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  Colored  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.

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).

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

If \( \beta \), a tunable parameter, is 1, then we only care about the true frame (independent frame colorization). If \( \beta = 0 \), then we only care about the previous frame (results in duplication of previous frame).

In practice, \( \beta = 0.8 \) was selected with the intuition that matching the true frame more closely is the more important metric.

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 training 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 \( (X=\text{true frame}, \text{y} = \text{true frame}) \).

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.

Results
Video Colorization of Training Example:

Challenges and Future Work
- Possibility of not having enough data to fully train our model
- Picking a value for \( \beta \) is highly subjective, and it’s difficult to know if the selected \( \beta = 0.8 \) is the optimal choice
- The majority of pixels in a video tend to have low values (especially if resized and left uncropped), which often push 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