TEXT-GUIDED IMAGE MANIPULATION USING CLIP

CADE BRUCE

Abstract. We implement a model for text-guided style manipulation. Given an input image and text prompt, our goal is to keep the image's identifiable structure and form intact while adopting style and local texture associated with the text. Our model is a DDIM (denoising diffusion implicit model) residual UNet guided by CLIP through the use of loss functions.

1. Background and Related Work

CLIP[RKH+21] is text-image embedding model that successfully connects the text and image domains in a shared representation space. Recent successful approaches to zero-shot text-guided image manipulation involve mapping the image and text into the latent space of a pre-trained CLIP model and then finding an appropriate vector direction in this representation that moves the image towards the text condition. Models seeking to transfer only texture must extract the semantic texture/style and regularize training through clever loss functions (so as not to corrupt or significantly perturb the non-textural elements of the image). The recent CLIPstyler[KY21] model by Gihyun Kwon and Jong Chul Ye takes this approach with impressive results by computing a loss over patches (among other losses) to deliver local spatially invariant information. Specifically, they spatially compare in the CLIP space by using a reference image. The CLIP losses then inform the update of a Residual UNet to produce the desired image.

Recently, Katherine Crowson designed a models to use a DDIM (Denoisng Diffusion Implicit Model) UNet with similar approach (there is no paper reference, but see: https://github.com/crowsonkb). There are also successful VQGAN+CLIP methods which invert a pre-trained GAN [XYXW21]. However, in the latter approaches, the non-local semantics of the image are not preserved as strictly.

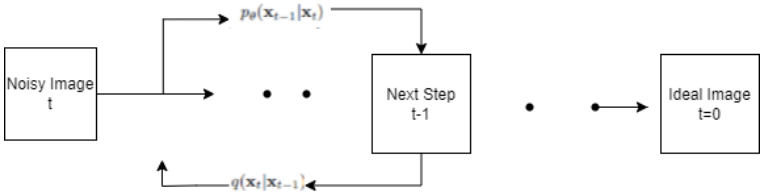

Figure 1. Standard DDIM Update Logic

Date: December 11th, 2021.

2. Our Approach

In this project, we implement a version of the existing Diffusion Guided CLIP model (https://github.com/crowsonkb). However our design has the goal of better preserving non-local information when compared to the original implimentation. To do so, we use loss functions and regularization introduced by the CLIPstyler model. Also inspired by the CLIPstyler model, we use also introduce a pretrained VGG16 and extract the features associated with our images. This enables us to compute an additional loss called the Content Loss to help ensure the semantic content of the input image isn't too significantly disturbed. We describe the content loss in the next section.

Notice that this is a zero-shot process where the images we input are not bound by a training setting.

3. Loss Functions

We describe our loss functions: The CLIP Loss is given generally by

$$L_{dir} = 1 - \frac{\Delta I \cdot \Delta T}{|\Delta I||\Delta T|};$$

Figure 2. Caption

Where ΔT is the difference in text CLIP representations and ΔI is the difference in image CLIP representations.

Firstly, we have a Patch Loss (as introduced by CLIPstyler) where we:

* Take random croppings of the image (outputted by the UNet) with perspective augmentations
* Map the patches into CLIP
* Compute the clip loss of these croppings with respect to the input image and text. Since there is no text part for the croppings, we use a general reference text. For example "A Picture" or "A Man" depending on the context of the input image.

We also have the global clip loss L_g, which is simply the CLIP loss: letting ΔI be between the input image and the entire image from our UNet, and ΔT be between our text prompt and the reference text.

The Content Loss L_c is given by the MSE (mean squared error) between the features associated with our input image and the image from our UNet associated with a pretrained VGG16. More specifically, the features are taken to be the values of the VGG's convolutional layers "conv4 2" and "conv5 2" .

The final loss is the total variation regularization loss given by the sum of the absolute differences for neighboring pixel-values.

Then our total loss is a weighted sum of all these losses (and the weight of each is manually adjusted).

4. Results

Taken very early in training for time-sensitive purposes:

Figure 3. Original Image

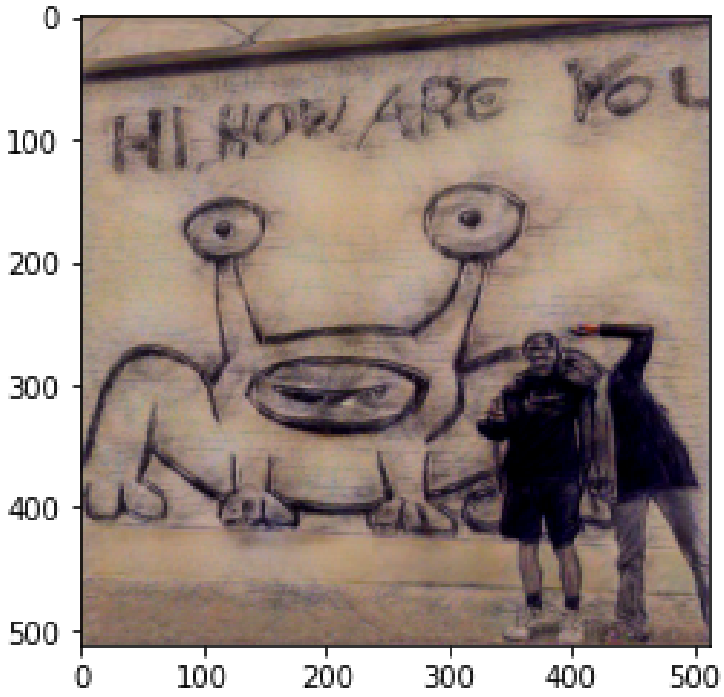

FIGURE 4. "Drawing with Black Ink"

The results can only really be qualitatively judged compared to other methods, so test your own images in the notebook:
https://colab.research.google.com/drive/1-Uw82qshrVfu37cuznEN2f7Dx8IIpxWp?usp=sharing

REFERENCES

[KY21] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition, 2021.

[RKH+21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[XYXW21] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation, 2021.