**Dataset**

1. CIFAR-10 dataset

The **CIFAR-10 dataset** is a well-known benchmark in the field of machine learning and computer vision, primarily used for image classification tasks.

* **Total Images**: 60,000
* **Image Size**: 32x32 pixels
* **Colour Type**: Colour images
* **Classes**: 10 distinct categories

Each class contains 6,000 images, distributed as follows:

A screenshot of a computer

Description automatically generated

Processing on the data set, loding and all is done in the notebook link provided below

<https://colab.research.google.com/drive/1_WG3jbajx5bZxcaon7CaMWYINZtK4A3I?usp=sharing>

Limitations of CIFAR-10

1. **Low Resolution**: The CIFAR-10 images are quite small, which may not provide enough detail for distinguishing between real and AI-generated images effectively. Higher-resolution images typically yield better performance in image classification tasks.
2. **Limited Classes**: CIFAR-10 contains only 10 classes, which may not provide a diverse enough set of examples for training a model to differentiate between real and AI-generated images.

CIFAR-100 dataset can be used but it has the same problem of low resolution.

Conclusion:

Using the **CIFAR-10 dataset** for differentiating between AI-generated images and real images may not be ideal due to its low resolution (32x32 pixels) and limited diversity.

1. CIFAKE Dataset

The **CIFAKE dataset** is specifically designed to aid in the differentiation between real images and those generated by artificial intelligence

* **Total Images**: 120,000 images
  + **Real Images**: 60,000 (sourced from the CIFAR-10 dataset)
  + **Fake Images**: 60,000 (generated using Stable Diffusion version 1.4)
* **Classes**: The dataset is divided into two classes:
  + **REAL**: Images from the CIFAR-10 dataset.
  + **FAKE**: AI-generated images that mimic the CIFAR-10 style.

Dataset Structure

* **Training Set**:
  + **50,000 Real Images**
  + **50,000 Fake Images**
* **Testing Set**:
  + **10,000 Real Images**
  + **10,000 Fake Images**

The CIFAKE dataset is intended for research and development in computer vision techniques aimed at discerning authenticity in images.

A screenshot of a calculator

Description automatically generatedA screenshot of a calculator

Description automatically generated

Since the dimensions of the both the fake images and real images is same we can use this data set for our project but the question which pops up now is what if there is some other input which if of high resolution, our model might fail.

This can be used temporarily but must be changed.

Processing on the data set, loding and all is done in the notebook link provided below

<https://colab.research.google.com/drive/1XUvwwB38zCwMLH1wh85VZVxcdD4Rr4ua?usp=sharing>

Conclusion:

This can be used for project if we didn’t get any other dataset, but I personally would like to search for some other dataset

1. **Flickr-Faces-HQ (FFHQ)**

The **Flickr-Faces-HQ (FFHQ)** dataset is a prominent collection of high-quality images of human faces, primarily created for research in face generation and related tasks.

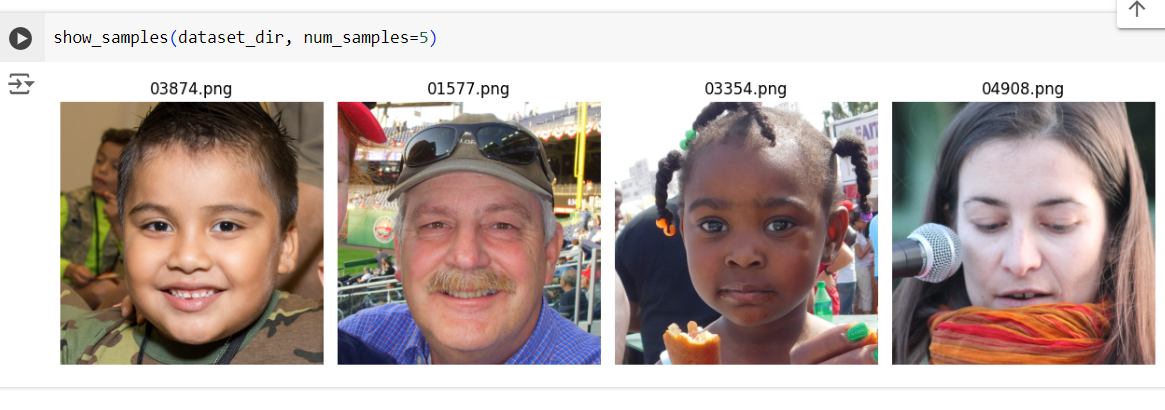
Overview

* **Total Images**: 70,000
* **Image Resolution**: 1024×1024 pixels
* **Unique Photos**: 67,646 unique Flickr photos
* **Source**: Images were crawled from Flickr, adhering to various Creative Commons licenses.

Licensing

Images in FFHQ are published under various Creative Commons licenses, allowing for free use and redistribution with proper attribution.

So we can use this dataset.



Processing on the data set, loding and all is done in the notebook link provided below

<https://colab.research.google.com/drive/1rMPeDYh7Xo3vkIw499h1Yv3w2vGYwOD-?usp=sharing>

Conclusion:

This data set can be used for our project as it has high quality images, and if our problem statement focuses on real vs AI generated image of human face our model could work fine.

1. Synthetic Faces High Quality (SFHQ)

The **Synthetic Faces High Quality (SFHQ)** dataset is a comprehensive collection of synthetic face images designed for various applications in computer vision, particularly in the realm of face generation and recognition.

* **Total Images**: Approximately **425,000** curated synthetic face images.
* **Image Resolution**: All images are **1024x1024 pixels**.
* **Creation Process**: The dataset was created by transforming various "inspiration" sources—such as paintings, drawings, 3D models, and text-to-image generators—into photorealistic images. This process involves encoding the images into the **StyleGAN2 latent space** and performing manipulations to enhance realism.

A screenshot of a computer

Description automatically generated

These images are high quality AI generated images and can be used for our project

https://colab.research.google.com/drive/15emlmJK4UacRf3xStgyDIEmRh\_DPI7NZ?usp=sharing

Finalized dataset :

Synthetic Faces High Quality (SFHQ) : For AI generated images

Flickr-Faces-HQ (FFHQ) : For Real images.