



Continuous control of latent diffusion model through token embedding interpolation

Henry Lin, Pracha Promthaw, Nikhil Gautam Mentor: Dmitry Petrov

UMassAmherst

Manning College of Information & Computer Sciences

Motivation

In this work, we are trying to investigate the encoded embedding space of certain attributes. To the best of our knowledge, no one has investigated the continuous changes of certain attributes before. Our goal is to change certain attributes continuously.

How does the realistic concept transform into Picasso?

Can we gradually change the level of realistic and Picasso in the following figures?

Realistic



Picasso



Background

The previous work by (Brack et. al, 2022) introduced Stable Artist which can steer semantics in diffusion latent space. The idea is that by finding the direction of certain attributes, they can steer the semantics of the image into the attributes that they picked. However, the limitation of their approach is that it **doesn't have continuous control over the attributes**.

- Start with an unconditioned image
- find the latent vectors that correspond to different concepts
- manipulate the unconditioned latent space
- steering the unconditioned to conditioned latent space based on the concepts



Brack, Manuel & Schramowski, Patrick & Friedrich, Felix & Hintersdorf, Dominik & Kersting, Kristian. (2022). The Stable Artist: Steering Semantics in Diffusion Latent Space. 10.48550/arXiv.2212.06013.

Methology

- Input: two texts that describes two different concepts.

We utilized the Stable Diffusion algorithm (Rombach et.al, 2021) and used the Clip model to encode the text, denoising loop, using a decoder from an auto-encoder to decode an embedding vector to an image. We pick two attributes that we want to interpolate and use the interpolation methods to shift between two concepts. The results show that the encoded embedding has the continuous attribute, so it is able to understand the continuous change in the vector space. This could be further adapted into continuous editing in latent vector space for more controllable image generation.

E: Clip encoder

D: Autoencoder's decoder

Embedding: Convert the input sequence of tokens into a continuous representation.

text1, text2 \rightarrow $\vec{t}_1 = \text{embedding}(\text{tokenizer}(\text{text1}))$
 $\vec{t}_2 = \text{embedding}(\text{tokenizer}(\text{text2}))$

Approach 1:

$\vec{t} = \text{Linear Interpolation}(\vec{t}_1) \rightarrow \vec{E} = E(\vec{t}) \rightarrow \vec{v} = \text{Denoising Loop}(\vec{E}) \rightarrow \text{image} = D(\vec{v})$

Approach 2:

$\vec{t}_1, \vec{t}_2 \rightarrow E_1 = E(\vec{t}_1) E_2 = E(\vec{t}_2) \rightarrow T = \text{Gradient Descent}(\vec{t}_1) \rightarrow \text{Non-Linear Interpolation } \vec{t} \subset T \rightarrow \vec{E} = E(\vec{t})$

Linear Interpolation:

$$\vec{t} = \alpha \vec{t}_1 + (1 - \alpha) \vec{t}_2$$

$$\alpha \in [0, 1]$$

Gradient Descent:

$$\vec{E}_{\text{diff}} = \vec{E}_2 - \vec{E}_1$$

$$L = \sqrt{\vec{E}_{\text{diff}}(\vec{E}_{\text{diff}})^T}$$

Algorithm 1 Gradient Descent

```

 $\vec{v}_1 = \text{Embedding}(\text{Tokenizer}(\text{text1}))$ 
 $\vec{\theta}_0 = \text{Encoder}(\vec{v}_1)$ 
 $\vec{v}_2 = \text{Embedding}(\text{Tokenizer}(\text{text2}))$ 
 $\vec{\gamma}_2 = \text{Encoder}(\vec{v}_2)$ 
 $\text{curves} = [\vec{\theta}_0]$ 
 $\text{while not converged do}$ 
 $\quad j = 1$ 
 $\quad \vec{\gamma}_1 = \text{Encoder}(\vec{v}_{j-1})$ 
 $\quad L = \sqrt{(\vec{\gamma}_1 - \vec{\gamma}_2)(\vec{\gamma}_1 - \vec{\gamma}_2)^T}$ 
 $\quad \vec{\theta}_j = \vec{\theta}_{j-1} - \alpha \frac{\partial L}{\partial \vec{\theta}_{j-1}}$ 
 $\quad \text{curves.append}(\vec{\theta}_j)$ 
 $\quad j = j + 1$ 
 $\text{end while}$ 
 $\text{return curves}$ 

```

Rombach, Robin and Blattmann, Andreas and Lorenz, Dominik and Esser, Patrick and Ommer, Björn. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. <https://arxiv.org/abs/2112.10752>

Conclusion/Discussion

From the results, we can continuously control attributes through two approaches: linear interpolation and non-linear interpolation. However, some change occurs suddenly and didn't depict a sign of continuous editing. For example, in figure 3, the middle figure didn't depict a continuous change. Also, we tried to form a Voronoi diagram on the text embedding space, but the dimension of it is too large. The future work of this project could be continuing to investigate the encoded latent space. For example, we can try to interpolate in the space that is formed by Voronoi Diagram or a space formed by the convex hull.

Results

