Disentangled Representation Learning for Controllable Image Synthesis: an Information-Theoretic Perspective

Shichang Tang
ShanghaiTech University
SIMIT, Chinese Academy of Sciences
University of Chinese Academy of Sciences

Xu Zhou
Shenzhen University

Xuming He
ShanghaiTech University

Yi Ma
University of California, Berkeley

Introduction

It is desirable to make the latent variables in a generative model disentangled and interpretable. However, the learned representation in a Variational AutoEncoder(VAE) or a Generative Adversarial Network(GAN) is usually entangled and not interpretable. Therefore, we propose to use the framework in Fig. 1 for learning disentangled representation and performing controllable image synthesis. The proposed framework is a variant of VAE that has its latent code partitioned: z_2 is the representation of the specified feature f(x) while z_1 is complementary to z_2. z_1 is encoded from the input image and z_2 is encoded from its feature f(x). If this disentangled representation can be learned, then one can control the properties of the output image by manipulating the two latent codes.

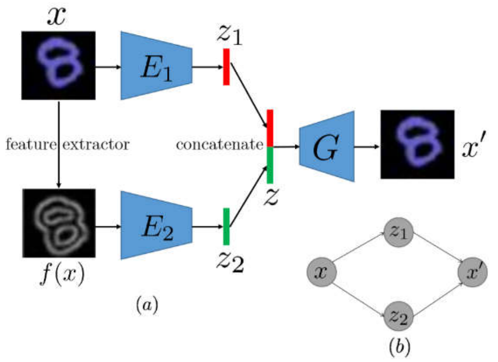
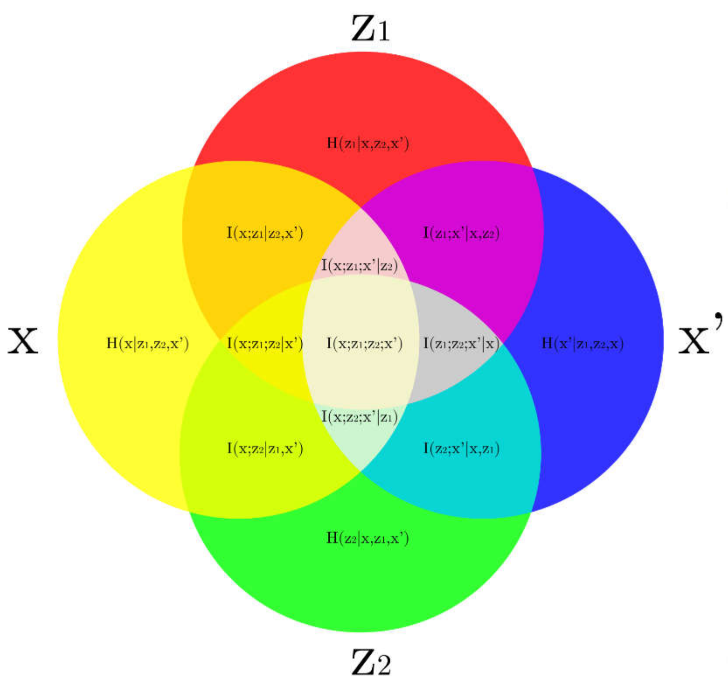


Fig. 1. model architecture and the relationship of x, z_1, z_2 and x' described as a probabilistic directed graphical model. The CMINST dataset is taken as an example.

Analysis

In the framework, since z_1 can contain complete information about the input, z_2 may be ignored in the training of VAE. We analyze the problem from the perspective of multivariate mutual information:

In order to make the Generator utilize its second input, maximizing The mutual information between z_2 and the output x' might be the solution. Unfortunately, this cannot help as the mutual information between z_2 and x' can be shared by z_1. Therefore, the conditional mutual information $I(z_2;x'|z_1)$ should be maximized instead.



Method

We derive lower bounds of the conditional mutual information when fusing two images(encode z_1 from input image x, encode z_2 from another independently sampled input image $\tilde{x}$, concatenate them and produce a fusion image $\hat{x}$) and incorporate them into the loss of the Encoders and the Generator:

$$I(\tilde{z}_2;\hat{x}|z_1) \geq \mathbb{E}[\log q(\tilde{z}_2|z_1,\hat{x})] + H(\tilde{z}_2|z_1)$$

$$I(\tilde{z}_2;\hat{x}|z_1) \geq \mathbb{E}[\log q^*(f(\tilde{x})|z_1,\hat{x})] + H(f(x))$$

When the second term in the bounds are constant, under a Bernoulli or Gaussian assumption, maximizing the bounds is equivalent to minimizing the reconstruction loss of $\tilde{z}_2$ or $f(\tilde{x})$:

$$L_{f1} = -\mathbb{E}[\log q(\tilde{z}_2|z_1,\hat{x})]$$

$$L_{f2} = \mathbb{E}[\|f(G(z_1,\tilde{z}_2)) - f(\tilde{x})\|_2^2]$$

For discrete features, we try to minimize L_{f1}; for continuous features, we try to minimize L_{f2}.

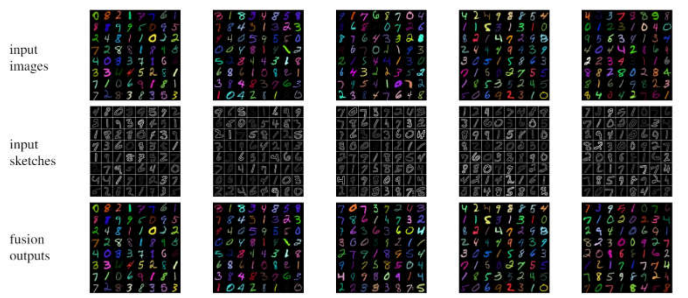
We also train a cGAN for better visual quality:

$$L_D = -\mathbb{E}[\log D(z_2,x') + \log(1-D(\tilde{z}_2,\hat{x}))] \quad L_{GAN} = -\mathbb{E}[\log(D(\tilde{z}_2,\hat{x}))]$$

The overall loss for the Encoders and the Generator is

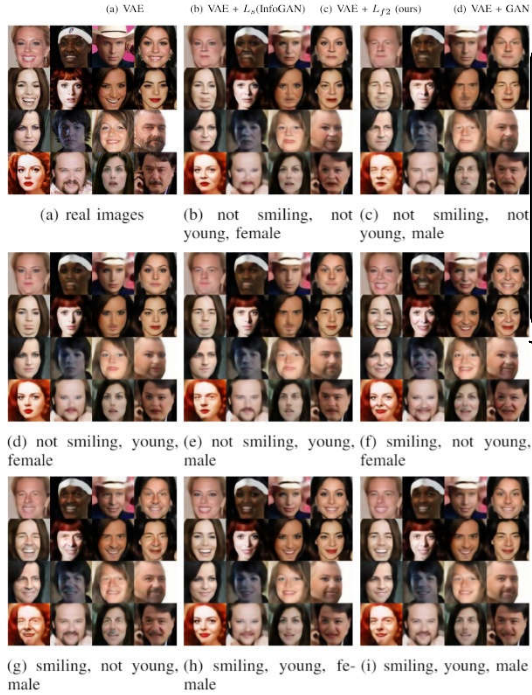
$$L = L_{VAE} + \lambda_3 L_f + \lambda_4 L_{GAN}$$

Results



(a) VAE (b) VAE + L_s(InfoGAN) (c) VAE + L_{f2} (ours) (d) VAE + GAN (e) VAE + GAN + L_{f2}

In the experiments on the CMNIST dataset, each input image is assigned with random color and f(x) is the sketch of the image. Our proposed approach succeeds in disentangling color and sketch. Since the color and the sketch are independent, adversarial training is unnecessary.



(a) real images (b) not smiling, not young, female (c) not smiling, not young, male (d) not smiling, young, female (e) not smiling, young, male (f) smiling, not young, female (g) smiling, not young, male (h) smiling, young, female (i) smiling, young, male

In the experiments on the CelebA dataset, f is a classifier and f(x) is one or more attributes of the image. Our proposed approach succeeds in disentangling and controlling multiple attributes simultaneously. For more results please refer to the paper and the supplementary material.