Summarization of Educational Videos with Transformers Networks

Leandro Massetti Ribeiro Oliveira
TeleMídia@MA Lab / PPGCC
Universidade Federal do Maranhão
São Luís, Brazil
massetti.leo@gmail.com

Li Chang Shuen
TeleMídia@MA Lab / DCCMAPI
Universidade Federal do Maranhão
São Luís, Brazil
li.chang@ufma.br

Allan Kássio Beckman Soares da Cruz
TeleMídia@MA Lab / DCCMAPI
Universidade Federal do Maranhão
São Luís, Brazil
allankassio@gmail.com

Carlos de Salles Soares Neto
TeleMídia@MA Lab / DCCMAPI / PPGCC
Universidade Federal do Maranhão
São Luís, Brazil
carlos.salles@ufma.br

ABSTRACT

This paper presents an approach to summarize educational videos using Deep Learning Transformers models. The approach focuses on educational content by summarizing captions and using the text results to summarize the videos. Tests were conducted using the EDUVSUM dataset, which improved upon the original paper's results, achieving an accuracy of 26.53% in a multi-class problem, with a mean absolute error of 1.49 per video frame and 1.45 per video segment. Transformer techniques for automatic text summarization have proven effective in creating multimedia learning objects. The results suggest that these techniques can generate more efficient and high-quality digital educational resources, reducing the time and effort required for their creation.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → **Learning management systems**; **E-learning**.

KEYWORDS

Machine learning, transformers, e-learning, video summarization.

ACM Reference Format:
Leandro Massetti Ribeiro Oliveira, Li Chang Shuen, Allan Kássio Beckman Soares da Cruz, and Carlos de Salles Soares Neto. 2023. Summarization of Educational Videos with Transformers Networks. In *Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia '23), October 23–27, 2023, Ribeirão Preto, Brazil.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3617023.3617042

1 INTRODUCTION

With the increase of video popularity, the variety of lengths, and the reduction of users' available time, a pattern of behavior has been observed: People tend not to watch a video in its entirety but to move on to the moment that interests them. This behavior called "skimming through" [1], shows how convenient it is to have ways to summarize or annotate the meaning of sections in an educational video to save time for the user who wants to find a key moment in that material. Video summarization can be defined as converting a long video into a shorter one that contains the essential segments and allows the viewer to consume that content in less time [2]. The selection of segments for summarization can be made by noting the importance of each segment concerning the content being covered in numerical values. The summary process can be complex, with different approaches that can use visual features [3], exhibit progression, and viewing time per segment [4], or features such as audio, video, and speech transcription [5].

Thanks to technological advances and easy access to web platforms, educational videos are becoming more popular and accessible. They come in various formats, content, run times, and presentation types, such as lesson recordings, slide presentations, animations, and more. One element that overlaps in this context is the speaker's speech. This type of video is usually formatted as a recording of a lesson or an explanation of a particular topic. Thus, the main element of these materials is the speaker's speech, which can be transcribed into subtitles.

In the decoupage process, subtitles are labeled with the time of their appearance and duration to highlight the existing segments in an instructional video. Thus, in the literature, tools for efficient summarization of texts using Deep Learning through Transformers [6] Networks that enable learning objects to be created by summarizing instructional texts while retaining the most relevant content are identified in the literature.

In this way, it is possible to correlate the summary of the text transcription with the instructional video segments to annotate and categorize their contextual meaning, as shown in works such as those by [7]. Alternatively, some works summarize the video based on the transcription of the text using the TF-IDF technique (Term Frequency - Inverse Document Frequency) [8].

In this paper, we present an automation proposal for summarizing educational videos based on a study of the feasibility of using text summarizers to summarize videos through captions. The article is organized as follows: In addition to the introduction and conclusions, we report the method, the experiment configurations

Publication rights licensed to ACM. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of a national government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.
WebMedia '23, October 23–27, 2023, Ribeirão Preto, Brazil

ACM ISBN 979-8-4007-0908-1/23/10…$15.00
https://doi.org/10.1145/3617023.3617042

with the training description using the EDUVSUM dataset, and the discussion of the results.

2 RELATED WORKS

Several studies are discussed in the related work section, each offering valuable perspectives in instructional video summarization. These works reveal that similar and complementary approaches are being explored to optimize content selection in enhanced instructional videos. While some studies focus on subjectivity analysis and selection criteria, others explore multimodal fusion techniques, the fusion of audio resources and external knowledge, and the use of language to guide the selection of relevant segments. In addition, image recognition and temporal information exploration strategies also prove relevant to improving instructional video summarization. This work demonstrates a variety of approaches aimed at making video content more accessible and effective for viewers and highlights the continued development and applicability of summarization methods in educational contexts.

The paper [9] presents a method for summarizing videos based on classifying the subjectivity of video transcripts. It is an approach that shares similarities with the present work, as both focus on approaches to summarize videos based on information extracted from audio transcripts. While the present work focuses on calculating the similarity between video segment captions and transcript summaries to determine the importance of segments, the aforementioned article uses subjectivity classification as a relevant criterion for selecting content in video summaries.

In addition, the paper [10] addresses the importance of automatic video summarization to cope with a large amount of video data, especially in areas such as news and education. It discusses content selection criteria as critical factors in creating summaries of multiple videos. It highlights that current approaches focus on redundancy and diversity but may not adequately capture content semantics. The research proposes using exclusion criteria based on human strategies to improve summaries of multiple videos. The focus is detecting and excluding subjective segments in news videos to produce more compact and informative summaries. Results show that the approach can create summaries that keep content relevant, are shorter, and take into account visual elements, motion, and diversity.

Theoretically and theoretically, [11] addresses the potential of Artificial Intelligence techniques, particularly Deep Learning methods, to improve video collaboration services, including video conferencing. The article highlights applying facial analysis, sentiment analysis, and video summarization techniques to improve the video conferencing experience. These techniques are also relevant in the context of instructional video summarization. Both studies recognize the importance of summarizing video content to facilitate comprehension and navigation. While the aforementioned work focuses on improving the video conferencing experience in general, this work focuses specifically on improving the comprehension of instructional video through summary methods. Both studies aim to make video content more accessible and effective for users, even though they address different application contexts.

When finding relevant content in video lectures to help viewers find interesting points, [12] proposes an approach that merges low-level and high-level audio resources, enriched with knowledge from open databases, to automatically segment topics into video classes. The goal is to use this structure to improve the accessibility and usability of video lecture content, allowing viewers to quickly identify points of interest without having to watch the entire lecture. Although the focus is on topic segmentation, merging audio resources and external knowledge has conceptual similarities with searching for important segments in educational videos, which also aims to summarize the content for a better user experience. Therefore, both works aim to optimize the navigation and understanding of instructional videos, each with specific emphases in their approaches.

Using a speech-driven approach to video summarization, [13] uses a multimodal transformer to create general and query-oriented summaries. The kinship with the work on summarizing instructional videos lies in the similarity of the goal of creating relevant summaries from long videos. Both studies aim to extract the most relevant and informative segments of videos so that users can understand the content more quickly and efficiently. In the context of instructional videos, this approach can provide insight into how language can be used to guide the selection of relevant video segments to create more targeted summaries of the topics covered in class. In addition, considering video subtitles as input resources and using pre-trained networks for image encoding can suggest ways to better integrate visual and linguistic information in summaries of instructional videos.

An algorithm that focuses on image recognition using deep learning techniques and convolutional neural networks (CNNs) to improve the accuracy and efficiency of image recognition is proposed by [14]. Although the article's main focus is in the area of computer vision and image recognition, combining CNNs with attention mechanisms to highlight relevant features in images may have relevance to the field of educational video summarization. Identifying highly visually relevant segments in summarizing educational videos is crucial for creating effective summaries. The technique proposed in the paper, which emphasizes extracting relevant features, could be applied to identify key moments in educational videos containing essential information. This could help improve the quality and effectiveness of educational video summaries, allowing viewers to quickly access the most important points of the content.

Another connection to instructional video summarization is in the work of [15], which focuses on merging multimodal information to create video summaries. While the work focuses on summarizing instructional videos based on caption transcriptions, the work proposes a model called Time-Aware Multimodal Transformer (TAMT) for summarizing multimodal videos. TAMT incorporates temporal information in videos to combine different modalities and create summaries that allow for a more comprehensive examination of temporal information in videos. Although their work focuses on multimodal video summarization, the TAMT model addresses the issue of information fusion in a similar way to their work, but with a broader focus on different modalities, including visual and audio information. The short-term, task-based attentional approach and the inclusion of date and time information may also be relevant to improving instructional video summarization by providing a better understanding of the temporal structure of videos and identifying important segments.

The comprehensive analysis of related work highlights the diversity of approaches and strategies used to address video summarization challenges, including those in education. Examining these studies makes it possible to identify commonalities and unique contributions that enrich the field. The diversity of approaches, ranging from subjectivity analysis to the exploration of temporal and multimodal information, reflects the complexity and importance of developing techniques to help viewers understand and effectively retain content in educational videos. As the field of video summarization continues to advance, the insights gained from these studies may be valuable in developing better approaches adapted to multimodal learning contexts and the growing demand for accessible and effective educational content.

3 EXPERIMENT METHOD AND SETUP

The proposed method to achieve the desired results is based on the correlation between the original subtitle of the video and the summarized text generated by the compositor. However, it is important to point out some limiting factors that were identified during the development of the method. First, the videos must be captioned, as we use these captions to transcribe and correlate with the generated summary.

In addition, it is important that the videos have a pedagogical nature and a clear focus on the interlocutor's speech so that there is transcribable content. If the video does not clearly focus on the interlocutor's speech, the amount of information transcribed may be insufficient.

Another limiting factor must be considered is the presence of moments with little or no language in the material. In these moments, the lack of available information makes it difficult to evaluate and understand the context of the video. It is important to be aware of these limiting factors when applying the proposed method to obtain more accurate results that correspond to the characteristics of the educational videos analyzed.

The proposed summary method includes a series of steps designed to ensure the efficiency and accuracy of the summary process. With them, it is possible to extract the most relevant information from the videos and create a concise and informative synthesis of the presented learning content. The sequence of structured steps consists of:

(1) Receiving the original subtitle (in English) of the video to be summarized with its *timestamp* annotated.
(2) The content of the subtitle is summarized in a large text which goes through the cleanup process to remove special characters.
(3) The cleaned text (without special characters) goes through a structuring phase: the material is divided into small sections of a maximum of 3500 characters. This procedure prevents the compactor from exceeding its capacity to create models.
(4) Each subdivision goes through the summarization process using the model *transformers: facebook/bart-large-cnn*, which has proven effective in summarizing texts with educational content [6].
(5) Each heading is correlated with the corresponding summary partition and text similarity techniques are used to determine a similarity score with values from 0-1.
(6) Segments without headings are noted with a minimum value (0). In this way, each segment with or without headings is assigned a corresponding importance value.

In step 1 of the proposed method, the original subtitle of the video in English is obtained along with the timestamps. The subtitle provides the textual transcription of the content spoken in the video, while the timestamps provide the exact time at which each part of the subtitle occurs. This combination of subtitles and timestamps is essential to match the text excerpts and the original video. This facilitates correlation and the creation of summaries that reflect the key points of the presented learning.

In step 2, the content of the captions is combined into a single comprehensive text. This text undergoes a cleanup phase in which unwanted special characters are removed. The purpose of the cleanup is to ensure that the text is free of irrelevant elements, such as unnecessary punctuation, symbols, or special characters. Removing these elements makes the resulting text more readable and suitable for subsequent processing, allowing for better analysis and extraction of essential information during the instructional video summarization process.

After cleaning the text, we proceed to step 3, in which the cleaned text is divided into smaller segments, each with a maximum of 3500 characters. This structuring prevents the Summarizer from exceeding its capacity to create models, thus ensuring efficient and high-quality summarization. By dividing the material into smaller segments, the Summarizer can process each segment more accurately, capturing the essential information and avoiding overload that could affect the summarization process.

Steps 4 and 5, shown sequentially in Figure 1, play a crucial role in determining the meaning of the segments. In step 4, the method uses the model "transformers: facebook/beard-large-cnn" to summarize the obtained texts. The goal is to produce summaries that focus on the transcribed speech's most important points and capture the class's most important aspects.

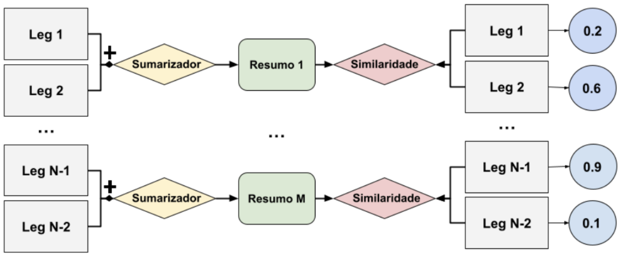

Figure 1: Stages of the proposal for making summaries and extracting meaning from the similarity of texts.

In the fifth step, a text similarity algorithm is used to score each segment and assign a score from 0 to 1 based on its similarity to the corresponding subtitles. This indication of importance helps identify the most important parts of the video.

Finally, in step 6, the video segments that do not have captions are noted. These segments are marked with a minimum importance value represented by the number 0. In this way, segments with and without subtitles are assigned an importance value. The reason for this annotation is to ensure that all segments of the video are considered in the importance rating.

Even if the segments without subtitles do not contain transcribed information, it is important to include them to avoid gaps in the overall context of the video. In this context, the problem of gaps refers to the possibility of a break in the cohesion and continuity of the video content due to the absence of subtitles in certain segments, which affects the comprehension and effectiveness of the summary. Imagine a video of a school class in which the teacher explains an important concept, but the transcription of that portion is unavailable in the captions. If this information gap is not adequately filled, the video summary would ignore the important, unsubtitled portion, resulting in a significant loss of content and context. This may reduce the effectiveness of the summary, as viewers would not have access to important information. Therefore, including the unsubtitled segments in the similarity analysis between the video and summary subtitles is critical to ensure that the importance of these segments is adequately considered when creating the summary to avoid gaps and maintain the integrity of the original content.

4 SUMMARIZATION TRAINING FROM THE EDUVSUM DATASET

The EDUVSUM (Educational Video Summarization) [5] dataset – a set of training videos for summary learning - was selected to conduct the experiments. The dataset contains videos from three popular e-learning platforms: Edx, YouTube, and TIB AV -portal, covering various computer science topics. These are 98 videos in English with the appropriate subtitles, each annotated in key segments (of 5 seconds). Each segment of the videos is scored on a scale of 1 to 10, with higher scores indicating greater importance of that segment in terms of information related to the video topic [5].

The EDUVSUM dataset used in this work consists of a training set with 83 videos and a test set with 15 videos. These datasets are used to train and test the proposed summary methods. To generate summaries of the subtitle texts, an approach that uses a pre-trained network was chosen. With this approach, it is possible to generate a summary for each subtitle time and a similarity value is assigned to each subtitle based on the comparison with the obtained summary. However, to allow comparison with the importance values noted by experts, it is necessary to adjust the similarity value output to the values defined by the authors of the dataset. This adjustment is essential to ensure a consistent and comparable evaluation of the results obtained during the summary process.

The method chosen in this work to compute text similarity was SentenceTransformers[1], a framework that provides access to pre-trained networks derived from BERT that can compute sentence similarity via cosine similarity [16]. The advantage of this text similarity algorithm is the ability of the framework to retrain the networks for the particular task context. It is possible to adapt the network output to the meaning annotations marked in the EDUVSUM dataset available in the training set.

Four models (pre-trained networks) were selected for the similarity task, with the importance of the segment annotated according to the expert's comment:

- **Model 1 - cross-encoder/stsb-roberta-base**[2]: trained on the STS benchmark set[3] to predict the similarity of two texts in values from 0 to 1, using the RoBERTa model, derived from BERT.
- **Model 2 - sentence-transformers/msmarco-roberta-base-v3**[4]: similar to Model 1, but trained on the MS MARCO dataset, composed of research documents and real questions from the Bing search engine[5] [17].
- **Model 3 - sentence-transformers/msmarco-bert-base-dot-v5**[6]: BERT model also trained with the MS MARCO dataset.
- **Model 4 - sentence-transformers/msmarco-distilbert-base-v4**[7]: DistilBERT model, derived from BERT , also trained with the MS MARCO dataset.

The 7591 pairs of similarity texts were extracted from the EDUVSUM training set, with scores ranging from 1 to 10. The scores were normalized to values of 0-1 to match the networks' output. The set was split into 6591 text pairs for training and 1000 pairs for validation. The networks were trained for 64 epochs (enough to have no further performance improvements) using the notes for each text pair from the EDUVSUM dataset.

Next, two different epochs were selected for each model, considering the best accuracy results from the validation set and the balanced accuracy. Balanced accuracy is important to avoid overestimated results in unbalanced datasets, as it aims to balance precision and sensitivity [18].

The balanced accuracy metric plays a fundamental role in evaluating the performance of similarity models used in research. Traditional accuracy often underestimates the true effectiveness of the model when dealing with unbalanced data sets, such as the importance ratings of instructional video segments. Balanced accuracy, on the other hand, provides a more comprehensive assessment that considers both the model's ability to correctly identify positive and negative cases. In this way, a balance is struck between the precision and sensitivity of the model, and biases resulting from the uneven distribution of classes of interest are mitigated.

Balanced accuracy, expressed as the harmonic mean of true positive (TP) and true negative (TN) rates, stands out as an objective and reliable measure for evaluating the performance of models in situations where data are disproportionately distributed [19]. In this study, its application enabled a more equitable evaluation of summarization techniques, especially for notes of varying importance. The selection of models based on balanced accuracy provided an accurate overview of these models' generalization capacity and helped mitigate the biases associated with data imbalance. Thus, balanced accuracy emerges as an essential metric for evaluating the effectiveness of summary approaches, leading to more robust results and a more comprehensive understanding of the performance

[1] https://www.sbert.net/

[2] https://huggingface.co/cross-encoder/stsb-roberta-base

[3] http: //ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark

[4] https://huggingface.co/sentence-transformers/msmarco-roberta-base-v3

[5] https://www.bing.com/

[6] https://huggingface.co/sentence-transformers/msmarco-bert-base-dot-v5

[7] https://huggingface.co/sentence-transformers/msmarco-distilbert-base-v4

of these models. Equation 1 describes the calculation of balanced accuracy [20], which provides an objective and reliable measure for evaluating the performance of the models in question.

$$balancedaccuracy = \frac{1}{2}\frac{TP}{TP+FN} + \frac{TN}{TN+FP} \quad (1)$$

In the context of the experiment conducted, the determination of balanced accuracy was based on the analysis of four essential components: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Each component plays a crucial role in evaluating the performance of similarity models compared to the importance ratings of learning video segments.

True positive (TP) refers to cases where the model correctly predicted the importance of segments, and these segments were actually rated as important by the experts who assigned the ratings. True negative (TN), on the other hand, refers to cases in which the model correctly predicted the importance of segments, and these segments were not considered important by the experts who gave the ratings.

False positive (FP) cases occur when the model incorrectly predicts that a segment is important, but the experts do not rate that segment as important. False negative (FN) cases occur when the model does not recognize the importance of a segment that was classified as important by the experts.

In the balanced accuracy analysis, combining these four components provides a more comprehensive and accurate assessment of model performance by considering the ability to correctly identify majority and minority classes. Balanced accuracy therefore provides a balanced and unbiased view of the effectiveness of similarity models, making it a critical metric in evaluating the summarization process in the context of instructional video segments.

After identifying the best networks for each architecture, they were used in the test set to predict the similarity of each caption to the summarized text. Then, the scores for each video frame are discretized from 1-10. The score per segment is calculated from the average of the frame scores. This approach enabled the creation of more meaningful summaries, facilitating a coherent evaluation of the performance of the summary models.

5 RESULTS

The first results relate to predicting the similarity of the models for the summarized texts and the subtitles in the test set. These can be seen in Table 1, which shows the training time in minutes, the accuracy (Acc), the balanced accuracy (Acc balance), and the mean absolute error (EAM) of the predicted results with the original results. For each model, two networks with the highest performance in accuracy (Acc) and balanced accuracy (b_acc) were selected by type.

According to the data described in Table 1, the model that managed to best adapt to the notes given by the dataset's author was Model 2 (msmarco-roberta-base-v3), which achieved an accuracy of 26.54%. Then, Model 1 had an accuracy of 24.49%. The performances are also similar when compared to EAM. Model 3 had slightly lower performance in terms of accuracy and EAM compared to the roBERTa models. This was expected, as the BERT model is state of the art and the roBERTa model, derived from BERT, has around 2-20% higher performance than BERT[8]. There is also a slight difference in models 1 and 2 regarding the dataset from which they were trained: the msmarco dataset is asymmetric; that is, it was trained to find similarity in texts of different sizes, which is similar to the cases trained in this experiment, presenting a have an advantage in accuracy compared to stsb.

Next, in Table 2, we present the EDUVSUM test set summarization process result. Following the same methodology, we have accuracy, average error per frame, and average error per 5-second segment. As the networks had sigmoid-shaped outputs, obtaining Top-2 and Top-3 accuracy results was impossible. Model 5 presents the two best results in the author's original experiment on the EDUVSUM set.

Initially, it is noted that all trained models were able to achieve a performance similar to that of the network trained by the author (VGG-16), having even been surpassed by Model 3 with 26.53% accuracy in the test set and by Model 2 in terms of EAM per frame and segment. It is worth mentioning that the method of the original article uses visual (video image), textual (subtitles), and audio (sound) [5] characteristics. The results were satisfactory because the approach described in this work uses only the summary of speech transcription from educational videos.

For a better visualization of the results, Figure 2 shows the prediction of model 3 (which presented the best result in Table 2) and the original value noted in the EDUVSUM dataset for the 5-second segments in two videos. In the first graph, we have a case in which the prediction obtained a high balanced accuracy (35.2%), and in the second, we have a prediction with low balanced accuracy (16.6%).

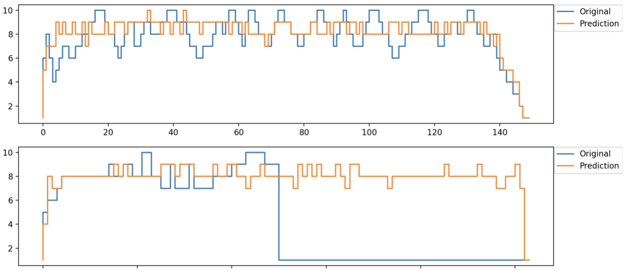

Figure 2: Model 3 predictions for two videos. Upper with high balanced accuracy (35.2%), lower with low balanced accuracy (16.6%).

Note in this case: the accuracy was relatively high, the prediction pattern behaved similarly to the original comment, and although it was wrong for much of the video, the results remained close, even at the beginning and end. Considering that the networks only use the similarity of the subtitle to the summary of the video transcription and no temporal character is added, the performance in predicting these elements is satisfactory.

On the other hand, in the prediction that showed low accuracy, we can notice a discrepancy in the last half of the video, where

[8] https://towardsdatascience.com/bert-roberta- distilbert-xlnet-which-one-to-use-3d5ab82ba5f8

Table 1: Performance of similarity networks for EDUVSUM grades

Model	Name	Training time	Type	Acc %	Acc balan. %	EAM
1	stsb-roberta-base	624 min	acc	26,49	**12,81**	**1,38**
			b_acc	24,74	11,42	1,41
2	msmarco-roberta-base-v3	354 min	acc	**26,54**	12,12	1,39
			b_acc	25,46	11,18	1,39
3	msmarco-bert-base-dot-v5	615 min	acc	25,00	11,50	1,44
			b_acc	24,74	12,42	1,48
4	msmarco-distilbert-base-v4	**178 min**	acc	25,98	11,45	1,42
			b_acc	23,87	11,62	1,51

Table 2: Performance of similarity networks for EDUVSUM grades

				EAM	
Model	Name	Type	Acc %	med_{fra}	med_{seq}
1	stsb-roberta-base	acc	24,83	1,52	1,47
		b_acc	22,54	1,55	1,51
2	msmarco-roberta-base-v3	acc	25,08	1,52	1,47
		b_acc	25,02	**1,49**	**1,45**
3	msmarco-bert-base-dot-v5	acc	**26,53**	1,52	1,48
		b_acc	25,25	1,59	1,55
4	msmarco-distilbert-base-v4	acc	25,20	1,55	1,49
		b_acc	23,85	1,60	1,52
5	VGG-16	acc	26,26	1,60	1,57
		b_acc	25,55	1,51	1,49

the expert rated the rest of the segments as 1, while the prediction remained between 7-9 on average, which is the main reason for the low accuracy in this video. This shows that the EDUVSUM set needs more raters to increase confidence in the segment ratings.

Given the need to transcribe speech, this approach is best suited for videos that contain a large amount of dialog, such as videos of classes and presentations. An alternative to improve results would be to include other video features (e.g., image and sound) in the methodology for summarizing the video presenter's speech transcription.

For a deeper analysis of the results, it is necessary to enrich the EDUVSUM dataset with the ratings of other domain experts to increase the reliability of the annotations. It is also necessary to reduce the rating scale to values from 1-5, thus reducing the variety of options for the annotator. In addition, entering the option for an average value (3) is desirable.

Based on the intrinsic nature of human communication and the understanding that the transcription of speech into subtitles is a textual synthesis that embodies the essential information of oral discourse, the underlying hypothesis is that the degree of similarity between the original subtitle of the video and the summarized subtitle can be a reliable indicator of the meaning of the video segment. The assumption is based on the premise that the summary captures the most important aspects of the instructional material by including the most relevant and representative excerpts of the discursive content. The similarity measure between these subtitles could therefore reflect the amount of information conveyed and, thus, the relevance of the corresponding segment in the video. A rigorous evaluation of this hypothesis could provide valuable insight into the ability of textual summarization methods to appropriately contextualize and categorize video segments in terms of their essential content, representing a significant advance in automating the importance assessment process in instructional videos.

6 CONCLUSION

This work aimed to verify the feasibility of using networks based on Deep Learning Transformers models to generate textual objects from educational videos, using automatic text summarization as the main tool. The results point to both the need to increase the number of expert evaluations in the EDUVSUM dataset to obtain greater reliability of the grades and to reduce the rating scale to values with less granularity, such as 1-5.

In the experiment described and analyzed here, summarization was done by classifying the importance of segments of an educational video. The result demonstrated the feasibility of using text summarizers to classify segments of this type of artifact. The performance was similar to the best multimodal model performed by [5], surpassed both by the BERT model pre-trained on the MS-Marco dataset with 26.53% accuracy and by the RoBERTa model pre-trained on the same dataset in the regression task, in which the average absolute error of the notes was 1.49 per frame and 1.45 per 5-second segment. Considering that only the textual component of the videos was used, without image, audio or temporal information, the results were satisfactory.

REFERENCES

[1] Potapov, D., Douze, M., Harchaoui, Z., & Schmid, C. (2014). Category-specific video summarization. In Springer (Ed.), European Conference on Computer Vision (pp. 540-555). [S.l.].
[2] Ghauri, J. A., Hakimov, S., & Ewerth, R. (2021). Supervised video summarization via multiple feature sets with parallel attention. In IEEE (Ed.), 2021 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1–6s). [S.l.]: IEEE.
[3] Song, Y., Vallmitjana, J., Stent, A., & Jaimes, A. (2015). Tvsum: Summarizing web videos using titles. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5179-5187).
[4] Mubarak, A. A., Cao, H., & Ahmed, S. A. (2021). Predictive learning analytics using deep learning model in MOOCs' courses videos. Education and Information Technologies, 26(1), 371-392.
[5] Ghauri, J. A., Hakimov, S., & Ewerth, R. (2020). Classification of important segments in educational videos using multimodal features. arXiv preprint arXiv:2010.13626.
[6] Oliveira, L. M. R., Busson, A. J. G., Salles, S. N. Carlos de, Santos, G. N. dos, & Colcher, S. (2021). Automatic generation of learning objects using text summarizer based on deep learning models. In SBC (Eds.), Anais do XXXII Simpósio Brasileiro de Informática na Educação (pp. 728-736). [S.l.].
[7] Alrumiah, S. S., & Al-Shargabi, A. A. (2022). Educational videos subtitles' summarization using latent dirichlet allocation and length enhancement. CMC-Computers Materials & Continua, 70(3), 6205–6221.
[8] Abhilash, R. K., Anurag, C., Avinash, V., & Uma, D. (2021). Lecture video summarization using subtitles. In EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing (pp. 83-92). Springer.
[9] Moraes, L., Marcacini, R. M., & Goularte, R. (2022, November). Video summarization using text subjectivity classification. In Proceedings of the Brazilian Symposium on Multimedia and the Web (pp. 133-141).
[10] de Souza Barbieri, T. T., & Goularte, R. (2020, November). Investigating Subjectivity Criterion for Multi-video Summarization. In Proceedings of the Brazilian Symposium on Multimedia and the Web (pp. 137-144).
[11] Mendes, P. R. C., Vieira, E. S., de Freitas, P. V. A., Busson, A. J. G., Guedes, Á. L. V., Neto, C. D. S. S., & Colcher, S. (2020, November). Shaping the Video Conferences of Tomorrow With AI. In Anais Estendidos do XXVI Simpósio Brasileiro de Sistemas Multimídia e Web (pp. 165-168). SBC.
[12] Soares, E. R., & Barrére, E. (2018, October). A framework for automatic topic segmentation in video lectures. In Anais Estendidos do XXIV Simpósio Brasileiro de Sistemas Multimídia e Web (pp. 31-36). SBC.
[13] Narasimhan, M., Rohrbach, A., & Darrell, T. (2021). Clip-it! language-guided video summarization. Advances in Neural Information Processing Systems, 34, 13988-14000.
[14] Huang, J. H., Murn, L., Mrak, M., & Worring, M. (2021, August). Gpt2mvs: Generative pre-trained transformer-2 for multi-modal video summarization. In Proceedings of the 2021 International Conference on Multimedia Retrieval (pp. 580-589).
[15] Shang, X., Yuan, Z., Wang, A., & Wang, C. (2021, October). Multimodal video summarization via time-aware transformers. In Proceedings of the 29th ACM International Conference on Multimedia (pp. 1756-1765).
[16] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. arXiv preprint arXiv:1908.10084.
[17] Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., & Deng, L. (2016). Ms Marco: A human generated machine reading comprehension dataset. In CoCo@ NIPs. [S.l.: s.n.].
[18] Mosley, L. (2013). A balanced approach to the multi-class imbalance problem (Doctoral dissertation). Iowa State University of Science and Technology, USA.
[19] de Freitas, P. V., Santos, G. N. D., Busson, A. J., Guedes, Á. L., & Colcher, S. (2019, October). A baseline for NSFW video detection in e-learning environments. In Proceedings of the 25th Brazillian Symposium on Multimedia and the Web (pp. 357-360).
[20] Balraj, B. (2021). Multilabel Active Learning for User Context Recognition In-the-Wild. North Carolina State University.