Aerial Scene Classification using ImageBind and Transformer-based Attention Mechanisms

September 2024

1 Dataset

AuDio Visual Aerial sceNe reCognition datasEt (ADVANCE) :

The AuDio Visual Aerial sceNe reCognition datasEt (ADVANCE) is a comprehensive multimodal dataset specifically designed for aerial scene classification tasks. It combines both visual and audio data to enhance the recognition of diverse environmental scenes from an aerial perspective. The dataset comprises 5,075 paired samples, each consisting of an aerial image and its corresponding environmental audio recording. These pairs are categorized into 13 distinct classes capturing the unique auditory and visual characteristics of various locations such as airports, beaches, urban areas, forests, and more. Figure 1 shows the number of paired samples per class.

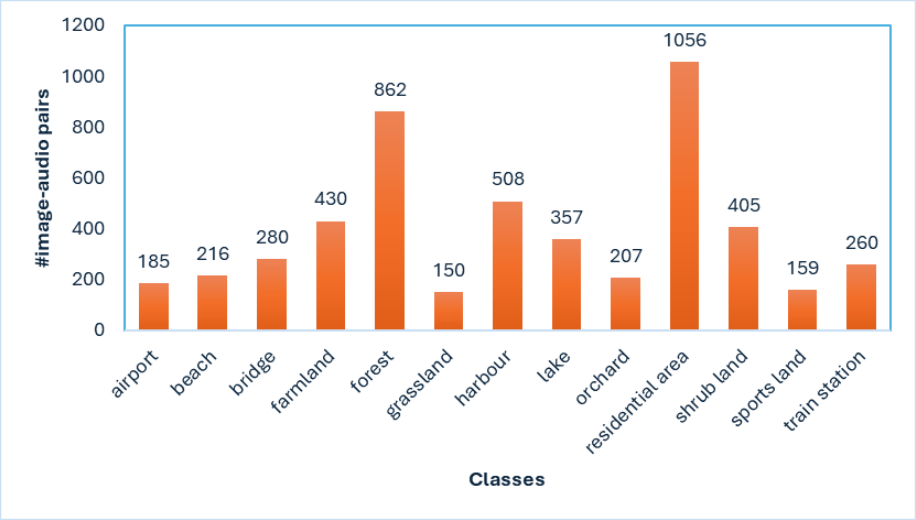

Figure 1: Distribution of Image-Audio pairs in ADVANCE dataset.

2 Methodology

2.1 ImageBind: A Multimodal Embedding Framework

ImageBind is a state-of-the-art multimodal model developed to generate unified embeddings from various data types, including images, audio, and text. The model achieves cross-modal alignment by leveraging modality-specific encoders, which map inputs into a shared embedding space using a contrastive loss function. The goal of the contrastive loss is to minimize the distance between embeddings of semantically similar data points across different modalities while maximizing the distance between embeddings of dissimilar data points. This alignment allows ImageBind to effectively capture semantic similarities across different modalities, making it a powerful tool for retrieval and classification tasks.

2.2 Embedding Extraction

In our research, we employ the pretrained ImageBind model to extract embeddings from images and their corresponding audio counterparts. For each aerial scene, the corresponding image and audio data are processed through the ImageBind model to obtain 1024-dimensional embeddings. These embeddings serve as the foundational representations for our classification task. By utilizing ImageBind, we benefit from its robust ability to handle diverse data types, leading to more accurate and semantically meaningful results.

2.3 Statistical Approach

2.3.1 Data Preparation

For each data point, the image embedding and corresponding audio embedding were concatenated to create a combined feature vector. This concatenation was done for all available embeddings, resulting in a dataset where each sample consisted of a concatenated feature vector of size 2048. The dataset was then split into training, validation, and testing sets using an 80-10-20 split.

2.3.2 Feature Selection

Given the high dimensionality of the concatenated embeddings, feature selection was performed to remove low-variance features, which might not contribute significantly to the classification task. Specifically:

- **Vision Embeddings:** A variance threshold of 0.00015 was applied to the image embeddings.
- **Audio Embeddings:** A more aggressive variance threshold of 0.65 was applied to the audio embeddings.

The thresholds were chosen empirically based on preliminary experiments. This process reduced the feature size from 2048 to 1087.

2.3.3 Feature Standardization

After feature selection, the remaining features were standardized to have zero mean and unit variance. This step was crucial to ensure that all features contributed equally to the model during training.

2.3.4 Dimensionality Reduction

To further reduce the feature space and address potential multicollinearity, Principal Component Analysis (PCA) was applied to the concatenated embeddings. The number of principal components was set to retain 50% of the variance in the data, effectively reducing the dimensionality by half.

2.3.5 Model Training

A Support Vector Machine (SVM) with a radial basis function (RBF) kernel was then trained on the reduced and standardized feature set. This model was chosen for its effectiveness in handling non-linear relationships between features. Figure 2 shows the detailed architecture of the statistical approach.

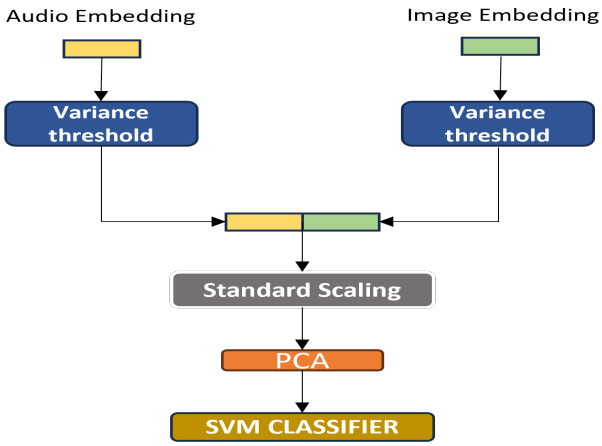

Figure 2: Architecture of statistical approach

2.4 Attention-Based Approach

This study introduces a Transformer-based model leveraging a Dual Attention mechanism for the classification of aerial scenes using both image and audio data. The following sections outline the design and implementation of our model, which combines self-attention on individual modalities with cross-modal attention to integrate information across modalities.

2.4.1 Dual Attention Mechanism

The core of our approach is the Dual Attention mechanism, which processes image and audio embeddings in parallel and integrates them through a cross-modal attention operation. This mechanism enhances the model's ability to capture both intra-modal and inter-modal dependencies.

Architecture:

- **Self-Attention on Image and Audio Embeddings:**
  - The image and audio embeddings are first processed independently using separate Multihead Attention layers. These layers consist of 8 heads, where each head computes attention scores between the elements of the respective embeddings, allowing the model to capture relationships within each modality.
  - The output of each Multihead Attention operation is followed by a Layer Normalization step and a residual connection, which helps in maintaining the original embedding information while also incorporating the attention-derived context.
- **Cross-Modal Attention:**
  - After processing the embeddings independently, the model applies a Cross-Modal Attention mechanism. The image embeddings are used as queries, while the audio embeddings serve as keys and values. This allows the model to integrate complementary information from both modalities.
  - A Layer Normalization step and a residual connection from the self-attention output of the image modality are again applied to ensure stable learning and preserve important information from the original image embeddings.

The output of the Dual Attention mechanism is a set of combined embeddings that encapsulate both intra-modal and inter-modal interactions.

2.4.2 Transformer Classifier

Following the Dual Attention mechanism, the model employs a Transformer-based architecture to further process and classify the combined embeddings.

Architecture:

- **Transformer Encoder:**
  - The combined embeddings are passed through a stack of 2 Transformer Encoder layers. Each encoder layer includes a Multihead Attention mechanism with 8 heads and a feedforward neural network.

  - The Multihead Attention within the encoder layers processes the combined embeddings to capture more complex dependencies across the entire input sequence.
  - The feedforward network, with a hidden dimension of 2048, applies dropout of 40% to prevent overfitting.

- **Global Average Pooling:** The output of the Transformer Encoder is reduced to a fixed-size vector through global average pooling. This operation computes the mean of the sequence elements, yielding a global context vector that summarizes the combined embeddings.
- **Fully Connected Layer:** The resulting global context vector is passed through a fully connected layer. The number of output units corresponds to the number of classes in the classification task.

Figure 3 shows the detailed architecture of the attention-based approach.

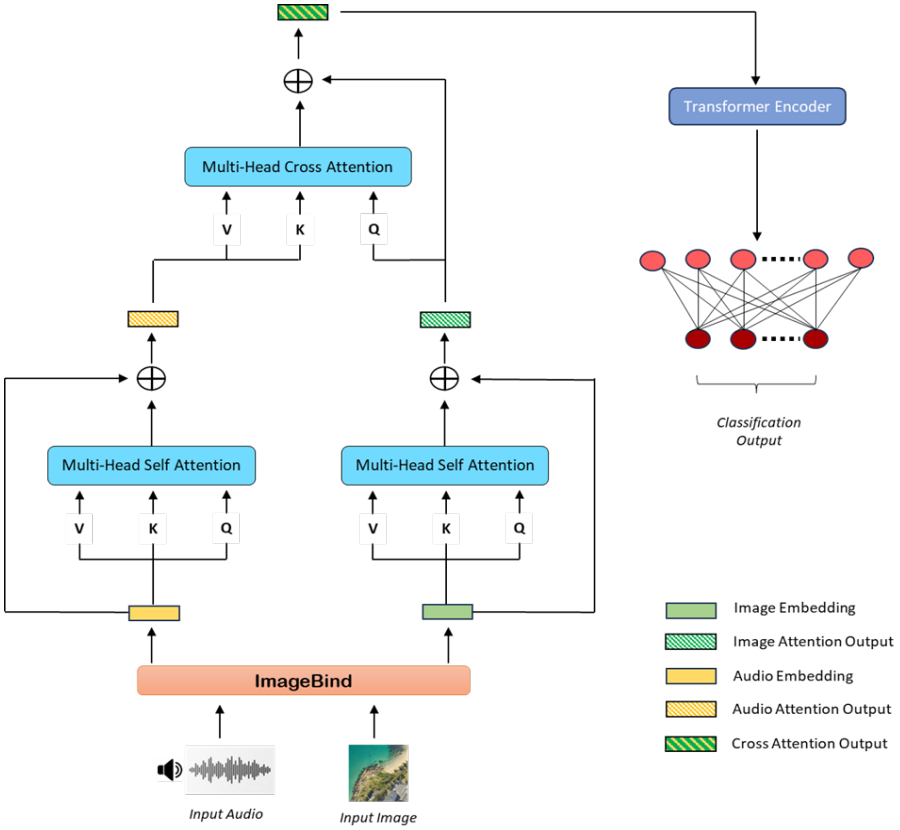

Figure 3: Architecture of attention based approach

The attention based approach is summarized as follows:

1. Dual Attention Mechanism

The `DualAttention` class applies self-attention to the image and audio embeddings separately and then performs cross-attention between the two.

Self-Attention on Image Embeddings

Given the image embeddings $\mathbf{X}_I$:

$$\mathbf{Q}_I = \mathbf{X}_I \mathbf{W}_Q^I, \quad \mathbf{K}_I = \mathbf{X}_I \mathbf{W}_K^I, \quad \mathbf{V}_I = \mathbf{X}_I \mathbf{W}_V^I$$

$$\text{Self-Attention on Image Embeddings: } \mathbf{A}_I = \text{softmax}\left(\frac{\mathbf{Q}_I \mathbf{K}_I^\top}{\sqrt{d_k}}\right)\mathbf{V}_I$$

The output after self-attention and normalization:

$$\mathbf{O}_I = \text{LayerNorm}(\mathbf{A}_I + \mathbf{X}_I)$$

Self-Attention on Audio Embeddings

Given the audio embeddings $\mathbf{X}_A$:

$$\mathbf{Q}_A = \mathbf{X}_A \mathbf{W}_Q^A, \quad \mathbf{K}_A = \mathbf{X}_A \mathbf{W}_K^A, \quad \mathbf{V}_A = \mathbf{X}_A \mathbf{W}_V^A$$

$$\text{Self-Attention on Audio Embeddings: } \mathbf{A}_A = \text{softmax}\left(\frac{\mathbf{Q}_A \mathbf{K}_A^\top}{\sqrt{d_k}}\right)\mathbf{V}_A$$

The output after self-attention and normalization:

$$\mathbf{O}_A = \text{LayerNorm}(\mathbf{A}_A + \mathbf{X}_A)$$

Cross-Attention Between Image and Audio Embeddings

The cross-attention is performed using the image embeddings as queries and the audio embeddings as keys and values:

$$\mathbf{Q}_{CA} = \mathbf{O}_I \mathbf{W}_Q^{CA}, \quad \mathbf{K}_{CA} = \mathbf{O}_A \mathbf{W}_K^{CA}, \quad \mathbf{V}_{CA} = \mathbf{O}_A \mathbf{W}_V^{CA}$$

$$\text{Cross-Attention: } \mathbf{O}_{CA} = \text{softmax}\left(\frac{\mathbf{Q}_{CA} \mathbf{K}_{CA}^\top}{\sqrt{d_k}}\right)\mathbf{V}_{CA}$$

The final output after cross-attention and normalization:

$$\mathbf{O}_{\text{final}} = \text{LayerNorm}(\mathbf{O}_{CA} + \mathbf{O}_I)$$

2. Transformer Classifier

The `TransformerClassifier` class uses the output from the `DualAttention` mechanism and passes it through a Transformer encoder and a classification layer.

Transformer Encoder

The output from the `DualAttention` mechanism, $\mathbf{O}_{\text{final}}$, is passed through a transformer encoder. For each encoder layer:

$$\mathbf{O}_{TE} = \text{TransformerEncoderLayer}(\mathbf{O}_{\text{final}})$$

The final output from the Transformer encoder:

$$\mathbf{O}_{\text{encoder}} = \text{TransformerEncoder}(\mathbf{O}_{\text{final}})$$

Classification Layer

The encoded output is averaged over the sequence dimension and passed through a fully connected layer for classification:

$$\mathbf{O}_{\text{mean}} = \text{mean}(\mathbf{O}_{\text{encoder}}, \text{dim} = 1)$$

$$\text{Classification Output: } \mathbf{y} = \text{Softmax}(\mathbf{O}_{\text{mean}}\mathbf{W}_{fc} + \mathbf{b}_{fc})$$

Where:
- $\mathbf{W}_Q^I$, $\mathbf{W}_K^I$, $\mathbf{W}_V^I$: Query, key, and value matrices for image embeddings.
- $\mathbf{W}_Q^A$, $\mathbf{W}_K^A$, $\mathbf{W}_V^A$: Query, key, and value matrices for audio embeddings.
- $\mathbf{W}_Q^{CA}$, $\mathbf{W}_K^{CA}$, $\mathbf{W}_V^{CA}$: Query, key, and value matrices for cross-attention.
- $\mathbf{W}_{fc}$, $\mathbf{b}_{fc}$: Weights and biases for the fully connected layer.
- d_k: Dimensionality of the keys.

2.4.3 Model Training

The entire pipeline, including data preparation, model training, and evaluation, is implemented in Python using the PyTorch framework. The model is trained using the cross-entropy loss function. The equation for Cross-Entropy Loss for a classification problem with N classes is:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

Where:

- y_i is the true label, which is 1 if the class i is the correct class and 0 otherwise.

- $\hat{y}_i$ is the predicted probability for class i, which is the output of the softmax function.
- The sum is taken over all classes N.

AdamW is used as the optimizer with a learning rate of 1×10^{-6}, and a weight decay of 1×10^{-5} to prevent overfitting. Training is conducted over 15 epochs, with the average loss per epoch calculated to monitor progress.

3 Experiments

3.1 Aerial Scene Recognition

3.1.1 Statistical Approach

The statistical approach involves dimensionality reduction techniques, primarily Principal Component Analysis (PCA) and a combination of Variance Thresholding with PCA, applied to the embeddings of images and audio in the aerial scene recognition task. PCA alone provides a moderate level of precision, recall, and F1 score. While PCA is effective in reducing the dimensionality of the embeddings, it seems that the variance captured by PCA alone may not fully represent the critical features needed for high performance in this task. Adding a variance threshold before applying PCA significantly improves performance across all metrics. This suggests that filtering out low-variance features before applying PCA helps retain the most informative components, leading to a more accurate and balanced model. The improvement of approximately 3-4 percentage points across all metrics indicates the effectiveness of combining these two techniques. The results are summarized in Table 1

Approach	**Precision**	**Recall**	**F1 score**
PCA only	84.39	83.54	82.57
Variance Threshold and PCA	87.75	87.21	86.55

Table 1: Performance metrics for the statistical approach.

3.1.2 Attention-Based Approach

This approach explores the use of self-attention mechanisms, particularly focusing on residual connections from the outputs of image and audio self-attention modules. The results summarized in Table 2 highlight the impact of different configurations of residual connections:

Residual Connection from Image Self-Attention Output Only: This configuration delivers the best performance across all metrics, indicating that retaining the residual connection from the image self-attention output is highly

Approach	Precision	Recall	F1 score
Residual connection from image self-attention output only	95.39	95.30	95.02
Residual connection from audio self-attention output only	65.48	66.32	63.21
Residual connection from both image and audio self-attention outputs	92.75	92.57	91.76
No residual connection	82.82	84.10	82.41

Table 2: Performance metrics for different configurations of residual connections in the attention-based approach.

beneficial. The model appears to gain significant information from image embeddings, leading to very high precision, recall, and F1 score.

Residual Connection from Audio Self-Attention Output Only: In contrast, relying solely on the residual connection from audio self-attention output results in a substantial drop in performance. The audio embeddings alone may not be as discriminative as the image embeddings in this task, which is reflected in the much lower scores. This could indicate that while audio features are useful, they may not be sufficient when used in isolation.

Residual Connection from Both Image and Audio Self-Attention Outputs: Combining residual connections from both image and audio self-attention outputs results in a slight decrease in performance compared to using the image self-attention output alone. While this approach leverages information from both modalities, it seems that the combined residuals introduce some redundancy or noise that marginally reduces the model's effectiveness.

No Residual Connection: Removing residual connections altogether leads to a noticeable drop in performance. This demonstrates the importance of residual connections in maintaining and enhancing the flow of information through the network, ensuring that critical features are preserved and effectively utilized in the final prediction.

3.1.3 Comparision

The proposed approach outperforms all other methods by a significant margin as shown in Table 3. The high precision, recall, and F1 score demonstrate that the attention-based method with a residual connection from image self-attention output is particularly well-suited for this task. This indicates that the model effectively captures the discriminative features of aerial scenes, leveraging both the spatial and contextual information present in the images, and to some extent, the audio data.

The experimental results reveal that attention mechanisms, particularly with residual connections from image self-attention outputs, are highly effective for aerial scene recognition. The proposed method not only surpasses traditional statistical approaches like PCA and Variance Thresholding but also outperforms existing state-of-the-art models. This highlights the importance of properly

Approach	**Precision**	**Recall**	**F1 score**
ADVANCE	75.25	74.79	74.58
SoundingEarth	89.59	89.52	89.50
TFAVCNet	89.90	89.85	89.83
Ours	95.39	95.30	95.02

Table 3: Comparison of performance metrics for different methods.

leveraging image features in conjunction with advanced attention mechanisms to achieve superior performance in complex recognition tasks.