Detecting AI generated images using a machine learning model

Andreas Jürgenson, Richard Jaarman

Business understanding

Artificial image generation tools like Midjourney, Google's Nano Banana, OpenAI's DALL-E and many others are currently more accessible than ever. These powerful models are capable of creating highly realistic images nearly indistinguishable from real ones. These tools have made identifying false news or just misleading images a lot harder than it used to be.

Our project addresses this problem by aiming to create a machine learning algorithm that would be able to distinguish real photographs from AI generated images. This tool could benefit people who are less technologically aware to tell apart the fake from the real and inhibit the spread of misleading news and clickbait.

Business success will be measured by how effective the final system is at identifying artificially generated images. We are aiming to achieve a reasonable level of classification accuracy of at least 75%. The project will be considered a success if we gain a clear understanding of the patterns that differentiate the two types of images and the model reliably identifies artificial images.

The resources available for this project are limited, but enough for our initial needs. The project team consists of university students with base knowledge of machine learning, neural networks and image processing, which will be needed. Data resources include 2 kaggle competition sets containing 13 GB of images, half of which are AI generated. The project team has access to 2 main computers for training and testing. Computer 1 has an AMD Radeon RX 9070 XT with 16 GB of memory and 32 GB of RAM. Computer 2 has an Nvidia GeForce RTX 3060 with 12 GB of memory and 32 GB of RAM. Software resources include Python and its various libraries like pandas, OpenCV, scikit-learn, Jupiter notebooks and GitHub for version control.

Requirements for this project are to develop a functioning machine learning model capable of distinguishing between AI-generated and real images by December 8th, in time for the poster session. The team assumes the data contains a good split of both AI-generated and real images and that there are patterns which exist to be learned by the model. Constraints include limited computing power due to not having access to high end Nvidia GPUs.

This project could face multiple risks during its completion. Firstly the dataset could contain unbalanced data. To mitigate this the team could gather additional data for training. Another risk is that computing limitations can slow training. To combat this, the team could use a smaller model or seek additional computing power. Finally the effectiveness of this model on new AI image generation tools may be low. This risk is mitigated by diversifying training data further.

The following terms are defined so everybody involved can have a common understanding of the project:

- AI - Artificial Intelligence.
- AI - generated image - image produced entirely by an artificial intelligence model.
- Real image - a photograph captured by a camera.
- Binary classification - putting things into one of two categories (in our case AI or real).
- Dataset - collection of data used for training and testing a model.
- Training set - example data used to train ML models to make predictions and find insights.
- Test set - data used to evaluate how well the model does with data outside the training set.
- Accuracy - proportion of all classifications that were correct, whether positive or negative.
- Precision - proportion of all the model's positive classifications that are actually positive.
- Recall - proportion of all actual positives that were classified correctly as positives.
- F1 score - harmonic mean (a kind of average) of precision and recall.

Costs include time spent on additional data collection, model development and evaluation, as well as computational costs associated with training and testing models. Benefits include gaining experience with machine learning and data science projects and gaining insight into patterns behind AI image generation.

The primary data mining goal of this project is to develop a machine learning model capable of classifying images as AI-generated or real. Additional deliverables include an evaluation report containing the models performance and visualizations which show the patterns the model uses to differentiate images. Data mining is successful if the model's accuracy is over 75%.

Data Understanding

To create a machine learning algorithm for detecting AI generated images, the project requires a large set of data inputs and supporting data for training. The main data set consists of AI generated images and stock photos taken by humans, both in .jpg format. The images should vary a lot in subject and style. All of the images have a corresponding value in a .csv file, if it’s AI generated or not.

The required data is available on two publicly accessible Kaggle datasets, both providing a large collection of AI generated images and stock photos taken by humans in a large variety of styles, subjects and quality. These datasets have the required .csv files that label each image as AI generated or real. This ensures that the machine learning model can be trained correctly and that the data fits the project requirements for training and testing.

The selection criteria for this project focuses on identifying the specific data sources and subsets within these sources that are relevant for training a machine learning model to detect AI generated images. The two Kaggle data sets chosen for this project will be the primary data sources, and the image files in .jpg format and their corresponding labels in the provided .csv files will be used. If there are any images with missing labels, they will be excluded from the main dataset to maintain quality. The data from .csv files have the binary classification indicating if the image is AI generated or human made.

The data that will be used for this project comes from two publicly available Kaggle datasets made for training a machine learning model to distinguish between AI generated images and real photographs. Both datasets contain images in .jpg format alongside .csv files, which include the image file name and binary labels that indicate whether each image is AI generated or taken by a human. Together the datasets have 105 thousand images. The training data is about 75% of the images and test data is 25% of the images. The images vary in style and subject, providing the necessary diversity to develop a model capable of generalizing across different types of images.

After acquiring the datasets, an initial exploration of the data was done to better understand the structure and quality of the available data. The two Kaggle datasets were examined by looking at the .csv files and reviewing samples from the corresponding .jpg images. This examination confirmed that both datasets contain clearly labeled classes AI generated images and real photographs. A comparison between the two datasets revealed that they are mostly the same, except that one has 20 thousand more test images, so the training will be done on one dataset and testing will be done using both datasets test images. Training images in both datasets are 50% AI generated and 50% real images. The examination of the data indicates that the data is of adequate quality and diverse enough for model development.

After exploring the data a verification of data quality was done. This process involved checking the completeness and consistency of the image files and their corresponding csv file entries. No major issues were found in the data quality and all the listed files were matched correctly. The data quality is high and only requires preprocessing for images to train the model. This makes it appropriate to train a model using this data.

Project plan

Task	**Description**	**Time**	**Team member responsible**
Repository setup	Setting up the repository for the models codebase.	1 hour	Andreas
Data collection	Download the data from Kaggle, inspect it for anomalies.	Up to 3 hours depending on the state of the data.	Richard
Data preprocessing	Resize and augment the images.	1 hour	Richard
Model selection	Test different models and architectures.	6 hours	Andreas & Richard
Model training*	Training the model on the training dataset.	1 day	Andreas & Richard
Tuning hyperparameters*	Tune the model's hyperparameters.	6 hours	Andreas & Richard
Evaluation	Evaluate the model's performance.	2 hours	Andreas & Richard
Writing a report	Writing a report containing the	3 hours	Andreas & Richard

	models performance and the team's findings.		
Poster	Making a poster for submission.	1 hour	Andreas

* Tasks may be repeated if model performance is low

These tasks add up to a total of 32 hours per team member.

Methods and tools

Methods:

- Data preprocessing - before training the model, the images should be standardized and prepared. Main steps:
  - Image resizing - Resize all images to a consistent size (256x256 pixels)
  - Data augmentation - Transformations such as rotations, flips, cropping and compression noise to improve generalization
  - Cleaning the dataset - Remove files with missing data.
- Training - Test different models to see which fits best for this project
- Model validation - Choose a Train/Validation/Test split to ensure reliability
- Evaluation Methods - To measure model performance we will use accuracy, precision and recall, Confusion Matrixes to see how well the model distinguishes between the two classes.

Tools:

- Github repository - https://github.com/Andreasest/ai-image-detection
- Programming language - Python
- Development environment - Jupyter notebook
- Libraries - Pandas, scikit-learn, OpenCV, Matplotlib…
- Hardware:
  - AMD Radeon RX 9070 XT with 16 GB of memory and 32 GB of RAM
  - Nvidia GeForce RTX 3060 with 12 GB of memory and 32 GB of RAM
- Two datasets from Kaggle (13GB) - including AI generated images and real photographs.