VAE の理解のための短いノート

林 祐輔

2019 年 6 月 10 日

概要

変分オートエンコーダー（VAE）の目的関数 $\mathcal{L}_{\mathrm{VAE-ELBO}}$ にあらわれる正則化項 $D_{\mathrm{KL}}\left(q_\phi\left(\psi|x\right)\|p\left(\psi\right)\right)$ が，入力データを無視した潜在表現の学習を促す「有害な正則化項」であることを紹介し，問題を回避する「無害な正則化項」を導入する方法を紹介する．

1　生成モデル

1.1　変分オートエンコーダー（VAE）

変分オートエンコーダー（VAE）は，入力データ x を潜在変数 ψ に変換するエンコーダー $q_\phi\left(\psi|x\right)$ と，潜在変数 ψ から入力データ x を推測するデコーダー $p_\varphi\left(x|\psi\right)$ によって構成される．VAE は，エンコーダーとデコーダー，それぞれの同時分布

$$q_\phi\left(x,\psi\right)=q\left(x\right)q_\phi\left(\psi|x\right). \tag{1}$$
$$p_\varphi\left(x,\psi\right)=p\left(\psi\right)p_\varphi\left(x|\psi\right). \tag{2}$$

が一致する方向にモデルのパラメータ $\{\phi,\varphi\}$ を最適化することで，入力データの生成過程をモデリングする＜具体的には，パラメータ ϕ で定まるニューラルネット（エンコーダー）が，入力データを潜在変数に符号化し，パラメータ φ で定まるニューラルネット（デコーダー）が，潜在変数から入力データを復号する＞．

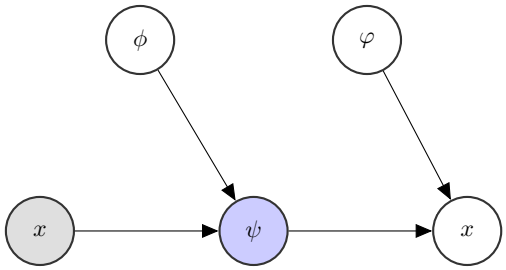

図 1　VAE のグラフィカルモデル

VAE の目的関数 $\mathcal{L}_{\mathrm{VAE-ELBO}}$ は，次のようにして導くことができる．

ELBO の導出法 1 （Jensen の不等式）：

$$\begin{aligned}
\mathbb{E}_{q(x)}\left[\log p_\varphi(x)\right] &= \mathbb{E}_{q(x)}\left[\log \int p_\varphi(x,\psi)\, d\psi\right] \\
&= \mathbb{E}_{q(x)}\left[\log \int q_\phi(\psi|x)\, \frac{p_\varphi(x,\psi)}{q_\phi(\psi|x)} d\psi\right] \\
&\geq \mathbb{E}_{q(x)}\left[\int q_\phi(\psi|x) \log \frac{p_\varphi(x,\psi)}{q_\phi(\psi|x)} d\psi\right] \\
&= \mathbb{E}_{q_\phi(x,\psi)}\left[\log p_\varphi(x|\psi)\right] - \mathbb{E}_{q(x)}\left[D_{\mathrm{KL}}\left(q_\phi(\psi|x) \,\|\, p(\psi)\right)\right]. \\
&=: \mathcal{L}_{\mathrm{VAE-ELBO}}
\end{aligned} \tag{3}$$

また，Eq. 3 は Jensen の不等式を使わなくても導くことができる．

ELBO の導出法 2 （同時分布の KL-ダイバージェンス）：

$$\begin{aligned}
-D_{\mathrm{KL}}\left(q_\phi(x,\psi) \,\|\, p_\varphi(x,\psi)\right) &= -\mathbb{E}_{q_\phi(x,\psi)}\left[\log \frac{q_\phi(x,\psi)}{p_\varphi(x,\psi)}\right] \\
&= H_{q_\phi}(x|\psi) + \mathbb{E}_{q_\phi(x,\psi)}\left[\log p_\varphi(x|\psi)\right] - \mathbb{E}_{q(x)}\left[D_{\mathrm{KL}}\left(q_\phi(\psi|x) \,\|\, p(\psi)\right)\right]. \\
&= H_{q_\phi}(x|\psi) + \mathcal{L}_{\mathrm{VAE-ELBO}}
\end{aligned} \tag{4}$$

ここで，Eq. 4 の第 1 項 $H_{q_\phi}(x|\psi)$ は定数になるため，目的関数の勾配には影響しないことに注意する（Appendix 参照）．ELBO の導出法 1, 2 に特別な優劣があるわけではないが，本稿では基本的に導出法 2 を使って ELBO を導くことにする．

さて，$\mathcal{L}_{\mathrm{VAE-ELBO}}$ の内訳をみると，まず第 1 項は，潜在変数 ψ から入力データ x をどれくらい正確に復号できるか，を測る指標（＝対数尤度関数，再構成損失）になっている．また，第 2 項は，潜在変数の学習に制限をかけるための正則化項だということがわかる．様々な先行研究 [Bowman et al., 2016, Snderby et al., 2016] から，潜在変数の正則化項 $D_{\mathrm{KL}}\left(q_\phi(\psi|x) \,\|\, p(\psi)\right) \geq 0$ は，入力データを無視した潜在変数の学習を促す「有害な正則化項」であると指摘されている．実際，この正則化項を相互情報量をつかって次のように書き直すと

$$\begin{aligned}
\mathcal{L}_{\mathrm{VAE-ELBO}} &= \mathbb{E}_{q_\phi(x,\psi)}\left[\log p_\varphi(x|\psi)\right] - \mathbb{E}_{q(x)}\left[D_{\mathrm{KL}}\left(q_\phi(\psi|x) \,\|\, p(\psi)\right)\right] \\
&= \mathbb{E}_{q_\phi(x,\psi)}\left[\log p_\varphi(x|\psi)\right] - I_{q_\phi}(x,\psi) - D_{\mathrm{KL}}\left(q_\phi(\psi) \,\|\, p(\psi)\right).
\end{aligned} \tag{5}$$

Eq. 5 より，$\mathcal{L}_{\mathrm{VAE-ELBO}}$ の最大化は，入力データと潜在変数の相互情報量 $I_{q_\phi}(x,\psi) \geq 0$（≒相関）を下げる方向に進むことがわかる（これは，$D_{\mathrm{KL}}\left(q_\phi(\psi|x) \,\|\, p(\psi)\right) = 0$ に近づくことで，$q_\phi(\psi|x) = p(\psi)$ が導かれ，エンコーダーが入力データを無視して潜在変数を生成するように学習が進むことを意味する）．

このような問題を解決するため，$\mathcal{L}_{\mathrm{VAE-ELBO}}$ に相互情報量を加えた新しい目的関数を導入する [Zhao et al., 2018, Zhao et al., 2018]．

$$\begin{aligned}
\mathcal{L}_{\mathrm{InfoVAE}} &=: \mathcal{L}_{\mathrm{VAE-ELBO}} + I_{q_\phi}(x,\psi) \\
&= \mathbb{E}_{q_\phi(x,\psi)}\left[\log p_\varphi(x|\psi)\right] - D_{\mathrm{KL}}\left(q_\phi(\psi) \,\|\, p(\psi)\right).
\end{aligned} \tag{6}$$

新しい目的関数を Eq. 6 のように設定してしまえば，学習が進んでも，潜在変数が入力データを無視して生成されることはなくなる．このように，潜在変数にかかる正則化項を新しい角度から見つめ直すことで，様々な VAE の派生モデルが提案されている．

例えば，最近では，潜在変数の各成分 $\psi = (\psi_1, \psi_2, \ldots, \psi_D)$ 毎の分布 $q_\phi(\psi_1), q_\phi(\psi_2), \ldots, q_\phi(\psi_D)$ を考え

$$\begin{aligned}
\mathcal{L}_{\mathrm{InfoVAE}} &= \mathbb{E}_{q_\phi(x,\psi)}\left[\log p_\varphi(x|\psi)\right] - D_{\mathrm{KL}}\left(q_\phi(\psi) \,\|\, p(\psi)\right) \\
&= \mathbb{E}_{q_\phi(x,\psi)}\left[\log p_\varphi(x|\psi)\right] - D_{\mathrm{KL}}\left(q_\phi(\psi) \,\|\, \prod_{i=1}^{D} q_\phi(\psi_i)\right) - \sum_{i=1}^{D} D_{\mathrm{KL}}\left(q_\phi(\psi_i) \,\|\, p(\psi_i)\right).
\end{aligned} \tag{7}$$

のように，Eq. 6 の正則化項をも分解するアプローチが提案されている．ここであらわれる Eq. 7 の第 2 項は Total correlation と呼ばれ，成分間 $\psi = (\psi_1, \psi_2, \ldots, \psi_D)$ の相関の強さをあらわす．Total correlation を下げる方向に学習を進めることで，潜在空間の各次元の意味をより解釈しやすいものにすることができる [Gao et al., 2019].

参考文献

[1] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In International Conference on Learning Representations (ICLR), 2014.

[2] Sachin Ravi and Alex Beatson. Amortized Bayesian Meta-Learning. In International Conference on Learning Representa-tions (ICLR), 2019

[3] Shengjia Zhao, Jiaming Song, Stefano Ermon. InfoVAE: Information Maximizing Variational Autoencoders.

[4] Tiancheng Zhao, Kyusong Lee, Maxine Eskenazi. Unsupervised Discrete Sentence Representation Learning for Interpretable Neural Dialog Generation. In proceedings of Association for Computational Linguistics (ACL), 2018.

[5] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, Samy Bengio. Generating sentences from a continuous space. In Proceedings of Conference on Computational Natural Language Learning (CoNLL), 2016.

[6] Casper Kaae Snderby, Tapani Raiko, Lars Maale, Sren Kaae Snderby, and Ole Winther. How to train deep variational autoencoders and probabilistic ladder networks. In Proceedings of the 33rd International Conference on Machine Learning (ICML), 2016.

[7] Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, Aram Galstyan. Auto-Encoding Total Correlation Explanation. In International Conference on Artificial Intelligence and Statistics (AISTATS), 2019.

2 Appendix

シャノン・エントロピー：

$$H_p(x) =: -\mathbb{E}_{p(x)}[\log p(x)] = -\int p(x) \log p(x) dx \tag{8}$$

$$H_{q_\phi}(x) =: -\mathbb{E}_{q_\phi(x)}[\log q_\phi(x)] = -\int q_\phi(x) \log q_\phi(x) dx \tag{9}$$

相互情報量：

$$I_p(x, y) =: \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dxdy \tag{10}$$

$$I_{q_\phi}(x, y) =: \int q_\phi(x, y) \log \frac{q_\phi(x, y)}{q_\phi(x)q_\phi(y)} dxdy \tag{11}$$