Food Category Transfer with Conditional CycleGAN and a Large-scale Food Image Dataset

Daichi Horita Ryosuke Tanno Wataru Shimoda Keiji Yanai
The University of Electro-Communications, Tokyo

ABSTRACT

This paper describes "food image transformation" based on a conditional cycleGAN[A3] (cCycleGAN) with a large-scale food image data collected from the Twitter stream. A cCycleGAN is an extension of CycleGAN, which enables "food category transfer" among 10 types of foods and retain the shape of a given food. We experimentally show that 200 and 30,000 food images with the cCycleGAN enable a very natural food category transfer among 10 types 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, CycleGAN, Food Image Generation

1 INTRODUCTION

In recent years, generative adversarial networks (GANs) have attracted particular attention. A GAN can generate an image that resembles 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 the GAN as the target domains. In addition, recently [3] proposed a new task of style transfer for clothes using GAN. Meanwhile, none have reported food image generation or transformation using the GAN thus far. In this work, we propose a novel paradigm on food image transformation to convert a given food image to another category of food image automatically.

Our objective is to create a system that takes a food image and a food category to be transferred as inputs, and subsequently outputs a new food image that corresponds to the given food category. We propose a food category transfer method by extending the CycleGAN, which converts an image into another domain image. To generate realistic images, the number of training images is key. We gathered 230,000 food images consisting of 10 types of food categories from Twitter streams for the food image transformation. We continuously gathered images from the Twitter stream for more than eight years, and we mined the images corresponding to any of the 10 food categories to create a large-scale food photo dataset for the food category transfer. We show that it enables high-quality mutual transformation on a food domain with a conditional CycleGAN (cCycleGAN). In addition, we show that the number of training images is important to obtain more realistic images.

CEA/MADiMa'18, July 15, 2018, Mässvägen, Stockholm, Sweden

ACM ISBN 978-1-4503-6537-6/18/07…$15.00
https://doi.org/10.1145/3230519.3230597

2 RELATED WORK

With the standard GAN, we cannot control the category of the generated images explicitly because the GAN uses only a noise vector v sampled from a uniform distribution or normal distribution as a seed. Meanwhile, 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, the cGAN cannot transfer an image to another image because the model does not have an encoder that converts an input image to a hidden representation. Pix2Pix [2] is an extension of the cGAN, which uses an image as a conditional vector. In Pix2Pix, an encoder–decoder network is used instead of a generator that generates an image from a seed vector. Because an encoder–decoder network uses an image as an input and outputs a transformed image, image transformation can be performed. To train the Pix2Pix network, many paired samples of raw images and the corresponding transformed images are required for training.

Zhu et al. proposed a method to train an image transformation network using unpaired training samples consisting two domains of image samples such as color images and the corresponding grayscale images [9]. They introduced a cycle consistency loss for training, and successfully trained an image transformation network that transforms an original-domain image to the other-domain image while retaining the rough shape structure. We herein use a cycle consistency loss to train a model. With this loss function, in the food domain, we can transfer a food image to the other food image category while retaining the original food image structure.

3 IMAGE TRANSFORMATION USING CONDITIONAL CYCLEGAN

In this section, we review two recent image transformation methods, Pix2Pix [2] and Cycle GAN [9]. Subsequently, we describe the cCycle GAN that we used in this study.

3.1 Pix2Pix

Isola et al. proposed Pix2Pix [2], which is an extension of the conditional GAN. Prior to this paper, only the L2 mean square loss was used for training 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 the conventional L1 loss function for training an encoder–decoder

network. This can be regarded as an image-conditioned version of the cGAN. This enabled a between-domain image transformation by the CNN. For example, it can transform edge images into color drawings.

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

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

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

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

3.2 CycleGAN

While Pix2Pix[2] requires many pairs of sample images of two domains, which are sometimes difficult to be prepared, the Cycle GAN [9] has solved this problem. It requires only unpaired samples.

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

$$L_{GAN}(G,D_Y,X,Y) = \mathbb{E}_{\boldsymbol{y}\sim p_{data}(\boldsymbol{y})}[\log D_Y(y)] + \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\log(1-D_Y(G(x))] \quad (3)$$

$$L_{cyc}(G,F) = \mathbb{E}_{\boldsymbol{x}\sim p_{data}(x)}[||F(G(x))-x||_1] + \mathbb{E}_{\boldsymbol{y}\sim p_{data}(y)}[||G(F(y))-y||_1] \quad (4)(5)$$

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

3.3 Conditional CycleGAN

We show the network of the cCycleGAN in Fig.1, which is a conditioned extension of the CycleGAN. The cCycleGAN can convert an image to another image that belongs to the selected category by adding a conditional input to the image transformation network of the CycleGAN [9]. To use a conditional vector effectively, in the cCycleGAN, we added an auxiliary classifier Loss [6] to the discriminator, similar to the approach of [1]. The discriminator of the cCycleGAN classifies not only a real or fake but also the category of the images. By the discriminator, a multiclass generator can be trained. Finally, the loss of cCycleGAN is represented by the following equation:

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

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

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

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

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

where λ_{ccl} and λ_{acl} are the weight biases for the auxiliary classifier loss.

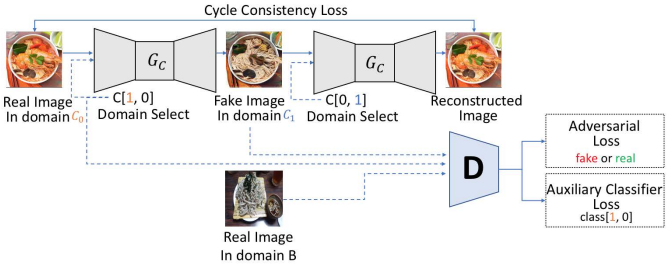

Figure 1: Architecture of the Conditional CycleGAN.

4 EXPERIMENTS

4.1 Dataset

By adding the cycle consistency loss, we can generate an image that retains the original image structure. Therefore, in this experiment, we imposed a constraint that uses images that have only the structure "bowl," such that the corresponding structure prompts the training of the cycle consistency loss. In fact, we selected 10 types of categories related to "bowl" foods from UECFOOD-100 [5]. We gathered the 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 images in the dataset [8] using the confidence scores obtained from a food classifier model, which was trained with the UECFOOD-100 dataset [5]. We show the 10 bowl food categories and the number of selected images from the re-ranked images in Table 1. In particular, for the "ramen" class, a specific ramen class that is a very big ramen exists, called "Jiro ramen" in Japanese. The ramen is far from the typical ramen in terms of appearance. Mixing significantly different appearances in the same category may negatively affect the GAN training. Therefore, we removed this class by clustering using a model pretrained with ImageNet. In fact, we used the VGG16 model of the compressed feature on the fc6 layer. It is noteworthy that we used k-means as a clustering method and the cluster number is eight. We removed the images belonging to the big "Jiro" ramen cluster. We separated all the selected images into a train set and a test set. The ratios of the train set and test set are 90% and 10%,

respectively, retarding the total amount of the 10 types of 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] for the network of the cCycleGAN. The generator is the same as FastStyleNet [4], which is added to several residual blocks of the conv–deconv network. The input image size is 256×256. As a conditional vector, we used a one-hot vector. After broadcasting the conditional vector to input the image size, we concatenated 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 updated the generator once. We used NVIDIA Quadro P6000 for the training, in which the bath size is 32, the optimization method is Adam, and the iteration epoch is 20. During testing, we generated images with 512×512 resolution.

4.3 Results of Food Image Transformation

We show the results by the proposed method in Fig.2. The leftmost image is the input image and the other 10 images are the transformed images of each of the 10 categories, respectively. Our proposed method can clearly transform one certain category of an input to any of the other 10 food categories. We transformed the given food images to the other food categories of images while retaining the shape structure of the cycle consistency loss. This means that the generator trained the concept of "bowl." In addition, the generator generated an image that did not only fool the discriminator but also minimized the classification error of the discriminator by the auxiliary classifier loss. We consider that the auxiliary classifier loss is also beneficial for generating higher quality images than the typical GAN. The images generated using the auxiliary classifier loss do not contain blurs that frequently appear in a simple GAN model. Additional results are shown 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 a smaller dataset for training the model. Here, we prepare the following three types of subsets of the dataset.

(1) 1000 image per category: 10,000 images.
(2) 10000 image per category: 100,000 images.
(3) All the images shown in Table1: 230,053 images.

In Fig.3 and Fig.4, we show the results obtained from the model trained with different number of images. The leftmost images are the input images, and the remaining six images are generated images. The transformed images are separated into two blocks by the food categories 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. The generated image quality becomes better as the number of training images increases. Although we obtained acceptable results by the model trained with a small training set, the details are not reconstructed. In Fig.3, the results that are transformed to the "chilled noodle" category are shown in the second block. As shown in Table1, a small margin on training image number exists between the subset of the second row and third row in the "chilled noodle" category. However, the third column results of the "chilled noodle" category shows higher quality than the second column results. We suspect that the generator learned additional information from the other category domain; consequently, the generated image quality becomes better with a small number of training images.

Figure 3: The leftmost images are the 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 the results of "chilled noodle." In each block, from left to right, we show the generated images trained with 10,000 images, 100,000 images, and 230,000 images, respectively.

5 CONCLUSIONS

Herein, we proposed a novel paradigm to transform a food image to another category of food image automatically using the convolutional neural network. We achieved the following results by adapting the cCycleGAN, which is an extended version of the CycleGAN.

Figure 2: The leftmost images are the input images, and the remaining images are generated images with each of the 10 categories.

Figure 4: The leftmost images are the 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 the results of "curry rice." In 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.

(1) A food category transfer that retains the shape structure.
(2) Improvement in the quality of food image transformation by mining a large number of training samples of the corresponding categories from the Twitter stream.

As future work, we plan to evaluate our method quantitatively as an extension to our qualitative results herein. Objective experiments are required that involve evaluation scores using the classification accuracy of the generated images such as the inception score.

Acknowledgements: This work was supported by JSPS KAKENHI Grant Number 15H05915, 17H01745, 17H05972, 17H06026 and 17H06100.

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.*