**DNA-GAN: Learning Disentangled Representations from Multi-Attribute Images**

Taihong Xiao, Jiapeng Hong & Jinwen Ma
Department of Information Sciences, School of Mathematical Sciences and LMAM, Peking University
xiaotaihong@pku.edu.cn, jphong@pku.edu.cn, jwma@math.pku.edu.cn

**Introduction**
Disentangling factors of variation has become a very challenging problem on representation learning. However, existing algorithms suffer from many limitations, such as unpredictable disentangling factors, unable to identify different styles of a certain attribute, lack of identity information, etc. Motivated by DNA double helix structure, in which different kinds of traits are encoded in different DNA pieces, we make a similar assumption that different visual attributes in an image are controlled by different pieces of encodings in its latent representations.

**Model**
As shown in Figure 1, DNA-GAN is mainly composed of three parts: an encoder (Enc), a decoder (Dec) and a discriminator (D).

We focus on one attribute in each iteration. Let’s say we are at the $i$-th attribute. $A$ and $B$ are required to have such labels, $y^A = (y^A_i, ... , y^A_J)$ and $y^B = (y^B_i, ... , y^B_J)$, respectively. The encoder maps the real-world images $A$ and $B$ into two latent disentangled representations.

$$
Enc(A) = [a_1, ... , a_i, ... , a_n, z_a] \quad \text{and} \quad Enc(B) = [b_1, ... , b_i, ... , b_n, z_b]
$$

We copy $Enc(A)$ directly as the latent representation of $A_1$, and annihilate $b_i$ in the copy of $Enc(B)$ as the latent representation of $B_1$. The annihilating operation means replacing all elements with zeros.

By swapping $a_i$ and $0_i$, we obtain two new latent representations $[a_1, ... , 0_i, ... , a_n, z_a]$ and $[b_1, ... , b_i, ... , b_n, z_b]$ that are supposed to be decoded into $A_2$ and $B_2$, respectively. Via a decoder Dec, we can get four newly generated images $A_1, B_1, A_2$ and $B_2$.

**Loss Functions**
The encoder and decoder receive two types of losses: (1) the reconstruction loss,

$$L_{\text{reconstruct}} = ||A - A_1||_1 + ||B - B_1||_1$$

(2) the standard GAN loss,

$$L_{\text{GAN}} = -\mathbb{E}[\log(D(A|y^A_i = 1))] - \mathbb{E}[\log(1 - D(B|y^B_i = 1))]$$

Omitting the coefficient, the loss function for the encoder and decoder is

$$L_{E} = L_{\text{reconstruct}} + L_{\text{GAN}}$$

The discriminator $D$ receives the standard GAN discriminator loss

$$L_{D_1} = -\mathbb{E}[\log(D(A|y^A_i = 1))] - \mathbb{E}[\log(1 - D(B_i|y^B_i = 1))]$$

$$L_{D_2} = -\mathbb{E}[\log(D(B|y^B_i = 0))] - \mathbb{E}[\log(1 - D(A_i|y^A_i = 0))]$$

$$L_D = L_{D_1} + L_{D_2}$$

**Experiments**

Experiments are conducted to verify the effectiveness of the DNA-GAN model. The input image is the input, and the other three images are the reference images of different styles of bangs.

**Figure 1:** DNA-GAN architecture.

**Figure 2:** Face image generation by adding different types of eyeglasses. The yellow box indicates the input images, the three on the top are reference images. Our model can add exact the same eyeglasses in the reference images to the input image.

**Figure 3:** Interpolation of different types of bangs. The top-left image is the input and images at the other three corners are three reference images of different styles of bangs.

**Figure 4:** Interpolation of bangs and mustache. The top-left image is the input. The bottom-left is the reference image of the bangs attribute and top-right image is the reference image of the mustache attribute.