms the classical CLIP model on various datasets. Section 4.3 provides exploration results for different training configurations. We also implement a proof-of-concept QCLIP on NISQ `IBM_Auckland` quantum computer and report its performance results in Section 4.4.

4.1. Results on QCLIP Representation Learning

We first verify whether the proposed hybrid QCLIP model can be successfully trained for representation learning. To this end, we train the QCLIP model using *CC3M* [86] as a proxy dataset for 70 batches and record the training loss after each batch in Figure 6a. Results show that the loss decreases from 6.075 to 4.152 over the course of training, indicating that our model is able to learn. It is notable that the training time for QCLIP is significantly less than classical QCLIP models, which would typically take several hundred epochs [53]. By comparison, QCLIP is more compute-efficient, which allows us to reach higher overall performance within a limited computing budget.

We further quantitatively study the representation learning ability of QCLIP. We adopt the widely used Hilbert–Schmidt distance as the evaluation metric and report results on several key distances, following the approach taken in related work [53, 83]. Figure 6b, 6c, and 6d respectively record the distance between positive

Table 2: Performance comparison between QCLIP and CLIP on representation transfer.

Dataset	Zero-Shot			One-Shot			Linear Probing		
	CLIP	QCLIP	Δ	CLIP	QCLIP	Δ	CLIP	QCLIP	Δ
MNIST [92]	17.97%	20.32%	+2.35%	20.03%	30.51%	+10.48%	59.12%	62.05%	+2.93%
Cifar10 [52]	44.41%	46.40%	+1.99%	46.82%	55.62%	+8.80%	70.82%	76.63%	+5.81%
OxfordPet [93]	17.95%	27.12%	+9.17%	19.32%	33.93%	+14.61%	67.73%	68.26%	+0.53%
Food101 [94]	31.19%	37.97%	+6.78%	32.25%	49.46%	+17.21%	59.76%	64.02%	+4.26%

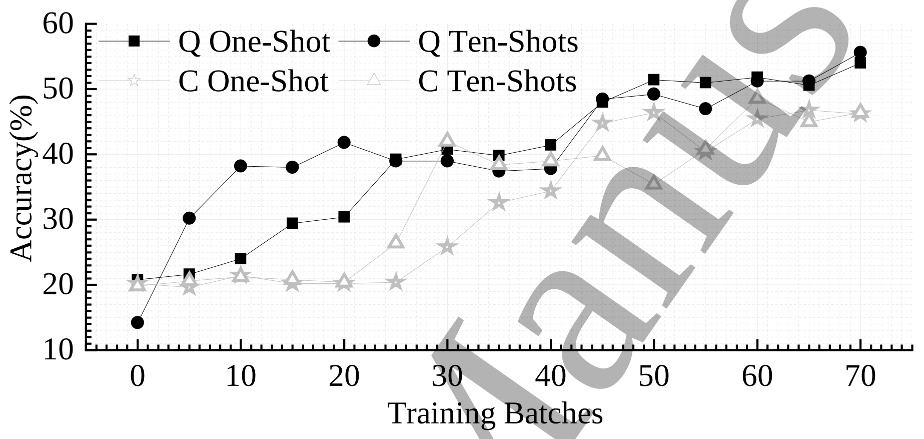


Figure 7: QCLIP performance results on few-shot learning.

and negative pairs (denoted as *Distance*), similarity within positive pairs (denoted as *Positive Similarity*), dissimilarity between negative pairs (denoted as *Negative Dissimilarity*). Throughout the training process, we observe that the measured similarity and dissimilarity undergo expected changes, indicating successful information propagation both forward and backward in the hybrid architecture of QCLIP. These quantitative results affirm that quantum components can effectively combine with classical resources to achieve meaningful and nontrivial representation learning tasks.

4.2. Results on QCLIP Representation Transfer

QCLIP is pre-trained to predict whether an image and a text prompt are paired together in a source dataset. This capability is then reused to perform *zero-shot inference*, *one-shot prompt learning*, and *linear-probing*, to study the representation transfer ability on downstream datasets. To demonstrate the robustness of QCLIP on various datasets with wide distributions, we evaluate QCLIP on four different target datasets including MNIST [92], Cifar10 [52], OxfordPet [93], and Food101 [94].

In Table 2, we summarize the performance of QCLIP on each task and highlight the accuracy improvement (denoted as Δ) provided by QCLIP compared to classical baselines. The quantitative results show that QCLIP is robust on all tested datasets and outperforms classical CLIP on all tasks. While supervised *linear probing* exhibits the upper bound on model transferability, QCLIP has the lowest performance improvement

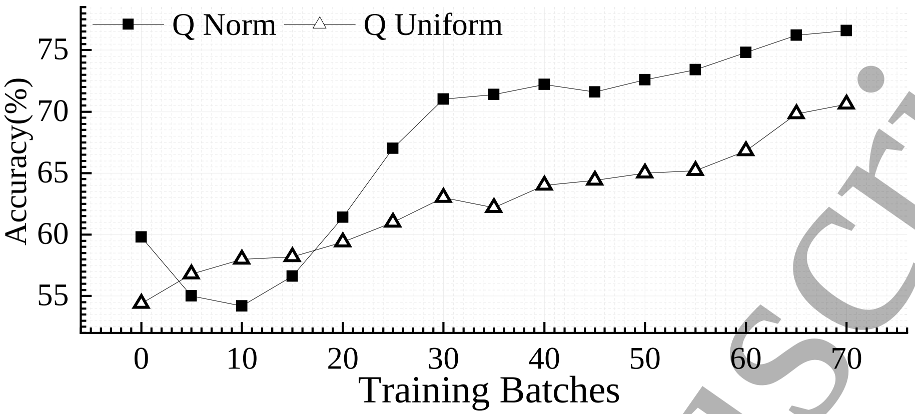


Figure 8: QCLIP performance using different initialization methods.

over CLIP on this task. Notably, *one-shot prompt learning* benefits the most from QCLIP with a performance improvement up to +17.21% on the Food101 dataset.

We further increase the shot number from one to ten for both classical CLIP and QCLIP and report the *few-shot* performance in Figure 7. The performance of few-shot prompt learning shows negligible improvement when the shot number increases from one to ten, indicating that the accuracy of the small domain prompt generators rapidly saturated with just very few (i.e., one per class) training data.

4.3. Results on Different Training Configurations

As discussed in Section 3.3, the QCLIP performance is greatly impacted by pre-defined loss functions and parameter initialization. Since *linear probing* represents an upper bound of QCLIP representation transferability, here we use it as a proxy task to explore the impact of different types of loss functions and parameter initialization methods.

Figure 8 compares the QCLIP model accuracy on *linear probing* by using normalized initialization (denoted as *Q Norm*) and uniform initialization (denoted as *Q Uniform*).

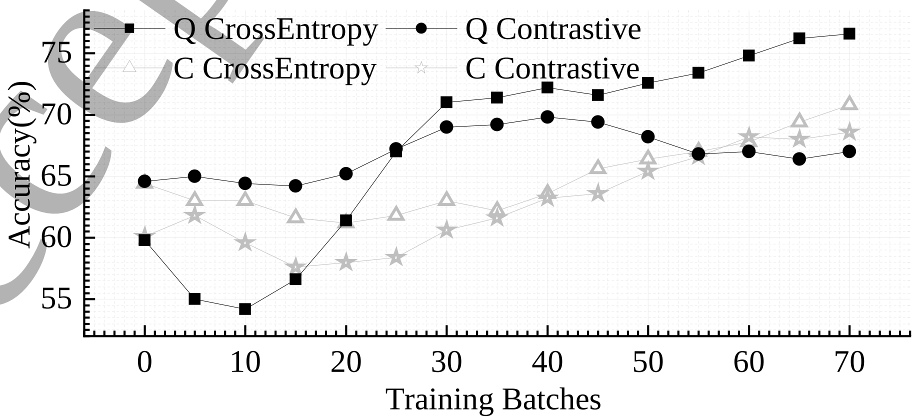


Figure 9: QCLIP performance using different loss functions.

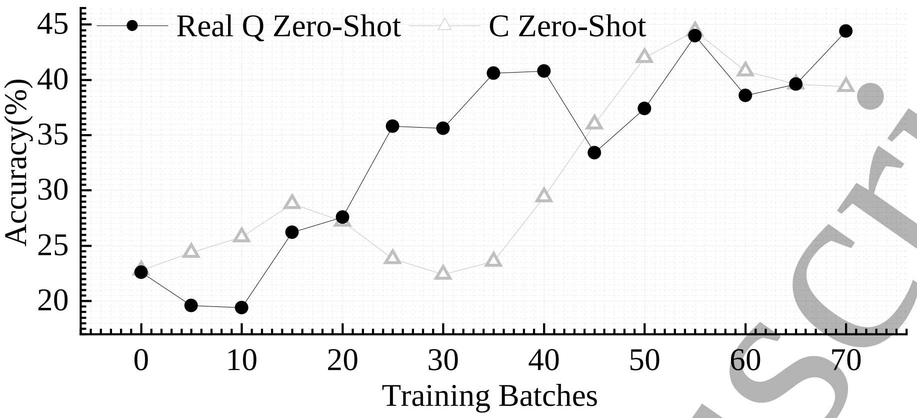


Figure 10: QCLIP performance on zero-shot inference using the `IBM_Auckland` quantum computer.

Results show that *Q Uniform* performs better in the first several training runs, while *Q Norm* provides better (8.2% higher than *Q Uniform*) final accuracy. These results are consistent with the observation reported in a previous work [95]. Therefore, normalized initialization is adopted in QCLIP training.

Figure 9 reports the performance on *linear probing* for classical CLIP and QCLIP by using respectively contrastive loss and cross-entropy loss. In general, cross-entropy loss improves the performance of both classical and quantum models. For classical CLIP training, the cross-entropy loss (denoted as *C CrossEntropy*) provides a 2.2% accuracy improvement compared to the contrastive loss (denoted as *C Contrastive*). For QCLIP, a significant 9.6% accuracy improvement is achieved when replacing the contrastive loss (denoted as *Q Contrastive*) with the cross-entropy loss (denoted as *Q CrossEntropy*). Recent work on quantum self-supervised learning [83] directly employs the contrastive loss function for QuNN training, whereas in this work we identify the cross-entropy loss function as an optimal option and used it for QCLIP training.

4.4. Results on NISQ Devices

In addition to the numerical simulation results reported in previous sections, we also implement a proof-of-concept QCLIP on real NISQ devices and report its performance to demonstrate the effectiveness of QCLIP. We use the `IBM_Auckland` quantum computer, which is a 27-qubit device with respective 0.022%, 1.164%, and 1.110% error rates for 1Q-Gate, 2Q-Gate, and SPAM. Compared with the state-of-the-art devices reported in Table 1, `IBM_Auckland` is a more practical NISQ device that is publicly available to average users. We adopt the pre-trained QCLIP model and implemented it on `IBM_Auckland` using only 8 qubits.

We perform *zero-shot inference* and *one-shot prompt learning* on real devices and report the results respectively in Figure 10 and Figure 11. Note that we exclude the fully fined-tuned *linear probing* on real devices due to its long training latency. In

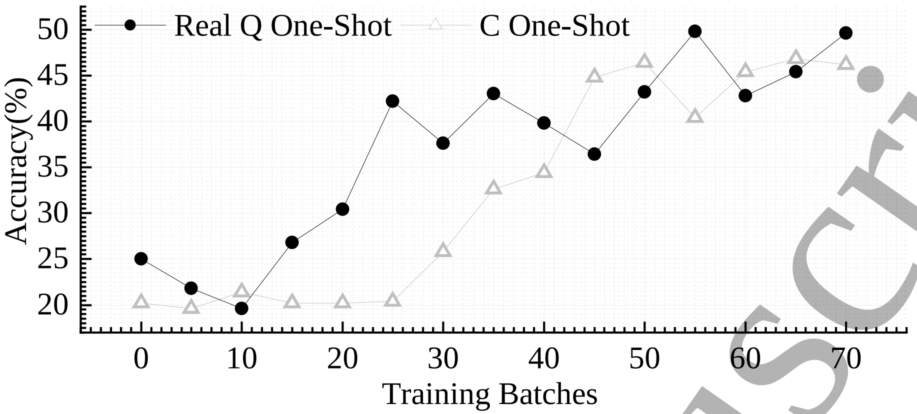


Figure 11: QCLIP performance on one-shot prompt learning using the `IBM_Auckland` quantum computer.

general, the performance of QCLIP on real devices is decreased due to the noisy qubits and imperfect control and measurement. Specifically, the QCLIP accuracy on *zero-shot inference* drops from 46.4% to 44.4% (i.e., the final accuracy for *Real Q Zero-Shot* in Figure 10), while the performance on *one-shot prompt learning* decreases from 55.6% to 49.6% (i.e., the final accuracy for *Real Q One-Shot* in Figure 11). However, the classical CLIP model only achieves respectively an accuracy of 39.0% and 46.2% for *zero-shot inference* and *one-shot prompt learning*. Therefore a quantum advantage (up to 5.4%) on representation transferability is still reserved for real-device results.

5. Conclusion

Current quantum machine learning models mainly focused on supervised classification tasks using down-sampled input data with a very small scale, i.e., labeled images with a 4×4 or even 2×2 size. Such models failed to solve practical problems and show limited generalization and transferability to unseen downstream datasets. In this work, we propose to advance the flagship Contrastive Language-Image Pre-Training (CLIP) method by proposing QCLIP, a quantum CLIP framework, to improve the performance of quantum machine learning algorithms on transfer representation learning tasks. The key idea is to leverage the quantum-enhanced transferability and generalization only efficiently accessible on quantum computers. However, current quantum computers are all Noisy Intermediate-Scale Quantum (NISQ) devices, which can only support 50∼100 qubits and a limited number of quantum gate operations. In order to leverage the limited NISQ resources to perform meaningful tasks, QCLIP combines quantum computing resources with classical computing power in a hybrid quantum-classical fashion, where classical neural networks are used to generate low-dimensional input embeddings in the classical feature space, and quantum neural networks are employed to enhance the model generalization in the quantum Hilbert space. We survey the mainstream quantum neural network implementation and study how different encoding methods, variational circuit ansatzes, and training configurations affect the final performance of the QCLIP model. We present a dense encoding method in this work, and also identify the optimal quantum circuit for each quantum component in QCLIP.

We implement a small-scale QCLIP and demonstrate the proposed hybrid quantum-classical neural network can be successfully trained for representation learning. We evaluate the transfer representation learning capability of QCLIP against the classical CLIP model using different datasets. Experimental results on numerical simulation and NISQ `IBM_Auckland` quantum computer both show that QCLIP model outperforms the classical CLIP model in all test cases.

Acknowledgments

This work was supported by the National Science Foundation CAREER Award (Grant No. CNS-2143120). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of grant agencies or their contractors.

Appendix A. Quantum Encoding Methods

Here we provide a detailed mathematical derivation of Equation 7. As illustrated in Figure 4, the proposed dense encoding method in this work is implemented by applying a layer of $\texttt{RY}(x_{2j-1})$ gates followed by a layer of $\texttt{U1}(x_{2j})$ gates to the ground quantum state of a N_Q-qubit system, where $\mathbf{x} = (x_0, x_1, \ldots x_{N_C-1})$ represents the classical N_C-dimensional input vector. The matrix representations for `RY` gate and `U1` gate are:

$$\texttt{RY} = \begin{bmatrix} \cos\frac{\theta}{2} & -\sin\frac{\theta}{2} \\ \sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{bmatrix} \qquad \texttt{U1} = \begin{bmatrix} 1 & 0 \\ 0 & e^{i\theta} \end{bmatrix} \tag{A.1}$$

The generated quantum input feature map is:

$$\mathbf{x} \rightarrow |\mathbf{x}\rangle = \mathbf{E}(\mathbf{x})|0\rangle^{\otimes N_Q} \tag{A.2}$$

$$= \bigotimes_{j=1}^{\lceil N_Q/2 \rceil} U1(x_{2j}) \cdot RY(x_{2j-1}) \cdot |0\rangle \tag{A.3}$$

$$= \bigotimes_{j=1}^{\lceil N_Q/2 \rceil} \begin{bmatrix} 1 & 0 \\ 0 & e^{i \cdot x_{2j}} \end{bmatrix} \cdot \begin{bmatrix} \cos\frac{x_{2j-1}}{2} & -\sin\frac{x_{2j-1}}{2} \\ \sin\frac{x_{2j-1}}{2} & \cos\frac{x_{2j-1}}{2} \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{A.4}$$

$$= \bigotimes_{j=1}^{\lceil N_Q/2 \rceil} \begin{bmatrix} \cos\frac{x_{2j-1}}{2} \\ e^{i \cdot x_{2j}} \sin\frac{x_{2j-1}}{2} \end{bmatrix} \tag{A.5}$$

$$= \bigotimes_{j=1}^{\lceil N_Q/2 \rceil} \cos\left(\frac{x_{2j-1}}{2}\right) \begin{bmatrix} 1 \\ 0 \end{bmatrix} + e^{i \cdot x_{2j}} \sin\left(\frac{x_{2j-1}}{2}\right) \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{A.6}$$

$$= \bigotimes_{j=1}^{\lceil N_Q/2 \rceil} \cos\left(\frac{x_{2j-1}}{2}\right) |0\rangle + e^{i \cdot x_{2j}} \sin\left(\frac{x_{2j-1}}{2}\right) |1\rangle \tag{A.7}$$

Therefore, we obtain the encoder function same as shown in Equation 7.

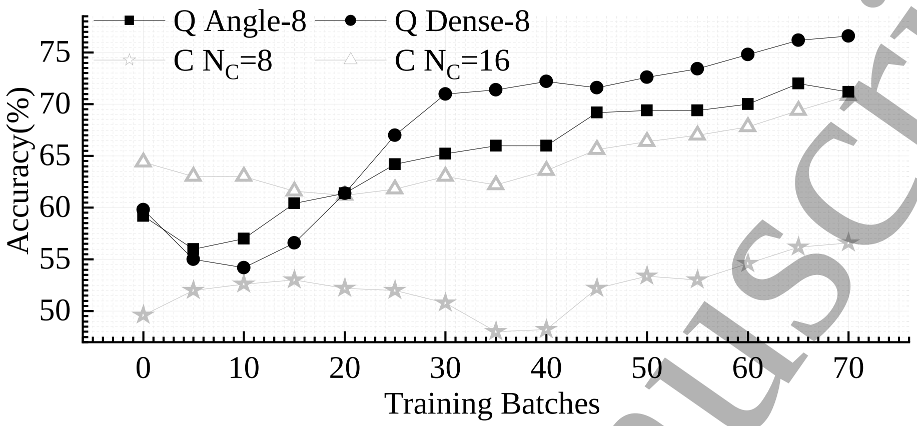


Figure A1: QCLIP performance using different encoding methods.

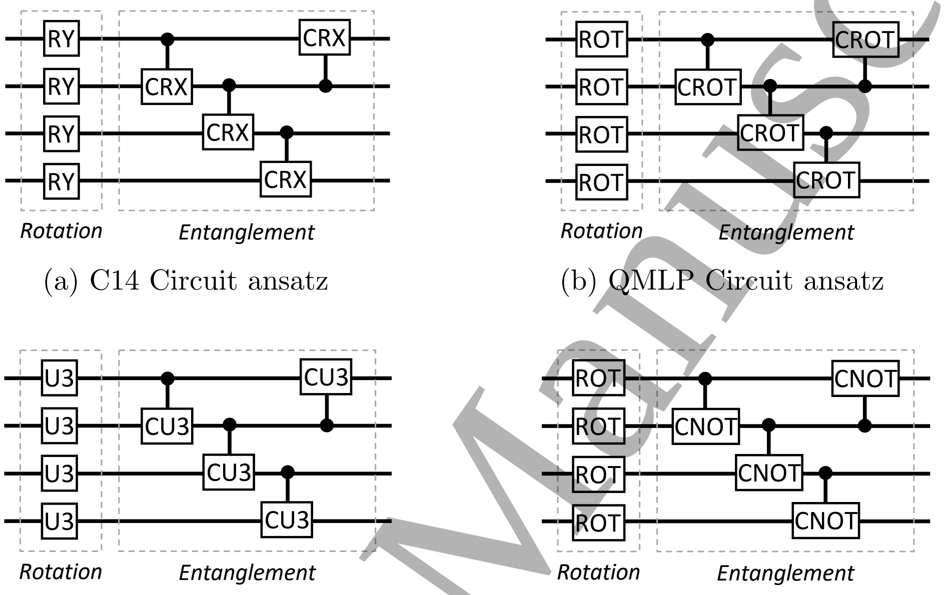
To evaluate the effectiveness of the proposed dense encoding method (Equation 7) against the baseline angle encoding method (Equation 1), we provide a performance comparison using *linear probing* as a proxy task. As shown in Figure A1, we denote performance results using dense encoding and the baseline angle encoding respectively as *Q Angle-8* and *Q Dense-8*, where 8 denotes the total number of qubits for a fair comparison. We also provide performance results for classical CLIP using different 512-to-N_C compression layers, denoted as $C\ N_C = 8$ and $C\ N_C = 16$. We find that QCLIP with $N_Q = 8$ matches the performance of a classical CLIP with $N_C = 16$ trained on the same dataset, indicating the improved representation learning capability enhanced by QuNNs. The proposed dense encoding provides an average of 5.4% accuracy improvement compared to the baseline angle encoding. Moreover, increasing the width of the QuNN (i.e., N_Q) improves the QCLIP accuracy, demonstrating the scalability of our approach.

We also explored *data re-uploading* [84] and *variational encoding* [85], which are two recently proposed encoding techniques to improve QuNN performance. The key idea of *data re-uploading* is to repeatedly apply the classical-to-quantum encoder, $\mathbf{E}(\mathbf{x})$, before each parameterized VQC ansatz, $\mathbf{U_k}(\theta_\mathbf{k})$. *Variational encoding* proposes to introduce trainable parameters to a classical-to-quantum encoder by defining a variational encoder function, $\mathbf{E}(\mathbf{x}{\cdot}\theta)$, where the parameter set θ is pre-trained to produce faithful quantum presentations in which data from different clusters are separated. We refer interested readers to [84] and [85] for a more detailed explanation and demonstration.

Appendix B. Quantum Neural Network Circuit ansatzes

In this work, we survey the recently proposed QuNN circuit ansatzes and identify four designs that have demonstrated state-of-the-art performance as shown in Figure B1. We denote these four designs respectively as *C14* [44], *QMLP* [45], *DAC22* [46], and

DATE22 [47]. These four ansatzes all follow the general structure summarized in Figure 2 with a single-qubit rotation layer followed by a two-qubit entanglement layer. Specifically, *DATE22* adopt the early designs [25] that utilize fixed two-qubit `CNOT` gates to force maximum entangling power, while *C14*, *QMLP* and *DAC22* explore to replace `CNOT` with trainable entanglement two-qubit gates.



(a) C14 Circuit ansatz

(b) QMLP Circuit ansatz

(c) DAC22 Circuit ansatz

(d) DATE22 Circuit ansatz

Figure B1: Four state-of-the-art candidate QuNN circuit ansatzes.

Appendix C. Performance with Different Encoding Methods and QuNNs

To investigate all the candidate quantum encoding methods (Appendix A) and VQC ansatzes (Appendix B) in the proposed QCLIP framework, we run various experiments using Cifar10 [52] as the downstream dataset. Based on the performance evaluation, we identify the optimal QuNN circuits for each quantum component in the proposed QCLIP framework, including the quantum image/text encoder neural networks and the quantum prompt adapter neural networks.

Taking the design of quantum image and text encoder as an example, we report the performance comparison for different circuit ansatz selections using *one-shot prompt learning*. As shown in Table C1, we make two conclusions: (1) the proposed dense encoding method followed by the QuNN circuit ansatz in figure 4 achieves the best performance and thus is identified as the optimal QuNN implementation for the quantum image and text encoder neural networks. (2) *data re-uploading* [84] and *variational encoding* [85] achieves negligible accuracy improvement or even performance degradation in such a hybrid quantum-classical framework. Therefore, we do not recommend using these two methods in QCLIP.

We follow the similar approach described above and identify the optimal QuNN circuit for the quantum prompt adapter neural networks shown in Figure 5.

Theoretical Insights. Theoretical research [96] on data encoding interprets quantum machine learning models as a Fourier-type sum, where the data encoding plays a crucial role in determining the functions the model can access and how these accessible functions can be combined. Consequently, the data encoding significantly influences the expressivity of the model. By applying this analysis, we propose that dense encoding outperforms single-qubit rotation-based angle encoding, likely due to the two-layer `RY-U1` gate in dense encoding, enabling access to frequency spectra with two frequencies, in contrast to angle encoding's single frequency. However, it is important to consider that increasing encoding density also leads to higher training complexity. Considering the problem set used in this work, we find the two-layer `RY-U1` dense encoding to be the optimal choice for our quantum machine learning models.

The theoretical analysis in VQC circuit ansatz [44] primarily explores the impact of circuit entanglement capacity on the expressivity of quantum machine learning models. As of now, there is no universally agreed-upon optimal VQC design, and VQC circuits are typically empirically designed. However, there is a common consensus that adaptive and trainable entanglement capabilities can be beneficial for QML algorithms compared to fixed maximized entanglement provided by fixed `CNOT` gates.

Table C1: QCLIP performance on one-shot prompt learning using different encoding methods and VQC circuit ansatzes.

Ansatz	**Encoding Scheme**		
	Dense	**Reuploading**	**Variational**
C14 [44]	**54.06%**	53.18%	53.31%
QMLP [45]	51.61%	**50.83%**	50.12%
DAC22 [46]	51.61%	51.42%	**52.25%**
DATE22 [47]	51.25%	**52.42%**	51.58%
QCLIP (**This Work**)	55.62%	53.65%	53.42%

Appendix D. Loss Functions

Here we formally define the contrastive loss [68,69] that has also been explored in QCLIP training. Contrastive loss defines two loss functions named image-to-text contrastive loss, i.e., $l_i^{(I\to T)}$, and text-to-image contrastive loss, i.e., $l_t^{(T\to I)}$. The image-to-text contrastive loss for the i_{th} image and the text-to-image contrastive loss for the t_{th} text can be calculated by the following equations D.1 and D.2, where i=1,2,..., B labels the input image feature in a batch and $\tau \in B^+$ represents a temperature parameter.

$$l_i^{(I\to T)} = -log\frac{exp(< I_i, T_t >)/\tau}{\sum_{t=1}^{B} exp(< I_i, T_t > /\tau)} \tag{D.1}$$

$$l_t^{(T\to I)} = -log\frac{exp(< T_t, I_i >)/\tau}{\sum_{i=1}^{B} exp(< T_t, I_i > /\tau)} \tag{D.2}$$

The final training loss is defined as the weighted sum of the above two losses. For batch training, the averaged loss is calculated using the following equation D.3, where $\lambda \in [0, 1]$ is a scaling hyperparameter.

$$loss = \frac{1}{B}\sum_{p=1}^{B}(\lambda l_p^{(I\to T)} + (1-\lambda) l_p^{(T\to I)}) \tag{D.3}$$

Appendix E. Parameter Initialization

We study both *uniform* initialization and *Gaussian* initialization in QCLIP training. Below we provide details for each initialization method.

Uniform initialization generates the initial values for the trainable parameters from Uniform distribution. The general formulation is shown in Equation E.1 with a minimal value a and a maximal value b. In QCLIP, we set the minimal and maximal values respectively to 0 and $\frac{\pi}{2}$, as shown in Equation E.2.

$$f(x) = \begin{cases} \frac{1}{a-b}, & a < x < b \\ 0, & else. \end{cases} \tag{E.1}$$

$$f(weight) = \begin{cases} \frac{2}{\pi}, & 0 < weight < \frac{\pi}{2} \\ 0, & else. \end{cases} \tag{E.2}$$

Gaussian initialization generates initial values from a Gaussian distribution. We show the general formulation for Gaussian distribution in Equation E.3 with a mean value μ and a standard deviation σ. Inspired by classical Xavier initialization [91], we utilize the information of QuNN structures and initialize parameters according to the network width N_Q. The Gaussian initialization used in QCLIP can be formalized by Equation E.4.

$$X \sim \mathcal{N}(\mu, \sigma^2), f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{E.3}$$

$$Weight \sim \mathcal{N}(0, N_Q), f(weight) = \frac{1}{\sqrt{2\pi N_Q}} exp(-\frac{weight^2}{2N_Q}) \tag{E.4}$$

References

[1] Preskill J 2018 Quantum Computing in the NISQ era and beyond *Quantum* vol 2 (Verein zur Förderung des Open Access Publizierens in den Quantenwissenschaften) p 79 ISSN 2521-327X URL https://doi.org/10.22331/q-2018-08-06-79

[2] "IonQ Forte" https://ionq.com/quantum-systems/forte/

[3] "IBM Quantum Heron" https://research.ibm.com/blog/ibm-quantum-roadmap-2025/

[4] Shor P W 1999 Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer *SIAM review* vol 41 (SIAM) pp 303–332

[5] Grover L K 1996 A fast quantum mechanical algorithm for database search *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing* pp 212–219

[6] Fowler A G, Mariantoni M, Martinis J M and Cleland A N 2012 Surface codes: Towards practical large-scale quantum computation *Physical Review A* vol 86 (APS) p 032324

[7] Childs A M, Maslov D, Nam Y, Ross N J and Su Y 2018 Toward the first quantum simulation with quantum speedup *Proceedings of the National Academy of Sciences* vol 115 (National Acad Sciences) pp 9456–9461

[8] Campbell E, Khurana A and Montanaro A 2019 Applying quantum algorithms to constraint satisfaction problems *Quantum* vol 3 (Verein zur Förderung des Open Access Publizierens in den Quantenwissenschaften) p 167

[9] Kivlichan I D, Gidney C, Berry D W, Wiebe N, McClean J, Sun W, Jiang Z, Rubin N, Fowler A and Aspuru-Guzik A 2020 Improved fault-tolerant quantum simulation of condensed-phase correlated electrons via trotterization *Quantum* vol 4 (Verein zur Förderung des Open Access Publizierens in den Quantenwissenschaften) p 296

[10] Gidney C and Ekerå M 2021 How to factor 2048 bit rsa integers in 8 hours using 20 million noisy qubits *Quantum* vol 5 (Verein zur Förderung des Open Access Publizierens in den Quantenwissenschaften) p 433

[11] Lee J, Berry D W, Gidney C, Huggins W J, McClean J R, Wiebe N and Babbush R 2021 Even more efficient quantum computations of chemistry through tensor hypercontraction *PRX Quantum* vol 2 (APS) p 030305

[12] Lemieux J, Duclos-Cianci G, Sénéchal D and Poulin D 2021 Resource estimate for quantum many-body ground-state preparation on a quantum computer *Physical Review A* vol 103 (APS) p 052408

[13] Shor P W 1995 Scheme for reducing decoherence in quantum computer memory *Phys. Rev. A* vol 52 pp R2493–R2496

[14] Lidar D A and Brun T A 2013 *Quantum error correction* (Cambridge university press)

[15] Terhal B M 2015 Quantum error correction for quantum memories *Rev. Mod. Phys.* vol 87 pp 307–346

[16] Bharti K, Cervera-Lierta A, Kyaw T H, Haug T, Alperin-Lea S, Anand A, Degroote M, Heimonen H, Kottmann J S and Menke T 2022 Noisy intermediate-scale quantum algorithms *Reviews of Modern Physics* vol 94 (APS) p 015004

[17] Farhi E, Goldstone J and Gutmann S 2014 A quantum approximate optimization algorithm *arXiv preprint arXiv:1411.4028*

[18] Cao Y, Romero J, Olson J P, Degroote M, Johnson P D, Kieferová M, Kivlichan I D, Menke T, Peropadre B and Sawaya N P 2019 Quantum chemistry in the age of quantum computing *Chemical reviews* vol 119 (ACS Publications) pp 10856–10915

[19] Endo S, Cai Z, Benjamin S C and Yuan X 2021 Hybrid quantum-classical algorithms and quantum error mitigation *booktitle of the Physical Society of Japan* vol 90 (The Physical Society of Japan) p 032001

[20] McArdle S, Endo S, Aspuru-Guzik A, Benjamin S C and Yuan X 2020 Quantum computational chemistry *Reviews of Modern Physics* vol 92 (APS) p 015003

[21] Peruzzo A, McClean J, Shadbolt P, Yung M H, Zhou X Q, Love P J, Aspuru-Guzik A and O'brien J L 2014 A variational eigenvalue solver on a photonic quantum processor *Nature communications* vol 5 (Nature Publishing Group) pp 1–7

[22] Kandala A, Mezzacapo A, Temme K, Takita M, Brink M, Chow J M and Gambetta J M 2017 Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets *Nature* vol 549 (Nature Publishing Group) pp 242–246

[23] Cerezo M, Sharma K, Arrasmith A and Coles P J 2020 Variational quantum state eigensolver *arXiv preprint arXiv:2004.01372*

[24] Huang H Y, Broughton M, Cotler J, Chen S, Li J, Mohseni M, Neven H, Babbush R, Kueng R, Preskill J and McClean J R 2022 Quantum advantage in learning from experiments *Science* vol

376 pp 1182–1186 URL https://www.science.org/doi/abs/10.1126/science.abn7293

[25] Biamonte J, Wittek P, Pancotti N, Rebentrost P, Wiebe N and Lloyd S 2017 Quantum machine learning *Nature* vol 549 (Nature Publishing Group) pp 195–202

[26] Lloyd S, Mohseni M and Rebentrost P 2013 Quantum algorithms for supervised and unsupervised machine learning *arXiv e-prints* p arXiv:1307.0411 (*Preprint* 1307.0411) URL https://ui.adsabs.harvard.edu/abs/2013arXiv1307.0411L

[27] Schuld M and Killoran N 2019 Quantum machine learning in feature hilbert spaces *Physical review letters* vol 122 (APS) p 040504

[28] Havlíček V, Córcoles A D, Temme K, Harrow A W, Kandala A, Chow J M and Gambetta J M 2019 Supervised learning with quantum-enhanced feature spaces *Nature* vol 567 (Nature Publishing Group) pp 209–212

[29] Huang H Y, Broughton M, Mohseni M, Babbush R, Boixo S, Neven H and McClean J R 2021 Power of data in quantum machine learning *Nature communications* vol 12 (Nature Publishing Group) pp 1–9

[30] Lloyd S and Weedbrook C 2018 Quantum generative adversarial learning *Physical review letters* vol 121 (APS) p 040502

[31] Dallaire-Demers P L and Killoran N 2018 Quantum generative adversarial networks *Physical Review A* vol 98 (APS) p 012324

[32] Havlícek V, Córcoles A D, Temme K, Harrow A W, Kandala A, Chow J M and Gambetta J M 2019 Supervised learning with quantum-enhanced feature spaces *Nature* vol 567 pp 209–212 URL https://doi.org/10.1038/s41586-019-0980-2

[33] Lloyd S and Weedbrook C 2018 Quantum generative adversarial learning *Physical Review Letters* vol 121 p 040502 URL https://booktitles.aps.org/prl/abstract/10.1103/PhysRevLett.121.040502

[34] Xia R and Kais S 2018 Quantum machine learning for electronic structure calculations *Nature communications* vol 9 (Nature Publishing Group) pp 1–6

[35] Choudhary K, Bercx M, Jiang J, Pachter R, Lamoen D and Tavazza F 2019 Accelerated discovery of efficient solar cell materials using quantum and machine-learning methods *Chemistry of materials* vol 31 (ACS Publications) pp 5900–5908

[36] Cao Y, Romero J and Aspuru-Guzik A 2018 Potential of quantum computing for drug discovery *IBM booktitle of Research and Development* vol 62 (IBM) pp 6–1

[37] Amin J, Sharif M, Gul N, Kadry S and Chakraborty C 2022 Quantum machine learning architecture for covid-19 classification based on synthetic data generation using conditional adversarial neural network *Cognitive Computation* vol 14 (Springer) pp 1677–1688

[38] Alcazar J, Leyton-Ortega V and Perdomo-Ortiz A 2020 Classical versus quantum models in machine learning: insights from a finance application *Machine Learning: Science and Technology* vol 1 (IOP Publishing) p 035003

[39] Coyle B, Henderson M, Le J C J, Kumar N, Paini M and Kashefi E 2021 Quantum versus classical generative modeling in finance *Quantum Science and Technology* vol 6 (IOP Publishing) p 024013

[40] Parsons D F 2011 Possible medical and biomedical uses of quantum computing *Neuroquantology* vol 9 (NeuroQuantology)

[41] Crawford S E, Shugayev R A, Paudel H P, Lu P, Syamlal M, Ohodnicki P R, Chorpening B, Gentry R and Duan Y 2021 Quantum sensing for energy applications: Review and perspective *Advanced Quantum Technologies* vol 4 (Wiley Online Library) p 2100049

[42] Focardi S, Fabozzi F J and Mazza D 2020 Quantum option pricing and quantum finance *The booktitle of Derivatives* vol 28 (Institutional Investor booktitles Umbrella) pp 79–98

[43] Bengio Y, Courville A C and Vincent P 2013 Representation learning: A review and new perspectives *IEEE Trans. Pattern Anal. Mach. Intell.* vol 35 pp 1798–1828 URL https://doi.org/10.1109/TPAMI.2013.50

[44] Sim S, Johnson P D and Aspuru-Guzik A 2019 Expressibility and entangling capability of

parameterized quantum circuits for hybrid quantum-classical algorithms *Advanced Quantum Technologies* vol 2 p 1900070 URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qute.201900070

[45] Chu C, Chia N, Jiang L and Chen F 2022 QMLP: an error-tolerant nonlinear quantum MLP architecture using parameterized two-qubit gates *ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED)* (ACM) pp 4:1–4:6 URL https://doi.org/10.1145/3531437.3539719

[46] Wang H, Gu J, Ding Y, Li Z, Chong F T, Pan D Z and Han S 2022 Quantumnat: quantum noise-aware training with noise injection, quantization and normalization *DAC '22: 59th ACM/IEEE Design Automation Conference, San Francisco, California, USA, July 10 - 14, 2022* (ACM) pp 1–6 URL https://doi.org/10.1145/3489517.3530400

[47] Patel T, Silver D and Tiwari D 2022 OPTIC: A practical quantum binary classifier for near-term quantum computers *Design, Automation & Test in Europe Conference & Exhibition (DATE)* (IEEE) pp 334–339 URL https://doi.org/10.23919/DATE54114.2022.9774707

[48] Schuld M, Bocharov A, Svore K M and Wiebe N 2020 Circuit-centric quantum classifiers *Phys. Rev. A* vol 101 (American Physical Society) p 032308

[49] Patel T, Silver D and Tiwari D 2022 Optic: a practical quantum binary classifier for near-term quantum computers *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)* (IEEE) pp 334–339

[50] Niu M Y, Zlokapa A, Broughton M, Boixo S, Mohseni M, Smelyanskyi V and Neven H 2022 Entangling quantum generative adversarial networks *Physical Review Letters* vol 128 (APS) p 220505

[51] Kübler J M, Arrasmith A, Cincio L and Coles P J 2020 An adaptive optimizer for measurement-frugal variational algorithms *Quantum* vol 4 (Verein zur Förderung des Open Access Publizierens in den Quantenwissenschaften) p 263

[52] Krizhevsky A, Hinton G *et al.* 2009 Learning multiple layers of features from tiny images (Toronto, ON, Canada)

[53] Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I 2021 Learning transferable visual models from natural language supervision *International Conference on Machine Learning (ICML* (*Proceedings of Machine Learning Research* vol 139) (PMLR) pp 8748–8763 URL http://proceedings.mlr.press/v139/radford21a.html

[54] Bengio Y, Courville A C and Vincent P 2013 Representation learning: A review and new perspectives *IEEE Trans. Pattern Anal. Mach. Intell.* vol 35 pp 1798–1828 URL https://doi.org/10.1109/TPAMI.2013.50

[55] Zeiler M D and Fergus R 2014 Visualizing and understanding convolutional networks *European conference on computer vision (ECCV)* vol 8689 (Springer) pp 818–833 URL https://doi.org/10.1007/978-3-319-10590-1_53

[56] Razavian A S, Azizpour H, Sullivan J and Carlsson S 2014 CNN features off-the-shelf: An astounding baseline for recognition *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE Computer Society) pp 512–519 URL https://doi.org/10.1109/CVPRW.2014.131

[57] LeCun Y, Bengio Y and Hinton G E 2015 Deep learning *Nature* vol 521 pp 436–444 URL https://doi.org/10.1038/nature14539

[58] Grover A and Leskovec J 2016 node2vec: Scalable feature learning for networks *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM) pp 855–864 URL https://doi.org/10.1145/2939672.2939754

[59] Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I and Abbeel P 2016 Infogan: Interpretable representation learning by information maximizing generative adversarial nets *Advances in Neural Information Processing Systems (NIPS)* pp 2172–2180 URL https://proceedings.neurips.cc/paper/2016/hash/7c9d0b1f96aebd7b5eca8c3edaa19ebb-Abstract.html

[60] Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D 2017 Grad-cam: Visual explanations from deep networks via gradient-based localization *IEEE International Conference on Computer Vision (ICCV)* (IEEE Computer Society) pp 618–626 URL https://doi.org/10.1109/ICCV.2017.74

[61] Hjelm R D, Fedorov A, Lavoie-Marchildon S, Grewal K, Bachman P, Trischler A and Bengio Y 2019 Learning deep representations by mutual information estimation and maximization *International Conference on Learning Representations (ICLR)* (OpenReview.net) URL https://openreview.net/forum?id=Bklr3j0cKX

[62] Kingma D P and Welling M 2014 Auto-encoding variational bayes *International Conference on Learning Representations (ICLR)* ed Bengio Y and LeCun Y URL http://arxiv.org/abs/1312.6114

[63] Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A C and Bengio Y 2014 Generative adversarial nets *Advances in Neural Information Processing Systems (NIPS)* pp 2672–2680 URL https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

[64] Doersch C, Gupta A and Efros A A 2015 Unsupervised visual representation learning by context prediction *IEEE International Conference on Computer Vision (ICCV)* (IEEE Computer Society) pp 1422–1430 URL https://doi.org/10.1109/ICCV.2015.167

[65] Zhang R, Isola P and Efros A A 2016 Colorful image colorization *European Conference on Computer Vision (ECCV)* (*Lecture Notes in Computer Science* vol 9907) ed Leibe B, Matas J, Sebe N and Welling M (Springer) pp 649–666 URL https://doi.org/10.1007/978-3-319-46487-9_40

[66] Gidaris S, Singh P and Komodakis N 2018 Unsupervised representation learning by predicting image rotations *International Conference on Learning Representations (ICLR)* (OpenReview.net) URL https://openreview.net/forum?id=S1v4N2l0-

[67] Bachman P, Hjelm R D and Buchwalter W 2019 Learning representations by maximizing mutual information across views *Advances in Neural Information Processing Systems (NeurIPS)* pp 15509–15519 URL https://proceedings.neurips.cc/paper/2019/hash/ddf354219aac374f1d40b7e760ee5bb7-Abstract.html

[68] Chen T, Kornblith S, Norouzi M and Hinton G E 2020 A simple framework for contrastive learning of visual representations *Proceedings of the 37th International Conference on Machine Learning (ICML)* (*Proceedings of Machine Learning Research* vol 119) (PMLR) pp 1597–1607 URL http://proceedings.mlr.press/v119/chen20j.html

[69] Chen T, Kornblith S, Swersky K, Norouzi M and Hinton G E 2020 Big self-supervised models are strong semi-supervised learners *Advances in Neural Information Processing Systems (NeurIPS)* URL https://proceedings.neurips.cc/paper/2020/hash/fcbc95ccdd551da181207c0c1400c655-Abstract.html

[70] Quattoni A, Collins M and Darrell T 2007 Learning visual representations using images with captions *Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE Computer Society) URL https://doi.org/10.1109/CVPR.2007.383173

[71] Srivastava N and Salakhutdinov R 2012 Multimodal learning with deep boltzmann machines *Annual Conference on Neural Information Processing Systems (NIPS)* ed Bartlett P L, Pereira F C N, Burges C J C, Bottou L and Weinberger K Q pp 2231–2239 URL https://proceedings.neurips.cc/paper/2012/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html

[72] Joulin A, van der Maaten L, Jabri A and Vasilache N 2016 Learning visual features from large weakly supervised data *European Conference on Computer Vision (ECCV)* (*Lecture Notes in Computer Science* vol 9911) (Springer) pp 67–84 URL https://doi.org/10.1007/978-3-319-46478-7_5

[73] Li A, Jabri A, Joulin A and van der Maaten L 2017 Learning visual n-grams from web data *IEEE International Conference on Computer Vision (ICCV)* (IEEE Computer Society) pp 4193–4202 URL http://doi.ieeecomputersociety.org/10.1109/ICCV.2017.449

[74] Desai K and Johnson J 2021 Virtex: Learning visual representations from textual annotations *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Computer Vision Foundation / IEEE) pp 11162–11173 URL https://openaccess.thecvf.com/content/CVPR2021/html/Desai_VirTex_Learning_Visual_Representations_From_Textual_Annotations_CVPR_2021_paper.html

[75] Locatello F, Bauer S, Lucic M, Rätsch G, Gelly S, Schölkopf B and Bachem O 2020 A sober look at the unsupervised learning of disentangled representations and their evaluation *booktitle of Machine Learning Research (JMLR)* vol 21 pp 209:1–209:62 URL http://jmlr.org/papers/v21/19-976.html

[76] Elsayed G F, Goodfellow I and Sohl-Dickstein J 2018 Adversarial reprogramming of neural networks *arXiv preprint arXiv:1806.11146*

[77] Li X L and Liang P 2021 Prefix-tuning: Optimizing continuous prompts for generation *arXiv preprint arXiv:2101.00190*

[78] Bahng H, Jahanian A, Sankaranarayanan S and Isola P 2022 Visual prompting: Modifying pixel space to adapt pre-trained models *arXiv preprint arXiv:2203.17274*

[79] LaRose R and Coyle B 2020 Robust data encodings for quantum classifiers *CoRR* vol abs/2003.01695 (*Preprint* 2003.01695) URL https://arxiv.org/abs/2003.01695

[80] McClean J R, Romero J, Babbush R and Aspuru-Guzik A 2016 The theory of variational hybrid quantum-classical algorithms *New booktitle of Physics* vol 18 (IOP Publishing) p 023023

[81] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I *et al.* 2019 Language models are unsupervised multitask learners *OpenAI blog* vol 1 p 9 URL https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[82] Mari A, Bromley T R, Izaac J A, Schuld M and Killoran N 2020 Transfer learning in hybrid classical-quantum neural networks *Quantum* vol 4 p 340 URL https://doi.org/10.22331/q-2020-10-09-340

[83] Jaderberg B, Anderson L W, Xie W, Albanie S, Kiffner M and Jaksch D 2022 Quantum self-supervised learning *Quantum Science and Technology* vol 7 (IOP Publishing) p 035005 URL https://dx.doi.org/10.1088/2058-9565/ac6825

[84] Pérez-Salinas A, Cervera-Lierta A, Gil-Fuster E and Latorre J I 2020 Data re-uploading for a universal quantum classifier *Quantum* vol 4 p 226 URL https://doi.org/10.22331/q-2020-02-06-226

[85] Chu C, Skipper G, Swany M and Chen F 2022 IQGAN: Robust Quantum Generative Adversarial Network for Image Synthesis On NISQ Devices *arXiv e-prints* p arXiv:2210.16857 (*Preprint* 2210.16857) URL https://ui.adsabs.harvard.edu/abs/2022arXiv221016857C

[86] Sharma P, Ding N, Goodman S and Soricut R 2018 Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)* (Association for Computational Linguistics) pp 2556–2565 URL https://aclanthology.org/P18-1238/

[87] Zhang Y, Jiang H, Miura Y, Manning C D and Langlotz C P 2020 Contrastive learning of medical visual representations from paired images and text *CoRR* vol abs/2010.00747 (*Preprint* 2010.00747) URL https://arxiv.org/abs/2010.00747

[88] PyTorch PyTorch https://pytorch.org/

[89] Pennylane https://pennylane.ai/

[90] Zhang K, Hsieh M, Liu L and Tao D 2022 Gaussian initializations help deep variational quantum circuits escape from the barren plateau *CoRR* vol abs/2203.09376 (*Preprint* 2203.09376) URL https://doi.org/10.48550/arXiv.2203.09376

[91] Kumar S K 2017 On weight initialization in deep neural networks *arXiv preprint arXiv:1704.08863*

[92] LeCun Y 1998 The mnist database of handwritten digits *http://yann. lecun. com/exdb/mnist/*

[93] Parkhi O M, Vedaldi A, Zisserman A and Jawahar C V 2012 Cats and dogs *2012 IEEE Conference on Computer Vision and Pattern Recognition* pp 3498–3505

[94] Bossard L, Guillaumin M and Van Gool L 2014 Food-101–mining discriminative components with random forests *European conference on computer vision* (Springer) pp 446–461

[95] Grant E, Wossnig L, Ostaszewski M and Benedetti M 2019 An initialization strategy for addressing barren plateaus in parametrized quantum circuits *Quantum* vol 3 (Verein zur Förderung des Open Access Publizierens in den Quantenwissenschaften) p 214

[96] Schuld M, Sweke R and Meyer J J 2021 Effect of data encoding on the expressive power of variational quantum-machine-learning models *Physical Review A* vol 103 (APS) p 032430