AI-vs-Real Image Classification: LeViT-192 Model

Abstract:

The task of distinguishing between AI-generated images and real images has gained increasing attention in recent years due to deep fake. In this report, I have explored the use of the LeViT-192 vision transformer model for classifying images into AI-generated or real categories. I have clearly outlined the methods utilized for data preprocessing, feature engineering, model training, and evaluation. Additionally, I have presented model explainability techniques such as Grad-CAM to gain insights into the model's decision-making process.

1. Introduction:

1.1 Background:

With the rise of AI-generated content, it has become very important to develop something to be capable of distinguishing between real and AI-generated images. This classification task is challenging due to availability of high quality modern ai providing images very close to real photos.

1.2 Objective:

The primary objective of this project is to develop an image classification model that can accurately differentiate between AI-generated and real images. We leverage the LeViT-192 model, a vision transformer-based architecture, and explore various techniques to enhance the model's performance, like data augmentation and Grad-CAM for model explainability.

2. Methodology:

2.1 Data Collection and Preprocessing

The dataset used in this project consists of images labeled as either AI-generated or real. The images were preprocessed to ensure consistency in size and format:

- **Image Resizing**: All images were resized to 224x224 pixels.
- **Normalization**: Images were normalized using a mean of `[0.5, 0.5, 0.5]` and standard deviation of `[0.5, 0.5, 0.5]`.

2.2 Data Augmentation

To improve model generalization and prevent overfitting, we employed the following data augmentation techniques:

- Random rotation
- Horizontal flipping
- Random color jitter
- Random cropping

These augmentations provided diverse variations of the training images, allowing the model to learn robust features.

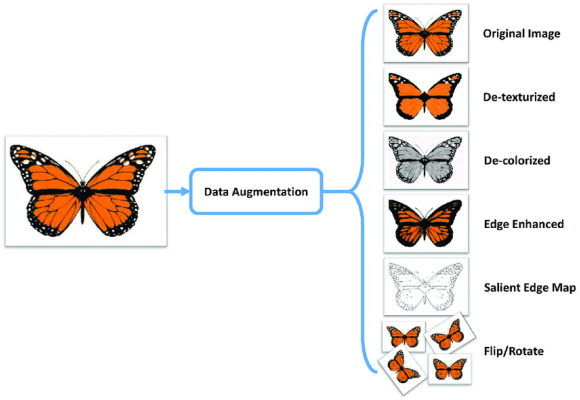
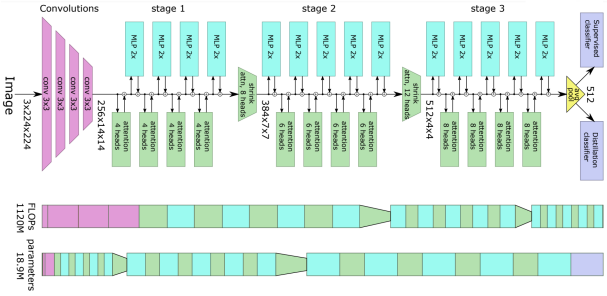


Figure 1: Data Augmentation

2.3 Model Architecture

The core of this project is the **LeViT-192** model, a vision transformer designed for efficient image classification tasks. Unlike traditional CNNs, the LeViT model processes images using self-attention mechanisms, capturing long-range dependencies in the data. The model was trained from scratch on the dataset to tailor it specifically to the task.

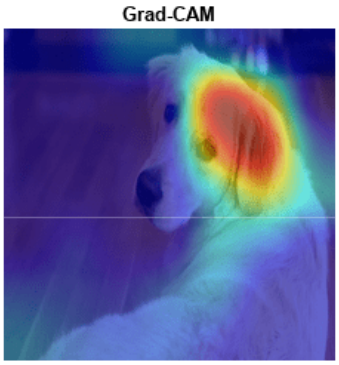


2.4 Training Process

The model was trained using the **Adam optimizer** with an initial learning rate of **1e-4**. Adam was chosen due to its efficient handling of sparse gradients and its ability to adapt the learning rate for each parameter. The loss function used was **binary cross-entropy**, as this is a binary classification task. The model's performance was monitored by tracking both the **loss** and **accuracy** during training.

2.5 Model Explainability (Grad-CAM)

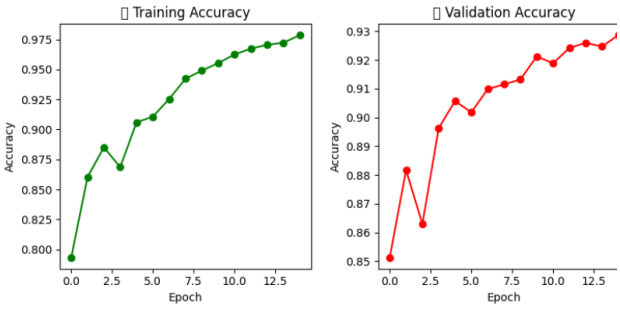
To gain insights into the model's decision-making process, we applied **Grad-CAM** (Gradient-weighted Class Activation Mapping). Grad-CAM allows us to visualize the regions of the image that the model focuses on when making its predictions, providing transparency and interpretability.

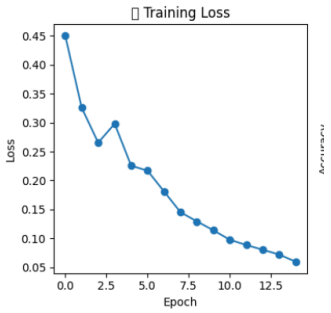


3. Results:

3.1 Training Performance

The model demonstrated a significant reduction in training loss over time, and the accuracy gradually improved with each epoch. The validation accuracy also followed a similar trend, confirming the model's ability to generalize to unseen data.





3.2 Test Performance

The final evaluation on the test set revealed the model achieved an impressive accuracy of **85.26%**. This result highlights the model's robustness in distinguishing AI-generated images from real ones.

4. Discussion:

4.1 Model Explainability

The use of Grad-CAM has significantly enhanced the explainability of the model's decisions. By providing a clear visualization of the areas the model attends to, we gained insights into the reasoning behind its predictions. This transparency is essential for understanding and improving model behavior, especially in high-stakes applications such as content authentication.

4.2 Potential Improvements

Incorporating Noise Perturbation Regularization (NPR) and Frequency Noise could have enhanced the model's robustness and generalization. NPR would help prevent overfitting by adding noise during training, while Frequency Noise could improve the model's ability to handle high-frequency variations, making it more resilient to subtle image artifacts in both real and AI-generated images.

5. Conclusion:

In this report, we presented an AI-vs-Real image classification task using the LeViT-192 model. The model demonstrated impressive performance in distinguishing between AI-generated and real images, with high classification accuracy on the test set. The application of Grad-CAM provided valuable insights into the model's decision-making process, enhancing its explainability. This work lays the foundation for future improvements in AI content detection systems and offers a transparent approach to understanding model predictions.

6. References:

- LeViT-192
- **Grad-CAM**
- **Data Augmentation Techniques**: