Week 3: Thursday Paper 1: Stable Diffusion

explained by The Illustrated Stable Diffusion (overview) and this blog with more math

Supplemental: High-Resolution Image Synthesis with Latent Diffusion Models

1. Explain the stable diffusion architecture from start to end in your own words.

Ans: The Stable Diffusion architecture is designed for generating high-quality images from textual descriptions. It operates in a latent space, where images are represented in a compressed form. The process begins with a random noise image in the latent space. Then, through a series of steps, it gradually applies changes to this image, guided by a text description, until it matches the desired output. The model consists of two main components: the diffusion model, which works in the latent space to add and then reverse noise from images, and a conditional model that ensures the generated images match the text description. The architecture progressively refines the image, moving from a noisy state to a clear, detailed image that aligns with the provided text.

2. What is a Latent Diffusion Model? Explain the process of diffusion in your answer. How are diffusion models trained?

Ans: A Latent Diffusion Model operates in a compressed, or latent, space rather than directly on the pixel space. This approach significantly reduces computational costs and memory requirements. The diffusion process involves two phases: the forward process gradually adds noise to the original data until it is completely random, and the reverse process aims to reconstruct the original data from the noisy data, guided by the learned data distribution. Diffusion models are trained by teaching the model to predict the original image from its noisy version, improving through iterations where the model learns to denoise images effectively.

3. What role does the image decoder play in the architecture? What will happen if you remove the decoder?

Ans: In the Stable Diffusion architecture, the image decoder plays a crucial role in translating the manipulated latent representations back into images. It takes the denoised latent representations and reconstructs them into visible images. If you remove the decoder, the model would not be able to convert the latent representations into images, leaving the process incomplete. Essentially, the decoder is what enables the abstract, compressed modifications in the latent space to be realized as tangible images.

4. How does stable diffusion use text embeddings for image generation? What model is used for text embedding?

Ans: Stable Diffusion uses text embeddings to guide the image generation process, ensuring that the generated images align with textual descriptions. These embeddings are generated using a separate model that is trained to understand and encode textual information into a format that the image generation model can utilize.

From the article, we got to know that the early Stable Diffusion models just plugged in the pre-trained ClipText model released by OpenAI. It's possible that future models may switch to the newly released and much larger OpenCLIP variants of CLIP (Nov2022 update: True enough, Stable Diffusion V2 uses OpenClip). This new batch includes text models of sizes up to 354M parameters, as opposed to the 63M parameters in ClipText.

5. What are some ethical implications of stable diffusion? How might people address them?

Ans: The most recent ethical concern that was discussed all over the world is the issue of AI generated images of Taylor Swift in obscene positions and the need for regulation for generating and publicizing images without consent. Other ethical concerns include copyright infringement and privacy concerns.

6. What questions do you have about the paper?

We'll cover in class a bit more of the math that you can see here, but no HW on it

Week 3: Thursday Paper 2: DALL·E 2

explained by DALL·E 2 Explained - model architecture, results and comparison

Supplemental: OpenAI's DALL-E 2 and DALL-E 1 Explained

1. How is DALL.E 2 different from Stable Diffusion?

Ans: DALL.E 2 and Stable Diffusion are both AI-driven image generation models, but they have distinct approaches and architectures. DALL.E 2, developed by OpenAI, focuses on generating images from textual descriptions using a two-part model involving a prior for generating CLIP image embeddings from text and a decoder for creating images from these embeddings. Stable Diffusion, on the other hand, uses a latent diffusion model conditioned on text embeddings for direct text-to-image generation, emphasizing efficiency and broad accessibility.

2. What is a prior? What are the different priors used in the paper? Can you think of a different prior?

Ans: A "prior" is a component that generates CLIP image embeddings based on text captions. The paper describes two types of priors: autoregressive (AR) and diffusion priors. The AR prior predicts CLIP image embeddings autoregressively from text captions, while the diffusion prior models the continuous CLIP image embedding vector directly, aiming for computational efficiency and quality. I am not able to think of a different prior at the moment.

3. How can the architecture be used for image manipulation? Explain in your own words.

Ans: The architecture allows for sophisticated image manipulations by encoding an image into a CLIP embedding and then applying transformations within this latent space. Manipulations such as variations, interpolations, and text-driven modifications are possible, enabling changes in style, content, or both, guided by textual descriptions.

4. The paper uses the CLIP latent space extensively. Is it possible to make the latent space explainable i.e. to map the features in the generated image to variables in the latent space? If yes, can you think about a scheme to make this latent space more explainable? If not, explain your reasoning.

Ans: Making the latent space explainable involves linking specific features in generated images to variables in the latent space. One approach could involve dissecting the latent space to identify directions corresponding to interpretable attributes (e.g., object types, colors) and then manipulating these directions to control specific aspects of the generated images.

5. What are the advantages/disadvantages of DALL.E 2 over Stable Diffusion? Can you think about changes in the respective models to address the disadvantages?

Ans: DALL.E 2 offers high-quality, diverse image generation with the ability for precise text-to-image alignment and manipulation. However, it may require more computational resources than Stable Diffusion, which is designed for efficiency and scalability. Improving the computational efficiency of DALL.E 2 and enhancing the diversity and flexibility of Stable Diffusion could address these disadvantages.

6. What questions do you have about the paper?