Visual Chain-of-Thought Diffusion Models

Overview:

The paper addresses the underdevelopment of unconditional image diffusion models in comparison to conditional counterparts. The research is focused on high-quality image generation, particularly in the context of unconditional settings and "lightly-conditional" settings where the input is low-dimensional, such as a class label. A novel two-stage sampling procedure is proposed to enhance the quality of unconditional image generation. This method samples an initial embedding which describes the image's semantic content, followed by a secondary sampling of the image conditioning on this embedding, showing a 25-50% improvement on the Frechet Inception Distance (FID) compared to existing models. The paper further introduces the Visual Chain-of-Thought Diffusion Models (VCDM), adding to the model's versatility in various approximate inference settings. Experimentation underscores the validity of the proposed model with the conclusion that the more information the Diffusion Generative Model (DGM) is conditioned on, the greater the realism in the generated image. The two-stage approach was demonstrated to effectively bridge the gap between conditional and unconditional image diffusion models.

Detailed Method:

The authors of this document propose a model called VCDM. It's created to improve Deep Generative Models (DGMs) using the CLIP technology. VCDM is designed to work in two scenarios - when the user doesn't wish to specify any input (unconditional setting), or when the input is small-scale, such as a class-label (lightly-conditional setting).

$$\mathbb{E}_{u(\sigma)p_\sigma(\mathbf{x}_\sigma|\mathbf{x},\sigma)p_{\mathrm{data}}(\mathbf{x},\mathbf{y})}\left[\lambda(\sigma)||\mathbf{x}-\hat{\mathbf{x}}_\theta(\mathbf{x}_\sigma,\mathbf{y},\sigma)||^2\right] \quad (1$$

They present a mathematical approach to illustrate how VCDM approximates the target distribution (which is pdata). It approximates the target data distribution using two DGM models - one for the CLIP embeddings and another for the image. To generate samples, they sample from the CLIP embeddings DGM (which they refer to as the auxiliary model) first, then use this to sample from the Image DGM (known as the conditional image model), while disregarding the initial CLIP sampling.

$$\begin{aligned} p_{\mathrm{data}}(\mathbf{x}|\mathbf{a}) &= \mathbb{E}_{p_{\mathrm{data}}(\mathbf{y}|\mathbf{a})}\left[p_{\mathrm{data}}(\mathbf{x}|\mathbf{y},\mathbf{a})\right] \\ &\approx \mathbb{E}_{p_\phi(\mathbf{y}|\mathbf{a})}\left[p_\theta(\mathbf{x}|\mathbf{y},\mathbf{a})\right] \end{aligned}$$

The auxiliary model operates on CLIP embeddings, which are features of an image in a 512-dimensional space. It uses a transformer architecture that takes in sequences of these 512-dimensional features. For better generalization and to prevent overfitting, the authors recommend image augmentation techniques particularly for AFHQ and FFHQ datasets, including rotations, color flips, and jittering.

The conditional image model, on the other hand, uses a diffusion process and builds on existing work by other researchers. The authors used different architectures depending on the dataset - a U-Net architecture for AFHQ and FFHQ datasets, and a slightly different U-Net architecture for the ImageNet dataset. This model combines the CLIP embeddings and extra input (if any), obtained from the auxiliary model into a single vector and then applies a learned projection.

So, in simpler terms, the authors propose a method of generating images from data where they first use an "auxiliary" model to generate a kind of blueprint, and then an image model to use this blueprint to create the final image.