International Journal of Engineering, Management, Humanities and Social Sciences Paradigms (IJEMHS)
Volume 38, Issue 01 and Publication Date: 27th April, 2026
An Indexed, Referred and Peer Reviewed Journal
ISSN (Online): 2347-601X
www.ijemhs.com

Question Answering & Youtube Video Summarization Bot (QA & YT Video Summarization Bot)

Neeraj Kumar[1], Preeti[2] and Preetishree Patnaik[3]

[1]M.Tech. Scholar, Department of Computer Science & Engineering, Sat Kabir Institute of Technology and Management Bahadurgarh, India
nirajkumar776.ns@gmail.com

[2]Assistant Professor, Department of Computer Science & Engineering, Sat Kabir Institute of Technology and Management Bahadurgarh, India
preetikataria007@gmail.com

[3]Assistant Professor, Department of Computer Science & Engineering, St. Andrews Institute of Technology & Management, Gurugram, India
patnaik.preetishree@gmail.com

Abstract

Bot is a computer program that performs automated activities on the internet. Our creation is a Question Answering and YouTube Video Summarization Bot, or in other words, a QA and YT Video Summarization Bot. Our Bot is a computer program intended to replicate the roles of a Question Answering System, including text and visual summarisation, as well as Video summarisation of YouTube Videos. The production process forms the basis of the assurance process of our bot, and it comprises two types: visual verification process and full verification process. In a full-fledged assurance system, the system produces the answer to the query the user poses out of the text presented to it. Our model has the power to answer a lot of language queries using a trained BERT model. The answers given by our model are only vague. We port our code to Gradio and run it on Hugging Face.

Keywords: *Transformers, BERT, BART, VILT, Gradio, Hugging face*

1. Introduction

Question Answering is a trending area of research in NLP, and the practical implications of the Question Answering system are many, including customer care chatbots and virtual assistants. NLP is becoming advanced at a rapid speed with the creation of large pre-trained language models like BERT and GPT-3. These models have given state of the art result.

Performance on over 11 applications of NLP, including language translation, sentiment analysis and question answering. This paper will discuss how these pre-trained models can be used in three different tasks, such as text-based question answering, visual question answering and video summarization. Visual QA (VQA) is a variant of QA that implies answering questions based on visual data. We introduce a question-answering tool in our Bot that applies BERT to find accurate answers to questions according to a specific context. We also introduce a visual Question answering tool, which is based on a pre-trained model VQA and a YouTube Video Summarization tool, which is based on a pre-trained summarization model to create summaries of available captioned YouTube videos.

To create a bot, we want to offer real answers in both text and images, rather than using search engines, like Google. Our model can be used by people who are pressed for time and researchers who need to find correct and quick answers. Answers to the questions that might not be visible to the user can be given by our bot. Besides this, we have also released a YouTube video summarisation bot that enables users to summarise long videos without necessarily having to watch them. In case the user is the one who creates the content, he or she can get feedback on the quality and content of the videos he or she uploads. They can also enhance their understanding by comparing their own perception of the video to that of the bot.

We have 3 tabs in our model. The first tab is the Visual QA tab, whereby the user can input an image and a question regarding the image, and our model will be able to find the answer to the question inside the image. On the same note, in the case of the Comprehensive QA system, the user must enter text and a question, and the response based on the context will be produced.
The video summarization model takes a video link on YouTube as input and produces a summary of the video. Nevertheless, the captions need to be included in the YouTube video to make the model functional.

2. Literature Review

Pre-trained models have been utilized in question- answering tasks in many studies. As an illustration, the BERT model has been trained to state-of-the-art on a variety of question-answering benchmarks (Devlin et al., 2018). Our bot is based on these studies, where it creates a question-answering bot, which relies on BERT to give the right answers to a question posed in relation to a particular situation.

The pre-trained models have been studied in a few studies concerning text summarization tasks. High-quality text summaries have been created using the BART model (Lewis et al., 2020). The application of various pre-trained models like T5 and PEGASUS to the task is also investigated by other researchers (Raffel et al., 2020; Zhang et al., 2020). In the model, we have a YouTube Video Summarization tool that relies on a pre-trained summarization model to create summaries of YouTube videos with captions.
The code described in this paper is based on a transformer model, which has demonstrated state-of-the-art results in several benchmark datasets.

3. Methodology

Our question-answering Bot was based on the BERT model, and our Video Summarization based on the BART model. The Question Answering Bot accepts context and a question as input and provides the answer to the question. YouTube Video Summarization is an input tool that accepts a YouTube video link and produces a summary of the video based on a pre-trained summarization model. Transformers and Gradio libraries were used to create these tools and were deployed to a web interface.
The adopted code employs transformers and Gradio libraries to generate a user-friendly interface of the QA and video summarization models. The QA model is founded on the BERT model that has been fine-tuned on the SQuAD dataset. The VQA model is run with the VILT architecture with a pre-trained model on a VQA dataset. The YouTube video summarization model is based on the Facebook BART model to summarize the videos.

3.1 Equations

3.1.1 For Visual Question Answering System

ViLT attention mechanism is one of the methods by which the model puts attention on various regions of the image and doubts when producing an answer. It computes a weight on each pair of image features and question tokens, indicating the importance of each pair to answer the question. After that, it takes the most significant pieces of the image and questions and processes them to come up with the answer.
Mathematically, attention weight is obtained as a dot product of the image feature and the question token's embeddings. The similarity score thus obtained is fed into a softmax to get a weight vector:

α=softmax(similarity(image features, question embeddings))

A weighted sum of the image features is then computed using the weight vector and combined with the question embeddings and fed through a sequence of Multi-Layer Perceptron (MLP)layers to produce the answer: answer=MLP([weighted sum(image features, 0 -1); question embeddings]).

3.1.2 For Text-Based Question Answering System attention mechanism

The attention mechanism in a text-based QA system is used to enable the model to pay

attention to the relevant portions of the input text when responding to a query. The attention mechanism mathematical equations in a text-based question-answer system are as follows.

1.Calculate both the query and the set of values:

- Query: q
- Set of values: V

2.Calculate the compatibility scores:

- Scores: s = q * V^T

3.Normalize the scores:

- Weights: w = softmax(s)

4.Calculate the weighted sum:

- Weighted sum: a = w * V

The weighted sum is the output of the attention mechanism and is the most relevant sections of the input text to the question. The model is then able to use this to produce an answer.

3.1.3 For YouTube Video Summarisation

The focus algorithm of an overview of a YouTube video with subtitles through the Facebook BART model, formulated in mathematical equations:

1. Calculate query and set of values:

Query: q (in this case, a summary of the video)

Set of values: V (in this case, the transcript of the video)

2. Calculate the compatibility scores:

Scores: s = q * V^T (where * denotes matrix multiplication and ^T denotes matrix transpose)

3. Normalize the scores:

Weights: w = softmax(s) (where softmax is a function that converts the scores into a probability distribution)

4. Calculate the weighted sum:

Weighted sum: a = w * V (where * denotes matrix multiplication)

The weighted sum is the output of the attention mechanism, and it reflects the most relevant bits of the transcript to the summary. The model can then use this to produce a summary of the video.

3.2 Architecture

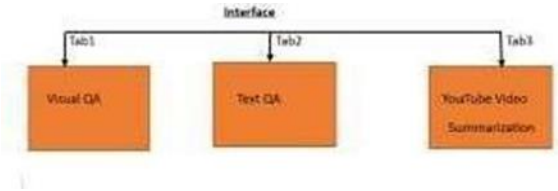

Fig. 1 User interface contains 3 tabs

Fig. 2 (a) Working of QA & YT video summarization

4. Analysis

Our Bot has been analysed using various examples, some of which are highlighted below.

4.1 For visual QA

Fig. 2 (b)

Table 1

Questions Asked By user	Answers ven by Bot
How many Horses are there?	8
What are horses doing?	running.
Where are horses?	Field
Is it raining?	no
is it sunny?	yes
What is the color of sky?	blue
What is the color of grass?	green
How many trees are there?	0
is there any grass?	yes

4.2 For text based QA

a) For Hindi Language

International Journal of Engineering, Management, Humanities and Social Sciences Paradigms (IJEMHS)
Volume 38, Issue 01 and Publication Date: 27th April, 2026
An Indexed, Referred and Peer Reviewed Journal
ISSN (Online): 2347-601X
www.ijemhs.com

Fig. 3

b) For English Language

Fig. 4

c) For Nepali Language

Fig. 5

English = Mount Everest is the highest mountain in the world, and its peak is 8,848 meters (29,029 feet) above sea level. It is found in the Mahalangur range along the Nepal-Tibet (China) border of the Himalayas.

Nepali = सगरमाथा संसारको सबैभन्दा अग्लो हिमाल हो, र यसको चुचुरो समुद्र सतहबाट ८,८४८ मिटर (२९,०२९ फिट) उचाइमा छ। यो हिमालयको नेपाल-तिब्बत (चीन) सीमामा रहेको महालंगुर पर्वत शृङ्खलामा पाइन्छ।

4.3 YouTube video summarization

Link provided by user-(147) Machine Learning In 2 Minutes | What Is Machine Learning? | Machine Learning Basics | Simplilearn – YouTube

Summary created by Bot-Machine learning is the programming machine science human beings not in the specific sense to think and do like programmed too. We already apply machine learning in our everyday life, unknowingly, email spam recognition, spell check, and even the YouTube video recommendation that has taken you here are applied through machine learning.

Machine learning involves the use of algorithms to learn tasks. Machine learning is bringing us to a new era where machines can think and learn, and has provided us with an entirely new plethora of job opportunities and developing machines which can sense, reason, act and adapt. Deep learning is a sub-branch of machine learning that is based on the way the human brain works. Watch other similar videos to nerd up and be certified by clicking here.

5. Code

https://huggingface.co/spaces/hema1/flag
https://huggingface.co/spaces/hema1/flag/tree/main

6. Result

We have tested our type of question answering bot that we have created, and it performs well on a variety of questions. The accuracy of bots will differ with the complexity of a question and the degree of context given. YouTube video editors can also be availed of subtitles using the YouTube Video Editor tool. The quality of the compilation created varies with the duration and the intricacy of the video. The applied model demonstrates encouraging outcomes in responding to written and pictorial questions. Text-based assurance models give correct answers to the situation and questions. The VQA model performs well when it comes to responding to questions that are grounded in visual information. It was also discovered that the video summary template offers an accurate and brief YouTube video summary. The model has a user-friendly user interface, whereby the user can interact with the model easily.

7. Discussion

We have demonstrated in our research that pre-trained language models may be very useful in the creation of natural language processing tools. Our Question Answering Bot and YouTube Video Summarization tool have many possible applications, such as in chatbots and customer support. Our bot however suffers some limitations including the extent of the text-based QA model input context and the fact that the VQA model relies on the quality of the image. Video summarization model also involves captions to produce the summary, and our models need some time to produce the answer and summary. Further research might overcome these shortcomings and discuss how the models can be enhanced. Also, it is possible to expand the video summarization model to support longer videos and provide a more accurate summary.

8. Conclusion

In this paper, we have provided a code where we have used state-of-the-art transformers as comprehensive QA solutions. The code has two models of QA, one that is able to respond to text-based questions and another that is able to respond to questions that are based on visual information. The code also has a video summarization model, which summarizes YouTube videos. The performance of the code can be helpful in different areas, and it is applicable in different fields. can be enhanced by fine-tuning the models using domain-specific data. On the whole, the code is an important contribution to the sphere of QA and the summarization of YouTube videos, and it could be incorporated into different applications and websites to enhance the user experience.

References

[1] S. Antol et al., "VQA: Visual Question Answering," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 2425-2433, doi: 10.1109/ICCV.2015.279.

[2] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel, “Visual Question Answering: A Survey of Methods and Datasets”, Computer Vision and Image Understanding, Vol. 163, 2017, pp. 21-40.

[3] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, and V. Stoyanov, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension", Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5568–5578.

[4] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision", arXiv:2102.03334, 2021.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, 2019, pp. 4171–4186.

[6] D. Gupta, S. Kumari, A. Ekbal, and P. Bhattacharyya, "MMQA: A Multi-domain Multilingual Question-Answering Framework for English and Hindi", arXiv preprint, 2021.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Multilingual BERT: Injecting Multilingual Information into Pretrained Models", arXiv preprint,2019.

[8] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", arXiv:1910.10683, 2019.

[9] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization", arXiv preprint, 2019.

[10] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks", arXiv preprint, Vol. X, No. X, 2019.

[11] A. Abid, A. Ahmad, and M. Trenary, "Gradio: Hassle-Free Sharing and Testing of

ML Models in the Wild", arXiv:1906.02569, 2019.

[12] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Davison, "The Hugging Face Transformers Library: A State-of-the-Art Natural Language Processing Library", Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.