The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

大量の Twitter 画像を用いた Conditional Cycle GAN による食事写真カテゴリ変換

堀田 大地 Daichi Horita　成冨 志優 Shu Naritomi　丹野 良介 Ryosuke Tanno　下田 和 Wataru Shimoda　柳井 啓司 Keiji Yanai

*1電気通信大学 情報理工学部
The University of Electro-Commnications, Tokyo

This paper proposes “Food Image Transformation” based on the Conditional Cycle GAN (cCycle GAN) with a large-scale food image data collected from the Twitter Stream for more than five years. By the experiments, we showed that two hundred and thirty thousand food images with cCycle GAN enabled very natural food photo transform among ten kinds of typical Japanese foods: ramen noodle, curry rice, fried rice, gyodon, cold Chinese noodle, spaghetti with meat source, white rice, eel bown, and yakisoba.

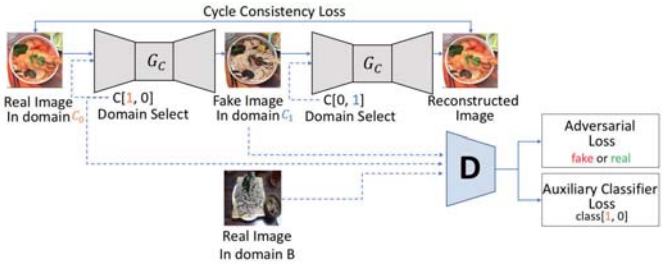

図 1: Conditional CycleGAN のネットワーク全体

1. はじめに

近年，生成モデルと深層学習を組合せた深層生成モデル Generative Adversarial Networks(GAN) が従来手法と比べてより本物らしい画像を生成できるとして注目を集めている．訓練データの分布に近似するよう最適化することで本物らしい画像の生成に成功している．GAN の研究において用いられるデータセットは CelebA データセットの顔画像や MNIST の数字文字画像，LSUN の居住画像など，ある程度パターンが限られる画像群が通常用いられる．また，最近では，[1] のように衣服画像へのデザイン転送タスクといった新しい課題を提案し，GAN や Neural Style Transfer のような深層学習技術を応用する研究がでてきている．一方で，本研究のような食事に限定した食事画像生成・変換に関する研究は未だ存在しないのが現状である．

本研究では，深層学習技術を用いて，自動的に食事画像を変換するという新しい問題に焦点を当てる．食事画像と変換先のカテゴリ情報を入力すると，リアルタイムに特定のカテゴリに変換された食事画像を生成することを目指す．深層学習による画像変換手法の 1 つである CycleGAN の手法を拡張し，1 つの変換ネットワークで複数のカテゴリへと変換可能とする conditional CycleGAN を用いた食事画像変換手法を提案し，10 種類 23 万枚の Twitter から収集した食事画像に適用することで，きわめて自然な 10 種類食事間での食事カテゴリの相互変換が可能であることを示す．

2. 関連研究

GAN は一様分布や正規分布などからノイズベクトル z をサンプリングするため，生成される画像のコントロールをすることができない．そこで，GAN の構造に条件付き信号 conditional vector を付与することで，条件付き確立分布を学習するモデルに拡張したものが cGAN である．一方で，cGAN には入力画像を潜在表現に落とし込む機構 (Encoder) が欠けているため，画像の変換は行うことができない．pix2pix は Adversarial Loss と ConvDeconvNet を組合せることで，画像のペア集合間の変換方法を学習することが可能となり，線画彩色や白黒画像のカラー化などの変換を学習させることができる．

[2] では学習データ間 X, Y の写像を学習する方法が提案された．通常の GAN で用いられる損失関数に再構築誤差である Cycle Consistency Loss を追加することで,「集合 X, Y に共通する構造を保って」変換する写像関数の学習に成功している．よって，本研究においても Cycle Consistency Loss による制約を設けることで,「集合 X, Y に共通する構造を保って」，つまりは，食事画像であるならば，食事の部分のみ別のカテゴリの食事に変換し，それ以外は，元の形状を保ったまま変換されることが可能になると考えた．

3. Conditional CycleGAN による画像変換

本節では，まず関連する画像変換技術である pip2pix [3], cycleGAN について説明し，その後，本研究で用いる Conditional CycleGAN (cCycleGAN) [2] について説明を行う．

3.0.1 pix2pix

pix2pix [3] は conditional GAN の一種であるが，通常の GAN では，一様分布や正規分布からサンプリングしたノイズベクトル z を Generator への入力とするが，pix2pix や後述する CycleGAN では，画像 x を Generator の入力とする点が大きく異なる点である．入力に用いていた乱数 z は直接サンプリングする代わりに Generator の複数の層に Dropout でノイズを加えるように代替されている．

pix2pix では，式 1 で表される cGAN の損失関数に加えて，より本物らしい画像を生成するために，式 2 の L1 正則化項の追加と Discriminator のベース構造に [4] で提案された PatchGAN を組合せた式 3 が最終的な pix2pix の損失関数と

なる．入力には変換前と変換後の画像のペアを必要とし，(変換元画像，変換先画像) or (変換元画像，Generator が生成した画像) のいずれのペアであるかを Discriminator に判断させるように学習する．

$$L_{cGAN}(G,D) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[\log D(x,y)] + \mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}[\log(1-D(x,G(x,z)))] \quad (1)$$

$$L_{L_1}(G) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}}[||y-G(x,z)||_1] \quad (2)$$

$$G^* = arg\min_G \max_D L_{GAN}(G,D) + \lambda L_{L_1}(G) \quad (3)$$

3.0.2 CycleGAN

pix2pix[3] では，変換前と変換後の画像の 1 対 1 ペアを必要とする制限があったが，CycleGAN[2] ではドメイン間の写像を学習できるように拡張することで，1 対 1 に対応せずとも学習が行えるのが特徴である．ここで，ドメイン X とドメイン Y があるとして，$X \to Y$ への写像を G，その逆写像 $Y \to X$ を F と定義する．また，入力が G によって生成された偽物の X か元の X のデータかを判別する D_Y，入力が Y によって生成された偽物の Y か元の Y のデータかを判別する D_X をそれぞれ定義する．この G, F, D_X, D_Y を式 4 と式 5 の 3 つの損失の和で表される式 6 を用いて学習する．式 4 は Vanilla GAN で用いられる Adversarial Loss そのままであるが，式 5 は Cycle Consistency Loss と呼ばれるもので，ドメイン X に属する x から生成された $\hat{Y}$ を再度，ドメイン X に属する $\hat{x}$ に戻しても元のドメイン X に一致するように制約をかけるものである．この Cycle Consistency Loss を小さくすることは，$G(F(x))$ により変換した結果がそれぞれ元のデータを再構築できるだけの情報を保持することを意味する．よって，学習に成功した場合は，$G(F(x))$ とした場合,「ドメイン X とドメイン Y に共通する構造を保ったまま，一方のドメインに属するデータをもう一方のドメインのデータに変換する」写像関数が得られることになる．

$$L_{GAN}(G,D_Y,X,Y) = \mathbb{E}_{\boldsymbol{y}\sim p_{data}(\boldsymbol{y})}[\log D_Y(y)] + \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\log(1-D_Y(G(x))] \quad (4)$$

$$L_{cyc}(G,F) = \mathbb{E}_{\boldsymbol{x}\sim p_{data}(x)}[||F(G(x))-x||_1] + \mathbb{E}_{\boldsymbol{y}\sim p_{data}(y)}[||G(F(y))-y||_1] \quad (5)$$

$$L(G,F,D_X,D_Y) = L_{GAN}(G,D_Y,X,Y) + L_{GAN}(F,D_X,Y,X) + \lambda L_{cyc}(G,F) \quad (6)$$

3.0.3 conditional CycleGAN

図 1 として conditional CycleGAN(cCycleGAN) の模式図を示す．CycleGAN [2] を conditional 化することで，1 つの Generator で複数のカテゴリへと変換可能とする conditional CycleGAN に拡張してある．CycleGAN の conditional 化には，[5] と同様に [6] で提案されている分類誤差項 Auxiliary Classifier Loss を Discriminator に追加することで実現する．本物か偽物かの判断をさせるだけでなく，Discriminator にどのカテゴリに属する画像かの識別も同時に学習させることで，複数のカテゴリに変換可能な Generator の学習を行った．こうすることで，Generator は単に Discriminator を欺くように画像を生成するだけでなく，Discriminator の識別エラーを最小限に抑えるように偽物のサンプルを生成できるようになる．つまり，各カテゴリのサンプルを生成できるように最適化されることを意味する．

よって，最終的な損失関数は，Adversarial Loss L_{adv} に式 7 で表される Cycle Consistency Loss と式 8，式 9 で表される Auxiliary Classifier Loss にそれぞれの重みバイアス項 λ_{ccl} 及び λ_{acl} を追加した式 10，式 11 を conditional CycleGAN の損失関数として用いた．

$$L_{ccl} = \mathbb{E}_{x,c,c'}[||x-G(G(x,c),c')||_1] \quad (7)$$

$$L_{acl}^{real} = \mathbb{E}[-\log D_{acl}(c'|x)] \quad (8)$$

$$L_{acl}^{fake} = \mathbb{E}_{x,c}[-\log D_{acl}(c|G(x,c))] \quad (9)$$

$$L_D = L_{adv} + \lambda_{acl} L_{acl}^{real} \quad (10)$$

$$L_G = L_{adv} + \lambda_{acl} L_{acl}^{fake} + \lambda_{ccl} L_{ccl} \quad (11)$$

4. 実験

4.1 学習データ

Cycle Consistency Loss を追加することで,「集合 X, Y に共通する構造を保って」変換することが可能である．そのため，学習データに「共通する構造」がある方が変換が上手くいくと推測される．よって今回は,「丼」という制約を設けて UECFOOD-100[7] の 100 カテゴリの食事の中から「丼」の構造をもつ 10 個のカテゴリを選出した．その 10 カテゴリについて高品質な食事画像の選別のために，2011 年より継続的に Twitter Stream より収集している食事画像データベース [8, 9] の中から UECFOOD-100[7] で学習した食事認識エンジンを用いて，各カテゴリ毎に認識精度が高い順にリランキングした結果から表 1 にある枚数分を学習データとした．この中で「ラーメン」のカテゴリに至ってはその種類の多様性が他のカテゴリと比べて高かったため（例えば,「二郎系のラーメン」は基本的に「丼」からはみ出るほどの具材が乗っているため，他のラーメンと比べて差が大きい．つまり,「共通する構造」が同カテゴリであるが，差が大きくなってしまい，学習が難しくなる恐れがある.），図 2 の処理を行った．8 万枚の「ラーメン」画像に対して，ImageNet で学習済みの VGG16 を特徴抽出器として用い，224x224x3(150,528 次元) を fc6 層 (4,096 次元) まで特徴量を圧縮して k-means により k 個のクラスタ (今回は，k=8 に設定した) に分割を行った．想定していた通り,「二郎系ラーメン」が大部分を占めるクラスタを得られたため，そのクラスタを除外した画像を「ラーメン」カテゴリの画像とした．全ての学習において，訓練 9 割，テスト 1 割となるように配分した．

4.2 学習モデル構造

conditional CycleGAN のネットワークは，基本的に CycleGAN [2] と同一である．変換ネットワーク (Generator) は [10] で提案された ConvDeconvNet の中間層に Residual Block を何層も積層する FastStyleNet の構造を用いて 256x256 の画像を学習に用いた．なお，conditional signal は one-hot vector で表現し，入力画像サイズにブロードキャストした後に，入力画像とチャネル方向に結合して，Generator に入力している．また，Discriminator には [4] で提案された PatchGAN を採用してある．重みの更新頻度は Discriminator を 5 回更新した後に Generator を 1 回更新するようにした．学習は NVIDIA Quadro P6000 を利用しバッチサイズ 32，最適化手法には Adam を用いて 20epoch 繰り返した．テスト時は 512x512 の画像を生成するようにした．

表 1: 学習データ

カテゴリ	学習枚数
冷やし中華	13,499
ミートスパゲティ	7,138
蕎麦	3,530
ラーメン	74,007
焼きそば	24,760
白飯	21,324
カレーライス	34,216
牛丼	18,396
うな重	5,329
炒飯	27,854
合計	230,053

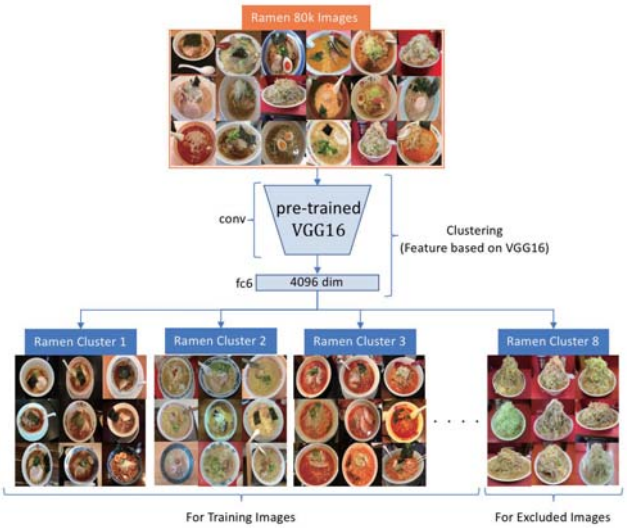

図 2: 多様性があるカテゴリに対するクラスタの構築

4.3 食事画像変換結果

本手法により変換した結果を図 3 に示す．最左列を入力画像として，最上部の 10 カテゴリのドメインへ同時に変換した例を示してある．食事が複数品目ある場合に対しても正確に食事領域のみ対象のドメインへと変換できていることがわかる．再構築誤差 Cycle Consistency Loss により「ドメイン X ドメイン Y に共通する構造を保ったまま，あるドメインに属するデータをもう一方のドメインのデータに変換する」写像関数の学習に成功し，「共通構造=丼，器，食器」の概念を Generator が獲得していることを意味する．また，分類誤差 Auxiliary Classifier Loss を導入することで Generator は単に Discriminator を欺くように画像を生成するだけでなく，Discriminator の分類エラーを最小限に抑えるように偽物のサンプルを生成できるようになり，各ドメインのサンプルを生成できるように最適化されることで，歪みや GAN に特有のブラーが掛かっていない高いクオリティで変換できていることがみてとれる．

4.4 学習に用いるデータ数の変化による変換結果のクオリティへの影響

学習に用いるデータ数が表 1 と比べて小規模な場合，変換結果のクオリティがどのように変化するかの考察を行った．学習に用いるデータセットは 3 種類あり，まとめると以下のようになる．「白飯」「冷やし中華」「牛丼」「カレー」の 4 カテゴリ

図 3: 食事画像変換結果

について，1 カテゴリあたりの画像枚数，総画像枚数がどのようなクオリティ変化を示すの考察を行った．

1. 1 カテゴリ 1 千枚．合計 1 万枚のデータセット．
2. 1 カテゴリ 1 万枚．合計 10 万枚のデータセット．
3. 表 1 の合計約 23 万枚のデータセット．

各条件により学習したモデルで変換した画像を総学習枚数が少ない順に左から並べたものを図 4，図 5 に示す．各カテゴリ千枚の比較的小規模なデータセットでも変換先ドメインの大域的特徴を捉えることには成功しているが，局所的にみると細かいディテールまでは再現できていないようにみえる．つまり，画像枚数が多ければ多いほど，大域的特徴に加えて局所的な特徴をもった細部の細かい部分まで正確に変換先のドメインに変換可能な写像関数の学習ができている結果となった．また，図 4 のカテゴリ「冷やし中華」の変換結果に着目すると，1 列目は 1 千枚，2 列目は 1 万枚，3 列目は表 1 にある通り，1.3 万枚と 2 列目と 3 列目で画像枚数は 3 千枚ほどしか変わらない．しかし，2 列目より 3 列目の方が細かい部分まで変換できていることがわかる．一方で，「冷やし中華」以外の画像枚数も考慮すると，2 列目は総学習枚数 10 万枚に対し，3 列目は 23 万枚の大規模データセットを用いている．他カテゴリの画像から得られた特徴も上手く変換結果に反映されていることがこの結果から伺えるが，これは，1 つの Generator で複数のカテゴリに変換可能にすることで，「食事変換」という共通特徴を Generator が獲得していることを意味する．つまり，1 つの生成器が複数のカテゴリへの変換を担うことで，画像枚数が少ない特定のカテゴリが存在した場合でも，どのカテゴリへも一定の質を保って変換することが可能としていることになる．

5. まとめ と 今後の課題

本研究では，深層学習技術を用いて，自動的に食事画像を生成・変換するという新しい問題に取り組み，CycleGAN の手法を拡張した conditional CycleGAN を用いることで，

1. 変換前と変換後で共通構造を保ったままの変換
2. 複数のカテゴリへの変換を行うことで，変換カテゴリの共通特徴の獲得による変換クオリティの向上

を実現し，実験により実際に高品質に食事画像が変換可能であることを示した．

今後の課題としては，現状，学習したモデルの有効性を示すために，主観的な定性評価しか行っていないため，他者による客観評価実験や変換した画像が期待するターゲットドメインへと変換できているかについて，食事画像分類問題を解くことで定量評価としたい．また，本研究で学習したモデルを用いてモバイルアプリとして実装する予定である．

謝辞 本研究は JSPS 科研費 15H05915, 17H01745, 17H05972, 17H06026, 17H06100 の助成を受けたものです．

参考文献

[1] S. Jiang and Y. Fu. Fashion Style Generator. In *Proc. of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017.

[2] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Proc. of IEEE International Conference on Computer Vision*, 2017.

[3] P. Isola, J. Zhu, T Zhou, and A. A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2017.

[4] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros. Context Encoders: Feature Learning by Inpainting. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[5] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. arXiv:1711.09020, 2017.

[6] A. Odena, C. Olah, and J. Shlens. Conditional Image Synthesis With Auxiliary Classifier GANs. In *Proc. of the 34th International Conference on Machine Learning*, 2017.

[7] Y. Matsuda, H. Hoashi, and K. Yanai. Recognition of Multiple-Food Images by Detecting Candidate Regions. In *Proc. of IEEE International Conference on Multimedia and Expo*, 2012.

[8] 柳井 啓司 河野 憲之. ラーメン vs カレー：2 年分のログデータと高速食事画像認識エンジンを用いた twitter 食事画像分析とデータセット自動構築. In **電子情報通信学会パターン認識・メディア理解研究会**（*PRMU*）, 2013.

[9] K. Yanai and Y. Kawano. Twitter food image mining and analysis for one hundred kinds of foods. In *Proc. of Pacifit-Rim Conference on Multimedia (PCM)*, 2014.

[10] J. Johnson, A. Alahi, and L.F. Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Proc. of European Conference on Computer Vision*, 2016.

図 4: 学習に用いるデータ枚数のクオリティへの影響結果 (1). 左から入力入力画像,「白飯」カテゴリへの 1 万枚学習モデルでの変換結果，10 万枚モデル結果, 23 枚万枚モデル結果,「冷やし中華」カテゴリへの 1 万枚モデル変換結果，10 万枚モデル結果, 23 枚万枚モデル結果.

図 5: 学習に用いるデータ枚数のクオリティへの影響結果 (2). 左から入力入力画像,「牛丼」の 1 万枚モデル結果，10 万枚モデル結果, 23 枚万枚モデル結果,「カレー」の 1 万枚結果，10 万枚モデル結果, 23 枚万枚モデル結果.