Text-to-Image Generation

Aleksandr Timashov
June 4, 2021, CS231n Project

Introduction

The **score function** of a distribution **p(x)** is defined as $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ and a model to approximate score function is called **score-based model**.

We should optimize **Fisher divergence**, using **score matching**:

$$\mathbb{E}_{p(\mathbf{x})}[\|\nabla_{\mathbf{x}} \log p(\mathbf{x}) - \mathbf{s}_\theta(\mathbf{x})\|_2^2] = \int p(\mathbf{x}) \|\nabla_{\mathbf{x}} \log p(\mathbf{x}) - \mathbf{s}_\theta(\mathbf{x})\|_2^2 \mathrm{d}\mathbf{x}.$$

Key challenge: Score function is not accurate in a low density regions.
Solution: Add noise to the data points and train score - based model on a noise perturbed data points.

Relation to **diffusion models**: Transition from finite number of noise scales to continuous scale. In this case we have **continuous-time stochastic process**.

Sample using **annealed Langevin dynamics**: $\bar{\mathbf{x}}_{\mathbf{i+1}} \leftarrow \bar{\mathbf{x}}_{\mathbf{i}} + \frac{\alpha_{\mathbf{i}}}{\mathbf{2}} \nabla_{\mathbf{x}} \log \mathbf{p}(\mathbf{x}) + \sqrt{\alpha_{\mathbf{i}}} \mathbf{z}_{\mathbf{i}}$

Problem Statement

The challenge of image generation **usually** accomplished with **adversarial learning**. However adversarial training often suffers from unstable training. I **propose** to use **score-based** generative models to solve this challenge.

Input: Text description.
Output: Image, based on the text description.

Main idea is based on Bayes' rule: $\nabla_{\mathbf{x}} \log p(\mathbf{x} \mid \mathbf{y}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x})$, where **x** is an image, conditioned on text description **y**.

Dataset

I used **COCO Captions Dataset**, containing over 330,000 images and more than 1.5 million captions describing these images.
The split for **train** / **validate** / **test** is default as follows:

- train: 120,000 images;
- validate: 5,000 images;
- test: 205,000 images;

Methods

I will approach this problem in **three steps**:

- Unconditional **score-based model** trained on COCO dataset;
- **Image captioning model**, based on transformer architecture;
- Inverse problem solving for **controllable generation**;

Score-based model details:

- Resizing images to **32 x 32** size and keeping 3 channels;
- Geometric progression of added noises;
- **U-Net architecture** with the same dimension of input and output;
- **Loss** function: $\frac{\mathbf{1}}{\mathbf{L}} \boldsymbol{\Sigma}_{\mathbf{i=1}}^{\mathbf{L}} [\|\sigma_{\mathbf{i}} \mathbf{s}_\theta(\mathbf{x}_{\mathbf{i}} + \sigma_{\mathbf{i}} \mathbf{z}_{\mathbf{i}}, \sigma_{\mathbf{i}}) + \mathbf{z}_{\mathbf{i}}\|_{\mathbf{2}}^{\mathbf{2}}]$

Image captioning model details:

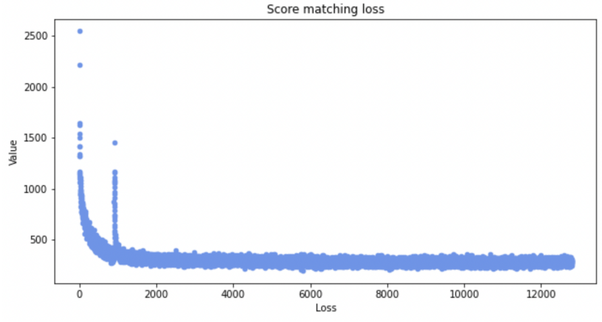
- Resizing images to 32 x 32 size and keep 3 channels;
- **GloVe** pretrained embeddings is used for text embeddings;
- I create a pair of time - dependent training data **(x(t), y)** by adding noise to **(x(0), y)**, sampled from COCO dataset;
- I use **2 headed self-attention** in a transformer-based encoder;
- To find score function, I use an idea from Bayes' rule:

$$\nabla_{\mathbf{x}} \log \mathbf{p}(\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_k} | \mathbf{x}) = \nabla_{\mathbf{x}} \log \mathbf{p}(\mathbf{y_1} | \mathbf{x}) + \boldsymbol{\Sigma}_{\mathbf{i=2}}^{\mathbf{k}} \nabla_{\mathbf{x}} \log \mathbf{p}(\mathbf{y_i} | \mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_{i-1}}, \mathbf{x})$$

Sampling:

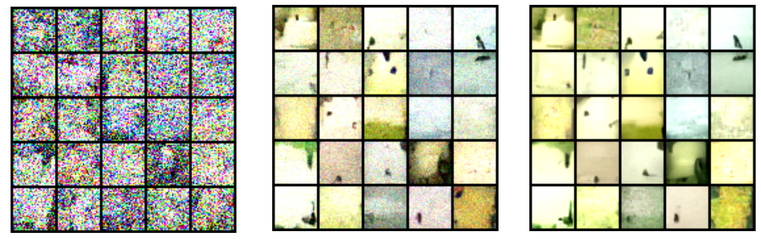
$$\bar{\mathbf{x}}_{\mathbf{i+1}} \leftarrow \bar{\mathbf{x}}_{\mathbf{i}} + \frac{\alpha_{\mathbf{i}}}{\mathbf{2}} [\nabla_{\mathbf{x}} \log \mathbf{p}(\mathbf{x}) + \nabla_{\mathbf{x}} \log \mathbf{p}(\mathbf{y} \mid \mathbf{x})] + \sqrt{\alpha_{\mathbf{i}}} \mathbf{z}_{\mathbf{i}}$$

Experiments & Analysis



- As we can see, **loss** is converging. Around step number 1,000 Adam optimizer lost local minimum, but after that started converging again.

Below are **some results** starting from noise and converging to clear pictures.



Conclusions & Future Work

My work shows the great opportunity to experiment **with image generation** using **score-based models**.
I experimented using different model architectures, including **DDPM** and **NCSN**.
Sampling was done using **Annealed Langevin dynamics**.

Future Work:

- Experiments with Conceptual Captions dataset;
- Experiments with different models architectures and different sampling techniques;
- Experiments with increased input size;

References

Yang Song et al. Score-based generative modeling through stochastic differential equations, 2021.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In Advances in Neural Information Processing Systems, pages 11895– 11907, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffu- sion probabilistic models, 2020.