情報処理学会論文誌 Vol.59 No.12 2261–2278 (Dec. 2018)

深層生成モデルを用いた半教師ありマルチモーダル学習

鈴木 雅大[1,a)] 松尾 豊[1]

受付日 2018年2月21日, 採録日 2018年8月9日

概要：マルチモーダル学習は，異なる種類の情報を入力として扱う機械学習の問題設定である．近年のマルチモーダル学習では，識別モデルとして深層ニューラルネットワークが使われる．通常，訓練のために大量のラベルありデータ集合が必要だが，ラベル情報の付与には人的コストがかかってしまう．したがって，ラベルなしデータ集合も利用して識別モデルの汎化性能を向上させる半教師あり学習が重要となる．半教師あり学習の中でも，variational autoencoder（VAE）などの深層生成モデルによる方法は，end-to-end に訓練でき，識別精度が高いことが知られている．そこで本稿では，深層生成モデルを用いた半教師ありマルチモーダル学習に取り組み，semi-supervised MVAE（SS-MVAE）と semi-supervised HMVAE（SS-HMVAE）という 2 つのモデルを提案する．しかしこれらのモデルは，すべてのモダリティ入力がテスト時に与えられることを想定しており，単一のモダリティ入力のみの場合に識別精度が落ちる可能性がある．半教師ありマルチモーダル学習の枠組みでは，複数のモダリティだけでなく，各モダリティを入力としたときの識別精度の向上も求められる．そこで既存の欠損補完の研究に基づき，SS-HMVAE を拡張したモデルである SS-HMVAE-kl を提案する．実験では，まず MNIST を用いて提案手法の性質について検証する．そしてマルチモーダルデータ集合である RGB-D データ集合を用いて，提案手法が従来の半教師あり単一モダリティおよびマルチモーダル学習を上回る性能となることを確認する．

キーワード：深層生成モデル，マルチモーダル学習，半教師あり学習

Semi-supervised Multimodal Learning with Deep Generative Models

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

Received: February 21, 2018, Accepted: August 9, 2018

Abstract: Multimodal learning is a problem setting of machine learning that treats different types of information as input. In recent years, deep neural networks are used as classifiers of multimodal learning. In general, we need a large amount of labeled dataset for training them, but it takes a human cost to label inputs. Therefore, semi-supervised learning, which improves the performance of the discriminator by using unlabeled dataset, becomes important. Among semi-supervised learning, methods based on deep generative models such as variational autoencoders (VAEs) are known to be trained end-to-end with high accuracy. In this paper, we propose two novel models of semi-supervised multimodal learning based on multimodal VAEs: semi-supervised MVAE (SS-MVAE) and semi-supervised HMVAE (SS-HMVAE). However, since these models assume that all modality inputs are given at testing, the accuracy of discriminator may be poor if the input is only a single modality. In the framework of semi-supervised multimodal learning, it is important to improve discriminator accuracy when inputting single modality as well as multiple modalities. Therefore, we propose SS-HMVAE-kl, which is an extended model of SS-HMVAE based on existing studies of complementation of missing values. From experimentation using the MNIST and RGB-D dataset, we confirm that the proposed models have higher performance than either conventional unimodal or multimodal semi-supervised learning.

Keywords: deep generative model, multimodal learning, semi-supervised learning

1. はじめに

我々は日々，様々な種類の情報に接している．こうした異なる種類の情報はそれぞれモダリティと呼ばれ，我々は

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

複数のモダリティ，すなわちマルチモーダルな情報を取り入れることで，より確実な情報処理を行っている．近年機械学習においても，画像や音声など複数の異なる情報を入力として扱うマルチモーダル学習 [1], [2] が研究されている．

マルチモーダル学習の最も一般的な問題設定は，複数のマルチモーダルデータを入力として，それに対応するラベルを予測するというものである．これは，マルチモーダル学習の中でも「融合（fusion）」の設定と呼ばれている [1], [2]．この問題設定が注目されているのは，異なるモダリティには相補性の性質があり，あるモダリティからは，ラベルを予測するうえで他のモダリティにはない付加価値（多様性）が得られるからである [1]．

たとえば，ロボットは画像情報だけでなく距離情報や感覚センサ情報など，様々なモダリティ情報を同時に取得している．ロボットが物体認識をする問題設定を考えると，距離情報やセンサ情報も利用することで，画像情報だけでは分からない物体の形や質感を知ることができる．したがってそれらの情報をすべて物体カテゴリを予測する識別モデルの訓練集合として利用することで，精度の高いカテゴリラベルの予測が期待できる．

近年のマルチモーダル学習は，識別モデル（識別器）として深層ニューラルネットワーク（深い階層構造を持つニューラルネットワーク）が使われることが多い [3]．これは，従来の識別モデルや特徴抽出器と比較して識別精度が高く，ネットワークの隠れ層で入力の良い特徴表現が獲得できるからである．また，識別モデルがネットワークで構成されているため，設計や結合が容易である．したがって，各モダリティを入力としてとるネットワークを用意し，それらのネットワークの最も上位の隠れ層を共有したものを訓練することで，複数のモダリティの情報を統合した表現（共有表現と呼ばれる）を獲得することができる [3]．

各モダリティのネットワークは，特徴空間に応じて自由に設定できるため，このアプローチは様々な問題設定で成功を収めてきた．たとえば，画像とそれに関する質問文が与えられた場合に解答を予測するという visual question answering では，画像のモダリティに対して畳み込みニューラルネットワーク，質問文に対して LSTM を設定して，それらを結合することで解答を予測するモデルを設計している [4]．

一般に，深層ニューラルネットワークの学習には大量のラベルありデータ集合が必要である．ニューラルネットワークの入力に対応する各モダリティの情報については，比較的安価に手にいれることができる．たとえば，前述したロボットの物体認識の場合，ロボットに搭載されている様々なセンサから，ある物体に関する複数のモダリティ情報を入手することができる．その一方で，出力に対応するラベル情報は人手で与える必要があるため，データが大量にある場合，ラベル付けのための人的コストがかかってしまうという課題がある．

これを解決する方法の 1 つが半教師あり学習の適用である．半教師あり学習は，人手でラベル付けしたデータだけでなく，大量のラベルなしデータも訓練に用いることで識別モデルの汎化性能を高める枠組みである．前述したように，マルチモーダルデータの組合せは安価に入手できることが多いので，これらをラベルなしデータ集合として利用することで，識別モデルの精度を向上させることが考えられる．

複数のモダリティを入力とした半教師あり学習，すなわち半教師ありマルチモーダル学習は，これまでもいくつか提案されている．Cheng らは，深層ニューラルネットワークを用いた共訓練（co-training）によって，各モダリティおよびすべて（ここでは 2 種類）のモダリティをそれぞれテスト集合としたときの識別精度が向上する枠組みを提案している [5]．この問題設定を示したものが**図 1** である．本研究では，これを半教師ありマルチモーダル学習の問題設定とする．

一方で近年，半教師あり学習のモデルとして深層生成モデルが注目されている [6], [7]．深層生成モデルとは，確率分布を深層ニューラルネットワークで表現した生成モデルの総称である．深層生成モデルを用いることで，ラベルあり集合とラベルなし集合を統一的に扱うことができる．したがって，上記の共訓練のアプローチなどとは異なり，end-to-end に半教師あり学習を実行することができる．深層生成モデルとしては，variational autoencoder（VAE）[6] と generative adversarial network（GAN）[7] が代表的な手法であり，いずれも既存の半教師あり学習と比べて高い精度となることが知られている [8], [9], [10]．特に VAE に基づく手法は，GAN と比較して学習が安定すること，そ

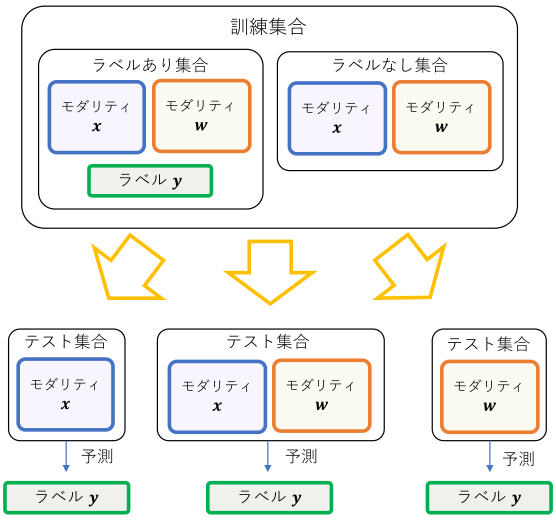

図 1　本研究での半教師ありマルチモーダル学習の問題設定の概要
Fig. 1　The outline of problem setting of semi-supervised multimodal learning in this paper.

して目的関数を確率分布で明示的に定義するため変数や分布（ネットワーク）を増やすなどの拡張が容易であることから，広く利用されている [11], [12]．しかし，これまで深層生成モデルを用いた，マルチモーダルデータのための半教師あり学習はほとんど提案されていない．

こうした背景から，本研究では深層生成モデルを用いた半教師ありマルチモーダル学習のモデルとして，次の 2 つを新たに提案する．1 つ目が，既存の VAE による半教師モデルを単純にマルチモーダルに拡張した semi-supervised MVAE（SS-MVAE）で，もう 1 つが，SS-MVAE に複数のモダリティを統合する共有表現（潜在変数）を追加した semi-supervised HMVAE（SS-HMVAE）である．

これらのモデルは，テスト集合として複数のモダリティが与えられた場合を想定しているが，図 1 で示したように，単一のモダリティのみが入力として与えられた場合でも精度の良い識別結果が求められる．これはすなわち，他のモダリティ入力が欠損した場合に対処する必要があるということである．前述の複数モダリティから物体認識をするロボットの例で考えると，あるモダリティに関するセンサが故障して入力が得られない場合でも，適切に動作できる必要があるということである．

そこで本研究では，入力の欠損に対処するために，マルチモーダル深層生成モデルの欠損補完手法である「KL 最小化の方法」[13], [14] を適用して SS-HMVAE を拡張した，SS-HMVAE-kl を提案する．

本研究の主な貢献は以下のとおりである．

- 深層生成モデルによる半教師ありマルチモーダル学習について新たに取り組み，SS-MVAE と SS-HMVAE を提案する．
- テスト時に単一のモダリティしか与えられない問題設定に対処するため，KL 最小化の方法を利用して SS-HMVAE を拡張した SS-HMVAE-kl を提案する．
- 提案手法の性質を評価するために，MNIST から作成した hmMNIST と nmMNIST という簡単なマルチモーダルデータを用いる．これらを用いた検証から，モダリティ間で情報量が異なり，関係性が確率的なマルチモーダルデータでは，SS-MVAE よりも SS-HMVAE のほうが有効であることを確認する．
- また，そうしたデータではモダリティ入力を欠損した場合に大きく予測精度が落ちることを確認し，SS-HMVAE-kl によってこの問題が大幅に改善されることを確認する．
- マルチモーダルデータ集合である RGB-D データ集合を用いて実験し，提案手法によって，従来の単一モーダルおよびマルチモーダルの半教師あり学習モデルよりも精度の高い結果が得られることを確認する．

2. 関連研究

2.1 半教師ありマルチモーダル学習の既存研究

Guillaumin らは，マルチモーダルデータ（画像とタグ）のラベルあり集合とラベルなし集合から，画像を入力してラベルを予測する識別モデルを学習する枠組みを提案している [15]．この手法は 2 段階からなっており，まずマルチカーネル学習によって，ラベルありデータを用いてラベルを予測する識別モデルを学習し，その識別モデルを使ってラベルなしデータのラベル情報を予測する．次に，ラベルありデータおよびラベルなしデータから予測したラベル情報を用いて，画像からラベルを予測するもう 1 つの識別モデルを学習する，という流れである．この手法をもとに，これまでいくつかの拡張が提案されている [16], [17]．しかしこの手法は，タグからラベルを予測する識別モデルについては，同時に学習することができない．

Cheng らは，深層ニューラルネットワークを用いた共訓練による半教師ありマルチモーダル学習を提案している [5]．共訓練とは，入力データを 2 つに分けて，それぞれに識別モデルを用意し，片方の識別モデルで高い確信度となった事例をもう片方の訓練事例とすることで訓練事例を増やす半教師あり学習の手法である．この手法は 2 つに分けたデータがラベルの下で条件付き独立である必要があるが，マルチモーダルデータの場合，前述した相補性の性質から，この要件を満たしている可能性が高い．Cheng らは，通常の共訓練に加えて，訓練データとして追加する事例が各カテゴリごとに多様性を保つように凸クラスタリング法を用いている．さらに各モダリティの識別モデルを訓練したあと，両方のモダリティを入力とした識別モデルを学習している．これによって，Guillaumin らの枠組みとは異なり，各モダリティを入力とした識別モデルと，両方のモダリティを入力とした識別モデルが得られる．ただし，この手法は上記のように精度向上のために複数の手法を併用しており，RGB-D データ集合に特化したものである．

2.2 深層生成モデルによるマルチモーダル学習

Srivastava らは，複数のモダリティを入力としてとることのできる深層生成モデルである multimodal restricted Boltzmann machine（MRBM）を提案した [18]．これは深層生成モデルの 1 つである deep Boltzmann machine（DBM）をモダリティごとに用意し，各モダリティの最も上位の潜在変数を共有するというものである．これを学習することで，共有した潜在変数で各モダリティの情報を融合した共有表現が得られる．単一モダリティよりも情報量が多くなるため，これを識別モデルの入力とすることで，目標ラベルの予測精度が向上することが期待される．

Sohn らは Srivastava らの DBM による手法を拡張し，variation of information（VI）最小化に基づく手法を提案

した [19]．VI は，2 種類のモダリティの双方向の負の条件付き尤度と見なせるので，これを最小化することで，あるモダリティを入力としたときに，もう片方のモダリティをより良く生成できるようになる．しかし，DBM の学習則は MCMC 法に基づくので，高次元のデータを入力として学習するのは困難という課題があった．

近年では，誤差逆伝播法を用いて学習できる深層生成モデルの研究が進んでおり，マルチモーダル学習に応用されるようになっている．異なるモダリティの共有表現を獲得する深層生成モデルとしては，variational autoencoder（VAE）[6] を用いた方法が主流である [13], [14], [20]．最も単純なものが，JMVAE [13], [14] である[*1]．

2.3 深層生成モデルによる半教師あり学習

深層生成モデルは，半教師あり学習にも優れた性能を示すことで知られている．VAE では，M2 モデル [8] が最初に提案され，それを拡張する形で ADGM と SGDM [9] が提案されている．

マルチモーダルデータで半教師あり学習を行う深層生成モデルは，これまでのところほとんど提案されていない．唯一，Du らによってマルチモーダル感情認識のための半教師あり深層生成モデルが提案されている [21]．Du らは，semiMVAE というモデルを提案しているが，これは本研究で提案する SS-MVAE（3.3 節）と同じ生成モデルの形をしている[*2]．一方本稿では，SS-MVAE を拡張した SS-HMVAE を提案し，さらに単一モダリティを入力とした識別モデルを同時に学習する方法を提案している．

なお，VAE とは別の深層生成モデルである generative adversarial network（GAN）[7] も半教師あり学習のモデルとして使われる [10]．しかし GAN に基づく手法では，マルチモーダル入力を統合する手法が確立していないため，マルチモーダル入力における半教師あり学習の手法も提案されていない．また本研究では，単一モダリティ入力におけるラベル予測も目的としているが，GAN ではモダリティ（または入力）の欠損を補完する手法についても提案されていない．したがって，本研究では GAN に基づく手法との比較は行わない．

3. 提案手法

本章では，本研究の問題設定について説明したあと，マルチモーダルデータを扱う VAE である JMVAE について説明する．次に本研究で提案する SS-MVAE と SS-HMVAE について説明する．そして，単一のモダリティしか入力として与えられない場合に，適切にラベルを予測できる手法について検討する．

3.1 問題設定

データ集合 $\mathcal{D}_{\mathcal{L}} = \{(\boldsymbol{x}_1, \boldsymbol{w}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_N, \boldsymbol{w}_N, \boldsymbol{y}_N)\}$ が訓練集合として与えられるとする．ただし，$\boldsymbol{x}$ と $\boldsymbol{w}$ は異なるモダリティであり[*3]，$\boldsymbol{y} \in \{0,1\}^K$ はそれらの目標カテゴリを表すラベル情報とする[*4]．また訓練集合の各事例 $(\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{y}_i)$ は同じ対象を表現しているものとする．

本研究で取り組む問題設定は，ラベルあり訓練集合の他にラベルなし訓練集合 $\mathcal{D}_{\mathcal{U}} = \{(\boldsymbol{x}_1, \boldsymbol{w}_1), \ldots, (\boldsymbol{x}_M, \boldsymbol{w}_M)\}$（ただし $N \ll M$）が与えられた下で，Cheng らの枠組みのように，$\boldsymbol{x}$ と $\boldsymbol{w}$ を入力とする識別モデル $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})$ だけでなく，各モダリティを入力とする識別モデル，すなわち $p(\boldsymbol{y}|\boldsymbol{x})$ および $p(\boldsymbol{y}|\boldsymbol{w})$ も獲得することである（図 1）．本研究ではこのタスクを**半教師ありマルチモーダル学習**（**semi-supervised multimodal learning**）と呼ぶ．

3.2 JMVAE

まず，マルチモーダルデータを入力として扱う VAE である JMVAE [13], [14] について説明する．合わせて，VAE の枠組みにおける学習方法や確率分布のパラメータ化の方法についても説明する．

2 つのモダリティ $\boldsymbol{x}$，$\boldsymbol{w}$ に共通する潜在的な概念 $\boldsymbol{z}$ を考え，各モダリティはその潜在的概念から（条件付き）独立に異なる分布から生成されるとする．すなわち，潜在変数 $\boldsymbol{z}$ および各モダリティ $\boldsymbol{x}$，$\boldsymbol{w}$ の生成過程を $\boldsymbol{z} \sim p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ および $\boldsymbol{x}, \boldsymbol{w} \sim p(\boldsymbol{x}, \boldsymbol{w}|\boldsymbol{z}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{z})$ とする（ただし，$\boldsymbol{\theta}$ は各分布のモデルパラメータである）．すると，潜在変数 $\boldsymbol{z}$ は 2 つのモダリティを統合する共有表現と考えることができる．深層生成モデルでは，各分布はそれぞれ異なる深層ニューラルネットワークで表現されると考えるので，$\boldsymbol{\theta}$ は，その分布を表すネットワークのパラメータに対応する．このとき，2 つのモダリティの同時分布は $p(\boldsymbol{x}, \boldsymbol{w}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z}$ となる．

深層生成モデルは，訓練集合が与えられた下での同時分布を最大化することで学習できる．しかし $p(\boldsymbol{x}, \boldsymbol{w})$ の計算には潜在変数の周辺化が必要なため，この最大化を直接行うことは困難である．したがって，代わりに以下のような対数尤度の変分下界 $\mathcal{U}_{JMVAE}(\boldsymbol{x}, \boldsymbol{w})$ を最大化することを考える．

$$
\begin{aligned}
&\log p(\boldsymbol{x}, \boldsymbol{w}) \\
&= \log \int p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z}) d\boldsymbol{z}
\end{aligned}
$$

*1 論文によって，joint VAE [20] や multi-view VAE [21] など，呼び方が異なっている．

*2 ただし Du らの論文を発表したのは 2017/4/25 であり，筆者らが初めて SS-MVAE を発表した時期とほぼ同じである [22]．また semiMVAE は，推論モデルに混合ガウス分布を仮定しており，単純にガウス分布を仮定している SS-MVAE とは厳密には異なる．

*3 本研究ではモダリティの数を 2 つに限定する．

*4 本研究ではマルチラベル（1 つの事例が複数の対象に該当する）を想定しないので，one-hot（1 つの要素のみが 1 で残りは 0）表現となる．

$$
\begin{aligned}
&\geq E_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{z})p(\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w})}\right] \\
&= -D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w})||p(\boldsymbol{z})) \\
&\quad + E_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})] + E_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{z})] \\
&\equiv \mathcal{U}_{JMVAE}(\boldsymbol{x},\boldsymbol{w}). \qquad (1)
\end{aligned}
$$

ただし $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w})$ は事後分布の近似分布であり，$\boldsymbol{\phi}$ はモデルパラメータである．事後分布についても深層ニューラルネットワークで表現されるので，$\boldsymbol{\phi}$ はネットワークパラメータとなる．また，JMVAE では $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w})$ をエンコーダ（推論モデル），$p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$，$p_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{z})$ をそれぞれデコーダ（生成モデル）と呼ぶ．式 (1) を目的関数として $\boldsymbol{\phi}$ と $\boldsymbol{\theta}$ について最大化することで，エンコーダとデコーダを学習することができる．なお式 (1) において，第 1 項は正則化項，第 2 項と第 3 項は各モダリティの負の再構成誤差を表している．

エンコーダ $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w})$ 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．

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

ただし，$f_{\boldsymbol{\mu}}$ と $f_{\boldsymbol{\sigma}^2}$ はそれぞれ線形の単層ニューラルネットワーク，$f_{\mathrm{MLP}}(\boldsymbol{x})$ は $\boldsymbol{x}$ を入力とした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 $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$，$p_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{z})$ については，$\boldsymbol{x}$ や $\boldsymbol{w}$ 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．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（ここでは $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ の例を示している）．

$$
\begin{aligned}
&p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = \mathrm{Bern}(\boldsymbol{x};\boldsymbol{\mu}), \\
&\boldsymbol{\mu} = \mathrm{Sigmoid}(f_{\boldsymbol{\mu}}(f_{\mathrm{MLP}}(\boldsymbol{z}))). \qquad (3)
\end{aligned}
$$

ただし Sigmoid はシグモイド関数とする．

なお訓練時には，デコーダについてサンプリング演算はせず，対数尤度（負の再構成誤差）の評価のみを行うので，入力が画像などの場合は，しばしば分布を仮定せずに，負の平均二乗誤差などで対数尤度の評価を代用することがある [14]．

通常の深層ニューラルネットワークと同様に確率的勾配降下法などで学習するためには，パラメータ $\boldsymbol{\theta}$，$\boldsymbol{\phi}$ [*5]について下界の勾配を計算する必要がある．しかし，式 (1) の負の再構成誤差項はパラメータ $\boldsymbol{\phi}$ を持つエンコーダの期待値になっている．したがって，$\boldsymbol{\phi}$ の勾配が期待値の中に入らないため，勾配を計算することができない．そこで，$\boldsymbol{z} \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w})$ を $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$（ただし $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I})$）と再パラメータ化（reparameterize）する．すると，$\boldsymbol{\theta}$ と $\boldsymbol{\phi}$ に関する勾配は式 (1) の負の再構成誤差項の期待値の中に入り，各モデルの勾配を計算できるようになる．これは再パラメータ化トリックと呼ばれる [6], [23]．

式 (1) の正則化項は，エンコーダがガウス分布であり，事前分布 $p(\boldsymbol{z})$ も標準ガウス分布 $\mathcal{N}(\mathbf{0},\mathbf{I})$ であることから，解析的に計算することができる．一方負の再構成誤差項は，再パラメータ化トリックによって $E_{q_{\phi}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\theta}(\boldsymbol{x}|\boldsymbol{z})] = E_{\mathcal{N}(\boldsymbol{\epsilon};\mathbf{0},\mathbf{I})}[\log p_{\theta}(\boldsymbol{z}|\boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon})]$ としたあと，モンテカルロサンプリングによって近似する．先行研究 [6] に従い，本研究ではサンプリング数を 1 としている．なお，以降の式に出てくる期待値についても，同様の方法で計算している．

JMVAE では，訓練後にエンコーダ $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w})$ を使って，複数のモダリティ情報を共有した潜在表現 $\boldsymbol{z}$ を推論できる．しかし一方のモダリティの入力が欠損した場合，推論する $\boldsymbol{z}$ は崩れてしまう可能性がある．この問題を解決する方法として，反復サンプリング手法による補完方法が提案されている [23]．しかし，欠損する情報が大きすぎる場合，この手法でも解決しないことが示されており，反復サンプリングを用いずに欠損問題に対処する方法として KL 最小化の方法 [13], [14] が提案されている．これらについては，3.5 節で説明する．

3.3 SS-MVAE

次に，本研究で提案する半教師ありモデルの説明をする．

モダリティ $\boldsymbol{x}$，$\boldsymbol{w}$ の他にラベル $\boldsymbol{y}$ を考え，生成過程を $\boldsymbol{y} \sim p(\boldsymbol{y}) = \mathrm{Cat}(\boldsymbol{y};\boldsymbol{\pi})$，$\boldsymbol{z} \sim p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0},\boldsymbol{I})$，$\boldsymbol{x},\boldsymbol{w} \sim p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{w}|\boldsymbol{z},\boldsymbol{y}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{y})p_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{y})$ とする．ここで，ラベル $\boldsymbol{y}$ は，先行研究 [8] に従い，潜在変数 $\boldsymbol{z}$ と独立であるとしている．この独立性によって，$\boldsymbol{z}$ では，ラベル情報に依存しない概念を獲得することができる．上記の生成過程から，全モダリティとラベルの同時分布は $p(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{y})p_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{y})p(\boldsymbol{z})p(\boldsymbol{y})d\boldsymbol{z}$ で与えられる．

この深層生成モデルは，JMVAE にラベル情報が加わったものとみることができる．もう 1 つの見方として，モダリティが 1 つだとすると，Kingma らによって提案された VAE の半教師あり学習モデルである M2 モデル [8] と同じ生成モデルになる．したがって，この深層生成モデルは，M2 モデルを単純にマルチモーダルデータに拡張したものと見なすことができる．本研究ではこのモデルを *Semi-Supervised Multimodal Variational AutoEncoder*（*SS-MVAE*）と呼ぶ．

*5 確率分布を深層ニューラルネットワークでパラメータ化したので，これらは深層ニューラルネットワークのパラメータである．

SS-MVAE の変分下界 $\mathcal{L}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})$ は次のようになる．

$$\begin{aligned}
&\log p(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})\\
&= \log \int p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{z},\boldsymbol{y})d\boldsymbol{z}\\
&\geq E_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{y})p_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{y})p(\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})}\right]\\
&\equiv \mathcal{L}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y}) \qquad (4)
\end{aligned}$$

ただし $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})$ は推論モデルである．

この目的関数はラベルあり集合において学習するが，半教師あり学習の枠組みではラベルなし集合も学習に利用する．そのために，ラベル情報を含まない同時分布 $p(\boldsymbol{x},\boldsymbol{w})$ を考えて目的関数を設計する．この目的関数は，識別モデル $q_{\boldsymbol{\phi}}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})$ を導入して次のように求められる．

$$\begin{aligned}
&\log p(\boldsymbol{x},\boldsymbol{w})\\
&= \log \iint p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{z},\boldsymbol{y})d\boldsymbol{z}d\boldsymbol{y}\\
&\geq E_{q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{y})p_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{y})p_{\boldsymbol{\theta}}(\boldsymbol{z})p_{\boldsymbol{\theta}}(\boldsymbol{y})}{q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})}\right]\\
&\equiv \mathcal{U}(\boldsymbol{x},\boldsymbol{w}) \qquad (5)
\end{aligned}$$

ただし，$q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w}) = q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})q_{\boldsymbol{\phi}}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})$ である．

なお識別モデルは，ラベル情報の定義域から（3.1 節を参照）カテゴリ分布となるが，通常離散分布に対して再パラメータ化トリックを使うことはできない．既存の半教師ありモデルでは，これをカテゴリ分布の積分を各ラベルの総和の形に直すことで対処している [8], [9]．しかし，これらの論文では 10 クラス程度のデータ集合を使用しているのに対して，本研究の実験では 51 クラスある RGB-D データ集合も用いるため，計算コスト的に総和をとるのは困難である．そこで，本研究では Gumbel-softmax [24] を用いて，近似的に再パラメータ化する．

さらに，ラベルあり集合において識別モデルを学習するために，以下のようにラベルあり集合における識別損失を式 (5) に加える．

$$\mathcal{L}_l(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y}) = \mathcal{L}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y}) + \alpha \cdot \log q_{\boldsymbol{\phi}}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w}) \qquad (6)$$

ただし，α は学習において識別モデルと生成モデルの割合を調節するパラメータである．

したがって，ラベルあり・なし集合の両方における目的関数 $\mathcal{J}_{MVAE}$ は，

$$\begin{aligned}
&\mathcal{J}_{MVAE}\\
&= \frac{1}{N}\sum_{(\boldsymbol{x}_i,\boldsymbol{w}_i,\boldsymbol{y}_i)\in\mathcal{D}_{\mathcal{L}}} \mathcal{L}_l(\boldsymbol{x}_i,\boldsymbol{w}_i,\boldsymbol{y}_i) + \frac{1}{M}\sum_{(\boldsymbol{x}_j,\boldsymbol{w}_j)\in\mathcal{D}_{\mathcal{U}}} \mathcal{U}(\boldsymbol{x}_j,\boldsymbol{w}_j) \qquad (7)
\end{aligned}$$

となる．

この目的関数をラベルあり・なし集合を用いて最大化することで，生成モデル，推論モデル，そして識別モデル $q_{\boldsymbol{\phi}}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})$ を同時に学習することができる．また式 (7) の計算は，ラベルあり集合だけでなく，ラベルなし集合でも利用できるので，end-to-end な半教師あり学習が実現できる．

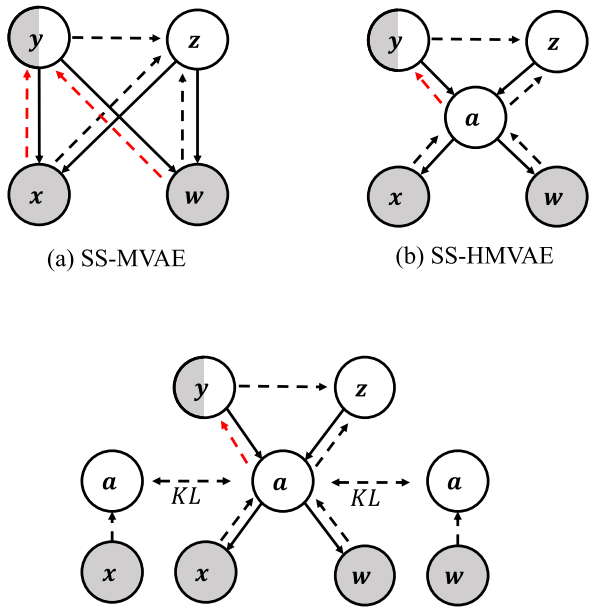

図 2　本研究で提案する半教師ありマルチモーダルモデルのグラフィカルモデル

Fig. 2　A graphical model of semi-supervised multimodal model proposed in this paper.

図 2 (a) は SS-MVAE のグラフィカルモデルである．グラフィカルモデルとは，確率変数間の依存関係をグラフで表現したモデルで，黒丸が観測変数，白丸が潜在変数を表している．ラベル $\boldsymbol{y}$ が黒と白の半分ずつになっているのは，ラベルありとラベルなしの両方（すなわち，観測変数と潜在変数の両方）として考えるためである．それぞれの矢印は異なる確率分布を表していて，実線が生成モデル，点線が推論モデル，そして赤い点線が識別モデルを表している．

3.4　SS-HMVAE

SS-MVAE では，推論モデルは $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})$ と設定し，ガウス分布を仮定していたが，変分推論において，推論モデルは複雑な形になっていることが望ましい．たとえば，観測変数 $\boldsymbol{x}$ と潜在変数 $\boldsymbol{z}$ を持つ生成モデル $p(\boldsymbol{x},\boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$ を考えたとき，変分下界は $\mathcal{L}(\boldsymbol{x}) = \log(\boldsymbol{x}) - D_{KL}(q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}|\boldsymbol{x}))$ となる（ただし $q(\boldsymbol{z}|\boldsymbol{x})$ は事後分布 $p(\boldsymbol{z}|\boldsymbol{x})$ の近似分布となっている推論モデルである）．このとき，推論モデル $q(\boldsymbol{z}|\boldsymbol{x})$ が複雑になればなるほど，真の事後分布に適合するようになるため，下界が真の対数尤度の良い近似となる．また，本研究では 2 つのモダリティを入力として扱うので，その

点からも，潜在変数を推論する分布はガウス分布のような単純な分布ではなく，より複雑な形となることが望ましい．

そこで，新たな潜在変数として $\boldsymbol{a}$ を導入し，推論モデルを $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w},\boldsymbol{y}) = \int q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})d\boldsymbol{a} = \int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{a},\boldsymbol{y})q_{\boldsymbol{\phi}}(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})d\boldsymbol{a}$ とする．また潜在変数 $\boldsymbol{a}$ を使って，生成モデルも $p(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y}) = \iint p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{a})p_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{a})p_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{z},\boldsymbol{y})p(\boldsymbol{z})p(\boldsymbol{y})d\boldsymbol{a}d\boldsymbol{z}$ とする．すると，ラベルあり集合における目的関数は，

$$\begin{aligned}&\log p(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})\\&\geq E_{q_{\boldsymbol{\phi}}(\boldsymbol{a},\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})}\left[\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{a})p_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{a})p_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{z},\boldsymbol{y})p(\boldsymbol{z})p(\boldsymbol{y})}{q_{\boldsymbol{\phi}}(\boldsymbol{a},\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})}\right]\\&\equiv\mathcal{L}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})\end{aligned}\tag{8}$$

となる．

また，識別モデルについても $\boldsymbol{a}$ を使って $q_{\boldsymbol{\phi}}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w}) = \int q_{\boldsymbol{\phi}}(\boldsymbol{y}|\boldsymbol{a})q_{\boldsymbol{\phi}}(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})d\boldsymbol{a}$ とする．すると識別モデルの対数尤度（負の損失関数）は $\log q_{\boldsymbol{\phi}}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w}) = \log E_{q_{\boldsymbol{\phi}}(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})}[q_{\boldsymbol{\phi}}(\boldsymbol{y}|\boldsymbol{a})]$ となるため，より柔軟な確率分布で表現することができる．この識別モデルを用いることで，ラベルなし集合における目的関数は次のようになる．

$$\begin{aligned}&\log p(\boldsymbol{x},\boldsymbol{w})\\&\geq E_{q_{\boldsymbol{\phi}}(\boldsymbol{a},\boldsymbol{z},\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})}\left[\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{a})p_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{a})p_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{z},\boldsymbol{y})p(\boldsymbol{z})p(\boldsymbol{y})}{q_{\boldsymbol{\phi}}(\boldsymbol{a},\boldsymbol{z},\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})}\right]\\&\equiv\mathcal{U}(\boldsymbol{x},\boldsymbol{w})\end{aligned}\tag{9}$$

ただし $q_{\boldsymbol{\phi}}(\boldsymbol{a},\boldsymbol{z},\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w}) = q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{a},\boldsymbol{y})q_{\boldsymbol{\phi}}(\boldsymbol{y}|\boldsymbol{a})q_{\boldsymbol{\phi}}(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})$ である．

なお，式 (9) の下界は

$$\begin{aligned}&E_{q_{\boldsymbol{\phi}}(\boldsymbol{a},\boldsymbol{z},\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})}\left[\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{a})p_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{a})p_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{z},\boldsymbol{y})p(\boldsymbol{z})p(\boldsymbol{y})}{q_{\boldsymbol{\phi}}(\boldsymbol{a},\boldsymbol{z},\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})}\right]\\&=E_{q_{\boldsymbol{\phi}}(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})}\left[E_{q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{y}|\boldsymbol{a})}\left[\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{z},\boldsymbol{y})p(\boldsymbol{z})p(\boldsymbol{y})}{q_{\boldsymbol{\phi}}(\boldsymbol{z},\boldsymbol{y}|\boldsymbol{a})}\right]\right]\\&\quad+E_{q_{\boldsymbol{\phi}}(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{a})]+E_{q_{\boldsymbol{\phi}}(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{a})]\\&\quad+\mathcal{H}(q_{\boldsymbol{\phi}}(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w}))\end{aligned}\tag{10}$$

と分解することができる．ここで，第 1 項は観測変数を $\boldsymbol{a}$ とした M2 モデルの $q_{\boldsymbol{\phi}}(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})$ における期待値であり，第 2 項と第 3 項は JMVAE（式 (1)）の再構成誤差項に等しい．また第 4 項は平均情報量である．このことから，SS-HMVAE の目的関数は JMVAE の学習と M2 モデルの両方の下界を含んでいることが分かる．したがって $\boldsymbol{a}$ では，JMVAE と同様に複数のモダリティを統合した共有表現が獲得され，それを入力とした半教師あり学習が行われると解釈できる．

このように，潜在変数が階層的な構造になっていることから，本モデルを *Semi-Supervised Hierarchical Multimodal Variational AutoEncoder*（*SS-HMVAE*）と呼ぶ．図 2 (b) がグラフィカルモデルである．

$\boldsymbol{a}$ のような潜在変数を追加してより柔軟な分布を表現する研究は，これまでもいくつか行われている [9], [25], [26]．2.3 節で述べた ADGM と SDGM [9] は，VAE に補助変数と呼ばれる潜在変数を導入した半教師あり学習モデルである．

3.5 単一モダリティ入力への対処

SS-MVAE や SS-HMVAE の識別モデルは，マルチモーダルデータが入力として与えられることが前提となっている．しかし，3.1 節や図 1 で示したように，テスト時に単一のモダリティデータしか与えられない場合でも適切にカテゴリラベルを予測したい．

テスト集合に単一のモダリティデータしかない場合，ラベルを推定する最も単純な方法は，そのモダリティ以外に該当する識別モデルの入力を欠損させることである[*6]．しかし，識別モデルは深層ニューラルネットワークで設計されており，ニューラルネットワークではモダリティの欠損に対して適切に対処できないという問題が指摘されている [2]．こうした問題の解決方法としては反復サンプリング手法 [23] がある．しかし，欠損モダリティの情報量が大きい場合，十分に補完できないことが指摘されている [14]．そのため本研究では，先行研究で提案された KL 最小化の方法 [13] を利用して SS-HMVAE を拡張したモデルである SS-HMVAE-kl を提案する．

以下では，反復サンプリング手法，そして KL 最小化の方法およびそれを利用した SS-HMVAE-kl について，それぞれ説明する．

3.5.1 反復サンプリング

反復サンプリング手法 [23] は，VAE において入力の欠損値を補完する方法である．この手法は，通常の VAE のように，入力から潜在変数への推論モデルと潜在変数から入力への生成モデルがある場合に，反復的にサンプリングを繰り返すことで入力の欠損を補完する方法である．

SS-MVAE の識別モデルには潜在変数が含まれないが，SS-HMVAE では $q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w}) = \int q(\boldsymbol{y}|\boldsymbol{a})q(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})d\boldsymbol{a}$ のように，識別モデルに潜在変数 $\boldsymbol{a}$ が含まれており，マルチモーダル入力から潜在変数 $\boldsymbol{a}$ を推論する推論モデル $q(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})$ と，潜在変数 $\boldsymbol{a}$ から入力への生成モデル $p(\boldsymbol{x},\boldsymbol{w}|\boldsymbol{a}) = p(\boldsymbol{x}|\boldsymbol{a})p(\boldsymbol{w}|\boldsymbol{a})$ がある．これらを用いて，たとえば識別モデルの入力 $\boldsymbol{x}$ が欠損した場合の遷移カーネルは次のように書ける．

$$T(\tilde{\boldsymbol{x}}|\boldsymbol{x},\boldsymbol{w}) = \int p(\tilde{\boldsymbol{x}}|\boldsymbol{a})q(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})d\boldsymbol{a}.\tag{11}$$

したがって，SS-HMVAE における反復サンプリングによる補完の手段は次のとおりである．まず，欠損モダリ

*6 なお，ここでの「欠損させる」とは，識別モデルのそのモダリティに対応する入力（本稿の場合は $\boldsymbol{x}$ か $\boldsymbol{w}$）を 0 値とすることである．

ティである $\boldsymbol{x}$ の初期値を $\boldsymbol{x} \sim p(\boldsymbol{x})$ のようにランダムなノイズとし，推論モデル $q(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})$ を使って潜在変数 $\boldsymbol{a}$ をサンプリングする．次に，生成モデル $p(\boldsymbol{x}|\boldsymbol{a})$ から $\boldsymbol{x}$ をサンプリングし，新たな入力値として置き換える．そして，この再構成の工程を n 回反復し，最終的に n 回目に得られた $\boldsymbol{x}$ を，ラベル $\boldsymbol{y}$ を予測する際の識別モデルの入力として用いる．

反復サンプリング手法では，反復の結果，最終的に補完したモダリティの値が収束することが想定されている．しかし欠損モダリティの情報量が大きすぎる場合は，文献 [14] で示されているように，収束せずに適切な値が補完できない可能性がある．

3.5.2 KL 最小化の方法

「KL 最小化の方法」は，上記の反復サンプリングでは対処できない欠損モダリティを補完するための手法である．この手法は，JMVAE の欠損問題を対処する方法として提案された [13], [14] [*7]．この手法も，入力から潜在変数への推論モデルと，潜在変数から入力への生成モデルがあるならば利用できる．したがって，反復サンプリングと同様 SS-HMVAE で適用可能である．

マルチモーダル入力から潜在変数 $\boldsymbol{a}$ への推論モデル $q_{\boldsymbol{\lambda}}(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})$ があるとき，それとは別に，モダリティごとの推論モデルである $q_{\boldsymbol{\lambda}}(\boldsymbol{a}|\boldsymbol{x})$ と $q_{\boldsymbol{\lambda}}(\boldsymbol{a}|\boldsymbol{w})$ を新たに用意する（ただし $\boldsymbol{\lambda}$ はモデルパラメータ）．適切に各モダリティの推論モデルが学習されていれば，これらを使って単一モダリティから潜在変数 $\boldsymbol{a}$ を直接推定できる．

KL 最小化の方法では，モダリティごとの推論モデルを，元の推論モデルとの距離（ここではカルバック・ライブラー（KL）ダイバージェンスを考える）が小さくなるように学習する．すなわち，以下の目的関数 $\mathcal{J}_{div}$ を最小化する．

$$\begin{aligned}\mathcal{J}_{div} = E_{p_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{w})}[&D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})||q_{\boldsymbol{\lambda}}(\boldsymbol{a}|\boldsymbol{x})) \\ &+ D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})||q_{\boldsymbol{\lambda}}(\boldsymbol{a}|\boldsymbol{w}))] \qquad (12)\end{aligned}$$

ただし，$p_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{w})$ はデータ集合 $\mathcal{D}$ におけるデータ分布である．SS-HMVAE の場合は，この期待値はラベルあり集合 $\mathcal{D}_{\mathcal{L}}$ とラベルなし集合 $\mathcal{D}_{\mathcal{U}}$ のマルチモーダルデータでの平均値となる．

さらに，元の推論モデルや生成モデルも同時に学習する必要があるため，SS-HMVAE に KL 最小化の方法を適用した場合の目的関数は

$$\mathcal{J}_{kl} = \mathcal{J}_{HMVAE} - \beta\mathcal{J}_{div} \qquad (13)$$

となる．ただし，$\mathcal{J}_{HMVAE}$ は SS-HMVAE の目的関数，β は KL 項の影響を調節するパラメータである．本研究では，このように SS-HMVAE に KL 最小化の方法を適用したモデルを *SS-HMVAE-kl* と呼ぶ．SS-HMVAE-kl では，単一モダリティ $\boldsymbol{x}$ を入力とする識別モデルを $q_{\boldsymbol{\phi},\boldsymbol{\lambda}}(\boldsymbol{y}|\boldsymbol{x}) = E_{q_{\boldsymbol{\lambda}}(\boldsymbol{a}|\boldsymbol{x})}[q_{\boldsymbol{\phi}}(\boldsymbol{y}|\boldsymbol{a})]$ のように求めることができる．したがって，反復サンプリング手法を適用する必要がない．

KL 最小化の方法は直感的なアプローチであるが，その妥当性についても検証されている．文献 [14] では，これが変分推論による VI 最小化と数式的に等価であることが示されている．VI は $-E_{p_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{w})}[\log p(\boldsymbol{x}|\boldsymbol{w}) + \log p(\boldsymbol{w}|\boldsymbol{x})]$ と表される．したがって，これを最小化するということは，モダリティ間の双方向変換がより適切になるように促されることを意味する．JMVAE や SS-HMVAE では双方向変換がうまくいくということは，欠損補完が適切に行われることなので，直感的な JMVAE-kl や SS-HMVAE-kl の効果と対応している．

また文献 [20] では，式 (12) の片方の KL 項，すなわち $E_{p_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{w})}[D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})||q_{\boldsymbol{\lambda}}(\boldsymbol{a}|\boldsymbol{x}))]$ から $D_{KL}(q_{\boldsymbol{\phi}}^{avg}(\boldsymbol{a}|\boldsymbol{x})||q_{\boldsymbol{\lambda}}(\boldsymbol{a}|\boldsymbol{x}))$ を導出できることが証明されている．ただし，$q_{\boldsymbol{\phi}}^{avg}(\boldsymbol{a}|\boldsymbol{x}) = E_{p_{\mathcal{D}}(\boldsymbol{w}|\boldsymbol{x})}[q_{\boldsymbol{\phi}}(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})]$，すなわち元の推論分布について，$\boldsymbol{x}$ に対応するすべての $\boldsymbol{w}$ で平均したものである．このことから，KL 最小化の方法には，本質的に一方のモダリティを補完する効果があることが分かる．

半教師ありマルチモーダル学習における KL 最小化の方法の利点は，単一モダリティ入力とマルチモーダル入力の識別モデルを同時に学習できることである．一方，Cheng らの半教師ありマルチモーダル学習では，各モダリティのネットワークを共訓練したあとに，2 つのモダリティネットワークの特徴量を入力とした識別モデルを学習する必要がある [5]．

図 2 (c) が SS-HMVAE-kl のグラフィカルモデルである．KL と書かれている部分は，KL ダイバージェンスを近づけているという意味である．

なお，モダリティの欠損を考慮する方法としては，SS-MVAE や SS-HMVAE におけるラベル情報 $\boldsymbol{y}$ のように，観測変数と潜在変数の両方としてあらかじめモデル化する方法も考えられる．しかし，この方法を適用するためには，欠損モダリティを推定するための推論モデルを新たに用意し，さらに各モダリティについて欠損あり/なしの両方での下界を目的関数に加える必要がある．そのため，モデルが複雑になってしまううえに，計算コストが大幅に増大してしまうことが懸念される．したがって，本研究ではこの方法はとらず，反復サンプリングや KL 最小化の方法といった，既存の手法を活用することで欠損モダリティ問題に対処した．

[*7] JMVAE に KL 最小化の方法を適用したモデルは JMVAE-kl と呼ばれる [13], [14]．ただし，文献 [20] では，JMVAE-kl のことを単に JMVAE と呼んでいる．

4. 実験

4.1 データ集合

本研究では，以下の 2 種類のデータ集合を用いて実験を行う．

4.1.1 MNIST

1 つ目のデータ集合は手書き数字画像の MNIST である．MNIST はマルチモーダルデータではないが，本研究では提案手法について様々な分析をするために利用する．MNIST は 10 クラス（それぞれのクラスが数字カテゴリに該当）のラベル情報を持ち，訓練事例数が 60,000 枚，テスト事例数が 10,000 枚である．

本研究では，以下の 2 つの方法で MNIST からマルチモーダルデータを作成した．

1) MNIST を $[0,1]$ に正規化したあと左右に分割し，それぞれを異なるモダリティと見立てる（$\boldsymbol{x} \in [0,1]^{392}$ および $\boldsymbol{w} \in [0,1]^{392}$）．これは文献 [19] でもトイプロブレムとして利用されており，本研究では half multimodal MNIST（hmMNIST）と呼ぶ．
2) MNIST を $[0,1]$ に正規化したあと，2 通りのノイズを付与して，それぞれを異なるモダリティのデータとする．MNIST の各事例ごとに $[-n,n]$（ただし n は角度）の範囲でランダムに回転させたものを 1 つ目のモダリティ $\boldsymbol{x} \in [0,1]^{784}$ とし，各事例にガウスノイズ（平均 0，標準偏差 σ）を付与したものをもう 1 つのモダリティ $\boldsymbol{w} \in [0,1]^{784}$ とする．こちらは文献 [27] での方法に基づくもので[*8]，本研究では，noisy multimodal MNIST（nmMNIST）と呼ぶ．

これら 2 つのデータ集合はラベル予測の難しさが異なっている．hmMNIST については 2 つのモダリティを結合することで元の MNIST を復元することができ，片方のモダリティからもう片方のモダリティを容易に推測できる．また，ラベル情報に対する 2 つのモダリティの情報量も同じくらいである．一方 nmMNIST の場合は，結合しただけでは冗長な表現になるうえ，それぞれのモダリティで異なるノイズが加わっていることから，決定論的な対応関係になっておらず，ラベルに対する情報量も異なる．したがって，nmMNIST は hmMNIST よりも困難な設定である．実際のマルチモーダルデータでも，異なるモダリティでは決定論的な対応関係になっておらず，多くの場合一方のモダリティの情報量がもう一方のモダリティの情報量と異なる．nmMNIST はそうした特性を人工的に再現したデータ集合である．

*8 なお，いくつかの点で文献 [27] とは異なる．まず，文献 [27] では，$\boldsymbol{w}$ は $[0,1]$ の範囲の一様ノイズとしてるが，本研究はガウスノイズに変更している．また，文献 [27] では同一カテゴリ内の事例について各モダリティでランダムにシャッフルしたのちにペアとしている（すなわちマルチモーダルデータのペアでラベル情報は同じだが，元の事例は異なる）が，本研究ではシャッフルせずに元が同一の事例をペアとしている．

4.1.2 Washington RGB-D

もう 1 つのデータ集合は，Washington RGB-D である．

RGB-D とは，カラー画像（RGB）とそれに対応する奥行き（Depth）画像で構成される情報である．本実験で用いる Washington RGB-D データ集合 [28] は Microsoft の Kinect によって撮られたものである．300 種類の家庭用品の画像で構成され，それらは 51 のカテゴリにグループ分けされている．先行研究 [29] と同様，それぞれのグループを画像のクラスとして扱う．

こちらのデータ集合は，モダリティ間で特徴空間も情報量も異なることから，先行研究でもマルチモーダルデータとして扱われている [29]．本研究では，実践的なデータ集合として，既存の半教師ありマルチモーダル学習の手法と比較するために用いる．**図 3** が，Washington RGB-D の例である．

51 クラスのそれぞれには 600 枚の事例があり，各クラスについて 5 枚飛ばしに事例を選択することで 41,877 枚のデータ集合を作った．41,877 枚のうち，文献 [28] が指定した 10 通りの分割方法に従い，訓練集合とテスト集合に分けた．平均しておよそ 35,000 枚が訓練集合，6,877 枚がテスト集合となっている．さらに訓練集合から 5%をランダムに選択してラベルあり集合とし，残りをラベルなし集合とした．

RGB-D データ集合の前処理について説明する．まず，RGB 画像と Depth 画像の両方を 148×148 にリサイズする．元画像は縦長や横長なので，元画像の長いほうを 148 で固定し，短いほうは端のピクセルを拡張する形で補間する．次に，Depth 画像の距離情報が欠損している部分を最近傍の距離の値で補間し，距離の値を 0 から 255 に正規化する．さらに，jet colormap 処理によって，単一チャネルの Depth 画像を 3 チャネルに拡張する．以上の前処理の

図 3 Washington RGB-D の例．左が RGB 画像，右が対応する Depth 画像である．ここでは Depth 画像の色が奥行きの値に対応していて，赤いほど手前の距離であることを表している

Fig. 3 Examples of Washington RGB-D dataset. The left figure is a RGB image, and the right is a corresponding depth image. In this figure of the depth image, the color corresponds to the depth value, i.e., the red color indicates the closer distance.

方法は，すべて文献 [29] に従ったものである．ただし画像サイズは文献 [5] に従い，文献 [29] とは異なりデータ拡張は行っていない．

また RGB-D の実験では，RGB-D データを深層生成モデルの入力として直接扱うのではなく，深層ニューラルネットワークで特徴抽出したものを使うとする．これは，本研究では画像生成が目的でないため，画像を直接入力として入れる必要がないからである．特徴抽出用の深層ニューラルネットワークは，ILSVRC2012 データ集合で事前学習済みの VGG16 [30] とし，fc1 層（4,096 次元）での出力値を特徴量とする．まず VGG16 をモダリティごとに用意し，訓練集合で再学習する．ただし半教師あり学習の設定なので，訓練集合のうちラベルあり集合のみを訓練に用いる．訓練には Adam [31] を用い，過学習を防ぐため学習率は 10^{-5} として 200 エポック学習した．前述のとおり，訓練・テストの分割方法は 10 通りあるので，それぞれの分割のラベルあり集合で訓練して特徴抽出する．

したがって，RGB 画像と Depth 画像の入力特徴量は，それぞれ $\boldsymbol{x} \in \mathcal{R}_{>0}^{4096}$ と $\boldsymbol{w} \in \mathcal{R}_{>0}^{4096}$ となる[*9]．

4.2 学習設定

MNIST と Washington RGB-D における，確率分布とそれをパラメータ化する深層ニューラルネットワークの構造の設定は付録 A.1 章に載せる．

学習の際の最適化アルゴリズムは Adam [31] を利用した．MNIST は，バッチサイズを 512 とし，学習率 10^{-3} で 500 エポック学習した．Washington RGB-D では，バッチサイズ 128，学習率 10^{-4} として，100 エポック学習した．その他，SS-HMVAE の反復サンプリング数は 100 回とした．また，特に言及がない場合は $\alpha = 1$（識別モデルのパラメータ）としている．

実装は Theano [32] と Lasagne [33] をベースとした深層生成モデルライブラリ Tars [*10] を使用した．

4.3 実験 1：MNIST

本実験では，MNIST を用いて，提案手法（SS-MVAE, SS-HMVAE，SS-HMVAE-kl）における半教師ありマルチモーダル学習の有効性や精度の違いなどを検証する．

4.3.1 実験 1-1：マルチモーダル入力における評価

表 1 は，SS-MVAE および SS-HMVAE を用いた，hmMNIST と nmMNIST における 2 つのモダリティ（$\boldsymbol{x}$，$\boldsymbol{w}$）を入力としたときの正解率を表している．ラベルあり集合数は，教師あり（全訓練集合をラベルあり集合にした場合），1,000，500，100 枚と変更した．なお，訓練集合のうち，ラベルあり集合以外はすべてラベルなし集合として半教師あり学習に利用した．

表 1 ラベルあり集合の数を変化させたときの，マルチモーダル入力における提案手法の正解率（%）．上が hmMNIST，下が nmMNIST での結果

Table 1 Accuracy (%) of the proposed methods on multimodal input, changing the number of labeled set. The upper represents the result of hmMNIST, and the lower shows that of nmMNIST.

モデル	教師あり	500	300	100
SS-MVAE	98.73	96.21	96.82	96.21
SS-HMVAE	**98.91**	**97.21**	**97.06**	**96.32**

	モデル	教師あり	500	300	100
$n = \pi/3$	SS-MVAE	97.43	94.75	94.52	88.15
$\sigma = 0.3$	SS-HMVAE	**98.17**	**95.00**	**95.42**	**93.35**
$n = \pi/4$	SS-MVAE	96.50	93.73	92.87	86.36
$\sigma = 0.5$	SS-HMVAE	**97.37**	**93.86**	**93.71**	**90.80**
$n = 2\pi/3$	SS-MVAE	95.64	91.15	90.52	70.57
$\sigma = 0.7$	SS-HMVAE	**96.52**	**91.37**	**91.71**	**75.67**

まず hmMNIST の結果を見る．教師あり学習について，SS-HMVAE のほうが SS-MVAE より精度が高いことが確認できる．また半教師あり学習の場合でも，SS-HMVAE のほうが高い精度となっている．これは，SS-HMVAE のほうが優れた半教師ありモデルであるということを示唆しているが，大きな差という訳ではない．このことから，hmMNIST の場合は，SS-HMVAE，すなわち潜在変数 $\boldsymbol{a}$ を含んだモデルの効果はそれほど大きくないことが分かる．

次に，nmMNIST の結果を確認する．この表では，下に行くほどモダリティ $\boldsymbol{x}$ の回転ノイズの角度 n が $\pi/3$，$\pi/4$，$2\pi/3$ と増え，$\boldsymbol{w}$ のガウスノイズの標準偏差 σ も 0.3，0.5，0.7 と大きくなっている．つまり表の下のほうが，モダリティ間の関係性がより非決定論的となっている．表の結果から，ラベル数が少なくなるほど，SS-HMVAE と SS-MVAE の差が開くことが分かる．ラベルありデータ数が少なくなるということは，それだけモダリティ間を対応付けるためのヒントがなくなるということである．したがって hmMNIST の結果と比較すると，確率的な $\boldsymbol{a}$ は，モダリティ間やラベル間の関係性がとりにくいときに，特に重要になるといえる．ただし，関係性が非決定論的になりすぎると（$n = 2\pi/3$，$\sigma = 0.7$ の場合），SS-HMVAE の効果も小さくなることが分かる．

ここで，提案手法が適切にマルチモーダル学習を行えていることを確認するために，SS-MVAE と，単純に 2 つのモダリティ $\boldsymbol{x}$ と $\boldsymbol{w}$ を結合して M2 モデル [8] で学習した場合と比較する．これら 2 つの違いは，異なるモダリティをそれぞれ異なる確率変数と見なすかどうかである．**表 2** は，nmMNIST（$n = \pi/4$，$\sigma = 0.5$）での，M2 モデルと SS-MVAE の結果を示している．この結果から，単純にモダリティを結合するだけでは精度が向上せず，SS-MVAE

*9 定義域が正の実数となるのは，fc1 層の活性化関数に ReLU を使っているためである．

*10 https://github.com/masa-su/Tars

表 2　マルチモーダル入力（nmMNIST，$n = \pi/4$，$\sigma = 0.5$）における，M2 モデル（既存の半教師ありモデル）と SS-MVAE の正解率（%）

Table 2　Accuracy (%) of M2 model, which is an existing semi-supervised model, and SS-MVAE on multimodal input (nmMNIST, where $n = \pi/4$ and $\sigma = 0.5$).

モデル	500	300	100
M2 [8]（$\boldsymbol{x}$ と $\boldsymbol{w}$ を結合したものを入力）	89.03	88.85	70.68
SS-MVAE	**93.73**	**92.87**	**86.36**

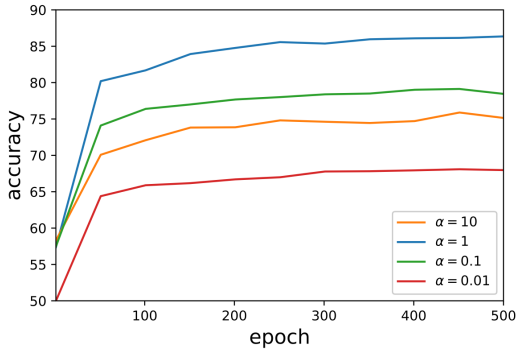

図 4　識別モデルのパラメータ α を変化させたときの学習曲線．データ集合が nmMNIST（$n = \pi/4$，$\sigma = 0.5$），ラベルあり集合数が 100 における結果

Fig. 4　Learning curves in the change of the parameter of the discriminative model, α. They show the results on nmMNIST, where $n = \pi/4$ and $\sigma = 0.5$, for dataset and 100 for the number of labeled set.

で適切に半教師ありマルチモーダル学習を行えていることが分かる．なお hmMNIST については，結合すれば元の MNIST を簡単に復元できるため，ここでは検証しない．

また，提案手法に含まれる識別モデルのパラメータである α を変化させたときの学習曲線を図 4 にプロットする．縦軸が正解率で，横軸がエポック数を表す．モデルは SS-MVAE を用いており，データ集合は nmMNIST（$n = \pi/4$，$\sigma = 0.5$）で，ラベルあり集合のデータ数が 100 のときの結果である．まず，α を 1 より小さくすると，正解率が大幅に落ちることが確認できる．これは，識別モデルの学習よりも生成モデルの学習が優先されるためと考えられる．また，α を 1 より大きくした場合についても，精度が $\alpha = 1$ よりも落ちていることが確認できる．これは α を大きくしすぎると，逆に識別モデルの効果が強くなりすぎて，半教師あり学習の効果が薄れてしまっているためと考えられる．したがって，本研究では以降も $\alpha = 1$ と設定する．

その他，SS-HMVAE における潜在変数 $\boldsymbol{a}$ の次元数を変化させたときの正解率も検証する．付録 A.1.1 節に記載のように，MNIST における SS-HMVAE の $\boldsymbol{a}$ の次元数は 128 に設定している．表 3 では，128 の他に，64，32，2 についての結果を示している．この結果から，次元が大きいうちはあまり大きな性能の差はないが，次元数を大きく減らすと精度も大きく下がることが分かる．

表 3　マルチモーダル入力（nmMNIST，$n = \pi/4$，$\sigma = 0.5$）における，潜在変数 $\boldsymbol{a}$ の次元数に対する正解率（%）．ラベルあり集合数は 100

Table 3　Accuracy (%) on multimodal input (nmMNIST, where $n = \pi/4$ and $\sigma = 0.5$) in the change of the number of dimensions of latent variable $\boldsymbol{a}$. The number of labeled set is 100.

$\boldsymbol{a}$ の次元数	2	32	64	128
SS-HMVAE	38.31	92.68	90.8	93.71

表 4　半教師ありマルチモーダル学習のテスト集合における正解率（%）．上が hmMNIST，下が nmMNIST の結果である

Table 4　Accuracy (%) on test set in semi-supervised multimodal learning. The upper represents the result of hmMNIST, and the lower shows that of nmMNIST.

モデル	$\boldsymbol{x}$	$\boldsymbol{w}$	$\boldsymbol{x}+\boldsymbol{w}$
SS-MVAE	76.81	67.02	96.21
SS-HMVAE	38.30	52.62	**96.32**
SS-HMVAE（反復）	77.60	72.33	**96.32**
SS-HMVAE-kl（$\beta = 0.1$）	**90.75**	**91.39**	94.12
SS-HMVAE-kl（$\beta = 1$）	90.42	91.14	94.30
SS-HMVAE-kl（$\beta = 10$）	88.69	89.53	91.56

モデル	$\boldsymbol{x}$	$\boldsymbol{w}$	$\boldsymbol{x}+\boldsymbol{w}$
SS-MVAE	83.87	21.03	86.36
SS-HMVAE	83.21	46.52	**90.80**
SS-HMVAE（反復）	85.06	59.71	**90.80**
SS-HMVAE-kl（$\beta = 0.1$）	**89.50**	84.18	90.21
SS-HMVAE-kl（$\beta = 1$）	88.71	83.95	89.21
SS-HMVAE-kl（$\beta = 10$）	88.55	**85.54**	89.12

4.3.2　実験 1-2：単一モダリティ入力における評価と欠損補完手法の有効性の検証

3.1 節で示したように，半教師ありマルチモーダル学習では，マルチモーダル入力だけでなく，単一モダリティ入力でも高い識別精度が求められる．そこで，ここでは単一モダリティ入力も含めた精度を確認し，さらに 3.5 節で示した欠損補完手法を適用することで，どのように精度が変化するかを検証する．

ここでは，ラベルあり集合数を 100 枚とし，hmMNIST と nmMNIST（$n = \pi/4$，$\sigma = 0.5$）の両方で実験した．モデルは，SS-MVAE と SS-HMVAE のほかに，反復サンプリングで補完した SS-HMVAE と，SS-HMVAE に KL 最小化の方法を適用した SS-HMVAE-kl を用いた．表 4 が実験結果である．

まず，SS-MVAE と SS-HMVAE については，単一モダリティ入力の正解率（$\boldsymbol{x}$ と $\boldsymbol{w}$）がマルチモーダル入力の正解率よりも低くなっている．これは，hmMNIST と nmMNIST で，その傾向が異なっている．hmMNIST では，$\boldsymbol{x}$ と $\boldsymbol{w}$ で同じくらいの精度となっているが，nmMNIST では，$\boldsymbol{w}$ のほうが大幅に精度が低い．これは，モダリティ $\boldsymbol{x}$ がモダリ

ティ $\boldsymbol{w}$ よりもラベル情報についての情報量が大きく[*11]，$\boldsymbol{x}$ を欠損させたことによりラベルをうまく推定できなかったためと考えられる．これが，3.5 節で述べた欠損モダリティ問題である．

次に，反復サンプリングを適用した SS-HMVAE の結果を見る．SS-MVAE と SS-HMVAE に比べると単一モダリティ入力の精度が向上しており，反復サンプリングの効果が確認できる．しかし，nmMNIST の $\boldsymbol{w}$ の精度は，マルチモーダル入力と比べて低いままである．これは，文献 [14] でも示されているように，情報量の異なるマルチモーダルデータで学習すると，情報量の大きいモダリティが欠損したとき，反復サンプリングでも潜在変数の崩壊が改善されないためと考えられる．

最後に，SS-HMVAE-kl の結果を見る．hmMNIST と nmMNIST の両方で，単一モダリティ入力の精度が向上していることが分かる．特に，SS-MVAE や SS-HMVAE では精度が低かった nmMNIST の $\boldsymbol{w}$ について，大幅に精度が向上している．ここで興味深いのは，nmMNIST のもう 1 つのモダリティである $\boldsymbol{x}$ についても，SS-HMVAE-kl によって精度が向上していることである．このことから，情報量が異なるマルチモーダルデータにおいて SS-HMVAE-kl を訓練した場合は，情報量が少ないほうに合わせて精度が低くなるのではなく，それらを組み合わせた情報，すなわちマルチモーダルにおける識別結果に向かって両方のモダリティの精度が向上するということが分かる．

なお，SS-HMVAE と比べると，SS-HMVAE-kl のマルチモーダル入力の結果が低くなっていることが分かる．これは，KL ダイバージェンスを小さくする際に，マルチモーダルの推論モデル $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w})$ のほうでも，各モダリティの推論モデル $q_{\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{x})$，$q_{\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{w})$ に近づくようにパラメータが更新されるためと考えられる．

また，SS-HMVAE-kl 内で KL 項の影響を調整するパラメータ β を変化させた場合の結果を比べると，β を大きくした場合，単一モダリティの精度がマルチモーダル入力に近づく一方で，マルチモーダル入力自体の精度が落ちてしまうことが分かる．これは，KL 項の影響を強めすぎると，元の推論モデルにも悪影響を及ぼしてしまうことを示している．このことから，β は大きすぎない値に調節するのが適当である．

次に，欠損補間手法による違いを定性的に確認するために，SS-HMVAE の潜在変数 $\boldsymbol{a}$ を介して，あるモダリティから別のモダリティを適切に推定できるかを確認する．具体的には，片方のモダリティから各手法の方法で潜在変数 $\boldsymbol{a}$ を推定し，その後生成モデル $p(\boldsymbol{x},\boldsymbol{w}|\boldsymbol{a}) = p(\boldsymbol{x}|\boldsymbol{a})p(\boldsymbol{w}|\boldsymbol{a})$ を使ってもう片方のモダリティを生成する．対応するモダリティを推定することができれば，単一モダリティ入力からの推論が適切にでき，共有表現も適切に学習されていることになる．hmMNIST と nmMNIST で実験し，ラベル数やパラメータなどは表 4 と同じである．

図 5 の上が hmMNIST の結果である．hmMNIST の結果では，生成したモダリティ画像が適切かどうかを判断しやすくするために，生成元モダリティの画像をつなげて薄く表示している．SS-HMVAE，すなわち欠損させて推論した場合は，うまく生成できていないが，反復サンプリングによって，適切な生成ができるようになっている．ただし $\boldsymbol{x}$（左半分）については，「8」や「3」，「4」や「1」などで生成元データ（右半分）が似た形になるので，わずかな不確定要素として適切に生成できない場合がある．一方 SS-HMVAE-kl は，反復サンプリングが間違えたような点もほとんど間違えずに，適切に補完できていることが分かる．

図 5 の下が nmMNIST である．SS-HMVAE では $\boldsymbol{w}$ については $\boldsymbol{x}$ からノイズがかった画像が生成できていることが分かる．また反復サンプリングを適用することで，一番左の 3 などがはっきり生成できるようになっている．一方，$\boldsymbol{w}$ から $\boldsymbol{x}$ の生成については，$\boldsymbol{x}$ の元データと比較して大幅に情報が欠落している．反復サンプリングを適用した結果を見ても，あまり大きな違いはない．これは，表 4 において，単一モダリティ $\boldsymbol{w}$ のときにあまり精度が向上していないことに対応している．この結果を見ても，$\boldsymbol{w}$ よりも $\boldsymbol{x}$ のほうがラベル情報に対する情報量が大きいことが明らかであり，情報量が小さいほうから大きいモダリティを生成するには，適切な単一モダリティ入力の推論と共有表現の獲得が重要であることが分かる．そして SS-HMVAE-kl の場合を見ると，依然文字が薄いものの，他の手法に比べて遥かに適切に $\boldsymbol{x}$ が補完できている．

このように，定性的な結果からも，hmMNIST と nmMNIST の両方で SS-HMVAE-kl が最も適切に補完できることが確認できた．これは，SS-HMVAE-kl で使われている KL 最小化の方法が，優れた欠損モダリティ補完手法であることを意味し，また半教師あり学習においても適切に動作することが示された．

4.4 実験 2：Washington RGB-D

本実験は，実際のマルチモーダルデータ集合である Washington RGB-D を用いる．本実験の目的は，提案手法を，単一モダリティおよびマルチモーダルにおける既存の半教師あり学習の結果と比較することである．

単一モダリティについては深層生成モデルの M2 モデル [8]，SDGM [9] と比較する．マルチモーダルについては，CT+SVM [34]，共訓練 [5] と比較する．共訓練の手法は，マルチモーダル入力の深層ニューラルネットワークを使い，事前学習に ILSVRC2012 データ集合を用いるなど，本

*11 実際，$\boldsymbol{x}$ はランダムに回転しているものの，画像にノイズが加わっているわけではないため，比較的ラベルを判読しやすいと考えられる．

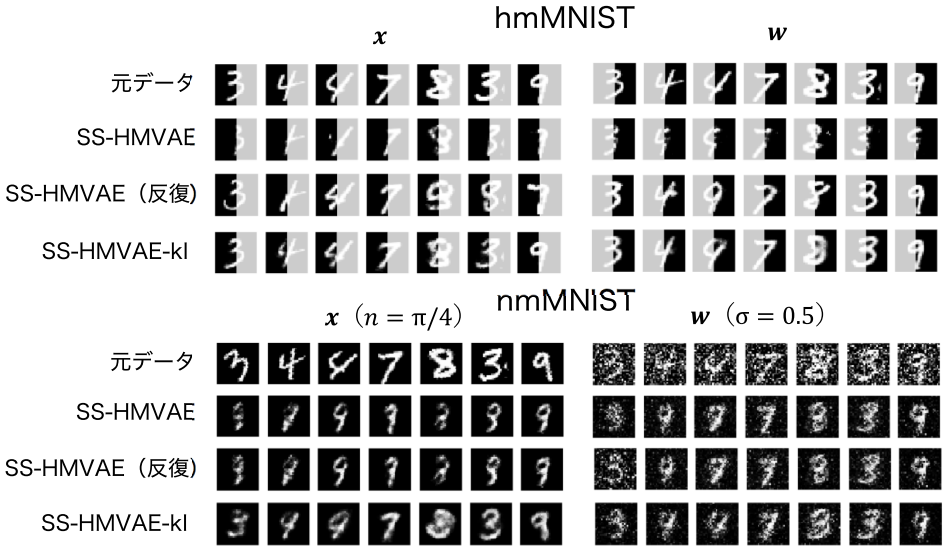

図 5　一方のモダリティから対応するモダリティを生成した結果．各データの結果の一番上の行が元データ，すなわち該当するモダリティの正しい補完結果であり，その下の行は，各手法（SS-HMVAE は単純に欠損させて推論）を使って，もう片方のモダリティの元データから生成した結果である．各モダリティの各列は，同一ペアの事例となっている

Fig. 5　Results of generating a modality from another modality. The top row on each data indicates the original data, i.e., the correct complement results of each modality. Moreover, the lower rows are the results generated from the original data of the other modality using each method (in SS-HMVAE, we simply infer from missing inputs). Each column of each modality is the same paired example.

表 5　Washington RGB-D を用いた単一モダリティおよびマルチモーダルにおける半教師あり学習．テスト集合における正解率（%）で評価．† の結果は，元論文からの引用

Table 5　Results of semi-supervised learning in single modality and multiple modalities on Washington RGB-D dataset. We evaluated them by accuracy (%) on the test set. † means that these results are quoted from the original papers.

	モデル	$\boldsymbol{x}$	$\boldsymbol{w}$	$\boldsymbol{x}+\boldsymbol{w}$
単一モダリティ	M2 [8]	85.6 ± 1.6	72.0 ± 1.7	-
	SDGM [9]	85.6 ± 1.9	75.8 ± 1.7	-
マルチモーダル	CT+SVM [34]†	78.7	75.4	83.7
	共訓練 [5]†	85.5 ± 2.0	**82.6** ± 2.3	89.2 ± 1.3
提案手法	SS-MVAE	79.6 ± 1.8	34.4 ± 7.0	89.9 ± 1.7
	SS-HMVAE	85.4 ± 2.2	41.5 ± 6.7	**90.6** ± 1.6
	SS-HMVAE（反復）	86.4 ± 2.1	54.8 ± 1.9	**90.6** ± 1.6
	SS-HMVAE-kl	**86.8** ± 2.2	81.1 ± 2.4	90.2 ± 1.4

研究での設定に近い．ただし，この研究では Depth 画像の前処理に surface normal を用いている．Surface normal とは距離情報から物体の法線ベクトルを求め，それを入力特徴量とする方法である．ただしこの方法は，先行研究 [5] に具体的な入力特徴量への変換方法が記載されておらず，また計算のための汎用的なライブラリも存在しない．したがって本研究では，surface normal よりは精度が劣ることが報告されているものの，処理が単純である jet colormap 処理を利用している [29]．

表 5 が実験結果である．それぞれの訓練・テスト分割での正解率を平均した値と標準偏差を載せている．また提案手法については，SS-MVAE, SS-HMVAE（反復サンプリングを適用した場合としなかった場合の両方），SS-HMVAE-kl（$\beta = 1$ に設定）での結果を載せた．

まず，提案手法内の結果を確認する．SS-MVAE と SS-HMVAE を比べると，MNIST での結果と同様，SS-HMVAE のほうが精度が高い結果となった．また，MNIST（特に nmMNIST）での結果と同じ傾向として，一方のモダリティ（Washington RGB-D では，Depth 画像（$\boldsymbol{w}$））について大幅に精度が低くなってしまっている．これは，Depth 画像は RGB 画像（$\boldsymbol{x}$）と比べてラベルについての情報量が小さく，その RGB 画像を欠損させたために，ラベルがうま

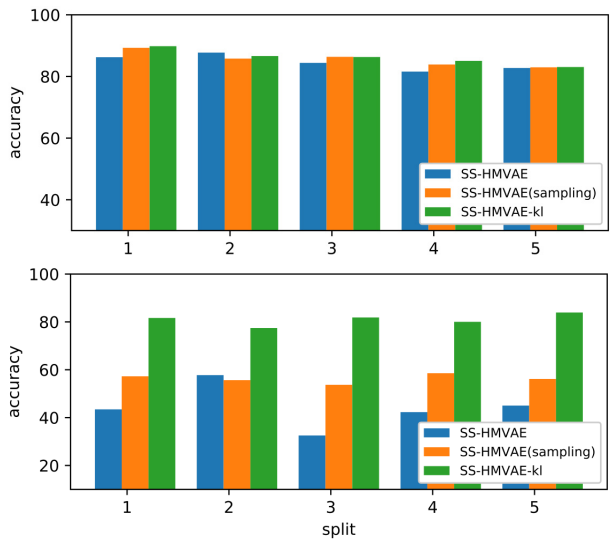

図 6　それぞれの訓練・テスト分割での RGB 画像と Depth 画像の正解率．上が RGB 画像の結果で下が Depth 画像の結果である

Fig. 6　Accuracy of RGB and depth images in each training/test split. The top figure is the result of the RGB images and the bottom is that of the depth images.

く推定されなかったと考えられる．反復サンプリング手法を適用した結果を見ると，Depth 画像の精度が向上しており，欠損補完の効果が見られるが，それでも RGB 画像の結果よりは大幅に低い．これも nmMNIST の場合と同様，情報量の違いから反復サンプリング手法でうまく補完できなかったと考えられる．

そして SS-HMVAE-kl の結果を見ると，SS-MVAE や SS-HMVAE と比べて，Depth 画像について大幅に精度が向上していることが分かる．また，**図 6** は訓練・テストの各分割における，SS-HMVAE と SS-HMVAE-kl の RGB 画像と Depth 画像の精度比較である．これを見ると，どの分割においても SS-HMVAE-kl での Depth 画像の精度が，SS-MVAE や SS-HMVAE と比べて大幅に向上していることが分かる．一方 RGB 画像での SS-HMVAE-kl の結果を見ると，SS-HMVAE の結果と比較して精度の減少はほぼ発生しておらず，むしろわずかに上昇していることが分かる．このことから，Washington RGB-D でも KL 最小化の方法の有効性が検証できたといえる．

次に，単一モダリティにおける半教師あり学習の既存手法と比較する．これらの既存手法では単一モダリティのみを入力にとるため，各モダリティで別々にモデルを用意して訓練・テストしている．

SS-HMVAE の結果を見ると，RGB 画像では既存手法よりも良い結果となっているが，Depth 画像については，大幅に精度が落ちてしまっている．その一方で，SS-HMVAE-kl については，RGB 画像および Depth 画像の両方で，M2 や SDGM を上回っていることが確認できる．この結果から，SS-HMVAE-kl では，訓練でマルチモーダルデータが与えられたときに，一方のモダリティ入力だけでも良い予測となるような知識を獲得できていることが分かる．

表 6　各設定における教師あり学習．訓練集合全体をラベルあり集合としている．テスト集合における正解率（%）で評価．† の結果は，元論文からの引用

Table 6　Results of supervised learning in each setting. In these results, we set the entire training set as a labeled set. We evaluated them by accuracy (%) on the test set. † means that these results are quoted from the original papers.

モデル	$\boldsymbol{x}$	$\boldsymbol{w}$	$\boldsymbol{x}+\boldsymbol{w}$
文献 [5] による設定 †	86.3	84.0	91.3
本研究による設定（SDGM [9]）	87.1	80.5	-
本研究による設定（SS-HMVAE）	82.4	42.8	91.4
本研究による設定（SS-HMVAE-kl）	87.2	81.1	90.3

そして，既存の半教師ありマルチモーダル学習の結果と比較する．これらの結果は，それぞれの論文の結果を引用したものである．また，単一モダリティでの既存研究とは異なり，これらの既存手法は深層生成モデルではない．

まず，マルチモーダルにおける結果を比較すると，提案手法が最も高い結果となっていることが分かる．このことから，マルチモーダル入力の半教師あり学習については，提案手法が最も優れているといえる．

各モダリティについて見ると，SS-HMVAE と SS-HMVAE-kl は，RGB 画像のときに既存手法の精度を上回っていることが分かる．その一方で，Depth 画像については，SS-HMVAE-kl でも，共訓練の結果 [5] よりわずかに低い結果となっている．この原因の 1 つとして，文献 [5] では，より精度の高い surface normal 処理を使って Depth 画像を前処理していることがあげられる．

ここで，前処理やネットワーク設定による違いを確認するために，文献 [5] での設定と本研究での設定のそれぞれで，全訓練集合をラベルあり集合として教師あり学習した場合の結果を確認する．

表 6 がその結果である．これは，それぞれの設定での半教師あり学習の精度の上限である[*12]．ここでは，本研究での設定として，SDGM，SS-HMVAE，SS-HMVAE-kl の識別モデルでの教師あり学習の結果を載せている．なお，SS-HMVAE-kl では，$q_{\boldsymbol{\lambda}}(\boldsymbol{a}|\boldsymbol{x})$ と $q_{\boldsymbol{\lambda}}(\boldsymbol{a}|\boldsymbol{w})$ が必要なため，$\mathcal{D}_{\mathcal{L}}$ 全体と式 (12) を目的関数に加えて学習する．つまり，推論モデルや生成モデルも同時に学習している．

SDGM の識別モデルの結果を見ると，RGB 画像については，本研究が VGG16 を用いて特徴抽出していることから，文献 [5] よりもわずかに高い精度となっている[*13]．しかし，Depth 画像のほうは，前処理の違いの影響によって，

*12 半教師あり学習の上限を確認する意図があるので，「本研究の設定」では，特徴抽出用ネットワークの再学習で，訓練集合すべてを用いずに，これまでと同様 5%のラベルあり集合のみを用いた．
*13 文献 [5] では AlexNet [35] を用いている

文献 [5] の設定のほうが高い結果となっている．

SS-HMVAE の識別モデルは，マルチモーダル入力のときに，文献 [5] の結果と同じくらいになることが分かる．しかし各モダリティについては，欠損問題のため，精度が落ちている．なお，ここでは識別モデルのみを学習しているので，反復サンプリングによる補完はできないことに留意されたい．

SS-HMVAE-kl の識別モデルの場合は，KL ダイバージェンスの効果によって，Depth 画像の精度が上がっているが，それでも前処理の違いの影響によって，文献 [5] の結果には及ばないことが確認できる．マルチモーダル入力の場合については，SS-HMVAE や文献 [5] よりも本研究の設定のほうが低い結果となっている．これは，SS-HMVAE-kl の識別モデルを学習する際に，$\mathcal{D}_{\mathcal{L}}$ と式 (12) の項を同時に学習している影響と考えられる．これは，SDGM や SS-HMVAE について，純粋に入力から出力の写像となる識別モデルのみを学習しているのとは異なることに留意されたい．

この SS-HMVAE-kl の結果を表 5 での結果と見比べると，半教師あり学習でも教師あり学習の場合とほぼ変わらない精度で学習できていることが分かる．このことから，半教師あり学習と KL 最小化の方法の効果が敵対するような事態は発生しておらず，SDGM [9] や文献 [5] の結果と比較して，高い性能で半教師ありマルチモーダル学習を実行できていることが分かる．

なお，文献 [5] の手法が，凸クラスタリング法を併用するなど，RGB-D データ集合に特化したアプローチであるのに対して，本研究の提案手法は，追加的な改善手法を用いておらず，汎用的な半教師ありモデルであることにも留意されたい．

4.5 欠損補完に関する考察

ここまでの実験で，提案手法が適切に半教師ありマルチモーダル学習を行えることを確認した．特に SS-HMVAE-kl は，情報量が異なるモダリティの場合，精度の低いほうを大幅に改善させることが分かった．

一方，反復サンプリング手法については，欠損モダリティの情報量が大きいときに精度が改善しないことが分かった．これは，情報量に偏りのない hmMNIST では，両方のモダリティの精度が反復サンプリングによって向上することからも実証される (表 4 および図 5 参照．ただし，KL 最小化の方法よりは精度が低い)．反復サンプリング手法の改善策の 1 つは，文献 [14] でも示されているように，潜在変数をさらに増やすことである．しかし，SS-HMVAE はすでに確率的層 (観測変数や潜在変数を層と考えた場合の層数) が 3 層になっており，深層生成モデルの確率的層は，単純に多く積み重ねるだけではうまく学習できないことが知られている [26]．さらに多層にするためには，probabilistic ladder network [26] や MatNet [36] のような手法を用いる必要がある．

5. まとめ

本稿では，半教師ありマルチモーダル学習について取り組み，深層生成モデルによる SS-MVAE と SS-HMVAE を提案した．また，入力に 1 つのモダリティしか与えられない問題設定に対処するため，SS-HMVAE を拡張した SS-HMVAE-kl を提案した．

実験ではまず，MNIST から作成した hmMNIST と nmMNIST という簡単なマルチモーダルデータを用いた．結果から，モダリティ間で情報量が異なり，関係性が決定論的でない nmMNIST では，SS-HMVAE の方が有効であることを確認した．また nmMNIST では，ラベル情報に対する情報量の大きいモダリティ入力を欠損した場合に，大きく予測精度が落ちることを確認した．さらに SS-HMVAE-kl によって，この問題が大幅に改善されることが分かった．

次に実践的なマルチモーダルデータである RGB-D データ集合を用いた実験を行い，単一モダリティおよびマルチモーダルにおける既存の半教師あり学習と比較して，高い性能で半教師あり学習を実行できることを確認した．

今後は，より大きく次元や構造が異なるマルチモーダルデータ集合を用いて検証する予定である．

参考文献

[1] Lahat, D., Adali, T. and Jutten, C.: Multimodal data fusion: An overview of methods, challenges, and prospects, *Proc. IEEE*, Vol.103, No.9, pp.1449–1477 (2015).

[2] Baltrušaitis, T., Ahuja, C. and Morency, L.-P.: Multimodal Machine Learning: A Survey and Taxonomy, arXiv preprint arXiv:1705.09406 (2017).

[3] 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).

[4] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C. and Parikh, D.: VQA: Visual question answering, *Proc. IEEE International Conference on Computer Vision*, pp.2425–2433 (2015).

[5] Cheng, Y., Zhao, X., Cai, R., Li, Z., Huang, K. and Rui, Y.: Semi-Supervised Multimodal Deep Learning for RGB-D Object Recognition (2016).

[6] Kingma, D.P. and Welling, M.: Auto-encoding variational Bayes, arXiv preprint arXiv:1312.6114 (2013).

[7] 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).

[8] 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).

[9] Maaløe, L., Sønderby, C.K., Sønderby, S.K. and Winther, O.: Auxiliary deep generative models, arXiv preprint arXiv:1602.05473 (2016).

[10] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. and Chen, X.: Improved techniques for training gans, *Advances in Neural Information Process-*

ing Systems, pp.2234–2242 (2016).
[11] Xu, W., Sun, H., Deng, C. and Tan, Y.: Variational Autoencoder for Semi-Supervised Text Classification, *AAAI*, pp.3358–3364 (2017).
[12] Yang, Z., Hu, Z., Salakhutdinov, R. and Berg-Kirkpatrick, T.: Improved variational autoencoders for text modeling using dilated convolutions, arXiv preprint arXiv:1702.08139 (2017).
[13] Suzuki, M., Nakayama, K. and Matsuo, Y.: Joint Multimodal Learning with Deep Generative Models, arXiv preprint arXiv:1611.01891 (2016).
[14] 鈴木雅大，松尾　豊：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，情報処理学会論文誌，Vol.59, No.3, pp.859–873 (2018).
[15] Guillaumin, M., Verbeek, J. and Schmid, C.: Multimodal semi-supervised learning for image classification, *2010 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.902–909, IEEE (2010).
[16] Luo, Y., Tao, D., Geng, B., Xu, C. and Maybank, S.J.: Manifold regularized multitask learning for semi-supervised multilabel image classification, *IEEE Trans. Image Processing*, Vol.22, No.2, pp.523–536 (2013).
[17] Xie, W., Lu, Z., Peng, Y. and Xiao, J.: Graph-based multimodal semi-supervised image classification, *Neurocomputing*, Vol.138, pp.167–179 (2014).
[18] Srivastava, N. and Salakhutdinov, R.R.: Multimodal learning with deep boltzmann machines, *Advances in Neural Information Processing Systems*, pp.2222–2230 (2012).
[19] 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).
[20] Vedantam, R., Fischer, I., Huang, J. and Murphy, K.: Generative Models of Visually Grounded Imagination, arXiv preprint arXiv:1705.10762 (2017).
[21] Du, C., Du, C., Li, J., Zheng, W.-L., Lu, B.-L. and He, H.: Semi-supervised Bayesian Deep Multi-modal Emotion Recognition, arXiv preprint arXiv:1704.07548 (2017).
[22] 鈴木雅大，松尾　豊：半教師ありマルチモーダル学習のための深層生成モデル，2017 年度人工知能学会全国大会（第 31 回）(2017).
[23] Rezende, D.J., Mohamed, S. and Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models, arXiv preprint arXiv:1401.4082 (2014).
[24] Jang, E., Gu, S. and Poole, B.: Categorical Reparameterization with Gumbel-Softmax, arXiv preprint arXiv:1611.01144 (2016).
[25] 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).
[26] Sønderby, C.K., Raiko, T., Maaløe, L., Sønderby, S.K. and Winther, O.: Ladder variational autoencoders, *Advances in Neural Information Processing Systems*, pp.3738–3746 (2016).
[27] Wang, W., Yan, X., Lee, H. and Livescu, K.: Deep variational canonical correlation analysis, arXiv preprint arXiv:1610.03454 (2016).
[28] 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).
[29] Eitel, A., Springenberg, J.T., Spinello, L., Riedmiller, M. and Burgard, W.: Multimodal deep learning for robust RGB-D object recognition, *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), pp.681–687, IEEE (2015).
[30] Simonyan, K. and Zisserman, A.: Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
[31] Kingma, D. and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[32] The Theano Development Team, 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).
[33] Dieleman, S., Schlüter, J., Raffel, C., Olson, E., Søderby, 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., Takács, G., de Rivaz, P., Crall, J., Sanders, G., Rasul, K., Liu, C., French, G. and Degrave, J.: Lasagne: First release (2015).
[34] Cheng, Y., Zhao, X., Huang, K. and Tan, T.: Semi-supervised learning and feature evaluation for RGB-D object recognition, *Computer Vision and Image Understanding*, Vol.139, pp.149–160 (2015).
[35] Krizhevsky, A., Sutskever, I. and Hinton, G.E.: Imagenet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems*, pp.1097–1105 (2012).
[36] Bachman, P.: An architecture for deep, hierarchical generative models, *Advances in Neural Information Processing Systems*, pp.4826–4834 (2016).
[37] Ioffe, S. and Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift, *International Conference on Machine Learning*, pp.448–456 (2015).

付　　録

A.1 確率分布とネットワーク構造の設定

ここでは，確率分布とネットワーク構造の設定について書く．

まず，モデル構造の表記方法について説明する．

出力が k ユニットの線形全結合層を `Dk` とし，それにバッチ正規化 [37] と ReLU（正規化線形関数）を加えた構造を `DkBR` と表記する．

また，ある処理 `I` のあとに別の処理 `J` を適用する場合は `I-J` と表記し，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．

A.1.1 MNIST

SS-MVAE の分布とそのネットワーク構造は以下のようにした．

- $p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{y})$，$p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{y})$（ベルヌーイ分布）
  - f_μ: `D392`（hmMNIST の場合）or `D784`（nmMNIST の場合）
  - f_{MLP}: `(z, y)-D512BR-D512BR`

- $q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})$（カテゴリ分布）
  - f_μ: D10
  - f_{MLP}: (x-D512BR-D512BR, w-D512BR-D512BR)-D512BR
- $q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})$（ガウス分布）
  - f_μ and f_{σ^2}: D64
  - f_{MLP}: (y, x-D512BR-D512BR, w-D512BR-D512BR)-D512BR

SS-HMVAE については次のように設定した．

- $p(\boldsymbol{x}|\boldsymbol{a})$，$p(\boldsymbol{w}|\boldsymbol{a})$（ベルヌーイ分布）
  - f_μ: D392（hmMNIST の場合）or D784（nmMNIST の場合）
  - f_{MLP}: a-D512BR-D512BR
- $q(\boldsymbol{y}|\boldsymbol{a})$（カテゴリ分布）
  - f_μ: D10
  - f_{MLP}: a-D512BR
- $q(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})$（ガウス分布）
  - f_μ and f_{σ^2}: D128
  - f_{MLP}: (x-D512BR-D512BR, w-D512BR-D512BR)-D512BR
- $q(\boldsymbol{a}|\boldsymbol{z},\boldsymbol{y})$（ガウス分布）
  - f_μ and f_{σ^2}: D128
  - f_{MLP}: (z, y)-D512BR-D512BR
- $q(\boldsymbol{z}|\boldsymbol{a},\boldsymbol{y})$（ガウス分布）
  - f_μ and f_{σ^2}: D64
  - f_{MLP}: (a, y)-D512BR-D512BR

さらに，SS-HMVAE-kl の各モダリティの推論モデルは次のように設定した．

- $q(\boldsymbol{a}|\boldsymbol{x})$（ガウス分布）
  - f_μ and f_{σ^2}: D128
  - f_{MLP}: x-D512BR-D512BR
- $q(\boldsymbol{a}|\boldsymbol{w})$（ガウス分布）
  - f_μ and f_{σ^2}: D128
  - f_{MLP}: w-D512BR-D512BR

A.1.2 Washington RGB-D

SS-MVAE の分布とそのネットワーク構造は以下のようにした．

- $p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{y})$，$p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{y})$（0 以上における負の平均二乗誤差）
  - f_μ: D4096
  - f_{MLP}: (z, y)-D1024BR-D1024BR
- $q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})$（カテゴリ分布）
  - f_μ: D51
  - f_{MLP}: (x-D1024BR-D1024BR, w-D1024BR-D1024BR)-D1024BR-Dropout0.5
- $q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})$（ガウス分布）
  - f_μ and f_{σ^2}: D1024
  - f_{MLP}: (y, x-D1024BR-D1024BR, w-D1024BR-D1024BR)-D1024BR

ただし，「0 以上における負の平均二乗誤差」は，対数尤度の評価として，f_μ に対して Softplus を適用して 0 以上にしたあと，負の平均二乗誤差を計算するものである．また，DropoutRate はドロップアウト率 Rate のドロップアウト層である．

SS-HMVAE については次のように設定した．

- $p(\boldsymbol{x}|\boldsymbol{a})$，$p(\boldsymbol{w}|\boldsymbol{a})$（0 以上における負の平均二乗誤差）
  - f_μ: D1024
  - f_{MLP}: a-D1024BR-D1024BR
- $q(\boldsymbol{y}|\boldsymbol{a})$（カテゴリ分布）
  - f_μ: D51
  - f_{MLP}: a-D1024BR-Dropout0.5
- $q(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})$（ガウス分布）
  - f_μ and f_{σ^2}: D1024
  - f_{MLP}: (x-D1024BR, w-D1024BR)
- $q(\boldsymbol{a}|\boldsymbol{z},\boldsymbol{y})$（ガウス分布）
  - f_μ and f_{σ^2}: D1024
  - f_{MLP}: (z-D1024BR, y-D1024BR)
- $q(\boldsymbol{z}|\boldsymbol{a},\boldsymbol{y})$（ガウス分布）
  - f_μ and f_{σ^2}: D1024
  - f_{MLP}: (a-D1024BR, y-D1024BR)

さらに，SS-HMVAE-kl の各モダリティの推論モデルは次のように設定した．

- $q(\boldsymbol{a}|\boldsymbol{x})$（ガウス分布）
  - f_μ and f_{σ^2}: D1024
  - f_{MLP}: x-D1024BR-D1024BR
- $q(\boldsymbol{a}|\boldsymbol{w})$（ガウス分布）
  - f_μ and f_{σ^2}: D1024
  - f_{MLP}: w-D1024BR-D1024BR

以上のネットワーク構造については，各モダリティについて同じ構造になるように，そして各層のユニット数が同じになるように設計した．RGB-D におけるユニット数については，先行研究 [5], [29] では 4,096 となっているが，計算量を削減する意図から，他のマルチモーダル学習研究 [18] でも設定されている 1,024 とした．

鈴木 雅大 （正会員）

2013 年北海道大学工学部卒業．2015 年同大学大学院修士課程修了．2018 年東京大学大学院工学系研究科博士課程修了．博士（工学）．2018 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特任研究員．人工知能，深層学習の研究に従事．

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 年人工知能学会理事・編集委員長，2014 年倫理委員長．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．