Exploring and Analyzing ControlNet for Image Generation

Sanjana Nambiar
svn9705@nyu.edu

Saamia Shafqat
ss14758@nyu.edu

Our research centers on ControlNet, an encompassing suite featuring nine models crafted for text-to-image synthesis. Our specific focus lies on the Canny model, a pivotal component nested within the Stable Diffusion model variant integrated into ControlNet. Renowned for its adeptness in incorporating conditional control, the Stable Diffusion model stands at the forefront of generative capabilities.

In this exploration, our attention is directed towards the utilization of the Canny model within the Stable Diffusion framework, specifically evaluating its performance in regenerating images from a curated dataset of horses. The research emphasizes the architecture and functionality of ControlNet within the context of our dataset, aiming to gauge its accuracy in image regeneration. Through this investigative journey, we aim to shed light on the practical utility of ControlNet within our horse dataset, identifying strengths and potential areas for improvement. Notably, our research unravels interesting findings concerning image resizing and its consequential impact on image generation, contributing nuanced insights within the realm of our distinct dataset.

Methodology

A dataset comprising equine images was employed, characterized by its diversity in capturing various poses, backgrounds, and lighting conditions. Each image within this dataset played a foundational role in the subsequent image generation process.

The application of Canny edge detection to each image in the dataset served the dual purpose of extracting essential structural information and significantly reducing the data to be processed. This technique facilitated a focused analysis of the edges and contours inherent in the visual elements of the images, resulting in the creation of what is conventionally referred to as a "Canny image."

Two pivotal inputs were introduced into the image generation process: a text prompt and the Canny image obtained through preprocessing. The text prompt functioned as a creative guide for the model, exerting influence over the stylistic and content aspects of the generated images. Concurrently, the Canny image contributed an enhanced representation of the structural features present in the original equine images. This amalgamation of textual guidance and enhanced visual representations constituted a key dynamic within the image generation paradigm.
In the initial phases of image generation, instances were observed wherein certain outputs received NSFW (Not Safe For Work) flags. These flagged outputs, accompanied by the generation of black images and associated error messages suggesting alterations to the prompt or seed input, failed to alleviate the occurrence of NSFW image generation. The persistence of this issue may be linked to variations in image size across the dataset, posing challenges to maintaining consistency throughout the generative process.

During the normalization of the dataset to ensure accurate LPIPS calculations, a standardized approach was adopted, resizing input images uniformly to dimensions of 500 by 500. Unexpectedly, this adjustment in image size yielded a substantial improvement in mitigating NSFW errors compared to the previous dataset. The standardized dimensions not only ameliorated issues associated with NSFW outputs but also markedly enhanced the overall quality of the generated images. This normalization process contributed to a more controlled and refined creative output. Remarkably, in certain instances, the generated images exhibited superior quality to their original counterparts. However, it is imperative to underscore that the primary objective is not solely to produce aesthetically pleasing images but rather to generate images that faithfully mirror the characteristics of the original input.

The evaluation of model accuracy in this study involved the exploration of FID (Fréchet Inception Distance) score and LPIPS (Learned Perceptual Image Patch Similarity) torch metric. The selected metric, LPIPS, operates by accepting two input tensors and providing a numeric representation indicative of the similarity between the corresponding images. The image transformation process was initiated to convert them into tensor objects, enabling the subsequent computation of their similarity.

The numeric output from the LPIPS metric serves as a quantitative measure, where a higher value, closer to zero, signifies lower similarity, whereas a lower value implies greater similarity between the compared images. This analysis extended to the investigation of the impact of image resizing on the quality of generated images. Notably, resizing resulted in not only improved image quality but also yielded images with a higher similarity index when compared to the original dataset.

An observed challenge emerged concerning outliers in similarity scores. Specifically, instances where NSFW (Not Safe For Work) images, characterized by an all-black composition, were generated from horse images that were originally black or possessed a higher black pixel ratio. In such cases, the similarity metric tended to produce higher scores, introducing an aspect of distortion in the overall similarity assessment.

Constructing the ControlNet Synthesis Pipeline

In our text-to-image synthesis research, we employ a ControlNet model pipeline tailored to our equine dataset. Acknowledging the complexity, we integrate a pre-trained ControlNet model from the "lllyasviel/sd-controlnet-canny" repository. This model, enriched through diverse datasets, serves as the foundation, facilitated by the ControlNetModel class.

Our synthesis pipeline incorporates the Stable Diffusion model, synergized with ControlNet capabilities, utilizing the "runwayml/stable-diffusion-v1-5" model. The orchestration of this pipeline involves a UniPCMultistepScheduler from the diffusers library, ensuring controlled evolution in synthesizing images.

Results

In the initial phase of our study, we conducted image generation from the original dataset, encompassing images of diverse sizes. Notably, this process yielded a significant number of NSFW (black) images. The mean accuracy score for this dataset of generated images stood at 0.79.

As a pivotal step forward, we standardized the image size to 500x500 pixels in the subsequent phase. Remarkably, this resizing operation not only alleviated the prevalence of NSFW (black) images but also led to a substantial enhancement in the overall quality of the generated images. The mean accuracy score for this resized image dataset recorded a notable improvement, reaching 0.72.

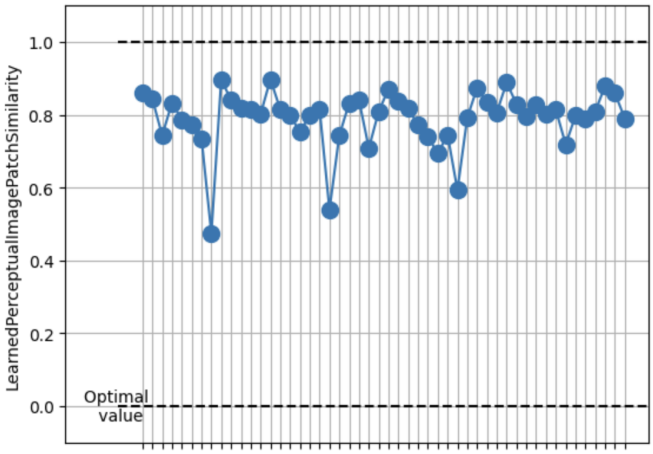

Fig 1: LPIP value plot for dataset without resizing

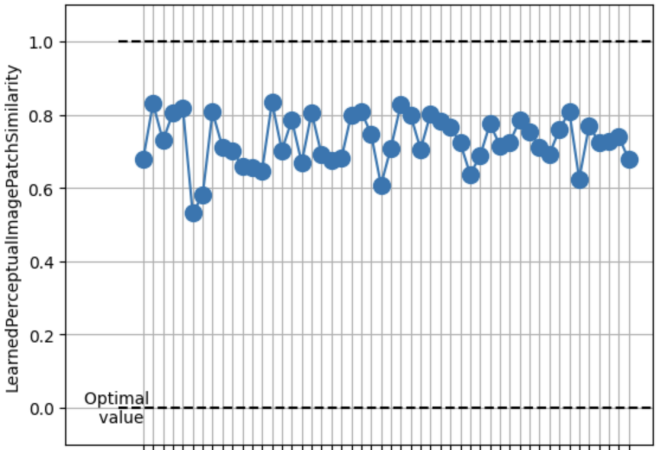

Fig 2: LPIP value plot for dataset with resizing

A visual depiction of the comparison between the generated images would provide a more detailed illustration of the significant differences observed.

Fig 1: Without Resizing

Fig 2: Original Image

Fig 3: With Resizing

In contemplating the future trajectory of this research, there exists the prospect of refining accuracy assessments. An avenue for consideration involves the pre-filtering of NSFW images and outliers before computing accuracy scores, offering a more nuanced evaluation.

Furthermore, our investigation uncovered a noteworthy insight regarding the limitations of the LPIPS method. While this method facilitates pixel-to-pixel image comparison, it lacks consideration for locality and shapes, focusing solely on pixel colors. This underscores the need for future enhancements that encompass a more comprehensive evaluation, taking into account the spatial aspects of image generation.

For further exploration with the analysis of this project:
https://github.com/Sanjana-Nambiar/controlNet_Analysis

Demonstration URL:
https://www.canva.com/design/DAF3oEAPac4/OjgAjg-kpfr7BEfs0Zgi7Q/view?utm_content=DAF3oEAPac4&utm_campaign=designshare&utm_medium=link&utm_source=recording_view

References

1. Lllyasviel. (n.d.). lllyasviel/ControlNet: Let us control diffusion models! Retrieved from https://github.com/lllyasviel/ControlNet
2. Sanagapati, P. (2018). Images Dataset. Retrieved from https://www.kaggle.com/datasets/pavansanagapati/images-dataset/
3. Mseitzer. (n.d.). mseitzer/pytorch-fid: Compute FID scores with PyTorch. Retrieved from https://github.com/mseitzer/pytorch-fid
4. Learned Perceptual Image Patch Similarity (LPIPS)¶. (n.d.). Retrieved from https://lightning.ai/docs/torchmetrics/stable/image/learned_perceptual_image_patch_similarity.html
5. Lim, B. (2023). Stable Diffusion on Mac: A Step-by-Step Guide to Local Installation. Retrieved from https://medium.com/@brucelim/stable-diffusion-on-mac-a-step-by-step-guide-to-local-installation-f3c8a64ee67a
6. (N.d.). Retrieved from https://huggingface.co/lllyasviel/sd-controlnet-canny
7. Ashvanth.S. (2023). SDXL Openpose: A Deep Dive into Effective Parameter Choices. Retrieved from https://blog.segmind.com/best-settings-for-sdxl-openpose/