Student Occasional Paper Series No. 4 | September 2023

Understanding Generative Adversarial Networks

By Peyton Ludwig | Edited by Aaron Brantly

Abstract

Generative adversarial networks (GANs) are a new technology impacting social, policy, and security discourses. The rise of GANs enables the creation of artificially generated hyper-realistic human faces. GANs have been around since 2014,[1] yet only recently has their quality risen to a level capable of fooling the average person. **StyleGANs** are GANs trained to manipulate or generate high-quality images.[2] StyleGANs are becoming increasingly publicly accessible and enable users to generate human faces with ease. The styleGAN program is open access;[2] while this allows for usage in positive ways, it also leads to easy accessibility for those that want to use this technology for malicious purposes. Fake accounts with AI-generated faces as their profile pictures plague social media sites and are often used as tools for misinformation. As styleGANs improve, it becomes more difficult to spot these fake faces. It is important to understand how these styleGANs work in order to best combat these disinformation attempts and understand what to do moving forward.

GANs

Understanding how styleGANs work is a key step to understanding how to identify an AI-generated image and how to counter fake profiles. To explain this process, it can be broken down into its most basic form: the original GAN structure.

A GAN consists of two neural networks competing against each other: the **generator** and the **discriminator**.[1] *As their names suggest, the generator's role is to create images, and the discriminator's role is to determine if images given to it are real or fake.* The generator attempts to fool the discriminator into thinking the images it generates are real. The process involves training both the discriminator and the generator in an iterative process. As training progresses the discriminator finds it increasingly difficult to detect the generator-created images. Once the generator adapts and improves in quality to the point at which the discriminator is unable to differentiate between real and generated images, the image has a real and natural quality.

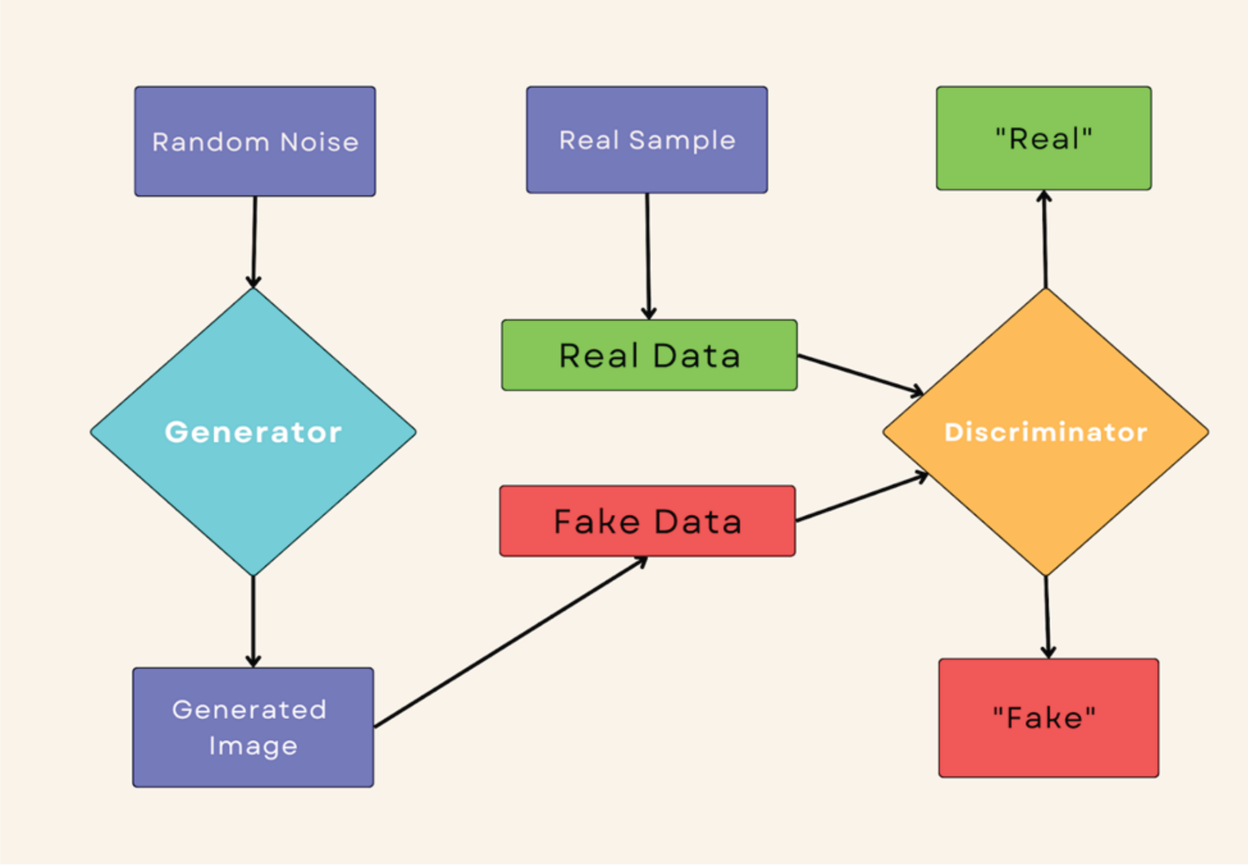

Figure 1: Diagram showing the generator's input of random noise and output of a generated image, and discriminator's input of real and fake data and output of labeling it as real or fake.

How GANs are Trained

The GAN training process begins with the training of the discriminator. Discriminators are trained from two sources: real data used as positive examples, and generated data used as negative examples.[3] Positive examples include items such as images of real people, whereas negative examples include fake data created by the generator. During training, the GAN goes through a two-pass cycle. In the first pass, known as the forward pass, the discriminator classifies the data given from the generator as real or fake. The discriminator is penalized by its **loss**, or a measure (probability of accuracy) of how far its prediction is from the correct label of real or fake.[3] The second pass, known as the backward pass, constitutes the discriminator adjusting accordingly through **backpropagation**.[4] Backpropagation is the process by which a system reduces its loss through adjusting the **weights**, or the values comprising the weighting system, to improve image fidelity.[4] In essence, the discriminator is progressively tweaked through an iterative process which penalizes it for making mistakes. The negative weighting alters the values associated with specific image attributes over time like a pendulum seeking its equilibrium point.

Once the discriminator has finished training, it remains unchanged (unless new training data is provided) as the generator begins its training. It begins with the generator sampling random noise as its input. From this random noise, it generates an image as an output. This output is then judged by the discriminator as real or fake. The generator is penalized if it fails to get a "real" label from the discriminator. Similar to the discriminator, the generator then changes its weighting of image attributes depending on its success in fooling the discriminator.[5]

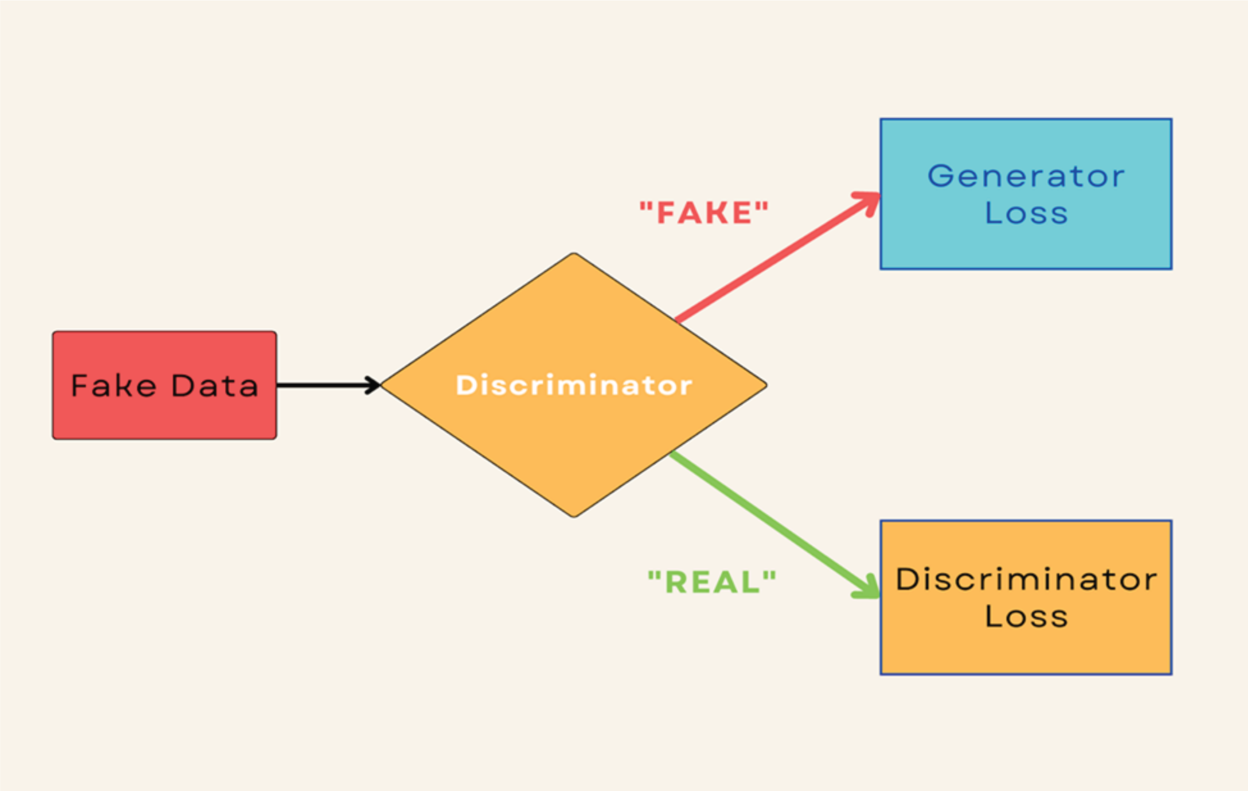

Figure 2: Diagram showing how generators and discriminators are penalized. If fake data is labeled as fake, it's a generator loss; if fake data is labeled as real, it's a discriminator loss.

GAN training alternates between training the generator and discriminator. The discriminator trains for one or more **epochs** (number of passes of a training dataset), then the generator, and this process repeats. As the discriminator trains, the generator remains constant, and vice versa. As the generator improves, the accuracy of the discriminator decreases. If the generator fully succeeds, the discriminator has a 50% accuracy, the same as the discriminator flipping a coin to decide whether an output is real or fake.[5]

To continue training the generator, in many instances the discriminator is made stronger by changing the program architecture. This can be done through different techniques, such as using multiple discriminators for a diversification of feedback,[6] multiresolution discrimination to judge at different levels of resolution,[7] or self-attention so inputs can interact with each other to determine where they should put their attention.[8] These techniques increase the relative level of difficulty for the generator to fool the discriminator, which leads to the generator becoming increasingly effective.

GANs use a system of adversarial networks to develop an iterative and evolving system that strives to improve its methods of developing images. This technique enables StyleGANs to generate highly accurate human faces that improve over time and become increasingly difficult to discern from real images.

Why We Should Care

Because of this GAN architecture, these styleGANs have improved rapidly over time. While many of these faces are still detectable because of certain giveaways, such as odd shapes, clear inconsistencies, or abstract backgrounds, these flaws could very well be eliminated as this technology continues to improve.

It is important to understand how these programs work on a base level to further understand how to break them down and spot styleGAN-created images. In comprehending how these images are created, it is significantly easier to find where its weak points in generation lie, which gives us a greater ability to detect these generated images.

The Future: StyleGANs

Developed by NVIDIA in December of 2018, **styleGAN** builds upon and modifies the traditional GAN structure.[9] Alongside this, it borrows from **style transfer literature**, or changing the style of an image in one domain to the style of an image in another.[10] In particular, styleGAN specializes in **adaptive instance normalization**, which aligns the style features with the mean and variance of the content features.[11] Through these GAN modifications and use of style transfer literature, styleGAN can generate high-fidelity human faces.

By themselves, GANs operate as **black boxes**, or a process where the program being executed is not well examined.[12] In the case of GANs, the source of randomness in generating images is not well understood, and there is a lack of control over the images it generates. There are few tools to control style properties such as the background, foreground, and features. This is where styleGAN comes in.

StyleGAN serves as an extension to GAN architecture and gives control of these **disentangled style properties**, which allows the program to have control over different properties of the image.[13] It has control over these style properties for every layer, so the strength of features can be controlled at different scales.

It begins with the baseline progressive GAN.[13] The GAN is trained first by small images, such as 4 pixels by 4 pixels. The GAN stabilizes, adds new blocks of layers that can take in larger images, and is trained again. This process repeats, gradually growing in size by doubling the width and height of the image. This process repeats until it hits the target image size, such as 1024 pixels by 1024 pixels.[13]

From there, the styleGAN architecture has many modifications from the traditional GAN structure. The first of these is tuning and bilinear upsampling. After an image is generated, the

activations, or the output values, are filtered using a **second-order binomial filter**.[13] This is a type of smoothing filter that removes noise and other unwanted artifacts while maintaining its sharpness and details.[14] Through doing this, styleGAN can produce images that are high quality alongside eliminating the noise that comes with the generative process.

The second modification is the addition of a mapping network and adaptive instance normalization. Instead of the generator directly taking in latent space as its input, the latent space is passed through 8 connected layers, resulting in the output of the **style vector,** which captures high-level features such as color and texture. This vector is incorporated into each block of the synthesis network through using adaptive instance normalization, or **AdaIN**. This technique normalizes the mean and variance of the activations of each layer to match the style vector.[13] Through AdAIN, styleGAN has more flexibility and control over the style of the images, leading to images with a wide range of styles and characteristics.

The third modification is the removal of the traditional input of latent space to the generator. Instead, a fixed value is used as input. When starting image synthesis, the model has a 3 x 4 x 512 constant.[15] By using a constant instead of a random point in the latent space, styleGAN has more control over aspects of the generated images.

The fourth modification is the addition of noise to each block. Gaussian noise is fed to each layer of the generator, and a different sample of noise is generated for each block. This noise is broadcast to all feature maps using per-layer scaling factors.[15] By adding this noise, styleGAN introduces random variations that lead to a diverse number of images.

The fifth modification is the addition of mixing regularization, which prevents the network from assuming adjacent styles are related. Two random latent space nodes are trained instead of one, which means the mapping network generates two style vectors. These vectors are then used to control different aspects of the image. To use these vectors, styleGAN chooses a split point in the synthesis network, separating the network into two halves. The first half uses the first style vector to generate the initial structure of the image, while the second half uses the second style vector to control the finer details and texture of the image.[15]

StyleGAN Usage

Training and Generation Requirements

To train a styleGAN model, powerful GPU is required; with v2 of the program, it is recommended to have at least 16 GB of memory.[16] Depending on the dataset size and complexity of the model, the training of a model from scratch can take days to weeks. For example, training for the output of 128 x 128 images with 1 GPU takes 4 days, whereas training for the output of 1024 x 1024 images with 1 GPU takes 46 days. This rate can be increased with the addition of more GPUs and higher processing power.[17] The original styleGAN trained for one week using NVIDIA DGX-1 with 8 Tesla V100 GPUs.[15]

Legitimate Usage

Companies are beginning to see styleGAN generations as lucrative opportunities. Some examples of legitimate use cases are below:

Icons8 is a design marketplace that has begun utilizing AI-generated photos to offer to companies to use as models, with clients paying to download 10,000 photos a month. Their clients have included a dating app, an American university, and a human-resources planning firm. While the company photographed models to train the styleGAN themselves, none of these models will be paid any residuals from the AI-generated images trained on their likenesses.[18]

Similarly, styleGAN created faces have also begun being utilized in the game development world. The site Generated Photos offers "worry-free" and "diverse" models by using generated human faces and bodies,[19] and these include partnerships with several games and game creation software. Headshot, a plugin to Reallusion's Character Creator that converts a 2D image to a 3D model, has begun to use AI to generate a 2D photo to convert into a 3D face. This allows a 3D model to be created without it having a likeness to any real person.[19] Another instance of this is the game *Beyond Humanity: Colonies,* which uses a neural network to simulate thousands of virtual citizens, it uses AI-generated faces to represent these citizens.[19]

While these are legitimate uses of styleGAN, it does raise its own complications about the ethics of using technology in place of human models.

Illegitimate Usage

*T*here are many scenarios where these generated images have been used illegitimately. Due to the easy accessibility of styleGAN, many have used these fake faces as ways to spread disinformation, especially as profile pictures for fake social media accounts. Since styleGAN can generate a face that does not exist, using the face as a profile picture cannot be traced back to a stock image or real person that would easily indicate the profile as being fake.

Instances of styleGAN-created images can be seen in fake profiles spreading Russian propaganda during the Ukraine conflict. It was reported that Twitter removed over a dozen accounts and Facebook removed 40 accounts that were associated with spreading propaganda.[20] Those with AI-generated faces were almost indistinguishable from normal profiles, making them harder to identify.

Alongside this, this technology is also beginning to be used in espionage operations. Espionage efforts run rampant on the networking platform LinkedIn; officials across several countries have issued warnings of how thousands have been contacted by foreign spies over this site.[21] One of these fake profiles was recently identified as having an AI-generated profile picture,[21]and this phenomenon is likely to grow over time.

Conclusion

StyleGANs build upon the iterative and evolving process of GANs by using modifications to utilize style transfer literature. Through these modifications, hyper realistic images, especially of human faces, can be generated. The program is open access, and while this technology has many beneficial and legitimate applications, its easy accessibility leads to its common use for malicious purposes. In comprehending how these GANs and styleGANs function, we have a significantly better method of understanding the best way to identify these generated faces.

Endnotes

[1] Goodfellow et al., “Generative Adversarial Nets,” 2014.

[2] Karras, Tero, Samuli Laine, and Timo Aila. “A Style-Based Generator Architecture for Generative Adversarial Networks.” *arXiv*, 2018.Click here to enter text.

[3] Goodfellow et al., “Generative Adversarial Nets,” 2014.

[4] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. “Learning Representations by Back-Propagating Errors.” Nature 323, no. 6088 (1986): 533–36. doi:10.1038/323533a0.

[5] Goodfellow et al., “Generative Adversarial Nets,” 2014.

[6] Xu, Depeng, et. al. “FairGAN: Fairness-Aware Generative Adversarial Networks.” arXiv, 2018. doi:10.48550/arxiv.1805.11202.

[7] You, Jaeseong, et. al. “GAN Vocoder: Multi-Resolution Discriminator Is All You Need.” arXiv, 2021. doi:10.48550/arxiv.2103.05236.

[8] Zhang, Han, et. al. “Self-Attention Generative Adversarial Networks.” arXiv, 2018. doi:10.48550/arxiv.1805.08318.

[9] Karras, Tero, Samuli Laine, and Timo Aila. “A Style-Based Generator Architecture for Generative Adversarial Networks.” *arXiv*, 2018.

[10] Jing, Yongcheng, et. al. “Neural Style Transfer: A Review.” arXiv, 2017. doi:10.48550/arxiv.1705.04058.

[11] Huang, Xun, and Serge Belongie. “Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization.” arXiv, 2017. doi:10.48550/arxiv.1703.06868.

[12] Petch, Jeremy, Shuang Di, and Walter Nelson. “Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology.” Canadian Journal of Cardiology 38, no. 2 (2022): 204–13. doi:10.1016/j.cjca.2021.09.004.

[13] Karras, Tero, Samuli Laine, and Timo Aila. “A Style-Based Generator Architecture for Generative Adversarial Networks.” *arXiv*, 2018.

[14] Derpanis, Konstantinos G. “Overview of Binomial Filters.” York University, 2005.

[15] Karras, Tero, Samuli Laine, and Timo Aila. “A Style-Based Generator Architecture for Generative Adversarial Networks.” *arXiv*, 2018.

[16] Martinelli, Alex. “StyleGAN v2: Notes on Training and Latent Space Exploration.” Medium, 2020.

[17] “StyleGAN2-ADA — Official PyTorch Implementation.” *GitHub*, 2021.

[18] Harwell, Drew. “Dating Apps Need Women. Advertisers Need Diversity. AI Companies Offer a Solution: Fake People.” The Washington Post, 2020.

[19] “How Generated Photos Are Used.” Generated Photos, n.d.

[20] Rauwerda, Annie. “Fake Social Media Users with AI- Generated Photos Are Spreading Lies for Russia.” Inverse, 2022.

[21] Satter, Raphael. “Experts: Spy Used AI-Generated Face to Connect with Targets.” AP News, 2019.