LREC-COLING 2024

Longform Multimodal Lay Summarization of Scientific Papers: Towards Automatically Generating Science Blogs from Research Articles

LREC-COLING 2024 - The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation

Sandeep Kumar[1], Guneet Singh Kohli[2], Tirthankar Ghosal[3], Asif Ekbal[1]

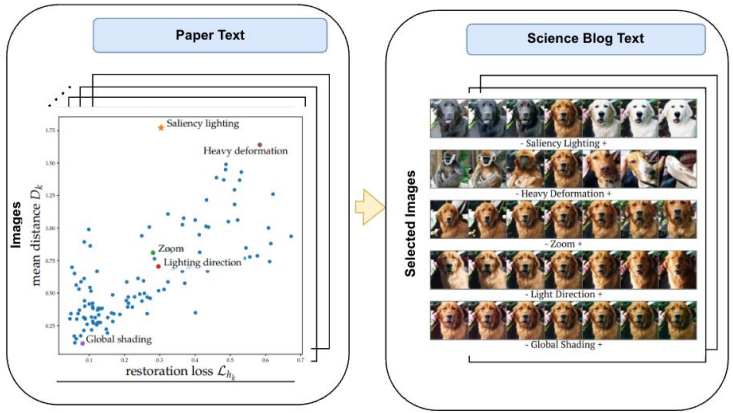
[1] Indian Institute of Technology Patna ,
[2] Thapar Institute of Engineering and Technology, India
[3] National Center for Computational Sciences, Oak Ridge National Laboratory, USA

Contributions

- We propose a novel task of automatically generating science blogs from research articles.
- We create an annotated dataset of nearly 3k papers, which includes science blogs generated using presentation transcriptions and annotated figures from the academic research articles.
- We introduce a pipelined multimodal multi-output framework for the task.
- We evaluate our approach both quantitatively, as well as qualitatively using human evaluation metrics.

Motivation for Science Blogs

Scientific blogging bridges the gap between intricate research and its comprehension by the general public, policymakers, and other researchers. It offers a platform for experts to articulate findings in layperson's terms, making scientific knowledge more accessible.



Dataset

- Papertalk website (3k papers)
- Extraction of Text and Figures from Paper (Science Parse, PDFFigures 2.0)
- Video to Transcription Generation (OpenAI Whisper model)
- Transcription to Science Blog Generation
- Figure Labelling
- Finally, we found that the average **length of papers** is approximately 5.6k words, while the **average summary length** is 732 words. Furthermore, there is an average of **8.8 figures or tables per paper** and an average of **4.1 figures/tables per slide**.

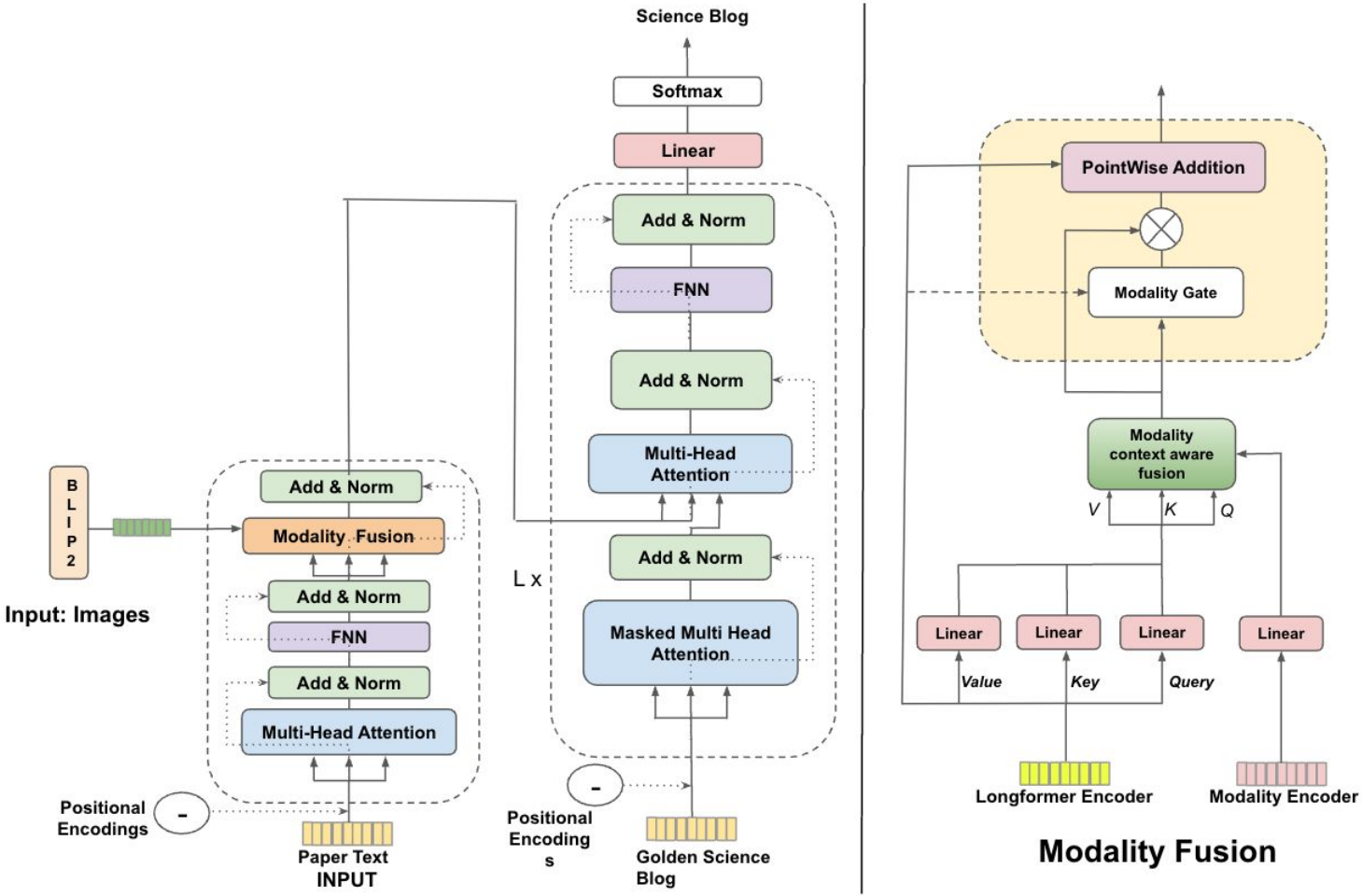


Fig 1: Methodology

Relevance Figure Selection

- image-text retrieval task
- Split the summary text T S into paragraphs
- Image-Text Matching (ITM)

$$ITM_j = \text{mean}\Big(\text{ITMScore}(I_j, TS_1), \ldots, \text{ITMScore}(I_j, TS_n)\Big)$$

- Text-Text Matching SS_j = mean (Cosine(E_j , E_1), . . . , Cosine(E_j , E_n))
- $R_j = w_1 \cdot ITM_j + w_2 \cdot SS_j$

Results

Mode	Model	R1	R2	RL	BS
Textual	Transformers (Vaswani et al., 2017)	42.07	5.78	26.77	73.84
Textual	DANCER (Gidiotis and Tsoumakas, 2020)	43.17	6.35	27.97	74.21
Textual	BigBird (Zaheer et al., 2020)	45.37	4.83	32.46	75.90
Textual	LED-large (Beltagy et al., 2020)	47.72	14.98	33.49	76.03
Multimodality	SITA (Jiang et al., 2023)	46.66	14.12	34.67	74.92
Multimodality	**MCF-TVa**	**48.42**	**15.02**	**37.03**	**77.06**
Multimodality	**MCF-TVc**	**48.69**	**15.17**	**37.38**	**77.47**

Table 1: Experimental results. (Abbreviation: R1/2/L: ROUGE1/2/L; BS: BERT Score

Human Evaluation Metrics

- Q1 (Readability): determines which of the blog are most readable?
- Q2 (Diversity): determines which of the blog contains the least amount of repetitive information?
- Q3 (Informativeness): determines how much useful information about the reviews does the blog provide? You need to skim through the original reviews to answer this.
- Q4 (RI: Relatedness to images): determines how much information of the figures does the blog provide?

Human Evaluation Result

Model	RI	Readability	Diversity	Informativeness
LED-Large	3.0	4.0	4.25	3.0
MCF-TVc	**3.5**	4.0	4.25	**3.75**

Table 2: Human evaluation results. Here, RI denotes relatedness to images

Error Analysis

- Error Propagation from Textual Science Blogs
- Varying Length of Science Blogs
- Irrelevant Words

Thank You!