**GANBot testing datasets**

**Project goal:** Building a system to identify Twitter accounts with GAN-generated human faces.

**Project background:** We noticed that some bad actors started to use fake human faces generated by Style-GAN as the profiles of fake Twitter accounts and we want to detect them.

**Presentation Link:** [GANBots: Detecting Twitter bots with GAN generated human faces.](https://docs.google.com/presentation/d/1gJtLcAjxICr03xUug9VxggC0TYNenyZSLHuabFW7MRQ/edit?usp=sharing)

**Your Job?:** As we talked about in the presentation on (3/6), the position of the eyes looks like a good key feature to engineer and exploit in the images. We’d like you to take a deep learning-based approach (*something that can classify images based on their eye position wrt to the rest of the image*). Once you have a running model for the eye position, you’re more than welcome to explore other features that you think might be more effective or pair well together with the “eye position feature”. The practical goal is simply to identify GAN faces on Twitter (*which happen to be very scarce in nature*). We’re proposing some experiments to perform once we have a running model to emulate a real-world scenario.

**Testing datasets:**

*3 datasets will be provided to run 2 experiments:*

A : 25k+ StyleGAN images

B : 25k+ Random Twitter profile images - **NOTE: may contain some NSFW pictures**

C : 250+ TwitterGAN images - GAN faces that have been already used by bots on Twitter (done)

The data will be shared through: [Datasets Link](https://indiana-my.sharepoint.com/:f:/r/personal/singhdan_iu_edu/Documents/ganbot-Danishjeet%E2%80%99s%20MacBook%20Pro/testing_datasets?csf=1&web=1&e=hzRHwJ)

**Experiments:**

We the following 2 experiments

Experiment 1:

**To test model accuracy**

* A dataset comprised of dataset A (all labeled 1) and dataset B (all labeled 0)
* This is a binary classification task with balanced classes
* Use accuracy (or F1, or AUC) to measure the model performance

Experiment 2:

**To emulate a real-world scenario**

* A dataset comprised of dataset B and dataset C
* This experiment emulates the real-world scenario where accounts with GAN-generated profiles are scarce
* Build a machine learning model that can rank the images based on their likelihood of being generated by GAN, use hits@k (k=100 or 200) to measure the model performance

**Relevant links:**

* [**https://arxiv.org/pdf/1903.06836.pdf**](https://arxiv.org/pdf/1903.06836.pdf)
* [**https://www.semanticscholar.org/reader/533ce1a733a2ff5ba9d33d62577f2007f84abe43**](https://www.semanticscholar.org/reader/533ce1a733a2ff5ba9d33d62577f2007f84abe43)
* [**https://doi.org/10.3390/jimaging7080128**](https://doi.org/10.3390/jimaging7080128)