Towards Adversarially Robust AI-Generated Image Detection

Annan Zou
Department of Computer Science, Vanderbilt University, Nashville, U.S.A.

Keywords: Artificial Intelligence Generated Content (AIGC), Adversarial Robustness, Image Classification.

Abstract: Over the last few years, Artificial Intelligence Generated Content (AIGC) technology has rapidly matured and garnered public attention due to its ease of use and quality of results. However, due to these characteristics, forged images generated by AIGC technology have a high potential of being misused and causing negative social consequences. While AI-based tools can identify AI-generated images with reasonable accuracy, these models did not consider the factor of adversarial robustness or resistance against intentional attacks. This paper empirically evaluated several existing AIGC detection models' adversarial robustness under select attack setups. Overall, it is discovered that even naked-eye unnoticeable perturbation of source images can consistently cause sharp drops in performance in all the models in question. This study proposes constructing a Convolutional Neural Network (CNN) based AIGC classifier model with additional adversarial training using a combination of transformation-based and l-based adversarial examples constructed with existing AIGC data. This paper uses clean and adversarial data sets to test the performance of the resulting model. The results show the model's remarkable robustness to the adversarial attack techniques described above while maintaining relative accuracy on clean datasets.

1 INTRODUCTION

Artificial Intelligence Generated Content (AIGC) is creating content via artificial intelligence-based systems. The topic of AIGC has garnered enormous public and academic attention for the past few years due to its rapid improvement, proliferation, and commercial adoption. AIGC systems can produce various content types from different inputs. However, text-to-image generation models have also been a particular focus of controversy due to their ample potential for misuse. Some worried they were threatening the art industry and stifling creativity in the visual arts.

In contrast, others pointed out that they could be used to disseminate misinformation by generating photorealistic scenes of non-existent events (Z. Sha et al., 2022). Moreover, text-to-image AIGC has been, and is still, improving at an impressive pace, to the point that some are worried that it will soon reach the point where it becomes altogether impossible for human eyes to distinguish between AI-generated images and natural photographs. This uncertain future necessitates the development of tools that can reliably discern AI-generated images.

There have been several works that focused on AIGC image classification. Bird & Lotfi used a customized Convolutional Neural Network (CNN) with a Gradient class activation map (Grad-CAM) to offer a more explainable approach to classify AI-generated images and generalize AIGC artifacts (J. Bird Jordan, and L. Ahmad, 2023). Xi developed a novel dual-stream network for pure image classification that synthesizes AIGC artifacts in high and low-frequency regions of a given image with a cross-attention module (Z. Xi et al., 2023). All the models achieved impressive classification accuracy that far exceeds regular forensic models. However, no studies have focused on AIGC classifiers' adversarial robustness – or how well they perform with inputs intentionally engineered to misguide or otherwise disrupt their classification (A. Madry et al., 2017).

All the methods above utilize Convolutional Neural Network (CNN) architectures, which have been proven to be especially susceptible to adversarial attacks such as projected gradient descent (PGD) in previous studies (A. Madry et al., 2017), (L. Engstrom et al., 2019).

This paper aims to address these issues by (1) assessing the adversarial robustness of existing models and (2) creating a model that can achieve a

Zou, A.
Towards Adversarially Robust AI-Generated Image Detection.
DOI: 10.5220/0012816300003885

In *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, pages 380-385
ISBN: 978-989-758-705-4

higher degree of robustness against both adversarial attacks and naturally occurring transformations. The first goal is achieved using a combination of the CIFAKE database developed by Bird & Lotfi (J. Bird Jordan, and L. Ahmad, 2023), adversarial perturbation methods utilizing the Adversarial Robustness Toolbox, and input transformation methodology inspired by Engstrom et al. to generate an adversarial dataset based on the CIFAKE database and then testing the performance of existing models on the dataset (L. Engstrom et al., 2019), (M. I. Nicolae, and M. Sinn, 2018). The second goal is constructing a CNN-based model that achieves adversarial robustness utilizing data augmentation and adversarial training. To ascertain that goal (2) is fulfilled, this study comprehensively evaluates the resulting model's performance with the clean CIFAKE database and those above adversarially perturbed datasets, including comparative and ablation studies. In summary, the contributions that this paper have made include:

- Constructed an adversarial dataset of real and T2I (Text-to-image) AI-generated images using different adversarial perturbation techniques, totaling a 120,000-image count with two 60,000-image groups corresponding to clean and attacked input data.
- Evaluated the adversarial robustness of existing models using adversarial databases and preprocessing scripts that perform rotations and translations. This evaluation shows that existing models are vulnerable to adversarial perturbation and preprocessing.
- Introduced an adversarially robust model for T2I AIGC detection using a combination of preprocessors and adversarial training. Using the aforementioned adversarial datasets, this model is verified to have superior adversarial and spatial robustness than existing models while maintaining comparable accuracy on clean datasets.

2 METHOD

This section can be divided into three sections. The first section (II. A, II. B, II. C) describes the source of the data, the $l-p$ bounded adversarial perturbation methods to generate the adversarial datasets, and the adversarial spatial transformation procedures. The second section (II. D) briefly describes the models used in the robustness evaluation and the methods and metrics used. The third section (II. E) describes, in detail, the overall architecture of an adversarially robust AIGC detection model, including preprocessors, the classifier proper, the adversarial training procedures, and the loss function.

2.1 Base Dataset

This paper utilizes the CIFAKE database which consists of 120,000 images as the base dataset. 60,000 of these are photographic images taken from the CIFAR-10 dataset, a database widely used for image classification tasks (J. Bird Jordan, and L. Ahmad, 2023). These are 32 × 32 resolution images of real subjects divided into ten classes, in RGB channels. The other 60,000 are RGB images generated by Stable Diffusion v1.4, a popular, public T2I model that utilizes the principle of latent diffusion to generate synthetic images. The AIGC dataset is formatted the same as the CIFAR data: ten classes of 32 x 32 images of objects in RGB channels. For this study, the dataset is divided into two equal-sized subsets each consisting of 30,000 pairs of natural and AIGC images. For each of the subsets, 83.3% (25,000 pairs) are used for training and 16.7% (5,000 pairs) are used for testing.

2.2 Gradient-Based Adversarial Methods

Adversarial attacks on classifiers are considered optimization problems that maximize the given classification's loss while minimizing the perturbation to the input. Formally, given any classifier $f(x): x \rightarrow y$ that maps input *x* to label *y*, an adversarial attack seeks out an adversarial perturbation δ such that:

$$L(f(x+\delta;\theta), y), \|\delta\|_p \leq \epsilon \quad (1)$$

Where $\|\cdot\|_p$ is a l_p norm and ϵ is the given perturbation budget.

Many state-of-the-art adversarial attacks are based on the method of Fast Gradient-Sign Method (FGSM) first proposed by Goodfellow, Shlens and Szegedy. in 2014 where the perturbed input x' is given by (I. J. Goodfellow et al., 2014):

$$x' = x + \varepsilon \cdot sgn\big(\nabla_x L(\theta, x, y)\big) \quad (2)$$

Such that $\nabla_x L$ is the gradient of the original model's loss function with respect to the model parameters, input x and output y. An improved and much more powerful derivative of the FGSM is an iterative version that breaks the problem down into

several smaller maximization problems of step size α and step count $x + S$. This variation, known as Projected Gradient Descent (PGD), is formalized by the expression:

$$x'_{t+1} = \prod_{x+S}[x' + sgn(\nabla_x L(\theta, x, y))] \quad (3)$$

Where denotes the projection operator that projects the finished iterations back to the constraint space $\|\delta\|_p \leq \epsilon$ and therefore clipping δ to the interval $[\epsilon, -\epsilon]$.

For this study's purpose, the Auto-PGD attack first introduced by Croce et al. implemented by the Adversarial Robustness Toolbox (ART) (M. I. Nicolae, and M. Sinn, 2018), (F. Croce, and M. Hein, 2022) is chosen. Its key improvement over the base PGD attack is the ability to dynamically adapt its step size based on the rate of learning, allowing it to use larger step sizes to find good starting points over the whole attack space and smaller step sizes for more aggressive search of local maxima (F. Croce, and M. Hein, 2022). It achieves this by setting N checkpoints to decide if the step size should be halved from the initial size, and whenever the step size is halved, it starts from the best previously found parameters. An iteration count of 500 is used for the attack on the adversarial dataset.

2.3 Adversarial Spatial Transformations

Apart from the gradient-based methods, Engstrom et al. proposed an alternative view on adversarial perturbations, specifically, they questioned the concept of "perturbation budget" defined based on using l_p norms as the sole metric of image similarity (L. Engstrom et al., 2019). They argue that human perception often defines images with large l_p norm variations as visually similar, specifically, images that have undergone small rotation or translation operations. The optimization view of this spatial translation based adversarial perturbation is given by:

$$L(T(x; \delta u, \delta v, \theta_p), y) \quad (4)$$

Where each pixel at position *(u, v)* in the given image undergoes the spatial operation *T*:

$$[coscos\theta_p \theta_p sinsin\theta_p coscos\theta_p] \cdot [uv] + [\delta u \delta v] \quad (5)$$

The author had proposed several methods of solving the maximization problem. These include (1) first order minimization towards the gradient of the loss function from a random choice, (2) grid search over all the possible combinations over the attack parameter space, and (3) generating k different random choices of attack parameters and searching among these. Engstrom et al. had concluded that the third method, dubbed the *worst-of-k method*, achieves a balance between computational performance and loss maximization, while having the advantage of not requiring full knowledge of the target model's loss function, unlike gradient-based attacks (L. Engstrom et al., 2019).

2.4 Target Models of Robustness Evaluation

Four models that can be considered state-of-the-art in AIGC detection are evaluated. These include the ResNet-18 image classification residual neural network, the customized light CNN architecture by Bird et al., the cross-attention enhanced dual-stream network proposed by Xi et al., and an ensemble-based CG detection network developed by Quan et al. that employs a modified FGSM adversarial training method similar to the ones described here (Z. Sha et al., 2022), (J. Bird Jordan, and L. Ahmad, 2023), (Z. Xi et al., 2023), (W. Quan et al., 2020). All these models use cross-entropy loss as their loss functions.

2.5 Adversarial Training

This section introduces the author's attempt at training an adversarially robust AI generated image detection model. The ResNet-18 architecture is used as a base classifier for this task (Z. Sha et al., 2022). The core of the training stage is the technique of Fast Adversarial Training introduced by Wong, Rice, and Kolter (E. Wong et al., 2020). Theoretically, adversarial training is a minimax or saddle point problem such that:

$$\rho(\theta) where \rho(\theta) = L(f(x + \delta; \theta), y) \quad (6)$$

The training technique of Wong, Rice, and Kolter is based on an FGSM adversary (E. Wong et al., 2020). While the base FGSM adversarial technique had been described as not empirically robust against PGD attacks, Fast Adversarial Training utilizes random non-zero initialization of the FGSM perturbations to achieve robustness on par with PGD adversarial training while being computationally much less costly due to removing the iterative factor (E. Wong et al., 2020).

Besides the gradient-based adversarial technique, the aforementioned adversarial spatial transformations are then introduced into the adversary image generation process, since the two types of attacks had been proven to be orthogonal to each other and their effects are simply additive.

Engstrom et al. proposed using the same worst-of-k method as described above to generate adversarial samples and adding extra degrees of translation and rotation $(\delta u, \delta v, p)$ had been proven to help with generalizing across different attack landscapes while not affecting the clean accuracy by much (L. Engstrom et al., 2019). Hence, the choice parameter *k* is set as 10, and the maximum rotation and translation are set to 30° and 5 pixels.

Finally, inspired by Wang et al., all training data are augmented with a 20% probability of either Gaussian blur with σ ∼ Uniform [0, 3] and JPEG compression with quality ∼ Uniform$\{30,31,\ldots,100\}$ (R. Wang et al., 2019).

3 RESULTS

This section presents the results of the robustness evaluation of the aforementioned models as well as the evaluation of the performance of the paper's proposed model.

3.1 Robustness Evaluation of Existing Models

The adversarial datasets are attacked using a white-box attack scheme, as the target model gradients are input into the AutoPGD procedure to generate the adversarial datasets. For this purpose, the adversarial dataset is copied for each target model, and each copy is attacked individually using the gradient information from the target.

A toggleable preprocessing script is used to perform a worst-of-k attack on input images prior to classification tasks. The choice parameter k is fixed to 10, and the maximum rotation and translation parameters are set to 15° and 20% (3 pixels) in any given direction, respectively.

According to Engstrom et al., spatial translations and gradient based attacks occupy orthogonal attack spaces and reduce classification accuracy in an additive manner (L. Engstrom et al., 2019). Therefore, this paper attempts both attack models individually and then combined. The results are then compared to the accuracy obtained from the clean CIFAKE dataset. The results of the evaluation are presented in Table 1.

As shown in the table, the classifiers universally experienced significant accuracy degradation with any form of adversarial attack. Gradient-based AutoPGD significantly outperforms Worst-of-10 in reducing classification accuracy, and the degradation effects are indeed roughly additive to each other. The cross-attention-based networks by Xi and Quan show higher natural accuracy as well as higher robustness spatial transformation-based adversarial attacks. The ENet model in particular shows significantly higher robustness against AutoPGD attacks than the model by Xi et al. due to incorporating gradient-based adversarial training. Though, the model by Xi et al. still displays a higher degree of resistance to PGD attack than "simple" CNN models.

3.2 Robustness Evaluation of the Proposed Model

This section employs the same methodology described in the above sector to evaluate the model constructed with the forms of adversarial training proposed by this paper. Table 2 displays the accuracy of the proposed model, along with ablation experiments performed with each individual adversarial training method as well as disabling the Gaussian-based and JPEG artifact-based data augmentation by Wang et al. (R. Wang et al., 2019).

The proposed model showed significantly improved adversarial robustness against both PGD attacks and adversarial spatial transformations than any of the existing models described above. The natural accuracy is comparable to the unmodified ResNet-18 model but worse than the cross-attention-based models described above.

The ablation study of removing either adversarial training components clearly highlights the orthogonal nature of the spatial transformation attacks and the gradient based perturbations. The absence of either training modules completely nullify the resistance to the corresponding attacks and degrade the combined adversarial robustness accordingly. Though, removing the additional augmentation step did not significantly affect the adversarial robustness of the complete model but slightly increased the model's natural accuracy. This result goes against the common understanding that harder training data typically generate better classifier accuracies.

4 DISCUSSION

Rodriguez et al. proposed that more complex deep learning models are more susceptible to adversarial perturbation attacks. However, the results shown in Table 1 suggest otherwise. While Model-Centric ENet showed a higher degree of robustness due to incorporating adversarial training, even the non-adversarially trained dual-stream network by Xi et al. is surprisingly more robust against the AutoPGD

Table 1: Comparative Accuracy of Evaluated Models Under Different Adversarial Attacks.

Defense Model	Attack Type	Accuracy
ResNet-18	Natural	84.40%
	Auto-PGD l_∞ε=0.033	4.10%
	Worst-of-10 transformation	31.10%
	APGD + W-10	2.20%
Bird & Lotfi (J. Bird Jordan, and L. Ahmad, 2023)	Natural	83.30%
	Auto-PGD l_∞ε=0.033	3.70%
	Worst-of-10 transformation	28.20%
	APGD + W-10	2.80%
Model-Centric ENet (W. Quan et al., 2020)	Natural	92.70%
	Auto-PGD l_∞ε=0.033	39.30%
	Worst-of-10 transformation	43.10%
	APGD + W-10	31.50%
Xi et al. (Z. Xi et al., 2023)	Natural	93.30%
	Auto-PGD l_∞ε=0.033	18.00%
	Worst-of-10 transformation	45.40%
	APGD + W-10	13.10%4

Table 2: Proposed Model's Accuracy Under Different Adversarial Attacks.

Training Mode	Attack Type	Accuracy
Fast Adversarial Training + W-10	Natural	82.40%
	Auto-PGD l_∞ε=0.033	56.10%
	Worst-of-10 transformation	79.80%
	APGD + W-10	55.20%
Worst-of-10 augmentation only	Natural	86.00%
	Auto-PGD l_∞ε=0.033	5.90%
	Worst-of-10 transformation	85.50%
	APGD + W-10	7.10%
Fast Adversarial Training	Natural	82.90%
	Auto-PGD l_∞ε=0.033	58.50%
	Worst-of-10 transformation	25.10%
	APGD + W-10	41.20%
Fast Adversarial Training + W-10 (No Aug)	Natural	82.80%
	Auto-PGD l_∞ε=0.033	55.50%
	Worst-of-10 transformation	81.00%
	APGD + W-10	53.90%

attack than either plain ResNet-18 or even Bird & Lotfi's network that only has 6 convolutional layers in total (J. Bird Jordan, and L. Ahmad, 2023), (Z. Xi et al., 2023), (W. Quan et al., 2020). This discrepancy might have stemmed from the difference in the nature of the tasks. The study by Rodriguez et al. focused on medical image detection where the features are more concentrated and aligned to human perception. Therefore, deeper neural networks might create unnecessarily complex decision boundaries that are more sensitive to adversarial perturbations – in other words, close to overfitting. However, unlike traditional image forgery techniques or shape classification tasks, artifacts of AI-generated images are not limited to high-frequency areas or primary features, and there is in fact evidence of major differences in the overall statistical distribution of the image (J. Bird Jordan, and L. Ahmad, 2023), (Z. Xi et al., 2023). These two factors might mean that AIGC detection tasks necessitate deeper networks for better extraction of latent features since it is harder to determine whether a given feature is robust or strongly relevant to the prediction outcome. Consequently, higher complexity models learned for AIGC detection might be less susceptible to adversarial perturbations of small magnitude/budget. However, the accuracy

drop caused by adversarial examples is still severe on the more complex models and warrants actual robust training techniques.

Time and computational constraints are the main limitations of this study. This study is performed by an individual researcher utilizing Google Colab instances with Tesla T4 GPUs, which necessitated choices such as utilizing 32x32 color images from CIFAR-10 and the ResNet-18 architecture, which ensures training performance but might cause the resulting model to have difficulties with generalizing across different AIGC models. In particular, 32x32 pixels is a significantly lower resolution than the majority of natural and T2I AI-generated images currently available on the internet. This may be the cause of the lower prediction accuracy as there is less space for AI generation artifacts, like the ones hypothesized by Sha et al. to manifest (Z. Sha et al., 2022). However, increasing the resolution of the training samples or model depth would cause a multiplicative increase in all training costs, including T2I image generation, adversarial attacks, and general model training. If there are fewer resource constraints, investigating the interplay of adversarial training, cross-attention-based ensemble models, and higher resolution samples is a promising future area, as both latter factors are shown to improve the natural accuracy of models (J. Bird Jordan, and L. Ahmad, 2023), (W. Quan et al., 2020).

5 CONCLUSION

This study focuses on the problem of the adversarial robustness of models that detect AI-generated images. It aims to (1) evaluate the adversarial robustness of existing models and (2) construct a model that can achieve a higher degree of robustness against adversarial attacks that are either gradient or spatial transformation-based. For purpose (1), several state-of-the-art AIGC detection models are evaluated against both PGD attacks and adversarial translations and rotations. Both attacks are proven to be highly effective at reducing the classification accuracy of all models. For purpose (2), adversarial training and data is utilized along with a convolutional image classifier model, which has an improved degree of robustness against both kinds of adversarial attacks while preserving the accuracy of the base classifier.

In conclusion, this study proves the susceptibility of CNN-based AIGC detection models to adversarial attacks and the possibility of enhancing these models' robustness with adversarial training. As AIGC technology continues to improve and proliferate at an unprecedented pace, AI-based classification technology might be the best solution for combating their abuse. Based on this paper's results, future models that detect AIGC should also take the issue of adversarial robustness in consideration, especially when it comes to distinguishing between what is real and what is fake.

REFERENCES

Z. Sha, Z. Li, N. Yu, and Zhang, Y, De-fake: Detection and attribution of fake images generated by text-to-image diffusion models, arXiv preprint arXiv: 2210.06998, 2022.

J. Bird Jordan, and L. Ahmad, CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images, arXiv preprint arXiv: 2303.14126, 2023.

Z. Xi, W. Huang, K. Wei, W. Luo, and P. Zheng, AI-Generated Image Detection using a Cross-Attention Enhanced Dual-Stream Network, arXiv preprint arXiv: 2306.07005, 2023,

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, Towards deep learning models resistant to adversarial attacks, arXiv preprint arXiv:1706.06083, 2017.

L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, Exploring the landscape of spatial robustness, In International conference on machine learning, PMLR, 2019, pp. 1802-1811.

M. I. Nicolae, and M. Sinn, Adversarial Robustness Toolbox v1.2.0. CoRR, 1807.01069, arXiv preprint arXiv: 1807.01069, 2018.

I. J. Goodfellow, J. Shlens, and C. Szegedy, Explaining and harnessing adversarial examples, arXiv preprint arXiv:1412.6572, 2014.

F. Croce, and M. Hein, Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks, In International conference on machine learning, 2022, pp. 2206-2216.

W. Quan, K. Wang, D. M. Yan, X. Zhang, and D. Pellerin, Learn with diversity and from harder samples: Improving the generalization of CNN-Based detection of computer-generated images, Forensic Science International: Digital Investigation, 2020, vol. 35, pp. 301023.

E. Wong, L. Rice, and J. Z. Kolter, Fast is better than free: Revisiting adversarial training, arXiv preprint arXiv:2001.03994, 2020.

R. Wang, L. Ma, F. Juefei-Xu, X. Xie, J. Wang, and Y. Liu, Fakespotter: A simple baseline for spotting AI-synthesized fake faces, arXiv preprint arXiv:1909.06122, 2019.