Name: Mukund Dhar
UCF ID: 5499369

Report on **"Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding"**

The paper introduces Imagen, a model that combines the strengths of transformer language models with high-fidelity diffusion models, to provide text-to-image synthesis with an unmatched level of photorealism and language understanding. It consists of a frozen T5-XXL text encoder to map input text into a sequence of embeddings and a 64x64 image diffusion model, along with two super-resolution diffusion models for generating the high-resolution images of 256x256 and 1024x1024. Effective text conditioning in all the diffusion models vitally depends on classifier-free guidance. It is discovered that large frozen language models trained on text only are better encoders for text-to-image generation. A new diffusion sampling technique called dynamic thresholding is introduced to avoid oversaturated and unnatural images, especially when using very large guidance weights. There were many changes in the neural network architecture introduced. For the base model, the U-Net architecture is adapted and conditioned on the text embeddings and for the super-resolution models, a new variant called Efficient U-Net is used with several modifications for both 64x64 to 256x256 and 25x256 to 1024x1024 super-resolution steps. Without ever having trained on COCO, Imagen obtains a new state-of-the-art FID score of 7.27 on the dataset and in terms of image-text alignment, human raters consider Imagen to be comparable to the reference pictures. A brand new comprehensive and challenging evaluation benchmark for the text-to-image problem, named DrawBench, is also introduced that contains a set of prompts to test different capabilities of models. Imagen outperforms all other work, including the work of DALL-E 2, on DrawBench's human evaluation. Overall, Imagen is made up of two components: a text encoder that converts text into a series of embeddings, and a cascade of conditional diffusion models that map these embeddings into images with progressively higher resolutions.

Strengths: Imagen's key discovery is that text embeddings from huge language models that have been pretrained on text-only corpora are surprisingly more effective for text-to-image synthesis. More model quality improvement results are generated from scaling the frozen text encoder model than the U-Net model. It is also easier to achieve since the text embeddings can be computed and stored offline during training. It is also found that Efficient U-Net converges notably faster than U-Net and obtains better overall performance. The Efficient U-Net is also two-three times faster at sampling. The use of dynamic thresholding results in more natural looking images when using high guidance weights compared to static thresholding.

Weaknesses: Since the datasets used might contain stereotypes, oppressive viewpoints, and derogatory, or otherwise harmful, associations to marginalized identity groups, Imagen results might cause representational harm due to these social biases. Imagen displays significant limits when creating images of humans. Imagen also struggles sometimes with certain prompts, such as, from the Conflicted category.

Questions:

1. Why was swish used in the Efficient U-Net ResNetBlock in place of ReLU? How does it differ in performance?
2. Why is there an Alignment-Fidelity trade off when iterating over guidance weight?

Possible ideas:

More research on social bias evaluation methods for text-to-image models could be done so that the social and cultural bias of these models is evaluated more accurately. A conceptual vocabulary around potential harms of text-to-image models could be made so that the release of a model is well informed. Image synthesis and editing in a personalized way can also be done so that given as input just a few images of a subject, a pretrained text-to-image model is fine-tuned and then photorealistic images of the subject contextualized in different scenes are generated.