Image Colorization - Vision and Cognitive Services

Jacopo Guidolin
jacopo.guidolin@studenti.unipd.it

Giovanni Vedana
giovanni.vedana.1@studenti.unipd.it

Abstract

This project aims to analyze in depth the topic of Image Colorization. We studied and compared different techniques, from the simplest model to the state-of-the-art approaches. Moreover we improved an already existing model by applying a suitable transformation that modify the saturation of the colorized images. We also proposed our pipeline that makes use of a pre-trained high-level feature extractor.

1. Introduction

The aspect we decided to tackle is the Image Colorization problem. This is notably a challenging problem in the field of Vision and Cognitive Services since it implicitly incorporates also other relevant computer vision tasks such as Image Segmentation and Object Recognition.

The Image Colorization challenge consists in building an algorithm that is able to reconstruct a probable and believable colored image by working only with its grayscale version.

This challenge can be also extended to the more complex problem of Video Colorization. This task results to be even more rough since it is required to solve also the problem of temporal inconsistency between adjacent frames. This task could be of notable interest for old photographic material since thanks to this technique old images or films could be restored in order to bring new life to them and a realistic look.

In addition to this Image Colorization has also more scalable and interesting application such as improving image compression as seen in [1] or correcting images imperfections like chromatic aberration as seen in [2].

2. LAB Colorspace and Formulation of the Problem

The natural setting to develop our framework appears to be the LAB Colorspace. This colorspace provides a particular representation for images in which every pixel is described by three channels (L, a, b). L corresponds to the lightness of the pixel (also known as the grayscale of the image) and the values a and b represent the relative position of the color in respect to four reference points: red, green, blue and yellow. In particular the a channel represents the relative position of the color between red and green and the b the relative position between blue and yellow.

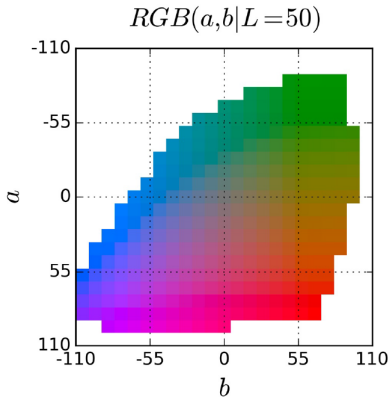

Figure 1. An approximate representation of the RGB equivalent of a slice (L=50) of the LAB Colorspace.

This colorspace equips us with what seems to be a natural representation of images in the contest of our task. Since our goal is to colorize grayscale images, our task becomes trying to provide a feasible estimate of the a and b channels of our pixels when receiving as input only the lightness L channel.

3. Difficulties

This task turned out to be very demanding to face for different reasons. To begin with, the relatively large dimension of the images we considered demands for large models and large models need to be trained on a set that is substantial enough in order to grant the possibility of broad generalization.

Moreover the task is arduous per se. Indeed the luminosity channel furnishes a limited amount of information about the real image. Due to this fact the model should be complex enough (namely 10-50 millions of parameters in the case of neural networks) to understand and unravel a complex signal from a relatively poor input. Since we are asking

to our net to increase the channel dimension of a signal from 1 to 3 we cannot expect realistically to find an easy way out.

This issue inevitably reflects on the training phase.

In order to achieve finer results we would need more time and more resources: we partially solved the space problems (RAM and Memory) by splitting the training images in more batches and loading them one at the time, saving gradually the weights of the model. For what concerns the time problems we alleviated them by training the models on a high-performance GPU (still very weak with respect to those used to train state-of-the-art models).

4. Dataset

To train and test the models we will present in 7.1 we used the MIRFLICKR Dataset[1], a collection of 25000 images of size 224×224×3. Due to RAM and Memory issues we were not able to take advantage of the whole Dataset, but we were forced to select a smaller subset, whose size depended on the complexity of the network.

For what concerns the model discussed in 7.2 instead, we decided to test it on the ImageNet database since it has been trained on it by its authors.

For training purposes we filtered the images and removed the ones that were already in grayscale. Black and White images show no use during the training of our models, on the contrary they could be harmful and damage the process. The other data manipulation we performed before training is a classic normalization of the features.

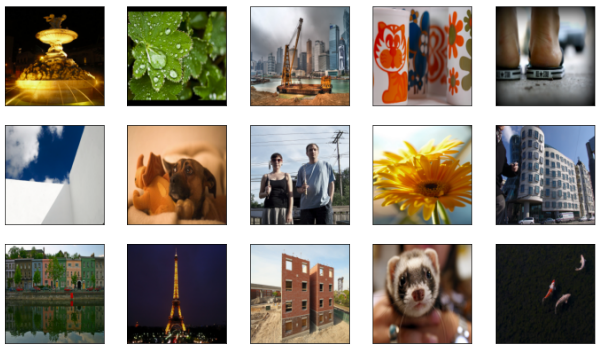

Figure 2. Random images from the MIRFLICKR Dataset

5. Contribution

In brief, our contribution in this report can be summarized as follows.

- A naive CNN-based initial approach aimed at understanding the limits of the problem.
- Blending a CNN with a pre-trained high-level feature extractor (*Inception-ResNetv2*) as seen in [3] to enhance the coloring process.
- Implementation and analysis of a higher quality model proposed in [4].
- Creation of a simple transformation on images aimed to correct one of the classical flaws of artificial image colorization.
- A glance into the state-of-the-art image colorization techniques.

[1] *http://press.liacs.nl/mirflickr/mirdownload.html*

6. Related Work

For what concerns the different techniques and approaches we implemented in order to perform Image Colorization we can refer to [3] for our model paired with the Inception-ResNet. Instead the source for model described in 7.2 can be found in [4].

Many other approaches were explored to tackle this challenge, such as probabilistic models like Generative Adversarial Networks (GAN), as presented in [5]. GANs are particular networks based on a game theory approach. They consist of two networks, one called the generator and the other the discriminator. The generator aims to fool the discriminator by creating artificially colored images aiming to imitate real ones. The discriminator on the other hand aims to detect which images are real and which are artificially colored.

This approach is valuable because the produced images tend to look more and more real as the training proceeds. Even if this technique is very popular nowadays we decided to consider different approaches.

7. Method and Experiments

We decided to face this problem by building several Deep Neural Networks, that differ in complexity and in the formulation of the objective function. All our models take as input a signal $X \in \mathbb{R}^{224\times224\times1}$, that corresponds to the grayscale of the image, or more precisely the L channel of the image's LAB encoding. Then they try to learn a map $Y = f(X) \in \mathbb{R}^{224\times224\times2}$ that associate to X the values a and b in the LAB color space. This means that all of the images we feed to our network must be resized beforehand to the mentioned input size.

7.1. Euclidean loss based CNNs

For our first trails we adopted an euclidean loss function, since this choice looks valid and this type of objective function is easy to handle. We defined the loss as follows

$$f(\hat{Y}, Y) = L_2(\hat{Y}, Y) = \frac{1}{2}\sum_{h,w} \|\hat{Y}_{h,w} - Y_{h,w}\|_2^2$$

where $\hat{Y}_{h,w}$ is the couple of predicted a and b values for the pixel in position h, w, while $Y_{h,w}$ are the true values.

We firstly decided to implement a simple CNN with 7 convolutional layers and approximately 30k parameters in order to see if this task can be handled by a small network. After several experiments and hyperparameters tuning we noticed that this strategy was providing us with results absolutely far-off from our objective. In particular our network used only red and blue based colors and it was barely able to understand the outlines of the figures appearing in the images. This neural network indeed is not able to retain much information about the visual features of the images since its restricted size and the minimal amount of training did not allow it to generalize adequately.

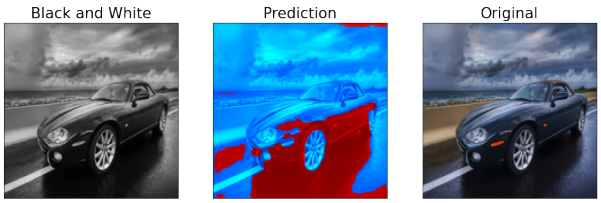

Figure 3. A result obtained with the simple CNN.

Therefore considered the results obtained when trying to work with a modest model, we switched to a more complex method exploiting a model pretrained on the whole ImageNet dataset. In particular we made use of the Inception Res-Net V2, a traditional architecture that has yielded convincing performance in the 2015 ILSVRC challenge: it is a very deep CNN with 55 million parameters. In our case it will work as an a high-level feature extractor in order to provide information about the image contents.

As described in [3] we built an encoder-decoder neural network, where the encoder maps the initial image into a tensor of shape $(28, 28, 256)$. This tensor is then concatenated with the encoding obtained with the feature extraction derived from *Inception*. The resulting encoding corresponds to a tensor of shape $(28, 28, 1000)$ and it carries the high-level information about the image. The obtained tensor is then given as input to the decoder, which outputs the predicted AB channels for the image.

We trained our model on 5000 images using Adam as optimizer. We let the model train for several epochs and this phase took approximately 5 hours of GPU-time on a NVIDIA Tesla P-100. We were forced to train the algorithm with 1000 images a time: this approach indeed needs to compute and store in memory also the Inception Net embedding of the images used during training, resulting to be expensive in terms of resources.

Considering the obtained colorized images we can say that our model is able to identify correctly the sky and grass. This could be also due to the fact that this patterns present themselves several times in the training set. The model indeed seems to overfit sightly on these images since it is a common feature for the predictions to have a blue or green blur on uniform surfaces.

Due to the fact that this model was trained on a limited number of images the net does seem not understand the color of unusual objects, but despite this flaw it can gave pleasing results, as shown in Figure 4.

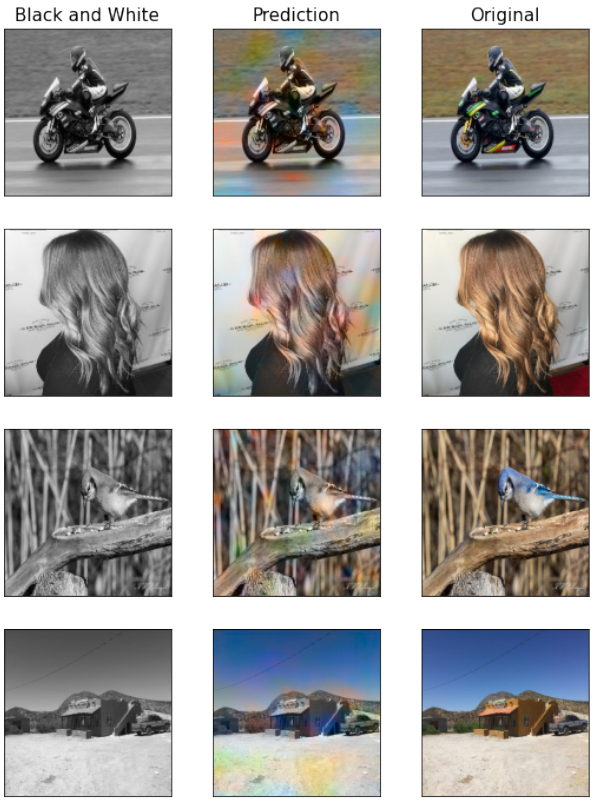

Figure 4. Some results obtained with our model.

7.2. Colorful Image Colorization[4]

In this paper *Zhang et al.* tried a novel approach to solve the image colorization problem. They mention that foregoing works did not address the obstacles that arise when choosing a Loss Function carelessly.

In particular, the problem of antecedent works was that the colored images would often result "dull" or blandly colored.

This unwanted feature typically arises when choosing loss functions based on the Euclidean Error due to the fact that this type of loss functions tend to encourage conservative predictions. This behavior is reflected on the results under the form of under-saturated sparsely colored images in which the only visible colors are brown, grays, light blues and light greens.

To address this problem *Zhang et al.* propose three key procedures. To begin with they propose to select a loss func-

tion carefully tailored to fit the image colorization problem. They implemented a VGG based network with this loss function, and after some training, they propose to re-weight the loss in order to emphasize rare colors. Finally when producing the final colorization they take an "annealed mean" over the color distribution.

The most interesting one of the three for our purposes is the selected loss function.

7.2.1 AB Space Quantization

The proposed procedure to obtain a meaningful loss function is the following. To begin with the output space, namely the a and b channels, is quantized into bins with grid size 10, then 313 of the in-gamut values are selected. The aim now is, given an input X, to learn with training a mapping

$$\mathcal{G}(X) : [0,1]^{W\times H} \longrightarrow [0,1]^{W\times H\times Q}$$

where $Q = 313$ is the number of quantized ab values. As we can see the image colorization task has now been transformed into a classification task. Now to compare $\hat{Z} = \mathcal{G}(X)$ against the ground truth Y we have to encode Y using the quantized ab values. Once we obtained the quantized ground truth Z we can proceed by calculating the multinomial cross entropy between $\hat{Z}$ and Z:

$$L(\hat{Z}, Z) = -\sum_{h,w,q} Z_{h,w,q}\log(\hat{Z}_{h,w,q}).$$

As mentioned before this new approach will not be enough by itself but with the addition of class rebalancing and the final annealed mean it can deliver satisfactory results on several images. However the method proposed by *Zhang et al.* it is not perfect and struggles to provide plausible coloration on a large class of images even after massive amounts of training (as we can in Figure 5).

Figure 5. Some example of failed *colorful colorization*. (Predicted vs Ground Truth)

7.2.2 Artificial Saturation Enhancer

Although *Zhang et al.* actively tried to face the problem of "dull colored" images, the trained and rebalanced net still prefers, on a broad spectrum of images, safe colors, namely light grays, browns and undersaturated colors in general.

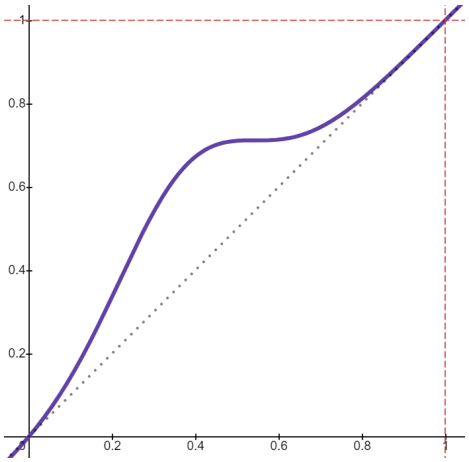

Figure 6. Visual representation of the transformation, with $w = 0.8$, $c = 0.2$, $\alpha = 2$, $\beta = 4$. (Rescaled in the interval $[0, 1]$)

To try to live up some of this images we implemented an artificial saturation enhancer that can be heavily tuned to provide a great gamut of modifications.

They key idea of this method is to selectively enhance the saturation of each pixel based on the current saturation value. In particular we want to boost the saturation of "mild" pixels while keeping nearly unchanged the pixels where the saturation is considerably low or, on the other hand, high enough.

To accomplish this goal we started by changing our setting. We took our colored images and considered them in the HSB Colorspace (*Hue Saturation Brightness*). This colorspace seems to be a natural choice for our task since it allows us to work comfortably only on the saturation channel.

Finally we implemented the following transformation on the saturation channel:

$$x' = x\left(1 + w\left(e^{-|\beta\left(\frac{x}{255}-c\right)|^{\alpha}}\right)\right)$$

where w is the *weight* of the enhancement (a weight of 0 would result in no change), c controls the center of the enhancement (that is to say for which values of x we want the change to be more pronounced), β is a parameter we called *bell tightness* that controls how large the area of affected values around the center is and finally α controls the shape of transformation curve (specifically lower values of α grant a smoother transformation).

8. A glance at State-of-the-art Image Colorization

To conclude our journey in the world of Image Colorization we decided to examine one of the latest published papers on the topic.

Figure 7. Example effect of enhancing. (*Zhang et al. vs Enhanced vs Ground Truth*)

Our choice has been *Wavelet Transform-assisted Adaptive Generative Modeling for Colorization* [6] published on July of 2021.

Jin Li et al. propose an unsupervised learning approach: they begin by carefully choosing a set of wavelet transforms in order to extract more explicit information from the starting images.

After applying these transformations to the input image and obtaining a tensor containing each wavelet transformed signal, they build a generative model that operates with these tensors.

Their ideas are motivated by the manifold hypothesis of machine learning and are backed up by theoretical analysis.

After this initial phase of prior learning from the input the final structure for colorization is built. *Jin Li et al.* implement two other key features in their algorithm. In particular they develop two consistency strategies, one for data consistency, to limit the uncertainty derived from the use of a generative model, and one for structure consistency to avoid some unnatural artifacts initially introduced by their model.

The obtained results deliver the promises and final images are stunning. Colored images do not present artifacts, the hallucinated colors are plausible and well-saturated resulting in overall pleasing outputs.

Although the proposed images seem to be almost perfect there is a caveat to consider. *Jin Li et al.* trained and tested their method on some specific thematic datasets. In particular they mainly focused on a dataset containing churches or pictures of home interiors like bedrooms, living rooms etc.

This bias introduced by the dataset can have some undesired effects on other images as we can see in Figure 8. The first two images look flawless while the second two *out-of-dataset* images have some common problems. The third image present a blue blur on the wall and the fourth image looks grayish, particularly on the sky.

Figure 8. Colorized images from *Jin Li et al.*.

9. Conclusions

With the development of this project we took our first steps towards the deep and complex task of Image Colorization.

Starting with one of the simplest approaches we worked our way up to analyzing state-of-the-art publications.

We comprehended the importance and intricacy of this challenge, the difficulties it poses and the main strategies to alleviate them.

We noticed that each individual framework introduces an unique algorithmic bias in the learning process. This bias needs to be recognized and addressed case by case in order to deliver distinct results.

In particular we can say that training image colorization models on specific datasets seems to be a common practice and almost mandatory in order to produce finer results at the expense of narrowing the ability of generalization of the model.

Furthermore when building an image colorization model it is crucial to try to squeeze the most amount of information possible form the starting image in order to provide a realistic and coherent final result.

Having said this we have to remember that this task demands for major amount of resources in terms of training data and computing power.

References

[1] Mohammad Haris Baig and Lorenzo Torresani. "Multiple hypothesis colorization and its application to image compression". In: *Computer Vision and Image Understanding* 164 (2017). Deep Learning for Computer Vision, pp. 111–123. ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2017.01.010. URL: https://www.sciencedirect.com/science/article/pii/S1077314217300267.

[2] Yinxu Bian et al. "Deep learning virtual Zernike phase contrast imaging for singlet microscopy". In: *AIP Advances* 11.6 (2021), p. 065311. DOI: 10.1063/

5.0053946. eprint: https://doi.org/10.1063/5.0053946. URL: https://doi.org/10.1063/5.0053946.

[3] Federico Baldassarre, Diego González Morın, and Lucas Rodés-Guirao. "Deep Koalarization: Image Colorization using CNNs and Inception-ResNet-v2". In: *CoRR* abs/1712.03400 (2017). arXiv: 1712.03400. URL: http://arxiv.org/abs/1712.03400.

[4] Richard Zhang, Phillip Isola, and Alexei A. Efros. *Colorful Image Colorization*. 2016. arXiv: 1603.08511 [cs.CV].

[5] Phillip Isola et al. "Image-to-Image Translation with Conditional Adversarial Networks". In: *CoRR* abs/1611.07004 (2016). arXiv: 1611.07004. URL: http://arxiv.org/abs/1611.07004.

[6] Jin Li et al. "Wavelet Transform-assisted Adaptive Generative Modeling for Colorization". In: *CoRR* abs/2107.04261 (2021). arXiv: 2107.04261. URL: https://arxiv.org/abs/2107.04261.