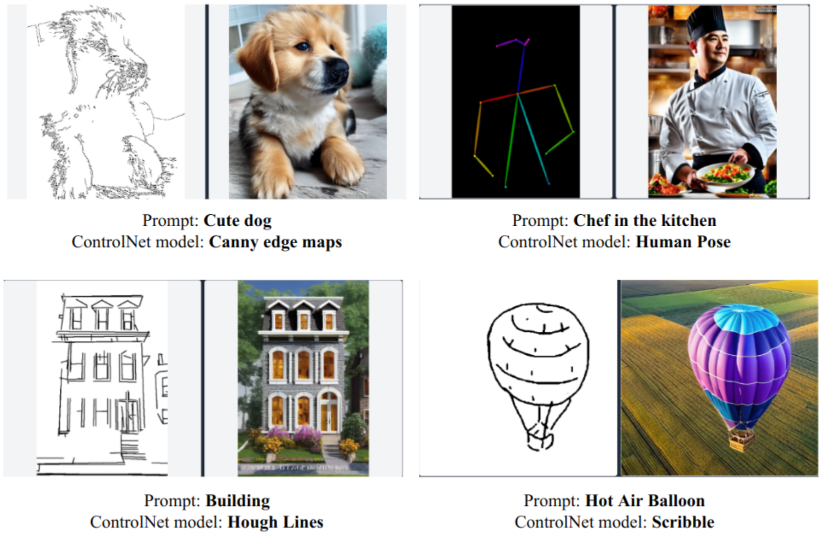
Training a ControlNet for Stable Diffusion

Spring 2023



Prompt: **Cute dog**
ControlNet model: **Canny edge maps**

Prompt: **Chef in the kitchen**
ControlNet model: **Human Pose**

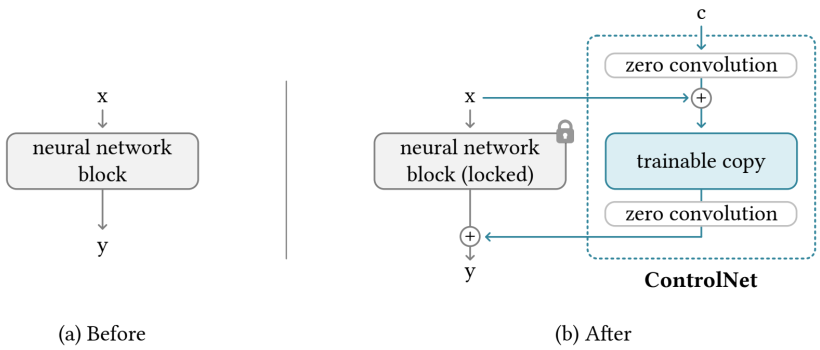
Prompt: **Building**
ControlNet model: **Hough Lines**

Prompt: **Hot Air Balloon**
ControlNet model: **Scribble**

Figure 1: Example outputs from different ControlNets trained on various conditions. ControlNets are fed the input prompt and the conditioning image (left on each pair), and produce a high quality image (right on each pair). Separate ControlNets are trained for each condition.

1 Introduction

The goal of this project is to train a ControlNet [2] to control Stable Diffusion [1] on a new condition. ControlNet is a deep learning algorithm that can be used for controlling image synthesis tasks by taking in a control image and a text prompt, and producing a synthesized image that matches the prompt and follows the constraints imposed by the control image. For example, ControlNet allows you to generate an image based not only on a prompt, but also on a basic



(a) Before

(b) After

Figure 2: Visualization of the ControlNet setup.

sketch that defines the general shape and position of the objects in your image (Figure 1).

ControlNet essentially proposes to freeze the original Stable Diffusion UNet, while instantiating a set of trainable copies for particular blocks. The trainable copies, alongside ”zero convolution” blocks, are trained to receive a condition and integrate that information into the main model (Figure 2).

This project proposes to train a new condition and qualitatively analyze the results in terms of prompt fidelity, condition fidelity and quality of the resulting imagery. This can be done either through an available toy dataset, Fill50k[1], through a dataset that you might find online, or through your own synthetically created dataset.

We wish to document the training process, categorize challenges found, and thoroughly analyze the resulting model in terms of quality and condition fidelity.

2 Objectives

The main objectives of this project are:

1. To train a ControlNet on a condition of your own. This can be either done with the toy Fill50k dataset, which contains 50,000 images of circles with prompts and corresponding filled circle images, or through your own dataset following either one of the existing ControlNet conditions (scribble, pose, canny edge maps, etc) or a new condition of your choosing.

2. To evaluate the performance of the trained ControlNet on a properly defined test set. The performance analysis should include a qualitative analysis of prompt fidelity, condition fidelity and image quality.

[1] Dataset can be found here: https://github.com/lllyasviel/ControlNet/blob/main/docs/train.md

3 Methodology

3.1 Suggested steps

: You are free to follow any strategy to achieve the aforementioned goals. However, here are a set of steps that we suggest you to follow:

1. Dataset preparation: Either download the Fill50K dataset or find/create your own. In both cases, ensure that you have train and test splits.

2. ControlNet training: Train a ControlNet on the training set using the PyTorch framework. The ControlNet will take in a control image and a text prompt and output a synthesized image that matches the prompt.

3. ControlNet evaluation: evaluate the performance of the trained ControlNet on the test set. Qualitative evaluation is sufficient, but feel free to explore the literature for quantitative metrics as well.

4. Result analysis: Analyze the results and identify potential areas for improvement. Enumerate challenges and distill conclusions.

3.2 Resources

Here is a list of useful links and resources that can help you get started:

1. Main ControlNet Github

2. ControlNet training instructions

3. Colab with ControlNet examples

3.3 Challenges

Training stable diffusion with ControlNet will require significant computational resources. We recommend you to use Colab, Runpod or cloud compute to facilitate this work. Feel free to use resources such as https://github.com/jehna/stable-diffusion-training-tutorial to guide your setup.

4 Expected Results

We expect to obtain a ControlNet that can effectively control generations with the given conditions, and a report that indicates how the process was conducted. You should describe your dataset decisions, in terms of choice of data, preprocessing and splitting; you should explain your training process and challenges found; and you should qualitatively analyze the model, distill conclusions and find potential areas for improvement.

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[2] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.