Literature Review on Learning on Gradients: Generalized Artifacts Representation for GAN - GeneratedImages Detection [1]
https://openaccess.thecvf.com//content/CVPR2023/papers/Tan_Learning_on_Gradients_Generalized_Artifacts_Representation_for_GAN-Generated_Images_Detection_CVPR_2023_paper.pdf

Literature review by: Kenny, Siaa, Skye

Introduction

As image generation technology GAN (generative adversarial networks) generates realistic fake images, the research article investigates Learning on Gradients (LGrad) in order to detect these GAN- generated images. The research uses a pretrained CNN model; gradients are used in the process and are sent as input to a classifier. Some methods depend on local region aritifacts,or global textures, which may not work for unseen GAN categories/ or when unknown test images are used. This makes a challenge to develop a generalized detector to detect AI generated images.

The research paper uses a "pretrained discriminator of ProGAN to extract gradients of images produced by Celeba-HQ , ProGAN, StyleGAN, StyleGAN2".

The images are filtered out at this point, and only pixels relevant to the target of the pretrained discriminator of ProGAN are considered. The model used to convert images to gradients in he research article is the transformation model. This results in the pretrained model being more efficient which does not require extensive use of trining resources and will be bale to perform given unknown images. ProGAN trains the detectors and evaluate the performance of the LGrad model using cross-category, crossmodel, and cross-model and category. The goal of the research article is to create a model that detect fake AI generated images in a broad range of categories and acts appropriately when faced with an unseen image.

According to the research, after training, the GAN model has unique blueprints or fingerprints and one may extract them in order to identify these GAN model generated images. The methods that used large amounts of training data were not able to produce a generalized detector of unseen fake images. Some of these methods involved pixel or frequency level artificats.

As a result, transformation to gradients model is used; a pretrained CNN model (transformation model) converts all training image and test data to gradients. This gives a more generlalized representation and improves the detection of GAN generated images.

$$l = M(I_i),$$

The following equations shows that images (I_i) are fed into the transformation model M (fixed in the framework)

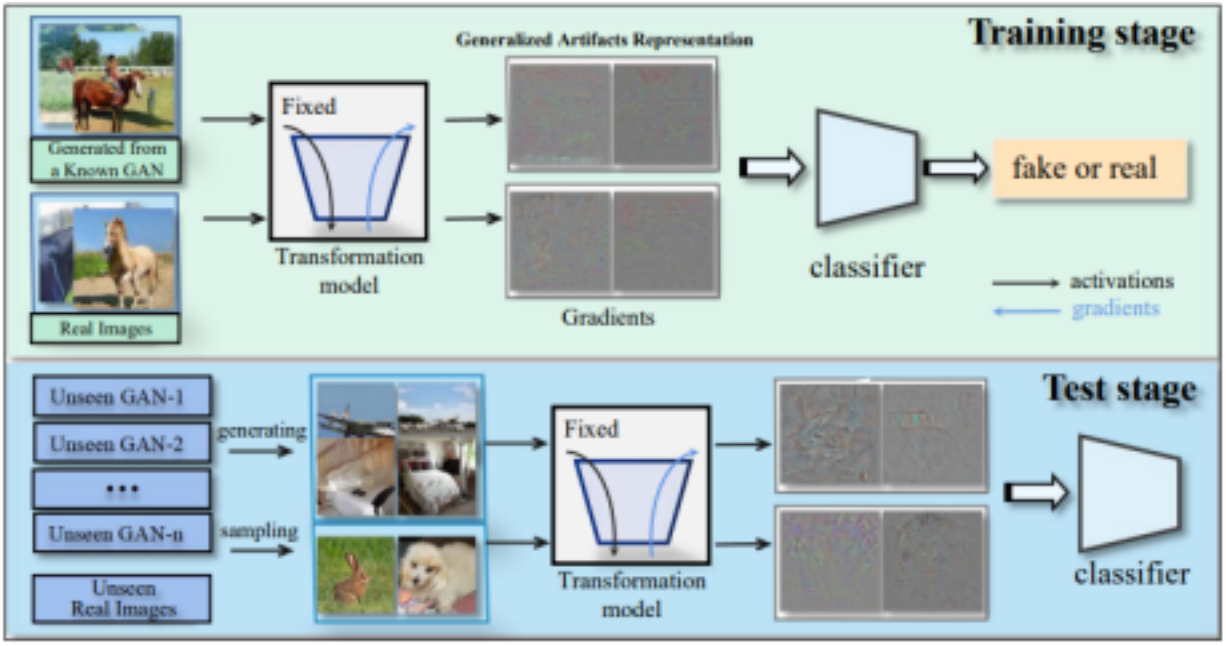
And the gradients are the sum of the all the l.

$$G = \frac{\partial sum(l)}{\partial I_i}.$$

Popular CNN models to implement the transformation model have been used in the research that consist of discriminator of GAN, inversion, contrastive learning model, classification model, and segmentation model. In order to detect fake images and train the binary classification network, up to 255 normalized gradients were used.

General information

- New detection framework - LGrad - Learning on gradients
- Pre-trained CNN model is used as a transformation model to convert images into gradients. Gradients are fed into a classifier to determine the authenticity of an image. ● GAN stands for Generative Adversarial Network
- Other models have led to failure because they depend on the pre-trained data set. ● A generalized model needs to avoid overfitting of data, and furthermore, failure of detection
- The gradients of a trained CNN model can highlight important pixels that can detect fake images



Other models discussed:

- Image-based detection - using spatial information such as colors and global texture. The model relies highly on its training data, resulting in failure with unseen images. ● Frequency-based detection - GAN architecture depends on upscaling, and generated images have a “frequency” compared to images that are not fake. However, GAN models can vary in the output of “frequency” patterns which is why this model does not work effectively.

What is LGrad?

We want to use a labeled data set to produce a general set of gradients that can detect a variety of GAN generated images.

1. Prepare a labeled data set with 1 as real, and 0 as fake. Dataset used in the article included images from several different GAN generated face images.
2. Train a model
    a. Transform images into gradients using a CNN. Reduce the dimensions through the pooling step, highlighting the essential pixels. Use normalized gradients between 0 and 255 to train a binary classification network (classifying between real and fake).
        i. Article used ResNet50 - CNN with 50 layers and compared this with several other CNNs, such as VGG16, InceptionV3, etc.
    b. Use a classifier to classify the data - article used ImageNet classifier.
    c. Use a cross-entropy loss function to optimize the model for the highest accuracy

Model evaluation

The article conducted evaluations in four different settings: cross-model images, cross-category images, cross-model & category images, and perturbed images.

Trans. Models	ProGAN-CelebaHQ		StyleGAN-CelebaHQ		StyleGAN2-CelebaHQ	
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
Input Image	99.6	100.0	76.8	98.9	64.2	96.2
ProGAN-RandomInit	91.4	97.3	57.4	66.3	52.4	56.9
VGG16	96.9	99.8	79.6	96.3	68.2	91.6
CLIP-RN50	99.5	100.0	99.4	100.0	99.2	100.0
ProGAN-bedroom	98.8	100.0	98.4	99.9	96.5	99.6
ProGAN-bridge	96.4	99.5	84.8	95.0	81.9	93.6

Test-categorys	CLIP Trans.		Bedroom Trans.		Bridge Trans.	
	ACC	AP	ACC	AP	ACC	AP
airplane	92.9	97.0	99.4	100.0	99.4	100.0
bicycle	80.5	94.6	94.5	98.5	94.8	98.5
bird	85.6	94.4	99.2	100.0	97.5	99.7
boat	90.6	96.6	99.6	100.0	98.5	99.9
bottle	94.7	98.2	97.7	100.0	99.2	100.0
bus	81.5	94.9	98.1	100.0	96.7	99.8
car	84.0	95.7	99.7	100.0	99.1	100.0
cat	90.2	96.1	99.7	100.0	98.7	99.9
chair	94.5	98.5	96.2	99.9	99.1	100.0
cow	85.5	95.0	99.5	100.0	97.7	99.8
diningtable	92.4	97.7	98.5	99.8	96.9	99.5
dog	89.8	96.1	99.6	100.0	98.4	99.9
horse	89.0	96.3	100.0	100.0	99.8	100.0
motorbike	79.2	94.8	97.0	99.7	96.9	99.3
person	93.6	97.1	98.9	100.0	99.1	100.0
pottedplant	77.9	95.6	97.5	99.7	95.7	99.1
sheep	83.2	94.2	99.2	100.0	95.8	99.3
sofa	94.1	98.6	99.1	100.0	99.2	100.0
train	83.6	94.4	94.4	99.7	96.1	99.6
tvmonitor	93.0	98.7	96.5	100.0	99.2	100.0
Mean	87.8	96.2	98.2	99.9	97.9	99.7

Table 2. Cross-category Performance on the ProGAN models trained on different LSUN [43] object datasets.

- 25,000 real and fake face images used as evaluation
- Cross-model (unseen data). The results show that the model eliminates the content of the image but keeps the discriminative (classification of fake/real) regions in the gradients. The highest accuracy was 99.4% achieved by ResNet50.
- Cross-category (data across several different classes) - used 18,000 fake and real images to train. These models performed well with the highest accuracy being 98.2%. The article compares LGrad to other models. As shown from the data, the accuracy is much higher compared to others.

Methods	Settings		Test Models																	
	Input	#class	ProGAN		StyleGAN		StyleGAN2		BigGAN		CycleGAN		StarGAN		GauGAN		Deepfake		Mean	
			Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
Wang [42]	Image	1	50.4	63.8	50.4	79.3	68.2	94.7	50.2	61.3	50.0	52.9	50.0	48.2	50.3	67.6	50.1	51.5	52.5	64.9
Frank [12]	Freq	1	78.9	77.9	69.4	64.8	67.4	64.0	62.3	58.6	67.4	65.4	60.5	59.5	67.5	69.1	52.4	47.3	65.7	63.3
Durall [11]	Freq	1	85.1	79.5	59.2	55.2	70.4	63.8	57.0	53.9	66.7	61.4	99.8	99.6	58.7	54.8	53.0	51.9	68.7	65.0
BiHPF [18]	Freq	1	82.5	81.4	68.0	62.8	68.8	63.6	67.0	62.5	75.5	74.2	90.1	90.1	73.6	92.1	51.6	49.9	72.1	72.1
FrePGAN [19]	Image	1	95.5	99.4	80.6	90.6	77.4	93.0	63.5	60.5	59.4	59.9	99.6	100.0	53.0	49.1	70.4	81.5	74.9	79.3
LGrad (ProGAN-bedroom)	Grad	1	98.4	99.9	82.6	95.6	83.3	98.4	76.2	81.8	82.3	90.6	99.7	100.0	71.7	75.0	52.8	57.8	80.9	87.4
LGrad (StyleGAN-bedroom)	Grad	1	99.4	99.9	96.0	99.6	93.8	99.4	79.5	88.9	84.7	94.4	99.5	100.0	70.9	81.8	66.7	77.9	**86.3**	**92.7**
Wang [42]	Image	2	64.6	92.7	52.8	82.8	75.7	96.6	51.6	70.5	58.6	81.5	51.2	74.3	53.6	86.6	50.6	51.5	57.3	79.6
Frank [12]	Freq	2	85.7	81.3	73.1	68.5	75.0	70.9	76.9	70.8	86.5	80.8	85.0	77.0	67.3	65.3	50.1	55.3	75.0	71.2
Durall [11]	Freq	2	79.0	73.9	63.6	58.8	67.3	62.1	69.5	62.9	65.4	60.8	99.4	99.4	67.0	63.0	50.5	50.2	70.2	66.4
BiHPF [18]	Freq	2	87.4	87.4	71.6	74.1	77.0	81.1	82.6	80.6	86.0	86.6	93.8	80.8	75.3	88.2	53.7	54.0	78.4	79.1
FrePGAN	Image	2	99.0	99.9	80.8	92.0	72.2	94.0	66.0	61.8	69.1	70.3	98.5	100.0	53.1	51.0	62.2	80.6	75.1	81.2
LGrad (ProGAN-bedroom)	Grad	2	99.5	100.0	85.8	99.3	83.5	99.4	78.9	87.7	78.8	89.0	99.6	100.0	70.5	77.6	51.9	52.7	81.1	88.2
LGrad (StyleGAN-bedroom)	Grad	2	99.8	100.0	94.8	99.7	92.4	99.6	82.5	92.4	85.9	94.7	99.7	99.9	73.7	83.2	60.6	67.8	**86.2**	**92.2**
Wang [42]	Image	4	91.4	99.4	63.8	91.4	76.4	97.5	52.9	73.3	72.7	88.6	63.8	90.8	63.9	92.2	51.7	62.3	67.1	86.9
Frank [12]	Freq	4	90.3	85.2	74.5	72.0	73.1	71.4	88.7	86.0	75.5	71.2	99.5	99.5	69.2	77.4	60.7	49.1	78.9	76.5
Durall [11]	Freq	4	81.1	74.4	54.4	52.6	66.8	62.0	60.1	56.3	69.0	64.0	98.1	98.1	61.9	57.4	50.2	50.0	67.7	64.4
BiHPF [18]	Freq	4	90.7	86.2	76.9	75.1	76.2	74.7	84.9	81.7	81.9	78.9	94.4	94.4	69.5	78.1	54.4	54.6	78.6	77.9
FrePGAN [19]	Image	4	99.0	99.9	80.7	89.6	84.1	98.6	69.2	71.1	71.1	74.4	99.9	100.0	60.3	71.7	70.9	91.9	79.4	87.2
LGrad (ProGAN-bedroom)	Grad	4	99.7	100.0	87.8	99.1	91.7	99.7	80.9	89.3	78.2	89.0	99.8	100.0	73.5	78.6	53.1	55.0	83.1	88.8
LGrad (StyleGAN-bedroom)	Grad	4	99.9	100.0	94.8	99.9	96.0	99.9	82.9	90.7	85.3	94.0	99.6	100.0	72.4	79.3	58.0	67.9	**86.1**	**91.5**

The article further uses image perturbations such as blur, cropping, noise, and jpeg files to test the model. All cases except jpeg and noise receive high accuracy.

Perturbed	ProGAN		StyleGAN		StyleGAN2		BigGAN		CycleGAN		StarGAN		GauGAN		Deepfake		Mean	
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
No	99.4	99.9	96.0	99.6	93.8	99.4	79.5	88.9	84.7	94.4	99.5	100.0	70.9	81.8	66.7	77.9	86.3	92.7
blur	90.1	96.5	90.6	95.7	87.5	93.4	71.1	76.3	70.6	73.6	93.6	98.0	63.7	69.5	61.2	62.6	78.5	83.2
cropping	99.2	99.9	95.9	99.6	94.0	99.6	80.1	88.0	78.5	88.1	94.4	99.4	70.6	78.1	67.5	82.3	85.0	91.9
jpeg	76.2	90.0	74.4	90.2	72.6	89.1	66.0	74.6	72.7	83.2	76.0	89.5	60.1	67.6	58.0	65.2	69.5	81.2
noise	77.1	87.7	73.3	84.8	74.3	84.9	68.2	77.4	66.4	77.2	76.0	88.8	60.4	69.1	57.5	65.1	69.1	79.4
combined	86.3	94.6	83.5	93.0	82.4	92.3	71.2	78.9	73.0	80.4	84.7	94.3	64.7	71.5	61.2	67.4	75.9	84.1

Effect of Models and Training Data on the Study

Instead of depending on data, the transformation model was used to improve accuracy. Images were converted into a gradient.

Trans. Model	ProGAN		StyleGAN		StyleGAN2		BigGAN		CycleGAN		StarGAN		GauGAN		Deepfake		Mean	
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
VGG16 [39]	96.0	99.5	65.5	88.8	74.7	93.8	73.0	78.1	77.8	86.1	99.8	100.0	60.7	63.1	60.4	67.8	76.0	84.7
InceptionV3 [41]	64.9	74.4	58.8	66.9	65.4	73.8	50.9	52.1	59.1	68.4	52.6	58.5	54.0	56.5	49.8	50.1	56.9	62.6
Resnet50 [15]	86.4	95.0	81.0	92.2	83.7	93.5	57.2	56.2	68.3	75.8	96.4	99.5	51.2	52.7	63.4	70.4	73.4	79.4
CLIP-Resnet50 [35]	87.6	95.8	80.2	90.0	78.9	91.0	60.1	61.1	84.2	87.9	88.5	95.1	72.6	71.6	64.9	64.4	77.1	82.1
ViT [10]	51.2	71.1	51.7	65.7	52.4	67.9	50.6	53.3	54.1	75.2	51.6	64.0	50.6	61.1	50.0	53.7	51.5	64.0
DeeplabV3 [5]	81.6	91.6	68.7	80.4	70.6	84.5	54.5	55.7	66.2	71.0	87.9	94.7	51.7	53.1	59.3	58.9	67.6	73.7
Idinvert [48]	97.4	99.8	71.6	95.3	71.2	95.4	86.6	94.8	78.7	85.7	97.4	99.7	72.0	82.1	60.1	72.1	79.4	90.6
ProGAN-bedroom [21]	98.4	99.9	82.6	95.6	83.3	98.4	76.2	81.8	82.3	90.6	99.7	100.0	71.7	75.0	52.8	57.8	80.9	87.4
ProGAN-bridge [21]	97.8	99.7	86.4	97.5	85.7	97.3	72.5	78.7	76.8	87.5	94.1	99.9	62.5	75.8	53.2	61.3	78.6	87.2
StyleGAN-bedroom [22]	99.4	99.9	96.0	99.6	93.8	99.4	79.5	88.9	84.7	94.4	99.5	100.0	70.9	81.8	66.7	77.9	**86.3**	**92.7**
StyleGAN-cats [22]	97.4	99.7	83.4	97.3	77.4	96.4	69.8	74.6	79.3	90.2	97.8	99.8	68.0	77.4	65.9	72.9	79.9	88.5
StyleGAN2-church [23]	99.1	100.0	88.2	97.7	91.9	99.6	70.1	71.7	80.6	89.1	95.6	99.8	60.8	68.9	72.7	76.5	82.4	87.9

Table 5. Performance of different transformation models.

As a result, the quality of the transformation will effect the detector performance. This is shown in table 5. Form this table, the performance of the gradients that were obtained from GAN discriminator are better on unseen data unlike other models such as classification or contrastive learning models.
As a result, “StyleGAN-bedroom obtains the best performance”, according to the research.

Methods	Num. of Train. data	ProGAN		StyleGAN		StyleGAN2		BigGAN		CycleGAN		StarGAN		GauGAN		Deepfake		Mean	
		Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
FrePGAN [19]	36k	95.5	99.4	80.6	90.6	77.4	93.0	63.5	60.5	59.4	59.9	99.6	100.0	53.0	49.1	70.4	81.5	74.9	79.3
LGrad	4k	95.9	99.5	88.9	98.0	91.7	99.0	59.1	55.2	58.7	58.4	87.8	99.8	59.2	66.2	65.9	83.5	75.9	82.4
LGrad	9k	98.3	99.9	92.4	99.4	92.2	99.4	71.7	74.5	73.0	76.8	98.1	100.0	64.0	74.2	68.0	79.3	82.2	87.9
LGrad	18k	98.8	100.0	95.1	99.8	91.4	99.7	76.8	87.1	79.8	90.7	99.2	100.0	68.9	80.4	63.9	73.0	84.2	91.3
LGrad	36k	99.4	99.9	96.0	99.6	93.8	99.4	79.5	88.9	84.7	94.4	99.5	100.0	70.9	81.8	66.7	77.9	**86.3**	**92.7**
LGrad	72k	99.8	100.0	94.8	99.7	92.4	99.6	82.5	92.4	85.9	94.7	99.7	99.9	73.7	83.2	60.6	67.8	86.2	92.2

Table 6. Results with variance in number of training data.

Similarly, to test the training data amount, 4,000, 18,000, and 36,000 images were used. It is important to note that equal numbers of real and fake images were used in these three trials and the discriminator of StyleGAN-bedroom is used in the process. The detector achieves similar data performance with 18,000, 36,000, and 72,000 training images; the results indicate that the LGrad 4,000 training data performs better than the FrePGAN with 36,000 images used as training data.

Conclusion

The article concludes also that the amount of data used in training does not heavily impact the LGrad model either. The model also acknowledges that it does not perform well in non-GAN models such as deepfake.

References

[1]C. Tan, Y. Zhao, S. Wei, G. Gu, and Y. Wei, ‘Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12105–12114.