UAM-VPULab - Internship Evaluation Task

Arindam Ghosh
Barcelona Institute of Science and Technology (BIST)
Email: aghosh@bistgraduatecentre.com

I. Stage 1: Text-to-Image Generation

Approach used for this Task:

- **Choosing the Category - Beaches:** For this task, I chose the category of images to be beaches. I selected this category because I believed that beach images would have minimal clutter and provide a clean canvas for studying the divergence in generated images effectively.
- **Selecting Latent Diffusion Models:** To generate realistic images from text descriptions, I began by searching for latent diffusion models. I was already familiar with stable diffusion models, so I explored the Hugging Face model hub for "Stable Diffusion v2," [1] which seemed suitable for my task.
- **Setting Up the Environment:** I set up my notebook and used the stable diffusion model from the Diffuser library [2] by Hugging Face.
- **Initial Prompt and Iterations:** Initially, I used the prompt "Realistic beach" to generate images. However, the results were cartoonish and did not meet the desired level of realism as shown in figure 1.

Figure 1. Initial image generated with the caption "Realistic Beach" for the beach category.

- **Experimenting with Prompts:** To improve image quality, I experimented with various prompts and visually verified the results. After several iterations, I settled on the prompt: "A photo of a beach, ultraHD, high quality, photo-realistic, outdoors." This prompt consistently generated high-quality, realistic beach images.
- **Generating Images:** Using the finalized prompt, I generated a total of 100 beach images. These images were of ultra-high resolution and met the criteria of being photo-realistic and outdoors.

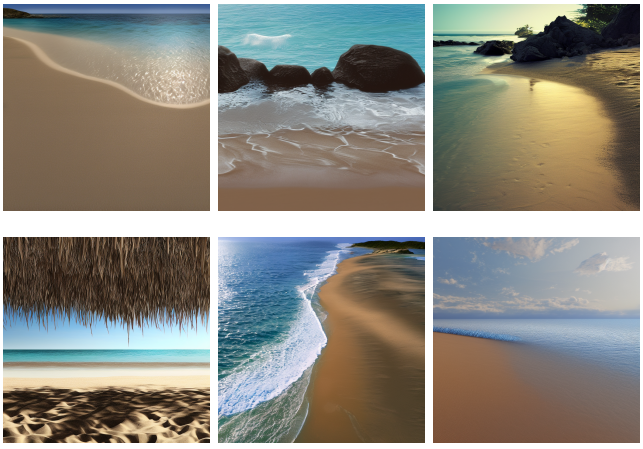

Figure 2. Generated images of the category "Beach". The final prompt that was used to generate these images was - "A photo of a beach, ultraHD, high quality, photo-realistic, outdoors".

II. Stage 2: Quantify Semantic and Style Divergence

Approach used for this Task:

In assessing the semantic and style divergence of the generated beach images, it's important to understand the concepts of semantic and style divergence from a practical standpoint. Semantic divergence refers to how different the content or objects in the images are, while style divergence pertains to artistic characteristics, such as colors and textures.

To evaluate these divergences, I researched more on the attribute disentanglement for generated images, this indicates the model's ability to change specific attributes while keeping others constant. This is crucial because it allows us to assess how well the model can generate diverse images while maintaining certain aspects unchanged.

During my research, I came across a paper that discussed attribute disentanglement for stable diffusion models [3]. This paper provided valuable insights into understanding the capabilities of a generative model.

For quantifying semantic divergence among the 100 generated beach images, I tried two metrics:

- **Pixel-wise Difference:** One straightforward approach to measure semantic similarity is to calculate the pixel-wise difference between grayscale versions of the images. This method indicates whether the images have objects in similar positions.
  To do this, I treated each image as a reference and

computed a dissimilarity value for all other images. This process was repeated for each image as a reference.

- **Edge Detection Difference:** I also assessed semantic divergence by comparing edge-detected versions of the generated images. This method focused on highlighting prominent semantic objects within the images. Using the Sobel edge detection technique to get edge-detected images, I calculated divergence values between these edge-detected images in a similar fashion as the previous metric and constructed a histogram for reference, providing insights into the variations in semantic content among the generated images.

Style divergence metric

- Histogram-Based Metrics (e.g., Histogram Intersection and Bhattacharyya Distance [4]): These metrics analyze the color histograms of the images. Larger values indicate greater style divergence in terms of color distribution.
- Another metric that would take into consideration is the Gabor filter which would help us find the difference in texture which could be a good metric for style divergence. It's worth noting that the metrics mentioned above do not take into account object localization. A more advanced approach could involve object detection for each image and then assessing the similarity of individual objects. This would provide a more nuanced understanding of semantic divergence, considering object-level variations.

III. Stage 3: Enlarging Image Diversity

Approach used for this Task:

The initial idea for this was to use different prompts which would keep the images in the same category but also significantly make the images more diverse in the sense of style and semantics.

Method - 1

During my research, I came across a paper authored by M. Brack et al. [5], which highlighted a common challenge in one-shot image generation—achieving results that precisely align with the user's intent can be exceedingly difficult. Even small alterations to the input prompt often yield significantly different images, leaving users with limited semantic control.

To address this issue and empower users to have more control over the image generation process, the authors introduced a method detailed in their notebook [6]. This approach allows for interactive steering of the diffusion process along semantic directions. The outcomes of this method, which offer enhanced user control, are illustrated in Figure 3.

Method - 2

In this method, I utilized the hyperparameter known as 'guidance_scale' within the stable diffusion model. This parameter plays a crucial role in adhering the images to the text prompt. Essentially, it allows for fine-tuning the balance between adhering closely to the given text prompt and maintaining the overall quality of the generated samples.

Figure 3. Semantic Guided image editing to create divergence [6]. The image on the right and left are similar except for the crowd is present in one.

Figure 4. Images generated using Method 3 for stage 3 where different prompts were used to have a significant semantic and style divergence.

To put it simply, a lower 'guidance_scale,' such as 3 in this instance, leans towards classifier-guided generation, which tends to result in more diverse outputs. On the other hand, using larger values, like 7 or 8.5 as recommended by the authors in the notebook [1] [7], can lead to images that closely match the prompt but may exhibit less diversity.

This adjustment of the 'guidance_scale' hyperparameter allowed for a controlled exploration of the trade-off between fidelity to the text prompt and the diversity of generated images.

Method - 3

In this method, I employed a strategy to diversify the generated images both semantically and stylistically by using a variety of prompts. I initiated the process by providing an initial prompt to the OpenAI GPT-3.5 API and then requested different variations of it. These variations were intended to elicit diverse image outputs while staying within the chosen category. To ensure that the generated prompts focused on the intended category and style, I instructed the model not to introduce objects or elements from unrelated categories. This approach aimed to produce a range of stylistic variations while keeping the core content of the initial prompt intact.

IV. Code Availability

The code used for this Technical report is available for reference. You can access the code on the following GitHub repository:

https://github.com/aryndam9/Diffusion-Models-Exp

All the images displayed in this report are included in the repository, and a readme file is provided to assist in reproducing the work.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 684–10 695.

[2] P. von Platen, S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, and T. Wolf, "Diffusers: State-of-the-art diffusion models," https://github.com/huggingface/diffusers, 2022.

[3] Q. Wu, Y. Liu, H. Zhao, A. Kale, T. Bui, T. Yu, Z. Lin, Y. Zhang, and S. Chang, "Uncovering the disentanglement capability in text-to-image diffusion models," 2022.

[4] Bhattacharyya distance for histograms. Accessed on 2023-09-18. [Online]. Available: https://stats.stackexchange.com/questions/51848/bhattacharyya-distance-for-histograms

[5] M. Brack, F. Friedrich, D. Hintersdorf, L. Struppek, P. Schramowski, and K. Kersting, "Sega: Instructing diffusion using semantic dimensions," *arXiv preprint arXiv:2301.12247*, 2023.

[6] Semantic image editing with semanticguidance example notebook. Accessed on 2023-09-18. [Online]. Available: https://colab.research.google.com/github/ml-research/semantic-image-editing/blob/main/examples/SemanticGuidance.ipynb

[7] H. Face. (2023) Stable diffusion: Training diffusion models with the diffusion probabilistic models library. Accessed on 2023-09-18. [Online]. Available: https://colab.research.google.com/github/huggingface/notebooks/blob/main/diffusers/stable_diffusion.ipynb