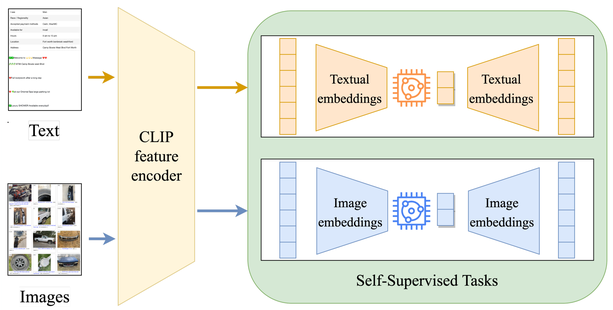
CLIP-ACQUA 📎💧 CLIP Autoencoder-based Classic-Quantum Latent Space Reduction

Pablo Rivas[1], *Senior Member, IEEE* Liang Zhao[2]
[1]Baylor University [2]St. Ambrose University

Problem Definition and Contribution

Goal: Find smaller text and image embeddings that preserve contrastive-learning distances using hybrid variational quantum machine learning



Motivation:

- Fine-tuning CLIP to produce low-dimensional embeddings is expensive.
- Quantum machine learning algorithms are largely understudied.

Key Contributions:

- A CLIP-ACQUA 📎💧 model can be trained from CLIP embeddings to reduce the latent space while preserving distances using *quantum variational circuits* in a self-supervised configuration.
- 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 robustness and manifolds.
- The data used as a demonstration aids in the modeling of consumer-to-consumer online marketplaces for the detection of illicit activities.

Designing the CLIP-ACQUA 📎💧 Model

Main idea: We train a CLIP-based model to reduce the dimensionality of text-image embedding pairs. The process of dimensionality reduction preserves the distances of the original latent space and takes advantage of variational quantum circuits. This model is trained to minimize reconstruction and distance losses:

image-text quantum autoencoder hyperparameters

$$\mathcal{L}(\theta_i, \theta_t\,;\,\mathrm{x}_i, \mathrm{x}_t, d_{\mathrm{x}}) = \alpha_i||\mathrm{x}_i - \widehat{\mathrm{x}}_i||_1 + \alpha_t||\mathrm{x}_t - \widehat{\mathrm{x}}_t||_1 + \alpha_d|d_{\mathrm{x}} - ||\mathrm{z}_i - \mathrm{z}_t||_2|$$

CLIP image-text embeddings and their distance

where $\mathrm{x} \in \mathbb{R}^{512}$ is the input, $\widehat{\mathrm{x}}$ is the reconstruction, and θ are the model parameters, 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.

CLIP-ACQUA 📎💧 Model

- The model uses a pre-trained model to produce image-text embedding pairs and their distance. The model uses a vision transformer with patches of size 32, available on HuggingFace as `ViT-B/32`.
- The model then is pre-trained using an autoencoder configuration using the embeddings. A variational quantum circuit is used to find the desired low-dimensional embeddings using gradient descent.

Variational Quantum Machine Learning

Variational quantum circuits have been recently studied in combination with different models, including neural networks, support vector machines, and other linear classifiers [McClean 2016, Schuld 2019]. Researchers [e.g. Mari 2020] define a quantum layer as a unitary operation, U, implemented as a variational circuit on an input state $|\hat{\mathrm{x}}\rangle$, that produces 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. It can be also found in the literature as a block diagram as follows:

$$|\hat{\mathbf{x}}\rangle \text{ — } U \text{ — } |y\rangle$$

The proposed quantum hybrid model is based on a variational circuit with many gates and operators represented in the following quantum layers:

- **Hadamard operators layer**. The Hadamard operator on a qubit is:

$$H = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

  Its primary purpose is to create superpositions.
- **Single qubit Y rotation layer**. A rotation of a qubit makes a qubit change the spin based on the rotation angle, ϕ, as follows:

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

  The rotation angle ϕ in our research is a trainable parameter.
- **CNOT qubit entangling layer**. The CNOT operation, defined as:

$$\text{CNOT} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

  is aimed at linking qubits, combining them, and propagating superposition across layers.
- **Expectation layer over Pauli Z operators**. Finally, the output of the circuit is a measurement that is calculated over many observations returning the expected value. In our case the measurements are applied after the Pauli Z operator defined as follows:

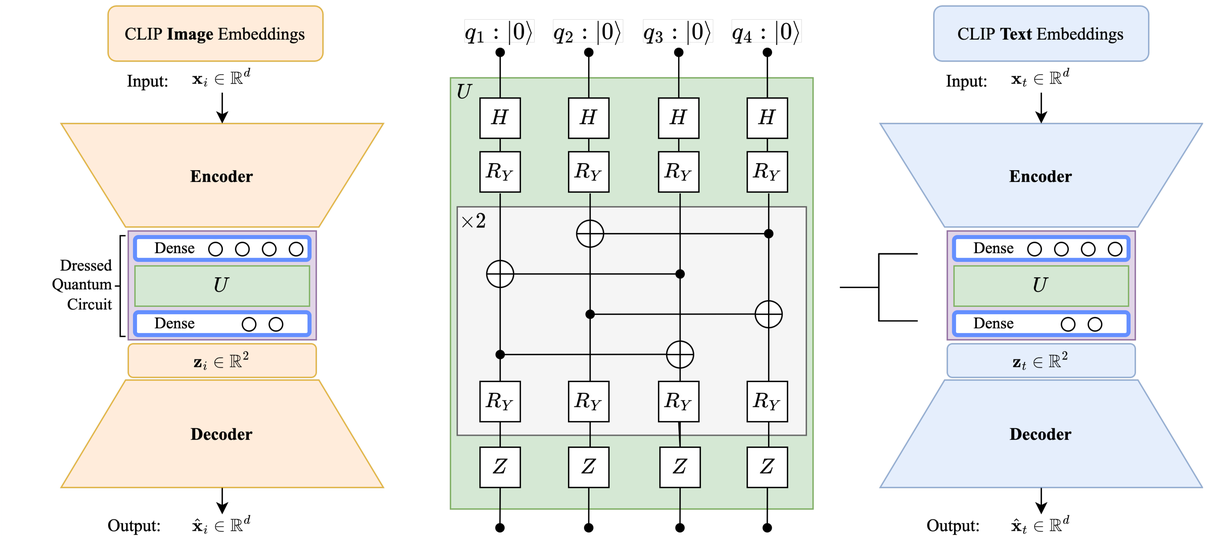
$$\sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

- **Quantum dressed circuit**. Once the circuit is defined, it needs to be dressed up to be combined with the classic autoencoder model. This process involves a simple process that adds a single dense layer before and after the quantum circuit [Mari 2020, Rivas 2021].
  The number of neurons in the input layer is set to match the number of qubits, in this study is four. The last layer has two neurons, as set to match our visualization intent in two dimensions. This is because we are interested in inspecting the latent space in two dimensions. However, this can be arbitrarily set to any latent space dimension as desired.

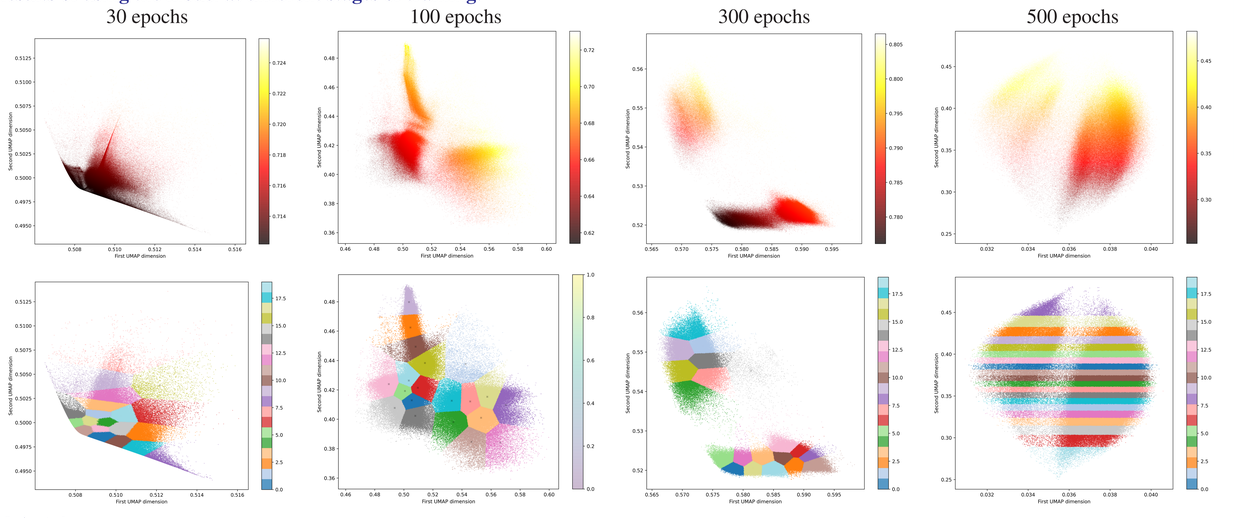
More details at: https://baylor.ai/?tag=quantum-ml
Contact: Pablo_Rivas@Baylor.edu

Experimental Architecture and Preliminary Results

Detailed architecture including the two-layered variational quantum circuit:



Results of using the model at different stages of training:



Main take aways

- 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 observe immediate reconstruction gains and progressive distance enforcement, as shown above.
- After the model is trained, it can be used to produce lower-dimensional CLIP-based embeddings for specific applications or datasets. Quantum advantage can be achieved upon deployment for real-time applications.

Dataset disclosures

- The data was collected from publicly available ads on consumer-to-consumer online platforms.
- 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.
- The dataset used supports research to identify illicit online activity such as trafficking of stolen goods and sex.

Acknowledgments

Part of this work was done while P. Rivas was funded by the National Science Foundation under grants CHE-1905043, CNS-2136961, and CNS-2210091.