Model Details

• Developed by: GRIP-University Federico II of Naples
• Model date: 29/01/2023
• Model version: 1.0.

• Processing pipeline:
– Download the image from the input URL.
– Preprocess the image by applying the necessary transformations or adjustments to align it with the detector's requirement.
– Feed the image to the detector to get a fake probability score in the range of (0,1).
• Model input: Image URL or image file.
• Model output: The probability for the image to have been generated by a Generative Adversarial Network (GAN) or a similar architecture. Probabilities below 50% and closer to 0% mean real and above 50% and closer to 100% mean fake.
• Model type: ResNet-50 with no downsampling in the first layers.
• Method: Avoiding downsampling in initial layers of the convolutional neural networks to preserve subtle artifacts and employing intense data augmentation. Training on ProGAN data.
• Related Paper: On the detection of synthetic images generated by diffusion models
• Feedback and Contact: Riccardo

– The model cannot detect if an image has been tampered with using manipulations or other forgeries (e.g. splicing, copy-move, in-painting).
Metrics

• Model performance measures:
– Accuracy: It evaluates the overall correctness of the model's classifications, measuring the ratio of correctly predicted instances to the total instances evaluated, with possible values ranging from 0 to 1. Higher is better.
– AUC: defined as the Area Under Curve (AUC) based on the Receiver Operating Characteristic (ROC) curve with possible values ranging from 0 to 1. Higher is better.
• Metrics decision: Accuracy and AUC are popular metrics used in the related literature. Accuracy offers a straightforward understanding of the model's performance, while AUC does not rely on any predetermined threshold, making it suitable for assessing the robustness and generalization of the detector.

Relevant Datasets

• COCO: The COCO dataset contains a large collection of images that are richly labeled with object annotations, segmentation masks, and keypoints. It offers a diverse range of object categories and complex real-world scenes.

• ImageNet: The ImageNet dataset is one of the most influential datasets in computer vision. It consists of millions of labeled images across thousands of categories.

• Large-scale Scene Understanding (LSUN): The LSUN dataset contains images of 10 scene categories, such as dining room, bedroom, chicken, outdoor church, and so on, and twenty object classes.

• UCID: The Uncompressed Color Image Database, offers a collection of 1338 uncompressed images paired with ground truth data.
• The generated images are acquired by employing a pretrained GAN or Diffusion model to generate artificially generated images for different classes to evaluate the cross-concept scenario.

(riccardo.corvi@unina.it), (davide.cozzolino@unina.it), (verdoliv@unina.it).

Intended Use

Davide Luisa

Corvi Cozzolino Verdoliva

• Primary intended use: Detect whether the input image has been generated using a Generative Adversarial Network (GAN).
• Primary intended users: Journalists, media verification companies / organizations / groups and researchers working on the problem of Synthetic Image Detection.

• Out-of-scope uses:
– The model cannot detect whether the faces present in the input image (if any) have been manipulated using Deep Learning methods (DeepFake).

GRIP ProGAN Detector - Model Card “progan_r50_grip”

Training Data

– Datasets: Real images obtained from the LSUN dataset and synthetic images generated by the ProGAN model. (20 models each trained on a different LSUN object category are used to generate 18K train images each.)

– Preprocessing: A set of various augmentations is applied as a preprocessing step during training, including blurring, and compression.

Evaluation Data

– Datasets: Real images obtained from COCO, ImageNet and UCID datasets and synthetic images generated by several state-of-the-art generative models including GANs, Transformers, and Diffusion models: ProGAN, StyleGAN2, StyleGAN3, BigGAN, EG3D, Taming Transformer, DALL·E Mini, DALL·E 2, GLIDE, Latent Diffusion, Stable Diffusion and ADM (Ablated Diffusion Model).

– Setup: Training on LSUN and ProGAN data and evaluating on the rest of the data.

Caveats and Recommendations

– General performance: The performance of the detector highly depends on the generative model it has seen during training. For example, if a detector is trained exclusively on one generative model, it may perform poorly on most images generated by different models. The generalization to novel manipulations is an open research issue that almost all approaches suffer from, including ours. Thus, we cannot guarantee good performance on unseen manipulations.

– Image quality: It is recommended that the input media be of the best quality possible since factors like compression and blur can lead to erroneous detection.

– Adversarial attacks: an adversarial attacker might affect detection accuracy. Even though these attacks might not be visible to the naked eye, they can fool a synthetic image detector into assessing that a synthetic image is real.

Relevant Factors

• Factors affecting model performance include:
– Whether the models have been trained with the presented generative method or not. (Refer to the Training Data section for more information.)
– Image quality: blurry or low quality images can lead to erroneous detections (false positives).
–Adversarial Attacks: alterations in the images to evade detection are detrimental to detection accuracy.

Quantitative Analysis

GRIP ProGAN Detector - Model Card “progan_r50_grip”

Acc./ AUC%	progan_r50_grip trained on ProGAN	Fusion with ldm_r50_grip trained on Lattent Diffusion

ProGAN	99.9/100	90.2 / 100
StyleGAN2	63.3 / 94.8	56.6 / 94.6

StyleGAN3	58.3 / 94.4	55.4 / 93.9
BigGAN	79.0 / 99.1	59.3 / 98.5
	56.8 / 96.6	54.4 / 97.7

EG3D		
Taming Tran.	56.2 / 94.3	61.5 / 98.2
	62.3 / 95.4	65.9 / 97.7

DALL·E Mini		
DALL·E 2	50.0 / 64.4	50.0 / 72.5
	51.8 / 90.0	52.5 / 95.9

GLIDE		
Latent Diff.	52.4 / 89.4	84.9 / 99.8
	58.1 / 93.7	92.5 /100

Stable Diff.		
ADM	50.6 / 77.2	50.8 / 80.6

Table 1: Accuracy and AUC comparison for the progan_r50_grip detector (trained exclusively on ProGAN data) and its fusion with the ldm_r50_grip detector (trained on Latent Diffusion data). The fusion strategy involves selecting a soft maximum of the two probabilities at each instance.

Performance Intuition

GRIP ProGAN Detector - Model Card “progan_r50_grip”

– Cross-Concept Evaluation: Training the model on ProGAN data tends to excel with GANs but a significant performance drop is observed with Diffusion models. The fusion approach demonstrates a balanced performance, offering adequate accuracy and AUC scores across both GAN and Diffusion models. The fusion strategy selects a soft maximum of the probabilities at each instance, leveraging the complementary strengths of both detectors. In many cases, the fusion results in substantial performance gains, with accuracy and AUC scores surpassing those achieved by gan_r50_mever alone.

Explanation: The veraAI detector of the verification plugin finds weak evidence (non conclusive) suggesting that this image could be synthetic. The absence of detection does not guarantee the image is not post-edited.

What do the gauge colors mean? <50% means Weak evidence (non-conclusive). ≥50% means Moderate evidence (suspicious but non-conclusive). ≥70% means Strong evidence, i.e., image was AI-generated. ≥90% means Very strong evidence, i.e. image is AI-generated.

If one or more algorithms returned errors, the analysis may be incomplete. Make sure the image has dimensions between 128x128 and 2Mpx.

This is a beta version tool aiming to detect synthetic images generated through Dall-E, Midjourney or other kind of Diffusion models. A probability equal or above 70% is considered a detection.