1C-03

# Variational Autoencoder を用いたマルチモーダル情報の統合

青木　達哉 †‡　長井 隆行 †
電気通信大学大学院 †　日本学術振興会特別研究員 DC‡

## 1　はじめに

近年，実社会のサービスに対し，機械学習の技術を活用することに関心が集まっており，医療，経済活動，ロボットなど応用例は多岐に渡る．しかし，機械学習の技術はまだまだ発展途中であり，実社会の情報に対応しきれてはいない．その一例として，マルチモーダル情報の学習が挙げられる．実環境の事象は複雑であるため，複数種のセンサを組み合わせ，得られた情報を統合することで対象を正確に把握できる．今後，機械学習で扱う対象が多様化するにつれ，マルチモーダル情報処理の必要性は高くなると予想される．

本研究では，マルチモーダル情報に対し，対象を表現する潜在情報が存在すると仮定する．マルチモーダル情報の学習において，観測情報から適切な潜在情報を抽出することが重要と考える．このような理由からマルチモーダル情報の学習には生成モデルを用いることが適切と考え，Variational Autoencoder(VAE) に基づいた深層生成モデルを検討する．先行研究として，文献 [2]，[3] で画像情報とラベル情報の 2 種のマルチモーダル情報に対する学習が提案されている．本研究では，別のモダリティの組み合わせで構成されるデータに対し同様に学習が可能であるかを検証する．また，学習結果の活用と解析のために，入力情報が部分的に観測された場合における汎用的な潜在情報の推論法を提案する．

## 2　理論

### 2.1　Joint Multimodal Variational Autoencoder

Joint Multimodal Variational Autoencoder (JMVAE) は，Suzuki らが文献 [2] で提案した深層学習を用いたマルチモーダル情報に対する生成モデルである．JMVAE は，Kingma らが文献 [1] で提案した VAE をベースとし，1 つの潜在情報より複数のモダリティ情報が生成される過程をモデル化している．グラフィカルモデル及びネットワーク構造は図 1 のように表される．$\boldsymbol{X}_n$ は $n$ 種目の観測情報，$\boldsymbol{Z}$ は潜在情報を意味する．このモデルは赤枠で示した Encoder と青枠で示した Decoder の 2 つの部分に分割して考えられる．

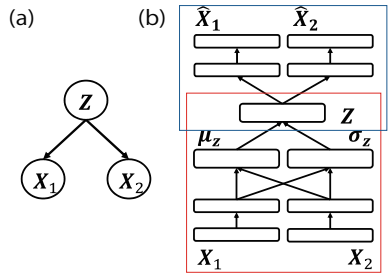


図 1　Joint Multimodal Variational Autoencoder
(a) グラフィカルモデル (b) ネットワーク構造

$$q_\phi(\boldsymbol{z}|\boldsymbol{X}_1,\boldsymbol{X}_2)=\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_z,\boldsymbol{\Sigma}_z) \tag{1}$$

$$P_\theta(\boldsymbol{X}_1,\boldsymbol{X}_2|\boldsymbol{z})=P_{\theta_1}(\boldsymbol{X}_1|\boldsymbol{z})P_{\theta_2}(\boldsymbol{X}_2|\boldsymbol{z}) \tag{2}$$

ネットワーク全体を最適化するために，Encoder は，式 (1)，Decoder は，式 (2) の確率分布のパラメータ $\phi$ 及び $\theta$ をそれぞれ推定する．

### 2.2　部分観測時における潜在情報 $\boldsymbol{z}$ の推論

入力情報のうち，一部のモダリティ情報が観測された場合に，$\boldsymbol{z}$ が適切に推論することができれば，学習結果を未観測モダリティ情報の予測のために利用できる．また，潜在情報を介した予測結果から潜在情報の学習結果の解析も可能になる．しかし，図 1 で示すように，学習された Encoder を用いて，$\boldsymbol{z}$ を推論するには，全てのモダリティ情報が不可欠である．そこで，別の推論法として Decoder を用いる方法を考える．Decoder で，$\boldsymbol{z}$ を決めることは，$P(\boldsymbol{X}_0|\boldsymbol{z}),P(\boldsymbol{X}_1|\boldsymbol{z})$ の確率分布を決めることと等しい．そのため，観測情報に対する最適な $\boldsymbol{z}$ は，尤度 $P(\boldsymbol{X}_0|\boldsymbol{z})P(\boldsymbol{X}_1|\boldsymbol{z})$ を最大化すると考えられる．さらに，部分観測な情報 $\boldsymbol{X}_0$ からの潜在情報の推定は，$\boldsymbol{X}_1$ について積分消去を考えると，$P(\boldsymbol{X}_0|\boldsymbol{z})$ を最大化する $\boldsymbol{z}$ の推定となる．この方法であれば，推論時に観測されるモダリティ数が変化しても，観測情報に対する尤度を最大化する $\boldsymbol{z}$ の推定問題として推論を考えられる．

---

The integration of multimodal information using Variational Autoencoder
Tatsuya Aoki†‡, Takayuki Nagai†
†The University of Electro-Communications
‡JSPS Research Fellow DC

表 1　未観測モダリティ情報の予測誤差

| 聴覚情報から視覚情報の平均累積予測誤差 | | | |
|---|---|---|---|
| モデル | 提案法 | 比較法 1 | 比較法 2 |
| JMVAE | 28239.33 | **26214.75** | - |
| JMVAE-kl | **28444.71** | 37210.02 | 32694.21 |

| 視覚情報から聴覚情報の平均累積予測誤差 | | | |
|---|---|---|---|
| モデル | 提案法 | 比較法 1 | 比較法 2 |
| JMVAE | **112.01** | 176.61 | - |
| JMVAE-kl | **121.00** | 132.44 | 127.52 |

表 2　観測情報の復元誤差

| モデル | 聴覚情報 | 視覚情報 |
|---|---|---|
| JMVAE | 110.08 | 26214.75 |
| JMVAE-kl | 119.61 | 24350.01 |

## 3　実験

検証用に，物体から視覚情報 (DSIFT+ ベクトル量子化 500 次元)，聴覚情報 (MFCC+ ベクトル量子化，50 次元) の 2 種のモダリティの組み合わせのデータを用意した．データ数は 67 であり，観測対象の物体は 11 種に分類される．学習のためのモデル構造として，JMVAE 及び文献 [2] において提案された JMVAE-kl の 2 種類を用いた．JMVAE-kl はマルチモーダル情報の学習時に同時に単モーダルに対する補助 Encoder を学習するモデルである．学習結果を用いた潜在情報の推定は，Decoder を用いる方法 (提案法)，比較方法として，未観測モダリティ情報は要素が全て 0 であるベクトルとして Encoder を用いる方法 (比較法 1 )，JMVAE-kl のみ補助 Encoder を用いる方法 (比較法 2) の 3 通りで行った．

推定結果の評価のために，推定時に未観測モダリティとして扱った情報を目標値とし，$z$ を介した予測値との累積誤差を算出した．表 1 に累積誤差の平均値，表 2 に完全観測時の復元誤差の平均値を示す．また，推定した潜在情報の適切さを確認するため，全情報を観測した場合の Encoder による推定結果 (a) と各手法の推定結果 (b)(c)(d) を図 2, 図 3 で比較した．図中の各点の色は物体の種類，座標は潜在情報の値を表す．

平均累積予測誤差と推定結果の比較から Decoder を用いた潜在情報の推定法は，他の推定法に対し同等以上の精度で推定が行えると考えられる．加えて，特別な構造がない JMVAE でも同様に推定できたことは，提案法の汎用性を示す結果といえる．また，全観測時の復元誤差と部分観測時の予測誤差の差を見ると，低次元情報から高次元情報を予測する場合に相対的に誤差が大きくなっている．これは観測情報の次元数が低い場合，推論すべき情報が多くなることが影響していると予想される．

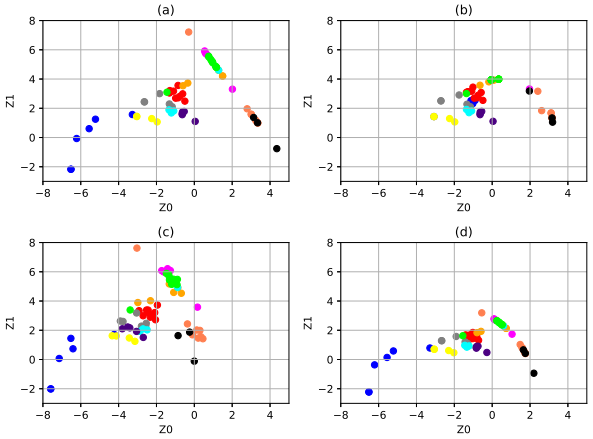


図 2　JMVAE-kl による視覚情報からの推定結果 (a) 潜在空間 (b) 提案法 (c) 比較法 1 (d) 比較法 2

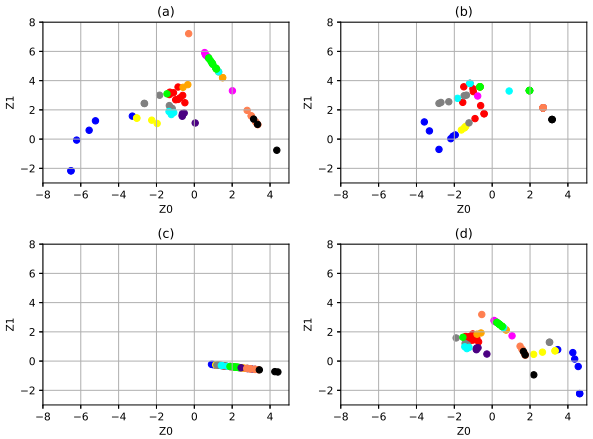


図 3　JMVAE-kl による聴覚情報からの推定結果 (a) 潜在空間 (b) 提案法 (c) 比較法 1 (d) 比較法 2

## 4　まとめ

本稿では，VAE を拡張したマルチモーダル情報に対する深層生成モデルの潜在情報の推論において部分観測下での新しい推論法を提案した．検証実験を通して，提案した推定法の有効性と汎用性が確認できた．今後，提案した推論法を利用し，潜在情報の学習結果を解析し，VAE を用いたマルチモーダル情報の統合学習のための最適なモデルの検討を進める予定である．

## 謝辞

本研究の一部は，JST CREST(JPMJCR15E3) 及び JSPS 科研費 (17J10512) の支援を受けて実施した．

## 参考文献

[1] D.P.Kingma et al. ”Autoencoding variational bayes.”, arXiv preprint arXiv:1312.6114, 2013.
[2] M.Suzuki et al. ”Joint multimodal learning with deep generative models.”, arXiv preprint arXiv:1611.01891, 2016.
[3] R.Vedantam et al. ”Generative models of visually grounded imagination.”, arXiv preprint arXiv:1705.10762, 2017.