広島県立総合技術研究所東部工業技術センター研究報告　No.36（2023）
技術ノート

深層学習による異常検知手法の簡単な比較（第 2 報）

2 次元畳み込み変分オートエンコーダ

廣川　勝久

Anomaly Detections using Deep Learning (Ⅱ)

2D-Convolutional Variational Autoencoder

HIROKAWA Katsuhisa

全結合型の深層学習モデルを異常検知に用いる場合，時間的ずれや位置ずれの事前補正を行うか，ずれのあるデータを含めた学習を必要とする。位置ずれに対応した学習には，多くの場合，畳み込み深層学習が使われる。本報告では，深層学習の生成モデルに，局所結合を持つ畳み込み変分オートエンコーダの実装を行った。実験により畳み込み変分オートエンコーダのクラスタリングと典型的なデータへの修復・復元能力の可能性の評価を行い，その結果について述べる。

キーワード：畳み込みニューラルネットワーク，変分オートエンコーダ，VAE，Convolutional Neural Network

1. 緒　　言

パーセプトンから始まったニューラルネットワークの研究は，現在の畳み込みニューラルネットワークの原型であるネオコグニトロン [1]により画像処理へと応用されるようになった。畳み込みニューラルネットワーク [2]（**Convolutional Neural Network: CNN**）は，生物の視覚情報処理と同様に，位置によらず，入力画像のどの位置に認識したい物体があっても検出が可能である。これは，ニューラルネットワークの各層を局所受容野と呼ばれる大きさを限定した重みにより結合した効果である。畳み込みニューラルネットワークの初段の層では，入力画像の局所的な特徴が抽出され，後段に進むにつれ，局所的な特徴を組み合わせた特徴の学習が行われる。

一方，オートエンコーダ（**Auto Encoder: AE**）などの生成モデル [3]では，入力の多次元データの次元数を減らしながら，各データに共通する典型的な特徴部分を低次元の潜在変数へと学習する。生物の視覚神経も生成モデルとよく似ており，網膜の受光細胞と比較して，目から脳に情報を伝達する視神経の経路数は著しく少ない。これは，網膜に入力された情報全てを脳に送るのでは無く，低次元化した情報を送っていると考えられる。

前報では，生成モデルの一種であるオートエンコーダの高い汎化能力による異常検知について報告した [4]。また，生物の視覚情報処理を模した学習型画像認識システムの汎化能力についても研究を行った [2]。本稿では，生物の視覚情報処理に類似した畳み込みニューラルネットワークをオートエンコーダに実装することにより，より生物に近いニューラルネットワークを構築し，このニューラルネットワークの教師なしクラスタリング能力の可能性を確認した。また，異常検知に必要とする入力情報の修正・復元能力についても報告する。

2. 2 次元畳み込による異常検知

2.1 2 次元生成モデルによる異常検知

異常信号の検知には，温度センサーや振動センサーなどからの 1 次元データやイメージセンサーなどからの 2 次元データ，さらにレンジファイダーなどの 3 次元データなど様々なデータが用いられる。これらのデータは実用上，時間的なずれや位置のずれなどが生じる。全結合型の深層学習をこれらのデータの異常検知に用いる場合，事前のずれ補正，またはずれデータを含めた学習が必要となるため，実用的であるとは言えない。畳み込みニューラルネットワークは，人間の視覚を模したモデルであり，位置が変化した入力データに対しても学習したデータと同じ値を出力することが可能である。

一方，オートエンコーダなどの深層学習による生成モデルでは，低次元の潜在変数から成る特徴空間の座標から，入力データに対応した典型的なデータを復元することによって，2 つのデータ比較により良否の判定を行う。これらのモデルを組み合わせた異常検知の可能性を明らかにする。

2.2 畳み込みニューラルネットワーク

畳み込みニューラルネットワークは，現在，画像処理分野へ広く応用が試みられている。これは従来からニューラルネットワークが全結合型の重みで構成されるのに対して，受容野と呼ばれる局所結合型の重みにより，画像等の入力位置ずれに対応できるからである。

2023.7.12　加工技術研究部　受理

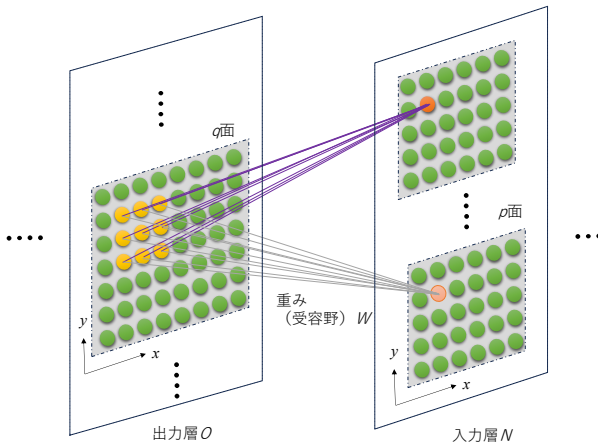


図 1 畳み込みニューラルネットワーク

図 1 に畳み込みニューラルネットワークの構造を示す。各層は複数面（クラスタ）から構成されている。この場合，出力層 O の q 面から入力層 N の p 面への結合は，

$$N_p(x,y) = \sum_q \sum_j \sum_i W_{p.q}(i,j) O_q(x+i, y+j) + b_p \quad (1)$$

と表せる。$W_{p.q}(i,j)$は q 面から p 面へ結合重みを表し，q 面内では同じ重みを用いる。同じ重みによる局所結合を持つ構造により，画像の位置ずれに対しても同じ出力や認識能力を与えることができる。b_pは重みのバイアス値である。

2.3 畳み込み変分オートエンコーダ

変分オートエンコーダに畳み込みニューラルネットワークを組み合わせたものが，畳み込み変分オートエンコーダである。オートエンコーダでは，入力データの低次元化を行うエンコーダ部とデータの生成を行うデコーダ部分から構成されている。この，低次元化やデータの生成に対して畳み込み学習を行う。

オートエンコーダの汎化能力を高めた深層学習モデルが変分オートエンコーダ（**V**ariational **A**uto **E**ncoder: **VAE**）である。変分オートエンコーダでは，潜在変数 z の計算に確率分布を用いて汎化能力を高めている。エンコーダーから平均μと標準偏差σが出力され，その変数の正規分布確率からサンプリングを行い，潜在変数 z を計算する。この方法により，学習時には潜在変数にノイズが加えられた状態となり潜在変数のわずかな変化に対してもデコーダから同じような出力が得られ，汎化能力が結果的に高められる。この学習には，正規分布からサンプリングに対して，微分計算がとぎれないよう誤差を逆伝搬させるため，reparametrization trick と呼ばれる手法が用いられる。次式のように標準正規分布からのサンプリングした値εを使って，潜在変数 z を求め，誤差逆伝搬を可能としている。

$$z = \mu + \varepsilon \cdot \sigma \quad (2)$$

変分オートエンコーダの学習は，入力データと出力データを一致させるため両データの差分を損失関数とする。加えて，潜在変数部分の平均μと標準偏差σの確率分布についても拘束条件が付加されている。

２つの離散確率分布の差異を評価する手法として KL ダイバージェンスがある。離散確率分布P, Qが与えられた場合，KL ダイバージェンスD_{KL}は

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \quad (3)$$

により求められる。変分オートエンコーダでは，エンコーダからの平均μ，標準偏差σの確率分布 P と平均 0，標準偏差 1 の正規分布 Q の KL ダイバージェンスが最小値となるよう学習を行う。すなわち，2 つの確率分布関数を

$$P = N(\mu, \sigma) \quad (4)$$
$$Q = N(0,1)$$
$$\because N(\ \ ) \text{ 正規分布}$$

とすると，損失関数

$$loss = D_{KL}[N(\mu,\sigma) \parallel N(0,1)] \quad (5)$$

を最小化するよう学習を行う。

3. 畳み込みオートエンコーダの実装

3.1 畳み込みオートエンコーダの構成

28×28 画素の 2 次元データから構成されている手書き数字データベース MNIST : Modified National Institute of Standards and Technology[5] を畳み込みオートエンコーダの学習に用いた。**図 2** に示すように，28×28 画素の入力画像は，エンコーダ部分により，まず 14×14 画素の 4 枚の画像に変換される。更に 7×7 画素の 16 枚へと変換され，最後には 1 次元の潜在変数へと変換される。特徴空間に写像された潜在変数は，デコーダ部分により徐々に特徴値に対応した画像の復元が行われ，28×28 画素の典型的な画像が出力される。

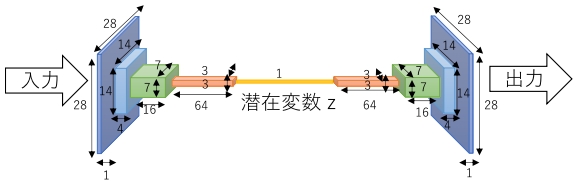


図 2 畳み込みオートエンコーダ

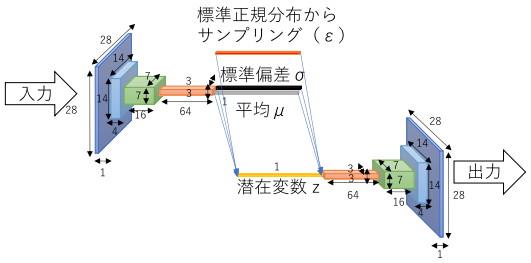


図 3 畳み込み変分オートエンコーダ

3.2 畳み込みオートエンコーダの実装

図 2 の畳み込みオートエンコーダ実装のために各層を次のように定義した。2 層と 3 層では正規化を行った。

```
self.conv1 = nn.Conv2d(1, 4, 4, 2, 1, bias=False)  # 14x14

self.conv2 = nn.Conv2d(4, 16, 4, 2, 1, bias=False)  # 7X7
self.bn2 = nn.BatchNorm2d(16)

self.conv3 = nn.Conv2d(16, 64, 3, 2, 0, bias=False)  # 3x3
self.bn3 = nn.BatchNorm2d(64)

self.conv4 = nn.Conv2d(64, zsize, 3, 2, 0, bias=False)  # 1x1

self.relu = nn.LeakyReLU()
```

1 層から 2 層と，2 層から 3 層への重みのサイズを 4×4 とし，3 層から 4 層と 4 層から潜在変数への重みサイズは 3×3 とした。全層ストライドは 2 とした。

```
def forward(self, x):

    x = self.conv1(x)
    x = self.relu(x)

    x = self.conv2(x)
    x = self.relu(self.bn2(x))

    x = self.conv3(x)
    x = self.relu(self.bn3(x))

    x = self.conv4(x)

    return x
```

各層からの出力に対する非線形関数には LeakyReLU を用いた。デコーダについては，エンコーダとほぼ同じ構造とした。

```
self.conv4 = nn.ConvTranspose2d(zsize, 64, 3, 1, bias=False)  # 3x3
self.bn4 = nn.BatchNorm2d(64)

self.conv3 = nn.ConvTranspose2d(64, 16, 3, 2, bias=False)  # 7X7
self.bn3 = nn.BatchNorm2d(16)

self.conv2 = nn.ConvTranspose2d(16, 4, 4, 2, 1, bias=False)  # 14x14
self.bn2 = nn.BatchNorm2d(4)

self.conv1 = nn.ConvTranspose2d(4, 1, 4, 2, 1, bias=False)  # 28x28

self.relu = nn.LeakyReLU()
```

これらのプログラムを組み合わせ，畳み込みオートエンコーダのプログラムを構成し，評価のため潜在変数 z を 1 次元の戻り値としている。

```
class AutoEncoder(nn.Module):
    def __init__(self, features):
        super().__init__()
        self.encoder = Encoder(features)
        self.decoder = Decoder(features)
        self.flat = nn.Flatten()

    def forward(self, x):
        y = self.encoder(x)
        x = self.decoder(y)
        z = self.flat(y)
        return x, z
```

3.3 畳み込み変分オートエンコーダ

全結合型オートエンコーダと同様に畳み込みオートエンコーダも汎化能力を高めた畳み込み変分オートエンコーダとすることが可能である。**図 3** はその畳み込み変分オートエンコーダの構成を示す。4 層のエンコーダから，まず，1 次元の平均 μ と標準偏差 σ を出力する。1 次元の平均 μ と標準偏差 σ が定まれば，全結合型変分オートエンコーダと全く同様に潜在変数を計算すれば良い。標準偏差が正値となるよう softplus 関数を出力段に用いた。また，PyTorch の Normal 関数により出力値の平均と標準偏差から正規分布の計算を行った。

```
    x = self.conv3(x)
    x = self.relu(self.bn3(x))

    mean = self.flat(self.conv_mean(x))
    div = self.flat(F.softplus(self.conv_div(x)))
    q_z = Normal(mean, div)

    return q_z, mean, div
```

正規分布からのサンプリングが行われた場合においても，rsample 関数を使うことで出力段の損失逆伝搬を可能とした。

```
    def forward(self, x):
        q_z,_,_ = self.encoder(x)
        x = self.decoder(q_z.rsample())
        return x, q_z
```

サンプリングされた潜在変数は，畳み込みオートエンコーダのデコーダに入力され，潜在変数に対応した画像が復元される。

4. 畳み込みオートエンコーダの評価

4.1 畳み込みオートエンコーダによるクラスタリング

まず，各層全てを 2 次元の畳み込みによる結合で構成した畳み込みオートエンコーダのクラスタリング能力について確認を行った。手書き文字の学習後，未学習の 1

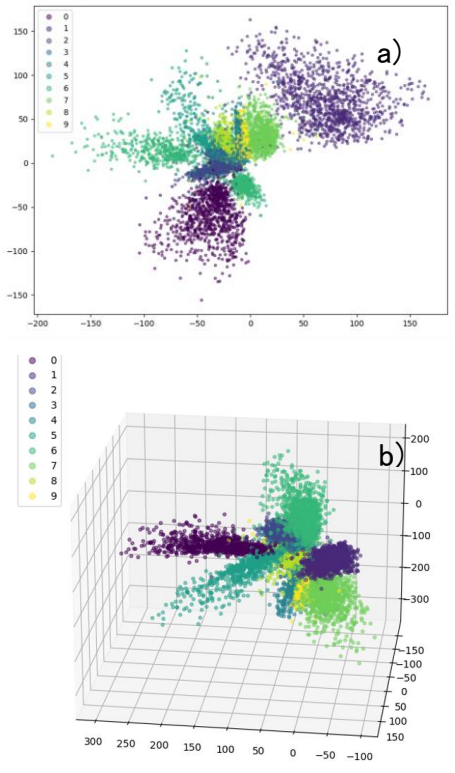


図 4 オートエンコーダのクラスタリング
a) 潜在変数 2 次元，b) 潜在変数 3 次元

万文字の入力に対してクラスタリングされた状態をグラフ化したものを**図 4** に示す。**図 4 a)** は 2 次元空間への潜在変数への写像を表す。全て畳み込みによる局所結合の構造であっても，数字ごとにクラスタリングが行われている。また，**図 4 b)** は 3 次元空間への写像である。2 次元と同様に，この場合もクラスタリングが行われている。但し，2 次元と 3 次元のどちらの写像についても，数字の種類によりクラスタサイズは異なり，各クラスタ間にはデータの存在しない大きな隙間がある。これは，全結合のオートエンコーダと同様に，未学習の領域が広範囲に存在し，汎化能力が低いことが想定される。

4.2 畳み込み変分オートエンコーダによるクラスタリング

同様に畳み込み変分オートエンコーダのクラスタリング能力についても評価を行った。**図 5 a)，b)** は潜在変数が 2 次元と 3 次元の場合を示す。2 次元，3 次元グラフともに**図 4** と比較して，各文字に対するクラスタは狭い範囲に圧縮されている。また各クラスタの大きさの差が小さい。これらの傾向は，以前報告した全結合型変分オートエンコーダと同じである。また，畳み込みオートエンコーダに見られたクラスタ間の隙間は，畳み込み変分オートエンコーダでは，隙間を埋めるようにクラスタが存在し，汎化能力を高めていることが見て取れる。

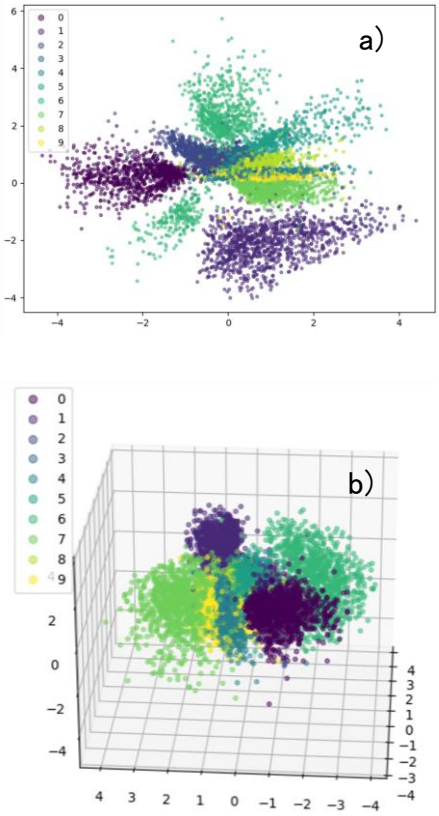


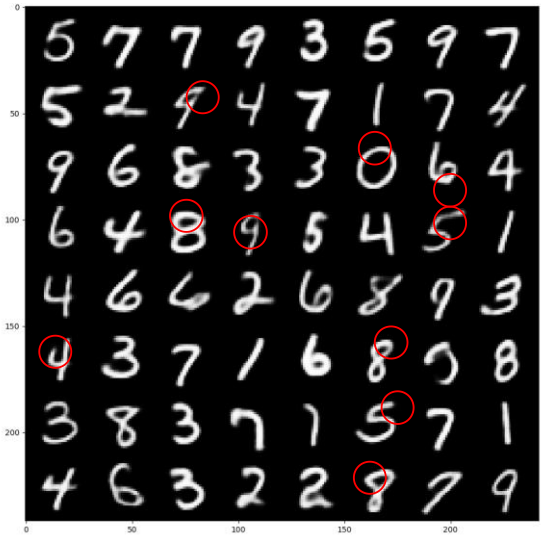
図 5 変分オートエンコーダのクラスタリング
a) 潜在変数 2 次元，b) 潜在変数 3 次元

4.3 畳み込み変分オートエンコーダによる異常検知能力

畳み込み変分オートエンコーダの異常検知能力を評価した。64 次元の潜在変数から成る畳み込み変分オートエンコーダの学習を行い，未学習画像がどのように修正されるか評価した。**図 6 a)** は用いた未学習の入力画像，**図 6 b)** は入力画像に対応した畳み込み変分オートエンコーダからの生成画像である。両画像を比較すると，生成画像の赤丸部分が，畳み込み変分オートエンコーダにより修正されている。変分オートエンコーダでは，学習による潜在変数の次元削減の結果，典型的な画像の生成を行う。従って，入力画像にその画像のみの特異な部分があれば，生成データからは削除され，入力画像に欠損部分があれば，画像が補われ典型的な画像に近づくよう復元される。**図 6** の両画像を比較すると，数字の誤認識は無く，文字のかすれや，小さなはみ出しなど局所的な修正が行われている。このことから，入出力画像の比較により，局所部分に発生した異常を検知することが可能となる。

図 7 には学習画像と未学習画像に対して，畳み込み変分オートエンコーダが生成した画像との誤差と画像数をグラフに表した。青の棒グラフは学習画像，オレンジ色の棒グラフは未学習画像による誤差を示す。学習画像，未学習画像に限らず，生成画像には一定程度の誤差が生じる。但し，**図 7 b)** に示すように誤差が大きい領域では，

a)



b)

図 6 変分オートエンコーダの修正能力
a) 入力画像，b) 出力画像

未学習画像の割合が増加する傾向にある。従って，このことからも，学習画像と大きく異なるような画像が入力された場合には，一定程度以上の誤差を異常として検知出来る可能性が高い。

4.4 畳み込み変分オートエンコーダの誤認識

畳み込み変分オートエンコーダに 64 次元の潜在変数を用いた場合，生成画像には局所的な修正能力があることを示したが，次に潜在変数が極端に少なく低次元化された場合の結果を示す。この場合，学習画像の特徴量は各数字に共通するより典型的なもののみに制限される。**図 8** に潜在変数を 3 次元空間に制限した場合の入出力画像を示す。64 次元潜在変数の局所的な生成画像修正と比較すると，3 次元の潜在変数からの生成画像は文字全体が大きく修正され，乱れが少なく，どれも似通った文字となっている。また，全体的に画像はぼやける。生成画像

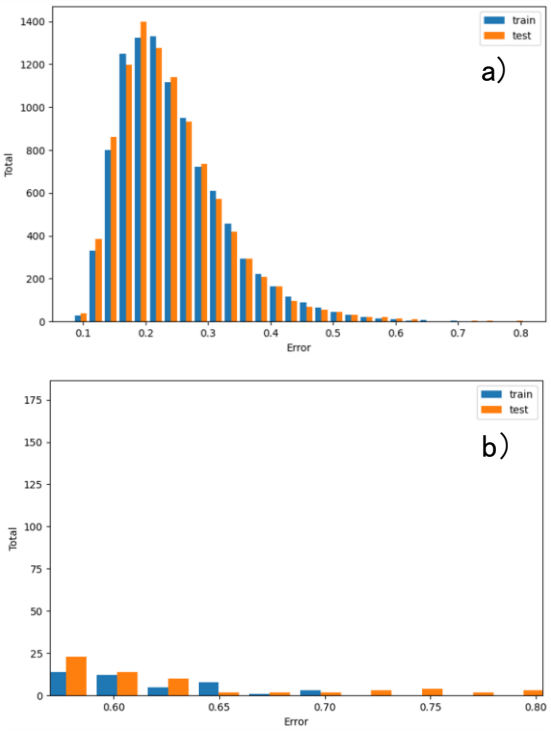


図 7 学習画像と未学習画像の生成誤差
a) 学習画像と未学習画像の誤差比較
b) 誤差の大きい領域の拡大

の赤丸部分は入力画像より見やすい画像に修正されているが，一方で青四角の部分では文字の誤認識が発生した。特徴値の低次元化では，生成画像に大幅な修正が行われ典型的な画像に近づくものの，誤認識の割合も増加する。

4. 結　　言

生物の視覚情報処理に類似した畳み込みニューラルネットワークをオートエンコーダに実装し，より生物の構造に近い畳み込み変分オートエンコーダの汎化能力や異常部分の修正能力について報告した。畳み込み変分オートエンコーダにおいても，全結合型のニューラルネットワークと同様に，教師なしクラスタリングが行われることを確認した。また，情報の低次元化が行われた潜在変数から典型的な画像を復元することも確認した。これは，畳み込み変分オートエンコーダが異常検知に必要とする入力情報の修正・復元能力を備えており，正常信号のみの学習から，異常信号を検出する能力を持つことを意味する。

但し，潜在変数の極端な低次元化を行った場合，文字の誤認識が発生した。これは，極端な低次元化により各文字のクラスタに重なりが生じていると思われる。

a)

b)

図 8　変分オートエンコーダの修正能力
a) 入力画像，b) 出力画像

文　　献

1）　福島　邦彦：神経回路と情報処理, 朝倉書店(1989).

2）　K Hirokawa, K Itoh, Y Ichioka: Invariant pattern recognition by neural networks combined with optical wavelet preprocessor. Opt. Rev. 7(4), 284-293 (2000).

3）　毛利　拓也　他: GAN ディープラーニング実装ハンドブック, 秀和システム（2021）.

4）　廣川　勝久：広島県立総合技術研究所東部工業技術センター研究報告，深層学習による異常検知手法の簡単な比較（第 1 報）35（2022）.

5）　THE MNIST DATABASE of handwritten digits：http://yann.lecun.com/exdb/mnist/