GMR-29 Identification of AI-Generated Images

Abstract

With the quick rise of Artificial Intelligence (AI), generative AI models have greatly increased the volume and velocity of data creation. Among that data, AI-generated images have become a highly discussed topic, especially when discussing the potential dangers of these AI models. Due to these dangers, being able to distinguish AI-generated art from human-made art is becoming a necessity. Additionally, as these AI-models improve, it is becoming increasingly difficult for humans to determine whether art is AI-generated or human-made. This paper proposes the further exploration of the effectiveness of a current state of the art AI-image identification model.

Introduction

Over the past few years, AI has arisen in many different fields. What was once only used for the purposes of dynamic calculations has now become capable of producing its own media. From speech synthesis to video and music and, for the purposes of our project, art and photography. With AI image generation becoming more refined, people are becoming more weary as to how its application can affect those that make legitimate art. More and more people are looking for ways to identify these AI-generated images.

One such way people are looking into AI identification is by using AI to detect these images. This ‘fighting fire with fire’ approach allows a program to detect specific attributes involving the image that people may not be able to notice. This would allow not only an accurate way of determining AI-generated imagery, it also provides a fast method of classification. There are already instances of AI in use to identify AI-generated images, yet there still remains avenues to improve their accuracy and efficiency with other AI-generation models.

Research Question(s)

The benchmark CNN’s performance was tested on images that were generated with a single model. Does the model’s performance change when the generation technique is different?

If so, what needs to be changed about the model?

Materials and Methods

Our approach involved improving on an already proposed model. In that model it employed a Convolutional Neural Network (CNN) that classifies images as “real” (human-made) or “fake” (AI-generated). It also utilized an explainable AI method called Gradient Class Activation Mapping (Grad-CAM) for the interpretation of classifications. This model was able to accurately classify images into their correct categories, but was only trained and tested using a single AI-generative model.

Before beginning our testing, we needed to evaluate the base performance of the originally proposed model. To do this, we implemented the CNN with the highest F-1 score, then trained and evaluated its performance on the CIFAKE dataset. We would then train, test, and evaluate it using datasets produced using various generative AI models. Additionally, the CNN’s architecture would be compared to current state-of-the-art CNN models to help identify possible improvements that could be made.

Before analyzing the CNN’s performance on image datasets, we first searched for the optimal hyperparameters, which were not mentioned in the original model. We assume regularization was not used, a 2x2 kernel, and do not utilize mini-batching during training. We leave these parameters to be optimized in non-benchmark algorithms.

We investigated different optimization techniques with various learning rates, monitoring the network’s validation loss in order to optimize the number of epochs used. Training is stopped early if validation loss does not improve over 5 epochs and returns the weights that produced the minimum validation loss.

Each optimizer was implemented with every argument set to default values and with learning rates varying from 0.01 to 0.00001. Once the optimal optimizer is identified, we will then evaluate its performance using various models of AI-generated imagery.

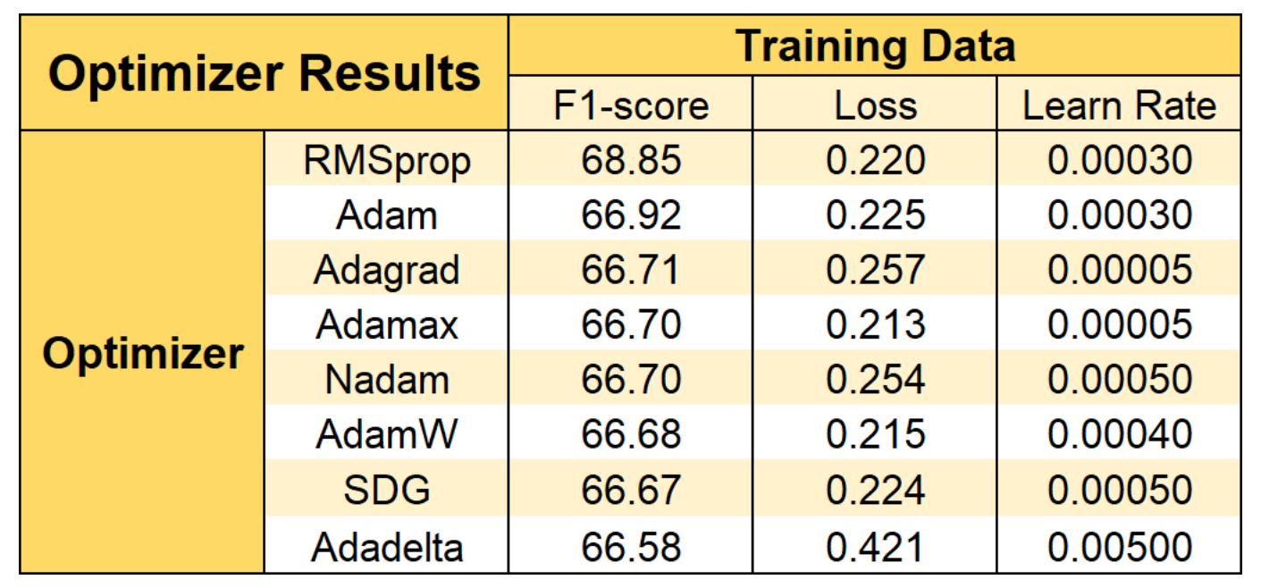
The datasets we chose for training and testing our model includes:

- CIFAKE – Images of cars, trucks, ships, planes, and various animals
    - 60,000 real images from CIFAR-10
    - 60,000 fake images generated with CompVisSD (Diffusion)
- DALLE – Images of artwork
    - 3,781 real images made by artists
    - 2,575 fake images made with DALLE (Gated CNN)
- PEOPLE – Images of people
    - 15,292 real images of people from a facial recognition dataset
    - 77,182 fake images of people generated using Midjourney (Diffusion)

We will train and test our model using various combinations of these datasets and evaluate their performance.

Results

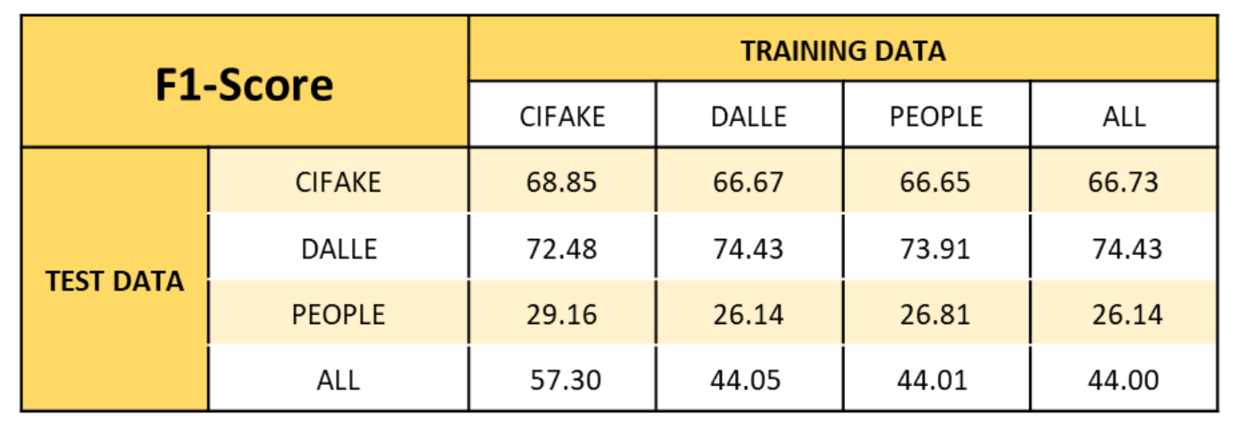
We tested the CNN model using different optimization techniques. During our analysis, we found that RMSprop produced the best F1-score of 68.85 with a loss of 0.22 using a learning rate of 0.0003. In addition, Adamax produced the lowest loss of 0.213 with an optimal F1-score of 66.70



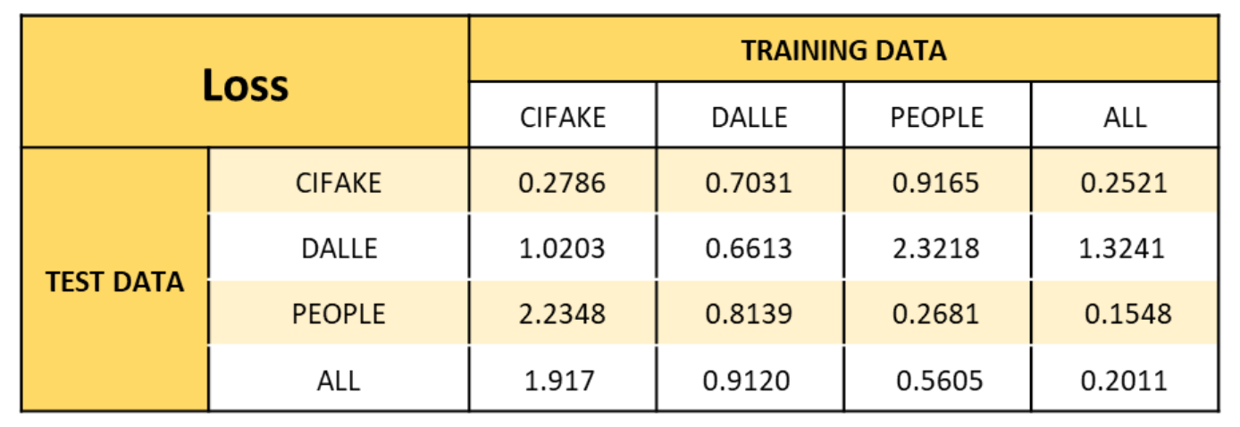
Optimizer Results		Training Data		
		F1-score	Loss	Learn Rate
Optimizer	RMSprop	68.85	0.220	0.00030
	Adam	66.92	0.225	0.00030
	Adagrad	66.71	0.257	0.00005
	Adamax	66.70	0.213	0.00005
	Nadam	66.70	0.254	0.00050
	AdamW	66.68	0.215	0.00040
	SDG	66.67	0.224	0.00050
	Adadelta	66.58	0.421	0.00500

Fig. 1 Table of the highest F1 Scores and associated losses of various optimizers

With the best performing optimizer identified, we proceeded to use a CNN model utilizing the RMSprop optimizer. For each of the datasets used, we chose to train with each individual dataset then evaluate its F1-Score and Loss when tested using each possible dataset. In addition, we trained and tested the model using a combination of all datasets to evaluate overall performance.



F1-Score		TRAINING DATA			
		CIFAKE	DALLE	PEOPLE	ALL
TEST DATA	CIFAKE	68.85	66.67	66.65	66.73
	DALLE	72.48	74.43	73.91	74.43
	PEOPLE	29.16	26.14	26.81	26.14
	ALL	57.30	44.05	44.01	44.00



Loss		TRAINING DATA			
		CIFAKE	DALLE	PEOPLE	ALL
TEST DATA	CIFAKE	0.2786	0.7031	0.9165	0.2521
	DALLE	1.0203	0.6613	2.3218	1.3241
	PEOPLE	2.2348	0.8139	0.2681	0.1548
	ALL	1.917	0.9120	0.5605	0.2011

Fig. 2 (Left) Table of model F1-Scores when using various training and testing datasets.
Fig.3 (Right) Table of model Loss values when using various training and testing datasets.

From our testing, we found that our F1-Score was at its highest when we trained and tested using the same dataset. Despite this, the F1-Score maintained a relatively similar value across different training and testing combinations. When combining all datasets into a single one, it maintained a comparable F1-Score with the CIFAKE and DALLE datasets.

When evaluating our Loss values, similarly to F1-Score, we found it was at its lowest when trained and tested using the same dataset. Unlike F1-Score, we found that Loss performed marginally better when trained using all datasets and tested using CIFAKE.

Conclusions

For our testing, we looked into the benchmark performance of CNN and compared it to the performance of different generation techniques. In our analysis, we found that not only did the performance change, it also improved in cases. Through analysis of various optimization methods, we have found that the best performing optimizer was RMSprop with a learning rate of 0.0003.

When testing the model using different generation techniques, we found it generally performed best when using only one generation technique via training and testing with the same dataset. While F1-Scores were comparable when using multiple different generation techniques, Loss values had a significant difference in performance when using different generation techniques.

Contact Information

Chris Foster – cfoste71@students.kennesaw.edu

Joshua Brock – jbrock42@students.kennesaw.edu

Srilatha Korrapati – skorrap3@students.kennesaw.edu

Harini Kottala – hkottala@students.kennesaw.edu

References

Jordan J. Bird and Ahmad Lotfi. 2024. CIFAKE: Image classification and explainable identification of AI-generated synthetic images. *IEEE Access* 12 (January 2024), 15642–15650. DOI:http://dx.doi.org/10.1109/access.2024.3356122

A. Krizhevsky and G. Hinton, ‘‘Learning multiple layers of features from tiny images,’’ 2009
COMPVIS/stable-diffusion: A latent text-to-image diffusion model. https://github.com/CompVis/stable-diffusion

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 45, 9 (September 2023), 10850–10869. DOI:http://dx.doi.org/10.1109/tpami.2023.3261988

Krichen, M. Convolutional Neural Networks: A Survey. Computers 2023, 12, 151. https://doi.org/10.3390/computers12080151

https://www.kaggle.com/datasets/birdy654/cifake-real-and-ai-generated-synthetic-images
https://www.kaggle.com/datasets/gauravduttakiit/dalle-recognition-dataset
https://www.kaggle.com/datasets/atulanandjha/lfwpeople
https://www.kaggle.com/datasets/ahmadahmadzada/images2000

KENNESAW STATE
UNIVERSITY
COLLEGE OF COMPUTING AND
SOFTWARE ENGINEERING

Author(s) – Chris Foster, Joshua Brock, Srilatha Korrapati, Harini Kottala
Advisors(s) – Dan Lo