Graphics Generation with User Inputs

ndjoshi@andrew.cmu.edu

1. Motivation

One might argue that creating graphics for given user requirements is best done by an artist. As an example, shown in Fig. __ is a personal branding attempt by the artist where they try to create graphics for the hypothetical situation, “if big tech companies were to sell cookies” (©Yassine Chouk). Clearly, the reproduction is highly creative and witty, as one would expect from an intelligent human being.

Figure 1: Human rendering of “if Google were to sell cookies”

Alternatives to purely encoder-decoder based models are diffusion models like Imagen [2] and stable diffusion models like [3] that have mechanism to upsample denoised images to render very high quality in terms of pixel density. Another advanced stable diffusion model creates a depth image from the input image and uses denoising, constrained to the shallow (focused) area to render visualisation, conditioned upon given text prompt.

The main component from which all the above models start, is the tokenizer. This is a pretrained encoder that creates hidden tokens out of text or image prompts, use these as initial reference or else as conditional priors for denoising. This is where we should aim to fuse custom user inputs into the generative models.

The intention of this project is to seek a black-box architecture that may take custom user inputs to output a stylised graphic, that may be used for branding and in advertising campaigns. A parallel study is to gauge how capable generative AI has become till date and whether it is at a stage to effectively replace human creativity and brilliance.

2. Proposed Methods

Text-to-image generation models such as DALL-E [1] produce highly plausible visualisations conditioned on the text input and a prior. Here, the art of manufacturing a perfect sentence as the input, lies in the hands of the user. The inherent randomness leads to variations still conditioned on the input text, not desirable for the current task. Also, it can be made more user-friendly by providing user inputs for expected qualities such as colour, font, icon, etc. rather than putting all these as a single text prompt.

It is conjectured that the model being trained on data gathered from open sources, will perform well in case of well known brands (i.e. Google, Microsoft, etc.). The main task to address is how to fuse user inputs at the appropriate stage so as to reduce the reliance on detailed text prompts.

3. Experiments

Owing to package dependency issues, the experiments were conducted using pretrained weights, without fusing user inputs to the hidden tokens. A comprehensive text prompt was provided to the effect of “Packet of Cookies branded by Company_XYZ”.

4. Results

Some results are included in Figures 2 and 3. The top figure in Figure 2 is the result for YouTube. Clearly, the model has done efforts to cast the branding colours (red, white, black) into the image, although that expands beyond the packet. The bottom image is the output for Microsoft where it tries to draw the Windows logo and juxtapose it with picture of cookie. These aspects seem to improve in the stable diffusion model outputs (Figure 3).

To see more such results from both the models, refer Appendix I.

Figure 2: diffusion model output for (top) YouTube, (bottom) Microsoft

5. Further Work

- Using stable diffusion with depth map
- Training fresh with fused tokens from the user inputs
- Creating UI that can be easily used to give range of inputs to the model

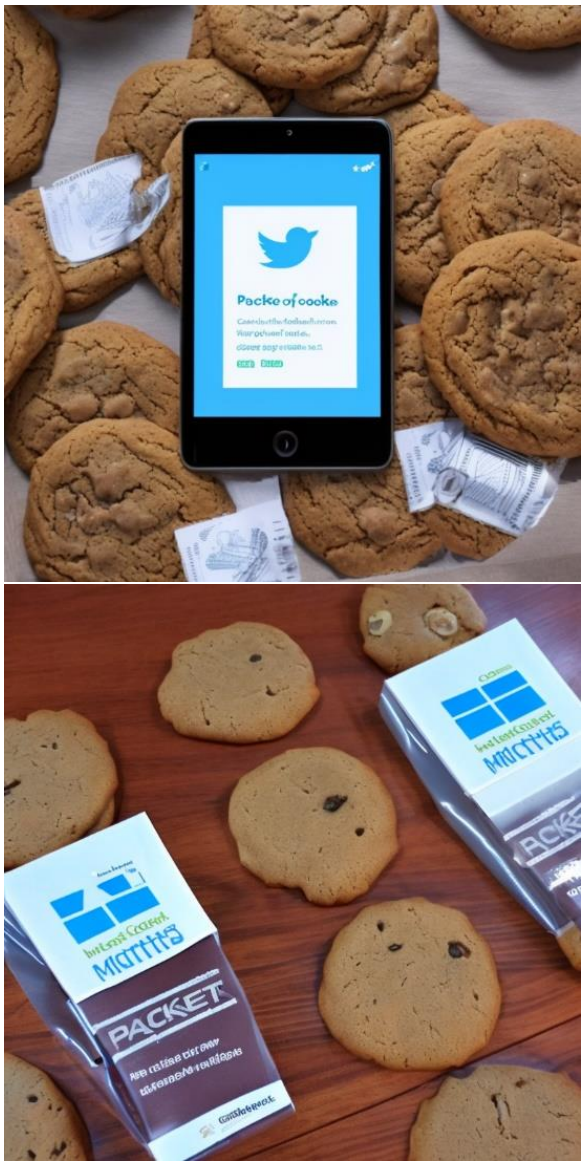

Figure 3: Stable diffusion model output for (top) Twitter, (bottom) Microsoft

References

[1] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. and Sutskever, I., 2021, July. Zero-shot text-to-image generation. In *International Conference on Machine Learning* (pp. 8821-8831). PMLR.

[2] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G. and Salimans, T., 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487*.

[3] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684-10695).

Appendix I

Stable diffusion results for Amazon, Apple, Google and LinkedIn (clockwise).