딥러닝의 통계적 이해 프로젝트 결과 보고서: Causal Representation Learning via WAE

노요한* 이현종* 정승필†

December 21, 2022

1 개요

Variational autoEncoder(VAE, [1]) 기반의 생성 모델에서 disentangled representation은 생성된 표본의 품질과 downstream task로의 활용성 측면에서 많은 장점을 가지고 있다고 알려져있다. 이때 특성 사이에 관계성을 고려한 잠재변수를 학습하는 과정을 causal representation learning이라 한다. [2]에서 VAE 구조를 활용하여 causal representation을 찾는 방법, CausalVAE를 제안했다. CelebA 데이터셋을 활용하여 4가지 특성(gender, smile, eyes open, mouth open)을 나타내고 특성 사이의 인과관계도 가지고 있는 잠재변수를 학습하였다. 이후 각 특성을 조절하였을 때 특성 사이의 관계를 기반으로 얼굴 사진이 생성됨을 보였다.

Wasserstein autoencoder(WAE, [3])는 VAE 구조를 활용하여 데이터 분포와 생성된 데이터의 분포 사이의 Wasserstein distance를 최소화하는 것을 목표로 하는 생성 모델이고, 일반적으로 VAE보다 더 선명한 사진을 생성해낸다고 알려져 있다. 그리고 VAE는 우도함수(likelihood)를 기반으로 하는 반면 WAE는 데이터가 특이 분포(singular distribution)을 가지는 경우에도 적용될 수 있다. 이와 같은 성질로부터 우리는 WAE를 활용하면 더욱 좋은 결과를 얻을 수 있을 것으로 기대하였다. 언급한 두 가지 방법으로 다양한 데이터(pendulum, flow, CelebA)에 대해 개입(intervention) 결과를 재현해 보았다.

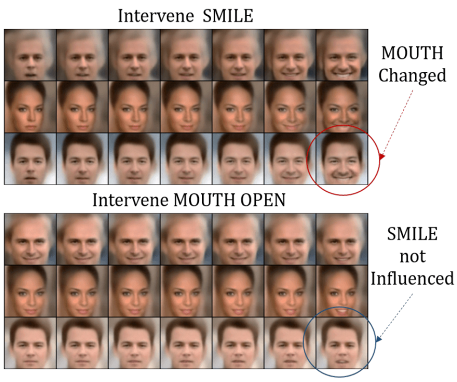

Figure 1: [2, Figure 4]. 개입 결과. 위쪽 그림에서 smile과 관련된 잠재변수를 변화시킬 때 입도 벌어지게 된다. 아래쪽 그림에서 mouth open과 관련된 잠재변수를 변화시킬 때 웃게 되지는 않는 경우가 존재한다.

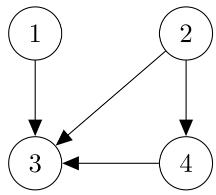

Figure 2: Ground truth Causal Graph on CelebA
1:Gender, 2:Smile, 3:Eyes Open, 4:Mouth Open

*Department of Statistics, SNU
†Graduate School of Public Health, SNU

2 CausalVAE

먼저, CausalVAE의 학습 과정을 살펴보면 [2]는 잠재표현 Z에 인과 관계를 반영하기 위해 VAE 구조를 활용하여 인과 표현 학습을 했다. d개의 특성을 나타내는 인과 표현 $\mathbf{Z} \in \mathbb{R}^d$에 대해 선형 structured causal model(SCM) 모형을 따르는 인과 관계를 고려하자.

$$\mathbf{z} = \mathbf{A}^T\mathbf{z} + \epsilon = (I - \mathbf{A}^T)^{-1}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{1}$$

여기서, $\mathbf{A} \in \mathbb{R}^{d \times d}$는 특성 사이의 directed acyclic graph(DAG) 구조를 나타내는 가중치 인접행렬을 나타낸다. 식 1의 구조방정식 모형을 따르는 인과 표현 Z를 학습하기 위해 Figure 3과 같은 생성 모형을 제시하였다.

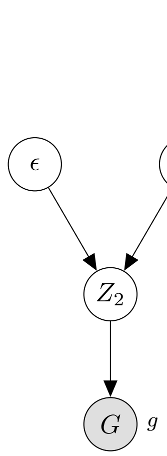

Figure 3: CausalVAE Decoder Structure

생성 과정은 아래와 같다. $\mathbf{u}$는 성별, 나이와 같이 실제 얼굴 사진의 특성을 나타내는 범주형 변수이고 $\mathbf{z}$의 각 차원이 특성을 나타내도록 $z_{1i}|u_i$ 조건부 분포를 $\mathcal{N}(\lambda_1(u_i), \lambda_2^2(u_i))$와 같이 가정한다. λ_i는 임의의 함수를 나타낸다.

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad z_{1i}|u_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\lambda_1(u_i), \lambda_2^2(u_i)), \ i = 1, \ldots, d$$
$$\mathbf{z}_2 = \mathbf{A}^T\mathbf{z}_1 + \epsilon \overset{d}{=} \mathbf{z}_1, \quad \mathbf{x} \sim p^{\theta}_{\mathbf{x}|\mathbf{z}_2}$$

[2, Theorem 1]에 의하면 $\mathbf{u}$가 모두 관측된 지도학습의 경우 모델의 identifiability가 보장된다. VAE는 log likelihood를 최대화하는 방향으로 학습이 진행되는데 이를 위해 인코더 또는 variational posterior($q^{\phi}_{\mathbf{z},\epsilon|\mathbf{x},\mathbf{u}} = \delta(\mathbf{z} = (\mathbf{I} - \mathbf{A}^T)^{-1}\epsilon) q^{\phi}_{\epsilon|\mathbf{x},\mathbf{u}}$)를 도입하여 ELBO를 정리하면 아래와 같다.

$$\begin{aligned}\sum_{i=1}^{N} \log p(\mathbf{x}_i|\mathbf{u}_i) \geq \text{ELBO} = \sum_{i=1}^{N} \Big\{ & \mathbb{E}_{Z^{\phi} \sim q^{\phi}_{\mathbf{z}|\mathbf{x}_i,\mathbf{u}_i}} [\log p^{\theta}_{X|Z}(\mathbf{X}|Z^{\phi})] \\ & - \mathcal{D}(q^{\phi}_{\epsilon|\mathbf{x}_i,\mathbf{u}_i}(\cdot|\mathbf{x}_i, \mathbf{u}_i) || p_{\epsilon}(\cdot)) \\ & - \mathcal{D}(q^{\phi}_{\mathbf{z}|\mathbf{x}_i,\mathbf{u}_i}(\cdot|\mathbf{x}_i, \mathbf{u}_i) || p^{\theta}_{\mathbf{z}|\mathbf{u}_i}(\cdot|\mathbf{u}_i)) \Big\}\end{aligned}$$

[2]에서는 deterministic한 인코더를 고려하였고, 식 1을 이용하여 잠재표현 $\mathbf{z}$를 $(\mathbf{I} - \mathbf{A}^T)^{-1}\epsilon$로 계산하였다. 여기서 $\mathbf{A}$는 학습되는 파라미터로 활용되었는데 실제 데이터의 특성 $\mathbf{u}$를 반영하고, DAG의

인접행렬이 되도록하는 제약 조건이 추가된다.

$$\sum_{i=1}^{N} \|\mathbf{u}_i - \sigma(\mathbf{A}^T \mathbf{u}_i)\|_2^2 \le \kappa_1 \tag{2}$$

$$tr\left[\left(\mathbf{I} + \frac{c}{d}\mathbf{A} \circ \mathbf{A}\right)^d\right] - d = 0 \tag{3}$$

$$\sum_{i=1}^{N} \mathbf{E}_{Z^\phi \sim q^\phi_{\mathbf{z}|\mathbf{x}_i,\mathbf{u}_i}} \|Z^\phi - g(A^T Z^\phi)\|^2 \le \kappa_2 \tag{4}$$

σ는 시그모이드 함수를 나타내고, κ_1, κ_2는 충분히 작은 양수, c는 임의의 양수, $\circ$은 행렬 원소 별로 곱하는 연산을 나타낸다. 식식 2는 identifiability를 보장하고, 특성 사이의 관계를 $\mathbf{A}$가 나타내도록 하는 조건이고, 식 3는 DAG의 가중치 인접행렬이 가져야하는 조건, 식 4은 SCM 모형을 만족하도록 하는 조건을 의미한다.

정리하면, CausalVAE의 목표는 다음과 같다.

$$\begin{aligned} &\underset{\phi,\theta}{\text{minimize}} - \text{ELBO} \\ &\text{s.t.} \quad (2), (3), (4) \end{aligned}$$

$\mathbf{A}$가 관측된 데이터의 특성 $\mathbf{u}$의 인과관계를 따르도록, $\mathbf{z}$가 SCM를 만족하면서 $\mathbf{x}$와 유사한 분포를 생성하도록 학습이 진행된다. 정확히 손실함수는 라그랑즈 승수 $\gamma_i, i = 1, 2, 3$을 도입하여,

$$\begin{aligned} \underset{\phi,\theta}{\text{minimize}} - \text{ELBO} + \gamma_1 \sum_{i=1}^{N} \|\mathbf{u}_i - \sigma(\mathbf{A}^T \mathbf{u}_i)\|_2^2 + \gamma_2 \left(tr\left[\left(\mathbf{I} + \frac{c}{d}\mathbf{A} \circ \mathbf{A}\right)^d\right] - d\right) \\ + \gamma_3 \sum_{i=1}^{N} \mathbf{E}_{Z^\phi \sim q^\phi_{\mathbf{z}|\mathbf{x}_i,\mathbf{u}_i}} \|Z^\phi - g(A^T Z^\phi)\|^2 \end{aligned}$$

로 나타낼 수 있고 β-VAE와 유사하게 추가로 ELBO의 쿨백-라이블러 발산 항의 계수를 하이퍼파라미터로 조절한다.

3 CausalWAE

3.1 Wasserstein Autoencoder(WAE)

$\mathbb{R}^D$ 상의 확률변수 X의 분포를 P_X, $\mathbb{R}^d$ 상의 잠재변수 Z의 분포를 P_Z라 하자. ($d << D$) 데이터를 함수(decoder) g를 이용하여 $G = g(Z)$로 생성한다고 하자. 이때, WAE는 두 분포 P_X와 $g_\sharp P_Z$ 사이의 Wasserstein distance를 g에 대해 최소화한다. p-Wasserstein distance는 아래와 같이 정의된다.

$$W_p^p(P_X, P_G) := \inf_{\Gamma \sim \mathcal{P}(X \sim P_X, Y \sim P_G)} \mathbb{E}_{(X,Y)\sim\Gamma}[d^p(X, Y)] \tag{5}$$

Γ는 P_X와 P_G를 marginal distribution으로 가지는 (X, Y)의 joint distribution을 의미하고 d는 X의 공간 상의 metric이다. 2-Wasserstein distance에 대해 다음 정리가 성립한다 [4, Theorem A.5].

Theorem 1. *Let $d(x, y) = \|x - y\|_2$ for $x, y \in \mathcal{X}$. If P_X has a density with respect to the Lebesgue measure, and the measurable function $g : \mathcal{Z} \to \mathcal{X}$ is injective, then*

$$W_2^2(P_X, g_\sharp P_Z) = \inf_{f \in \mathcal{Q}} \mathbb{E}_{P_X} d^2(X, g(f(X))), \tag{6}$$

where $\mathcal{Q}$ is the set of all measurable functions from $\mathcal{X}$ to $\mathcal{Z}$ such that $f_\sharp P_X = P_Z$.

Remark 1. *[4, Theorem A.5] 증명에서 $(\tilde{g} \circ g)_\sharp P_Z(F) = P_Z(\tilde{g}^{-1}(g^{-1}(F))$와 같이 right inverse를 잘못 계산하여 g가 injective라는 가정이 필요하다. 증명은 Appendix에서 확인할 수 있다.*

Theorem 1([4, Theorem A.5])를 보면, injective decoder가 있을 때 Wasserstein distance를 (6)와 같이 나타낼 수 있고 WAE의 목적함수가

$$D_{WAE}(P_X, g_\sharp P_Z) := \inf_{f\in\mathcal{Q}} \mathbb{E}_{P_X} d^2(X, g(f(X))) + \lambda\mathcal{D}(f_\sharp P_X \| P_Z) \tag{7}$$

와 같이 정의된다. f는 인코더로 생각할 수 있고 디코더 g와 함께 neural network로 매개화한다. Theorem 1은 디코더가 주어졌을 때 데이터 분포 P_X와 디코더로 생성되는 분포 $g_\sharp P_Z$ 사이의 Wasserstein distance를 인코더를 도입하여 간단히 정리해준다. 앞서 2절에서 CausalVAE의 디코더 구조를 살펴봤는데 theorem 1을 통해 WAE 프레임 워크에 적용해보았다.

3.2 CausalWAE: Theorem 1 적용

단사함수(injective function)인 디코더를 생각하기 위해 $\mathbf{u}$를 도입하자. Figure 3으로부터 디코더 $g : \mathcal{Z}\times\mathcal{U} \to \mathcal{X}$를 다음과 같이 정의하자. CausalVAE에서 특성 $\mathbf{u}$에 대한 잠재표현 $\mathbf{z}$의 조건부 분포를 가정하고 SCM 조건($\mathbf{z} = (\mathbf{I} - \mathbf{A}^T)^{-1}\epsilon$)을 만족하도록 잠재표현을 정의한 점에서 착안하여 CausalWAE에서의 잠재표현 $\mathbf{z}_1$을 아래와 같이 정의하였다.

$$\mathbf{z}_1 = (\mathbf{I} - \mathbf{A}^T)^{-1}\epsilon + g_1(\mathbf{u})$$
$$g(\boldsymbol{\epsilon}, \mathbf{u}) = g_3(g_2(\mathbf{A}^T\mathbf{z}_1) + \boldsymbol{\epsilon})$$

여기서 $g_1 : \mathcal{U} \to \mathcal{Z}$은 임의의 함수, $g_3 : \mathcal{Z} \to \mathcal{X}$는 neural network를 나타내고, $g_2 : \mathcal{Z} \to \mathcal{Z}$는 mask layer[5]이다. 특히 g_3는 leakyReLU 또는 sigmoid와 같이 단사함수인 활성화 함수를 사용한 neural network를 가정한다. 디코더 g를 확장시켜 $\tilde{g} : \mathcal{Z}\times\mathcal{U} \to \mathcal{X}\times\mathcal{U}$, $\tilde{g}(\boldsymbol{\epsilon}, \mathbf{u}) = (g(\boldsymbol{\epsilon}, \mathbf{u}), \mathbf{u})$를 생각하면, g_3가 단사함수일 때, $\tilde{g}$는 단사함수가 된다.

$\tilde{g}$에 대해 Theorem 1을 적용하기 위해 $\mathcal{U}$ 공간에서의 거리를 d'으로 두고, $\mathcal{X}\times\mathcal{U}$ 공간에서의 거리를 $\tilde{d} = \sqrt{d^2 + d'^2}$로 정의하자. 이제 P_{XU}와 $P_{\boldsymbol{\epsilon}} \otimes P_U$, $\tilde{g}$에 대해 Theorem 1을 적용하면,

$$\begin{aligned}
W_2^2(P_{XU}, \tilde{g}_\sharp(P_{\boldsymbol{\epsilon}} \otimes P_U)) &= \inf_{\tilde{f}\in\tilde{\mathcal{F}}} \mathbb{E}_{P_{XU}} \tilde{d}^2\left(\begin{pmatrix} X \\ U \end{pmatrix}, \tilde{g}(\tilde{f}(X,U))\right) \\
&= \inf_{f\in\mathcal{F}} \mathbb{E}_{P_{XU}} \tilde{d}^2\left(\begin{pmatrix} X \\ U \end{pmatrix}, \begin{pmatrix} g_3(g_2(\mathbf{A}^T((\mathbf{I} - \mathbf{A}^T)^{-1}\epsilon + g_1(\mathbf{u}))) + f(X,U)) \\ U \end{pmatrix}\right) \\
&= \inf_{f\in\mathcal{F}} \mathbb{E}_{P_{XU}} d^2(X, g_3(g_2(\mathbf{A}^T((\mathbf{I} - \mathbf{A}^T)^{-1}\epsilon + g_1(\mathbf{u}))) + f(X,U))),
\end{aligned}$$

$\tilde{\mathcal{F}} = \{\tilde{f} : \mathcal{X}\times\mathcal{U} \to \mathcal{Z}\times\mathcal{U} | \tilde{f}_\sharp P_{XU} = P_{\boldsymbol{\epsilon}} \otimes P_U)\}$, $\mathcal{F} = \{f : \mathcal{X}\times\mathcal{U} \to \mathcal{Z} | (f, \Pi_U)_\sharp P_{XU} = P_{\boldsymbol{\epsilon}} \otimes P_U\}$, $\Pi_U(X,U) = U$를 의미한다. 두 번째 등호는 $\tilde{f}(X,U) = (f(X,U), \Pi_U(X,U))$를 생각하면 성립함을 알 수 있다. $\mathcal{F}$의 조건은

$$f(X,U) \stackrel{d}{=} \boldsymbol{\epsilon}, \quad f(X,U) \perp\!\!\!\perp U,$$

2가지 조건으로 생각할 수 있다.

두 분포가 얼마나 다른지 측정하는 $\mathcal{D}$와 독립성을 측정하는 $\mathcal{H}$에 대해 위의 2가지 조건을 패널티 항으로 추가하여 WAE objective를 정리하면,

$$\min_g \min_f \mathbb{E}_{P_{XU}} d^2(X, g_3(g_2(\mathbf{A}^T g_1(U)) + f(X,U))) + \lambda_1\mathcal{D}(f_\sharp P_{XU} \| P_{\boldsymbol{\epsilon}})) + \lambda_2\mathcal{H}(f(X,U), U). \tag{8}$$

예를 들어, $\mathcal{D}$는 [3]에서처럼 MMD 또는 GAN loss를 사용할 수 있고, $\mathcal{H}$는 HSIC(Hilbert-Schmidt Independence Criterion) [6]를 사용할 수 있다.

Appendix

Proof of Theorem 1

Proof. Let $P_G = g_\sharp P_Z$. Under the conditions of the theorem, the Monge-Kantorovich equivalence holds:

$$W_2^2(P_X, P_G) = \inf_{T:\mathcal{X}\to\mathcal{X}: T_\sharp P_X = P_G} \mathbb{E}_{P_X} d^2(X, T(X)).$$

Hence it suffices to show that

$$\inf_{f:\mathcal{X}\to\mathcal{Z}:f_\sharp P_X=P_Z}\int_{\mathcal{X}} d^2(x,g(f(x)))dP_X = \inf_{T:\mathcal{X}\to\mathcal{X}:T_\sharp P_X=P_G}\int_{\mathcal{X}} d^2(x,T(x))dP_X$$

or equivalently

$$\{g\circ f : f:\mathcal{X}\to\mathcal{Z}, f_\sharp P_X = P_Z\} = \{T:\mathcal{X}\to\mathcal{X} : T_\sharp P_X = P_G\}.$$

First,

$$\{g\circ f : f:\mathcal{X}\to\mathcal{Z}, f_\sharp P_X = P_Z\} \subset \{T:\mathcal{X}\to\mathcal{X} : T_\sharp P_X = P_G\}$$

since for any measurable $f:\mathcal{X}\to\mathcal{Z}$ such that $f_\sharp P_X = P_X f^{-1} = P_Z$ we have $g\circ f:\mathcal{X}\to\mathcal{X}$ measurable and for any Borel set $E\subset\mathcal{X}$

$$(g\circ f)_\sharp P_X(E) = P_X(g\circ f)^{-1}(E) = P_X(f^{-1}\circ g^{-1})(E) = P_Z(g^{-1}(E)) = g_\sharp P_Z(E) = P_G(E).$$

Second, suppose $T:\mathcal{X}\to\mathcal{X}$ is measurable and satisfies $T_\sharp P_X = P_G$. There exists a set $A\subset\mathcal{X}$ with $P_X(A)=1$ such that $g:\mathcal{Z}\to T(A)$ is surjective. Otherwise, there exists $B\subset\mathcal{X}$ with $P_X(B)>0$ and $\tilde{g}^{-1}(T(B))=\emptyset$, and hence $0=\tilde{g}_\sharp P_Z(T(B)) = T_\sharp P_X(T(B)) = P_X(B)>0$ that is a contradiction. In addition, since $g:\mathcal{Z}\to\mathcal{X}$ is injective, there exists a left inverse $g^\dagger:\mathcal{X}\to\mathcal{Z}$. Note that $g^\dagger|_{T(A)}:T(A)\to\mathcal{Z}$ is an inverse function of $g:\mathcal{Z}\to T(A)$. Let $f=g^\dagger\circ T$. Then $g\circ f = g\circ g^\dagger\circ T = T$ almost surely in P_X and also for any Borel set $F\subset\mathcal{Z}$

$$\begin{aligned}
f_\sharp P_X(F) &= P_X(g^\dagger\circ T)^{-1}(F)\\
&= P_X(T^{-1}(g^\dagger)^{-1}(F))\\
&= P_G((g^\dagger)^{-1}(F))\\
&= P_Z(g^{-1}((g^\dagger)^{-1}(F)))\\
&= P_Z((g^\dagger\circ g)^{-1}(F)) = P_Z(F).
\end{aligned}$$

Therefore,

$$\{g\circ f : f:\mathcal{X}\to\mathcal{Z}, f_\sharp P_X = P_Z\} \supset \{T:\mathcal{X}\to\mathcal{X} : T_\sharp P_X = P_G\}$$

□

References

[1] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[2] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9593–9602, 2021.

[3] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

[4] Giorgio Patrini, Rianne van den Berg, Patrick Forre, Marcello Carioni, Samarth Bhargav, Max Welling, Tim Genewein, and Frank Nielsen. Sinkhorn autoencoders. In *Uncertainty in Artificial Intelligence*, pages 733–743. PMLR, 2020.

[5] Ignavier Ng, Shengyu Zhu, Zhuangyan Fang, Haoyang Li, Zhitang Chen, and Jun Wang. Masked gradient-based causal structure learning. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 424–432. SIAM, 2022.

[6] Romain Lopez, Jeffrey Regier, Michael I Jordan, and Nir Yosef. Information constraints on auto-encoding variational bayes. *Advances in neural information processing systems*, 31, 2018.