Name: Mukund Dhar
UCF ID: 5499369

Report on **"Cascaded Diffusion Models for High Fidelity Image Generation"**

The paper shows that generating high fidelity images is possible using cascaded diffusion models without the use of any auxiliary classifiers. Cascaded Diffusion Models (CDMs) consists of a pipeline of multiple diffusion models that generate images of increasing resolution. A standard diffusion model is used at the lowest resolution to generate an initial image that is passed to one or more super diffusion models that upsample the image successively and add higher resolution details. The paper shows that images generated by CDMs are superior to BigGAN-deep and VQ-VAE-2 in terms of the FID score and the classification accuracy score. This is achieved using pure generative models that are not combined with any classifier. In conditional generation, the diffusion model is modified to include a conditioning signal as input to the reverse process. The data and conditioning signal are sampled jointly from the data distribution and the forward process remains unchanged. A U-Net architecture is used for all the diffusion models and for the first model, scalar conditioning is provided by adding embeddings for a class label and a diffusion timestamp into the intermediate layers of the network. For the remaining diffusion models, the reverse input image along with the class label is concatenated channelwise with the lower resolution image. It is also discovered that the sample quality of the cascaded pipeline depends on conditional augmentation. The paper also explores different augmentation polices for the super resolution models to find Gaussian augmentation to be key for low resolution upsampling, and Gaussian blurring for high resolution upsampling. The sample quality is improved using conditional augmentation since the compounding error in cascading pipelines is avoided. The paper also proposes methods to train and test models amortized over varying levels of conditional augmentation.

Strengths:

- The paper focusses on improving sample quality by cascading diffusion models without classifier guidance and beats the models such as BigGAN-deep and VQ-VAE-2 on the ImageNet class-conditional generation benchmark.
- Conditioning augmentation experiments done in the paper provide a major insight on their effects on improving the sample quality.
- The hyperparameters of the models and the architecture used in the ImageNet cascading pipelines are given in a detailed way in the paper.

Weaknesses:

- The CDMs do not outperform the concurrent work of ADM with classifier guidance (Dhariwal and Nichol, 2021) in terms of FID and Inception scores.
- It takes thousands of diffusion timesteps in the low-resolution models to generate images, hence consuming a lot of time and being computationally expensive.
- There is no mention of how to tackle the negative impact of generating high fidelity images for malicious uses.

Questions:

- Under what conditions would the truncated conditioning augmentation be used during sampling instead of non-truncated one since they perform similarly, and the author recommends using the non-truncated augmentation for its practical benefits?
- The paper mentions about diffusion models with classifier guidance running the risk of cheating metrics such as FID and Inception scores since they themselves are computed on activations of an image classifier trained on ImageNet. What is a better metric to evaluate models in such cases that use classifier guidance?

Possible ideas:

- The cascading techniques shown in the paper can be combined with classifier guidance which might further improve the results on the sample quality.
- The Cascaded Diffusion models can be utilized in further research on improving the results on downstream tasks such as data compression.