Mission Statement
Automatically select an appealing thumbnail from within the frames of a video.

Background
- The thumbnail of a YouTube video is the image that a user sees before clicking on the video. Naturally, this has a large affect on the success of the video.
- Experienced YouTubers often create and upload custom thumbnails, but newer content creators often let YouTube choose a thumbnail for them, in which case it comes from within the video.
- Yang and Tsai, 2015 [1] used CNNs to select good thumbnails.
- Liu et al., 2015 [2] also investigated thumbnail selection, but focused on thumbnail-query relevance.

Dataset
In order to train a network that could rate the quality of a thumbnail, we put together a dataset labelled with two classes: good and bad.

Good: We define a good video as one with 1 million or more views. In order to find these videos, we downloaded (at most) 5 videos with a million or more views from the 2500 most-subscribed YouTube channels. It is worth noting that this set of channels is skewed towards the categories most popular on YouTube, like music and sports.

Bad: We define a bad video as one with 100 or fewer views. In order to find these, we looked at videos selected by a pseudorandom algorithm [3], of which about half were under 100 views. Unlike the “good” videos, we take these to be a representative sample of what is on YouTube.

We ended up with ~5000 thumbnails of each class. Every image was cropped and scaled down to 45 pixels by 80 pixels before being fed into our model.

Approach
1. First, we use our dataset of “good” and “bad” thumbnails to train a convolutional 2-class classifier.
2. In order to choose the thumbnail for a video, we push each frame of the video through the classifier and select the frame that receives the highest probability of being in the “good” class.

Model Architecture
Our best model is based on the AlexNet architecture [4] with the following modifications:
1. We removed the batch normalization layers because they did not help learning.
2. We decreased the filter size in the 1st convolutional layer from 11x11 to 5x5 because our images have about half as many pixels as ImageNet, which AlexNet was trained on.
3. We reduced the size of the dense layers from 4096 to 1000 because we are only performing binary classification.

Model Evaluation
- Two of our group members classified 212 thumbnails. Both achieved an accuracy of 81.6%, so our model almost has a human level of accuracy on the classification task.
- We have a 7.7% false positive rate and a 10.7% false negative rate on the validation set. Below are some examples illustrating some forgivable and unforgivable mistakes that our model makes.
- Our saliency maps show some features our model has fit to. The left one shows its preference for hands, which makes sense as something to focus on in a thumbnail. However, the right one shows its focus on logos/labels which appear often in the thumbnails from our “good” set. This is something to mitigate since having a logo does not indicate a high quality image.

Training
- To select hyperparameters, we ran the following experiments (which except for the last two were run in different categories on YouTube. 2015.

Model

Approach

Background

Dataset

Model Architecture

Model Evaluation

Training

References