Image Colorization with Convolution Block Attention Modules

Chang Liu and Yujie Tu

New York University

Abstract

This paper introduces a new technique for image colorization. Convolution block attention modules (CBAM), as well as a classification network, are innovatively incorporated in a colorization model based on patch generative adversarial networks (PatchGAN). The effectiveness of CBAMs in improving the model's performance is proven.

Introduction

Old photos are fascinating. While people are obsessed with the distinction in black and white pictures, they are curious about how the missing colors may redefine the vision and mood of an image. Such "curiosity" is crucial to humanity and social sciences researchers, who have tried hard approximating the colorful reality behind the monotony of historical photos.

Image colorization, a possible solution to the problem addressed above, has been of great interest to computer vision scientists for over a decade. The earliest fully automatic approach is presented by the paper "Colorful Image Colorization" in 2016. The authors regard colorization as a classification task and structure their model mainly using convolutional neural networks (CNN). Nowadays, one of the best and most recognized colorization algorithms is DeOldfy, which manages to colorize both pictures and videos with high efficiency and low bias. DeOldfy is creative in bravely harnessing a NoGAN technique, where both the generator and the

discriminator are generally trained with the most conventional methods and only with GAN techniques later for a short time.

Undeniably, prior work on this topic is outstanding and has provided excellent solutions for the problem. We, however, believe that new features, integrated with the current achievements, can further enhance reliability, sensibility, and vibrance of image colorization.

Datasets

We use a 6 GB dataset of 50,000 images from ImageNet and a 20+ GB dataset, named Places365, of 1,800,000 images. Given the limited computational capacity of our laptops, all images are resized to 128×128 pixels before input into our model. We also change the color space used by the model from RGB to LAB, since we believe that LAB, capable of keeping exact contours of a picture, is a better fit for the training process.

We first train the model on the ImageNet dataset, but the results after 100+ epochs are still disappointing. We carefully examine the datasets used by previous colorization research and find that all successful models are trained on much larger datasets. Hence we change the dataset our model is trained on to Place365, and only after four epochs, the output has been greatly improved.

Methodology

The model consists of three structural parts: PatchGANs, a classification network, and CBAMs. PatchGANs act as the model's skeleton, strengthened by the classification network and CBAMs.

The concept of PatchGANs is borrowed from the paper "Image-to-Image Translation with Conditional Adversarial Networks." While a GAN discriminator outputs a single binary value, a PatchGAN one generates an $n \times n$ matrix that outputs the binary values of every patch. To better fit our model, We slightly modify PatchGAN discriminators to make them output an $n \times n \times 2$ tensor whose two layers correspond to the two LAB color channels respectively. We also make changes to the loss function proposed by the paper:

$$L = loss_{gan} + loss_{classification} + pixelDsitance \times 200$$

By examining the outputs of each epoch, we find this loss function ineffective since it attaches fixed, excessive importance to pixel distance. Inspired by the NoGAN technique in DeOdify, we make the function gradually highlight the GAN loss by subtracting a fixed number from the coefficient of pixel distance after each epoch.

After constructing the PatchGAN skeleton, we inject a classification network capable of identifying the main object in an image. With the add-on providing additional information, the model, knowing exactly what object it is coloring, is expected to output a more plausible colorization result. The paper "Let There Be Color!: Joint End-to-End Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification" helps concretize our idea by providing an implementation framework for the classification network. Eventually, We manage to implement the network and have it output an information map that perfectly fits the feature tensor.

Furthermore, We boldly plug CBAMs in the PatchGAN skeleton. Attention mechanisms are originally proposed and have been widely used in natural language processing, and we try to examine whether they will work wonders for image colorization as well. In the feature network, every channel concentrates on a specific feature, and a CBAM, by focusing on a single feature each time, can simulate how a human being observes and learns in real life. With the CBAMs between every two layers of the network creating an attention map, the model can pay the most attention to the principal object while colorizing an image. Computational resources assigned to CBAMs are negligible, for these additional modules are lightweight compared to the whole architecture.

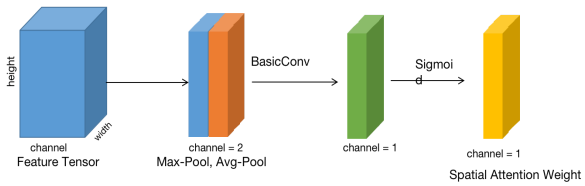

Figure 1.1: Spatial Attention Module

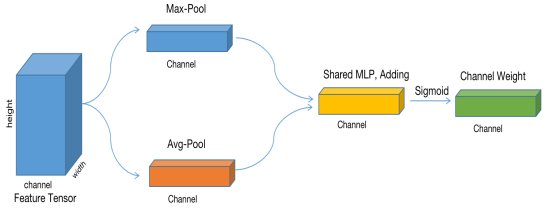

Figure 1.2: Channel Attention Module

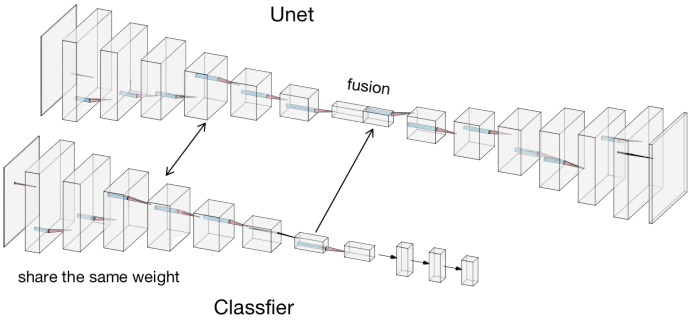

Figure 2: Complete Architecture of the Model

Results

After equipping the PatchGAN skeleton with the classification network and CBAMs, we repeat the training process and compare the results with those generated by the initial model. An exciting increase in the performance, especially on lanscape images, is clearly shown.

Since this time we use the Places365 dataset, most of the images the model has been trained on are featured by landscapes. Green and blue are omnipresent in landscape photos, thus the model now does an excellent job in recovering these two colors. The model, nevertheless, is fairly irresponsive to red. For instance, color changes on objects that are supposed to be red, such as apples, lanterns, and lips, can barely be perceived in the output. These objects usually remain pale after colorization. This should not be a predicament, though, as we have proved that the number of training samples can greatly influence training results, thus images containing abundant red elements can certainly help improve the model's sensitivity to red.

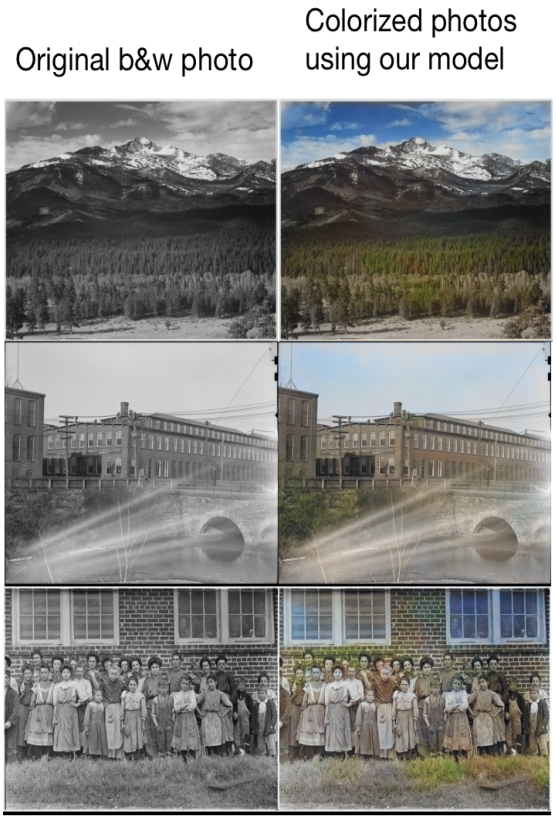

Figure 3: Input versus Output

Future Work

If we have a few more weeks to work on this project, we would train the model for more epochs. Four epochs are insufficient for a model to learn to complete such a complex task, even though quite a few satisfying results have been generated even after such a short learning time. Then we would further adjust the model's parameters to maximize the increase in the performance due to CBAMs. And to improve the performance in a more general way, it is important to train the model on a more balanced number of images of different contents, including but not limited to landscapes, still objects, and portraits.

Reference

Iizuka, Satoshi, Simo-Serra, Edgar, and Ishikawa, Hiroshi, "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, 2016.
http://iizuka.cs.tsukuba.ac.jp/projects/colorization/en/

Isola, Phillip, Zhu, Jun-Yan, Zhou, Tinghui, and Efros, Alexei A., "Image-to-Image Translation with Conditional Adversarial Networks." CVPR, 2017. https://arxiv.org/abs/1611.07004

Woo, Sanghyun, Park, Jongchan, Lee, Joon-Young, and Kweon, In So, "CBAM: Convolutional Block Attention Module." ECCV, 2018. https://arxiv.org/abs/1807.06521

Zhang, Richard, Isola, Phillip, and Efros, Alexei A., “Colorful Image Colorization.” ECCV, 2016. https://arxiv.org/abs/1603.08511

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. “Places: A 10 Million Image Database for Scene Recognition.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017. http://places2.csail.mit.edu/index.html