Picture Colorization

Junhui Li, Yisen Wang, Zhen Zhong, Ziyuan Huang
University of Michigan, Ann Arbor
{opheelia,yisenw,zhongzh,ziyuanh}@umich.edu

Abstract

In this project, we addressed the greyscale picture colorizing problem with the approach proposed in [3] by rebuilding the model and training it on a small portion of the ImageNet dataset [1]. Unlike the approach in the paper [3], we trained our model directly upon a pre-trained ResNet-50 model [2]. we tried two different loss functions proposed by [3] during training and compared their effect on the corresponding colorized output. After training the models, We evaluated our model by testing the images on real human. We achieved a relatively acceptable result with smaller size of training data.

1. Introduction

When photography was first invented in the 19th century, photos taken by the cameras were in the form of black and white image. Due to the limitation of technology, lots of old pictures are only available in grayscale pictures. Since the only colors of grayscale images are shades of gray, less information needs to be provided for each pixel of them. However, there are times when we need to restore the huge amount of information lost due to transformation of colored scenes to grayscale pictures, which leads to the motivation to colorize these grayscale images.

Humans can hallucinate colors of grayscale images because the texture as well as the semantics of the scene provide abundant clues for the assignment of colors. Thus our goal is, instead of predicting the ground truth color, producing semantically reasonable coloring for grayscale image-which can be formalized as follow: given input images with only lightness channel, our model predicts for each pixel the corresponding color. We focused on the approach in [3] where new methods of loss function design and evaluation of model are introduced. Since [3] did not provide their source code, we reimplement the training process and two loss functions based on the paper. Instead of training from scratch, we trained our models directly upon a pre-trained ResNet-50 model [2]. After training, we use the Perceptual Realism Test [3], which require us to test our generated images on real human. Though our train dataset is much smaller, our final result is comparable with [3].

2. Related Work

Zhang et al. [3] first utilized CNN model with L2 loss to solve the colorization problem. And then they reformulated the problem as multinomial classification by dividing the output ab space into discrete bins of size 10. The class rebalancing term in the training objective, effectively oversamples rarer, more vibrant colors relative to their representation in the training set. Applying the combination of these modifications, which is to use multinomial classification loss instead of L2 loss, and add the class-rebalancing enabled their results to be qualitatively more colorful than previous and concurrent work, and achieved 32% accuracy in the "Perceptual Realism Test", which asked human participants to choose between a generated and ground truth color image. 32% accuracy shows Zhang et al.'s algorithm is the state-of-the-art colorization algorithm.

3. Method

We utilize a simple feed-forward VGG structure as our model, despite some alternations of expanded dimensions at the tail, extended depth, and additional dilated convolutions. The overall structure of the model is illustrated in Fig. 1. The model takes in the l (lightness) channel as input and outputs the learned ab (chromaticity) channel, which will then be combined with the l channel to construct the colorized image.

Objective function: According to Zhang et al.[3], myriad objective functions are applied to image colorization tasks in the state-of-the-art algorithms. We tested two objective functions mainly discussed in [3]: L2 Loss and Multinomial Classification Loss.

3.1. L2 Loss

The basic formula for L2 Loss in this problem is

$$L(\hat{Y}, Y) = \sum_{h,w} ||\hat{Y} - Y||_2^2$$

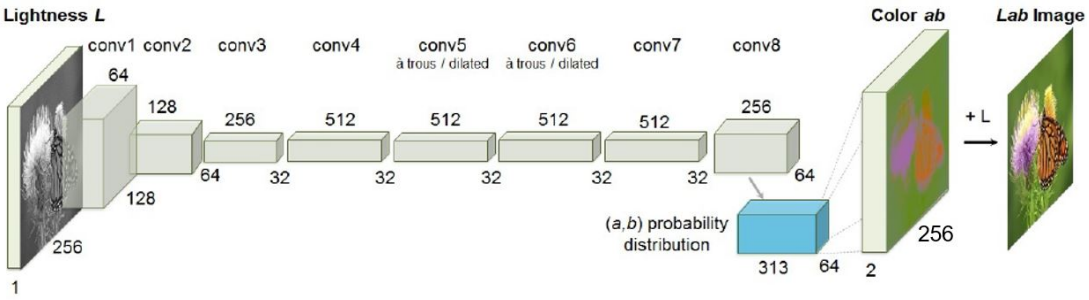


Figure 1. Structure of the model. Each block represents a duplicated or triplicated sequence of a convolution layer and a ReLU layer, ended with a Batch Normalization layer. [3]

where $\hat{Y}$ stands for the model's prediction and Y represents the ground truth (both in shape of $[H, W, 2]$ where H, W are shapes of the image and 2 denotes the ab channels).

However, this objective function has potential problems. This function learns a deterministic mapping between the l channel and the ab channels which is implausible given the nature of possible variations in colors of same objects, and results in far-fetched and desaturated images (*e.g.* in Section 4 A man with a fish water are colorized as blue where it actually looks a little green and grayish). Another drawback proposed by Zhang et al.[3] is that when the colorization becomes non-convex, L2-Loss will fall out of the set.

3.2. Weighted Multinomial Classification Loss

To better capture the nature of ambiguity and object color variations, Zhang et al.[3] proposed the Weighted Multinomial Classification Loss. Following their approach, we quantize the in-gamut ab channel values in our dataset into $Q = 313$ bins with grid stride 10. The model learns the probability distribution of the color of each pixel over the Q bins and makes prediction by summing over the ab values in the bins weighted by the probability.

To construct the loss, given $\hat{Z} \in [0, 1]^{H \times W \times Q}$ the predicted probability distribution and Z the ground truth value. The objective function is

$$\mathrm{L}_{cl}(\widehat{\mathbf{Z}}, \mathbf{Z}) = -\sum_{h,w} v\left(\mathbf{Z}_{h,w}\right) \sum_{q} \mathbf{Z}_{h,w,q} \log\left(\widehat{\mathbf{Z}}_{h,w,q}\right)$$

where $v(Z_{h,w})$ is the weight to rebalance the loss according to color class rarity (to be discussed in the next section).

3.3. Class Rebalancing

Due to the ubiquity of background colors, such as sky, clouds, and grasslands, these color values tend to aggregate in orders of magnitude higher than other color values. To rebalance different color classes, we applied the methods introduced in Zhang et al.[3] that resamples the training space using the weighted loss. The result against different quantized ab values are illustrated in Fig. 2.

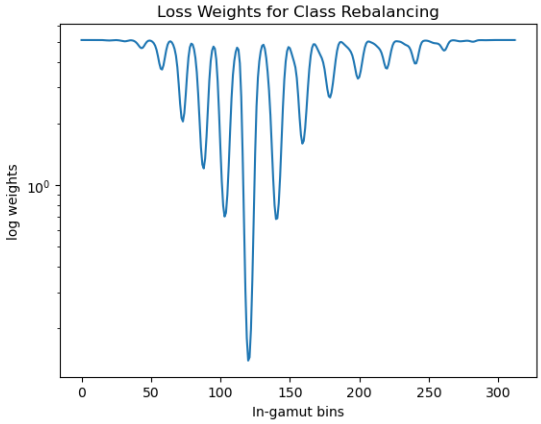


Figure 2. loss weights for each quantized ab value.

4. Experiments

Due to the limitation of our computer, we could not train the whole ImageNet [1] as Zhang et al. [3]. Instead, we use 200,000 images from the ImageNet as our dataset. We split our dataset as in table 1.

Dataset	Number of images
Training	190,000
Validation	8,400
Testing	1,600

Table 1. Dataset splitting

We trained two models separately: one was trained with L2 loss and the other with multinomial classification loss (MCL for short). We trained each model for 15 epochs, and we monitored their training losses and validation losses at the same time (recorded in figure 3 and figure 4).

In order to prevent overfiting, we choose the models with the least validation loss. Therefore, for the model with L2 loss, we choose epoch 8 as our final model; for the model with MCL loss, we choose epoch 10 as our final model.

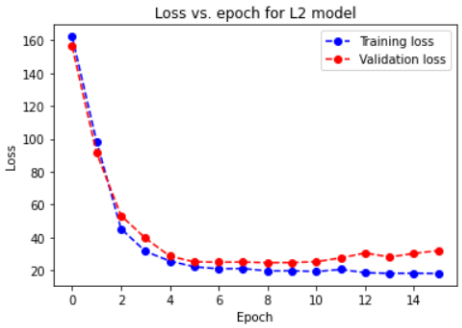


Figure 3. Loss vs epoch for the model with L2 loss

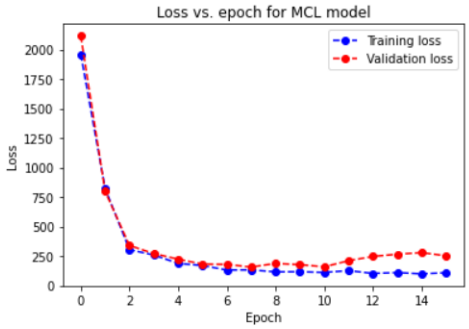
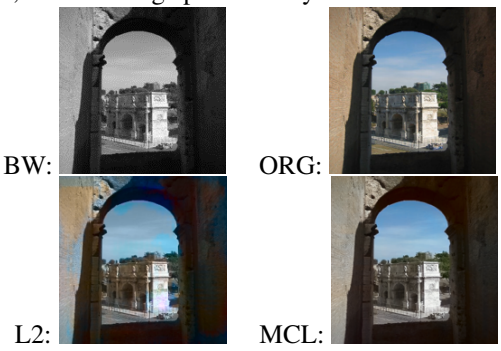


Figure 4. Loss vs epoch for the model with MCL loss

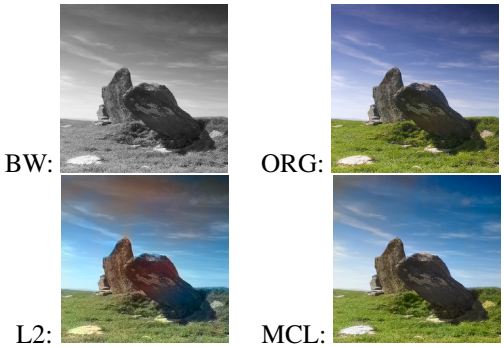
After training, we test our models on the testing dataset (which is composed of 1,600 images). In the following subsections, we'll evaluate the two models generated by us.

4.1. Outputs

After running the test dataset on our own models, we observe the following points:
(1) In most cases, the pictures generated by the model with MCL loss are more similar to the ground truth pictures than the model with L2 loss. Please see the following examples. (Note: BW = black and white image; ORG = original ground truth image; L2 = image predicted by model with L2 loss; MCL = image predicted by model with MCL loss)



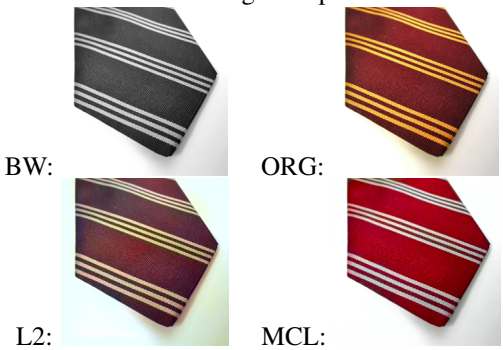
Arc de Triomphe



A lawn with a stone on it

A man with a fish

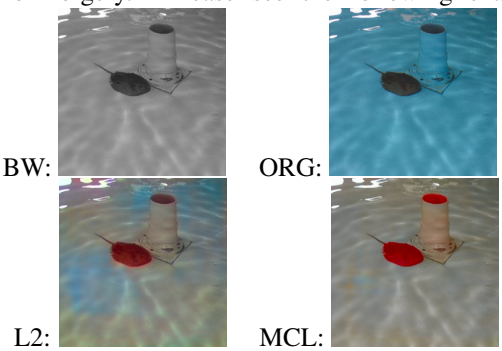
(2) In some cases, the pictures generated by our models are not similar to the ground truth. But our predictions are semantically reasonable, which could also fool human's eyes. Please see the following examples.



A tie

A mother and her baby

(3) Of course, in a few cases, our models fail to generate semantically reasonable pictures. The combination of the generated colors would clearly show the traces of forgery. Please see the following example.



A leaking pool

Pumpkins

4.2. Perceptual Realism Test

Just like [3], we also perform the Perceptual Realism Test on the two models generated by us. We asked our participants to distinguish the generated figure from the ground truth figure. Each test include 20 sets of images from each model (i.e. each subject need to identify 40 sets of images). The test images are randomly chosen from the outputs of the testing dataset (which includes a total of 1600 images, same as [3]). After releasing our survey for a week, we received results from 588 participants. Figure 5 includes both the results of our models and the results from [3].

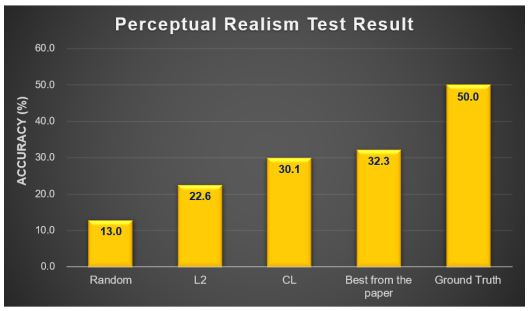


Figure 5. Result of the Perceptual Realism Test

From figure 5, we can observe that the model based on L2 loss does not provide an adequate result (22.6%), just as stated by [3]. As for the model based on the MCL loss, our result (30.1%) is close to [3]. Please note that we trained on a much smaller dataset (190,000 training images), and [3] trained on the whole ImageNet dataset (about 1,300,000 images). Therefore, it seems **our method of using a pre-trained ResNet-50 model shows great potential**.

4.3. Now, Let's Predict the History!

After the experiment mentioned above, we believed we have the ability to predict some historical images now! The following pictures (and the Appendix) show some very interesting results.

Chairman Zedong Mao

Sophia Loren (left) and Jayne Mansfield

Old Summer Palace (a.k.a. Yuanming Yuan, before destruction)

5. Conclusions

Despite the relative small size of training data, our model fooled human participants on 30.1% of the instances, showing that our implementation basically reproduced the results of [3] (which achieved 32.3%), and predicted colors for historical images with good quality. Moving forward, we hope to train our model on the whole ImageNet dataset to improve the performance. We would also like to attempt more suitable loss functions for training since 30% is faraway from 50%, the expected result for ground truth.

References

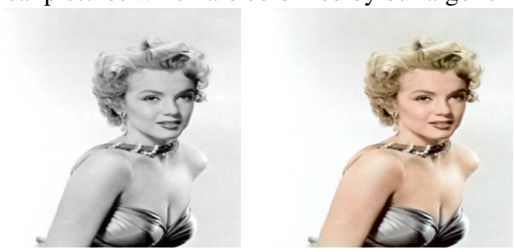
[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. 2009.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE*, abs/1512.03385, 2015.

[3] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. *CoRR*, abs/1603.08511, 2016.

Appendix

In the appendix of this paper, we will show more historical pictures which are colorized by our algorithms.



Marilyn Monroe (pic 1)

Marilyn Monroe (pic 2)

Marilyn Monroe (pic 3)

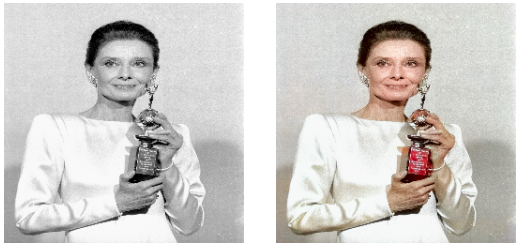
Albert Einstein

Albert Einstein and his friend

Taj Mahal (under construction)

Chairman Mao and President Nixon

President Abraham Lincoln



Audrey Hepburn (pic 1)


Audrey Hepburn (pic 2)


Audrey Hepburn (pic 3)


Broadway at the United States Hotel, New York, 1915


Nanjing Massacre (pic 1)


Nanjing Massacre (pic 2)

We mourn the 83rd anniversary of the Nanjing Massacre (12/13/1937 - 12/13/2020)