Let there be color: Deep Learning Image Colorization

Justin Olah and Jenny Yang
Department of Computer Science, Stanford University

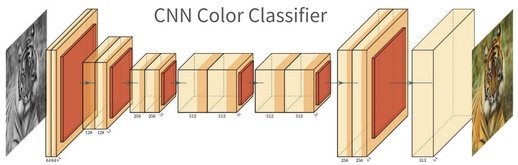
Motivation

Ever since the dawn of black and white photography, people have been searching to add color to decades of black and white photographs, videos, and sketches. With the improvement of artificial intelligence, colorization has become possible as a machine learning task. Models that learn to color well often have learned other extremely relevant features about the image, such as segmentation, texture, and depth.

Objective

Given a black and white image, we want to produce a colorized version of the image. We convert our images into the AB color space. Our models take as input a black and white image and outputs the predicted AB channels. We then add the AB channels to the black and white image to produced an image in color.

Methods



Regression: As a baseline, we train a regression model with L2 loss.
Classification: We treat colorization as a classification task. We quantize the AB color spectrum into 313 color bins. For each pixel we then predict a color bin. Our base model is a CNN in two parts, the encoder portion that downsamples the input dimensions and learns features, and a decoder portion that upsamples the dimensions and adds color.
U-Net: We modify our CNN with the addition of skip connections from the encoder portion to decoder to create a U-Net.
Class rebalancing weights: To counteract the model predicting dull colors to minimize loss, we weight the loss of each pixel based on the rarity of the quantized color bin. The updated cross entropy loss with class weights is shown below.

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

The loss now includes a weighting term w, and sums over each quantized color bin q.

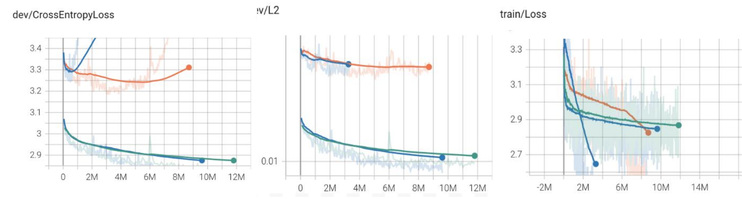
Experiments

We ran each of the following models on each dataset. We trained using the Adam optimizer with learning rate 3e-5, until loss plateaus:

1. Baseline Regression
2. CNN Classifier
3. CNN Classifier with Class Weights
4. Unet Classifier
5. Unet Classifier with Class Weights

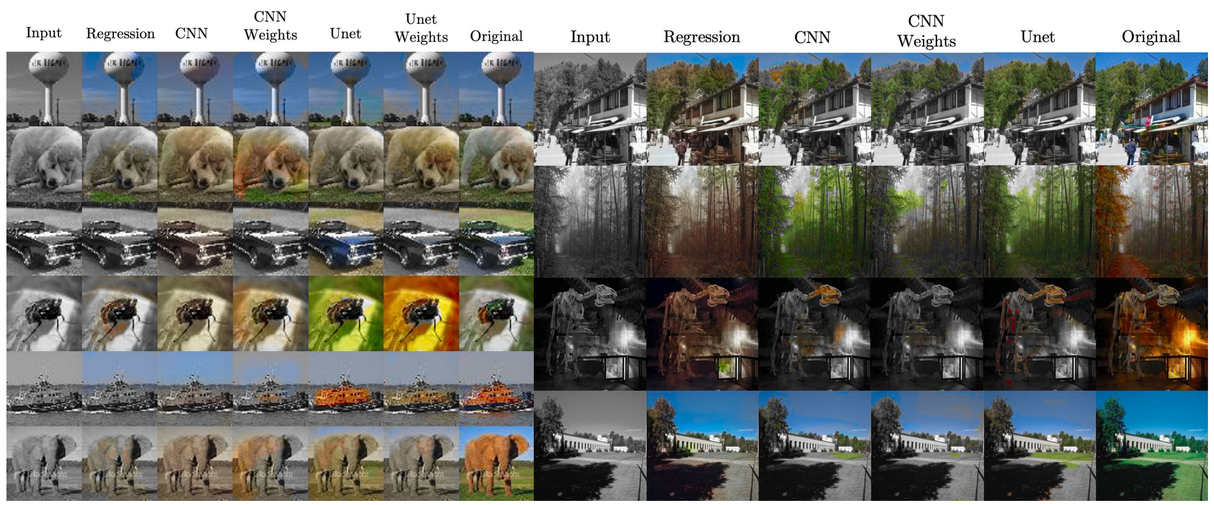
Best Model: Unet Classifier trained on MIT places.

Loss curves for all models without class weights:



Upper blue: CNN on Tiny Imagenet, Orange: Unet on Tiny Imagenet, Lower blue: Unet on MIT Places, Green: CNN on MIT Places.

Tiny ImageNet | **MIT Places**



Analysis

TIny Imagenet photos colorized by U-Net trained on MIT Places.

Overfitting: We observed strong overfitting while training, especially when training CNN on TinyImagenet. However, train loss and visual results on the val set continued improving. Resolved by training on MIT Places and with U-Net.
Robustness: We applied models trained on one dataset onto the other, found that TinyImagenet models generalized best due to diversity of images in dataset.
Common Errors: MIT Places learned to color green and blue frequently due to prevalence of outdoor scenes in dataset.
Quantitative metrics not good evaluation metric: Loss metric does not fully capture performance of model and visual improvement. We compared relative model performance by spot checking output rather than comparing loss values.

Bad colorization samples from regression and classification models trained on MIT Places.

Datasets

We experimented on Stanford's Tiny Imagenet dataset and the MIT Places dataset. Tiny imagenet is a 100,000 image dataset of 64x64 scaled down Imagnet Images. MIT Places is a 1.8 million image dataset of 256x256 images.

Conclusions

We conclude that the U-Net without class weights performs the best. This suggests that the skip connections were helpful in accelerating learning by improving upsampling and preventing neurons from dying. While the baseline regression produced smoother colors they were often less saturated and more dull. We also found that the class weights did not have a significant effect in predicting brighter colors.

Faces colorized by CNN trained on MIT Places.

Semantics: While learning to color, the models also learned to identify and classify certain objects like the faces above , meaning colorization models can be leveraged to perform other tasks like semantic segmentation.

Future Work

Training data: Colorization requires a large amount and variety of data. In future experiments, we could train on the MIT Places365 Challenge 2016 dataset which contains 8 million training images. We could also use an ensemble of datasets to improve robustness.
Evaluation: Another evaluation method could involve observing semantic information learned in our models by using out models' autoencodings to train a classifier in an object recognition or classification task.

References

R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In ECCV, 2016.
Y.Le and X.S. Yang .Tiny image net visual recognition challenge. 2015.
B. Zhou, A.Lapedriza, A.Khosla, A.Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.