# Country Classification from Street-Level Images Using Vision Transformers

Teba Suh , Reman Mahameed

February 28, 2026

**Abstract**

This project presents an image classification system designed to distinguish between three countries—Japan, France, and Mexico—using street-level imagery. A balanced subset of 1,200 images (400 per country) was extracted from the GeoGuessr dataset. The model is based on a pre-trained Vision Transformer (ViT) architecture fine-tuned for a 3-class classification task.

## 1 Introduction

Image-based country classification is a challenging computer vision task that requires recognizing subtle geographical cues such as architecture styles or road markings. unlike object classification (e.g., cat vs. dog), this task requires contextual scene understanding rather than focusing on a single object. In this project, we focus on classifying images from: Japan, France, Mexico. These countries were chosen because they are visually distinguishable but still share some global scene similarities (roads, buildings, vehicles). The goal is to determine whether a transformer-based architecture can effectively learn discriminative geographical features from a relatively small dataset from the .geoguessr dataset. To accomplish this, we fine-tune a pre-trained Vision Transformer model, rather than training from scratch, leveraging transfer learning to achieve high performance with limited data.

## 2 Model Architecture

- Overview:

The model used in this project is the Vision Transformer (ViT), based on the transformer architecture originally introduced in Attention Is All You Need. Unlike convolutional neural networks (CNNs), Vision Transformers rely entirely on self-attention mechanisms to process images. The specific model used was `google/vit-base-patch16-224`, pre-trained on ImageNet.

- How the Vision Transformer Works?

First, the Vision Transformer divides the image into fixed-size patches of $16 \times 16$ pixels.(instead of applying convolutional filters), each patch is flattened into a vector and projected into a 768-dimensional embedding space using a linear layer. Thus, the image becomes a sequence of 196 patch embeddings. This process converts the image into a structure similar to a sentence in NLP, where each patch acts like a token.

Then, a special learnable classification token (CLS token) is added at the beginning of the sequence. This token is responsible for aggregating global information from all patches. After adding the CLS token, the input becomes: 197 tokens $\times$ 768 dimensions ,the final representation of this CLS token will later be used for classification.

The next step, is Positional Encoding: Transformers do not inherently understand spatial position, therefore, positional embeddings are added to each patch embedding to preserve information about the patch's location in the image. without positional encoding, the model would treat the image as an unordered set of patches, losing spatial structure.

The last step is Transformer Encoder Layers: The ViT-Base model contains: 12 Transformer encoder layers, 12 attention heads, hidden dimension size of 768 and Feed-forward dimension of 3072 Each encoder layer consists of: Multi-head self-attention, Feed-forward neural network, Residual connections, Layer normalization. Self-attention allows each patch to interact with all other patches. Instead of focusing locally like CNNs, the model computes relationships globally.

The attention mechanism calculates: $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$ Where: Q = Query matrix, K = Key matrix, V = Value matrix.

This allows the model to determine which regions of the image are important relative to others. After attention, each token passes through a fully connected feed-forward network with non-linear activation. This increases representational capacity and enables complex feature transformations. The original ViT model was trained for 1000 ImageNet classes. In this project, the final classification layer was replaced with a new linear layer mapping from 768 dimensions to 3 output classes.

# 3 Training Strategy

- 3.1 Data Augmentation

To improve generalization and prevent overfitting, the following augmentations were applied to the training set: Random horizontal flip (p=0.5), Random rotation (15), Color jitter (brightness=0.2, contrast=0.2, saturation=0.2), Random resized crop (224, scale=(0.8, 1.0)). the parameters we chose for these augmentations are industry-standard default values, especially for Vision Transformer fine-tuning and ImageNet-style training. These transformations improves generalization by exposing the model to slightly modified versions of the same images. This encourages the model to learn invariant and robust features rather than memorizing specific pixel patterns.

- 3.2 Optimization

Training configuration:

Learning rate: 2e-5 (This is the standard value for ViT, lower value might underfit, higher value the training was unstable).

Epochs: 15 (Since the dataset is relatively small so it converges quickly).

Batch size: 16 (larger batch required more GPU memory because the ViT is heavy).

Gradient accumulation steps: 2 (Memory constraints prevent real batch so now the effective batch = $16 \times 2 = 32$.).

Weight decay: 0.01 (the model is large and we suffered from a larger overfitting so we used in to minimize the gap between training and validation loss).

Warmup steps: 100 (it gradually increases learning rate at beginning,it's important for transformers, because early layers are sensitive and large early updates can damage pretrained features.) Cosine learning rate scheduler(gradually reduces the learning rate over time, improving convergence stability.).

The optimizer used was AdamW, which combines adaptive learning rates with weight decay regularization.

note: without warmup steps and weight decay the maximum accuracy was 82% (with playing with the other parameters values).

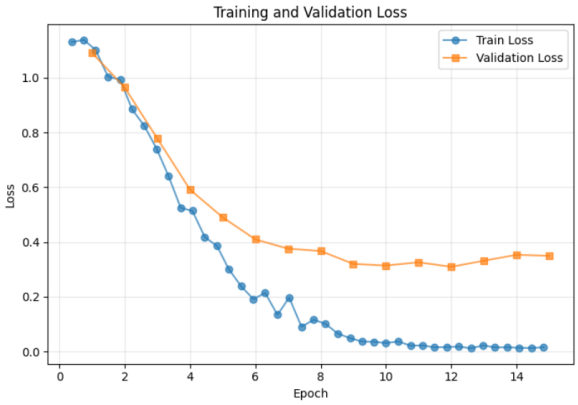


Figure 1: Training and Validation loss

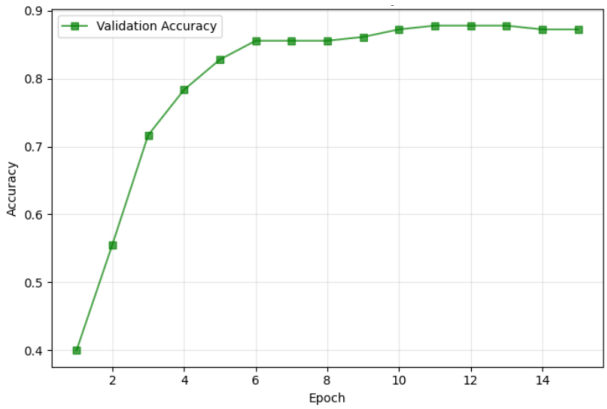


Figure 2: Validation accuracy

# 4 Results

We got these training plots:

- 4.1 Training Convergence and Optimization Behavior

The Vision Transformer model was trained for 15 epochs using AdamW optimization and a cosine learning rate scheduler. The evolution of training and validation loss demonstrates stable and efficient convergence.

At the beginning of training, both training and validation losses were approximately 1.1, reflecting the random initialization of the classification head. During the first five epochs, a sharp decline in both losses was observed, indicating rapid adaptation of the pre-trained backbone to the new task. This behavior highlights the effectiveness of transfer learning: instead of learning low-level features from scratch, the model leveraged pre-trained visual representations and adjusted them to the country classification task.

Between epochs 6 and 9, the training loss continued decreasing substantially, while the validation loss declined at a slower rate. This pattern suggests that the model was refining discriminative features specific to geographical patterns while maintaining generalization capability.

After approximately epoch 10, training loss approached near-zero values (0.015), whereas validation loss stabilized around 0.31–0.35. A slight increase in validation loss was observed in later epochs, indicating mild overfitting. However, the magnitude of this increase remained

limited and did not significantly affect validation accuracy. Overall, the optimization process was stable, and no signs of divergence or unstable gradient behavior were observed.

The final evaluation on the held-out test set produced the following Accuracy: 88.9%.

The close agreement between validation accuracy ( 88% ) and test accuracy (88.9%) demonstrates strong generalization and indicates that the model did not overfit to the validation data.

- 4.2 Interpretation of Model Behavior

The strong performance achieved in this task can be attributed to the architectural properties of the Vision Transformer. Unlike convolutional neural networks, which rely primarily on local receptive fields, the transformer architecture models global relationships through self-attention mechanisms. Country classification from street-level imagery requires contextual reasoning rather than object recognition. The model must integrate information such as: architectural structure, road layout and markings, vegetation patterns The self-attention mechanism enables each image patch to attend to all other patches, allowing the model to capture long-range spatial dependencies. This capability is particularly advantageous in scene-level classification tasks, where relationships between distant regions (e.g., road signs and surrounding buildings) provide critical information.

The training curves further support this interpretation. The early rapid accuracy gain indicates that the pre-trained attention layers already encoded general visual patterns, which were then refined for geographical discrimination.

- 4.3 Overfitting Analysis

Although training loss continued decreasing toward zero in later epochs, validation loss remained stable and validation accuracy did not decline significantly. This behavior indicates mild overfitting but not severe memorization. Given the relatively small dataset size (1,200 images) and the high capacity of the Vision Transformer ( 86 million parameters), such behavior is expected. The use of data augmentation, weight decay, and a cosine learning rate scheduler likely helped control excessive overfitting.

Overall, the gap between training and validation performance remains moderate and acceptable.

# 5 Summary of Findings

The experimental results demonstrate that fine-tuning a pre-trained Vision Transformer is highly effective for country-level scene classification. The model achieved strong performance (88.9% test accuracy) with stable convergence and controlled overfitting.

The findings support the hypothesis that transformer-based architectures are particularly well-suited for contextual image classification tasks, where global spatial relationships are more informative than isolated local features.