Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/41/20230741

Predicting video popularity based on video covers and titles using a multimodal large-scale model and pipeline parallelism

Jie Qin[1,4], Bei'an Wang[2,5], Tianyu Zhu[3,6]

[1]Xi'an Jiaotong University
[2]Wuhan University
[3]South China University of Technology

[4]1503490056@stu.xjtu.edu.cn
[5]843531329@qq.com
[6]Zty781004@126.com

Abstract. In the era of traffic, controlling traffic is equivalent to mastering influence and economic benefits. At the video level, under the premise of the same video content, it is very important to study what kind of cover and title can be more attractive to people. Unlike most previous studies that focused on YouTube videos, our data came from Bilibili's videos. This paper tried to use two neural network models, ViT and Bert, combined with GPipe and backend fusion multimodal data fusion methods, to predict the possible click-through rate and popularity of a specific video based on its existing video cover and title. In the process, we switched to different visual and language models to complete the same training task, with the goal of comparing the impact of different models on the results. By adjusting the weight of two models, we finally achieved a good result of up to 62% accuracy.

Keywords: ViT, Bert, Gpipe, Pipeline Parallelism, Covers on Video Views.

1. Introduction

This section will introduce three parts: Background, Research value and significance and Our work.

1.1. Background

The traffic era is based on the current developed Internet era. The essence of the traffic era is to use the influence on the public to generate economic benefits. Take advantage of some of the public's preferences to quickly accumulate popularity and turn into "cash flow" through certain platforms.

So that it is particularly important to study how various factors of a video affect the number of views of the video, thereby indirectly affecting the traffic and influence of the video.

1.2. Research value and significance

In recent years, there have been many related studies on how traffic is converted into economic gains, whether public preferences are controllable, and what kind of video content attracts people:

1. Millionz.Co Offers Multi-Tiered Service to help influencers, businesses, and creatives gain massive increases in followers, likes, video views, music plays, and shares [1]. This paper shows that

DOI: 10.54254/2755-2721/41/20230741

people's preferences are controllable and predictable. It also directly shows the fact that influence is power. It is in line with the use of influence to create economic benefits in the traffic era.

2. Combining recent theorising on affect with insights from sociolinguistic research, the study investigates how the YouTube users' affective investments contribute to a (re)evaluation of the two minoritised languages - Irish and Sámi [2]. The study shows that when watching a video, viewers are emotionally and thoughtfully engaged in the process. This affects their cognition and thus some of their real-life behaviors, such as reviews and purchases. This is also part of the basic principle of using mass influence to generate economic benefits in the traffic era.

3. This paper investigates the impact of marketers' video optimization practices on video views. Unlike previous studies that focused on audience interaction behaviors such as liking, commenting, sharing videos, etc., this experiment focused on elements of video views and video metadata [3]. This experiment shows that artificially changing video optimization can affect video views to a certain extent. After clarifying that people's preferences are controllable, the number of video views is also controllable to a certain extent. This paper lays the foundation for the study of how to artificially control various factors of the video, such as the cover, title, partition, etc., so as to study their impact on the number of views.

1.3. Our work

Studies contrast:

In the era of traffic, controlling traffic is equivalent to mastering influence and economic benefits. At the video level, under the premise of the same video content, it is very important to study what kind of cover and title can be more attractive to people. The goal is to grasp the traffic.

There have been some studies that have targeted grasp the traffic in the past. In the past, the elements of research, such as video topics, tended to be more figurative and data-based. No special attention was paid to the cover elements. So this paper decided to look at the impact of covers on video views.

Due to the influence of regional, cultural and other factors, most of the previous studies have focused on YouTube videos, and very little about Bilibili. But Bilibili is also a very common long-form video platform in China. Considering the relatively high production cost of long-form videos and the high expected return value, the need to increase the number of views is urgent and strong. So the video dataset this paper chose was taken from Knowledge Section of Bilibili.

Many previous studies have favored direct access to user experience or judgment through metrics. The research methods are questionnaires, interviews, etc. What this paper achieve is to train the model with data, and then let the model make a predictive judgment on the video cover. At the same time, the parallel assembly line is also a feature of our experiment.

Models realization:

The models this paper use are ViT and BERT:

The ViT model is the vision transformer. The idea is to use the self-attention mechanism-based transformer model in the NLP field for image tasks.

BERT: The whole is a self-coding language model, and it is designed with two tasks to pre-train the model.

At the same time, this paper also use Gpipe to achieve pipeline parallelism to achieve model parallelism. GPipe uses a model parallelism scheme to divide the model into a series of stages. Each stage is placed on a separate device (GPU/TPU) to support hyperscale models.

In the stage of post-fusion, this paper adopted a multimodal approach called Back-end fusion. To make the data more balanced, this paper use normal normalization to normalize video views.

Finally, a model that can judge the number of video views based on the cover is realized. And achieved a relatively high accuracy rate - about 62%.

2. Methods

This section will introduce three parts of our work: Dataset, Model and Experiment.

DOI: 10.54254/2755-2721/41/20230741

2.1. Dataset

2.1.1. Dataset Collection

In the dataset collection part, after thorough research and analysis, this paper decides to conduct data collection on the Knowledge Section of Bilibili. The primary reasons for this choice are as follows:

1. Bilibili website boasts a large user base and significant traffic in China [4], and the collected data, after careful filtering, exhibits high quality. Moreover, the video data from Bilibili is publicly available data from video bloggers, and the data collection process strictly adheres to 'robots' protocol.
2. The Knowledge Section exhibits a relatively balanced user preference. In contrast to other sections, the majority of content in the Knowledge Section revolves around popular science videos, catering to a diverse audience and attracting viewers with a wide range of interests. However, sections like Technology and Entertainment may have more subjective content, driven by the focused preferences of their respective audiences, thus potentially introducing bias in the collected data.
3. Bilibili website hosts numerous video bloggers, making it an ideal data source with practical applications in the field of video creation.

In Bilibili website, we collect more than 27000 pieces of data, including video cover, title, views number, likes number, coins number (coins are a unique measure of video popularity on Bilibili platform, where each user can use only one coin per day) [5], favorites number, and shares number. Then the data is sent to Dataset Filtering section.

2.1.2. Dataset Filtering

After collecting the dataset, it is necessary to perform data filtering. The filtering rules are as follows

The number of followers for video authors is limited to below 500,000. Besides, the author must have been on the hot list. For the first rule, excessive followers of video authors can lead to inflated data results, affecting model training, and the "celebrity effect" may influence the authenticity of the data. Moreover, video platforms tend to grant additional exposure to high-follower content creators, potentially impacting the universality of the data. The second condition ensures that there are not too few high scores in the dataset. By doing so, the model can effectively predict higher scores instead of all the predictions are lower scores.

2.1.3. Dataset Processing

A quantitative standard is required to evaluate the video popularity. In the model, video popularity is calculated using the following formula:

$$\text{Final Score} = w_1 * Views\ Score + w_2 * Likes\ Score + w_3 * Coins\ Score + w_4 * Favor\ Score + w_5 * Shares\ Score$$

The Views Score, Likes Score, Coins Score, Favorites Score, and Shares Score contribute to the overall score. In this paper, the weights are set sequentially as follows: 0.75, 0.1, 0.1, 0.05, 0, as the cover and title have a bigger impact on views score and the other scores are more influenced by the video's content.

Then, how to convert these numbers to scores? To prevent biases arising from differences in data scales across different dimensions and to narrow the gap between regular videos and exceptional ones, Z-score normalization is applied to each score:

$$\text{Each score} = \frac{\text{Each number} - \mu}{\text{Max} - \text{Min}}$$

The μ in this formula means mean value, the Max means the highest score in each dimension, and the Min means the lowest one.

Processed final scores are then segmented into five groups based on the following formula, yielding the label for each data instance.

DOI: 10.54254/2755-2721/41/20230741

$$\text{label} = \begin{cases} 0, & score < 0.00025 \\ 1, & 0.00025 \leq \text{score} < 0.0015 \\ 2, & 0.0015 \leq \text{score} < 0.005 \\ 3, & 0.005 \leq \text{score} < 0.02 \\ 4, & 0.02 \leq \text{score} \end{cases}$$

Using above-mentioned formula, the dataset will be divided into five groups. Groups 0 to 2 have a larger number of data instances, while groups 3 to 4 have fewer instances. This hierarchical policy allows for better discrimination between data instances and facilitates the model in making more accurate predictions, when getting the better data.

2.2. Model

2.2.1. Model Architecture

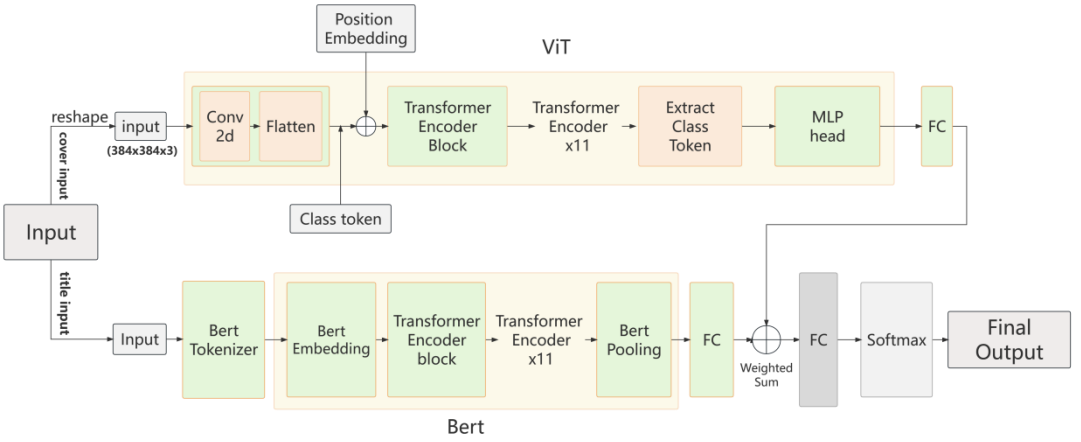

Figure 1. Model Architecture. FC means Fully Connected Layer.

Firstly, the input data will be divided into two parts: video cover data and title data. These data will be separately fed into the visual recognition module and the language recognition module for feature extraction. Subsequently, the outputs from both modules will be combined through a weighted summation process. Finally, the fused features will be passed through a fully connected layer to generate the final output used for prediction. The entire model structure is depicted in Figure 1.

Visual Model:

For video cover recognition, considering the significant amount of information contained in the cover, a deeper and larger model is required to extract cover features effectively. Therefore, this paper have chosen ViT(Vision Transformer)[6], and adapted it for the cover recognition task. Furthermore, the ViT model exhibits a robust global contextual mechanism, which is particularly advantageous for scenarios like video covers that encompass entire images. The adoption of ViT enables simultaneous consideration of all regions within an image, in contrast to the local perception of traditional CNN models. This attribute proves highly beneficial for cover recognition tasks. ViT also demonstrates exceptional scalability and adaptability to transfer learning. It could be pretrained on a dataset and then applied transfer learning, resulting in a substantial enhancement of dataset utilization.

Since covers are typically smaller, and viewers see them in a reduced form while browsing, this paper employs a 12-encoder-layer ViT model (as shown in Figure 1). Here, the model own 768 hidden size, 3072 MLP size and 12 multi-attention heads, allowing better attention to the overall cover, enabling the extraction of more comprehensive features. Furthermore, ViT's scalability and extensibility are well-suited for multi-GPU parallel training, enhancing training speed.

DOI: 10.54254/2755-2721/41/20230741

Firstly, the video cover is reshaped to a size of 383x384 and then sent to the ViT model. After going through convolutional and flatten layers, the output is concatenated with the class token. Then the positional embedding is added, and the result is fed into the transformer encoder blocks of ViT. Following this, a dropout layer is applied before the output passes through LayerNorm layer, Extract Class Token layer, and an MLP head layer is employed. Finally, the output from the last fully connected layer serves as the output of the visual model.

Language Model:

Similar to video cover recognition, video titles also carry a substantial amount of information, often encapsulating the theme and main points of the video within 20 or so words. Thus, extracting title features requires a deeper and larger model as well. After considering model performance and resource consumption, this paper adopts the Bert-base language model (as shown in Figure 1) [7], comprising 12 Transformer Encoder layers,768 hidden size, 12 multi-attention heads, one Embedding layer, and a final Pooling layer.

Firstly, the video titles are tokenized using Bert Tokenizer, and fed into 12 layers of Bert transformer encoder blocks. After passing through Bert pooling and a fully connected layer, the output serves as the output of the language model.

Model Fusion:

Reasons for the adoption of a multimodal model are as follows:

Inherent Relationship: There exists a significant connection between video covers and titles. The correlation between the textual elements on the cover and the content of the title, such as alignment and relevance, profoundly influences the initial impression users form about the video.

Complementary Nature: Video covers and titles are mutually supportive components. The absence or inadequacy of either element's design can diminish the video's appeal and, in some cases, render it entirely unengaging.

Front-Facing Information: The primary information presented to users on video platforms, whether on the homepage or within dedicated sections, consists of video covers and titles. Since these two components are integral to user engagement, simultaneously considering both cover and title information is of paramount importance.

The incorporation of both visual and textual cues through a multimodal model enables a comprehensive analysis that captures the intricate relationship between video covers and titles, ultimately leading to more accurate and effective predictions of video popularity.

This paper adopts the late-fusion approach, where the visual model and language model independently output their respective feature tensors. The two sets of features are combined through weighted summation and passed through one connected layer before producing the final predictions.

In this paper, the weights for the video and title information are set to 0.75, respectively, as users tend to prioritize video cover information before focusing on the title information. By assigning a higher weight to the video (0.75), the model can emphasize the importance of visual cues captured from the cover during the prediction process.

2.2.2. Model Training

Pretraining:

ViT: The ViT pretrained model is sourced from the "pytorch_pretrained_vit" library in PyTorch. It undergoes pretraining on the ImageNet-21k dataset and is subsequently fine-tuned on the ImageNet-1k dataset.

Bert: The Bert pretrained model is sourced from the "transformers" library in PyTorch. It utilizes the "bert-base-chinese" option from the BertConfig for pretraining.

Training:

The dataset contains 27000 videos' covers and titles, it is divided into two parts: the training set and the test set, with a ratio of 10:1.

Training Configuration: The training environment consists of four machines equipped with NVIDIA GeForce RTX 3080 GPUs.

DOI: 10.54254/2755-2721/41/20230741

Parallel Strategy: The Gpipe parallel strategy is employed to perform pipeline parallelism for the model [8]. In the Gpipe parallel strategy, the training process is divided into several micro-batches, and each micro-batch is processed independently by different GPUs in a pipeline-like manner.

The AdamW optimizer is chosen for the training process [9], with an initial learning rate of 5e-5. The cross-entropy loss function is used during the optimization.

3. Results

We introduced ordinal multinomial logistic regression to redefine the calculation of accuracy, making it more reasonable and realistic. We set the weight of the results that are completely correct to 1, and the weight of the results that are wrong but in the adjacent position category of the correct result to 0.5.

Table 1. When there are variations in the weights of ViT and BERT, the changes in training set accuracy and test set accuracy are as shown in the table below.

	Epoch 1		Epoch 2		Epoch 3		Epoch 4		Epoch 5	
ViT/BERT	Train ACC	Valid ACC	Train ACC	Valid ACC	Train ACC	Valid ACC	Train ACC	Valid ACC	Train ACC	Valid ACC
0.3/0.7	0.5456	0.5603	0.6323	0.6037	0.7116	0.6282	0.7933	0.6199	0.8564	0.5967
0.45/0.55	0.5538	0.5767	0.6354	0.6044	0.7217	0.6160	0.8047	0.6107	0.8683	0.6169
0.6/0.4	0.5525	0.5715	0.6343	0.6156	0.7158	0.6112	0.7960	0.5986	0.8551	0.6161
0.75/0.25	0.5510	0.5737	0.6346	0.5897	0.7089	0.6242	0.7909	0.6150	0.8556	0.6109
0.9/0.1	0.5499	0.5850	0.6276	0.5782	0.7071	0.6156	0.7882	0.6075	0.8505	0.6180
1.0/0	0.5387	0.5613	0.6195	0.5977	0.6898	0.5988	0.7665	0.6031	0.8393	0.5980
0/1.0	0.4551	0.4659	0.4569	0.4659	0.4579	0.4659	0.4578	0.4659	0.4577	0.4659

According to the table, when the weight of ViT is set to 0.75, we can achieve the highest testing accuracy (0.6242). Both overemphasizing and underemphasizing the contribution of ViT lead to a decrease in accuracy. When training with either the ViT or BERT unimodal model, the training performance is noticeably inferior to that of the backend-fused multimodal model. This further validates the necessity of using a multimodal approach.

However, when using a multimodal model, the variation in weights between ViT and BERT doesn't have a significant impact on accuracy. This might be due to the strong adaptability of ViT. When comparing the performance of the two unimodal models, it's also evident that training solely with the ViT model yields better results. This could be attributed to the weaker feature extraction capability of BERT for video titles and the relatively low correlation between video titles themselves and their popularity.

Next, we will switch to different visual and language models to explore the performance differences of various models in completing this task.

DOI: 10.54254/2755-2721/41/20230741

Table 2. When using VGG as the visual model and either RNN or BERT as the language model, the training results are as shown in the table below.

Models	Epoch 1		Epoch 2		Epoch 3		Epoch 4		Epoch 5	
	Train ACC	Valid ACC	Train ACC	Valid ACC	Train ACC	Valid ACC	Train ACC	Valid ACC	Train ACC	Valid ACC
VGG/RNN=0.75/0.25	0.5083	0.5013	0.5337	0.5245	0.5431	0.5141	0.5465	0.5283	0.5510	0.5471
VGG/BERT=0.75/0.25	0.5064	0.5247	0.5286	0.5288	0.5391	0.5447	0.5429	0.5337	0.5470	0.5345

We switched to different visual and language models to complete the same training task, and the results are shown in the table. In terms of language models, the training performance of using an RNN model is not significantly different from using BERT. However, the variation in visual models has a substantial impact on training effectiveness. This once again confirms the dominant role of visual models in this type of training task. In conclusion, using ViT as the visual model and BERT as the language model yields the best training results.

The accuracy of several training sessions with better results has reached more than 60%, which is in line with expectations.

4. Discussion

4.1. Difficulties and solutions

In the experiment process, we encountered some difficulties and adjusted and optimized the research content and techniques used accordingly. At first, we chose CoAtNet as the image model, which was proposed by Google Brain in 2021, as it enjoys both good generalization like ConvNets and superior model capacity like Transformers, achieving state-of-the-art performances under different data sizes and computation budgets [10]. However, due to the small size of the dataset and limited computing power, it was very difficult to achieve a satisfactory accuracy in a short time even using pipeline parallel techniques such as GPipe. Based on this situation, we replaced the image model with a pre-trained ViT, and added ordinal multinomial logistic regression when calculating accuracy, giving some weight to the results that were wrong but in the adjacent position category of the correct result to increase its rationality. After optimization, our model performance and training results were significantly improved.

4.2. Limitations and future work

In this study, we obtained a small dataset, which did not fully exploit the performance of the two neural network models and other techniques used. In addition, we paid less attention to the adjustment of hyperparameters, which should be an important factor for further optimizing model performance and training results. In future work, we will improve these two aspects.

4.3. Predictable application scenarios

We hope that after this technology is developed and improved, it can help video creators optimize video content and promotion strategies, increase video exposure and attractiveness, and thus increase video traffic and revenue. This technology can also provide more accurate recommendation systems for video platforms, enhance user experience and satisfaction, and promote platform development and growth.

In the future, this technology should have great development prospects. Of course, with the explosive growth of video data, the demand and challenges of video prediction technology will also increase. In order to improve the accuracy and efficiency of video prediction, we still need to constantly improve and innovate artificial intelligence models, introduce more video features and information, such as spatio-temporal motion, action information, texture information, etc. in videos. At the same time, we also need to consider how to combine video prediction technology with other artificial intelligence

DOI: 10.54254/2755-2721/41/20230741

technologies, such as object detection, semantic segmentation, behavior recognition, etc., to achieve better video analysis and adapt to more application scenarios.

5. Conclusion

In this study, we tried to use some neural network models, such as ViT, Bert, VGG and RNN, combined with GPipe and backend fusion multimodal data fusion methods, to predict the possible click-through rate and popularity of a specific video based on its existing video cover and title. During the training process, we tested several visual and language models and performed multiple adjustments and tests on their weight ratios to achieve the best training performance. In our computation of results, we also introduced ordinal multinomial logistic regression to better align with the specific application of predicting video popularity, aiming to mitigate the inherent inconsistencies caused by rating segments. By keeping records of results under different variables and conducting comparisons, we analyzed some potential factors that may lead to differences in training performance, and eventually achieved a good result of up to 62% accuracy.

References

[1] Millionz.Co Offers Multi-Tiered Service To Help Influencers, Businesses, and creatives gain massive increases in followers, likes, video views, music plays, and shares. Influence is power! [J] M2 Presswire. Volume , Issue . 2021

[2] Kati Dlaske. Music video covers, minoritised languages, and affective investments in the space of YouTube [J] Language in Society. Volume 46 , Issue 4 . 2017. PP 451-475

[3] Wondwesen Tafesse. YouTube marketing: how marketers video optimization practices influence video views [J] Internet Research. Volume ahead-of-print , Issue ahead-of-print . 2020. PP 1689-1707

[4] Ya L I , University A A .Research on Development Status and Optimization Measures of Agricultural Products Cross-border E-commerce Logistics in Anhui Province[J].Journal of Anhui Agricultural Sciences, 2019.

[5] Dai X, Wang J. Effect of online video infotainment on audience attention[J]. Humanities and Social Sciences Communications, 2023, P3.

[6] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[7] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[8] Huang Y, Cheng Y, Bapna A, et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism[J]. Advances in neural information processing systems, 2019, 32.

[9] Shazeer N, Stern M. Adafactor: Adaptive learning rates with sublinear memory cost[C]//International Conference on Machine Learning. PMLR, 2018: 4596-4604.

[10] Dai, Z., Liu, H., Le, Q. V., & Tan, M. CoAtNet: Marrying Convolution and Attention for All Data Sizes. arXiv preprint arXiv:2106.04803, 2021.