H-024

U-Net による顔画像の回転変換での潜在変数の特性

Feature of Latent Variables in Rotational Transformation of Face Images by U-Net

岡本 紗季[1)]　神野 健哉[1)]
Saki Okamoto　Kenya Jin'no

概要

与えられた画像から教師なし学習により特徴量を抽出するモデルに Auto Encoder(AE) がある．AE で与えられた画像から特徴量が精度高く抽出することができるようになると，学習に用いていない画像からも特徴を抽出でき，これを利用することで画像修復などができる．AE は Encoder で抽出した特徴量を元に Decoder で画像を復号している．Encoder で抽出した特徴量を元に入力とは異なる画像を復号できるように Decoder を学習すれば画像変換が可能となる．我々はこの性質に着目し，同一人物の顔の向きを変換する AE モデルの検討を行ってきた．そして AE に異なる解像度ごとに Contracting Path(Concat) を導入した U-Net を利用することで出力画像の品質が向上することを実験的に確認している．一般的に入力と出力を同一にした場合の U-Net では Concat の利用により勾配消失を抑制させることができ，高精度な出力画像が得られるとされている．しかし入力と出力が異なる場合にはそれぞれの結合がどのような役割を果たすのかは明らかではない．そこで各 Concat で伝える潜在変数の役割とその効果を確認する実験を行った．その結果，位置によって役割が変化し，色の情報や顔の情報，顔の回転情報を伝えていることを実験的に明らかにした．またカラー入力画像の RGB 成分を分解して確認した結果，我々が提案するモデルでは入出力間で色情報が混合することが無いことも確認した．

1　まえがき

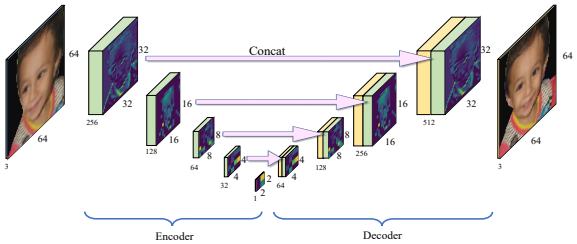
与えられた画像などの情報から特徴を抽出するための方法に Auto Encoder(AE)[1] がある．AE は入力画像を入力画像の次元よりも低次元の潜在変数空間に Encoder で符号化し，これを Decoder で入力画像と等しい出力画像が復元できるように学習をするものである．AE によって入出力が等しくなるように学習を行うことができれば低次元の潜在変数に，画像に含まれる特徴量が抽出されることになる．

Encoder で抽出した特徴量を元に入力とは異なる画像を復号できるよう Decoder を学習すれば画像変換が可能となる．我々はこの性質に着目し，同一人物の顔の向きを変換する AE モデルの検討を行ってきた．しかしながら入出力が異なる場合，そのままの AE では出力画像の品質の向上が難しい [2]．このような入出力が異なる場合には U-Net[3] と呼ばれる構造のモデルが有効であることが知られている．

U-Net は Olaf らによって生物医学のために提案されたセマンティックセグメンテーション用のモデルであり，Encoder と Decoder 間の同じ解像度のデータ間に Contracting Path(Concat) と呼ばれるスキップ構造が存在するモデルである．我々は U-Net を利用して，横向きの顔画像を入力，正面の顔画像を出力となるように学習を行ったモデルによって出力画像の品質が非常に高くなることを実験的に明らかにした [2]．そして，Concat の場所によって伝えている特徴が異なることを確認した [4]．しかしながら，Concat がどのような役割を果たしているかについての検討は不十分である．そこで本稿では，Concat で伝達される潜在変数に着目し，その特性を考察する．

2　モデル

U-Net[3] は学習時の勾配消失を防ぎ，高周波成分を保存するために，異なる解像度ごとに抽出された特徴を加算させることを意図した図 1 に示すような Concat が含まれるモデルである．



層	入力	Conv	Conv	Conv	Conv	Conv	Deconv	Concat	Deconv	Concat	Deconv	Concat	Deconv	Concat	Deconv(出力)
ch	3	256	128	64	32	1	32	64	64	128	128	128	256	512	3

図 1　U-Net

図 1 に示すモデルは特徴を抽出する Encoder とその特徴を利用して画像を再構成する Decoder を合わせて 10 層からなる．Encoder では画像の特徴抽出に畳み込みニューラルネットワーク (CNN) を使用し，畳み込み演算によってデータを縮小させることで特徴抽出を行う．図 1 の Encoder は 5 層の畳み込み層で構成され，64 × 64 次元の入力画像を 2 × 2 次元の潜在変数空間に符号化する．そして Decoder では 2 × 2 次元の潜在変数から 5 層の逆畳み込み層によって出力画像を生成する．このとき精度の高い出力画像が得られれば，潜在変数にはこの画像の特徴量が埋め込まれていると言える．ただし，U-Net で

1)　東京都市大学知能情報工学科

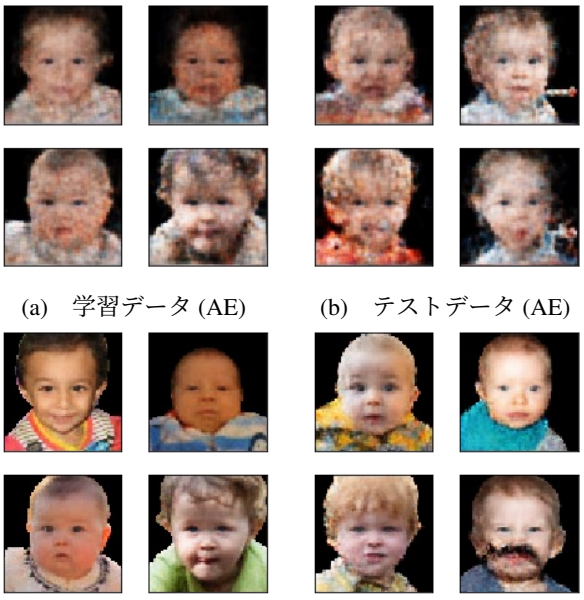
は Encoder の各畳み込み層で生成される異なる解像度ごとに抽出されたと考えられる潜在変数を Decoder の対応した解像度の逆畳み込み層に伝える Concat が存在する．このような Concat の存在により出力画像の品質が向上することが期待できる．

3　顔画像の向き変換

我々は図 2(a) に示す横向きの顔画像を入力とした際に図 2(b) の正面の顔画像が出力されるように学習をさせる．図 1 の U-Net から Concat を全て取り除いた AE モデルと U-Net モデルで学習した結果の出力画像例を図 3 に示す．図 3(a)(b) は AE モデルによる結果，図 3(c)(d) は U-Net モデルによる結果である．(a)(c) は学習に用いたデータに対する出力結果，(b)(d) は学習に用いていないデータに対する出力結果である．またそれぞれの出力結果の品質を確認するため，システムの出力結果と教師画像との平均二乗誤差 (MSE) を表 1 に示す．これらの結果から AE モデルに比べ，U-Net モデルの出力画像は品質が大きく改善されることが確認できる．出力結果から学習に用いた画像では顔の向きが教師画像通りに正しく向きが変換されていることを確認できる．学習に用いていない画像に関しても顔の向きが正しく変換されていることが確認できる．

(a)　横向きの顔画像　　(b)　正面の顔画像

図 2　顔画像



(a)　学習データ (AE)　　(b)　テストデータ (AE)

(c)　学習データ (U-Net)　　(d)　テストデータ (U-Net)

図 3　出力画像

表 1　生成画像と教師画像との平均二乗誤差

	Train	Test
AE	0.0102	0.0488
U-Net	**0.0007**	**0.0089**

4　Concat の役割の確認

前節の実験結果が示す通り，U-Net モデルでは出力画像の品質が改善される．AE と U-Net の違いは Concat の存在であるため，これらが出力画像の品質改善に貢献していることは間違いない．我々が用いた U-Net モデルでは 4 種類の解像度での情報を伝達する Concat が存在する．これらの Concat の役割を確認するため以下のような実験を行う．

図 4 のモデルを Baseline とし，U-Net に含まれる Concat が伝達する情報および最小解像度に対応した潜在変数を latent1，latent2，latent3，latent4，latent5 と名前をつける．Baseline モデルの入力画像に図 2(a) に示す横向きの顔画像を，出力に図 2(b) の正面の顔画像を与え学習を行う．

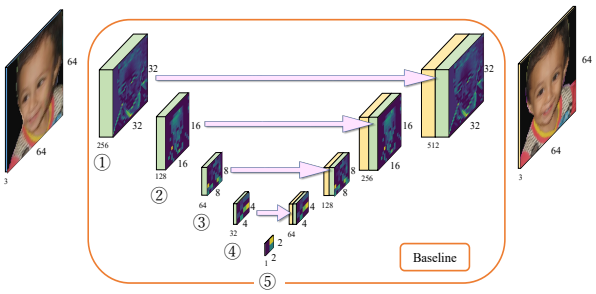


図 4　Baseline

Baseline の出力結果は図 3(c)(d) に示したものである．学習したモデルのうち特定の latent 情報のみを残し，その他は 0 にすることで特定の latent 情報の役割を検討する．特定の latent 情報のみを用いた場合の出力結果を図 5 に示す．図 5(a) から (e) は学習に用いた画像を入力した際の出力結果，図 5(f) から (j) は学習に用いなかった画像を入力した際の出力結果を表す．図 5 より，latent1 のみの情報では入力画像にローパスフィルタが乗じられた状態の画像，latent2 のみの情報では横向きに変換した顔の情報が伝えられていることが確認できる．latent3 から latent5 ではどのような情報を伝えているかはこの結

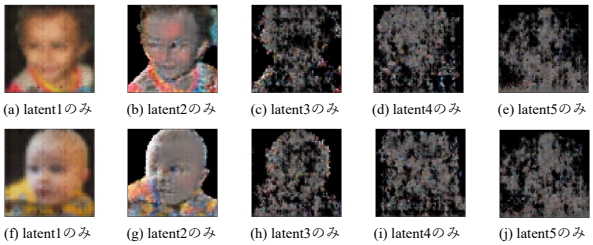


図 5　各 Concat

果からは不明であるが，色に関する情報を伝達している様子は抽出されなかった．

latent3 から latent5 が伝達する情報を確認するため，Baseline から各 latent 情報を除去した場合の出力結果を比較する．特定の latent 情報を除去するため，該当する latent 情報を 0 とした．この時の出力結果を図 6 に示す．図 6(a) から (e) は学習に用いた画像を入力した際の出力結果，図 6(f) から (j) は学習に用いなかった画像を入力した際の出力結果を表す．図 6 の各キャプションの「-n」が latent n を削減したことを意味する．

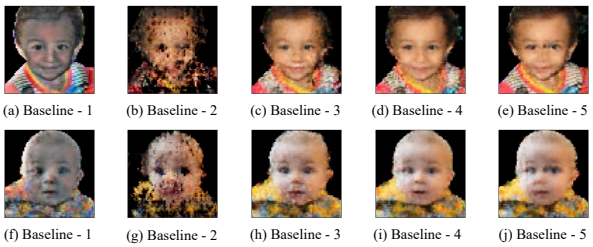


図 6　潜在変数を 1 つずつ減らした場合

図 6(a) が示すように latent1 の情報を取り除くと色の情報が失われ，(b) のように latent2 を取り除いた場合は顔の情報が失われた [4]．(c)(d) の latent3，latent4 を取り除いた場合，出力画像の品質は低下したものの，顔画像の回転や顔の生成への影響は見られない．(e) に示す latent5 を取り除いた場合，目付近に着目すると横向きの顔画像での目の位置と正面の目の位置の情報があらわれているように見える．この結果からは顔画像の向きを回転させる情報は latent5 に含まれるように考えられる．

U-Net の学習では初期値依存性が存在するため複数回試行を行った結果，latent3 に関しては異なる結果が得られる場合を観測した．baseline から latent3 を取り除いた場合に試行によって異なる結果が得られた際の出力結果を図 7 に示す．

図 7　latent3 を取り除いた際の試行による違い

図 7 より試行によって精度にのみ影響を与える場合と，目の回転情報を伝えている結果が得られる場合がある．一方 latent1, latent2 に関しては試行によらず同じ結果が得られた．

5　latent1, latent2 の役割

前節の結果から，latent1 は主に色に関する情報を伝達し，latent2 は顔の情報及び色情報も伝達していると考えられる．これを確認するため，カラー入力画像の RGB の 3 チャンネルのうち特定の 1 チャンネルのみを入力した場合にどのような出力が得られるのかを検討する．実験では学習に使用した画像のみを用いた．RGB のうちそれぞれ 1 チャンネルのみを入力した Baseline モデルから各 latent 情報を削減した際の出力結果を図 8 に示す．図 8 の各キャプションは図 6 と同様である．

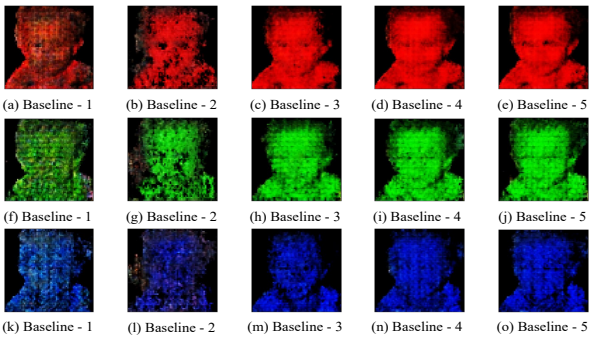


図 8　カラー別潜在変数 (単色)

図 8 から分かる通り RGB の単色画像のみを入力した場合，Baseline からどの latent 情報を削減しても入力した色情報のみを有した出力結果が得られた．このことから，latent 情報によらず，入力の各色情報は他の色情報と混じることなく伝達されることが確認できた．

次に入力の色情報が保存されていることを利用し，入力画像の位置情報が出力画像にどのように伝達しているかを確認する．具体的には入力画像を縦方向横方向共に三分割し，9 分割された各領域を隣り合う領域で色が異なるように RGB いずれかの色情報のみを残し入力画像とした．この画像を入力した際にどのような出力が得られるのかを検討する．結果を図 9 に示す．

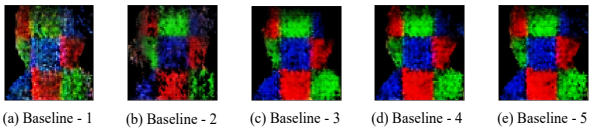


図 9　カラー別潜在変数 (RGB)

図 9 の結果より，異なる色が隣り合っていても境界線ははっきりし，色が混ざり合うことはない．このことから我々が提案する U-Net モデルでは顔の位置のみでな

く色の位置もずらさずに学習ができていることを確認した．

6　Concat 別での学習

図 4 のモデルのうち Concat を 1 つのみ残し，他の Concat は削減した状態で学習を行った．この時の出力結果と教師画像との MSE を表 2 に示す．

表 2　Concat 別 MSE

	Concat1	Concat2	Concat3	Concat4
Train	0.0084	**0.0014**	0.0015	0.0019
Test	0.0132	**0.0087**	0.0108	0.0151

表 2 より，Concat2 のみ残して学習を行った場合，学習データ，テストデータ共に MSE が小さくなった．

Concat2 のみ繋げることで精度が向上することを確認したため，図 4 のモデルのうち Concat2 を含む複数の Concat を繋げたモデルで学習を行う．Concat2 と他の Concat を 1 つ組み合わせ，Concat を 2 つ利用した場合と Concat2 と他の Concat を 2 つ組み合わせ，Concat を 3 つ利用した場合で学習を行う．表 2 と同様に MSE を求め，Concat を 2 つ利用した場合の結果を表 3，Concat を 3 つ利用した場合の結果を表 4 に示す．

表 3　Concat を 2 つを利用した場合の MSE

	Concat1,2	Concat2,3	Concat2,4
Train	**0.0014**	0.0010	0.0010
Test	**0.0087**	0.0089	0.0089

表 4　Concat を 3 つ利用した場合の MSE

	Concat1,2,3	Concat1,2,4	Concat2,3,4
Train	0.0009	**0.0009**	0.0008
Test	0.0090	**0.0087**	0.0088

表 3 と表 4 より，Concat1 と Concat2 を繋げた場合と Concat2 のみを繋げた場合では少数第 4 位までの MSE は同じであるが，わずかに Concat2 のみを繋げた方が MSE が小さくなった．Concat3 を繋げることにより，MSE が少し上がり精度が下がることを確認した．表 1 の U-Net(Concat 全てを繋げた場合) と比較しても Concat3 が入ることにより精度が悪くなっているといえる．また，繋げる Concat の数が多いほど学習データでの MSE とテストデータでの MSE が乖離していき，過学習が起きてしまっている．

数値のみでなく画像でも比較を行う．Concat2 のみ繋げた場合の出力結果を図 10 に示す．

(a)　学習データ　　(b)　テストデータ

図 10　Concat2 のみでの出力結果

図 10 より，全ての Concat を利用した場合である図 3(c)(d) と比較すると全てを利用した方が学習データの精度は上がるが，テストデータの精度は変化しない．

これらの結果から Encoder と Decoder の全ての層を Concat で繋げると精度が向上するということはなく，特に latent2 が精度を上げる上で重要であるといえる．

7　まとめ

U-Net を用いた顔画像の向きを変化させる変換システムにおいて，Concat で伝達される情報がどのようなものであるかについて実験的に考察を行った．その結果，入力画像の解像度に近い情報が伝達される latent1, latent2 では画像に含まれる色情報，顔画像の形状に関する情報が伝達されていることを確認した．色に関しては入力画像の色情報が出力にそのまま伝えられ，色が混ざらないことを確認した．しかしながら顔の回転に関する情報がどのように伝達，生成されているかは十分には解明できていない．また，今までは latent1 から latent4 まで全ての情報を Concat で伝えていたが，テストデータにおいては latent2 のみを伝える場合と変化がなく，latent3 の情報を伝えることで精度が下がった．今後は Concat2 を繋げたときの潜在変数について考え，回転変換がどこで伝えられているかについて検討を進める．

謝辞

本研究の一部は JSPS 科研費 JP20K11978 の助成，および 東北大学 電気通信研究所 共同プロジェクト研究によるものです．

参考文献

[1] Geoffrey E. Hinton; R. R. Salakhutdinov,“Reducing the Dimensionality of Data with Neural Networks”, Science 313 (5786), pp. 504-507, 2006. DOI: 10.1126/science.1127647

[2] 岡本 紗季，神野 健哉，”AutoEncoder による顔の向き変換”，電子情報通信学会 2022 年総合大会，N-1-12，2022.

[3] Jonathan Long, Evan Shelhamer, Trevor Darrell, ”Fully convolutional networks for semantic segmentation”, IEEE CVPR2015, pp. 3432-3440, 2015. DOI:10.1109/CVPR.2015.7298965.

[4] 岡本 紗季，神野 健哉，”U-Net による顔画像の回転変換”，電子情報通信学会 2022 年 NOLTA ソサイエティ大会，NLS-40，2022.