QMLHEP: Task VIII

Task: Comment on potential ideas to extend this classical vision transformer architecture to a quantum vision transformer and sketch out the architecture in detail.

Response:

Vision Transformer (ViT) emerged as a competitive alternative to convolutional neural networks (CNNs) that are currently state-of-the-art in computer vision and widely used. The researchers are trying to extend this classical vision transformer architecture to a quantum vision transformer. In the "Quantum Vision Transformer" paper, A El Cherrat et al. have proposed some architectures of quantum vision transformers and seen promising results using those. The architectures are-

- Quantum Orthogonal Transformer
- Quantum Attention Mechanism
- Quantum Compound Transformer

The first two methods are quantum translation of the classical vision transformer but the Quantum Compound Transformer is novel and natively quantum. The architecture of the model is explained below-

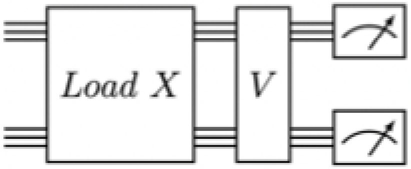

Fig. 1: Quantum Compound Transformer

In Fig. 1, "Load X" is a data loader for matrix X $\in \mathbf{R}^{nxd}$ and V is a quantum orthogonal layer. This architecture requires (n+d) number of qubits and circuit depth is O(log(n) + n * log(d) + log(n+d)) . The input to the circuit X corresponds to the patches of an image. The matrix data loader creates the state |X⟩, and after applying the quantum orthogonal layer on all n + d qubits, the resulting state is |Y⟩ = |$V^{(2)}$X⟩, where $V^{(2)\,i}$s the 2nd-order compound matrix.

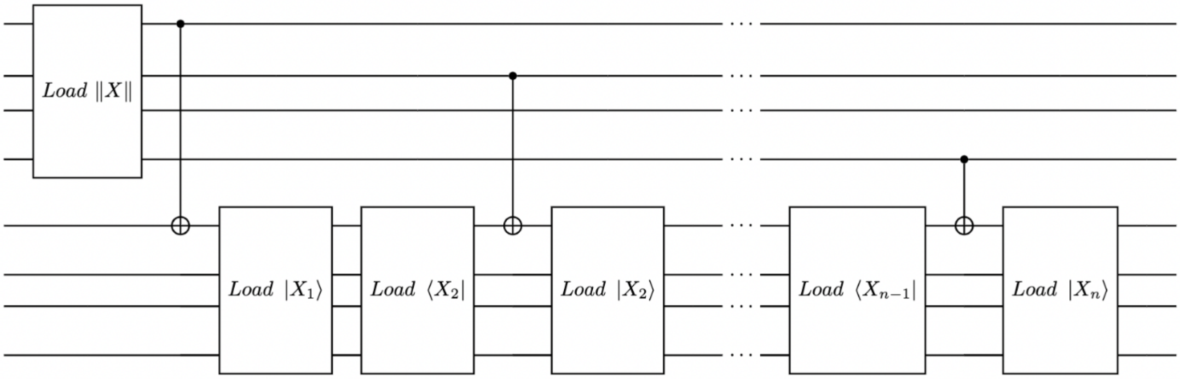

Fig. 2: Data loader circuit (Load X) for a matrix X $\in$ $\mathbf{R}^{\mathrm{nxd}}$

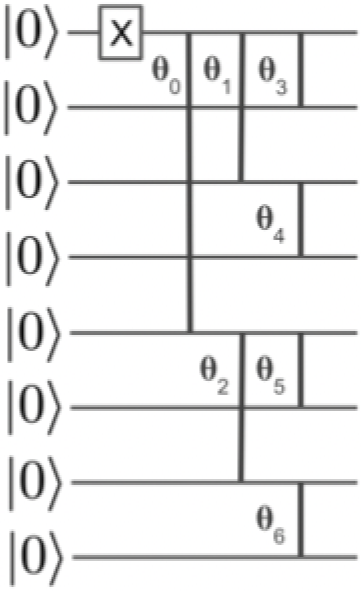

Fig. 3: Load Circuit for d-dimensional vectors (d =8). The X gate represents the Pauli X gate, and the vertical lines represent RBS gates with tunable parameters.

In Fig. 2, the top register of the data loader has n qubits and the bottom register has d qubits. And, as shown in Fig. 3, amplitude encoding is used in the Load circuit and RBS gates are parameterised two-qubit gate given by the following unitary matrix-

$$RBS(\theta) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & \sin\theta & 0 \\ 0 & -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Again in Fig. 1, there is a Quantum Orthogonal Layer named **V** which is a parameterised circuit made of RBS gates. In this Quantum Compound Transformer, Butterfly circuit (Fig. 4) is used which has depth of log(n) and ((n/2) * log(n)) number of gates for n qubits.

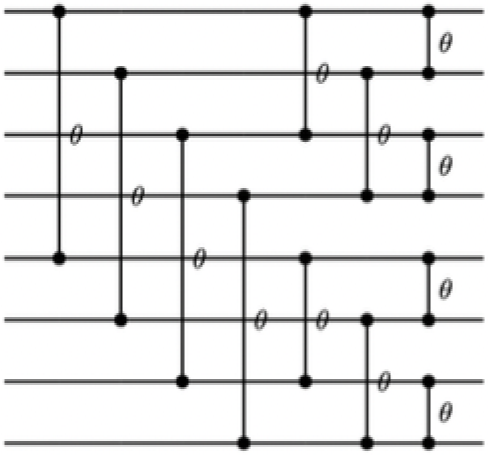

Fig. 4: Butterfly Circuit on 8 qubits for Quantum Orthogonal Layer

In the "Quantum Vision Transformers" paper, authors have also benchmarked this model on each test dataset of MedMNIST using AUC and ACC. The performance analysis is given below and it's worth mentioning that the paper is still under review -

Dataset	Metric	Vision Transformer	Compound Transformer
PathMINST	AUC	**0.957**	0.957
	ACC	**0.755**	0.735
BreastMINST	AUC	0.824	**0.859**
	ACC	0.833	**0.846**
DermaMINST	AUC	0.895	**0.901**
	ACC	0.727	**0.734**
OCTMINST	AUC	**0.879**	0.867
	ACC	**0.608**	0.545
OrganAMINST	AUC	0.968	**0.975**
	ACC	0.77	**0.789**

Reference: Quantum Vision Transformers