Food Category Transfer with Conditional Cycle GAN and a Large-scale Food Image Dataset*

ABSTRACT

This paper describes "Food Image Transformation" based on a Conditional Cycle GAN (cCycle GAN) with a large-scale food image data collected from the Twitter Stream. Conditional Cycle GAN is an extension of CycleGAN, which enables "Food Category Transfer" among ten kinds of foods mutually keeping the shape of a given food. By the experiments, we show that two hundred and thirty thousand food images with cCycle GAN enables very natural food category transfer among ten kinds of typical Japanese foods: ramen noodle, curry rice, fried rice, beef rice bowl, chilled noodle, spaghetti with meat source, white rice, eel bowl, and fried noodle.

KEYWORDS

Food Image Transformation, Food Category Transfer, Cycle GAN, Food Image Generation

1 INTRODUCTION

In recent years, Generative Adversarial Networks (GAN) is drawing a lot of attention. GAN can generate an image which looks like a real image. A human face image dataset such as CelebA and a numeric character image dataset such as MNIST have been used for training GAN as target domains. In addition, recently [3] proposed a new task which is a style transfer for clothes using GAN. On the other hand, there exists no work for a food image generation or transformation using GAN so far. In this work, we propose a novel paradigm on food image transformation to convert a given food image to another category of a food image automatically.

Our objective is to make a system which takes a food image and a food category to be transferred as inputs, then outputs a new food image which corresponds to the given food category. We propose a food category transfer method by extending CycleGAN which converts an image into another domain image. To generate realistic images, the number of training images is the important key. We have gathered 230,000 food images which consist of 10 kind of food categories from Twitter stream for food image transformation. We have been keeping gathering images from the Twitter stream for more than eight years, and we mined the images corresponding to any of the ten food categories to create a large-scale food photo dataset for food category transfer. We show that it enabled high quality mutual transformation on a food domain with conditional Cycle GAN (cCycle GAN). In addition, we show the number of the training images is important to get more realistic images.

*Food Category Transfer

2 RELATED WORK

With a standard GAN, we cannot control the category of generated images explicitly, since GAN uses only a noise vector v sampled from uniform distribution or normal distribution as a seed. On the other hand, with Conditional GAN (cGAN) which is an extension of GAN by adding cognitional inputs, we can control the category of generated images by providing a conditional vector in addition to a random noise vector. On the contrary, cGAN cannot transfer an image to an image so that the model does not have an encoder which converts an input image to hidden representation. Pix2Pix [2] is an extension of cGAN which takes an image as a conditional vector. In Pix2Pix, an Encoder-Decoder network is used instead of a generator which generates an image from a seed vector. Since an encoder-decoder network takes an image as an input and output an transformed image, image transformation can be done. To train the network of Pix2Pix, many paired samples of raw images and corresponding transformed images are required for training.

Zhu et al. proposed a method to train an image transformation network using unpaired training samples which consists of two domains of image samples such as color images and corresponding grey-scale images [9]. They introduced a cycle consistency loss for training, and successfully trained an image transformation network which transform an original-domain image to the other-domain image keeping rough shape structure unchanged. In this paper, we use a cycle consistency loss to train a model. In a food domain, with this loss function, we can transfer a food image to the other category of a food image keeping the original food image structure unchanged.

3 IMAGE TRANSFORMATION USING CONDITIONAL CYCLEGAN

In this section, we review two recent image transformation methods, Pix2Pix [2] and Cycle GAN [9], and then we describe a Conditional Cycle GAN we used in this paper.

3.1 Pix2Pix

Isola et al. proposed Pix2Pix [2] which is an extension of a conditional GAN. Before publishing this paper, only L2 mean square loss was used for training of an encoder-decoder-based image transformation network, which was unable to transform images between different domains clearly. In Pix2Pix, they proposed to use an adversarial loss in addition to conventional L1 loss function for training of a encoder-decoder

network. This can be regarded as image-conditioned version of conditional Generative Adversarial Network (cGAN). This enabled between-domain image transformation by CNN. For example, it can transform edge images into color drawings.

In Pix2Pix, they used Eq.1 as an adversarial loss, Eq.2 as a L1 normalization term. Eq.3 shows the loss function of Pix2Pix, which is minimized for training a generator and is maximized for training a discriminator. Pix2Pix needs paired samples which consists of an original image and the corresponding image which is transformed from a domain A to a domain B. Although this setting is possible for a pair of edge images and color drawings where an image of one domain is easily able to be converted into the corresponding image of the other domain, it is impossible for category conversion such as a pair of horse images and zebra images.

$$\begin{aligned} L_{cGAN}(G,D) &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[\log D(x,y)] + \\ &\quad \mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}[\log(1-D(x,G(x,z)))] \quad (1) \\ L_{L_1}(G) &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}}[||y-G(x,z)||_1] \quad (2) \\ G^* &= arg\min_G \max_D L_{GAN}(G,D) + \lambda L_{L_1}(G) \end{aligned}$$

3.2 CycleGAN

While Pix2Pix[2] needs many pairs of sample images of two domains, which are sometimes difficult to be prepared, Cycle GAN [9] solved this problem. It does need not paired samples but unpaired samples.

We denote two types of the domain space as X and Y. We represent the mapping $X \to Y$ as G and the inverse mapping as F. Discriminator for domain Y is D_Y, discriminator for domain X is D_X. The loss is defined as Eq.6 using Eq.3 and Eq.4. Eq.3 is general loss for Adversarial network. Eq.4 is called as a Cycle Consistency Loss. Here, we denote $\hat{y}(x)$ as generated image from x and $\hat{x}(\hat{y})$ as generated image. Cycle Consistency Loss constrains the value of x to be $\hat{x}(\hat{y})$. If we minimize this loss the converted results by $G(F(x))$ keeps information for reconstruction. Hence we can obtain a map which converts images belonging to the domain X to images belonging to the domain Y and converted images keep their original image structure.

$$\begin{aligned} L_{GAN}(G,D_Y,X,Y) &= \mathbb{E}_{\boldsymbol{y}\sim p_{data}(\boldsymbol{y})}[\log D_Y(y)] + \\ &\quad \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\log(1-D_Y(G(x))] \quad (3) \\ L_{cyc}(G,F) &= \mathbb{E}_{\boldsymbol{x}\sim p_{data}(x)}[||F(G(x))-x||_1] \quad (4) \\ &\quad + \mathbb{E}_{\boldsymbol{y}\sim p_{data}(y)}[||G(F(y))-y||_1] \\ &\quad (5) \end{aligned}$$

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

3.3 Conditional CycleGAN

We show the network of Conditional CycleGAN (cCycleGAN) in 1 which is an conditioned extension of CycleGAN. cCycleGAN can convert an image to an image which belongs to the selected category by adding a conditional input to the image transformation network of CycleGAN [9]. To use a conditional vector effectively, in cCycleGAN we added Auxiliary Classifier Loss [6] to the discriminator which is the same approach to [1]. The discriminator of cCycleGAN classifies not only real or fake but also category of images. By the discriminator, a multi-class generator can be trained. Finally the loss of cCycleGAN is represented by a following equation:

$$\begin{aligned} L_{ccl} &= \mathbb{E}_{x,c,c'}[||x-G(G(x,c),c^{'})||_1] \quad (7) \\ L_{acl}^{real} &= \mathbb{E}[-\log D_{acl}(c^{'}|x)] \quad (8) \\ L_{acl}^{fake} &= \mathbb{E}_{x,c}[-\log D_{acl}(c|G(x,c))] \quad (9) \\ L_D &= L_{adv} + \lambda_{acl} L_{acl}^{real} \quad (10) \\ L_G &= L_{adv} + \lambda_{acl} L_{acl}^{fake} + \lambda_{ccl} L_{ccl} \quad (11) \end{aligned}$$

where λ_{ccl} and λ_{acl} are bias of weight for Auxiliary classifier loss.

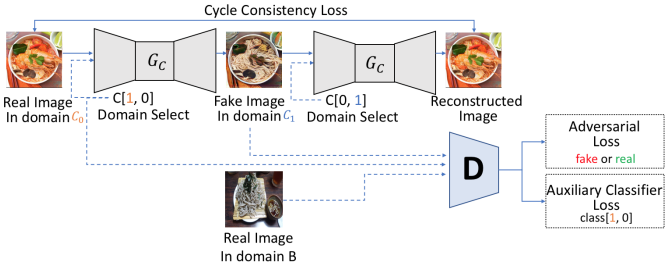

Figure 1: The architecture of the Conditional CycleGAN.

4 EXPERIMENTS

4.1 Dataset

By adding Cycle Consistency Loss, we can generate an image which keeps the original image structure. Therefore, in this experiments, we use a constrain to use images which have only same structure "bowl" so that corresponding structure prompts training of Cycle Consistency Loss. Actually we selected ten kind of categories related to "bowl" foods from UECFOOD-100 [5]. We gathered images from the large-scale food image dataset [8] which was created by mining food images from the twitter stream for more than eight years continuously. We sorted the iamges in the dataset [8] by using confidence scores obtained by a food classifier model which was trained with UECFOOD-100 dataset [5]. We show the ten bowl food categories and the number of selected images from re-ranked images in Table 1. In particular, in the case of "ramen" class, there is a specific ramen class which is a very big ramen called as "Jiro ramen" in Japanese. The ramen is far from usual ramen in terms of appearance. Mixing greatly different appearances in the same category may cause bad effect for training GAN. Therefore, we removed this class by clustering using a model pre-trained with ImageNet. Actually we used VGG16 model of compressed feature on fc6 layer. Note that we used k-means as a clustering method and the cluster number is 8. We removed the images belong to the big "Jiro" ramen cluster. We separated all the selected images into a train set and a test set dis-jointly. The ratio of the train set is 90% and the ratio of the test set is 10% retarding the total amount of ten kinds of the bowl food images.

Table 1: training data

category	image number
chilled noodles	13,499
meat spaghetti	7,138
buckwheat noodle 3,530	
ramen	74,007
fried noodles	24,760
white rice	21,324
curry rice	34,216
beef bowl	18,396
eel bowl	5,329
fried rice	27,854
TOTAL	230,053

4.2 Network and training setting

We followed [9] a network of conditional CycleGAN. The generator is the same to FastStyleNet [4] which is added several Residual block to Conv-Deconv Network. The input image size is 256×256. As a conditional vector, we use a one-hot vector. After broadcasting the conditional vector to input image size, we concatenate it with an input image in the middle of the encoder part. As a discriminator, we used PatchGAN [7]. After updating the discriminator five times, we update the generator one time. We used NVIDIA Quadro P6000 for training, bath size is 32, optimization method is Adam and iteration epoch is 20. On testing, we generate images with 512×512 resolution.

4.3 Results of food image transformation

We show the results by the proposed method in Fig.2. The left end image is the input image and other 10 images are the transformed images of each of the ten categories, respectively. Our proposed method can transform one certain category of an input to any of the other ten food categories clearly. We transformed given food images to the other food categories of images with keeping shape structure the Cycle Consistency Loss. This means that the generator trained the concept of "bowl". In addition, the generator generated an image which did not only fool the discriminator but also minimized the classification error of discriminator by Auxiliary Classifier Loss. We consider that Auxiliary Classifier Loss is also helpful for generating higher quality image than usual GAN. The images generated by using Auxiliary Classifier loss do not have blur which is frequently appeared if we use a simple GAN model. Note that additional results can be see at https://negi111111.github.io/FoodTransferProjectHP/.

4.4 Relation between image quality and the number of training images

We show the food image transformation results, when we used smaller dataset for training the model. Here, we prepare following three types of the subset of dataset.

(1) 1000 image per category : 10,000 images.
(2) 10000 image per category : 100,000 images.
(3) The number of training images follows Table1 : About 230,000 images.

In Fig.3 and Fig.4, we show the results which are obtained from the model trained with the different number of image. The leftmost images are the input images, and remaining six images are generated images. The transformed images are separated into two blocks by food categories which was used for the conditional vector. In each block, we used 10,000 images for the first column, 100,000 images in the second column, and 230,000 images in the third column, for training, respectively. The generated image quality becomes better as increasing training image number. Though we obtain not bad results by the model trained with small training set, the detail is not reconstructed. In Fig.3 the results which are transformed to "chilled noodle" category are shown on the second block. As shown on the Table1, there is a small margin on training image number between subset of second row and third row on "chilled noodle" category. However, the third column results of "chilled noodle" category shows higher quality than second column results clearly. We expect that the generator learned additional information from other category domain, then the generated image quality become high with small number of training images.

Figure 3: The leftmost image is input images. The remaining six images are separated into two blocks. The left blocks show the results of "white rice" and the right blocks show results of "chilled noodle". In the each block, from left to right, we show the generated images trained with with 10,000 images, 100,000 images and 230,000 images, respectively.

5 CONCLUSIONS

In this paper, we tackled a novel paradigm to transform a food image to another category of a food image automatically using Convolutional Neural Network. We have achieved the following results by adapting conditional CycleGAN which is an extended version of CycleGAN.

(1) A food category transfer keeping the shape structure unchanged.
(2) Improvement on the quality of food image transformation by mining a large number of training samples of corresponding categories from the Twitter Stream.

As a future work, we plan to evaluate our method quantitatively, although we show only qualitative results in this paper. We need objective experiments by defining evaluation

Figure 2: The leftmost images are input images, and the rest images are generated images with each of the ten categories.

Figure 4: The leftmost image is input images. The remaining six images are separated into two blocks. The left blocks show the results of "beef bowl" and the right blocks show results of "curry rice". In the each block, from left to right, we show the generated images trained with with 10,000 images, 100,000 images and 230,000 images, respectively.

scores using classification accuracy on generated images such like inception score.

REFERENCES

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

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

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

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

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

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

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

[8] K. Yanai and Y. Kawano. 2014. Twitter Food Image Mining and Analysis for One Hundred Kinds of Foods. In *Proc. of Pacifit-Rim Conference on Multimedia (PCM).*

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