The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Wasserstein Autoencoderを用いた画像スタイル変換

A Style Transfer Method using Wasserstein Autoencoder

中田 秀基 *1　　麻生 英樹 *1
Hidemoto NAKADA　　Hideki ASOH

*1産業技術総合研究所 人工知能研究センター
Artifical Intelligence Research Center, National Institute of Advanced Institute of Technology

We propose a image style transfer method based on Wasserstein Autoencoder. Style transfer is an area of image generation technique that generates an image that shows content taken from one image with a style taken from another image. While there are extensive researches in this area, most of them requires some 'training' time to generate images with specific style. The proposed method does not require training time which trains a network that can be used for any style and content image. The network encode images into content latent variables and style latent variables. We can transfer image style by simply replacing style latent variables. We tested the proposed method with images from CelebA and confirmed that it can generate style transferred images.

1. はじめに

画像スタイル変換とは、コンテント画像に対してスタイル画像から抽出したスタイルを適用することで、任意のコンテントを任意のスタイルで描画することを実現する技術である。Gatys ら [Gatys 15][Gatys 16] の手法は、画像そのものを最適化することでスタイル変換を実現するため、スタイル変換に時間がかかるという問題がある。

これに対して、Johnson ら [Johnson 16] はフィードフォワードネットワークのみでスタイル変換を行う手法を提案した。この手法は、スタイル変換自体は瞬時に行うことができるが、スタイルにあわせてフィードフォワードネットワークを訓練する必要があり、これには一定の訓練時間が必要となる。

われわれは、Variational Autoencoder を用いたスタイル変換の研究を進めてきた [中田 18]。本稿では、Wesserstein Autoencoder[Tolstikhin 18] を用いてスタイル変換を行う手法を提案する。スタイルを表す隠れ変数へのエンコーダ、コンテントを表す隠れ変数へのエンコーダ、2 つの隠れ変数からのデコーダを同時に学習させることで、任意のスタイル画像と任意のコンテント画像からの変換済み画像の生成を実現する。学習した結果のネットワークは、任意のスタイル画像に対するスタイル変換を行う汎用のフィードフォワードネットワークとなるので、

本論文の構成は以下の通りである。2. で本研究の背景を概説し、3. で提案手法について述べる。4. で実験結果を示す。5. で関連する研究を議論する。6. では、まとめと今後の課題について述べる。

2. 背景

2.1 画像スタイル変換

Gatys ら [Gatys 15][Gatys 16] は、物体を識別するように訓練された CNN の各レイヤのレスポンスからスタイルを表す特徴量を構成することが可能であることを示した。この特徴量は、各レイヤを構成する各フィルタ（チャンネル）におけるレスポンスの相関をとったグラム行列で、レイヤ l のフィルタ i, j に対して特徴マップをベクタにしたものを $\mathbf{F_i}$,$\mathbf{F_j}$ として、$G^l_{i,j} = \mathbf{F_i} \cdot \mathbf{F_j}$ と表される。ここで $\cdot$ はベクタの内積を表す。以降これをスタイル特徴量と呼ぶ。

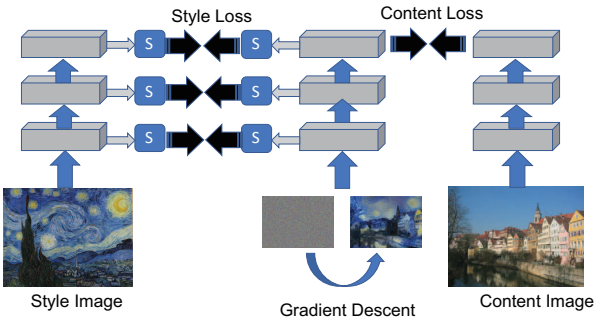

図 1: Gatys らの手法。図中の画像は [Gatys 15] より。

1）スタイルロス：生成画像とスタイル画像のスタイル特徴量の差、2) コンテントロス：コンテント画像と生成画像の上位レイヤでのレスポンスの差、とし、この 2 つのロスの線形和をロス関数とする。Gatys らの手法は、このロス関数が最小になるように、ノイズで初期化した生成画像を勾配降下法で最適化する。図 1 にこの様子を示す。左右に 3 つ並んでいるネットワークはすべて事前に ImageNet を学習した同一のネットワークで、固定されている。画像生成時には、スタイル画像を左のネットワークに、コンテント画像を右のネットワークに、ノイズで初期化した生成画像を中央のネットワークに与え、ロス関数を用いて、生成画像を最適化する。この手法は、非常に解像度の高い画像が得られる一方で、画像の生成に勾配降下法を用いるため、時間がかかる。

この問題点を解決するために Johnson ら [Johnson 16] はフィードフォワードネットワークのみでスタイル変換を行う手法を提案した。この手法は Gatys らの提案したものと同じロス関数を用いて画像を変換するネットワークを訓練する。スタイル変換はフィードフォワードネットワークに一度通すだけで行われるため、変換時の計算負荷は小さい。一方で、画像変換ネットワークの訓練にはコストが掛かり、個々のスタイルに対して、GPU を用いても数時間の訓練が必要となる。

丹野ら [丹野 17] は、一つのネットワークに複数のスタイルを学習させる方法を提案しているが、スタイルごとに学習が必要であるという意味では、本質的な解決とはなっていない。

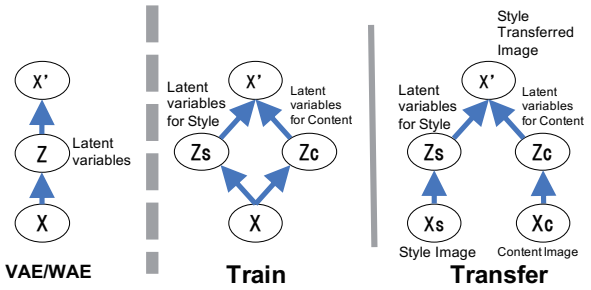

図 2: 左: VAE の構造。右：提案手法の概要

2.2 Variational Autoencoder と Wesserstein Autoencoder

Variational Autoencoder[Kingma 14][Doersch 16]（変分オートエンコーダ以下 VAE）は、隠れ変数 $\mathbf{z}$ から学習データ $\mathbf{x}$ を生成する確率分布 $p(\mathbf{x}|\mathbf{z})$ を学習する教師なし学習の一種である。具体的には、隠れ変数 $\mathbf{z}$ がある確率分布に従うことを仮定し、学習データ $\mathbf{x}$ から $\mathbf{z}$ へ変換するエンコーダネットワークと $\mathbf{z}$ から再構成データ $\mathbf{x}'$ へ変換するデコーダネットワークを同時に学習する。この際、$\mathbf{x}$ と $\mathbf{x}'$ の差と、隠れ変数 $\mathbf{z}$ と仮定した確率分布の差（KL ダイバージェンス）をロス関数として、エンコーダとデコーダを最適化する (図 2 左)。一般には仮定する $\mathbf{z}$ の事前確率分布として、平均 0 分散 1 の正規分布を用いる。

Wesserstein Autoencoder（以下 WAE）は、VAE が隠れ変数と仮定した確率分布の差の KL ダイバージェンスを最小化するのに対して、Wesserstein 距離を最小化する。生成画像を比較すると、WAE は VAE よりも良好な FID(Frechet Inception Distance) スコアを示すことが示されている。このため本稿では、WAE を用いた。

3. 手法

3.1 概要

本手法の基本的な発想は、スタイルとコンテントを隠れ変数として分離して学習させることにある (図 2 中央)。これを実現できれば、図 2 右のようにスタイル画像からスタイル隠れ変数を、コンテント画像からコンテント隠れ変数を取り出したものをデコーダネットワークに与えることで、スタイル変換を実現できる。

3.2 提案手法

提案手法では、スタイル隠れ変数を Gatys らのスタイル行列から抽出し、コンテント隠れ変数は通常の WAE で学習する。図 3 に提案手法の学習時のネットワークを示す。

左辺のネットワークには、Imagenet を用いて事前学習された VGG16[Simonyan 14] ネットワークを用いる。このネットワークを用いてスタイル行列を作り、それからニューラルネットワークを介してスタイル隠れ変数を構成する。このネットワークは学習の対象ではない。右辺のネットワークは通常のコンボリューションを用いたエンコーディングネットワークでこれからコンテント隠れ変数を構成する。これら 2 つの隠れ変数から、デコーダネットワークを用いて、画像を再構成する。

学習時のロスとしては、通常の WAE と同様に再構成画像と入力画像の誤差と、隠れ変数の事前確率と事後確率の Wesserstein 距離を用いる。これらを用いてエンコーダネットワーク、デコーダネットワーク、スタイル行列から隠れ変数を構成するネットワークの 3 つを同時に学習する。

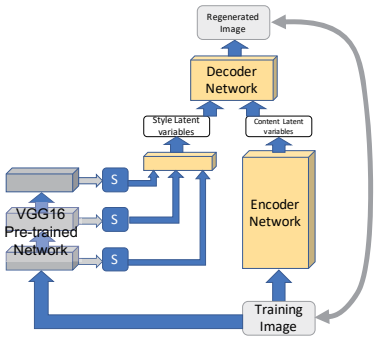

図 3: WAE を用いたスタイル変換の学習

表 1: エンコーダネットワークの構成

層	種類	入力	出力	カーネル	ストライド
1	Conv	64x64x3	64x64x8	3	1
2	Conv	64x64x8	32x32x32	4	2
3	Conv	32x32x32	16x16x128	4	2
4	Conv	16x16x64	8x8x256	4	2
5	Conv	8x8x128	4x4x512	4	2
6	Full	4x4x256	128 x 2	-	-

変換時には、左辺のスタイル隠れ変数導出ネットワークにスタイル画像を、右辺のエンコーダネットワークにコンテント画像を入力する。フィードフォワードネットワークの実行一回で任意のスタイル画像に対して任意のコンテント画像を適用する事ができる。

4. 実験

4.1 ロス

前述の通り、ロスは入力画像と再構成画像の誤差と、隠れ変数の事前確率と事後確率の Wesserstein 距離を用いる。前者の画像誤差としては、通常のピクセルの自乗誤差に加えて、スタイルネットワークの 3 層目までの出力の自乗誤差と、スタイル行列そのものの自乗誤差を用いた。これは、より強くスタイルを反映するようにするためである。各項の寄与はほぼ同等となるように重みで調整してある。

4.2 ネットワーク

図 3 に示したネットワークを、Tensorflow[ten] を用いて実装した。隠れ変数の数は、スタイル側を 1 層当たり 64 変数とし、3 層分まで用いた。したがって隠れ変数の数は 192 となる。スタイルとコンテントの隠れ変数の数の比率の影響を調べるため、コンテント側の隠れ変数の数を 512, 256, 128, 64, 32 と変化させて実験を行った。

エンコーダネットワークは標準的なコンボリューションネットワークである。このネットワークのパラメータを表 1 に示す。コンボリューション層では、max pooling 等は用いず、ストライドでサイズを小さくしていっている。活性化関数にはエンコーダでは eLU(Exponential Linear Unit) を、デコーダでは ReLU(Rectified Linear Unit) を用いた。

スタイル行列からスタイル隠れ変数を生成するネットワークには完全結合ネットワークを用いた。各層のレスポンスから生成したスタイル行列に対して、出力を 64x2 に設定した完全結合ネットワークを 3 段適用して、その結果を隠れ変数とした。

デコーダネットワークの構成を表 2 に示す。スタイル隠れ変数をサンプルしたものと、コンテンツ隠れ変数をサンプルした

表 2: デコーダネットワークの構成

種類	出力	カーネル	ストライド
Full	4x4x1024		
Subpixel conv	8x8x512	3x3	1
ResBlock	8x8x512		
ResBlock	8x8x512		
ResBlock	8x8x512		
Subpixel conv	16x16x256	3x3	1
Subpixel conv	32x32x128	3x3	1
Subpixel conv	64x64x64	3x3	1
Conv	64x64x3	5x5	1

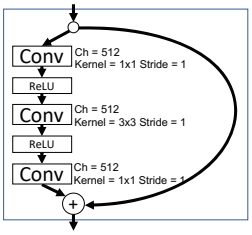

図 4: ResBlock の構造

ものを一本のベクトルにし、全結合ネットワークで、4x4x1024 に形を整えてから Subpixel Convolution[Shi 16] で 8x8x512 に変形してから図 4 に示す ResBlock[He 16] を 3 度通し、それをさらに Subpixel Convolution[Shi 16] で 3 回拡大して、64x64 の画像を得ている。

また、学習を高速化するために WeightNormalization[Salimans 16] を用いている。

4.3 データ

学習データ、変換前のコンテントデータとしては CelebA[cel] を用いた。スタイル画像としては図 5 に示す 8 枚の画像を用いた。また、学習データの多様性を確保するため、アニメ顔画像データセット [ani] と ImageNet[Russakovsky 15] の画像を一部用いた。

CelebA に関しては、OpenCV を用いて顔の中心を識別し顔のサイズが大まかにおなじになるようクロップして 64x64 とした。顔が小さい画像は排除したため、193,800 枚となった。

アニメ顔画像データセットは縦横いずれか長い方が 160 ピクセルの画像である。一辺が短辺の長さの正方形領域を切り出し、64x64 にリサイズして利用した。14,490 枚を利用した。

ImageNet に関しては色数が 3 で幅、高さがともに 64 以上のものを選択し、中心部 64x64 の領域を切り出して利用した。枚数は 196,371 枚を用いた。

4.4 結果

4.4.1 コンテント画像の再構成

スタイル変換の結果を示す前に、提案ネットワークの再現性能を示すために、コンテント画像（CelebA）を再構成したものを図 6(左) に示す。1 行目がオリジナル画像、2 行目-6 行目は CelebA のみで学習した場合、7 行目-11 行目はアニメ画像、IamgeNet を加えた学習した場合である 2 行目-6 行目、7 行目-11 行目はそれぞれ、コンテント画像の隠れ変数の数がそれぞれ、512, 256, 128, 64, 32 の場合の結果を示している。

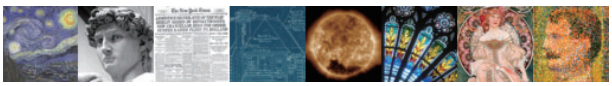

図 5: スタイル画像

図 6: 再構成画像。左:コンテント, 右:スタイル

学習データによる影響は見られない。コンテント画像の隠れ変数の数が増えるにしたがって、再現が正確になる傾向があることがわかる。

4.4.2 スタイル画像の再構成

同様にスタイル変換対象画像を再構成した結果を図 6(右) に示す。各行の意味は前項と同様である。

学習データの多様化による影響は大きく、アニメ画像、ImageNet を加えた場合のほうが良好な再構成ができている。また、CelebA のみで学習したケースでは、入力にない顔のような構造が再構成画像に現れている。これは特にコンテント画像の数が少ない場合に顕著である。これは CelebA のみで学習した結果、顔を出力するよう生成ネットワークが訓練されてしまったためであると考えられる。

4.4.3 スタイル変換結果

スタイル変換の結果を図 7 に示す。1 行目がコンテント画像、2 行目がスタイル画像、上段（3 行目-7 行目）は CelebA のみで学習した場合、下段（8 行目-12 行目）はアニメ画像、IamgeNet を加えた学習した場合の結果で、コンテント画像の隠れ変数の数がそれぞれ、512, 256, 128, 64, 32 の場合を示している。いずれの場合も、基本的にスタイル画像の特徴を再現したスタイル変換ができているといえる。

また、上段、下段いずれの場合もコンテント隠れ変数が減少すると、スタイル画像による影響がおおきくなることがわかる。また上段の場合には、CelebA の平均的な顔画像に引きずられる傾向がある。

9 行目、10 行目（コンテント隠れ変数 256, 128）近辺で特に良好なスタイル変換を実現できているといえる。

5. 関連研究

Li ら [Li 17] は、コンテント画像に対して whitening という処理を施した後に、スタイル画像に基づく coloring という操作を各レイヤで施すことでスタイル変換を行う。この手法は、

図 7: WAE を用いたスタイル変換

提案手法同様に任意のスタイルを任意のコンテントに適用でき、非常に高品質な画像生成を達成している。提案手法とは全く異なるアプローチであるため、比較を行う必要がある。

6. おわりに

本稿では、Wesserstein Autoencoder を用いて、スタイル変換を行う手法を提案し、ネットワーク学習時に多様な画像を与えることで良好なスタイル変換が行えることを確認した。また、スタイル隠れ変数とコンテント隠れ変数の比率を変化させることでスタイルの寄与度を変更できることを確認した。

今後はより多様な画像への適用を行い、従来手法との比較を行う予定である。

謝辞

実装をお手伝いいただいた井上辰彦氏に感謝する。この成果の一部は、国立研究開発法人新エネルギー・産業技術総合開発機構 （ＮＥＤＯ）の委託業務の結果得られたものである。本研究は JSPS 科研費 JP16K00116 の助成を受けたものである。

参考文献

[ani] AnimeFace Character Dataset : `http://www.nurs.or.jp/~nagadomi/animeface-character-dataset/` Accessed: 2018-02-01

[cel] Large-scale CelebFaces Attributes (CelebA) Dataset, `http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html` Accessed: 2018-02-01

[Doersch 16] Doersch, C.: Tutorial on Variational Autoencoders, arXiv:1606.05908v2 [stat.ML] (2016)

[Gatys 15] Gatys, L. A., et al.: A Neural Algorithm of Artistic Style, arXiv:1508.06576 (2015)

[Gatys 16] Gatys, L. A., et al.: Image Style Transfer Using Convolutional Neural Networks, in *Proc. of IEEE Computer Vision and Pattern Recognition* (2016)

[He 16] He, K., et al.: Deep Residual Learning for Image Recognition, in *Computer Vision and Pattern Recognition (CVPR)* (2016)

[Johnson 16] Johnson, J., et al.: Perceptual Losses for Real-Time Style Transfer and Super-Resolution, in *Proc. of European Conference on Computer Vision*, pp. 694–711 (2016)

[Kingma 14] Kingma, D. P. and Welling, M.: Auto-encoding variational Bayes, in *Proceedings of the 2nd International Conference on Learning Representations* (2014)

[Li 17] Li, Y., et al.: Universal Style Transfer via Feature Transforms, in *Advances in Neural Information Processing Systems 30*, pp. 385–395 (2017)

[Russakovsky 15] Russakovsky, O., et al.: ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision (IJCV)*, Vol. 115, No. 3, pp. 211–252 (2015)

[Salimans 16] Salimans, T. and Kingma, D. P.: Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks, *CoRR*, Vol. arXiv/1602.07868, (2016)

[Shi 16] Shi, W., et al.: Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network, *CoRR*, arXiv/1609.05158, (2016)

[Simonyan 14] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *CoRR*, arXiv/1409.1556, (2014)

[ten] TensorFlow : `https://tensorflow.org/` Accessed: 2015-11-04

[Tolstikhin 18] Tolstikhin, I., et al.: Wasserstein Auto-Encoders, in *International Conference on Learning Representations* (2018)

[丹野 17] 丹野 良介, 下田 和, 柳井 啓司:複数スタイルの融合と部分的適用を可能とする Multi-style Feed-forward Network の提案, 人工知能学会全国大会予稿集 (2017)

[中田 18] 中田 秀基, 麻生 英樹 : Variational Autoencoder を用いた画像スタイル変換, 信学技報, vol. 117, no. 514, PRMU2017-192, pp. 121–126 (2018)