Text-guided Image Generation with Diffusion Model and CLIP

Kao Kitichotkul
rkitichotkul@stanford.edu
Department of Electrical Engineering

Patin Inkaew
pinkaew@stanford.edu
Department of Computer Science

Summary

- Diffusion models are latent variable models that map data points onto a latent space via a Markov chain. For some tasks, diffusion models have been shown to outperform GANs in terms of sample quality.
- Song et al. [1] showed that diffusion models are equivalent to score-based models, which can perform conditional generation.
- CLIP [2] is a model that predicts the likelihood of text captions of images. We can leverage CLIP for text-guided image generation from diffusion models without the need to train any additional model.
- **Contribution 1**: We used CLIP to model the conditional probability for conditional generation using diffusion models. Given multiple text captions, one target and the rest distractors, CLIP provides the likelihood of the target caption which can be optimized.
- **Contribution 2**: Following DiffusionCLIP [3], we used a CLIP loss between the sampled image and the target text to guide unconditional generation by optimizing the inputs to the diffusion models.

Background

Denoising Diffusion Probabilistic Model (DDPM)

- DDPMs map a data point $\mathbf{x}_0$ to noise using the forward process:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t, \mathbf{I});\ t = 1, \ldots, T$$

  for some variance schedule $\{\beta_t\}_{t=1}^T$.
- To sample, we initialize $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and follow the reverse process

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z};\ t = T, \ldots, 1$$

  for some $\{\sigma_t\}_{t=1}^T$ where $\epsilon_\theta(\mathbf{x}_t, t)$ is the model, $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$, and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Denoising Diffusion Implicit Model (DDIM)

- DDIM is a method to sample from DDPM *deterministically* by using a different forward process from DDPM that shares the same marginals:

$$\mathbf{x}_{t+1} = \frac{\sqrt{\bar{\alpha}_{t+1}}}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, t)\right) + \sqrt{1-\bar{\alpha}_{t+1}}\epsilon_\theta(\mathbf{x}_t, t);\ t = 0, \ldots, T-1$$

- The reverse process of DDIM is

$$\mathbf{x}_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, t)\right) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(\mathbf{x}_t, t);\ t = T, \ldots, 1$$

References

- [1] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. arXiv preprint, 2021.
- [2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. CoRR, abs/2103.00020, 2021.
- [3] Gwanghyun Kim and Jong Chul Ye. Diffusionclip: Text-guided image manipulation using diffusion models. arXiv preprint, 2021.

Method

Approach 1: Contrastive Conditional Sampling

- Song et al. [1] showed that DDPM is related to the variance-preserving score-based model as follow:

$$s_\theta(\mathbf{x}_t, t) = \nabla_{\mathbf{x}_t} \log p_\theta(\mathbf{x}_t) = -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)$$

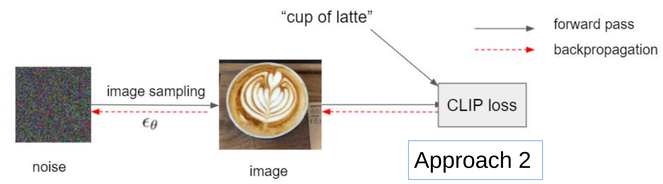
- To sample from $p(\mathbf{x}|\mathbf{y})$, we need to compute $\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \nabla_{\mathbf{x}} \log p(\mathbf{y}|\mathbf{x})$
- For a text caption $\mathbf{y}$, we can calculate $p(\mathbf{y}|\mathbf{x})$ using CLIP. The domain of $\mathbf{y}$ is the set of all possible text captions. To make computation tractable, we choose only a few captions and use CLIP to calculate the normalized probability of the target caption given the image. With a scaling factor γ, the reverse process for DDPM is

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t) + \gamma\beta_t\nabla_{\mathbf{x}_t}\log p_{\text{CLIP}}(\mathbf{y}|\mathbf{x}_t)\right) + \sigma_t\mathbf{z}$$

- The additional texts contrastively assists the guiding.

Approach 2: CLIP-guided Unconditional Sampling

- Inspired by DiffusionCLIP [3], we calculate a CLIP loss between the unconditionally sampled image and the target caption, and then optimize the input $\mathbf{x}_T$ to reduce this loss.
- CLIP maps a text and an image to similar vectors if they share semantic meanings. We choose the cosine distance between the text and image encodings as the CLIP loss.
- Since stochasticity of the DDPM reverse process makes the optimization difficult, we use DDIM instead in this approach.



Experimental Results

Approach 1: Contrastive Conditional Sampling

- We used a DDPM pretrained on the CIFAR10 dataset.
- We compare contrastive conditional sampling with two methods. By starting with the same noisy latent $\mathbf{x}_T$, the results are comparable.
  - Unconditional sampling.
  - Conditional sampling with spherical distance loss using single caption, inspired by *Quick CLIP Guided Diffusion* Colab notebook.
- For all sampling, target prompt is "airplane." Contrastively-assisted prompts are the remaining 9 classes of the CIFAR10 dataset.
- We compute Kernel Inception Distance (KID) between the generated images and the real images in the airplane class from CIFAR10.
- Our experiment shows a potential for better performance of contrastive sampling over baseline method.

← Samples from different methods: unconditional (row 1), spherical distance(row 2), and contrastive (row 3)

↓ Metrics calculated on different sampling methods

Sampling Method	Contrastive	Spherical distance	Unconditional
Time to generate 100 samples (minutes)	90	117	20
KID	0.0618 ± 0.0210	0.0653 ± 0.0236	0.1184 ± 0.0333

Approach 2: CLIP-guided Unconditional Sampling

- DiffusionCLIP solves the image-editing problem. From experiment, we see that starting with a structured data point like an image is essential to generate meaningful images.
- For text-to-image generation, we do not have starting images, so our model struggle to generative meaningful images from completely random noise $\mathbf{x}_T$.
- Since we need to compute gradient with respect to noisy $\mathbf{x}_T$, we need to store gradients in various steps. Because of the memory constraint, we can diffuse only for ~100 steps, while T = 1000. As a result, CLIP might not correctly predict probability with this intermediate noisy image.

↓ Probability for each class computed by CLIP model for contrastive (left) and spherical distance (right)