卒業論文

離散的な潜在空間の
変分オートエンコーダの学習

指導教員 唐堂 由其 准教授

2017年2月17日

金沢大学 理工学域

電子情報学類 情報システムコース

脳型情報処理研究室

新田 一稀

卒業論文概要 2016 年度（平成 2017 年 2 月 17 日年度）

離散的な潜在空間の
変分オートエンコーダの学習

深層ニューラルネットの最適化には, 誤差逆伝播法により出力層から微分を繰り返し勾配を求め, 勾配降下法でパラメータを更新している. 変分オートエンコーダ (Variational AutoEncoder; VAE) などの微分可能を満たさない確率ユニットを扱うモデルでは, 再パラメータ化トリックという変数変換を施すことで勾配伝播の要請を満たす. しかしこの手法は潜在空間が連続的である場合のみ適用され, 離散的な場合における手法は未だ確立されていない. 本研究ではターゲット伝播法による誤差信号の代わりに近似逆関数を用いた伝播法を用いて離散的な潜在空間における VAE の実装を提案する.

キーワード： ニューラルネットワーク, 教師なし学習, 生成モデル

目 次

第1章　はじめに

1.1　背景

深層ニューラルネットの各層のパラメータは, 最小化すべき誤差関数を設定し, 各層のパラメータの更新を確率的勾配降下法により最適化する. その際, 各パラメータの勾配を誤差逆伝播法 [1] を用いて出力層から一階微分を連鎖律により求める方法が一般的である.

深層ニューラルネットは非線形かつ階層的な生成モデルにも応用が出来, その一例として, 変分オートエンコーダ (Variational AutoEncoder; VAE) [2] がある. 変分オートエンコーダは変分ベイズ学習の枠組みにニューラルネットを適用し, 潜在変数が任意の確率分布による生成モデルを導く. この場合, 微分可能の要請を満たさない確率的なユニットを導入するため, 誤差逆伝播法によるパラメータ勾配の算出は途切れてしまう. その対策として再パラメータ化トリック [2][3] という変数変換の手法が利用されている. しかしこの手法はガウス分布のような連続的な確率分布のみに適用ができ, 離散的な確率分布の場合は未だ手法が確立していない.

本研究では, 離散的な確率分布の潜在空間をもつ変分オートエンコーダを対象とし, 確率的ユニットの箇所に対し勾配ではなく近似逆関数による伝播を行うターゲット伝播法 [4] の導入を試みる. 離散的な確率分布としてカテゴリカル分布を潜在空間とした実験を行い, ターゲット伝播法による変分オートエンコーダについて考察をする.

1.2　本論文の構成

本論文の構成を説明する.

第 2 章は基本的なニューラルネットワークの定式化を行い, 近年の深層ニューラルネットに広く利用されている最適化方法を解説した. また本研究で扱うニューラルネットワークモデルの大元として教師なし学習のオートエンコーダの解説も含めた. 第 3 章ではターゲット伝播法を解説した. 第 4 章では変分オートエンコーダについて解説し, ガウス分布を例とした連続的な潜在変数における問題とカテゴリカル分布を例とした離散的な潜在変数における問題について解説し, 後者では本研究が試みるターゲット伝播法の適用について述べた. 第 5 章はカテゴリカル分布の潜在空間をもつ変分オートエンコーダに対し本研究によるモデリングと論文 [5][6] のモデリングを実装し比較した. そして得られた考察と今後の展望を第 6 章に記した.

第2章　ニューラルネットワーク

ニューラルネットワークは非線形写像の積層化であり, 各パラメータを調整し, 近似的な解の収束を目指す機械学習アルゴリズムである. この章ではニューラルネットの定式化と基本的な学習方法について, そして基本的なオートエンコーダについて解説する.

2.1　定式化

ニューラルネットを定式化する. 入力を $\mathbf{x}$, 出力を $\mathbf{y}$, 層数を $i = 0, ..., M$ とし各層を $\mathbf{h}_i$ と置く.$\mathbf{h}_0 = \mathbf{x}$, $\mathbf{h}_M = \mathbf{y}$ である. 各層の活性化関数 s_i とパラメータ $\boldsymbol{\theta}_i = \{\mathbf{W}_i, \mathbf{b}_i\}$ より入力層除く各層は以下のように定式化できる.

$$\mathbf{h}_i = s_i(\mathbf{W}_i\mathbf{h}_{i-1} + \mathbf{b}_i) \tag{2.1}$$

$$= f_i(\mathbf{h}_{i-1}) \tag{2.2}$$

この時 i から j までのパラメータを $\boldsymbol{\theta}^{i,j}_{\mathbf{W}} = \{\mathbf{W}_k, k = i+1, ..., j\}$ とすると,$\mathbf{h}_j$ は $\mathbf{h}_i$ についての関数で表すことができる.

$$\mathbf{h}_j = \mathbf{h}_j(\mathbf{h}_i; \boldsymbol{\theta}^{i,j}_{\mathbf{W}}) \tag{2.3}$$

逐次的に入力層から出力層まで写像を繰り返す演算方式を順伝播ネットワークと呼ぶ.(図 2.1)

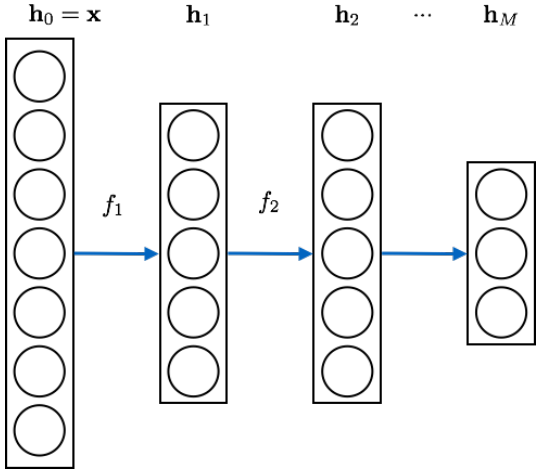

図 2.1: ニューラルネットの順伝播

訓練事例が $(\mathbf{x}, \mathbf{y})$ の時, 誤差関数 $L(\mathbf{h}_M(\mathbf{x}; \boldsymbol{\theta}^{0,M}_{\mathbf{W}}), \mathbf{y})$ は全体誤差となる.i 層に着目すると, 全体誤差は次のように記述できる.

$$L(\mathbf{h}_M(\mathbf{x}; \boldsymbol{\theta}^{0,M}_{\mathbf{W}}), \mathbf{y}) = L(\mathbf{h}_M(\mathbf{h}_i(\mathbf{x}; \boldsymbol{\theta}^{0,i}_{\mathbf{W}}); \boldsymbol{\theta}^{i,M}_{\mathbf{W}}), \mathbf{y}) \tag{2.4}$$

2.2 確率的勾配降下法

ニューラルネットは誤差関数を最小化することを目的に, 各層 i のパラメータ $\boldsymbol{\theta}_i = \{\mathbf{W}_i, \mathbf{b}_i\}$ を最適化する. 最適化で用いられる手法は, 誤差関数による各パラメータの一階微分を求め, 更新を繰り返す勾配降下法である. また, 多くの設計において, 入力データを一つずつ用いるのではなく, ある程度のまとまりとしてミニバッチ単位でデータの学習を行う確率的勾配降下法が広く一般的に用いられる.

ミニバッチのデータ数のことをバッチサイズと呼ぶ. バッチサイズを M とおくと, 求まる誤差関数 $L(\mathbf{x})$ は $L_1, ..., L_M$ である. これらを全て求めた上で, 学習の更新に用いる誤差関数を平均から導く.

$$L = \sum_{i}^{M} L_i \tag{2.5}$$

確率的勾配降下法は, このようにバッチサイズ毎に誤差関数を計算し最適化を行う.

確率的勾配降下法の基本式は以下となる.

$$\mathbf{W}_i \leftarrow \mathbf{W}_i - \alpha \frac{\partial L}{\partial \mathbf{W}_i} \tag{2.6}$$

$$\mathbf{b}_i \leftarrow \mathbf{b}_i - \alpha \frac{\partial L}{\partial \mathbf{b}_i} \tag{2.7}$$

ここで α は学習率であり, 勾配降下によるパラメータの更新量を調整するハイパーパラメータである.

収束性能の向上を目指し, 確率的勾配降下法は様々な派生が提案されている.

2.3 誤差逆伝播法

学習には誤差関数の各層のパラメータによる勾配を計算する必要がある. 誤差逆伝播法はこの勾配を効率よく求めるアルゴリズムである.

2.3.1 アルゴリズム

誤差逆伝播法は今もなおニューラルネットワークのパラメータ更新を行うために多く利用されている手法である. ここでは教師あり学習と仮定し, 出力 $\mathbf{h}_M$ と正解データ $\mathbf{y}$ の誤差関数 $\mathcal{L}(\mathbf{h}_M, \mathbf{y})$ を定義し, パラメータの最適化を行う問題を扱い誤差逆伝播法を紹介する. また, 勾配の最適化手法は前節に述べた確率的勾配降下法に準拠する.

$$\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}_i - \alpha \frac{\partial L}{\partial \mathbf{h}_i} \frac{\partial \mathbf{h}_i}{\partial \boldsymbol{\theta}_i} \tag{2.8}$$

出力層 M のパラメータ $\boldsymbol{\theta}_M = \{\mathbf{W}_M, \mathbf{b}_M\}$ を更新する場合は以下となる.

$$\boldsymbol{\theta}_M \leftarrow \boldsymbol{\theta}_M - \alpha \frac{\partial L}{\partial \mathbf{h}_M} \frac{\partial \mathbf{h}_M}{\partial \boldsymbol{\theta}_M} \tag{2.9}$$

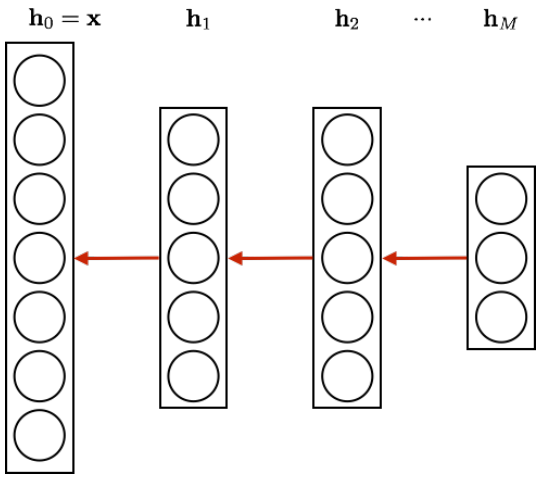

図 2.2: ニューラルネットの誤差逆伝播法

その下の層 M-1 の場合は, 微分の連鎖律を利用する.

$$\boldsymbol{\theta}_{M-1} \leftarrow \boldsymbol{\theta}_{M-1} - \alpha \frac{\partial L}{\partial \mathbf{h}_{M-1}} \frac{\partial \mathbf{h}_{M-1}}{\partial \boldsymbol{\theta}_{M-1}} \tag{2.10}$$

$$= \alpha \frac{\partial L}{\partial \mathbf{h}_M} \frac{\partial \mathbf{h}_M}{\partial \mathbf{h}_{M-1}} \frac{\partial \mathbf{h}_{M-1}}{\partial \boldsymbol{\theta}_{M-1}} \tag{2.11}$$

$$\tag{2.12}$$

このように連鎖律により上層のヤコビアンの線形な乗算で勾配を求めることができる. ここで $\delta_M = \frac{\partial L}{\partial \mathbf{h}_M}$, $\delta_{M-1} = \delta_M \frac{\partial \mathbf{h}_M}{\partial \mathbf{h}_{M-1}}$ と誤差信号 δ として表現することができ,i 層の更新式は

$$\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}_i - \alpha \delta_i \frac{\partial \mathbf{h}_i}{\partial \boldsymbol{\theta}_i} \tag{2.13}$$

となり $\delta_i = \delta_M \delta_{M-1} ... \delta_{i+1}$ である.

このように誤差信号が出力層から伝播していき, これが誤差逆伝播法と呼ばれる由縁である.(図 2.2)

計算量も順伝播と同等な線形演算であり, 多くのニューラルネットのフレームワークが誤差逆伝播法による計算を採用している.

2.3.2 誤差逆伝播法の性質

活性化関数の例の一つであるシグモイド関数は以下の式で表される.

$$s(\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{x})} \tag{2.14}$$

図 2.3 はシグモイド関数をプロットしたものである. 青線が $f(x)$ であり, 緑線がその導関数 $f'(x)$ である. 活性化関数にシグモイド関数などの値域が有界な非線形関数を使用した場

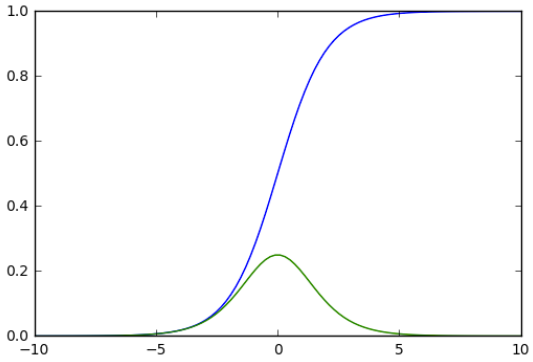

図 2.3: シグモイド関数のプロット

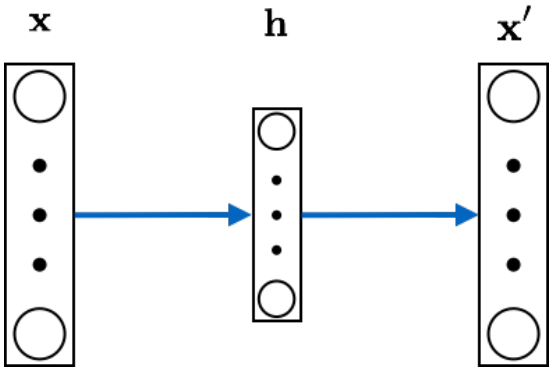

図 2.4: オートエンコーダ

合, 飽和領域においてその微分値は 0 に近い小数点以下の数値になる. そのため誤差逆伝播において誤差信号を乗算すればするほど小さな値となり, 出力層から離れた層のパラメータの更新が途絶えてしまう. この問題を勾配消失問題という.

勾配消失問題の対策として, 事前学習によるチューニングや, 値域が有界ではない活性化関数 ReLU[7] の利用, ニューラルネット特有の正規化手法である DropOut[8] や Batch Normalization[9] などが誤差逆伝播法を出力層から入力層まで満遍なく適用すべく提案されている.

また, 微分の連鎖律を用いるため, 出力層から入力層までの経路全てにおいて微分可能な必要がある. そのため, 入力において出力が一意に求まる決定的計算とは異なる, 出力が一意に定まらない確率的計算の振る舞いを持つユニットは勾配を求められないケースであり, 逆伝播がそこで途切れてしまう.

2.4 オートエンコーダ

オートエンコーダはニューラルネットによる教師なし学習モデルである. 図 2.4 が示すように入力層 $\mathbf{x}$ と同じ次元数をもつ出力層 $\mathbf{x}'$ を定義する. 中間層の次元数は通常入力層より少なく設定する. そしてオートエンコーダの出力が入力と同値になることを目的とする. つまり誤差関数は出力層を入力層へ近づけるような $\mathcal{L}oss(\mathbf{x}', \mathbf{x})$ を定義する.

入力層から中間層までの写像を特徴量の符号化の働きを持つためエンコーダと呼び, 中間層から出力層までを符号化されたデータの復号化の働きを持つためデコーダと呼ぶ. 活性化関数が恒等写像であれば順伝播は線形な演算であり主成分分析と同等と解釈できる. 非線形写像を行うオートエンコーダではパラメータの更新は確率的勾配降下法を用いる.

入力が連続値な場合は連続値である必要があるため出力層の活性関数を恒等写像に設定し, 誤差関数は最小二乗誤差を定義する. 入力がバイナリ値の場合はシグモイド関数を設定し, 交差エントロピーを誤差関数に定義することが一般的である.

オートエンコーダは中間層を多層にすることが出来, 通常のニューラルネットと同様に深層化により復元性能が向上する.

第3章　ターゲット伝播法

この章では本研究の目的のために用いるターゲット伝播法について解説する. 各層の理想的なターゲットという概念を導入し, 誤差信号ではなく近似逆を用いており, 誤差逆伝播法とは異なるアプローチで学習を行うアルゴリズムである.

3.1　ターゲット

訓練事例はデータ $\mathbf{x}$ と正解ラベル $\mathbf{y}(\mathbf{x}, \mathbf{y})$ の教師あり学習とする.

各層 $\mathbf{h}_i$ に対し全体の誤差が小さくなるようなターゲット $\hat{\mathbf{h}}_i$ を定義する. (式 3.1)

$$L(\mathbf{h}_M(\hat{\mathbf{h}}_i; \boldsymbol{\theta}_{\mathbf{W}}^{i,M}), \mathbf{y}) < L(\mathbf{h}_M(\mathbf{h}_i(\mathbf{x}; \boldsymbol{\theta}_{\mathbf{W}}^{0,i}); \boldsymbol{\theta}_{\mathbf{W}}^{i,M}), \mathbf{y}) \tag{3.1}$$

これにより各層の目標はターゲットに近づくことであり, 各層において誤差関数 $\mathcal{L}oss(\mathbf{h}_i, \hat{\mathbf{h}}_i)$ が定義される. 確率的勾配降下法により $\mathbf{h}_i$ のパラメータ $\boldsymbol{\theta}_i$ を更新するが, ターゲットは定数として扱える.

$$\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}_i - \alpha_i \frac{\mathcal{L}oss(\mathbf{h}_i, \hat{\mathbf{h}}_i)}{\partial \mathbf{h}_i} \frac{\partial \mathbf{h}_i}{\partial \boldsymbol{\theta}_i} \tag{3.2}$$

ここで α_i は層ごとの学習率である.

出力層 $\mathbf{h}_M$ のターゲットは, 教師あり学習では正解ラベル $\mathbf{y}$ であり, オートエンコーダのような入力を復元することを目的とする場合は入力 $\mathbf{x}$ がターゲットとなる.

中間層のターゲットは, 次に解説する近似逆で求める.

3.2　近似逆

$\mathbf{h}_{i-1}$ から $\mathbf{h}_i$ への写像が関数 f_i により $\mathbf{h}_i = f_i(\mathbf{h}_{i-1})$ であることから, 逆関数 f_i^{-1} があれば $\mathbf{h}_{i-1} = f_i^{-1}(f_i(\mathbf{h}_{i-1}))$ である. しかし, 完全な逆関数を求めることは難しいため, f_i^{-1} を近似する非線形関数 g_i を導入する.

$$g_i(\mathbf{h}_i) = \bar{s}_i(\mathbf{V}_i \mathbf{h}_i + \mathbf{c}_i) \tag{3.3}$$

ここで $\bar{s}_i, \mathbf{V}_i, \mathbf{c}_i$ はそれぞれ活性関数, 重み, バイアスである. これにより $\mathbf{h}_{i-1} = g_i(f_i(\mathbf{h}_{i-1}))$ を満たすように g_i を更新する. 更新に用いる誤差関数は最小二乗誤差となる.

$$||g_i(f_i(\mathbf{h}_{i-1})) - \mathbf{h}_{i-1}||_2^2 \tag{3.4}$$

図 3.1 のように近似逆 g_i により上の層のターゲットを逆向きに写像し下の層のターゲットが求まる.

$$\hat{\mathbf{h}}_{i-1} = g_i(\hat{\mathbf{h}}_i) \tag{3.5}$$

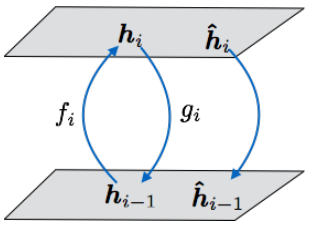

図 3.1: ターゲット

しかし式 3.5 では最適化が不十分なため, ベクトルのズレを調整する線形補正を行う.(式 3.6)

$$\hat{\mathbf{h}}_{i-1} - \mathbf{h}_{i-1} = g_i(\hat{\mathbf{h}}_i) - g_i(\mathbf{h}_i) \tag{3.6}$$

この線形補正によるターゲットの導出を利用した伝播法は差分的ターゲット伝播法 (difference target propagation) と呼ばれている. 図 3.2 はベクトル補正の直感的な図示である.

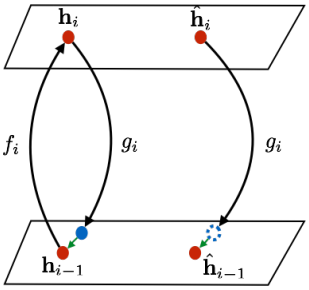

図 3.2: ターゲット

3.3 オートエンコーダとしてのターゲット伝播法

図 3.3 のように入力層 $\mathbf{x}$, 中間層 $\mathbf{y}$, 出力層 $\mathbf{x}'$ のオートエンコーダを定義する. $\mathbf{y} = f(\mathbf{x})$ かつ $\mathbf{x}' = g(\mathbf{y}) = g(f(\mathbf{x}))$ である. 出力層のターゲット $\hat{\mathbf{x}}'$ は入力 $\mathbf{x}$ 自身である. f の近似逆

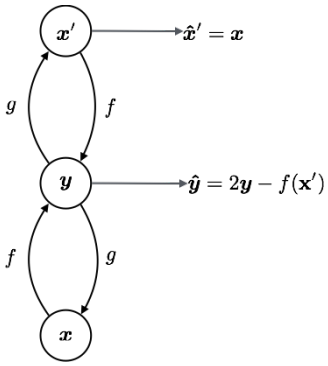

図 3.3: オートエンコーダ

には g を,g の近似逆には f を用いることができる. そのため差分的ターゲット伝播法による中間層のターゲットは次式が導出される.

$$\hat{\mathbf{y}} = 2\mathbf{y} - f(\mathbf{x}') \tag{3.7}$$

出力層と中間層に対する 2 つの誤差関数を定義する. 多層化したオートエンコーダの場合では, 中間層のみにターゲットを導入することで, 多層的なエンコーダ f, デコーダ g としてこれらを異なる誤差関数で最適化することが可能となる. つまり誤差逆伝播法のような出力層から入力層までのワンライナーな伝播ではなく, 中間層のターゲットに含まれる出力層 $\mathbf{x}'$ のみがエンコーダの外部の情報として利用されている.

第4章　変分オートエンコーダ

4.1　変分推論

図 4.1 のような潜在変数 $\mathbf{z}$ を含んだグラフィカルモデルで表現される生成課程を考える. 観測されたデータ $\mathbf{x}$ に対して潜在変数 $\mathbf{z}$ を定義し, 潜在変数が表す抽象的な表現をもとに具体的に複雑化されたデータが観測されることをモデリングしている.

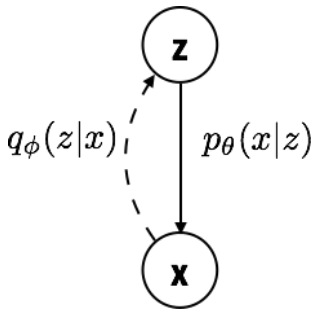

図 4.1: VAE のグラフィカルモデル

仮定とされる事前分布 $p(\mathbf{z})$ を観測されたデータ $\mathbf{x}$ を用いて更新するベイズ的なアプローチを用いる.

$$p(\mathbf{z}|\mathbf{x}) \propto p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) \tag{4.1}$$

真の事後分布 $p(\mathbf{z}|\mathbf{x})$ は推定が困難なため近似分布 $q_\phi(\mathbf{z}|\mathbf{x})$ を考え, この分布を真の事後分布と近づけることを考える.

$$\mathrm{KL}\big[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})\big] \tag{4.2}$$

しかし式 4.2 には推定困難な $p(\mathbf{z}|\mathbf{x})$ が含まれるため直接最適化ができない. そこで対数尤度

が式 4.2 を項に持つ分解が可能なことを利用し, 変分下限を導出する.

$$\mathrm{KL}\big[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})\big] = \int q_\phi(\mathbf{z}|\mathbf{x}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z}|\mathbf{x})} d\mathbf{z} \tag{4.3}$$

$$= \int q_\phi(\mathbf{z}|\mathbf{x}) \times \big(\log q_\phi(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{z}|\mathbf{x})\big) d\mathbf{z} \tag{4.4}$$

$$= \int q_\phi(\mathbf{z}|\mathbf{x}) \times \big(\log q_\phi(\mathbf{z}|\mathbf{x}) - \log \frac{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}\big) d\mathbf{z} \tag{4.5}$$

$$= \int q_\phi(\mathbf{z}|\mathbf{x}) \times \big(\log q_\phi(\mathbf{z}|\mathbf{x}) - \log p_\theta(\mathbf{x}|\mathbf{z}) - \log p(\mathbf{z}) + \log p(\mathbf{x})\big) d\mathbf{z} \tag{4.6}$$

$$= \log p(\mathbf{x}) + \int q_\phi(\mathbf{z}|\mathbf{x}) \times \big(\log q_\phi(\mathbf{z}|\mathbf{x}) - \log p_\theta(\mathbf{x}|\mathbf{z}) - \log p(\mathbf{z})\big) d\mathbf{z} \tag{4.7}$$

$$= \log p(\mathbf{x}) + \int q_\phi(\mathbf{z}|\mathbf{x}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})} d\mathbf{z} - \int q_\phi(\mathbf{z}|\mathbf{x}) \log p_\theta(\mathbf{x}|\mathbf{z}) d\mathbf{z} \tag{4.8}$$

$$= \log p(\mathbf{x}) + \mathrm{KL}\big[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\big] - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] \tag{4.9}$$

$$\log p(\mathbf{x}) - \mathrm{KL}\big[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})\big] = -\mathrm{KL}\big[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\big] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] \tag{4.10}$$

ここで KL ダイバージェンスは非負であるため,

$$\log p(\mathbf{x}) \geq -\mathrm{KL}\big[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\big] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] = \mathcal{L}(\theta, \phi; \mathbf{x}) \tag{4.11}$$

が常に成り立つ. 右辺の最大化の代わりに左辺である変分下限 $\mathcal{L}(\theta, \phi; \mathbf{x})$ の最大化を行う. これにより真の事後分布 $p(\mathbf{z}|\mathbf{x})$ の直接な最適化を行わない問題に帰着できる. また生成モデルにおいて観測データ $\mathbf{x}$ の生起確率が最大であることは尤もらしいという考えにより誤差関数は負の対数尤度を設定することからも負の変分下限を誤差関数に定義可能だと解釈できる.

4.2 ガウス分布の変分オートエンコーダ

例として, 事前分布 $p(\mathbf{z})$ がガウス分布である変分オートエンコーダをモデリングする. $q_\phi(\mathbf{z}|\mathbf{x})$ がパラメータ ϕ のエンコーダに該当し, $p_\theta(\mathbf{x}|\mathbf{z})$ がパラメータ θ のデコーダに該当する.

4.2.1 定義

多変量ガウス分布は平均ベクトル $\boldsymbol{\mu}$, 共分散行列 $\boldsymbol{\Sigma}$ を母数とする確率分布である. 事前分布 $p(\mathbf{z})$ として平均ベクトル $\mathbf{0}$, 共分散行列 $\mathbf{I}$ を仮定する.

式 4.11 の各項について, 第一項の KL ダイバージェンスは潜在表現がガウス分布の場合解析可能である. 第二項は解析的に計算できないためモンテカルロ推定を行う.

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] \sim \frac{1}{L} \sum_{l=1}^{L} \log p_\theta(\mathbf{x}|\mathbf{z}^l) \tag{4.12}$$

$\mathbf{z}^l \sim q_\phi(\mathbf{z}|\mathbf{x})$ は近似事後確率からのサンプリングであり, θ に関する推定量は総和の中の式を微分して勾配を得られる.

確立的勾配降下法で用いるミニバッチ数 M が十分大きければ $L = 1$ でよいとされている.[2] 改めて変分下限は次式に求まる.

$$\mathcal{L}(\theta, \phi; \mathbf{x}) \simeq -\mathrm{KL}\big[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\big] + \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\mathbf{x}|\mathbf{z}^l) \tag{4.13}$$

第一項の KL 項は単純な事前分布への依存を強める正則化項としての役割をもつ. 第二項は観測データ $\mathbf{x}$ がエンコーダ $q_\phi(\mathbf{z}|\mathbf{x})$ を通り, $\mathbf{z}^l$ をサンプリングし, デコーダ $p_\theta(\mathbf{x}|\mathbf{z})$ を通り元の観測データ $\mathbf{x}$ と比較される. これは通常のオートエンコーダ同様の復元精度を表し, 変分下限による誤差関数の定義は正則化オートエンコーダと解釈ができる.

4.2.2 再パラメータ化トリック

ガウス分布の母数は平均 $\boldsymbol{\mu}$ と分散 $\boldsymbol{\Sigma}$ であるからエンコーダ部のニューラルネットの出力層は二つの $\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\Sigma}(\mathbf{x})$ となる. 図 4.2 はモデルの外観図である. 左の入力層 $\mathbf{x}$ から右に向かって順に順伝播していく. 潜在空間の層は平均 $\boldsymbol{\mu}$, 分散 $\boldsymbol{\Sigma}$ のガウス分布による潜在変数 $\mathbf{z}$ のサンプリングが行われる. 逆伝播の際, 定義された誤差関数を基に勾配伝播が行われる. しかしサンプリング箇所は微分可能の要請を満たさないため, 誤差勾配の伝播が打ち止めになる.

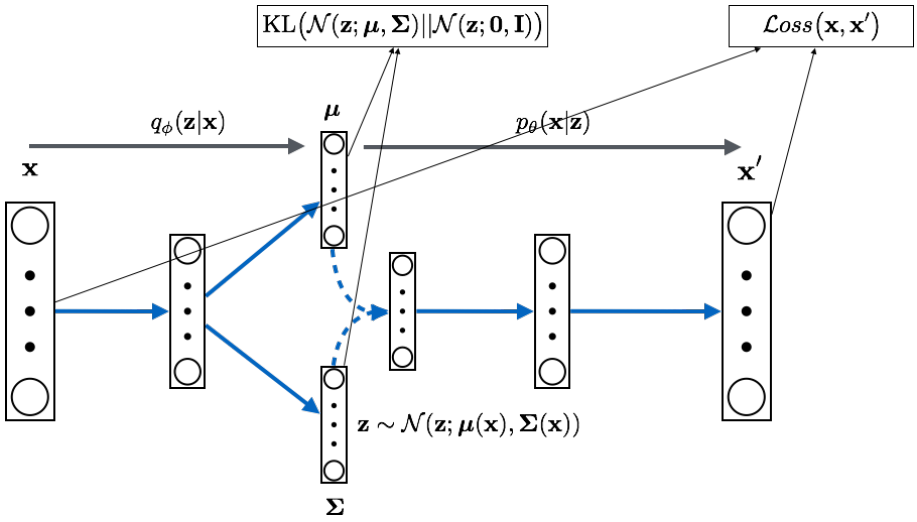

図 4.2: ガウス分布 VAE の外観図

サンプリング箇所を決定的な関数に変換する手法を再パラメータ化トリック [2][3] という.

$$\mathbf{z} \sim p(\mathbf{z}|\mathbf{x}) \rightarrow \mathbf{z} = g(\boldsymbol{\epsilon}, \mathbf{x}) \tag{4.14}$$

$$where\ \boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}) \tag{4.15}$$

ガウス分布の場合 $p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \mathbf{I})$ である. ガウス分布の変数変換は以下のように導出される.

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{\frac{1}{2}} \odot \boldsymbol{\epsilon} \tag{4.16}$$

再パラメータ化トリックを行うことで左の出力層から右の入力層まで誤差信号が伝播されるようになる.(図 4.3)

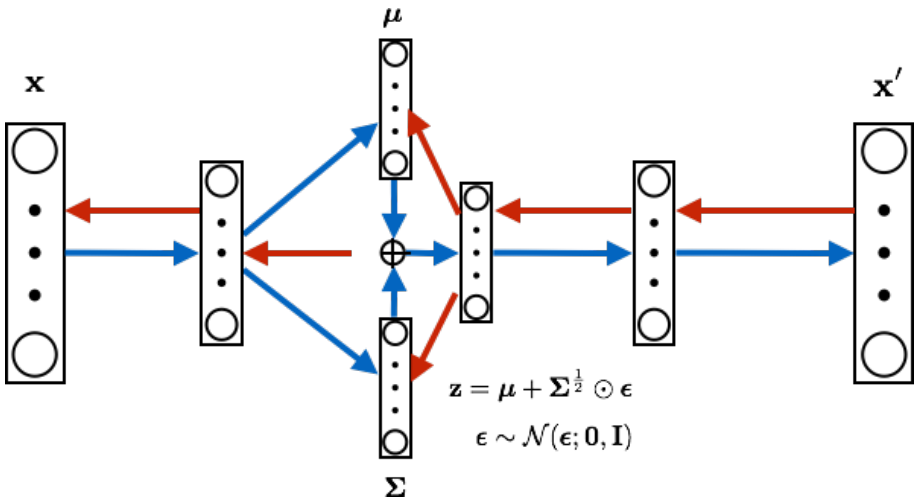

図 4.3: ガウス分布の再パラメータ化トリック適用後の VAE 外観図

4.3 カテゴリカル分布の変分オートエンコーダ

再パラメータ化トリックは潜在空間が連続的な確率分布に限定される. 世の中には潜在表現として離散値を扱うべき事象があるため, 離散的な確率分布も導入したい.

本研究では次節で解説するターゲット伝播法を導入し, 微分可能でなくてはいけないという誤差逆伝播法の制約条件を根本的に解決することを試みる.

その他にも離散的な確率分布に対しての研究は [5][6] がある. この 2 つの論文はお互いの提案内容が同等であり, カテゴリカル分布の潜在空間が対象となっている. 離散変数と連続変数のギャップを緩和するアプローチから, 離散空間のサンプリングを微分可能な要請を満たすように, つまり誤差逆伝播法の制約条件に適合するようなモデリングの提案を行っている.

本論文も同様に潜在空間をカテゴリカル分布としてモデリングをしこれらの論文と比較し考察する.

4.3.1 定義

潜在変数 $\mathbf{y}$ は one-hot ベクトル (1-of-K 表現) と定義する. 例えばクラス数 K が 6 の場合, 数値 2 は $\mathbf{y} = [0, 1, 0, 0, 0, 0]$ というように one-hot ベクトルで表される. このような離散値を生成する確率分布にカテゴリカル分布がある.

エンコーダの出力は確率ベクトル $\boldsymbol{\pi}$ であるため, 出力時の活性化関数がソフトマックス関数である関数 $\boldsymbol{\pi}_\phi(\mathbf{x})$ となる.

ガウス分布時と同様に, カテゴリカル分布においても $\mathbf{y} \sim q(bfy|\mathbf{x})$ は関数的なサンプリングを行うためにガンベルマックストリック [10] によるサンプリングを行う. ガンベル分布は一様分布から生成可能である極値分布である.(式 4.17)

$$\mathcal{G}_*(0,1) = -\log\left(-\log\left(u\right)\right) \; where \; u \sim \mathcal{U}(0,1) \tag{4.17}$$

乱数 u をクラス数 K 個分生成し, エンコーダの出力である確率ベクトルの対数と足しあわせ, 最大値となる要素を 1 とし, その他の要素を 0 とする one-hot ベクトルをサンプリングする.

$$\text{one-hot}\left(\text{argmax}\left(\log \boldsymbol{\pi} + G_k(0,1)\right)\right) \tag{4.18}$$

式 4.18 は, ガンベル分布によるノイズを除けば何度 argmax 操作を行っても同じクラスが出力されるが, ノイズを加えることで分布の形状を変え, 違うクラスも出力されるように擬似的なサンプリングが行える関数である.

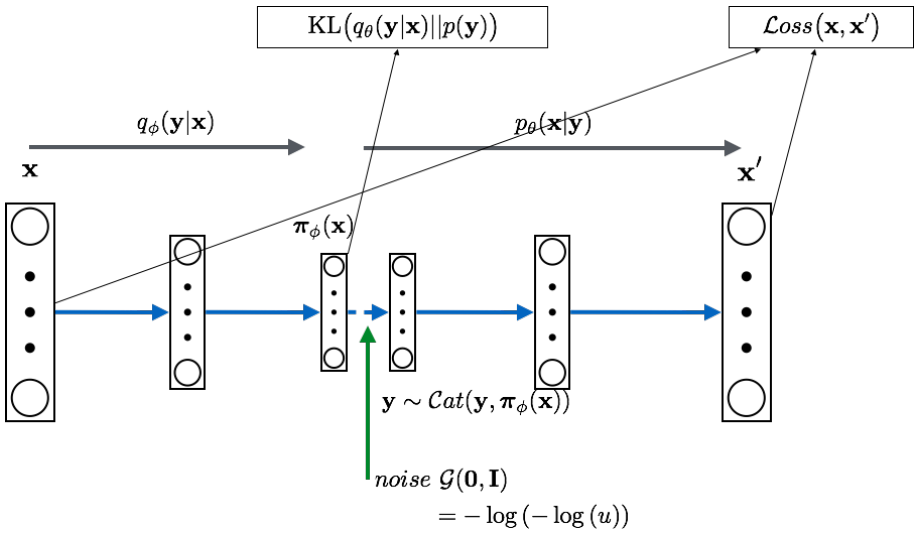

図 4.4: カテゴリカル分布の VAE 外観図

one-hot ベクトルでは勾配は導けない. よってこのサンプリング箇所で誤差逆伝播は途切れてしまう.

4.3.2 ターゲット伝播法の適用

本研究では逆伝播が不可能な中間層から入力層までパラメータの更新を行うために, この変分オートエンコーダのモデルに第 3 章で述べた (差分的) ターゲット伝播法を適用する.

モデルの外観図は図 4.5 となる. 誤差関数は前述のように出力層と中間層にそれぞれ定義される. 変分オートエンコーダの変分下限の二項より復元誤差項は出力層に定義され, KL 正則化項は中間層にターゲット誤差と共に定義される.

4.3.3 ガンベルソフトマックス

比較するモデル [5][6] について解説する. このモデルではサンプリングにおいて,argmax を取り出し one-hot ベクトル化することは行わずに温度 τ というハイパーパラメータの概念を導入し確率ベクトルを直接ソフトマックス関数に入力するガンベルソフトマックスという手法を提案している.

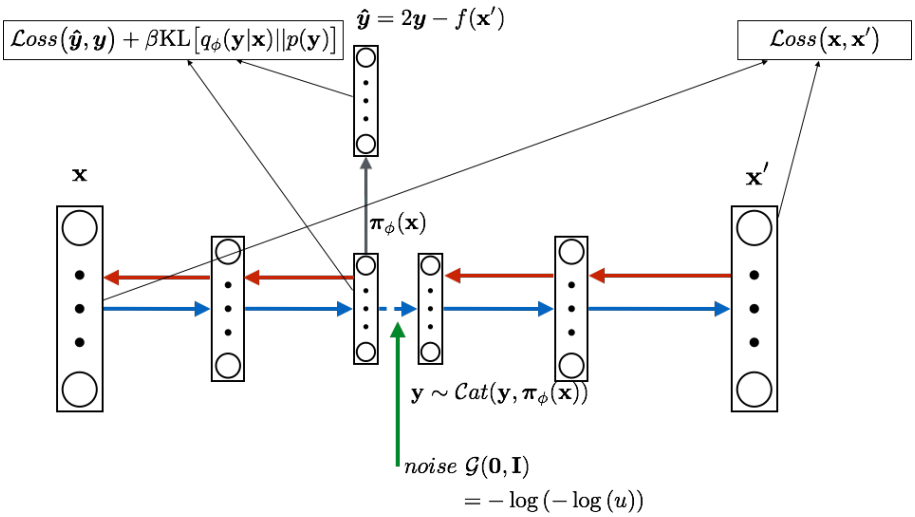

図 4.5: ターゲット伝播法を適用したカテゴリカル分布 VAE 外観図

$$\mathrm{softmax}\left(\left(\log \boldsymbol{\pi} + \mathbf{G}_k(0,1)\right)/\tau\right) \tag{4.19}$$

式 4.19 は温度 τ が高いほど出力される分布形状は一様分布に近づき, 低いほど one-hot な形状の分布となる. 図 4.6 は $\tau = \{0.1, 1, 5, 100\}$ において各 100 回サンプリングして得た出力の平均をプロットしたものである.

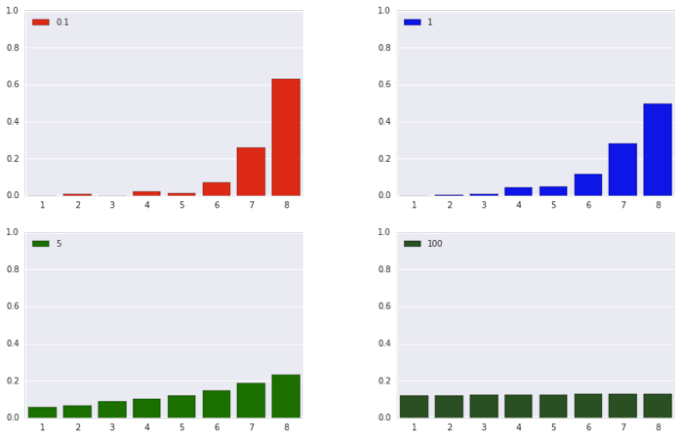

図 4.6: 左上:$\tau = 0.1$, 右上:$\tau = 1$, 左下:$\tau = 5$, 右下:$\tau = 100$

このモデルは潜在空間の離散変数と潜在空間以外の連続変数の連続性を緩和し, このサンプリングによる出力は微分可能であるため誤差逆伝播法が出力層から入力層まで適用することができる.

なお, 温度 τ の最良な値はモデル構築時に探索する必要があるパラメータであり, 低いほど one-hot なサンプリングが行えるが勾配が高分散となり, 高いほどなだらかな一様分布が出力されるが勾配は低分散というトレードオフの関係がある.

第5章　実験

第 4.3 節で述べた本研究提案モデリングとガンベルソフトマックスによるモデリングを手書き数字画像の表現学習をタスクとして実装する.

5.1　問題設定

学習するタスクは手書き数字データセット MNIST[11] とする. 訓練データ 50000 個, 検証データ 10000 個の画像を用いて教師なし学習を行う. 検証データは学習に利用せずに誤差のみを出力させるために用いており, 過学習を定量的に検証できるものである.

5.1.1　前処理

MNIST の手書き数字画像は縦 28*横 28 ピクセル (784 次元) であり要素は実数値となる. 前処理を施し値を 0 1 に正規化した後に 0.5 未満を 0, 0.5 以上を 1 に変換し, 要素がバイナリ値な手書き数字画像 (白黒) とする.

5.1.2　ネットワーク

各モデリングのネットワークのパラメータ $\boldsymbol{\theta}$ の次元数を統一する. 入力層から 784-500-300(30*10)-500-784 とする. ここで潜在空間の次元数はカテゴリカル分布の数 30, クラス数 $\mathbf{K}$ は 10 とする. エンコーダ, デコーダともに中間層の活性化関数はソフトプラス関数を用いる. 入力データがバイナリ値の要請から出力層の活性化関数はシグモイド関数にする. よってデコーダの出力 $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{z})$ はベルヌーイ分布によるものと解釈ができ, 変分下限の復元誤差項はベルヌーイ分布の対数尤度となる.

5.1.3　共通設定

学習エポック数は 50 とし, 学習率は 0.0003 とする.

5.2　実装

定義されている変分下限は以下である.

$$\mathcal{L}(\theta, \phi; \mathbf{x}) \simeq -\mathrm{KL}\big[q_\phi(\mathbf{y}|\mathbf{x})||p(\mathbf{y})\big] + \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\mathbf{x}|\mathbf{y}^l) \tag{5.1}$$

バッチサイズを 100 とし, サンプリング試行回数は $L = 1$ とする. KL 正則化項は $p(\mathbf{y}) = \frac{1}{\mathbf{K}}$ から次式を導く.

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = -\sum_{k}^{K} \pi_k \log \mathbf{K}\pi_k + \log p_\theta(\mathbf{x}|\mathbf{y}) \tag{5.2}$$

5.2.1 ターゲット伝播法

本研究が試みるターゲット伝播法の適用を実装する.

出力誤差は負のベルヌーイ分布の対数尤度となり中間層誤差はターゲット誤差項と KL 正則化項を定義する. ターゲット誤差は二乗誤差とする.

KL 正則化項にはスケール係数 β を導入する. 学習はじめの事前分布に強く依存させてしまわないよう 0 からスケールさせる.

$$\mathcal{L}oss(\mathbf{x}, \mathbf{x}') = -\log p_\theta(\mathbf{x}|\mathbf{y}) \tag{5.3}$$

$$\mathcal{L}oss(\mathbf{x}, \hat{\mathbf{y}}) = ||\boldsymbol{\pi}(\mathbf{x}) - \hat{\boldsymbol{\pi}}||_2^2 + \beta \sum_{k}^{K} \pi_k \log \mathbf{K}\pi_k \tag{5.4}$$

5.2.2 ガンベルソフトマックス

ハイパーパラメータである温度 τ を 0.5 と 0.1 の 2 通りで検証する.

τ と学習率は 1 エポックごとに減衰していくアニーリング方式を採用し, 元論文 [6] の設計に従った.

5.3 結果

各モデルの復元誤差の対数尤度を表 5.1 にまとめた.

本研究が提案したターゲット伝播法適用によるモデルはガンベルソフトマックスによるモデルに比べ復元誤差の対数尤度は高くみられるが, 完全に one-hot な離散値の潜在空間を導いたため 0 か 1 のバイナリ値で構成される潜在表現を獲得できている.

表 5.1: 復元誤差の対数尤度

モデル	訓練データ対数尤度	検証データ対数尤度
ターゲット伝播法	-70.102	-72.545
ガンベルソフトマックス ($\tau = 0.5$)	-63.876	-64.401
ガンベルソフトマックス ($\tau = 0.1$)	-68.883	-68.623

図 5.1 はターゲット伝播法モデルの確率ベクトル $\boldsymbol{\pi}$ のヒートマップである. 行が分布数 30, 列がクラス数 10 を表し, 値域が [0,1] の数値が大きいほど色が濃い. 図 5.2 はターゲット伝播法モデルに 5 つのテストデータ画像を入力し出力された再生成画像である.

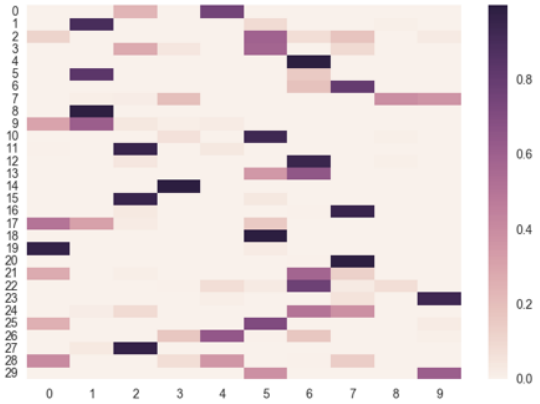

図 5.1: ターゲット伝播法モデルの確率ベクトル $\boldsymbol{\pi}$ ヒートマップ

図 5.2: 上が入力画像, 下が出力画像

第6章　おわりに

6.1　まとめ

離散的な確率分布の潜在空間を仮定した変分オートエンコーダの学習方法としてターゲット伝播法を適用したモデリングを提案した.

例としてカテゴリカル分布で実験を行い, 完全な one-hot の離散値表現を学習することができた. 離散表現は完全にバイナリ値なため, 連続値に比べパラメータの保存が低容量でありデータの潜在表現獲得において検出の高速化が期待できる.

本手法はカテゴリカル分布以外の離散的な確率分布でも適用が可能であり, 潜在表現が離散的構造を持つと仮定されるタスクにおいて数多くの適用が期待できる.

また, カテゴリカル分布を連続緩和し誤差逆伝播法の微分可能の要請を満たしたガンベルソフトマックスとの比較も行った. 表現力については連続性を備えたガンベルソフトマックスに譲るが, 先述のように完全なバイナリ値の潜在表現の獲得が本研究手法の特徴である.

本研究では確率ベクトルを扱うにあたって差分的ターゲット伝播法の線形補正について考察を得た. 差分的ターゲット伝播法は近似逆と完全逆をユークリッド幾何的な”ズレ”と解釈し, 線形補正を行う. しかし, 今回のような確率ベクトル空間を扱う場合, 幾何的な確率分布の性質を考慮した”ズレ”の補正を行うことが望ましい.

ターゲット伝播法において確率ベクトルのターゲットを導入する際は, 確率分布の性質に合わせた幾何を定義し行う差分的な補正の発見を今後の課題として見つけることができた.

参考文献

[1] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." Cognitive modeling 5.3 (1988): 1.

[2] Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. Proceedings of the 2nd International Conference on Learning Representations, 2014.

[3] Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. Proceedings of The 31st International Conference on Machine Learning. 2014.

[4] Lee, Dong-Hyun, et al. "Difference target propagation." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer International Publishing, 2015.

[5] Maddison, Chris J., Andriy Mnih, and Yee Whye Teh. "The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables." In Proceedings of the Second International Conference on Learning Representations (ICLR 2017).

[6] Jang, Eric, Shixiang Gu, and Ben Poole. "Categorical Reparameterization with Gumbel-Softmax." In Proceedings of the Second International Conference on Learning Representations (ICLR 2017).

[7] Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. "Deep Sparse Rectifier Neural Networks." Aistats. Vol. 15. No. 106. 2011.

[8] G. E. Hinton, N. Srivastava, A. Krizhevsky,I. Sutskever and R. Salakhutdinov, " Improving neural networks by preventing co-adaptation of feature detectors," CoRR, abs/1207.0580, 2012.

[9] Ioffe, Sergey, and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." Proceedings of the 32nd International Conference on Machine Learning (ICML-15). 2015.

[10] Maddison, Chris J., Daniel Tarlow, and Tom Minka. "A* sampling." Advances in Neural Information Processing Systems. 2014.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." Proceedings of the IEEE, 86(11):2278-2324, November 1998.