Predicting YouTube View Counts

Jon Braatz & Lance Lamore

October 26, 2018

Abstract

We propose to create a model for predicting YouTube views for a video posted to an uploader's channel based on a canditate video title and statistics from previously uploaded videos to the uploader's channel. Time permitting, we will additionally incorporate thumbnails as a feature. Previous work in this area has found that channel statistics are better predictors of a video's views than metadata about the video itself, so for a baseline prediction we will use the average of the views of up to five of that channel's most recently uploaded videos, and for an oracle we will use the candidate title as a search term and return the view count of the search result whose channel view count most closely matches the uploader's.

1 Overview

The premise of this project revolves around the idea of predicting the view count of YouTube videos given features like the title, thumbnail, description, and similar information from previous videos uploaded on the same channel. We had the initial idea of predicting views based just on a given title and thumbnail and then using this predictor to generate video titles that are optimized for maximizing views given a list of keywords that the title should contain. However, previous work in this area found that the metadata of individual videos were far less predictive of views than channel information like subscriber count, channel view count, and channel video count. Therefore, the input to our system will be a YouTube channel ID, a title, and maybe a thumbnail image if time allows, and the output will be a prediction of the number of views a video with that title and thumbnail would get if it were posted to that channel. For example, the input "Deep Reinforcement Learning" and channel ID "UCdKG2JnvPu6mY1NDXYFfN0g" (the ID of the Stanford Engineering channel) should give an output of rougly 120000. If time allows, we could use this view count predictor as a subsystem of a title generator that given a channel ID, thumbnail, and set of keywords, will will generate a video title corresponding to those keywords that is optimized for getting the most views for that channel.

2 Previous Work

Another team of researchers attempted a similar project that sought to predict view counts based on an input thumbnail image and title [1]. They performed gradient boosted regression on features from a dataset of YouTube videos from the Fitness & Health category, which they obtained from the YouTube-8M dataset that YouTube released to the research community [2]. They also used this data to scrape information from more videos using the YouTube API [3]. The features they used included: a title clickbait score output from a pre-trained neural network, a "NSFW" score of the thumbnail image output from another pre-trained neural network, channel subscriber count, view count, and video count, and the view count of the channel's previous uploaded video, among other features. They found that the most important features were features extracted from channel information and not the features extracted from either the title or image of the video.

3 Approach

Since previous research found that coarse channel information like total view count and number of videos published were the most predictive of a video's future view count, we will extend that approach by including a higher quantity of more precise channel data in our gradient boosted regression. In particular, we will take into account the titles and view counts of all videos in the channel, as well as video tags and descriptions. We might want to use GloVe embeddings to present natural language data as input to a recurrent neural network for use in the regression. Using these embeddings, we might also be able to perform a "nearest neighbor" search for videos on the channel with similar titles to the input title, and weight those video's view counts more. If we have time, we might use CNNs as feature extractors on an input thumbnail. Alternatively, if we have time we could use our prediction model as a subcomponent of a title generation system that, given a set of keywords a title should contain, constructs a title for a YouTube video optimized for views using search algorithms similar to the ones used in the "reconstruction" homework assignment, with the output title being contrained to have a certain minimum fluency score as scored by an n-gram fluency model.

4 Baseline and Oracle

For a baseline prediction for how many views a video with a certain title will get if it's uploaded to a particular channel, we compute the average number of views for up to 5 of the most recent uploads to the same channel. The oracle, on the other hand, will search YouTube using the input title as a search term, find the video in that set uploaded by the channel with the closest number of views to the input channel (that isn't the video being searched), and return the view count for that video. The loss function for these predictors is the difference squared of the logarithms (base 10) of the prediction and true view counts. We take logs in the loss function because the values being predicted range over many orders of magnitude.

After running these algorithms on twenty random YouTube videos from a Kaggle dataset [4] with views counts spread across 5 orders of magnitude, the baseline's average loss was found to be approximately 2.40, while the oracle's average loss was 0.22. The baseline prediction of taking the average view counts of the channel's most recent videos was thus off by an average factor of 250 from the true view counts, while the oracle was off by an average factor of less than 2. This gap suggests that using NLP to find videos with similar titles to the input title, like our approach will do, has the potential to greately outperform predictors that rely heavily on coarse channel statistics like total view count or an average of the view counts for the most recently uploaded videos.

References

[1] Aravind Srinivasan, *YouTube Views Predictor*, https://towardsdatascience.com/youtube-views-predictor-9ec573090acb

[2] YouTube 8M dataset: https://research.google.com/youtube8m/

[3] YouTube Data API: https://developers.google.com/youtube/v3/

[4] *Statistics Observations of Random YouTube Videos*, https://www.kaggle.com/nnqkfdjq/statistics-observation-of-random-youtube-video/version/1