複数の潜在変数を用いた表現学習モデル MPVAE の提案

MPVAE: using multi latent variables for representation learning

石川 敬規 [1], 妹尾 卓磨 [2], 今井 倫太 [3]

[1] 慶應義塾大学 理工学部 情報工学科
[2] 慶應義塾大学院 理工学研究科 開放環境科学専攻
[3] 慶應義塾大学 理工学部

Abstract: 表現学習はデータの表現を決定づける潜在表現を獲得し、その表現は分類などの教師あり学習の性能を向上させる。本論文では、表現学習の新たなフレームワークである Multi Path VAE(MPVAE) を提案する。既存の VAE を用いた手法では 1 組のエンコーダとデコーダから潜在表現を獲得していたのに対し，MPVAE では複数の異なる層のエンコーダとデコーダから潜在表現を獲得する．実験の結果、獲得した表現で識別タスクの精度を向上できた。

1 導入

表現学習とは, データをそのままの表現ではなく, データが持つ特徴量を潜在変数として抽出することである. 表現学習により獲得された潜在変数は分類や回帰などの学習に使用される [1].

表現学習における良い表現とは, 潜在変数が教師なし学習や半教師あり学習に有用な表現になっていることである [2]. 人手による特徴エンジニアリング [3] と同じように深層学習による表現学習 [4] はデータの性質に大きく影響される. そのため, 全ての潜在的要因を捉え, 原因因子を紐解くような表現の獲得は困難である.

教師なし学習を利用して, 広範なタスクに適用できる表現を学習することが提案されている [2]. 良い表現の獲得のために, データ拡張や補助タスクの利用などの手法が提案されてきた. また, ディスエンタングルな表現は良い表現の獲得に役に立つと主張されている [2]. ディスエンタングルな表現とは, 1 つの潜在変数が 1 つの生成因子の変化に影響を受けるが, 他の生成因子の変化には不変である表現として定義される [2]. たとえば, 顔のデータセットで訓練されたモデルの場合, それぞれの潜在変数に対して, 顔の表情, 目の色, 髪型, 髪色, 眼鏡の有無などの単一の独立な生成因子を学習することが可能である. そのため, ディスエンタングルな表現は解釈可能性が高い. 顔認識や物体認識などのデータの属性に関する特徴を必要とするタスクにおいて, ディスエンタングルな表現を使用することで精度を向上させている [5]. Variational Autoencoder(VAE)[6, 7] や generative adversarial network(GAN)[8] を拡張したディスエンタングルな表現を獲得するためのモデルが提案されている [9, 10].

β-VAE[10] は, 画像に関連するディスエンタングルな表現の教師なし学習モデルである. β-VAE は,VAE の最適化関数にハイパーパラメータ β を追加する. これにより, 潜在変数の学習に正則化が行われ, ディスエンタングルな表現を獲得しやすくなる. β-VAE などの VAE に基づいた手法などは, 単一の層から獲得された潜在変数のみが扱われる. コンピュータビジョンの分野である領域や物体の認識においては, 異なる階層から複数の特徴量を利用することにより, 大きな成果を上げている [11, 12]. これは深層学習において, 層の深さにより獲得される表現が異なることを利用している [13].

本論文において, 異なる深さの層から潜在変数を獲得できるように複数のエンコーダを持つように変更を加えた VAE,Multi Path VAE(MPVAE) を提案する. この変更により,MPVAE は異なる深さの層から独立な潜在変数を獲得し, 潜在的に抽象度の異なる表現を獲得する. また, 本論文ではそれぞれの潜在変数がディスエンタングルな表現を獲得するように β-VAE に倣い最適化関数に正則化項を加えるように修正を加えた. MPVAE で獲得された潜在変数を特徴量として, クラス分類タスクを解いた. その結果,VAE により潜在変数の特徴量よりも精度を向上させることができた.

本論文の章の構成を記しておく. 次章では, 本論文に関連する背景技術について言及する. 次に, 提案手法とその実装について述べる. 最後に, 実験では提案手法と既存の手法を比較する.

2 背景

2.1 VAE

VAE[6, 7] は, エンコーダーとデコーダを対にした生成モデルである. 限界対数尤度に関して最尤推定を直

接実行するかわりに, 変分下限 (evidence lower bound) を最適化をすることによって学習を行う. 真の潜在変数 z をパラメータに持つ分布からサンプルされたデータセット x があると仮定する. そのような生成過程におけるデータの周辺尤度を学習することを目的とする.

$$\max_{\phi,\theta} \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] \tag{1}$$

ϕ と θ は VAE のエンコーダおよびデコーダのパラメータである. これは次のように書き直せる.

$$\begin{aligned} \log p_\theta(x|z) &= D_{KL}(q(z|x)||p(z)) \\ &+ \mathcal{L}(\theta, \phi; x, z) \end{aligned} \tag{2}$$

$L(\theta, \phi; x, z)$ を最大化することは, 式 (2) の下限を最大化することに等しい.

$$\begin{aligned} \log p_\theta(x|z) &\geq \mathcal{L}(\theta, \phi; x, z) \\ &= \mathrm{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] \\ &- D_{KL}(q(z|x)||p(z)) \end{aligned} \tag{3}$$

2.2 β-VAE

β-VAE は,VAE の拡張である. ハイパーパラメータ β を VAE の目的関数に導入する.

$$\begin{aligned} \mathcal{L}(\theta, \phi; x, z, \beta) &= \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] \\ &- \beta D_{KL}(q_\phi(z|x)||p(z)) \end{aligned} \tag{4}$$

上手く選択された β の値は, 通常の VAE よりディスエンタングルな潜在変数 z をもたらす. β-VAE の目的関数から導出されるガウス事前分布 $p(z)$ に加える β による制約は, 潜在的なボトルネックに追加の制約を与えることになる. その制約は, データを再構築するのに十分である状態で, 潜在変数をディスエンタングルになるように分解する. ディスエンタングルな表現を獲得するために大きな β を選択することは, 再構成とのトレードオフとなる.

2.3 異なる階層から獲得される表現

FCN[14] や Hypercolumns[13] は, 深層学習が異なる層からの特徴量は異なる抽象度の表現を獲得していることを利用している. 浅い層の畳み込みでは, エッジ, コーナーなどの低レベルの空間的視覚情報を符号化し, 深い層からの畳み込みでは, オブジェクトまたはカテゴリを表現するような, 高レベルの意味情報を符号化する.

3 Multi Path VAE(MPVAE)

3.1 MPVAE の概要

本研究では, 独立な複数の潜在変数を用いて表現学習を行う Multi Path VAE(MPVAE) を提案する. MPVAE のモデルを図 3.1 に示す. MPVAE は,VAE を独立な複数の潜在変数を持つように拡張したモデルであり, エンコーダからの出力とデコーダへの入力が複数存在する. エンコーダは複数存在するがデコーダは,1 つである. 独立な潜在変数を複数持つことにより, 各々にある程度特有の潜在的な表現を学習することが可能になる.

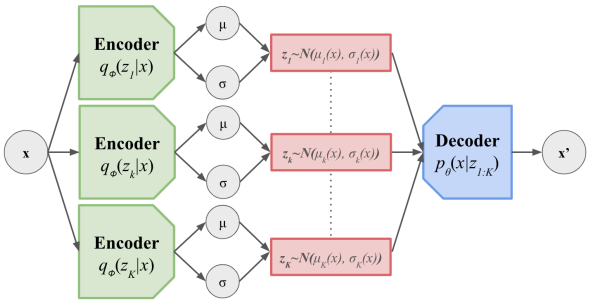

図 1: MPVAE

3.2 モデルの構成

エンコーダ側では, 畳み込み層や全結合層と活性化関層の組み合わせが複数層続き, 潜在変数が獲得される. デコーダ側では, 異なる潜在変数が結合層でまとめられ, 逆畳み込み層や全結合層と活性化層の組み合わせが複数層続き, 元の入力データが再構成される.

3.3 損失関数の修正

VAE の損失関数は,2 章の (3) である. 提案手法では, 複数の潜在変数を用いる. そのため, 次のように修正する.

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_{q_\phi(z_1,...,z_n|x)}[\log p_\theta(x|z_1, ..., z_n)] \\ &- \sum_{k=1}^{n} D_{KL}(q(z_k|x)||p(z_k)) \end{aligned} \tag{5}$$

第 1 項は, 入力と出力による再構成誤差である. 第 2 項は,n 個の潜在変数それぞれに対するカルバック・ライブラーダイバージェンスの和である. また,z_n は

$z_{n-1}(n \geq 2)$ よりも深い層から獲得される潜在変数を表す.

本論文では, 潜在変数が 2 つの場合を扱う. また, 潜在変数がディスエンタングルな特徴を獲得しやすくなるように β-VAE に倣い正則化項 β を加える. この制約は, カルバック・ライブラーダイバージェンスへの正則化項である. そのため, 最終的な損失関数は以下のように定義される.

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_{q_\phi(z_1,z_2|x)}[\log p_\theta(x|z_1,z_2)] \\ &- \sum_{k=1}^{2} \beta_k D_{KL}(q(z_k|x)||p(z_k)) \end{aligned} \tag{6}$$

3.4 モデルの特徴

MPVAE では, 潜在変数への経路が複数存在する. 複数の経路が存在することにより, それぞれから互いに潜在的に表現力の異なる潜在変数を獲得することができる. 層の深さを変えることにより, 浅い層は大域的な変化を表現する潜在変数を獲得し, 深い層はより意味のある高次元の情報が獲得することができる.

4 実装

MPVAE のエンコーダは, カーネルサイズ 4 × 4, ストライド 2, パディング 1 の畳み込み層 4 層の後に, 全結合層 1 層とカーネルサイズ 4 × 4, ストライド 2, パディング 1 畳み込み層と全結合層 1 層を合わせた 1 層がそれぞれ続く. MPVAE のデコーダは, 全結合層 1 層と全結合層 1 層と逆畳み込み層 1 層を合わせた 1 層がそれぞれ続き, 連結層 1 層によりチャネル方向に結合される. その後は, カーネルサイズ 3 × 3, ストライド 1, パディング 1 の畳み込み層 2 層の後に, カーネルサイズ 4 × 4, ストライド 2, パディング 1 の逆畳み込み層 4 層が続く. 活性化関数には, 全て ReLU を用いる.

5 実験

5.1 定量的な評価

VAE と MPVAE によってそれぞれ獲得された潜在変数を使用して, クラス分類器を学習することにより潜在変数の特徴量としての有用性を評価した. 1 つのクラス分類器につき,5 つの訓練データと検証データの組み合わせと 5 つの乱数シードの 25 通りを学習して平均を計算した. 潜在変数の獲得には,CelebA では,32 次元の VAE と 8 次元と 24 次元の MPVAE を用いた. また,dSprites では,12 次元の VAE と 3 次元と 9 次元の MPVAE を用いた.

5.1.1 CelebA

アノテーションデータに含まれている 40 種類のラベルについて, それぞれのクラスを予測する分類器を学習した. MPVAE・VAE 共に潜在変数の次元数は 32 である. MPVAE は, 浅い層から獲得された潜在変数が 8 次元・深い層から獲得された潜在変数が 24 次元である. 表 1 に VAE と MPVAE の全ラベルに対する精度の平均を示す. MPVAE の精度は,VAE の制度よりも 1%程高くなった.

表 1: VAE と MPVAE の全ラベルに対する精度の平均.

	Accuracy
VAE	82.02%
MPVAE(z1)	81.87%
MPVAE(z2)	82.09%
MPVAE(z1+z2)	83.02%

5.2 dSprites

アノテーションデータにある潜在クラスの形, スケール, 角度, 位置 x・y の 5 種類のラベルについて分類モデルを学習した. 表 5.2 に各クラスに対する平均の ACC を示す. β-MPVAE は, 全てのクラスで β-VAE を上回る精度を示した. 異なる階層の潜在変数を利用することにより, 潜在的に抽象度の異なる表現を学習できているためであると考えられる.

6 結論

本論文では, 階層的な特徴量を用いた VAE である MPVAE を提案した. VAE の階層性を持たせることにより, 特徴量としての有用性を向上させることができた. 本論文では, 階層性の有用性を特徴量という形で評価した. 将来研究として, 階層間の特徴量のディスエンタングル度を評価することが考えられる.

表 2: β-VAE と β-MPVAE の各ラベルに対する ACC の平均と全クラスラベルに対する ACC の平均.

	shape	scale	orientation	posX	posY	mean
β-VAE	55.19%	46.97%	7.37%	47.49%	38.52%	39.11%
β-MPVAE(z1=3)	40.75%	22.40%	3.37%	5.54%	7.24%	15.86%
β-MPVAE(z2=9)	60.70%	46.73%	**8.91**.%	72.07.%	69.73%	51.63%
β-MPVAE(z1+z2)	**63.10**%	**47.02**%	8.86%	**72.37**%	**69.75**%	**52.22**%

参考文献

[1] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pp. 3581–3589, 2014.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, Vol. 35, No. 8, pp. 1798–1828, 2013.

[3] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, Vol. 55, No. 10, pp. 78–87, 2012.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[5] Jimei Yang, Scott E Reed, Ming-Hsuan Yang, and Honglak Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *Advances in Neural Information Processing Systems*, pp. 1099–1107, 2015.

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

[7] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[9] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pp. 2172–2180, 2016.

[10] Matthey Higgins, Burgess Pal, Botvinick Glorot, and Lerchner Mohamed. β-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Representaion Learning*, 2017.

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, Vol. 1, p. 3, 2017.

[12] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, Vol. 40, No. 4, pp. 834–848, 2018.

[13] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 447–456, 2015.

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.