U-net을 사용한 자연 경관 흑백 이미지의 색채화

Tram-Tran Nguyen Quynh, 김수형*, Nhu-Tai Do
광주광역시 전남대학교 전자컴퓨터공학과
*교신 저자: shkim@chonnam.ac.kr

Colorization of Natural Scene Image by Using a U-net

Tram-Tran Nguyen-Quynh, Soo-Hyung Kim, Nhu-Tai Do
School of Electronics and Computer Engineering, Chonnam National University, Gwangju

요 약

In this paper, we present the scene-context classification for the fully colorization image problem. The scene-context helps our model to exploit the global information context. From there, our model can transfer learning the experience from scene domain to colorization domain. Moreover, we apply the weighted instance classification for the uncertainty. It helps not only labeling smoothing on classification scene but also discovering the relations among scene labels. The experiments training in the Coco-Stuff dataset show that our results are very encouraging.

Keyword: image colorization, soft-encoding, u-net, scene-context classification

1. Introduction

We tend to simplify the way that an image will be full automatically colorized with the machines. Instead of using image-editing software and taking a lot of time to cover colors to a gray image, current studies give many methods to predict colors for an image that only depends on the context of its gray channel [1]. Generally, the system uses deep learning to learn the gray channel and the corresponding color channel of each color image in a large database of pictures before deciding on predicted colors. However, besides colors contrast, naturalness are also key factors influencing on the quality of an image [2]. Thus, naturalness elements also need to be studied to integrate into the system.

Most recent studies have been resolved based on deep convolution neural networks that have a powerful ability for learning. Typically, the authors in [3] built their model, which fuses two convolutional neural network paths to combine information of small image patches and global features of that image. The benefit of such a merger can transfer the scene style into another grayscale image. The method of [3] is completely automatic and focuses on the semantics of an image, but the dataset is only limited in categories about landscapes. Similarly, Larsson et al. in [4] also integrate semantic features of the scene into their model. In addition, to extend dataset for training process, they pre-train a network on ImageNet. In lieu of pre-training on ImageNet, [5] directly trains on the ImageNet. They use a huge dataset and propound the classification formula with 313 color bins for each pixel of the image. It the main reason to encourage the rare colors and tackle the desaturated and grayish problem in the fully automatic colorization.

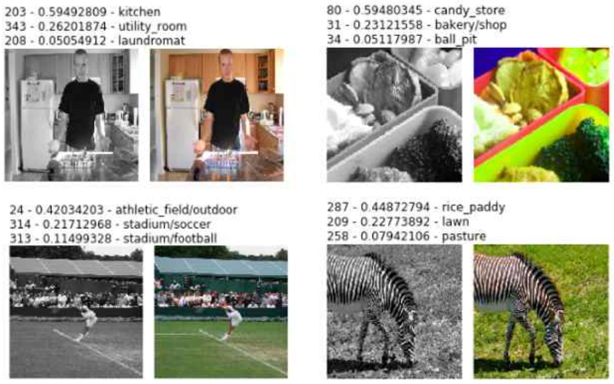
203 - 0.59492809 - kitchen
343 - 0.26201874 - utility_room
208 - 0.05054912 - laundromat
80 - 0.59480345 - candy_store
31 - 0.23121558 - bakery/shop
34 - 0.05117987 - ball_pit
24 - 0.42034203 - athletic_field/outdoor
314 - 0.21712968 - stadium/soccer
313 - 0.11499328 - stadium/football
287 - 0.44872794 - rice_paddy
209 - 0.22773892 - lawn
258 - 0.07942106 - pasture

그림 1 Scene-context colorization

To colorize the grayscale image more efficiently, we recognize the scene-context in the image plays an important role. The models of human cognition hold and process the information especially the color information based on previous experiences. A scene brings the observers some basic attributes such as indoor/outdoor, the places of images, etc. In Figure 1, the top-left image is a scene indoor in the kitchen room. Besides, the right-left image is the out-door scene in the football field. It brings about different color sensations based on references from the observer's experience. Moreover, the observers also determine what the objects are in the image and the reference object models. In the top-right image in Figure 1 describes the things in the food tray and the bottom-right image reference to a zebra eating grass on the field.

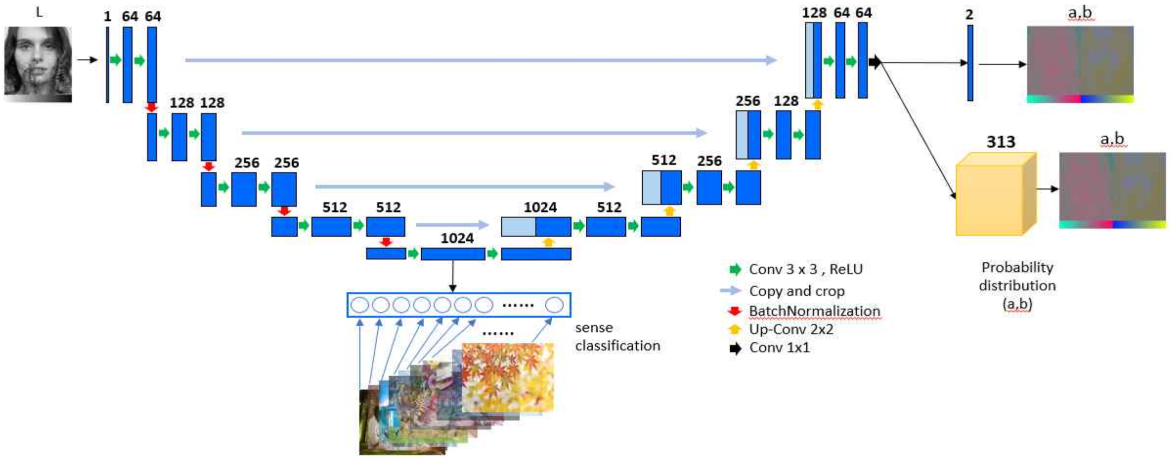

그림 2 Scene-context Image Colorization U-Net

So, we propose the colorized image solutions emphasizing the scene-context and the uncertainty in the scene classification. We use U-Net as the basic architecture with a classification branch for determining the kind of scene, pixel-wise classification branch to output the soft-encoding of the 2D color histogram from ab channel, and the regression branch for ab channel. Multi-task learning helps our model to integrate the global semantic information from the scene classification (indoor/outdoor, kitchen rooms/football field, etc.) under the uncertainty by the weighted instance classification. The soft encoding of the 2D color histogram from Zhang et al. [5] helps our model to encourage the rare colors, and keeps balance with the accuracy from the content by the regression branch.

We describe our proposed method in Section 2. Then we show our experiments and discussion in Section 3. Finally, we conclude our research and suggest works for further improvement in Section 4.

2. Proposed Method

2.1 Method overview

The overview of our proposed method is described in Figure 2. Initially, all images are converted to CIE Lab color space that includes three values, L represents lightness, a* axis represents green-red component, and blue-yellow component belongs to b* axis. The system receives input $I_L \in R^{h\times w\times 1}$ of the image. Our network is built as in Figure 2 based on U-Net [6] in which the contracting path encodes the lightness channel to the color features, and the expanding path decodes the color features into the ab channel $I_{ab} \in R^{h\times w\times 2}$ for regression branch and the probability distribution $Z_{ab} \in R^{h\times w\times n} \in [0,1]$ of the 2D color histogram in ab channels of CIE Lab color space, where n is the number of bins for the encoding the 2D color space [5]. Between two main paths, we embed one more branch to classify scene under the uncertainty using the weighted instance classification $Y_{class} \in R^{h\times w\times c} \in [0,1]$, where c is the number of the scene types. This branch plays the role of adding the weighted scene classification from the pre-trained path. We think that the scene classification could support the system to exploit the naturalness of images and the global semantic matters during the training process.

2.2 The uncertainty scene-context classification

Rafael et al. [7] shows the important improvement in generalization and learning speed of the network when using the label smoothing. It helps the network to prevent the over-confident in the classification problem. The traditional cross-entropy classification needs the one-hot vector only contains "1" for the correct class and "0" for the incorrect class. With the label smoothing, we replace the one-hot encoded label y_{hot} with the mixture of y_{hot} and the uniform distribution:

$$y_{smooth} = (1-\alpha)y_{hot} + \frac{\alpha}{c} \qquad (1)$$

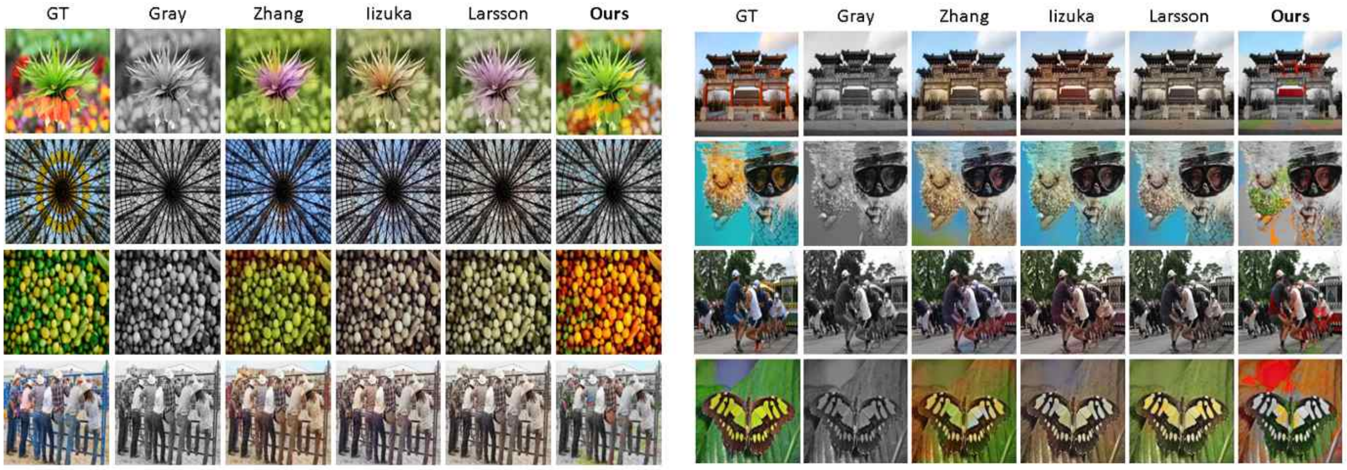

그림 3 (Left) The best cases and (Right) the worst cases in our model.

Where c is the number of label classes, and is the smoothing factor.

We use the label smoothing technique as the uncertainty from transfer knowledge from the scene domain network to our colorization model. By using pre-trained weight from Places [8], we extract the top-5 probabilities for every training image and normalize them. Moreover, the normalize probabilities shows the relation among class labels to help the model in generalization during training process.

2.3 The mixture of the regression and classification branch

In the previous work [9], we use the regression for ab channel and the classification for the quantized color encoding of ab channel independently. We recognize the quantized color encoding from Zhang et al. [5] nearly same as the 2D color histogram from Larsson et al. [4].

Both approach to the problem by replacing the predicting exactly the ground-truth color as a regression from softly predicting its color histogram. The main advantages of this approach are to tackle the over-confident in colorization by only trying to generate the plausible colorization most suitable from the data experience.

In this paper, we integrate the regression branch with the classification branch to take advantage of the flexible in the generation of the plausible colors with the accuracy in the regression.

3. Experimental Results

3.1 Datasets

COCO-Stuff dataset is a large-scale object detection, segmentation, and captioning dataset. It involves 118.000 images for training and 5.000 images for validation set in which includes 172 classes containing 80 thing classes, 91 stuff classes and 1 class unlabeled. We make input by converting images to grayscale images and rescale each image to 224×224.

Places365-Standard is the latest subset of Places dataset [8] with train set containing about 1.8 million images from the 365 different classes of scene/location with the image number per class from 3,068 to 5,000. We use the pre-trained weight from Place dataset to help our model to exploit the global semantic matters in the image. Our pre-trained scene network is the standard weights from the paper authors supply using the VGG16 architecture.

3.2 Environment, training process and metric evaluation

We used Keras Tensorflow 2.0 to build the model. For data augmentation, we used random contrast, brightness; random horizontal flip, rotate, scale and translation. We trained the model with two steps: the first step training by step decay 0.95 at every 5 epochs with learning rate 0.001, and after that fine tuning again using Cycle Learning Rate with range in [0.0008, 0.0002] in the period 8 epoch. We used Adam, SGD and RMSProp optimization for training and recognized the Adam and RMSProp giving the better result.

In this paper, we used PSNR, SSIM, and L2 distance for ab channel to evaluate the quantitative metric. But the PSNR metric cannot reflect the visual quality, we will evaluate by comparison the visual performance manually.

3.3. Comparison and discussion

For comparison the performance, we used 100 validation in DIV2K dataset [10]. It is the dataset of NTIRE 2019 colorization challenge [11]. The images of dataset are downloaded from the Internet with high-resolution with the variety of the scene-context.

표 1 Comparison methods

Method	Name	Train data	Test data
1	Our method	Coco-Stuff	DIV2K
2	Iizuka et al. [3]	Place	DIV2K
3	Larsson et al. [4]	ImageNet	DIV2K
4	Zhang et al. [5]	ImageNet	DIV2K

Besides, we compared three robust image colorization from Iizuka et al. [3], Larsson et al. [4] and Zhang et al. [5]. We used the pre-trained weight from the authors to generate the color version of image and compared to our result as in Table 1.

Figure 3 shows our quality result comparing with the other methods. Our results show the color diversity and lightness more than the other methods. But some cases in the right side gave the bad. Our results gave a little bit noise in the images.

표 2 The results of comparison methods

Method	PSNR	SSIM	$L2_{ab}$
1	Soft:19.961 Reg: 22.263	Soft: 0.785 Reg: 0.867	Soft: 0.584 Reg: 0.605
2	23.492	0.912	0.620
3	23.809	0.914	0.585
4	21.173	0.885	0.630

Table 2 shows our results to the other methods. Our network has three branches with regression and soft-encoding branch outputting the colorized image. So, we compare two outputs from our model to the other methods.

Our method with regression branch outputting the colorized images greater than the results from the soft-encoding branch in the quantitative metrics such as PSNR, SSIM (the higher the better). And $L2_{ab}$ of the soft-encoding branch is better than regression branch (the lower the better). It showed that the soft-encoding branch gave the better result when comparing the color attribute. Our result $L2_{ab}$ also gave the better result more than Iizuka et al., Zhang et al., and Larsson et al.

4. Conclusion

In this paper, the colorization method was considered as a scene-context classification integrated in the pixel-wise soft-encoding classification of 2D color histogram from ab channel and regression ab values. We used the original U-Net architecture adding the scene-context classification based on label smoothing to achieve the performance of transferring information well from the scene dataset. We achieve some encourage results when training on Coco-Stuff dataset and validating our model on DIV2K dataset.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2018R1D1A3A03000947).

참고문헌

[1] A. Deshpande, J. Rock, and D. Forsyth, “Learning large-scale automatic image colorization,” In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 567-575.

[2] S. Y. Choi, M. R. Luo, M. R. Pointer, and P. A. Rhodes, “Investigation of Large Display Color Image Appearance I: Important Factors Affecting Perceived Quality,” Journal of Imaging Science and Technology 52, 2008, vol. 52, no. 4, pp. 040904-1.

[3] S. Iizuka, E. Simo-Serra, and H. Ishikawa, “Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification,” ACM Transactions on Graphics (ToG), vol. 35, no. 4, pp. 1-11, 2016.

[4] G. Larsson, M. Maire, and G. Shakhnarovich, “Learning Representations for Automatic Colorization,” In European Conference on Computer Vision (ECCV), pp. 577-593, Springer, Cham, 2016.

[5] R. Zhang, P. Isola, and A. A. Efros, “Colorful Image Colorization,” In European conference on computer vision, pp. 649-666, Springer, Cham, 2016.

[6] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” In International Conference on Medical image computing and computer-assisted intervention, pp. 234-241, Springer, Cham, 2015.

[7] R. Müller, S. Kornblith, and G. Hinton, “When Does Label Smoothing Help?,” In Advances in Neural Information Processing Systems (NeurIPS), pp.4696-4705, 2019.

[8] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 Million Image Database for Scene Recognition,” IEEE transactions on pattern analysis and machine intelligence (TPAMI), vol. 40, no. 6, pp. 1452-1464, 2018.

[9] T. T. Nguyen-Quynh, N. T. Do, and S. H. Kim, “MLEU: Multi-level embedding u-net for fully automatic image colorization,” In Proceedings of the 4th International Conference on Machine Learning and Soft Computing (ICMLSC), pp. 119-121, 2020.

[10] E. Agustsson and R. Timofte, “NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp.126-135, 2017.

[11] S. Gu, R. Zhang, A. N. S, C. Chen, and A. P. Singh, “NTIRE 2019 Challenge on Image Colorization: Report,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019.