

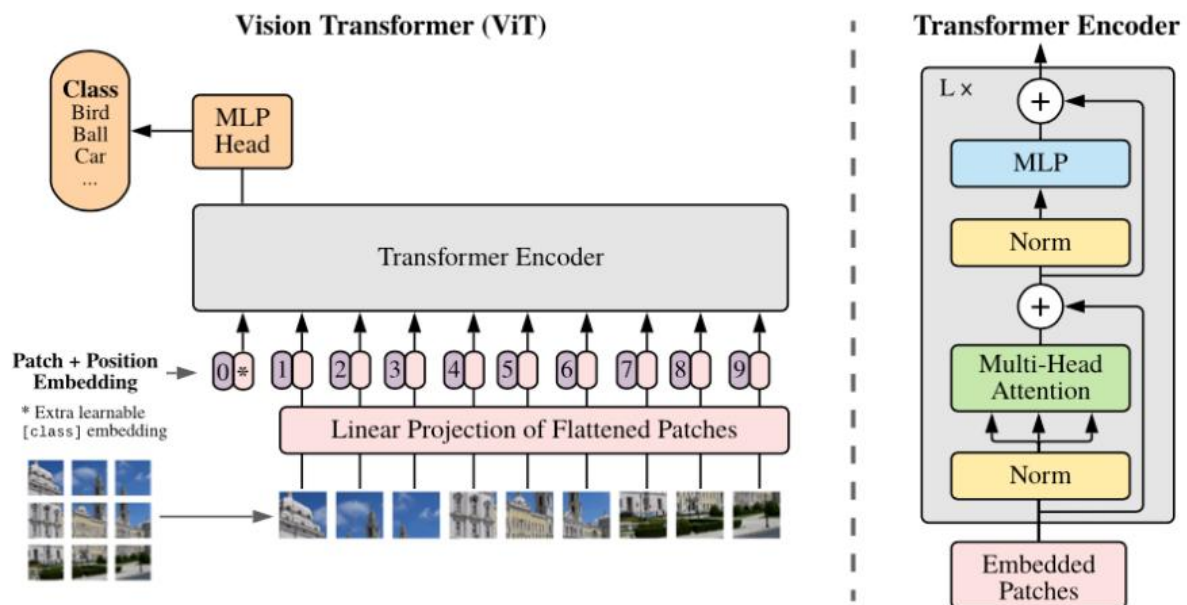
Team ID: CHP_5

| | |
|----------------------|------------|
| Haneen Akram Ahmed | 2022170812 |
| Menna Ali Thabet | 2022170844 |
| Mohamed Ashraf Fathy | 2022170919 |
| Reem Ahmed Ismail | 2022170815 |
| Zeina Shawkat | 2022170818 |
| Islam Hesham | 2022170802 |

Vision Transformer (ViT) Implementation for Fruit Quality Assessment

Model Architecture Overview

The implemented model is a Vision Transformer (ViT) designed specifically for a fine-grained fruit quality assessment task. The ViT architecture represents a departure from traditional convolutional neural networks (CNNs) by applying a transformer-based approach to image classification.



Key Components:

1. Patch Creation Layer:

- Divides input images ($224 \times 224 \times 3$) into patches of size 8×8
- Resulting in 784 patches (28×28) per image

2. Embedding Layers:

- Patch Embedding: Projects each patch into a 128-dimensional embedding space
- Position Embedding: Adds positional information to maintain spatial relationships

3. Transformer Encoder:

- 8 transformer layers with multi-head self-attention
- Each with 8 attention heads and layer normalization
- GELU activation function in feed-forward networks
- Dropout rate of 0.01 for regularization

4. Classification Head:

- Global average pooling to aggregate feature representations
- MLP with two hidden layers (2048 → 1024 → 7 classes)
- Softmax activation for final class probabilities

Model Parameters:

- Total parameters: 7,772,551 (29.65 MB)
- All parameters are trainable

Training Configuration

- **Image Size:** 224×224×3
- **Batch Size:** 32
- **Epochs:** 50 (with early stopping at epoch 43)
- **Optimizer:** AdamW with weight decay (1e-5)
- **Initial Learning Rate:** 1e-4
- **Loss Function:** Sparse Categorical Cross-Entropy
- **Random Seeds:** 42 for reproducibility

Data Augmentation:

- Rotation (up to 20°)
- Zoom (±10%)
- Width and height shifts (±10%)
- Shear transformation (10%)
- Horizontal flipping

Training Strategies:

- Class weights to handle significant class imbalance
- Learning rate reduction on plateau
- Early stopping with patience of 10 epochs
- Model checkpointing to save best weights

Dataset Analysis

The dataset focuses on banana and tomato quality classification with 7 classes:

- banana_overripe: 1,395 samples
- banana_ripe: 1,440 samples
- banana_rotten: 1,987 samples
- banana_unripe: 1,370 samples

- tomato_fully_ripened: 50 samples
- tomato_green: 334 samples
- tomato_half_ripened: 81 samples

Severe Class Imbalance:

- Imbalance ratio: 39.74 (largest/smallest class)
- Tomato classes significantly underrepresented
- Class weights applied: from 0.48 for banana_rotten to 19.02 for tomato_fully_ripened

Training Results and Analysis

Performance Metrics:

- **Final Training Accuracy:** 96.25%
- **Final Validation Accuracy:** 95.39%
- **Best Model:** Saved at epoch 41

Training Progression:

1. **Initial Phase** (Epochs 1-4):
 - Rapid improvement from 20.93% to 81.03% training accuracy
 - Validation accuracy reached 88.08% by epoch 4
2. **Mid Training** (Epochs 5-22):
 - Slower but steady improvements
 - Notable performance drop at epoch 21 (validation accuracy: 67.75%)
 - Learning rate reduced to 2e-5 at epoch 22
3. **Fine-tuning** (Epochs 23-43):
 - Two more learning rate reductions (to 4e-6 at epoch 28, 1e-6 at epoch 38)
 - Validation accuracy plateaued around 94-95%
 - Early stopping triggered after 10 epochs without improvement

Learning Dynamics:

- The learning rate scheduler effectively managed training progression
- Model showed good resilience to overfitting with validation loss generally tracking training loss
- The temporary performance drop at epoch 21 suggests the model encountered a challenging optimization landscape

Comparison with Previous Attempts

1. Google ViT (Pre-trained on ImageNet):

- Despite leveraging transfer learning, this approach yielded inferior accuracy
- The specialized architecture in the custom ViT proved more effective for the fruit quality task
- Pre-trained weights may have been less relevant for the specific fine-grained distinctions required

2. Data Augmentation to Balance Classes:

- Previous attempt to augment underrepresented classes to 2,200 samples each
- This approach did not yield satisfactory accuracy
- Current implementation with class weights appears more effective than synthetic oversampling

Strengths of the Current Model

1. Custom ViT Architecture:

- Specifically designed for the fruit quality assessment task
- Appropriate patch size (8×8) captures relevant texture details

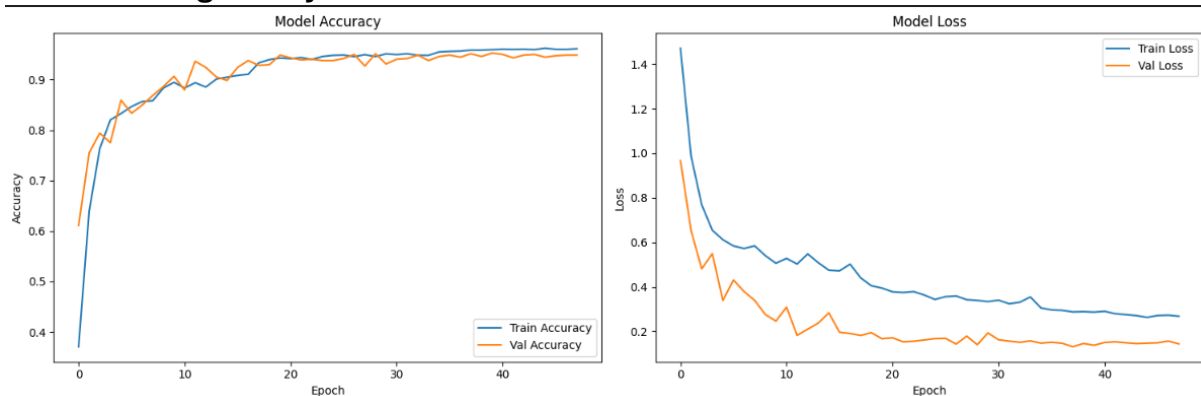
2. Effective Training Strategy:

- Class weights better addressed imbalance than synthetic oversampling
- Learning rate scheduling prevented convergence to poor local minima
- Early stopping and checkpointing ensured optimal model selection

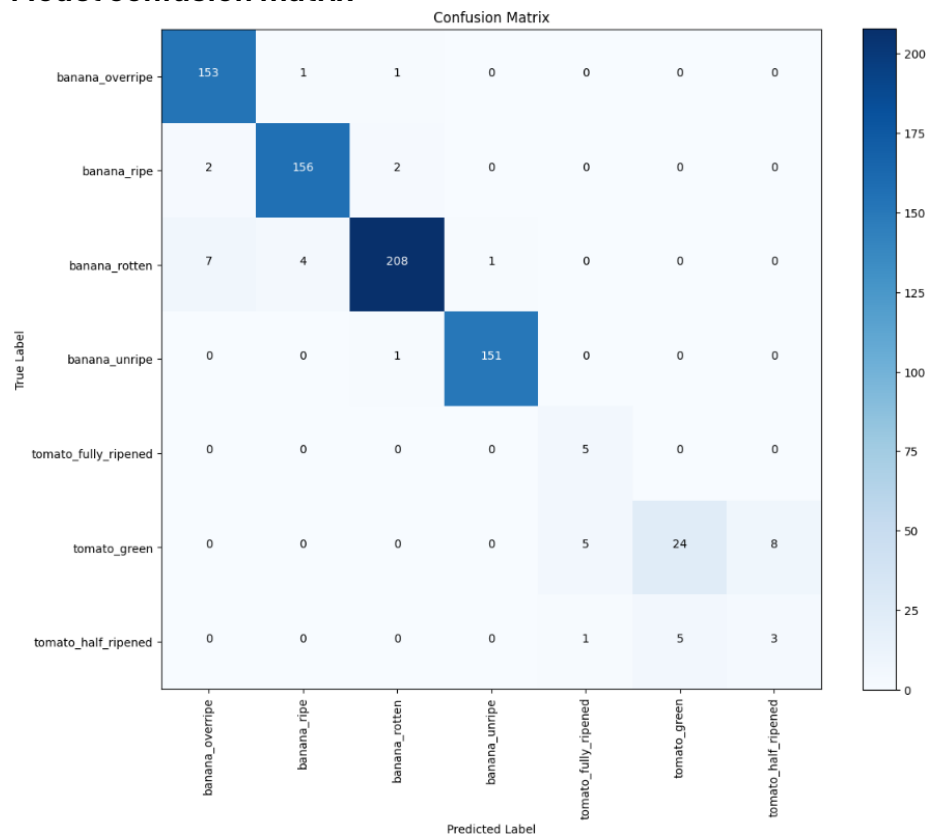
3. High Performance:

- ~95% validation accuracy shows strong generalization capability
- Consistent performance across both banana and tomato categories despite imbalance

Model training history



Model confusion matrix



Model weights

| Layer (type) | Output Shape | Param # |
|--|---------------------|-----------|
| input_layer_19 (InputLayer) | (None, 224, 224, 3) | 0 |
| patch_creation_3 (PatchCreation) | (None, None, 768) | 0 |
| patch_embedding_3 (PatchEmbedding) | (None, None, 128) | 96,432 |
| position_embedding_3 (PositionEmbedding) | (None, 196, 128) | 25,088 |
| transformer_encoder_block_17 (TransformerEncoderBlock) | (None, 196, 128) | 659,712 |
| transformer_encoder_block_18 (TransformerEncoderBlock) | (None, 196, 128) | 659,712 |
| transformer_encoder_block_19 (TransformerEncoderBlock) | (None, 196, 128) | 659,712 |
| transformer_encoder_block_20 (TransformerEncoderBlock) | (None, 196, 128) | 659,712 |
| transformer_encoder_block_21 (TransformerEncoderBlock) | (None, 196, 128) | 659,712 |
| transformer_encoder_block_22 (TransformerEncoderBlock) | (None, 196, 128) | 659,712 |
| transformer_encoder_block_23 (TransformerEncoderBlock) | (None, 196, 128) | 659,712 |
| transformer_encoder_block_24 (TransformerEncoderBlock) | (None, 196, 128) | 659,712 |
| layer_normalization_52 (LayerNormalization) | (None, 196, 128) | 256 |
| global_average_pooling1d_2 (GlobalAveragePooling1D) | (None, 128) | 0 |
| dense_60 (Dense) | (None, 2048) | 264,192 |
| dense_61 (Dense) | (None, 1024) | 2,098,176 |
| dense_62 (Dense) | (None, 7) | 7,175 |

Conclusion

The custom Vision Transformer implementation demonstrates excellent performance on the fruit quality assessment task, achieving 95.39% validation accuracy despite significant class imbalance. The model successfully outperformed previous attempts using pre-trained Google ViT and data augmentation approaches. The combination of appropriate architecture design, effective handling of class imbalance through weighting, and careful training strategies contributed to the model's success.