How to Know When Image Synthesis Systems Are Producing Genuinely ‘Original’ Material

Originally published at https://www.unite.ai/how-to-know-when-image-synthesis-systems-are-producing-genuinely-original-material/

'Teddy bears working on new AI research underwater with 1990s technology' – Source: https://www.creativeboom.com/features/meet-dall-e/

A new study from South Korea has proposed a method to determine whether image synthesis systems are producing genuinely novel images, or ‘minor’ variants on the training data, potentially defeating the objective of such architectures (such as the production of novel and original images).

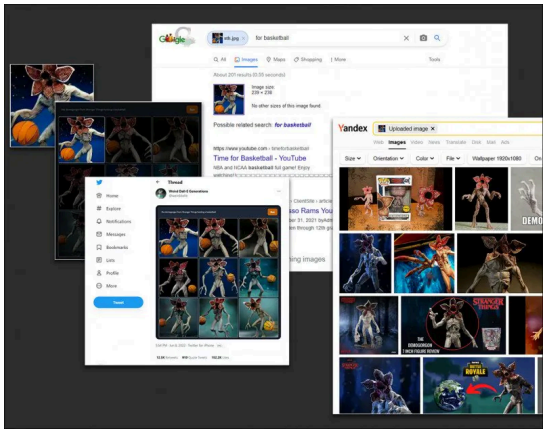
Very often, the paper suggests, the latter is true, because the existing metrics that such systems use to improve their generative capacities over the course of training are forced to favor images that are relatively close to the (non fake) source images in the dataset.

After all, if a generated image is ‘visually close’ to the source data, it is inevitably likely to score better for ‘authenticity’ than ‘originality’, since it’s ‘faithful’ – if uninspired.

In a sector too nascent and untried for its legal ramifications to be yet known, this could turn out to be an important legal issue, if it transpires that commercialized synthetic image content does not differ enough from the (often) copyrighted source material that is currently allowed to perfuse the research sector in the form of popular web-scraped datasets (the potential for future infringement claims of this type has come to prominence fairly recently in regard to Microsoft’s GitHub Co-Pilot AI).

In terms of the increasingly coherent and semantically robust output from systems such as OpenAI’s DALL-E 2, Google’s Imagen, and China’s CogView releases (as well as the lower-specced DALL-E mini), there are very few *post facto* ways to reliably test for the originality of a generated image.

Indeed, searching for some of the most popular of the new DALL-E 2 images will often only lead to further instances of those same images, depending on the search engine.

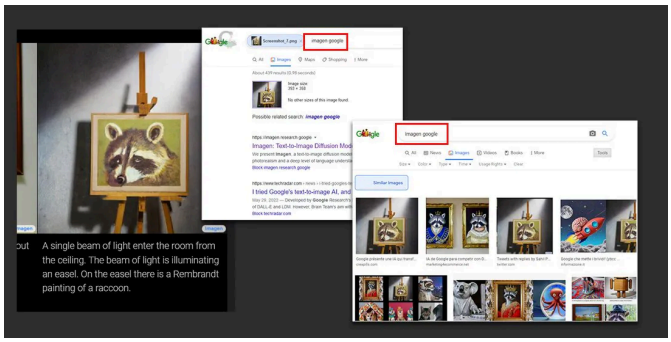


Uploading a complete 9-image DALL-E 2 output group only leads to more DALL-E 2 output groups, because the grid structure is the strongest feature. Separating and uploading the first image (from this Twitter post of 8th June 2022, from the ‘Weird Dall-E Generations’ account) causes Google to fixate on the basketball in the picture, taking the image-based search down a semantic blind alley. For the same image-based search, Yandex seems at least to be doing some actual pixel-based deconstruction and feature-matching.

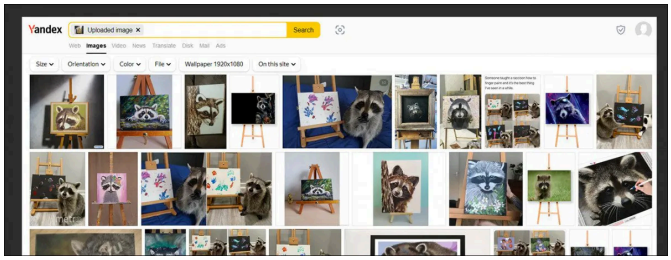
Though Yandex is more likely than Google Search to use the actual *features* (i.e. an image's derived/calculated features, not necessarily facial features of people) and *visual* (rather than semantic) characteristics of a submitted image to find similar images, all image-based search engines either have some kind of agenda or practice that may make it difficult to identify instances of *source>generated* plagiarism via web searches.

Additionally, the training data for a generative model may not be publicly available in its entirety, further hobbling forensic examination of the originality of generated images.

Interestingly, performing an image-based web-search on one of the synthetic images featured by Google at its dedicated Imagen site finds absolutely nothing comparable to the subject of the image, in terms of actually looking at the image and impartially seeking similar images. Rather, semantically fixated as ever, the Google Image search results for this Imagen picture will not permit a pure image-based web-search of the image without adding the search terms 'imagen google' as an additional (and limiting) parameter:

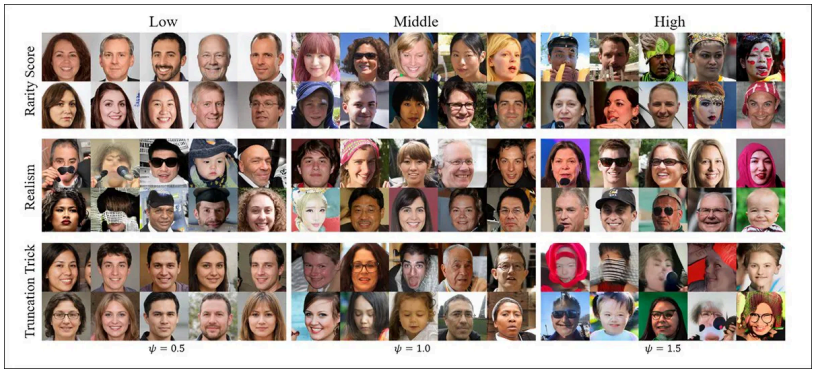


Yandex, conversely, finds a multitude of similar (or at least visually-related) real-world images from the amateur artistic community:



In general, it would be better if the novelty or originality of the output of image synthesis systems could in some way be measured, without needing to extract features from every possible web-facing image on the internet at the time the model was trained, or in non-public datasets that may be using copyrighted material.

Related to this issue, researchers from the Kim Jaechul Graduate School of AI at the Korea Advanced Institute of Science and Technology (KAIST AI) have collaborated with global ICT and search company NAVER Corp to develop a *Rarity Score* that can help to identify the more original creations of image synthesis systems.



Images here are generated via StyleGAN-FFHQ. From left to right, the columns indicate worst to best results. We can see that the 'Truncation trick' metric (see be their own agendas, whilst the new 'Rarity' score (top row) is seeking out cohesive but original imagery (rather than just cohesive imagery). Since there are image-see the source paper for better detail and resolution. Source: https://arxiv.org/pdf/2206.08549.pdf

The new paper is titled *Rarity Score: A New Metric to Evaluate the Uncommonness of Synthesized Images*, and comes from three researchers at KAIST, and three from NAVER Corp.

Beyond the 'Cheap Trick'

Among the prior metrics that the new paper is seeking to improve on are the 'Truncation trick' suggested in 2019 in a collaboration between the UK's Heriot-Watt University and Google's DeepMind.

The Truncation Trick essentially uses a different latent distribution for sampling than was used for training the generative model.

The researchers who developed this method were surprised that it worked, but concede in the original paper that it reduces the variety of generated output. Nonetheless, the Truncation Trick has become effective and popular, in the context of what could arguably be re-described as a 'cheap trick' for obtaining authentic-looking results that don't really assimilate all the possibilities inherent in the data, and may resemble the source data more than is desired.

Regarding the Truncation Trick, the new paper's authors observe:

> *'[It] is not intended to generate rare samples in training datasets, but rather to synthesize typical images more stably. We hypothesize that existing generative models will be able to produce samples richer in the real data distribution if the generator can be induced to effectively produce rare samples.'*

Of the general tendency to rely on traditional metrics such as Frechet Inception Distance (FID, which came under intense criticism in December 2021), inception score (IS) and Kernel Inception Distance (KID) as 'progress indicators' during the training of a generative model, the authors further comment*:

> *'This learning scheme leads the generator not to synthesize much rare samples which are unique and have strong characteristics that do not account for a large proportion of the real image distribution. Examples of rare samples from public datasets include people with various accessories in FFHQ, white animals in AFHQ, and uncommon statues in Metfaces.*
>
> *'The ability to generate rare samples is important not only because it is related to the edge capability of the generative models, but also because uniqueness plays an important role in the creative applications such as virtual humans.*
>
> *'However, the qualitative results of several recent studies seldom contain these rare examples. We conjecture that the nature of the adversarial learning scheme forces generated image distribution similar to that of a training dataset. Thus, images with clear individuality or rareness only take a small part in images synthesized by the models.'*

Technique

The researchers' new Rarity Score adapts an idea presented in earlier works – the use of K-Nearest Neighbors (KNNs) to represent the arrays of genuine (training) and synthetic (output) data in an image synthesis system.

Regarding this novel method of analysis, the authors assert:

> *'We hypothesize that ordinary samples would be closer to each other whereas unique and rare samples would be sparsely located in the feature space.'*

The results image above shows the smallest nearest neighbor distances (NNDs) over to the largest, in a StyleGAN architecture trained on FFHQ.

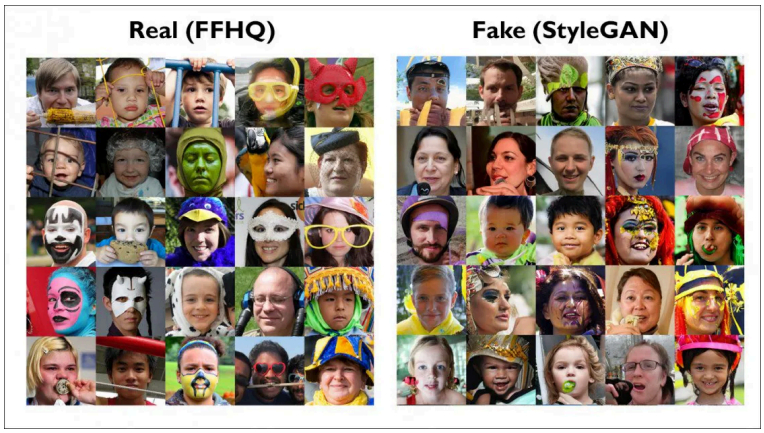
> *'For all datasets, samples with the smallest NNDs show representative and typical images. On the contrary, the samples with the largest NNDs have strong individuality and are significantly different from the typical images with the smallest NNDs.'*

In theory, by using this new metric as a discriminator, or at least including it in a more complex discriminator architecture, a generative system could be steered away from pure imitation towards a more inventive algorithm, whilst retaining essential cohesion of concepts that may be critical for authentic image production (i.e. *'man'*, *'woman'*, *'car'*, *'church'*, etc.).

Comparisons and Experiments

In tests, the researchers conducted a comparison of the Rarity Score's performance against both the Truncation Trick and NVIDIA's 2019 Realism Score, and found that across a variety of frameworks and datasets, the approach is able to individuate 'unique' results.

Though the results featured in the paper are too extensive to include here, the researchers seem to have demonstrated the ability of the new method to identify rarity in both source (real) and generated (fake) images in a generative procedure:



Select examples from the extensive visual results reproduced in the paper (see source URL above for more details). On the left, genuine examples from FFHQ tha neighbors (i.e. are novel and unusual) in the original dataset; on the right, fake images generated by StyleGAN, which the new metric has identified as truly novel. image-size limits in this article, please see the source paper for better detail and resolution.

The new Rarity Score metric not only allows for the possibility of identifying 'novel' generative output in a single architecture, but also, the researchers claim, allows comparisons between generative models of various and varying architectures (i.e. autoencoder, VAE, GAN, etc.).

The paper notes that Rarity Score differs from prior metrics by concentrating on a generative framework's capability to create unique and rare images, in opposition to 'traditional' metrics, which examine (rather more myopically) the diversity between generations during the training of the model.

Beyond Limited Tasks

Though the new paper's researchers have conducted tests on limited-domain frameworks (such as generator/dataset combinations designed to specifically produce pictures of people, or of cats, for example), the Rarity Score can potentially be applied to any arbitrary image synthesis procedure where it's desired to identify generated examples that use the distributions derived from the trained data, instead of increasing authenticity (and reducing diversity) by interposing foreign latent distributions, or relying on other 'shortcuts' that compromise novelty in favor of authenticity.

In effect, such a metric could potentially distinguish truly novel output instances in systems such as the DALL-E series, by using identified distance between an apparent 'outlier' result, the training data, and results from similar prompts or inputs (i.e., image-based prompts).

In practice, and in the absence of a clear understanding of the extent to which the system has truly assimiliated visual and semantic concepts (often impeded by limited knowledge about the training data), this could be a viable method to identify a genuine 'moment of inspiration' in a generative system – the point at which an adequate number of input concepts and data have resulted in something genuinely inventive, instead of something overly derivative or close to the source data.

** My conversions of the authors' inline citations to hyperlinks.*

First published 20th June 2022.

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

Discover More