# A Novel Approach For Analysis and Understanding of Video Content

**James Huang**

International School of Beijing (ISB), Shunyi，Beijing，China 101318
Email: James.Huang@student.isb.bj.edu.cn

**Abstract:** Leveraging recent neural network studies in image caption and audio translation, we introduce a model that concatenate different clips in a video, based on their relatedness score to a set of given keywords, to form a trailer that summarizes the video. Compared to the traditional video analysis that relies heavily on meta data such as subtitles, abstracts, etc., this approach focuses on the textual information generated from each scene. Thus, the video content itself is refined and understood. Firstly, the input video is processed and divided into two types of data: image and audio. Secondly, the corresponding text data that best describes them are obtained through processing the images and audios via neural networks, like Resnet 101 and attention LSTM, as well as audio API. Thirdly, video scenes are segmented by calculating their frames and audios' corresponding text data's relatedness score to a set of input keywords. Then the scenes that best summarize each video are ua trailer. This way, a high accuracy end-to-end image analysis neural network is trained. Then a similar method is employed to rank a list of videos based on their respective trailer rankings. Comparison and analysis are performed on the ranking of the generated videos, ranking of the YouTube videos, and the ranking by surveyed participants as well as their satisfaction feedbacks. The results showed that the algorithm can better rank the videos for the consumers and generate summary trailers that provide better accuracy and higher efficiency.

## 1. Introduction

There is a common problem in today's video search. Because most videos are classified and assessed by their title/description and number of visitors, most of the times the videos show up in the search result are either the ones with similar title or description or the ones that are ranked by "hotness". The main drawback of this mechanism is that the actual content of the videos is not a factor in the relevance ranking process.

The method presented in this article aims to alleviate such problems. A model is developed to understand the hidden/untold information behind a video title/abstract – through visual and audio analysis of the video itself, utilizing neural networks and deep learning. Furthermore, with a set of input keywords, the model can synthesize a trailer from the relevant frames of the video. Preliminary study and assessment have shown effectiveness and potential usefulness for this approach.

## 2. Methodology

*2.1. ResNet 101, CoCo API 2014, HMM, Attention LSTM and BEAM*

A ResNet [1] of 101 layers is used in the study. Attention LSTM [2, 5] is employed to analyze the video. CoCo data set [6] is used for object recognition training. HMM is used to analyze the audio clips. BEAM [3] search is used to index the generated sentences.

*2.2. Video key frame extraction*

Frame-to-frame difference method [4, 7] is chosen for its adaptability and simplicity. It works to obtain the contour of a moving target by performing a difference operation on two consecutive frames of a video image sequence. Other methods considered include video-frame-clustering and lens-based key frame extraction.

*2.3. Cosine Similarity*

The cosine formula of the angle between vector A (x1, y1) and vector B (x2, y2) in two-dimensional space:

$$\cos\theta = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2}\sqrt{x_2^2 + y_2^2}}$$

The formula reflects the relative distance between the two words, so the greater the value is, the more relative these two words are with each other. It is chosen for its adaptability toward dimensionality.

*2.4. Discounted Cumulative Gain (DCG)*

DCG, iDCG, and nDCG are used to measure quality of search result list. The formula of these accumulated at a particular rank position p are defined as below, respectively.

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

$$\mathrm{IDCG_p} = \sum_{i=1}^{|REL_p|} \frac{rel_i}{\log_2(i+1)}$$

$$\mathrm{nDCG_p} = \frac{DCG_p}{IDCG_p},$$

where reli is the graded relevance of the result at position i, and relp represents the list of relevant documents (ordered by their relevance) in the corpus up to position p.

**3. Experiment & Result Analysis**

Python 3.8 and Tensorflow 2.4.0 were used. The model was trained on a Ubuntu20.04 Linux GPU server with 64 GB memory. Source code and results are stored at https://jameshuang2004.github.com/isef/ (access available upon request)

*3.1. Overall approach*

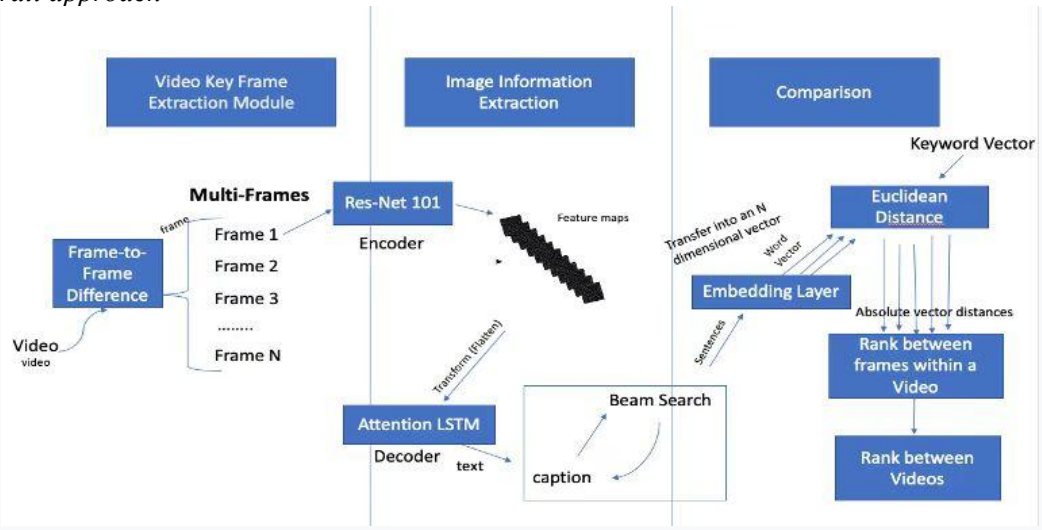


**Figure 1.** illustration of the overall approach

*3.2. Image to Text Model Training Processing*

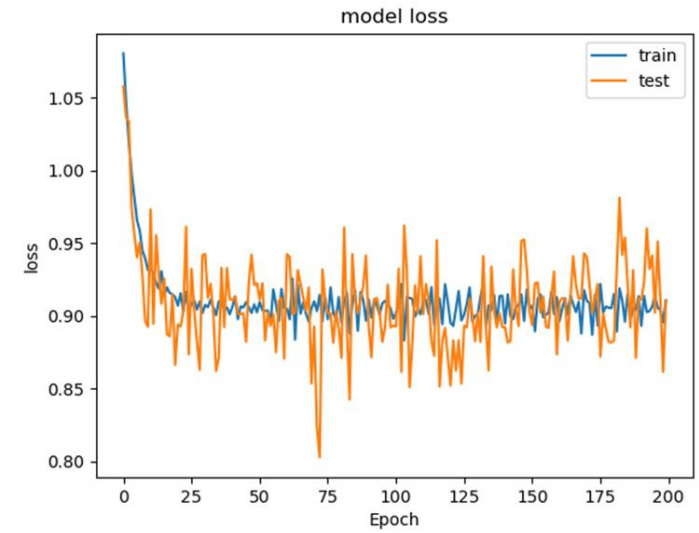


**Figure 2.** Loss vs. Epoch curve of the model

The coco data set was used as the training and testing datasets (train: test, 9: 1). The training duration was set to 200 epochs. Loss value was used to supervise the training process of the model. It can be clearly seen from the loss curve that the train set's loss has stabilized after 40 epochs.

*3.3. Model Result*

To test the effectiveness of the model, I then applied it to some test photos. Shown below are the two sets of photos.

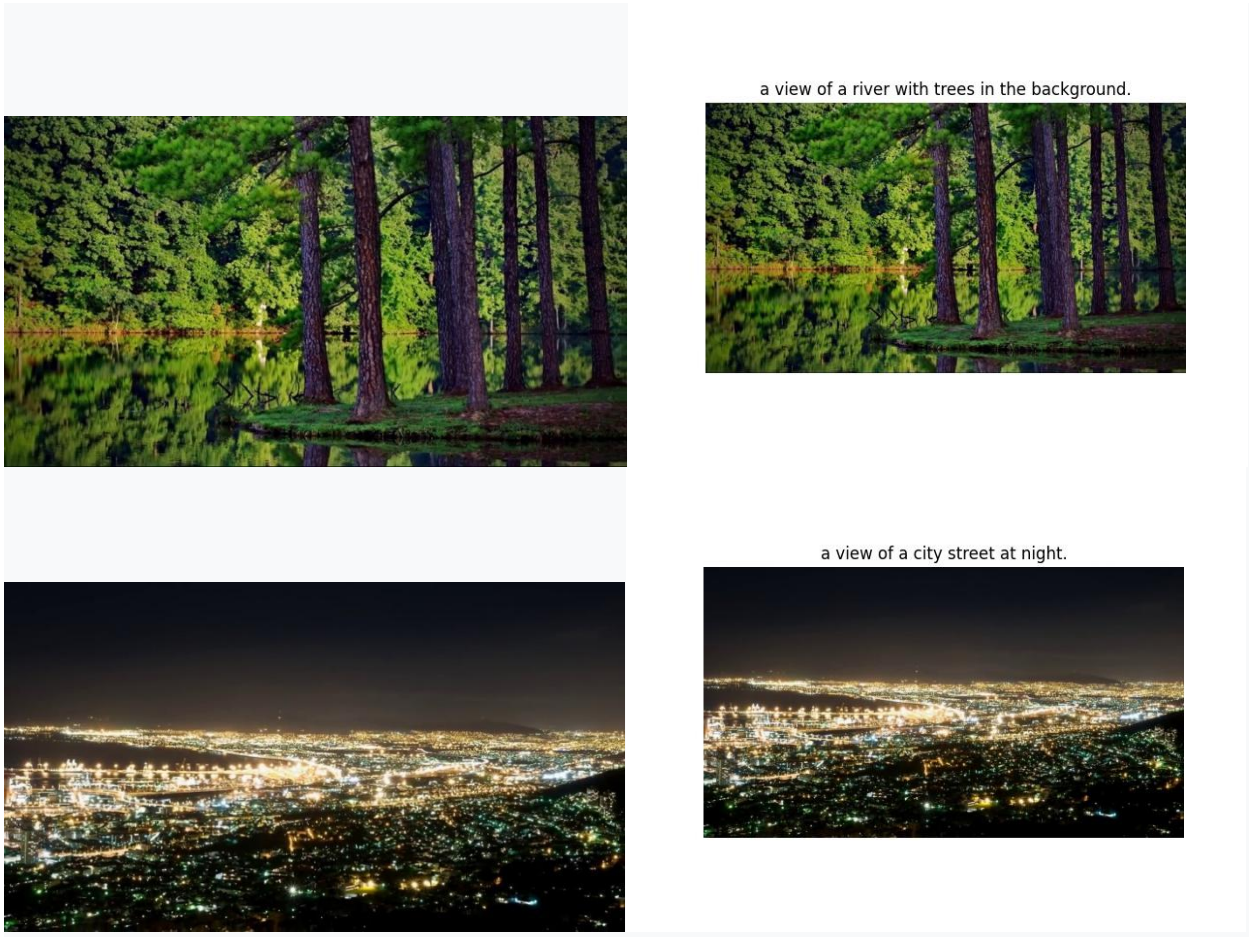


**Figure 3.** Visual to text generation by the model

The results suggest that the visual to text generation model can effectively catch the important features from the images.

*3.4. Video Segmentation*

The relationship is determined according to the distance between the keyword and each frame of the video. By setting a threshold (0.5), the greater the value of Cosine Similarity is the stronger the relationship between the two, so the frames above 0.5 will be extracted and converted into corresponding visualized time points, which will make the segment and extract new videos process easier. This way the model is able to generate a trailer from a video, if given a set of input keywords.

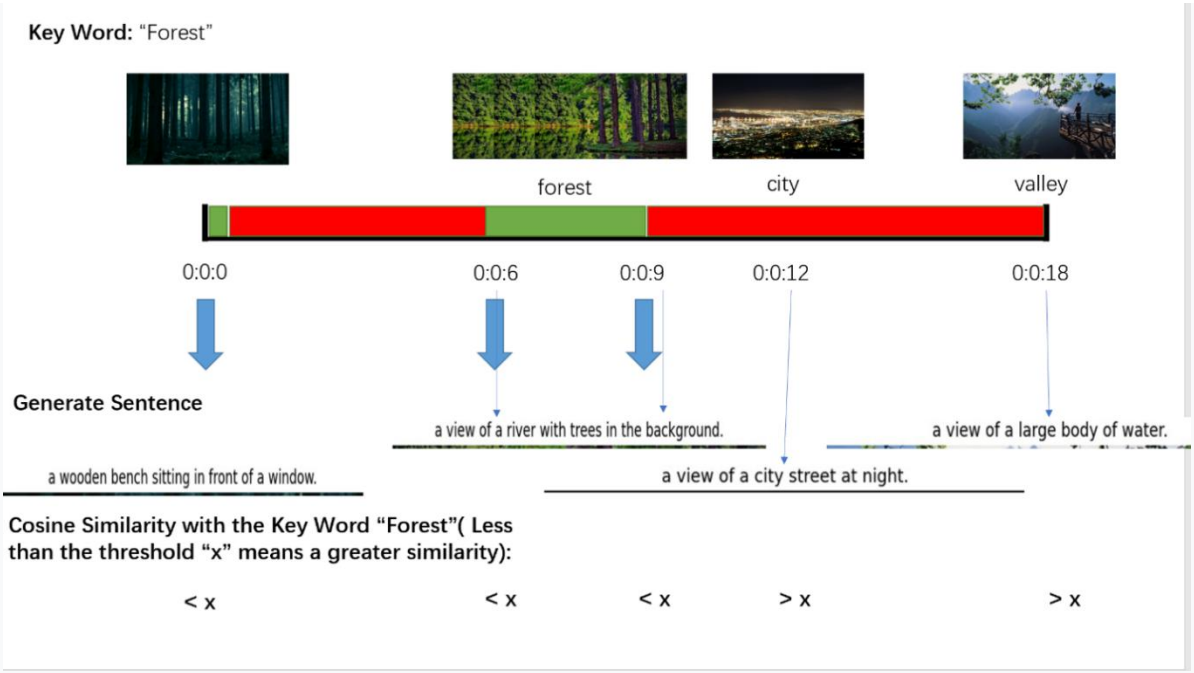


**Figure 4.** How video segmentation, visual to text generation, and keyword similarity ranking work together.

*3.5. Ranking Videos*

Currently the most commonly practiced SEO video ranking system, also used by the world's biggest video search engine YouTube, is heavily based on videos' metadata (video title, description, comments) and user engagements (number of times being clicked by users). A survey for 40 randomly chosen participants was used to rank, from 1 to 3, the usefulness of a total of 30 videos related to 3 different keywords (these videos are the top 10 search result for each keyword in YouTube). Then the survey's each response rank was compared to YouTube's original rank as well as the newly generated rank (based on extracted videos' relativeness with the keyword) through the utilization of DCG (Discounted Cumulative Gain) to determine which one fits the audience' tastes more. For example, Table 1 shows the users' ranking data on the keywords "Cadillac car".

**Table 1.** User ranking data. IDCG = sum(Rel i / Log i) = 9.20

**Participants' Ranking**

| Rank (i) | Video # | Reli | Log2i | Reli / Log2i |
|---|---|---|---|---|
| 1 | 10 | 3 | 0 | N/A |
| 2 | 9 | 3 | 1 | 3 |
| 3 | 1 | 3 | 1.58 | 1.90 |
| 4 | 4 | 2 | 2 | 1 |
| 5 | 3 | 2 | 2.32 | 0.86 |
| 6 | 2 | 2 | 2.58 | 0.78 |
| 7 | 8 | 2 | 2.81 | 0.71 |

| | | | | |
|---|---|---|---|---|
| 8 | 6 | 1 | 3 | 0.33 |
| 9 | 7 | 1 | 3.17 | 0.32 |
| 10 | 5 | 1 | 3.32 | 0.30 |
| | | | total | 9.20 |

**YouTube's Ranking**

**Table 2.** YouTube ranking data. nDCG = sum(Rel i / Log i) / IDCG = 8.61 / 9.20 = 0.936. nDCG indicates the similarity between the tested rank and the reference rank. In this case, the NDCG of the Youtube rank to People's Tendency rank is about 0.936, meaning that there's a 93.6% similarity between the two.

| Rank (i) | Video # | Rel i | Log i | Rel i / Log i |
|---|---|---|---|---|
| 3 | 1 | 3 | 1.58 | 1.90 |
| 6 | 2 | 3 | 2.58 | 1.16 |
| 5 | 3 | 3 | 2.32 | 1.29 |
| 4 | 4 | 2 | 2 | 1 |
| 10 | 5 | 2 | 3.32 | 0.60 |
| 8 | 6 | 2 | 3 | 0.67 |
| 9 | 7 | 2 | 3.17 | 0.63 |
| 7 | 8 | 1 | 2.81 | 0.36 |
| 2 | 9 | 1 | 1 | 1 |
| 1 | 10 | 1 | N/A | N/A |
| | | | total | 8.61 |

**Generated Video's Ranking**

**Table 3.** Generated video ranking data. nDCG = sum(Rel i / Log i) / IDCG = 9.09 / 9.20 = 0.988

| Rank (i) | Video # | Rel i | Log i | Rel i / Log i |
|---|---|---|---|---|
| 5 | 3 | 3 | 2.32 | 1.29 |
| 3 | 1 | 3 | 1.58 | 1.90 |
| 4 | 4 | 3 | 2 | 1.5 |
| 6 | 2 | 2 | 2.58 | 0.78 |
| 2 | 9 | 2 | 1 | 2.00 |
| 1 | 10 | 2 | N/A | N/A |
| 9 | 7 | 2 | 3.17 | 0.63 |
| 7 | 8 | 1 | 2.81 | 0.36 |
| 8 | 6 | 1 | 3 | 0.33 |
| 10 | 5 | 1 | 3.32 | 0.30 |
| | | | total | 9.09 |

The nDCG of the Generated Video rank is about 0.988, suggesting 98.8% similar to what most people think, and has a 5% jump on the accuracy than the Youtube Rank. This shows that our approach can better predict users' tastes than the original SEO method based on video popularity and metadata. More experiments were done on different set of keywords, all results have shown similar trends. Another survey with some random group of people on their opinions on the 2 sets of videos -- the original long ones and the newly generated clips – revealed that the majority agreed that they could make sense of the autogenerated clips.

## 4. Conclusion

In conclusion, a novel approach that can analyze and understand the content of videos has been developed. Using a keyword approach, the model can generate meaningful trailer clips from videos, based on the analysis and understanding of the video frames. Side-by-side comparison of ranking data suggests it could result in higher user satisfaction than the YouTube metadata rank. The accuracy effectiveness of the model can be further improved if more time, data and resource are available. Furthermore, this model can be useful in video understanding, video content indexing and search, video recommendation, as well as other scenarios such as content filtration, video recognition and advertisement, etc.

## 5. References:


[1] He et al. ,"Deep Residual Learning for Image Recognition", https://arxiv.org/abs/1512.03385

[2] Phi M., "Illustrated Guide to LSTM's and GRU's: A step by step explanation", https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

[3] Doshi K., "Foundations of NLP Explained Visually: Beam Search, How It Works", https://towardsdatascience.com/foundations-of-nlp-explained-visually-beam-search-how-it-works-1586b9849a24

[4] Amanwalia123. (n.d.). AMANWALIA123/KeyFramesExtraction: This repository contains script to divide a video into key frames. GitHub. Retrieved February 7, 2022, from https://github.com/amanwalia123/KeyFramesExtraction

[5] DeepRNN. (n.d.). DeepRNN/image_captioning: Tensorflow implementation of "show, attend and tell: Neural image caption generation with visual attention". GitHub. Retrieved February 7, 2022, from https://github.com/DeepRNN/image_captioning

[6] Lin T., "Microsoft CoCo: Common Objects in Context", https://arxiv.org/abs/1405.0312

[7] Nikhilroxtomar. (n.d.). Nikhilroxtomar/extract-frame-from-videos: A simple piece of code written in python that will help you to extract the frame from a video or a set of video at once. GitHub. Retrieved February 7, 2022, from https://github.com/nikhilroxtomar/Extract-Frame-from-Videos