Automated Notes and Question Generation

Siddhi Zanwar
dept. of CSE-AIML,
PES University
Bengaluru, India
siddhizanwar03@gmail.com

Shusrith Srikanth
dept. of CSE-AIML,
PES University
Bengaluru, India
shusrith03@gmail.com

Anushka Ghei
dept. of CSE,
PES University
Bengaluru, India
anushkaghei07@gmail.com

Dhriti Rajesh Krishnan
dept. of CSE, PES University
Bengaluru, India
dhriti.krishnan@gmail.com

Srinivas K S
dept. of CSE-AIML, PES University
Bengaluru, India
srinivasks@pes.edu

***Abstract*— This research paper focuses on augmenting the process of learning a concept through the automated generation of notes and MCQs from educational videos. Our system works in three steps: (i) It transcribes and analyzes the video coupled with extracting keywords and concepts, (ii) then it generates detailed and informative notes, personalized to the video content and (iii) it creates Multiple Choice Questions (MCQs) concerning the content. The system currently uses Machine Learning algorithms and Natural Language Processing to carry out the above-mentioned. This research holds relevance for online courses, flipped classrooms, and self-directed learning along with promoting better understanding of a concept among learners.**

Keywords*— Education, Learning, Machine learning, Natural Language Processing

I. Introduction

In today's generation, where the primary source of learning has become the internet, educational videos tend to turn into a staple in both classrooms and online learning platforms. These videos offer a more visual understanding along with provision of dynamic content which reaches a more diverse set of students with different learning preferences. However, unlike a traditional classroom setup where teachers and educators can provide prompt doubt clarification and feedback, educational videos pose problems such as (i)Passive Learning, while common, can hinder academic performance and engagement [1], where students simply consume information without having any active engagement with the content and (ii)Information Overload, the imbalance between the demands of information processing and the capacity to handle it [2], where students are immensely impacted by the amount of information put forward. These problems overall lead to poor retention of the concept in focus that result in gaps during learning. In addition to this, the lack of relevant study material, tailored to the video content, poses a notable challenge for students. Without informative notes and practice questions, one may find it burdensome to merge their understanding of the concept and assess themselves on the same. Taking all the above into consideration, automated systems that are based on Machine Learning and Natural Language Processing techniques offer a promising approach to enhance the overall learning curve for students that rely on educational videos. This paper focuses on the development of such automated systems and also illustrates their potential to bring about a change in the current learning experience for students.

II. Literature survey

According to Radford et al. [3], Whisper is resilient in speech recognition, without requiring dataset-specific fine-tuning. As the model size increases, Whisper demonstrates its capacity to generalize well across datasets. Whisper also excels at multilingual voice recognition and is resistant to noise. However, it suffers from long-form transcription limitation and language bias. For our system, Whisper's model has been used to transcript audio from the videos. According to Lewis et al. [4], Bidirectional and Auto-Regressive Transformer (BART) performs well on a variety of tasks, such as text production, comprehension, and machine translation. The model is extremely adaptable, allowing for a wide variety of noising schemes during pre-training. However, its uni-directional decoder layers may lower performance on selective tasks. For our system, BART has been used to get the abstractive summary of the audio transcription. According to Devlin et al. [5], "Bidirectional Encoder Representations from Transformers (BERT) is intended to pre-train deep bidirectional representations from unlabelled text", allowing it to collect context in both the left and right directions across all layers. BERT may be fine-tuned with just one additional output layer, without requiring significant task-specific architecture changes. BERT may demand more computing power; a lot of training time and it suffers from hyperparameter sensitivity. For our system, BERT has been used to get the extractive summary of the audio transcription. According to He et al. [6], Mask Region-based Convolutional Neural Network (Mask R-CNN) produces excellent segmentation masks for each occurrence of an item while also effectively detecting objects in an image. The framework's versatility is demonstrated by how easily it can be applied to

other tasks, like human position estimation. The design of Mask R-CNN is not speed optimized and it is highlighted that domain knowledge may be a complement, although the framework's method for human pose estimation does not make extensive use of it. For our system, MASK R-CNN has been used for masking undesired objects and filling the spaces with relevant content. According to Jiang et al. [7], You Only Look Once (YOLO) is prominent for its quick inference speed and when compared to certain other object identification algorithms, the YOLO architecture is more easily understood and put into practice due to its plain and simple architecture. YOLO exhibits a high degree of generalization and encodes global information and lowers background detection errors by taking the full image into account during the detection process. However, YOLO may lose some accuracy when compared to slower, more intricate algorithms, particularly when handling small or closely clustered items.

III. Methodology

Using a YouTube video's URL, we extract the audio and the video separately. The audio obtained is then processed by Whisper, leaving us with an unstructured text corpus containing the transcribed audio.

A. Notes Generation

This unstructured corpus undergoes conventional preprocessing, i.e., tokenization, removal of stop words, and lemmatization. To structure the pre-processed corpus further for notes generation, we use Non-Matrix Factorization (NMF) for topic modeling. The process starts with vectorization of the pre-processed text using Term Frequency- Inverse Document Frequency (TF-IDF). The term frequency (TF) constituent measures how frequently a term occurs in a document, whereas the inverse document frequency (IDF) constituent reduces the weight of the frequently occurring terms in the corpus, highlighting more unique terms. The process of vectorization is followed by application of NMF, which is imported from the SciKitLearn library, to extract topics. Here, the TF-IDF matrix is decomposed by NMF into two lower-dimensional matrices- W(the document-topic matrix) and H(the topic-term matrix) such that the product of W and H will approximately give back the original matrix. The number of topics is user-defined and can be changed according to requirements. Next, we extract the keywords from the topics obtained through NMF. The keywords are determined on the basis of their weights in the H matrix, where higher the weight - greater the importance within a topic. We then calculate the importance of the keywords based on their frequency in the topics. Next, we associate each document with its dominant topic and organize documents by topic, followed by generating semantically meaningful topic names. The number of words in the topic name is again user-defined and can be adjusted based on requirements. Finally, the documents are printed with their respective semantically meaningful topic names.

B. Summarization

The unstructured text corpus is also used for the generation of summaries. This is done by using BART and BERT models. Where BERT provides extractive summaries, BART provides abstractive summaries. The summaries generated by these two models are combined to make one meaningful summary that will used for the generation of our Multiple-choice Choice Questions (MCQs)

C. Key-frame extraction

To visualize the concept better, we also add images from the video in the notes. This is done by setting up a video capture object to read the video files and then defining two thresholds, one for black frames and another for changes between consecutive frames. We iterate through each frame of the video and calculate the average pixel intensity of each. We skip frames with intensities below a certain threshold (black frames). We also compute the absolute difference between the frame and its preceding frame and calculate the mean intensity difference. If the difference of a frame exceeds a specified threshold, it is identified as a keyframe and saved as a PNG image.

D. Image Generation

Assuming videos that have a person in front of a screen explaining a concept, the images need to contain only the information on the screen and not the person. For this, a Mask R-CNN model is used for semantic segmentation and object detection, to find and remove human figures from the images. Every loaded image is passed through the Mask R-CNN model, which identifies the presence and location of objects in the frame, creates binary masks for every object instance, and produces other vital details such as class probabilities and bounding boxes. From the Mask R-CNN output, if a person is present in the keyframe, the corresponding binary mask of the person is extracted. This mask is modified by replacing the pixels inside the mask with pixels from the surrounding area. Finally, to ensure a flawless transition between the modified pixels, we perform blending or smoothening operations.

E. Question Generation

The question generation involves fine-tuning a T5 transformer on the sQuad 2.0 dataset. The T5 transformer is a text-to-text transformer which is trained on a denoising objective, to predict shuffled words in a sentence. It can be easily fine-tuned for specific tasks, such as question generation since it can learn the patterns in the structure of a question very well. The model comprises two trained T5 transformers. The first transformer takes a paragraph of text as context and forms MCQs based on a given answer keyword. The paragraphs to be used as context are selected from the summary by applying the same preprocessing that was applied for the generation of notes. NMF and TF-IDF give us the most important keywords that can be used as answers to the questions. When a [MASK] element is provided as an answer input, it can generate a question-answer pair from the context. The question generator can be trained on any text corpus to generate more context aware and relevant questions.

F. Distractor

The second T5 transformer in the model is trained on the RACE dataset and generates the distractors by taking in the context, question, and answer as inputs. The distractor is fine-tuned using sense2vec to generate the most semantically similar phrase to the given input. The generated phrase is modified with false information and replaced keywords,

generating a set of incorrect options. The distractor can be fine-tuned on any text material to generate incorrect options closer to the actual answer.

IV. EXPERIMENTAL SETUP

PyTube is a Python library that provides an interface to access and download YouTube videos.

PyTorch Lightning's LightningModule is the framework used extensively for question generation. The LightningModule is a high-level framework over PyTorch which provides a high level of abstraction and a very scalable structure to training of models. It integrates effortlessly with any available hardware like CPU, GPU or TPU. The module provides a workflow for all the stages of model development, while providing fine grained control and flexibility. The integration with the DataLoader makes it very easy to train, test, and validate the data. It worked seamlessly with the T5 transformers, allowing efficient training because of its modularity. The vast customization also allowed us to fine-tune the model in the way we wanted.

OpenAI Whisper is used for producing the audio transcription of the video. Whisper is a state-of-the-art automatic speech recognition model. It is a transformer-based model with an encoder and a decoder. The encoder converts log-mel spectrograms into a set of patterns, while the decoder converts these patterns into a set of probabilities for a sequence of tokens that give the final transcript. It is trained over a large number of languages and can successfully filter out background noise and handle various accents of speech as well. It proved to be one of the best models for automatic speech recognition and had a very simple and usable API. The model is fully open source as well, reducing the operating cost of the workflow and increasing the availability to prospective users. It has one of the lowest error rates compared to other state-of-the-art models.

The Hugging Face Transformers library consists of many pre-trained models – (i) The **BART model** is a transformer based state-of-the-art seq2seq model that uses bidirectional encoders and auto regressive decoders. BART is trained on a denoising task which makes it well suited for tasks like text summarization, comprehension and generation. The system uses the BART model for abstractive summarization of the transcript generated by the speech recognition model. (ii) The **BERT model** is a bidirectional, transformer based seq2seq model. It uses a random masking training technique and combines sentences for context aware predictions. It is well suited for sentence classification, text summarization and named entity recognition. The system uses the BERT model for extractive summarization of the transcript generated by the speech recognition model - A combination of the abstractive and extractive summaries using a few NLP techniques forms the corpus for question generation.

The **OpenCV** library is used for video processing along with **NumPy** for numerical computations, and **PIL** for image manipulation to extract keyframes from the video.

V. RESULTS AND CONCLUSIONS

The application is capable of generating a well-structured document, with relevant headers and images. Given a lack of other metrics, human evaluation shows that most of the text and images extracted are legible and meaningful. The questions being generated were evaluated against rouge, relevance, and grammaticality. Relevance was calculated by calculating the ratio of the cosine similarity of the question generated against the text corpus by the cosine similarity of the test questions. The relevance indicates how meaningful the questions are concerning the text corpus and the test questions.

The generation of questions is subjective because the same question can be framed in many ways and still carry meaning. This renders NLP metrics like Bleu and Rouge ineffective in showing the true accuracy of the predictions. The relevancy shows that the model can capture most of the important features of the context and produce questions of good quality. The human evaluation involved 20-25 students using the model before their exams to check how helpful the notes and questions generated were. Human evaluation of the questions also revealed that the model understood the relationship between various objects in the sentence and was able to generate questions that test the understanding of the content to a good extent. Human evaluation of the distractors also showed that the generated distractors were similar to the correct option and the multiple choice question as a whole resembled the way a human would frame it.

VI. DISCUSSION

The results of our evaluation metrics depict a promising performance of the question generation system. While traditional NLP metrics like Bleu and Rouge fell short of capturing the quality of the generated questions, metrics including Relevance, Grammaticality, and Rouge scores, provide a more comprehensive assessment. Rouge scores, namely, Rouge1 and RougeL, indicate very close alignment with corpus, though Rouge2 on the other hand suggests room for improvement when it comes to capturing word relationships. Relevance is measured by cosine similarity and it shows a high correlation with the original text corpus. This indicates effective capture of information. The grammaticality score ensures that the majority of the questions generated are grammatically correct.

Comparing our system with the existing conventional methods of note-taking and question generation, our system offers a few notable advantages as our system reduces the amount of effort and time required for manual note-taking and question generation, hence allowing students to focus more on understanding and grasping the topic more efficiently.

By providing personalized study materials that are tailored to the content of the video, our system encourages active engagement among students. Moreover, with the accessibility of detailed notes and relevant questions, our system can also be used in a classroom setting as well where teachers/educators can utilize it to support and enrich their teaching practices.

VII. FUTURE WORK

Looking ahead, there are several paths for enhancing our system. One area of future work is the refinement of content extraction to improve its accuracy and reliability across widespread video content. Furthermore, future research could emphasize expanding the applicability of our system to different educational domains and contexts. This could also include exploring the effectiveness of our system in multilingual settings. Additionally, exploring the potential of incorporating other multimedia and interactive components in the notes and questions is suggested as this would lead to more engagement of students with the content.

Our current architecture uses Mask R-CNN and T5 transformers for image generation and question generation respectively. Though effective, incorporation of newer technologies such as Detectron2 and Retrieval-Augmented Generation (RAG) can offer notable improvements in the concerned fields. Building on Mask R-CNN, Detectron encompasses advanced optimizations and features for more precise, accurate and detailed object segmentation, leading to better image generation. Meanwhile, RAG well combines retrieval-based and generative methods, thereby augmenting question generation. It can also work effectively on topic modeling and

However, both Detectron and RAG are computationally expensive due to which their integration into the current system has been restricted. Addressing these challenges is unavoidable to exploit their complete potential.

Lastly, another considerable area of future work would be extending the system's pipeline further for the generation of slides and questions of varied difficulty levels for the provision of better assessment with regard to the topic in focus.

VIII. TABLES AND FIGURES

TABLE I. EVALUATION METRICS

Metric	**Score**
Rouge1	0.456
Rouge2	0.233
RougeL	0.481
Relevance	0.935
Grammaticality	0.844

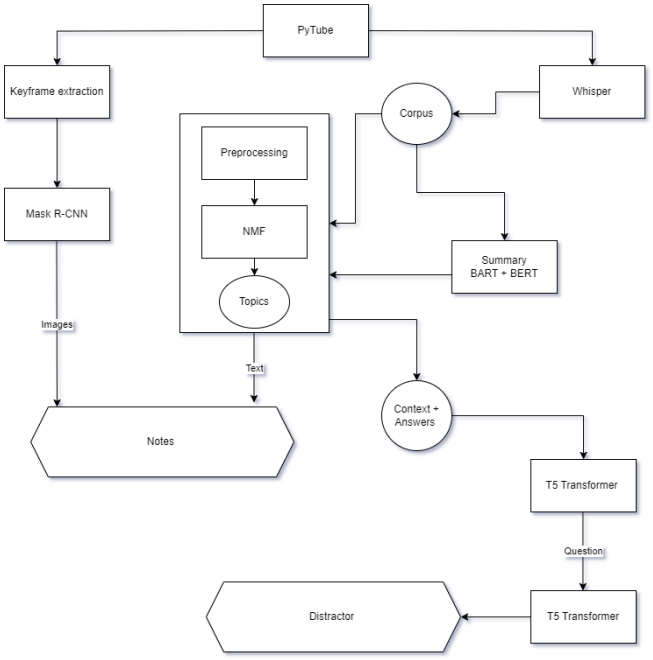

Fig. 1. Pipelined workflow of methodical question, answer, and distractor generation

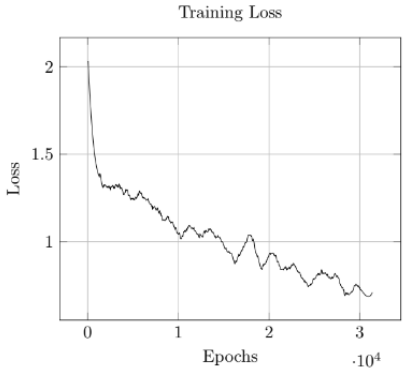

Fig. 2. Learning curve

REFERENCES

[1] S. Dolan, D. Mallott, and J. A. Emery, "Passive learning: a marker for the academically at risk," *Medical Teacher*, vol. 24, pp. 648-649, 2002.

[2] A. G. Schick, L. A. Gordon, and S. F. Haka, "Information overload: A temporal approach," *Accounting, Organizations and Society*, vol. 15, pp. 199-220, 1990.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, Jul. 2023, pp. 28492-28518.

[4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, ... and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in Proceedings *of the IEEE International Conference on Computer Vision*, 2017, pp. 2961-2969.

[7] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A Review of Yolo algorithm developments," *Procedia Computer Science*, vol. 199, pp. 1066-1073, 2022.

[8] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*, Jun. 2022, pp. 12888-12900.

[9] J. Haddock, L. Kassab, S. Li, A. Kryshchenko, R. Grotheer, E. Sizikova, C. Wang, T. Merkh, R. W. Madushani, M. Ahn, D. Needell, and K. Leonard, "Semi-supervised NMF Models for Topic Modeling in Learning Tasks," *arXiv preprint arXiv:2010.07956*, 2020.

[10] K. Vachev, M. Hardalov, G. Karadzhov, G. Georgiev, I. Koychev, and P. Nakov, "Leaf: Multiple-choice question generation," in *European Conference on Information Retrieval*, Cham: Springer International Publishing, Apr. 2022, pp. 321-328.

[11] P. Laban, C. Wu, L. Murakhovs'ka, W. Liu, and C. Xiong, "Quiz Design Task: Helping Teachers Create Quizzes with Automated Question Generation," in *NAACL-HLT*, 2022.

[12] A. B. Abacha, W. Yim, G. Michalopoulos, and T. Lin, "An Investigation of Evaluation Metrics for Automated Medical Note Generation," *arXiv preprint arXiv:2305.17364*, 2023.

[13] M. R. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[14] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," *Frontiers in Sociology*, vol. 7, 2022.

[15] J. Araki, D. Rajagopal, S. Sankaranarayanan, S. Holm, Y. Yamakawa, and T. Mitamura, "Generating Questions and Multiple-Choice Answers using Semantic Analysis of Texts," in *International Conference on Computational Linguistics*, 2016.

[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.

[17] M. Koppam, R. Mehta, R. Kulkarni, S. Pai, and K. S. Srinivas, "Edvent—Speech to Presentation Automation," in *World Conference on Information Systems for Business Management*, Singapore: Springer Nature Singapore, Sep. 2023, pp. 305-325.

[18] N. Ballier, A. Méli, M. Amand, and J. Yunès, "Using Whisper LLM for Automatic Phonetic Diagnosis of L2 Speech, a Case Study with French Learners of English," in *International Conference on Natural Language and Speech Processing*, 2023.

[19] J. R. Terven and D. M. Esparza, "A Comprehensive Review of YOLO: From YOLOv1 to YOLOv8 and Beyond," *arXiv preprint arXiv:2304.00501*, 2023.

[20] L. Cui, Y. Wu, J. Liu, S. Yang, and Y. Zhang, "Template-Based Named Entity Recognition Using BART," *Findings*, 2021.

[21] A. Trask, P. Michalak, and J. Liu, "sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings," *arXiv preprint arXiv:1511.06388*, 2015.