Breaking MAD: Generative AI could break the internet

July 30 2024, by Silvia Cernea Clark

Generative artificial intelligence (AI) models trained on synthetic data generate outputs that are progressively marred by artifacts. In this example, the researchers trained a succession of StyleGAN-2 generative models using fully synthetic data. Each of the six image columns displays a couple of examples generated by the first, third, fifth and ninth generation model, respectively. With each iteration of the loop, the cross-hatched artifacts become progressively amplified. Credit: Digital Signal Processing Group/Rice University

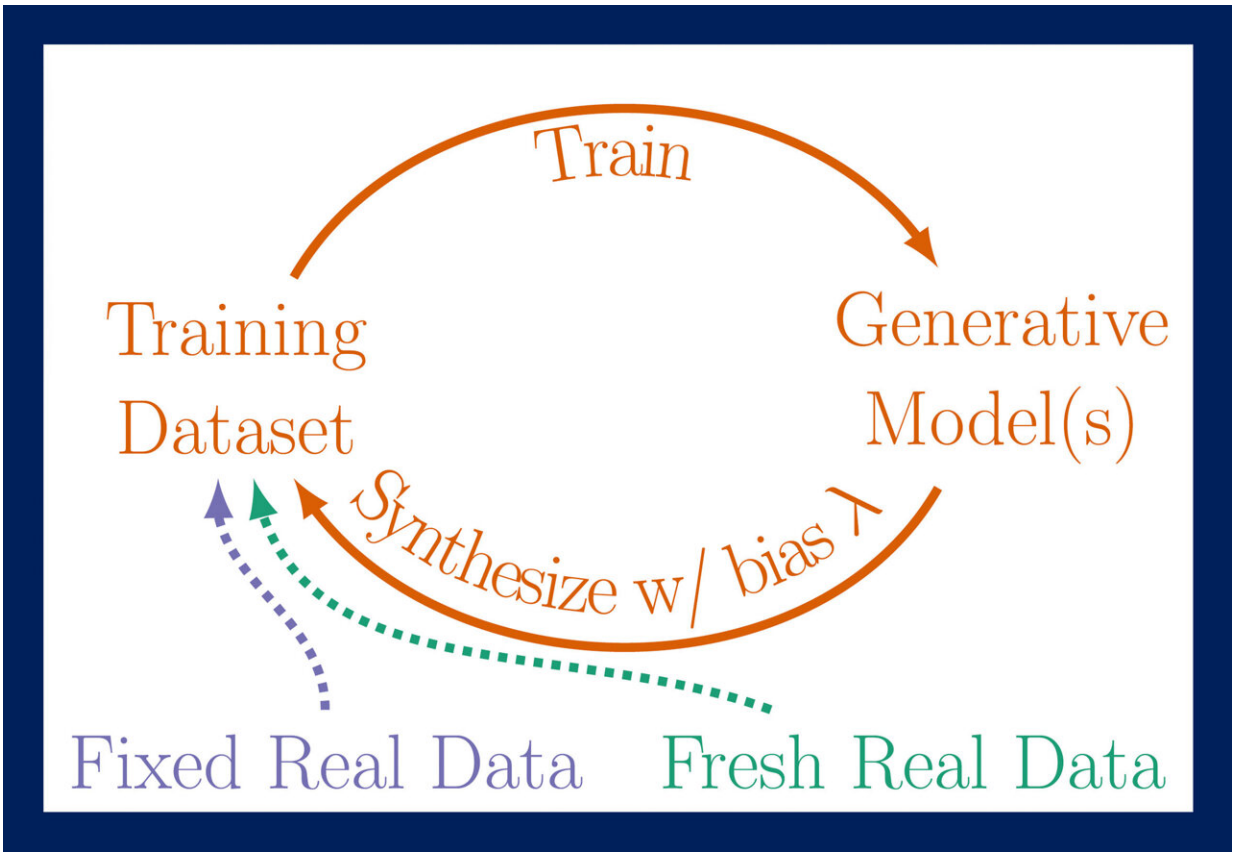
Generative artificial intelligence (AI) models like OpenAI's GPT-4o or Stability AI's Stable Diffusion are surprisingly capable at creating new text, code, images and videos. Training them, however, requires such

vast amounts of data that developers are already running up against supply limitations and may soon exhaust training resources altogether.

Against this backdrop of data scarcity, using synthetic data to train future generations of the AI models may seem like an alluring option to big tech for a number of reasons, including: AI-synthesized data is cheaper than real-world data and virtually limitless in terms of supply; it poses fewer privacy risks (as in the case of medical data); and in some cases, synthetic data may even improve AI performance.

However, recent work by the Digital Signal Processing group at Rice University has found that a diet of synthetic data can have significant negative impacts on generative AI models' future iterations.

"The problems arise when this synthetic data training is, inevitably, repeated, forming a kind of a feedback loop--what we call an autophagous or 'self-consuming' loop," said Richard Baraniuk, Rice's C. Sidney Burrus Professor of Electrical and Computer Engineering. "Our group has worked extensively on such feedback loops, and the bad news is that even after a few generations of such training, the new models can become irreparably corrupted. This has been termed 'model collapse' by some--most recently by colleagues in the field in the context of large language models (LLMs). We, however, find the term 'Model Autophagy Disorder' (MAD) more apt, by analogy to mad cow disease."



Richard Baraniuk and his team at Rice University studied three variations of self-consuming training loops designed to provide a realistic representation of how real and synthetic data are combined into training datasets for generative models. Schematic illustrates the three training scenarios, i.e. a fully synthetic loop, a synthetic augmentation loop (synthetic + fixed set of real data), and a fresh data loop (synthetic + new set of real data). Credit: Digital Signal Processing Group/Rice University

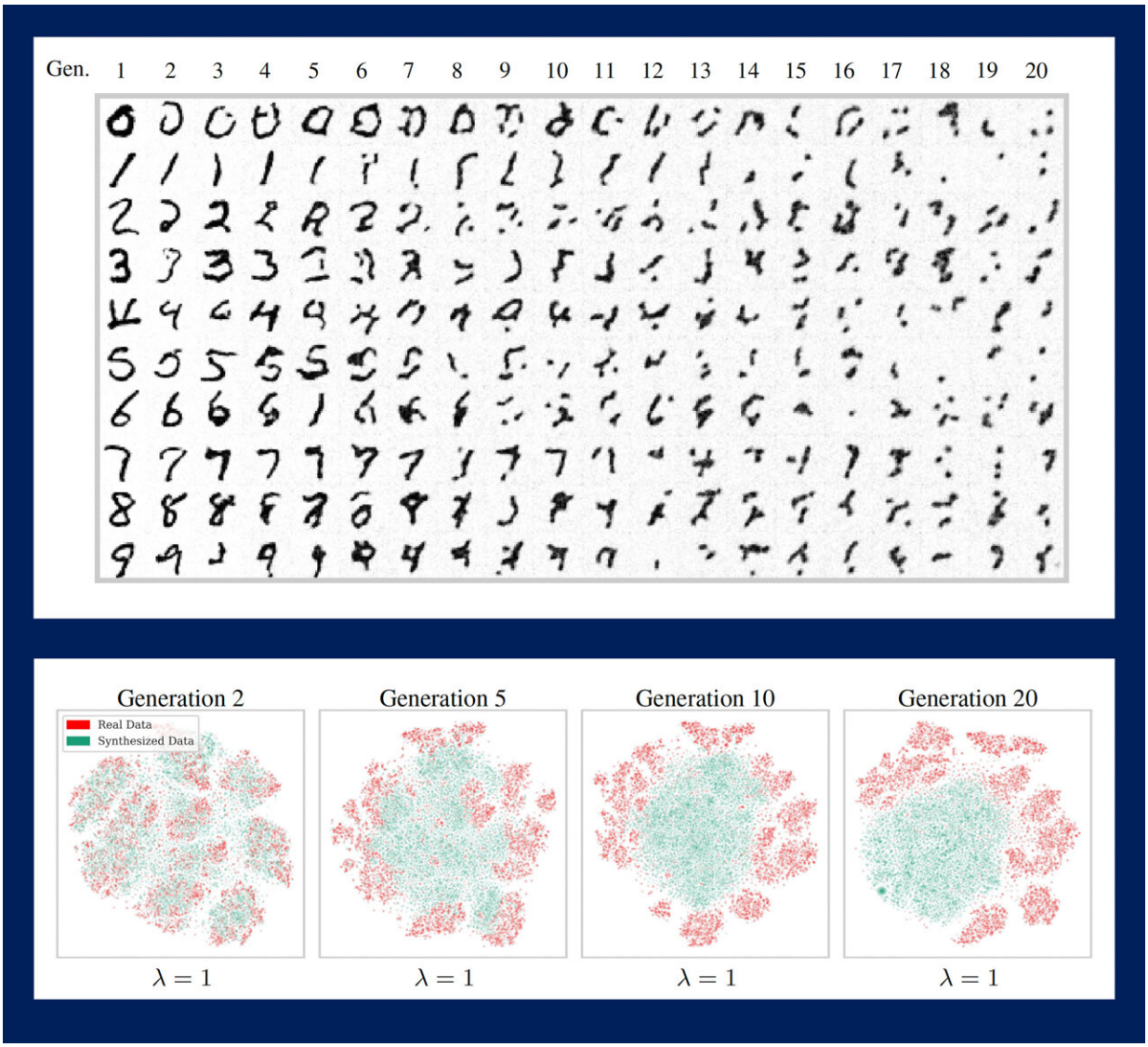
Mad cow disease is a fatal neurodegenerative illness that affects cows and has a human equivalent caused by consuming infected meat. A major outbreak in the 1980-90s brought attention to the fact that mad cow disease proliferated as a result of the practice of feeding cows the processed leftovers of their slaughtered peers--hence the term "autophagy," from the Greek auto-, which means "self,"' and phagy--"to

eat."

"We captured our findings on MADness in a paper presented in May at the International Conference on Learning Representations (ICLR)," Baraniuk said.

The study, titled "Self-Consuming Generative Models Go MAD," is the first peer-reviewed work on AI autophagy and focuses on generative image models like the popular DALL·E 3, Midjourney and Stable Diffusion.

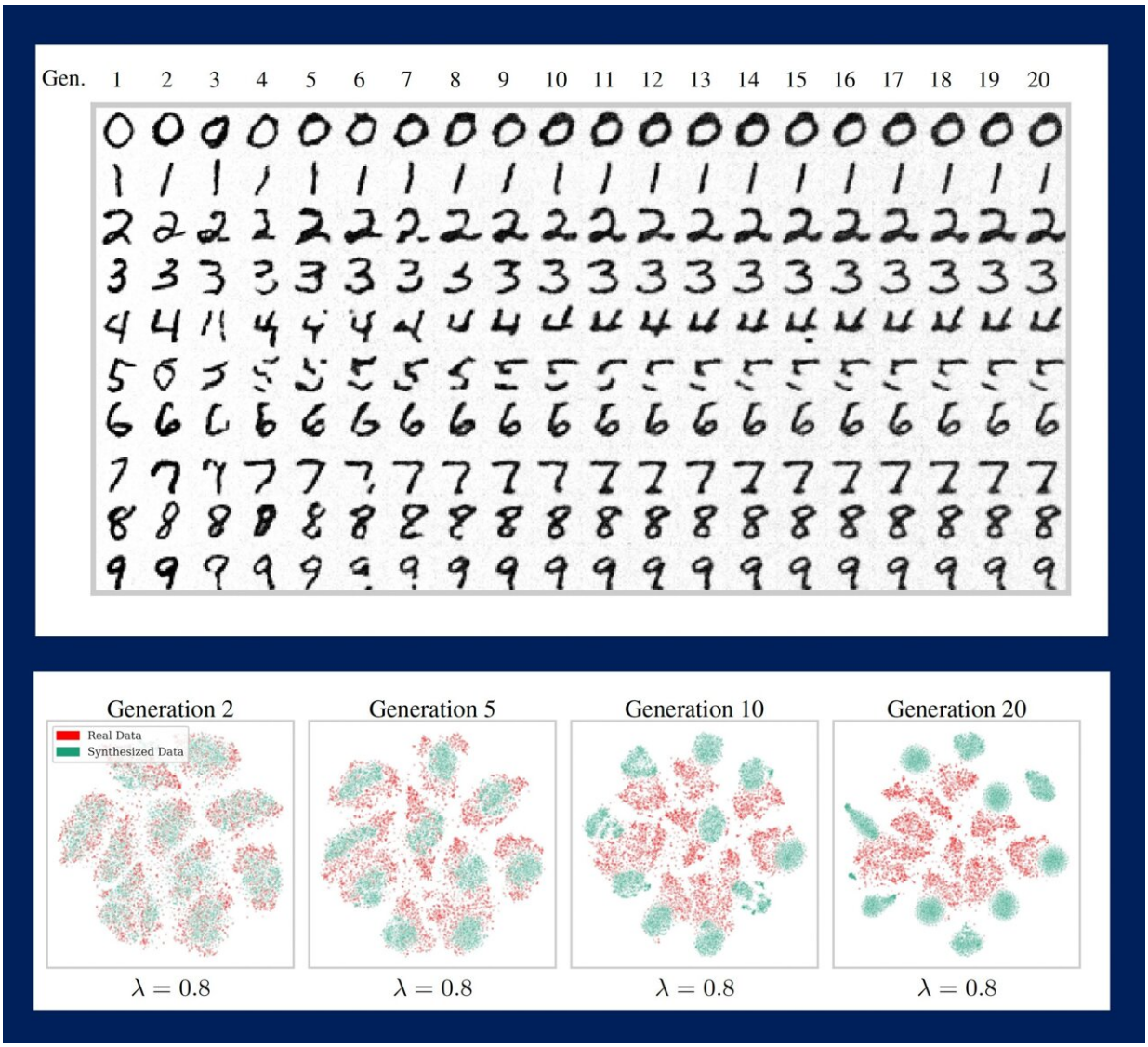
"We chose to work on visual AI models to better highlight the drawbacks of autophagous training, but the same mad cow corruption issues occur with LLMs, as other groups have pointed out," Baraniuk said.



Progressive transformation of a dataset consisting of numerals 1 through 9 across 20 model iterations of a fully synthetic loop without sampling bias (top panel), and corresponding visual representation of data mode dynamics for real (red) and synthetic (green) data (bottom panel). In the absence of sampling bias, synthetic data modes separate from real data modes and merge. This translates into a rapid deterioration of model outputs: If all numerals are fully legible in generation 1 (leftmost column, top panel), by generation 20 all images have become illegible (rightmost column, top panel). Credit: Digital Signal Processing Group/Rice University

The internet is usually the source of generative AI models' training datasets, so as synthetic data proliferates online, self-consuming loops are likely to emerge with each new generation of a model. To get insight into different scenarios of how this might play out, Baraniuk and his team studied three variations of self-consuming training loops designed to provide a realistic representation of how real and synthetic data are combined into training datasets for generative models:

- Fully synthetic loop--Successive generations of a generative model were fed a fully synthetic data diet sampled from prior generations' output.
- Synthetic augmentation loop--The training dataset for each generation of the model included a combination of synthetic data sampled from prior generations and a fixed set of real training data.
- Fresh data loop--Each generation of the model is trained on a mix of synthetic data from prior generations and a fresh set of real training data.



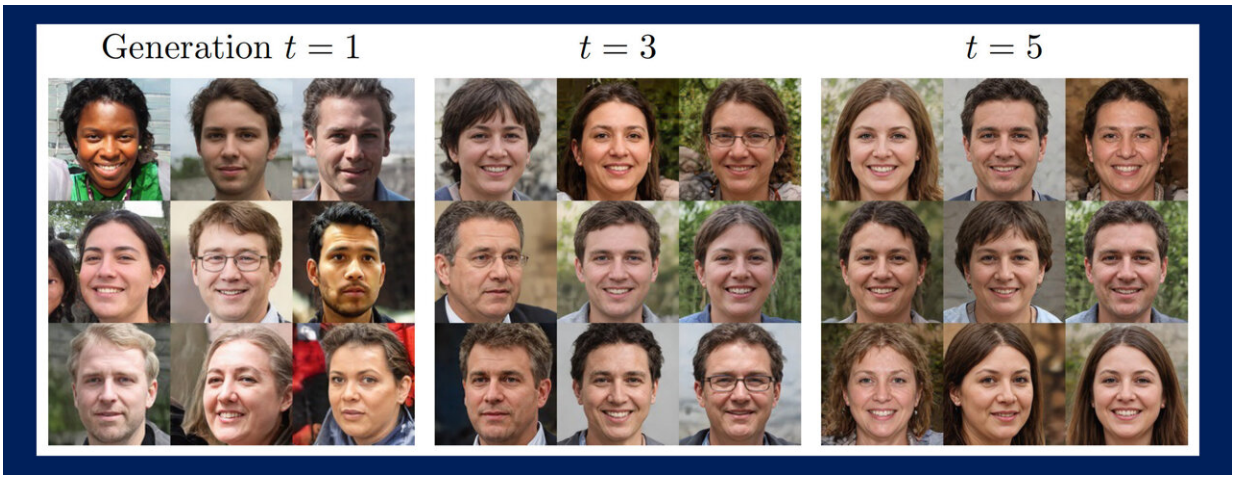
Progressive transformation of a dataset consisting of numerals 1 through 9 across 20 model iterations of a fully synthetic loop with sampling bias (top panel), and corresponding visual representation of data mode dynamics for real (red) and synthetic (green) data (bottom panel). With sampling bias, synthetic data modes still separate from real data modes, but, rather than merging, they collapse around individual, high-quality images. This translates into a prolonged preservation of higher quality data across iterations: All but a couple of the numerals are still legible by generation 20 (rightmost column, top panel). While sampling bias preserves data quality longer, this comes at the expense of data diversity. Credit: Digital Signal Processing Group/Rice University

Progressive iterations of the loops revealed that over time and in the absence of sufficient fresh real data, the models would generate increasingly warped outputs lacking either quality, diversity or both. In other words, the more fresh data, the healthier the AI.

Side-by-side comparisons of image datasets resulting from successive generations of a model paint an eerie picture of potential AI futures. Datasets consisting of human faces become increasingly streaked with gridlike scars--what the authors call "generative artifacts"--or look more and more like the same person. Datasets consisting of numbers morph into indecipherable scribbles.

"Our theoretical and empirical analyses have enabled us to extrapolate what might happen as generative models become ubiquitous and train future models in self-consuming loops," Baraniuk said. "Some ramifications are clear: Without enough fresh real data, future generative models are doomed to MADness."

To make these simulations even more realistic, the researchers introduced a sampling bias parameter to account for "cherry picking"--the tendency of users to favor data quality over diversity, i.e., to trade off variety in the types of images and texts in a dataset for images or texts that look or sound good.



The incentive for cherry picking — the tendency of users to favor data quality over diversity — is that data quality is preserved over a greater number of model iterations, but this comes at the expense of an even steeper decline in diversity. Pictured are sample image outputs from a first, third and fifth generation model of fully synthetic loop with sampling bias parameter. With each iteration, the dataset becomes increasingly homogeneous. Credit: Digital Signal Processing Group/Rice University

The incentive for cherry picking is that data quality is preserved over a greater number of model iterations, but this comes at the expense of an even steeper decline in diversity.

"One doomsday scenario is that if left uncontrolled for many generations, MAD could poison the data quality and diversity of the entire internet," Baraniuk said. "Short of this, it seems inevitable that as-to-now-unseen unintended consequences will arise from AI autophagy even in the near term."

In addition to Baraniuk, study authors include Rice Ph.D. students Sina Alemohammad; Josue Casco-Rodriguez; Ahmed Imtiaz Humayun; Hossein Babaei; Rice Ph.D. alumnus Lorenzo Luzi; Rice Ph.D. alumnus

and current Stanford postdoctoral student Daniel LeJeune; and Simons Postdoctoral Fellow Ali Siahkoohi.

More information: Self-Consuming Generative Models Go MAD (2024)

Provided by Rice University