Ensemble of Neural Networks to Predict YouTube Video Popularity

Benjamin Edwards **Evan Schwartz** **Nithin Sivakumaran**
Department of Computer Science
University of North Carolina at Chapel Hill
{bkedwar, eschw, nsivaku}@unc.edu

Abstract

YouTube is an American-based online video-sharing platform. Alongside features like titles and categories, all YouTube videos feature a thumbnail image that represents the video. The goal of our research is to predict view counts for YouTube videos when given the title, category, and the thumbnail image. Our results can help YouTube creators identify image and title styles that generate higher view counts. Thumbnails and titles, as the first features that a user sees, are key to capturing interest. Gathering data from top-viewed creators, we trained an ensemble neural network that processed text features and thumbnail images separately, then combined them to form a predicted class for each category. We find that our model shows significant results, and achieves an accuracy of over 99% on the testing and validation set.

1 Introduction

Predicting popularity of video content has become a very important task for many content creators. The majority of this content is on YouTube, a media sharing platform that boasts almost 2.5 billion monthly active users, and is the second most used social media site in the world [10]. For YouTube publishers, known as creators, revenue is generated based on view counts. Therefore, any dedicated creator desires to use effective styles to garner user interest to maximize their profit. Importantly, the first metrics of a YouTube video that any user sees are the title and thumbnail. A thumbnail is a small image displayed to the user, that is either chosen by the creator or by a YouTube thumbnail algorithm that is designed to select a representative frame for the video [9].

There have been many efforts towards building models to predict online popularity, using techniques like early-view patterning, stochastic modeling, and various regression learning models [6]. We propose a new view count prediction model, one that classifies view counts in four tiers for each category. Our model is an ensemble neural network, that combines a network trained on thumbnails, a network trained on titles, and one of 19 category inputs. We chose to form our classes within each category, as creators want to determine styles that pertain to their channel category. The model is trained on a combined set of Kaggle [5] data and data scraped from YouTube, where we collected video id's, channel names, categories, titles, view counts, and thumbnails. We find that our model exhibits significantly accurate results, producing a validation accuracy up to 99.5% and a testing accuracy up to 99.7%.

2 Related Works

In this section, we explore the background of our research. Several efforts have been made to predict online popularity. For example, one early method presented two models to help predict future popularity of YouTube content based on early popularity [6]. Their first model was a multivariate linear

Code available at https://github.com/nsivaku/562project.

regression model that incorporated information about historical patterns. Their second prediction model, MRBF, built upon the first model and exploited patterns more explicitly to improve prediction [6]. They frequently compared their model to Szabo and Huberman's long-term popularity model [8], and saw vast improvements in reducing error predictions, with the second model performing extremely well on specific patterns [6].

Previous attempts have various machine learning regression models have concluded that RandomForestRegression outperforms the others[3]. Others proposed a thumbnail feature extraction method that uses AIME to predict the number of views on YouTube [1]. They used AIME to extract image features on a blackbox model across seven different countries to examine desirable thumbnail features. Miyamoto et al. proposed that thumbnail images may actually be a major factor for users in selecting a video. Their classification model was trained specifically on the "Nataro-channel" on YouTube, and classified videos as either being in the "low", "middle", or "high" categories of access counts [4].

3 The Dataset

We use a dataset consisting primarily of data found on Kaggle [5]. In addition, the dataset itself was later supplemented by roughly an additional 900 videos that we scraped using the ytp-dl API. All videos are given one of the nineteen categories that we selected. Although YouTube does have its own category system of fifteen labels, not all videos have categories and many of them are overall general (e.g. a blog about surviving on an island and a video about an unsolved mysteries both are listed in the nebulous category of "entertainment." We chose to stick with the categorization provided by the original data source, using those descriptions to guide the categorization of further data. It should be noted all videos were made for and by citizens of English-speaking countries. It should also be noted a small number of videos are incorrectly categorized. We attempted to fix as many as possible, but given that this model would practically require an additional model to categorize videos, it seems reasonable that there be some errors.

3.1 Characteristics of the data

The dataset has several key characteristics. The median number of videos per channel is 28, and the mean is around 28.1. The variance of the quartile for each channel has a mean of 0.29 and a median of 0.25. There are 3144 valid data points. It should be noted that there will in most cases be some channels that appear in at least 2 of the training, validation, and test datasets. One potential issue for expanding this model is the categories themselves are imperfect, some to specific and others to general. If the dataset were to be expanded to many more data points there would likely need to be a reworking of the category system. All title lengths were under 25, with most of the data residing between around 5-10 words.

4 The Model

When considering the information available to any YouTube user seeking out new content on the platform, the two most important points of reference are the thumbnail and title. These two pieces of information serve as key indicators of the topic, quality, and tone of a video. Thus, it follows that an interesting question is to devise a method of estimating the success of a video based on these two pieces of information. Our model does exactly that by combining a convolutional neural network to

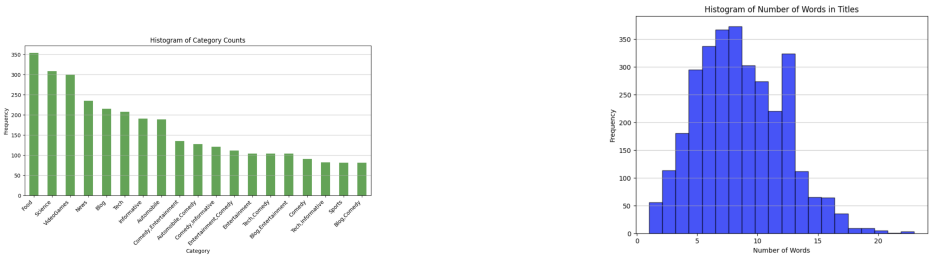

Figure 1: Bar Graph of Category Count and Histogram of Title Counts

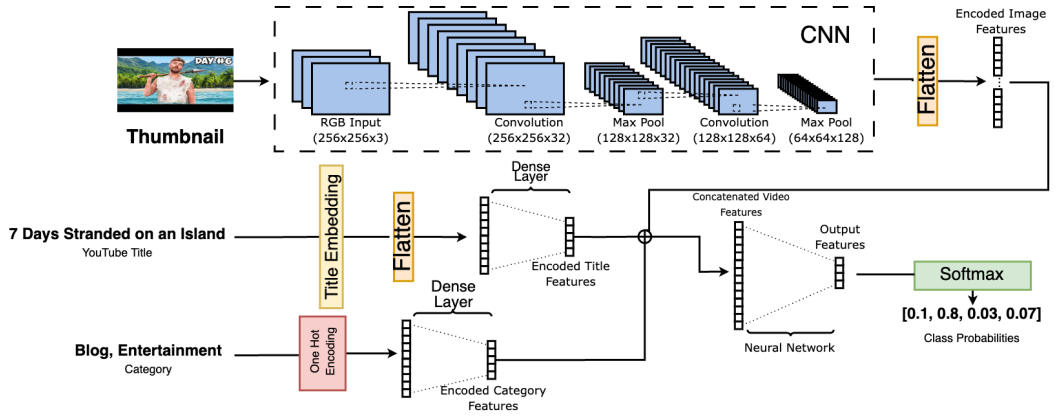

Figure 2: The thumbnail, title, and categories are passed into their own network to get their encoded features. These are then concatenated together before being passed through their own network to obtain the output features. Applying a softmax to these output features gives us our class probabilities.

process the image and additional hidden layers to develop connections between the features. Given that users all have varying preferences on what content they prefer and the popularity of various genres can vary widely, we argue it is most important to consider performance given video category. This leads to the conclusion that our label should be which quartile of views for its category a specific video achieved. In effect, we attempt to model the distribution

$$\mathbb{P}(\text{quartile} = k \mid \text{Image, Category, Title})$$

.

4.1 Architecture and Training Details

The Convolutional Neural Network (CNN) is a known robust method for image analysis, a combination of image convolution for feature identification and fully connected layers for identifying relationships between the features that lead to improved accuracy. We built a simple CNN model for image analysis and merged it with the two other features, title and category, each of which is passed through a 32-node dense layer before the merge. After merging the features, there are a further 2 dense layers with regularization methods in each. More details about the specific architecture are available in Figure 2.

We use a stochastic gradient descent optimizer with a momentum of 0.9 and an initial learning rate of 0.001, using AlexNet as a source of inspiration for the optimizer [2]. We also implemented the learning rate reduction method used in the aforementioned networks, as well as a stop-loss method for training. From Figure 3, we determined that the model had converged at epoch 19 as the validation accuracy dropped drastically at epoch 20.

5 Results

The accuracy of our model ranged between 93% and 99% on the test data over the various training sessions. The model's main weakness is built on the limitations of the project. We lack the computational resources to analyze the vast quantity of data points necessary to generalize the model. These limitations may mean the model is more trained to identify successful channels and differentiate between the view counts of their videos, rather than evaluate the success of random videos based on the first things the user will see. However, this is not to say the model is deeply limited. Compared to

	Model		
	Logisitic Regression	Random Forest	Ours
Acc.	0.497	0.489	**0.997**

Table 1: Accuracy of models on dataset. Logistic regression and random forest were ran on CLIP [7] features from the title and images as well as the raw one hot encoding of categories.

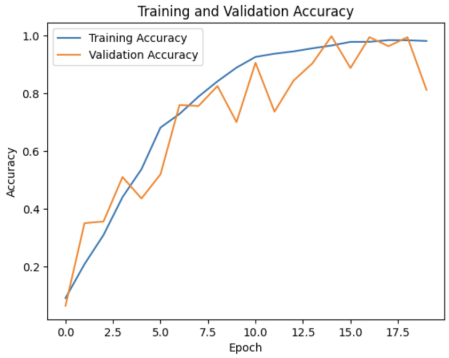

Figure 3: Training Accuracy v. Validation Accuracy curves over 20 epochs

classical techniques like logistic regression and random forest, our deep learning framework achieves stronger results. This suggests that our model is more capable of handling the high-dimensional data provided by the thumbnail image and the semantic meaning on the title. Our comparisons of the tested models are available in Table 1.

5.1 Examples

As seen in Figure 4, the model does make differentiation between channels. In the case of all channels represented, video titles and thumbnails tend to be quite homogeneous. However, for the two incorrect videos, the model does correctly identify that these videos should outperform the channel's standard videos, which in both cases are mostly 2s. The correctly identified video also shows positive signs of deeper model understanding, as the related channel has relatively high variance and almost identical thumbnails. Note the variance of the correctly identified video channel is 0.57 and the variance of the incorrect identification are from channels with 0.17 and 0.23, from left to right.

6 Conclusion

In this work we presented an ensemble neural network that predicts classes of view counts of YouTube videos given title, category, and thumbnail features. We propose a convolutional neural network to process image thumbnails with additional hidden layers to support text and category encoding. We find that the performance of the model excels on validation and testing data, and is a promising approach to determining thumbnail and feature styles that capture viewer interest the best. It would be interesting to perform further analysis on thumbnail and title styles to see which styles generate higher view counts. For future work on this model, we believe it would be beneficial to create a true regression model instead of a classifier and see how well a similar model can minimize prediction error. It would also be interesting to expand upon the work of and generalize our model to different countries and languages to examine cross-cultural differences in styles [1].

Figure 4: Examples of View Count Classifications

References

[1] Okada R. Minematsu A. Nakanishi T. Fukui, R. Country by country comparison of thumbnail features contributing to views using aime for youtube. *2023 15th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter).*, 2023.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

[3] Manimuthu A. Sharmila J. R. Sathya K. N. Manikandan, P. Prediction of youtube view count using supervised and ensemble machine learning techniques. *2022 International Conference on Automation, Computing and Renewable Systems (ICACRS)*, 2022.

[4] Y. Miyamoto. Prediction of number of accesses by thumbnail image classification on video sites. *2023 7th International Conference on Information Technology(InCIT).*, 2023.

[5] Praneh Mukhopadhyay. YouTube Thumbnail Dataset, 2022.

[6] Almeida J. M. Gonçalves M. A. Pinto, H. Using early view patterns to predict the popularity of youtube videos. *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining.*, 2013.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[8] Huberman B. A. Szabo, G. Predicting the popularity of online content. *SSRN Electronic Journal.*, 2008.

[9] Drost Video. How are youtube thumbnails chosen, Aug. 2017.

[10] J. Zote. 25 youtube stats marketers should know in 2024 [updated]. Available at `https://sproutsocial.com/insights/youtube-stats/` (2024/03/20).