Contrastive Representation Learning For Image Attribute Manipulation

Wonyl Choi, Duc Minh Nguyen, and Quan Pham
Boston University
{wonyl,nguymi1,phamquan}@bu.edu

Abstract

The recent advancement in image manipulation is implemented with the help of to generative adversarial network. Currently, we can mention a few popular implementation, AttGAN [3], and StarGAN [2]. While these models perform well on the current dataset, such as CelebA, CelebA-HQ dataset, we want to further enhance the accuracy, as well as the realistic systhesis images produced of such state-of-the-art models. In this paper, we proposed a GAN based solution integrated with contrastive learning to understand the embedding space of the attributed manipulated, and with that, to improve such existing models.

1. Introduction

An Image may contain a vast number and variety of attributes. An agent with the capability of processing visual information can categorize attributes, classify images with respect to the category, and even can imagine a scenery having a certain collection of visual attributes.

Image attribute manipulation is the task of translating an input image to a new realistic image that has a certain subset of attributes modified as desired, while the rest remains as the original. In general, it is a multi-domain task, which requires learning the translation function for numerous attributes.

One of the challenges of the image attribute manipulation task rises from the fact that different parts of the image can contribute to different attributes. Another challenge is that the quality of synthesized visual content is heavily reliant on the relations among its components, so an attempt of manipulating a single attribute may be generating unrealistic images. For example, baldness or moustache may be a local facial attribute, especially for a classification task, while it exhibits different patterns in each gender, and is more expressive among males. Such feature dependency essentially calls to learn different translations on the embedding space for each attribute combination which potentially has exponentially many possibilities.

Despite the difficulties of the task, there has been great progress from the generative adversarial network (GAN) based models, due to its efficacy in generating realistic images. The generator module of many GAN-based attribute manipulation models contains the encoder-decoder structure. Here, the sample image and target attribute labels are combined to produce the latent representation vector (as in AttGAN [3] or StarGAN [2]) which is then fed into the decoder for generating the output image.

Taking the target image representation as a vector in the latent embedding space sampled with the condition of desired attributes seems to promise better interpretability and finer controllability.

In this project, we extend the previous work by studying the attribute representation learning method and its impact on the image attribute modification task. More specifically, we suggest an idea of how to incorporate contrastive learning into the GAN-based face image attribute manipulation. The models using three diffrent variants of GAN, AttGA [3], StarGAN [2], and StyleGAN2 [5] have trained using the CelebA [7] dataset. The organization of the report is as follows. First, we review the GAN-based image attribute manipulation focusing on the common structure that is shared by many pre-existing models. After introducing infoNCE loss, we suggest a GAN-based attribute manipulation model and training strategy. Experiment results and the limitations and issues arise upon the application of contrastive learning will be discussed as well.

2. Related works

2.1. GAN-based image attribute manipulation

In image processing, vision and graphics communities, GAN-based methods have shown high performance in generating realistic images, rapidly gaining popularity and much attention [10] [4]. Among many applications of the GAN, image attribute manipulation has been recognized as one of the first important examples. In general, GAN architecture consists of generator and discriminator modules. When the distribution of the sample data is unknown, the discriminator is trained to distinguish the true data points from the generated ones, while the generator, mapping in-

put variable to the sample domain, is trained to against the discriminative prediction.

Formally, considering a simple image generation task, the generator G maps the latent space $\mathcal{W}$ to the image space $\mathcal{X}$, while the discriminator D maps the sample images $x \in \mathcal{X}$ to their probability values $D(x) \in [0, 1]$ distinguishing the real and generated images. Then the parameters of the generator G are optimized to ensure the realistic image synthesis based on the following loss function

$$\mathcal{L}_{\text{adv}}^{(G)} = \mathbb{E}_{w \sim p_{\mathcal{W}}}[\log(1 - D(G(w)))]; \tag{1}$$

while the discriminator aims to distinguish the synthesized images from the real samples, and hence subjected to the following maximization problem:

$$\begin{aligned} \mathcal{L}_{\text{adv}} =& \mathbb{E}_{x \sim p_{\mathcal{X}}}[\log D(x)] \\ &+ \mathbb{E}_{w \sim p_{\mathcal{W}}}[\log(1 - D(G(w)))]. \end{aligned} \tag{2}$$

Note that the loss function for the discriminator can be expressed as $\mathcal{L}_{\text{adv}}^{(D)} = -\mathcal{L}_{\text{adv}}$ and the minimization problem for the generator, $\mathcal{L}_{\text{adv}}^{(G)}$, is equivalent to the minimization of $\mathcal{L}_{\text{adv}}$.

When GAN is applied to the image attribute manipulation task, it is desirable to have the generator encoding the conditional representation of the sample image with respect to the original attribute properties. This motivates introducing the encoder-decoder structure to the generator. Since the image attribute manipulation aims to modify a part of the input information to be aligned with the target attribute labels, Attr_T, the generator can be trained to diminish the effect of the original attribute properties contained in the sample image x by aligning the predicted labels $\text{Attr}_P^{(m)}$ of the generated image $x_m := G(x, \text{Attr}_T)$ to the target labels. Here, the classifier C is trained by comparing the original labels Attr_S to the predicted labels $\text{Attr}_P^{(r)}$ of the reconstructed image of the sample, x_r. The visual plausibility of the generated image can be encouraged further by employing the reconstruction loss that compares the sample image x with x_r.

The classifier may be trained together with the discriminator and so from now on we consider C as a submodule of the discriminator, together with D, the function distinguishing the real and synthesized images. The classification loss $\mathcal{L}_{\text{clsf}}^{(G)}$ to train the generator can be expressed as

$$\begin{aligned} \mathcal{L}_{\text{clsf}}^{(G)} =& \text{cross-entropy}(\text{Attr}_T; \text{Attr}_P^{(m)}) \\ =& \text{cross-entropy}(\text{Attr}_T; C(G(x, \text{Attr}_m))); \end{aligned} \tag{3}$$

and the reconstruction loss is given as

$$\mathcal{L}_{\text{rec}} = d(x, x_r) = d(x, G(x, \text{Attr}_S)); \tag{4}$$

where d is an appropriate similarity function. The typical choice of d includes the pixel-wise ℓ_2 norm.

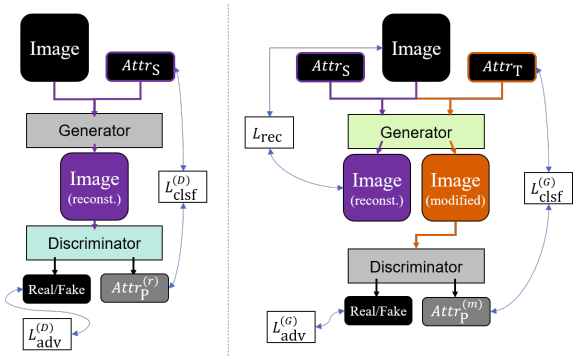


Figure 1. Schematic Diagram of General GAN-based Attribute Manipulation Model

The loss that the classifier C is subjected to takes the form of the cross-entropy, similar to the classification loss,

$$\begin{aligned} \mathcal{L}_{\text{clsf}}^{(D)} =& \text{cross-entropy}(\text{Attr}_S; \text{Attr}_P^{(r)}) \\ =& \text{cross-entropy}(\text{Attr}_S; C(G(x, \text{Attr}_S))). \end{aligned} \tag{5}$$

Altogether, the generator and discriminator losses can be expressed as combinations of the defined loss functions as follows:

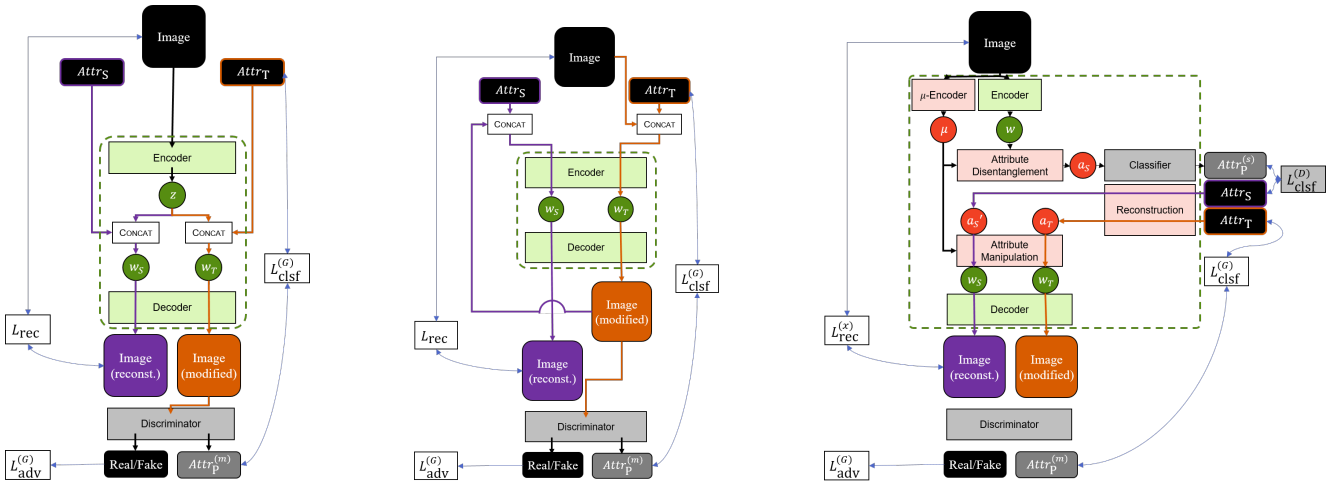
$$\mathcal{L}_G = \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{clsf}}^{(G)} + \mathcal{L}_{\text{rec}}; \quad \mathcal{L}_D = -\mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{clsf}}^{(D)} \tag{6}$$

The general GAN-based attribute manipulation method discussed here is illustrated in Fig. 1.

In the remaining part of the section, we discuss some details of AttGAN [2] and StarGAN [3], the models on which we have focused for the project.

AttGAN. In AttGAN, the attribute information is combined with $z = E(x)$, the latent representation of the sample image x. In particular, the final conditional representation of the sample image, w, with respect to the given attribute can be expressed as $w = (z||a)$, where a is the vector representation of the attribute labels, which is taken to be the one-hot encoding of the labels and $||$ denotes the concatenation operation. In order to overcome the instability of GAN model training, the authors have utilized the Wasserstein distance for the adversarial loss, following WGAN [1]. The AttGAN model uses U-Net [8] for the base structure for both encoder E and decoder F of the generator. The schematic diagram for the model structure and encoder training pipeline is illustrated in Fig. 2a.

StarGAN. In contrast to AttGAN, StarGAN takes a view of image translation to the attribute manipulation, similar to CycleGAN [11]. As illustrated in Fig. 2b, the generator ingests the sample image concatenated with the target attribute labels depth(channel)-wise and produces the modified image. The modified image is then used as a new source



(a) Architecture of AttGAN. (b) Architecture of StarGAN. (c) Architecture design of the suggested model.

Figure 2. Schematic Diagram of GAN-based Attribute Manipulation Models

image and fed to the generator with the original attribute labels, to reconstruct the sample image.

2.2. InfoNCE loss

Contrastive learning is a general framework of learning representation space of the given data. The main idea of contrastive learning is that the metric or the distance measure must correspond to the similarity of data points.

InfoNCE [9] is a type of contrastive objective which aims to identify the positive sample that agrees with the given context from a set of random samples. Formally, considering a set of samples $X = \{x_i\}_{i=1}^{N}$ and the positive sample x_0, infoNCE loss with respect to the given condition c can be expressed as

$$\mathcal{L}_{\text{contra}} = -\mathbb{E}_{x_0,c}\left[\log\left(\frac{f(x_0,c)}{f(x_0,c)+\sum_{i=1}^{N-1} f(x_i,c)}\right)\right]; \tag{7}$$

where the score function $f(x,c)$ estimates the density ratio $\frac{\Pr(x|c)}{\Pr(x)}$. When f is selected to be the exponential function, infoNCE loss is equivalent to the log-softmax function. It is worth mentioning that the connection between infoNCE loss and mutual information $I(x;c)$ is well-established.

3. Our Method

Here we describe the idea of incorporating the contrastive learning approach to GAN-based attribute manipulation. The full design of the model suggested is illustrated in Fig. 2c. Then we describe the partial modification made to each variety of GAN used for the experiments.

In general, unlike the variational method or the flow-based techniques, GAN learns the distribution underlying the data implicitly. While GAN is advantageous to learn complex distribution functions, it is hard to control and navigate through the latent space of the conditional representation. To overcome the issue, we suggest learning the identity representation μ with an additional encoder and disentangling the attribute-specific information to the attribute embedding a, obtained by the disentanglement function g; i.e. $a(w) := g(w;\mu)$. The attribute embedding a also needs to be reconstructed from the given attribute labels. Taking the inverse of the partial function of g, w can be retrieved from μ and a. Through our experiments, the operation g is taken to be element-wise subtraction. The attribute representation of a sample, a_S, can be used to train the attribute label classifier C contained in D. Ideally, if C is invertible, attribute vector a corresponding with the given attribute label Attr can be obtained easily. Since it is the usual case that the number of attributes is much smaller than the dimension of the latent space, the reconstruction function h mapping the attribute label to the latent representation must be trained. For example, the reconstruction loss $\mathcal{L}_{\text{rec}}^{(a)} = \|a_S - a'_S\|_2^2$ can be used, where $a_S = g(w;\mu)$ and $a'_S = h(\mathrm{Attr}_S)$ are the attribute representation obtained from the sample image and label, respectively.

Here, the independence of the attribute representation a and the identity representation μ can be ensured by utilizing InfoNCE loss. Consider a set of distinct attribute labels, $\{\mathrm{Attr}_i\}_{i=1}^{N}$ and a set of paired sample images $\{(x_i^{(1)}, x_i^{(2)})\}_i$ of the same attribute Attr_i but with different individuals. The attribute embeddings of a pair must be similar to each other, independent of the identity of individuals who appears in those images. Taking the first image of the pair i as the condition c and the collection of the second images of all the other pairs as the random samples, the additional infoNCE loss term to train the generator using

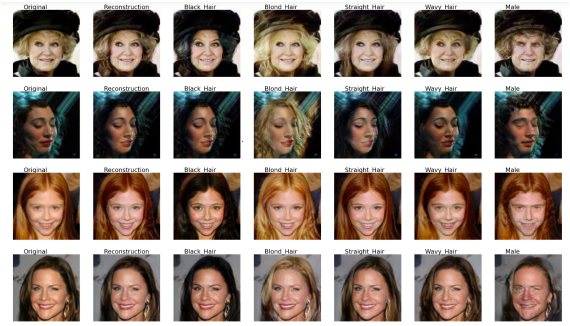


Figure 3. Image Manipulation Result Generated by StarGAN trained with infoNCE

Figure 4. Image Manipulation Result Generated by Vanila StarGAN

the set of positive pairs is given as in Eq. (7). More rigorously, letting f to be the exponential of the cosine similarity, $\mathcal{L}_{\text{contra}} = -\text{Tr}\log(\text{Softmax}M)$, where M is the similarity matrix and $\log$ is taken element-wise.

4. Experiments

4.1. Dataset

For the training and evaluation of the model, CelebFace Attributes (CelebA) [7] is used. CelebA is probably the most widely used large-scale facial image dataset in attribute manipulation studies. The dataset contains local, global, and abstract attributes of 40 kinds annotated on 202,599 face images. The dataset can be downloaded at the site [6]. The dataset is officially partitioned into training, validation and test set. Here we only used the training set for the training. Each image is centre-cropped, resized, and normalized before the training and testing.

4.2. Implementation Details

Each of pre-trained AttGAN[1], StarGAN[2], and StyleGAN2[3] are taken from repositories and fine-tuned onto the CelebA dataset.

In each model, μ-encoder is added, taking the architecture of the encoder structure of the corresponding model. Since the styleGAN2 does not have the encoder structure, a custom encoder using the pre-trained VGG16 model is constructed and trained.

For the optimization, we set the learning rates to 0.0001 and utilized beta values of 0.5 and 0.99 for the Adam optimizer.

4.3. Results

We computed two classification metrics as a benchmark for our proposed solution with the baseline version of the model.

Model	Accuracy Score	F1 Score
Original StarGAN	0.544	0.807
InfoNCE StarGAN	0.515	0.782
Original AttGAN	0.503	0.780
InfoNCE AttGAN	0.556	0.750

Table 1. Classification Metrics

Qualitative analyses. When dealing with hair color, the original StarGAN sometimes tends to over-colorize the images. An example of this is with the blond hair feature with a black background, the surrounding of the person gets filtered out by yellow color. When dealing with bright-colored hair, the base StarGAN seems to have problems with modifying it to be black. Both of these issues seem to be lessen after

5. Conclusion

In the project, we have studied the effect of contrastive learning on GAN-based image attribute manipulation. The unstable training of GAN models, in particular, the phenomenons such as the gradient explosion and the mode collapse are the hard obstacles. Besides, the property of contrastive learning requiring the large batch size or the large number of random samples makes the incorporation difficult.

[1] https://github.com/elvisyjlin/AttGAN-PyTorch
[2] https://github.com/yunjey/stargan
[3] https://github.com/NVlabs/stylegan2-ada-pytorch

6. Contributions

W.C. conceived of the project idea. Each team member selected one of the GAN-based models for encoding and image generation and explore various contrastive learning loss and strategies. W.C., D.N., and Q.P. studied StyleGAN2, StarGAN, and AttGAN respectively. The possible choices of the base model and contrastive learning losses are discussed above. All project members will discuss the results and contribute to the final manuscript.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017. 2

[2] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *CoRR*, abs/1711.09020, 2017. 1

[3] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Arbitrary facial attribute editing: Only change what you want. *CoRR*, abs/1711.10678, 2017. 1

[4] He Huang, Philip S. Yu, and Changhu Wang. An introduction to image synthesis with generative adversarial nets. *CoRR*, abs/1803.04469, 2018. 1

[5] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *CoRR*, abs/1912.04958, 2019. 1

[6] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Celeba dataset, 2015. https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html. 4

[7] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 1, 4

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 2

[9] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. 3

[10] Hui Ying, He Wang, Tianjia Shao, Yin Yang, and Kun Zhou. Unsupervised image generation with infinite generative adversarial networks. *CoRR*, abs/2108.07975, 2021. 1

[11] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017. 2