Improving Relational Regularized Autoencoders with Spherical Sliced Fused Gromov Wasserstein

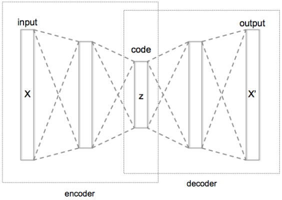
Yue Xiang

2021. 1. 15

Outlines

Autoencoder

- An autoencoder is a type of artificial neural network used to learn efficient data codings in an unsupervised manner.
- Autoencoders consist of two components, namely, an encoder E_ϕ and a decoder G_θ.
- Major task: to obtain a decoder G such that its induced distribution p_G and the data distribution are very close under some discrepancies.
- Two popular instances of autoencoders
  - variational autoencoder(VAE): KL divergence
  - Wasserstein autoencoder(WAE): Wasserstein distance



Learning Problem

$$min_{\theta,\phi} \underbrace{\mathbb{E}_{p_d(x)}\mathbb{E}_{q_\phi(z|x)}[d(x, G_\theta(z))]}_{reconstruction\ loss} + \underbrace{\lambda D(q_\phi(z)\|p(z))}_{distance(posterior,\ prior)} \quad (1)$$

d: the ground metric of Wasserstein distance.
D: a discrepancy between distributions.
$q_\phi(z|x)$: a distribution for encoder $E_\phi : X \to Z$, parameterized by ϕ.

WAE and RAE

- WAE
  D: MMD/GAN, $p(z)$: Gaussian.
  Cons: WAE suffers from either over-regularization or under-regularization problem.
- Relational regularized AutoEncoder(RAE)
  $p(z)$: mixture of Gaussian $p_{\mu_{1:k},\Sigma_{1:k}}(z)$.
  The state-of-the-art version of RAE: deterministic relational regularized autoencoder(DRAE).

Notations

- $\mathbb{S}^{d-1}$: d-dimensional hypersphere.
- $\mathcal{U}(\mathbb{S}^{d-1})$: uniform distribution on $\mathbb{S}^{d-1}$.
- $\Pi(\mu, v)$: the set of all transport plans between μ and v.
- $\theta\#\mu$: the pushforward measure of μ through the mapping $\mathcal{R}_\theta$ where $\mathcal{R}_\theta(x) = \theta^T x$ for all x.

Deterministic Relational Regularized Autoencoder (DRAE)

Object function:

$$min_{\theta,\phi,\mu_{1:k},\Sigma_{1:k}} \mathbb{E}_{p_d(x)} \mathbb{E}_{q_\phi(z|x)}[d(x, G_\theta(z))] + \lambda \mathbb{E}_{q_\phi(z), p_{\mu_{1:k},\Sigma_{1:k}}(z)} SFG[(\hat{q}_N(z) \| \hat{p}_N(z))]. \quad (2)$$

where $\hat{q}_N(z)$ and $\hat{p}_N(z)$ are the empirical distributions of $q_\phi(z)$ and $p_{\mu_{1:k},\Sigma_{1:k}(z)}$ respectively.

Sliced Fused Gromov Wasserstein(SFG)

Let μ, $v \in \mathcal{P}(\mathbb{R}^d)$ be two probability distributions, β be a constant in [0,1], and $d_1 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ be a pseudo-metric on $\mathbb{R}$.
The sliced fused Gromov Wasserstein (SFG) between μ and is defined as:

$$SFG(\mu, v; \beta) := \mathbb{E}_{\theta \sim \mathcal{U}(\mathbb{S}^{d-1})}[D_{fgw}(\theta\#\mu, \theta\#v; \beta, d_1)] \tag{3}$$

where the fused Gromov Wasserstein D_{fgw} is given by:

$$\begin{aligned} D_{fgw}(\theta\#\mu, \theta\#v; \beta, d_1) := min_{\pi \in \Pi(\theta\#\mu, \theta\#v)} \Big\{ (1-\beta) \int_{\mathbb{R}^d \times \mathbb{R}^d} d_1(\theta^T x, \theta^T y) d\pi(x, y) + \\ \beta \int_{(\mathbb{R}^d)^4} [d_1(\theta^T x, \theta^T x') - d_1(\theta^T y, \theta^T y')]^2 d\pi(x, y) d\pi(x', y') \Big\}. \end{aligned} \tag{4}$$

Sliced Fused Gromov Wasserstein(SFG)

- relational regularization: sliced fused Gromov Wasserstein(SFG).
- Cons: SFG uses the uniform distribution over the unit sphere ($\theta \sim \mathcal{U}(\mathbb{S}^{d-1})$) to sample projecting directions, which leads to the underestimation of the discrepancy between the two distributions.
- Pros:
  - SFG is a linear combination of sliced Wasserstein (SW) and sliced Gromov Wasserstein (SG), so takes advantages of both of them.
  - If μ and v have n supports and uniform weights and $d_1(x, y) = (x - y)^2$, SFG has good computational complexity ($\mathcal{O}(nlogn)$).

Von Mises-Fisher Distribution

The von Mises–Fisher distribution (vMF) is a probability distribution on the unit sphere $\mathbb{S}^{d-1}$ where its density function is given by:

$$f(x|\epsilon, \kappa) := C_d(\kappa) exp(\kappa \epsilon^T x), \tag{5}$$

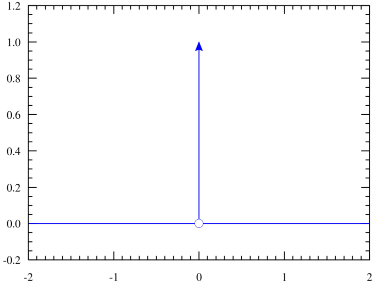
where $\kappa \geq 0$ is the concentration parameter , $\epsilon \in \mathbb{S}^{d-1}$ is the location vector , and $C_d(\kappa) := \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} \mathcal{I}_{d/2-1}(\kappa)}$ is the normalization constant. Here, $I_{d/2-1}$ is the modified Bessel function of the first kind at order $d/2-1$. It is possible to define the Bessel function by its series expansion around $x = 0$ as:

$$\mathcal{I}_{\frac{d}{2}-1}(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{m!\Gamma(m+\frac{d}{2})} \left(\frac{\kappa}{2}\right)^{2m+\frac{d}{2}-1} \tag{6}$$

The vMF concentrates around mode ϵ and its density decreases when x goes away from ϵ.

Von Mises-Fisher Distribution

- By changing κ from 0 to infinity, the vMF family could interpolate from the uniform distribution to any Dirac distribution on the sphere.
- When $\kappa \to 0$, vMF converges to the uniform distribution.
- When $\kappa \to \infty$, vMF approaches to the Dirac distribution centered at ϵ.



Spherical Sliced Fused Gromov Wasserstein(SSFG)

Definition 3. *(SSFG) Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be two probability distributions, $\kappa > 0$, $\beta \in [0,1]$, $d_1 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ be a pseudo-metric on $\mathbb{R}$. The* ***spherical sliced fused Gromov Wasserstein*** *(SSFG) between μ and ν is defined as follows:*

$$SSFG(\mu, \nu; \beta, \kappa) := \max_{\epsilon \in \mathbb{S}^{d-1}} \mathbb{E}_{\theta \sim \mathrm{vMF}(\cdot|\epsilon,\kappa)} \left[D_{fgw}(\theta \sharp \mu, \theta \sharp \nu; \beta, d_1) \right], \tag{6}$$

where the fused Gromov Wasserstein D_{fgw} is defined at equation (3).

Complexity:

- General case of d_1: $\mathcal{O}(n^4)$
- $d_1(x, y) = (x - y)^2$: $\mathcal{O}(nlogn)$

Key Properties of SSFG

SSFG is a pseudo-metric in the probability space and does not suffer from the curse of dimensionality.

Theorem 1. *For any $\beta \in [0,1]$ and $\kappa > 0$, SSFG$(.,.;\beta,\kappa)$ is a pseudo-metric in the space of probability measures, namely, it is non-negative, symmetric, and satisfies the weak triangle inequality.*

Theorem 2. *For any probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, the following holds:*

(a)

$$\lim_{\kappa \to 0} SSFG(\mu,\nu;\beta,\kappa) = SFG(\mu,\nu;\beta),$$

$$\lim_{\kappa \to \infty} SSFG(\mu,\nu;\beta,\kappa) = \max_{\theta \in \mathbb{S}^{d-1}} D_{fgw}(\theta\sharp\mu, \theta\sharp\nu;\beta) := max\text{-}SFG(\mu,\nu;\beta).$$

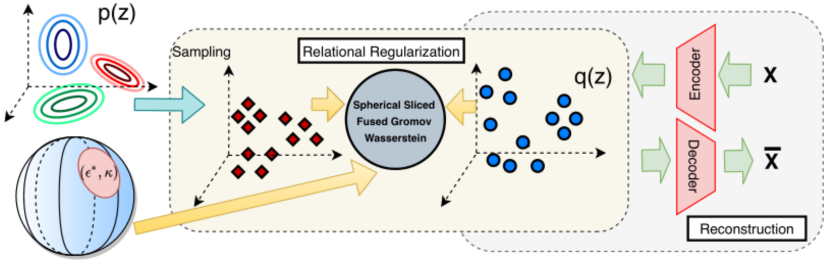
(b) For any $\kappa > 0$, we find that

$$\exp(-\kappa)C_d(\kappa)SFG(\mu,\nu;\beta) \leq SSFG(\mu,\nu;\beta,\kappa) \leq \exp(\kappa)C_d(\kappa)SFG(\mu,\nu;\beta),$$

$$SSFG(\mu,\nu;\beta,\kappa) \leq max\text{-}SFG(\mu,\nu;\beta).$$

s-DRAE

$$min_{\theta,\phi,\mu_{1:k},\Sigma_{1:k}} \mathbb{E}_{p_d(x)} \mathbb{E}_{q_\phi(z|x)}[d(x, G_\theta(z))] + \lambda \mathbb{E}_{q_\phi(z), p_{\mu_{1:k},\Sigma_{1:k}}(z)} SSFG[(\hat{q}_N(z) \| \hat{p}_N(z))].$$



2 Variants of SSFG

- mixture spherical sliced fused Gromov Wasserstein(MSSFG)

$$\begin{aligned} &\mathit{MSSFG}(\mu, \nu; \beta, \{\kappa_i\}_{i=1}^{\kappa}, \{\alpha_i\}_{i=1}^{\kappa}) \\ &\qquad := \max_{\epsilon_{1:k} \in \mathbb{S}^{d-1}} \mathbb{E}_{\theta \sim \mathit{MovMF}(\cdot|\epsilon_{1:k}, \{\kappa_i\}_{i=1}^{k}, \{\alpha_i\}_{i=1}^{k})} \left[D_{fgw}(\theta \sharp \mu, \theta \sharp \nu; \beta, d_1) \right], \end{aligned} \tag{7}$$

where D_{fgw} is defined in equation (3) and the mixture of vMF distributions is defined as $\mathit{MovMF}(\cdot|\epsilon_{1:k}, \{\kappa_i\}_{i=1}^{k}, \{\alpha_i\}_{i=1}^{k}) := \sum_{i=1}^{k} \alpha_i \mathrm{vMF}(\cdot|\epsilon_i, \kappa_i)$.

- power SSFG (PSSFG)
  PSSFG uses power spherical distribution instead of vMF and its mixtures as the slicing distribution to improve the computational time of (M)SSFG.

FID Score and Reconstruction Loss

Method	MNIST FID	MNIST Reconstruction	CelebA FID	CelebA Reconstruction
VAE	71.55 ± 26.65	18.59 ± 2.22	59.99(*)	96.36(*)
GMVAE	75.68 ±11.95	18.19 ± 0.14	212.59 ± 18.15	97.77± 0.19
Vampprior	138.03 ± 34.09	29.98± 4.09	-	-
PRAE	100.25 ±41.72	16.20 ±3.14	52.20 (*)	**63.21(*)**
WAE	80.77 ± 11	11.53 ±0.33	52.07 (*)	63.83(*)
SWAE	80.28 ±19.22	14.12 ± 2.06	86.53 ± 2.49	89.71±2.15
DRAE	58.04 ± 20.74	14.07 ± 4.31	50.09 ±1.33	66.05 ± 2.56
m-DRAE (ours)	52.92 ±13.81	13.13 ± 0.33	49.05 ± 0.93	66.30 ± 0.22
s-DRAE (ours)	47.97 ± 13.83	11.17 ±1.73	46.63 ±0.83	66.62 ± 0.51
ps-DRAE (ours)	49.15 ± 12.93	11.71 ±1.21	48.21 ±1.02	66.31 ± 0.43
mps-DRAE (ours)	44.67 ±9.98	11.01 ±1.32	46.61 ± 1.01	66.23 ± 0.56
ms-DRAE (ours)	**43.57 ± 10.98**	**11.12 ±0.91**	**46.01 ±0.91**	65.91± 0.4

Computational Time

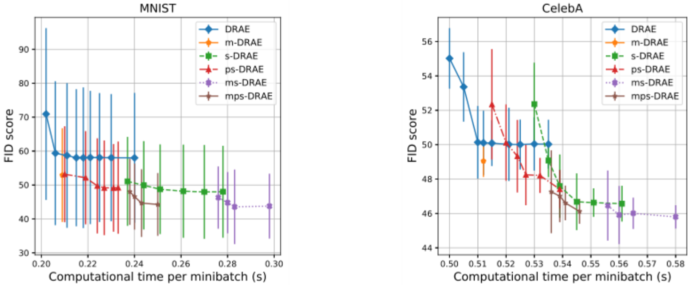


Figure 2: Each dot represents the computational time per minibatch and FID score. For DRAE, we vary the number of projections $L \in \{1, 5, 10, 20, 50, 100, 200, 500, 1000\}$; for s-DRAE we set $\kappa = 10$, $L \in \{1, 5, 10, 20, 50, 100\}$; for ps-DRAE we set $\kappa = 50, L \in \{1, 5, 10, 20, 50, 100\}$; and for m(p)s-DRAE we set $L = 50$, the number of vMF components $k \in \{2, 5, 10, 50\}$ (for each k, we find the best $\kappa \in \{1, 5, 10, 50, 100\}$).