StarGAN

Before StarGAN proposed, many successful studies show the image-to-image translation for two domains. For example, the Pix2Pix model solves the image translation problem with Paired data; CycleGAN solves the image translation problem under Unpaired data. But both of them solve the one-to-one problem, that is, the conversion from one field to another. The weakness is the limited scalability and robustness when handling more than two domains. The existing GAN model needs to construct k∗(k−1) generators, and cannot be trained across datasets.

Basic principle

StarGAN is proposed to solve the training across multiple domains and multiple data sets. In StarGAN, the traditional fixed translation is quit, but the domain c information and pictures are input together for training, and the mask vector is added to the domain label, which is convenient for different training sets to do joint training.

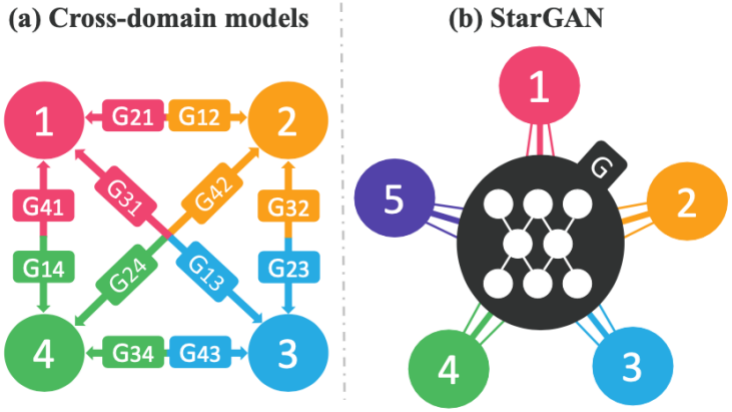


Figure 1: Comparison between cross-domain models and our pro- posed model, StarGAN. (a) To handle multiple domains, cross-domain models should be built for every pair of image domains. (b) StarGAN is capable of learning mappings among multiple do- mains using a single generator. The figure represents a star topology connecting multi-domains.

As the figure 1 shows, there are k(k – 1) generators need to be trained in order to learn all mappings among k domains. It is an ineffective model and failure utilizing training data. As a solution, StarGAN model takes multiple domains within a single generator with a label to represent domain information.

Contribution

This paper written by Choi et al (2018) can be concluded into three contributions including the novel GAN network, the using of mask vector method and superior qualitative and quantitative results on facial attribute transfer and facial expression synthesis tasks.

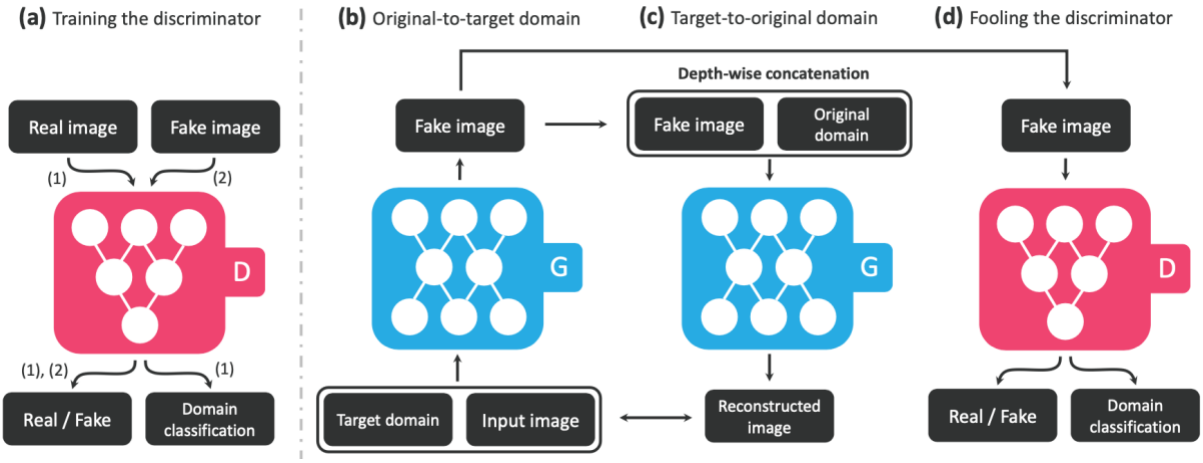


Figure 2: Overview of StarGAN, consisting of two modules, a discriminator D and a generator G. (a) D learns to distinguish between real and fake images and classify the real images to its corresponding domain. (b) G takes in as input both the image and target domain label and generates an fake image. The target domain label is spatially replicated and concatenated with the input image. (c) G tries to reconstruct the original image from the fake image given the original domain label. (d) G tries to generate images indistinguishable from real images and classifiable as target domain by D.

Framework

Choi et al (2018) proposed a novel framework to train a single generator to map multiple domain image-to-image translation. As shown in figure 2, there are two modules in total. The discriminator D is trained by inputting real and fake image to classify them into corresponding domain. The Generator G is trained by adding the target domain to construct image fooling the discriminator. To achieve ideal generator and discriminator, this StarGAN paper introduced three loss function.

Adversarial Loss.

$$\mathcal{L}_{adv} = \mathbb{E}_x\left[\log D_{src}(x)\right] + \mathbb{E}_{x,c}[\log\left(1 - D_{src}(G(x,c))\right)],$$

The first is the general loss function of the GAN network to judge the output image. As the paper describe (Choi et al., 3), "G generates an image G(x, c) conditioned on both the input image x and the target domain label c, while D tries to distinguish between real and fake images." $D_{src}(x)$ refers to a probability distribution.

Domain Classification Loss.

$$\mathcal{L}_{cls}^{r} = \mathbb{E}_{x,c'}[-\log D_{cls}(c'|x)],$$

$$\mathcal{L}_{cls}^{f} = \mathbb{E}_{x,c}[-\log D_{cls}(c|G(x,c))].$$

The second loss function is divided into two parts. When training the discriminant network D, use the supervision signal of the real picture in the original field for training; and when training the generation network G, use the supervision of the generated picture in the target field.

Reconstruction Loss.

$$\mathcal{L}_{rec} = \mathbb{E}_{x,c,c'}[||x - G(G(x,c),c')||_1],$$

It is designed to preserve the content of the input images while style transfer because minimizing the adversarial loss and domain classification loss cannot guarantee the content integrity.

Full Objective.

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{cls}\,\mathcal{L}^r_{cls},$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{cls}\,\mathcal{L}^f_{cls} + \lambda_{rec}\,\mathcal{L}_{rec},$$

Finally, the formula combining all the loss function as full objective is shown above where λ_{cls} and λ_{rec} are hyper-parameters affecting the relative importance of domain classification and reconstruction losses.

Pros & cons

One of the most important advantage of StarGAN is that it contains multiple labels within different datasets and easily control all the labels. However, there is a latten issue as Choi et al., (2018) said, “the label information is only partially known to each dataset”. To solve the problem, the paper introduced “mask vector” which allows StarGAN to ignore unspecified labels, focus on the known label and define them into a vector.

MUNIT

Introduction

In computer vision, unsupervised image-to-image conversion is an important and challenging problem because the model tries to learn the condition distribution of target domain without any paired domain of source and target. Nowadays, many popular computer vision models are designed to solve those transfer problems from one domain to another, such as super-resolution[1], colorization[2], inpainting[3], attribute transfer[4] and style transfer[5]. However, the fact is many cross-domain mapping of interest is multimodal. The question is whether unsupervised image-to-image conversion can be applied into a multi-modal domain. To address this limitation, Huang et al proposed Multimodal Unsupervised Image-to-image Translation (MUNIT) framework. Their primary assumptions are that an image can be decomposed into a content space and a style space, where the style spcae can capture the style information and different domain share the same content space. For the conversion of a source domain image to the target domain, the only thing needs to do is recombining the content code and style code.

They analyze this framework and develop several theoretical results. Extensive experiments, as well as comparisons with state-of-the-art models, demonstrate the strengths of the framework while maintaining style control.

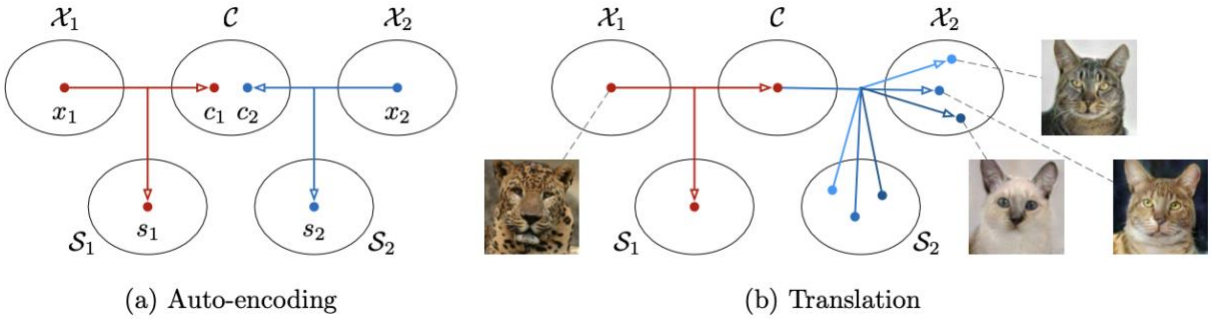
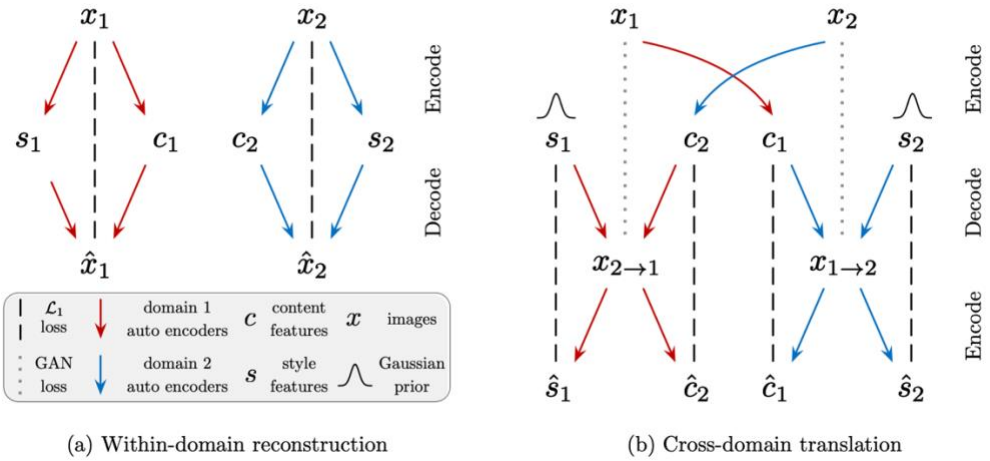


Figure 3: An illustration of our method. (a) Images in each domain Xi are encoded to a shared content space C and a domain-specific style space Si. Each encoder has an inverse decoder omitted from this figure. (b) To translate an image in X1 (e.g., a leopard) to X2 (e.g., domestic cats), we recombine the content code of the input with a random style code in the target style space. Different style codes lead to different outputs.

The figure 3 shows an abstract encoding part of MUNIT. X_1 refers to all the input images, which are encoded into a domain-shared content space C and a domain unique style space S. During the image conversion process, the content code of the input image should remain unchanged, and the style code represents the main change information of the image. By combining with different style codes, the model has the ability to generate a variety of images. The style code can be given randomly, or it can be extracted from a picture in the target domain.

Methodology



(a) Within-domain reconstruction (b) Cross-domain translation

Figure 4: Model overview. Our image-to-image translation model consists of two auto- encoders (denoted by red and blue arrows respectively), one for each domain. The latent code of each auto-encoder is composed of a content code c and a style code s. We train the model with adversarial objectives (dotted lines) that ensure the translated images to be indistinguishable from real images in the target domain, as well as bidirectional reconstruction objectives (dashed lines) that reconstruct both images and latent codes.

The figure 4 is the mode process overview in which each domain X_i contains both Encoder and Decoder. Each latent code extracted will be decomposed into content code c_i and style code s_i. In the process of image conversion, the encoder and decoder are implemented in pairs. For a sample $x_1 \in X_1$ transferred to domain x2, the content latent conde c_1 and style code s2 selected from Gaussian prior distribution are extracted. After model training, multimodal images are eventually generated.

To train the MUNIT model, Huang et al proposed two loss functions, bidirectional reconstruction loss and adversarial loss. The bidirectional reconstruction loss was designed to confirm the encoders and decoders are inverse, which means both image to latent to image and latent to image to latent directions can be constructed. The formula is shown as follow:

$$\mathcal{L}_{\text{recon}}^{c_1} = \mathbb{E}_{c_1 \sim p(c_1), s_2 \sim q(s_2)}[||E_2^c(G_2(c_1, s_2)) - c_1||_1]$$
$$\mathcal{L}_{\text{recon}}^{s_2} = \mathbb{E}_{c_1 \sim p(c_1), s_2 \sim q(s_2)}[||E_2^s(G_2(c_1, s_2)) - s_2||_1]$$

where q(s2) is the prior N (0, I), p(c1) is given by c1 = E1c(x1) and x1 $\sim$ p(x1).

Adversarial loss was used like GAN to match the Gaussian distribution between translated images and target data. The formula is shown as follow:

$$\mathcal{L}_{\text{GAN}}^{x_2} = \mathbb{E}_{c_1 \sim p(c_1), s_2 \sim q(s_2)}[\log(1 - D_2(G_2(c_1, s_2)))] + \mathbb{E}_{x_2 \sim p(x_2)}[\log D_2(x_2)]$$

where D2 is a discriminator that tries to distinguish between translated images and real images in X2. The discriminator D1 and loss Lx1 are defined similarly.

Future development and Summary

In this review, we mainly discussed the GAN basic mode, GAN-based style migration mode and relevant realistic applications. The purpose is to summarize the development tendency of image style transfer based on GAN method and predict future method immigration.

~~Future research on neural style transfer, promising directions mainly focus on two aspects. One is to solve the problems faced by the current algorithm mentioned above, namely, the parameter fine-tuning problem, the stroke direction control problem, and the "fast" and "faster" problems in neural style transfer. A description of these challenges and their corresponding possible solutions is found in Section 7. A second promising direction focuses on new extensions of neural style transfer (e.g., fashion style transfer and character style transfer), and there have been some preliminary results in this direction, such as the recent work by Yang et al. [47] Research on Text Effects Transfer. These interesting expansions may turn into trends in future research topics, which in turn may create new related fields.~~