1A1-OS-05a-5

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

Deep Generative Models for Semi-Supervised Multimodal Learning

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 multimodal learning settings, various modal information can be obtained at relatively low cost. However, obtaining labels is time-consuming because these information need to be labeled by human beings. In order to solve this issue, we propose novel semi-supervised multimodal learning models with deep generative models, SS-MVAE and SS-HMVAE. We validate semi-supervised learning by the proposed models through experiments using multimodal datasets.

1. はじめに

マルチモーダル学習は，複数のモーダル情報を入力として取り入れることで，単一のモーダル情報の場合よりも精度の良い識別を行う手法である．例えば，ロボットの物体認識タスクにおいては，画像だけでなく音声情報やセンサ情報など，様々なモーダル情報を得られるため，それらを識別モデルの学習に利用することができる．しかし，認識している物体が何か，即ち物体のラベル情報は通常人手で付与する必要があり，人的にも時間的にもコストがかかってしまう．半教師あり学習は，人間が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学習に用いることで，こうしたコストを削減し，かつ識別モデルの汎化性能を高める手法である．これまでもマルチモーダル情報を使った半教師あり学習の手法は複数提案されている [Guillaumin 10, Cheng 16] が，本研究では，深層生成モデルを用いた新たなマルチモーダル半教師あり学習の手法を検討する．

深層生成モデルは，データの生成過程を明示的あるいは暗黙的に深層ニューラルネットワークによってモデル化する手法であり，variational autoencoder（VAE）[Kingma 13] や generative adversarial nets（GAN）[Goodfellow 14] などが提案されている．これらのモデルは，画像の生成過程を学習し，新たな画像を生成できることで知られているが，半教師あり学習の手法としても注目されており，従来の半教師あり学習と比較して高い汎化性能が得られることが報告されている [Kingma 14a, Maaløe 16, Salimans 16]．これまで深層生成モデルによるマルチモーダルデータの半教師あり学習は提案されていないが，筆者らはマルチモーダルデータを扱える深層生成モデルとして joint multimodal variational autoencoder（JMVAE）[Suzuki 16][鈴木 16] を提案している．

本研究は，マルチモーダルデータを使って end-to-end に学習可能な，深層生成モデルを用いた半教師あり学習の手法を提案する．具体的には，JMVAE をベースに半教師あり学習のモデルとして，semi-supervised MVAE（SS-MVAE）と semi-supervised HMVAE（SS-HMVAE）をそれぞれ提案する．

本稿の構成は次の通りである．まず，2 章でマルチモーダルデータを用いた半教師あり学習及び深層生成モデルによる半教師あり学習の先行研究について述べる．そして 3 章で本稿で提案モデルについて説明する．4 章では，半教師ありマルチモーダル学習の実験と考察をする．最後に 5 章でまとめと今後の展望について述べる．

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

2. 先行研究

本章では，マルチモーダルデータを使った半教師あり学習の先行研究と，深層生成モデルによる半教師あり学習の先行研究について述べる．

2.1 マルチモーダルデータによる半教師あり学習

Guillaumin らは，画像とタグからなるマルチモーダルデータに少数のラベルしかない場合に，ラベルのないマルチモーダルデータを用いて，画像からラベルをより高精度に予測する半教師あり学習の枠組みを提案した [Guillaumin 10]．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（multiple kernel learning, MKL）によって，ラベルありデータを用いてラベルを予測する分類器を学習し，その分類器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，ラベルありデータ及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一つの分類器を学習する，という流れである．

その他には，Cheng らによって RGB-D データを使っ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 [Cheng 16]．

これらの手法はメタ的な学習アルゴリズムと捉えることもでき，学習が複数の段階に分かれているのが特徴である．本研究で提案するモデルは，ラベルありデータとラベルなしデータを用いて，生成モデルと識別モデルを 1 つの目的関数で同時に学習することができる．

2.2 深層生成モデルによる半教師あり学習

Kingma らは，深層生成モデルである variational autoencoder（VAE）[Kingma 13] を半教師あり学習に適用した手法を提案した [Kingma 14a]．その後も，VAE もしくは generative adversarial nets（GAN）を用いた半教師あり学習が提案されており [Kingma 14a, Maaløe 16, Salimans 16]，半教師あり学習において深層生成モデルが有用であることが示されている．

3. 提案手法

3.1 問題設定

ラベルあり訓練事例集合 $\mathcal{D}_{\mathcal{L}} = \{(\mathbf{x}_1, \mathbf{w}_1, \mathbf{y}_1,), ..., (\mathbf{x}_M, \mathbf{w}_M, \mathbf{y}_M)\}$ が与えられているとする．ここで，$\mathbf{x}_i$ と $\mathbf{w}_i$ は同じカテゴリを表す異なるモーダル情報であり，それぞれ異なる次元や構造を持つものとし，$\mathbf{y} \in \{0,1\}^C$ は対応するカテゴリのラベル情報を表す [*1]．また，その他にラベルなし訓練事例集合 $\mathcal{D}_{\mathcal{U}} = \{(\mathbf{x}_1, \mathbf{w}_1, \mathbf{y}_1,), ..., (\mathbf{x}_N, \mathbf{w}_N)\}$ も与えられているとする．本研究では，$M << N$ のときに，より適切なマルチモーダル学習，すなわち汎化性能の高い識別モデル $q(\mathbf{y}|\mathbf{x}, \mathbf{w})$ を得ることが目標である．

3.2 JMVAE

本節では，マルチモーダルデータの教師なし学習モデルである joint multimodal variational autoencoder（JMVAE）[Suzuki 16][鈴木 16] について説明する．

潜在変数 $\mathbf{z}$ が与えられたとき，モーダル情報 $\mathbf{x}$ と $\mathbf{w}$ を観測変数とすると，同時分布は $p(\mathbf{x}, \mathbf{w}) = \int p_\theta(\mathbf{x}|\mathbf{z}) p_\theta(\mathbf{w}|\mathbf{z}) p_\theta(\mathbf{z}) d\mathbf{z}$ となる．VAE 及び JMVAE では，生成モデルはパラメータ θ をもつ深層ニューラルネットワークでモデル化される．パラメータ θ は変分下界 $-\mathcal{U}_{JMVAE}(\mathbf{x}, \mathbf{w})$ を最大化することで学習でき，その下界は次のようになる．

$$\begin{aligned} \log p(\mathbf{x}, \mathbf{w}) &= \log \int p_\theta(\mathbf{x}, \mathbf{w}, \mathbf{z}) d\mathbf{z} \\ &\geq E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})}[\log \frac{p_\theta(\mathbf{x}|\mathbf{z}) p_\theta(\mathbf{w}|\mathbf{z}) p_\theta(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{w})}] \\ &\equiv -\mathcal{U}_{JMVAE}(\mathbf{x}, \mathbf{w}) \end{aligned} \quad (1)$$

ただし，$q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{w})$ は近似事後確率であり，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推論モデルとしてモデル化される．推論モデルが期待値をとる部分にあるため，式 1 の ϕ に関する勾配はそのままでは求められないが，推論モデルがガウス分布 $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2))$ ならば，$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma}^2 \odot \boldsymbol{\epsilon}$（ただし $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$）とすることで，勾配が期待値の中に入り，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 [Kingma 14a, Rezende 14]．なお本研究の提案モデルでは，推論モデルがベルヌーイ分布となることがある（式 12 を参照）が，この場合は Gumbel softmax[Jang 16, Maddison 16] によって再パラメータ化する．したがって，ネットワークのパラメータ ϕ と θ は確率的勾配降下法によって式 1 を最大化することで，同時に学習できる．

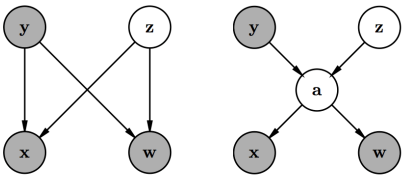
3.3 半教師あり学習モデル：SS-MVAE

次に，本稿の提案手法である，マルチモーダルデータによる半教師あり学習モデルについて説明する．

JMVAE は教師なしモデルだが，半教師あり学習ではラベル情報 $\mathbf{y}$ を含んだ生成モデルを考える．本研究では生成モデルを $p(\mathbf{x}, \mathbf{w}, \mathbf{y}) = \int p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{y}) p_\theta(\mathbf{w}|\mathbf{z}, \mathbf{y}) p(\mathbf{z}) p(\mathbf{y}) d\mathbf{z}$ とする．図 1(a) は，この生成過程をグラフィカルモデルで表したものである．このとき，変分下界は以下のようになる．

$$\begin{aligned} \log p(\mathbf{x}, \mathbf{w}, \mathbf{y}) &= \log \int p_\theta(\mathbf{x}, \mathbf{w}, \mathbf{z}, \mathbf{y}) d\mathbf{z} \\ &\geq E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w},\mathbf{y})}[\log \frac{p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{y}) p_\theta(\mathbf{w}|\mathbf{z}, \mathbf{y}) p_\theta(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{w}, \mathbf{y})}] \\ &\equiv -\mathcal{L}(\mathbf{x}, \mathbf{w}, \mathbf{y}) \end{aligned} \quad (2)$$

*1 本研究ではラベル情報として C カテゴリのマルチラベルを想定しているため，ベクトル表記としている．



(a) SS-MVAE (b) SS-HMVAE

図 1: 提案手法のグラフィカルモデル

この下界は，ラベルあり訓練事例集合によって学習する．一方で，ラベルなし訓練事例集合を学習に用いるには，ラベル情報を含まない下界を求める必要がある．これは，ラベル情報の識別モデル $q_\phi(\mathbf{y}|\mathbf{x}, \mathbf{w})$ を導入して次のように求められる．

$$\begin{aligned} \log p(\mathbf{x}, \mathbf{w}) &= \log \int \int p_\theta(\mathbf{x}, \mathbf{w}, \mathbf{z}, \mathbf{y}) d\mathbf{z} d\mathbf{y} \\ &\geq E_{q_\phi(\mathbf{z},\mathbf{y}|\mathbf{x},\mathbf{w})}[\log \frac{p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{y}) p_\theta(\mathbf{w}|\mathbf{z}, \mathbf{y}) p_\theta(\mathbf{z})}{q_\phi(\mathbf{z}, \mathbf{y}|\mathbf{x}, \mathbf{w})}] \\ &\equiv -\mathcal{U}(\mathbf{x}, \mathbf{w}) \end{aligned} \quad (3)$$

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

本研究の目標は，汎化性能の高い識別モデル $q_\phi(\mathbf{y}|\mathbf{x})$ を求めることであるが，ラベルあり事例集合の目的関数である式 2 には，識別モデルに関する部分が存在しない．よって，以下のようにラベルあり事例集合による識別損失を加える．

$$\mathcal{L}_l(\mathbf{x}, \mathbf{w}, \mathbf{y}) = \mathcal{L}(\mathbf{x}, \mathbf{w}, \mathbf{y}) - \alpha \cdot \log q_\phi(\mathbf{y}|\mathbf{x}, \mathbf{w}) \quad (4)$$

ただし，α は識別モデルと生成モデルの学習の割合を調節するパラメータであり，本研究では $\alpha = 0.5 \cdot \frac{M+N}{M}$ とした．

したがって，ラベルあり・なし事例集合の目的関数は以下のようになる．

$$\mathcal{J} = \sum_{(\mathbf{x}_i, \mathbf{w}_i, \mathbf{y}_i) \in \mathcal{D}_{\mathcal{L}}} \mathcal{L}_l(\mathbf{x}_i, \mathbf{w}_i, \mathbf{y}_i) + \sum_{(\mathbf{x}_j, \mathbf{w}_j) \in \mathcal{D}_{\mathcal{U}}} \mathcal{U}(\mathbf{x}_j, \mathbf{w}_j) \quad (5)$$

この目的関数を JMVAE と同様パラメータ ϕ，θ に関して最適化することで，識別モデル $q_\phi(\mathbf{y}|\mathbf{x})$ を学習することができる．本稿では，この提案モデルを *Semi-Supervised Multimodal Variational AutoEncoder*（*SS-MVAE*）と呼ぶ．

3.4 半教師あり学習モデル：SS-HMVAE

本節では，SS-MVAE の他に，生成モデルを変更した別の半教師あり学習のモデルを提案する．潜在変数として新たに $\mathbf{a}$ を導入し，ラベル情報を含んだ生成モデルを $p(\mathbf{x}, \mathbf{w}, \mathbf{y}) = \int \int p_\theta(\mathbf{x}|\mathbf{a}) p_\theta(\mathbf{w}|\mathbf{a}) p_\theta(\mathbf{a}|\mathbf{z}, \mathbf{y}) p(\mathbf{z}) p(\mathbf{y}) d\mathbf{a} d\mathbf{z}$ とする．図 1 の (a) と (b) を比べると，SS-MVAE では $\mathbf{y}$ と $\mathbf{z}$ が直接 $\mathbf{x}$，$\mathbf{w}$ と接続しているが，本モデルでは，$\mathbf{y}$ と $\mathbf{z}$ は $\mathbf{a}$ を介して $\mathbf{x}$，$\mathbf{w}$ に情報が伝わるようになっている．さらに $\mathbf{x}$ と $\mathbf{w}$ は $\mathbf{a}$ について条件付き独立となっている．

推論モデルは $q(\mathbf{a}, \mathbf{z}|\mathbf{x}, \mathbf{w}, \mathbf{y}) = q_\phi(\mathbf{z}|\mathbf{a}, \mathbf{y}) q_\phi(\mathbf{a}|\mathbf{x}, \mathbf{w})$ とおく．これによって $q(\mathbf{z}|\mathbf{x}, \mathbf{w}, \mathbf{y}) = \int q(\mathbf{a}, \mathbf{z}|\mathbf{x}, \mathbf{w}, \mathbf{y}) d\mathbf{a} d\mathbf{z}$ となるため，より複雑な分布が表現できる．これは，より真の事後分布 $p(\mathbf{z}|\mathbf{x}, \mathbf{w}, \mathbf{y})$ に適合できることを意味し，同じことが様々な先行研究でも示されている [Maaløe 16, Sønderby 16, Gulrajani 16].

このように，生成モデルと推論モデル共に，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．

SS-HMVAE の目的関数は，ラベルありとラベルなしそれぞれにおいて，以下のとおりである．

$$\begin{aligned}
&\log p(\mathbf{x},\mathbf{w},\mathbf{y}) \\
&\geq E_{q_\phi(\mathbf{a},\mathbf{z}|\mathbf{x},\mathbf{w},\mathbf{y})}[\log \frac{p_\theta(\mathbf{x}|\mathbf{a})p_\theta(\mathbf{w}|\mathbf{a})p_\theta(\mathbf{a}|\mathbf{z},\mathbf{y})p(\mathbf{z})p(\mathbf{y})}{q_\phi(\mathbf{a},\mathbf{z}|\mathbf{x},\mathbf{w},\mathbf{y})}] \\
&\equiv -\mathcal{L}(\mathbf{x},\mathbf{w},\mathbf{y}) \qquad (6)
\end{aligned}$$

$$\begin{aligned}
&\log p(\mathbf{x},\mathbf{w}) \\
&\geq E_{q_\phi(\mathbf{a},\mathbf{z},\mathbf{y}|\mathbf{x},\mathbf{w})}[\log \frac{p_\theta(\mathbf{x}|\mathbf{a})p_\theta(\mathbf{w}|\mathbf{a})p_\theta(\mathbf{a}|\mathbf{z},\mathbf{y})p(\mathbf{z})p(\mathbf{y})}{q_\phi(\mathbf{a},\mathbf{z},\mathbf{y}|\mathbf{x},\mathbf{w})}] \\
&\equiv -\mathcal{U}(\mathbf{x},\mathbf{w}) \qquad (7)
\end{aligned}$$

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

4. 実験

4.1 データセット

本研究では，マルチモーダルの研究でベンチマークとして利用される MIR Flickr データセット [Huiskes 08] を用いた．このデータセットは，写真共有コミュニティサイト Flickr[*2] で得られた 1,000,000 の画像と各画像に付けられたタグで構成されており，そのうち 25,000 の画像とタグには 38 カテゴリのラベルがつけられている．ただし，1 枚の画像に複数のカテゴリが割り当てられることもある．本実験では [Srivastava 12][*3] で抽出された特徴量を利用する．この特徴量は，画像は 3,857 次元，タグは 2,000 次元となっている．本実験では，25,000 のうち 15,000 をラベルあり訓練事例集合，10,000 をテスト事例集合とした．また，ラベルの付いていない 975,000 の画像及びタグは，本実験ではラベルなし訓練集合として全て用いる．したがって，$M = 15,000$ 及び $N = 975,000$ となる．

4.2 実験設定

本実験では，画像を $\mathbf{x}$，タグを $\mathbf{w}$ とする．画像とタグの定義域はそれぞれ $\mathcal{R}^{3857}$ と $\{0,1\}^{2000}$ なので，それぞれの生成モデルの分布を

$$\begin{aligned}
p_\theta(\mathbf{x}|\mathbf{z},\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\theta(\mathbf{z},\mathbf{y}),\mathrm{diag}(\boldsymbol{\sigma}^2_\theta(\mathbf{z},\mathbf{y}))) \qquad (8)\\
p_\theta(\mathbf{w}|\mathbf{z},\mathbf{y}) &= Ber(\mathbf{w}|\boldsymbol{\pi}_\theta(\mathbf{z},\mathbf{y})) \qquad (9)\\
p_\theta(\mathbf{x}|\mathbf{a}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\theta(\mathbf{a}),\mathrm{diag}(\boldsymbol{\sigma}^2_\theta(\mathbf{a}))) \qquad (10)\\
p_\theta(\mathbf{w}|\mathbf{a}) &= Ber(\mathbf{w}|\boldsymbol{\pi}_\theta(\mathbf{a})) \qquad (11)
\end{aligned}$$

とする．また，ラベル情報は $\mathbf{y}$ とし，$\{0,1\}^{38}$ なので，識別モデルを

$$q_\phi(\mathbf{y}|\mathbf{x},\mathbf{w}) = Ber(\mathbf{y}|\boldsymbol{\pi}_\theta(\mathbf{x},\mathbf{w})) \qquad (12)$$

とする．本実験では SS-MVAE と SS-HMVAE で同じネットワーク構造の識別モデルを用いる．

それ以外の生成モデル，推論モデルの分布は

$$\begin{aligned}
p(\mathbf{z}) &= \mathcal{N}(\mathbf{z}|0,\mathbf{I}) \qquad (13)\\
p(\mathbf{y}) &= Ber(\mathbf{y}|\boldsymbol{\pi}) \qquad (14)\\
p_\theta(\mathbf{a}|\mathbf{z},\mathbf{y}) &= \mathcal{N}(\mathbf{a}|\boldsymbol{\mu}_\theta(\mathbf{z},\mathbf{y}),\mathrm{diag}(\boldsymbol{\sigma}^2_\theta(\mathbf{z},\mathbf{y}))) \qquad (15)\\
q_\phi(\mathbf{a}|\mathbf{x},\mathbf{w}) &= \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\theta(\mathbf{x},\mathbf{w}),\mathrm{diag}(\boldsymbol{\sigma}^2_\theta(\mathbf{x},\mathbf{w}))) \qquad (16)\\
q_\phi(\mathbf{z}|\mathbf{a},\mathbf{y}) &= \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\theta(\mathbf{a},\mathbf{y}),\mathrm{diag}(\boldsymbol{\sigma}^2_\theta(\mathbf{a},\mathbf{y}))) \qquad (17)
\end{aligned}$$

とした．各分布のネットワーク構造については，ページ数の都合上省略する，

ネットワークの各層の活性化関数には rectified linear unit を用い，最適化アルゴリズムに Adam [Kingma 14b] を利用した．また各層でバッチ正則化 [Ioffe 15] を行い，識別モデルには過学習を防ぐため dropout による正則化を行った．実装は Theano[*4] と Lasagne[*5] をベースとした深層生成モデルライブラリ Tars[*6] を使用した．

訓練事例集合で学習した識別モデルをテスト事例集合で検証し，mean average precision (MAP) で評価する [*7]．

本実験では，**1. ラベルあり事例集合のみで教師あり学習**，**2. SS-MVAE で半教師あり学習**，**3. SS-HMVAE で半教師あり学習**をそれぞれ行い比較する．また学習の際に，下界のモンテカルロサンプリングの数を変更した場合も検証した．

4.3 実験結果

表 1 が実験結果である．まず，通常の教師あり学習と SS-MVAE による半教師あり学習の結果を比較すると，MC=1 のときには教師あり学習の方が MAP が高い結果となった．しかし，モンテカルロサンプリング数を増やし MC=10 とすると教師あり学習を上回る精度となることが確認できる．モンテカルロサンプリング数を増やすと識別モデルの汎化性能が向上することは他の深層生成モデルによる半教師あり学習の結果 [Maaløe 16] でも示されている．

次に SS-HMVAE の結果をみると，MC=1 の場合でも教師あり学習の結果を上回っていることが確認できる．しかし，MC=10 の場合は逆に精度が落ちてしまっていることがわかるが，それでも SS-MVAE と同じか僅かに高い結果となっている．

図は，エポックごとの MAP の値を各モデルについてプロットしたものである．この図をみると，教師あり学習と SS-HMVAE は立ち上がりは同じくらいの精度だが，教師あり学習が 10 エポックあたりで MAP が変わらなくなっているのに対して，SS-HMVAE はそこからさらに精度が向上しているのが確認できる．このことから，ラベルなしデータが識別モデルの汎化性能向上に貢献していることがわかる．ただし，学習したラベルなしデータの数を考えると，結果として大きな改善にはなっていない．

5. まとめ

本稿では，マルチモーダルデータによる深層生成モデルを用いた半教師あり学習の手法を検討し，SS-MVAE と SS-HMVAE の 2 つのモデルを提案した．実験から，提案モデルによる半教師あり学習の一定の有効性は検証できたが，大きな汎化性能の向上にはならなかった．考えられる理由の 1 つとして，デー

∗2 http://www.flickr.com
∗3 http://www.cs.toronto.edu/~nitish/multimodal/index.html
∗4 https://github.com/Theano/Theano
∗5 https://github.com/Lasagne/Lasagne
∗6 https://github.com/masa-su/Tars
∗7 [鈴木 16] では LRAP で評価していたが，本実験ではより一般的な指標である MAP で評価する．

表 1: 教師あり学習と提案手法による半教師あり学習の比較（MC はモンテカルロサンプリング数を表している）

モデル	MAP
教師あり学習（ベースライン）	0.618
SS-MVAE (MC=1)	0.612
SS-HMVAE (MC=1)	**0.632**
SS-MVAE (MC=10)	0.626
SS-HMVAE (MC=10)	0.628

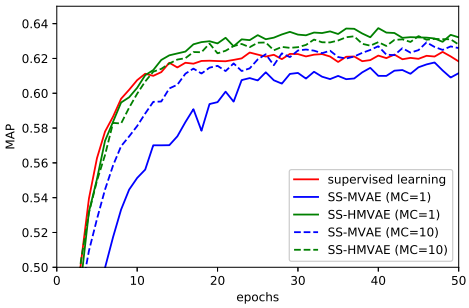


図 2: エポックごとの MAP の評価

タセット自体の困難さが挙げられる．本実験で利用したデータセットは，すでに特徴抽出されたものだが，深層ニューラルネットワークで学習するとすぐに過学習してしまうことを確認している．また，ある画像とタグに対するラベルが one-hot ではなく複数のラベルに属しうるので，半教師あり学習の効果を直接的には確認しづらい（単純な 2 クラスや多クラス分類の方がタスクとして簡単になるため，半教師あり学習の評価をしやすいと思われる）．今後は，より大規模な画像データなどを本手法で学習した場合の有効性を検証したい．

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

[Cheng 16] Cheng, Y., Zhao, X., Cai, R., Li, Z., Huang, K., and Rui, Y.: Semi-Supervised Multimodal Deep Learning for RGB-D Object Recognition (2016)

[Goodfellow 14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y.: Generative adversarial nets, in *Advances in neural information processing systems*, pp. 2672–2680 (2014)

[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)

[Gulrajani 16] 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)

[Huiskes 08] Huiskes, M. J. and Lew, M. S.: The MIR Flickr retrieval evaluation, in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pp. 39–43 (2008)

[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)

[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)

[Sønderby 16] Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O.: Ladder variational autoencoders, in *Advances in Neural Information Processing Systems*, pp. 3738–3746 (2016)

[Srivastava 12] Srivastava, N. and Salakhutdinov, R. R.: Multimodal learning with deep boltzmann machines, in *Advances in neural information processing systems (NIPS)*, pp. 2222–2230 (2012)

[Suzuki 16] Suzuki, M., Nakayama, K., and Matsuo, Y.: Joint Multimodal Learning with Deep Generative Models, *arXiv preprint arXiv:1611.01891* (2016)

[鈴木 16] 鈴木 雅大, 松尾 豊 F 深層生成モデルを用いたマルチモーダル学習 (2016)