A Deep Learning Approach to Improved Video Colorization

Yang Fang, Vedi Chaudhri
yangfang@stanford.edu, vchaudhr@stanford.edu
Department of Computer Science, Stanford University

CS 230, Winter 19

Overview

This project applies deep learning to the task of grayscale video colorization. One current technique is to independently color each individual frame using an image colorization neural network.

B/W Colorized True

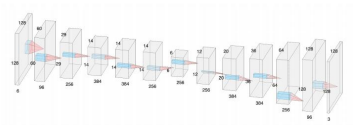
This, however, leads to temporal inconsistency in the coloring of consecutive frames. This project proposes using the previously colorized frame as input to help colorize the current frame in a more positionally consistent manner.

Model

Dataset:

This project uses the MIT-IBM Watson AI Lab's open source Moments in Time dataset, a collection of 1 million labeled 3 second videos. The hiking sub-dataset was first converted to grayscale and colorized using a pretrained image colorization network, and then center-cropped to produce 128 x 128 colorized videos for training. We then constructed a dataset consisting of labeled examples of (X=[prev. colorized frame, cur.colorized frame], Y=[cur. true frame]).

AlexNet-based CNN:



Input shape: 128 x 128 x 6, Output shape: 128 x 128 x 3
Fully connected (FC) layers replaced with deconvolutional layers

VGG16-ImageNet-based CNN:

Fine-tune the Keras VGG16 ImageNet pretrained model.
All VGG16 layers are frozen, concatenation and deconvolutional layers added to correct for different input/output shapes.

Objective Function

Weighted MSE:

Essentially, we are calculating the weighted sum of MSE(cur. pred. frame, true frame) and MSE(cur. pred. frame, prev. pred. frame).

$$\beta * \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 + (1-\beta) * \frac{1}{N}\sum_{i=2}^{N}(\hat{y}_i - \hat{y}_{i-1})^2$$

If β, a tunable parameter, is 1, then we only care about the true frame (independent frame colorization); if β=0, then we only care about the previous frame (results in duplication of previous frame). In practice, β=0.8 was selected, with the intuition that matching the true frame more closely is the more important metric.

Results

Video Colorization of Training Example:

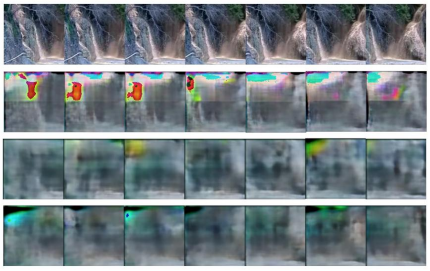
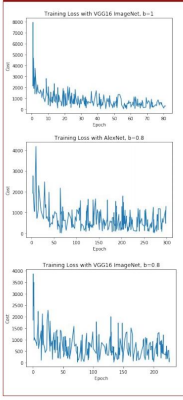


Figure 2: Video Colorizations. Top to bottom: (1) True colorization, (2) AlexNet CNN, β=0.8, (3) VGG16 ImageNet CNN, β=1, (4) VGG16 ImageNet CNN, β=0.8

Analysis



The VGG16 ImageNet with β = 1 model converges most quickly, is less noisy, and has a smaller stable cost, resulting in colors closer to the ground truth than the other results (see Figure 2). In all cases, there were likely insufficient training examples for the models to learn both object shapes as well as to optimize for a similar colorization between frames, resulting in blurry frames.

The VGG16 likely performed better because it fine tunes a pre-trained ImageNet network, and thus performs better given few training examples. Additionally, a model might perform better when β = 1 compared to when β = 0.8 because there's a slight mismatch between the input and the loss function since the input is using the previously colorized frame, whereas the loss is using the generated previous frame.

Challenges and Future Work

- Possibility of not having enough data to fully train our model
- Picking a value for β is highly subjective, and it's difficult to know if the selected β=0.8 is the optimal choice
- The majority of pixels in a video tend to have low values (especially if resized and left uncropped), which often pushed models towards predicting dark / black frame
- To perform better in the future, we could train with a larger data set, train for more epochs, or use a different model architecture

References

[1] Divyahans Gupta, Sanjay Kannan. *ColorNN Book: A Recurrent-Inspired Deep Learning Approach to Consistent Video Colorization.* 2016, http://cs229.stanford.edu/proj2016/report/KannanGupta-ColorNNBook-report.pdf.
[2] Gael Colas, Kevin Lee, Rafael Rafailov. *Consistent Video Colorization.* 2018, cs230.stanford.edu/projects_fall_2018/reports/12444081.pdf.
[3] MIT-IBM Watson AI Lab. "moments in time" dataset. http://moments.csail.mit.edu
[4] Richard ZHang, Phillip Isola, and Alexei Efros. Colorful image colorization. 9907:649-666, 10 2016. https://github.com/richzhang/colorization.