Name: Mukund Dhar
UCF ID: 5499369

Report on **"GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models"**

The paper investigates using diffusion models to solve the problem of text-conditional image generation and contrasts two alternative guidance methods: classifier-free guidance and CLIP guidance. The paper discovers that the latter frequently yields photorealistic samples and is favored by human evaluators for both photorealism and caption similarity. Text-driven image editing is made possible by the models' ability to be tuned to perform image inpainting. The system is referred to as GLIDE, which stands for Guided Language to Image Diffusion for Generation and Editing. At first, by gradually introducing Gaussian noise to a sample taken from the data distribution, a Markov chain of latent variables is created. The final image generated can well be approximated by N(0; I) if the magnitude of the total noise added throughout the chain is large enough. Hence, a model to approximate the true posterior starting with Gaussian noise and progressively lowering the noise in a series of steps is suggested. The model is trained to predict the added noise using a standard mean-squared error loss after it generates samples by applying Gaussian noise. Two guidance methods are introduced for image generation: Classifier free and CLIP guidance. To implement classifier-free guidance with generic text prompts, it replaces the text captions with an empty sequence occasionally so that output of the model is extrapolated further in the direction of the class-conditional diffusion model by adding the difference using a guidance scale during sampling. For CLIP guidance, since the CLIP model provides a score of how close an image is to a caption, the classifier is replaced with it in classifier guidance. In particular, the reverse-process mean is perturbed with the gradient of the dot product of the image and caption encodings with respect to the image. Visually contrasting samples from classifier-free guidance with those created with CLIP guidance, it is found that the samples from classifier-free guidance frequently appear more realistic. Even in configurations that greatly favor DALL-E by letting it use a much larger amount of test-time compute (via CLIP reranking) while lowering GLIDE sample quality (by VAE blurring), the GLIDE model is preferred by the human evaluators in all settings.

Strengths: The paper introduces the GLIDE model which applies guided diffusion to the problem of text-conditional image synthesis. The guided diffusion models generate photorealistic images and are capable to handle free-form prompts. While the model can produce realistic images for a wide range of text prompts zero-shot, it may have some trouble rendering sophisticated prompts with realistic images. As a result, in addition to zero-shot generation, it is also given editing features, allowing users to iteratively enhance model samples until they correspond to increasingly intricate prompts. It was also observed that the smaller models in comparison often fail at binding attributes to objects and perform worse at compositional tasks.

Weaknesses: The paper points one of the limitations of the introduced model being when the text prompts are describing highly odd items or situations, it occasionally misses to capture them and outputs unexpected images. In addition, the unoptimized model also takes longer to sample as compared to GAN related methods and hence, become less favorable for real-time implementations. GLIDE also exhibits some biases when generating and filtering images that go beyond the image datasets: for example, the hate symbol classifier used only considers Western and American symbols.

Questions:

- If the classifier-free approach of the model gives such good results for the unconditional image generation, it hints to make questions on whether we are utilizing the potential of using the data information completely.
- What are real-world applications of using this model that takes more time to train than others? What are the different computer vision tasks that can utilize this model?

Possible ideas:

The GLIDE model decodes embeddings to generate images. This can be done in a better way by using model learning image-text representations. Also, instead of training from scratch as done for GLIDE, a frozen Transformer model trained on a massive corpus can be used. GLIDE needs to also modify the model architecture if they want to perform image inpainting and editing, by using an image or a text embedding as conditioning, it might be possible to combine image guidance with conditioning to fill in the missing areas of an image and edit the image in the way that is required.