Name: Mukund Dhar

UCF ID: 5499369

Report on **“DreamBooth: Fine Tuning Text-to-Image Diffusion Models**

**for Subject-Driven Generation”**

The paper introduces a way to mimic the appearance of subjects in a given reference set and synthesize novel renditions of them in different contexts. A pretrained text-to-image model (Imagen) is fine tuned so as to learn a unique identifier of a subject given its few images. The unique identifier is used to synthesize novel images of the subject with high fidelity contextualized in various scenes after the subject has been embedded in the model’s output domain. This is achieved in two steps: a low-resolution image from text is generated and then super-resolution diffusion models are applied to it. The tasks such as subject recontextualization, text-guided view synthesis, appearance modification, and artistic rendering can be challenged using this technique. The paper also proposes an autogenous, class-specific prior preservation loss that makes use of the model's embedded class's semantic prior to urge it to produce a variety of instances of the same class as that of the subject. This is done to prevent overfitting and language drift. The super-resolution component is also fine-tuned using the low resolution and high-resolution versions of the input images.

**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.