Project idea for the lecture "Learning from Images": Applications of advanced techniques like DreamBooth, Prompt-to-Prompt and Conceptualizer to enhance the performance of Stable Diffusion model on generating images from text.

Students: Dennis Fast, Amin Suaad, Manuel Freistein

1. About Stable Diffusion model in general

Stable Diffusion is a deep learning, text-to-image model released in 2022. It is primarily used to generate detailed images conditioned on text descriptions, though it can also be applied to other tasks such as inpainting, outpainting, and generating image-to-image translations guided by a text prompt. Stable Diffusion uses a variant of diffusion model (DM), called latent diffusion model (LDM). Stable Diffusion was trained on pairs of images and captions taken from LAION-5B, a publicly available dataset derived from Common Crawl data scraped from the web, where 5 billion image-text pairs were classified based on language, filtered into separate datasets by resolution, a predicted likelihood of containing a watermark, and predicted "aesthetic" score (e.g. subjective visual quality).

2. Application of Stable Diffusion

Evaluation metrics for all applications: In order to evaluate the results of the generative process, we are going to use some metrics to assess the image quality (e.g. inception score), the text relevance (e.g. R-precision) and the object accuracy (e.g. Semantic Object Accuracy).

DreamBooth (Dennis Fast)

Short description: Fine-tuning a model using DreamBooth technique enables generation of different images of the a subject instance in different environments, with high preservation of subject details and realistic interaction between the scene and the subject. For my part of the project, I want to use the DreamBooth technique to fine-tune Stable Diffusion model in order to create a digital avatar of myself and put it to different scenes using text prompts.
Used data: The input data are the images of myself in diverse clothes, in various poses, at different ages and in wide-ranging enviroments. In order to improve the quality of the synthesized images, the input images should have the greatest possible variation so that the model learns only the features of the person and not particular clothes or environment.
Sources: https://arxiv.org/pdf/2208.12242.pdf, https://dreambooth.github.io/

Prompt-to-Prompt (Amin Suaad)

Short description: Editing a generated image is challenging. Often, small changes in prompt makes a huge change in the image and this localized editing or controlled editing is a problem in situations. Prompt-to-Prompt technique is a solution is such cases. Here, Cross attention layers are key to establish the relation between the image and each word of the

prompt. Prompt-to-prompt allows text level control. Some examples of Prompt-to-prompt technique: Localized editing by replacing a word, global editing by adding a specification, and even controlling the extent to which a word is reflected in the image. I will be trying to use prompt-to-prompt technique in specific situations where it makes more sense to have a text level control in editing.
Used data: Set of prompts
Sources: https://github.com/google/prompt-to-prompt, https://prompt-to-prompt.github.io/ptp_files/Prompt-to-Prompt_preprint.pdf

Conceptualizer (Manuel Freistein)

Short description:
I will try to "teach" Stable Diffusion the concept of a few particular fine art styles via textual-inversion. Textual Inversion is a technique for capturing novel concepts from a small number of example images in a way that can later be used to control text-to-image pipelines. It does so by learning new ‘words’ in the embedding space of the pipeline’s text encoder. These special words can then be used within text prompts to achieve very fine-grained control of the resulting images. Using only 3-5 images of a user-provided concept, like an object or a style, Stable Diffusion will learn to represent it through new "words" in the embedding space.
Used data:
From the portfolios of a few famous artists I will select public domain images that best represent a distinctive style. I will choose the images myself and have the evaluation metrics help me determine the “best” choice.
Evaluation metric:
Fréchet Inception Distance (FID) is a performance metric that calculates the distance between the feature vectors of real images and the feature vectors of fake images. It measures if the similarity between “real” (in my case a number of authentic art pieces by the particular artist) and generated images (Stable Diffusion images using my concept prompt) is close. My idea is that feature vectors will pick up on a style's particular attributes (some of which even the best art critics wouldn't be able to detect). I can then compare these results with human (my own or survey data) judgements. As textual inversion does not need a lot of data and artists often specialize on a certain subject matter (so that the metric can really concentrate on style rather than subject), I will have enough testing data available from the selected artist's portfolios. This is just an idea. It might not work in practice and I might have to try something else, but I can determine that best while I am working on the project.
Sources:
- https://huggingface.co/spaces/sd-concepts-library/stable-diffusion-conceptualizer
- Rinon Gal, Yuval Alaluf, Yuval Atzmon et al.: An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion, 2022 online: https://arxiv.org/abs/2208.01618 (retrieved: 27 November).
- Ali Borji: Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney and DALL-E 2, 2022, online: https://arxiv.org/abs/2210.00586 (retrieved: 27 November).
- Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh et al.: Reliable Fidelity and Diversity Metrics for Generative Models, 2020, online: https://arxiv.org/abs/2002.09797 (retrieved: 27 November).