Title: Mode collapse

URL: https://en.wikipedia.org/wiki/Mode_collapse

PageID: 78998190

Categories: Category:Artificial intelligence, Category:Generative artificial intelligence, Category:Machine learning

Source: Wikipedia (CC BY-SA 4.0).

In machine learning , mode collapse is a failure mode observed in generative models , originally noted in Generative Adversarial Networks (GANs) . It occurs when the model produces outputs that are less diverse than expected, effectively "collapsing" to generate only a few modes of the data distribution while ignoring others. This phenomenon undermines the goal of generative models to capture the full diversity of the training data.

There are typically two times at which a model can collapse: either during training or during post-training finetuning.

Mode collapse reduces the utility of generative models in applications, such as in

image synthesis (repetitive or near-identical images);

data augmentation (limited diversity in synthetic data);

scientific simulations (failure to explore all plausible scenarios).

Distinctions

Mode collapse is distinct from overfitting , where a model learns detailed patterns in the training data that does not generalize to the test data, and underfitting , where it fails to learn patterns.

Memorization is where a model learns to reproduce data from the training data. Memorization is often confused with mode collapse. However, a model can memorize the training dataset without mode collapse. Indeed, if a model is severely mode-collapsed, then it has failed to memorize large parts of the training dataset.

Model collapse is one particular mechanism for the phenomenon of mode collapse, i.e. when a generative model 2 is pretrained mainly on the outputs of model 1, then another new generative model 3 is pretrained mainly on the outputs of model 2, etc. When models are trained in this way, each model is typically more mode-collapsed than the previous one. However, there are other mechanisms for mode collapse.

In GANs

Training-time mode collapse was originally noted and studied in GANs, where it arises primarily due to imbalances in the training dynamics between the generator and discriminator in GANs. In the original GAN paper, it was also called the "Helvetica scenario". [1] [2]

Common causes include: [3]

If the discriminator learns too slowly, the generator may exploit weaknesses by producing a narrow set of outputs that consistently fool the discriminator.

Traditional GAN loss functions (e.g., Jensen-Shannon divergence) may be too lenient on generating same-looking outputs.

The adversarial training process can lead to oscillatory behavior, where the generator and discriminator fail to converge to a stable equilibrium, but instead engage in a rock-beats-paper-beats-scissors kind of cycling. The generator would generate just "rock" until the discriminator learns to classify that as generated, then the generator switch to generating just "scissors", and so on. The generator would always be mode-collapsed, though the precise mode in which it collapses to would change during training.

Several GAN-specific strategies were developed to mitigate mode collapse:

Two time-scale update rule. [4]

Mini-batch discrimination [5] allows the discriminator to evaluate entire batches of samples, encouraging diversity.

Unrolled GANs [6] optimize the generator against future states of the discriminator.

Wasserstein GAN uses Earth Mover's distance to provide more stable gradients. [7]

Use a big and balanced training dataset. [8]

Regularization methods such as gradient penalty and spectral normalization . [9]

Finetuning

The large language models are usually trained in two steps. In the first step ("pretraining"), the model is trained to simply generate text sampled from a large dataset. In the second step ("finetuning"), the model is trained to perform specific tasks by training it on a small dataset containing just the task-specific data. For example, to make a chatbot in this method, one first pretrains a large transformer model over a few trillion words of text scraped from the Internet, then finetunes it on a few million words of example chatlogs that the model should imitate.

Mode collapse may occur during finetuning, as the model learns to generate text that accomplishes the specific task, but loses ability to generate other forms of text. It may also be able to generate a smaller subset of texts that accomplish the specific task. It is hypothesized that there is a tradeoff between quality and diversity. Given a single pretrained model, one may finetune it to perform a specific task. More finetuning would result in higher average task performance, but less diverse outputs. Less finetuning would result in lower average performance, but more diverse outputs. [10] A similar tradeoff has been observed in image generation models [11] and GAN-based text generators. [12]

Similarly, mode collapse may occur during RLHF , via reward hacking the reward model or other mechanisms. [13] [14]

See also

Variational autoencoder

Generative model

Generative artificial intelligence

Generative pre-trained transformer

Overfitting

References