HIERARCHICAL TEXT-CONDITIONAL IMAGE GENERATION WITH CLIP LATENTS

Chiara Piccolo (619224)
2023/2024
Intelligent Systems for Pattern Recognition

Introduction

The field of computer vision has **advanced significantly** due to the development of large-scale models trained with extensive collections of captioned images. Among these innovations, CLIP stands out as a **robust learner** capable of understanding complex image and text correlations with impressive zero-shot performance.

Simultaneously, diffusion models have shown **outstanding performance** in the generation of images and videos, utilizing guidance techniques to increase photorealism

The proposed model, **unCLIP**, combines innovatively the robust learning framework of CLIP with advanced diffusion techniques to enable text-driven image generation.

Model Description – Architecture

There are 3 main components of the unCLIP model:

1. **CLIP**: it establishes a joint representation space for text and images.
2. **Prior**: it generates CLIP image embeddings conditioned on a given caption.
3. **Diffusion decoder**: it outputs the final image, conditioned on both the predicted image embedding and the caption.

Mathematical formulation: let us consider a training dataset consisting of pairs (x, y), where x is an image and y is its caption. Additionally, let z_i and z_t be respectively the CLIP image and text embedding of x:

$$P(x|y) = P(x, z_i|y) = P(x|z_i, y)P(z_i|y)$$

The first holds because z_i is a deterministic function of x (CLIP associates a unique embedding z_i to every image x).
The second equation follows from the chain rule.

This framework establishes a generative model $P(x|y)$, that computes the likelihood of producing an image x given the caption y.

Model Description - Prior

- $\boldsymbol{P(z_i|y)}$: it produces a CLIP image embedding z_i conditioned on caption y
- For the experiments, the prior is conditioned on both the caption y and the CLIP text embedding z_t. To improve sample quality, classifier-free guidance is used, randomly dropping the text conditioning 10% of the time during training.
- Two different classes were explored for the prior model:

Autoregressive Prior

This model component transforms the CLIP image embeddings into a sequence of discrete codes. It starts by reducing the dimensionality of the image embeddings from 1,024 to 319 principal components using PCA. The principal components are then sorted based on their eigenvalues and quantized into 1,024 discrete buckets.

The AR prior is conditioned on both the text caption and the CLIP text embedding, which are encoded together as a sequence prefix. Additionally, a token representing the quantized dot product between the text and image embeddings ($z_i * z_t$) is prepended to the sequence.

Diffusion prior

This model is trained on a decoder-only transformer with causal attention mask, on an input sequence that consists of: the encoded text, the CLIP text embedding, the diffusion timestep embedding, the noised CLIP image embedding, and a final embedding for predicting the unnoised CLIP image embedding.

During the sample phase, the quality is enhanced by generating two samples of z_i and selecting the one with a higher dot product with z_t. This ensures that the generated image is more aligned with the textual description.

The model is then trained to predict directly the unnoised z_i using a mean-squared error loss.

Model Description - Decoder

- $P(x \mid z_i, y)$: it produces an image x conditioned on CLIP image embedding z_i (and optionally also on the text caption y)

- The decoder is based on a modified version of the GLIDE architecture, incorporating additional information in 2 ways:

    1) Adding CLIP embeddings to the existing timestep embedding

    2) Creating 4 tokens of context, concatenated to the output sequence of the GLIDE text encoder

- To promote robustness and diversity, the model employs classifier-free guidance during training by randomly setting the CLIP embeddings to zero and dropping the text captions.

- Images are initially generated at 64x64 resolution and then upsampled twice to produce a final image with a resolution of 1024x1024. During training, images are slightly degraded (e.g., using Gaussian blur) to teach the model to handle and correct imperfections, thus enhancing the photorealism of generated images.

Experiments and Results

Image Manipulation: The unCLIP model allows comprehensive image manipulation by encoding images into a **bipartite latent** representation (z_i, x_T), where z_i encodes **aspects of the image** recognized by CLIP and x_T encodes **residual information** necessary for the decoder to reconstruct the image x accurately.

z_i is obtained by **simple encoding** with the CLIP image encoder, while x_T is obtained by applying **denoising diffusion implicit model** (DDIM) to x using the decoder, conditioned on z_i .

1. Variations: by combining z_i and x_T, the input image can be manipulated to generate different variations that maintain essential content but vary in aspects like shape and orientation. This is achieved by applying the decoder to the bipartite representation (z_i, x_T) using DDIM.

2. Interpolations: blending two different images is possible by using spherical interpolation between their respective CLIP embeddings z_{i1}, and z_{i2} and between x_{T1} and x_{T2}.

Text Diffs: CLIP embeds both images and text in the same space, allowing for **text-guided** image manipulations.
To modify an image based on a new text description y, the first thing to do is to obtain both the CLIP text embedding z_t for the **new** description and z_{t0} for the **current** description of the image. Then **a text difference vector** $z_d = norm(z_t - z_{t0})$ is computed.
After that, the **spherical interpolation** is applied between the image's CLIP embedding z_i and the text diff vector z_d: this generates intermediate embeddings that represent **different stages** of transformation from the original image's concept to the new description.

Experiments and Results

Probing the CLIP latent space: The decoder model allows **direct visualization** of what the CLIP image encoder perceives. It helps explore cases where CLIP makes **incorrect predictions**, such as typographic attacks, where text on an image misleads the model. For example, an apple with "iPod" text may be misclassified by CLIP, but the decoder still generates apple images. PCA reconstructions show that early PCA dimensions encode **broad semantic information**, like object types, while later dimensions capture **detailed features**, such as specific shapes.

Importance of the prior: Training a prior to generate CLIP image embeddings from captions is **useful** but **not essential** for caption-to-image generation. The decoder can be conditioned on **captions alone** or on **CLIP text embeddings**, with reasonable results. However, comparisons show the unCLIP approach performs **best**.

Human Evaluation: unCLIP excels in generating **diverse** and **realistic** images, often outperforming other models like GLIDE, especially in **diversity**. Human evaluations show that unCLIP is slightly less preferred for photorealism but strongly favored for diversity. Automated evaluations reveal that unCLIP produces **superior** artistic illustrations and photographs. Key findings highlight unCLIP's ability to **maintain scene semantics** during guidance, unlike GLIDE. Overall, unCLIP demonstrates **superior capabilities** in generating diverse, high-quality images and excels in zero-shot generalization.

Conclusions

While conditioning the image generation on CLIP embeddings seems to improve diversity, this choice still has certain **limitations**:

- unCLIP is **less effective** than other models (like GLIDE) at accurately **binding** attributes to specific objects, resulting in attribute and object mix-ups in generated images
- unCLIP **struggles** with producing **coherent text** due to the non-explicit encoding of spelling information in CLIP embeddings
- the initial **low resolution** of images generated by unCLIP's decoder (64x64) limits the **detail** achievable in complex scenes

Moreover, improvements in unCLIP's image generation make it harder to distinguish AI-generated images from real ones. This raises concerns related to deception and bias, highlighting the need for careful risk evaluation and proper safety measures when using this technology.

My opinion: the unCLIP model represents a **significant advancem**ent in text-to-image generation. The evolution of technologies like this one opens **new frontiers** for collaboration between humans and AI. However, it is important to consider the **ethical implications** of such powerful technology. Responsible use is essential to prevent the generation of misleading or harmful content and to address any inherent biases in the training data. Overall, unCLIP stands out as a **powerful tool** for creative and practical applications, pushing the boundaries of what is possible in AI-driven image generation.