Youtube Video View Prediction

Brett Bass[1] and Evan Ezell[1]

Abstract— **Youtube has increasingly become a prime space for content creators to distribute their content. There are many different factors that go into a video going "viral". Work has been done in predicting views using metadata of the video (category, title...etc). The goal of this analysis is to predict Youtube video views using convolutional neural networks (CNNs) in an effort to utilize the acutal videos in addition to any metadata associated with the videos.**

I. INTRODUCTION

Since it was founded in 2005, Youtube has increased in popularity and the quantity of views a creator obtain is becoming more and more lucrative. The exact value of a view is unknown due to the complex cost structure that Youtube utilizes to pay content creators (ad views, skipped ads, video topic... etc). However, there are several Youtubers that are making a fortune from producing content.

There are many factors that can be used to predict the success of a video, which we define as the number of views. Among these factors are the channel popularity, title, date published, category, etc. While these are important factors, they do not capture the actual content of the video.

CNNs in recent years have taken the lead in image and video processing. They are used in such cases as facial recognition, video analysis and natural language processing. We propose using CNNs to capture the content of Youtube videos to predict the number of views.

II. DATA

Since the video content of Youtube varies so vastly, we are restricting this study to specific subsets of Youtube videos. We will consider subsets of videos from a different channels, different categories, and different lengths.

It is important to obtain data from a wide variety of sources. It will be explore the difference between typical videos from popular channels and videos that have gone viral for their respective categories.

All Youtube videos have associated metadata that will be considered for the analysis. This data includes author, category, description, likes, dislikes, length, keywords, title, and viewcount. The data will be scraped from Youtube using python and stored in a database (Database type unknown).

III. WORK DELEGATION

A. Dr. Michel Ballings

Dr. Ballings is the principal investigator of this project. He will meet with the group twice per month to assess progress of the group as a whole and individually. During these meetings, he will provide assistance for any challenges encountered over the previous weeks work and point us in the right direction for the coming weeks. He has set up and is in charge of overseeing the project github repository. Dr. Ballings has also agreed to set up any computing power necessary to fully utilize CNNs, specifically graphical processing units (GPUs) which will increase performance of analyzing large video files.

B. Brett Bass

Brett will help in the effort to scrape Youtube to collect the metadata and store this data in the database. He will also use python to wrangle the data, turning the data into a form that will be well-suited for data science. Brett will complete some exploratory data analysis (EDA) to find trends and insights in the data. He is in charge of developing a baseline model and metrics as to determine if the integration of CNNs on video data increase accuracy. He will have to research CNNs and assist the team in the modeling process which includes model selection, model creation, and model evaluation. Brett will write the sections of the final report that relate to the scraping of Youtube, wrangling the data, and EDA.

C. Evan Ezell

I will help Brett with process of scraping Youtube. We may encounter problems of API rate limits. In this case I will integrate API scraping scripts across all of the groups machines in order to grab as much data as possible. I will decide which database we want to use and engineer the design to work with our specific project. I will work specifically with literature regarding the types of videos that are most likely to be successfully described using CNNs. I will look at literature regarding the types of CNN techniques that have shown the most promise in the recent years. I will work with Brandon for validation using his experience in CNNs to see what models he believes show the most promise.

D. Brandon

Brandon is an undergraduate computer science student who has received a scholarship to conduct research with a University of Tennessee, Knoxville faculty member. He has experience using CNNs in some of his other projects and will help implement the CNN models the group decides to use. Brandon will help tweak the models if the group sees room for improvement in any of the models.

University of Tennessee, Knoxville

[2]P. Brett and Evan are graduate research assistants in the Bredesen Center, University of Tennessee, Knoxville, TN 379,02 USA

IV. Objective

We hope to gain insight into video content on our chosen domain of Youtube by using CNN. We hope this insight adds something to the field of social content curation for creators, hosts, and consumers. If we are able to capture the content of the video then we will have essentially understood a popular Youtube video - popularity being measured by number of Youtube views. We will compare how much a CNN adds versus a baseline model which tries to predict views given standard metadata. Given our results we will determine future directions and potential for new research.

V. Tentative Schedule

- 9/28 - Determine categories, channels, and frequency of data to be collected.
- 10/5 - Decide on type of database and configuration to store data
- 10/19 - Write code to scrape, clean, and store in database.
- 10/26 - Complete EDA and baseline model.
- 11/2 - Research CNNs and determine optimal implementation for our data.
- 11/16 - Implement CNN on Youtube video data and determine effectiveness.
- 11/23 - Complete report and prepare to present findings.