COSE474-2024F: Final Project Report
“Distinguishing AI-Generated and Real Images Using CLIP: A Fine-Tuning Approach”

2019320006 HeeWoong Ahn

1. Introduction

1.1. Motivation

With the rapid advancements in AI image generating models, synthetic images are becoming increasingly indistinguishable from real-world images. While these advancements provide new opportunities in content creation, they also introduce challenges in identifying AI-generated images, particularly in contexts where authenticity is critical. Human observers often struggle to differentiate AI-generated images from real ones, highlighting the need for automated and robust methods to address this growing challenge.

1.2. Problem Definition

As AI-generated images become increasingly realistic, even well-performing classification models face difficulties in accurately distinguishing between synthetic and real images. This issue arises from the high-quality rendering of modern AI techniques, which make detection a non-trivial task. Moreover, the accessibility of these tools amplifies the challenge, as they can be used maliciously to spread misinformation or counterfeit media. Effective methods are required to reliably classify and interpret such images in diverse real-world scenarios.

1.3. Concise description of Contribution

In this work, we aim to evaluate the capability of the CLIP model in detecting AI-generated images. Initially, we assess its baseline performance on distinguishing synthetic images from real ones using its pre-trained text and image encoders. Subsequently, we fine-tune both the text and image encoders with a targeted dataset of AI-generated and real images to enhance its classification accuracy. This research contributes to understanding how effectively CLIP can address the challenge of AI-generated image detection and explores the impact of fine-tuning on CLIP in improving its performance on this critical task.

2. Related Works

3. methods

3.1. Significance and Novelty

Distinguishing AI-generated images from real images presents a significant challenge, particularly as synthetic images become increasingly indistinguishable from authentic visuals. One of the primary hurdles encountered in this study was the inability of the pre-trained CLIP model to reliably differentiate between fake and real images out of the box. This necessitated improving the model’s ability to handle this binary classification task.

Initially, we considered prompt tuning as a potential solution, given its computational efficiency and the constraints of a limited GPU environment provided by Google Colab. However, prompt tuning is better suited for tasks involving class-specific discrimination, such as assigning images to specific predefined categories (e.g., ”a picture of class”). In contrast, our task required a binary classification across all possible classes, aiming to determine whether an image was AI-generated or real, regardless of its content or category.

Given this unique challenge, fine-tuning was selected as the preferred approach. Fine-tuning allowed us to directly adapt both the text and image encoders of CLIP to the CIFAKE dataset, allowing the model to enhance its distinguishing ability from general to the specific task.

3.2. figure

3.3. reproducibility-algorithm

4. Experiments

4.1. Datasets

The *CIFAKE* dataset, designed as a benchmark for distinguishing between AI-generated and real images, consists of 20,000 images (32x32 pixels) evenly split into 10,000 synthetic (FAKE) and 10,000 real (REAL) samples. The real images are derived from the CIFAR-10 dataset, representing real-world objects across ten classes, such as airplanes, cats, and cars (Krizhevsky & Hinton, 2009). In contrast, the synthetic images are generated using AI(Stable Diffusion Version 1.4) (Bird & Lotfi, 2023). The dataset is further divided into 16,000 training and 4,000 test samples, providing a standardized format for evaluating binary classification models. CIFAKE serves as a valuable resource for tasks involving GAN evaluation, and the detection of AI-generated images.

4.2. Computer Resource Experimental Design

4.3. Quantitative Results

4.4. Qualitative Results

4.5. Analysis

4.6. Discussion

5. Future Direction

References

Bird, J. and Lotfi, A. Cifake: Image classification and explainable identification of ai-generated synthetic images. *arXiv preprint arXiv:2303.14126*, 2023.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. 2009.