AI and OG image Classification

This report outlines the development and evaluation of a binary image classification system that distinguishes between AI-generated and real images. The model leverages a pre-trained LeViT-128S transformer as a feature extractor, employs dimensionality reduction layers, and integrates explainability methods such as Grad-CAM for insight into its decision-making process.

1. Introduction

In this project, we focus on distinguishing AI-generated images from real ones. The key innovation lies in combining the power of transformer-based architectures with classical convolutional layers to achieve both efficiency and interpretability.

2. Methodology:

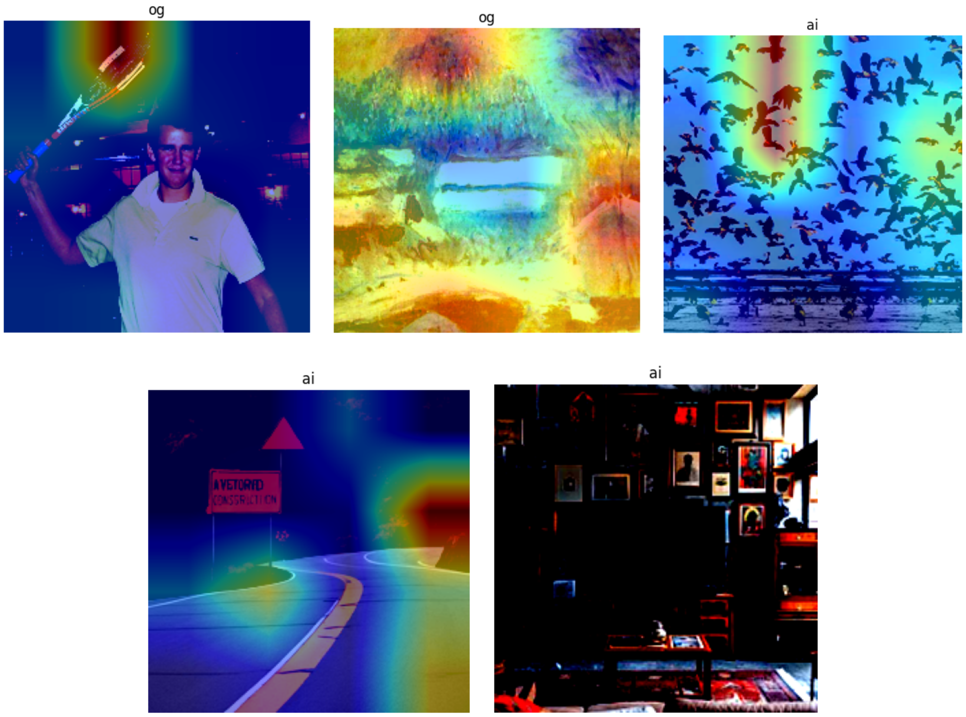
- **Transfer Learning with LeViT-128S**
  The LeViT-128S model from the TIMM library is used as the backbone. Its transformer architecture is pre-trained on a large-scale dataset of 4 million images specifically for AI-generated versus real image classification[1]. To avoid overfitting on our relatively small dataset and to accelerate training, the backbone is frozen. This allows the model to leverage rich, pre-learned representations while focusing the training on subsequent layers.

- **Model Architecture Enhancements**
  An intermediate Conv2D layer, followed by a ReLU activation, is added to reduce the channel dimensionality from 384 to 64. This step is crucial to lower computational complexity and help the model focus on the most discriminative features. An AdaptiveAvgPool2d layer is incorporated to reduce spatial dimensions. This ensures that the network can handle inputs of varying sizes without the need for a fixed-size input. A linear layer is used to map the pooled features to the output. Notably, the model outputs raw logits, which are then passed to the binary cross-entropy loss function with logits (BCEWithLogitsLoss). This approach circumvents the need for an explicit sigmoid activation during inference. The Adam optimizer is used to train the final layers of the model efficiently.

3. Explainability with Grad-CAM:

- **Rationale for Explainability**
  Understanding where a model “looks” to make its prediction is critical for building trust and ensuring the reliability of the system, especially in applications with high stakes such as content verification.
- **Implementation of Grad-CAM**
  Grad-CAM generates visual explanations by highlighting the regions in an image that significantly influence the model’s predictions[2]. This is achieved by computing the gradients of the target class with respect to the feature maps of the convolutional layers.

- **Case Study**



Grad-cam heat plot of images from my model

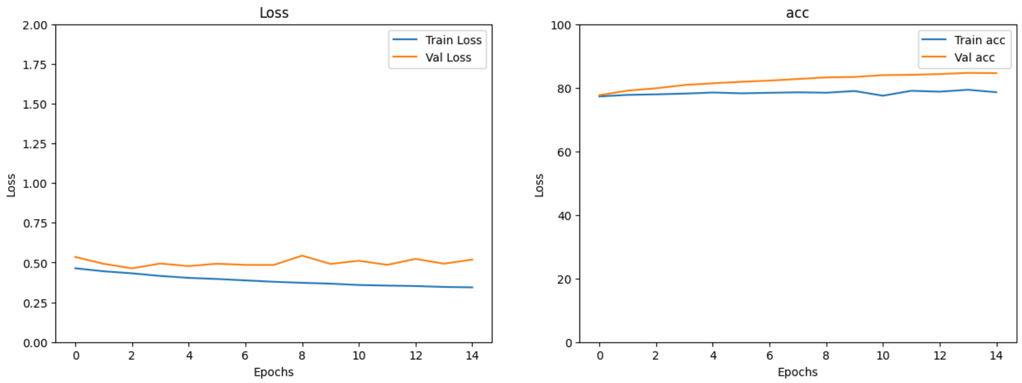
Here for an image of a bird which is AI generated we can see that the model focused most on **specific bird clusters** to determine the class AI. The bottom part is relatively ignored by the model. In the first image of the tennis player we can see that the model is focused on the racket hand. This suggests the **model is attending heavily to the racket** for predicting the class og. In the second image of tree and land, the heat map shows that the model focuses on the central horizontal structure This shows that the model is using spatial and color patterns for classification.

.

This image of the road shows that the model focuses heavily on the road curvature. Here the second image shows that the model reveals weak and diffuse activation, indicating uncertainty in the model's decision-making. Such outputs are **valuable** in identifying scenes where the model may misinterpret or over-generalize indoor environments as being AI-related.

4. Performance analysis:

- **Validation Accuracy and Loss:**
  During training, significant improvements in validation accuracy were observed, alongside a marked decrease in loss. These trends indicate effective learning and generalization capabilities of the model.
- **Training Time:**
  The model required approximately 275 minutes to train over 15 epochs. The use of a frozen backbone contributed to reduced training time by minimizing the number of trainable parameters.



Loss and accuracy vs epochs plot

5. Conclusion:

The binary image classification system presented in this report demonstrates a robust combination of modern transformer architectures and classical convolutional techniques. The approach effectively differentiates between AI-generated and real images while maintaining interpretability through Grad-CAM. The successful integration of transfer learning and explainability tools not only boosts performance but also builds confidence in the system's predictions.

This project highlights the benefits of leveraging pre-trained models for rapid development and fine-tuning on specialized tasks, providing a strong foundation for further research and practical applications in the field of image classification.

References

[1]: https://huggingface.co/docs/transformers/model_doc/levit
[2]:https://medium.com/@codetrade/grad-cam-in-pytorch-a-powerful-tool-for-visualize-explanations-from-deep-networks-bdc7caf0b282