The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

マルチモーダル学習のための階層ニューラルトピックモデル

Hierarchical Neural Topic Models for Multimodal Learning

青木 達哉 *1　南坂　雅人 *1　長井 隆行 *1*2
Tatsuya Aoki　Masato Minamisaka　Takayuki Nagai

*1電気通信大学　*2大阪大学
The University of Electro-Communications　Osaka University

In this paper, we propose novel hierarchical neural topic models for learning multimodal sensory data. The first one is the neural version of multilayered multimodal LDA (mMLDA), which we call "Deep-mMLDA". The use of the neural inference network makes the Bayesian topic model scalable and powerful. Therefore, the model can deal with very large scale multimodal sensor data in real world. The second one is further extension of the Deep-mMLDA to nonparametric Deep-mMLDA, which can infer the number of categories from the learning data. The idea behind the model is to use the recurrent stick breaking process (RSBP) that uses recurrent neural networks for implementing the stick breaking process. We conduct an experiment using multimodal data of human activities collected in a smart house environment. As the result of comparison in the performance of the proposed and baseline models, the validity of the proposed models is confirmed.

1. はじめに

近年，機械学習の性能が向上し，実社会での応用が進んでいる．機械学習の役割の 1 つは観測データから有益な情報を予測することである．実世界の観測情報は一般に，複数のモダリティからなるマルチモーダル情報であり，マルチモーダル情報から潜在情報を抽出することは重要な要素となる．機械学習において，データの潜在的な情報を抽出する研究は数多く行われてきた．その中でも階層ベイズに基づくトピックモデルである Latent Dirichlet Allocation (LDA) [Blei03] は文書分類から始まり，情報検索，画像分類など様々な分野へと応用されている．マルチモーダル情報処理においても，例えば中村らは画像，音，触覚といった複数のセンサ情報を同時に扱う Multimodal LDA (MLDA) を提案し，ロボットにおける有効性を示した [Nakamura11]．さらに [Attamimi16] は，MLDA を多層構造とした multilayerd MLDA (mMLDA) を提案し，マルチモーダルな観測情報に関するトピック同士の関係性を，上位層としてモデル化することを可能とした．しかし MLDA や mMLDA には，大規模な入力データに対して計算コストが高いといった問題がある．特にギブスサンプリングを用いた推論には，多大な時間を要する．また，学習データに対応したトピック数を事前に設定する必要があるが，マルチモーダルな情報に対して適切なトピック数を決定するのは困難である．

一方で，近年の深層学習の発展により，確率的生成モデルを深層学習によって推論する深層生成モデルに注目が集まっている．その中の 1 つが，変分近似と深層学習を援用して推論を行う Variational Autoencoder (VAE) [Kingma13] である．VAE の推論手法をベースとして，様々な確率的生成モデルの推論が提案されており，先に述べたトピックモデルについても，[Srivastava17] や [Srivastava18] などの研究がある．VAE をベースとした学習法の利点として，GPU を使った並列計算による推論が挙げられ，大規模な学習データを扱いやすいという強みをもつ．

本研究ではこのような背景に基づき，両分野の先行研究の利点を兼ね備えたより効率的に大規模なマルチモーダル情報を学習できるトピックモデルの実現を目指す．具体的には，[Srivastava18] で提案されたトピックモデルの推論手法に基づき，[Attamimi16] で提案された階層構造を持つトピックモデル mMLDA のニューラルネットワークによる推論モデルを提案する．そして，上位層のトピック数を推定できるトピックモデルへと拡張を行う．

連絡先: 青木達哉，電気通信大学，東京都調布市調布ヶ丘 1-5-1，aoki@apple.ee.uec.ac.jp

2. mMLDA

はじめに本研究で提案するニューラルトピックモデルのベースとなる mMLDA について概要を述べる．mMLDA は，マルチモーダルデータに対する階層的な構造を持つトピックモデルである．mMLDA の下位層には，複数モダリティ情報を含む観測情報を 1 つのトピックに統合するモデルが複数配置され，上位層には，それらのモデルのトピックの関係を捉えるモデルが存在する．mMLDA は，$\boldsymbol{x}_d^{cm}$ データ数 D の観測データ集合 $\boldsymbol{X} \in \{\boldsymbol{x}_d^{cm} | m = M^1, \cdots, M^c, c = 1, \cdots, C\}$ が上位トピック数 K, 下位トピック数 K^c の確率分布の混合分布から生成されたとして以下のようにモデル化する．

$$p(\boldsymbol{X}) = \prod_{d=1}^{D} \prod_{c=1}^{C} \prod_{m=1}^{M^c} p(\boldsymbol{x}_d^{cm} | \boldsymbol{\phi}^{cm}, \beta^{cm}, \boldsymbol{\theta}^c) p(\boldsymbol{\theta}^c | \boldsymbol{\theta}, \alpha^c) p(\boldsymbol{\theta} | \alpha)$$

β^*，α^*，α はハイパーパラメータ，C は下位層の MLDA のモデル数，M^c は c 番目の下位層の MLDA のモデルに含まれるモダリティ数を表す．mMLDA の学習とは，観測情報 $\boldsymbol{X}$ の尤度を最大とするパラメータ $\boldsymbol{\phi}$，$\boldsymbol{\theta}$ を推定することであり，[Attamimi16] ではギブスサンプリングによって推定を行っている．

3. 提案手法

3.1 Deep-mMLDA

本研究では，前節で述べた mMLDA のように階層的なトピック構造を持つモデルをニューラルネットワークで構築し，そのパラメータの推論手法を提案する．図 1 が提案するモデルの構造である．このモデルを "Deep-mMLDA" と呼ぶ．図 1 中の Z^* のユニットは上位のノードの出力を受け，平均 $\boldsymbol{\mu}^*$，分

散共分散行列 $\boldsymbol{\Sigma}^*$ を算出し，混合比 $\boldsymbol{\theta}^*$ を出力するノード全体を表す．このネットワークは，全モダリティの観測情報を入力とし，全結合層を経て上位層のトピック混合比 $\boldsymbol{\theta}$，下位層のトピック混合比 $\boldsymbol{\theta}^c$ と観測情報の出力分布 $p(\boldsymbol{x}^{cm}|\boldsymbol{\theta}^{cm})$ をニューラルネットワークの前向き計算により算出する．

$$\begin{aligned}
&\boldsymbol{\theta} = \mathrm{softmax}(\boldsymbol{\mu} + \boldsymbol{\epsilon}\boldsymbol{\Sigma}) \\
&\boldsymbol{\theta}^c = \mathrm{softmax}(\boldsymbol{\mu}^c + \boldsymbol{\epsilon}^c\boldsymbol{\Sigma}^c) \quad (c = 1, \cdots, C) \\
&p(\boldsymbol{x}^{cm}|\boldsymbol{\theta}^c) = \mathrm{softmax}(\boldsymbol{\theta}^c \cdot \boldsymbol{w}^{cm}) \quad (m = 1, \cdots, M^c) \qquad (1)
\end{aligned}$$

ただし，C は下位層の MLDA のモデルの数，M^c は c 番目の下位層のモデルのモダリティ数を表す．

次に，Deep-mMLDA の損失関数となる変分下限 $\mathcal{L}$ を導出する．ここでは，$C = 2$，$M^1 = 1$，$M^2 = 1$ の場合について述べるが，他の場合も同様に考えることができる．観測情報 $\boldsymbol{X} = [\boldsymbol{x}^{11}, \boldsymbol{x}^{21}]$ に対する周辺対数尤度より，$\mathcal{L}$ は以下のようになる．なお，$q(*)$ は近似分布，$\boldsymbol{H} = [\alpha, \alpha^1, \alpha^2]$ はハイパーパラメータの集合，$\boldsymbol{\Theta} = [\boldsymbol{\theta}, \boldsymbol{\theta}^1, \boldsymbol{\theta}^2]$ は混合比の集合を表す．

$$\begin{aligned}
\log p(\boldsymbol{X}|\boldsymbol{H}) &= \log \int_{\boldsymbol{\Theta}} q(\boldsymbol{\Theta}) \frac{p(\boldsymbol{X}, \boldsymbol{\Theta}|\boldsymbol{H})}{q(\boldsymbol{\Theta})} d\boldsymbol{\Theta} \\
&\leq \int_{\boldsymbol{\Theta}} q(\boldsymbol{\Theta}) \log \frac{p(\boldsymbol{X}, \boldsymbol{\Theta}|\boldsymbol{H})}{q(\boldsymbol{\Theta})} d\boldsymbol{\Theta} = \mathcal{L} \quad (2)
\end{aligned}$$

式 (2) の同時確率 $p(\boldsymbol{X}, \boldsymbol{\Theta}|\boldsymbol{H})$ は，mMLDA の場合，

$$\begin{aligned}
p(\boldsymbol{X}, \boldsymbol{\Theta}|\boldsymbol{H}) &= p(\boldsymbol{x}^{11}|\boldsymbol{\theta}^1)p(\boldsymbol{x}^{21}|\boldsymbol{\theta}^2)p(\boldsymbol{\theta}^1|\boldsymbol{\theta}, \alpha^1) \\
&\quad p(\boldsymbol{\theta}^2|\boldsymbol{\theta}, \alpha^2)p(\boldsymbol{\theta}|\alpha) \qquad (3)
\end{aligned}$$

である．さらに，deep-mMLDA では近似分布 q を次のように仮定する．これらのパラメータは，式 (1) のようにニューラルネットワークで計算される．

$$q(\boldsymbol{\Theta}) = q(\boldsymbol{\theta}|\boldsymbol{x}^{11}, \boldsymbol{x}^{21})q(\boldsymbol{\theta}^1|\boldsymbol{x}^{11})q(\boldsymbol{\theta}^2|\boldsymbol{x}^{21}) \qquad (4)$$

式 (3)，式 (4) を用いて最終的に，下限 $\mathcal{L}$ は次のようになる．$D_{KL}(\cdot||\cdot)$ は KL ダイバージェンスを表す．

$$\begin{aligned}
\mathcal{L} &= -D_{KL}(q(\boldsymbol{\theta}|\boldsymbol{x}^{11}, \boldsymbol{x}^{21})||p(\boldsymbol{\theta}|\alpha)) \\
&\quad -\mathbb{E}_{\boldsymbol{\theta}\sim q(\boldsymbol{\theta}|\boldsymbol{x}^{11}, \boldsymbol{x}^{21})}[D_{KL}(q(\boldsymbol{\theta}^1|\boldsymbol{x}^{11})||p(\boldsymbol{\theta}^1|\alpha^1))] \\
&\quad -\mathbb{E}_{\boldsymbol{\theta}\sim q(\boldsymbol{\theta}|\boldsymbol{x}^{11}, \boldsymbol{x}^{21})}[D_{KL}(q(\boldsymbol{\theta}^2|\boldsymbol{x}^{21})||p(\boldsymbol{\theta}^2|\alpha^2))] \\
&\quad +\mathbb{E}_{\boldsymbol{\theta}^1\sim q(\boldsymbol{\theta}^1|\boldsymbol{x}^{11})}[\log p(\boldsymbol{x}^{11}|\boldsymbol{\theta}^1)] \\
&\quad +\mathbb{E}_{\boldsymbol{\theta}^2\sim q(\boldsymbol{\theta}^2|\boldsymbol{x}^{21})}[\log p(\boldsymbol{x}^{21}|\boldsymbol{\theta}^2)] \qquad (5)
\end{aligned}$$

つまり，$\mathcal{L}$ は上位層のトピックに対する KL ダイバージェンス，下位層のトピックに対する KL ダイバージェンス及び各モダリティ毎の観測情報に対する対数尤度の期待値の 3 種類の項で構成される．この式 (5) を目的関数とし，値を最大化するパラメータを決定する．

3.2 RSB-mMLDA

ここでは，Deep-mMLDA を学習データに応じて上位層のトピック数が調整可能なモデルへ拡張する．具体的には，上位層のトピックの確率分布に対し，ディリクレ過程から生成される無限次元ディリクレ分布を事前分布として導入する．本研究では Stick-Breaking Process (SBP) をニューラルネットワークで計算するための手法として，[Miao17] で提案された Recurrent Stick Breaking Process (RSBP) を利用する．

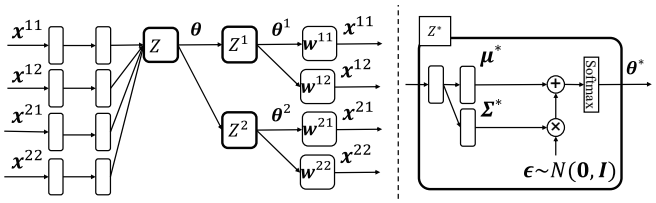


図 1: Deep-mMLDA のネットワーク構造

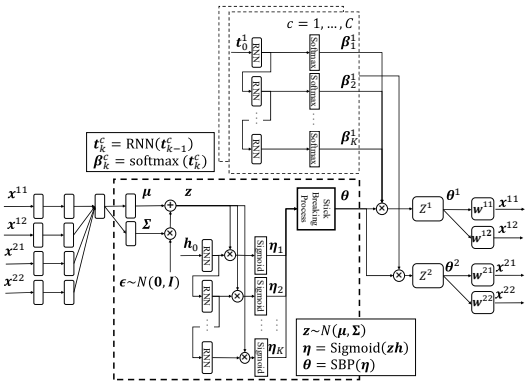


図 2: RSB-mMLDA のネットワーク構造

RSBP は，Recurrent Neural Network (RNN) の再帰的な計算により SBP を再現する手法である．

図 2 に RSBP を導入し，上位層のトピック数を可変にした Deep-mMLDA のネットワーク構造を示す．本研究では，このモデルを “RSB-mMLDA” と呼ぶ．このネットワークでは，全モダリティの観測情報を入力とし，全結合層を経て RSBP により上位層のトピックの混合比 $\boldsymbol{\theta}$ を出力する．また，$\boldsymbol{t}^0$ を初期入力値とする RNN を用いたネットワークにより，上位層のトピックの出力分布のパラメータに相当する $\boldsymbol{\beta}^c$ を出力する．最終的に，$\boldsymbol{\theta}$ 及び $\boldsymbol{\beta}^c$ を入力とし，下位のトピック混合比 $\boldsymbol{\theta}^c$ 及び観測情報の出力分布 $P(\boldsymbol{x}^{cm}|\boldsymbol{\theta}^c)$ を算出する．これらの計算は Deep-mMLDA と同様，全てニューラルネットワークの前向き計算により実行される．

$$\begin{aligned}
&\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\epsilon}\boldsymbol{\Sigma},\ \boldsymbol{\eta} = \boldsymbol{h}\boldsymbol{z},\ \boldsymbol{\theta} = \mathrm{SBP}(\boldsymbol{\eta}) \\
&\boldsymbol{\beta}^c = \mathrm{RNN}(\boldsymbol{t}^c),\ \boldsymbol{\theta}^c = \mathrm{softmax}(\boldsymbol{\theta}^c \cdot \boldsymbol{\beta}^c)\ (c = 1, \cdots, C) \\
&p(\boldsymbol{x}^{cm}|\boldsymbol{\theta}^c) = \mathrm{softmax}(\boldsymbol{\theta}^c \cdot \boldsymbol{w}^{cm})\ (m = 1, \cdots, M^c) \qquad (6)
\end{aligned}$$

RSB-mMLDA の損失関数は Deep-mMLDA と同様であり，式 (5) を最小化するパラメータを推定する．また，パラメータの推定と同時に，トピック数を増加させるべきかを判断する．そのために，トピック数が $i-1$ の場合の損失 $\mathcal{L}^{i-1}$ と i の場合の損失 $\mathcal{L}^i$ を考え，トピック数を増やしたときの尤度の増加量 $\mathcal{I}$ を算出する．

$$\mathcal{I} = \sum_d^D [\mathcal{L}_d^i - \mathcal{L}_d^{i-1}] / \sum_d^D [\mathcal{L}_d^i] \qquad (7)$$

トピック数の増加しやすさを決めるハイパーパラメータを γ とし，$\mathcal{I} > \gamma$ であればトピック数を 1 つ増加させる．

3.3 学習における工夫

提案モデルの学習を安定化させるために，2 つの工夫を学習に導入する．

1 つ目は，事前学習による下位層の重みの初期値の決定である．提案モデルでは，潜在変数の階層化などのため複雑な

ネットワーク構造になったことで，十分にパラメータが学習されない可能性がある．そこで，事前にモジュール毎に下位層の学習を行う．その後，学習された重みを Deep-mMLDA，RSB-mMLDA の下位層の重みの初期値としてネットワーク全体を学習する．

2 つ目は欠損情報の予測学習である．これは観測情報のうち意図的に一部分の情報を全ての要素が 0 であるベクトルと置換し，他の観測情報から予測する学習を混入させる．この学習を通常の学習と合わせて行うことでモダリティ間の関係性の学習を促進する．

4. 評価実験

提案手法である Deep-mMLDA，RSB-mMLDA を用いて適切にマルチモーダルデータのトピックが学習可能であるかを評価した．また，ベースライン手法として [Attamimi16] で提案されたギブスサンプリングを用いる mMLDA(Gibbs-mMLDA) でも同様に学習を行い，性能を比較した．

4.1 実験設定

評価実験に用いたデータセットは，[Attamimi16] の実験で使用されたスマートハウス内で人が物体を操作する活動の場面を観測したマルチモーダルデータである．観測情報には，操作物体の画像情報 (Object)，被験者の関節角情報 (Motion)，LRF で計測した被験者の位置情報 (Place)，画像から推定した被験者の性別情報 (Person Gender)，年齢情報 (Person Age) とそれらに対応する単語情報 (Object Word, Motion Word, Place Word, Person Word) の 9 種類の情報が含まれる．各情報は学習のための前処理として Bag of Features 表現への変換を行った．データの総数は 446 個である．

上述のマルチモーダルデータをもとに，次の 2 通りの情報の組み合わせに対してトピック推定を行った．

1. 4 入力条件
   - 使用した情報
     (Object，Object Word，Motion，Motion Word)
   - mMLDA の構造
     $C = 2$ (Object，Motion)
     $C1 = (\boldsymbol{x}^{\mathrm{Object}}, \boldsymbol{x}^{\mathrm{ObjectWord}})$
     $C2 = (\boldsymbol{x}^{\mathrm{Motion}}, \boldsymbol{x}^{\mathrm{MotionWord}})$
2. 9 入力条件
   - 使用した情報
     (Object，Object Word，Motion，Motion Word，Place，Place Word，Person Gender，Person Age，Person Word)
   - mMLDA の構造
     $C = 4$ (Object, Motion, Place, Person)
     $C1 = (\boldsymbol{x}^{\mathrm{Object}}, \boldsymbol{x}^{\mathrm{ObjectWord}})$
     $C2 = (\boldsymbol{x}^{\mathrm{Motion}}, \boldsymbol{x}^{\mathrm{MotionWord}})$
     $C3 = (\boldsymbol{x}^{\mathrm{Place}}, \boldsymbol{x}^{\mathrm{PlaceWord}})$
     $C4 = (\boldsymbol{x}^{\mathrm{PersonGender}}, \boldsymbol{x}^{\mathrm{PersonAge}}, \boldsymbol{x}^{\mathrm{PersonWord}})$

本実験では，学習を行うにあたり 3 種類の手法に共通で下位層のカテゴリ数を，Object カテゴリ数=25，Motion カテゴリ数=17，Place カテゴリ数=5，Person カテゴリ数=3 とした．また，上位層のカテゴリ数について，Gibbs-mMLDA と Deep-mMLDA は 17 とし，RSB-mMLDA は初期カテゴリ数を 1，$\gamma = 0.05$ とした．

表 1: 4 入力条件の学習後の対数尤度の比較

	Deep-mMLDA	RSB-mMLDA
Object	-28219	-28349
Motion	-13155	-13649

表 2: 9 入力条件の学習後の対数尤度の比較

	Deep-mMLDA	RSB-mMLDA
Object	-28746	-29153
Motion	-13852	-14734
Place	-7658	-8062
Person	-7581	-7940

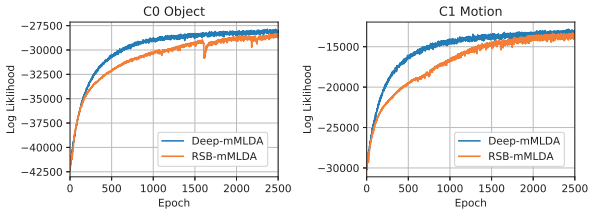


図 3: 4 入力条件における学習データに対する対数尤度

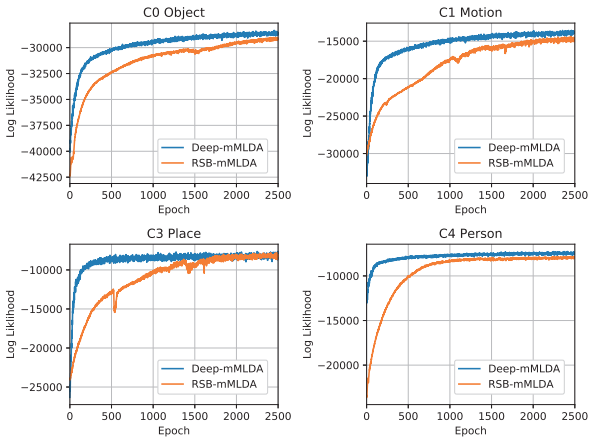


図 4: 9 入力条件における学習データに対する対数尤度

4.2 結果

まず，提案手法である Deep-mMLDA 及び RSB-mMLDA の学習中の挙動を確認した．図 3 に 4 入力条件の場合の対数尤度の挙動，図 4 に 9 入力条件の場合の対数尤度の挙動を示す．どちらの条件においても，Deep-mMLDA は急激に損失を下げるように学習が進んだ．一方，RSB-mMLDA は Deep-mMLDA と比較すると多くの更新回数が必要であるが，対数尤度の変化は徐々に小さくなり，無限にトピックを増やし続けることはないことが確認できる．学習終了時の最終的な対数尤度を，表 1，表 2 に示す．また，各種法の学習時間の実測値を表 3 に示す．本実験では，Gibbs-mMLDA のサンプリング回数を 2000，Deep-mMLDA 及び RSB-mMLDA のエポック数を 2500 と設定した．表 3 に示すように，大きく学習時間が削減された．この学習時間の短縮は並列計算によるパラメータ推定に起因する．

次に，それぞれのモデルが推定したトピックの構造を評価した．評価のために，人手により学習データに対して Object，Motion，Place，Person の 4 つの項目のラベルづけを行い，ラベルによる分類と学習結果を用いた分類の一致度を分類精度と

表 3: 各手法で学習にかかった時間

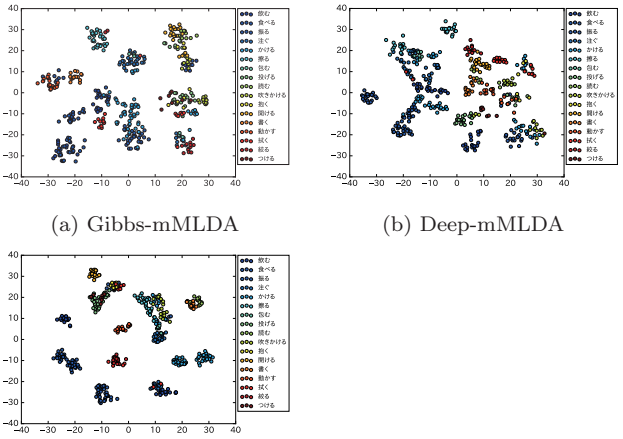
	Gibbs-mMLDA	Deep-mMLDA	RSB-mMLDA
4 入力条件 [秒]	62085	356	1091
9 入力条件 [秒]	70718	559	2141

表 4: 4 入力条件における分類精度の比較

	Gibbs-mMLDA	Deep-mMLDA	RSB-mMLDA
Object	0.484	0.748	0.647
Motion	0.562	0.600	0.567

表 5: 9 入力条件における分類精度の比較

	Gibbs-mMLDA	Deep-mMLDA	RSB-mMLDA
Object	0.443	0.576	0.457
Motion	0.502	0.674	0.504
Place	1.00	0.982	0.746
Person	0.807	0.959	0.695



(a) Gibbs-mMLDA (b) Deep-mMLDA

(c) RSB-mMLDA

図 5: 4 入力における上位層の状態

定義する．表 4 に 4 入力条件，表 5 に 9 入力条件の結果をそれぞれ示す．

まず，Deep-mMLDA についてみると，表 4 で示すようにベースラインである Gibbs-mMLDA と比較して，同等もしくはそれ以上の分類精度であった．また，表 5 に示すように，入力情報の種類を 4 から 9，下位概念数を 2 から 4 に増加させても，Gibbs-mMLDA と同等以上の分類精度があった．これらの結果は Deep-mMLDA が Gibbs-mMLDA と同等の学習をすることができ，入力情報の種類数や下位層のモデル数に対して拡張性を持つことを示す．

続いて，RSB-mMLDA についてみると 4 入力条件では，表 4 で示すように，RSB-mMLDA の分類精度は Gibbs-mMLDA の値を上回っているが，上位トピック数を固定した Deep-mMLDA には劣る．また，9 入力条件では，表 5 で示すように Gibbs-mMLDA よりも劣る結果となった．

最後に，推定した上位層のトピック混合比 $\boldsymbol{\theta}$ を t-SNE で 2 次元へ圧縮し，各モデルの上位層のトピック空間を定性的に調べた．図 5 に，それぞれの上位層のトピック空間を示す．各データ点は，学習データに付与された Motion ラベルで色付けを行った．予めカテゴリ数を与えた Gibbs-mMLDA，Deep-mMLDA と同様に，トピック数をデータに応じて推定した RSB-mMLDA の上位層のトピック空間も，Motion ラベルに応じたクラスタが形成されていることが分かる．これらの結果は，提案手法にマルチモーダル情報に対して有効な上位層のトピックを推定できることを示唆する．

5. まとめ

本稿では，より効率的に大規模なマルチモーダル情報を学習できるトピックモデルの実現を目指し，ニューラルネットワークを用いた 2 種類のマルチモーダル情報のためのトピックモデルを提案した．Deep-mMLDA は，文書分類のために提案された階層トピックモデルをマルチモーダル情報が扱えるように拡張したモデルであり，学習のための損失関数の導出を行った．RSB-mMLDA は，Deep-mMLDA にディリクレ過程を導入し上位層のカテゴリ数推定を可能にしたモデルである．評価実験として，最大 9 種のモダリティ情報で構成されるマルチモーダルデータに対して，Deep-mMLDA，RSB-mMLDA で学習を行い，その分類精度を先行研究と比較した．Deep-mMLDA は全条件において，RSB-mMLDA は 4 入力条件において従来の mMLDA と同等以上の分類精度を得る学習が可能であることが確認できた．また深層学習の並列化という特徴を生かし，学習時間を大幅に削減した．

今後は，本研究で提案した 2 つの提案モデルの学習結果が異なった要因を明らかにし，RSB-mMLDA においても全ての学習条件で従来の mMLDA と同等以上の学習結果が得られるようにモデルの改善を行う予定である．

謝辞

本研究は，JST CREST(JP-MJCR15E3) 及び JSPS 科研費 (17J10512) の支援を受けて実施した．

参考文献

[Blei03] D M.Blei, A Y.Ng, M I.Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research 3, pp. 993-1022, 2003

[Nakamura11] T Nakamura, T Araki, T Nagai, N Iwahashi, “Grounding of Word Meaning in Latent Dirichlet Allocation-Based Multimodal Concepts”, Adavanced Robotics 25, pp. 2189-2206, 2011

[Attamimi16] M Attamimi, Y Ando,T Nakamura, T Nagai, D Mochihashi, I Kobayashi, H Asoh, “Learning word meanings and grammar for verbalization of daily life activities using multilayered multimodal latent Dirichlet allocation and Bayesian hidden Markov models”, Advanced Robotics, pp. 806-824, 2016

[Kingma13] D P Kingma, M Welling, “Auto-Encoding Variational Bayes”, arXiv:1312.6114, 2013

[Srivastava17] A Srivastava, C Sutton, “Autoencoding Variational Inference For Topic Models”, ICLR, 2017

[Srivastava18] A Srivastava, C Sutton, ”Variational Inference In Pachinko Allocation Machines”, arXiv:1804.07944, 2018

[Miao17] Y Miao, E Grefenstette, P Blunsom, “Discovering Discrete Latent Topics with Neural Variational Inference”, ICML, 2017