Automatic Sketch Colourization

Yuanfang Li (yli03@stanford.edu), Xin Li (xinli16@stanford.edu)

Introduction

- Sketches have various applications in our daily life. Besides being a form of art, sketching is a good way to graphically record scenes and individuals, as well as demonstrate an idea.

- Colourizing black-and-white sketches is challenging and time-consuming. Our project aims at automatic landscape colourization, which can be applied practically to fields like CGI to reduce the amount of time spent on colouring backgrounds and allow artists to focus on characters.

Related Work

- Non-deep learning approaches to colourization typically use either colour scribbles or colour transfer from an example image.

- Traditional CNN methods use convolution/deconvolution layers and minimize a L1 or L2 pixel loss, which can lead to desaturated colours and blurry images.

- Using conditional GANs overcomes this by allowing the network to learn a suitable loss for the high-level objective of generating a realistic image given some input.

Data

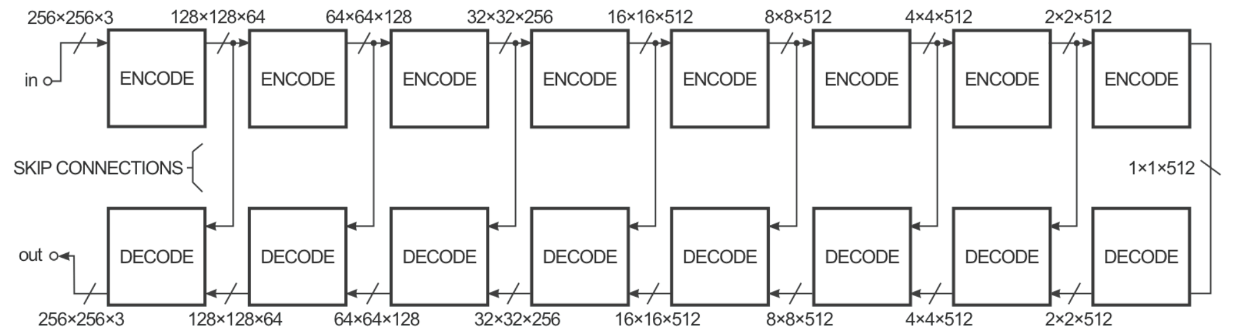
- We created our own dataset of 5000 images from flips and crops of 2500 images. We further split this into 3500 training images, 1000 validation images and 500 test images. These images were obtained through Google image search by modifying a crawler script to download the images returned.

- To obtain the sketch, the input image is first converted to greyscale, and then converted to double precision data. We use image gradient to locate breaks in uniform regions. If the gradient is not zero, then this pixel is located on the edge. Fuzzy inference system is defined and evaluated for edge detection. Finally, a image that resembles a pencil sketch is generated based on the result.

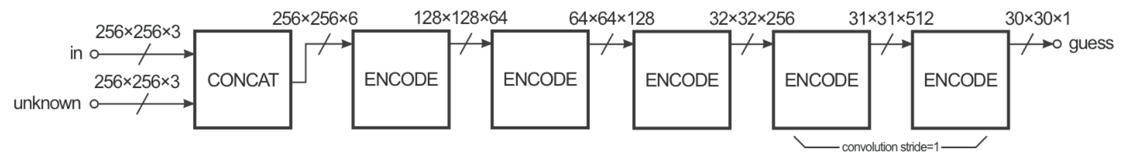
Methodology

- Architecture

- Generator



- Discriminator



We apply the conditional GAN model proposed by Isola et al. trained on our collected dataset of landscape sketches.

$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x,y\sim p(x,y)}[\log D(x,y)] + \mathbb{E}_{x\sim p(x),z\sim p(z)}[\log(1-D(x,G(x,z)))]$$

The model adds an additional loss to force the output and target images to be similar. We compare the performance of our model using L1-loss, L2-loss and Huber loss.

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda\mathcal{L}_L(G)$$

- Evaluation

One possible evaluation method is structural similarity (SSIM) [2], which measures the perceptual correctness by comparing contrast, luminance and structure of two images. In order to quantitatively evaluate the model, SSIM of the original images and the colourized images are computed and compared.

$$S(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

- Experiment

We experimented with training the network on various subsets of the complete dataset:

- Line sketch input, coloured target on the complete dataset
- Line sketch input, greyscale target + greyscale input, coloured target on the complete dataset
- Line sketch input, coloured target on each subcategory of the dataset.

Reference

[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. arXiv preprint arXiv:1611.07004, 2016.

[2] Z. Wang and E. P. Simoncelli. Translation insensitive image similarity in complex wavelet domain. In Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on, volume 2, pages ii–573. IEEE, 2005.

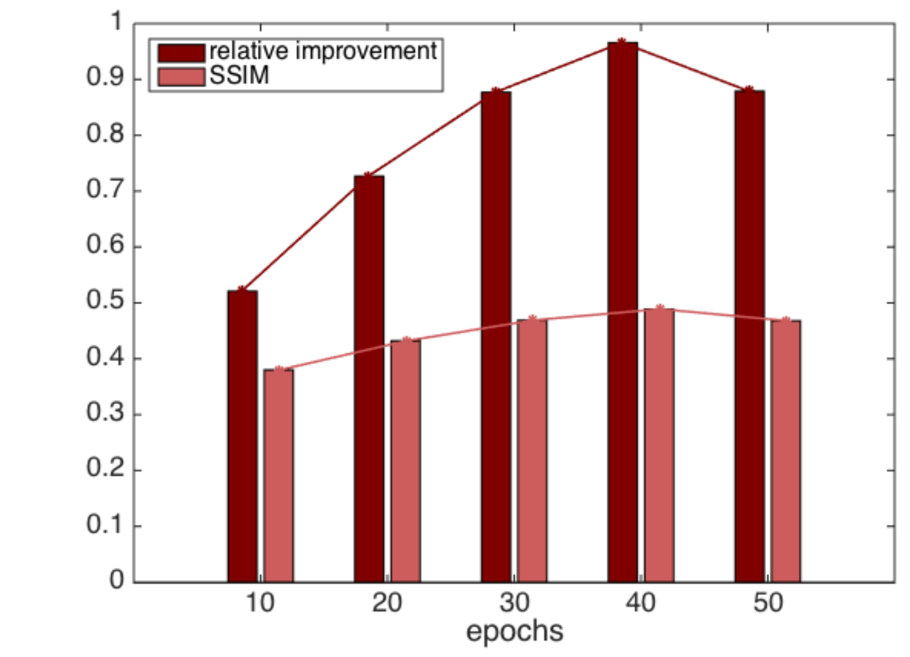
[3] C. Hesse. Image-to-image translation in tensor- flow. https://github.com/affinelayer/pix2pix-tensorflow, 2017.

Results

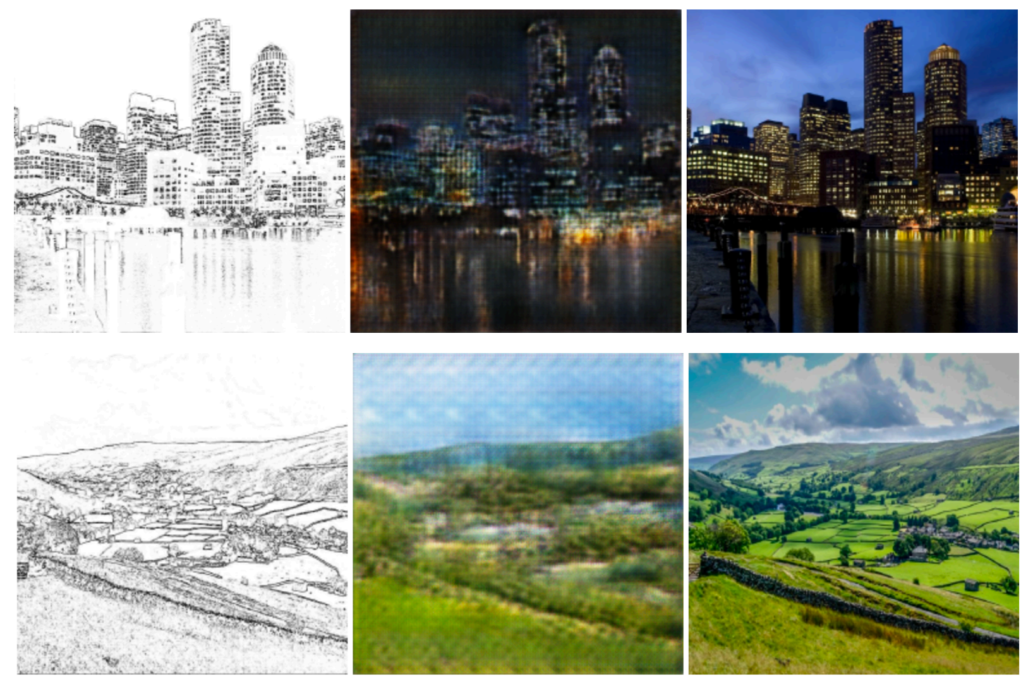
- Quantitative results

Loss Type	Relative improvement
L1-loss, 10	0.056
L1-loss, 100	**0.471**
L1-loss, 1000	0.294
L2-loss, 10	0.080
L2-loss, 100	0.450
L2-loss, 1000	0.236
Huber, 0.25	0.235
Huber, 0.5	0.252
Huber, 1	0.154

Number of epochs	SSIM (base: 0.249)	Relative improvement
10 epochs	0.380	0.521
20 epochs	0.432	0.727
30 epochs	0.467	0.877
40 epochs	0.489	0.966
50 epochs	0.468	0.879



- Qualitative results



- Analysis

Comparison of different losses shows that L1 loss with weight of 100 gives the greatest SSIM relative improvement. Low L1 weight caused the model to generated incorrectly coloured images while high weight caused blurring. Images generated using L2 loss were significantly blurrier while Huber loss led to patchy images.

Our model generates the correct colourings for most test images. As for human perception, the outputs are quite reasonable.

Conclusion

- The limited dataset and similarity of sketches made it difficult for the generator to learn colourings from the complete dataset, but the model performed significantly better when trained on data from the same landscape category. The SSIM is lower compared to state-of-the-art but still improves relative to the input. The model is able to learn some interesting features, such as inferring the presence of clouds not in the original linesketch and removing watermarks.

- The current model attempts to generate photo-realistic images from sketches but practical applications usually require artistic images. A possible extension is to use the generated image as a colouring template and apply style transfer to obtain a more artistic image.