AE & GAN

Adversarial autoencoders

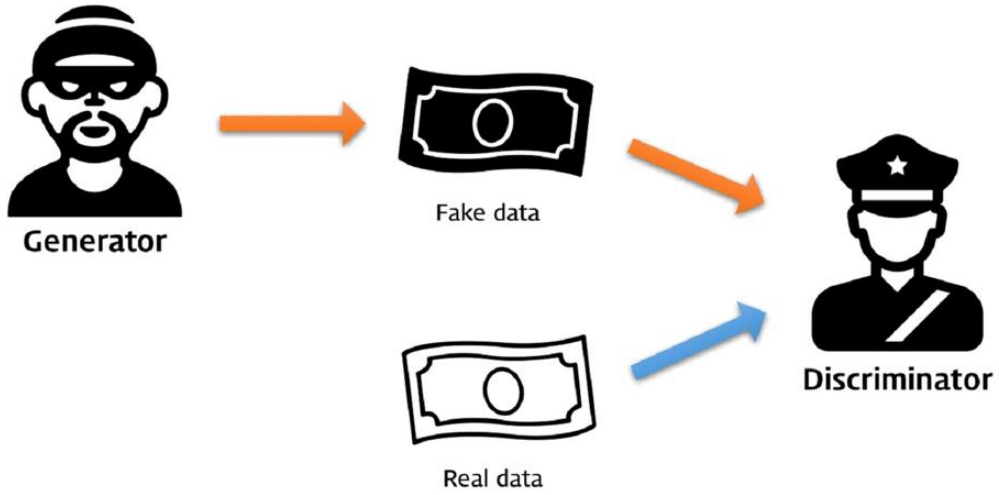
Makhzani, Alireza, et al. "Adversarial autoencoders."
arXiv preprint arXiv:1511.05644 (2015).

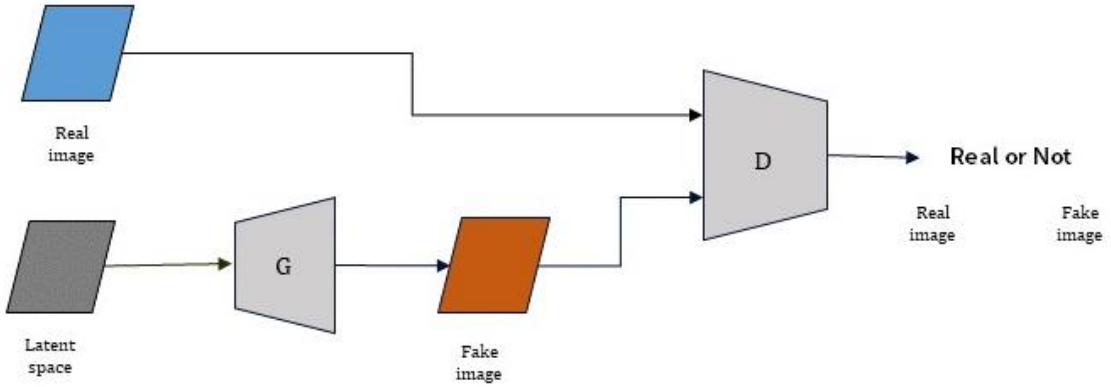
김수형

목차

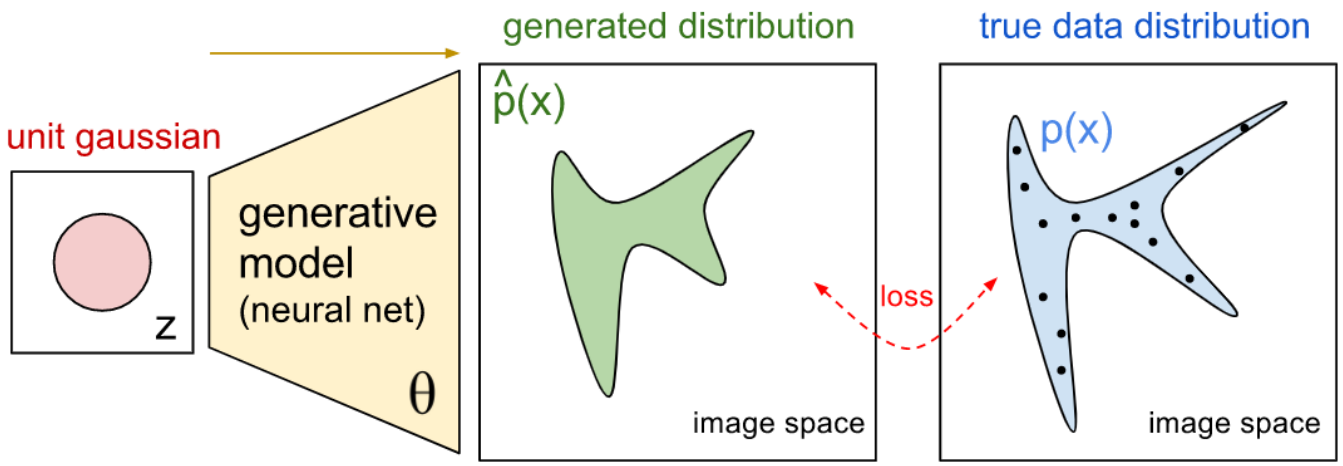
- What is GAN?
  (feat. Generative model)
- What is AE?
  (feat. manifold Hypothesis)
- VAE(variational autoencoder)
- AAE(adversarial autoencoders)

- **GAN(generative adversarial network)**
  - ✓ 그럴듯한 가짜를 만드는 적대적 생성 모델

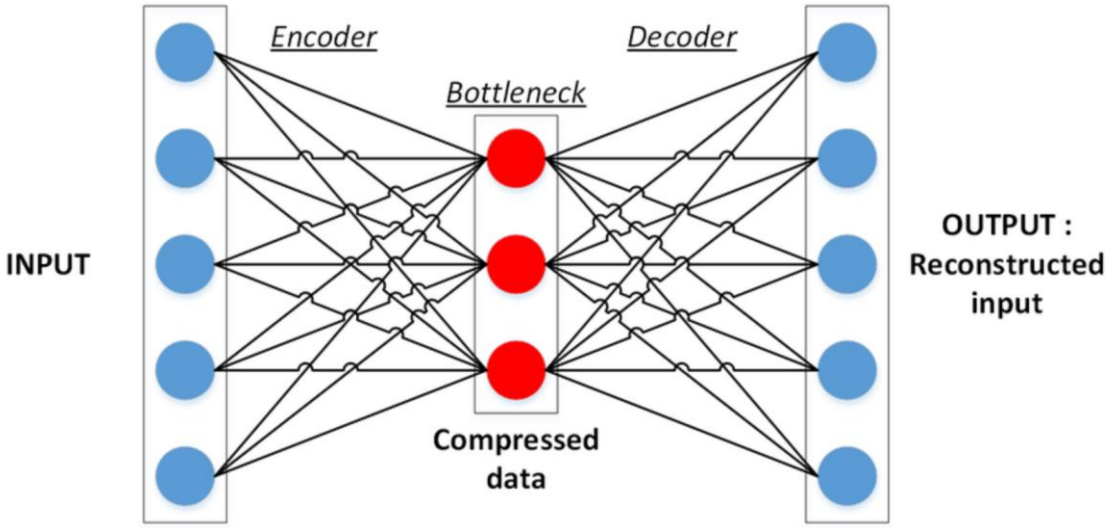


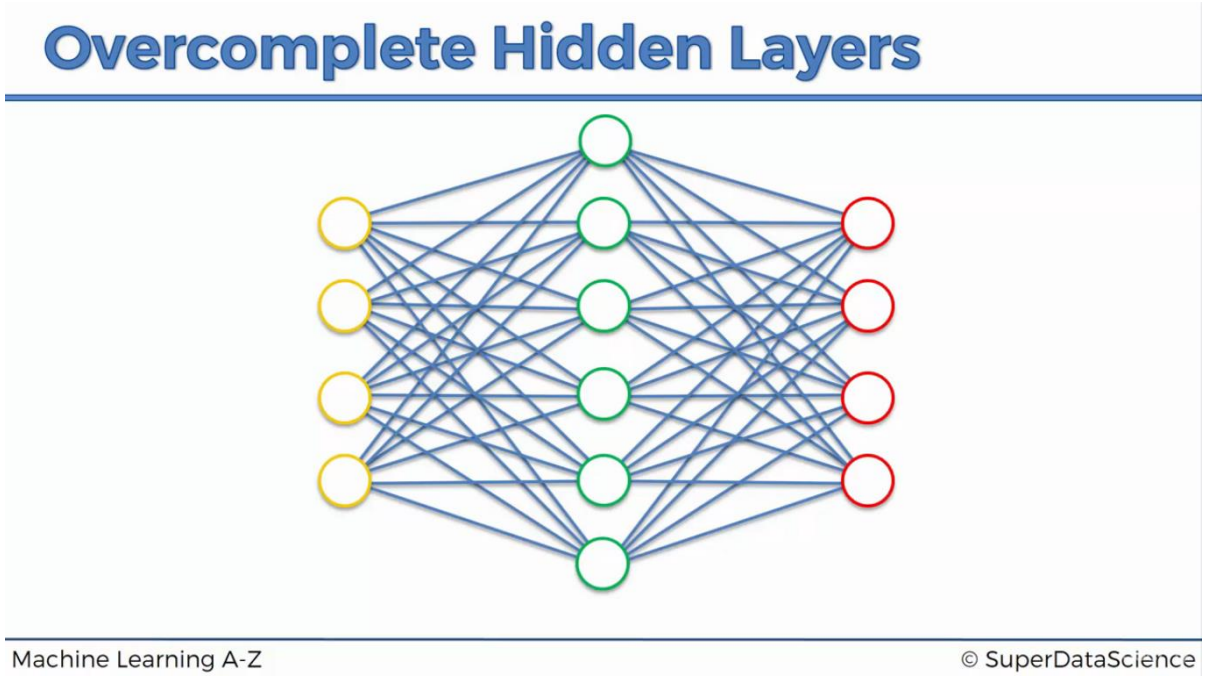


- **Generative model**
  - ✓ Target data의 분포를 학습하는 모델

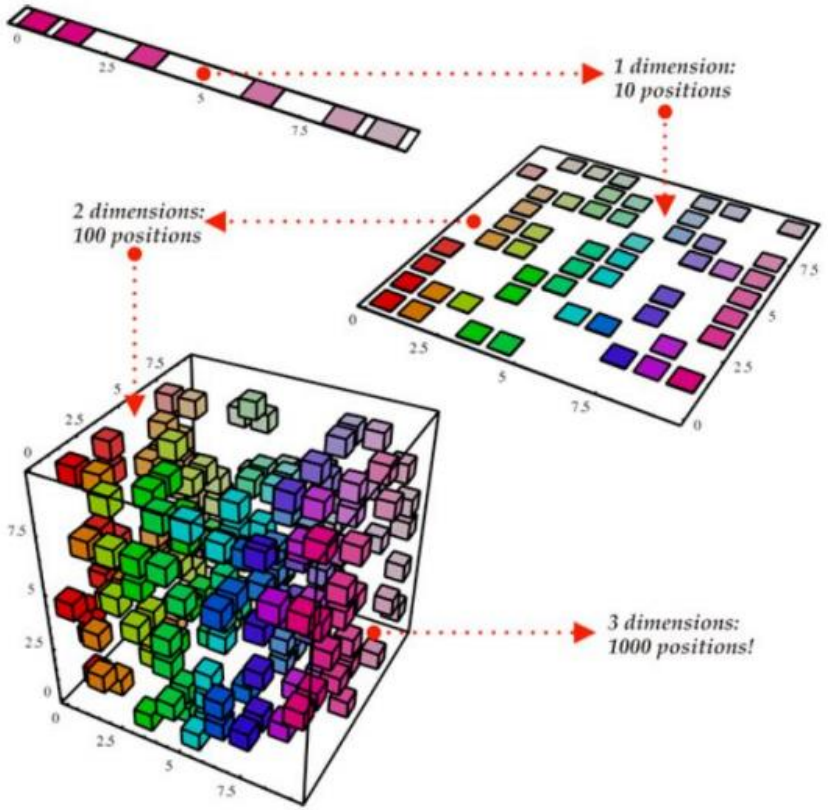


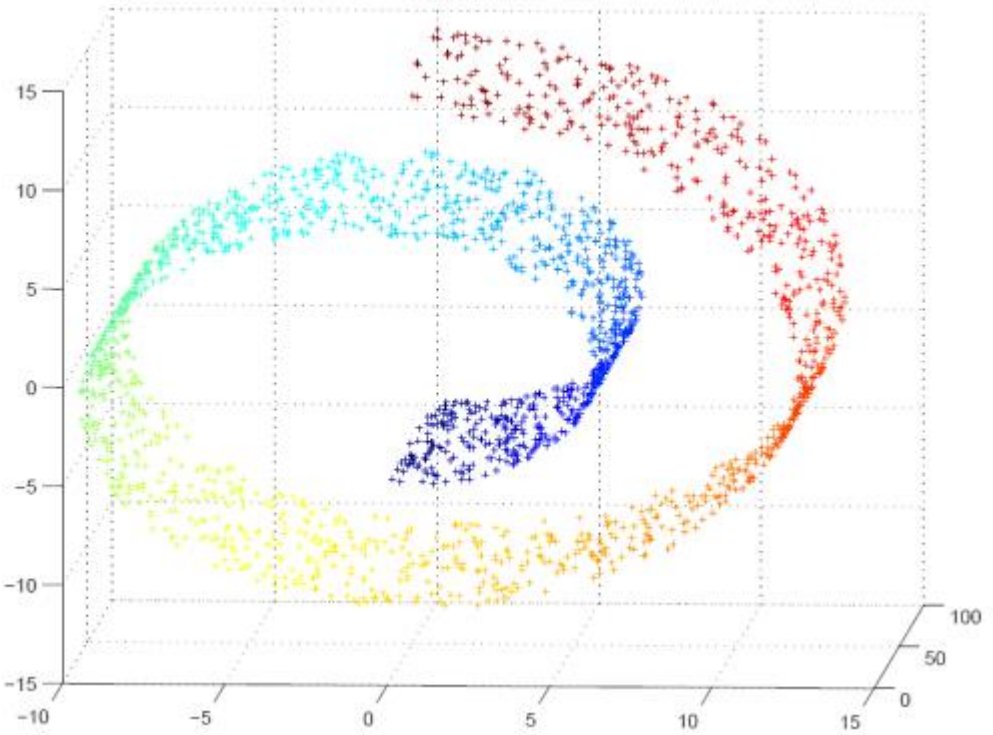
- **AE(auto-encoder)**
  - ✓ input을 넣었을 때 input과 똑같은 output을 출력하도록 만든 모델(신경망)



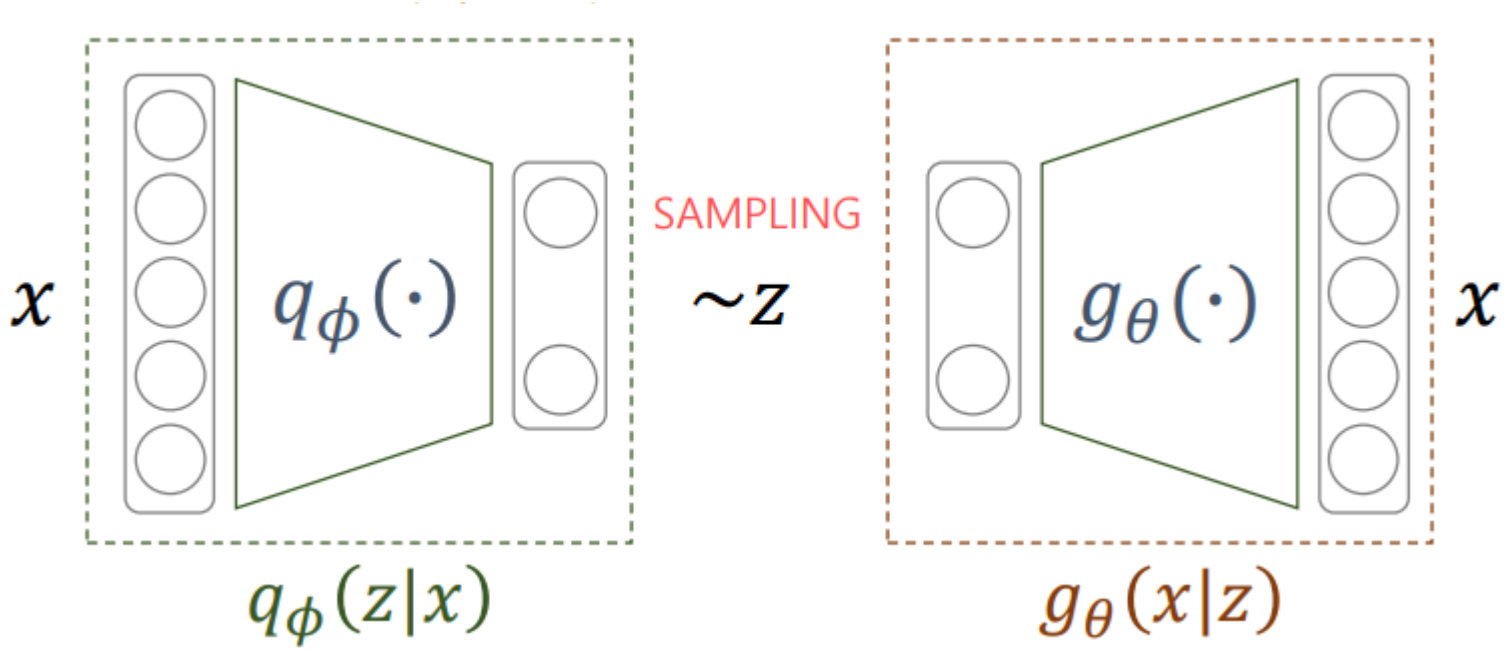


- **Manifold Hypothesis**
  - ✓ 고차원의 데이터의 밀도는 낮지만, 이들의 집합을 포함하는 저차원의 manifold(영역)이 존재한다. (차원이 높아질수록 밀도가 급격히 낮아진다.)





- **VAE(variational autoencoder)**
  - ✓ 데이터의 분포를 다루기 쉬운 확률 분포로 정규화 하여 해당 분포로부터 input과 비슷한 다른 output을 출력(generative model)



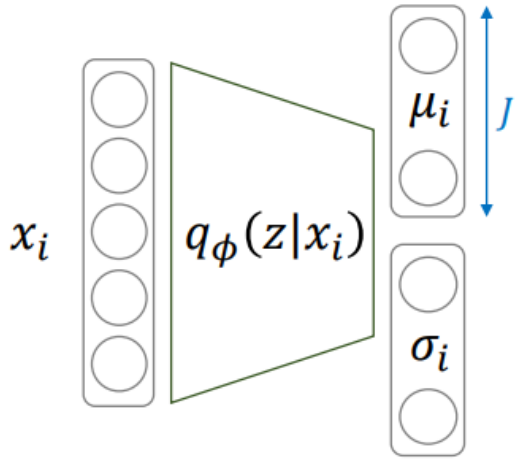
$$L_i(\phi,\theta,x_i) = \underbrace{-\mathbb{E}_{q_\phi(z|x_i)}\left[\log\left(p(x_i|g_\theta(z))\right)\right]}_{\text{Reconstruction Error}} + \underbrace{KL\left(q_\phi(z|x_i)||p(z)\right)}_{\text{Regularization}}$$

- **VAE(variational autoencoder)**
  - ✓ 데이터의 분포를 다루기 쉬운 확률 분포로 정규화 하여 해당 분포로부터 input과 비슷한 다른 output을 출력(generative model)

LOSS FUNCTION | Regularization

KL divergence

$$L_i(\phi,\theta,x_i) = -\mathbb{E}_{q_\phi(z|x_i)}[\log(x_i|g_\theta(z))] + \underbrace{KL\big(q_\phi(z|x_i)||p(z)\big)}_{\textbf{Regularization}}$$



$$KL\big(q_\phi(z|x_i)||p(z)\big) = \frac{1}{2}\left\{tr(\sigma_i^2 I) + \mu_i^T\mu_i - J + \ln\frac{1}{\prod_{j=1}^{J}\sigma_{i,j}^2}\right\}$$

$$= \frac{1}{2}\left\{\sum_{j=1}^{J}\sigma_{i,j}^2 + \sum_{j=1}^{J}\mu_{i,j}^2 - J - \sum_{j=1}^{J}\ln(\sigma_{i,j}^2)\right\}$$

$$= \frac{1}{2}\sum_{j=1}^{J}(\mu_{i,j}^2 + \sigma_{i,j}^2 - \ln(\sigma_{i,j}^2) - 1)$$ Easy to compute!

Kullback–Leibler divergence [edit]

The Kullback–Leibler divergence from $\mathcal{N}_0(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ to $\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, for non-singular matrices Σ_0 and Σ_1, is:[8]

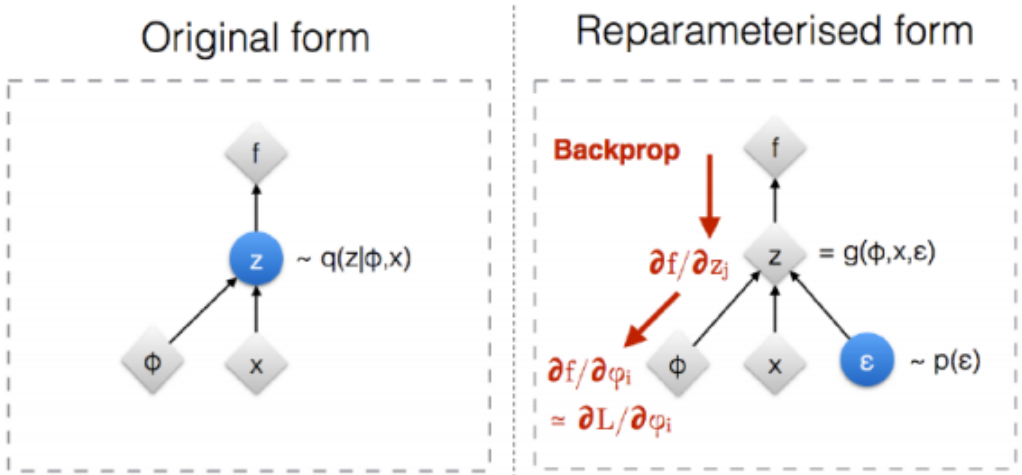
$$D_{\mathrm{KL}}(\mathcal{N}_0\|\mathcal{N}_1) = \frac{1}{2}\left\{\operatorname{tr}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0\right) + (\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0) - k + \ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|}\right\},$$

(posterior, prior)

where k is the dimension of the vector space.

- **VAE(variational autoencoder)**
  - ✓ 데이터의 분포를 다루기 쉬운 확률 분포로 정규화 하여 해당 분포로부터 input과 비슷한 다른 output을 출력(generative model)

Reparameterization Trick



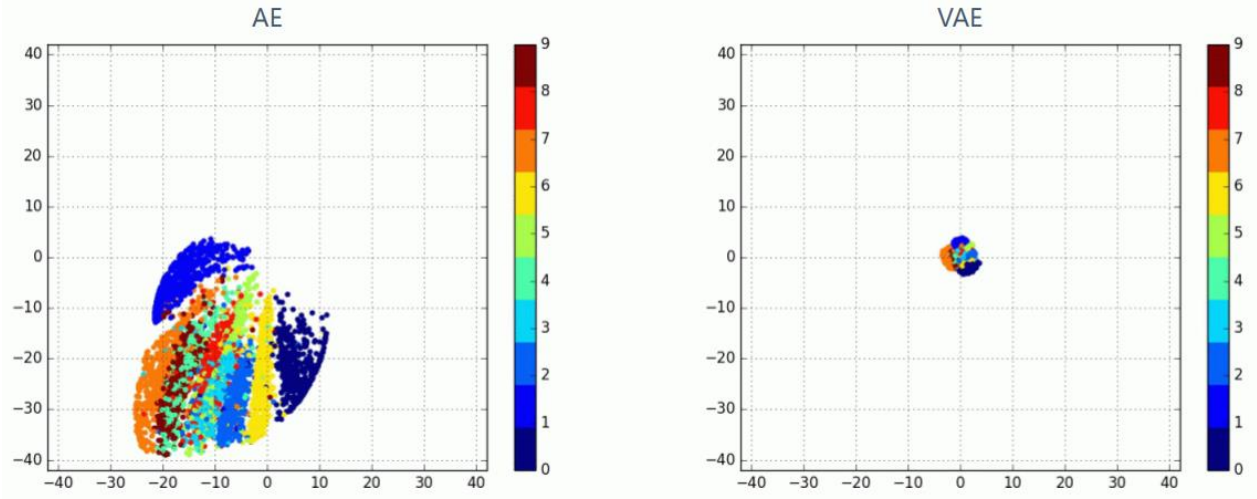
[Kingma, 2013]
[Bengio, 2013]
[Kingma and Welling 2014]
[Rezende et al 2014]

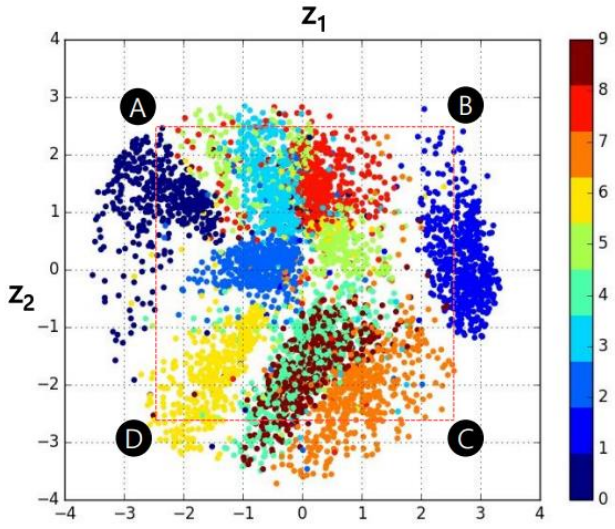
Sampling Process

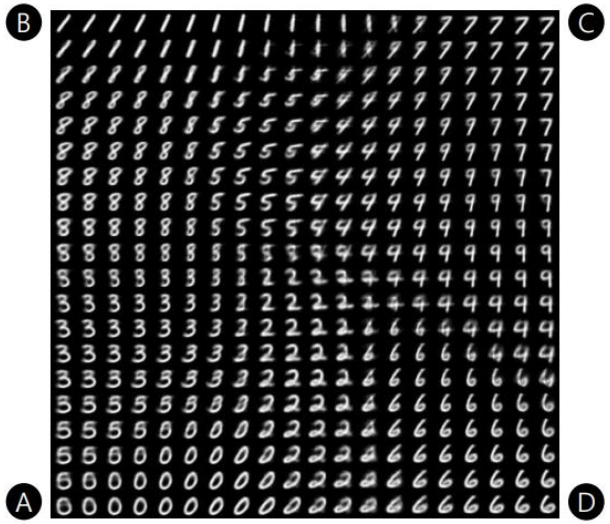
$$z^{i,l} \sim N(\mu_i, \sigma_i^2 I)$$

$$z^{i,l} = \mu_i + \sigma_i^2 \odot \epsilon$$
$$\epsilon \sim N(0, I)$$

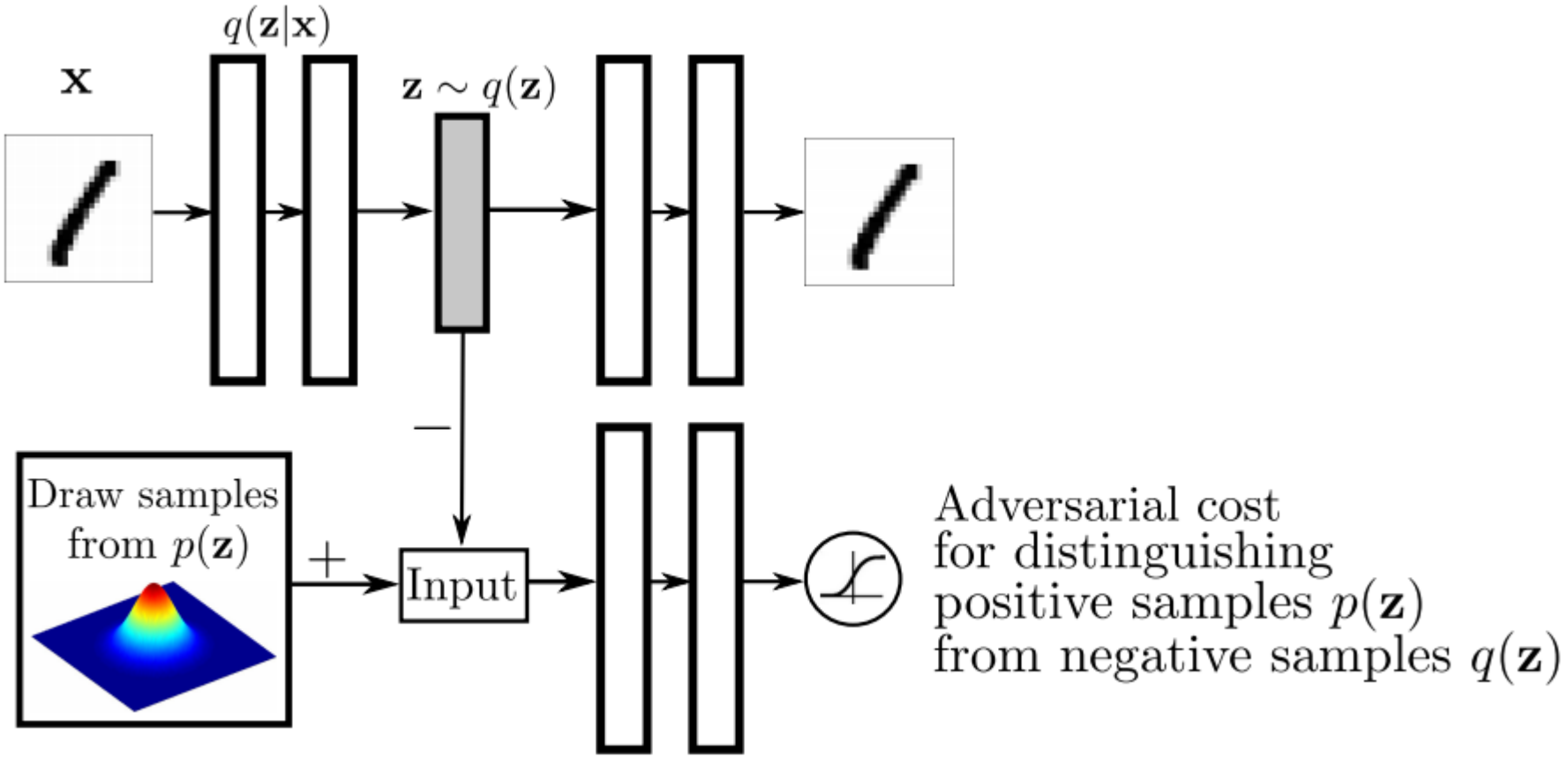
- **VAE(variational autoencoder)**
  - ✓ 데이터의 분포를 다루기 쉬운 확률 분포로 정규화 하여 해당 분포로부터 input과 비슷한 다른 output을 출력(generative model)







- **AAE(adversarial autoencoder)**
  - ✓ 데이터의 분포를 원하는 확률 분포로 정규화 하여 해당 분포로부터 input과 비슷한 다른 output을 출력(generative model)



- **AAE(adversarial autoencoder)**
  - ✓ 데이터의 분포를 원하는 확률 분포로 정규화 하여 해당 분포로부터 input과 비슷한 다른 output을 출력(generative model)

Loss Function

GAN loss

$$V(D,G) = \mathbb{E}_{z\sim p(z)}[\log D(z)] + \mathbb{E}_{x\sim p(x)}\left[\log\left(1 - D\left(q_\phi(x)\right)\right)\right]$$

Let's say G is defined by $q_\phi(\cdot)$ and D is defined by $d_\lambda(\cdot)$

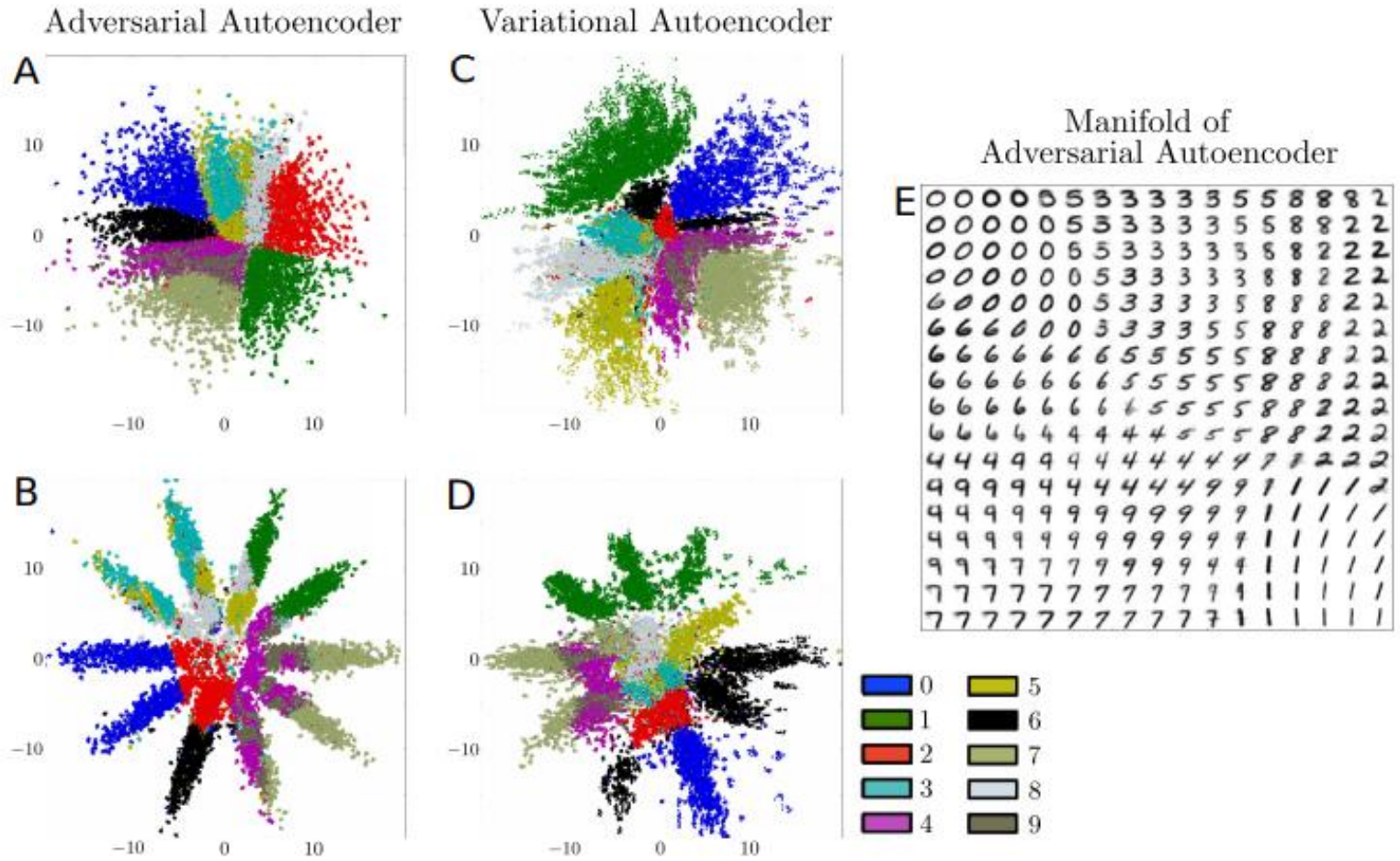
$$V_i(\phi,\lambda,x_i,z_i) = \log d_\lambda(z_i) + \log\left(1 - d_\lambda\left(q_\phi(x_i)\right)\right)$$

*논문에는 로스 정의가 제시되어 있지 않아 새로 정리한 내용

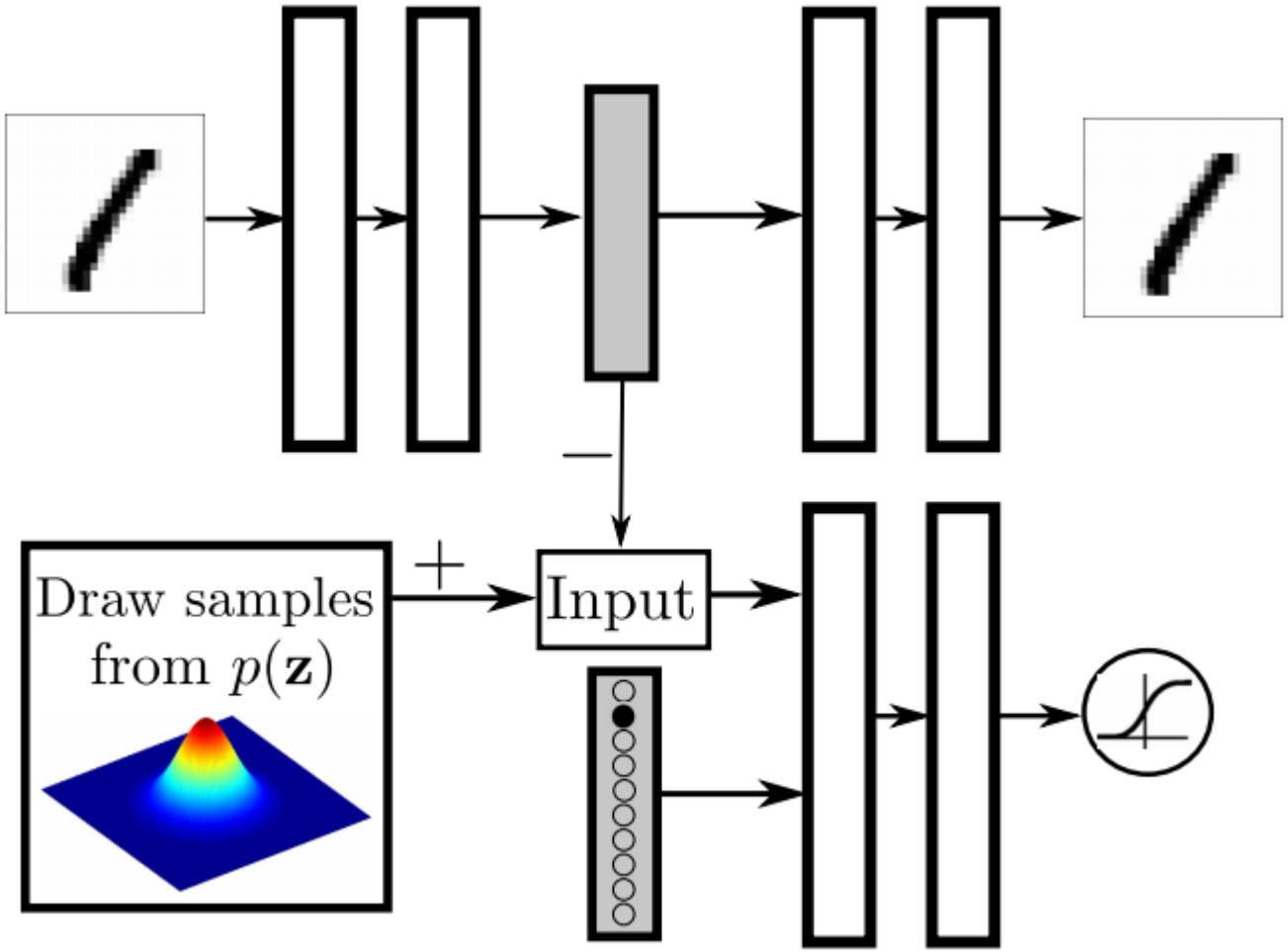
VAE loss

$$L_i(\phi,\theta,x_i) = -\mathbb{E}_{q_\phi(z|x_i)}\left[\log\left(p_\theta(x_i|z)\right)\right] + KL\left(q_\phi(z|x_i)||p(z)\right)$$

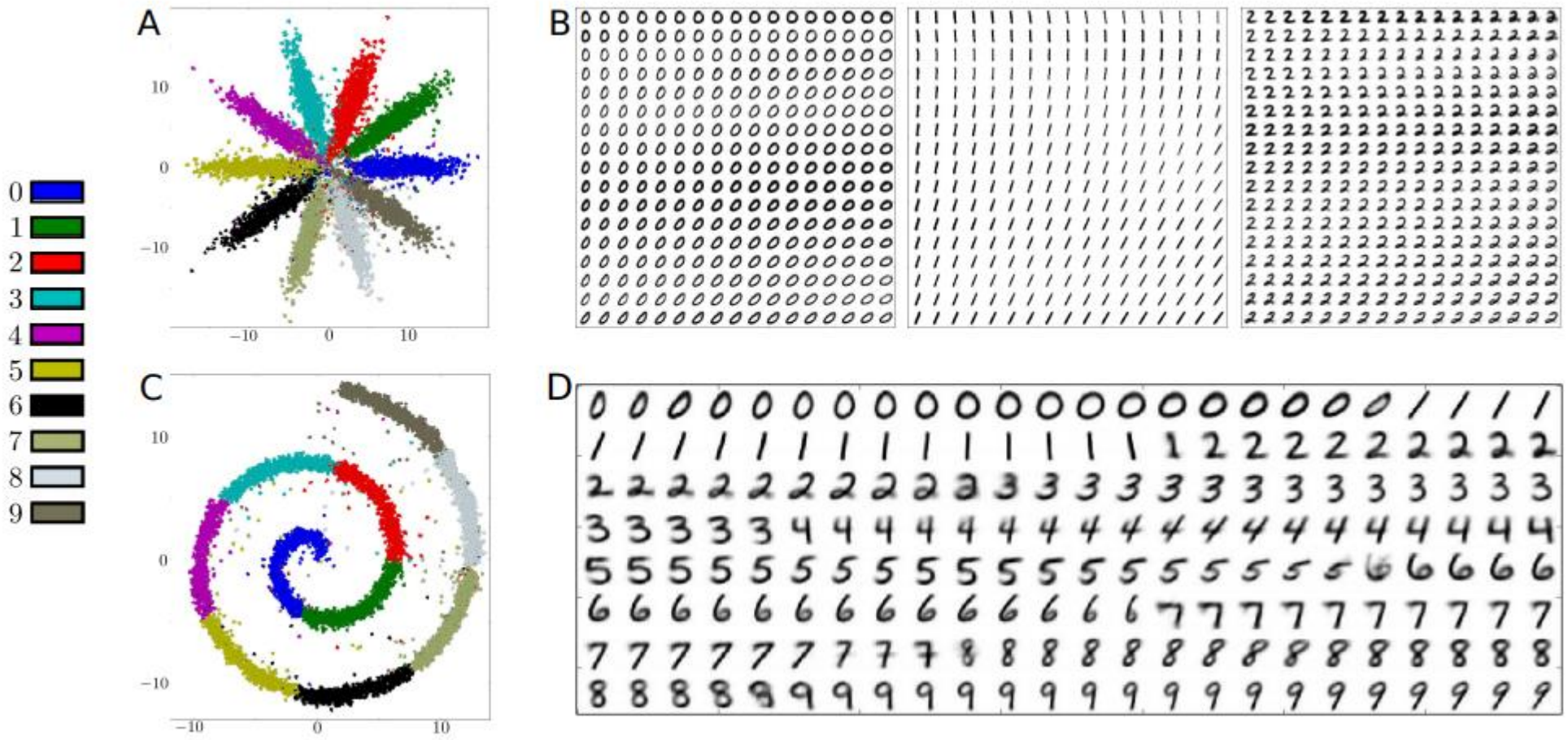
- **AAE(adversarial autoencoder)**
  - ✓ 데이터의 분포를 원하는 확률 분포로 정규화 하여 해당 분포로부터 input과 비슷한 다른 output을 출력(generative model)



- **AAE(adversarial autoencoder) + condition**
  - ✓ 데이터의 분포를 원하는 확률 분포로 정규화 하여 해당 분포로부터 input과 비슷한 다른 output을 출력(generative model)



- **AAE(adversarial autoencoder)**
  - ✓ 데이터의 분포를 원하는 확률 분포로 정규화 하여 해당 분포로부터 input과 비슷한 다른 output을 출력(generative model)



Reference

- Makhzani, Alireza, et al. "Adversarial autoencoders." *arXiv preprint arXiv:1511.05644* (2015).
- https://www.youtube.com/watch?v=rNh2CrTFpm4(오토인코더의 모든것)
- https://www.slideshare.net/NaverEngineering/ss-96581209?from_action=save(오토인코더의 모든것)

End

Appendix