The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

半教師ありマルチモーダル深層生成モデルにおける共有表現の有効性と単一モダリティ入力への拡張

The Effectiveness of Joint Representation and the Extension to Unimodal Input on Semi-Supervised Multimodal Deep Generative Models

鈴木 雅大 *1　　松尾 豊 *1
Masahiro Suzuki　　Yutaka Matsuo

*1東京大学工学系研究科技術経営戦略学専攻
Graduate School of Technology Management for Innovation, the University of Tokyo

In recent multimodal learning, deep neural networks are increasingly used as discriminators. In general, we need a large amount of labeled dataset for training them, but it takes a human cost to label multimodal inputs. Therefore, semi-supervised learning on multimodal data becomes important. Among these methods, semi-supervised multimodal learning with deep generative models has recently been proposed. In this study, we first compare these methods and show that SS-HMVAE, which is a method with latent variables corresponding to joint representation, have high performance when different modalities have no deterministic relation in particular. Next, to predict labels from a unimodal data, we propose SS-HMVAE-kl that is an extended model of SS-HMVAE. We confirmed that this method greatly improves the performance when inputting a single modality compared with the conventional models.

1. はじめに

近年，画像や音声など複数の異なる情報，すなわちマルチモーダルな情報を入力として扱うマルチモーダル学習（multimodal learning）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．この問題設定が注目されているのは，異なる情報（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，ラベルを予測する上で他のモダリティにはない付加価値が得られるからである [Lahat 15]．一般に，融合設定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ことが多いが，学習のためにラベルが大量に必要となるという課題から，半教師あり学習手法を利用したマルチモーダル学習，すなわちマルチモーダル半教師あり学習が研究されている．

従来の半教師ありマルチモーダル学習は，共訓練（co-training）を用いた方法が主流であった *1．一方で近年，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おり，深層生成モデルを用いたマルチモーダル半教師あり学習手法も提案されている [Du 17,Suzuki 17]．しかし，深層生成モデルを用いた手法については，入力モダリティの性質に応じて，どのようなモデルが有効かについて充分に検証されていない．また，共訓練に基づく手法では，各モダリティに対する識別器があるため，単一のモダリティからラベルを予測できるが，深層生成モデルによる手法では全モダリティが与えられる場合しか想定していないという課題があった．

こうした背景から，本研究では次の 2 つについて取り組む．1) 深層生成モデルによる半教師ありマルチモーダル学習を比較し，SS-HMVAE が最も高い性能となることを確認する．SS-HMVAE は，異なるモダリティ情報を統合した共有表現に該当する潜在変数を含んだモデルである．特に異なるモダリティ間に決定論的な関係性がないようなマルチモーダルデータにおいて，SS-HMVAE のような共有表現を明示的に含んだモデルが有効であることを確認する．2) テスト時に単一モダリティしか与えられない場合に対処するため，マルチモーダル深層生成モデルの欠損補完手法 [Suzuki 18] に基づき，SS-HMVAE を拡張する形で，SS-HMVAE-kl というモデルを提案する．

連絡先: 鈴木雅大，東京大学工学系研究科技術経営戦略学専攻，〒 113-8656 東京都文京区本郷 7-3-1，masa@weblab.t.u-tokyo.ac.jp

*1　この文脈では，multi-view learning と呼ばれる事が多い．

2. 関連研究

近年，深層生成モデルを用いたマルチモーダル学習の研究が進められており，特に異なるモダリティの共有表現を獲得するモデルとしては，variational autoencoder（VAE） [Kingma 13] を用いた方法が主流である [Suzuki 18, Vedantam 17]．最も単純なものが，JMVAE [Suzuki 18] である *2．また VAE は，半教師あり学習にも優れた性能を示すことが知られている．M2 モデル [Kingma 14a] が最初に提案され，それを拡張する形で ADGM と SGDM [Maaløe 16] が提案されている．

こうしたことから，近年，VAE を用いた半教師ありマルチモーダル学習の手法が提案されている．Du らは，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として semiMVAE を提案している [Du 17]．この手法は M2 モデルをマルチモーダル入力に拡張したものである．また同時期に，筆者らは semiMVAE とほぼ同じモデルである SS-MVAE の他，マルチモーダル入力とラベルの間に潜在変数を追加した SS-HMVAE を提案している [Suzuki 17]．文献 [Suzuki 17] では，SS-HMVAE の方が高い性能となる傾向が確認されたものの，マルチモーダル入力がどのような場合に有効性が示されるのかについては充分に検証されていない．また，これらの手法はいずれもマルチモーダル入力を想定しているため，単一のモダリティからラベルを予測することができない．

3. 問題設定

データ集合 $\mathcal{D}_{\mathcal{L}} = \{(\boldsymbol{x}_1, \boldsymbol{w}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_N, \boldsymbol{w}_N, \boldsymbol{y}_N)\}$ 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

*2　論文によって，joint VAE [Vedantam 17] や multi-view VAE [Du 17] など，呼び方が異なっている．

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 $(\boldsymbol{x}_n, \boldsymbol{w}_n, \boldsymbol{y}_n)$ 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 $\mathcal{D}_\mathcal{U} = \{(\boldsymbol{x}_1, \boldsymbol{w}_1), ..., (\boldsymbol{x}_M, \boldsymbol{w}_M)\}$（ただし $N << M$）が与えられた下で，$\boldsymbol{x}$ 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 $p(\boldsymbol{y}|\boldsymbol{x})$ 及び $p(\boldsymbol{y}|\boldsymbol{w})$ も獲得することである．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．

4. 半教師ありマルチモーダル VAE

本節では，鈴木らが提案した半教師ありマルチモーダル VAE である SS-HMVAE 及びその他の半教師ありマルチモーダル VAE について簡単に説明する．

4.1 SS-HMVAE

モダリティ $\boldsymbol{x}$，$\boldsymbol{w}$ とラベル $\boldsymbol{y}$ の生成過程を $\boldsymbol{y} \sim p(\boldsymbol{y}) = \mathrm{Cat}(\boldsymbol{y}; \boldsymbol{\pi})$, $\boldsymbol{z} \sim p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, $\boldsymbol{a} \sim p_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{z}, \boldsymbol{y})$, $\boldsymbol{x}, \boldsymbol{w} \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{w}|\boldsymbol{a})$ とする．ただし，$\boldsymbol{a}$ と $\boldsymbol{z}$ は潜在変数である．また $\boldsymbol{\theta}$ はモデルパラメータであり，深層生成モデルでは各分布を深層ニューラルネットワークでパラメータ化する．このとき，全モダリティとラベルの同時分布は $p(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{y}) = \int\int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{a}) p_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{a}) p_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{z}, \boldsymbol{y}) p(\boldsymbol{z}) p(\boldsymbol{y}) d\boldsymbol{a} d\boldsymbol{z}$ で与えられる．

この同時分布の下界は以下の式で与えられる．

$$\begin{aligned}
&\log p(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{y}) \\
&\geq E_{q_{\boldsymbol{\phi}}(\boldsymbol{a}, \boldsymbol{z}|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{y})}[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{a}) p_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{a}) p_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{z}, \boldsymbol{y}) p(\boldsymbol{z}) p(\boldsymbol{y})}{q_{\boldsymbol{\phi}}(\boldsymbol{a}, \boldsymbol{z}|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{y})}] \\
&\equiv \mathcal{L}(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{y}). \qquad (1)
\end{aligned}$$

ただし，$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{y}) = \int q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{a}|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{y}) d\boldsymbol{a} = \int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{a}, \boldsymbol{y}) q_{\boldsymbol{\phi}}(\boldsymbol{a}|\boldsymbol{x}, \boldsymbol{w}) d\boldsymbol{a}$ は推論モデルであり，ϕ をパラメータとしてもつ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．

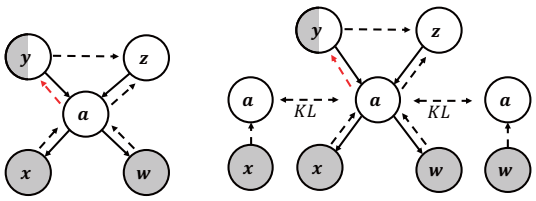
これを目的関数とみなしてラベルあり集合 $\mathcal{D}_\mathcal{L}$ において，パラメータ θ, ϕ について最大化することで各モデルを学習できる．なお ϕ が期待値部分に含まれるため，推論モデルに対して再パラメータ化トリック [Kingma 13] を用いることで勾配を求める．

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ため，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 $q_{\boldsymbol{\phi}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = E_{q_{\boldsymbol{\phi}}(\boldsymbol{a}|\boldsymbol{x}, \boldsymbol{w})}[q_{\boldsymbol{\phi}}(\boldsymbol{y}|\boldsymbol{a})]$ を導入して次のように求められる．

$$\begin{aligned}
&\log p(\boldsymbol{x}, \boldsymbol{w}) \\
&\geq E_{q_{\boldsymbol{\phi}}(\boldsymbol{a}, \boldsymbol{z}, \boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})}[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{a}) p_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{a}) p_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{z}, \boldsymbol{y}) p(\boldsymbol{z}) p(\boldsymbol{y})}{q_{\boldsymbol{\phi}}(\boldsymbol{a}, \boldsymbol{z}, \boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})}] \\
&\equiv \mathcal{U}(\boldsymbol{x}, \boldsymbol{w}). \qquad (2)
\end{aligned}$$

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．なお識別モデルは離散分布となるため再パラメータ化トリックを適用できないが，Gumbel-softmax [Jang 16] を用いて，近似的に再パラメータ化する．

*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（一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．



(a) SS-HMVAE (b) SS-HMVAE-kl

図 1: SS-HMVAE と SS-HMVAE-kl のグラフィカルモデル．黒丸は観測変数，白丸は潜在変数を表す．

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 $\mathcal{J}_{HMVAE}$ は，

$$\begin{aligned}
&\mathcal{J}_{HMVAE} \\
&= \frac{1}{N} \sum_{(\boldsymbol{x}_n, \boldsymbol{w}_n, \boldsymbol{y}_n) \in \mathcal{D}_\mathcal{L}} \mathcal{L}_l(\boldsymbol{x}_n, \boldsymbol{w}_n, \boldsymbol{y}_n) + \frac{1}{M} \sum_{(\boldsymbol{x}_j, \boldsymbol{w}_j) \in \mathcal{D}_\mathcal{U}} \mathcal{U}(\boldsymbol{x}_j, \boldsymbol{w}_j) \qquad (3)
\end{aligned}$$

となる．SS-HMVAE のグラフィカルモデルを図 1(a) で示す．

ここで，潜在変数 $\boldsymbol{a}$ について着目する．この変数は，推論モデル $q_{\boldsymbol{\phi}}(\boldsymbol{a}|\boldsymbol{x}, \boldsymbol{w})$ によって，マルチモーダル入力から推論され，もう一つの潜在変数を推論する入力 $q_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{y}|\boldsymbol{a})$ として使われる．また生成モデルでも，ラベルから $\boldsymbol{a}$ を介してマルチモーダルデータが生成される形となっている．したがって，各モデルが適切に学習されれば，$\boldsymbol{a}$ には複数のモダリティを統合した共有表現が獲得されることになる．実際，文献 [Suzuki 18] でも示されているように，VAE に基づくマルチモーダル学習は，対応関係が決定論的でないマルチモーダル入力からでも共有表現を獲得することができる．

これは，ラベルを予測する上で大きな効果があると考えられる．SS-HMVAE の識別モデルが $q_{\boldsymbol{\phi}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = E_{q_{\boldsymbol{\phi}}(\boldsymbol{a}|\boldsymbol{x}, \boldsymbol{w})}[q_{\boldsymbol{\phi}}(\boldsymbol{y}|\boldsymbol{a})]$ という形になっていることから，マルチモーダル情報を統合した共有表現からラベルを予測することができる．また，訓練時には少数のラベルあり集合を用いて識別モデルも同時に学習することから，JMVAE のような教師なしモデルよりもラベルの情報を含んだ良い共有表現が獲得できることが期待される．

4.2 SS-MVAE，semiMVAE

SS-MVAE と semiMVAE は，同じ生成モデル $p(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{y}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y}) p_{\boldsymbol{\theta}}(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{y}) p(\boldsymbol{z}) p(\boldsymbol{y}) d\boldsymbol{z}$，推論分布 $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{y})$，及び識別モデル $q_{\boldsymbol{\phi}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})$ をもつ．唯一の違いは，semiMVAE では推論分布を $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{y}) = \lambda_{\boldsymbol{x}} q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}) + \lambda_{\boldsymbol{w}} q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{w}, \boldsymbol{y})$ のような混合分布と考えている点である．本研究の実験では，SS-MVAE の方を用いる．

SS-HMVAE と比較すると，これらのモデルには共有表現にあたる $\boldsymbol{a}$ が存在しない．したがって，これらのモデルでは識別モデルにおける $\boldsymbol{x}$ と $\boldsymbol{w}$ を結合する部分が深層ニューラルネットワークによって決定論的になるため，$\boldsymbol{x}$ と $\boldsymbol{w}$ の分布の違いに適切に対処できない可能性がある．

5. 提案手法：単一モダリティ入力に対処した半教師ありマルチモーダル VAE

4 章で述べた深層生成モデルの識別モデルはすべて，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 章の問題設定でも示したように，単一のモダリティデータから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．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，決定論的な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．

こうした問題の解決方法の一つが反復サンプリング手法 [Rezende 14] の利用である．これは，入力の欠損値を潜在変数の間との再構成を繰り返すことで補完する方法である．SS-HMVAE は共有表現を含んでいるので，入力と共有表現の間で再構成を繰り返して欠損したモダリティを補完し，それを識別モデルの入力とする．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場合があることが指摘されている [Suzuki 18]．そのため本稿では，単一モダリティにおける識別モデルを明示的に学習する方法として，SS-HMVAE を拡張したモデルである **SS-HMVAE-kl** を提案する．

SS-HMVAE-kl のアイディアは，JMVAE の欠損問題への対処方法として提案された JMVAE-kl [Suzuki 18][*5] と同じである．

まず，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．そして，SS-HMVAE の推論モデル $q_{\boldsymbol{\lambda}}(\boldsymbol{a}|\boldsymbol{x},\boldsymbol{w})$ 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）を近づけるように学習する．したがって，SS-HMVAE-kl の目的関数 $\mathcal{J}_{kl}$ は，

$$\mathcal{J}_{kl} = \mathcal{J}_{HMVAE} - \frac{\beta}{M+N}\mathcal{J}_{div} \tag{4}$$

となる．ただし $\boldsymbol{\lambda}$ は各モダリティの推論モデルのモデルパラメータ，β は KL 項の影響を調節するパラメータ，そして

$$\mathcal{J}_{div} = \sum_{(\boldsymbol{x}_n,\boldsymbol{w}_n)\in\mathcal{D}_{\mathcal{L}}\bigcup\mathcal{D}_{\mathcal{U}}} [D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{a}|\boldsymbol{x}_n,\boldsymbol{w}_n)||q_{\boldsymbol{\lambda}}(\boldsymbol{a}|\boldsymbol{x}_n)) + D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{a}|\boldsymbol{x}_n,\boldsymbol{w}_n)||q_{\boldsymbol{\lambda}}(\boldsymbol{a}|\boldsymbol{w}_n))] \tag{5}$$

である．この式を最適化することで，SS-HMVAE のモデル及び各モダリティの推論モデルを同時に学習できる．適切に各モダリティの推論モデルが学習できれば，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．SS-HMVAE-kl のグラフィカルモデルを図 1(b) で示す．

6. 実験

6.1 設定

本実験のために，手書き数字のデータ集合 MNIST から次の 2 つの方法でマルチモーダルデータ集合を作成した．1) MNIST を左右に分割し，それぞれを異なるモダリティ（$\boldsymbol{x}$ と $\boldsymbol{w}$）と見立てる．本研究では half multimodal MNIST（hmMNIST）と呼ぶ．2) MNIST に 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とする．MNIST にガウスノイズ（平均 0，標準偏差 0.3）を付与したものを 1 つめのモダリティ$\boldsymbol{x}$ とし，各事例ごとに $[-\pi/4,\pi/4]$ の範囲でランダムに回転させたものを 2 つめのモダリティ$\boldsymbol{w}$ とする．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る．また，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．一方 nmMNIST の場合は，結合しただけでは冗長な表現になる上，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いない．したがって，nmMNIST は hmMNIST よりも困難な設定である．

訓練集合を 60,000，テスト集合を 10,000 とし，訓練集合のうちの 500 をラベルあり集合として，残りをラベルなし集合とした．ネットワークの各層の活性化関数には rectified linear unit を用い，各層でバッチ正則化 [Ioffe 15] を行った（その他詳しいネットワークの設定は省略する）．最適化アルゴリズムに Adam [Kingma 14b] を利用した．実装は深層生成モデルライブラリ Tars[*6] を使用した．訓練集合で学習した識別モデルをテスト集合で検証し，正解率で評価する．

6.2 実験結果

表 1 は，既存の深層生成モデルによる半教師ありマルチモーダル学習（SS-MVAE，SS-HMVAE）と，本研究で提案する SS-HMVAE-kl の，各データ集合における正解率を示している．

まず既存手法において，$\boldsymbol{x}+\boldsymbol{w}$，すなわちマルチモーダル入力の場合に SS-HMVAE が最も高い精度となっている．特に，nmMNIST の場合に大きな精度向上となっている．このことから，モダリティ間に決定論的な関係性がないデータに対して，共有表現を持つ SS-HMVAE が有効であることがわかる．

ここで，共有表現の分布を可視化する．図 2 が，nmMNIST における共有表現の可視化結果である．各色が数字ラベルに対応している．また，比較のために JMVAE [Suzuki 18] で獲得した共有表現も載せている．JMVAE は各ラベルについてある程度は分離しているものの，教師なしモデルであることもあり，十分ではない．一方 SS-HMVAE は，JMVAE よりもラベルごとに分離した分布が獲得されている．これは，SS-HMVAE が少数のラベルを元に共有表現を向上させていることを示している．また，JMVAE に比べて大きく広がった分布となっているが，これは JMVAE のように標準ガウス分布に近づけるという制約がないためである．

次に SS-HMVAE-kl の結果をみると，各単一モダリティ（$\boldsymbol{x}$,$\boldsymbol{w}$）において，既存の半教師ありマルチモーダルモデルを大幅に上回っていることがわかる．SS-HMVAE で反復サンプリングを適用した場合も大幅に精度が向上しているが，それでも SS-HMVAE-kl のほうが高い精度となっている．このことから，SS-HMVAE-kl によって単一モダリティ入力の識別器も適切に学習できることがわかった．なお，マルチモーダル入力については SS-HMVAE より精度が落ちていることから，単一モダリティとマルチモーダルの精度はトレードオフの関係になっていることがわかる．

7. まとめ

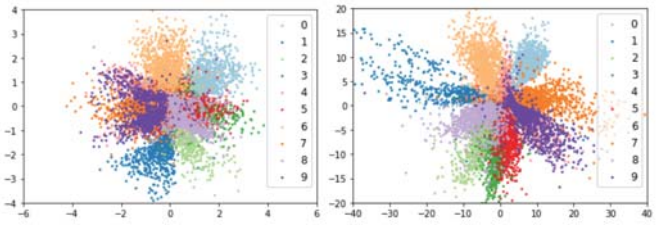
本稿では，深層生成モデルを用いた半教師ありマルチモーダル学習について検証した．まず，既存のマルチモーダル半教師あり学習の中でも，共有表現を含んだ SS-HMVAE が高い精度となることを確認した．特に決定論的な関係がない異なるモダリティを入力とした場合に大幅な向上が見られた．また，単一モダリティ入力に対処するため，SS-HMVAE を拡張した SS-HMVAE-kl を提案した．実験から，従来の半教師ありマル

*5 文献 [Vedantam 17] では，JMVAE-kl のことを単に JMVAE と呼んでいる．

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

表 1: 半教師ありマルチモーダル学習の比較．テスト集合における正解率（%）で評価．

	hmMNIST			nmMNIST		
モデル	$\boldsymbol{x}$	$\boldsymbol{w}$	$\boldsymbol{x}+\boldsymbol{w}$	$\boldsymbol{x}$	$\boldsymbol{w}$	$\boldsymbol{x}+\boldsymbol{w}$
SS-MVAE	64.45	72.91	96.20	70.00	46.42	84.45
SS-HMVAE	71.16	56.90	**97.02**	61.48	31.86	**94.82**
SS-HMVAE（反復サンプリング）	85.12	84.16	**97.02**	87.72	87.86	**94.82**
SS-HMVAE-kl（$\beta = 0.1$）	**94.11**	**93.47**	96.74	93.17	**93.97**	94.60
SS-HMVAE-kl（$\beta = 1$）	93.93	93.03	96.47	**93.35**	93.87	94.57
SS-HMVAE-kl（$\beta = 10$）	92.93	92.25	95.01	92.37	92.70	93.56



(a) JMVAE　(b) SS-HMVAE

図 2: nmMNIST での JMVAE と SS-HMVAE における共有表現の可視化．共有表現に該当する潜在変数の次元を 2 次元にして訓練したのち推論したものをプロットしている（軸のスケールが異なることに注意されたい）．

チモーダルモデルよりも高い精度で，単一モダリティからラベルを予測できることがわかった．

今回は，検証のため簡単なデータ集合に留めたが，今後はより複雑で大規模なデータ集合に対して検証したい．

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

[Du 17] 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)

[Guillaumin 10] Guillaumin, M., Verbeek, J., and Schmid, C.: Multimodal semi-supervised learning for image classification, in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 902–909IEEE (2010)

[Ioffe 15] Ioffe, S. and Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift, *arXiv preprint arXiv:1502.03167* (2015)

[Jang 16] Jang, E., Gu, S., and Poole, B.: Categorical Reparameterization with Gumbel-Softmax, *arXiv preprint arXiv:1611.01144* (2016)

[Kingma 13] Kingma, D. P. and Welling, M.: Auto-encoding variational bayes, *arXiv preprint arXiv:1312.6114* (2013)

[Kingma 14a] Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M.: Semi-supervised learning with deep generative models, in *Advances in Neural Information Processing Systems (NIPS)*, pp. 3581–3589 (2014)

[Kingma 14b] Kingma, D. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014)

[Lahat 15] Lahat, D., Adali, T., and Jutten, C.: Multimodal data fusion: an overview of methods, challenges, and prospects, *Proceedings of the IEEE*, Vol. 103, No. 9, pp. 1449–1477 (2015)

[Maaløe 16] Maaløe, L., Sønderby, C. K., Sønderby, S. K., and Winther, O.: Auxiliary deep generative models, *arXiv preprint arXiv:1602.05473* (2016)

[Maddison 16] Maddison, C. J., Mnih, A., and Teh, Y. W.: The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables, *arXiv preprint arXiv:1611.00712* (2016)

[Rezende 14] Rezende, D. J., Mohamed, S., and Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models, *arXiv preprint arXiv:1401.4082* (2014)

[Salimans 16] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X.: Improved techniques for training gans, in *Advances in Neural Information Processing Systems*, pp. 2226–2234 (2016)

[Suzuki 17] Suzuki, M. and Matsuo, Y.: Deep Generative Models for Semi-Supervised Multimodal Learning, *The 31st Annual Conference of the Japanese Society for Artificial Intelligence, 2017* (2017)

[Suzuki 18] Suzuki, M., Nakayama, K., and Matsuo, Y.: Improving Bi-directional Generation between Different Modalities with Variational Autoencoders, *arXiv preprint arXiv:1801.08702* (2018)

[Vedantam 17] Vedantam, R., Fischer, I., Huang, J., and Murphy, K.: Generative Models of Visually Grounded Imagination, *arXiv preprint arXiv:1705.10762* (2017)