Improving Parametric Colorization Methods Using Discriminators

Emily Mu
Massachusetts Institute of Technology
6.869
emilymu@mit.edu

Ka Wai Lee
Massachusetts Institute of Technology
6.869
kwlee@mit.edu

Abstract

We utilize a discriminator trained on top of a state-of-the-art colorization generator to improve plausible colorization of specifically difficult grayscale inputs. Determining a plausible colored version of a grayscale photograph is inherently an underconstrained problem so early data-driven attempts at producing plausible colored often resulted in undersaturated non-realistic images. Current state-of-the-art approaches are trained on millions of colored examples and incorporate intermediate steps and tailored loss functions to mediate undersaturation. Inspired by the success of generative adversarial networks on image-to-image translation problems, we built a discriminator on the results of a state-of-the-art colorizer to demonstrate the improvements that can still be made in this field. We demonstrate that discriminators can still determine the difference between computer-generated and actual colored photographs with a high-degree of accuracy, picking up on the same color confusions and inconsistencies that human users do. We then use the results of our discriminator to fine-tune the original generator to produce more vibrant and plausible colorizations for failure cases.

1. Introduction

Coloring a grayscale image in a vivid and realistic way is simple for the human imagination (Figure 1). However, the range of realistic colorizations can be very difficult to determine since many possibilities exist and a third of the information has been lost. Our imagination fills in some of this information with previous experience, we know logically that grass tends to be green and the sky tends to be blue. Much progress has been made already in generating photo-realistic images (Sources). In fact, the colorization task has been found to be a competitive method for self-supervised representation learning (Source).

Current methods of evaluating synthesized images include testing human observers by asking observers to distinguish computer generated and photo-realistic images. Human observers are presented with the ground truth colorization and the synthesized colorization and asked to identify the fake image. Current state-of-the-art colorizers are able to fool participants over 30% of the time, with 50% being equivalent to ground truth colorization (Sources). Some of these high-performing colorizer architectures will be discussed in related work, but it is important to note that many of these incorporate loss functions specifically to reduce saturation loss (Sources).

Relatedly, conditional adversarial networks (cGANs) have been shown to be flexible general-purpose solutions for image-to-image translation problems (Source). These translation problems are defined by translating one image representation to another and colorization is such and example. Conditional adversarial networks consist of training a generative model and a discriminator concurrently. The discriminator attempts to determine if the output of the generator model is real or fake. Consequently, cGANs have been shown to be able to learn a loss function that adapts automatically to the data and perform well on many image-to-image translation problems. By the same method of evaluation described above, a cGAN trained with additional L1 loss was able to fool observers 22.5% of the time.

Although cGANs have been shown to work well on image-to-image translation problems, state-of-the-art colorizers do not train with cGAN loss. Rather, they tend to define specific loss functions for more traditional convolutional neural networks to reduce saturation loss. In this paper, we explore training a discriminator on top of the results of the best generators to look at how to improve existing colorizers. We had a two-fold goal for this project: (1) to understand and (2) to improve failure cases for existing state-of-the-art colorizers. We used the generator designed by Zhang et al. for this project. After training a discriminator on ground truth and images generated by the algorithm, we find that the discriminator can determine the difference between fake and real images with an error of only 2.8%. We examine these error cases and the results of the discriminator to demonstrate that the discriminator can find the similar differences between fake and real images that human

perception can, including ambiguous coloration, inaccurate long-range consistency, and unnatural tones (Source). We then utilize our discriminator to identify and fine-tune the existing generator for the most difficult failure cases and demonstrate that we can construct more feasible colorizations for these cases (footnote: code).

Our contributions in this paper are as follows: to allow for better understanding of areas for improvement in image colorization, to provide a framework to automatically identify and tackle failure cases of existing generators, and to demonstrate the success of this framework to improve these cases.

1.1. Related Work

Colorization algorithms usually take one of two forms: non-parametric and parametric. Non-parametric models tend to use user-guided inputs or references from scribbles to determine accurate image colorization (Sources). Parametric models utilize large datasets by learning how to predict color as either a continuous or quantized output and tend to be more self-supervised or automatic (Sources). This paper will be focused of improving this second class of parametric models.

We define state-of-the-art parametric colorizers as methods that fool observers with over 30% reported accuracy, as defined in the introduction. Several recently developed architectures exist in this category. One method developed by Zhang et. al poses the colorization problem as a classification tast and used class-rebalancing at training time to capture a wider variety of colors. Concurrently, a method developed by Larsson et al. uses a modified VGG network to predict per-pixel histograms. Iizuka et al. also developed a similarly well-performing method by using a two-stream architecture to capture both local and global features of the image.

Conditional adversarial networks (cGANs) have been shown to work well for general image-to-image translation problems, achieving a 22.5% accuracy on the colorization problem. Traditional cGANs train a generator (conditioned on some input) and a discriminator concurrently with the generator attempting to minimize discriminator accuracy on real and fake data. Inspired by the cGAN architecture, we train a discriminator on top of an already tuned parametric generator in order to evaluation generation results and learn how to improve training.

We selected Zhang et al as the example generator for improvement because the source code and examples are open source and well documented. Comparisons made in the Zhang et al paper show that the results of their method are comparable to the other methods described (Source). Furthermore, their method is hosted on Algorithmia, allowing for easy data generation and testing (Source). The specifics of their generator model will be described in the next section.

2. Approach

We train a CNN to classify generated and ground-truth colored imaged using the architecture shown in Figure 2. In this section, we discuss the open source architecture of the generator designed and implemented by Zhang et al. and the architecture of the discriminator we implemented which can be found on our public repository (code).

2.1. Generator Architecture

The generator method was designed by Zhang et al., and we provide a brief description of method here. Further analysis of their method and its performance can be found in their paper. Given an input grayscale image, one potential method trains a CNN to map that image input channel X to two associated color channels $\hat{Y} = F(X)$. The most natural loss function for CNNs is the Euclidean loss function between ground truth and predicted values:

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

The problem with just using this loss function, however, is that the optimal solution is a mean of all plausible values, which results in grayish, unsaturated images. Consequently, to improve this, the problem is treated as a classification instead, where the input is mapped to a distribution over the quantized ab color space. Instead of learning $\hat{Y}$, we learn $\hat{Z} = G(X)$ where $\hat{Z}$ is a probability distribution over possible colors. We can then convert our ground truth color Y to a vector Z by function $Z = H_{gt}^{-1}(Y)$, using a soft-encoding scheme. Given this definition of Z, we can now use multinomial cross entropy loss to define our new loss function

$$L_{cl}(\hat{Z}, Z) = -\sum_{h,w} v(Z_{h,w}) \sum_{q} Z_{h,w,q} \log \hat{Z}_{h,w,q}$$

v is a weight that allows for the rebalancing of the loss baes on color-class during training. This is necessary because natural images tend towards desaturated or low ab values which dominate the loss function. Thus, in order to prevent this, we weight the loss of each pixel based on color rarity, analogous to resampling the training space. We define function v as follows in terms of its closest ab bin $q*$.

$$v(Z_{h,w}) = w_{q*}, \text{where } q* = argmax_q Z_{h,w,q}$$

In order to estimate the weighting factor w, we take the smoothed distribution of colors and weight it weight it with a uniform distribution. We then make the weighting factor

inversely related to this new distribution. We then normalize so that the expected value of the weighting factor is 1. The specific experiments to calculate and determine this weighting can be found in the original paper.

Finally, we must compute the function H that maps the color distribution $\hat{Z}$ to a single estimate in ab space $\hat{Y}$. Taking the mean has a similar problem to using the Euclidean loss–results tend to be desaturated. However, taking the mode of the distribution for each pixel also results in very splotchly, spatially inconsistent results. Thus, we compute a combination anneal-mean of the distribution as follows

$$H(Z_{h,w}) = \mathbb{E}[f_T(Z_{h,w})], f_T(z) = \frac{\exp(\log(z)/T)}{\sum_q \exp(\log(z)/T)}$$

Experiments with annealed mean demonstrated that a temperature of $T = 0.38$ worked well for this function. The final system composes the CNN G with this annealed-mean operation H to construct a final determined image output.

2.2. Discriminator Architecture

3. Results

Make figures

3.1. Training the discriminator

3.2. Discriminator Performance

3.3. Qualitative Observations

3.4. Fine-Tuning the Generator

4. Conclusion and Future Work

5. Acknowledgements

The authors would like to thank Professor Freeman and Professor Torralba for teaching this course and the rest of the 6.869 course staff for all their help in understanding the material and for the inspiration and review of this project.

References

[1] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. Places2: A large-scale database for scene understanding. *Arxiv, 2015.* places2