Fine-tuning Stable Diffusion for using DreamBooth

...

Tomasz - Rory - Brodie - Hank

Objective

- Generate images with specific styles or objects
  - Eg. images from within movie scenes
  - Images with known faces
- Traditional Approach
  - Standard GANs/Diffusion
  - Text-to-Image (eg. Stable Diffusion)
  - Can’t transfer generated objects to new contexts
  - Struggle with specificity
- Solution
  - Fine Tuning Stable Diffusion
    - Very little data required
    - Specificity learned quickly
  - DreamBooth

Motivation

Traditional Approaches for *specific* image generation

- GANs
  - Difficult to train
  - Generally domain specific
  - Can't transfer images well
  - Mode collapse – generator starts producing same image that fools discriminator
- Diffusion Models
  - Struggle with specificity
  - Can't preserve objects to new contexts
  - Too general
- In general you have to choose object consistency, or context portability
  - Dreambooth allows both

Ruiz et. al (2022)

Example

Input Image

Stable Diffusion output Finetuned with DreamBooth

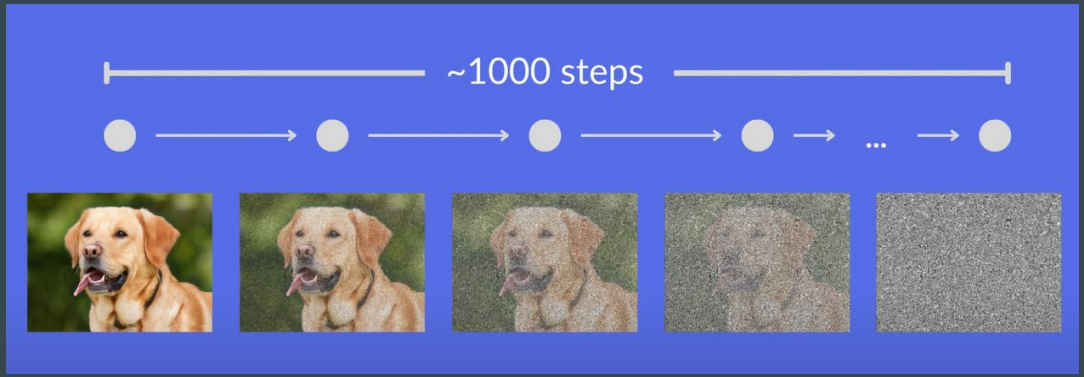
Can this be applied to the surrounding environment/style?

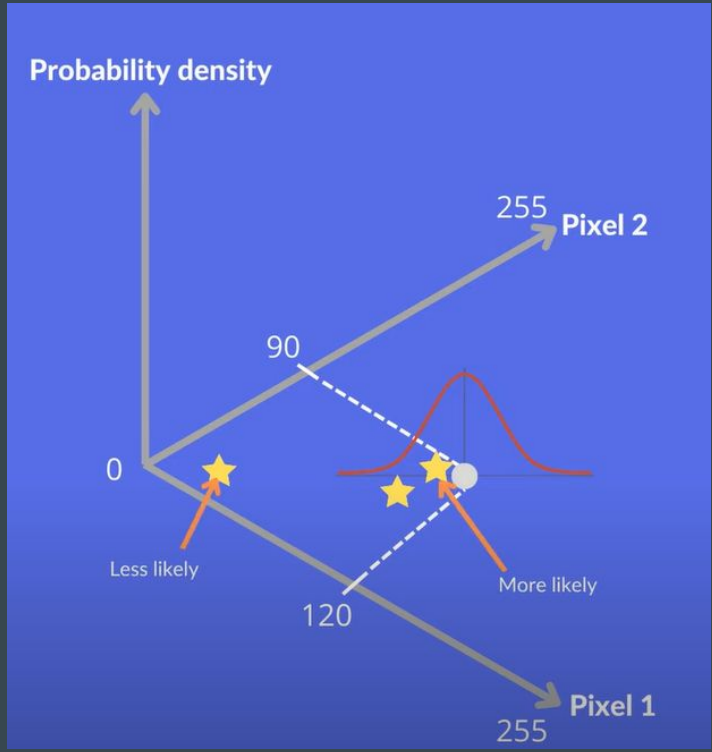
How does diffusion actually work?

...

Forward Process: Adding Gaussian Noise

- Markov Chain
- Ex: image of only 2 pixels (120, 90)
- Draw r.v. from Normal distribution (115, 84) → new value
- Little change at each timestep

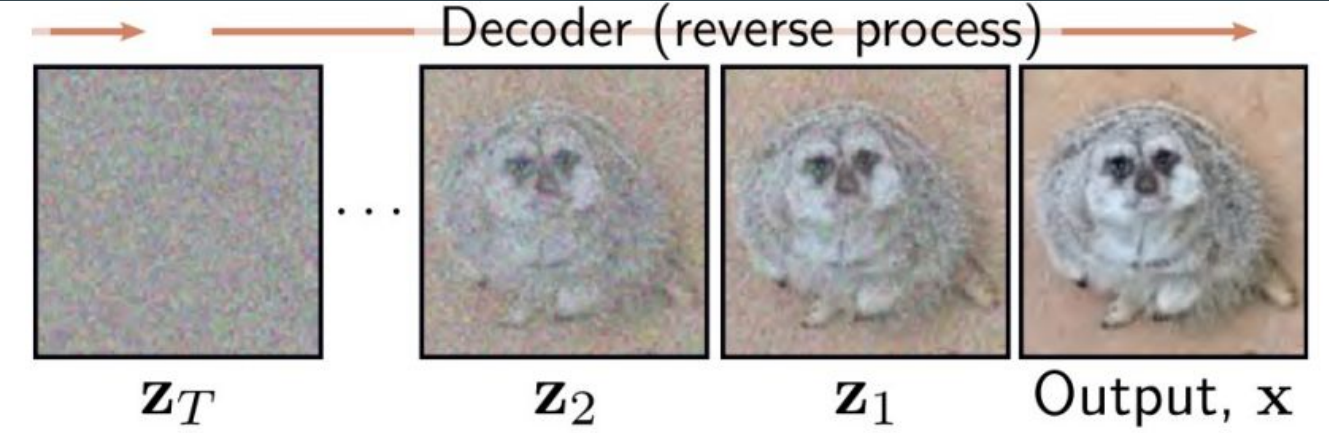




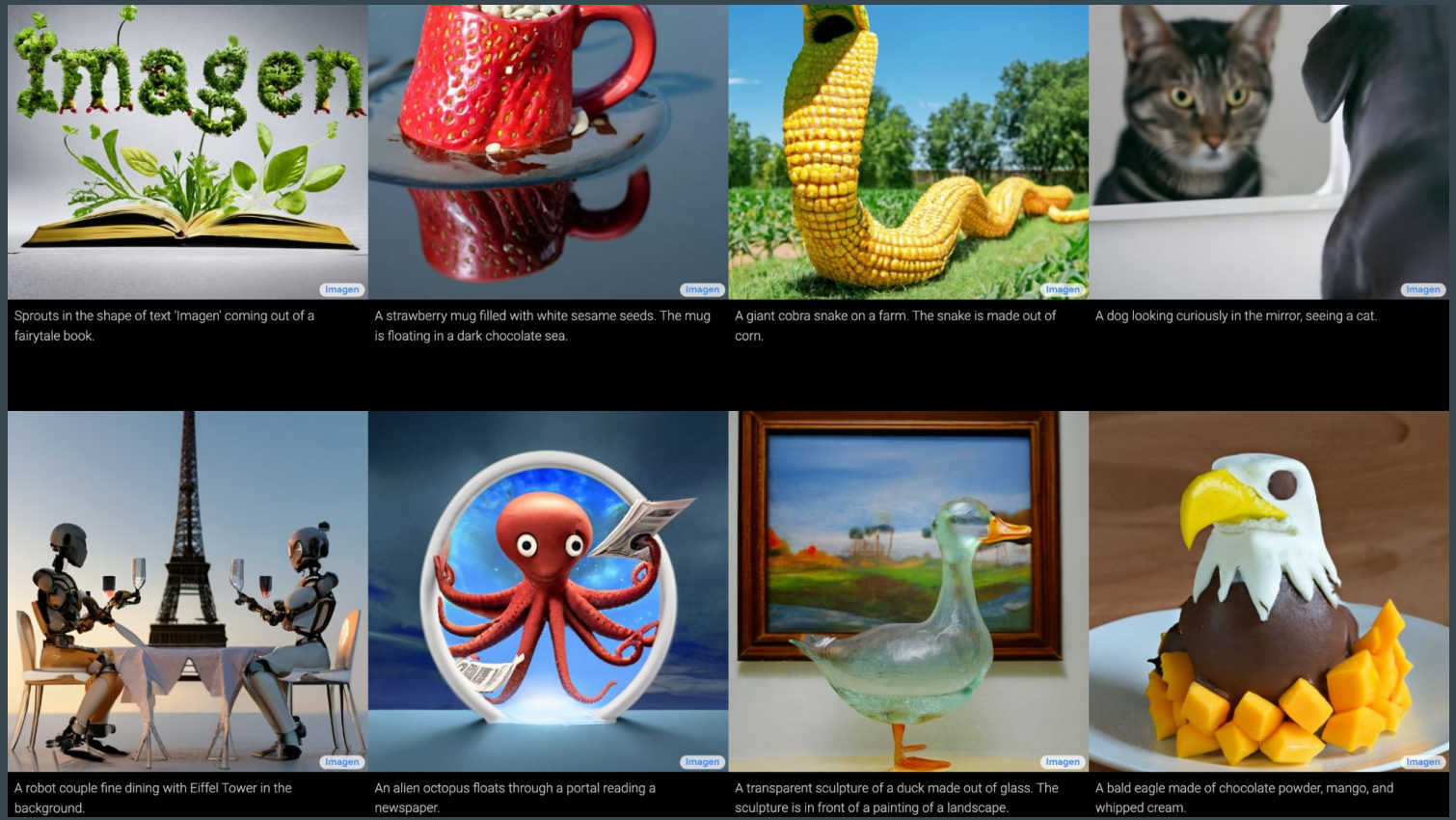
Images from AssemblyAI via YouTube

Reverse Process: Denoising

- Model learns to predict reverse of noise addition, iteratively reducing noise in img
- Predicts amount of noise added → subtract from current img → less noisy img
- Parameters (ie. temperature) can be adjusted to balance between diversity and fidelity of generated images, influencing their coherence and creativity



How does diffusion perform?



From Imagen, a large scale diffusion model released by Google. Does a good job generating various complex or impossible prompts. (Cobra made out of corn, or dog seeing a cat in the reflection of the mirror)

Stable Diffusion - An Overview

Stable Diffusion is a large latent text-to-image generation model published in August 2022, it expands on and offers the following advantages over simple diffusion:

- Runs diffusion process in latent space rather than pixel space
- Low training cost, better inference speeds
- Open-source, lightweight model
- Improved stability, and image quality
- Trains on 512x512 images, with 860M UNet

Stable Diffusion XL

Further **EXPANSION** on the Stable Diffusion model was released in July 2023

Contains over 3.5 billion parameters, over 3x more than any previous stable model:

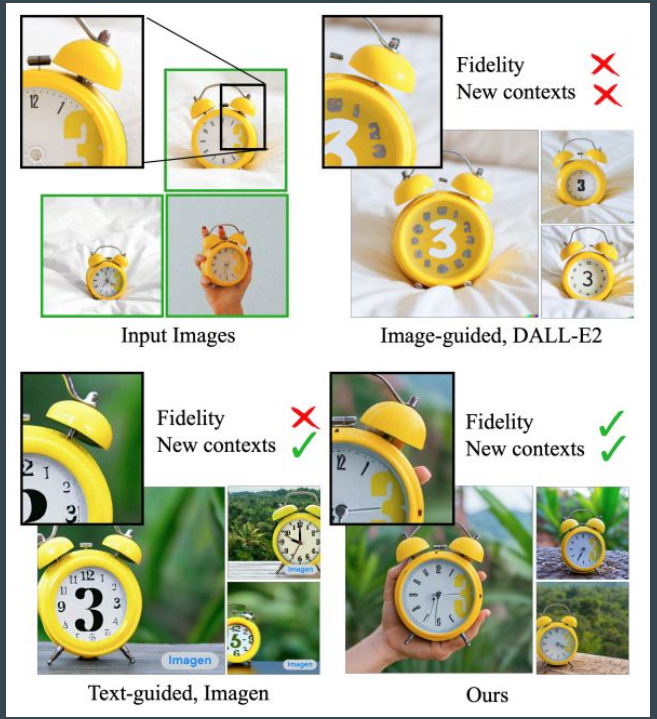
- Consistently generates higher quality, more photorealistic images
- Better generates human hands, faces, and legible text
- 1024x1024 px images, and a 3x larger UNet backbone than stable diffusion
- Covers a wider range of visual styles, and better detects fine image features
- Operates better on short prompts
- State-of-the-Art performance on benchmark datasets

Diffusion notes and Fine-tuning intuition

These models can suffer in performance when asked to generate images of a certain style, or to mimic the appearance of objects in a given reference set.

With a standard Text-2-Image model, when given a subject image to generate stylistically from, it is only able to create variations of the image content, instead of novel reconstructions.

Instead, introducing specific images into the models capacity requires finetuning



DreamBooth – High Level Overview

PROBLEM with Text-to-Image Diffusion Models:

"yellow tiny dog" → multiple correct outputs

- Model isn't designed to store a certain one
- No matter how detailed prompt is!

ENTER DREAMBOOTH:

- Store a constant subject as a token [V]

  "a [V] dog in <environment>"

- Now, insert constant subject [V] across multiple diffusions without describing!
  - With only 3-5 subject images

[V] + <Class Name> + <Environment>

- ***[V]*** – unique identifier that language model has weak prior for → add [V] to dictionary
- ***Class Name*** – model can use prior knowledge about class (ie. dog)
  - Without → increase in training time, worse model performance, language drift

“a [V] man standing in the desert in front of several camels and mountains”

The Delicate Art of Fine-Tuning DreamBooth

Assumption:

Maximize subject fidelity by fine-tuning *all layers* of model

Problems:

1. Language Drift: hyper specific task → LLM loses semantic & syntactic knowledge
2. Reduced Output Diversity: 3-5 imgs → reduced variability in poses and views

Solution:

Supervise fine-tune using prior-generated img's→ output diversity + no language drift

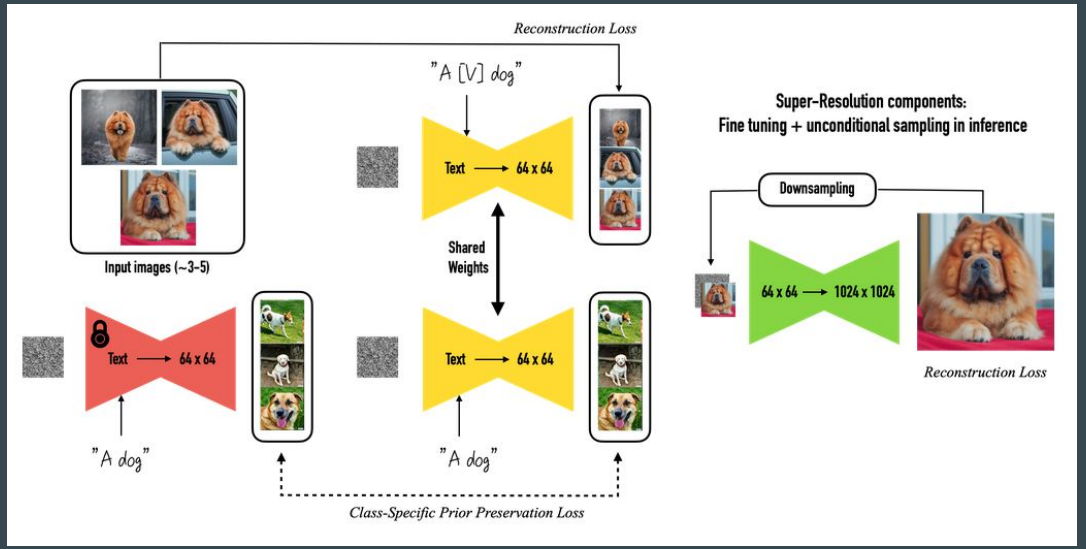
DreamBooth Internals

Solve:

1. Language drifting
2. Overfitting

Solution:

autogenous class-specific prior preservation loss



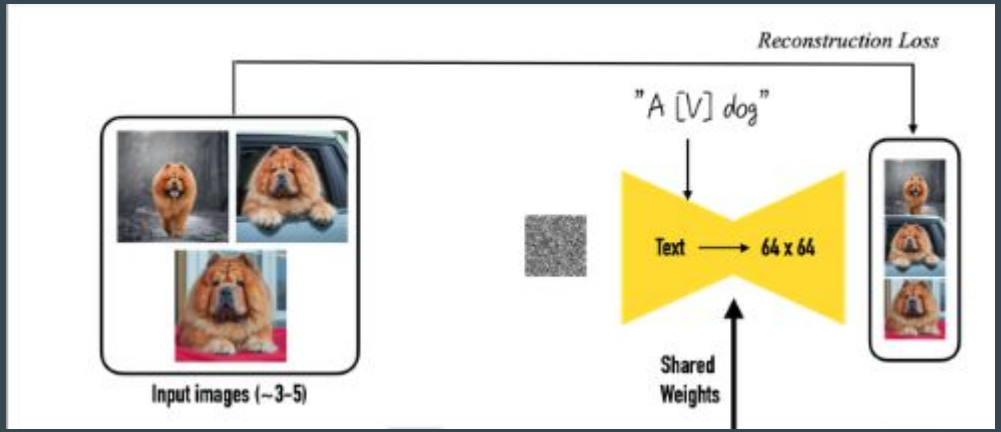
$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},\boldsymbol{\epsilon}',t}\left[w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\sigma_t\boldsymbol{\epsilon},\mathbf{c})-\mathbf{x}\|_2^2+\lambda w_{t'}\|\hat{\mathbf{x}}_\theta(\alpha_{t'}\mathbf{x}_{\text{pr}}+\sigma_{t'}\boldsymbol{\epsilon}',\mathbf{c}_{\text{pr}})-\mathbf{x}_{\text{pr}}\|_2^2\right],$$

DreamBooth Internals

Diffusion Loss

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},t}\left[w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\sigma_t\boldsymbol{\epsilon},\mathbf{c})-\mathbf{x}\|_2^2\right]$$

Motive: Fine-tuning



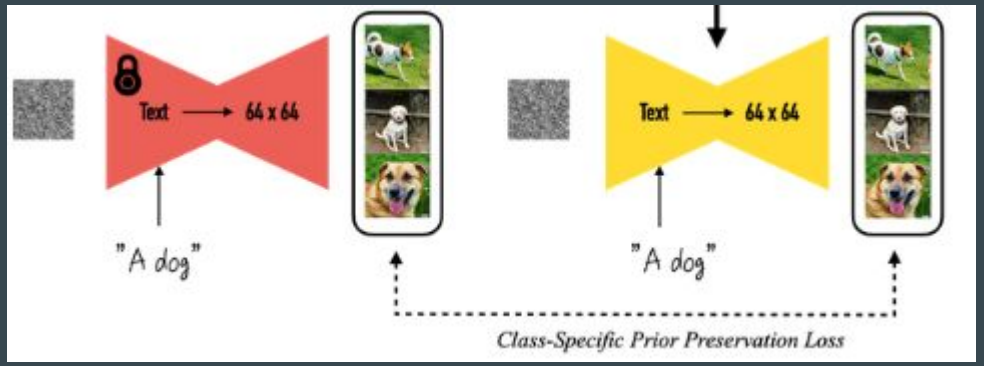
DreamBooth Internals

Class-Specific Prior Preservation Loss

$$w_{t'} \left\| \hat{\mathbf{x}}_\theta(\alpha_{t'}\mathbf{x}_{\text{pr}} + \sigma_{t'}\boldsymbol{\epsilon}', \mathbf{c}_{\text{pr}}) - \mathbf{x}_{\text{pr}} \right\|_2^2]$$

Motive:

Preserves the prior info



Autogenous Class-Specific Prior Preservation Loss

Equation

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},\boldsymbol{\epsilon}',t}\left[w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\sigma_t\boldsymbol{\epsilon},\mathbf{c})-\mathbf{x}\|_2^2+\lambda w_{t'}\|\hat{\mathbf{x}}_\theta(\alpha_{t'}\mathbf{x}_{\text{pr}}+\sigma_{t'}\boldsymbol{\epsilon}',\mathbf{c}_{\text{pr}})-\mathbf{x}_{\text{pr}}\|_2^2\right],$$

Motive

preserving the generative model's ability to produce diverse and high-quality outputs across its original range of classes, while also adapting it to generate new, specific types of content accurately.

training and results

1000 iterations, lambda 1 , lr 10 -5 (section 3)

experiments section 4

comparison with Textural Inversion

evaluation metric:

1. Subject fidelity: DINO instead of CLIP (cosine similarity) to encourage spotting difference between objects of same class (the whole point of DreamBooth)
2. prompt fidelity : cosine similarity between prompt and image CLIP embedding

Custom Subjects

Training Images:

- Stable Diffusion v1-5
- Diffusers Library and DreamBooth training script
- Nvidia A100 on colab
- Trained for 3 minutes
- 400 Training Steps
- Also finetuned text encoder
    - Very important for faces

Subject Results

A photo of Rory in the desert

A photo of Rory standing in the desert in front of several camels and mountains

A photo of Rory frowning in the desert in front of large snow capped mountains

A photo of Rory holding a trophy in times square but hes not smiling

Application to Environments

Future Plans

- Instead of introducing a custom object, introduce a custom environment
  - Eg. create image in the style of a certain movie
- Investigate Textual Inversion
- More experimentation with Stable Diffusion XL

Questions?