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

訓練済み深層生成モデルの潜在変数間相互条件付きサンプリングによるマルチモーダル双方向生成モデル

Bi-directional multimodal generetion via estimating conditional distribution of latent variables obtained from pre-trained generative models

今給黎 成彬 Shigeaki Imakiire　大知 正直 Masanao Ochi　森 純一郎 Junichiro Mori　坂田 一郎 Ichiro Sakata

東京大学大学院工学系研究科技術経営戦略学専攻
University of Tokyo School of Engineering Department of Technology Management for Innovation

In recent years, research on multimodal generation that mutually converts between different data such as images and sentences has attracted attention from the viewpoint of applicability to real service such as automatic annotation of images and subtitles of audio.

Meanwhile, in the field of machine learning research, reusable trained models trained using large-scale data sets are being opened to the public, and the number is expected to increase in the future.

Therefore, in this research, we aim to realize multimodal generation with small data by utilizing this trained model. In this paper, we propose a multimodal generation method using a trained generation model in which latent variables of individual modality can be inferred and a small amount of data set. We realized multimodal generation by estimating the conditional distribution of latent variables obtained from trained models using small number of train data.

1. はじめに

近年，画像や文章などの異なるデータ間を相互に変換するマルチモーダル生成の研究が実サービスへの応用可能性の観点から注目されている．マルチモーダル生成は，画像の自動注釈，音声の字幕生成, など最近でも大きな成果を得ており産業応用上重要な技術である一方で, 大規模なデータセットが必要であるという課題がある．

一方で，機械学習研究分野は，大規模なデータセットを用いて訓練した再利用可能な訓練済みモデルを一般に公開するようになってきており，その数は今後増加すると見込まれる．

そこで本研究では，この訓練済みモデルを活用することで，小規模なデータでマルチモーダル生成を実現することを目的とする．本稿では，個別モダリティの潜在変数が推論可能な訓練済み生成モデルと少量のデータセットによってマルチモーダル生成を行う手法を提案する．

本稿では画像とラベルからなるマルチモーダルデータに提案手法を適用する実験を行い，与えられた訓練済み生成モデルが十分訓練されていれば提案手法によって少数の訓練データでそれらのモデルと同等程度の精度でマルチモーダル双方向生成が可能なことを示した．

2. 関連研究

2.1 前提となる代表的な深層生成モデル

Kingma らによる Variational Autoencoders[1](以下，VAE) は代表的な深層生成モデルである．VAE は入力データから潜在変数を推論する Encoder と潜在変数から入力データを復原する Decoder という 2 つのニューラルネットワークを組み合わせた深層生成モデルである．VAE の対象とするデータ構造は画像だけでなく，テキスト [2] や音声 [3] などに対するモデルも派生的に提案されている．

Genereative Adversarial Networks[4](以下，GAN) は，潜在変数からデータを生成する Generator とその生成データの真偽を判定する Discriminator を互いに学習させることで，現実のデータに近いデータ生成を可能にする深層生成モデルである．Goodfellow らが提案した元の GAN には入力データから潜在変数の分布を推論する過程は備わっていないが，Dumoulin らによる ALI[5] などそれを可能にするモデルが提案されている．

連絡先: 今給黎 成彬 imakiire@utac.u-tokyo.ac.jp

2.2 単方向のマルチモーダル生成モデル

単方向のマルチモーダル生成モデルが扱う問題は画像からテキストの生成 [6]，テキストから画像の生成 [7]，テキストから音声の生成 [8] など多岐に渡っている．単方向のマルチモーダル生成の主要なアプローチは，一方のモダリティのデータの特徴を抽出する Encoder と抽出された特徴を入力として対応モダリティを出力する Decoder を構築するという方法である．特徴抽出とそのデコードをいかに行うかが重要であるため，モデルが複雑化し，一般的には大規模なデータセットが必要となる．

2.3 双方向のマルチモーダル生成モデル

双方向のマルチモーダル生成モデルは単方向のマルチモーダル生成モデルよりも学習のコストが低いという点で優れている．双方向マルチモーダル生成モデルの主なアプローチは異なるモダリティの共有表現を獲得する方法である．共有表現の獲得の方法としては Deep Boltzman Machine を用いるモデル [9] や AutoEncoders を用いるモデル [10], VAE を用いるモデル [11] などが存在する．

2.4 本研究の位置づけ

本研究ではマルチモーダルデータについて各モダリティについて潜在変数の推論が可能な生成モデルの訓練済みモデルが与えられた時に，それらの潜在変数の相互の条件付きサンプリングを通じてマルチモーダル双方向生成を行うモデルを提案する．

3. 提案手法

本節では，各モダリティ毎に潜在変数の推論が可能な訓練済み生成モデルが所与であるものとし，一方のモダリティの潜在変数で条件付けた異なるモダリティの潜在変数の分布を推定し，その推定分布からサンプリングされた潜在変数と所与のモ

デルの生成過程によって異なるモダリティデータの生成を行う手法を提案する．

(x, y) という複数のモダリティを持つデータセット $D = \{(x_1, y_1),...,(x_N, y_N)\}$ が存在し，それぞれのモダリティについて潜在変数 z_x, z_y が推論可能な訓練済みの生成モデルが所与であるとする．本研究では，x のみが観測された時に対応する y を獲得する問題と y のみが観測された時に対応する x を獲得する問題の 2 つの問題を考える．本稿では，この問題に潜在変数 z_x, z_y の相互の条件付き分布のモデリングするというアプローチを提案する．

3.1 訓練済み生成モデルの潜在変数間の条件付き分布推定によるマルチモーダル生成

x と y は対称性があるため，まずは x を観測した下で y を獲得する問題を考える．

y の生成過程を考慮すれば，x が与えられた下での z_y の分布 $p(z_y|x)$ を考えればよいが，x は低次元の z_x により生成されたものであるため z_x を z_y に対応付ける方が複雑さが小さい．したがって z_x が与えられた下での z_y の分布 $p(z_y|z_x)$ を考える．すると，以下の手順でモダリティ x のあるデータ x_i が観測された時に以下の手順で対応する y_i を生成できる．

1. 観測した x_i で条件付けた z_x の分布 $p(z_x|x_i)$ から z_{x_i} をサンプリングする
2. サンプリングした z_{x_i} で条件付けた $p(z_y|z_{x_i})$ から z_{y_i} をサンプリングする
3. サンプリングした z_{y_i} で条件付けた $p(y|z_{y_i})$ から y_i をサンプリングする

$p(z_x|x)$ は潜在変数 z_x の推論過程，$p(y|z_y)$ は y の生成過程であり，それぞれ所与の訓練済みモデルで十分よく近似されているとする．よって本稿で取り扱うのは $p(z_y|z_x)$ の推定である．

より具体的に $p(z_y|z_x)$ をモデル化する方法として，$q_{\psi_x}(z_y|z_x)$ を真の $p(z_y|z_x)$ の近似分布として，z_x を入力として q の分布パラメータを出力する関数 h_{ψ_x} を考える．ただし，q は所与のモデルにて z_y の事後分布に仮定されている分布と同一形状の分布とする．すなわち，例えば $p(z_y|y)$ に多変量正規分布が仮定されているならば，$q_{\psi_x}(z_y|z_x)$ も多変量正規分布を仮定する．ここで，その分布パラメータを γ_y とすると $q_{\psi_x}(z_y|z_x)$ は以下のように書き表すことができる．

$$q_\psi(z_y|z_x) = Dist_y(z_y|\gamma_y = h_{\psi_x}(z_x; \psi_x)) \tag{1}$$

h のパラメータ ψ_x を最適化すれば所与の訓練済みモデルと組み合わせて x を観測した下で対応する y の生成が可能になる．同様に y を観測した下で対応する x を獲得したい場合は，z_y を入力として z_x が従う分布のパラメータを出力する関数 h_{ψ_y} を考える．

3.2 最適化問題

まずは h_{ψ_x} の最適化を考える．(x_i, y_i) というペアのデータのうち x_i のみが観測された場合に，関数 h_{ψ_x} がどのような性質を満たせば y_i を生成できるかを考える．3.1 節で記した通り，y_i を生成するならば，分布 $p(z_y|y_i)$ からサンプリングされた潜在変数 z_{y_i} が必要である．よって，x_i によって z_{x_i} が推論された下で z_{y_i} を獲得するためには，$q_{\psi_x}(z_y|z_{x_i})$ と $p(z_y|y_i)$ を一致させればよい．すなわち，h_{ψ_x} は z_{x_i} を入力として $p(z_y|y_i)$ に近い分布のパラメータを出力する関数であればよい．

$D[p||q]$ を確率分布 p と q の間に定義され，分布間の距離を評価できるダイバージェンス関数とすると，パラメータ ψ_x はそのダイバージェンス関数を最小化するような値であればいい．これは以下の式を満たすことと等しい．

$$\min_{\psi_x} D[q_{\psi_x}(z_y|z_{x_i})||p(z_y|y_i)] \tag{2}$$
$$\text{ただし，} z_{x_i} \sim p_(z_x|x_i)$$

$D[q_{\psi_x}(z_y|z_{x_i})||p(z_y|y_i)]$ は x_i, y_i, ψ_x の関数として表せるためこれを $L_i(\psi_x, x_i, y_i)$ とする．ここで，目的は特定の x_i だけでなく，データセット $\{(x_1, y_1), ..., (x_N, y_N)\}$ の各点において ψ_x が式 2 を満たすことである．よって，$L_i(\psi_x, x_i, y_i)$ の期待値の最小化を考える．

$$\min_{\psi_x} \mathbb{E}_{(x_i,y_i)\sim p_{data}} \left[L_i(\psi_x, x_i, y_i) \right] \tag{3}$$

式 3 の期待値は $X = \{x_1, x_2, ..., x_N\}, Y = \{y_1, y_2, ..., y_N\}, \psi_x$ の関数として表されるためこれを $L(\psi_x; X, Y)$ とする．N が十分に大きいと L の計算はメモリ効率上困難となるため，バッチ学習を提案する．すなわち，データセット $\{(x_1, y_1), ..., (x_N, y_N)\}$ からランダムな M 個のサンプル X^M, Y^M を抽出し，以下の関数 L^M で L を近似する．

$$\begin{aligned} L(\psi_x, X, Y) &\simeq L^M(\psi_x, X^M, Y^M) \\ &= \frac{N}{M} \sum_{i=1}^{M} L_i(\psi_x, x_i, y_i) \end{aligned} \tag{4}$$

ダイバージェンス関数やモデルパラメータの最適化には様々な選択が考えられるが，本稿においてはダイバージェンス関数には KL ダイバージェンスを選択し，モデルパラメータの最適化には勾配降下法を用いる．

4. 実験の概要

本実験では，第 3. 節のモデルを画像とラベルからなるデータセットに適用してそれらの相互生成を行った．本実験の目的は提案手法の有効性を検証するために以下の 3 点を確認することである．

1. 提案手法によって異なるモダリティの潜在変数間の条件付き分布が正しく推定され，所与の訓練済み生成モデルと組み合わせてマルチモーダル双方向生成が可能となること
2. 提案手法においてモデルパラメータの最適化が少数のデータセットでも可能であること
3. 双方向の生成が与えられた訓練済み生成モデルと同程度の精度で行われること

4.1 データセット

使用したデータは MNIST データセットである．MNIST は 70,000 件の手書き数字画像データとその画像に書かれた数字を表すラベルデータからなるデータセットである本実験では全データセットのうち 60,000 件を各モダリティの生成モデルの訓練に利用し，N 件を提案モデルの訓練に利用し，残りの $(10,000 - N)$ 件を評価で利用した．

4.2 訓練済み生成モデル

本実験では事前に各モダリティの VAE の訓練を行い，その後に各モダリティの潜在変数間の条件付き分布の推定を行う提案モデルを訓練した．

どちらのモダリティについても潜在変数の分布は正規分布とし，潜在変数の次元は画像は 64 次元，ラベルは 2 次元とした．どちらのモデルも十分に学習させた．

4.3 提案モデルの訓練詳細

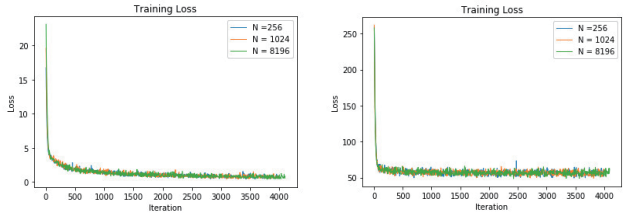
提案モデルに用いた構造は，2 層の BatchNormalization 及び LeakyReLU 活性化関数付きの線形全結合層と潜在変数の事後分布パラメータを出力する全結合層からなる多層パーセプトロンである．隠れ層の出力ユニット数は順に 64，128 とし，最終層の出力ユニット数はターゲットのモダリティの潜在変数の次元数とした．また，LeakyReLU 関数の negative slope は 0.2 とした

モデルのパラメータ最適化に必要なデータ数を確認するために訓練データ数 $N = \{256, 1024, 8192\}$ で訓練を行った．

5. 結果と考察

5.1 学習の進展

図 1 は提案モデルを訓練中の目的関数の値の推移である．



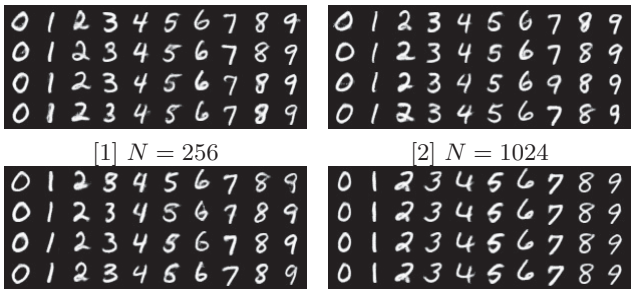
[1] image to label　　[2] label to image

図 1: 目的関数の値の推移

目的関数の値は学習が進むにつれて減少していることから異なるモダリティの潜在変数間の条件付き分布が推定されていることが分かる．また，学習の速度と最終的な収束値は訓練データ数によって変わっていないことから訓練データ数の増減が提案モデルの学習に及ぼす影響は小さいと考えられる．

5.2 ラベルデータから画像データの生成

0 ～ 9 のラベルデータから画像データを生成した結果と訓練済み画像 VAE による生成をまとめたものが図 2 である．



[1] $N = 256$　　[2] $N = 1024$

[3] $N = 8196$　　[4] 訓練済み画像 VAE

図 2: ラベルから生成した画像と訓練済み画像 VAE によって生成した画像

どの訓練データ数においても入力ラベルに合致した画像が生成され，生成された画像の質も訓練済み VAE によるものと同等精度であることが分かる．

次に，画像データ x の対数尤度 $logp(x)$ の推定値による生成精度の定量評価を行う．本実験の目的は提案手法によるデータの生成が所与の訓練済みモデルと同程度であることの確認であるため，対数尤度の比較は所与の訓練済みモデルによる生成と提案手法の生成の間で行う．表 1 はその結果である．

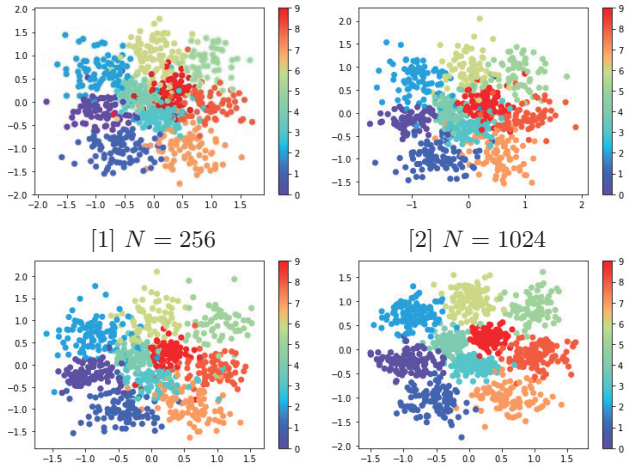
表 1: 提案手法と訓練済み画像 VAE による対数尤度の比較

モデル	$logp(x)$
$N = 256$	-351.52
$N = 1024$	-354.59
$N = 8192$	-355.56
訓練済み画像 VAE	-89.99

どの訓練データ数においても提案手法は元の訓練済みモデルよりも低い対数尤度となった一方で，訓練データ数による大きな差は見られない．

5.3 画像データからラベルデータの生成

潜在変数の分布が正しく推定されているかを検証するために，画像データの潜在変数からラベルデータの潜在変数を二次元平面上に可視化した結果を確認する．図 3 はその結果である．



[1] $N = 256$　　[2] $N = 1024$

[3] $N = 8192$　　[4] 訓練済みラベル VAE

図 3: 画像から推定したラベルの潜在変数と訓練済みラベル VAE で推定した潜在変数のプロット

どのデータ数においても訓練済み VAE と同様のプロットとなっていることが分かる．

次に，ラベルデータ y の対数尤度 $logp(y)$ の推定値による生成精度の定量評価を行う．結果は表 2 である．対数尤度には推定値を用いているため，値の大小関係が変動しうることに注意されたい．

表 2: 提案手法と訓練済みラベル VAE による対数尤度の比較

モデル	$logp(y)$
$N = 256$	-0.0456
$N = 1024$	-0.0456
$N = 8192$	-0.0457
訓練済みラベル VAE	-0.0456

訓練データ数間での差はなく，かつ訓練済みモデルと同等精度であることが分かる．

5.4 所与の訓練済みモデルの学習が不十分な場合

双方向生成が所与の訓練済みモデルと同等精度で行われることを確認するために，画像データの VAE に 3 エポックのみしか訓練していない学習途中の重みを使って提案手法を学習した．目的は双方向生成が所与の訓練済みモデルと同等精度で行われるかの確認なので，訓練データ数の比較は行わず $N = 1024$ のみで実験を行った．

図 4 はラベルデータから画像データを生成した結果である．ラベルと対応した画像が生成されているが．生成の質は画像 VAE よりもわずかに劣っている．

[1] 提案手法による生成結果

[2] 学習途中の画像 VAE からの生成結果．上段は真の画像である．

図 4: 学習途中の画像 VAE を用いた場合の生成結果

図 5 は画像データからラベルデータの潜在変数を推定した結果である．十分に訓練した場合である図 3 と比較すると異なるラベルの潜在変数が重なりあっていることが分かる．

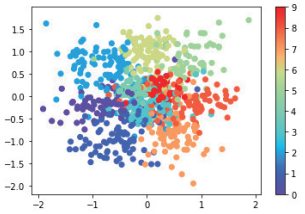


図 5: 提案手法によって推定した潜在変数のプロット

5.5 考察

以上の実験結果に対する考察を行う．5.2 節においてラベルデータから対応する画像データが生成されたこと，5.3 節において多くの画像から対応するラベルデータが正しく生成されたこと，5.3 節において提案手法によって画像データから推定したラベルデータの潜在変数が訓練済みラベル VAE によって推定した潜在変数と同様の分布をしたこと，これら三点によって提案手法によってモダリティの潜在変数間の条件付き分布の推定を介してマルチモーダル生成が可能になったことが分かった．

5.1 節における学習の進展速度と目的関数の収束値にデータ数による違いがないこと，5.2 節と 5.3 節における定性及び定量評価にデータ数による差が見られなかったこと，これらによって提案手法は少数の訓練データで学習が可能なことが分かった．

5.2 節において提案手法によってラベルから生成した画像が質の面で訓練済み画像 VAE による生成画像と同等程度であったこと，5.3 節において提案手法の尤度評価が訓練済み VAE と同等程度なこと，5.4 節において与えられたモデルが未学習ならば生成の質が落ちること，これら三点からは所与の訓練済み生成モデルの潜在変数の推論過程が十分パラメータ化されていれば提案手法による生成は与えられたモデルによる生成と同等程度の精度で行われることが分かった．

6. 結論

様々なモダリティについて個別に高精度な生成モデルが提案され，かつ訓練済みモデルの公開が一般的となりつつある現状に着目して，訓練済み生成モデルと小規模なデータセットを組み合わせることで十分に高精度なマルチモーダル双方向生成を達成することが本研究の目的であった．

まず複数モダリティについてそれぞれ訓練済みの生成モデルを所与として，その潜在変数間の相互の条件付き分布を推定しその分布からのサンプリングによって得られた潜在変数と所与のモデルの生成過程を用いることによってマルチモーダル生成を行うモデルを提案した．そして，特に所与の訓練済み生成モデルが Variational Autoencoders である場合に，潜在変数間の相互の条件付き分布を多層パーセプトロンによって推定することでたしかにマルチモーダル生成が可能であることを実験によって示した．また，従来手法とは異なり潜在変数間の条件付き分布を推定するのみなので，モデルを簡略化することによってパラメータ数を減らし小規模なデータセットによってモデルパラメータの最適化が可能であることを確認した．さらに，所与の訓練済み生成モデルが十分に訓練されているならばそれと同等程度の生成が可能なことを確認した．

以上より，本研究はマルチモーダルデータについての訓練済み生成モデルが得られた一方で対応する訓練データセットは少数しか得られない場合に十分よく双方向生成を達成する手法を提案し，その有効性を示した．

参考文献

[1] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[2] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. *arXiv preprint arXiv:1703.00955*, 2017.

[3] Wei-Ning Hsu, et al. Learning latent representations for speech generation and transformation. In *Interspeech*, pp. 1273–1277, 2017.

[4] Ian J. Goodfellow, et al. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pp. 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

[5] Vincent Dumoulin, et al. Adversarially learned inference. *CoRR*, Vol. abs/1606.00704, , 2016.

[6] Quanzeng You, et al. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4651–4659, 2016.

[7] Elman Mansimov, et al. Generating images from captions with attention. In *ICLR*, 2016.

[8] Aäron van den Oord, et al. Wavenet: A generative model for raw audio. In *SSW*, 2016.

[9] Nitish Srivastava, et al. Multimodal learning with deep boltzmann machines. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pp. 2222–2230. Curran Associates, Inc., 2012.

[10] Jiquan Ngiam, et al. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 689–696, USA, 2011. Omnipress.

[11] Masahiro Suzuki, et al. Improving bi-directional generation between different modalities with variational autoencoders. *arXiv preprint arXiv:1801.08702*, 2018.