Documenting the Human Ability to Correctly Identify Fake Facial Images Produced by Generative Adversarial Networks

By Booker Martin

Mentor: Dr. Andrew Forney

December 8 2019

Abstract:

Generative Adversarial Networks, or GANs, build upon the foundation of machine learning by introducing an "adversarial" network that contrasts sample data with generated data, pushing the generative network to yield realistic results. With NVIDIA's addition of open-source "style transfer" technology, programs that generate realistic facial images are being made that are accessible to anyone. This new technology comes real-world consequences, such as the ability to abuse these generated facial images online through fake social media accounts. Yet, there is little research that focuses on the human ability to determine whether an image is real or generated. This can be documented through a website that identifies how accurately visitors can distinguish between generated and real images and why they are able to spot fakes. In addition, the website will showcase this data, providing new insight on our ability to spot fake facial images as well as the most common flaws of GAN-generated facial images.

Research Proposal: Narrative

1. Introduction

In the past decade, great advancements have been made in the field of machine learning. Consequently, there has been a shift towards heavy use of machine learning in many different aspects of our lives: self-driving cars, smartphone assistants, online shopping recommendations, facial image recognition, etc. All of these programs share something in common: they "learn" from the data that they are exposed to in order to become better at their target objective. Generative Adversarial Networks, or GANs, build upon the concept of machine learning by introducing an adversarial network that contrasts generated data and real data. According to Ian Goodfellow, one of the researchers who introduced GAN technology, the addition of the adversarial network pushes the generator network to produce data that is nearly "indistinguishable" from the sample data. Consequently, the rise of GANs is ushering in an age where artificially generated content is more convincing than ever, and this poses certain repercussions. In particular, facial image generation technology using GANs has become so realistic that we must answer the following questions: How accurately can people identify whether an image of a face is randomly generated or real? What are some common flaws that people spot in generated images that reveal they are fake?

2. Background/Related Work and Motivation

Though GANs build upon decades of concepts and advancements in the field of artificial intelligence, they are a relatively new type of program. GANs were first detailed in a 2014 paper written by researchers including Ian Goodfellow. GANs consist of a generator network and a

discriminator network (Goodfellow). The job of the generator network is to generate accurate data samples (Goodfellow). These samples are then determined to be real or fake by the discriminator network, the generator network's "adversary" (Goodfellow). The generator network becomes better at generating realistic samples based upon which ones "trick" the discriminator network (Goodfellow). In addition, the discriminator network analyzes authentic samples to become more accurate in spotting fakes (Goodfellow). Each model depends on the other in a cycle of growth and feedback.

After Goodfellow introduced GAN technology, several major tech companies have contributed by implementing new concepts into traditional GAN programs. In December 2018, NVIDIA published the paper *A Style-Based Generator Architecture for Generative Adversarial Networks* which details a technique for GANs to independently distinguish different attributes of images. The technology "borrows from style transfer literature" to allow for greater control over generated images (Karras). For example, the method allows for the modification of specific facial features such as the eyes, face shape, or hairstyle without impacting the rest of the face (Karras).

In February 2019, NVIDIA published their "style-based generator architecture" under the name "StyleGAN" on GitHub, allowing for public use and access. In addition, NVIDIA has uploaded their dataset of 70,000 images of human faces taken from Flickr. The dataset was used as a benchmark for their GAN technology — it provided the "real" pictures of faces that were contrasted against generated images by the discriminator network in order to improve the generator.

Making StyleGAN open-source has allowed the generator architecture to be implemented on websites such as thispersondoesnotexist.com, which uses the architecture to produce a random image of a face. Consequently, anyone capable of browsing the internet has access to a new, unique generated image of a realistic face. While this demonstrates the power of GAN technology, it also makes it easier to assume a false identity online using a photo that cannot be traced to an existing person. There have been documented instances of fake accounts on websites such as LinkedIn that utilize generated images as profile pictures (Satter). This suggests that the convincing generated images resulting from GAN technology have real consequences, such as contributing to more convincing online bots that may steal information or spread propaganda.

The unethical use of realistic generated facial images, such as fake accounts, illustrates the importance of being able to recognize fake images. There are currently websites such as whichfaceisreal.com that allow users to test whether they can distinguish these randomly generated faces from real photos as while as provide techniques to better spot fakes. However, despite the likelihood that an increasing number of people may encounter fake facial images online every day, there is currently little documentation of how accurately humans can identify whether an image of a face is real or fake. This incentivizes further research on the human capability to correctly identify fake images.

3. Methods

While there are currently websites that use NVIDIA's StyleGAN to introduce GAN technology to a wider audience and demonstrate how realistic image generation has become, more documentation can be done on the human capability to correctly identify fake images. The

next step is to collect data on how accurately people can determine whether an image of a face is randomly generated or authentic and what specific facial features this image has. This will be accomplished by the creation of a new website with two primary goals: to test visitors' ability to correctly identify photos as real and fake and to record and display the data.

The first page of the website will function as the "data collection" page. Visitors will be presented with either an authentic photo or a generated image, chosen at random. Real pictures of people will be chosen from NVIDIA's published dataset of 70,000 photos. Fake images generated by NVIDIA's StyleGAN will be pulled from thispersondoesnotexist.com. Visitors will be prompted to choose whether the image is real or fake. If the visitor correctly identifies a fake image, they will be asked how they knew the image was fake: they will be presented with a checklist of common issues with generated facial images to choose from, including "water-splotches," background problems, unrealistic hair, unrealistic teeth, fluorescent bleed, asymmetrical features, and an option to write any unlisted reasons (West). The data will be saved to the website's database, all of which will be viewable on the results page.

Once the website has been created, steps must be taken to drive traffic to the website in order to collect data. This may be done in the form of emails to students in relative departments, links to the website posted online, posters and pamphlets being distributed on LMU campus, and possibly a small ad campaign to target a wider audience.

4. Expected Results

The website will function as both the method of collecting the data and the method of showcasing the data. The second page of the website will function as the "data exhibition" page;

this page will provide a live update of the current data that has been gathered. The page will display what percentage of all real and generated images were correctly identified by visitors. In addition, the reasons that people were able to identify fake images will be listed below in order from most to least common to allow for comparison. As data is compiled, it will reveal the general accuracy of visitors' ability to detect fake images. In addition, it will indicate the most significant problems with generated facial images that reduce the probability of a fake image fooling someone. The data can be viewed within various time frames, such as in the past week, month, year, and since the site's creation in order to see different dataset sizes.

5. Conclusion

Generative Adversarial Networks, or GANs, build upon a foundation of machine learning by introducing an "adversarial" network that contrasts sample data with generated data. Since the adversarial network pushes the generative network to create data indistinguishable from the provided samples, the technology succeeds in creating convincing generated content.

With NVIDIA's addition of "style transfer" technology to their open-source GAN-framework, websites and programs that generate realistic facial images are being made that are accessible to anyone. Alongside this new technology comes real-world consequences, such as the ability to abuse these generated facial images online through fake social media accounts. This highlights the importance of the human ability to determine whether an image is a real photo of a face or a product of a GAN. Yet, there is little research that focuses on this.

The creation of a new website will document how accurately visitors can distinguish between generated and real images and why they are able to spot fakes. In addition, the website

organizes and showcases the collected data. Once a sufficient number of visitors participate, the data will provide new insight on our ability to spot fake facial images and the most common flaws of GAN-generated faces.

Works Cited

Bresnick, Jennifer. “Deep Learning for Medical Imaging Fares Poorly on External Data.” *Health IT Analytics*, 7 Nov. 2018, healthitanalytics.com/news/deep-learning-for-medical-imaging-fares-poorly-on-external-data.

Chintapalli, Karthik. “Generative Adversarial Networks for Text Generation - Part 1.” *Medium*, Becoming Human: Artificial Intelligence Magazine, 28 Mar. 2019, becominghuman.ai/generative-adversarial-networks-for-text-generation-part-1-2b886c8cab10.

“GAN Dissection.” *GAN Dissection*, gandissect.csail.mit.edu/.

Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems*. 2014.

Grubb, Jeffrey, director. *Google Duplex: A.I. Assistant Calls Local Businesses To Make Appointments*. *YouTube*, 8 May 2018, www.youtube.com/watch?v=D5VN56jQMWM.

Jain, Anant. “Breaking Neural Networks with Adversarial Attacks.” *Medium*, Towards Data Science, 9 Feb. 2019, towardsdatascience.com/breaking-neural-networks-with-adversarial-attacks-f4290a9a45aa.

Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

“Loebner Prize.” *Wikipedia*, Wikimedia Foundation, 7 Oct. 2019, en.wikipedia.org/wiki/Loebner_Prize.

Lundervold, Arvid, and Alexander Lundervold. “An Overview of Deep Learning in Medical Imaging Focusing on MRI.” *Zeitschrift Für Medizinische Physik*, Urban & Fischer, 13 Dec. 2018, www.sciencedirect.com/science/article/pii/S0939388918301181.

Russell, Stuart J., and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3rd ed., Pearson, 2009.

Satter, Raphael. “Experts: Spy Used AI-Generated Face to Connect with Targets.” AP NEWS. Associated Press, June 13, 2019. https://apnews.com/bc2f19097a4c4fffaa00de6770b8a60d.

“StyleGAN - Official TensorFlow Implementation.” GitHub. NVIDIA, December 3, 2019. https://github.com/NVlabs/stylegan.

Synced. “GAN 2.0: NVIDIA's Hyperrealistic Face Generator.” *Medium*, SyncedReview, 15 Dec. 2018, medium.com/syncedreview/gan-2-0-nvidias-hyperrealistic-face-generator-e3439d33ebaf.

"This Person Does Not Exist." *This Person Does Not Exist*, thispersondoesnotexist.com/.

West, Jevin, and Carl Bergstrom. "Which Face Is Real?" *Which Face Is Real?*, whichfaceisreal.com/.

Wiggers, Kyle. "DeepMind's AI Learns to Generate Realistic Videos by Watching YouTube Clips." *VentureBeat*, VentureBeat, 19 July 2019, venturebeat.com/2019/07/19/deepminds-ai-learns-to-generate-realistic-videos-by-watching-youtube-clips/.

Research Proposal: Budget

The website creation will have certain costs and time-commitments. First, the registration of a relevant domain, such as "fakeorrealfaces.com," will cost around $12 a year according to Google Domains. To ensure the domain name is maintained, the website should be registered for about six years ($72). In addition, the web hosting service will cost around $360 for three years, or $720 for six years, based on the costs of A2Hosting, a competitively-priced web host. The website will likely take two to three months of work.

Once the website has been created, it will be important to drive traffic to the site to attract visitors such that a sufficient amount of data is collected. Active steps to publicize the website will be taken the following two to three months. A budget of $70 will be sufficient for printing posters and pamphlets for the website that can be placed around the LMU campus. In order to significantly increase the number of visitors, however, an online ad campaign may be necessary. According to Google Ads, reaching an estimated 5,260 to 8,810 people would cost around $14 day, or a max of $426 per month.

The total cost, including six years of domain registration and web hosting, printing, and one month of online advertising, is around $1,288.