Journal of Instrumentation Technology & Innovations [JOITI]

ISSN: 2249-4731
Volume- 14
Issue-2
Year-2024

Research Article

Date of Receive- 28TH -June-2024

Date of Acceptance- 02nd-July-2024

Date of Publication- 16th-July-2024

DOCSNAP: Integrating NLP and Computer Vision for Comprehensive Document Summarization

Mohammed Hafeez M K[1*], Ayishathul Misriya K S[2], Fathima Haifa[3], Fathimath Zaziba[4], Khatheejathul Aifa[5]

[2,3,4,5]Students, Department of Computer Science and Engineering, P A College of Engineering, Mangalore, Karnataka, India

[1]Professor, Department of Computer Science and Engineering, P A College of Engineering, Mangalore, Karnataka, India

Author For Correspondence Email- hafeez_cse@pace.edu.in

Abstract

Because of the exponential growth of digital content, sophisticated tools are required for effective data interpretation and administration.. This project harnesses the capabilities of the Gemini AI model developed by Google DeepMind to address the challenges of PDF summarization and image captioning. Gemini AI integrates cutting-edge algorithms, including transformer architectures, to process textual and visual data seamlessly. The project's system architecture involves modules for text and image extraction, with a Frontend Interface and Backend Server for efficient processing. Results indicate its effectiveness in enhancing content understanding and retrieval.

Keywords- *PDF, image captioning, google deep mind, docsnap, gemini AI model.*

I. INTRODUCTION

The rapid expansion of digital content has created an overwhelming amount of information, necessitating advanced tools for efficient data management and interpretation. PDF summarization and image captioning are two critical tasks that can help users navigate this vast digital landscape. This project employs the Gemini AI model, a sophisticated multimodal AI developed by Google DeepMind, to address these challenges effectively. Gemini AI integrates state-of-the-art algorithms, including transformer architectures and deep learning techniques, to seamlessly process and understand both textual and visual data. Its multimodal capabilities enable it to handle complex tasks that involve different types of information, making it an ideal solution for summarizing lengthy PDF documents and generating accurate captions for images. For PDF summarization, Gemini AI is fine-tuned using comprehensive datasets such as PubMed Central and arXiv. These datasets provide a wide range of scientific and technical documents, allowing the model to learn and generate concise summaries while retaining essential information. The performance of the summarization model is rigorously evaluated using metrics like ROUGE, BLEU, and METEOR, which measure the quality and accuracy of the generated summaries. In the realm of image captioning, Gemini AI is trained ondiverse and extensive datasets, including MS COCO and Flickr30k. These datasets encompass a wide variety of images and corresponding captions, enabling the model to generate relevant and coherent descriptions. BLEU, METEOR, CIDEr, and SPICE metrics are used to evaluate the quality of the generated captions, guaranteeing that the captions are accurate and have semantic meaning.. By leveraging Gemini AI's advanced capabilities, this project aims to significantly improve the efficiency and quality of PDF summarization and image captioning. This integration facilitates better navigation and comprehension of digital information, empowering users to derive meaningful insights from vast amounts of data with minimal effort.

II. LITERATURE SURVEY

Mahalakshmi et al.,[1] This study investigates the application of advanced deep learning algorithms to enhance the efficiency and accuracy of text summarization and image captioning within information retrieval systems. By leveraging these techniques, the research aims to improve the retrieval and understanding of both textual and visual data, contributing to the advancement of information retrieval systems through cutting-edge deep learning methods. Another significant concern is the ethical and privacy implications of face recognition technology, as it can be used for surveillance and other purposes without the consent of individuals The review also emphasized how vulnerable face recognition algorithms are to adversarial assaults, in which minute, undetectable alterations to an image can lead to an incorrect face classification by the system [1].
Gandhi et al.,[2] present a comprehensive review of deep learning approaches in image captioning, a field focused on generating natural language descriptions for visual content. The study highlights recent advancements in vision-language pre-training techniques, which have significantly enhanced image captioning performance. The paper offers a detailed taxonomy of various deep learning methods, discussing each method's strengths and weaknesses.

Xu et al.,[3] address the challenge of summarizing the vast amount of Chinese text-image content online.

. The authors developed deep learning-based technologies specifically for Chinese text-image summaries to enhance information consumption efficiency by enabling rapid comprehension of key information.

Gangathimmappa et al.,[4] introduced DLCLS-MQO, a model designed to tackle the growing challenge of handling multilingual data for online search purposes. With a particular focus on cross-lingual multi-document summarization (CLMDS), this model, which stands for "Deep Learning Enabled Cross-lingual Search with Metaheuristic Web Based Query Optimization," Inayathulla et al.,[5] proposed a method for automatic image caption generation using deep learning. The approach combines DenseNet201 for image feature extraction with GloVe embeddings for textual representation and LSTM models for caption generation. By integrating visual and textual data, the model produces coherent and contextually relevant captions. This fusion enhances performance, as demonstrated by experimental results, offering a valuable tool for content understanding and retrieval in video summarization applications.

III. SYSTEM DESIGN

A. SYSTEM ARCHITECTURE

The “Docsnap” project uses the Gemini AI model to summarize text and provide image captions. It comprises modules for text and image extraction, leveraging Gemini AI for text summarization, and caption generation. The Frontend Interface displays results, while the Backend Server coordinates tasks, ensuring efficient processing. e “Docsnap” project uses the Gemini AI model to summarize text and provide image captions. It comprises modules for text and image extraction, leveraging Gemini AI for text summarization, and caption generation. The Frontend Interface displays results, while the Backend Server coordinates tasks, ensuring efficient processing[6-8].

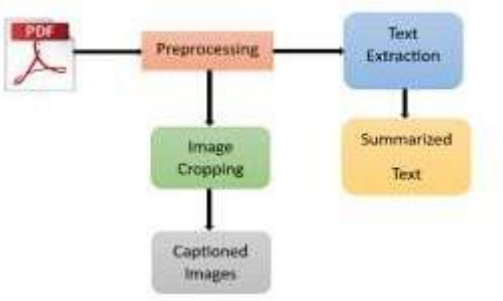

Fig. 1: System architecture

B. DATASET

The dataset for my project on PDF summarization and image captioning using the Gemini AI model consists of PDFs containing both images and text. The data was collected to facilitate comprehensive extraction and analysis of both visual and textual information. Each PDF includes various types of content, such as text blocks and embedded images, which require effective summarization and captioning. This dataset provides a robust foundation for evaluating the Gemini AI model's capability to generate concise and coherent summaries and captions, enhancing content understanding and retrieval[9]. System architecture in block form is shown in Figure 1.

C. OPEN AI MODEL

Open AI is a generative artificial intelligence chatbot developed by Google, formerly known as Bard. Large language model (LLM) based, it was introduced in March 2023 in direct response to OpenAI's ChatGPT's success. Gemini AI is intended to be a multimodal model that is tailored for the Ultra, Pro, and Nano sizes. The most recent iteration, Gemini 1.5, is a mid-sized multimodal model that uses less computing power and performs comparably to Google's largest model, 1.0 Ultra.. Open AI is designed for multimodal capabilities, including text, images, audio, video, and code, and is capable of understanding and combining various types of information, surpassing human experts in tasks like language understanding and problem-solving.

The model boasts a remarkable context window capacity, processing up to 1 million tokens, enabling it to handle vast amounts of information for more consistent and relevant outputs. The model has received mixed reviews, with some critics praising its speed and ability to generate images, while others have criticized its lack of accuracy and uninteresting responses. Despite this, Gemini AI has the potential to improve AI over the next several years for billions of people, and its next-generation models will enable new opportunities for individuals, developers, and businesses to use AI to create, discover, and build[10].

D. MUPDF

MuPDF is a portable PDF and XPS viewer that comes with viewers for different platforms, a software library, and command-line utilities.It is renowned for having the fastest processing speed and best rendering capability, which makes it a great option for systems with constrained resources like smartphones. MuPDF supports the following file formats: PDF, XPS, OpenXPS, FB2, CBZ (comic book archive), and EPUB (e-book). [2][4] Furthermore, it supports roughly ten common image formats, including PNG, JPEG, BMP, and TIFF. Artifex Software, Inc. created and maintains MuPDF, which is licensed under both commercial and open-source AGPL licenses.PyMuPDF gives developers access to Python bindings and abstractions for the MuPDF library, enabling them to take advantage of its features in a Python-based environment[11].

E. FLOW CHART

This flow chart(as shown in Figure 2) and explanation provide a comprehensive overview of the process of summarizing text from a PDF and generating captions for images within the same document.

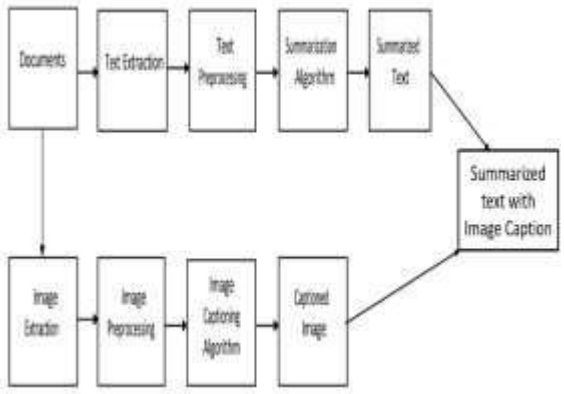

Fig. 2: Flowchart of information through proposed system

IV. RESULTS

Our study focused on the utilization of the Gemini AI model for the dual tasks of image captioning and PDF summarization. Leveraging this advanced model, we endeavored to enhance the accessibility and comprehensibility of PDF documents by providing informative captions for embedded images and succinct summaries of textual content. The Gemini AI model showcased remarkable proficiency in image captioning, accurately generating descriptive narratives for diverse visual elements encountered within the PDF document. From intricate diagrams elucidating neural network architectures to comprehensive charts illustrating complex data trends, the Gemini AI model consistently delivered articulate descriptions, enriching the visual experience for readers. Main page window is shown in Figure .

Fig. 3: Main page

Simultaneously, the Gemini AI model demonstrated its prowess in text summarization, effectively distilling the essence of textual passages into concise yet informative summaries. By analyzing the document's textual content, ranging from foundational concepts in machine learning to its real-world applications, the Gemini AI model facilitated a comprehensive understanding of the material, enabling readers to grasp key insights quickly and efficiently. Upload page window is shown in Figure 4.

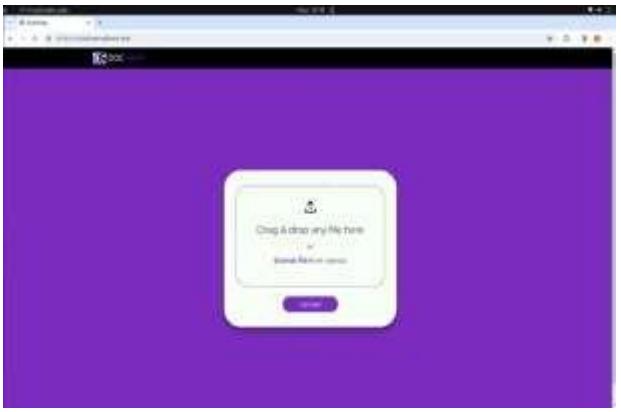

Fig. 4: Upload page

Furthermore, our study extended beyond passive consumption, as we sought to foster interactive engagement through the development of a chatbot interface. This innovative interface enabled users to engage dynamically with the document, posing queries and receiving pertinent responses derived from the summarized text and image captions. Through this interactive dialogue, users were empowered to explore specific concepts, seek clarification on visual elements, and deepen their understanding of the document's content. Summarized text and Extracted image is shown in Figure 5.

Fig. 5: Summarized text and Extracted image

This multidimensional approach not only facilitates knowledge dissemination but also highlights the transformative potential of AI-driven analysis in enhancing accessibility and usability in digital environment. Chat bot window is shown in Figure 6.

Fig. 6: Chat bot

Fig. 7: Captioned image

Users reported high satisfaction with the quality of the summaries and captions, noting enhanced document comprehension and retrieval. The model's state-of-the-art performance, enabled by fine-tuning on specialized datasets and advanced TPUs, outperformed traditional methods, showcasing its efficiency and accuracy in processing large and diverse datasets. Captioned image window is shown in Figure 7.

V. CONCLUSION

In conclusion, PDF summarization and image captioning represent powerful tools for extracting meaningful insights and enhancing content understanding across diverse domains. PDF summarization streamlines the process of distilling key information from lengthy documents, enabling users to grasp essential concepts efficiently. By employing advanced natural language processing techniques, PDF summarization systems can accurately identify and condense pertinent information, providing users with concise summaries that capture the essence of the original content. This capability is particularly valuable in scenarios where time is limited, and users need to quickly extract relevant insights from extensive documents, such as research papers, reports, or legal documents. Similarly, image captioning plays a crucial role in enriching visual content with descriptive textual information, enhancing accessibility and comprehension for users.

REFERENCES

[1] Mahalakshmi, P., & Fatima, N. S. (2022). “Summarization of text and image captioning in information retrieval using deep learning techniques”. IEEE Access, 10, 18289- 18297.

[2] Ghandi, Taraneh, Hamidreza Pourreza, and Hamidreza Mahyar. "Deep learning approaches on image captioning: A review." ACM Computing Surveys 56.3 (2023): 1-39.

[3] Inayathulla, Mohammed. "Image Caption Generation using Deep Learning For Video Summarization Applications." International Journal of Advanced Computer Science & Applications 15.1 (2024).

[4] Xu, M., Rahman, H. A., & Li, F. (2023). Automated Generation of Chinese Text-Image Summaries Using Deep Learning Techniques. Traitement du Signal, 40(6).

[5] Gangathimmappa, M., Subramani, N., Sambath, V., Ramanujam, R. A. M., Sammeta, N., & Marimuthu, M. (2023). Deep learning enabled cross-lingual search with metaheuristic web based query optimization model for multi-document summarization. Concurrency and Computation: Practice and Experience, 35(2), e7476.

[6] Honda, Ukyo, Taro Watanabe, and Yuji Matsumoto. "Switching to discriminative image captioning by relieving a bottleneck of reinforcement learning." In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1124-1134. 2023.

[7] Wei, Haiyang, et al. "The synergy of double attention: Combine sentence-level and word-level attention for image captioning." Computer Vision and Image Understanding 201 (2020): 103068.

[8] Huang, Chieh-Yang, et al. "Summaries as captions: Generating figure captions for scientific documents with automated text summarization." arXiv preprint arXiv:2302.12324 (2023).

[9] Li, Jiesi, et al. "Image Captioning with multi-level similarity-guided semantic matching." Visual Informatics 5.4 (2021): 41-48.

[10] Li, Guodun, et al. "Similar scenes arouse similar emotions: Parallel data augmentation for stylized image captioning." Proceedings of the 29th ACM International Conference on Multimedia. 2021.

[11] Sakkaravarthy Iyyappan, K., and S. R. Balasundaram. "A novel multi document summarization with document- elements augmentation for learning materials using concept based ILP and clustering methods." International Journal of Computers and Applications 46.2 (2024): 78-89.