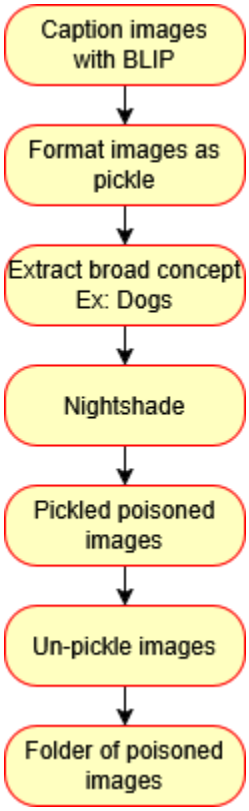
Nightshade Documentation

Tips and Tricks:

- If you look at our repo, the requirements.txt will tell you exactly how to set up your conda environment.
- There is an environment.yml for creating a quick conda environment on pre-blackwell Nvidia GPUS.

System Breakdown

Problem: We don't want our photos scraped and used to train image diffusion models, so we are going to poison our public images to disrupt unauthorized training.



Step 1: Gather a set of images and run the following commands to caption and pickle the image.

```
Bash ./prepare_a_directory.bash input_directory output_directory
```

Step 2: Extract broad image concept. TBD

Step 3: Pass prepared images through Nightshade

```
python3 data_extraction.py --directory example-data/data/ --concept dog --num 100
--outdir example-data/selected_data/
python3 gen_poison.py --directory example-data/selected_data/ --target_name cat
--outdir example-output/ --eps 0.04
```

Step 4: Run the unpickle script to extract a folder of poisoned images. [WIP]

Testing

This attack is designed to poison an image diffusion model. So we recommend fine-tuning an open-source model on poisoned data.

Example: https://github.com/CompVis/stable-diffusion

Step-by-Step

Expected results

Example 1:

Prompt: “a photo of a building with lots of windows”
Poisoned concept: building

AWS Integration [WIP]