Iterated training of Wasserstein GANs

Ben René Bjørsvik, Jonas Båtnes og Carl Fredrik Nordbø Knutsen

Problem Description

We have a model for approximating a probability distribution.

We are interested in what happens, when we do the following:

- We fit the model
- We sample from the fitted model.
- We fit a new model on the data we just sampled
- Repeat.

Motivating example: LLMs

Article

AI models collapse when trained on recursively generated data

https://doi.org/10.1038/s41586-024-07566-y

Received: 20 October 2023

Accepted: 14 May 2024

Published online: 24 July 2024

Open access

Check for updates

Ilia Shumailov[1,8]✉, Zakhar Shumaylov[2,8]✉, Yiren Zhao[3], Nicolas Papernot[4,5], Ross Anderson[6,7,9] & Yarin Gal[1]✉

Stable diffusion revolutionized image creation from descriptive text. GPT-2 (ref. 1), GPT-3(.5) (ref. 2) and GPT-4 (ref. 3) demonstrated high performance across a variety of language tasks. ChatGPT introduced such language models to the public. It is now clear that generative artificial intelligence (AI) such as large language models (LLMs) is here to stay and will substantially change the ecosystem of online text and images. Here we consider what may happen to GPT-{*n*} once LLMs contribute much of the text found online. We find that indiscriminate use of model-generated content in training causes irreversible defects in the resulting models, in which tails of the original content distribution disappear. We refer to this effect as 'model collapse' and show that it can occur in LLMs as well as in variational autoencoders (VAEs) and Gaussian mixture models (GMMs). We build theoretical intuition behind the phenomenon and portray its ubiquity among all learned generative models. We demonstrate that it must be taken seriously if we are to sustain the benefits of training from large-scale data scraped from the web. Indeed, the value of data collected about genuine human interactions with systems will be increasingly valuable in the presence of LLM-generated content in data crawled from the Internet.

Wasserstein Metric

Measures distance between probability distributions.

See: optimal transport theory.

Gaussian Model Collapse Theorem

Theorem 3.1 (Gaussian model collapse). Assume the original data are sampled from distribution $\mathcal{D}_0$ (not necessarily Gaussian), with non-zero sample variance. Assume X^n are fit recursively using the unbiased sample mean and variance estimators from the previous generation, $X_j^n | \mu_n, \Sigma_n \sim \mathcal{N}(\mu_n, \Sigma_n)$, with a fixed sample size. Then,

$$\mathbb{E}[\mathbb{W}_2^2(\mathcal{N}(\mu_n, \Sigma_n), \mathcal{D}_0)] \to \infty; \quad \Sigma_n \overset{\text{a.s.}}{\to} 0 \text{ as } n \to \infty,$$

in which $\mathbb{W}_2$ denotes the Wasserstein-2 distance between the true distribution and its approximation at generation n.

High-level empirical results

Training on generated data gives worse models.

Wasserstein GANs

GAN: Generative Adversarial Network

Loss:

$$L_{WGAN}(\mu_G, D) := \mathbb{E}_{x\sim\mu_G}[D(x)] - \mathbb{E}_{x\sim\mu_{ref}}[D(x)].$$

Discriminator:

- Optimal reply gives a loss that is proportional to Wasserstein distance.

If optimal discriminator is picked, then generator minimizes Wasserstein distance.

Goal

Perform iterated training of Wasserstein GANs

- Use a simple dataset, eg. MNIST

Explain qualitatively how this performs. Compare with models implemented in Shumailov paper.

A technicality: there is a requirement on the norm on the weight matrices in Wasserstein GANs. How should we enforce this requirement?