私のブックマーク

深層ベイズ学習と音環境理解†1

坂東　宜昭（産業技術総合研究所，https://ybando.jp/）

1. は じ め に

非線形写像を効率的に学習できる深層学習は，音声認識・音イベント検出や音響信号処理でも圧倒的な性能を達成している．特に，**雑踏下での読上げ音声認識** [1] や**ホームパーティでの口語音声認識** [2]，**日常・市街地での音イベント検出** [3] などは，国際技術評価会も開催され活発に研究されている．また信号処理においても，**不要な雑音を除去する音声強調** [4] や，**複数話者の同時発話から個別の音声を抽出する音声分離** [5]，**多種多様な音イベントの分離** [6] などで，目覚ましい発展を遂げている．

多くの深層学習に基づく枠組みでは，非線形写像の入力と出力を定義するために教師付きデータを必要とする．そのため，“教師データに含まれない未知環境では性能が劣化”したり，“十分な量の教師データを収集するには膨大なコストが必要”だったりといった課題が存在する．特に，日常環境の音響信号は複数の音を観測した混合音として収集されるため，雑踏などの複雑環境の音ラベルや，自然界の偶発的な音イベントの音源信号といったデータでは，教師データの準備コストが高く事実上不可能な場合もある．できるだけ少ない教師データから，音環境を認識・理解できる技術が望ましい．

従来，**統計的信号処理** [7] や**ベイズ学習** [8] の枠組みでは，まず観測の生成過程を表現する確率モデルを構築し，そのパラメータの最尤値や事後分布を推論することで，教師なしでさまざまなタスクを解いていた．確率モデルを構築するときに導入した仮定が妥当である限り，学習データを用いずとも高精度に所望の情報を抽出できる．例えば，マイクロホンアレーを用いる多チャネル音源分離では，音の空間的な特性が時間周波数領域での瞬時混合でよく近似できることが知られている．また，音楽や多くの騒音のパワースペクトログラムは，低ランク近似によりある程度表現できる．このような統計的信号処理の知見を深層学習とうまく組み合わせられれば，深層学習の高い表現性とベイズ学習の高いロバスト性が両立できると期待できる．

深層ベイズ学習は，ベイズ学習の枠組みにうまく深層学習を導入することで，両者の利点を活用するハイブリッドな技術を実現する枠組みである．ここでは，深層ベイズ学習の紹介から始めて，その音響信号処理や音環境理解への応用事例について紹介する．

2. ベイズ深層学習と深層ベイズ学習

深層学習とベイズ学習を融合する研究は，ベイズ深層学習（Bayesian deep learning）として広く知られている．例えば，**ベイズニューラルネットワーク** [9] や**深層生成モデル** [10]，**変分ドロップアウト** [11] などが有名である．これらの一部は，深層ベイズ学習（deep Bayesian learning）とも呼ばれ，ベイズ学習を深層学習で拡張する立場（深層ベイズ学習）と，深層学習をベイズ的に拡張・解釈する立場（ベイズ深層学習）に区別されることがある．筆者は，前者の立場に立つことが多く，深層ベイズ学習の用語を用いている．以下では，深層学習とベイズ学習を融合するうえで重要な概念である深層生成モデルと償却変分推論について紹介する．

- **深層生成モデル**

  深層生成モデルは，**従来の主成分分析** [12] に代表される線形な確率モデルを，ニューラルネットワークを用いて非線形に拡張したモデルである．以降で説明する**変分オートエンコーダ**（VAE）[13] や，**敵対的生成ネットワーク** [14]，Flow [15] などが知られている．

- **変分償却推論**

  償却変分推論は，確率モデルの事後分布推論を学習データセットを用いて深層ニューラルネットワーク（DNN）に記憶させることで，ベイズ推論を効率化する枠組みである．この枠組みは VAE でも用いられており，推論が

†1 http://www.ai-gakkai.or.jp/my-bookmark_vol35-no6

難しい深層生成モデルの事後分布推論をエンコーダネットワークとして同時学習している．

3. 変分オートエンコーダ

深層ベイズ学習における最も重要な技術の一つである，VAE について紹介する．

・ Kingma, D. P., et al.: Auto-encoding variational Bayes, *arXiv:1312.6114*（2013）[16]

VAE には，深層生成モデルと償却変分推論をうまく組み合わせた，深層ベイズ学習のエッセンスが詰まっている．この枠組みでは，まず観測の特徴を表す潜在変数を考える．そしてこの潜在変数から観測が生成される過程を DNN を用いて構成する．この生成モデルのパラメータ（DNN の重み）は，観測（学習データ）に対する周辺尤度を最大化するように学習するが，この周辺尤度は直接計算が難しい．そこで VAE では，生成モデルをデコーダとみなし，潜在変数の事後分布を推定するエンコーダ（推定ネットワーク）をもつオートエンコーダ型の学習を行う．推定された近似（変分）事後分布を補助変数として周辺尤度の変分下限を導出し，周辺尤度の代わりにこの変分下限を最大化する．変分下限と周辺尤度のギャップは変分事後分布と真の事後分布の間のカルバック・ライブラーダイバージェンスと対応しており，変分下限を最大化することで，エンコーダネットワークは生成モデルの事後分布推論を獲得する．このように，償却変分推論を活用することで，深層生成モデルを効率的に学習できるようになった．

VAE にはさまざまな拡張が研究されているが，以下に主要な研究を紹介する．

・ Dilokthanakul, N., et al.: Deep unsupervised clustering with Gaussian mixture variational autoencoders, *arXiv:1611.02648*（2016）[17]

例えばこの論文では，潜在変数の事前分布を正規分布から，混合正規分布（GMM）に変更することで，深層生成モデルとクラスタリングの同時学習を実現した．

・ Jang, E., et al.: Categorical reparameterization with Gumbel-softmax, *Proc. of ICLR*（2017）[18]

・ Maddison, C. J., et al.: The concrete distribution: A continuous relaxation of discrete random variables, *arXiv:1611.00712*（2016）[19]

これらの論文では，微分可能な離散潜在変数として，離散分布を近似するガンベル・ソフトマックス分布が提案されている．

・ Chung, J., et al.: A recurrent latent variable model for sequential data, *Proc. of NeurIPS*（2015）[20]

またこの論文では，音響信号処理や自然言語処理において特に重要となる，時系列信号に対する VAE が提案されている．

このような VAE の成果は，定式化が難しい観測に対する生成モデルとして応用が進んでいる．例として，単眼カメラの画像列からカメラ軌跡と周辺物体の形状を同時推定する simultaneous localization and mapping（SLAM）[21] への応用を紹介する．

・ Bloesch, M., et al.: CodeSLAM — Learning a compact, optimisable representation for dense visual SLAM, *Proc. of CVPR*（2018）[22]

カメラで観測される輝度画像と各画素に対応する深度画像の関係は不良設定問題のため，従来の確率モデルでは扱いが難しかった．CodeSLAM では，VAE を用いて輝度画像と深度画像の関係を表す生成・推論モデルを事前学習し，深度画像を表現する低次元の潜在表現（code）を獲得する．推論時は，輝度画像列に対する投影誤差を最小化するように潜在変数列を更新することで，カメラ軌跡と密な物体形状を同時推定する．このように，扱いが難しい生成過程を VAE で表現しながら，幾何制約といった定式化しやすい制約を活用することで，DNN や確率モデル単体ではできなかったさまざまな応用が期待できる．

4. 音環境理解への展開

深層ベイズ学習の音環境理解・音響信号処理への応用を紹介する．

4·1 変分オートエンコーダの音響信号処理への応用

VAE を音響信号処理へ応用した事例をいくつか紹介する．

・ Wu, Y., et al.: Semi-supervised neural chord estimation based on a variational autoencoder with latent chord labels and features, *arXiv:2005.07091* [23]

この研究では，音楽信号からコード進行を推定する DNN を学習する．コード認識は，楽曲を計算機が理解す

るために不可欠であるが，大量の楽曲にコード進行をアノテーションするのは専門知識を要し容易ではない．そこで，コード進行から音響信号（クロマベクトル）を生成する深層生成モデルとその推論ネットワークを同時学習する枠組みを提案し，すべてのデータに教師がついていなくとも，推論ネットワークを学習できる半教師あり学習を実現している．特に，コード進行のマルコフ性に基づく正則化を導入することで，高い性能を達成している．

- Kawachi, Y., et al.: Complementary set variational autoencoder for supervised anomaly detection, *IEEE/ICASSP*（2019）[24]

  他の分野でも用いられているように，VAE は学習データに対する密度関数を獲得できるため，異常検知にも応用されている．この研究では，正常値の潜在変数には標準正規分布を，異常値には標準正規分布と補完的になる（ドーナツ状の）分布を仮定することで，高精度に異常値を検出する枠組みを提案している．このほかにも，異常検知の性能評価が行われた DCASE 2020Task2 [25] では，VAE に基づく枠組みがいくつか提案されている．

4·2　深層音声事前分布に基づく半教師あり音声強調

VAE の応用の中でも特に，従来の確率的生成モデルと組み合わせた，半教師あり音声強調を紹介する．

- Bando, Y., et al.: Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization, *Proc. of IEEE ICASSP*（2018）[26]

  音声強調とは，雑音と音声が混合された信号から音声のみを抽出・分離するタスクで，音声コミュニケーションや音声認識のフロントエンドに不可欠な技術である．近年，**深層学習に基づく枠組み** [27] が高い性能を達成しているが，学習データに含まれていない**未知環境では性能が劣化** [28] する問題があった．そこでこの研究では，音のスペクトログラムを基底スペクトル行列（テンプレート）とそのアクティベーション行列（時間パターン）へ分解する**非負値行列因子分解**（NMF）[29] に着目した，VAE-NMF が提案されている．このモデルでは，事前学習が不要で環境ロバスト性が高い NMF で雑音信号を表現し，事前学習により精緻な表現ができる VAE で音声信号を定式化する．観測混合音はこれらの和として定式化され，推論時は NMF と VAE の潜在変数を一挙にベイズ推定することで，音声と雑音を分離する．このように従来の統計モデルと深層生成モデルを単一の統計モデルとして組み合わせることで，教師なしモデルと教師ありモデルを組み合わせた半教師ありの手法を構築することができる．

この枠組みは，いくつかの拡張が提案されている．

- Leglaive, S., et al.: A recurrent variational autoencoder for speech enhancement, *Proc. of IEEE ICASSP*（2020）[30]

  この論文では，VAE の代わりに再帰型 VAE を用いて音声の時間フレーム間の依存性を陽に扱っている．

- Pariente, M., et al.: A statistically principled and computationally efficient approach to speech enhancement using variational autoencoders, *Proc. of INTERSPEECH*（2019）[31]

  近似推論による高速化も提案されている．

- Sadeghi, M., et al.: Audio-visual speech enhancement using conditional variational auto-encoders, *IEEE/ACM TASLP*（2020）[32]

  ほかにも，視覚情報と併用することで，雑音環境に対し，より頑健な音声強調も提案されている．

- Sekiguchi, K., et al.: Semi-supervised multichannel speech enhancement with a deep speech prior, *IEEE/ACM TASLP*（2019）[33]

  また，マイクロホンアレーで観測した多チャネル混合音に対する空間的確率モデルを活用することで，より高い性能も達成できている．

- Kameoka, H., et al.: Supervised determined source separation with multichannel variational autoencoder, *Neural Computation*（2019）[34]

  VAE と NMF の組合せだけでなく，すべての音源を VAE で表現し，音声分離問題に適用した多チャネル VAE も提案されている．

4·3　償却変分推論に基づく教師なし音源定位・分離

教師なしで音源定位や音源分離を実現する手法として，マイクロホンアレーで観測される多チャネル混合音の空間的確率モデルを考える，ビームフォーマやブラインド音源分離が広く研究されている．このような多チャネル音響信号処理の確率モデルを，DNN の学習に活用する事例を紹介する．

・Drude, L., et al.: Unsupervised training of neural mask-based beamforming, *Proc. of Interspeech*（2019）[35]

この手法は，多チャネル音響信号を表現する確率的空間モデルの一つである**混合角中心複素ガウスモデル**（cACGMM）[36] の周辺尤度をコスト関数とし，音声強調する推論 DNN を教師なし学習する．この学習は cACGMM の事後分布を DNN が記憶する償却変分推論とみなすことができる．また，推論 DNN をエンコーダとし空間モデルをデコーダとする VAE としてみなすこともできる．この研究では，音の物理的な伝搬過程というデータの構造を記述した確率モデルを用いることで，観測混合音のみから音声強調を教師なし学習できることが示されている．従来の音声強調の教師あり学習では，クリーンな音声信号と雑音信号を数値的に混合したシミュレーションデータセットを用いることが多かったが，この枠組みにより正解信号のない実収録混合音からも直接学習できるようになった．また，一般に確率モデルの推論では，統計量が十分集まるまでデータをためる必要があったが，再帰型ニューラルネットワーク（RNN）などで償却変分推論すれば，オンラインリアルタイムの推論も期待できる．

・Bando, Y., et al.: Deep Bayesian unsupervised source separation based on a complex Gaussian mixture model, *Proc. of IEEE MLSP*（2019）[37]

上記と類似の枠組みを用いて，複数の音声信号を観測した混合音から個別の音声を抽出する音声分離を教師なし学習する手法が提案されている．一般に時間周波数領域の確率的空間モデルでは，周波数間の音源インデックスに曖昧性が生じるパーミュテーション問題が存在する．Drude らの手法では，音源信号のパワーが周波数間で相関をもつ特性に基づいた**パーミュテーションソルバ** [38] を DNN に導入していた．この研究では，マイクロホンアレーの配置情報を活用し，**音源の到来方向（DoA）に基づいてパーミュテーション問題を解決** [39] するアプローチを取っている．このような DoA に基づく音源分離の確率モデルは，音源数のわからない実環境でも頑健に動作する**ノンパラメトリックベイズの枠組み** [40] へ発展しており，このような知見も変分償却推論に導入できるようになると期待できる．

・Masuyama, Y., et al.: Self-supervised neural audio-visual sound source localization via probabilistic spatial modeling, *Accepted to IEEE/RSJ IROS*（2020）[41]

この研究では，視聴覚情報に基づく音環境理解システムへ，確率的空間モデルの償却変分推論を応用している．画像中の物体と音イベントの対応関係を**自己教師あり（教師なし）学習する枠組み** [42] が広く研究されている．このような枠組みは，音環境を画像情報も併用して多角的に理解するために不可欠である．多くの従来手法では，単チャネル音響信号と動画から学習しており，視覚的に識別が難しい音源では性能が劣化する原理的な問題があった．この研究では，多チャネル音響信号の空間的確率モデルを活用することで，音と画像の対応関係を空間情報から教師なし学習する．この確率モデルは，ベイズ複素混合ガウスモデルとして定式化されており，十分な個数の音源クラスを準備しておけば，実際の音源数に合わせて適宜音源クラスが縮退するようになっている．シミュレーションデータを用いて有効性を確認しているうえ，複数の展示物や来館者が常に存在する科学館での実収録データに対する適用事例も示されている．このようにベイズ学習をうまく深層学習と統合することで，未知情報の多い環境でも，頑健に教師なし学習できる枠組みを構築できる．

5. 教 科 書 な ど

・**須山敦志：ベイズ深層学習** [43]

ベイズ推論の基礎からベイズニューラルネットワークや深層生成モデルまで，ベイズ深層学習を日本語で学習するうえで最適な教科書である．

・Bishop, C. M.: *Pattern Recognition and Machine Learning* [44]（日本語訳版[45]）

・Murphy, K. P.: *Machine Learning: A Probabilistic Perspective* [46]

・**須山敦志 著，杉山 将 監修：ベイズ推論による機械学習入門** [47]

これらの教科書は，ベイズ学習をより深く理解するために大いに役立つ．

統計的音響信号処理に関しては以下の書籍が役立つ．

・**日本音響学会 編，浅野 太 著：音のアレー信号処理** [48]

アレー信号処理におけるバイブル的な本であり，クラシックな解釈に基づいているが，分野を広範に理解するうえで大きく役立つ．

・**戸上真人：Python で学ぶ音源分離** [49]

2020 年 8 月に出版されたばかりの書籍であるが，音響信号処理の基礎から応用までその実装例も含めて解説

されており，これからの必読書になると確信している．

6. 主 要 な 学 会

音響信号処理としてのコミュニティと，ロボット聴覚としてのコミュニティを紹介する．

6·1　国際会議

§ 1　音響信号処理系国際会議

- IEEE International Conference on Acoustics, Speech and Signal Processing（ICASSP）[50]
- INTERSPEECH [51]

§ 2　ロボティクス系国際会議

- IEEE/RSJ International Conference on Intelligent Robots and Systems（IROS）[52]
- IEEE International Conference on Robotics and Automation（ICRA）[53]

6·2　国内会議

§ 1　音響信号処理系国内会議

- 日本音響学会 研究発表会 [54]
- 電子情報処理学会・日本音響学会 音声研究会 [55]

§ 2　ロボティクス系国内会議

- 日本ロボット学会 学術講演会 [56]
- 人工知能学会 AI チャレンジ研究会 [57]

7. 主な国際技術評価会

- CHiME Challenge [58]
  音声分離・認識の評価会である．
- REVERB Challenge [59]
  残響下音声強調・音声認識の評価会である．
- DCASE [60]
  音イベント検出を中心に，音環境認識全般の評価会である．
- LOCATA Challenge [61]
  音源定位の評価会である．

8. お　わ　り　に

ここでは，深層学習とベイズ学習を組み合わせた深層ベイズ学習とその音環境理解への応用事例を紹介した．深層学習とベイズ学習はそれぞれ異なる利点と欠点をもっており，両者を組み合わせることで現状のさまざまな課題が解決できるようになると考えている．特に，データの性質を陽に記述することで，教師データがなくとも効率的な学習が実現できる点は，資源の少ないタスクや原理上得られないタスクにおいて有用な道具である．本稿をご覧いただき，深層ベイズ学習や音環境理解に興味をもっていただければ幸いである．

[1] https://chimechallenge.github.io/chime6/overview.html
[2] http://spandh.dcs.shef.ac.uk/chime_challenge/CHiME4/
[3] http://dcase.community/challenge2020/task-unsupervised-detection-of-anomalous-sounds-results
[4] https://ieeexplore.ieee.org/abstract/document/7471664/
[5] https://www.merl.com/demos/deep-clustering
[6] https://ieeexplore.ieee.org/abstract/document/8937253
[7] https://book.impress.co.jp/books/1119101154
[8] https://www.maruzen-publishing.co.jp/item/b294524.html
[9] https://www.mitpressjournals.org/doi/abs/10.1162/neco.1992.4.3.448

[10] https://www.jstage.jst.go.jp/article/sicejl/58/3/58_195/_article/-char/ja/
[11] http://papers.nips.cc/paper/5666-variational-dropout-and-the-local-reparameterization-trick
[12] https://www.kspub.co.jp/book/detail/1538320.html
[13] https://arxiv.org/abs/1312.6114
[14] http://papers.nips.cc/paper/5423-generative-adversarial-nets
[15] http://proceedings.mlr.press/v37/rezende15.html
[16] https://arxiv.org/abs/1312.6114
[17] https://arxiv.org/abs/1611.02648
[18] https://openreview.net/forum?id=rkE3y85ee
[19] https://arxiv.org/abs/1611.00712
[20] https://papers.nips.cc/paper/5653-a-recurrent-latent-variable-model-for-sequential-data.pdf
[21] http://www.probabilistic-robotics.org/
[22] https://openaccess.thecvf.com/content_cvpr_2018/papers/Bloesch_CodeSLAM_--_Learning_CVPR_2018_paper.pdf
[23] https://arxiv.org/abs/2005.07091
[24] https://ieeexplore.ieee.org/document/8462181
[25] http://dcase.community/challenge2020/task-unsupervised-detection-of-anomalous-sounds-results
[26] https://ieeexplore.ieee.org/document/8461530/
[27] https://ieeexplore.ieee.org/abstract/document/7471664
[28] https://ieeexplore.ieee.org/document/8673623
[29] https://openreview.net/forum?id=SJWQv3buWS
[30] https://ieeexplore.ieee.org/abstract/document/9053164
[31] https://www.isca-speech.org/archive/Interspeech_2019/pdfs/1398.pdf
[32] https://ieeexplore.ieee.org/abstract/document/9110765/
[33] https://ieeexplore.ieee.org/abstract/document/8861142
[34] https://www.mitpressjournals.org/doi/abs/10.1162/neco_a_01217
[35] https://www.isca-speech.org/archive/Interspeech_2019/pdfs/2549.pdf
[36] https://ieeexplore.ieee.org/abstract/document/7760429/
[37] https://ieeexplore.ieee.org/abstract/document/8918699
[38] https://ieeexplore.ieee.org/abstract/document/4253371/
[39] https://ieeexplore.ieee.org/abstract/document/6680684/
[40] https://ieeexplore.ieee.org/abstract/document/6680684/
[41] https://arxiv.org/abs/2007.13976
[42] https://openaccess.thecvf.com/content_ECCV_2018/html/Relja_Arandjelovic_Objects_that_Sound_ECCV_2018_paper.html
[43] https://www.kspub.co.jp/book/series/S043.html
[44] https://www.springer.com/jp/book/9780387310732
[45] https://www.maruzen-publishing.co.jp/item/b294524.html
[46] https://www.cs.ubc.ca/~murphyk/MLbook/
[47] https://www.kspub.co.jp/book/detail/1538320.html
[48] https://www.coronasha.co.jp/np/isbn/9784339011166/
[49] https://book.impress.co.jp/books/1119101154
[50] https://2021.ieeeicassp.org/
[51] http://www.interspeech2020.org/
[52] https://www.iros2020.org/
[53] http://www.icra2021.org/
[54] https://acoustics.jp/

[55] https://www.ieice.org/~sp/jpn/
[56] https://www.rsj.or.jp/
[57] http://www.osaka-kyoiku.ac.jp/~challeng/
[58] https://chimechallenge.github.io/chime6/
[59] https://reverb2014.dereverberation.com/
[60] http://dcase.community/
[61] https://www.locata.lms.tf.fau.de/