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

While fine-tuning models like CNNs can achieve high accuracy in distinguishing fake and real images, concerns about their generalization remain. CNN models typically rely on filters learned to extract features specific to the classes present in the training dataset, rather than leveraging broader notions of similarity between classes. Unlike CLIP, which benefits from its training on a vast and diverse dataset and aligns text and image features in a shared embedding space, CNNs are limited to classification based on predefined classes. The learning mechanism in CNNs focuses on extracting features tightly coupled with the training dataset's specific labels, which may hinder performance when encountering unseen datasets or classes. This limitation raises questions about their ability to generalize effectively beyond the trained domain.

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(CIFAKE, excluding class 'CAT') of AI-generated and real images to enhance its classification accuracy. Lastly, we test the novel dataset with the fine-tuned CLIP model to investigate the accuracy on non-trained datasets(CIFAKE, only including class 'CAT'). 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 both base and novel datasets.

2. Related Works

In prior research, CIFAKE dataset was used to fine-tune a CNN model to distinguish between fake and real images. The CNN model was trained to classify images into two binary classes, with fake and real images labeled as 0 and 1, respectively. Binary cross-entropy loss was utilized during fine-tuning. After training, the model achieved an accuracy exceeding 90%. Through Grad-CAM analysis, it was revealed that fake images often exhibit distinctive fake characteristics not in the objects themselves but in their backgrounds (Bird & Lotfi, 2023).

In this research, we aim to fine-tune the CLIP model to differentiate between fake and real images. Unlike CNN models, which require additional training on a new dataset to produce meaningful metrics for novel datasets, CLIP leverages image-text similarity and supports zero-shot classification. This allows it to generalize effectively to unseen datasets. By fine-tuning CLIP to understand the similarity between images and the text prompts of synthetic and real images, we hypothesize that it can not only accurately predict well for trained datasets, but also predict well for datasets it has never encountered before.

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 differentiate between fake and real images. 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 computational power 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. Main Figure of the Experiment

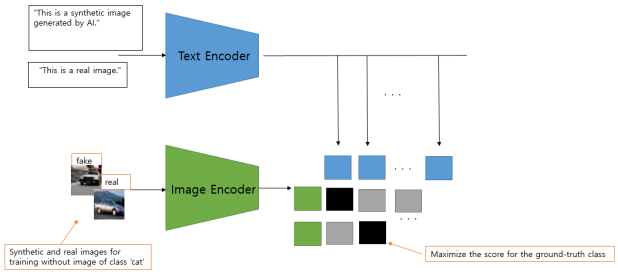

Figure 1. CLIP MODEL

3.3. Reproducibility-Algorithm

Dataset Preparation: CIFAKE Dataset

- Test datasets
  - **test-real-without-cat**: The first 100 real test images excluding any containing cats.
  - **test-fake-without-cat**: The first 100 fake test images excluding any containing cats.
  - **test-real-with-cat**: The first 100 real test images consisting only of cat images.
  - **test-fake-with-cat**: The first 100 fake test images consisting only of cat images.
- Train datasets
  - **train-real-without-cat**: The first 450 real images excluding cat images from the training set.
  - **train-fake-without-cat**: The first 450 fake images excluding cat images from the training set.

Model Configuration and Baseline Experiment

- Pre-trained CLIP model (ViT-B/32 variant)
- Prompts for classification:
  - "This is a synthetic image generated by AI."
  - "This is a real image."

Baseline classification was performed on the **test-real-without-cat** and **test-fake-without-cat** datasets to evaluate the model's initial performance.

Fine-Tuning CLIP Model

Both the image encoder and text encoder of CLIP were fine-tuned using contrastive learning.

- Training settings:
  - **Batch Size**: 32
  - **Epochs**: 5
  - **Loss Function**: Contrastive loss with temperature set to 0.07.
  - **Optimizer**: AdamW
  - **Learning Rate**: 1×10^{-5}

Post-Fine-Tuning Experiments

test-real-without-cat and **test-fake-without-cat** to assess the improvement in distinguishing real and fake images without cat images.

test-real-with-cat and **test-fake-with-cat** to evaluate the model's performance on datasets consisting exclusively of cat images.

4. Experiments

4.1. Datasets

The *CIFAKE* dataset, designed as a benchmark for distinguishing between AI-generated and real images, consists of 120,000 images (32x32 pixels) evenly split into 60,000 synthetic (FAKE) and 60,000 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 100,000 training and 20,000 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

Computer Resource

- **Environment**: Colab (Free)
- **OS**: Ubuntu 22.04.3 LTS
- **CPU**: Intel(R) Xeon(R) CPU @ 2.20GHz

Experimental Design

Baseline Evaluation Baseline experiment was designed to assess the inherent capability of the CLIP model to distinguish between fake and real images without any fine-tuning.

Fine-Tuning and Evaluation on Known Data The CLIP model was fine-tuned using the training datasets (**train-real-without-cat** and **train-fake-without-cat**), which excluded cat images.

After fine-tuning, the model was re-evaluated on the same test datasets (**test-real-without-cat** and **test-fake-without-cat**) to determine how well the fine-tuned model could distinguish fake from real images in a dataset similar to the training data.

Evaluation on Novel Data (Generalization Test) To evaluate the generalization ability of the fine-tuned CLIP model, additional experiments were conducted using test datasets containing only cat images (**test-real-with-cat** and **test-fake-with-cat**). This experiment assessed how effectively the CLIP model, fine-tuned on a specific dataset (excluding cat images), could classify fake and real images in a novel dataset (cat images only).

4.3. Experimental Results

Table 1. Performance Metrics for Fake and Real Image Detection on Test Dataset (Without Cat) Using Fine-Tuned CLIP Model

Class	Precision	Recall	F1-Score	Support
FAKE	0.97	0.83	0.89	100
REAL	0.85	0.97	0.91	100
Accuracy			0.90	200

There exists no well-established state-of-the-art (SOTA) benchmark in the field of synthetic image detection yet. Therefore, in this study, we aim to compare our results with a related study, which effectively improved accuracy by fine-tuning a CNN model on the same dataset (Bird & Lotfi, 2023).

The experimental results on our test dataset demonstrate an accuracy of 90% in distinguishing synthetic images from real images using the base datasets trained on the CLIP model. This accuracy is comparable to the 92.98% reported in a related study, which utilized a fine-tuned CNN model on the same dataset (Bird & Lotfi, 2023). This result suggests that the CLIP-based approach performs similarly to the CNN-based method in this context.

4.4. Analysis

Table 2. Performance Metrics for Fake and Real Image Detection on Test Dataset (Without Cat) Using Base CLIP Model

Class	Precision	Recall	F1-Score	Support
FAKE	0.50	1.00	0.67	100
REAL	0.00	0.00	0.00	100
Accuracy			0.50	200

Table 3. Performance Metrics for Fake and Real Image Detection on Test Dataset (With Cat) Using Fine-Tuned CLIP Model

Class	Precision	Recall	F1-Score	Support
FAKE	1.00	0.40	0.57	100
REAL	0.62	1.00	0.77	100
Accuracy			0.70	200

Precision is important when minimizing False Positive.

$$Precision = \frac{TruePositives(TP)}{TruePositives(TP) + FalsePositives(FP)}$$

Recall is important when minimizing False Negative.

$$Recall = \frac{TruePositives(TP)}{TruePositives(TP) + FalseNegatives(FN)}$$

F1-Score is important when both Precision and Recall are important.

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Accuracy is used to measure how well the model predicts data overall.

$$Accuracy = \frac{TruePositives(TP) + TrueNegatives(TN)}{TotalInstances}$$

According to Table 2, every image (fake or real) was classified as a fake image. 0 precision and 0 recall values of class REAL indicates that the base CLIP model was not able to distinguish between fake and real images.

In contrast, according to Table 1, fine-tuned CLIP model showed impressive enhancement in distinguishing fake and real images. F1-Score went up to around 0.90, which shows that the model performs well on minimizing both False Positives and False Negatives. Also, accuracy went up to 0.90 from 0.50.

Furthermore, the fine-tuned CLIP model also showed some notable enhancement on distinguishing fake and real images from novel datasets. According to Table 3, F1-Score of class REAL went up to 0.77 from 0.00 while F1-Score of class FAKE changed only in a moderate manner. Furthermore, accuracy went up to 0.70.

4.5. Discussion

The base CLIP model exhibited no significant ability to distinguish between synthetic and real images. In contrast, the fine-tuned CLIP model achieved an impressive F1-score of 0.90 and an accuracy of 0.90, demonstrating its effectiveness in learning the associations between prompts and images from the training dataset. This indicates that the fine-tuning process was successful in enabling the model to differentiate between fake and real images within the scope of the learned data.

Moreover, the model's performance on a novel dataset consisting of cat images, which were not included in the training data, further supports its generalization capability. While not perfect, the model was able to differentiate between synthetic and real cat images to a certain degree. These results highlight the potential of CLIP as a generalizable model capable of zero-shot classification by leveraging learned associations between images and textual prompts.

5. Future Direction

The ability to distinguish between synthetic and real images is crucial in addressing issues such as crimes involving fake images and broader authentication challenges. While the fine-tuned CLIP model demonstrated nearly 90% accuracy on the training dataset, this is not sufficient in high-stakes scenarios where even a single misclassification (e.g., a fake image being misclassified as real) can have severe consequences. Therefore, improving the model's accuracy further is an essential direction for future research.

Although the fine-tuned model showed some capability in classifying images from a novel dataset, its performance remains insufficient for practical applications. Particularly, the precision for the Real class on the unseen dataset is relatively low (0.62), which is concerning as misclassifying fake images as real can lead to significant risks. Thus, it is crucial to address this issue and improve the model's performance, especially on unseen data.

Lastly, evaluating the performance of a fine-tuned CNN model on the same novel dataset under identical conditions would provide valuable insights. Comparing its results with those of the fine-tuned CLIP model could reveal the extent to which the CLIP model's generalization capability outperforms CNN models. This comparison would underscore the importance of enhancing pre-trained models to maximize their utility in handling diverse datasets without extensive fine-tuning.

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.