MLEU: Multi-Level Embedding U-Net for Fully Automatic Image Colorization

Tram-Tran Nguyen-Quynh, Nhu-Tai Do, Soo-Hyung Kim

School of Electronics and Computer Engineering, Chonnam National University
77 Yongbong-ro, Buk-gu, Gwangju 500 – 757, Korea
tramtran2@gmail.com, donhutai@gmail.com, shkim@jnu.ac.kr

ABSTRACT

This paper presents the method for tackling the challenge of fully automatically image colorization. We improve U-net by fusion multi-level feature from the pre-trained ImageNet to enhance the model under the small datasets. Furthermore, we reduce the unbalance colors by the enhancement distribution over quantized colors based on the smoothness of the prior distribution. The experiments in the DIV2K dataset show that our results are very encouraging. Our method improves PSNR as well as colorizes the images under complex textures.

CCS Concepts

• **Computing methodologies~Neural networks.**

• **Computing methodologies~Scene understanding**

Keywords

Unet, Fully Automatic Image Colorization, Multi-Level Embedding, Regression, Classification, weighted category cross-entropy, weighted one-hot vector

1. INTRODUCTION

A color image always represents much semantic more than a gray-scale image, so humans can easily distinguish and recognize the objects by vision [1]. Coloring old historical photos and videos will help recreate historical events more clearly, which is one of the laborious and costly works, so humans have always yearned to find methods that reduce costs and time.

Previous approaches about image colorization often fall into three kinds: the first requires partial human intervention to specify some colors on definite regions like as: [2] uses the inputted annotations for the image with some color scribbles by the users. Then, prior declared colors will be propagated automatically under the assumption that neighbor pixels should have the same color. Further progress, [3] uses texture similarity and continuity of intensity to increase coloring efficiency. The result can still be refined then.

Transferring colors from a reference image to a target image is the second kind. Charpiat et al. [4] deal with multimodality in colorization by using the distribution probability of all possible colors on every pixel. They use graph-cut to maximize the probability and discretization of the color space to transfer color between images having similar semantic structures. In the same purpose, [4],[5],[6] also propose some algorithms using corresponding descriptors.

Conference'10, Month 1–2, 2010, City, State, Country.

DOI: http://dx.doi.org/10.1145/12345.67890

The third kind is fully automatic colorization, and this is also our goal. Recently, several works in automatic colorization have been performed based on deep convolutional neural networks. Typically, [7] uses the model to learn knowledge from the set of the given images to predict colors for a new greyscale image. The image colorization is considered as a regression problem, images are distributed to adaptive clusters according to global information. Every neural network is trained on a specify cluster for colorization with L2 regression loss, using joint bilateral filtering for post-processing. Although colorization is fully automatic, [7] cannot apply to synthetic images as well as the loss of color information due to transformation from a color image to a grayscale image. Hence, coloring results tend to look desaturated.

The problem is that they apply a loss function that promotes conservative predictions. Realistically, an object does not only have a single color. [8] tackles this mistake by considering the colorization problem as the multinomial classification. For each pixel of the image, [8] suggests the distribution of plausible colors. Their results are brilliantly colored images, and their classification formula with 313 color bins is very creative, so we are aware of this effective direction. Nevertheless, the synthetic data for the training task is so huge to keep an amount of information, so the training process is slow and costly.

Meanwhile, [9] uses un-rebalanced classification loss, builds on hyper-columns on a VGG network, trains on ImageNet dataset, and evaluates PSNR, RMSE. They reach many improvements, yet color bleeding still occurs in the images that have many homogeneous textures. A few images are not be fully covered color because of a lack of rare colors

Most of the recent approaches solve the question of how to color a gray-scale image automatically, but there are still some common restrictions: i) use a huge database to train, ii) return desaturated and pale results, iii) do not show rare colors in images.

In order to circumvent issues related to image colorization, we use Multi-Level Embedding UNet (MLEU) to transfer learning from the pre-trained ImageNet weight as well as the advantage in gradient flow enhancing by skip connections between the contracting path and expanding path. Moreover, we improve the distribution over quantized color by interpolation and smoothness for reducing the unbalance colors and focusing on rare colors. We also make experiments on DIV2K [10] dataset and compare it with the state-of-the-art method from Zhang et. al. work [8].

The remainder of the paper includes three sections. Section 2 will propose a method and its analysis. Experimental results, as well as the discussion, are described in Section 3. Finally, we conclude our results and discuss further works in Section 4.

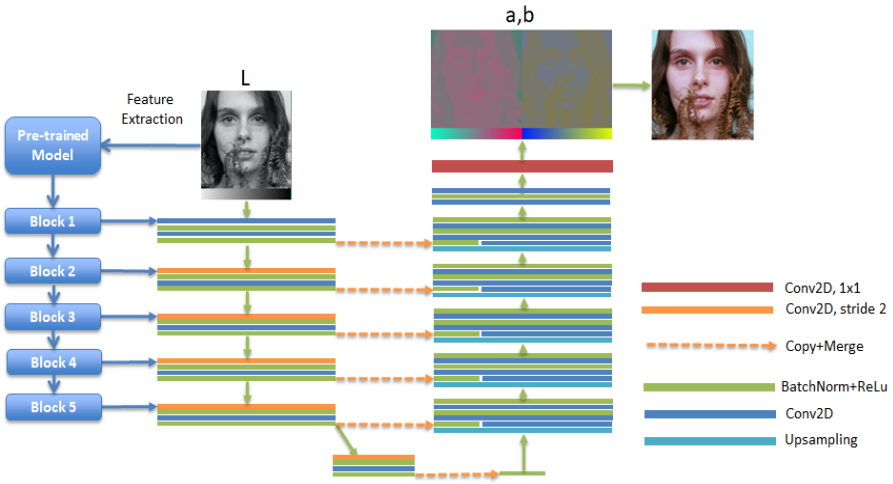

Figure 1. Multi-Level Embedding U-Net for Fully Automatic Image Colorization.

2. PROPOSED METHOD

In this section, we describe the design of MLEU model in detail. Firstly, we briefly define our fully automatic image colorization problem. Second, we mention enhancing the quantization process when focusing on a classification problem. Next, we explain our model with improving the U-Net model suitable for the colorization issue. Finally, we discuss the loss functions used in the paper.

2.1 Method overview

Given an input lightness image $I_L \in \mathcal{R}^{h\times w\times 1}$, our target is to predict the ab channel in CIE Lab color space $I_{ab} \in \mathcal{R}^{h\times w\times 2}$ where h, w is respectively height and width of images as Fig.1. From an integrated lightness channel in input and ab channel in output, we can transform from the gray image to the color image.

There are two main approaches to tackle this problem. Firstly, this problem is to solve from the regression view. It means that at every pixel of the input images, the model needs to learn a regression function to return two values a and b in the ab channel at the corresponding pixel [11]. In the second approach, every lightness value of the pixel will be classified into q bins receiving from the ab channel quantization process [8].

We build our network as in Fig.1 based on U-net [12] with the contracting path for encoding the lightness channel to the color features and the expanding path for decoding the color features into ab channel flexible. Two main approaches are only different at last Conv2D of the expanding path by the number of output and activation function.

From this advantage, we will be easy to do experiments on the classification and regression approach. Besides, we utilize the transfer learning to integrate pre-trained features from the conventional ImageNet models such as VGG16 [13], ResNet [14] into every encoding block. It helps the network easy to learn color image features from the small datasets instead of training the big dataset as Zhang et. al. [8]. Moreover, in the quantization process of ab channel into the discrete color bins, we apply the smoothness transform on prior color distribution to enhance the quality in the classification approach.

2.2 Preprocessing Input and Output data

For the regression approach, the input data is the lightness $I_L \in \mathcal{R}^{h\times w\times 1}$ and output is $I_{ab} \in \mathcal{R}^{h\times w\times 2}$ in CIE color space Lab. I_L and I_{ab} are normalized in the range [0, 1]. The role of the regression model is learning the mapping function $F_r(I_L) = \hat{I}_{ab}$, where $\hat{I}_{ab}$ is the predicted image in ab channel of CIE Lab color space as in Fig.2.

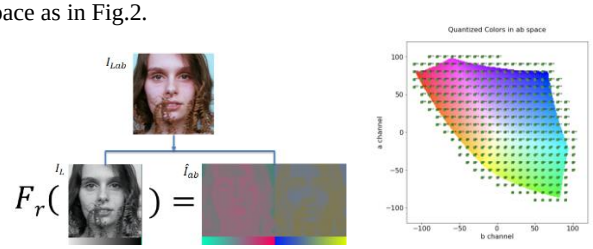

Figure 2. (Left) Regression mapping function and input/ output data and (Right) Quantized colors in ab space.

In Fig.2 (Left), the right image $\hat{I}_{ab}$ has the values in the range smaller than the range of the left image I_L. It means that the cells in our eyes determine brightness mainly, only very fewer for colors, which leads to the lightness layer is a lot sharper than the color layers. So, the regression approach has the mean effect to return the desaturated results [8].

For the classification approach, it is different from the regression approach at replacing the I_{ab} to the probability distribution $Z_{ab} \in \mathcal{R}^{h\times w\times n} \in [0,1]$ over quantized color $Y_{ab} \in \mathcal{R}^{h\times w\times 1} \in [0, n-1]$ in ab channels of CIE Lab color space where n is the number of bins for discretization. We achieve Y_{ab} by discretizing the ab channel into n bins. It is simply divided into the 2D grid by bins on the equal grid size as Fig. 2 (Right).

After that, we build $F_q(\mathrm{Y}_{ab})$ to transform Y_{ab} into Z_{ab}. Firstly, Y_{ab} will be converted into a one-hot vector as the common output of the classification problems. Different from the common classification output using one-hot encoding, the classification labels in this problem have relation together on 2D space. We need to express these relations by the weighted one-hot vector by applying the k-nearest searching for 5 nearest neighbors. All of them will be weighted by their distance using the Gaussian kernel with $\sigma = 5$. The role of the classification model is learning the mapping function $F_c(I_L) = \hat{Z}_{ab}$, where $\hat{Z}_{ab}$ is the predicted distribution over quantized color.

2.3 Smoothness prior distribution

The next problem in the classification approach, we need to tackle the unbalance among the labels. Also, we choose the number of bins n = 313 and carry out the statistical analysis to make sense

of the quantized color distribution on DIV2K dataset [10] [15] as Fig. 3.

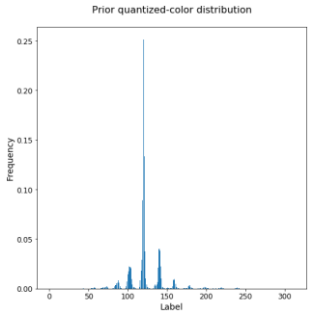

Figure 3. Prior quantized-color distribution.

Fig. 3 shows the highest imbalance with the highest frequency in the range [100-150] of 25% and almost the remain range nearly equal of zero.

To decrease the distance among the label frequency, we make the interpolation on the prior probability distribution, apply the 1D-Gaussian kernel with σ =5 and normalize to unit-length as Eq. 1. After that, we use the equation of Zhang et. al [8] to calculate the color-label weight based on smoothness prior distribution as Eq. 2:

$$\boldsymbol{P_s} = Interp(\boldsymbol{P}) * \mathbf{N}_\sigma, |\boldsymbol{P_s}| = 1 \quad (1)$$

$$W \propto \left((1-\boldsymbol{\lambda})\boldsymbol{P_s} + \frac{\boldsymbol{\lambda}}{\boldsymbol{n}}\right)^{-\alpha}, \mathrm{E}[\boldsymbol{W}] = \sum_{i=1}^{n} \boldsymbol{P_{si}} \boldsymbol{W}_i = 1 \quad (2)$$

where $*$ is convolution operator, N_σ is Gaussian kernel with, P is prior distribution, $\mathrm{P_s}$ is smoothness distribution, λ and α are control parameters with $\lambda = 0.5$ and $\alpha = 1$ in this paper.

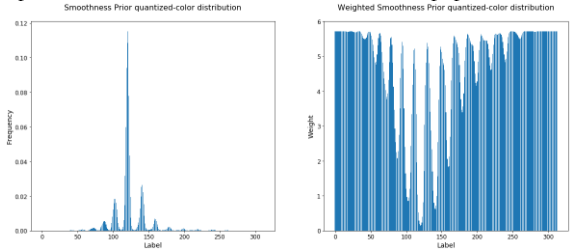

Figure 4. Smoothness (Left) and Weighted (Right) Prior quantized-color distribution.

Fig. 4 shows the different distance of the smoothness distribution is 12% lower than 25% in the prior distribution. They help to decrease unbalance among labels.

The weighted one-hot vector in the previous section helps our model to focus on the color space relation. In this section, the weighted smoothness prior distribution will be enhanced on the rarity colors to make the image look natural.

2.4 Network structure

Our model is depicted in Fig. 1. Its goal is aimed to transform the lightness image in pixel-level to the ab image in the CIE Lab color space. The contracting path produces the coarse feature maps by 5 encoding blocks. Every encoding block consists of the convolution layers, batch normalization layers, and ReLU activation functions.

They will connect to 5 decoding blocks of convolutional, up-sampling layers, normalization layers, and activation functions at the same level in the expanding path by skip connection. The role of skip connection will help to prevent the loss of information by down-sampling from stride 2 from the last convolution layer in the encoding block. In the image colorization problem, the model uses stride 2 to down-sampling instead of the max-pooling layer at the end of the block to capture more details.

Besides, the pre-trained models from ImageNet will extract the features from the gray image. They will be integrated into the encoder block at the same level. The features will combine with the encoding feature after every block, which helps the model to learn more general features by the classification knowledge domain in ImageNet.

The last convolution layer 1x1 after the expanding path plays the role of a multi-layer perceptron network in pixel-level to classify or regress by the output values. In the classification, we use the soft-max activation function, and the sum of filters are the number of quantized-color bins. Otherwise, the convolution layer uses the tanh function and two filters.

2.5 Loss function

Finally, we use the mean square error loss for the regress approach as below:

$$\boldsymbol{L_{MSE}}(I, \hat{I}) = \frac{1}{hw}\sum_{h,w}\left\|I - \hat{I}\right\|_2^2 \quad (3)$$

Next, for the classification approach, we use the weighted category cross-entropy as below:

$$\boldsymbol{L_{WCCE}}(Z, \hat{Z}) = -\sum_{h,w} W_{h,w} \sum_{n} Z_{h,w,n} log(\hat{Z}_{h,w,n}) \quad (4)$$

where $W_{h,w}$ is the weighted smoothness prior distribution of quantized color with the highest probability at location (h, w), I ($\hat{I}$) is the ground-truth (predicted result) of the ab channel normalized in [-1,1], Z ($\hat{Z}$) is the ground-truth (predicted result) of the distribution over quantized color.

3. EXPERIMENTAL RESULTS

3.1 Dataset, Metrics, and Training

We used the dataset DIV2K [10] in NTIRE 2019 colorization challenge [15]. DIV2K has 800 images for training and 100 images for validation. All images are collected from the Internet with high-resolution with the diversity of the scene as Fig. 5.

The quantitative measures in the challenge are Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM). In this paper, we only used PSNR to evaluate the quantitative metric. But the PSNR metric cannot reflect the visual quality, the challenge offers the perceptual quality conducted in manual evaluation by a human. In this paper, we will evaluate by comparison the visual performance manually.

Figure 4. DIV2K Dataset.

We applied Keras Tensorflow on the environment Python 3.7 to build MLEU model. For data augmentation, we incorporated the flipping, rotating and random cropping of the images. About the optimization algorithm, we tried to train on Adam, SGD, and RMSProp with learning rate 0.004 and selected the best model based on the quantitative and visual metrics. For the pre-trained model to combine features, we used pre-trained VGG-16 on ImageNet.

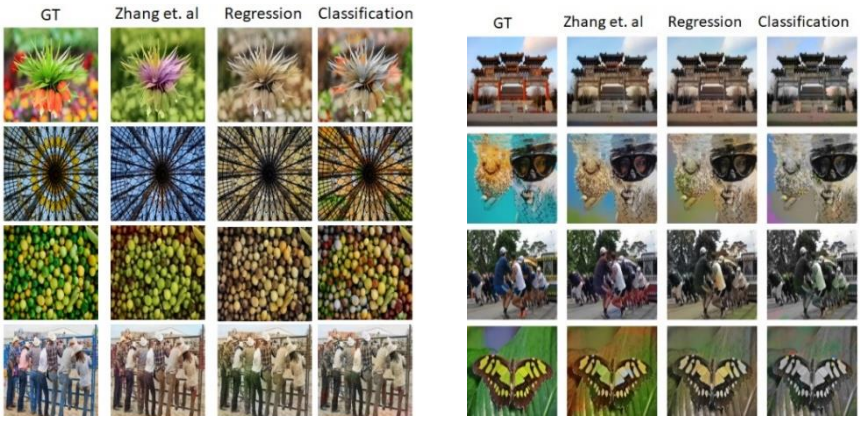

Figure 6. (Left) The best cases and (Right) the worst cases in our model.

We compared our models in regression and classification approaches with Richard Zhang et. al. [8] in Table 1. With Richard Zhang et. al. method, we used their pre-trained weighted with training data on ImageNet.

Table 1. Comparison methods

Method	Name	Type	Train Data
1	MLEU Regression	Regression	DIV2K
2	MLEU Classification	Classification	DIV2K
3	Richard Zhang et. al. . [8]	Classification	ImageNet

3.2 Results

Fig.5 shows our training history on MLEU Regression (Left) and classification (Right). The regression model converges slower than the classification.

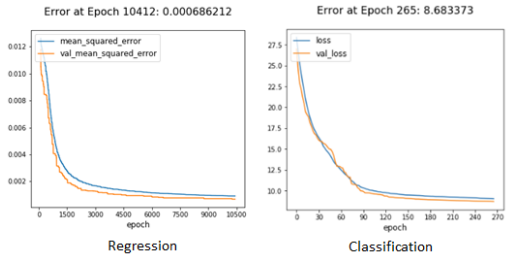

Figure 5. Training history on MLEU Regression (Left) and Classification (Right).

Table 2 shows our MLEU classification archive the better performance than Zhang et. al. method at the quantitively metric. Although about the visual metric, we only achieved the visual performance nearly as Richard Zhang et. al.

Table 2. Comparison of PSNR metric among methods

Method	Name	PSNR
1	MLEU Regression	20.9
2	MLEU Classification	**23.1**
3	Richard Zhang et. al. . [8]	22.8

In Fig. 6, we have the best cases in the left columns. It shows that the complex textures such as in picture 2, and 3 at the column left presented more details than Zhang et. al. method. However, we also met some errors in fail cases. The picture 2 in column right has some noises, and the remain pictures favor grayish, desaturated results.

4. CONCLUSION AND FUTURE WORKS

In this paper, the colorization method was considered as a classification based on color quantization or regression method for comparison. We improved the original U-Net architecture to achieve the performance of transferring information well on a small dataset. Nevertheless, our results remain some limitations which we plan to study in future work. One is the coloring for foregrounds regions is generally accurate than for background. We plan to cluster the color quantization space to avoid ambiguity in the quantizing process. Moreover, we need to combine the regression and classification approaches to enhance quality.

ACKNOWLEDGMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education(NRF-2017R1A4A1015559)

REFERENCES

[1] V. Lavrenko, R. Manmatha, and J. Jeon, “A model for learning the semantics of pictures,” *Advances in Neural Information Processing Systems*, 2004.

[2] A. Levin, D. Lischinski, and Y. Weiss, “Colorization using optimization,” in *ACM SIGGRAPH 2004 Papers on - SIGGRAPH ’04*, 2004, p. 689.

[3] Q. Luan, F. Wen, D. Cohen-Or, L. Liang, Y.-Q. Xu, and H.-Y. Shum, “Natural Image Colorization.,” *Rendering Techniques*, pp. 309–320, 2007.

[4] P. Galvis-Assmus, “ACM SIGGRAPH,” *ACM SIGGRAPH Computer Graphics*, vol. 38, no. 4, p. 2, Nov. 2004.

[5] M. He, J. Liao, D. Chen, L. Yuan, and P. V. Sander, “Progressive Color Transfer With Dense Semantic Correspondences,” *ACM Transactions on Graphics*, vol. 38, no. 2, pp. 1–18, Apr. 2019.

[6] A. Y. S. Chia *et al.*, “Semantic Colorization with Internet Images,” *ACM Transactions on Graphics*, vol. 30, no. 6, pp. 1–8, 2011.

[7] Z. Cheng, Q. Yang, and B. Sheng, “Deep Colorization,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, vol. 2015 Inter, pp. 415–423.

[8] R. Zhang, P. Isola, and A. A. Efros, “Colorful Image Colorization,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial*

Intelligence and Lecture Notes in Bioinformatics), 2016, pp. 649–666.

[9] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning Representations for Automatic Colorization," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9908 LNCS, 2016, pp. 577–593.

[10] E. Agustsson and R. Timofte, "NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.

[11] 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*, vol. 35, no. 4, pp. 1–11, Jul. 2016.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, 2015, pp. 234–241.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] S. Gu, R. Zhang, A. N. S, C. Chen, and A. P. Singh, "NTIRE 2019 Challenge on Image Colorization : Report," 2019.

Authors’ background

Your Name	Title*	Research Field	Personal website
Tram-Tran Nguyen Quynh	Master Student	Pattern Recognition, Deep Learning	
Nhu-Tai Do	Ph.D Candidate	Pattern Recognition, Deep Learning	
Soo-Hyung Kim	Full Professor	Pattern Recognition, Deep Learning	http://pr.jnu.ac.kr/shkim/