Modeling Multimodal Distribution with Improved BicycleGAN
CIS 6800 Final Project

Yukai Yang, Zhuolun Zhao
School of Engineering and Applied Science,
University of Pennsylvania
{yukaiy, alanzhao}@seas.upenn.edu

Abstract

Image-to-image translation aims to transform images from one domain to have the characteristics of another domain while preserving the content representations. Generative adversarial networks (GAN) [3] have made tremendous progress in recent years to enable photo-realistic image-to-image translation, which has applications in synthesis, restoration, and style transfer. In this project, we want to explore multimodal conditional synthesis based on BicycleGAN [15]. While most GAN-based approaches suffer from mode collapse in conditional synthesis, BicycleGAN proposes a hybrid model that encourages invertible mapping between the output and the latent code, which should improve generation diversity while maintaining realism. We aim to further improve BicycleGAN's performance through architectures and loss function modifications.

1. Introduction

Generative adversarial network (GAN) has shown great potential to capture complex distributions and generate high-dimensional samples since its first introduction in 2014. In particular, deep learning techniques have made rapid progress in conditional image generation. Pix2pix [4] and Cycle-GAN [14] translated images across two characteristic domains to change their visual styles under paired and unpaired settings, respectively.

Real-world scenarios expect the synthetic samples to be diverse and able to manipulate flexibly. However, all the applications mentioned above suffer from the problem of mode collapse to some extent, even though randomly sampled latent codes are added as additional inputs. To deal with this shortcoming, many works attempted to enhance the correlation between input latent codes and output images to ensure that the latent codes have control over the generated images. BicycleGAN adopted a latent regression loss term, which encourages the model to recover the input latent code from the generated images. However, the results were far from ideal and there have been multiple related works contributing to tackling the mode collapse problem, notably MS-GAN [9] and DivCo [8]. We aim to improve the performance of BicycleGAN such that the synthesized images are more (1) realistic and (2) diverse, all while remaining faithful to the input.

2. Related Works

In image synthesis tasks, generative adversarial networks [3] have been largely successful in both modeling natural image statistics and efficient training by mapping random values from an easy-to-sample distribution (e.g. a low-dimensional Gaussian) to output images in a single feedforward pass of a network. During training, a discriminator network distinguishes between samples from the target distribution and the generator network. BicycleGAN builds on the conditional version of VAE [7] and InfoGAN [1] or latent regressor models [2] by jointly optimizing their objectives.

2.1. Conditional image generation

Image-to-image conditional GANs have made significant progress in the quality of the results compared to conditional VAE [12] and autoregressive models [11]. However, the generator often learns to largely

ignore the random noise vector when conditioned on a relevant context, which leads to a loss of multimodality in the generated results.

2.2. Multimodality

In other related works, multimodality has been explicitly encoded as an additional input to the input image. Color and shape scribbles and other interfaces were used as conditioning in iGAN [13] and pix2pix. Others have experimented with using a mixture of models. These methods are unable to produce continuous changes. Before BicycleGAN, there usually is a trade-off between conditionality and multimodality. Post-BicycleGAN works such as MSGAN [9] explicitly maximizes the ratio of the distance between generated images with respect to the corresponding latent codes, thus encouraging the generators to explore more minor modes during training.

2.3. Bijective mapping

Although similar in names, CycleGAN and BicycleGAN tackle two different problems. CycleGAN was proposed to tackle the unpaired image translation problem by training two cross-domain transfer GANs at the same time, utilizing cycle consistency loss. BicycleGAN still uses paired images as input but models one-to-many mappings. CycleGAN encourages a bijective mapping between input and output spaces, whereas BicycleGAN encourages a bijective mapping between the latent and output spaces.

3. Methods

The training of BicycleGAN can be divided into two modules, cVAE-GAN and cLR-GAN. cVAE-GAN encodes the ground truth target B to a latent encoding z. A generator then tries to reconstruct B using the z concatenated with the paired input image A. cLR-GAN samples a latent encoding from a prior Gaussian distribution $\mathcal{N}(0, 1)$ and generates output $\hat{B}$. An encoder then attempts to reconstruct the latent encoding using $\hat{B}$. The overall objective is formulated as the following:

$$G^*, E^* = \arg\min_{G,E} \max_{D} \mathcal{L}_{\text{GAN}}^{\text{VAE}}(G, D, E) + \lambda \mathcal{L}_1^{\text{VAE}}(G, E) + \mathcal{L}_{\text{GAN}}(G, D) + \lambda_{\text{latent}} \mathcal{L}_1^{\text{latent}}(G, E) + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}}(E)$$

$\mathcal{L}_{\text{GAN}}^{\text{VAE}}(G, D, E)$ is the adversarial loss that encourages the generated image $\hat{B}$ in cVAE-GAN to be realistic. $\mathcal{L}_1^{\text{VAE}}(G, E)$ encourages the reconstructed image $\hat{B}$ to be similar to B. This also results in a bijective mapping that alleviates mode collapse. The KL divergence loss enforces the latent space to follow Gaussian distribution. For cLR-GAN, $\mathcal{L}_1^{\text{latent}}(G, E)$ is introduced to make the latent encodings similar. Again, an adversarial loss $\mathcal{L}_{\text{GAN}}(G, D)$ is used to encourage realistic generation.

Network architectures For the generator G, we experiment with two networks: the ResNet generator from CycleGAN [14] and U-Net. The ResNet generator consists of 2 downsampling blocks, 6 residual blocks, and 2 upsampling blocks, each with Convolution-InstanceNorm-ReLU layers. We use a U-Net with 7 downsampling blocks in the encoder and 7 upsampling blocks in the decoder followed by Tanh, where each downsampling block has Convolution-InstanceNorm-LeakyReLU layers and each upsampling block has TransposedConvolution-InstanceNorm-ReLU layers. Downsampled and upsampled features maps with the same spatial dimensions make pairs and are concatenated to form skip connections in the decoder. We also study the choice of normalization: InstanceNorm vs BatchNorm in U-Net.

For the discriminator D, we adopt PatchGAN [5], which classifies whether local image patches are real or fake. The PatchGAN follows the C64-C128-C256-C512 architecture from CycleGAN [14]. We experiment with using one discriminator and separate discriminators for cVAE-GAN and cLR-GAN during optimization. In addition, we use PatchGAN at different scales. We start with 70×70 receptive field and add a second PatchGAN that doubles the receptive field size by average pooling the input image.

The encoder E in cVAE-GAN uses a standard ResNet18 followed by 2 linear layers to encode mean and log variance, respectively.

4. Hypothesis

It is important for the synthesized images to be both photo-realistic and diverse. cVAE-GAN alone should have more variation since the latent space encodes more information about the ground truth outputs B.

However, KL divergence may not be sufficient to populate the latent space. cLR-GAN alone encourages the generated $\hat{B}$ to densely populate the latent space once encoded but suffers from mode collapse and fewer variations. Combining both approaches should give us the best of both worlds and obtain both realistic and diverse outputs. Since it is notoriously hard to balance the generator and discriminator during GAN training, it would be beneficial to study whether the tricks mentioned above can stabilize GAN training.

To improve diversity for multi-modal generation, BicycleGAN can be augmented with auxiliary loss functions such as mode seeking loss [9] and contrastive loss [8]. In particular, after sampling positive and negative latent vectors in cLR-GAN, we can feed these vectors into the generator and encoder and encourage the network to associate encoded positive samples and pull away from negative samples via a contrastive loss. However, due to a limited training budget, we do not have time to explore these options.

5. Experiments

Data We train and test our methods on the Edges2shoes dataset, which is a paired image-to-image translation task. The Edges2shoes dataset is a subset of the pix2pix dataset that contains 50,000 images from the UT Zappos50K[1]. We split the dataset into 49,800 train images and 200 test images and scale them to 128×128.

Metrics We use FID score [2] and Learned Perceptual Image Patch Similarity (LPIPS) [3] as our evaluation metrics. We compute Fréchet Inception Distance (FID) between generated images and real images in the test set to quantify the photo-realistic quality. To better match human judgments, LPIPS computes distance in AlexNet feature space pre-trained on Imagenet with linear weights. We compute the average pairwise LPIPS distance between 10 samples generated from 10 random latent vectors for each test image and then average across the entire test set.

Training details The training details are generally similar to that of BicycleGAN. The model is built on the Least Squares GANs (LSGANs) variant [10], which uses a mean square error loss (MSE) instead of a cross entropy loss. LSGANs generally produce high-quality results with stable training. Similar to BicycleGAN, we also did not condition the discriminator D on input A. We set the parameters $\lambda_{image} = 10$, $\lambda_{latent} = 0.5$, and $\lambda_{KL} = 0.01$ in all our experiments. We choose latent dimension $|z| = 8$. We only update G for the l_1 loss $L_1^{latent}(G, E)$ on the latent code, while keeping E fixed. Optimizing G and E simultaneously for the loss would encourage G and E to hide the information of the latent code without learning meaningful modes. We train our networks from scratch using the Adam [6] optimizer with a batch size of 8 and with a learning rate of 0.0002 for 20 epochs. Only the encoder E is initialized from ImageNet pre-trained weights. Training 20 epochs on Colab takes 5 hours.

5.1. Qualitative Evaluation

We selected multiple domain A images from the test dataset and 4 random latent vectors to generate domain B images. As shown in Figure 1, we found that our trained generator is able to produce diverse styles of domain B images with different latent vector inputs while maintaining the realism of the images. We also noticed that mode collapse still exists for some images.

5.2. Quantitative Evaluation

The baseline method in Table 1 uses the ResNet Generator and a single PatchGAN discriminator. The baseline achieves the highest realism (lowest FID). Separating the discriminators improve the LPIPS score by 29% compared to sharing weights. However, discriminators at multiple scales degrade both FID and LPIPS scores. We hypothesize that the generator is not able to catch up with the discriminator as the discriminator becomes more robust, which introduces instability in our setting.

Switching the generator to U-Net results in the highest diversity score, improving LPIPS by 52.1%. However, FID is worse. Further adding InstanceNorm does not improve the results. We also observed relatively more unstable generator and discriminator losses for all U-Net experiments.

[1] https://vision.cs.utexas.edu/projects/finegrained/utzap50k/

[2] https://github.com/mseitzer/pytorch-fid

[3] https://github.com/richzhang/PerceptualSimilarity.

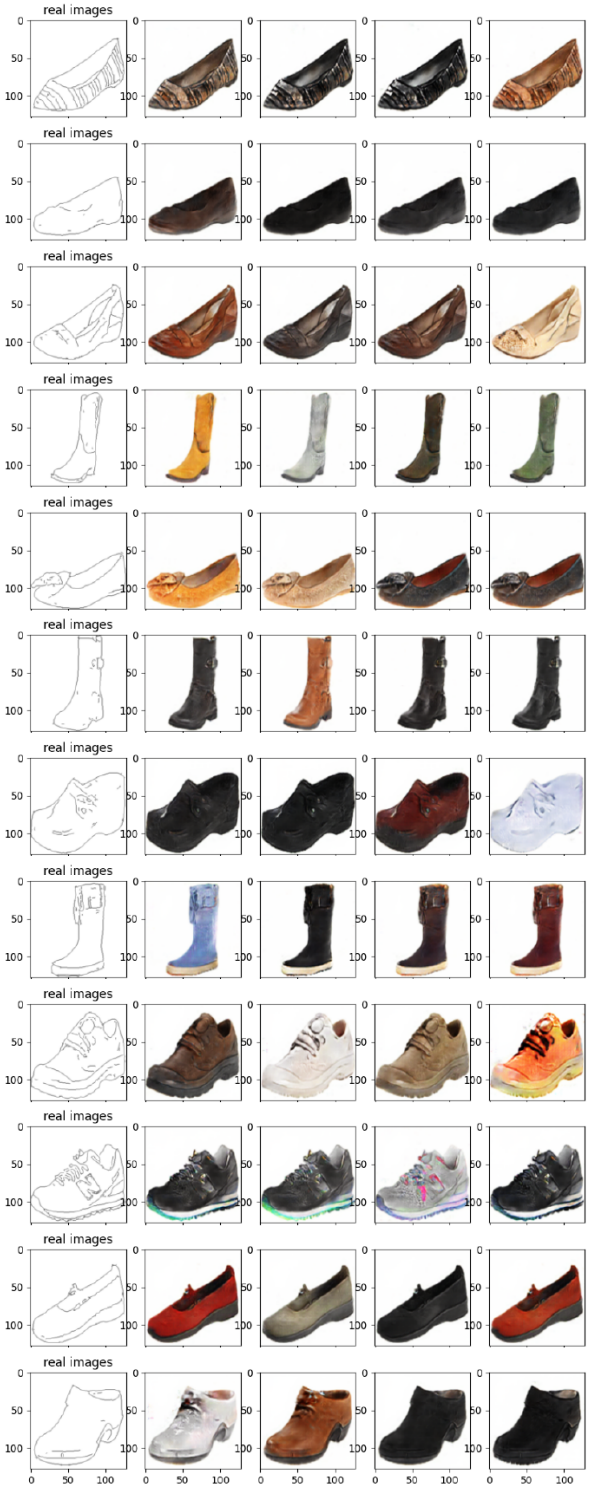

Figure 1. Inference results of multiple random latent vectors with U-Net-In generator with separate discriminators

Model	FID ↓	LPIPS ↑
baseline	**78.84**	0.110
sep-disc	79.44	0.142
sep-multi-disc	95.76	0.133
sep-disc + U-Net-gen	95.34	**0.216**
sep-disc + U-Net-In-gen	84.36	0.123

Table 1. Comparison of generation realism and diversity across different model choices.

6. Conclusions

In conclusion, we have evaluated a few methods for combating the problem of mode collapse in the conditional image generation setting. We find that by combining multiple objectives for encouraging a bijective mapping between the latent and output spaces, we obtain results that are more realistic and diverse. We observe a trade-off to some extent between realism and diversity with increased generator and discriminator complexity. However, due to the unstable nature of GAN training, it is relatively hard to achieve consistency. Future work includes experimenting with additional contrastive loss functions and directly enforcing a distribution in the latent space that encodes semantically meaningful attributes to allow for image-to-image transformations with user-controllable parameters.

References

[1] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets, 2016.

[2] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning, 2016.

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2016.

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2016.

[6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

[8] Rui Liu, Yixiao Ge, Ching Lam Choi, Xiaogang Wang, and Hongsheng Li. Divco: Diverse conditional image synthesis via contrastive generative adversarial network, 2021.

[9] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis, 2019.

[10] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks, 2016.

[11] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks, 2016.

[12] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models, 2015.

[13] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold, 2016.

[14] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2017.

[15] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation, 2017.