Comparative Analysis of Text-to-Image Models

COGS 185 Final Spring ‘24
Valeria Gonzalez Perez

Generative models have marked a distinctive transition in the field of AI where one is now able to automatize ideas, personalize artwork, and generate high quality images and content. In this article we will conduct a literature review on *Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney and DALL-E 2*[1] by Ali Borji in order to summarize and synthesize the most important findings on three state-of-the-art generative models more generally known as text-to-image models. A brief description of each model will be provided followed by details of the data used to evaluate such models, evaluation criteria, and results. With this article we hope to provide a concise understanding of some of the major capabilities and limitations of Stable Diffusion, Midjourney, and DALL-E 2 - and also offer a well rounded comparison of these three text-to-image models.

Generative text-to-image models work through diffusion or by iteratively adding Gaussian noise to an image and then removing the noise from the image/denoising by predicting the noise present and then subtracting it in order to create a clear image based on the text prompt. This text prompt acts as a guide to ensure the generated image matches the input description. One must also note that these models work in a latent space which is a compressed high dimensional space that captures the essential features or internal representation of an image.

Stable Diffusion is known for its generation of high quality detailed images and is used for inpainting and outpainting. In painting is used to paint in or fill in the gaps of missing areas in the picture or to construct a picture based on partial sections of the picture. Outpainting on the other hand is to paint beyond the picture or to extend an image beyond its original edges. Midjourney is known for producing surrealistic images and for its community efforts to allow users to provide feedback on shared artwork. DALL-E 2 created by OpenAI “is trained on approximately 650 million image-text pairs scraped from the Internet”[1] and is capable of combining different concepts and ideas into one single generated image. Due to its closed source nature, there is limited capability of producing images at a wide scale.

The data used by Borji to evaluate the models consists of a face dataset with two sets of faces, one with generated faces and one with real faces. To avoid capturing potential biases into the generated faces database, Borji chose the COCO dataset which contains captions that were used as prompts to generate images. Borji then selected captions that contained characteristic descriptive words related to faces such as ‘person’, ‘man’, ‘woman’, ‘kid’ among others and ran them through Stable Diffusion, Midjourney, and DALL-E 2. Next, he pruned false positives and the detected faces were resized to 100 x 100 with a collection of a total of 15,076 generated

faces. More specifically, " including 8,050 by Stable Diffusion, 6,350 by Midjourney, and 676 by DALL-E 2"[1] .

The three text-to-image models mentioned above were compared using the Frechet Inception Distance (FID) score. FID works by quantitatively evaluating the similarity between a real image dataset and a generated image dataset based on their high level meaningful feature representations. First, embeddings are extracted from both real and generated image datasets, then both mean and covariance of the feature representations for both types of images are computed, and finally the Frechet distance is measured between the mean vectors and the covariance matrices. The closer two vectors or in this case two probability distributions are in high dimensional space, the more similar they are. So, a lower FID score would indicate a better quality of generated images (distributions of features are closer to each other) as they are really close to the 'real thing' or real faces images.

To obtain a more accurate representative FID score, Borji shuffled both generated and real faces, selected 5,000 faces from each set, got the FID score, and repeated this 10 times to at last obtain an averaged FID score over the 10 runs. Results show that Stable Diffusion obtained the lowest FID score, generating better more realistic looking faces, followed by DALL-E 2 and then by Midjourney with the highest FID score. Borji did point out that the high FID score of Midjourney might be due to the surrealistic and anime type of generated looking faces. Interestingly, Borji also mentioned that one of the potential reasons why DALL-E 2 performs worse than Stable Diffusion is that "DALL-E 2 is optimized for images with a single focus of attention" [1] meaning it is optimized for generating images with a clear subject with minimal competing distractors - relying on the simplicity for visual clarity. Consequently, this optimization might make it difficult for DALL-E 2 to match generated face images to real face images in *complex* scenes.

By quantitatively comparing Stable Diffusion, DALL-E 2 and Midjourney through comparison of their FID score, Stable Diffusion demonstrated a higher quality of generated faces "in the wild". In fact, Borji reveals, "to the best of our knowledge, we are the first to evaluate the quality of generated faces in the wild"[1] - demonstrating the importance and relevance of applying this quantitative metric to generative models such as text-to-image models which now play a more pervasive and crucial role in our lives. And although "it is possible for a human to tell whether a face is real or generated"[1], utilizing FID score as a discrimination metric between real and generated images will be essential and prove even more helpful as generative models become better at producing real human faces over time.

Two of the several future points of improvement or additional work that Borji suggests when quantitatively comparing these text-to-image models are cross referencing other scores such as SSIM (Structural Similarity Index) and LPIPS (Learned Perceptual Image Patch

Similarity), and comparing these models based on specific facial features such as emotional facial expressions, different angles, and exploring other categories extending beyond faces. SSIM compares the *structural* similarity of images by considering for example luminance and contrast along with the spatial arrangement in an image at the pixel level. Higher SSIM score indicates closer/higher similarity. LPIPS evaluates the *perceptual similarity* between the *local patches* of images rather than focusing on the global structure by extracting feature representation using deep neural networks. Lower LPIPS score indicates closer/higher similarity. Implementing additional metrics and evaluating other aspects of a generated image can potentially render a better image quality assessment and serve useful in the advancement of generative AI models.

Conducting a literature review on Borji's work is one step closer to training Stable Diffusion model on a fashion dataset. Concluding Stable Diffusion generates higher quality pictures, specifically with faces as illustrated in Borji's work, it highlights the computational efficiency and capabilities of Stable Diffusion in generating realistic images - and a potential gateway to helping fashion designers come up with unique designs with multiple generated fashion looks to be inspired from. Stable Diffusion's architecture with an autoencoder with variation for encoding and decoding an image from pixel to latent space, forward dispersion for gradually adding Gaussian noise to an image, reverse diffusion for reversing forward dispersion's doings, and a U-Net noise predictor as the main driver behind noise predicting and denoising photos is what makes Stable Diffusion such an effective and transformative model. Exploring the capabilities of text-to-image models such as DALL-E 2, Midjourney, and especially Stable Diffusion is an opportunity to step into the leading field of AI responsible for many of the technological advancements of today.

Resources

[1] Borji, A. (2022). Generated Faces in the Wild: Quantitative comparison of stable diffusion, midjourney and DALL-E 2. *arXiv (Cornell University)*.
https://doi.org/10.48550/arxiv.2210.00586