

Capstone Project Report

James Cook University Cairns

Hunter Kruger-Ilingworth (14198489)

August 4, 2025

Contents

1 Introduction **2**

2 Dataset **3**

3 References **4**

List of Figures

1 AI generated images [1] 2

List of Tables

1 Dataset features and their descriptions. 3

2 Counts of each class label in the training and testing sets, where it can be seen that the testing set has the class labels withheld due to the competition setting 3

Introduction

"AI as a creative tool has improved leaps and bounds since its early incarnations. Generating still images that are interesting, sophisticated and photorealistic is now an easy process that can be done by anybody with an interest, some patience and determination" [2]. The prevalence of AI among the general public has led to extensive dialogue about their ethical implications in producing artwork, or for manufacturing misinformation. One such example of the use of AI is tools like *DALL-E* and *Stable Diffusion*, which generate images from text prompts. In the current digital age, we cannot rely on the transparency of the source of an image, so there exists a need to be able to identify whether an image has been generated by an AI, or a real artist. Figure 1 shows some examples of AI generated images, which are becoming increasingly difficult to distinguish from real images as the technology improves, underscoring the need for reliable detection methods.



Fig. 1. AI generated images [1]

[1] evaluates both human and AI capabilities in detecting fake images, showing that humans are often deceived by advanced image generation models, while AI-based detection algorithms outperform humans but still misclassify 13% of images. The study introduction of the Fake2M dataset and new benchmarks (HPBench and MPBench) aims to drive further research and improve the reliability of AI-generated content detection.

[3] presents a computer vision approach to distinguishing AI-generated images from real ones, utilising a synthetic dataset created with Latent Diffusion, classification via Convolutional Neural Networks, and interpretability through Grad-CAM. Achieving nearly 93% accuracy, the study also introduces the CIFAKE dataset, a large collection of real and synthetic images, to support further research on the detection of AI-generated imagery.

It is therefore evident that past research has shown that it is possible to classify images as AI-generated or not, with a high degree of accuracy. This report will explore the process of training a neural network to classify images as either AI-generated or not, and then deploying this model to AWS SageMaker for inference, and evaluating it against comparable models.

Dataset

The dataset used for this project was a competition dataset from Hugging Face, held in 2023 [4]. The dataset consists of 62,060 images, and is 2.37GB in size, being pre-split into training and testing sets, as summarised in tables 1 and 2, where it can be seen that the testing set has the class labels withheld due to the competition setting, restricting this analysis to the 18,618 training images, which we can sub-divide and validate with known labels.

Feature	Description
id	Index filename 34.jpg
image	The Image object (rgb 512x512 resolution)
label	Binary class label [1=AI, 0=not AI]

Table 1: Dataset features and their descriptions.

Class Label	Train Count	Test Count
AI (1)	10,330 (55.5%)	NA
Not AI (0)	8,288 (45.5%)	NA
Total	18,618	43,442

Table 2: Counts of each class label in the training and testing sets, where it can be seen that the testing set has the class labels withheld due to the competition setting

Listing 1 shows the code used to load the dataset, and split it into training, validation, and test sets [have a look at this later](#). The code demonstrates how the data is imported. Since Hugging Face datasets are not natively compatible with TensorFlow, the dataset is converted to a TensorFlow dataset using the `to_tf_dataset()` method. [will need to mention s3 buckets](#)

Listing 1: Data Wrangling Script

```

1
2 with open("token.txt", "r") as file:
3     hugging_face_token = file.read().strip() # must have credentials to access the dataset
4
5 login(token=hugging_face_token)
6 raw_dataset = load_dataset('competitions/aiornot') # hugging face method to import the data
7
8 dataset = raw_dataset['train'] # remove the unused 'test' set with no labels
9
10 # split dataset into:
11 # - small_train (10% of original)
12 # - train (70% of original)
13 # - test (20% of original)
14
15 # First split: small_train (10%) and remainder (90%)
16 split_1 = dataset.train_test_split(train_size=0.1, seed=RANDOM_SEED)
17 small_train = split_1["train"]
18 remainder = split_1["test"]
19 split_2 = remainder.train_test_split(train_size=0.7 / 0.9, seed=RANDOM_SEED)
20 train = split_2["train"]
21 test = split_2["test"]
22
23 is_any_data_unused = not(small_train.num_rows + train.num_rows + test.num_rows == dataset.num_rows)
24 assert not is_any_data_unused, "Some data is unused in the splits."
25
26 def format_dataset(hugging_face_dataset, batch_size=32, shuffle=True):
27     """convert a hugging face dataset to a TensorFlow dataset"""
28     dataset = hugging_face_dataset.with_format(type='tf', columns=['image', 'label'], output_all_columns=True)
29     dataset = dataset.to_tf_dataset(columns='image', label_cols='label', batch_size=batch_size, shuffle=shuffle)
30
31     # normalise the image channels to be in the range [0, 1] rather than [0, 255]
32     dataset = dataset.map(lambda image, label: (tf.cast(image, tf.float32) / 255.0, label), num_parallel_calls=tf.data.AUTOTUNE)
33     # return prefetched dataset
34     return dataset.prefetch(buffer_size=tf.data.AUTOTUNE)
35
36 small_train_tf = format_dataset(small_train, batch_size=32, shuffle=True)
37 train_tf = format_dataset(train, batch_size=32, shuffle=True)
38 test_tf = format_dataset(test, batch_size=32, shuffle=False)

```

References

- [1] Z. Lu, D. Huang, L. Bai, J. Qu, C. Wu, X. Liu, and W. Ouyang, “Seeing is not always believing: Benchmarking human and model perception of ai-generated images,” 2023. [Online]. Available: <https://arxiv.org/abs/2304.13023>
- [2] C. Scott *et al.* (2024) Exploring the ethics of artificial intelligence in art. Accessed: 22 July 2025. [Online]. Available: <https://www.artshub.com.au/news/opinions-analysis/exploring-the-ethics-of-artificial-intelligence-in-art-2694121/>
- [3] J. J. Bird and A. Lotfi, “Cifake: Image classification and explainable identification of ai-generated synthetic images,” 2023. [Online]. Available: <https://arxiv.org/abs/2303.14126>
- [4] Hugging Face Datasets, “competitions/aionot,” <https://huggingface.co/datasets/competitions/aionot>, 2023, accessed: 22 July 2025.