Ramen Spoon Eraser : CNN-based photo transformation for improving attractiveness of ramen photos

Daichi Horita, Jaehyeong Cho, Takumi Ege, Keiji Yanai
Department of Informatics, The University of Electro-Communications, Tokyo
{horita-d,cho,ege-t,yanai}@mm.inf.uec.ac.jp

ABSTRACT

In recent years, a large number of food photos are being posted globally on SNS. To obtain many views or "likes", attractive photos should be posted. However, some casual foods are served with utensils on a plate or a bowl at restaurants, which spoils attractiveness of meal photos. Especially in Japan where ramen noodle is the most popular casual food, ramen is usually served with a ramen spoon in a ramen bowl in a ramen noodle shop. This is a big problem for SNS photographers, because a ramen spoon soaked in a ramen bowl extremely degrades the appearance of ramen photos. Then, in this paper, we propose an application called "ramen spoon eraser" that erases a spoon from ramen photos with spoons using a CNN-based Image-to-Image translation network. In this application, it is possible to automatically erase ramen spoons from ramen photos, which extremely improve the attractiveness of ramen photos. In the experiment, we train models in two ways as CNN-based Image-to-Image translation networks with the dataset consisting of ramen images with / without spoons collected from the Web.

KEYWORDS

DeepLearning, CNN, Image-to-Image transformation

ACM Reference Format:
Daichi Horita, Jaehyeong Cho, Takumi Ege, Keiji Yanai. 2018. Ramen Spoon Eraser :, CNN-based photo transformation, for improving attractiveness of ramen photos. In *VRST 2018: 24th ACM Symposium on Virtual Reality Software and Technology (VRST '18), November 28-December 1, 2018, Tokyo, Japan.* ACM, USA, Article 4, 2 pages. https://doi.org/10.1145/3281505.3281622

1 INTRODUCTION

Nowadays, in social media such as Twitter and Instagram, a large number of images are being posted, including many food images. In the case of food images, utensils such as spoons and forks often cover the food, so that the attractiveness of food images may decrease. Especially in ramen images, the spoon are often provided on the soup of a ramen, it is hide ramen noodles and soup (Figure 1). And even if you try to remove a spoon for photography, since spoons are often soaked in soup, it is hesitant to put the spoon out of the plate feared that the table becomes dirty.

In the image recognition field, the CNN-based method achieves a great improvement [3] in some major tasks, and high-accurate image recognition is enabled. In addition, high-quality image generation and transforming are possible with Generative Adversarial Networks (GAN) [1]. For example, pix2pix [2] can perform high-quality image conversion by learning a pair image.

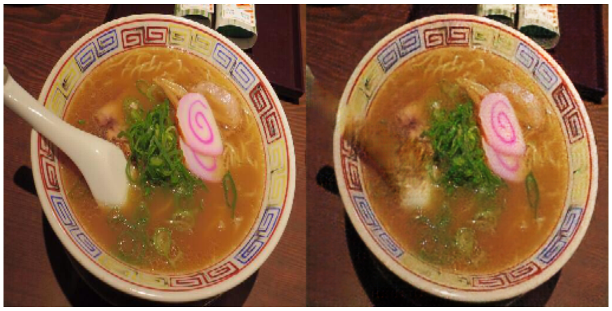

Figure 1: The left ramen image is partly hidden because of the spoon. In comparison, the right ramen image looks cleaner and more attractive since there is no spoon.

In this paper, we propose an application to erase spoons from given ramen images using CNN based Image-to-Image translation network. In the network, we transform the ramen of the input image to a spoon-erased ramen, so it is possible to obtain a more attractive ramen image.

Major contributions of this paper can be summarized as follows:

- We use the Image-to-Image translation network to erase the spoon from the ramen image and generate a more attractive ramen image.
- We compare Mean Squared Error (MSE) loss with pix2pix for the problem.

2 PROPOSE METHOD

In this section, we present a description of our approaches. The architecture of image translation with CNN can obtain image feature map by inputting image into encoder. Then, decode the feature map to generate an image. Ronneberger et al. propose U-Net [4] which has the encoder-to-decoder structure shown in Figure 2 and we use this architecture. In Section 2.1, we present the method we use Mean Squared Error(MSE) loss as L2 loss function. In Section 2.2, we present the method we use L1 loss and adversarial loss.

VRST '18, November 28-December 1, 2018, Tokyo, Japan

ACM ISBN 978-1-4503-6086-9/18/11.
https://doi.org/10.1145/3281505.3281622

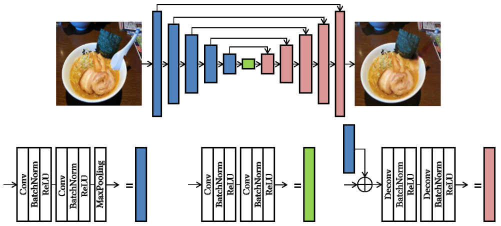

Figure 2: Architecture of U-Net [4] of the convolution-to-deconvolution structure. In comparison [4], Batch Normalization is introduced after convolutions and deconvolutions.

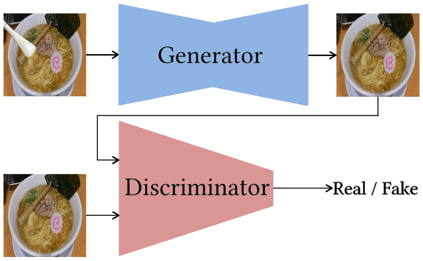

Figure 3: Schematic diagram of pix2pix [2]. In GAN training, the generator is trained to fool the discriminator trying to distinguish between fake generated and real samples.

2.1 Mean Squared Error (MSE) Loss

Based on [4], we use Batch Normalization after the convolutions and the deconvolutions. In training phase, we use Adam with learning rate of 10^{-4} and MSE loss.

2.2 Pix2pix

Isola et al. propose pix2pix [2] which is an extension of a conditional GAN shown in Figure 3. In GAN [1], noise vector z sampled from uniform distribution and normal distribution is input to the generator, but [2] uses image x as input, random number z is not sampled directly but is added by the dropout at multiple layers of the generator.

In [2], they used Eq.1 as an adversarial loss and Eq.2 as a L1 normalization term. Eq.3 shows the loss function of pix2pix, which is minimized for training the generator and is maximized for training the discriminator.

$$\begin{aligned}\mathcal{L}_{cGAN}(G,D) &= \mathbb{E}_{x,y}\left[\log D(x,y)\right] + \\ &\quad \mathbb{E}_{x,y}\left[\log(1 - D(x,G(x,z)))\right]. \quad (1)\\ \mathcal{L}_{L_1}(G) &= \mathbb{E}_{x,y,z}\left[\|y - G(x,z)\|_1\right]. \quad (2)\\ G^* &= \arg\min_G \max_D \mathcal{L}_{GAN}(G,D) + \lambda\mathcal{L}_{L_1}(G). \quad (3)\end{aligned}$$

We use the same architecture as in 2.1, adding the generator and the discriminator to the loss function Eq.3.

3 EXPERIMENTAL RESULTS

We implement MSE loss and pix2pix [2]. To erase spoons, we collect the training and testing data as presented in Section 3.1. In Section 3.2, we present the comparison between the MSE loss and pix2pix.

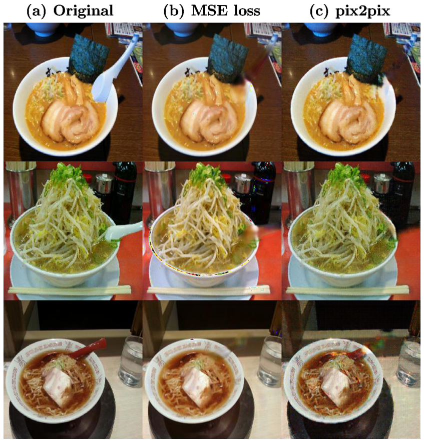

Figure 4: Some results of different ramen images generated by MSE loss and pix2pix.

3.1 Datasets

We crawl ramen images from the Web, and prepare 6000, 1000 images without / with spoon images. In addition, in order to create a pair image of a ramen, an image with a spoon is pasted on the collected an image without a spoon. We obtain spoon images by cutting out spoons from the collected images with spoons. In this way, we get paired ramen images and use them for training. In testing, we use images with spoons.

3.2 Results

Qualitative results are presented in Figure 4.

The network trained by MSE loss in Figure 4 (b) can erase spoons, though the results are blurred so that MSE loss cannot generate completely natural pixels.

On the contrary, the network trained by pix2pix can generate the natural spoon erased images as shown in Figure 4 (c). The adversarial loss is useful to avoid this complex transforming. We consider that the failure case in the Figure 4 (c) of the third line is caused by lacking training data for the red spoons.

4 CONCLUSION

We have presented an application that transforms into non-spooned ramen images with CNN. With this application, we make ramen images more attractive. In the future, we hope to apply our models to diverse utensils and transform to make it more attractive.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *NIPS*.

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. *CVPR* (2017).

[3] A. Krizhevsky, I. Sutskever, and G. E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proc.of Advances in Neural Information Processing Systems 25*. 1097–1105.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *CoRR* (2015).