Improving Data-Trained Colorization 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

The human ability to recognize and understand an entire scene at initial glance is an incredibly amazing process that has yet to be replicated by computational techniques. Scene recognition can help provide context for more complex computer vision tasks, such as image-to-image or image-to-description translation problems.

Currently, deep convolutional neural networks have been found to have high performance on object recognition tasks, approximately equivalent to human performance. However, this level of performance has yet to be replicated for scene classification. In order to help remedy this, the Places2 dataset was created with over 10 million images in over 400 categories to aid in the training of deep neural nets for scene classification tasks [1, 2]. In this paper, we utilize a subset of this dataset, 100,000 images for training, 10,000 each for validation and testing from 100 different scene categories, to train different deep network architectures.

We tested a variety of different architectures of different sizes, including VGG11, VGG16, AlexNet, and ResNet architectures. In order to reduce overfitting, we incorporated batch training, data augmentation, dropout, and parameter reduction techniques. In section 2, we describe our approach to the scene recognition problem, including our pipeline, architecture designs, and methods to reduce overfitting. In section 3, we present and analyze our experimental results.

1.1. Current Architectures

Both of us discussed, designed, and implemented the different architectures and methods to reduce model overfitting. Ka Wai (Joanne) headed the training of the different architectures and Emily led the writing of this paper. We believe the overall division of work and implementation was fair and equal.

2. Approach

Since the MiniPlaces dataset is limited in size and we were limited in computation time, we selected a few different architectures to test and augmented each with different overfitting techniques in order to compare model performance. We first performed a literature search to select CNNs to implement and test and to select optimal overfitting reduction methods. We also implemented several methods to reduce training convergence time.

2.1. Generator Architecture

2.2. Discriminator Architecture

From our initial literature search, we found that deep networks, including the VGG16/VGG19 networks and the GoogleNet performed very well on related image classification problems [3, 4]. Due to the limited scope of our problem, we found that very deep networks tended to easily overfit. Consequently, we tested three main networks, the original AlexNet, a VGG16 network, a smaller VGG11 network, and a ResNet . The input to all of these networks are images of size $128 \times 128 \times 3$, corresponding to the RGB image sizes from the Mini Places Challenge dataset.

3. Results

3.1. Training the discriminator

Originally, we used Keras because it supplied us with advanced data augmentation techniques such as feature-wise normalization. However, these techniques ended up taking too much memory and proved to be infeasible for our implementation.

3.2. Discriminator Performance

We measured the performance of our networks based on top-5 and top-1 error. For each image, the network predicts the top 5 classes that the image is most likely in. We mark down whether the correct class is the most likely predicted class, or whether the top class is within the predicted top 5. We compute these values for the entire validation set to get the top-5 or top-1 error. We chose the best network first based on top-5 error, followed by top-1 error.

3.3. Qualitative Observations

Generally, deeper networks (VGG16) tend to perform better than more shallow networks (VGG11, AlexNet), but because the size of our training data is limited, deeper networks tended to overfit and thus actually performed worse than shallow networks.

4. Conclusion

In this paper, we explore the application of a variety of different deep network architectures and reducing network overfitting techniques on the MiniPlaces Challenge Dataset for scene classification. We find that a moderately deep CNN with a limited number of parameters performs best in this case with a top-1 error of 58% and a top-5 error of 27%. Deeper architectures may work even better with longer periods of training and larger training sets. However, we were unable to explore these options due to the limited size of our training set and the limited computation time of this project.

5. Acknowledgements

References

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