DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

Nataniel Ruiz\textsuperscript{1,2}, Yuanzhen Li\textsuperscript{1}, Varun Jampani\textsuperscript{1}, Yael Pritch\textsuperscript{1}, Michael Rubinstein\textsuperscript{1}, and Kfir Aberman\textsuperscript{1}

\textsuperscript{1}Google Research \hspace{1cm} \textsuperscript{2}Boston University

Presented By: Shane Davis, Joe Fioresi, Mitchell Klingler, & Nyle Siddiqui
Agenda

- Introduction
- Research gap
- High-level methodology
- Solution specifics
- Results/Examples
- Ablations
- Limitations
- Experimentation
Few-Shot Image Generation

Pre-trained generator
(large source domain)

Identified Gaps

● Most text-to-image generation models are unable to perform few-shot learning
  ○ There is no efficient method to “inject” subjects into the model training
    ■ Overfitting
    ■ Language Drift
    ■ Best ex.: GAN reproducing the same face given ~100 images [1]

Contributions

- New efficient few-shot **fine-tuning technique** that preserves semantic class knowledge
  - Tuning requires only 3-5 examples
- Created new problem to explore diffusion model capabilities – subject-driven generation
  - Goal: maintain fidelity in new contexts
Improvements over Existing Work

“retro style yellow alarm clock with a white clock face and a yellow number three on the right part of the clock face in the jungle”

Subject-Driven Generation – Personalization

Input images

Methodology Overview

- Few-Shot text-guided diffusion
  - Fine-tune existing diffusion models
Identified Gaps

- Most text-to-image generation models are unable to perform few-shot learning
  - There is no efficient method to “inject” subjects into the model training
    - Overfitting
    - Language Drift
    - Best ex.: GAN reproducing the same face given ~100 images [1]

Rare-Token Identifier

- How to refer to a new subject?
  - Image is straightforward
- Diffusion model already has a ‘vocabulary’
  - Goal: Implant the subject into this vocabulary
Rare-Token Identifier: Naive Approaches

- Use English words to describe the new class
  - English lexicon has strong priors

- Generate a string of random characters
  - Nonsense identifier leads to literal artifacts

“unique” or “special”

(e.g. “xy5xyt00”)
Rare-Token Identifier: Solution

1. Perform a rare-token lookup in the vocabulary
2. Invert the rare-token, resulting in plain text (1-3 characters)
3. Use this plain text as the unique identifier
Identified Gaps

- Most text-to-image generation models are unable to perform few-shot learning
  - There is no efficient method to “inject” subjects into the model training
    - **Overfitting**
    - **Language Drift**
  - Best ex.: GAN reproducing the same face given ~100 images [1]

Autogenous Class-Specific Prior Preservation Loss

1. Overfitting

1. Language Drift
Autogenous Class-Specific Prior Preservation Loss

\[
\mathbb{E}_{x, c, \varepsilon, \varepsilon', t} \left[ w_t \| \hat{x}_\theta (\alpha_t x + \sigma_t \varepsilon, c) - x \|_2^2 \right] + \lambda w_{t'} \| \hat{x}_\theta (\alpha_{t'} x_{pr} + \sigma_{t'} \varepsilon', c_{pr}) - x_{pr} \|_2^2
\]
Results

"A [V] dog in the beach"
Results: Recontextualization

A [V] backpack in the Grand Canyon

A [V] vase in the ocean

A [V] teapot pouring tea
Results: Art Renditions

“A painting of a [V] dog in the style of Vincent Van Gogh”

“A sculpture of a [V] dog in the style of Michelangelo

“A painting of a [V] dog in the style of Leonardo Da Vinci”
Results: Expression Manipulation

“A [state] [V] dog”

depressed

screaming

barking

sleeping
Results: View Synthesis

“[V] cat seen from the [direction]”
Results: Accessorization

Input images

“A [V] dog wearing a [accessory]”
Results: Property Modification

“A [color] [V] car”

purple

“A cross of a [V] dog and a [target species]”

hippo
Ablations - class noun prior

Input images

Fine-tuning

No class noun: "A [V]"

Incorrect class noun: "A [V] dog"

Correct class noun: "A [V] sunglasses"

Inference

A [V]

A [V] on top of blue fabric

A [V] with a river in the background

A [V] dog

A [V] dog on top of blue fabric

A [V] dog with a river in the background

A [V] sunglasses

A [V] sunglasses on top of blue fabric

A [V] sunglasses with a river in the background
Main Failures

(a) Incorrect context synthesis
- [...] in the ISS
- [...] on the moon

(b) Context-appearance entanglement
- [...] in the Bolivian salt flats
- [...] on top of a blue fabric

(c) Overfitting
- [...] in the forest
Limitations

- Model learned certain subjects better than others
- Image noise ("hallucinated features") on the result images
- No suggestions on preventing malicious usage
- No quantitative results to compare with other models
Experimentation

- Used Hugging Face’s [diffusers](https://huggingface.co) implementation of Dreambooth
  - Stable Diffusion v1.5
  - Rare Word Token Chosen: “sks” ([List](https://example.com))
- Followed recommended hyperparameters
Results - Joe’s Dog

Input Images:

Regularization Images (96 total):
“A photo of a dog”

“A photo of a sks dog at disney world”:

“A photo of a sks dog running in the forest”: 
Results - Knightro

Input Images:

Regularization Images (200 total):
“A photo of a knight”

“A photo of a sks knight at the beach”:

“A photo of a sks knight riding a bicycle”: 
Questions?