# Image Classification with Vision Transformer on Food101 Dataset

Jun Li
Department of Statistics, University of Michigan
lju@umich.com

**Abstract**

This report presents an image classification pipeline leveraging the Vision Transformer (ViT) architecture on the Food101 dataset. The project demonstrates data preprocessing, fine-tuning of a pretrained ViT model, and evaluation of model performance. The pipeline achieves competitive accuracy, showcasing the potential of transformer-based architectures for image classification tasks.

## Introduction

The Food101 dataset provides a diverse collection of food images across 101 classes. With the growing demand for accurate image classification systems, transformer-based models like Vision Transformers (ViT) have emerged as state-of-the-art solutions. This project aims to fine-tune a pretrained ViT model on the Food101 dataset to evaluate its performance and applicability in food image classification.

## Method

### Dataset Description

The Food101 dataset contains 101,000 images, with 750 training images and 250 testing images per class. Due to computational constraints, we extract 7575 training samples and 2525 validation samples from the original dataset for training and validation.

### Data Preprocessing

Data augmentation techniques such as random resizing, horizontal flipping, and normalization were applied to enhance model generalization. The pixel values were normalized using the mean and standard deviation of ImageNet.

### Model Architecture

The Vision Transformer (ViT) model, pretrained on ImageNet21k, was fine-tuned for 101 food classes. The architecture divides images into patches, embeds them, and processes the embeddings through transformer layers.

## Training Setup

The model was trained for 5 epochs using the AdamW optimizer with a learning rate of $1 \times 10^{-4}$ and a cosine annealing learning rate scheduler. Mixed precision training was employed to improve efficiency.

| **Hyperparameter** | **Value** |
|---|---|
| Device | `cpu` |
| Learning Rate | 0.0002 |
| Training Batch Size | 32 |
| Evaluation Batch Size | 32 |
| Optomizer | AdamW with beta=(0.9,0.999) |
| Optomizer | AdamW with epsilon = $1 \times 10^{-8}$ |
| Epochs | 5 |

Table 1: Hyperparameters used in training

# Results

## Performance Metrics

The fine-tuned ViT model achieved a validation accuracy of 69.04% on the sampled Food101 dataset. Table summarizes the dataset sizes and accuracy.

Table 2: Training and Validation Results

| **Epoch** | **Training Loss** | **Validation Loss** | **Accuracy** |
|---|---|---|---|
| 1 | 1.3253 | 1.1283 | 0.7172 |
| 2 | 1.0036 | 1.1552 | 0.7073 |
| 3 | 0.8771 | 1.1416 | 0.7180 |
| 4 | 0.7424 | 1.0653 | 0.7287 |
| 5 | 0.6745 | 1.1179 | 0.7244 |

The results demonstrate the model's effectiveness in distinguishing between diverse food classes, despite the reduced dataset size.

# Conclusion

This project demonstrates the application of the Vision Transformer for image classification on the Food101 dataset. The model achieved a validation accuracy of 69.04% with only 10% of the original dataset, showcasing its efficiency and adaptability. Future work could involve exploring larger datasets and implementing domain-specific augmentations to further improve accuracy.

# Acknowledgment

The author thanks Hugging Face and the Food101 dataset contributors for their resources and support.