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

In developing a Quantum Vision Transformer (QViT), I draw inspiration from the recent advances outlined by Cherrat et al. [1]. This QViT architecture blends the effective aspects of classical transformers, such as attention mechanisms, with quantum amplitude encoding and orthogonal quantum transformations, designed explicitly for compatibility with near-term quantum devices.

The first operation in this architecture is preprocessing the classical input data. In this initial classical step, each image is segmented into smaller patches. These patches are individually converted into classical feature vectors. Unlike classical transformers, which directly process these vectors using standard linear algebra, the quantum approach first encodes these classical vectors into quantum states through amplitude encoding. This method efficiently maps classical vectors into high-dimensional Hilbert spaces.
In amplitude encoding, each classical feature vector, say $x_i \in R^d$, is normalized and encoded into the amplitudes of a quantum state. Formally, a classical vector x_i with dimension d is mapped to a quantum state

$$|x_i\rangle = (1/\|x_i\|) \sum_{j=1}^{d} (x_i)_j |j\rangle$$

where the computational basis states $|j\rangle$ form a basis in a d-dimensional Hilbert space. This mapping efficiently encodes high-dimensional classical data into quantum states using exponentially fewer qubits compared to classical bits, thus leveraging quantum computing's intrinsic high-dimensional feature representation capabilities without proportionally increasing computational resources.

After being encoded into quantum states, the core of the quantum transformer, the quantum attention layer, is applied. For classical vision transformers, attention is achieved through the calculation of interaction coefficients between patches directly through dot-product operations. Specifically, the attention mechanism first linearly projects each input patch into a query ($Q = XW_Q$), key ($K = XW_K$), and value ($V = XW_V$) vectors. The attention coefficient Aij representing how much the query patch i attends to key patch j is then computed using the scaled dot-product attention equation given by:

$$A_{ij} = q_i^T k_j / \text{sqrt}(d_k)$$

where d_k is the dimension of the key vectors. These attention coefficients are subsequently normalized through a softmax function and multiplied with the value vectors to obtain the output representations, which form the input to subsequent layers.

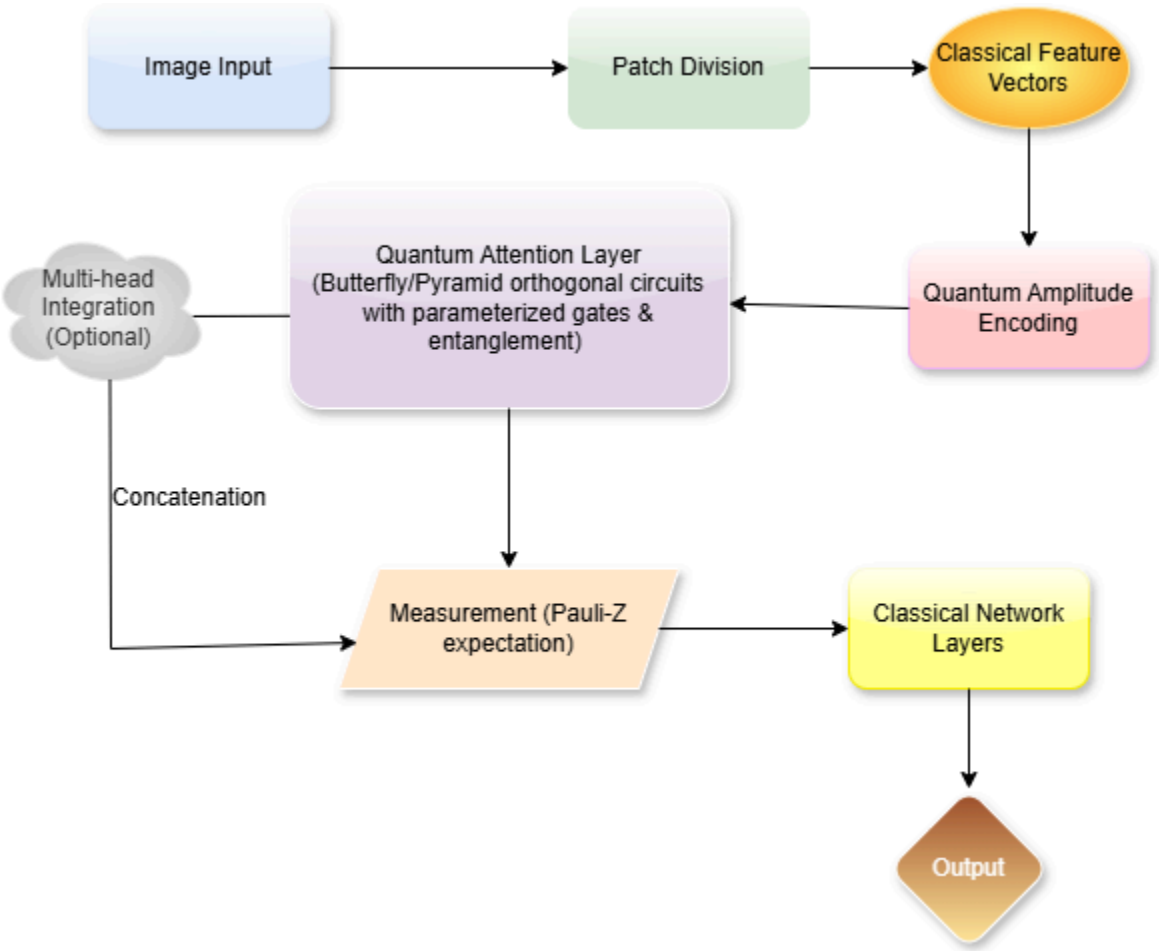
In contrast, the quantum attention layer substitutes these classical linear algebra operations with quantum orthogonal layers and controlled entangling gates. This quantum attention mechanism uses parameterized quantum circuits to implicitly and efficiently capture interactions between encoded patches. Orthogonality is crucial in this quantum implementation because quantum circuits must be unitary operations, that keep the norm of quantum states intact, prevent information loss, and ensure stable gradient updates during training, significantly mitigating the barren plateau problem common in variational quantum circuits.

In practice, I propose implementing the quantum attention layers with a structured method—quantum orthogonal layers built with parameterized entangling gates, specifically the Reconfigurable Beam Splitter (RBS) gates. RBS gates, or reconfigurable beam-splitter gates, effectively implement orthogonal transformations and entangle qubits with minimal quantum resources. The quantum attention layer can particularly use either a pyramid or butterfly circuit topology. The Butterfly circuit is a quantum orthogonal layer inspired by the classical butterfly architecture, providing logarithmic depth and all-to-all qubit connectivity, ideal for capturing global interactions with minimal quantum resources. Conversely, the Pyramid circuit leverages nearest-neighbor connectivity, offering efficient implementation on hardware with limited qubit connectivity, though at the cost of slightly increased circuit depth and reduced expressivity compared to the butterfly structure. Both parameterized rotation gates and entanglement gates within these quantum circuits are optimized to implicitly promote inter-patch attention interactions within quantum states, thereby replicating—and potentially enhancing—the classical attention mechanism.

Quantum state measurement comes after the use of quantum orthogonal attention layers. Each quantum state, which carries processed patch information, is measured through the expectation values of Pauli-Z operators. These expectation values capture meaningful inter-patch correlations embedded through the quantum circuit's learned parameters. The resulting expectation values form a quantum-enhanced feature vector, effectively encoding global relationships between image patches with fewer parameters and operations than classical dot-product-based attention.

For scalability and improved model capacity, multi-head quantum attention can also be used. Similar to classical multi-head attention, several different quantum circuits with different sets of parameters ("heads") are simultaneously applied to the quantum-encoded data. Each head uses a separately parameterized quantum orthogonal layer or quantum attention circuit. The resulting outputs of the different quantum attention layers (heads) are combined, usually concatenated, yielding richer feature representations that capture various views of inter-patch relationships.

The output of the quantum attention layer—single-head or multi-head—is then measured via expectation values and passed into a classical fully connected feed-forward neural network. These final classical layers then transform the quantum-improved features to yield the final output (classification probabilities, regression values, etc.).



Sketch of the Quantum ViT architecture

References

1) Cherrat, E. A., Kerenidis, I., Mathur, N., Landman, J., Strahm, M., & Li, Y. Y. (2024). *Quantum Vision Transformers*. Quantum. arXiv:2209.08167.