情報処理学会論文誌 Vol.59 No.3 859–873 (Mar. 2018)

異なるモダリティ間の双方向生成のための深層生成モデル

鈴木 雅大[1,a)] 松尾 豊[1]

受付日 2017年6月11日, 採録日 2017年12月8日

概要：本稿では，異なる種類のモダリティ間を双方向に生成できる深層生成モデルについて研究する．双方向とは，たとえば画像から対応する文書を生成するだけでなく，文書から画像も生成できるということである．近年，variational autoencoder（VAE）のような深層生成モデルで異なるモダリティを扱う研究が行われている．しかし，これらは条件づけられた従属的な関係しかモデル化していないため，あるモダリティから別のモダリティに 1 方向しか生成できない．双方向で生成するためには，すべてのモダリティ間の高レベルな概念をとらえるような共有表現を抽出し，それを通じて複数のモダリティを双方向に生成する必要がある．本研究では，各モダリティが共有表現に独立に条件づけられた下での全モダリティの同時分布をモデル化した joint multimodal variational autoencoder（JMVAE）を提案する．一般的に，あるモダリティから別のモダリティを生成する際には，入力では生成先のモダリティは欠損させる必要がある．もし生成先のモダリティの次元が生成元のモダリティより大きい場合，推論した潜在変数や生成したモダリティが崩れてしまう可能性がある．本研究では，既存の欠損値補完の手法でも解決できないことを明らかにし，この問題を解決するために，JMVAE-kl と階層的 JMVAE という追加的な手法を提案する．実験から，これらの手法によって，欠損モダリティ問題が解決すること，すべてのモダリティを統合した適切な共有表現が獲得されること，従来の 1 方向しか生成できないモデルと比較して，同等以上の精度で双方向に生成できることを確認した．

キーワード：深層生成モデル，マルチモーダル学習

Deep Generative Models for Bi-directional Generation between Different Modalities

Masahiro Suzuki[1,a)] Yutaka Matsuo[1]

Received: June 11, 2017, Accepted: December 8, 2017

Abstract: We investigate deep generative models that can exchange multiple modalities bi-directionally, e.g., generating images from corresponding texts and vice versa. Recently, some studies handle multiple modalities on deep generative models such as variational autoencoders (VAEs). However, these models typically assume that modalities are forced to have a conditioned relation, i.e., we can only generate modalities in one direction. To achieve our objective, we should extract a joint representation that captures high-level concepts among all modalities and through which we can exchange them bi-directionally. As described herein, we propose a joint multimodal variational autoencoder (JMVAE), in which all modalities are independently conditioned on joint representation. In other words, it models a joint distribution of modalities. In general, when generating another modality from one modality, the modality which we want to generate must be missing on input. If the missing modality is high-dimensional is larger in dimension than other modalities, then the inferred latent variable and generated samples might be collapsed. We found that this issue cannot prevent even using the conventional missing value complementation. In this study, we introduce two independent methods, JMVAE-kl and hierarchical JMVAE, which can prevent this issue. Our experiments showed the following results: our models can solve the missing modality problem; we can obtain appropriate joint representations which contain all modalities by our models; and our models can generate multiple modalities bi-directionally as same or better than the conventional models which can generate only one direction.

Keywords: deep generative model, multimodal learning

1. はじめに

実世界では，情報は様々な種類で表現されている．たとえば，画像はピクセル情報で表現される一方で，タグ情報

[1] 東京大学大学院工学系研究科
Graduate School of Engineering, The University of Tokyo, Bunkyo, Tokyo 113–8654, Japan
[a)] masa@weblab.t.u-tokyo.ac.jp

でも表現される．このような異なる種類の情報は，それぞれ**モダリティ**と呼ばれる．我々人間はこうしたモダリティを双方向に変換することができる．たとえば「笑っている眼鏡をかけていない若い女性」で表される顔がどのようなものかを想像するだけではなく，対応する顔写真に対して，このキャプションを追加することもできる．本研究の目標は，人間のように，異なるモダリティを双方向に変換するモデルを設計することである．本稿では，これをモダリティの双方向生成と呼ぶことにする．

異なるモダリティは，それぞれ異なる種類の次元や構造を持つ．そのため，モダリティ間の関係は強い非線形性を持つ．このような関係をモデル化するために，近年深層ニューラルネットワークが広く使われている [1], [2]．

モダリティを双方向に生成するための主要な手法は，モダリティごとに深層ニューラルネットワークを作成し，各モダリティの最も上位の隠れ層を共有するというものである [1]．この手法の利点は，すべてのネットワークを end-to-end に学習でき，また共有した隠れ層で共有表現（joint representation）が獲得できることである．各モダリティの情報を統合しているため，共有表現では，単一のモダリティよりも情報量の多い特徴量が得られる．また，この共有表現を介することで，あるモダリティから別のモダリティを容易に生成することができる．一方で，双方向生成のためのもう 1 つの単純な手法として，1 方向に生成するネットワークを別々に学習するということが考えられる．モダリティを 1 方向に生成するモデルについては，これまでも複数提案されている [3], [4], [5]．しかし，モダリティを双方向に生成する場合，このアプローチでは，必要なネットワークの数がモダリティの数に対して指数関数的に増大してしまう．さらに，それぞれの方向のネットワークは独立に学習されるので，隠れ層は共有されず，それぞれ異なる表現が獲得されてしまう．したがって，双方向に生成するという目標に対して，この単純な方法はあまり適さない．

異なるモダリティを生成するためには，共有表現を確率的な潜在変数としてモデル化することが重要である．これは，上記で述べたように，異なるモダリティは異なる種類の次元や構造を持つため，それらの関係性は決定論的にならないためである．これを実現する手法としては，深層生成モデルの deep Boltzmann machine（DBM）[2], [6] を用いたモデルが知られている．しかし，DBM の学習則はマルコフ連鎖モンテカルロ（MCMC）法に基づいており，大規模で高次元なデータを入力として学習するのは困難である．

近年，変分推論によって柔軟に深層生成モデルを学習できるモデルとして，variational autoencoder（VAE）[7], [8] が提案されている．このモデルは通常の深層ニューラルネットワークと同様，学習時に誤差逆伝播法を用いることができるため，従来の DBM のような MCMC 法による学習モデルに比べて，大規模で高次元のデータセットを学習することができる．これまでも，VAE を使った複数のモダリティを学習する手法が提案されているが，これらはすべて条件付き分布のモデルであり，1 方向でしかモダリティを生成できない [3], [4], [5]．

本稿では，VAE を用いた複数のモダリティを扱うモデルとして，joint multimodal variational autoencoder（JMVAE）を提案する．JMVAE では，各モダリティに対応する生成モデルの潜在変数が共有されるようにモデル化されており，これは上記の深層ニューラルネットワークや DBM でのアプローチと同じである．したがって，JMVAE は VAE で双方向生成を実現するための単純な拡張といえる．確率モデルとして考えると，JMVAE は潜在変数の下での各モダリティの条件付き分布によって，全モダリティの同時分布を構成している．この同時分布から，あるモダリティの下での別のモダリティの条件付き分布を求めることができるので，これを用いて双方向にモダリティを生成することができる．

あるモダリティから対応する別のモダリティを生成するとき，生成したいモダリティは入力では欠損として扱われる．しかし，欠損モダリティの次元が他のモダリティの次元よりも大きい場合，潜在表現や生成したサンプルが崩れてしまう可能性がある．本稿では，この問題が実際に生じることを実験的に確認し，従来提案されていた欠損値補完の手法を単純に用いるだけでは解消できないことを示す．本研究では，この問題を解決するための追加的な手法として，階層的 JMVAE と JMVAE-kl という 2 つの異なる手法を提案する．階層的 JMVAE は，潜在変数を確率的多層構造にすることで，潜在表現の崩壊を防ぎ，適切なサンプルを生成できる．JMVAE-kl は，本研究で新たに提案するアプローチで，各モダリティを単一で入力とする新たな近似分布を用意し，全モダリティを入力とする JMVAE の近似分布との距離を近づけることで学習する．これらの手法によって，欠損モダリティによる問題は解消され，異なるモダリティ間を双方向に生成できるようになる．**図 1** では，

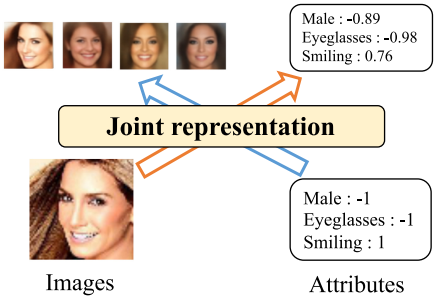

図 1 JMVAE による，共有表現を介した異なるモダリティ間の双方向生成

Fig. 1 Bi-directional generation between different modalities via a joint representation with JMVAE.

JMVAE によって共有表現となる潜在変数を介して，画像から属性，属性から画像のように，次元も構造も異なるモダリティ間を双方向に生成できることを示している．さらに，画像は属性よりも情報量が大きいので，同じ属性から対応する複数の画像を生成することができる．

本研究の主な貢献は以下のとおりである．

- 異なる種類のモダリティを双方向に生成するために，VAE でマルチモーダル学習を実現する JMVAE を提案する．JMVAE は VAE を用いて全モダリティが対等な形で同時分布をモデル化した初めての手法であり，双方向にモダリティを生成できる．
- 入力で欠損させるモダリティが他のモダリティよりも次元が大きい場合，生成したサンプルが崩れてしまう問題があること明らかにし，従来の欠損値補完手法を単純に用いるだけでは解決できないことを実験的に示す．
- 欠損モダリティ問題を解決するために，階層的 JMVAE および JMVAE-kl という追加的な提案手法を導入する．
- 上記の手法によって欠損モダリティ問題が解決し，すべてのモダリティを統合した共有表現が獲得され，適切に異なるモダリティ間の双方向の生成ができることを定量的・定性的実験によって示す．

2. 関連研究

深層ニューラルネットワークで複数のモダリティを扱う主要なアプローチは，モダリティごとのネットワークの最上位の隠れ層を共有することである．Ngiam らは深層オートエンコーダを用いてこのアプローチをとるモデルを提案し，共有表現で単一のモダリティよりも良い特徴が得られることを示した [1]．しかし，このモデルは決定論的な共有表現をモデル化しているため，異なる次元や構造を持つモダリティ間で生成することは困難である．

Srivastava らは上記のアイディアを DBM [9] に適用し，全モダリティの同時尤度が最大となるように学習するモデルを提案した [2]．また，Sohn らは Srivastava らの DBM による手法を拡張し，variation of information 最小化に基づく手法を提案した [6]．これらの DBM に基づくモデルは確率モデルとして設計されているので，双方向に異なるモダリティを生成できる．しかし，DBM の学習則は MCMC 法に基づくので，これらのモデルは自然画像のような高次元のデータを入力として学習できないという問題があった．本研究で提案する JMVAE は，大規模で高次元のデータを学習可能な深層生成モデルである VAE [7], [8] を拡張したモデルなので，このような DBM ベースのモデルが持つ問題を解決している．

VAE で異なる 2 つのモダリティをモデル化する手法としては，すでに Kingma らや Sohn らによって conditional VAE（CVAE）が提案されている [3], [4]．このモデルは，変分推論によって条件付き尤度を最大化するように学習される．近年の VAE を用いた 2 つのモダリティを学習する手法のほとんどが，CVAE の枠組みに基づいている．例として，ラベルから手書き数字 [3], [4]，回転角度から物体画像 [10]，属性から顔画像 [11], [12]，そしてキャプションから画像 [13]，といったものがある．CVAE の最も大きな特徴は，モダリティ間の関係が 1 方向であること，そして潜在変数が条件づけられたモダリティの情報を含んでいないことである[*1]．このため，CVAE は双方向に生成できないだけでなく，複数のモダリティを統合した共有表現を獲得できない．

Pandey らは CVAE と同様，VAE で条件付き対数尤度を最大化するモデルとして conditional multimodal autoencoder（CMMA）を提案している [5]．CVAE との違いは，潜在変数が 2 つのモダリティに対応する変数と接続しているため，その潜在変数で 2 つのモダリティ情報を統合した共有表現が得られる可能性があるということである．しかし，CMMA も 1 方向でしか生成できないため，双方向にモダリティを生成する場合は，各方向のモデルを別々に用意して学習しなければならない．このとき，各モデルの潜在変数は共有されないため，方向によって異なる共有表現が獲得されてしまう．一方 JMVAE では，全モダリティのネットワークを end-to-end に学習できるため，モダリティを統合した 1 つの共有表現が得られる．

VAE 以外の深層生成モデルの枠組みでは generative adversarial net（GAN）[15] を用いたマルチモーダル学習が多数提案されている．GAN は生成分布の尤度を明示的に指定せずに学習できるため，VAE よりも鮮明な画像を生成できることが特徴である．CVAE のように，別のモダリティで条件づけた conditional GAN（CGAN）[16] を用いた研究が一般的で，キャプションから画像を生成する研究も複数提案されている [17]．しかし，この手法も条件付き分布をモデル化しているので，1 方向しか生成できない．

最近では本研究と同様に，同時分布をモデル化して，画像から画像について双方向に生成できる GAN のモデルが提案されている [18], [19], [20]．これらのモデルは，同一サイズの画像間のピクセル単位での完全な対応関係をモデル化しており，生成の際に確率的要素はなるべく無視する傾向にある．しかし，本研究で着目するような異なる種類のモダリティでは，図 1 で示したように 1 対 1 関係になることはない．本研究ではすべてのモダリティ情報を統合した確率的な共有表現をモデル化しているため，モダリティ間の確率的関係を学習できる．

[*1] Louizos らは，CVAE のエンコーダには潜在変数と条件づけられた変数との間に依存関係が残っているため，潜在変数と条件づけられた変数は厳密に独立になるわけではないと指摘している [14]．

3. 既存手法と提案手法

本章では，最初に VAE のアルゴリズムについて簡単に紹介し，次に本稿の提案手法である JMVAE について述べる．

3.1 Variational autoencoder

観測変数 $\mathbf{x}$ が与えられたとき，潜在変数 $\mathbf{z}$ を考え，生成過程を $\mathbf{z} \sim p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ および $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{z})$ とする．ただし，θ は p のモデルパラメータである．VAE の目標は，周辺分布 $p(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$ を最大化するように生成分布を学習することである．この分布は計算困難であるため，代わりに次のような変分下界 $\mathcal{L}_{VAE}(\mathbf{x})$ を最大化することで，モデルを学習する．

$$\begin{aligned}&\log p(\mathbf{x})\\&\quad \geq -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) + E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]\\&\quad = \mathcal{L}_{VAE}(\mathbf{x})\end{aligned} \tag{1}$$

ただし，$q_\phi(\mathbf{z}|\mathbf{x})$ は事後分布 $p(\mathbf{z}|\mathbf{x})$ の近似分布であり，ϕ はモデルパラメータである．それぞれの分布について，$q_\phi(\mathbf{z}|\mathbf{x})$ を**エンコーダ**，$p_\theta(\mathbf{x}|\mathbf{z})$ を**デコーダ**と見なせるので，このモデルは variational autoencoders（VAE）[7], [8] と呼ばれる．式 (1) において，第 1 項は正則化項，第 2 項は負の再構成誤差を表している．

エンコーダ $q_\phi(\mathbf{z}|\mathbf{x})$ をガウス分布とすると，次のように深層ニューラルネットワークでパラメータ化できる．

$$\begin{aligned}q_\phi(\mathbf{z}|\mathbf{x}) &= \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2))\\ \boldsymbol{\mu} &= f_\mu(f_{\mathrm{MLP}}(\mathbf{x}))\\ \boldsymbol{\sigma}^2 &= \mathrm{Softplus}(f_{\sigma^2}(f_{\mathrm{MLP}}(\mathbf{x})))\end{aligned} \tag{2}$$

ただし，f_μ と f_{σ^2} はそれぞれ線形の単層ニューラルネットワーク，f_{MLP} は任意の層数を持つ深層ニューラルネットワークを表す．また，Softplus はベクトルの各要素に対してソフトプラス関数を活性化関数として適用することを意味する．

デコーダ $p_\theta(\mathbf{x}|\mathbf{z})$ については，$\mathbf{x}$ の各要素が実数値をとるときはガウス分布として式 (2) と同様に深層ニューラルネットワークでパラメータ化することができる．$\mathbf{x}$ の各要素がそれぞれ独立に 2 値をとる場合はベルヌーイ分布とし次のようにパラメータ化できる．

$$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{B}(\mathbf{x}; \boldsymbol{\mu}), \boldsymbol{\mu} = \mathrm{Sigmoid}(f_\mu(f_{\mathrm{MLP}}(\mathbf{z}))) \tag{3}$$

ただし Sigmoid はシグモイド関数とする．2 値の場合でも，$\mathbf{x}$ が one-hot（1 つの要素のみが 1 で残りは 0）のときはカテゴリ分布とし，次のようにパラメータ化できる．

$$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{C}(\mathbf{x}; \boldsymbol{\mu}), \boldsymbol{\mu} = \mathrm{Softmax}(f_\mu(f_{\mathrm{MLP}}(\mathbf{z}))) \tag{4}$$

ただし Softmax はソフトマックス関数である．

パラメータ θ，ϕ について下界 $\mathcal{L}(\mathbf{x})$ を最大化する際，式 (1) における負の再構成誤差項が ϕ を持つエンコーダの期待値になっているので，そのままでは勾配を求められない．そこで，$\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ を $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$（ただし $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$）と再パラメータ化（reparameterize）することで，負の再構成誤差項の θ と ϕ に関する勾配は，$\nabla_{\theta,\phi} E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] = E_{\mathcal{N}(\boldsymbol{\epsilon};\mathbf{0},\mathbf{I})}[\nabla_{\theta,\phi} \log p_\theta(\mathbf{z}|\boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon})]$ と推定でき，期待値部分はモンテカルロサンプリングによって近似できる．正則化項の勾配は解析的に求められるので，式 (1) は通常の確率的勾配降下法を用いて最適化することができる．

3.2 Joint multimodal variational autoencoder

本研究では，データセット $\{(\mathbf{x}_1, \mathbf{w}_1), \ldots, (\mathbf{x}_N, \mathbf{w}_N)\}$ を考える．ただし，$\mathbf{x}$ と $\mathbf{w}$ はそれぞれ異なる種類の次元や構造を持つとし，それぞれを異なる**モダリティ**と呼ぶ．データセットの各事例のモダリティの組 $(\mathbf{x}_i, \mathbf{w}_i)$ は同じ対象を表現しているものとする．本研究の目標は，これら 2 種類のモダリティ間を双方向に生成することである．ここでの「双方向に生成する」とは，$\mathbf{x}$ から $\mathbf{w}$ の生成と $\mathbf{w}$ から $\mathbf{x}$ の生成の両方を行うことを指す．

提案手法では，これらは同一の潜在的な概念 $\mathbf{z}$，すなわち共有表現の下で条件付き独立であるとし，各モダリティは異なる分布から生成されると仮定する．したがって，潜在変数および各モダリティの生成過程は $\mathbf{z} \sim p(\mathbf{z})$ および $\mathbf{x}, \mathbf{w} \sim p_\theta(\mathbf{x}, \mathbf{w}|\mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{w}|\mathbf{z})$ となる．

このモデルはすべてのモダリティの同時分布（**joint** distribution）をモデル化していることから，この提案モデルを **Joint Multimodal Variational AutoEncoder**（**JMVAE**）と呼ぶ．

近似事後分布を $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{w})$ とすると，対数尤度 $\log p(\mathbf{x}, \mathbf{w})$ の変分下界は次のようになる．

$$\begin{aligned}\mathcal{L}_{JM}(\mathbf{x}, \mathbf{w}) &= E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})}[\log \frac{p_\theta(\mathbf{x}, \mathbf{w}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{w})}]\\ &= -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{w})||p(\mathbf{z}))\\ &\quad + E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})}[\log p_\theta(\mathbf{x}|\mathbf{z})]\\ &\quad + E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})}[\log p_\theta(\mathbf{w}|\mathbf{z})]\end{aligned} \tag{5}$$

この式には，各モダリティに対応した 2 つの負の再構成誤差項がある．VAE と同様に $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{w})$ をエンコーダ，$p_\theta(\mathbf{x}|\mathbf{z})$ と $p_\theta(\mathbf{w}|\mathbf{z})$ をデコーダと呼ぶ．

式 (5) のエンコーダとデコーダは，深層ニューラルネットワークでパラメータ化し，VAE と同様に各パラメータ（θ と ϕ）に関して最適化できる．各モダリティは異なる特徴表現を持つので，デコーダ $p_\theta(\mathbf{x}|\mathbf{z})$ と $p_\theta(\mathbf{w}|\mathbf{z})$ に対して異なる分布や異なる構造のネットワークを設定する必要がある．分布やネットワーク構造の種類は，データセットにおける各モダリティに依存する．たとえば，あるモダリティの事例の各次元要素が連続値をとるならばガウス分布，2

値をとるならばベルヌーイ分布，2 値かつ one-hot ならばカテゴリ分布となる．エンコーダ $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})$ はガウス分布とし，各モダリティに対して異なるネットワークを用意し，最終層で結合することでパラメータ化する．

JMVAE は，CVAE や CMMA とは異なり全モダリティの同時分布をモデル化し，各モダリティは潜在変数の下で条件付き独立となっている．そのため，すべてのモダリティを含んだ共有表現を抽出できることが期待される．さらに，同時分布について各モダリティで条件付けることで，双方向の条件付き分布が得られるため，テキストから画像，画像からテキストのように，双方向のモダリティの生成も可能となる．加えて JMVAE は $p(\mathbf{x},\mathbf{w}_1,\mathbf{w}_2,\ldots)$ のように 3 つ以上のモダリティも入力として扱うことができる．

3.3 欠損モダリティの推定

JMVAE では，訓練後のテスト段階に，エンコーダ $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})$ を使って，複数のモダリティから共有された潜在表現を推論できる．本稿の目標は双方向に異なるモダリティを生成することなので，対応する生成したいモダリティの事例は手元にない．図 **2** (a) は JMVAE で $\mathbf{x}$ から $\mathbf{w}$ を生成する流れを示しており，エンコーダにおいて $\mathbf{w}$ は欠損しているものとして扱われる．従来の識別的なマルチモーダル学習の設定でも，あるモダリティから別のモダリティを推定する場合は，0 やランダムなノイズが設定される [1]．

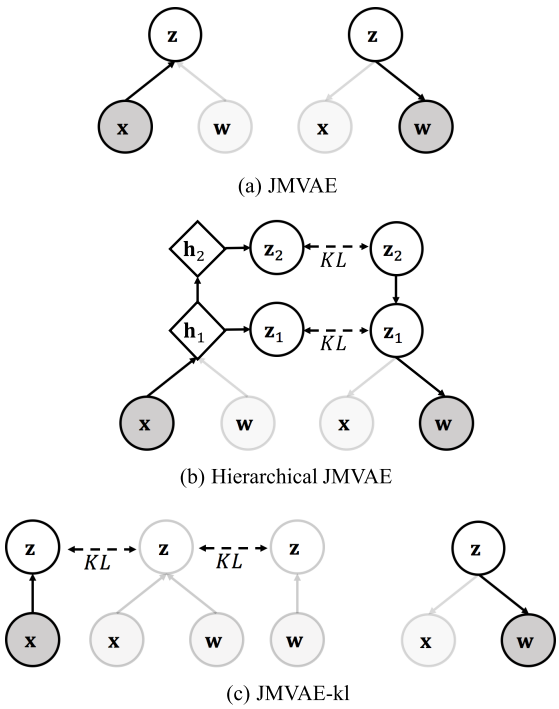

図 **2** (a) JMVAE，(b) 階層的 JMVAE，および (c) JMVAE-kl の推論分布（エンコーダ，左）と生成分布（デコーダ，右）．各手法での $q(\mathbf{z}|\mathbf{x})$ と $p(\mathbf{w}|\mathbf{z})$ のモデル化を表している．丸は確率的変数，菱形は決定論的変数を表す

Fig. 2 Inference (or encoder, left figures) and generative (or decoder, right figures) distributions for (a) JMVAE (b) hierarchical JMVAE, and (c) JMVAE-kl. These figures represent how $q(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{w}|\mathbf{z})$ are modeled on each approach. Circles represent stochastic variables and diamonds represent deterministic variables.

VAE において，入力が欠損している場合の補完方法としては，遷移カーネルを用いたマルコフ連鎖による反復サンプリングの方法が提案されている [8]．JMVAE の場合，$\mathbf{x}$ が欠損したときの遷移カーネル $T(\tilde{\mathbf{x}}|\mathbf{x},\mathbf{w})$ は次のようになる．

$$T(\tilde{\mathbf{x}}|\mathbf{x},\mathbf{w}) = \int p(\tilde{\mathbf{x}}|\mathbf{z})q(\mathbf{z}|\mathbf{x},\mathbf{w})d\mathbf{z} \tag{6}$$

$\mathbf{x}$ の初期値を $\mathbf{x} \sim p(\mathbf{x})$ のようにランダムなノイズとし，式 (6) を用いて反復的にサンプリングすることで，欠損モダリティを推定できる．本稿では，この手法を反復サンプリング手法と呼ぶ．

欠損モダリティが，他のモダリティと比べて高次元で複雑な構造の場合，エンコーダによって推論された潜在変数は不完全となり，デコーダで生成（補完）したサンプルは崩れてしまう可能性がある．本稿では，単純に反復サンプリング手法を用いるだけでは，この現象は防げないことを実験で示す．

本研究では，この問題を解決するために，階層的 JMVAE と JMVAE-kl という 2 つの異なる提案手法を導入する．

3.4 階層的 JMVAE

近年，VAE の潜在変数を確率的な階層構造に拡張して，モデルの表現力や尤度を向上させる手法がいくつか提案されている [21], [22], [23]．本研究で提案する JMVAE は，すべてのモダリティの情報を統合した潜在変数をモデル化しているため，確率的階層構造に容易に拡張できる．潜在変数を L 層の階層 $\mathbf{z}_1,\ldots,\mathbf{z}_L$ とすると[*2]，JMVAE の同時分布は次のようになる．

$$\begin{aligned} &p(\mathbf{x},\mathbf{w}) \\ &\quad = \int \cdots \int p(\mathbf{z}_L)p_\theta(\mathbf{z}_{L-1}|\mathbf{z}_L) \\ &\qquad \ldots p_\theta(\mathbf{z}_1|\mathbf{z}_2)p_\theta(\mathbf{x}|\mathbf{z}_1)p_\theta(\mathbf{w}|\mathbf{z}_1)d\mathbf{z}_1 \ldots d\mathbf{z}_L \end{aligned} \tag{7}$$

ただし，それぞれの条件付き確率分布 $p_\theta(\mathbf{z}_{l-1}|\mathbf{z}_l)$ はすべてガウス分布とし，深層ニューラルネットワークによってパラメータ化されるとする．

階層的な潜在変数を持つ VAE における，近似分布の因数分解の方法は様々提案されているが，本稿では Gulrajani らの分解方法 [23] に従った．この方法では，近似分布は次のように分解される．

$$q_\phi(\mathbf{z}_1,\ldots,\mathbf{z}_L|\mathbf{x},\mathbf{w}) = q_\phi(\mathbf{z}_1|\mathbf{x},\mathbf{w})\ldots q_\phi(\mathbf{z}_L|\mathbf{x},\mathbf{w}) \tag{8}$$

[*2] ここでの確率的階層は，深層ニューラルネットワークにおける決定論的な階層構造とは異なる．

各条件付き分布はガウス分布とし，式 (8) からそれぞれ独立となっている．本研究では，Gulrajani らの手法と同様，入力から最終層まで決定論的な写像で構成され（それぞれの写像は深層ニューラルネットワークでパラメータ化される），各階層の決定論的な出力 $\mathbf{h}_l$ から確率的な出力 $\mathbf{z}_l$ が得られるとした（図 2 (b) を参照）．したがって，確率的階層化した JMVAE の下界は次のようになる．

$$\begin{aligned}&\mathcal{L}_{JM_h}(\mathbf{x},\mathbf{w})\\&=-\sum_{l=1}^{L}E_{q_\phi(\mathbf{z}_{l+1}|\mathbf{x},\mathbf{w})}[D_{KL}(q_\phi(\mathbf{z}_l|\mathbf{x},\mathbf{w})||p_\theta(\mathbf{z}_l|\mathbf{z}_{l+1}))]\\&\quad+E_{q_\phi(\mathbf{z}_1|\mathbf{x},\mathbf{w})}[\log p_\theta(\mathbf{x}|\mathbf{z}_1)]+E_{q_\phi(\mathbf{z}_1|\mathbf{x},\mathbf{w})}[\log p_\theta(\mathbf{w}|\mathbf{z}_1)]\end{aligned}\tag{9}$$

第 1 項は，エンコーダとデコーダの各確率的層間のカルバック・ライブラー距離を最小化するようになっている．

潜在変数の確率的階層化によって，式 (8) のように複雑な近似分布を構成することができるので，より入力での欠損に対して頑健な潜在表現が得られ，さらに反復サンプリングによって適切な欠損モダリティの生成が可能となると期待される．本稿ではこの手法を階層的 JMVAE（hierarchical JMVAE）と呼び，実験を通して，本手法が欠損モダリティの問題を緩和できることを示す．図 2 (b) は階層的 JMVAE で $\mathbf{x}$ から $\mathbf{w}$ を生成する流れを示している．

3.5 JMVAE-kl

確率的階層構造の手法では，欠損モダリティの生成のために，依然として反復サンプリング手法が重要となる．しかし高次元のサンプルを生成する際には，時間がかかってしまうという問題がある．したがって，反復サンプリング手法を用いずに適切な欠損サンプルを生成できる手法として，JMVAE-kl を提案する．

単一のモダリティ入力をとるエンコーダ $q_\lambda(\mathbf{z}|\mathbf{x})$, $q_\lambda(\mathbf{z}|\mathbf{w})$ を考える（ただし λ はモデルパラメータ）．もしこれらを適切に得ることができれば，これらのうち生成元のモダリティに対応する分布を使って $\mathbf{z}$ を直接推論できる．たとえば，テスト時に $\mathbf{x}$ のみから潜在変数を推論したい場合は，$q_\lambda(\mathbf{z}|\mathbf{x})$ を使って $\mathbf{z}\sim q_\lambda(\mathbf{z}|\mathbf{x})$ のように推論できる（図 2 (c) 左参照）．この際，入力に欠損値をとらないので，階層的 JMVAE と異なり，反復サンプリングをせずに異なるモダリティ間を双方向に生成することが可能となる．

JMVAE-kl では，単一のモダリティ入力をとるエンコーダを JMVAE のエンコーダ $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})$ に近づけるようにして学習する（図 2 (c) 左参照）．分布間の距離をカルバック・ライブラー距離とすると，JMVAE-kl の目的関数は次のようになる．

$$\begin{aligned}\mathcal{L}_{JM_{kl}}(\mathbf{x},\mathbf{w})=\ &\mathcal{L}_{JM}(\mathbf{x},\mathbf{w})\\&-[D_{KL}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})||q_\lambda(\mathbf{z}|\mathbf{x}))\\&+D_{KL}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})||q_\lambda(\mathbf{z}|\mathbf{w}))]\end{aligned}\tag{10}$$

他の観点からみると，式 (10) を最大化することは，パラメータ化された分布において変分推論による variation of information の最小化と見なせる．この証明として付録 A.2 を参照されたい．Variation of information は 2 つの変数間の距離を計測する指標であり，p_D をデータ分布とすると $-E_{p_D(\mathbf{x},\mathbf{w})}[\log p(\mathbf{x}|\mathbf{w})+\log p(\mathbf{w}|\mathbf{x})]$ のように，2 つの負の条件付き対数尤度の和で表される．この指標を最小化するということは，双方向にモダリティ間の生成が適切に行われるように学習していることになる．Sohn らによるマルチモーダル学習の手法では，variation of information を最小化によってモデルを学習する方法をとっている [6]．しかし，彼らの手法では MCMC 法によって学習する DBM をモデルとして用いているため，本研究のように自然画像のような次元の大きいデータを直接学習することは困難である．

JMVAE-kl は，モダリティの数が増えると，元の JMVAE のエンコーダのネットワークが増えるだけでなく，モダリティが欠損した場合を考慮したエンコーダも必要となるので，ネットワーク数が膨大になってしまうという問題がある．モダリティが 2 個の場合，JMVAE-kl では 3.5 節で述べたように，元の JMVAE のエンコーダ $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})$ と各モダリティが欠損した場合のエンコーダ $q_\lambda(\mathbf{z}|\mathbf{x})$, $q_\lambda(\mathbf{z}|\mathbf{w})$ が必要となる．ここで，各モダリティから潜在変数への写像を表現したネットワークがそれぞれ同じパラメータ数であると仮定し，そのパラメータ数を M とする．すると，$q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})$ では $2M$，$q_\lambda(\mathbf{z}|\mathbf{x})$ と $q_\lambda(\mathbf{z}|\mathbf{w})$ ではそれぞれ M のパラメータが必要なので，エンコーダに必要なパラメータ数は合計 $4M$ となる．モダリティが 3 個になると，元の JMVAE のエンコーダ $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w}_1,\mathbf{w}_2)$ のほかに，全モダリティのすべての欠損の組合せを考慮したエンコーダ $q_\lambda(\mathbf{z}|\mathbf{x})$, $q_\lambda(\mathbf{z}|\mathbf{w}_1)$, $q_\lambda(\mathbf{z}|\mathbf{w}_2)$, $q_\lambda(\mathbf{z}|\mathbf{x},\mathbf{w}_1)$, $q_\lambda(\mathbf{z}|\mathbf{w}_1,\mathbf{w}_2)$, $q_\lambda(\mathbf{z}|\mathbf{w}_2,\mathbf{x})$ が必要になる．よって，ネットワークのパラメータ数の合計は $12M$ になる．モダリティ数が K 個の場合では，エンコーダのパラメータ数は $2^{K-1}KM$ となり，モダリティ数に対してネットワークのパラメータ数が指数的に増大してしまう．

一方階層的 JMVAE では，モダリティの数が増えても，対応する元の JMVAE のエンコーダのネットワークが増えるだけである．階層的 JMVAE のエンコーダが式 (8) のように因数分解され，それぞれの確率的階層間の決定論的な写像がパラメータ数 M のネットワークで表現されていると仮定すると，K 個のモダリティにおける L 層の階層的 JMVAE では，エンコーダのパラメータ数が合計 $(K+L-1)M$ となる．よって階層的 JMVAE では，JMVAE-kl と比べて，モダリティの数に対するパラメータ数を大幅に抑えられる．

したがって，階層的 JMVAE と JMVAE-kl にはそれぞ

れ長所と短所がある．階層的 JMVAE はモダリティが 3 種類以上で欠損モダリティの次元がそれほど大きくない場合に有効で，JMVAE-kl はモダリティが 2 種類のみで欠損モダリティが高次元の場合に有効であるといえる．

4. 実験

4.1 データセット

本実験の目的は，(1) 欠損モダリティ問題が確かに発生し提案手法でそれが改善されること，(2) 異なるモダリティを統合した共有表現が獲得されていること，(3) 1 方向の生成と同等（もしくはそれ以上）の精度で双方向の生成ができること，の 3 点について確認することである．

既存のマルチモーダル学習のデータセットとしては，MIR Flickr25k [24] やワシントン大学の RGB-D データセット [25] などがある．しかし，これらのデータセットは異なるモダリティでも次元や構造が同じである例が多く，特に RGB-D データセットは物体画像の RGB 情報と深さの情報を異なるモダリティとしていて，画像サイズはモダリティ間で同じとなっている．よって (1) を検証するには不向きである．一方，次元や構造が異なっている例として，一般物体画像とタグを異なるモダリティとする MIR Flickr25k などがあるが，一般物体画像の生成自体が今現在も困難なタスクであるため (2) や (3) の検証に向かない．そもそも，通常のマルチモーダル学習では，通常複数のモダリティからより良い表現を獲得して識別精度を高めることが目的であり，本研究の目的とは異なる．そのため，既存のマルチモーダル学習のデータセットは本実験では用いないこととした．

その代わり，本実験では上記の目的を達成するため，MNIST と CelebA [26] の 2 つのデータセットを用いることにした．

MNIST は，本来マルチモーダル学習のためのデータセットではない．しかし，ラベルを one-hot のタグと考えて 1 つのモダリティとすると，手書き数字画像とラベルでは次元や構造が大きく異なること，one-hot の情報なので潜在空間で多様体学習ができているか判断しやすいこと，そしてデータセットのサイズが小さく生成しやすいことから，上記の目的に適していると考えられる．前処理として，数字画像の各ピクセル値は $[0, 1]$ になるよう正規化した．全データセットのうち 50,000 を訓練事例集合とし，残りの 10,000 をテスト事例集合とした．

CelebA は 202,599 枚のカラー顔画像と対応する 40 の 2 値属性（男性，メガネ，髭など）によって構成される，より一般的なマルチモーダルデータセットである．モダリティ間で次元や構造がより大きく異なるため，MNIST よりも困難な設定だが，画像は顔に限定されているため，一般物体画像などと比べて生成しやすいデータセットと考えられる．前処理として，各画像を顔を中心に正方形に切り取り，64×64 にリサイズしたあと標準化した．本実験では，OpenCV によって顔を特定できた 191,899 をデータセットとして用いた．全データセットのうち 90%を訓練事例集合とし，残りの 10%をテスト事例集合とした．

4.2 モデル構造

4.2.1 MNIST

数字画像を $\mathbf{x} \in [0, 1]^{784}$，対応するラベルを $\mathbf{w} \in \{0, 1\}^{10}$ とした．ここでは構造の表記のため，出力が k ユニットの線形全結合層–ReLU（正規化線形関数）を `DkR`，`DkR` から ReLU を除いた構造を `Dk` とした．また，2 つのネットワーク `I`，`J` の最終層を連結して 1 つの層とする場合は `(I,J)` と表記する．これは深層ニューラルネットワークでは concatenate と呼ばれる処理で，本研究では複数のモダリティのネットワークを結合する場合に用いる．

エンコーダの式 (2) の f_{MLP} を `(D512R-D512R, D512R-D512R)` とし，f_μ と f_{σ^2} は，それぞれ `D64` とした．デコーダは $p(\mathbf{x}|\mathbf{z})$ をベルヌーイ分布，$p(\mathbf{w}|\mathbf{z})$ をカテゴリ分布とした[*3]．いずれのモダリティについても f_{MLP} を `D512R-D512R` とし，$p(\mathbf{x}|\mathbf{z})$ の f_μ は `D784`，$p(\mathbf{x}|\mathbf{z})$ の f_μ は `D40` とした．

本実験では階層的 JMVAE は 2 層まで（$L = 2$）とし，2 層目のエンコーダとデコーダはいずれも f_{MLP} が `D512R-D512R`，f_μ と f_{σ^2} を `D64` とした．JMVAE-kl の $q_\lambda(\mathbf{z}|\mathbf{x})$ の f_μ は `D512R-D512R`，$q_\lambda(\mathbf{z}|\mathbf{w})$ の f_μ は `D512R-D512R` とし，f_μ と f_{σ^2} は，いずれの場合もすべて `D64` とした．

4.2.2 CelebA

顔画像を $\mathbf{x} \in \mathbb{R}^{32\times32\times3}$，対応する属性を $\mathbf{w} \in \{-1, 1\}^{40}$ とする．フィルタのサイズが 4×4 でチャネル数が k，ストライドが 2 の畳込み層–バッチ正規化–ReLU を `CkBR` とし，`CkBR` からバッチ正規化を除いた構造を `CkR`，前述と同じフィルタ構造の逆畳込み層–バッチ正規化–ReLU を `DCkBR`，`DCkBR` からバッチ正規化を除いた構造を `DCkR`，さらに ReLU を除いた構造を `DCk` とする．また，k ユニットの線形全結合層–バッチ正規化–ReLU を `DkBR`，平坦化層を `F` と表記する．

エンコーダの f_{MLP} を `(C64R-C128BR-C256BR-C256BR-F, D512R-D512BR)-D1024R` とし，f_μ と f_{σ^2} は，それぞれ `D128` とした．デコーダは $p(\mathbf{x}|\mathbf{z})$ と $p(\mathbf{w}|\mathbf{z})$ はどちらもガウス分布とした．ただし式 (2) のパラメータ化と異なり，$\boldsymbol{\mu}$ は $\mathrm{Tanh}(f_\mu(f_{\mathrm{MLP}}(\mathbf{z})))$（ただし Tanh は双曲線正接関数）とし，$\boldsymbol{\sigma}^2$ の各要素を 1 に固定した．ネットワーク構造は，$p(\mathbf{x}|\mathbf{z})$ の f_{MLP} を `D4096R-DC256BR-DC128BR-DC64BR`，f_μ を `DC3` とし，$p(\mathbf{w}|\mathbf{z})$ の f_{MLP} を `D4096R-D512BR`，f_μ を `D40` とする．

[*3] MNIST は $[0, 1]$ の実数値をとるが，訓練時とテスト時は要素の実数値に応じて確率的に 0 または 1 に離散化している．これは，過学習を防ぐ役割も果たしている [21], [22].

階層的 JMVAE の 2 層目は，MNIST と同じネットワーク構造とした．JMVAE-kl の $q_\lambda(\mathbf{z}|\mathbf{x})$ の f_{MLP} は C64R-C128BR-C256BR-C256BR-F-DR1024，f_μ と f_{σ^2} はそれぞれ D128 とし，$q_\lambda(\mathbf{z}|\mathbf{w})$ の f_{MLP} は D512R-D512BR-D1024R，f_μ と f_{σ^2} はそれぞれ D128 とした．

4.3 学習パラメータ設定

最適化アルゴリズムは Adam [27] を使い，学習率は MNIST で 10^{-3}，CelebA で 10^{-4} とした．式 (5) の正規化項によって学習の初期に局所解に陥ることを防ぐために，ウォームアップ法 [22], [28] を用いた．これは，学習の初期は再構成誤差のみを学習し，N_t エポックまで線形に正規化項を大きくしていく方法である．MNIST では $N_t = 200$ とし，訓練エポック数を 2,000 とした．CelebA では，$N_t = 20$ とし，訓練エポック数を 50 とした．

モデルの実装には，Theano [29] と Lasagne [30] に基づく深層生成モデルライブラリ Tars [*4] を用いた．

4.4 評価指標

本実験では，モデルの評価にテスト条件付き対数尤度 $\log p(\tilde{\mathbf{x}}|\mathbf{x},\mathbf{w})$（もしくは $\log p(\tilde{\mathbf{w}}|\mathbf{x},\mathbf{w})$）を用いた．条件付き対数尤度 $\log p(\tilde{\mathbf{x}}|\mathbf{x},\mathbf{w})$ は，次のように近似できる．

$$\begin{aligned}\log p(\tilde{\mathbf{x}}|\mathbf{x},\mathbf{w}) &\simeq \log \int q(\mathbf{z}|\mathbf{x},\mathbf{w})p(\tilde{\mathbf{x}}|\mathbf{z})d\mathbf{z} \\ &\geq \int q(\mathbf{z}|\mathbf{x},\mathbf{w}) \log p(\tilde{\mathbf{x}}|\mathbf{z})d\mathbf{z} \\ &\simeq \frac{1}{N}\sum_{i=1}^{N} \log p(\tilde{\mathbf{x}}|\mathbf{z}^{(l)}) \qquad (11)\end{aligned}$$

ただし，$\mathbf{z}^{(l)} \sim q(\mathbf{z}|\mathbf{x},\mathbf{w})$ である．式 (11) では，イェンセンの不等式で下界を求めてからサンプル近似を行っているが，これは対数尤度で直接サンプル近似を行うとサンプル数に対して偏るためである [*5]．この下界は負の再構成誤差と考えることができるので，高い値になるほどモダリティが適切に再構成できることを意味する．本実験では，サンプル数は $N = 10$ とした．

JMVAE では，一方のモダリティしか与えられない場合，すなわち生成するモダリティが欠損している場合の条件付き対数尤度 $p(\tilde{\mathbf{x}}|\mathbf{w})$（または $\log p(\tilde{\mathbf{w}}|\mathbf{x})$）も考えられる．この対数尤度を求めるためには，近似分布 $q(\mathbf{z}|\mathbf{w})$（または $q(\mathbf{z}|\mathbf{x})$）を求め，式 (11) のように潜在変数をサンプリングする必要がある．この近似分布を求める方法は，JMVAE の各手法によって異なる．JMVAE や階層的 JMVAE では，生成するモダリティを欠損値として扱うため，まず欠損モダリティの初期値を 0 とする．次に反復サンプリングを複数回行うことで欠損モダリティを補完し，その後補完したモダリティをエンコーダに入力して潜在変数をサンプリングする．このため，条件付き尤度は反復サンプリングの回数や補完能力に依存し，適切に補完ができれば尤度は高くなる．JMVAE-kl では近似分布 $q_\lambda(\mathbf{z}|\mathbf{x})$（または $q_\lambda(\mathbf{z}|\mathbf{w})$）を学習の際に求めているので，これを用いて直接条件付き尤度を求める．

本稿では便宜上，$\log p(\tilde{\mathbf{x}}|\mathbf{x},\mathbf{w})$ を**複数**条件付き対数尤度もしくは単に条件付き対数尤度と呼び，$\log p(\tilde{\mathbf{x}}|\mathbf{w})$ を**単数**条件付き対数尤度と呼ぶ．単数条件付き対数尤度と複数条件付き対数尤度の両方を用いて，既存研究であり条件付き分布をモデル化する CVAE [3], [4] および CMMA [5] と比較する．適切に双方向に異なるモダリティを生成できるかを検証する際は，単数条件付き対数尤度を用いる．複数条件付き対数尤度では，モダリティの再構成の精度を評価する．階層的 JMVAE や CVAE，CMMA での単数および複数条件付き対数尤度の近似方法は付録 A.1 を参照されたい．

4.5 実験 1：MNIST

4.5.1 実験 1-1：欠損モダリティ問題と提案手法による改善の確認

本節では最初に，JMVAE で双方向にモダリティを生成する際，欠損モダリティが他のモダリティと比べて高次元の場合，生成したサンプルが壊れてしまうことを確認する．また，反復サンプリング手法 [8] のような従来の欠損値補完手法を用いるだけでは，この問題は解決できないことを示す．そして，本研究で提案する階層的 JMVAE と JMVAE-kl によってこの問題が解決されることを示す．

図 3 (a) は，JMVAE で $\mathbf{w}$ から $\mathbf{x}$ を生成した結果を示している．一番上の行，つまり 1 回だけ反復サンプリングを適用して生成した画像は，不鮮明で適切に生成されていない．反復サンプリング数が増えるにつれ，多少鮮明になるが，数字画像は明らかにラベルに対応していない．この結果から，反復サンプリング手法でも欠損モダリティの問題は解決できないことが分かる．

次に，階層的 JMVAE と JMVAE-kl が，この問題を解決できることを示す．図 3 (b) は階層的 JMVAE の場合の結果を示している．反復サンプリングが 1 回の場合は，(a) と同様ラベルに対応した数字が生成されていないが，サンプリング数を増やすと，ラベルにほぼ対応した数字画像が生成されることが確認できる．図 3 (c) は，JMVAE-kl の場合の結果である．3.5 節で述べたように，JMVAE-kl の場合は反復サンプリングせずに，数字に条件付けられた数字を生成することができる．また，JMVAE-kl は階層的 JMVAE よりもはっきりとした数字を生成できることが確認できる．

[*4] https://github.com/masa-su/Tars

[*5] 文献 [21] によると，対数尤度のサンプル近似は，サンプル数が十分に大きければ，期待値が真の対数尤度に近づく．一方下界は不偏推定量であるため，サンプル数に対して偏らない．

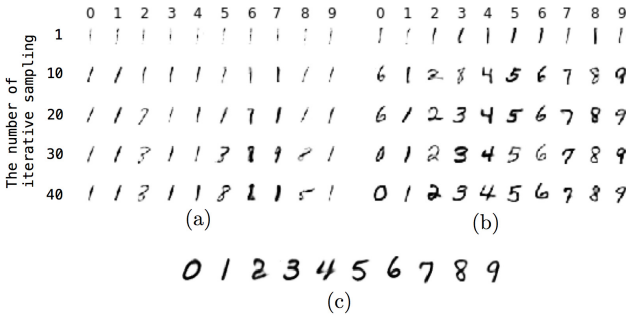

図 3 MNIST におけるラベル（$\mathbf{w}$）から画像（$\mathbf{x}$）の生成．各列は $\mathbf{w}$ 空間の各要素，すなわち 0 から 9 のラベルに対応している．下の行にいくにつれ，$\mathbf{x}$ を生成するための反復サンプリングの回数が増えている．(a) JMVAE．(b) 階層的 JMVAE．(c) JMVAE-kl

Fig. 3 Images ($\mathbf{x}$) generation from labels ($\mathbf{w}$) on MNIST dataset. Each column corresponds to each element of space of $\mathbf{w}$, i.e., labels from 0 to 9. As going to the bottom of the row, the number of iterative sampling for generating $\mathbf{x}$ increases. (a) JMVAE. (b) Hierarchical JMVAE. (c) JMVAE-kl.

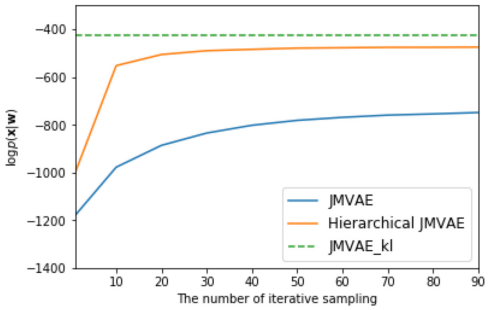

図 4 MNIST データセットにおける異なる反復サンプリング数での JMVAE，階層的 JMVAE，および JMVAE-kl の単数条件付き対数尤度の値

Fig. 4 Single conditional log-likelihood values for JMVAE, hierarchical JMVAE, and JMVAE-kl model with different number of iterative sampling on MNIST dataset.

図 4 は，JMVAE，階層的 JMVAE，JMVAE-kl のそれぞれの単数条件付き尤度 $\log p(\tilde{\mathbf{w}}|\mathbf{x})$ をプロットしたもので，図の横軸は反復サンプリング数を表している（ただし，JMVAE-kl は反復サンプリング手法をとらないので点線で横一直線で表している）．サンプリング数が増えると，JMVAE と階層的 JMVAE の両方の場合で対数尤度は高くなり，反復サンプリングの手法が精度向上に貢献していることが分かる．しかし通常の JMVAE では，JMVAE-kl の尤度よりもはるかに低く，サンプリング回数をかなり増やさないと尤度が高くならない．階層的 JMVAE の場合は，JMVAE よりも少ないサンプリング回数で尤度が大きくなり，最終的な尤度も通常の JMVAE より高くなることが分かる．一方 JMVAE-kl の場合は，反復サンプリングせずに高い尤度を得ることができる．

なお，階層的 JMVAE の条件付き尤度の評価値は，他のモデルの評価値と比較すると過小評価されている可能性があることに留意されたい．これは，条件付き尤度の近似式が，他のモデルよりも下界をおさえているためである（付録 A.1 を参照）．よって階層的 JMVAE は，サンプリング回数を増やした場合に，実際は JMVAE-kl よりも高い尤度となっている可能性もある．

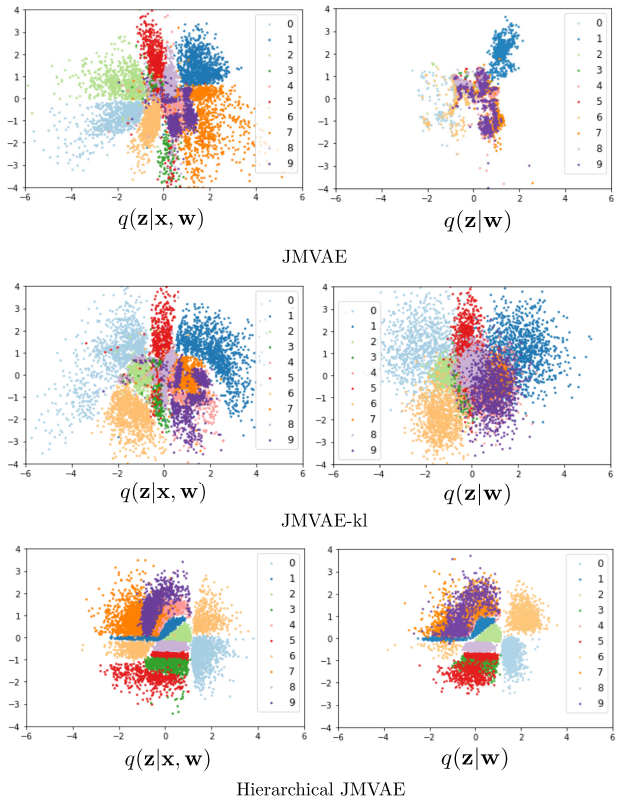

図 5 2-D の潜在表現の可視化．異なる色の点は数字ラベルに対応している．JMVAE と階層的 JMVAE の反復サンプリング数は 100 に設定した

Fig. 5 Visualizations of 2-D latent representation. The number of iterative sampling on JMVAE and hierarchical JMVAE is 100.

4.5.2 実験 1-2：潜在空間での共有表現の確認

続いて，潜在空間で異なるモダリティの共有表現が獲得されていることを確認する．

図 5 は，各手法での潜在表現を可視化したものである．図 5 の左がすべてのモダリティを入力としたエンコーダからサンプリングしたもので，右が単一のモダリティ $\mathbf{w}$ からサンプリングした潜在表現である．なお，階層的 JMVAE の場合は一番上の確率的層でサンプリングしている．ここでは，エンコーダの f_μ と f_{σ^2} を D2 として訓練したものを用いている．

まず図 5 左に着目すると，JMVAE と JMVAE-kl は，いずれもラベルごとに分かれて分布しており，2 つのモダリティを含んだ共有情報が獲得できていることが確認できる．一方，階層的 JMVAE では，よりラベルごとにまとまった表現になっていて，確率的階層化の効果が確認できる．

表 1 MNIST におけるテスト条件付き対数尤度の評価
Table 1 Evaluation of test conditional log-likelihood on MNIST.

	$\log p(\tilde{\mathbf{x}}\|\mathbf{w})$	$\log p(\tilde{\mathbf{w}}\|\mathbf{x})$	$\log p(\tilde{\mathbf{x}}\|\mathbf{w},\mathbf{x})$	$\log p(\tilde{\mathbf{w}}\|\mathbf{x},\mathbf{w})$
CVAE	−448.8	−5.293	−70.42	−5.304
CMMA	−451.1	−0.2971	−69.89	**−0.002574**
JMVAE	−747.1	**−0.2286**	**−69.57**	−0.2026
JMVAE-kl	**−422.35**	−0.2628	−76.94	−0.1874
Hierarchical JMVAE	−475.5	−2.640	−196.9	−0.4456

次に図 5 右に着目する．JMVAE の結果を見ると，左と比べてかなり小さい領域にサンプルの分布が押しつぶされているのが分かる．また，ラベルごとのまとまりもほとんど見られなくなっている．この結果から，画像情報が欠損されると潜在表現が崩れてしまうことが実際に確認できる．一方 JMVAE-kl の結果を見ると，左のすべてのモダリティからのサンプリングとほぼ変わらない潜在表現が得られていることが分かる．右の図では楕円にまとまった表現となっているが，これは $\mathbf{w}$ からの情報量が $\mathbf{x}$ と $\mathbf{w}$ の両方の情報量と比べて小さいため，不確かさが大きくなり，比較的単純な分布となっているためである．JMVAE の場合のように，小さい領域にまとまっていないことから，明らかに欠損モダリティの問題が解決されていることが分かる．最後に階層的 JMVAE の結果を見ると，欠損していない左の結果とほぼ同様の，ラベルごとに分離した分布が獲得されている．この結果から，階層的 JMVAE でも潜在表現が崩れてしまう問題点は解消されたといえる．

4.5.3 実験 1-3：条件付き対数尤度による評価

本項では，モダリティの双方向の生成と再構成を定量的に評価するため，単数条件付き対数尤度と複数条件付き対数尤度で評価する．また，対数尤度の評価を既存の条件付き VAE である CVAE や CMMA と比較する．ここで，CVAE と CMMA は条件付き分布をモデル化していることに注意されたい．つまり，これら従来のモデルは 1 方向でしかモダリティを生成できないため，双方向生成のためには，各生成方向についてモデルを用意して独立に学習する必要がある．そのため，学習時間のコストが増えるだけでなく，実験 1-2 で示したようなすべてのモダリティを統合した共有表現を獲得することができない．一方 JMVAE は，潜在変数が与えられた下での各モダリティの条件付き分布によって同時分布をモデル化しているので，潜在変数で共有表現を獲得でき，その表現を介して双方向にモダリティを生成することができる．

表 1 は両方向のモダリティにおける単数条件付き対数尤度と複数条件付き対数尤度の評価である．この評価では，JMVAE と階層的 JMVAE の反復サンプリング数を 100 とした．既存手法と提案手法，特に通常の JMVAE を比較すると，複数条件付き対数尤度については，既存手法と同等もしくはそれ以上の結果となっている．既存モデルが各方向の生成に別々のモデルを用意して学習する必要があることを考えると，それらと同等の精度で双方向に生成できるという結果は十分であると考えられる．また単数条件付き尤度についても，$\mathbf{x}$ から $\mathbf{w}$ の生成は，他の既存モデルよりも適切に生成できていることが分かる．一方 $\mathbf{w}$ から $\mathbf{x}$ の尤度は低くなっているが，これは実験 1-1 でも示したとおり，欠損モダリティ問題のためである．JMVAE-kl や階層的 JMVAE によって，この問題は解決され，特に JMVAE-kl では従来のモデルよりも高い尤度となることが分かる．

次に表 1 における提案手法内での比較をする．まず，$\mathbf{w}$ の単数条件付き尤度で評価した元の JMVAE と JMVAE-kl の差は，$\mathbf{x}$ の場合の差と比べると，あまり大きくなっていない．このことから，欠損モダリティ問題が発生するのは，生成したいモダリティ，すなわち欠損モダリティがそれ以外のモダリティと比較して高次元の場合だけであることが分かる．次に，複数条件付き尤度の結果から，JMVAE-kl はそれぞれのモダリティの再構成の精度向上には必ずしも貢献しないことが分かる．これは JMVAE-kl が，variation of information 最小化と等価であることから，同一のモダリティの再構成ではなく，異なる種類のモダリティ間をより適切に生成するようにモデル化されているためと考えられる．なお，階層的 JMVAE の結果を見ると，いずれも JMVAE-kl よりも低い結果となっているが，これは実験 1-1 でも述べたとおり，階層的 JMVAE の尤度の評価値が過小評価されているためと考えられる．つまり，階層的 JMVAE については他のモデルとの数値による厳密な比較は難しいということに留意されたい．

4.6 実験 2：CelebA

4.6.1 実験 2-1：欠損モダリティ問題と提案手法による改善の確認

図 6 は，属性から顔画像を生成した結果を示している．まず JMVAE の場合（図 6 (b)）では，属性に対応する顔画像が適切に生成されないことが分かる．さらに，MNIST の場合と異なり，反復サンプリング数を増やすと，生成した顔画像が崩れてしまうことが確認できる．このことから，CelebA の顔画像のような高次元のモダリティの場合では，反復サンプリング手法がまったくうまく働かないことが分かる．次に，階層的 JMVAE の場合（図 6 (c)）を見ると，反復サンプリング数を増やすことで，より綺麗な顔画像が生成されることが分かる．しかし，属性に対応した顔画像

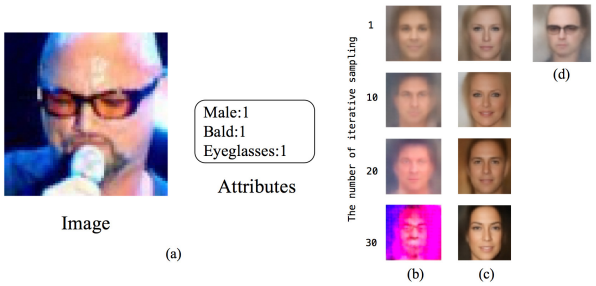

図 6 CelebA データセットにおける属性（$\mathbf{w}$）から顔画像（$\mathbf{x}$）の生成．(a) は CelebA 画像のテスト事例集合の中の一事例．(b) から (d) は，それぞれのモデルで (a) の事例の属性から生成した顔画像であり，下の行にいくにつれ反復サンプリング回数が増えている．(b) JMVAE．(c) 階層的 JMVAE．(d) JMVAE-kl

Fig. 6 Face images ($\mathbf{x}$) generation from attributes ($\mathbf{w}$) on CelebA dataset. (a) is an example of the test set on CelebA. (b) to (d) are face images which are generated from attributes of the example (a). As going to the bottom of the row, the number of iterative sampling for generating $\mathbf{x}$ increases. (b) JMVAE. (c) Hierarchical JMVAE. (d) JMVAE-kl.

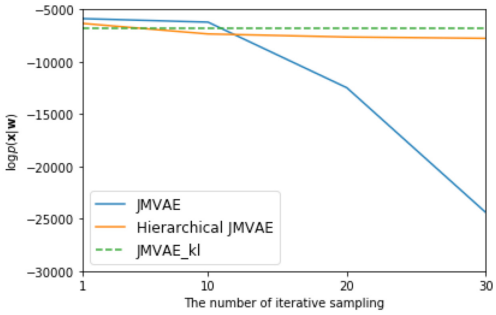

図 7 CelebA データセットにおける異なる反復サンプリング数での JMVAE，階層的 JMVAE，および JMVAE-kl の単数条件付き対数尤度の値

Fig. 7 Single conditional log-likelihood values for JMVAE, hierarchical JMVAE, and JMVAE-kl model with different number of iterative sampling on CelebA dataset.

にはなっていないことも確認でき，特に Eyeglasses という特徴的な属性が生成された顔画像にはまったく反映されていない．CelebA でこのような結果になる理由の 1 つとして，反復サンプリングの際，固定される属性よりも，サンプルごとにランダムに変化する画像の方が次元や情報量が大きいため，潜在空間で大きく移動してしまい，その結果属性とは異なる画像が生成されてしまう，ということが考えられる．また確率的階層が深くなると，潜在変数におけるモード（今回の場合は属性の分布）の隔たりが小さくなることが知られており [31]，このことも，潜在変数において生成したい属性とは異なる属性の分布に移動しやすくなる原因かもしれない．一方 JMVAE-kl の場合（図 6 (d)）は，反復サンプリングをせずに，属性に対応した顔画像を生成することができる．

図 7 は，CelebA において，反復サンプリング数を変更した場合の提案モデルにおける単数条件付き尤度をプロットしたものである．この結果を見ると，図 4 と異なり，反復サンプリング数が 1 のときに JMVAE や階層的 JMVAE は JMVAE-kl よりも高い尤度となり，サンプリング数が増えるごとに尤度が下がることが分かる．これは，CelebA の実験設定では，単数条件付き尤度 $\log p(\tilde{\mathbf{x}}|\mathbf{w})$ が，元の顔画像と対応する属性から生成した顔画像との平均二乗誤差と実質的に等価（生成分布 $\log p(\mathbf{x}|\mathbf{w})$ が分散 1 で固定されたガウス分布であるため）であることと関連している．平均二乗誤差で測った場合，ぼやけた平均的な顔画像と比較した方が全サンプルで比較したときの平均誤差は小さくなるため，サンプリング数が 1 のときに尤度が高くなる．一方，サンプリング数を増やすと，はっきりした顔画像が生成されるようになるので，元画像との誤差が大きくなるサンプルが複数現れ，結果的にサンプリング数が 1 の場合よりも尤度が低くなると考えられる．図 7 を見ると，階層的 JMVAE の場合は，サンプリング数が増えるにつれて JMVAE-kl よりも少し低い尤度で収束することが確認できる．しかし JMVAE の場合は，大きく尤度が下がってしまっている．これは図 6 の結果と対応しており，高次元モダリティの場合，反復サンプリング手法では欠損モダリティ問題が解決できないことを示している．

4.6.2 実験 2-2：条件付き対数尤度による評価

表 2 は条件付き尤度の評価を示している．JMVAE と階層的 JVMAE の反復サンプリング数は 40 とした．この表から，MNIST の場合と同様，既存の条件付きモデルと比べて，JMVAE による条件付き尤度の評価値が同等もしくはわずかに高い結果となった．また，実験 2-1 で示したように，JMVAE-kl と階層的 JMVAE の単数条件付き尤度の値が JMVAE よりも大幅に高くなっていることから，欠損モダリティ問題が解消されることが分かる．しかし，これらの値は CVAE の尤度よりわずかに低くなっていることも確認できる．これは CelebA は異なるモダリティ間での情報量の違いが大きいため，双方向の生成が MNIST より困難なためと考えられる．

提案手法内での違いについては，MNIST の場合（表 1）とほぼ同様の結果となった．

4.6.3 実験 2-3：属性から顔画像の生成と共有表現の確認

次に，JMVAE が CelebA データセットの属性から画像を生成できることを確認する．以降の実験では，実験 2-1 の結果から JMVAE-kl を用いることとし，また鮮明な画像を生成するために，GAN と組み合わせることとする．具体的には，$p(\mathbf{x}|\mathbf{z})$ のネットワークを GAN における**生成器**と見なし，JMVAE-kl の下界とともに GAN の誤差関数を最適化する．これは，VAE-GAN モデル [11] と同じ方法である．なお，学習に用いる GAN の識別器のネットワークは

表 2 CelebA におけるテスト条件付き対数尤度の評価
Table 2 Evaluation of test conditional log-likelihood on CelebA.

	$\log p(\tilde{\mathbf{x}}\|\mathbf{w})$	$\log p(\tilde{\mathbf{w}}\|\mathbf{x})$	$\log p(\tilde{\mathbf{x}}\|\mathbf{w},\mathbf{x})$	$\log p(\tilde{\mathbf{w}}\|\mathbf{x},\mathbf{w})$
CVAE	**−6825**	−44.28	−4031	−44.28
CMMA	−6920	−44.57	**−4026**	**−38.74**
JMVAE	−48763	**−43.97**	**−4026**	−43.43
JMVAE-kl	−6852	−44.13	−4089	−43.83
Hierarchical JMVAE	−7355	−47.61	−4901	−47.32

図 8 (a) 平均顔とランダムな顔画像の生成．各行は凡例の属性に対応しており，ランダムな顔画像の各列は同じバリエーションを持つ．(b) 潜在表現の PCA による可視化．それぞれの色は各サンプルが条件付けられている属性に対応している
Fig. 8 (a) Generation of average faces and corresponding random faces. Each row corresponds to same attribute according to legend. Each column in random faces has the same variation. (b) PCA visualizations of latent representation. Colors indicate which attribute each sample is conditioned on.

C64R-C128BR-C256BR-C256BR-F-D1024R-D1S とした（ただし DkS は Dk の活性化関数をシグモイド関数としたもの）．

図 8 (a) では，様々な属性で条件付けて生成した顔画像を示している．ここでは，まずすべての属性を $\{-1,1\}$ からランダムに選択したものを Base とし，Base の設定したい属性（ここでは男性，禿げている，笑っている，に該当する属性）の値を 2（Not の場合は −2）とすることで，各属性における $\mathbf{w}$ を設定した．各属性の平均顔は，$p(\mathbf{x}|\mathbf{z}_{mean})$（ただし $\mathbf{z}_{mean}$ は $q(\mathbf{z}|\mathbf{w})$ の平均）からサンプリングした．さらに，$\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})$（ただし，$\mathbf{z} = \mathbf{z}_{mean} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$ および $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\zeta})$ であり，この図では $\boldsymbol{\zeta} = 0.6$ とした）のように，同じ属性から様々なバリエーションの顔を生成することができる．結果から，各属性に応じて適切に顔が生成できていることが分かる．

図 8 (b) では，生成した各顔画像の潜在空間における位置をプロットしている．この図を見ると，顔画像のサンプルが，対応する属性ごとにまとまって配置されていることが分かる．また，属性ごとのまとまりの中では，平均顔がほぼ中心に位置し，その周りにランダムな複数の顔画像が配置されており，さらにこれらの配置は，Base と Not Male でほぼ同じとなることが確認できる．これらの結果から，顔画像と属性を含んだ共有表現の多様体学習が適切に行われていることが分かる．

4.6.4 実験 2-4：顔画像と属性の双方向の生成

最後に，JMVAE が顔画像と属性間を双方向に生成できることを示す．**図 9** では，訓練事例集合に含まれない画像から属性を生成し，さらに属性値を変更することで様々な顔画像を生成できることを示している．これらの画像は次のようにして作成している．まず，JMVAE でラベルのない画像 $\mathbf{x}$ から対応する属性 $\mathbf{w}$ を生成する．次に，生成した属性 $\mathbf{w}$ から平均顔 $\mathbf{x}_{mean}$ を生成する．そして，変更したい属性の値を変更した $\mathbf{w}$ から $\mathbf{x}'_{mean}$ を生成する．最後に，$\mathbf{x} + \mathbf{x}'_{mean} - \mathbf{x}_{mean}$ とすることで，変更した属性値に対応した顔画像 $\mathbf{x}'$ を得ることができる．なお図 9 の Reconstruction は，入力画像のみから再構成によって生成した画像である．

図 9 から，モダリティの次元が大きく異なるデータセットにおいても，提案手法が両方のモダリティを双方向に生成できることが分かる．なお，上記の属性から画像への生成方法は CMMA [5] と類似しているが，CMMA は 1 方向しか生成できないため，属性情報のない画像を変化させることはできない．

5. まとめ

本稿では，異なる次元や構造を持つモダリティ間を双方向に生成する問題を考え，それを実現するために，VAE を複数の異なるモダリティが扱えるように拡張した JMVAE を提案した．本モデルでは，すべてのモダリティは共有表現で条件付けられており，全モダリティの同時分布をモデ

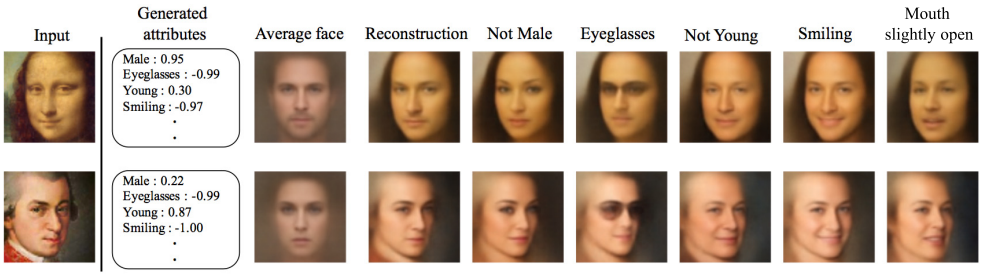

図 9 モナリザ*6（上）とモーツァルト*7（下）の肖像画と，それらの属性の生成，および変更した属性で条件づけて再構成した画像

Fig. 9 Portraits of the Mona Lisa (upper) and Mozart (lower), generated their attributes, and reconstructed images conditioned on varied attributes, according to the legend.

ル化している．このため，既存の条件付き分布をモデル化した手法とは異なり，双方向に異なるモダリティを生成することができる．また，異なる次元のモダリティ間で生成する際，次元の大きいモダリティを欠損させるとうまく補完できないという問題があることを確認した．これを解決するために，新たに階層的 JMVAE と JMVAE-kl という追加的手法を提案した．実験によって，これらの手法で欠損モダリティ問題が解決されることを確認した．さらに，全モダリティを統合した共有表現が適切に獲得され，既存の 1 方向しか生成できないモデルと比較して，同等もしくはそれ以上に適切にモダリティ間を生成できることを確認した．

今後は，画像と文書情報といった，より大きく次元や構造が異なるマルチモーダルデータセットを用いて，双方向の生成を検証する予定である．

謝辞 本研究は JSPS 科研費 JP25700032, JP15H05327, JP16H06562 の助成を受けたものです．

参考文献

[1] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. and Ng, A.Y.: Multimodal deep learning, *Proc. 28th International Conference on Machine Learning* (*ICML-11*), pp.689–696 (2011).
[2] Srivastava, N. and Salakhutdinov, R.R.: Multimodal learning with deep Boltzmann machines, *Advances in Neural Information Processing Systems*, pp.2222–2230 (2012).
[3] Kingma, D.P., Mohamed, S., Rezende, D.J. and Welling, M.: Semi-supervised learning with deep generative models, *Advances in Neural Information Processing Systems*, pp.3581–3589 (2014).
[4] Sohn, K., Lee, H. and Yan, X.: Learning structured output representation using deep conditional generative models, *Advances in Neural Information Processing Systems*, pp.3483–3491 (2015).
[5] Pandey, G. and Dukkipati, A.: Variational methods for conditional multimodal learning: Generating human faces from attributes, arXiv preprint arXiv:1603.01801 (2016).
[6] Sohn, K., Shang, W. and Lee, H.: Improved multimodal deep learning with variation of information, *Advances in Neural Information Processing Systems*, pp.2141–2149 (2014).
[7] Kingma, D.P. and Welling, M.: Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).
[8] Rezende, D.J., Mohamed, S. and Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models, arXiv preprint arXiv:1401.4082 (2014).
[9] Salakhutdinov, R. and Hinton, G.E.: Deep Boltzmann machines, *AISTATS*, Vol.1, p.3 (2009).
[10] Kulkarni, T.D., Whitney, W.F., Kohli, P. and Tenenbaum, J.: Deep convolutional inverse graphics network, *Advances in Neural Information Processing Systems*, pp.2539–2547 (2015).
[11] Larsen, A.B.L., Sønderby, S.K. and Winther, O.: Autoencoding beyond pixels using a learned similarity metric, arXiv preprint arXiv:1512.09300 (2015).
[12] Yan, X., Yang, J., Sohn, K. and Lee, H.: Attribute2image: Conditional image generation from visual attributes, arXiv preprint arXiv:1512.00570 (2015).
[13] Mansimov, E., Parisotto, E., Ba, J.L. and Salakhutdinov, R.: Generating images from captions with attention, arXiv preprint arXiv:1511.02793 (2015).
[14] Louizos, C., Swersky, K., Li, Y., Welling, M. and Zemel, R.: The variational fair auto encoder, arXiv preprint arXiv:1511.00830 (2015).
[15] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative adversarial nets, *Advances in Neural Information Processing Systems*, pp.2672–2680 (2014).
[16] Mirza, M. and Osindero, S.: Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784 (2014).
[17] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B. and Lee, H.: Generative adversarial text to image synthesis, *Proc. 33rd International Conference on Machine Learning*, Vol.3 (2016).
[18] Liu, M.-Y. and Tuzel, O.: Coupled generative adversarial networks, *Advances in Neural Information Processing Systems*, pp.469–477 (2016).
[19] Liu, M.-Y., Breuel, T. and Kautz, J.: Unsupervised image-to-image translation networks, arXiv preprint arXiv:1703.00848 (2017).
[20] Zhu, J.-Y., Park, T., Isola, P. and Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adver-

*6 https://en.wikipedia.org/wiki/Mona_Lisa
*7 https://en.wikipedia.org/wiki/Wolfgang_Amadeus_Mozart

sarial networks, arXiv preprint arXiv:1703.10593 (2017).
[21] Burda, Y., Grosse, R. and Salakhutdinov, R.: Importance weighted autoencoders, arXiv preprint arXiv:1509.00519 (2015).
[22] Sønderby, C.K., Raiko, T., Maaløe, L., Sønderby, S.K. and Winther, O.: Ladder variational autoencoders, arXiv preprint arXiv:1602.02282 (2016).
[23] Gulrajani, I., Kumar, K., Ahmed, F., Taiga, A.A., Visin, F., Vazquez, D. and Courville, A.: PixelVAE: A latent variable model for natural images, arXiv preprint arXiv:1611.05013 (2016).
[24] Huiskes, M.J. and Lew, M.S.: The MIR flickr retrieval evaluation, *Proc. 1st ACM International Conference on Multimedia Information Retrieval*, pp.39–43, ACM (2008).
[25] Lai, K., Bo, L., Ren, X. and Fox, D.: A large-scale hierarchical multi-view RGB-D object dataset, *2011 IEEE International Conference on Robotics and Automation* (*ICRA*), pp.1817–1824, IEEE (2011).
[26] Liu, Z., Luo, P., Wang, X. and Tang, X.: Deep learning face attributes in the wild, *Proc. IEEE International Conference on Computer Vision*, pp.3730–3738 (2015).
[27] Kingma, D. and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[28] Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A.M., Jozefowicz, R. and Bengio, S.: Generating sentences from a continuous space, arXiv preprint arXiv:1511.06349 (2015).
[29] Team, T.T.D., Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A. et al.: Theano: A Python framework for fast computation of mathematical expressions, arXiv preprint arXiv:1605.02688 (2016).
[30] Dieleman, S., Schlter, J., Raffel, C., Olson, E., Snderby, S.K., Nouri, D., Maturana, D., Thoma, M., Battenberg, E., Kelly, J., Fauw, J.D., Heilman, M., de Almeida, D.M., McFee, B., Weideman, H., Takcs, G., de Rivaz, P., Crall, J., Sanders, G., Rasul, K., Liu, C., French, G. and Degrave, J.: Lasagne: First release (2015).
[31] Bengio, Y., Mesnil, G., Dauphin, Y. and Rifai, S.: Better mixing via deep representations, *Proc. 30th International Conference on Machine Learning* (*ICML-13*), pp.552–560 (2013).

付　　録

A.1 階層的JMVAE, CVAE, CMMAにおける条件付き対数尤度の導出

CVAEの複数条件付き対数尤度 $\log p(\tilde{\mathbf{x}}|\mathbf{x},\mathbf{w})$ は，次のようになる．

$$\begin{aligned}\log p(\tilde{\mathbf{x}}|\mathbf{x},\mathbf{w}) &\simeq \log \int q(\mathbf{z}|\mathbf{x},\mathbf{w})p(\tilde{\mathbf{x}}|\mathbf{z},\mathbf{w})d\mathbf{z}\\ &\geq \int q(\mathbf{z}|\mathbf{x},\mathbf{w})\log p(\tilde{\mathbf{x}}|\mathbf{z},\mathbf{w})d\mathbf{z}\\ &\simeq \frac{1}{N}\sum_{i=1}^{N}\log p(\tilde{\mathbf{x}}|\mathbf{z}^{(l)},\mathbf{w}) \qquad \text{(A.1)}\end{aligned}$$

ただし，$\mathbf{z}^{(l)} \sim q(\mathbf{z}|\mathbf{x},\mathbf{w})$ である．

CMMAの複数条件付き対数尤度は，JMVAEと同様式(11)で計算できる．

階層的JMVAEの複数条件付き対数尤度は

$$\begin{aligned}\log p(\tilde{\mathbf{x}}|\mathbf{x},\mathbf{w}) &\simeq \log \int q(\mathbf{z}_2|\mathbf{x},\mathbf{w})\int p(\mathbf{z}_1|\mathbf{z}_2)p(\tilde{\mathbf{x}}|\mathbf{z}_1)d\mathbf{z}_1 d\mathbf{z}_2\\ &\geq \int q(\mathbf{z}_2|\mathbf{x},\mathbf{w})\log \int p(\mathbf{z}_1|\mathbf{z}_2)p(\tilde{\mathbf{x}}|\mathbf{z}_1)d\mathbf{z}_1 d\mathbf{z}_2\\ &\geq \int q(\mathbf{z}_2|\mathbf{x},\mathbf{w})\int p(\mathbf{z}_1|\mathbf{z}_2)\log p(\tilde{\mathbf{x}}|\mathbf{z}_1)d\mathbf{z}_1 d\mathbf{z}_2\\ &\simeq \frac{1}{N}\sum_{i=1}^{N}\int p(\mathbf{z}_1^{(l)}|\mathbf{z}_2)\log p(\tilde{\mathbf{x}}|\mathbf{z}_1)d\mathbf{z}_1 \qquad \text{(A.2)}\end{aligned}$$

と計算できる．ただし，$\mathbf{z}_2^{(l)} \sim q(\mathbf{z}_2|\mathbf{x},\mathbf{w})$ であり，積分部分については次のように近似できる．

$$\int p(\mathbf{z}_1^{(l)}|\mathbf{z}_2)\log p(\tilde{\mathbf{x}}|\mathbf{z}_1)d\mathbf{z}_1 \simeq \frac{1}{N}\sum_{j=1}^{N}\log p(\tilde{\mathbf{x}}|\mathbf{z}_1^{(k)})$$

ただし，$\mathbf{z}_1^{(k)} \sim q(\mathbf{z}_1|\mathbf{z}_2^{(l)})$．

式(A.2)の近似式は，式(11)や式(A.1)よりも下界をおさえた式となっている．したがって他の評価値よりも，実際の尤度と比べたときの過小評価の度合いが高くなる可能性があることに留意されたい．

単数条件付き対数尤度の近似式は，CVAEと階層的JMVAEについては，通常のJMVAEと同様，片方の入力を欠損させたエンコーダからサンプリングすることで導出する．CMMAについては $p(\mathbf{z}|\mathbf{x})$ または $p(\mathbf{z}|\mathbf{w})$ からサンプリングして導出する．

A.2 JMVAE-klの目的関数とvariation of informationの関係について

Variation of information は $-E_{p_{\mathcal{D}}(\mathbf{x},\mathbf{w})}[\log p(\mathbf{x}|\mathbf{w}) + \log p(\mathbf{w}|\mathbf{x})]$（ただし $p_{\mathcal{D}}$ はデータ分布）と表現される．以降の導出では，この式の負の対数尤度の和に着目し，期待値については考慮しない．

対数尤度の和 $\log p(\mathbf{x}|\mathbf{w}) + \log p(\mathbf{w}|\mathbf{x})$ の変分下界は次のように計算できる．

$$\begin{aligned}&\log p(\mathbf{x}|\mathbf{w}) + \log p(\mathbf{w}|\mathbf{x})\\ \geq\ & E_{q(\mathbf{z}|\mathbf{x},\mathbf{w})}[\log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z}|\mathbf{w})}{q(\mathbf{z}|\mathbf{x},\mathbf{w})}] + E_{q(\mathbf{z}|\mathbf{x},\mathbf{w})}[\log \frac{p(\mathbf{w}|\mathbf{z})p(\mathbf{z}|\mathbf{x})}{q(\mathbf{z}|\mathbf{x},\mathbf{w})}]\\ =\ & E_{q(\mathbf{z}|\mathbf{x},\mathbf{w})}[\log p(\mathbf{x}|\mathbf{z})] + E_{q(\mathbf{z}|\mathbf{x},\mathbf{w})}[\log p(\mathbf{w}|\mathbf{z})]\\ & -D_{KL}(q(\mathbf{z}|\mathbf{x},\mathbf{w})||p(\mathbf{z}|\mathbf{x})) - D_{KL}(q(\mathbf{z}|\mathbf{x},\mathbf{w})||p(\mathbf{z}|\mathbf{w}))\\ =\ & \mathcal{L}_{JM}(\mathbf{x},\mathbf{w})\\ & -[D_{KL}(q(\mathbf{z}|\mathbf{x},\mathbf{w})||p(\mathbf{z}|\mathbf{x})) + D_{KL}(q(\mathbf{z}|\mathbf{x},\mathbf{w})||p(\mathbf{z}|\mathbf{w}))]\\ & +D_{KL}(q(\mathbf{z}|\mathbf{x},\mathbf{w})||p(\mathbf{z})) \qquad \text{(A.3)}\end{aligned}$$

もし，すべての確率分布がニューラルネットワークなどでパラメータ化されていれば，同じネットワーク構造で表現できるので，$p(\mathbf{z}|\mathbf{x})$ と $p(\mathbf{z}|\mathbf{w})$ のそれぞれを，$q(\mathbf{z}|\mathbf{x})$ と $q(\mathbf{z}|\mathbf{w})$ と置き換えることができる．したがって，上記の置き換えをした式(A.3)は，

$$
\begin{aligned}
&\mathcal{L}_{JM}(\mathbf{x},\mathbf{w}) \\
&-[D_{KL}(q(\mathbf{z}|\mathbf{x},\mathbf{w})||q(\mathbf{z}|\mathbf{x}))+D_{KL}(q(\mathbf{z}|\mathbf{x},\mathbf{w})||q(\mathbf{z}|\mathbf{w}))] \\
&+D_{KL}(q(\mathbf{z}|\mathbf{x},\mathbf{w})||p(\mathbf{z})) \\
=\ &\mathcal{L}_{JM_{kl}}(\mathbf{x},\mathbf{w})+D_{KL}(q(\mathbf{z}|\mathbf{x},\mathbf{w})||p(\mathbf{z}))\geq\mathcal{L}_{JM_{kl}}(\mathbf{x},\mathbf{w})
\end{aligned}
\tag{A.4}
$$

となる．

したがって，式 (10) を最大化することは，パラメータ化された分布による対数尤度の和の変分下界の最大化，すなわち変分推論における variation of information の最小化に等しい．

鈴木 雅大 （学生会員）

2013 年北海道大学工学部情報エレクトロニクス学科卒業．2015 年同大学大学院修士課程修了．同年東京大学工学系研究科博士課程入学．人工知能，機械学習の研究に従事．

松尾 豊 （正会員）

1997 年東京大学工学部卒業．2002 年同大学大学院博士課程修了．博士（工学）．産業技術総合研究所，スタンフォード大学を経て，2007 年より東京大学大学院工学系研究科技術経営戦略学専攻准教授．2012 年より人工知能学会理事・編集委員長，2014 年より倫理委員長．人工知能学会論文賞，情報処理学会長尾真記念特別賞，ドコモモバイルサイエンス賞等受賞．専門は，Web 工学，Deep Learning，人工知能．