Data Selection Proposal

Javier A. Pollak (260446737)

February 8, 2021

1 Dataset Choice

Kaggle Dataset: `https://www.kaggle.com/wchaktse/data-of-5132-youtube-videos`

- This dataset contains the thumbnail urls, view counts, comment counts, and like and dislike counts of 24052 videos on youtube from channels ranging from 1010 up to 38200000 subscribers. This is all the data that I need to carry out my project.

2 Methodology

2.1 Data Preprocessing

The dataset chosen is feasible as it is large enough and contains all the information needed on the videos. The preprocessing steps needed will be:

1. Extracting the needed columns from the csv: thumbnail url, and view, comment, like and dislike count.
2. Converting the thumbnail urls into urls to lower resolution versions of the images. (This is done by simply changing the "maxresdefault" part of the url to "mqdefault")
3. Downloading the thumbnail images.

2.2 Machine Learning Model

The goal is to estimate the view and comment count of a video by analyzing it's thumbnail's visual features. I will be using a Convolutional Neural Network for this project as it is the most appropriate model to use for training on image data.

2.3 Evaluation Metric

The evaluation metric will be the Mean Squared Error of the predicted view and comment counts to the respective counts in the training set.

2.4 Final Conceptualization

See question 3

3 Application

I intend to make a simple and pretty single-page webapp using React where the user would upload an image of a video thumbnail and be presented with the model's predictions on the success of their thumbnail. The information would include predicted view and comment counts as well as some example videos that are within that view and comment count range in order to compare yourself with them.