Some Challenges around Retraining Generative Models on Their Own Data

Quentin Bertrand

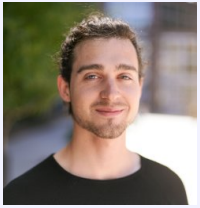
Inria – UJM – QB3.github.io

D. Ferbach

J. A. Bose

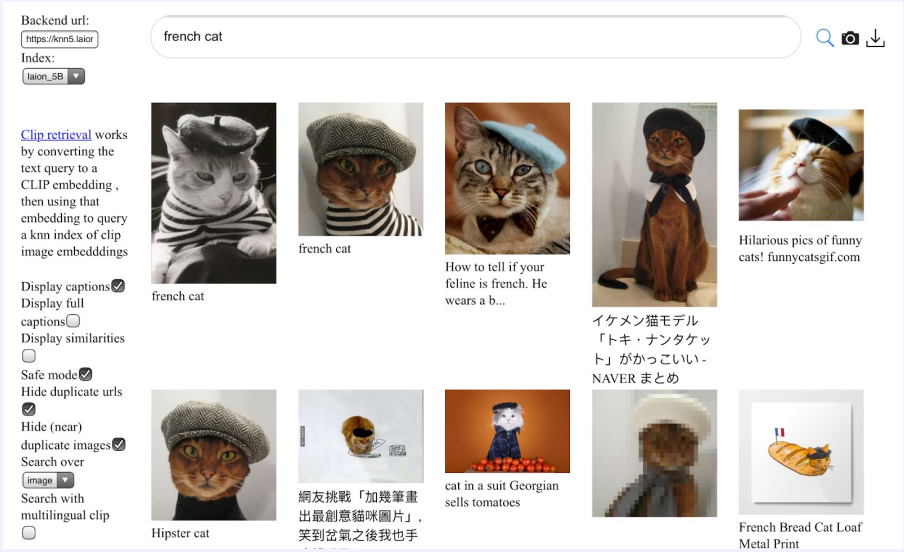
A. Duplessis

M. Jiralerspong



G. Gidel

LAION-5B[1][2]



[1] C. Schuhmann et al. "Laion-5B: An open large-scale dataset for training next generation image-text models". In: *NeurIPS* (2022).

[2] https://paperswithcode.com/dataset/laion-5b

Generative Models Today

- Powerful deep generative models
  - ↪ *e.g.* Diffusion trained on LAION-5B
- Easy access (Midjourney, Stable Diffusion, DALL·E)

Generative Models Today

- Powerful deep generative models
  - ↪ *e.g.* Diffusion trained on LAION-5B
- Easy access (Midjourney, Stable Diffusion, DALL·E)
- Populates the WEB with **synthetically generated images**

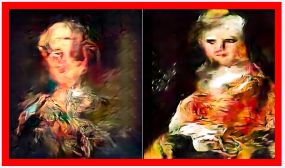
Generative Models Today

- Powerful deep generative models
  - ↪ *e.g.* Diffusion trained on LAION-5B
- Easy access (Midjourney, Stable Diffusion, DALL·E)
- Populates the WEB with **synthetically generated images**

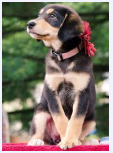
Inevitably Train on Synthetic Data

The LAION-5B[3] dataset already contains synthetically generated images[4]

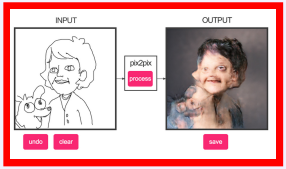








[3] C. Schuhmann et al. "Laion-5B: An open large-scale dataset for training next generation image-text models". In: *NeurIPS* (2022).

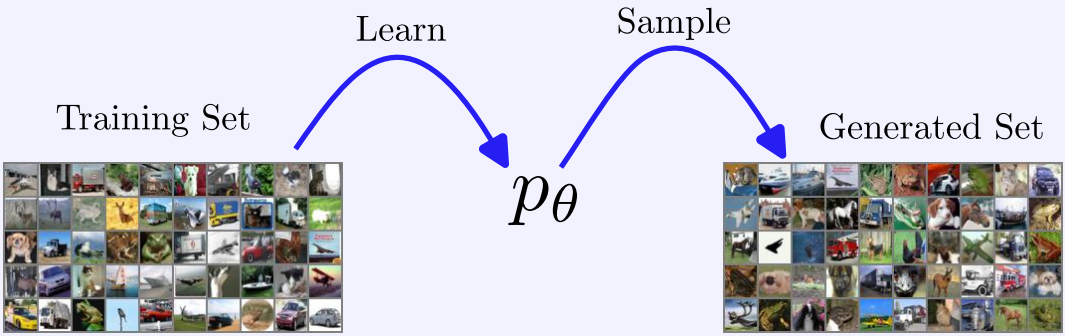
[4] S. Alemohammad et al. "Self-Consuming Generative Models Go MAD". In: *ICLR* (2024).

Retraining Generative Models on their Own Data

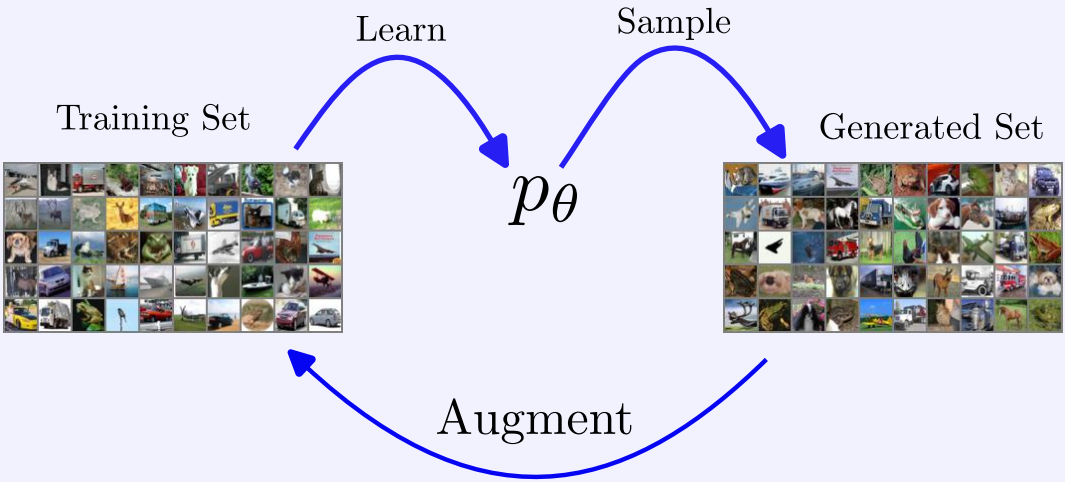
Training Set

Retraining Generative Models on their Own Data

Retraining Generative Models on their Own Data



Retraining Generative Models on their Own Data



Training on Synthetic Data, Good or Bad?

Iterative Retraining is **Bad**

- **The curse of recursion**: Training on generated data makes models forget[a]
- Self-Consuming Generative Models Go **MAD**[b]
- When **A.I.´s Output Is a Threat to A.I.** Itself (N.Y. Times article)

---

[a] I. Shumailov et al. "The Curse of Recursion: Training on Generated Data Makes Models Forget". In: (2023). arXiv: 2305.17493 [cs.LG].

[b] S. Alemohammad et al. "Self-Consuming Generative Models Go MAD". In: *ICLR* (2024).

Training on Synthetic Data, Good or Bad?

Iterative Retraining is **Bad**

- **The curse of recursion**: Training on generated data makes models forget[a]
- Self-Consuming Generative Models Go **MAD**[b]
- When **A.I.´s Output Is a Threat to A.I.** Itself (N.Y. Times article)

---

[a] I. Shumailov et al. "The Curse of Recursion: Training on Generated Data Makes Models Forget". In: (2023). arXiv: 2305.17493 [cs.LG].

[b] S. Alemohammad et al. "Self-Consuming Generative Models Go MAD". In: *ICLR* (2024).

Will generative models collapse?!

Training on Synthetic Data, Good or Bad?

Iterative Retraining is **Bad**

- **The curse of recursion**: Training on generated data makes models forget[a]
- Self-Consuming Generative Models Go **MAD**[b]
- When **A.I.´s Output Is a Threat to A.I.** Itself (N.Y. Times article)

[a] I. Shumailov et al. "The Curse of Recursion: Training on Generated Data Makes Models Forget". In: (2023). arXiv: 2305.17493 [cs.LG].
[b] S. Alemohammad et al. "Self-Consuming Generative Models Go MAD". In: *ICLR* (2024).

Will generative models collapse?!

Training on Synthetic Data is **Good**

- Data augmentation for downstream tasks
  - ↪ Adversarial training[a]
  - ↪ Classification with imbalanced datasets[b]
  - ↪ Generative modelling[c]

[a] Z. Wang et al. "Better diffusion models further improve adversarial training". In: *ICML*. 2023.
[b] R. A. Hemmat et al. "Feedback-guided Data Synthesis for Imbalanced Classification". In: *TMLR* (2024).
[c] C. Gulcehre et al. "Reinforced self-training (REST) for language modeling". In: (2023). arXiv: 2308.08998 [cs.CL].

Training on Synthetic Data, Good or Bad?

Iterative Retraining is **Bad**

- **The curse of recursion**: Training on generated data makes models forget[a]
- Self-Consuming Generative Models Go **MAD**[b]
- When **A.I.´s Output Is a Threat to A.I.** Itself (N.Y. Times article)

[a] I. Shumailov et al. "The Curse of Recursion: Training on Generated Data Makes Models Forget". In: (2023). arXiv: 2305.17493 [cs.LG].
[b] S. Alemohammad et al. "Self-Consuming Generative Models Go MAD". In: *ICLR* (2024).

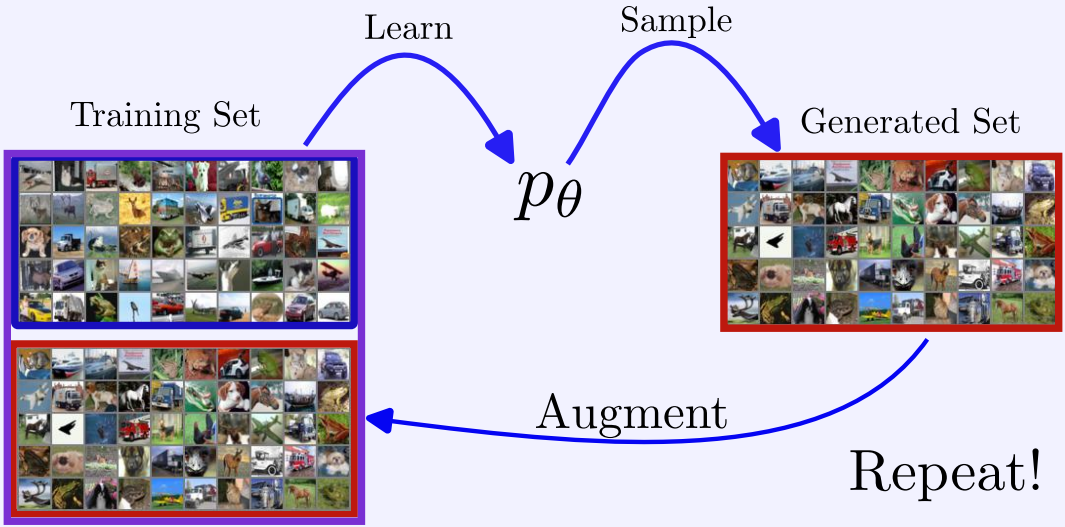
Will generative models collapse?!

Training on Synthetic Data is **Good**

- Data augmentation for downstream tasks
  - ↪ Adversarial training[a]
  - ↪ Classification with imbalanced datasets[b]
  - ↪ Generative modelling[c]

[a] Z. Wang et al. "Better diffusion models further improve adversarial training". In: *ICML*. 2023.
[b] R. A. Hemmat et al. "Feedback-guided Data Synthesis for Imbalanced Classification". In: *TMLR* (2024).
[c] C. Gulcehre et al. "Reinforced self-training (REST) for language modeling". In: (2023). arXiv: 2308.08998 [cs.CL].

Iterative Retraining / Self-Consuming Generative Models



$$\underbrace{\text{Training set}}_{\text{time } t+1} = \text{Real Data} + \underbrace{\text{Generated Set}}_{\text{time } t}$$

Iterative Retraining / Self-Consuming Generative Models

Learn

Training Set

p_θ

Sample

Generated Set

Augment

Repeat!

$$\underbrace{\text{Training set}}_{\text{time } t+1} = \text{Real Data} + \underbrace{\text{Generated Set}}_{\text{time } t}$$

Setting

Notation

- $\hat{p}_{\text{data}}$ Empirical data distribution
  - $\hookrightarrow$ n Data points
- p Likelihood of the model
  - $\hookrightarrow$ Parametrized by $\theta^n \in \Theta$

Iterative Retraining

$$p_0 \in \arg\max_{p \in \mathcal{P}_\Theta} \mathbb{E}_{x \sim \hat{p}_{\text{data}}} [\log p(x)]$$

Setting

Notation

- $\hat{p}_{\text{data}}$ Empirical data distribution
  - $\hookrightarrow$ n Data points
- p Likelihood of the model
  - $\hookrightarrow$ Parametrized by $\theta^n \in \Theta$

Iterative Retraining

$$p_0 \in \arg\max_{p \in \mathcal{P}_\Theta} \mathbb{E}_{x \sim \hat{p}_{\text{data}}} [\log p(x)]$$

$$p_{t+1} \in \arg\max_{p \in \mathcal{P}_\Theta} \underbrace{\mathbb{E}_{x \sim \hat{p}_{\text{data}}} \log p(x)}_{\text{Real data}} + \lambda \cdot \underbrace{\mathbb{E}_{\tilde{\mathbf{x}} \sim \hat{p}_t} \log p(\tilde{\mathbf{x}})}_{\text{Synthetic data}}$$

Setting

Notation

- $\hat{p}_{\text{data}}$ Empirical data distribution
  - $\hookrightarrow$ n Data points
- p Likelihood of the model
  - $\hookrightarrow$ Parametrized by $\theta^n \in \Theta$

Iterative Retraining

$$p_0 \in \arg\max_{p \in \mathcal{P}_\Theta} \mathbb{E}_{x \sim \hat{p}_{\text{data}}} [\log p(x)]$$

$$p_{t+1} \in \arg\max_{p \in \mathcal{P}_\Theta} \underbrace{\mathbb{E}_{x \sim \hat{p}_{\text{data}}} \log p(x)}_{\text{Real data}} + \lambda \cdot \underbrace{\mathbb{E}_{\tilde{\mathbf{x}} \sim \hat{p}_t} \log p(\tilde{\mathbf{x}})}_{\text{Synthetic data}}$$

Warm Up: Only Retrain on your Own Data 1/3

Iterative Retraining

$$p_0^n \in \arg\max_{p \in \mathcal{P}_\Theta} \mathbb{E}_{x \sim \hat{p}_{\text{data}}} [\log p(x)]$$

$$p_{t+1}^n \in \arg\max_{p \in \mathcal{P}_\Theta} \underbrace{\mathbb{E}_{x \sim \hat{p}_{\text{data}}} \log p(x)}_{\text{Real data}} + \cancel{1} \cdot \underbrace{\mathbb{E}_{\tilde{\mathbf{x}} \sim p_t} \log p(\tilde{\mathbf{x}})}_{\text{Synthetic data}}$$

Q: What will happen?

Warm Up: Only Retrain on your Own Data 1/3

Iterative Retraining

$$p_0^n \in \arg\max_{p \in \mathcal{P}_\Theta} \mathbb{E}_{x \sim \hat{p}_{\text{data}}} [\log p(x)]$$

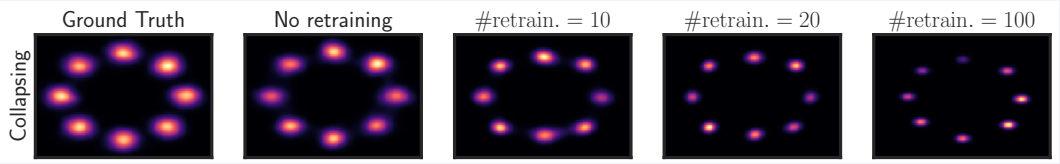
$$p_{t+1}^n \in \arg\max_{p \in \mathcal{P}_\Theta} \underbrace{\mathbb{E}_{x \sim \hat{p}_{\text{data}}} \log p(x)}_{\text{Real data}} + \cancel{1} \cdot \underbrace{\mathbb{E}_{\tilde{\mathbf{x}} \sim p_t} \log p(\tilde{\mathbf{x}})}_{\text{Synthetic data}}$$

Q: What will happen?

Warm Up: Only Retrain on your Own Data 2/3

Q: What will happen?

A: Mode Collapse



Setup

- Data: 8 Gaussians, $x \in \mathbb{R}^2$
- Algorithm: Diffusion (DDPM[a])

[a] J. Ho, A. Jain, and P. Abbeel. "Denoising diffusion probabilistic models". In: *NeurIPS* (2020).

Warm Up: Only Retrain on your Own Data 3/3

Single unidimensional Gaussian, unbiased estimator

Data: $x_j^0 = \mu_0 + \sigma_0 Z_j$, with $Z_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{0,1}$, $1 \leq j \leq n$

Learning step:
$$\begin{cases} \mu_{t+1} &= \frac{1}{n} \sum_j \tilde{\mathbf{x}}_j^t \\ \sigma_{t+1}^2 &= \frac{1}{n-1} \sum_j \left(\tilde{\mathbf{x}}_j^t - \mu_{t+1} \right)^2 \end{cases}$$

Sampling step: $\left\{ \tilde{\mathbf{x}}_j^{t+1} = \mu_{t+1} + \sigma_{t+1} \cdot Z_j^{t+1}, \text{ with } Z_j^{t+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{0,1}, 1 \leq j \leq n \right.$

Result

$$\mathbb{E}(\sigma_t) \leq \alpha^t \sigma_0 \underset{t \to +\infty}{\longrightarrow} 0, \ 0 \leq \alpha < 1$$

Warm Up: Only Retrain on your Own Data 3/3

Single unidimensional Gaussian, unbiased estimator

Data: $x_j^0 = \mu_0 + \sigma_0 Z_j$, with $Z_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{0,1}$, $1 \leq j \leq n$

Learning step:
$$\begin{cases} \mu_{t+1} &= \frac{1}{n} \sum_j \tilde{\mathbf{x}}_j^t \\ \sigma_{t+1}^2 &= \frac{1}{n-1} \sum_j \left(\tilde{\mathbf{x}}_j^t - \mu_{t+1} \right)^2 \end{cases}$$

Sampling step: $\left\{ \tilde{\mathbf{x}}_j^{t+1} = \mu_{t+1} + \sigma_{t+1} \cdot Z_j^{t+1}, \text{ with } Z_j^{t+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{0,1}, 1 \leq j \leq n \right.$

Result

$$\mathbb{E}(\sigma_t) \leq \alpha^t \sigma_0 \underset{t \to +\infty}{\longrightarrow} 0, \ 0 \leq \alpha < 1$$

Same type of results holds for a single multidimensional Gaussian

Warm Up: Only Retrain on your Own Data 3/3

Single unidimensional Gaussian, unbiased estimator

Data: $x_j^0 = \mu_0 + \sigma_0 Z_j$, with $Z_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{0,1}$, $1 \leq j \leq n$

Learning step:
$$\begin{cases} \mu_{t+1} &= \frac{1}{n} \sum_j \tilde{\mathbf{x}}_j^t \\ \sigma_{t+1}^2 &= \frac{1}{n-1} \sum_j \left(\tilde{\mathbf{x}}_j^t - \mu_{t+1} \right)^2 \end{cases}$$

Sampling step: $\left\{ \tilde{\mathbf{x}}_j^{t+1} = \mu_{t+1} + \sigma_{t+1} \cdot Z_j^{t+1}, \text{ with } Z_j^{t+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{0,1}, 1 \leq j \leq n \right.$

Result

$$\mathbb{E}(\sigma_t) \leq \alpha^t \sigma_0 \underset{t \to +\infty}{\longrightarrow} 0, \ 0 \leq \alpha < 1$$

Same type of results holds for a single multidimensional Gaussian

General Case

Iterative Retraining

$$p_0 \in \arg\max_{p \in \mathcal{P}_\Theta} \mathbb{E}_{x \sim \hat{p}_{\text{data}}} [\log p(x)]$$

$$p_{t+1} \in \arg\max_{p \in \mathcal{P}_\Theta} \underbrace{\mathbb{E}_{x \sim \hat{p}_{\text{data}}} \log p(x)}_{\text{Real data}} + \lambda \cdot \underbrace{\mathbb{E}_{\tilde{\mathbf{x}} \sim p_t} \log p(\tilde{\mathbf{x}})}_{\text{Synthetic data}} := \mathcal{G}(p_t)$$

Idea

- Fixed-point iteration $p_{t+1} = \mathcal{G}(p_t)$

General Case

Iterative Retraining

$$p_0 \in \arg\max_{p \in \mathcal{P}_\Theta} \mathbb{E}_{x \sim \hat{p}_{\text{data}}} [\log p(x)]$$

$$p_{t+1} \in \arg\max_{p \in \mathcal{P}_\Theta} \underbrace{\mathbb{E}_{x \sim \hat{p}_{\text{data}}} \log p(x)}_{\text{Real data}} + \lambda \cdot \underbrace{\mathbb{E}_{\tilde{\mathbf{x}} \sim p_t} \log p(\tilde{\mathbf{x}})}_{\text{Synthetic data}} := \mathcal{G}(p_t)$$

Idea

- Fixed-point iteration $p_{t+1} = \mathcal{G}(p_t)$
- p_{data} is a fixed point of $\mathcal{G}$
  - Study the stability of $\mathcal{G}$ around p_{data}

General Case

Iterative Retraining

$$p_0 \in \arg\max_{p \in \mathcal{P}_\Theta} \mathbb{E}_{x \sim \hat{p}_{\text{data}}} [\log p(x)]$$

$$p_{t+1} \in \arg\max_{p \in \mathcal{P}_\Theta} \underbrace{\mathbb{E}_{x \sim \hat{p}_{\text{data}}} \log p(x)}_{\text{Real data}} + \lambda \cdot \underbrace{\mathbb{E}_{\tilde{\mathbf{x}} \sim p_t} \log p(\tilde{\mathbf{x}})}_{\text{Synthetic data}} := \mathcal{G}(p_t)$$

Idea

- Fixed-point iteration $p_{t+1} = \mathcal{G}(p_t)$
- p_{data} is a fixed point of $\mathcal{G}$
  - Study the stability of $\mathcal{G}$ around p_{data}
- Link with performative prediction!

General Case

Iterative Retraining

$$p_0 \in \arg\max_{p \in \mathcal{P}_\Theta} \mathbb{E}_{x \sim \hat{p}_{\text{data}}} [\log p(x)]$$

$$p_{t+1} \in \arg\max_{p \in \mathcal{P}_\Theta} \underbrace{\mathbb{E}_{x \sim \hat{p}_{\text{data}}} \log p(x)}_{\text{Real data}} + \lambda \cdot \underbrace{\mathbb{E}_{\tilde{\mathbf{x}} \sim p_t} \log p(\tilde{\mathbf{x}})}_{\text{Synthetic data}} := \mathcal{G}(p_t)$$

Idea

- Fixed-point iteration $p_{t+1} = \mathcal{G}(p_t)$
- p_{data} is a fixed point of $\mathcal{G}$
  - Study the stability of $\mathcal{G}$ around p_{data}
- Link with performative prediction!

Retrain of Generative Models: Informal

Assumptions

- Regularity of the log-likelihood
  - $\hookrightarrow$ Local Lipschitzness and strong convexity
- The first generative model is "good enough"
  - $\hookrightarrow$ $\mathcal{W}(p_{\text{data}}, p_{\theta_0}) < \epsilon$

Retrain of Generative Models: Informal

Assumptions

- Regularity of the log-likelihood
  $\hookrightarrow$ Local Lipschitzness and strong convexity
- The first generative model is "good enough"
  $\hookrightarrow$ $\mathcal{W}(p_{\text{data}}, p_{\theta_0}) < \epsilon$
- Infinite Data

Retrain of Generative Models: Informal

Assumptions

- Regularity of the log-likelihood
  $\hookrightarrow$ Local Lipschitzness and strong convexity
- The first generative model is "good enough"
  $\hookrightarrow$ $\mathcal{W}(p_{\text{data}}, p_{\theta_0}) < \epsilon$
- Infinite Data

Results

- Regularity + good enough model + infinite data

---

[2] Q. Bertrand et al. "On the stability of iterative retraining of generative models on their own data". In: ICLR (2024).

[3] D. Ferbach et al. "Self-Consuming Generative Models with Curated Data Provably Optimize Human Preferences". In: NeurIPS (2024).

Retrain of Generative Models: Informal

Assumptions

- Regularity of the log-likelihood
  $\hookrightarrow$ Local Lipschitzness and strong convexity
- The first generative model is "good enough"
  $\hookrightarrow \mathcal{W}(p_{\text{data}}, p_{\theta_0}) < \epsilon$
- Infinite Data

Results

- Regularity + good enough model + infinite data
  $\implies$ Stability: if θ_0 is close enough to $\theta^\star$, then $\mathrm{KL}(p_{\theta^\star} | p_{\theta_t}) \to 0$ linearly[a] [b]

[a] Q. Bertrand et al. "On the stability of iterative retraining of generative models on their own data". In: *ICLR* (2024).

[b] D. Ferbach et al. "Self-Consuming Generative Models with Curated Data Provably Optimize Human Preferences". In: *NeurIPS* (2024).

Retrain of Generative Models: Informal

Assumptions

- Regularity of the log-likelihood
  - $\hookrightarrow$ Local Lipschitzness and strong convexity
- The first generative model is "good enough"
  - $\hookrightarrow$ $\mathcal{W}(p_{\text{data}}, p_{\theta_0}) < \epsilon$
- Infinite Data

Results

- Regularity + good enough model + infinite data
  - $\implies$ Stability: if θ_0 is close enough to $\theta^\star$, then $\mathrm{KL}(p_{\theta^\star} | p_{\theta_t}) \to 0$ linearly[a] [b]
  - $\hookrightarrow$ Interplay between prop. of real data and $\mathcal{W}(p_{\text{data}}, p_{\theta_0}) < \epsilon$

[a] Q. Bertrand et al. "On the stability of iterative retraining of generative models on their own data". In: *ICLR* (2024).

[b] D. Ferbach et al. "Self-Consuming Generative Models with Curated Data Provably Optimize Human Preferences". In: *NeurIPS* (2024).

Retrain of Generative Models: Informal

Assumptions

- Regularity of the log-likelihood
  - ↪ Local Lipschitzness and strong convexity
- The first generative model is "good enough"
  - ↪ $\mathcal{W}(p_{\text{data}}, p_{\theta_0}) < \epsilon$
- Infinite Data

Results

- Regularity + good enough model + infinite data
  - $\implies$ Stability: if θ_0 is close enough to $\theta^\star$, then $\mathrm{KL}(p_{\theta^\star} | p_{\theta_t}) \to 0$ linearly[a] [b]
  - ↪ Interplay between prop. of real data and $\mathcal{W}(p_{\text{data}}, p_{\theta_0}) < \epsilon$
  - ↪ Finite sample extension

---

[a] Q. Bertrand et al. "On the stability of iterative retraining of generative models on their own data". In: *ICLR* (2024).

[b] D. Ferbach et al. "Self-Consuming Generative Models with Curated Data Provably Optimize Human Preferences". In: *NeurIPS* (2024).

Retrain of Generative Models: Informal

Assumptions

- Regularity of the log-likelihood
  - ↪ Local Lipschitzness and strong convexity
- The first generative model is "good enough"
  - ↪ $\mathcal{W}(p_{\text{data}}, p_{\theta_0}) < \epsilon$
- Infinite Data

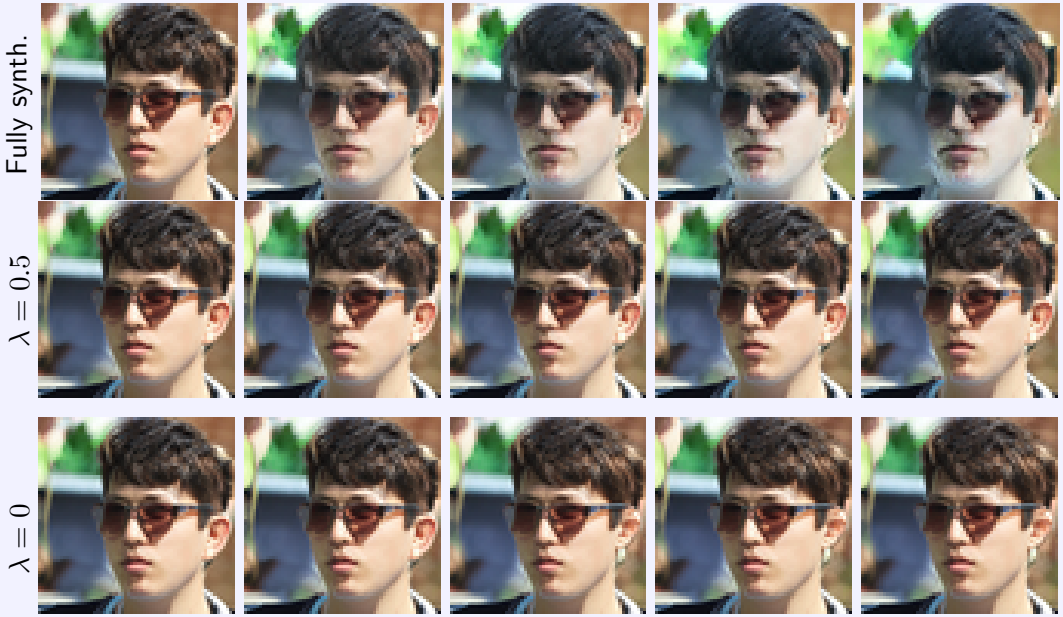
Results

- Regularity + good enough model + infinite data
  - $\implies$ Stability: if θ_0 is close enough to $\theta^\star$, then $\mathrm{KL}(p_{\theta^\star} | p_{\theta_t}) \to 0$ linearly[a] [b]
  - ↪ Interplay between prop. of real data and $\mathcal{W}(p_{\text{data}}, p_{\theta_0}) < \epsilon$
  - ↪ Finite sample extension

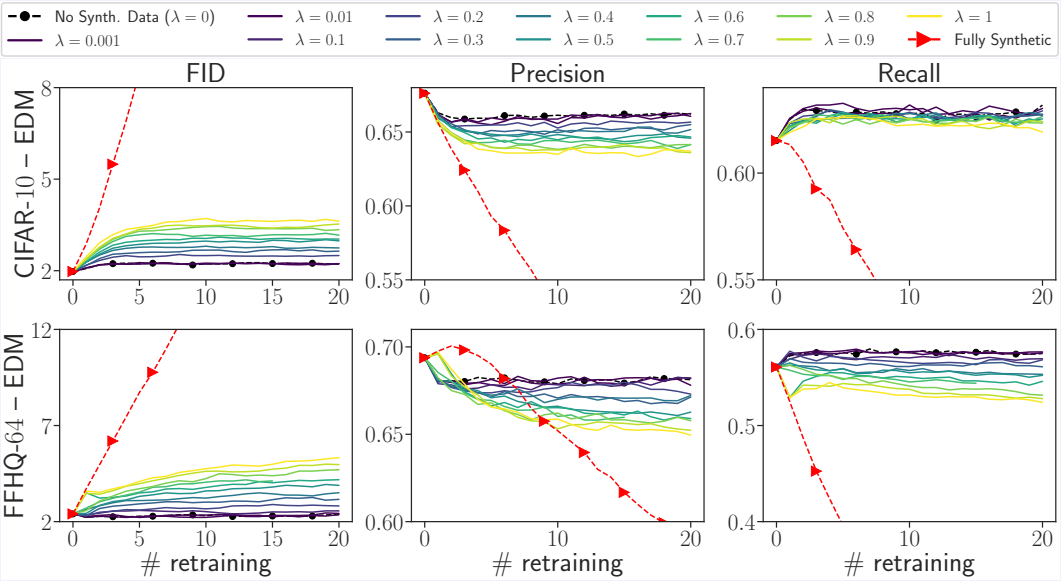
---

[a] Q. Bertrand et al. "On the stability of iterative retraining of generative models on their own data". In: *ICLR* (2024).

[b] D. Ferbach et al. "Self-Consuming Generative Models with Curated Data Provably Optimize Human Preferences". In: *NeurIPS* (2024).


Fully synth.
$\lambda = 0.5$
$\lambda = 0$
0 retrain.
5 retrain.
10 retrain.
15 retrain.
20 retrain.

Experiments



Intermediate Conclusions

Self-consuming generative models

- No collapse/MADness (if "enough" real data)[a][b]

---

[a] Q. Bertrand et al. "On the stability of iterative retraining of generative models on their own data". In: *ICLR* (2024).

[b] R. Hataya, H. Bao, and H.J. Arai. "Will Large-scale Generative Models Corrupt Future Datasets?" In: *ICCV*. 2023.

Intermediate Conclusions

Self-consuming generative models

- No collapse/MADness (if "enough" real data)[a][b]
- No improvements either

[a] Q. Bertrand et al. "On the stability of iterative retraining of generative models on their own data". In: *ICLR* (2024).

[b] R. Hataya, H. Bao, and HJ. Arai. "Will Large-scale Generative Models Corrupt Future Datasets?" In: *ICCV*. 2023.

Intermediate Conclusions

Self-consuming generative models

- No collapse/MADness (if "enough" real data)[a][b]
- No improvements either

[a] Q. Bertrand et al. "On the stability of iterative retraining of generative models on their own data". In: *ICLR* (2024).

[b] R. Hataya, H. Bao, and HJ. Arai. "Will Large-scale Generative Models Corrupt Future Datasets?" In: *ICCV*. 2023.

What if you already retrain on curated/filtered synthetic data?

Intermediate Conclusions

Self-consuming generative models

- No collapse/MADness (if "enough" real data)[a][b]
- No improvements either

[a] Q. Bertrand et al. "On the stability of iterative retraining of generative models on their own data". In: *ICLR* (2024).
[b] R. Hataya, H. Bao, and HJ. Arai. "Will Large-scale Generative Models Corrupt Future Datasets?" In: *ICCV*. 2023.

What if you already retrain on curated/filtered synthetic data?

Already Training on Curated/Filtered Data 1/2

LAION-Aesthetics[a]

[a] https://github.com/LAION-AI/laion-datasets/blob/main/laion-aesthetic.md

- Filter the LAION-5B dataset[a]
  - ↪ Filter: reward model[b]

[a] C. Schuhmann et al. "Laion-5B: An open large-scale dataset for training next generation image-text models". In: *NeurIPS* (2022).

[b] https://github.com/LAION-AI/aesthetic-predictor

[c] J. D. Pressman, K. Crowson, and Simulacra Captions Contributors. *Simulacra Aesthetic Captions*. Version 1.0. url https://github.com/JD-P/simulacra-aesthetic-captions. Stability AI, 2022.

Already Training on Curated/Filtered Data 1/2

LAION-Aesthetics[a]

[a] https://github.com/LAION-AI/laion-datasets/blob/main/laion-aesthetic.md

- Filter the LAION-5B dataset[a]
  - ↪ Filter: reward model[b]
    - ↪ Trained on synthetic data ¨Simulacra Aesthetic Captions¨ dataset[c]

[a] C. Schuhmann et al. "Laion-5B: An open large-scale dataset for training next generation image-text models". In: *NeurIPS* (2022).

[b] https://github.com/LAION-AI/aesthetic-predictor

[c] J. D. Pressman, K. Crowson, and Simulacra Captions Contributors. *Simulacra Aesthetic Captions*. Version 1.0. url https://github.com/JD-P/simulacra-aesthetic-captions. Stability AI, 2022.

Already Training on Curated/Filtered Data 1/2

LAION-Aesthetics[a]

[a] https://github.com/LAION-AI/laion-datasets/blob/main/laion-aesthetic.md

- Filter the LAION-5B dataset[a]
  - ↪ Filter: reward model[b]
    - ↪ Trained on synthetic data ¨Simulacra Aesthetic Captions¨ dataset[c]
- Filter 8M/120M sample subset of LAION-5B

[a] C. Schuhmann et al. "Laion-5B: An open large-scale dataset for training next generation image-text models". In: *NeurIPS* (2022).

[b] https://github.com/LAION-AI/aesthetic-predictor

[c] J. D. Pressman, K. Crowson, and Simulacra Captions Contributors. *Simulacra Aesthetic Captions*. Version 1.0. url https://github.com/JD-P/simulacra-aesthetic-captions. Stability AI, 2022.

Already Training on Curated/Filtered Data 1/2

LAION-Aesthetics[a]

[a] https://github.com/LAION-AI/laion-datasets/blob/main/laion-aesthetic.md

- Filter the LAION-5B dataset[a]
  - ↪ Filter: reward model[b]
    - ↪ Trained on synthetic data ¨Simulacra Aesthetic Captions¨ dataset[c]
- Filter 8M/120M sample subset of LAION-5B

[a] C. Schuhmann et al. "Laion-5B: An open large-scale dataset for training next generation image-text models". In: *NeurIPS* (2022).

[b] https://github.com/LAION-AI/aesthetic-predictor

[c] J. D. Pressman, K. Crowson, and Simulacra Captions Contributors. *Simulacra Aesthetic Captions*. Version 1.0. url https://github.com/JD-P/simulacra-aesthetic-captions. Stability AI, 2022.

Already Training on Curated/Filtered Data 2/2

JourneyDB Dataset[a]

[a] P. Junting et al. "JourneyDB: A Benchmark for Generative Image Understanding". In: *NeurIPS* (2023).

- Midjourney[a]: text-to-image model
- Midjourney "*discord*" server

[a] Midjourney. https://www.midjourney.com/home/. Version 5.2. Accessed: 2023-09-09. 2023.

Already Training on Curated/Filtered Data 2/2

JourneyDB Dataset[a]

[a] P. Junting et al. "JourneyDB: A Benchmark for Generative Image Understanding". In: *NeurIPS* (2023).

- Midjourney[a]: text-to-image model
- Midjourney "*discord*" server
  - ↪ User can request prompts

[a] Midjourney. https://www.midjourney.com/home/. Version 5.2. Accessed: 2023-09-09. 2023.

Already Training on Curated/Filtered Data $2/2$

JourneyDB Dataset[a]

[a] P. Junting et al. "JourneyDB: A Benchmark for Generative Image Understanding". In: *NeurIPS* (2023).

- Midjourney[a]: text-to-image model
- Midjourney "*discord*" server
  - ↪ User can request prompts
  - ↪ Midjourney proposes $K = 4$ images

[a] Midjourney. https://www.midjourney.com/home/. Version 5.2. Accessed: 2023-09-09. 2023.

Already Training on Curated/Filtered Data $2/2$

JourneyDB Dataset[a]

[a] P. Junting et al. "JourneyDB: A Benchmark for Generative Image Understanding". In: *NeurIPS* (2023).

- Midjourney[a]: text-to-image model
- Midjourney "*discord*" server
  - ↪ User can request prompts
  - ↪ Midjourney proposes $K = 4$ images
  - ↪ Users vote for which image to upscale

[a] Midjourney. https://www.midjourney.com/home/. Version 5.2. Accessed: 2023-09-09. 2023.

Already Training on Curated/Filtered Data $2/2$

JourneyDB Dataset[a]

[a] P. Junting et al. "JourneyDB: A Benchmark for Generative Image Understanding". In: *NeurIPS* (2023).

- Midjourney[a]: text-to-image model
- Midjourney "*discord*" server
  - ↪ User can request prompts
  - ↪ Midjourney proposes $K = 4$ images
  - ↪ Users vote for which image to upscale

[a] Midjourney. https://www.midjourney.com/home/. Version 5.2. Accessed: 2023-09-09. 2023.

Remarks

- No access to pairwise comparisons
  - ↪ Only access to the ¨winning¨ samples
  - ↪ As opposed to RHLF[a b]

[a] D. M. Ziegler et al. "Fine-tuning language models from human preferences". In: *arXiv preprint arXiv:1909.08593* (2019).

[b] R. Rafailov et al. "Direct preference optimization: Your language model is secretly a reward model". In: *NeurIPS* (2024).

Already Training on Curated/Filtered Data $2/2$

JourneyDB Dataset[a]

[a] P. Junting et al. "JourneyDB: A Benchmark for Generative Image Understanding". In: *NeurIPS* (2023).

- Midjourney[a]: text-to-image model
- Midjourney "*discord*" server
  - ↪ User can request prompts
  - ↪ Midjourney proposes $K = 4$ images
  - ↪ Users vote for which image to upscale

[a] Midjourney. https://www.midjourney.com/home/. Version 5.2. Accessed: 2023-09-09. 2023.

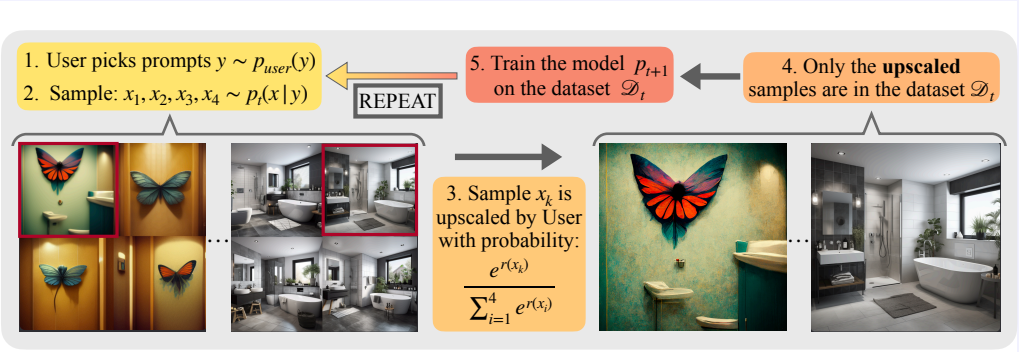
Remarks

- No access to pairwise comparisons
  - ↪ Only access to the ¨winning¨ samples
  - ↪ As opposed to RHLF[a b]

[a] D. M. Ziegler et al. "Fine-tuning language models from human preferences". In: *arXiv preprint arXiv:1909.08593* (2019).

[b] R. Rafailov et al. "Direct preference optimization: Your language model is secretly a reward model". In: *NeurIPS* (2024).

A Simple Curation Model 1/2



A Simple Curation Model $2/2$

Curation Model

- Suppose the existence of a reward model r
  $\hookrightarrow$ score $r(x)$ to each sample x
- $$\text{Sample } \tilde{\mathbf{x}}_1 \sim p_t, \ldots, \tilde{\mathbf{x}}_K \sim p_t \text{ , } i.i.d.$$
- $$\text{Pick } \hat{\mathbf{x}} \sim \mathcal{BT}(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_K) \text{ , } i.e.,$$
$$\mathbb{P}(\hat{\mathbf{x}} = \tilde{\mathbf{x}}_k | \tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_K) = \frac{e^{r(\tilde{\mathbf{x}}_k)}}{\sum_{j=1}^{K} e^{r(\tilde{\mathbf{x}}_j)}} \text{ , } 1 \leq k \leq K$$

Iterative Retraining

$$p_{t+1} = \arg\max_{p \in \mathcal{P}} \mathbb{E}_{x \sim \hat{p}_{\text{data}}} \Big[\log p(x) \Big] + \lambda \cdot \mathbb{E}_{\substack{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_K \sim p_t \\ \hat{\mathbf{x}} \sim \mathcal{BT}(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_K)}} \Big[\log p(\hat{\mathbf{x}}) \Big]$$

A Simple Curation Model 2/2

Curation Model

- Suppose the existence of a reward model r
  $\hookrightarrow$ score $r(x)$ to each sample x
- $$\text{Sample } \tilde{\mathbf{x}}_1 \sim p_t, \ldots, \tilde{\mathbf{x}}_K \sim p_t \text{ , } i.i.d.$$
- $$\text{Pick } \hat{\mathbf{x}} \sim \mathcal{BT}(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_K) \text{ , } i.e.,$$
$$\mathbb{P}(\hat{\mathbf{x}} = \tilde{\mathbf{x}}_k | \tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_K) = \frac{e^{r(\tilde{\mathbf{x}}_k)}}{\sum_{j=1}^{K} e^{r(\tilde{\mathbf{x}}_j)}} \text{ , } 1 \leq k \leq K$$

Iterative Retraining

$$p_{t+1} = \arg\max_{p \in \mathcal{P}} \mathbb{E}_{x \sim \hat{p}_{\text{data}}} \Big[\log p(x) \Big] + \lambda \cdot \mathbb{E}_{\substack{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_K \sim p_t \\ \hat{\mathbf{x}} \sim \mathcal{BT}(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_K)}} \Big[\log p(\hat{\mathbf{x}}) \Big]$$

Self-Consuming Loop

$$p_{t+1} = \arg\max_{p \in \mathcal{P}} \mathbb{E}_{x \sim \hat{p}_{\text{data}}} \Big[\log p(x) \Big] + \lambda \cdot \mathbb{E}_{\substack{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_K \sim p_t \\ \hat{\mathbf{x}} \sim \mathcal{BT}(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_K)}} \Big[\log p(\hat{\mathbf{x}}) \Big]$$

Q: What will happen?

Self-Consuming Loop

$$p_{t+1} = \arg\max_{p \in \mathcal{P}} \mathbb{E}_{x \sim \hat{p}_{\text{data}}} \Big[\log p(x) \Big] + \lambda \cdot \mathbb{E}_{\substack{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_K \sim p_t \\ \hat{\mathbf{x}} \sim \mathcal{BT}(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_K)}} \Big[\log p(\hat{\mathbf{x}}) \Big]$$

Q: What will happen?

Results

$$\mathbb{E}_{p_t} \Big[e^{r(x)} \Big] \xrightarrow{t \to \infty} e^{r_*} \quad \text{and} \quad \mathrm{Var}_{p_t} \Big[e^{r(x)} \Big] \xrightarrow{t \to \infty} 0 \ .$$

Self-Consuming Loop

$$p_{t+1} = \arg\max_{p \in \mathcal{P}} \mathbb{E}_{x \sim \hat{p}_{\text{data}}} \Big[\log p(x) \Big] + \lambda \cdot \mathbb{E}_{\substack{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_K \sim p_t \\ \hat{\mathbf{x}} \sim \mathcal{BT}(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_K)}} \Big[\log p(\hat{\mathbf{x}}) \Big]$$

Q: What will happen?

Results

$$\mathbb{E}_{p_t} \Big[e^{r(x)} \Big] \xrightarrow{t \to \infty} e^{r_*} \quad \text{and} \quad \text{Var}_{p_t} \Big[e^{r(x)} \Big] \xrightarrow{t \to \infty} 0 \ .$$

Other Results

- Equivalent to do RLHF if $K \to \infty$
- Can be extended with a mix of real and synthetic data

Self-Consuming Loop

$$p_{t+1} = \arg\max_{p \in \mathcal{P}} \mathbb{E}_{x \sim \hat{p}_{\text{data}}} \Big[\log p(x) \Big] + \lambda \cdot \mathbb{E}_{\substack{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_K \sim p_t \\ \hat{\mathbf{x}} \sim \mathcal{BT}(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_K)}} \Big[\log p(\hat{\mathbf{x}}) \Big]$$

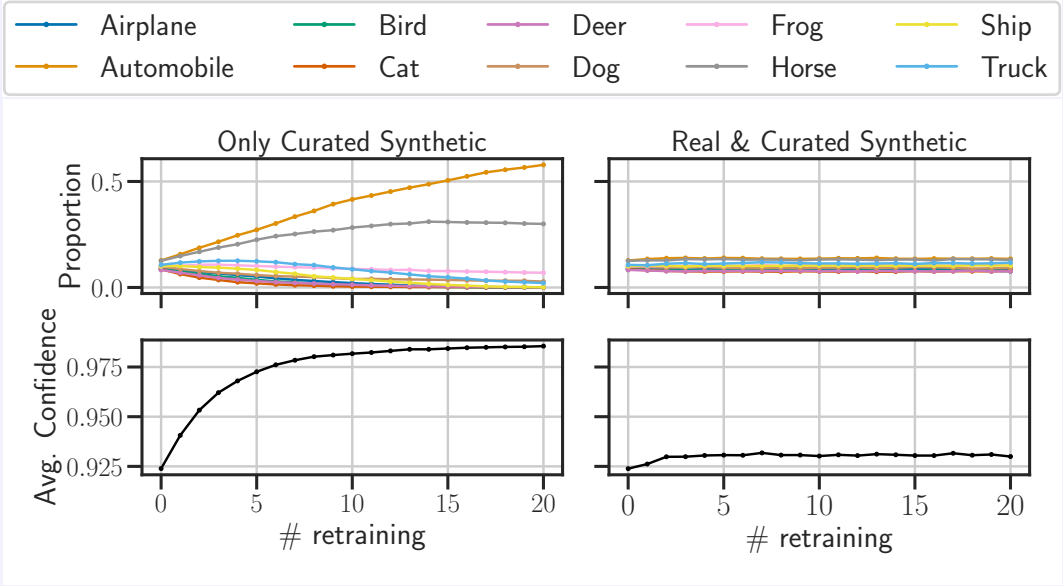
Q: What will happen?

Results

$$\mathbb{E}_{p_t} \Big[e^{r(x)} \Big] \xrightarrow{t \to \infty} e^{r_*} \quad \text{and} \quad \text{Var}_{p_t} \Big[e^{r(x)} \Big] \xrightarrow{t \to \infty} 0 \ .$$

Other Results

- Equivalent to do RLHF if $K \to \infty$
- Can be extended with a mix of real and synthetic data



- reward(x) = confidence of a pretrained classifier for the image x
- $\lambda = 1/2$

Conclusion and Future Work

Future Work

- Filtering without Human-Feedback?
  - ↪ Score per sample? Downstream-task specific?

---

[a] M. Jiralerspong et al. "Feature Likelihood Score: Evaluating Generalization of Generative Models Using Samples". In: *NeurIPS* (2023).

[b] R. A. Hemmat et al. "Feedback-guided Data Synthesis for Imbalanced Classification". In: *TMLR* (2024).

[c] T. Karras et al. "Guiding a Diffusion Model with a Bad Version of Itself". In: *arXiv preprint arXiv:2406.02507* (2024).

[d] S. Alemohammad et al. "Self-Improving Diffusion Models with Synthetic Data". In: *arXiv preprint arXiv:2408.16333* (2024).

Conclusion and Future Work

Future Work

- Filtering without Human-Feedback?
  - ↪ Score per sample? Downstream-task specific?
    - ↪ Feature Likelihood Score (FLS)[a]
    - ↪ Classifier to score the samples

[a] M. Jiralerspong et al. "Feature Likelihood Score: Evaluating Generalization of Generative Models Using Samples". In: *NeurIPS* (2023).

[b] R. A. Hemmat et al. "Feedback-guided Data Synthesis for Imbalanced Classification". In: *TMLR* (2024).

[c] T. Karras et al. "Guiding a Diffusion Model with a Bad Version of Itself". In: *arXiv preprint arXiv:2406.02507* (2024).

[d] S. Alemohammad et al. "Self-Improving Diffusion Models with Synthetic Data". In: *arXiv preprint arXiv:2408.16333* (2024).

Conclusion and Future Work

Future Work

- Filtering without Human-Feedback?
  - ↪ Score per sample? Downstream-task specific?
    - ↪ Feature Likelihood Score (FLS)[a]
    - ↪ Classifier to score the samples
      - ↪ Correlation between accuracy and sample quality?[b]

[a] M. Jiralerspong et al. "Feature Likelihood Score: Evaluating Generalization of Generative Models Using Samples". In: *NeurIPS* (2023).

[b] R. A. Hemmat et al. "Feedback-guided Data Synthesis for Imbalanced Classification". In: *TMLR* (2024).

[c] T. Karras et al. "Guiding a Diffusion Model with a Bad Version of Itself". In: *arXiv preprint arXiv:2406.02507* (2024).

[d] S. Alemohammad et al. "Self-Improving Diffusion Models with Synthetic Data". In: *arXiv preprint arXiv:2408.16333* (2024).

Conclusion and Future Work

Future Work

- Filtering without Human-Feedback?
  - ↪ Score per sample? Downstream-task specific?
    - ↪ Feature Likelihood Score (FLS)[a]
    - ↪ Classifier to score the samples
      - ↪ Correlation between accuracy and sample quality?[b]
  - ↪ Score per distribution?
    - ↪ Computationally intensive

[a] M. Jiralerspong et al. "Feature Likelihood Score: Evaluating Generalization of Generative Models Using Samples". In: *NeurIPS* (2023).

[b] R. A. Hemmat et al. "Feedback-guided Data Synthesis for Imbalanced Classification". In: *TMLR* (2024).

[c] T. Karras et al. "Guiding a Diffusion Model with a Bad Version of Itself". In: *arXiv preprint arXiv:2406.02507* (2024).

[d] S. Alemohammad et al. "Self-Improving Diffusion Models with Synthetic Data". In: *arXiv preprint arXiv:2408.16333* (2024).

Conclusion and Future Work

Future Work

- Filtering without Human-Feedback?
  - ↪ Score per sample? Downstream-task specific?
    - ↪ Feature Likelihood Score (FLS)[a]
    - ↪ Classifier to score the samples
      - ↪ Correlation between accuracy and sample quality?[b]
  - ↪ Score per distribution?
    - ↪ Computationally intensive
  - ↪ Use bad samples/models to improve[c,d]

---

[a] M. Jiralerspong et al. "Feature Likelihood Score: Evaluating Generalization of Generative Models Using Samples". In: *NeurIPS* (2023).

[b] R. A. Hemmat et al. "Feedback-guided Data Synthesis for Imbalanced Classification". In: *TMLR* (2024).

[c] T. Karras et al. "Guiding a Diffusion Model with a Bad Version of Itself". In: *arXiv preprint arXiv:2406.02507* (2024).

[d] S. Alemohammad et al. "Self-Improving Diffusion Models with Synthetic Data". In: *arXiv preprint arXiv:2408.16333* (2024).

Conclusion and Future Work

Future Work

- Filtering without Human-Feedback?
  - ↪ Score per sample? Downstream-task specific?
    - ↪ Feature Likelihood Score (FLS)[a]
    - ↪ Classifier to score the samples
      - ↪ Correlation between accuracy and sample quality?[b]
  - ↪ Score per distribution?
    - ↪ Computationally intensive
  - ↪ Use bad samples/models to improve[c,d]

---

[a] M. Jiralerspong et al. "Feature Likelihood Score: Evaluating Generalization of Generative Models Using Samples". In: *NeurIPS* (2023).

[b] R. A. Hemmat et al. "Feedback-guided Data Synthesis for Imbalanced Classification". In: *TMLR* (2024).

[c] T. Karras et al. "Guiding a Diffusion Model with a Bad Version of Itself". In: *arXiv preprint arXiv:2406.02507* (2024).

[d] S. Alemohammad et al. "Self-Improving Diffusion Models with Synthetic Data". In: *arXiv preprint arXiv:2408.16333* (2024).

Conclusion and Future Work

Future Work

- Filtering without Human-Feedback?
  - ↪ Score per sample? Downstream-task specific?
    - ↪ Feature Likelihood Score (FLS)[a]
    - ↪ Classifier to score the samples
      - ↪ Correlation between accuracy and sample quality?[b]
  - ↪ Score per distribution?
    - ↪ Computationally intensive
  - ↪ Use bad samples/models to improve[c,d]

---

[a] M. Jiralerspong et al. "Feature Likelihood Score: Evaluating Generalization of Generative Models Using Samples". In: *NeurIPS* (2023).

[b] R. A. Hemmat et al. "Feedback-guided Data Synthesis for Imbalanced Classification". In: *TMLR* (2024).

[c] T. Karras et al. "Guiding a Diffusion Model with a Bad Version of Itself". In: *arXiv preprint arXiv:2406.02507* (2024).

[d] S. Alemohammad et al. "Self-Improving Diffusion Models with Synthetic Data". In: *arXiv preprint arXiv:2408.16333* (2024).

Thank You!

- Alemohammad, S. et al. “Self-Consuming Generative Models Go MAD”. In: *ICLR* (2024).
- Alemohammad, S. et al. “Self-Improving Diffusion Models with Synthetic Data”. In: *arXiv preprint arXiv:2408.16333* (2024).
- Bertrand, Q. et al. “On the stability of iterative retraining of generative models on their own data”. In: *ICLR* (2024).
- Ferbach, D. et al. “Self-Consuming Generative Models with Curated Data Provably Optimize Human Preferences”. In: *NeurIPS* (2024).
- Gulcehre, C. et al. “Reinforced self-training (REST) for language modeling”. In: (2023). arXiv: 2308.08998 [cs.CL].
- Hataya, R., H. Bao, and HJ. Arai. “Will Large-scale Generative Models Corrupt Future Datasets?” In: *ICCV*. 2023.
- Hemmat, R. A. et al. “Feedback-guided Data Synthesis for Imbalanced Classification”. In: *TMLR* (2024).
- Ho, J., A. Jain, and P. Abbeel. “Denoising diffusion probabilistic models”. In: *NeurIPS* (2020).

- Jiralerspong, M. et al. “Feature Likelihood Score: Evaluating Generalization of Generative Models Using Samples”. In: *NeurIPS* (2023).
- Junting, P. et al. “JourneyDB: A Benchmark for Generative Image Understanding”. In: *NeurIPS* (2023).
- Karras, T. et al. “Guiding a Diffusion Model with a Bad Version of Itself”. In: *arXiv preprint arXiv:2406.02507* (2024).
- Midjourney. https://www.midjourney.com/home/. Version 5.2. Accessed: 2023-09-09. 2023.
- Pressman, J. D., K. Crowson, and Simulacra Captions Contributors. *Simulacra Aesthetic Captions*. Version 1.0. url https://github.com/JD-P/simulacra-aesthetic-captions. Stability AI, 2022.
- Rafailov, R. et al. “Direct preference optimization: Your language model is secretly a reward model”. In: *NeurIPS* (2024).
- Schuhmann, C. et al. “Laion-5B: An open large-scale dataset for training next generation image-text models”. In: *NeurIPS* (2022).
- Shumailov, I. et al. “The Curse of Recursion: Training on Generated Data Makes Models Forget”. In: (2023). arXiv: 2305.17493 [cs.LG].

- Wang, Z. et al. “Better diffusion models further improve adversarial training”. In: *ICML*. 2023.
- Ziegler, D. M. et al. “Fine-tuning language models from human preferences”. In: *arXiv preprint arXiv:1909.08593* (2019).