Advertisement Recommendation Engine - Improving YouTube Advertisement Services

Shanmuga Skandh Vinayak E,Venkatanath A G S,Shahina A, Nayeemulla Khan A

Abstract: ***Ever since its early inception in the year 2005, YouTube has been growing exponentially in terms of personnel and popularity, to provide video streaming services that allow users to freely utilize the platform. Initiating an advertisement-based revenue system to monetize the site by the year 2007, the Google Inc. based company has been improving the system to provide the users with advertisements on them. In this article, 7 recommendation engines are developed and compared with each other, to determine the efficiency and the user specificity of each engine. From the experiments and user-based testing conducted, it is observed that the engine that recommends advertisements utilizing the objects and the texts recognized, along with the video watch history, performs the best, by recommending the most relevant advertisements in 90% of the testing scenario.***

Keywords: Advertisement, Recommendation Engine, Objects, Texts, Audio, Recognition, Detection, YOLO, Tesseract.

I. INTRODUCTION

YouTube, a video streaming platform with the highest number of users visiting each day amongst the websites that stream videos, accounting to 1.78 billion users in 2020 [1]. Ever since its gradual increase of users from a few hundred to a million users every hour, the platform has always been on the prowl to improve user experience and efficiently generate revenue, utilizing the high user engagement towards the site. The method with which the site can generate its major portion of the revenue is by broadcasting advertisements of companies subscribed to the YouTube advertisement model. Although this seems to be straightforward, often the users skip advertisements irrelevant to them, thereby generating lesser revenue than scenarios when the advertisements are watched completely. This shows the importance of targeted advertisements to users, who influence the revenue system of the site. The YouTube video recommendation has been gradually improving over the years by recommending similar videos based on the users' watch history. Although the method usedin recommending YouTube videos may be similar in recommending advertisements, in theory, the same may not be as effective in a real-world scenario. The advertisement recommendation needs to be tailored such that, the user engagement towards the advertisement is high and is not skipped.

Revised Manuscript Received on October 20, 2020.

Shanmuga Skandh Vinayak E*, Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India. Email: shanmugaskandhvinayak16095@it.ssn.edu.in

Venkatanath A G S, Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India. Email: venkatanathags16120@it.ssn.edu.in

Shahina A, Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India. Email: shahinaa@ssn.edu.in

Nayeemulla Khan A, School of Computing Sciences and Engineering, Vellore Institute of Technology, Chennai, India. Email: nayeemulla.khan@vit.ac.in

The patterns amongst the user of similar interest and the popularity of an advertisement amongst them are to be considered. To tackle this dilemma, the YouTube platform openly mines data from the user activity its parent platform (Google Inc.), to efficiently and accurately determine the interests of the user. The architecture of the YouTube recommendation system is given in figure 1.

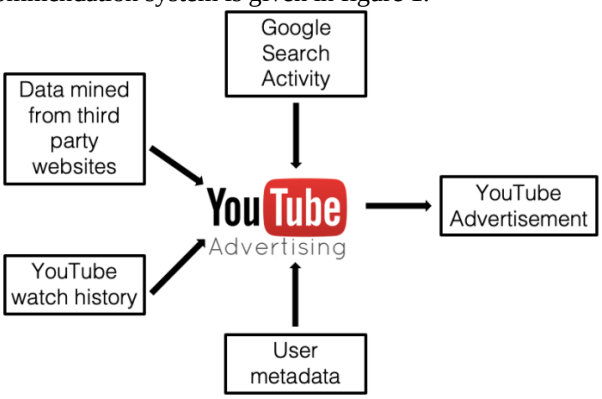

Fig. 1. YouTube advertisement engine

Although this is effective, a smaller platform that may not have access to the users' entire web search pattern, but can utilize their site data, can use the proposed solution to monetize their sites. The method used by YouTube can be improved for a small-scale regional video streaming platform to provide the users with targeted advertisements, by eliminating the third-party user data. The solution proposed in this article utilizes object and text detection techniques to generate data, that correlates with the genre and the intent of the advertisement. The architecture of the proposed solution that gave the best recommendation when tested with users, is given in figure 2.

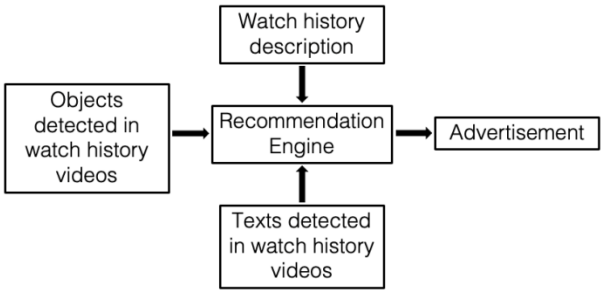

Fig. 2. Proposed Recommendation Engine

This article is organized as follows. Section II reviews the contemporary studies carried out in the field of advertisement recommendation using machine learning techniques. The Experimental setup and results are discussed in section III. The limitation of this work along with further directions is discussed in the concluding section IV.

II. RELATED WORKS ON RECOMMENDATION ENGINES

R Vinit Kaushik et al., in their work "Ad Analysis using Machine Learning" [3], perform content-based advertisement classification, utilizing the intent of the advertisement. The content of the advertisement is recognized using concepts such as object detection and speech recognition, to reveal aspects of the video such as the name of the product, usage of the product, nature of the product, etc. The authors realize object detection using edge detection algorithms, that utilize TensorFlow technology and speech recognition using application programming interfaces such as, SphinxBase. The recognized objects and speech are converted into text, which is classified to appropriate intents using Naïve Bayes algorithm.

In the work done by Jinqiao Wang et al., an interactive scheme for recommending advertisements based on the hierarchy concept, was developed to improve the recommendation of advertisement in their article, "Interactive ads recommendation with contextual search on product topic space" [4]. The authors implement a content-based advertisement data generation for each user, based on the content of the websites visited and the occasional videos watched by the user. The videos watched are analysed for advertisement tags, obtained using a per-sample multiple kernel learning method (PS-MKL). The collected data, comprising of advertisement tags from websites and product tags from videos, are compiled to a hierarchal structure. Based on domain specificity and similarity, the Group-Sensitive Multiple Kernel Learning (GS-MKL) algorithm, classifies the collected data. This data structure is utilized in suggesting advertisements, based on the search pattern and search history of the user on a specific website.

In this article, a similar approach to the works of Kaushik et al. (2017) is proposed with the extent of utilizing more aspects of the video such as texts and descriptions of the user watch history to recommend advertisements. Also, similar to the works of Wang et al. (2011), the obtained data are considered as content-related information, but rather than classifying the data, the system obtains the similarity amongst the available advertisement corpus to recommend the most similar advertisements to the user watch patterns.

III. EXPERIMENT AND RESULTS

All the experiments are conducted using an HP Z4 G6 Tower workstation, having a 2.1 GHz octa-core Xeon Haswell – EP processor, that a 64 GB DDR4 DRAM utilizes, with an NVIDIA GeForce GeForce® GTX 1080 Ti consisting of 3584 CUDA cores, clocking at 1.5 GHz. The datasets are generated using this workstation, as it enables efficient use of computing components in processing the videos.

A. Dataset

All the data are obtained from the YouTube platform, which consists of advertisement videos and test videos, from which the data for the recommendation engine (RE) are generated. 4 types of preliminary data are generated from the collected videos, which are mentioned below.

1. The video descriptions of the users' video watch history.
2. The texts present in the video frames.
3. The audio of the video.
4. The objects present in the video.

Using the data generated, 7 types of input datasets, comprising of all the combinations of the data generated, is prepared, to serve as the input to the engines. The following section describes the algorithms and techniques utilized in generating the 4 types of preliminary data.

1. Description data

The video description allows the engine to perform a basic classification, that segregates the genre of the videos watched while also providing metadata of the video. For this experiment, the descriptions are collected when the advertisement videos and the test videos are downloaded. The video description is utilized in all the versions of the input data combinations, as descriptions allow the engine to perform initial classifications, based on genre.

2. Text data

The videos obtained, are subjected to a frame-by-frame evaluation to identify any texts present in the video. A majority of the advertisements, tend to display the text of the company name, brand name, product name, etc., to visually reinforce the viewers' memory on the product, for them to identify the product more quickly. Utilizing this factor, the RE can provide recommendations on products, that have high similarity to the texts obtained. Two methods to recognize texts are tested for their efficiency, performance, and accuracy.

a. K-Nearest Neighbour - The K-Nearest Neighbour (k-NN) is a machine learning classification algorithm, that classifies the clusters formed by representing the data in a graphical form, based on the minimal Euclidean distance amongst them. The k – NN is a supervised algorithm, hence the k-NN model is trained before utilizing it to generate data.

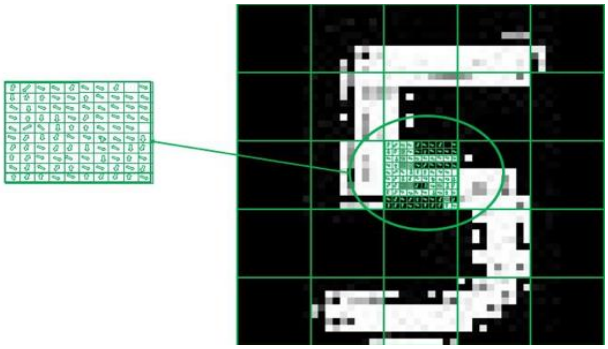

Fig. 3. k-NN cluster character classification on a bounded image

b. OpenCV - Optical Character Recognition - The Tesseract OCR [2], originally developed by Hewlett-Packard (HP) and open-sourced in 2005 is now maintained by Google Inc. since 2006. The software utilizes the OpenCV libraries, to identify texts in images, with higher accuracy and efficiency than a conventional machine learning algorithm, depending on the quality of the image. The pytesseract package, a python-based wrapper for the Tesseract OCR, is utilized for this experiment.

Similar to the k-NN method, the recognized images are saved as text files for each frame, and an aggregated set of words is compiled.

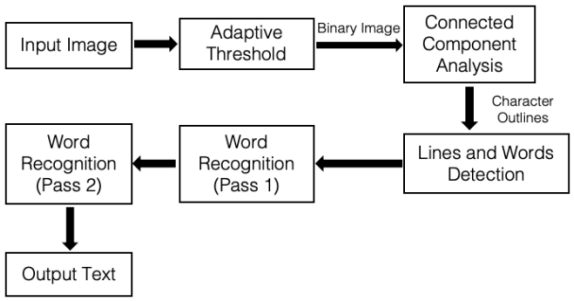

Fig. 4. Tesseract OCR - Text Extraction Architecture

The OpenCV Tesseract OCR performs with a faster average processing duration of 2.168s/frame. Whereas, the k-NN algorithm processes a single frame with an average of 3.629 seconds. Tesseract OCR provides higher accuracy with its detection accuracy of 78.48%, while the k-NN algorithm only provides 36% accuracy. The Tesseract OCR also exhibits efficient use of the computational resources, which are monitored using the CPU and RAM monitors for the operating system.

3. Audio data

The majority of the advertisements that promote products, tend to use actors in the advertisements, to provide an instructional insight on the type of the product, the benefits, the usage, etc., along with the reinforcement of the products' name using vocal communication. This is taken advantage of by the RE, to recommend products of similar description, usage, type, etc.

The video files are converted to the Waveform Audio file format (WAV) using FFmpeg software. The audio files are converted to text using the Google Text-to-speech (gTTS) Application Program Interface (API).

a. Google Text – To – Speech API - The gTTS server is a dynamic and periodically updated speech recognition repository, that allows users to access it using simple API to convert audio data to texts. For this experiment, the gTTS API for python is utilized. The API utilizes internet connectivity, to process the audio in the gTTS servers with a processing time of 0.3314s/1-minute audio.

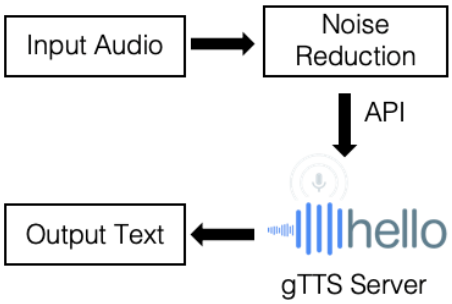

Fig. 5. Audio to Text conversion using gTTS

4. Objects data

The objects present in the video, play a crucial aspect in determining the type of advertisement a user may expect. Comparatively, the users are more reinforced on a product visually, than a textual or auditory stimulus. This allows the RE to recommend product advertisements, that are more inclined towards the products appearing in the users' watch patterns.

One of the most accurate and easy to implement algorithms to detect the objects in a video is the You-Only-Look-Once (YOLO) algorithm. This algorithm is effective in producing high accuracy results and performance with sufficient hardware. The training data corpus from which the algorithm detects the objects can be made to detect specific objects, based on the products available to a certain region.

a. YOLO Algorithm - YOLO is a computer vision algorithm, that is capable of detecting objects from an image, utilizing a known set of object data. For this experiment, the algorithm is extended to identify objects from videos in each frame. The YOLO algorithm sections each frame to an $S \times S$grid. This grid is then processed as smaller cells, to detect any object from the trained corpus. The object probability for each bounded cell margin is calculated and the margin with the highest confidence for a specific object is then labelled as the same. Multiple combinations of the cells are processed to obtain the maximum confidence for any object before the prediction is made.

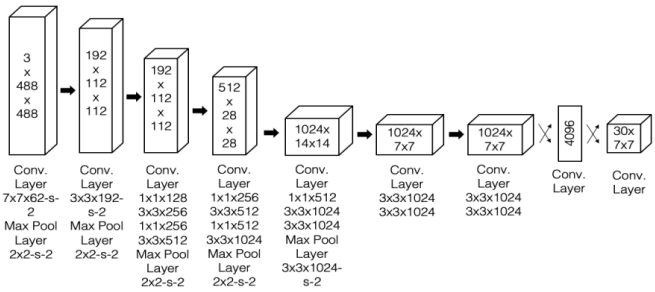

Fig. 6. YOLO Architecture

The convolution network used in the YOLO algorithm is of the following configurations.

TABLE I. CONFIGURATION OF YOLO CONVOLUTION NETWORK

Parameter	Configuration
Batch size	64
Learning rate	0.001
Decay rate	0.0005
Activation function	Leaky Rectified LinearUnit

The confidence for each margin cell is calculated using equation 1. This allows a comparative analysis amongst the cells and can derive the best combination with the highest score. Equation 2 shows the tensor format stored in the TensorFlow variables, during the execution of the algorithm. Similar to the other methods of dataset generation, the predicted labels are saved as text files.

$$\boldsymbol{P(class_i|Object) \times P(Object) \times IOU = P(class_i|Object) \times IOU} \quad \textbf{(1)}$$

Where,$P(class_i|Object)$ is the probability of $class_i$ for a detected object, $P(class_i)$ denotes the probability of a given object class,$IOU$is the intersection over union score for each cell.

$$S \times S = B * 5 + C \quad (2)$$

Where,S is the square root of the number of grids into which the image is divided,B is the confidence for a bounded box,C is the probability of the predicted class.

B. Recommendation Engine

The RE is classified into 7 versions, depending on the type of input data the engine uses. The accuracy of each version is tested with 5 user subjects for their ratings between 1, being the least relevant recommendation and 5, being the most relevant recommendation. The 7 versions are as follows.

1. **Version 1**-The RE version 1, depicts a similar working to the YouTube advertisement RE. This version utilizes the genre of the videos watched by the user and the description data of the same, to predict the probability of relevancy for a certain advertisement and suggest it to be played to the user.
2. **Version 2** - The RE version 2, uses the data generated by identifying the texts present test videos and the advertisements, along with the video description data of the same.
3. **Version 3** - Version 3 of the RE uses the dataset generated by identifying the words present in the audio of the test videos and the advertisements, along with the video description data of the same.
4. **Version 4** -RE version 4, uses the dataset generated by identifying the objects present in the test videos and the advertisements, along with the video description data of the same.
5. **Version 5**-RE version 5, uses the combination of versions 1, 2, and 3 datasets.
6. **Version 6**- REversion 6, uses the combination of versions 1, 2, and 4 datasets.
7. **Version 7** - RE version 7, uses the combination of versions 1, 3, and 4 datasets.

The engines use the cosine similarity algorithm, to calculate the similarity amongst the datasets of the videos. This similarity is taken as the relativity factor, that determines how close an advertisement is, to the users' watch pattern.

a. **Cosine Similarity**- The cosine similarity is a mathematical algorithm, that calculates the similarity between any two vectors. For this experiment, the dataset is used as vectors of texts present in the datasets. Utilizing the Natural Language Tool Kit (NLTK) dictionary package, the similarity between all the combinations of words in any two datasets are calculated. Equation 3 gives the mathematical representation of the cosine similarity calculation.

$$cosine - similarity(A, B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (3)$$

Where, A and B are the two-term frequency datasets,n is the length of the respective datasets.

Every version of the engines, recommend using the Cosine Similarity algorithm. The engines, calculate the cosine similarities between all the lemmatized texts of the description, frame texts, spoken words, and the objects, to find the data of an advertisement with the maximum cosine similarity. The advertisement with the maximum cosine similarity amongst all the genres of the users' watch history, is recommended to the user.

C. Results

For this experiment, a front-end program is implemented, to obtain the ratings for each engine form the users. The users are allowed to select and watch any number of test videos. Once the users are satisfied with the number of videos watched, the program recommends 7 advertisements, for which the users are prompted to enter a rating value between 1 and 5.

TABLE II. AVERAGE RE RATINGS (1-3)

Test Users	RE 1	RE 2	RE 3
User 1	2	3.5	3
User 2	2.3	2.3	2.6
User 3	3	2.6	3.3
User 4	1.6	2	3.6
User 5	2	2.5	3.1

TABLE III. AVERAGE RE RATINGS (4-7)

Test Users	RE 4	RE 5	RE 6	RE 7
User 1	4	4.5	5	3
User 2	3.3	4.3	4.6	4
User 3	4	4.6	5	4
User 4	3.8	4.4	4.8	3.8
User 5	3.5	4.5	4.8	4.3

These ratings reveal that the RE version 6, that utilizes the texts and objects detected in the video, recommends the most relevant advertisements. This is because, advertisements tend to visually engage the users and allocate the maximum duration of the advertisement, to show the product. The RE with the 2nd highest average user rating is version 5. This is because advertisements also tend to reinforce the product with the auditory stimulus. This audio data highly correlates with the characteristics of the product.

The sequence diagram of the RE version 6, is given in figure 7.

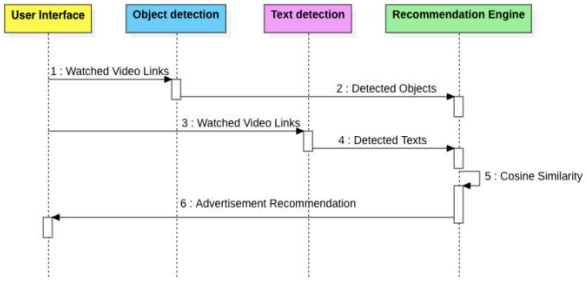

Fig. 7. RE version 6 - sequence diagram

IV. CONCLUSION

This article proposes an improved solution to the conventional YouTube advertisement RE by utilizing the features of the videos watched by the user such as the objects, texts present in the video frames, and the description data of the user watch history. From the conducted experiments, the following conclusions are drawn.

i. The texts present in the product advertisements are significant in influencing the advertisement recommendation, as it highly correlates with the nature of the product. The usage of this aspect increases the alignment of user interest in the recommended advertisement.
ii. The users tend to expect advertisement that coincides with the type of videos watched, that visually reminds them of their interests.
iii. Although the above two aspects rarely produce better results than the conventional YouTube RE separately, the combination of the two aspects consistently produces better results in recommending advertisements.
iv. The proposed RE can be used to provide targeted advertisements to monetize a small-scale video streaming website in the initial stages of revenue generation.
v. The targeted advertisements have a more significant impact on locally available products, than that of generic ones.

A. Limitations:

Although the proposed solution proves to be effective in recommending highly relevant advertisements that appeal to the user, the effectiveness is the result of a trade-off between effectiveness and cost. The YOLO algorithm utilizes sophisticated hardware components such as, high-end Graphics Processing Units (GPU) and computing components such as, Central Processing Units (CPU) and memory units that support such GPUs with high-end configuration. These types of systems are highly inefficient in an enterprise-level solution, that provides video streaming services to a significantly large user base.

B. Future Works:

Considering the cost inefficiency of the proposed solution, the RE can be used by advertisement agencies, to monetize book/movie renting platforms to recommending advertisements and books of similar genre and publication. The advertisements can be for users, of specific geographic locations, showing products available to that location.

REFERENCES

1. J. Clement. Statista, Feb 13, 2018. Accessed on: February 10, 2020. [Online]. Available: https://www.statista.com/statistics/805656/number-youtube-viewers-worldwide/
2. Google LLC. Tesseract OCR. Version 3.05.02. June 19, 2018. Accessed on: February 22, 2020. URL: https://tesseract-ocr.github.io/docs/
3. R. V.Kaushik, R.Raghu, L. M.Reddy, A.Prasad, & S. Prasanna.Ad analysis using machine learning: Classifying and recommending advertisements for a given category of videos, using machine learning. *2017 International Conference on Energy, Communication, Data Analytics, and Soft Computing (ICECDS)*, 2017 doi:10.1109/icecds.2017.8389887
4. J. Roh& S. Jin. Personalized advertisement recommendation system based on user profile in the smart phone. *2012 14th International Conference on Advanced Communication Technology (ICACT)*, PyeongChang, 2012, pp. 1300-1303.
5. J. Ruan &Z. Wang.An Improved Algorithm for Dense Object Detection Based on YOLO. *SpringerBriefs in Earth System Sciences*, 592–599, 2019. doi:10.1007/978-3-030-14680-1_65
6. J. Wang, B. Wang, L. Duan, Q. Tian & H. Lu.Interactive ads recommendation with contextual search on product topic space. *Multimedia Tools and Applications, 70(2)*, 799–820, 2011. doi:10.1007/s11042-011-0866-2
7. H.Zhang, X.Cao, J. K. L.Ho & T. W. S. Chow. Object-Level Video Advertising: An Optimization Framework. *IEEE Transactions on Industrial Informatics, 13(2)*, 520–531, 2017. doi:10.1109/tii.2016.2605629
8. X. Zhu,M. Liu,Y. Zhao, L. Dong, M. Hui& L. Kong. Product detection based on CNN and transfer learning. *Applications of Digital Image Processing XLII, 111371W*, 2019. https://doi.org/10.1117/12.2526236
9. R. Zhou, S. Khemmarat& L. Gao. The impact of YouTube recommendation system on video views. *Proceedings of the 10th Annual Conference on Internet Measurement – IMC '10*, 2010. doi:10.1145/1879141.1879193

AUTHORS PROFILE

Shanmuga Skandh Vinayak E, is a fourth-year Information Technology engineer at the SSN College of Engineering in Tamil Nadu, India. His current fields of work include image processing, signal processing and machine learning. He is interested in statistics, data science and automation.

Venkatanath A G S, is a fourth-year Information Technology engineer at the SSN College of Engineering in Tamil Nadu, India. His current fields of work include image processing and machine learning. He is interested in statistics, applied machine learning and algorithm design.

Dr. Shahina A, is a professor in the department of Information Technology at SSN. She has 20 years of teaching and research experience. She obtained her PhD from the department of Computer Science and Engineering at IIT-Madras, India. She also has an MTech from IIT-Madras. She has research interests in the areas of Machine Learning, Deep Learning and Speech Processing. She has more than 30 research publications, including in refereed international journals and international conferences.

Dr. Nayeemulla Khan A, is a Professor at the School of Computing Sciences and Engineering at VIT Chennai. He has 17 years of experience in the industry and 9 in teaching. He was the senior manager at the Airports Authority of India, when he took a break to finish his Ph.D. at IIT Madras. He then was a Research Scientist at Acusis India an MNC leading its speech recognition efforts. He has interest in the domains of Speech Recognition, Pattern Recognition and Machine Learning.