4G4-3

深層学習における敵対的ネットワークと注視を用いた画像生成の試み

Image generation using Generative Adversarial Nets and Attention in deep neural network

片岡裕介 *1　松原崇 *2　上原邦昭 *2
Yuusuke Kataoka　Takashi Matsubara　Kuniaki Uehara

*1神戸大学工学部情報知能工学科
Department of Computer Science and Systems Engineering, Kobe University

*2神戸大学大学院システム情報学研究科
Graduate School of System Informatics, Kobe University

For image generation, deep neural networks are trained to extract high-level features on natural images and to reconstruct the images from the features. However it is difficult to learn to generate images containing enormous contents. To overcome this difficulty, networks with an attention mechanism has been proposed. It is trained attending to parts of the image and generating images step by step. This enables the network to deal with the details and the rough structure of natural images. Additionally, the Generative Adversarial Nets (GANs) approach is effective method for training generative models with neural networks. In this study, we present image generation with the attention mechanism trained using the Generative Adversarial Nets approach. We show our method enables the iterative construction of images and more realistic image generation than standard GANs.

1. はじめに

良い自然画像の生成モデルを構築することは，深層学習による画像生成において根本的な問題である．そのようなモデルでは自然画像に対して適した確率分布を定義することで，データの潜在的な構造を獲得する必要がある．

深層学習における画像生成の手法としては，まず変分オートエンコーダー (VAE)[Kingma 14] を使ったものが挙げられる．VAE は変分ベイズ法を用いたネットワークと潜在変数の学習を行う．しかし VAE は画素単位の誤差を使用するため，生成した画像は大きくぼやけているという問題があった．VAE を元にした改良手法として DRAW[Gregor 15] が挙げられる．DRAW は注視のメカニズムを用いた画像生成手法であり，VAE に再帰ニューラルネットワークを組み合わせることで，注視の学習を行い画像を生成する．注視のメカニズムを使うことで，画像の各部分を注視し複数回に分けて画像を生成することができる．これによって画像全体を眺めることで全体のおおまかな特徴と構造を捉え，細部に注目することで詳細な特徴と構造について捉えることにより，全体の整合性が取れた画像が生成できるようになる．しかし DRAW は VAE を元にした手法であったため，生成される画像がまだぼやけているという問題点があった．VAE とは異なる画像生成の手法として敵対的ネットワーク [Goodfellow 14] がある．敵対的ネットワークはヒューリスティックな誤差に対して学習を行うため，生成される画像がぼやけているということはなくなった．特に深層畳み込み敵対的生成ネットワーク [Radford 15] は人間の顔画像の生成等で高い成果を挙げている．しかし敵対的ネットワークを用いた手法では生成画像にノイズがかかっていたり，構造が歪んだ画像が生成されるという問題があった．

そこで敵対的ネットワークに注視のメカニズムを組み合わせることで，画像がぼやける，ノイズがかかるといった問題を解決できると考えられる．DRAW で用いられる注視の手法は，微分可能なアルゴリズムを使用しており，深層学習で盛んに使われる誤差逆伝播法を用いたネットワークの学習に適していると考えられる．

本研究では注視を用いた画像生成手法の 1 つである DRAW の注視メカニズムを画像生成の学習手法の 1 つである敵対的ネットワークを用いて学習することで，複数回に分けて画像の各部分が生成されていき，最終的に全体の画像が生成されることで，より本物らしい画像が生成されることを示す．

連絡先: 片岡裕介，神戸大学工学部情報知能工学科，kataoka@ai.cs.kobe-u.ac.jp

2. 提案手法

2.1 注視のメカニズム

DRAW の注視メカニズムではフィルタの中心位置 (g_X, g_Y) と各格子点間の幅を表す δ により，フィルタの位置と適用する範囲を定めている．このフィルタには $N \times N$ の格子点を持つガウシアンフィルタを用いる．各点 (i,j) におけるフィルタの平均値 (μ_X^i, μ_Y^j) と等方的な分散 σ^2，密度計数 γ を使用し，フィルタの各点の値を決定する．入力画像のサイズが $A \times B$ の時，これら 5 つのパラメーターは各ステップ毎に導出され，再帰ニューラルネットワークの出力 h の線形変換として定義される．再帰ニューラルネットワークは入力 x_t, 出力 h_t, 前ステップの出力 h_{t-1} を用い以下の様な計算式を取る．

$$h_t = RNN[h_{t-1}, x_t] = \sigma(W_x(x_t) + W_h(h_{t-1})) \quad (1)$$

ここで W はバイアスの追加を含めた線形変換を表す．σ は活性化関数を表し主にシグモイド関数を使用する．

2.2 敵対的ネットワーク

敵対的ネットワークは生成部 G と識別部 D から構成される学習手法である．それぞれ内部に独立したニューラルネットワークを持つ．生成部 G はノイズの入力変数 z から元となるデータ x の事前分布 p_g の学習を行う．$D(x)$ は p_g から生成された x と，元データの x をそれぞれ個別に入力として取り扱う．D は入力データ x に対し，それが本物（データセットの一部）か偽物（生成されたもの）についての確率を出力する．

D は学習データが本物である確率と G の生成サンプルが偽物である確率を最大化するように学習を進め，同時に G は G

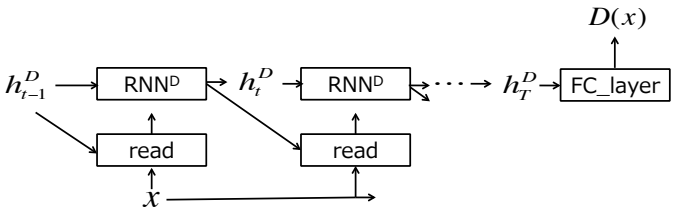

図 1: 識別部 D のネットワーク.

の生成サンプルが本物である確率を最大化するように学習する．学習が進むと，D は元データと生成されたサンプルを識別出来るようになり，G は D が元データとして識別するようなデータを生成するようになる．

2.3 識別部 D の定義

本手法の識別部 D のネットワークは DRAW のエンコーダー部分を元にした構成になっている．D のモデル図を図 1 に示す．読み取りの入力には入力画像 x と前のステップの D 中の再帰ニューラルネットワークの出力 h_{t-1}^D を用いる．ここで $t = 1, ..., T$ であり h_0^D はランダムな初期値を与えられ，このパラメータは学習される．

$$read(x, h_{t-1}^D) = \gamma[F_Y x F_X^T] \tag{2}$$

読み取った画像 r_t と h_{t-1}^D は結合され，再帰ニューラルネットワーク RNN^D の入力に使われる．

$$r_t = read(x, h_{t-1}^D) \tag{3}$$

$$h_t^D = RNN^D(h_{t-1}^D, [r_t, h_{t-1}^D]) \tag{4}$$

そして $t = T$ の時の RNN^D の出力 h_T^D を線形変換し，D の出力 $D(x)$ を決定する．

$$D(x) = \sigma(W(h_T^D)) \tag{5}$$

2.4 生成部 G の定義

本手法の生成部 G のネットワークは DRAW のデコーダー部分を元にした構成になっている．G のモデル図を図 2 に示す．G の各ステップの入力に使う入力変数 z_t には平均 μ_t，標準偏差 σ_t の正規分布によるノイズが与えられる．z_t は再帰ニューラルネットワーク RNN^G の入力として使われる．h_0^G はランダムな初期値を与えられる．

$$z_t \sim N(Z_t|\mu_t, \sigma_t) \tag{6}$$

$$h_t^G = RNN^G(h_{t-1}^G, z_t) \tag{7}$$

h_t^G は書き取りの入力に使われ，画像行列 c_{t-1} に加える事で c_t を出力する．c_0 はバイアス項として初期化される．

$$c_t = c_{t-1} + write(h_t^{dec}) \tag{8}$$

最終出力は $t = T$ の時の c_T であり，全体画像を表す．

2.5 学習アルゴリズム

学習アルゴリズムには敵対的ネットワークの学習法を使用する．D は入力 x の正解率を最大，入力 c_T の確率を最小化するように学習を進める．G は入力 c_T の時の D の確率を最大化するように学習を進める．

$$E_D(D, G) = E_{x \sim p_{data}(x)}[\log D(x)] + E_{z \sim p_z(z)}[\log(1 - D(c_T)] \tag{9}$$

$$E_G(D, G) = E_{z \sim p_z(z)}[\log(D(c_T)] \tag{10}$$

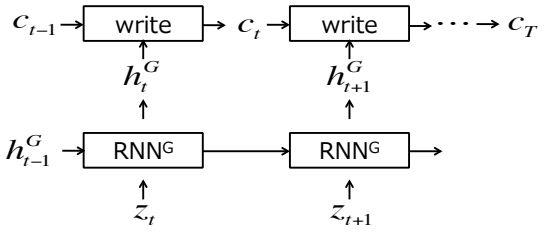

図 2: 生成部 G のネットワーク.

図 3: MNIST の数字画像サンプル.

3. 評価実験

3.1 画像の生成

実験には MNIST[*1] の画像データセットを使用した．MNIST は各画像に対し数字の 0,...,9 の内 1 つが描かれている，グレースケールの手書き文字によるデータセットであり，各画像のサイズは 28×28 で統一されている．図 3 に MNIST の画像サンプル 100 枚を示す．これを用いて生成部 G が本物らしい画像を出力するかどうか実験を行い，通常の敵対的ネットワークを使用した画像生成の結果と比較した．また注視に使用するフィルタの格子点の数を少なくした場合，生成画像がどのように変化するか実験を行った．さらに提案手法について学習後の生成部 G について，入力の各 z が出力結果にどのような影響を及ぼしているかの検証実験を行った．

3.2 実験設定

学習用データとして MNIST の数字画像 50,000 枚を利用し，バッチサイズ 100 枚による学習を行った．注視のフィルタの格子点の数はを $N \times N = 5 \times 5$ に設定し，注視の回数 (入力画像の読み取り回数と書き込み回数) は $t = 6$ 回に設定した．この注視の回数は再帰ニューラルネットワークの再帰回数と等しい．生成部 G の入力には平均 0, 標準偏差 1 の正規分布から独立にサンプリングした 100 次元のベクトル $\boldsymbol{z}$ を使用し，本手法における生成部 G と識別部 D の再帰ニューラルネットワークのユニット数は 256 に定めた．また c_0 の初期値として，学習用データの周辺分布からサンプリングした値を使用した．通常の敵対的ネットワークは [Goodfellow 14] で使用されているパラメータとネットワーク構造を使用した．最適化手法には ADAM[Kingma 15] を使用し．提案手法では生成部 G と識別部 D に対し学習係数を $\alpha = (3.75e-4, 3.75e-4)$, $\beta 1 = (0.75, 0.75)$, $\beta 2 = (0.999, 0.999)$ に設定し最適化を行い，敵対的ネットワークでは生成部 G と識別部 D に対し，学習係数を $\alpha = (1e-3, 1e-3)$, $\beta 1 = (0.5, 0.5)$, $\beta 2 = (0.999, 0.999)$ に設定し最適化を行った．提案手法の学習回数は 49 回，敵対的ネットワークの学習回数は 99 回で実験を行った．フィルタの格子点を少なくした場合の実験では，上記の提案手法の設定から

*1 http://yann.lecun.com/exdb/mnist/

図 4: 提案手法で生成した画像.

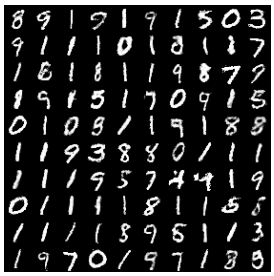

図 5: 敵対的ネットワークで生成した画像.

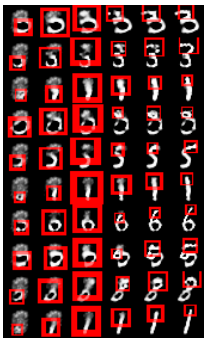

図 6: 各時間毎に生成した画像.

フィルタの格子点の数を 4×4 に変更した．生成部 G と識別部 D に対し学習係数を $\alpha = (0.5e-3, 0.5e-3)$，$\beta 1 = (0.9, 0.9)$，$\beta 2 = (0.999, 0.999)$ に設定し最適化を行い，学習回数は 71 回で実験を行った．

学習後の生成モデル G に対して入力 $\boldsymbol{z}_t$(t=1...,6) の検証のために以下の 6 つの条件で実験を行った．

1. 全ての $\boldsymbol{z}_t$ に標準正規分布からのサンプルベクトルを使用.
2. 全ての $\boldsymbol{z}_t$ に要素の値が 0 のベクトルを使用.
3. ある 1 つの $\boldsymbol{z}_t$ にのみ標準正規分布からのサンプルベクトルを使用，他の $\boldsymbol{z}_t$ には要素の値が 0 のベクトルを使用.
4. $\boldsymbol{z}_1$ にのみ平均 0, 標準偏差 10,000 の正規分布からのサンプルベクトルを使用，他の $\boldsymbol{z}_t$ には標準正規分布からのサンプルベクトルを使用.
5. $\boldsymbol{z}_3$ にのみ平均 0, 標準偏差 10,000 の正規分布からのサンプルベクトルを使用，他の $\boldsymbol{z}_t$ には標準正規分布からのサンプルベクトルを使用.
6. $\boldsymbol{z}_6$ にのみ平均 0, 標準偏差 10,000 の正規分布からのサンプルベクトルを使用，他の $\boldsymbol{z}_t$ には標準正規分布からのサンプルベクトルを使用.

条件 1 と条件 2 の実験結果を比較することで，各入力値の値がランダムである場合と，全て同じである場合の生成画像を変化について検証できる．また条件 3 の実験結果から，ある 1 ステップだけがランダムな入力値の場合，生成画像がどのように変化するか検証できる．更に条件 4〜条件 6 の実験結果を比較することで，出力結果に悪影響を及ぼす入力が最初のステップ，中間のステップあるいは最後のステップ与えられた場合に，生成画像がどのように変化するか検証できる．

3.3 実験結果と検証

以下に出力した画像サンプルを示す．図 4 が提案手法を用いた画像の生成結果であり，図 5 が敵対的ネットワークを用いた画像の生成結果である．図 6 は図 4 の画像サンプルの内，最上段にある 10 個の数字画像が，実際に各注視毎に段階的に作られていくところを表す．四角の赤枠が注視している領域を表し，四角の中心が注視の中心位置を示し，四角の大きさが注視の範囲を表している．枠の太さは各フィルタの注視領域に対するぼやけ具合を表している．左側から右側に掛けて，前の画像に部分対象を書き加えていくことで最終的に全体画像が生成されている．

図 7 はフィルタの格子点の数を 4×4 に減らした場合の生成画像である．図 4 に対し，出力される数字の種類は増えたが，形状が崩れた数字が生成されやすくなっている．

次に学習後の生成部 G について，各入力 $\boldsymbol{z}$ に対する生成画像の変化を示す．図 8 は上部が各入力 $\boldsymbol{z}_t$ を標準正規分布から生成した時の生成画像であり，下部が各入力 $\boldsymbol{z}_t$ の値が全て 0 の時の生成画像である．図 9 はある 1 つの $\boldsymbol{z}_t$ の入力にのみ標準正規分布からのサンプルベクトルを使用し，他の $\boldsymbol{z}_t$ には要素の値が 0 のベクトルを使用した場合の生成画像の結果である．上から順番にそれぞれ $t = 0, ..., 6$ についての結果を示している．

図 10 は上部が $\boldsymbol{z}_1$，中部が $\boldsymbol{z}_3$，下部が $\boldsymbol{z}_6$ の入力にのみ平均 0, 標準偏差 10,000 の正規分布からのサンプルベクトルを使用し，他の $\boldsymbol{z}_t$ には標準正規分布からのサンプルベクトルを使用した場合の生成画像の結果である．

3.4 考察

図 5 が一部ぼやけたような数字画像を生成しているのに対し，図 4 ではぼやけたように描かれた文字がほとんどない．注視によってまた図 7 では，出力される数字の種類がやや増えたものの，形状が歪んだ数字が生成されやすくなっている．フィルタの格子点の数が多い場合，数字の形状を保ったまま生成しやすくなり，少ない場合は他の数字との区別をしやすくなると考えられる．

今回の各実験では生成部 G の c_0 の初期値として，学習用データの周辺分布からのサンプリング値を用いた．この初期値の画像は MNIST の数字画像の中間的な状態を表している．この初期化により数字の大まかな形状と配置をあらかじめ学習することが出来，それにより数字画像をより高速かつ安定に学習することができた考えられる．

各入力の $\boldsymbol{z}_t$ の出力結果に与える影響についての検証結果では，図 8 と図 9 において前半のステップ ($t < 3$) で与えられる入力の影響が大きく，それ以降の入力の影響はほとんどない．

図 7: 格子サイズ 4 × 4 のガウシアンフィルタを使用した生成画像.

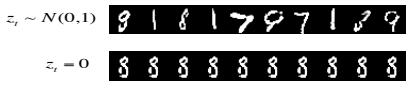

図 8: ランダムな潜在変数と固定した潜在変数での生成画像の比較.

前半のステップで入力された $\boldsymbol{z}_t$ の値が出力結果に与える影響が大きいのは，初期の入力で与えられた $\boldsymbol{z}_t$ の値が再帰ニューラルネットワークの内部で保持され，繰り返し用いられるためであると考えられる．後半の入力は内部で保持される期間の短いのでネットワーク全体に対する影響力が小さいと考えられる．また図 10 において，出力結果に悪影響を及ぼす入力に対しては，最終ステップで与えられた場合，生成画像の数字の形状はほぼ保たれており，他 2 つに対し変化の度合いが小さい．この事から再帰ニューラルネットワークを用いる場合，ネットワーク内部で保持されている情報の方が最終的な出力に対する影響が大きいと考えられる．

4. 結論

本研究では深層学習における画像生成の手法として，注視メカニズムによる敵対的画像生成を提案した．実験には MNIST の数字画像を学習データとして用い，その結果注視のメカニズムを用いることで，段階的に画像が作られ最終的に全体画像が生成されることを示し，また注視のメカニズムを適用したことで，人の目で見てより本物らしい画像が生成されることが分かった．またネットワークの各入力が出力に及ぼす影響について，注視の初期段階における入力が出力結果を左右し，後半の段階の入力は出力結果にほとんど影響を及ぼしていないことが分かった．

今後の課題として，今回は数字画像のデータを扱うのに留まったが，より注視の有効性が示せるように数字以外の自然画像や対象物が複数写った画像のデータセットについても本物らしい画像が生成できるよう取り組むことが挙げられる．また今後の実験では，より本物らしいかどうかを定量的に表す数値的な指標を用いた生成画像についての評価を行い，その評価指標に従った結果についても検討することが挙げられる．

参考文献

[Goodfellow 14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y.: Generative adversarial nets, in *Advances in Neural Information Processing Systems*, pp. 2672–2680 (2014)

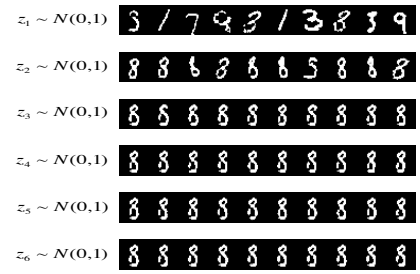

図 9: 潜在変数を部分的に固定した場合の生成結果の比較.

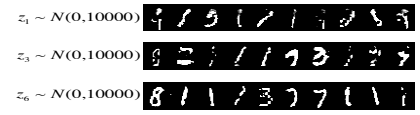

図 10: 不適当な潜在変数の分布に対する生成画像の比較.

[Gregor 15] Gregor, K., Danihelka, I., Graves, A., Rezende, D., and Wierstra, D.: DRAW: A Recurrent Neural Network For Image Generation, in *Proceedings of The 32nd International Conference on Machine Learning*, pp. 1462–1471 (2015)

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

[Kingma 15] Kingma, D. and Ba, J.: Adam: A method for stochastic optimization, in *Proceedings of the International Conference on Learning Representations* (2015)

[Radford 15] Radford, A., Metz, L., and Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, *arXiv preprint arXiv:1511.06434* (2015)