UrduMASD: a Multimodal Abstractive Summarization Dataset for Urdu

Ali Faheem, Faizad Ullah, Muhammad Sohaib Ayub, Asim Karim
Department of Computer Science, Lahore University of Management Sciences (LUMS),
Lahore, Pakistan 54792
{ali.faheem, faizad.ullah, sohaib.ayub, akarim}@lums.edu.pk

Abstract

In this era of multimedia dominance, the surge of multimodal content on social media has transformed our methods of communication and information exchange. With the widespread use of multimedia content, the ability to effectively summarize this multimodal content is crucial for enhancing consumption, searchability, and retrieval. The scarcity of such training datasets has been a barrier to research in this area, especially for low-resource languages like Urdu. To address this gap, this paper introduces "UrduMASD", a video-based Urdu multimodal abstractive text summarization dataset. The dataset contains $15,374$ collections of videos, audio, titles, transcripts, and corresponding text summaries. To ensure the quality of the dataset, intrinsic evaluation metrics such as Abstractivity, Compression, Redundancy, and Semantic coherence have been employed. It was observed that our dataset surpasses existing datasets on numerous key quality metrics. Additionally, we present baseline results achieved using both text-based and state-of-the-art multimodal summarization models. On adding visual information, an improvement of 2.6% was observed in the ROUGE scores, highlighting the efficacy of utilizing multimodal inputs for summarization. To the best of our knowledge, this is the first dataset in Urdu that provides video-based multimodal data for abstractive text summarization, making it a valuable resource for advancing research in this field.

Keywords: Urdu Summarization, Multimodal Abstractive Summarization, Multimodal Dataset

1. Introduction

The advent of multimedia content has transformed the way we consume and share information in the modern era. Videos, images, and audio recordings have emerged as popular mediums through which individuals convey their ideas across various online platforms. The rise of multimedia content has enabled people to create more engaging and interactive content, utilizing different modalities, making it easier to communicate complex ideas and thoughts. Technological advancements such as high-speed internet, smart mobile devices, and numerous online content-sharing platforms have been instrumental in driving this trend. These enable users to access, create, and consume multimedia content anytime, anywhere, thereby empowering users to express their emotions and serving as a valuable source of information.

Video-sharing platforms like TikTok, YouTube reels, and Instagram have gained immense popularity in recent years, hosting billions of active users who upload substantial amounts of multimedia content daily, often amounting to hundreds of hours of videos per day (McLachlan, 2022). These platforms often allow users to add text on videos to provide additional context, making it easier for viewers to understand the content and message being conveyed.

However, this transition to multimedia content also presents challenges, particularly in information extraction. The sheer volume and multi-modal nature of the data make traditional text-based information extraction techniques inadequate (Rupapara et al., 2020). Videos may contain music, spoken content, and visual elements that are not adequately represented in accompanying text descriptions. This can hinder users from discovering the videos relevant to their interests, posing a challenge for search engines (Hamroun et al., 2022). Therefore the development of an automated summarization system for such multimodal content has become a pressing need as it can significantly enhance the search and retrieval of relevant content for users. These systems can extract essential information, create concise text summaries of content, and enhance the efficiency of content search and consumption, especially in the context of multimedia marketing and businesses.

Multimodality refers to the integration of various modalities, such as text, images, videos, audio, and other forms of information. It involves the combined use of multiple sources of information, including both verbal and non-verbal cues, to provide a more comprehensive understanding of the subject matter. Human cognition naturally integrates various modalities, such as text, images, and videos, to comprehend and summarize complex content effectively. Also, various studies have shown that information extraction from multimodal content can provide rich information compared to text or other uni-modal content (Wang et al., 2022a,b). Audios and videos can provide non-verbal cues like the tone of the speaker, facial expressions, music, sound effects, and body language can provide additional information such

LREC-COLING 2024, pages 17245–17253
20-25 May, 2024.

as emotional context, emphasis on certain points, and the mood of the speaker. These cues play a pivotal role in conveying important details and aspects that might not be apparent from the text alone. While technology has enabled us to extract information from these modalities independently, there is a noticeable gap in effectively utilizing them together. This research gap primarily stems from the scarcity of multimodal training datasets, hindering progress in this field.

Automatic summarization holds a prominent position as one of the earliest and most important topics in the field of natural language processing (NLP). Automatic Summarization techniques are widely used to generate concise summaries of large-volume datasets. Such techniques can be broadly categorized into extractive and abstractive text summarization. Extractive summarization involves selecting the essential parts of the original content and combining them to create a concise summary. On the other hand, Abstractive summarization involves generating a summary, where the summarized output may contain information not part of the original input content. Abstractive summarization is a complex approach as it requires the system to understand the context and capture the main idea of the original content.

Multimodal Summarization (MMS) systems can also be classified on the modality of inputs and outputs. Researchers have recently shown a growing interest in MMS, acknowledging its significant potential across various applications. MMS integrates information from multiple modalities (text, video, audio, etc) to generate a summary. Furthermore, the generated summary can be presented in a multimodal format, such as a visual summary combining video and audio elements or a text summary with relevant images. The modalities can be categorized into two types: homogeneous and heterogeneous. Homogeneous modalities encompass similar data representations, such as video captions and transcripts, where the underlying information is comparable. In contrast, heterogeneous modalities involve different datatypes with distinct representations, such as images, text, and audio.

In recent years considerable research efforts have been dedicated to multimodal corpus creation and MMS for high-resource languages (Garg et al., 2022). However, due to the inherent complexity of the task and the scarcity of multimodal datasets, no significant advancements have been made for low-resource languages, such as Urdu. To bridge this gap, we have created a benchmark dataset for multimodal Urdu abstractive text summarization. These efforts seek to enhance the technology's alignment with human cognitive processes and its ability to handle a broader range of linguistic and visual content, thereby addressing the multifaceted nature of information consumption in the modern era.

We make the following contributions in this paper:

- We introduce a novel video dataset for Urdu multimodal abstractive summarization,
- We conduct a comprehensive intrinsic evaluation to assess dataset quality. Our dataset surpasses existing datasets on numerous key quality metrics.
- We perform comprehensive experiments to set a baseline for Urdu multimodal summarization, resulting in the performance improvement of 2.6% in ROUGE score when integrating visual information.

The rest of the paper is organized as follows: In Section 2, we survey recent research on multimodal abstractive summarization and available multimodal datasets. Section 3 introduces our proposed UrduMASD dataset. Section 4 outlines our experimental setup and the models used. The results and discussions are presented in Section 5.

2. Related Work

In the emerging field of multi-modal summarization (MMS), initial research efforts were focused on uni-modal data summarization, laying the groundwork for the subsequent integration of various data modalities into coherent summaries (Kintsch and Kozminsky, 1977). Early work in MMS utilized rule-based methods to preprocess individual modalities before their integration. The Video Skim system, for instance, stands as an early attempt at multimodal summarization, integrating images, text, and audio. It employs TF-IDF for textual filtration and uses comparative histogram analysis to distinguish scenes, enhanced by the incorporation of camera dynamics, object detection, and text recognition techniques, thus improving the accuracy of the summarization (Smith and Kanade, 1997). Furthermore, Shahraray and Gibbon (Shahraray and Gibbon, 1995) contributed to the field by creating pictorial video transcripts, selecting keyframes and adding textual captions, thereby generating compact multimedia summaries.

The advancements in neural networks have transformed the multimodal summarization paradigm, shifting it from rule-based techniques to supervised techniques, empowering models to decipher complex relationships among various input modalities. (Li et al., 2019) proposed an extractive MMS technique for news summarization. The study assessed transcriptions and their associated audio signals to rank them for summarization.

Subsequently, a neural network was employed to acquire combined text-image representations, guaranteeing that significant visual elements are integrated into the summary through text-image alignment. Recent advances in pre-trained transformer models and transfer learning have revolutionized machine learning, particularly for low-resource languages. These languages can now benefit from pre-trained models even with limited datasets (Ullah et al., 2023). (Yu et al., 2021) used the multilingual text-to-text pre-trained models, mT5 (Xue et al., 2020) and mBart'(Liu et al., 2020), and injected visual information into the model through attention-based add-on layers, achieving significant performance improvements. MultimodaL Article Summarization Kit (MLASK) (Krubiński and Pecina, 2023) introduced the MMS dataset for the Czech language and incorporated the cross-modal interaction module to combine visual and textual representations. The research also leveraged pretrained models for feature extraction in each modality. The findings demonstrated that additional pretraining of the text decoder substantially enhanced results, particularly for low-resource languages.

The development of effective MMS techniques largely depends on having access to diverse and comprehensive datasets that contain multiple modalities. Due to the flexible nature of MMS tasks with a large variety of different combinations of input and output modalities, researchers are exploring novel approaches to handle the challenges posed by heterogeneous modalities. In this context, we will focus on the Multimodal Abtractive Text Summarization (MAS) datasets with multiple modalities as input and abstractive text summaries as desired outputs.

The MAS research community has introduced a variety of datasets, with most of them featuring text and images as input modalities and a few incorporating text and video as inputs. (Liu et al., 2018) presented a corpus of 66000 samples, each comprising a sentence, an image, and a headline. Chen et al. (Chen and Zhuge, 2018) extended the standard DailyMail summarization corpora by extracting the images and captions from the source documents. This augmented dataset comprises a substantial $219,100$ samples. (Li et al., 2020a) curated a Chinese e-commerce dataset to provide a summary of Chinese e-commerce products. The dataset comprises 1.4 million instances from Home Appliances, clothing, cases, and bags categories, with 119, 86, and 33 product types, respectively. Instances consist of product information with images, titles, and descriptions, and high-quality product summaries. (Sanabria et al., 2018) compiled a dataset of instructional videos from YouTube for the MAS task. The dataset comprises $79,114$ instructional videos with English subtitles and human-written abstractive summaries, totaling $2,000$ hours. (Fu et al., 2020) introduced MM-AVS corpus, a novel dataset for multimodal summarization, combining articles, videos, and references. Data is collected from the Daily Mail and CNN, including titles, images, and captions. Following a similar approach, (Li et al., 2020b) created a multimodal summarization dataset sourced from Weibo. This dataset comprises articles with textual summaries and videos with cover pictures totaling $184,920$ samples. The majority of dataset creation efforts have been directed towards high-resource languages such as English and Chinese, with comparatively limited attention given to low-resource languages. (Liang et al., 2022) pioneered the development of the first multilingual multimodal dataset, MM-SUM, with a specific focus on low-resource languages. They expanded upon the XL-Sum dataset (Hasan et al., 2021), originally an abstractive text summarization dataset encompassing text article-summary pairs in 44 languages. The authors enhanced this dataset by incorporating images from the source pages, thereby creating a multimodal version. This extended dataset, encompassing various other low-resource languages, includes $40,672$ Urdu text and summary samples paired with $106,960$ associated images. The summary of the MAS datasets is provided in Table 1.

Although multi-modal summarization has evolved from rule-based techniques to advanced transformer-based methods, very few datasets have been proposed for low-resource languages. Our proposed MAS dataset for Urdu addresses this gap and contributes to the advancement of MAS in Urdu.

3. UrduMASD

In this section, we develop UrduMASD: a novel dataset designed for multimodal abstractive text summarization[1]. To the best of our knowledge, there is no publicly available dataset in Urdu that combines video, audio, and text for the purpose of multimodal summarization. The following subsections discuss the details of our data collection guidelines and provide comprehensive statistics for the dataset.

3.1. Data Collection and Preprocessing

We collected videos and their corresponding descriptions from Urdu news channels on YouTube. We focused on the collection of videos that consistently maintained human-written video descriptions and covered a wide array of topics, including

[1] https://github.com/AliFaheem/UrduMASD

Dataset	Language	Input Modalities	Samples
MMSS (Li et al., 2018)	English	Text, Image	$66,000$
Extended DailyMail (Chen and Zhuge, 2018)	English	Text, Image	$219,100$
Chinese e-commerce (Li et al., 2020a)	Chinese	Text, Image	$1,400,000$
How2 (Sanabria et al., 2018)	English	Text, Video	$79,114$
MM-AVS (Fu et al., 2020)	English	Text, Image, Video	2173
VMSMO (Li et al., 2020b)	Chinese	Text, Video, Image	$184,920$
MLASK (Krubiński and Pecina, 2023)	Czech	Text, Video, Image	$41,243$
MM-Sum (Liang et al., 2022)	Multilingual	Text, Image	$40,672$

Table 1: Multimodal Abstractive Text Summarization (MAS) Datasets

sports, politics, social issues, news, and entertainment. Our choice was motivated by the diversity of content they offered.

Our data collection process included extracting videos, their titles, and descriptions. Additionally, we generated text transcriptions from the videos, treating them as textual documents, while the human-written descriptions served as summaries for each video, as illustrated in Figure 1.

We ensured that all the collected videos had a time duration of 10 minutes or less. Additionally, we excluded videos that lacked descriptions. Many descriptions included template-based links and reference URLs that were not relevant to our task. To make the data cleaner and relevant, we systematically removed such links and hashtags using regular expressions.

To obtain the text transcriptions, we employed the Automatic Speech Recognition (ASR) model for Urdu (*ihanif/whisper-medium-urdu*[2]). This choice was driven by the sheer volume of the videos in our dataset, making manual transcriptions expensive and time-consuming. This ASR model reported a word error rate of 26.9 for Urdu, indicating its suitability for our language of interest. However, many modern ASR systems have a limitation of *hallucination* where it occasionally repeats the same words or phrases in transcriptions. While this introduces redundancy, it does not miss the text. We addressed this in the preprocessing phase by systematically removing such repeated patterns from the transcribed text.

The video transcriptions also contained some traces of English, as in informal settings, English and Urdu are spoken interchangeably. Additionally, some words are common between Urdu and English, which occasionally results in mixed language transcriptions. We opted to retain these traces as they were but removed the instances where the majority of the text was in another language. We checked all the transcriptions using *langdetect*[3] python library.

3.2. Dataset Statistics

Our dataset consists of a set of $15,374$ composed of videos, titles, transcripts, and their corresponding text summaries. On average, the length of the transcriptions is 282.5 words, while the summaries average 32.9 words. We compared these statistics with benchmark English datasets and text-based Urdu datasets, as presented in Table 2. Our dataset provides notably extensive average summary length producing comprehensive summaries for Urdu.

The dataset has been partitioned into training, validation, and test sets with sizes of $11,684$, 615, and $3,075$, respectively. These partitions correspond to aggregate video lengths of 606 hours for training, 33 hours for validation, and 158 hours for testing. Each video in the dataset has a duration of 10 minutes or less, with the majority falling within the range of 100 to 200 seconds. Figure 2 illustrates the distribution of video duration.

3.3. Intrinsic Evaluation of Dataset

The training dataset serves as the foundation for constructing the robust machine-learning model. To assess the quality of the dataset, we employed the following four intrinsic metrics: Abstractivity, Compression, Redundancy, and Semantic Coherence.

Abstractivity quantifies the degree of abstraction in the summary by computing the ratio of the length of the longest common sequence (LCS) between the document (transcriptions) and the summary (description).

$$ABS = 1 - \frac{LCS(\text{document}, \text{summary})}{|\text{summary}|} \quad (1)$$

Compression measures the reduction in content at the word level, where $|s|$ represents the number of words in sequence s.

[2] https://huggingface.co/ihanif/whisper-medium-urdu

[3] https://pypi.org/project/langdetect/

Title

پشتون موسیقی، رباب کے ماند پڑتے سُر

Pashton Music, The declining sounds of Rubab

Transcription

پشتوں مسیقی کا شین شاہ کہلانے والے عالِ مسیقی ربرباب گٹار سے ملتا جلتا آلی مسیقی ہے لیکن دونوں کی آواز اور شکل میں واضح فرق پایا جاتا ہے۔رباب کی تین قسمیں ہوتی ہیں بڑے سائس کارباب بائیس چھوٹے اور بڑے تاروں پر مشتمل ہوتا ہے جبکہ دووں میں رباب کے ایک ماہرائے تھے اس کا رباب بجانا مجھے بہت اچھا لگا اور پھر مجھے بھی سیکھنے کا شوق پیدا ہوا میں دز ب جانی والی رباب کی یہ عوامی محفلیں اس ساس کی مقبولیت میں اہم کردار ادا کرتی رہی ہیں لیکن گزش تک کچھ عرصہ سے یہ پروگرامات نہ ہونے کے برابر ہے اور شاید اسی وجہ سے رکزی حیثیت حاصل ہے باس تاریخدانوں کا کہنا ہے کہ جنوبی ایشیا مینرباب کی ابتدا سب سے پہلی بار اور وہ بھی رفیق غزنی بھی جو انڈیا اور پاکستان کا مشہور جس نے اس ساز کو استعمال کیا مختلف تاروں کے ساز میں سے اس کو ہمارے پورے ریجن میں اس سازوں کا بلکہ فوق کا بادشصہ سے پاپ میزیگ کے بڑھتے ہوئے رجحان اور شرد پسندی کے باعث پشتون علاقوں میں رواشتونوں کی اپنی سقافت سے لا علمی کی وجہ سے ایسے سازینوں کی مقبولیت معدوم ہوتی مسیقی کے مطالعک رباب کو سر میں لانا ایک سانتثفک عمل ہے اس میں جو چھوٹے تار ہوتے ہیں اس کو بڑے تاروں سے جوڑنا پڑتا ہے اور شاید یہی وجہ ہے کہ یہ ایک مشکل فن بھی سمجھا جاتا ہے.

Known as the Shin Shah of Pashto music, he is an expert musician in both the rabab and guitar. However, there is a clear distinction in their sound and appearance. The rabab comes in three types: the large-sized Saissorbab, the 21-string Carbab, and the smaller one with both big and small strings. Inspired by the mastery of a rabab player, I developed a keen interest in learning to play the rabab. These public gatherings featuring the Dz Rabab have played a significant role in enhancing its popularity. However, for some time now, such programs have dwindled, possibly contributing to its diminishing presence. Historians claim that the origin of the rabab in South Asia is attributed to Rafeeq Ghazni, who used this instrument in India and Pakistan. Among various string instruments, it has gained immense recognition not just in our region, but has seen an increase in its popularity worldwide, primarily due to its unique sound and charm. In the Pashtun regions, it has a deep connection to their cultural heritage, and people admire it because of its unexplored musical qualities. Learning to play the rabab is an intricate process; it involves connecting small strings to larger ones, which may be why it's considered a challenging art

Summary (Description)

پاپ میوزک کے بڑھتے ہوئے رحجان اور شدت پسندی کے باعث پشتون علاقوں میں رباب جیسے روایتی آلات موسیقی کا استعمال کم ہو رہا ہے۔

Due to the increasing popularity of pop music and a preference for intensity, the use of traditional musical instruments like the rabab is diminishing in Pashtun regions.

Figure 1: An illustrative example of UrduMASD showcasing a compilation of instructional videos accompanied by titles, transcriptions, and concise summaries.

Dataset	Total Docs	Input Modality	Language	Avg Doc Length	Avg Sum Length
CNN	92,539	Text	En	760.50	45.70
DailyMail	219,506	Text	En	653.33	54.65
XSum	226,711	Text	En	431.07	23.26
How2	79,114	Text+audio+video	En	282.57	32.99
UrduMASD	15,374	Text+audio+video	Ur	363.14	50.92

Table 2: Comparison of UrduMASD with other benchmark abstractive summarization datasets

$$CMP = 1 - \frac{|\text{summary}|}{|\text{document}|} \quad (2)$$

Redundancy is determined by averaging the ROUGE-L score across all unique sentence pairs in the summary. Here x and y represent the sentences in the summary.

$$RED = \max_{(x,y) \in s_i \times s_i, x \neq y} \text{ROUGE-L}(x, y) \quad (3)$$

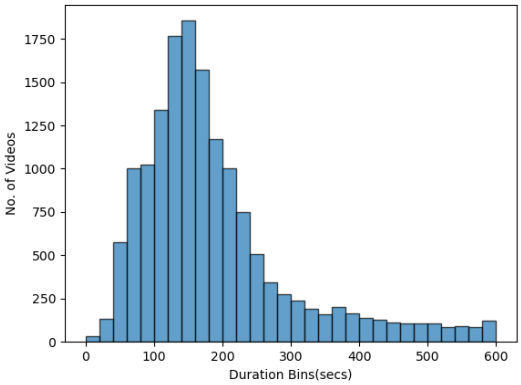

Figure 2: Distribution of video duration

semantic coherence measures the coherence in the semantic relation of consecutive sentences in the summary.

$$SC = \frac{\sum_{j=2}^{|q^j|} mBERT(q^j|q^{j-1})}{|\text{sentences}| - 1} \quad (4)$$

We used pre-trained multilingual BERT from HuggingFace (bert-base-multilingual-uncased[4]) to measure the probability of the next sentence q^j given the previous sentence q^{j-1} using all the consecutive sentence pairs in the multi-sentence summaries.

Table 3 provides the intrinsic evaluation of the UrduMASD, with a comparison to How2, the widely recognized English multimodal summarization dataset. How2 is a benchmark large-scale benchmark dataset that comprises both videos and descriptions. Notably, UrduMASD exhibits very high abstractivity and low redundancy scores, which are highly favorable attributes for an abstractive summarization dataset. The compression rate in UrduMASD is comparable to that of the How2 dataset. However, there was a notable decline in the semantic coherence scores. Further investigation revealed that this decrease is linked to limitations in the multilingual BERT model's comprehension of Urdu text. Testing the semantic coherence of well-written Urdu news articles produced a score of 0.75, highlighting the constraints of multilingual models for Urdu understanding.

4. Experimental Setup

To evaluate the quality of UrduMASD for the abstractive video summarization task and to present a robust baseline, we conducted thorough experiments across various settings. In this section, we present the baseline models and the experimental design.

[4] https://huggingface.co/bert-base-multilingual-cased

Dataset	ABS	CMP	RED	SC
How2	0.479	0.858	0.187	0.949
UrduMASD	0.958	0.732	0.111	0.527

Table 3: Intrinsic Evaluation of UrduMASD and How2 Datasets: ABS (Abstractivity), CMP (Compression), RED (Redundancy), and SC (Semantic Coherence)

4.1. Models

In this section, we present the models employed in our work. We selected these pre-trained models for their capacity to perform transfer learning and adapt to low-resource summarization tasks.

4.1.1. mT5

mT5, a multilingual extension of the T5 (Text-to-Text Transfer Transformer) model, is trained on the mC4 (Multilingual Common Crawl Categorization Corpus) dataset. The mC4 dataset is a vast collection of text data from web pages in 101 diverse languages, including Urdu. mT5 is known for its transfer learning capabilities, making it a suitable choice for generating video summaries, particularly in the case of low-resource languages like Urdu with limited training data.

4.1.2. MLASK

We used the architecture presented in MLASK (Krubiński and Pecina, 2023) to set the baseline for multimodal summarization. It employs a Transformer-based text encoder to convert input text into token embeddings. The model's weights are initialized with the pre-trained mT5 model, which is then fine-tuned for the specific task of text-to-text summarization. For processing video data, a 3D ResNeXt-101 model is employed to extract 2048-dimensional feature representations from non-overlapping sequences of 16 frames. These features are further processed using a Transformer encoder to capture long-term temporal dependencies. The model employs multi-head attention mechanisms for fusing visual and textual information. Through a combination of attention-based fusion and the use of a forget gate mechanism, the model effectively combines textual and visual inputs, enhancing the overall representation. The resulting representations guide the model in generating textual summaries.

Modality	Model	Input	ROUGE-1	ROUGE-2	ROUGE-L
Text	mT5-small	Transcripts	23.91	5.59	17.01
	mT5-base		22.12	4.80	15.96
Text	mT5-small	Transcripts + Image Captions	26.89	8.00	19.69
	mT5-base		23.27	6.10	17.03
Multimodal	MLASK	Transcripts+Video	23.86	6.34	17.40

Table 4: ROUGE-1, ROUGE-2, and ROUGE-L evaluation scores of mT5-small, mT5-base, and MLASK on different experimental settings

4.2. Experimental design

In our experimental setup for abstractive video summarization, we employed three different input configurations to assess the impact of visual information. The first configuration utilized only audio transcriptions as input, providing the initial results for the text-based features. We finetuned pretrained mT5-small and mT5-base for the summarization task.

In the second configuration, we enhanced the input by integrating visual data with transcriptions. We selected distinct frames from the video and employed 'blip-image-captioning-base'[5] model, a pretrained image captioning model that uses a multimodal mixture of Encoder-Decoder (MED), to generate English image captions that described the frame's content and actions. Since there is no equivalent model for Urdu, we translated these English captions into Urdu using the 'opus-mt-en-ur'[6] English to Urdu pretrained machine translation model. This homogeneous feature format allowed easy concatenation of image captions with the transcriptions, providing additional context to the model. We fine-tuned both mT5-small and mt5-base variants in this setting. In the third setting, we employed the MLASK model, which combines both the text transcripts and video features using multi-head attention to merge these elements, creating a comprehensive input for the summarization process.

These experiments were designed to assess how various input features influence the accuracy of the generated summary.

5. Results and Discussion

Table 4 presents the ROUGE-1, ROUGE-2, and ROUGE-L scores for various input modalities. When utilizing only audio transcriptions as input, the model achieved ROUGE scores of 23.91, 5.59, and 17.01 for ROUGE-1, ROUGE-2, and ROUGE-L, respectively. Similarly, under the same conditions, the use of the mt5-base model resulted in scores of 22.12, 4.80, and 15.96, respectively. These scores served as the baseline for assessing the impact of incorporating visual information on the outcomes.

The introduction of visual information by concatenating it with the transcriptions brought about a notable enhancement in ROUGE scores. Additionally, there were improvements in the ROUGE scores when employing multi-modal fusion through the MLASK model. However, these improvements were less pronounced when compared to using homogeneous text-based features. The superior performance of text-based features is largely due to the more robust training of text-based models compared to MLASK. For the MLASK model, the performance metrics achieved 23.86, 6.34, and 17.40 for ROUGE-1, ROUGE-2, and ROUGE-L, respectively.

These results highlight the impactful role of visual information in improving the summary quality, whether through image captions or vector-based video features. Furthermore, the current pre-trained models for Urdu generation perform more effectively when utilizing text input compared to complex feature fusion approaches.

6. Conclusion

We introduced the first dataset for Urdu abstractive text summarization for videos, paving the way for further research and advancements in this domain. Through a rigorous evaluation process, we substantiated the notion that the inclusion of visual information in the input significantly enhances the quality of the generated summaries. This work can be further extended by pretraining the mT5 model for the Urdu generation task as well as exploring more techniques for multimodal feature fusion.

[5] https://huggingface.co/Salesforce/blip-image-captioning-base

[6] https://huggingface.co/Helsinki-NLP/opus-mt-en-ur

Limitations

This study acknowledges several limitations. Firstly, the performance of out-of-the-box language models (mT5) for Urdu raises concerns, with the potential for improvement through dedicated pretraining for the language. Additionally, due to computational constraints, a limited variety of mT5 variants were explored. Furthermore, the MLASK model was trained for 10 epochs, offering room for enhanced results with extended training. Addressing these limitations in future research promises to unlock the full potential of multimodal abstractive text summarization, particularly in the Urdu language.

Dataset Availability

The dataset will be shared upon request, subject to signing the data sharing agreement, with the understanding that it will be used exclusively for research purposes.

7. Bibliographical References

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Jingqiang Chen and Hai Zhuge. 2018. Abstractive text-image summarization using multi-modal attentional hierarchical rnn. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4046–4056.

Xiyan Fu, Jun Wang, and Zhenglu Yang. 2020. Multi-modal summarization for video-containing documents. *arXiv preprint arXiv:2009.08018*.

Muskan Garg, Seema Wazarkar, Muskaan Singh, and Ondřej Bojar. 2022. Multimodality for NLP-centered applications: Resources, advances and frontiers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6837–6847.

Mohamed Hamroun, Karim Tamine, and Benoît Crespin. 2022. Multimodal video indexing (mvi): A new method based on machine learning and semi-automatic annotation on large video collections. *International Journal of Image and Graphics*, 22(02):2250022.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703.

Walter Kintsch and Ely Kozminsky. 1977. Summarizing stories after reading and listening. *Journal of educational psychology*, 69(5):491.

Mateusz Krubiński and Pavel Pecina. 2023. MLASK: Multimodal Summarization of Video-based News Articles. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 880–894.

Wendy G Lehnert. 1980. Narrative Text Summarization. In *AAAI*, pages 337–339.

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020a. Aspect-aware multimodal summarization for chinese e-commerce products. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8188–8195.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, Chengqing Zong, et al. 2018. Multimodal Sentence Summarization with Modality Attention and Image Filtering. In *IJCAI*, pages 4152–4158.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2019. Read, Watch, Listen, and Summarize: Multi-Modal Summarization for Asynchronous Text, Image, Audio and Video. *IEEE Transactions on Knowledge and Data Engineering*, 31:996–1009.

Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020b. VMSMO: Learning to Generate Multimodal Summary for Video-based News Articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9360–9369.

Yunlong Liang, Fandong Meng, Jinan Xu, Jiaan Wang, Yufeng Chen, and Jie Zhou. 2022. Summary-Oriented Vision Modeling for Multimodal Abstractive Summarization. *arXiv preprint arXiv:2212.07672*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256.

Stacey McLachlan. 2022. 23 YouTube Stats That Matter to Marketers in 2024. https://blog.hootsuite.com/youtube-stats-marketers/. [Accessed 26-03-2024].

Jielin Qiu, Jiacheng Zhu, William Han, Aditesh Kumar, Karthik Mittal, Claire Jin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Bo Li, et al. 2023. MultiSum: A Dataset for Multimodal Summarization and Thumbnail Generation of Videos. *arXiv preprint arXiv:2306.04216*.

Vaibhav Rupapara, Kaushika Reddy Thipparthy, Naresh Kumar Gunda, Manideep Narra, and Swapnil Gandhi. 2020. Improving video ranking on social video platforms. In *2020 7th International Conference on Smart Structures and Systems (ICSSS)*, pages 1–5. IEEE.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*.

Behzad Shahraray and David C. Gibbon. 1995. Automatic generation of pictorial transcripts of video programs. In *Multimedia Computing and Networking 1995*, volume 2417, pages 512 – 518. International Society for Optics and Photonics, SPIE.

Michael A Smith and Takeo Kanade. 1995. *Video skimming for quick browsing based on audio and image characterization*. School of Computer Science, Carnegie Mellon University Pittsburgh, PA, USA.

Michael A Smith and Takeo Kanade. 1997. Video skimming and characterization through the combination of image and language understanding techniques. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 775–781. IEEE.

Faizad Ullah, Ubaid Azam, Ali Faheem, Faisal Kamiran, and Asim Karim. 2023. Comparing prompt-based and standard fine-tuning for Urdu text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6747–6754, Singapore. Association for Computational Linguistics.

Josiah Wang, Josiel Figueiredo, and Lucia Specia. 2022a. MultiSubs: A Large-scale Multimodal and Multilingual Dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6776–6785.

Zhen Wang, X Shan, Xiangxie Zhang, and J Yang. 2022b. N24News: A New Dataset for Multimodal News Classification. In *2022 Language Resources and Evaluation Conference, LREC 2022*. European Language Resources Association (ELRA).

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision Guided Generative Pre-trained Language Models for Multimodal Abstractive Summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.