Improving Caption Similarity with GLIDE Generative Models

GM2
Harshil Bhullar
Manik Narang
William Santosa
Shaira Alam

Introduction

Brief Recap of GLIDE

"a hedgehog using a calculator"

"a corgi wearing a red bowtie and a purple party hat"

- An advanced AI model developed by OpenAI
- Combines language understanding with diffusion models for high-quality, photorealistic image generation
- Basically, text prompts are given to the model and it outputs an image based off the text provided
- Both CLIP and CFGs were used in the paper
- Important because it increased the amount of "texture" and detail relative to prior diffusion models

Goals & Objectives

Finetune on top of the original GLIDE model to increase accuracy.

Increase caption similarity when generating images

Why are we doing this?

Not all text details were accurate to caption

Caption: "small brown fluffy dog with black tongue standing on grass with autumn leaves on the ground" was clearly not accurate

Original "Real" Image

GLIDE Image

Methodology

OpenAI glide-text2im API

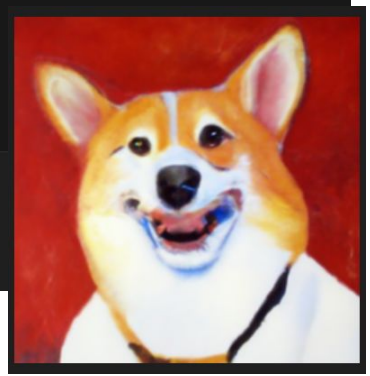
Input: Prompt

```
# Sampling parameters
prompt = "an oil painting of a corgi"
batch_size = 1
guidance_scale = 3.0

# Tune this parameter to control the sharpness of 256x256 images.
# A value of 1.0 is sharper, but sometimes results in grainy artifacts.
upsample_temp = 0.997
```

Output: Image

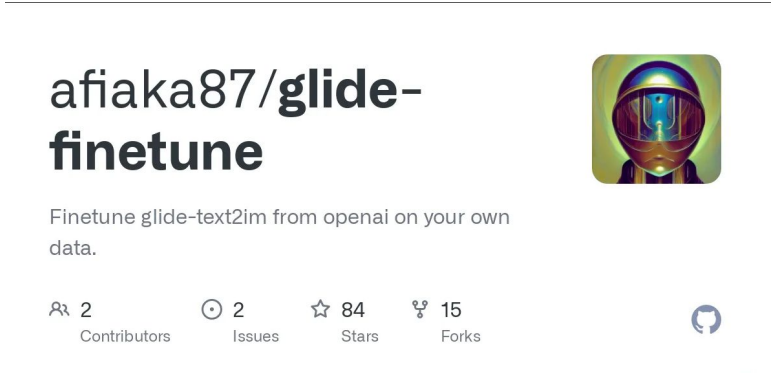
```
# Sample from the base model.
model.del_cache()
samples = diffusion.p_sample_loop(
    model_fn,
    (full_batch_size, 3, options["image_size"], options["image_size"]),
    device=device,
    clip_denoised=True,
    progress=True,
    model_kwargs=model_kwargs,
    cond_fn=None,
)[:batch_size]
model.del_cache()

# Show the output
show_images(samples)
```

```
100%|██████████| 100/100 [09:04<00:00,  5.44s/it]
```



https://github.com/openai/glide-text2im/tree/main/glide_text2im

Milestone 2: Approach



Utilized a GLIDE fine-tuning API

- created dataset that we believed could prove effective in improving GLIDE results
- set up dataloader for images and captions from this dataset
- trained and finetuned GLIDE's pretrained model

Datasets

Used Google Dreambooth's dog images https://github.com/google/dreambooth

Captioned them (each of team member captioned 25% of total Dreambooth's dog images): https://github.com/shairaalam19/cs245-fall2023-gm2-datasets

Consists of varvarying breeds, actions, and backgrounds

Milestone 3: Limitations of GLIDE

Although GLIDE is a great improvement on the original DALL-E model, it still suffers from common generative modeling problems

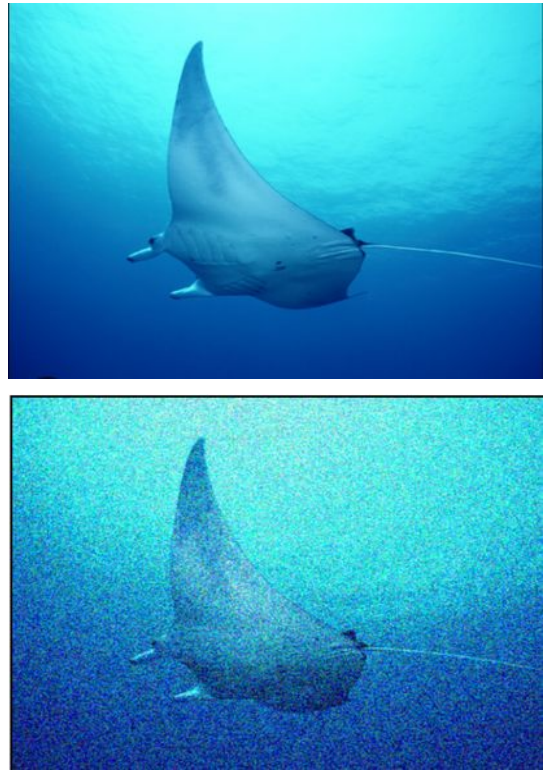
- Quality and Resolution
- Model Bias
- Training Data
- Evaluation Metrics

Quality and Resolution: Denoising

- Upsampler Model attempts to denoise image while also maintaining image features and quality
- Takes output image of base model and feeds into upsampler model → increase pixel amount of existing image
- Diffusion-based approach
- Conditioned on low-resolution image
- Iterative refinement
- Maintaining Semantic Consistency

Training Data & Removing Bias

- Diversifying Training Data
  - Diverse Sources
  - Varied Content Types
  - Data Auditing
- Noise Augmentation
  - Synthetic Noise Addition
  - Data Augmentation
  - Image Pairing



Results & Findings

Qualitative Results

Real photo

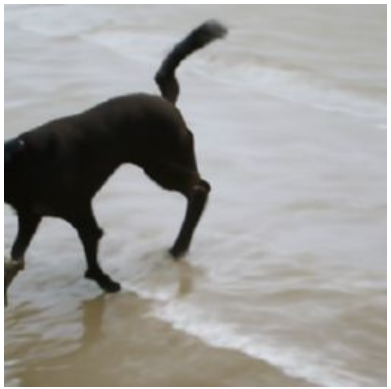
Original GLIDE

Finetuned GLIDE

Caption: medium sized brown fluffy dog posing while sitting on a couch

Qualitative Results (2)

Real photo



Original GLIDE

Finetuned GLIDE

Caption: brown dog with grey dogtags running through waves while smiling at camera

Qualitative Results (3)

Real photo

Original GLIDE

Finetuned GLIDE

Caption: small brown fluffy dog with black tongue standing on grass with autumn leaves on the ground

Qualitative Results (4)

Real photo

Original GLIDE

Finetuned GLIDE

Caption: a happy corgi with a big smile surrounded by cherry blossoms

Quantitative Results

Upsampler epochs	Models	64x64	256x256
20 epochs	Avg FID real images vs original GLIDE	281.75	237.73
	Avg FID real images vs finetuned GLIDE	217.36	191.12
40 epochs	Avg FID real images vs original GLIDE	281.75	237.73
	Avg FID real images vs finetuned GLIDE	217.36	189.55

Lower FID score is better

Improved FID score when upsampled with 40 epochs

Challenges and Solutions

Challenges

- Figuring out how to finetune OpenAI's GLIDE model
  - OpenAI's training scripts were not publicly released
- How to compare images and evaluate success
  - What metrics could be used to evaluate caption similarity?
- How to generalize training beyond specific category of images

Solutions (and new challenges!)

- Figuring out how to finetune OpenAI's GLIDE model
  - Found a repository (https://github.com/afiaka87/glide-finetune) for finetuning
- Getting the finetune model repository to work
  - Documentation is unclear
  - Installation dependencies
  - How to input the new dataset?
- What to train the model with?
  - What parameters should we use? (e.g how many epochs, learning rate, unconditional parameter rate)
- How to compare images and evaluate success
  - Decided to use Frechet Inception Distance (FID)
- How to generalize training beyond specific category of images
  - Currently working on this, will be discussed in the next few slides

Enhancing text-to-image Synthesis

- Linguistic discrepancy between captions of the same image → causes deviation from intended or “ground truth” representation
- Refinement process to provide consistency between captions and images
- Contrastic Learning Approach: 2 stages
  - Pre-training Stage
  - GAN Training Stage
- Proposes approach shows effective improvement
  - Inception Score (IS), Frechet Inception Distance (FID), R-precision

Paper: https://www.bmvc2021-virtualconference.com/assets/papers/0478.pdf

Improving Automatic Evaluation Metrics

- TIGEr: Text-to-image Grounding for Image Caption Evaluation
- Innovative metric that goes beyond traditional text-matching approaches used in existing metrics like BLEU and CIDEr
- Current metrics rely solely on text matching between reference captions and machine-generated captions
- TIGEr evaluates also on how accurately caption represents content of the image

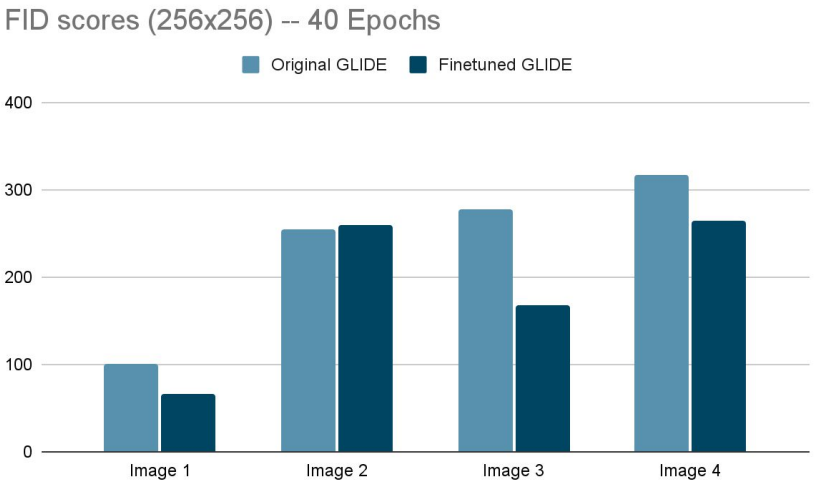
Future Improvements

- Increasing dataset
- Decreasing noise
- Generating more photorealism similar to original GLIDE
- Resources
  - https://arxiv.org/pdf/1802.08216.pdf (Adding dialogue to captions)
  - https://www.sciencedirect.com/science/article/pii/S2468502X21000590 (Image Captioning with multi-level similarity-guided semantic matching)
  - https://www.bmvc2021-virtualconference.com/assets/papers/0478.pdf (text-to-image Synthesis)
  - https://aclanthology.org/D19-1220.pdf (TIGEr)

Conclusions

Impact & Findings

- Training on a specific topic increases caption similarity
- Epochs
- Upsampling
- FID evaluation



Future Implications

- Fine-tuning and specialization
- Training strategies
- More advancements in upsampling techniques
- Increasing textual description training
- Caption development for images

Thank You and Q&A!