Colorful image colorization
(2016)

Richard Zhang, Phillip Isola, Alexei A. Efros
Notes

Contributions

The main objective in this paper is to predict the colorization of a grey scale image, the authors propose a novel classification based approach and using a class-rebalancing term in the loss function for more realistic colors. And propose to use crowd sourcing with mecanical turk to evaluate the colorized images.

They also use colorization as a pseudo-task in a self-supervised task for pretraining a classification network instead of using ImageNet.

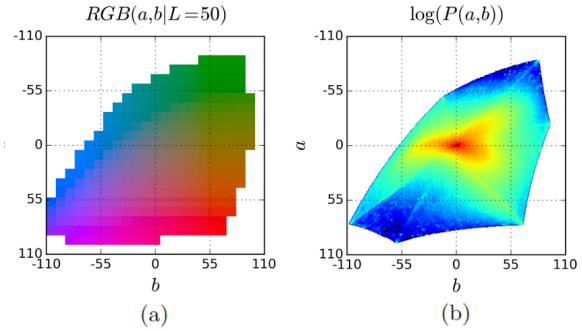
Quantifying the color space

The objective in a coloriztion task is given the lighness channel (i.e. the gray scale of a given image) as an input $\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$ we'd like to predict the ab values of each pixel $\mathbf{Y} \in \mathbb{R}^{H \times W \times 2}$. One possible way is to train the network using a simple L2 Loss:

$$\mathrm{L}_2(\widehat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \left\| \mathbf{Y}_{h,w} - \widehat{\mathbf{Y}}_{h,w} \right\|_2^2$$

The problem with this approach is that in a colorization task, we can have many plausible colorizations at the same time, in the case, the network will simply ouput the average of the set of the plausible solutions, which is of grayish color.

The authors propose to transform the problem into a classification problem, and the output is a probability distribution over all the possible pairs of (a, b) colors, which are 313 pairs quatified using a step of 10 for $(a, b) \in [-110, 110]$ and taking the pairs within the (a, b) color space (i.e. we end up with $22 \times 22 = 484$ in total, 313 are in the ab space).



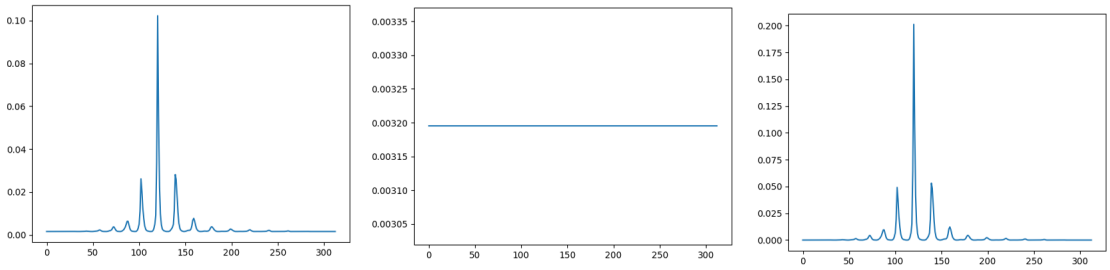
And the loss in this case is a categorical cross entropy between the output probabilities $\widehat{\mathbf{Z}}_{h,w,q}$ and groud truth $\mathbf{Z}_{h,w,q}$, with a weighting factor $v\left(\mathbf{Z}_{h,w}\right)$:

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

Generally, the ground truth $\mathbf{Z}_{h,w,q}$ are in the form of one hot vectors where we set the bin with closest ab values to one. But in this case the authors use soft-encoding, where we find the closest 5 ab bins, and weight them proportionnaly using a Gaussian kernel with $\sigma = 5$, so we'll endup with $\mathbf{Z}_{h,w,q}$ which is a matrix of size $H \times W \times 313$, where each vector $H \times W$ have 5 non-zero values.

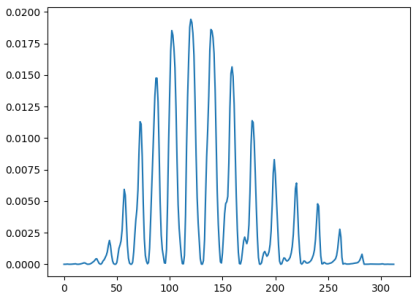
Class rebalancing

If we plot the ab color space (figure above, (b)) of the ImageNet images, we'll see that the colors are concentrated at low *ab* values, to avoid the clopse of the predict in this region and encourage the model to predict rare colors. We add a class rebalancing factor to the loss; so that the loss of the frequenst colors is reduced, and for the rare ones is amplified, this is done in the following, first we find the prior distribution of colors using the ImageNet images after applying a Gaussian kernel for smoothing $\widetilde{\mathbf{p}}$ (figure bellow left), and we contruct a uniform distribution Q based on this prior (figure bellow middle) and then mix both with a factor $\lambda = 0.5$ (figure bellow right) and normalize it to get an expectation of one.



The weighting factor is then the corresponding bin q^* of the new distribution $\mathbf{w}$ (figure bellow):

$$v\left(\mathbf{Z}_{h,w}\right) = \mathbf{w}_{q^*}, \text{ where } q^* = \arg\max_q \mathbf{Z}_{h,w,q}$$
$$\mathbf{w} \propto \left((1-\lambda)\widetilde{\mathbf{p}} + \frac{\lambda}{Q}\right)^{-1}, \quad \mathbb{E}[\mathbf{w}] = \sum_q \widetilde{\mathbf{p}}_q \mathbf{w}_q = 1$$



Class probabilities to ab colors

Now that we have an ouput in the form of probabilities of the 313 possible bins, we need to tranform them into estimates in the *ab* space, to do this we first use a softmax with a temperature T to transform the logits into probabilities:

$$f_T(\mathbf{z}) = \frac{\exp(\log(\mathbf{z})/T)}{\sum_q \exp\left(\log\left(\mathbf{z}_q\right)/T\right)}$$

and the estimate if the average of all of the probabilities $\mathcal{H}\left(\mathbf{Z}_{h,w}\right) = \mathbb{E}\left[f_T\left(\mathbf{Z}_{h,w}\right)\right]$. A first possible way is to simply take the argmax of the probabilites (when T=0), but this given inconsistent results

(some red patches in a green object), of we can take the mean of the distribution (T = 1), the authors found that with $T = 0.38$ we get good results.

Evaluation

For evaluating the results of the colorization, there is three metrics:

- Perceptual realism: Using Amazon Mecanical Turk, measuring the error rate when telling the colorized and the ground truth images apart, in this case the best results in random with 50% error rate.
- Semantic interpretability: In this metrics, we measure the classification accuracy of a VGG network trained on imagenet, using as inputs the colorized image instead of the ground truth.
- Raw accuracy: As a low-level test, this metric computes the percentage of predicted pixel colors within a thresholded L2 distance of the ground truth in ab color space.

Colorization Results on ImageNet

Method	Model Params (MB)	Model Feats (MB)	Runtime (ms)	AuC non-rebal (%)	AuC rebal (%)	VGG Top-1 Class Acc (%)	AMT Labeled Real (%)
Ground Truth	–	–	–	100	100	68.3	50
Gray	–	–	–	89.1	58.0	52.7	–
Random	–	–	–	84.2	57.3	41.0	13.0±4.4
Dahl [2]	–	–	–	90.4	58.9	48.7	18.3±2.8
Larsson et al. [23]	588	495	122.1	**91.7**	65.9	**59.4**	**27.2±2.7**
Ours (L2)	129	127	17.8	91.2	64.4	54.9	21.2±2.5
Ours (L2, ft)	129	127	17.8	91.5	66.2	56.5	23.9±2.8
Ours (class)	129	142	22.1	91.6	65.1	56.6	25.2±2.7
Ours (full)	129	142	22.1	89.5	**67.3**	56.0	**32.3±2.2**

Another usage of the colorarization taks, is to consider it as a pseudo task, and pretrain the model to predict the colors instead of using imagenet, and compare the results to the baseline which is imagenet trained model for classification / detection and segmentation using PASCAL VOC dataset:

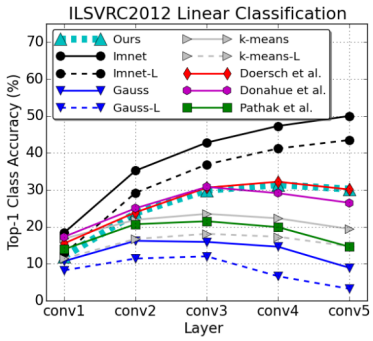


Fig. 7. ImageNet Linear Classification

Dataset and Task Generalization on PASCAL [37]

		Class. (%mAP)			Det. (%mAP)		Seg. (%mIU)	
fine-tune layers	[Ref]	fc8	fc6-8	all	[Ref]	all	[Ref]	all
ImageNet [38]	-	76.8	78.9	79.9	[36]	56.8	[42]	48.0
Gaussian	[10]	–	–	53.3	[10]	43.4	[10]	19.8
Autoencoder	[16]	24.8	16.0	53.8	[10]	41.9	[10]	25.2
k-means [36]	[16]	32.0	39.2	56.6	[36]	45.6	[16]	32.6
Agrawal et al. [8]	[16]	31.2	31.0	54.2	[36]	43.9	–	–
Wang & Gupta [15]	–	28.1	52.2	58.7	[36]	47.4	–	–
*Doersch et al. [14]	[16]	44.7	55.1	**65.3**	[36]	**51.1**	–	–
*Pathak et al. [10]	[10]	–	–	56.5	[10]	44.5	[10]	29.7
*Donahue et al. [16]	–	38.2	50.2	58.6	[16]	46.2	[16]	34.9
Ours (gray)	–	**52.4**	**61.5**	**65.9**	–	46.1	–	35.0
Ours (color)	–	**52.4**	**61.5**	**65.6**	–	46.9	–	**35.6**

Table 2. PASCAL Tests