Methodology

ImageBind: A Multimodal Embedding Framework

ImageBind is a state-of-the-art multimodal model designed to generate unified embeddings from diverse data types, including images, audio, and text. The model achieves **cross-modal alignment** by utilizing modality-specific encoders to map inputs into a shared embedding space. This alignment is driven by a **contrastive loss function** defined as:

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{N}\exp(\text{sim}(z_i, z_k)/\tau)} \tag{1}$$

where $\text{sim}(z_i, z_j)$ denotes the cosine similarity between embeddings z_i and z_j, and τ is a temperature scaling parameter. This loss minimizes the distance between embeddings of semantically similar data points while maximizing the distance between dissimilar ones. By aligning representations across modalities, ImageBind enables semantic understanding and facilitates retrieval and classification tasks.

Embedding Extraction

We utilize the pretrained **ImageBind model** to extract embeddings for both images and audio. For each aerial scene, the corresponding image and audio data are processed independently through ImageBind's modality-specific encoders, producing embeddings z_{image} and z_{audio}. These embeddings serve as foundational representations for the subsequent classification task, leveraging ImageBind's capability to capture semantic meaning across diverse modalities.

Architecture

This study presents a **Transformer-based model** for classifying aerial scenes using both image and audio data. The proposed model incorporates a **Dual Attention mechanism** to process intra-modal relationships and cross-modal dependencies, followed by a Transformer-based classifier for final predictions.

Dual Attention Mechanism

The **Dual Attention mechanism** combines **Self-Attention** for individual modalities with **Cross-Modal Attention** for integrating image and audio embeddings.

Self-Attention on Image and Audio Embeddings

Each modality's embeddings are processed independently through **Multi-Head Attention (MHA)** to capture intra-modal dependencies. Given an input X, the MHA operation is defined as:

$$\text{MHA}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2}$$

where each attention head computes:

$$\text{head}_i = \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_k}}\right)V_i \tag{3}$$

Here, Q, K, V are queries, keys, and values derived from X, and W^O is a learnable weight matrix.

Layer Normalization and **residual connections** are applied post-MHA to stabilize learning and retain original information.

Cross-Modal Attention

After independent processing, image embeddings (Q) are aligned with audio embeddings (K, V) using cross-modal attention:

$$\text{CrossAttention}(Q_{\text{image}}, K_{\text{audio}}, V_{\text{audio}}) \tag{4}$$

Similar to self-attention, this operation integrates complementary information from both modalities. Layer Normalization and residual connections ensure stable learning and preserve critical features.

The output of the **Dual Attention mechanism** is a set of unified embeddings representing both intra-modal and inter-modal interactions.

Transformer Classifier

The combined embeddings are further processed using a **Transformer-based classifier** to perform aerial scene classification.

Transformer Encoder

The unified embeddings are passed through a stack of Transformer encoder layers. Each layer consists of:

- **Multi-Head Attention (MHA)** to capture global dependencies.
- **Feedforward Neural Networks (FFN)** to introduce non-linear transformations.

The encoder output is defined as:

$$Z' = \text{FFN}(\text{MHA}(Z, Z, Z)) + Z \tag{5}$$

where Z is the input embeddings.

Global Average Pooling

To obtain a fixed-size vector, a **Global Average Pooling (GAP)** operation computes the mean over the sequence:

$$z_{\text{global}} = \frac{1}{N} \sum_{i=1}^{N} Z_i \tag{6}$$

Fully Connected Layer

The pooled vector z_{global} is passed through a fully connected layer:

$$\hat{y} = \text{softmax}(W z_{\text{global}} + b) \tag{7}$$

where W and b are learnable weights, and the number of output units corresponds to the number of classes.