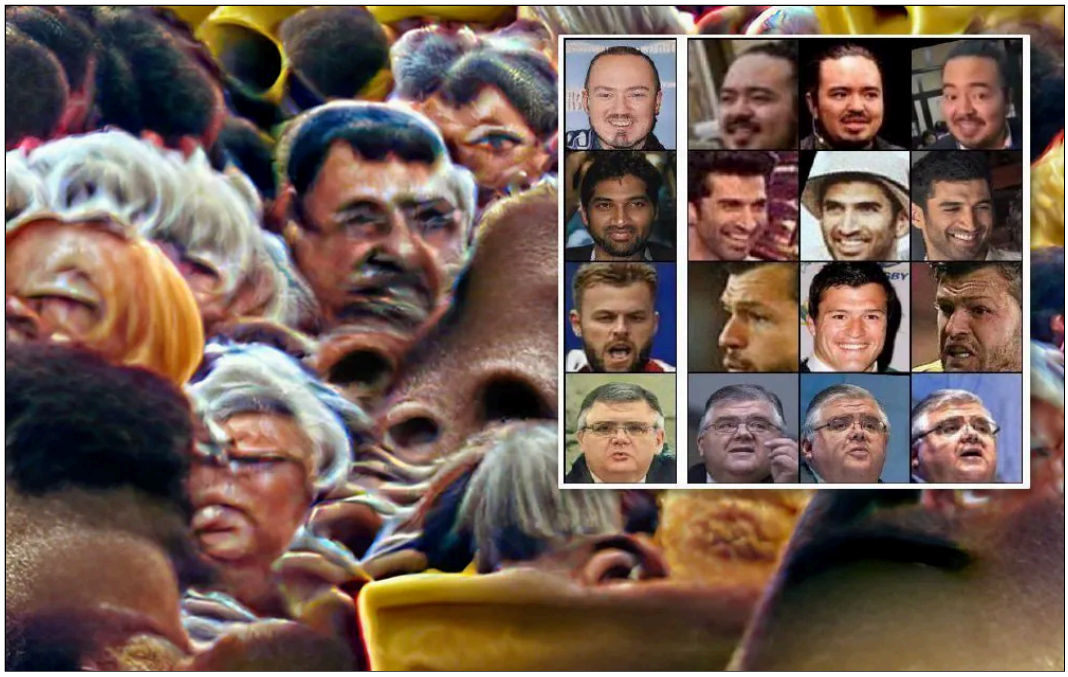
Re-Identifying Source Data for GAN Generators

Originally published at https://www.unite.ai/re-identifying-source-data-for-gan-generators/



New research from France has proposed a technique to 're-identify' source identities that have contributed to synthetically generated data, such as the GAN-generated 'non-existent people' at face-generating projects such as This Person Does Not Exist.

The method outlined in the paper, entitled *This Person (Probably) Exists. Identity Membership Attacks Against GAN Generated Faces*, does not require (unlikely) access to the training architecture or model data, and can be applied to a variety of applications for which the use of Generative Adversarial Networks (GANs) are currently being explored as methods to either anonymize personally identifiable information (PII), or as a means to generate synthetic data while protecting the source material.

The researchers have formulated a method called *Identity Membership Attack*, which evaluates the likelihood of a single identity appearing *frequently* in a contributing dataset, rather than attempting to key in on particular characteristics of an identity (i.e. on the pixel groups of an original image that was used to train the generative model).

Source: https://arxiv.org/pdf/2107.06018.pdf

In the image above, from the research, each row begins with a GAN-generated image created by StyleGAN. The left block of images was created from a database of 40,000 images, the middle from 80,000 and the right block from 46,000 images. All images come from the VGG2Face2 dataset.

Some samples have a fleeting resemblance, while others strongly correlate to the training data. The faces were successfully identified by the researchers using a face identification network.

More Than Value

Re-identification approaches of this nature have multiple implications across many research fields; the researchers, based at the University of Caen in Normandy, emphasize that their technique is not limited to face-sets and face-generating GAN frameworks, but is equally applicable to medical imaging datasets and biometric data, among other possible attack surfaces in image synthesis frameworks.

> *'We hold that if successful, such an attack would reveal as a serious hurdle for the safe exchange of GANs in sensitive contexts. For instance, in the context of paintings or other art pieces, distributing a non-private generator might well be ruled-out for obvious copyright issues. More importantly, consider a biometric company A releasing a generator exposing its consumer identity. Another company B could potentially detect which of their own consumers are also clients of company A. Similar situations can pose serious issues for medical data, where revealing a GAN could breach personal information about a patient disease.'*

Re-Identifying Illegitimately Web-Scraped or Private Data

Though the paper only touches lightly on the subject, the ability to identify original source data from abstracted output (such as GAN-generated faces, though this applies equally to encoder/decoder systems and other architectures) has notable implications for copyright-protection implementations over the next 5-10 years.

Currently most countries are operating a *laissez faire* approach to the scraping of public-facing web data in order not to be left behind in the developmental stage of the machine learning economies to come. As that climate commercializes and consolidates, there is significant potential for a new generation of 'Data Trolls' to present copyright claims on images confirmed to have been used historically in datasets that have contributed to machine learning algorithms.

As the developed algorithms mature and become more valuable with time, any non-permitted imagery that was used in their early development, and that can be inferred from their output by methods similar to those proposed in the new French paper, is a potential legal liability on the scale of SCO Vs IBM (a legendarily long-lived tech lawsuit that continues to threaten the Linux operating system).

Exploiting the Mexican Stand-off of Diversity vs. Frequency

The primary technique used by the French researchers exploits the frequency of original dataset images as a key to re-identification. The more frequently a particular identity is found in the dataset, the more likely that it will be possible to make an identification of that original identity, by correlating the results of the attack to publicly or privately available datasets.

The researchers note that this can be mitigated by including a far greater diversity of data (for instance, of faces) in the source dataset, and by not training the dataset so long that overfitting occurs. The problem with this is that the model must then achieve good abstraction in a much higher dimensional space, and with a much higher amount of data than is strictly necessary to obtain plausible synthetic results.

To achieve optimum generalization of this kind is expensive and time-consuming: the latent space (the formulaic analysis part of the machine learning model into which data is fed) will need more resources; the dataset will need more curation; and since the amount of data will need to be significant, batch sizes and rate scheduling will have to be optimized for quality and high levels of generalization, rather than speed-of-training and economy, making for higher development costs and longer development times.

Furthermore, overfitted generative algorithms can achieve highly realistic synthetic data, even if the output data (i.e. faces, maps, biomedical images, etc.) is not completely abstract, but features larger distinguishing traits from the source data than would be ideal – a tempting shortcut. In the current 'wild west' climate of the machine learning sector, where smaller initiatives are attempting to challenge FAANG's lead with scarcer resources (or else gain attention for a buy-out), it's questionable whether standards always rise this high.

The paper also observes that diversity of source data points (such as faces) is not enough by itself to prevent re-identification through these and similar methods, since early stopping of training can leave source identities insufficiently abstracted.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai