AE를 활용하여 AE의 latent space의 분포학습 및 낮은 recon loss 달성

NF를 활용하여, VAE보다 높은 성능을 보임.

* Motivations.

PAE는 Linear PCA와 PCA-based data likelihood를 측정하는 확률적 해석에서 동기부여 받았다.

고정된 latent space $k(k<N)$에서 PCA $x \in \mathbb{R}^N$은 orthogonal linear transform O로 찾을 수 있다.

$$\boldsymbol{O} : \mathbb{R}^K \to \mathbb{R}^N, \boldsymbol{z} \mapsto \boldsymbol{O}\boldsymbol{z}, \boldsymbol{O}\boldsymbol{O}^T = \mathbb{1}_N$$

↳ latent space의 data variance를 최대화.

↳ recon error 최소화.

↳ PCA는 analysis 하고, 주성분은 공분산행렬의 eigenvector로 주어짐.

적절한 k는 eigenvalue에 의해 결정되고, 큰 값에 해당하는 eigenvector만 남긴다

reconstruction error $\sigma^2_{recon} = \sum_{i=k+1}^{N} \lambda_i$

PCA에서 data model은 $x = Oz + \epsilon$ → 나머지

↳ 큰 eigenvalue의 eigenvector로 이뤄짐.

Probabilistic PCA(PPCA)의 residual은 gaussian으로 가정된다.

Implicit likelihood는

$$\ln \tilde{p}(\boldsymbol{x}|\boldsymbol{z}) = -\frac{1}{2}\left[N \ln(2\pi) + \ln\det\boldsymbol{\Sigma} + (\boldsymbol{x} - \boldsymbol{O}\boldsymbol{z})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{O}\boldsymbol{z})\right].$$

↳ $\Sigma = \sigma^2_{recon} I_N$

Prior의 경우, PPCA에선 latent가 $N(0, \Lambda)$를 따른다고 가정 (↱ diagonal matrix)

gaussian에서는 marginal을 analytically 계산 가능

$$\ln \tilde{p}(\boldsymbol{x}) = -\frac{1}{2}\left[N \ln(2\pi) + \ln\det\boldsymbol{C} + \boldsymbol{x}^T \boldsymbol{C}^{-1} \boldsymbol{x}\right], \quad \text{with } \boldsymbol{C} = \boldsymbol{O}\boldsymbol{\Lambda}\boldsymbol{O}^T + \boldsymbol{\Sigma}.$$

즉, PPCA는 recon error를 최소화할 basis를 찾고, latent의 distribution을 prior로 model을 구성.

논문은 AE로 이 과정을 비선형화. PCA → AE, gaussian prior → NF

* The probabilistic AE.

- PAE training

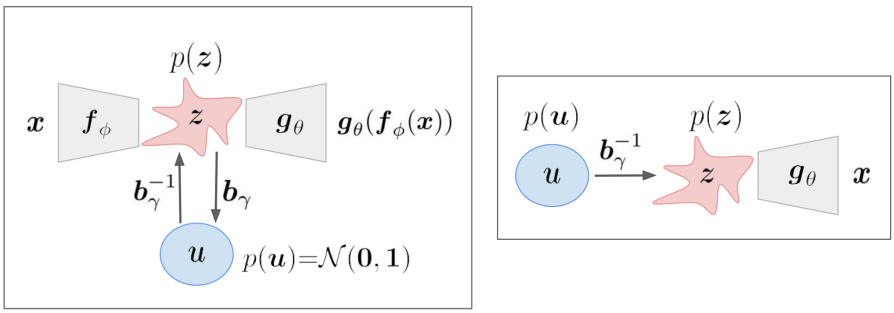

Figure 1: Schematic diagram of the PAE (left panel) and an illustration of the sampling procedure from the PAE (right panel). The autoencoder networks are depicted as gray trapezia, the normalizing flow is represented by black arrows and the latent spaces of the autoencoder and normalizing flow are shown in red and blue, respectively.

2 stage로 동작

stage 1.

enc : f_ϕ . dec : g_θ , latent dim : K

$$\mathcal{L}_{\mathrm{AE}} = \mathbb{E}_{p(\boldsymbol{x})} ||\boldsymbol{x} - \boldsymbol{g}_\theta(\boldsymbol{f}_\phi(\boldsymbol{x}))||_2^2.$$

확률 모델이 아닌 AE를 확률적으로 해석하기 위해 stage 2.

stage 2.

encoding 된 훈련 data에 대해 density estim을 하여 p(z) 근사. (NF 이용)

$$\boldsymbol{b}_\gamma : \mathbb{R}^K \to \mathbb{R}^K, \boldsymbol{z} \mapsto \boldsymbol{u}=\boldsymbol{b}_\gamma(\boldsymbol{z}).$$

$$p_\gamma(\boldsymbol{z}) = q(\boldsymbol{u})|\nabla_{\boldsymbol{z}} \boldsymbol{b}_\gamma(\boldsymbol{z})|.$$

↳ N(0,I) ↳ Jacobian.

$$\mathcal{L}_{\mathrm{NF}} = \mathbb{E}_{\tilde{p}(\boldsymbol{z})}[-\ln p_\gamma(\boldsymbol{z})] = \mathbb{E}_{\tilde{p}(\boldsymbol{z})} \left[-\ln p(\boldsymbol{u}) - \ln \left| \det \frac{\partial \boldsymbol{b}_\gamma^{-1}(\boldsymbol{u})}{\partial \boldsymbol{u}} \right| \right]_{\boldsymbol{u}=\boldsymbol{b}_\gamma(\boldsymbol{z})}.$$

↳ latent 를 N(0,I)로 mapping

$$\boldsymbol{x} = \boldsymbol{g}_\theta(\boldsymbol{b}_\gamma^{-1}(\boldsymbol{u})).$$

AE는 확률 모델이 아니기 때문에 encoding된 latent는 delta로 볼 수 있다.

NF는 이런 delta를 부드럽게 보간

완벽한 확률 모델은 전체 x에 대한 밀도추정이 필요하고, posterior에 대해 density estim이 필요.

$$p_{\text{model}}(\boldsymbol{z}) = \int \mathrm{d}\boldsymbol{x}\, p(\boldsymbol{x})\, p_{\text{model}}(\boldsymbol{z}|\boldsymbol{x}) = \mathbb{E}_{p(\boldsymbol{x})}\left[p_{\text{model}}(\boldsymbol{z}|\boldsymbol{x})\right],$$

NF는 낮은 차원에 대해서는 빠르고 효율적이다.

- Comparison to VAE

VAE는 ELBO에서 lower bound를 보장하지만, $\mathcal{L}_{\text{rec}}$과 $\mathcal{L}_{\text{kl}}$ 사이의 조정이 필요함

PAE는 개별적으로 최적화.

$$\mathcal{L}_{\beta-\text{VAE}} = -\,\mathbb{E}_{p(\boldsymbol{x})}\left[\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\ln p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - \beta\, \mathrm{D}_{\text{KL}}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right]\right].$$

$$\mathcal{L}_{\text{flow}-\text{VAE}} = -\,\mathbb{E}_{p(\boldsymbol{x})}\left[\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\ln p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - \mathrm{D}_{\text{KL}}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\gamma(\boldsymbol{z})\right]\right].$$

* Downstream task.

- Anomaly detection.

likelihood를 최적화하는 VAE 등은 anomaly detection에 실패할 수 있다.

하지만 PAE는 likelihood를 최적화하지 않고, reconㄹ prior를 따로한다.

∴ Anomaly detection에 유용.

$$\ln \tilde{p}(\boldsymbol{z}) = -\frac{1}{2}\left[K\ln(2\pi) + \sum_{i}^{K}\ln\lambda_i + \sum_{i}^{K} z_i\lambda_i^{-1}z_i\right].$$

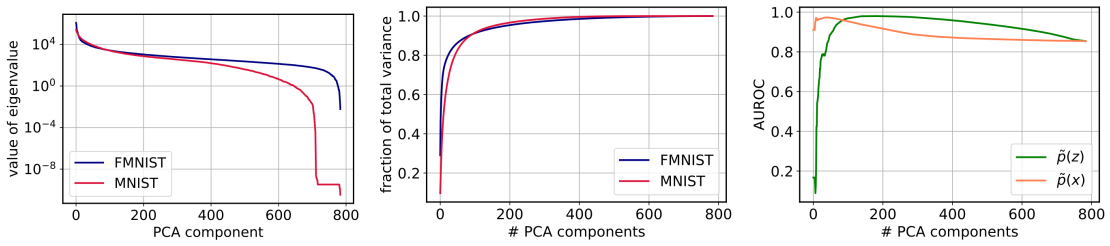

Figure 2: Principal component analysis of FashionMNIST and MNIST data sets and outlier detection accuracy (in-distribution: Fashion MNIST, out-of-distribution: MNIST) with equation 4 and equation 15 as a function of included number of PCA components. A higher AUROC value corresponds to a better separation between in- and out-of-distribution data.

* Experiments.