1A3-OS-27a-3

深層生成モデルを用いたマルチモーダル学習

Multimodal Learning by Deep Generative Model

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

Multimodal learning is a framework to discover joint representations from multiple different modalities (e.g., images, tags). Since deep neural networks have been successful in unsupervised feature learning, multimodal learning by deep neural network has been studied. In this work, we propose a novel multimodal learning model using the variational autoencoder which is deep generative model and has been proposed in recent years. We call this novel model the multimodal variational autoencoder (MVAE). We validate the MVAE on the MIR Flickr dataset, and confirm whether the MVAE can obtain better representations from multiple modalities.

1. はじめに

我々の住む実世界では，情報は様々なモーダル情報で表現されている．例えば，画像データはピクセル情報として表現される一方で，それらにタグ付けされている場合はタグデータ，すなわちテキストデータとしても表現される．人間はこれら複数のモーダル情報を 5 感から取り入れることで，単一のモーダル情報よりも確実な情報処理を行っている．機械学習においても，人間のように複数のモーダル情報を処理する方法としてマルチモーダル学習が研究されている．マルチモーダル学習では，複数のモーダル情報を入力として学習を行うが，それぞれのモーダル情報は表現が大きく異なるため，単純に全てを結合して入力とすることはできない．機械学習で適切な学習ができるようにするためには，それらの情報から共通する，より普遍的な特徴表現を獲得する必要がある．

深層ニューラルネットワークは，学習の過程において深い層で普遍的な特徴表現が獲得できることが知られている．そのため，これまで深層教師なし学習モデルである autoencoder（AE）や restricted Boltzmann machine（RBM）によるマルチモーダル学習 [Ngiam 11, Srivastava 12] が提案されている．これらは特徴抽出器や深層ニューラルネットワークの事前学習などに用いられるモデルで，2 つのネットワークの隠れ層，もしくは潜在変数を共有することで，異なるモーダル情報から共通かつより高次の特徴表現を獲得することを目的としている．

近年，深層ニューラルネットワークによる生成モデルとして，variational autoencoder（VAE）[Kingma 13] が提案された．VAE は生成モデルでありながら，モデルとして通常の多層ニューラルネットワークを用いることができ，学習が容易であること，さらに絵や顔写真などサイズの大きな画像などを生成できることから，RBM に代わる深層生成モデルとして注目されている．

本研究では，VAE を用いたマルチモーダル学習のモデルとしてマルチモーダル VAE（MVAE）を提案する．また MVAE を用いて MIR Flickr データセットによるマルチモーダル学習の実験を行い，MVAE で適切な表現が獲得できることを確認する．

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

本稿の構成は以下の通りである．2 章で深層ニューラルネットワークによるマルチモーダル学習と深層生成モデルに関する関連研究について述べる．そして 3 章で提案モデルである MVAE について説明する．4 章では MIR Flickr データセットを用いたマルチモーダル学習の実験をし，考察する．そして 5 章でまとめと今後の課題について述べる．

2. 関連研究

本章では，深層ニューラルネットワークによるマルチモーダル学習の先行研究と，本稿で用いる深層生成モデルである VAE について述べる．

2.1 深層ニューラルネットワークによるマルチモーダル学習

Ngiam らは，深層ニューラルネットワークが教師なし表現学習で幅広い成功を収めていることから，複数のモーダル情報による深層ニューラルネットワークの表現学習を提案した [Ngiam 11]．Ngiam らは，教師なし学習のモデルとして多層の autoencoder（AE）を用いている．各モーダル情報について AE を用意し，それらの最も深い隠れ層を共有することで，2 つのモーダル情報に共通する特徴表現を獲得できるようにした．様々な問題設定による比較検証によって，単一のモーダル情報よりも良い特徴表現を獲得できることを示している．

Srivastava らは，生成モデルである restricted Boltzmann machine（RBM）を多層にした deep Boltzmann machine（DBM）を用いたマルチモーダル学習を提案した [Srivastava 12]．生成モデルとは，データの生成過程を明示的に記述したモデルのことである．実験では Ngiam らの AE による手法と同じモデル構造ながら，DBM によってより良い特徴表現が得られることが示された．

2.2 Variational autoencoder

Variational autoencoder（VAE）は Kingma らによって提案された深層生成モデルである [Kingma 13]．

観測変数 $\mathbf{x}$ と潜在変数 $\mathbf{z}$ が与えられたとき，それらの生成過程を次のように考える．

$$\mathbf{z} \sim p(\mathbf{z}), \quad \mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{z}) \tag{1}$$

ただし θ はモデルパラメータである．

変分推論の枠組みでは，潜在変数の近似事後分布 $q_\phi(\mathbf{z}|\mathbf{x})$($\phi$ はモデルパラメータ) を考えて，次の周辺尤度の下界が最大になるように学習する．

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

本稿では，$q_\phi(\mathbf{z}|\mathbf{x})$ をエンコーダー，$p_\theta(\mathbf{x}|\mathbf{z})$ をデコーダーと呼ぶ．

下界 $\mathcal{L}(\mathbf{x})$ をパラメータ θ, ϕ について最適化する際，VAE では stochastic gradient variational Bayes（SGVB）[Kingma 13]，または stochastic backpropagation [Rezende 14] と呼ばれる手法を利用する．$z \sim q_\phi(z|x)$ がガウス分布 $\mathcal{N}(z|\mu, \sigma^2)$（ただし $\phi = \{\mu, \sigma^2\}$）のとき，$z = \mu + \sigma\epsilon$（ただし $\epsilon \sim \mathcal{N}(0,1)$）のように再パラメータ化（reparameterization）することができる．すると，期待値 $E_{q_\phi(z)}[f_\theta(z)]$ を $E_{\mathcal{N}(\epsilon;0,1)}[f_\theta(\mu+\sigma\epsilon)]$ と置き換えることができ，モンテカルロ法によって $\frac{1}{L}\sum_{l=1}^{L} f_\theta(\mu+\sigma\epsilon^{(l)})$（ただし $\epsilon^{(l)} \sim \mathcal{N}(0,1)$）として求めることができる．すなわち式 (3) の下界の推定量は次のように求まる．

$$\hat{\mathcal{L}}(\mathbf{x}) = -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) + \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\mathbf{x}|\mathbf{z}^{(l)})$$
$$\text{ただし } \mathbf{z}^{(l)} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}^{(l)}, \boldsymbol{\epsilon}^{(l)} \sim \mathcal{N}(0, \mathbf{I}) \quad (3)$$

式 (4) の第 1 項は正則化項，第 2 項は負の再構成誤差となっている．この下界を目的関数としてパラメーター ϕ, θ について最大化する．

VAE は SGVB を利用することで，他の推定法と比較して低バリアンスな推定量を求めることができる．その一方で，柔軟かつ計算が容易な近似事後分布を選ぶ必要もあり，これを解決するために normalizing flows [Rezende 15] や importance weighted VA [Burda 15] などが提案されている．

その他 VAE を拡張したモデルとして，conditional VAE [Kingma 14a]，deep Kalman filter [Krishnan 15]，variational Gaussian process [Dai 15] などが提案されている．

3. 提案モデル

次に本稿で提案するマルチモーダル VAE（MVAE）について説明する．

異なるモーダル情報のデータセットを，それぞれ $\{\mathbf{X}\}$ と $\{\mathbf{W}\}$ とし，潜在変数を $\mathbf{z}$ とする．また，それらの生成過程を

$$\mathbf{z} \sim p(\mathbf{z}), \quad \mathbf{x}, \mathbf{w} \sim p_\theta(\mathbf{x}, \mathbf{w}|\mathbf{z}) \quad (4)$$

とする．

$\mathbf{x}$ と $\mathbf{w}$ を観測変数とすると

$$p_\theta(\mathbf{x}, \mathbf{w}|\mathbf{z}) = p_{\theta_x}(\mathbf{x}|\mathbf{z}) p_{\theta_w}(\mathbf{w}|\mathbf{z}) \quad (5)$$

のように条件付き独立となる，ただし θ_x と θ_w は各分布のモデルパラメータである．

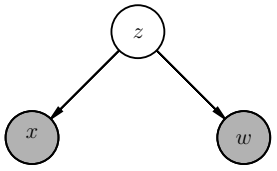

図 1: MVAE のグラフィカルモデル

この過程をグラフィカルモデルで表したものが図 1 である．グラフィカルモデルとは，生成モデルで設計した確率変数の依存関係を有効グラフで表現したもので，白丸が潜在変数，黒丸が観測変数を表す．

近似事後分布を $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{w})$ とすると，尤度 $p(\mathbf{x}, \mathbf{w})$ の変分下界は次のようになる．

$$\begin{aligned}\mathcal{L}(\mathbf{x},\mathbf{w}) &= \int q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w}) \log \frac{p_\theta(\mathbf{x},\mathbf{w},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})} d\mathbf{z} \\ &= -\int q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})}{p(\mathbf{z})} d\mathbf{z} \\ &\quad + \int q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})[\log p_{\theta_x}(\mathbf{x}|\mathbf{z}) + \log p_{\theta_w}(\mathbf{w}|\mathbf{z})] d\mathbf{z} \\ &= -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})||p(\mathbf{z})) \\ &\quad + E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})}[\log p_{\theta_x}(\mathbf{x}|\mathbf{z})] \\ &\quad + E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})}[\log p_{\theta_w}(\mathbf{w}|\mathbf{z})] \end{aligned} \quad (6)$$

SGVB によって，式 (6) の下界推定量は次のように求まる．

$$\begin{aligned}\hat{\mathcal{L}}(\mathbf{x},\mathbf{w}) &= -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})||p(\mathbf{z})) \\ &\quad + \frac{1}{L}\sum_{l=1}^{L} \log p_{\theta_x}(\mathbf{x}|\mathbf{z}^{(l)}) + \log p_{\theta_w}(\mathbf{w}|\mathbf{z}^{(l)}) \\ &\quad \text{ただし } \mathbf{z}^{(l)} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}^{(l)}, \boldsymbol{\epsilon}^{(l)} \sim \mathcal{N}(0, \mathbf{I}) \end{aligned} \quad (7)$$

式 (7) の第 1 項は正則化，第 2 項と第 3 項は各モーダル情報の再構成誤差となっている．

このグラフィカルモデルを深層ニューラルネットワークで表現したものが図 2 である．

エンコーダー $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{w})$ がガウス分布 $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\phi, \boldsymbol{\sigma}_\phi^2\mathbf{I})$ のときは，ネットワークの出力を次のようにして平均と分散を得る．

$$\begin{aligned}\mathbf{y}(\mathbf{x}) &= \text{MLP}_{\phi_x}(\mathbf{x}) \\ \mathbf{y}(\mathbf{w}) &= \text{MLP}_{\phi_w}(\mathbf{w}) \\ \boldsymbol{\mu}_\phi &= \text{Linear}(\mathbf{y}(\mathbf{x}), \mathbf{y}(\mathbf{w})) \\ \log \boldsymbol{\sigma}_\phi^2 &= \text{Tanh}(\text{Linear}(\mathbf{y}(\mathbf{x}), \mathbf{y}(\mathbf{w})) \end{aligned} \quad (8)$$

ただし，MLP_{ϕ_x} と MLP_{ϕ_w} はモーダル情報ごとの異なる多層ニューラルネットワークである．また Linear は線形層，Tanh は双曲線関数である．$\text{Linear}(a, b)$ はネットワークの入力として a と b の 2 つが与えられる構造を表す．

デコーダー $p_{\theta_x}(\mathbf{x}|\mathbf{z})$ と $p_{\theta_w}(\mathbf{w}|\mathbf{z})$ のモデルとして，それぞれ異なるネットワークを用意する．分布の形は，各モーダル情報のデータ形式に依存し，連続値の場合はガウス分布，$\{0, 1\}$ の 2 値のときはベルヌーイ分布とする．$p_{\theta_w}(\mathbf{w}|\mathbf{z})$ がベルヌー

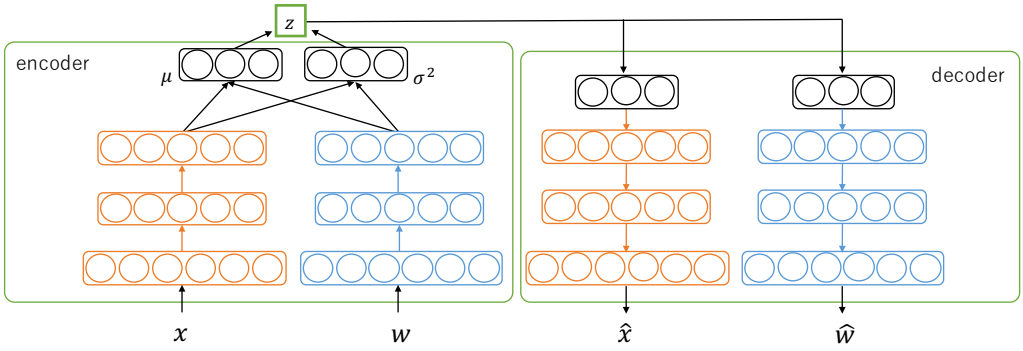

図 2: MVAE のネットワーク構造

イ分布 $\mathcal{B}(\mathbf{w}|\boldsymbol{\mu}_{\theta_w})$ のとき，ベルヌーイ分布のパラメータ $\boldsymbol{\mu}_{\theta_w}$ は次のように求める．

$$\begin{aligned} \mathbf{y}(\mathbf{z}) &= \mathrm{MLP}_{\theta_w}(\mathbf{z}) \\ \boldsymbol{\mu}_\theta &= \mathrm{Linear}(\mathbf{y}(\mathbf{z})) \end{aligned} \tag{9}$$

デコーダーがガウス分布のときは，式 (8) の Linear の入力が 1 つの場合と同様である．

VAE を利用したマルチドメインの研究として，Louizos らによって提案された variational fair autoencoder（VFAE）[Louizos 15] がある．この研究では，ドメイン変数と潜在変数が独立になるように Maximum Mean Discrepancy（MMD）による制約を設けている．しかし，VFAE は同一のモーダル情報を想定しており，異なる次元のデータを活用することはできない．一方提案手法では，複数のモーダル情報で学習することが可能である他，各モーダル情報の独立性を明示的に考慮することができる．また，Pandey らによって conditional multimodal autoencoder [Pandey 16] が提案されているが，VFAE と同様に次元の異なるようなマルチモーダル情報を利用するモデルではない．

4. 実験

4.1 データセット

MIR Flickr データセット [Huiskes 08] は，写真共有コミュニティサイト Flickr[*1] から抽出された 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（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][*2] 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 次元のベクトルで表現されている．本実験では，ラベルなしの 975,000 の画像及びタグは使用しない．25,000 のうち 15,000 を訓練事例集合，10,000 をテスト事例集合とした．

4.2 実験設定

本実験では，画像を $\mathbf{x}$，タグを $\mathbf{w}$ とし，マルチモーダル情報として扱う．各モーダル情報の定義域はそれぞれ $\mathcal{X} = \mathcal{R}^{3857}$，$\mathcal{W} = \{0,1\}^{2000}$ なので，$p_{\theta_x}(\mathbf{x}|\mathbf{z})$ をガウス分布，$p_{\theta_w}(\mathbf{w}|\mathbf{z})$ をベルヌーイ分布とした．

エンコーダー $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})$ のモデル構造及びユニット数は，$\mathbf{x}$ 部分が 3857-1024-1024，$\mathbf{w}$ 部分が 2000-1024-1024 で，それらを結合した層のユニット数は 2048 である．デコーダーのモデル構造は $p_{\theta_x}(\mathbf{x}|\mathbf{z})$ が 2048-1024-1024-3857，$p_{\theta_w}(\mathbf{w}|\mathbf{z})$ が 2048-1024-1024-2000 である．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．

適切な特徴表現が獲得されていることを確認するために，訓練事例集合でモデルを学習した後，テスト事例集合を与えたときの潜在変数 $\mathbf{z}$ をエンコーダー $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w})$ からサンプリングし，カテゴリ情報を出力 $\mathbf{y}$ とした写像 $f : \mathbf{z} \to \mathbf{y}$ を線形識別器であるロジスティック回帰モデルで学習する．訓練事例集合で学習したロジスティック回帰モデルをテスト事例集合で検証し，label ranking average precision（LRAP）で評価する．LRAP のスコアが高い時は，潜在変数の値が分類しやすい，即ちより適切な特徴表現が得られていることを意味する．

本実験では，提案手法である MVAE をマルチモーダル autoencoder（MAE）[Ngiam 11] と比較する．MAE のモデル構造，学習に用いるデータ，最適化手法は MVAE と同じとする．

評価実験は，それぞれ

設定 1 画像のみで訓練，テスト（AE と VAE を利用）

設定 2 画像とタグの両方で訓練，テスト（AE と VAE を利用）

設定 3 画像とタグの両方で訓練，テスト

設定 4 画像とタグで訓練，画像のみでテスト

の 3 つの設定で行う．設定 1 と設定 2 はマルチモーダルではない通常の AE 及び VAE を利用する．設定 2 は画像とタグの特徴ベクトルを結合して 1 つの入力とする．またデコーダーの分布はガウス分布とする．設定 4 は，テスト段階でエンコーダーの $\mathbf{w}$ の入力を全て 0 とすることで隠れ層の値を求める．

4.3 実験結果

表 1 と表 2 が実験結果である．評価値は，訓練事例とテスト事例の分け方を 5 回変更した平均値で求めている．

まず設定 1 の結果を見ると，VAE よりも AE の方が 6%以上高くなっていることが確認できる，この結果から，MIR Flickr データセットでは，VAE よりも AE の方が適切な特徴量を獲得できていることがわかる．通常，生成モデルの VAE の方が適切に特徴量を獲得できると思われるので，他のデータセットでも VAE の精度の方が低くなるのかどうか検証する必要がある．

設定 2 では画像とタグ情報の両方を与えているので，設定 1 の場合よりも情報が増え，さらに良い特徴を獲得できることが

*1 http://www.flickr.com
*2 http://www.cs.toronto.edu/~nitish/multimodal/index.html

表 1: 設定 1 と設定 2 の実験結果（LRAP で評価）

モデル	設定 1	設定 2
AE	**0.517**	**0.507**
VAE	0.455	0.457

表 2: 設定 3 と設定 4 の実験結果（LRAP で評価）

モデル	設定 2	設定 3
MAE	0.611	**0.517**
MVAE	**0.618**	0.455

期待される．しかし結果をみると，それほど精度が向上していないことがわかる．特に AE の場合は画像のみの場合より精度が落ちていることが確認できる．この結果から，異なるモーダル情報のデータは，単純に特徴ベクトルを連結するだけでは十分に活用できないことがわかる．

次に設定 3 の結果に着目する．表 2 から，設定 3 よりも精度が高くなっていることがわかる．このことから，MAE，MVAE ともにマルチモーダル学習によって，より適切な特徴表現を獲得できたことがわかる．また MAE と MVAE を比較すると，MVAE が MAE より僅かに高くなっていることが確認できる．設定 1 の結果で，VAE は AE より精度が低かったことを考えると，MVAE によるマルチモーダル学習によって大幅に精度が向上し，より適切な特徴量が獲得できるようになったといえる．

最後に，設定 4 の結果を確認する．表 1 と表 2 を比較すると，MAE，MVAE 共に設定 4 の結果が設定 1 の結果と変わらないことがわかる．つまりマルチモーダル学習のあと，単一のモーダル情報でテストをしても，単一のモーダル情報で学習した結果と変わらないという結果となった．実験では，テスト時にタグ情報 $\mathbf{w}$ の入力を 0，即ちエンコーダーのタグ情報 $\mathbf{w}$ 部分を取り除いていた．しかし，実際には画像情報部分とタグ情報部分のネットワークは隠れ層を共有しているため，単純に取り除くだけでは，共有されている隠れ層から訓練時のタグ情報が完全に失われてしまうと考えられる．よって，テスト時に単一のモーダル情報を利用するためには，$\mathbf{w}$ の入力を 0 ではなくランダムな値にする，などの工夫が必要である．

5. まとめ

本稿では，新たなマルチモーダル学習のモデルとしてマルチモーダル VAE（MVAE）を提案した．実験では，得られた潜在変数を線形識別器で識別することで，本手法がマルチモーダル情報を利用したときにより適切な表現を獲得できていることを確認した．また，マルチモーダル autoencoder と比較し，マルチモーダル情報を利用したときに MVAE の方が僅かに高い精度となることを確認した．しかし，テスト時に片方のモーダル情報のみを利用したとき，単一のモーダル情報で学習した場合と変わらない結果となることを確認した．今後は，マルチモーダル RBM などの他の既存手法との比較の他，様々なデータセットで検証を行う予定である．

参考文献

[Burda 15] Burda, Y., Grosse, R., and Salakhutdinov, R.: Importance weighted autoencoders, *arXiv preprint arXiv:1509.00519* (2015)

[Dai 15] Dai, Z., Damianou, A., González, J., and Lawrence, N.: Variational Auto-encoded Deep Gaussian Processes, *arXiv preprint arXiv:1511.06455* (2015)

[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)

[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)

[Krishnan 15] Krishnan, R. G., Shalit, U., and Sontag, D.: Deep Kalman Filters, *arXiv preprint arXiv:1511.05121* (2015)

[Louizos 15] Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R.: The Variational Fair Auto Encoder, *arXiv preprint arXiv:1511.00830* (2015)

[Ngiam 11] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y.: Multimodal deep learning, in *Proceedings of the 28th international conference on machine learning (ICML)*, pp. 689–696 (2011)

[Pandey 16] Pandey, G. and Dukkipati, A.: Variational methods for Conditional Multimodal Learning: Generating Human Faces from Attributes, *arXiv preprint arXiv:1603.01801* (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)

[Rezende 15] Rezende, D. J. and Mohamed, S.: Variational inference with normalizing flows, *arXiv preprint arXiv:1505.05770* (2015)

[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)