ColorNN Book

A Recurrent-Inspired Deep Learning Approach to Consistent Video Colorization

Divyahans Gupta (dgupta2)
Sanjay Kannan (skalon)

TL;DR

Can a neural network infer color from the spatial and temporal features of a video? We built a convolutional neural network (CNN) architecture, inspired by recurrent learning, to address this problem.

Data

Our data consist of two separate sets: a corpus of 3500 images with water features (coasts, lakes, and other seascapes), and 10 videos of beach scenes.

We trained our full architecture on public domain videos, which we obtained from Pixabay.com. We trained a convolutional neural network component on the water images, which we obtained from the McGill Calibrated Colour Image Database, MIT CVCL's Urban and Natural Scene Categories, and MIT CSAIL's SUN database.

Here, our models are constrained to a single image genre. Proving their feasibility on a more diverse data set would significantly increase the computational complexity of training.

Features

Every image and video frame was mapped from the RGB color space to YUV. The three channels were used as the inputs and desired outputs of our models. The Y channel was the grayscale input to our models, while our models predicted the U and V color channels.

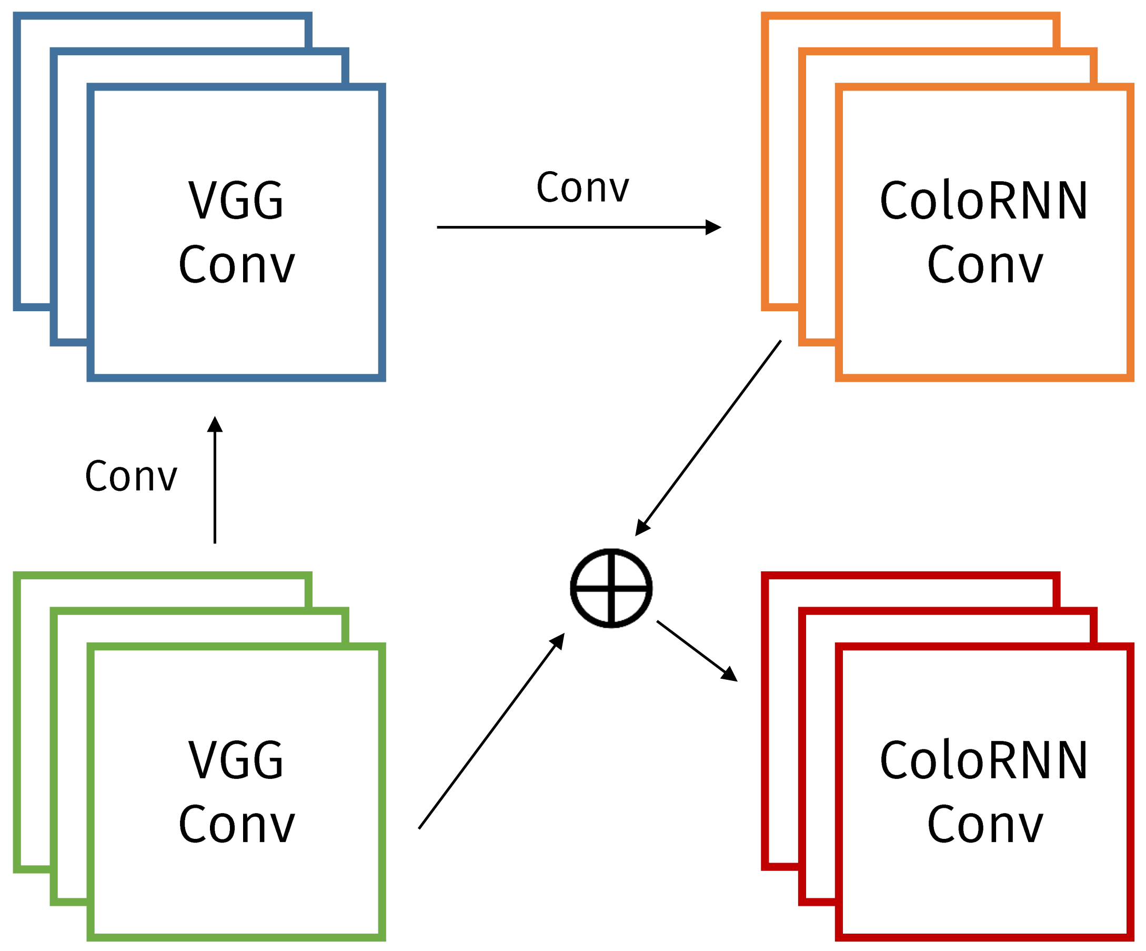
To employ transfer learning with the VGG-16 image classification network, we rescaled images to 224 by 224 pixels. Also, each color channel's range was quantized into 30 bins for classification.

Models

The first component of our architecture was a CNN trained to colorize static images. Given our computational constraints, it was infeasible to train a state-of-the-art model from scratch. Thus, we integrated a pre-trained instance of VGG-16 to extract a diverse set of latent features, and formed residual connections between the hidden layers of VGG-16 and those of our network.

Next, we applied our trained colorization network to sequential video data. Independently colorizing the frames of a video discards their sequential relationship, creating a jittery coloring. Instead, our network colors a frame independent of prior frames, and then recolors it based on the accurate binned coloring of its predecessor.

We framed our tasks as color classification, rather than regression, problems. This decision is intended to avoid desaturated, averaged colorings caused by minimizing the mean-squared error between pixel values. For this reason, we used a categorical cross-entropy loss on the quantized outputs of the U and V color channels.



Visual representation of the residual connections between ColoRNN and VGG-16.

Results

For our color model, we randomly allocated 3000 images to our train set and 500 images to our test set. Our color model trained for over 100 epochs but started to overfit after epoch 16.

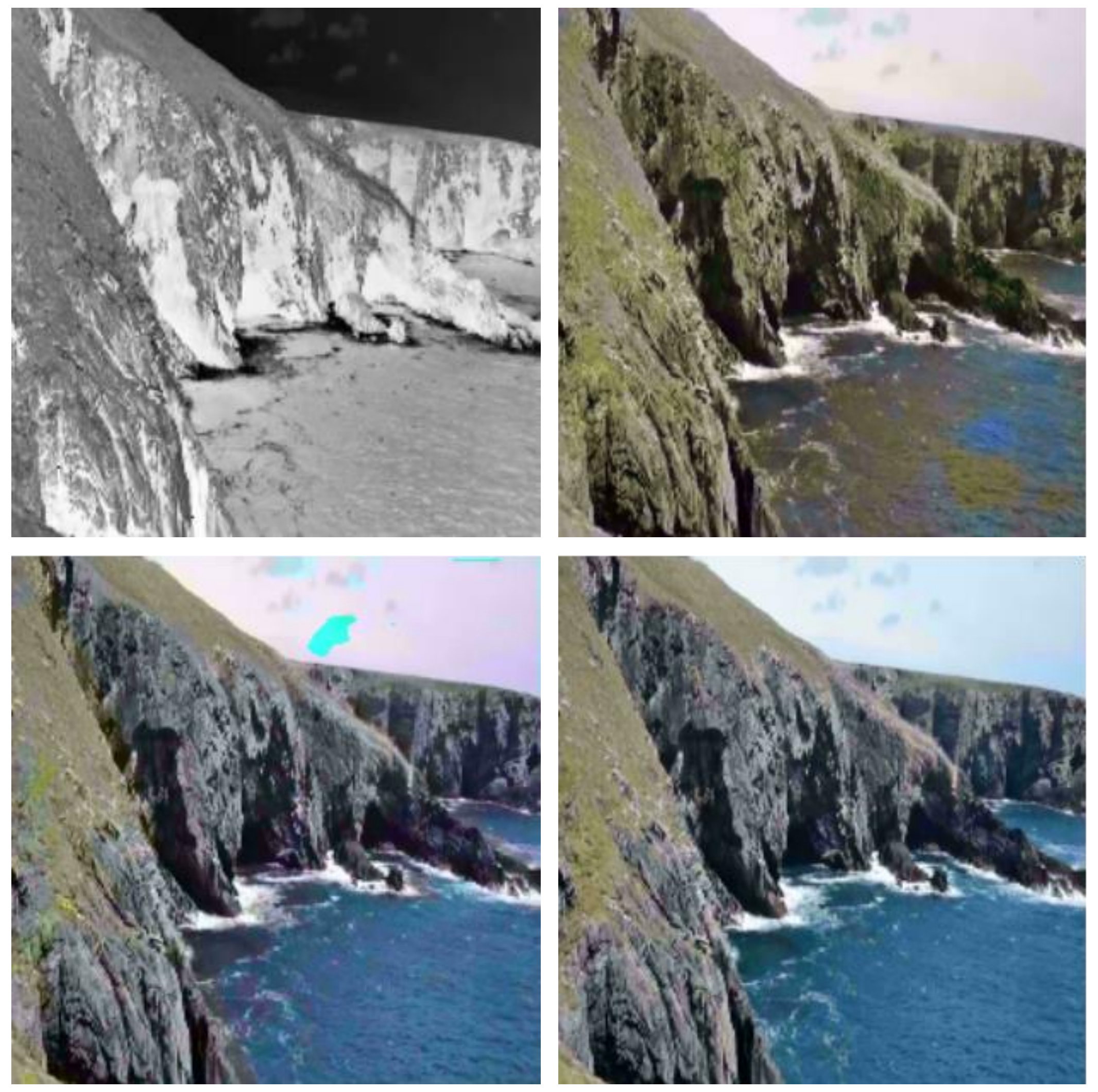
For our consistency model, we randomly allocated seven videos to our train set (300 frames per video, or 2100 frames total), and three videos to our test set (900 frames total). However, we actually use data from adjacent pairs of frames when running our model.

	Color Model Only	Full Consistency Model
U-Channel Train Accuracy	37.362%	46.608%
U-Channel Test Accuracy	36.198%	34.203%
V-Channel Train Accuracy	33.908%	85.414%
V-Channel Test Accuracy	32.474%	75.345%

Discussion

Given that we quantized the color ranges to 30 bins, our classification accuracies are noteworthy. Nevertheless, we should note that frames were extracted from our video set at a rate of 30 FPS. It is possible that the deviations between frames were minimal, so the model may bias heavily toward replicating a given frame's predecessor. Further work will explore this phenomenon.

Several other factors also deserve consideration. First, we are not certain that minimizing cross-entropy loss and maximizing subjective visual similarity are strongly correlated. It is possible that other cost functions may produce better visual results. Additionally, by framing our task as a multi-class classification problem, our model does not understand the closeness of similar colors. Finally, our training runs were limited largely by the number of parameters to learn. There are a multitude of relevant architectures we could not practically test.



Top: Grayscale Image, CNN Coloring
Bottom: Consistent Coloring, Actual Image

Future

We hope to design more nuanced architectures for both the frame coloring and temporal consistency models. In particular, we could likely improve consistency with an LSTM-based architecture, which would learn longer-term dependencies between the convolutional outputs of several prior frames. We found no pre-built architecture of this kind.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[2] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. *arXiv preprint arXiv:1603.08511*, 2016.

[3] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 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 (TOG)*, 35(4):110, 2016.