CLIP-ACQUA: CLIP Autoencoder-based Classic-Quantum Latent Space Reduction

Pablo Rivas , *Senior Member, IEEE*
Baylor University
Email: Pablo_Rivas@Baylor.edu

Liang Zhao
St. Ambrose University
Email: ZhaoLiang@sau.edu

***Abstract*—Applications of quantum machine learning algorithms are currently still being studied. Recent work suggests that classical gradient descent techniques can effectively train variational quantum circuits. We propose to train quantum variational circuits to find smaller text and image embeddings that preserve contrastive-learning distances based on CLIP large embeddings. This is a critical task since fine-tuning CLIP to produce low-dimensional embeddings is prohibitively expensive. We introduce CLIP-ACQUA, a model trained in a self-supervised configuration from CLIP embeddings to reduce the latent space. We use CLIP-ACQUA on a sizeable unlabelled corpus of text and images to demonstrate its effectiveness. Our experiments show that we can obtain smaller latent spaces that preserve the original embedding distances inferred during contrastive learning. Furthermore, using our model requires no fine-tuning of CLIP, preserving its original robustness and structure. The data used aids in modeling consumer-to-consumer online marketplaces.**

***Index Terms*—quantum machine learning, self-supervised learning, quantum variational circuits**

I. Introduction

Contrastive Language-Image Pre-training (CLIP) models have gained popularity in text-image pairs research, and it has motivated many applications [1]. CLIP can produce large text and image embeddings with a baseline model of 63 million parameters; however, training CLIP from scratch requires great compute resources, fine-tuning can be expensive, and for many applications, the size of the embeddings is large. Using hybrid variational quantum machine learning, we aim to find smaller text and image embeddings that preserve contrastive-learning distances. Although there have been recent advances [2]–[5], quantum machine learning applications are largely understudied.

In this paper, we introduce CLIP-ACQUA, a model trained from CLIP image-text embeddings to reduce the latent space while preserving distances using quantum variational circuits in a self-supervised configuration, as shown in Fig. 1. By applying this CLIP-ACQUA model to a large unlabelled corpus of text and images, we obtain smaller latent spaces that preserve the original embedding distances obtained during contrastive learning. Using our model requires no fine-tuning of CLIP, preserving its original latent structure. The data used as a demonstration aids in modeling consumer-to-consumer online marketplaces to detect illicit activities.

We discuss background material and methodology in Sec. II. Results and conclusions are in Sec. III.

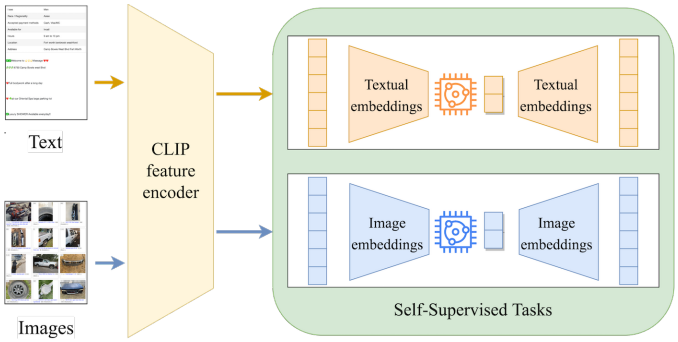

Fig. 1. A hybrid classic-quantum architecture to reduce the dimensions of CLIP image-text embeddings.

II. Background

Some of the most exciting work in the quantum machine learning area has occurred in recent years. The most related work in variational approaches can be found in [6]–[8]. Quantum machine learning research has been widely influenced by these works and our research continues the work of trainable variational quantum circuits.

The work by Mari *et al.* [9], is closely related to our work in the sense that both approaches combine classic and quantum approaches. However, the authors focus on ResNet-based transfer learning.

A. Variational quantum circuits

Variational quantum circuits [9], can be defined in terms of a unitary operation, U, implemented as a variational circuit on an input state $|\hat{\mathrm{x}}\rangle$, that produces the the output state $|y\rangle$ as follows: $|\hat{\mathrm{x}}\rangle \rightarrow |y\rangle = U(\mathrm{w})|\hat{\mathrm{x}}\rangle$, where w denotes the parameters of the variational circuit. Then decompose the unitary operation in the following quantum layers.

1) Hadamard operators layer: The Hadamard operator on a qubit facilitates superposition: $H = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$.

2) Single qubit Y rotation layer: The *trainable* rotation of a qubit makes it change the spin angle, ϕ, as follows:

$$R_Y(\phi) = e^{-i\phi\sigma_Y/2} = \begin{bmatrix} \cos(\phi/2) & -\sin(\phi/2) \\ \sin(\phi/2) & \cos(\phi/2) \end{bmatrix}. \quad (1)$$

3) CNOT qubit entangling layer: The CNOT operation, $\bigoplus$, links qubits and propagates superposition.

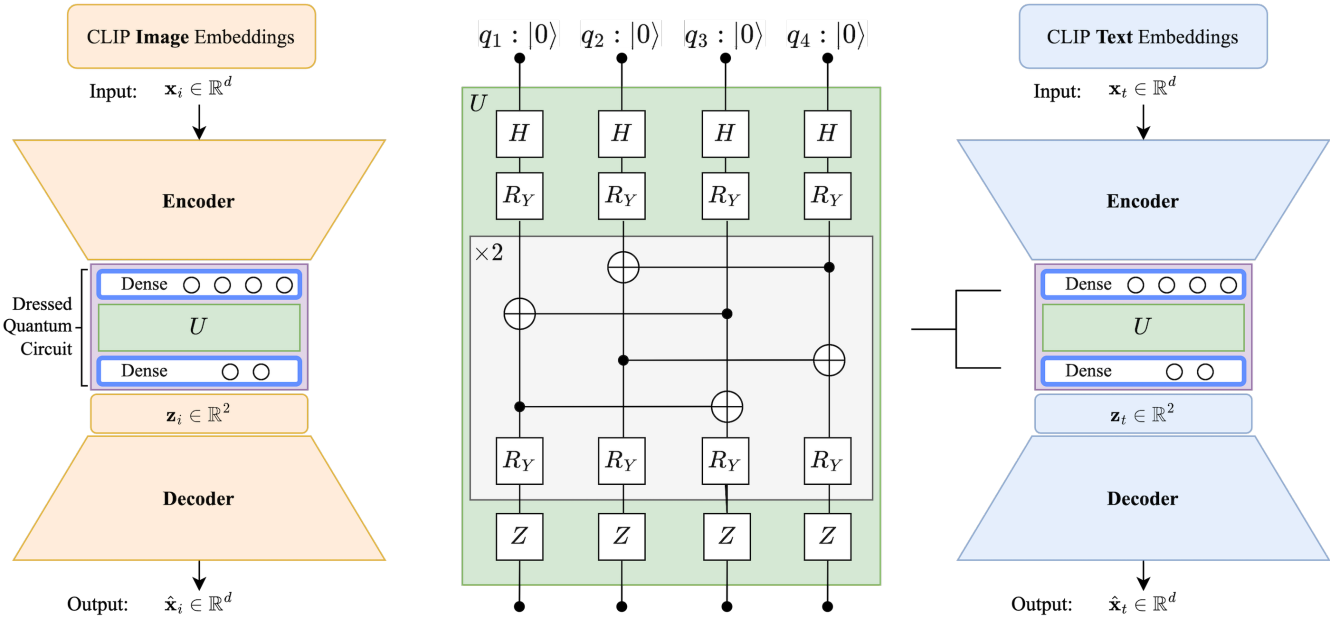

Fig. 2. Two hybrid autoencoders have dressed variational quantum circuits (middle). The autoencoder on the left trains on CLIP image embeddings, while the one on the right uses textual embeddings. Both autoencoders are conditioned to maintain original CLIP distances.

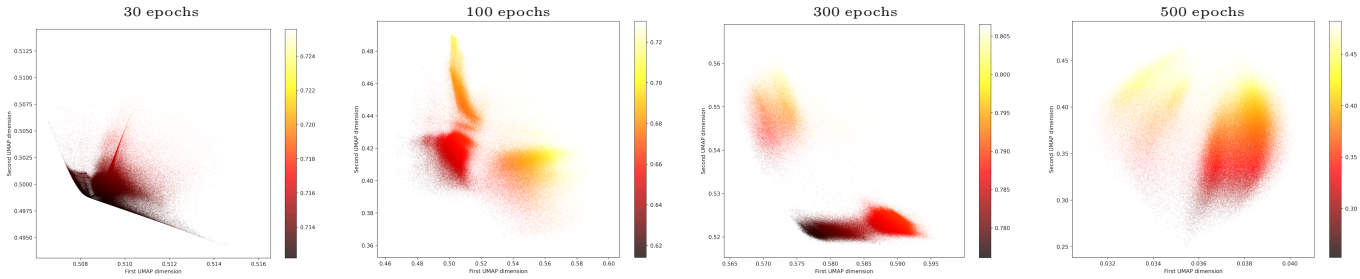

Fig. 3. Data visualization across different training stages, where the color indicates the norm of the embedding vector. From left to right, the quantum hybrid approach shows how the data moves in the new latent space to satisfy the distance constraints, facilitating clustering.

4) Expectation layer over Pauli Z operators: Finally, the output of the circuit is based on the expected value of several measurements. The measurements are applied after the Pauli Z operator defined as follows: $\sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$.

B. Quantum dressed circuit

The process of dressing a quantum circuit implies adding a single dense layer before and after the quantum circuit, as shown in Fig. 2 in the middle [2], [9]. The number of neurons in the input layer matches the number of qubits.

C. Loss function

For $\mathrm{x} \in \mathbb{R}^{512}$ as an input embedding, our loss is:

$$\mathcal{L}(\underbrace{\theta_i, \theta_t}_{\text{image-text quantum autoencoder parameters}}; \underbrace{\mathrm{x}_i, \mathrm{x}_t, d_\mathrm{x}}_{\text{CLIP image-text embeddings \& distance}}) = \alpha_i ||\mathrm{x}_i - \hat{\mathrm{x}}_i||_1 + \alpha_t ||\mathrm{x}_t - \hat{\mathrm{x}}_t||_1 + \alpha_d |d_\mathrm{x} - ||\mathrm{z}_i - \mathrm{z}_t||_2|$$

where $\hat{\mathrm{x}}$ is the reconstruction, and $\mathrm{z} = q_\theta(\mathrm{x})$ is the new low dimensional embedding achieved through an encoder $q(\cdot)$. Minimizing this loss yields a new latent space that minimizes embedding reconstruction loss and preserves original distances. Note that for $\alpha_i = \alpha_t = \alpha_d = \frac{1}{3}$, the loss is an average of the three components.

III. Results and Conclusions

We collected publicly available ads from consumer-to-consumer online platforms where trafficking of stolen goods and sex is common. The data consists of 82.71G of posts that contain images and text. Duplicate posts are ignored and all unique image-text pairs are used. We trained the model and monitored the learning process as shown in Fig. 3. As it can be observed, When the elements of the loss $\mathcal{L}(\theta_i, \theta_t; \mathrm{x}_i, \mathrm{x}_t, d_\mathrm{x})$ are treated as a classic average, we have immediate reconstruction gains and progressive distance enforcement, which satisfies the main goal.

After the model is trained, it can be used to produce lower-dimensional CLIP-based embeddings for specific applications or datasets. Quantum advantage occurs upon deployment for real-time applications, having a broader impact as quantum technology becomes more accessible.

Acknowledgment

This work was, in part, funded by the National Science Foundation under grants CNS-2136961, and CNS-2210091.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763. 1

[2] P. Rivas, L. Zhao, and J. Orduz, "Hybrid quantum variational autoencoders for representation learning," in *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2021, pp. 52–57. 1, 2

[3] M. Schuld, "Quantum machine learning models are kernel methods," *arXiv e-prints*, pp. arXiv–2101, 2021. 1

[4] M. Schuld and N. Killoran, "Quantum machine learning in feature hilbert spaces," *Physical review letters*, vol. 122, no. 4, p. 040504, 2019. 1

[5] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, "Circuit-centric quantum classifiers," *Physical Review A*, vol. 101, no. 3, p. 032308, 2020. 1

[6] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio *et al.*, "Variational quantum algorithms," *Nature Reviews Physics*, pp. 1–20, 2021. 1

[7] A. Khoshaman, W. Vinci, B. Denis, E. Andriyash, H. Sadeghi, and M. H. Amin, "Quantum variational autoencoder," *Quantum Science and Technology*, vol. 4, no. 1, p. 014001, 2018. 1

[8] X. Yuan, S. Endo, Q. Zhao, Y. Li, and S. C. Benjamin, "Theory of variational quantum simulation," *Quantum*, vol. 3, p. 191, 2019. 1

[9] A. Mari, T. R. Bromley, J. Izaac, M. Schuld, and N. Killoran, "Transfer learning in hybrid classical-quantum neural networks," *Quantum*, vol. 4, p. 340, 2020. 1, 2