STVAE

Spatial Transformation VAE

Zhenya Liu

Spatial Transformation

We can get a good generative model by implementing VAE.

However, the latent variable $\mathbf{z}$ does not have any interpretability.

For example, we hope to disentangle part of $\mathbf{z}$ to get latent variable $\mathbf{u}$ such that $\mathbf{u}$ provides position information of generated $\mathbf{x}$.

Spatial Transformation

Idea

Consider $\mathbf{u}$ is made of some variables controlling affine transformations.

Spatial Transformation

What is **Affine transformation**?

Easily speaking,

$$\mathbf{y} = f(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$$

Example

Translation $\begin{pmatrix} a \\ b \end{pmatrix}$

Rotation

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

Shearing

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} x + my \\ y \end{pmatrix} = \begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} x \\ mx + y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ m & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

Spatial Transformation

In the paper [1],
The affine transformation in 2 d space is given by the matrix below:

$$\begin{bmatrix} S_x \cos\psi & -S_x \sin\psi & S_x T_x \\ S_y \sin\psi & S_y \cos\psi & S_y T_y \end{bmatrix}$$

where ψ gives the rotation, (S_x, S_y) gives the shearing, and (T_x, T_y) is the translation. In the case of affine transformation, we use 6 latent variables to account for the 6 entries in the transformation matrix.

TVAE

Now consider training one VAE with two latent variables $\mathbf{z}$ and $\mathbf{u}$, and assume they are **independent**, then

$$\text{ELBO} = E_{(z,u)\sim Q_\Psi(\cdot|x)} \{\log P_\Phi(x \mid z, u)\} - D_{KL}\left(Q_\Psi(z \mid x) \| P_\Phi(z)\right) - D_{KL}\left(Q_\Psi(u \mid x) \| P_\Phi(u)\right)$$

TVAE

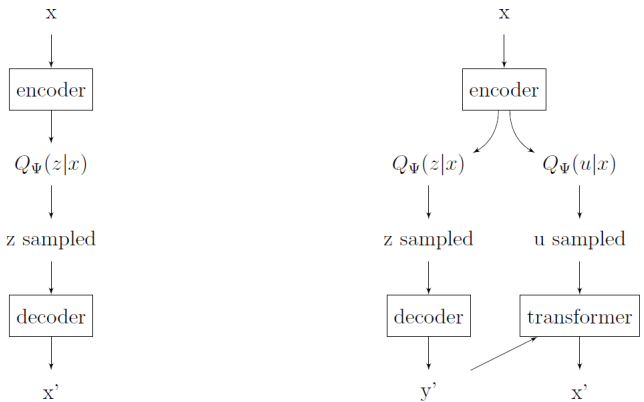


Figure 1: Architectures of VAE and T-VAE. Left: Vanilla VAE as in [7], x is the input image, x' is the reconstructed image; Right: T-VAE, y' is the reconstructed upright image

STVAE

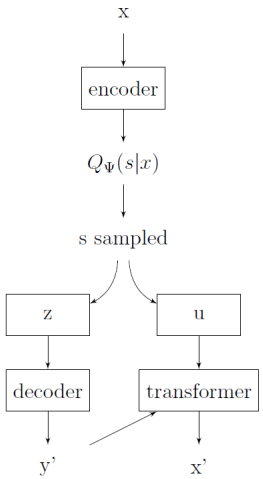
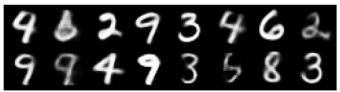
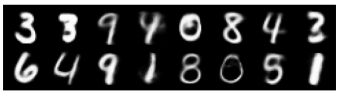


Figure 3: Architecture of S-TVAE

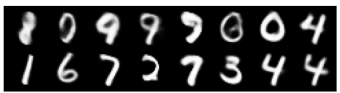
STVAE



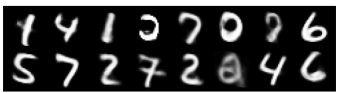
(a)



(b)



(c)



(d)

Figure: STVAE Generations

Reference

Heqing, Ye, 2019, Variational Auto-encoder with Spatial Transformations.