CS182 Project

Fine tuning in vision context

Ray Wang, Logan Ju, David Long, Daniil Samylovskikh

Advisor: Anant Sahai

Department of EECS
UC Berkeley

DATE SUBMITTED
[November 29, 2023]

Introduction

Urban environments are rich tapestries woven from varied elements such as architecture, street networks, and green spaces. While much of the existing work as such has taken land use and urban function–based measures as key variables to estimate neighborhoods' socioeconomic status, our project recognizes the profound impact of visual elements in urban settings. The project is grounded in existing research that links urban design to various social and economic outcomes. By fine-tuning a stable diffusion model, we propose to edit existing street views, adjusting elements, styles, and overall design quality to create more engaging and sustainable urban environments, and provide insights in the practise of urban planning design.

Data Generation

Goal

Let S be an image synthesized by a text-guided diffusion model utilizing the text prompt P and a random seed s. Our objective is to edit S, guided solely by modifications to the original prompt P, to produce a revised image I.

For instance, take an image that was rendered in response to the prompt: 'The image depicts a residential street lined with houses. Parked cars are visible along the road, and trees dot the sidewalks. The weather is sunny with a lone cloud adorning the sky. The road is neatly paved, and the houses are constructed with red brick, contributing to a serene and tranquil atmosphere.' Now, if the user's intention is to retain only the trees, the most user-friendly approach would be to adjust the text prompt by omitting any mention of the cars.

Figure 1: Streets with cars and without cars.

Method

Our dataset comprises two sets of textual prompts: the original set and a modified set, with the latter featuring slight alterations from the former. We have developed an algorithm that facilitates controlled image generation by concurrently executing the iterative diffusion process on both sets of prompts. The foundational algorithm for this technique is adapted from this source.

Algorithm 1 Prompt-to-Prompt image editing

Input: A source prompt P, a target prompt P^*, and a random seed s.
Output: A source image x_{src} and an edited image x_{dst}.
$z_T \sim \mathcal{N}(0, I)$ a unit Gaussian random variable with random seed s;
$z_T^* \leftarrow z_T$;
for $t = T, T-1, \ldots, 1$ **do**
    $z_{t-1}, M_t \leftarrow DM(z_t, P, t, s)$;
    $M_t^* \leftarrow DM(z_t^*, P^*, t, s)$;
    $M_t \leftarrow Edit(M_t, M_t^*, t)$;
    $z_{t-1}^* \leftarrow DM(z_t^*, P^*, t, s)\{M_t \leftarrow \hat{M}_t\}$;
return (z_0, z_0^*)

The main challenge is to preserve the original composition while also addressing the content of new prompt.

Figure 2: Streets with red trees and green trees

Fine-tuning

Given the limited compute resources, we initially decided to fine tune the diffusion model. However, after researching web for similar ideas, we stubled across ‘InstructPix2Pix‘ paper. It introduced the idea of fine tuning the diffusion model for the specific needs. In the original paper, the authors used a multi-modal training dataset. The examples include changing style of an image and adding small details - mostly related to color transformations and adding some texture. Out task, however, is related to big spacial transformations that InstructPix2Pix is not good at. Nonetheless, the architecture of the model - image + prompt as input and image as output, suits our needs well. All was left is to collect the needed data and fine tune the model. Experimentally, we found that fine-tuning InstructPix2Pix gave better results compared to fine-tuning plain diffusion model - as it probably learned to differentiate spacial objects and work better with them. Below are few results (now only trained on 1k iterations):

Figure 3: ”Change the green trees in the picture to Cherry Blossom trees” prompt

Figure 4: ”Remove the power lines on the top of the bridge” prompt

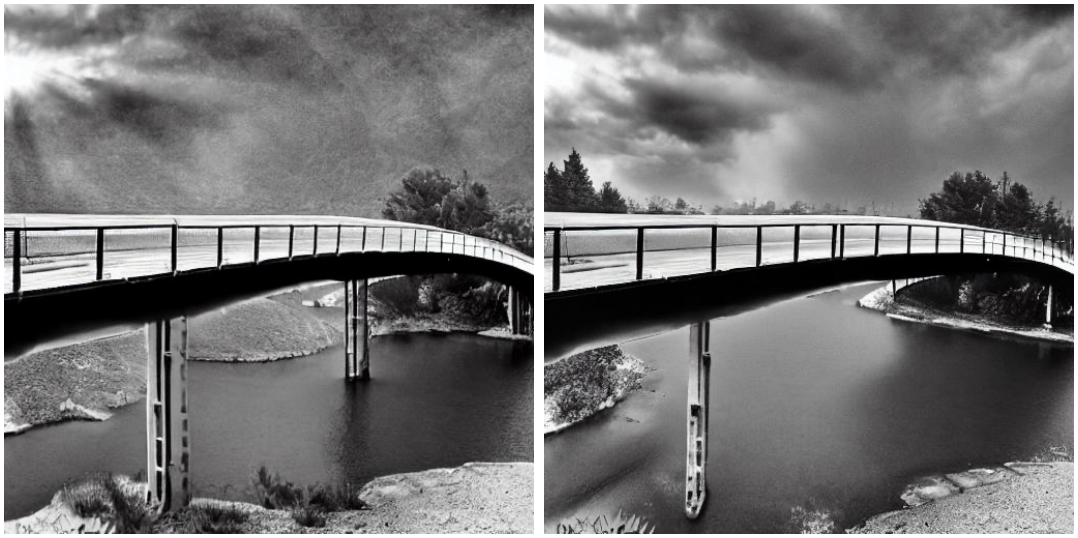

Figure 5: ”Remove the large gaps between the metal bars” prompt

Challenges

Token Limit Constraints

The intricacy of the generated images is contingent upon the length of the provided prompts. With more extensive prompts, it becomes challenging for stable diffusion models to encapsulate all the detailed aspects within the prompt. Furthermore, the token limit is capped at 77, which is exceeded by most of our prompts. Consequently, we have opted to utilize only those prompts where the token count remains under this threshold.

Overcoming Token Limitations through Embedding

To mitigate the issues imposed by token limits, token embedding presents a viable solution. This technique is widely adopted in facilitating stable diffusion models. Nonetheless, our evaluation indicates that the prompt-to-prompt method, as delineated in the work we have cited from this source, does not accommodate numerical inputs.

Links to work done

Code for fine tuning - https://github.com/Aldenysq/182proj
Generated dataset - https://huggingface.co/datasets/aldenn13l/182-fine-tune
Final model - https://huggingface.co/aldenn13l/geo-finetuned

Additional:self-review

Goal, Task, and Claims

In this paper, we aim to develop a model, InstructPix2Pix, capable of editing street view imagery through natural language instructions, primarily serving urban designers, planners, architects, and researchers. This tool facilitates the modification of urban elements within street views, streamlining the urban design and decision-making process. Our primary task involves fine-tuning InstructPix2Pix to accurately interpret and execute diverse linguistic commands into precise visual edits, thereby bridging the gap between verbal descriptions and their corresponding visual representations in urban planning contexts.

Dataset

Our dataset is organized into three distinct parts, each tailored to support the development and fine-tuning of advanced AI models.

1. **Manual Annotations:** The first section comprises manually crafted content, including pre-edited and post-edited descriptions of street view images, along with specific editing instructions. This curated dataset serves as the foundational layer for fine-tuning GPT-3, ensuring it learns from high-quality, human-generated data.
2. **GPT-3 Generated Content:** Building on the first part, the second segment features an expansive collection of caption pairs and corresponding editing instructions, all generated by the fine-tuned GPT-3 model. This large-scale, AI-generated dataset allows for a broader range of scenarios and linguistic variations, enhancing the model's versatility and adaptability.
3. **Stable Diffusion Image Pairs:** The third part leverages the caption pairs from the second section to generate corresponding street view image pairs using Stable Diffusion. This process synthesizes visual data that mirrors the textual transformations suggested in the captions, creating a rich visual dataset for further model training.

Finally, we utilize both the image pairs and the instructional text to fine-tune InstructPix2Pix. This comprehensive approach, integrating text and image data across multiple stages, ensures a robust and nuanced training process, ultimately leading to more sophisticated and accurate AI models.

Results and Strength

For a detailed understanding of how our results align with and support our overarching goal, please refer to the 'Fine-tuning' section. The primary strength of our fine-tuned model lies in its enhanced performance in tasks related to identifying, editing, and generating urban elements. Compared to the InstructPix2Pix model prior to fine-tuning, our refined version demonstrates a superior understanding of urban environments' perception, leading to more accurate and contextually relevant image modifications.

Limitations and Weakness

For an in-depth analysis of the limitations of our model, please see the 'Challenges' section. The primary weaknesses identified include the model's inability to capture fine details in street view images and its limited understanding of complex geographical contexts. Specifically, it struggles with concepts like place identity and land use and land cover (LULC), which are crucial for a comprehensive interpretation of urban spaces.

Future Work

In our approach to enhancing the model's performance, we plan to incorporate a diverse range of training sets featuring high-quality, detailed descriptions. This strategy is aimed at improving the model's precision in capturing intricate details. Additionally, integrating other advanced neural networks specialized in analyzing street view images could significantly augment the description generation capabilities of GPT, leading to more accurate and contextually relevant outputs.