**Intro**

This work introduces a Hybrid 3D CNN-Transformer Model, designed to combine the local feature extraction capabilities of 3D Convolutional Neural Networks (CNNs) with the long-range dependency modeling of Transformer-based architectures.

The proposed architecture effectively extracts multi-scale spatial and spectral representations via a 3D CNN backbone before processing them through a Transformer Encoder, which refines feature dependencies and enhances global context modeling. The architectural design ensures computational efficiency and robustness in handling high-dimensional input data.

**Overview**

The model comprises two key components:

* A 3D CNN Backbone for hierarchical feature extraction and dimensionality reduction.
* A Transformer Encoder for long-range dependency modeling and final classification.

**Structure of the 3D CNN Module**

The 3D CNN backbone consists of three convolutional blocks, each composed of:

* A 3D convolutional layer (Conv3D) with a kernel size of , enabling local feature extraction across spatial and spectral dimensions.
* Batch normalization (BatchNorm3D) to normalize activations and improve stability during training.
* A non-linearity (Tanh activation function), which normalizes values in the range .

The initial processing of the input occurs through a stack of 3D convolutional layers, defined as:

where is the input tensor, is the Tanh activation function and batch normalization ensures stability.

* A 3D Max Pooling operation, which progressively reduces spatial dimensions while preserving spectral depth:

Each convolutional block extracts increasingly abstract feature representations, refining the learned feature space. Importantly, the pooling operations are carefully designed to prevent the depth dimension from being reduced excessively. Specifically:

* The first two pooling layers use a kernel of (1,2,2), reducing only the height and width but keeping the spectral depth unchanged.
* The final layer employs Adaptive Average Pooling, ensuring a fixed-size feature representation of , regardless of input image size.

The final convolutional feature map is then flattened and projected into an embedding space suitable for Transformer processing.

**Advantages of Using a 3D CNN**

The adoption of a 3D CNN module provides several key advantages over traditional 2D CNN architectures:

1. Preserving Multispectral and Volumetric Information: Unlike 2D CNNs, which process each spectral channel independently or in stacked formats, 3D CNNs treat spectral and spatial dimensions jointly, maintaining inter-band relationships.
2. Reducing Computational Complexity for the Transformer: The 3D CNN acts as a preprocessing mechanism, reducing the input dimensions passed to the Transformer. This helps mitigate the computational overhead of self-attention mechanisms, which scale quadratically with sequence length.
3. Robust Spatial Feature Extraction: CNNs provide strong inductive biases for grid-structured data, capturing local correlations in a way that self-attention mechanisms alone struggle to replicate.
4. Adaptive Representation Learning: The use of adaptive average pooling guarantees that the Transformer receives a consistent input size, preventing issues with variable input resolutions.

**Transformer Encoder for Global Feature Refinement**

After feature extraction via the 3D CNN, the resulting feature vectors are fed into a Transformer Encoder, which refines global dependencies and long-range interactions within the extracted representations.

The Transformer module consists of six stacked Transformer Encoder Layers, each performing Multi-Head Self-Attention (MHSA) and Feedforward Processing.

Since Transformers do not inherently understand spatial relationships, we incorporate positional encodings that encode position information using sine and cosine functions:

The MHSA module allows the model to capture global dependencies across the feature space:

where:

* are linear projections of the input embeddings.
* Softmax normalization ensures that attention weights sum to 1.

Each attention head operates independently, and their outputs are concatenated and projected into the original embedding space.

Each Transformer layer contains a position-wise feedforward network (FFN):

Additionally, Layer Normalization and residual connections stabilize training.

**Conclusion**

The Hybrid 3D CNN-Transformer model successfully bridges the gap between local feature extraction and global contextual learning. The integration of CNN-based hierarchical representation learning and self-attention-based feature refinement results in an effective, scalable approach for handling multispectral datasets.