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

Sandeep Kumar†, Guneet Singh Kohli‡, Tirthankar Ghosal§, Asif Ekbal†
†Indian Institute of Technology Patna, India,
‡Thapar Institute of Engineering and Technology, India,
§National Center for Computational Sciences, Oak Ridge National Laboratory, USA
†{sandeep_2121cs29,asif}@iitp.ac.in

Abstract

Science communication, in layperson's terms, is essential to reach the general population and also maximize the impact of underlying scientific research. Hence, good science blogs and journalistic reviews of research articles are so well-read and critical to conveying science. Scientific blogging goes beyond traditional research summaries, offering experts a platform to articulate findings in layperson's terms. It bridges the gap between intricate research and its comprehension by the general public, policymakers, and other researchers. Amid the rapid expansion of scientific data and the accelerating pace of research, credible science blogs serve as vital artifacts for evidence-based information to the general non-expert audience. However, writing a scientific blog or even a short lay summary requires significant time and effort. Here, we are intrigued by the question: *What if the process of writing a scientific blog based on a given paper could be semi-automated to produce the first draft?* In this paper, we introduce a novel task of Artificial Intelligence (AI)-based science blog generation from a research article. We leverage the idea that presentations and science blogs share a symbiotic relationship in their aim to clarify and elucidate complex scientific concepts. Both rely on visuals, such as figures, to aid comprehension. With this motivation, we *create a new dataset of science blogs* using presentation transcript and its corresponding slides. We create a dataset containing a paper's presentation transcript and figures annotated from nearly 3000 papers. We then propose a multimodal attention model to generate a blog text and select the most relevant figures to explain a research article in layperson's terms, essentially a science blog. Our experimental results with respect to both automatic and human evaluation metrics show the effectiveness of our proposed approach and the usefulness of our proposed dataset.

Keywords: Multimodal, Research Article Summarization, Multioutput, Image Retrieval, Blog Generation

1. Introduction

Social media have given rise to new opportunities for science organizations to communicate with the public (Su et al., 2017). Online environments, such as blogs, social networks, and online forums enable a more immediate and democratized dissemination of scientific information. Scientists and science communicators have new tools to engage directly with various audiences, fostering dialogue and participation in ways that were previously challenging (Brossard and Scheufele, 2013). Many scientists-turned-communicators continue to see online communication environments mostly as tools for resolving information asymmetries between experts and lay audience (Krause et al., 2021). As a result, they blog, tweet, and post podcasts and videos to promote public understanding and excitement about science.

Openness in communication can pose challenges. The use of anecdotal data or reliance on certain scientific figures can promote misinformation (Brossard and Scheufele, 2013). Misinformation spreads quickly on social media, leading to widespread dissemination of false information. In areas like public health, where accuracy is vital, credible scientific communication is essential. Such communication fosters trust in scientific professionals and institutions. The public, when equipped with reliable information, is more inclined to trust and understand scientific experts and their advances.

Misinformation and conspiracy theories can distort public perception and understanding (Joshi et al., 2023). Credible scientific communication is key to debunking these fallacies by providing evidence-backed data. Such communication aids informed decision-making across sectors from public health to environmental conservation, enhancing societal well-being. To enhance credible scientific communication, scientists and experts should engage the public, especially through platforms like social media, presenting data clearly and transparently. Collaborative efforts between scientists, journalists, and communication specialists are crucial for effective information dissemination. Prioritizing this communication counters the adverse effects of misinformation, leading to a more scientifically informed society.

LREC-COLING 2024, pages 10790–10801

Over the past few years, blogging ('web logging') has become a major social movement, and as such includes blogs by scientists about science. Blogs are highly idiosyncratic, personal and ephemeral means of public expression, and yet they contribute to the current practice and reputation of science as much as, if not more than, any popular scientific work or visual presentation (Wilkins, 2008). A survey of over 600 science bloggers reveals that on the broadest level, science bloggers see themselves engaging most often as explainers of science and public intellectuals (Brown Jarreau, 2015). Scientific blogging, distinct from standard research summaries, offers experts to present scientific findings in lay language, serving as a bridge between complex research and the general public, policymakers, and other researchers. In the expanding digital landscape, where the volume of scientific research is vast, credible science blogs stand as beacons of reliable, evidence-based information. They do not just restate findings; they offer a deeper dive, similar in style to presentation transcripts, where authors elucidate technical content using a blend of text and visuals to reach a broader, non-expert audience. Inspired by this idea, our paper introduces a novel task of AI-based science blog generation.

Science blogs retain critical information while maintaining readability and appeal for a general audience. Also, presentation transcripts naturally encapsulate complex topics in lay-friendly language and structure which is very similar to science blogs. So, we show how presentation transcript could be utilized to train the model and can be converted easily in the format of science blogs. Figures play an important role of explaining a complex concept and has vital information. Motivated by this, we propose a multimodal system to generate the science blogs. Science blogs also contain the the important visual figures of papers. Motivated by this, we build our system as multimodal multioutput summarization. This multimodal approach ensures a balanced representation of data, enhancing the understanding of scientific concepts for the general (or lay) readers.

Inspired by this idea, we first generate the presentation transcript and then covert it into science blog. We introduce a novel dataset containing almost 3k papers, presentation transcript, slides, figures of the paper and annotation whether it is included in the presentation or not and from scientific conferences and an annotated images which they present in their slides. Further, we propose a transformer based multimodal framework for this task. Our proposed system leverages both text and visual elements, such as figures and tables, to create engaging, easily digestible summaries. Our proposed technique is both multimodal and multioutput, which uses the figures of the paper as well as text to train. It gives crucial images from the papers as an output. Our proposed system leverages both text and visual elements, such as figures and tables, to generate engaging, easily digestible summaries and also gives out the relevant images from the paper. The proposed framework surpasses the traditional multimodal fusion baselines and reports to have achieved the best performance on almost all metrics. Lastly, we carry out detailed analyses, both quantitatively and qualitatively to compare our results.

A scientific blogs are lay and long as compared to a normal summarization. It is similar to a presentation transcript where the author tries to explain the technical paper to a wider and non-native audience. Similar to science blogs, the author explains it through images and tables. Motivated by this idea, we introduce a novel dataset containing almost 3k presentation transcript from scientific conferences and annotated images which they present in their slides. Further, we propose a new task, known as Longform Multimodal Lay Summarization (LMLS), to bridge the gap between highly technical scientific literature and the general public. LMLS aims to automatically generate comprehensive, yet accessible, summaries of scientific papers in the form of science blogs.

Experiments conducted on various scientific domains demonstrate the system's ability to produce high-quality summaries that retain critical information while maintaining readability and appeal for a general audience. The generated summary preservation of the original authors' intent. In conclusion, the Longform Multimodal Lay Summarization offers a promising pathway towards making scientific knowledge more accessible. The findings of this research have broad implications for science communication, education, and public engagement with science, and mark a significant step forward in the automated summarization of technical content.

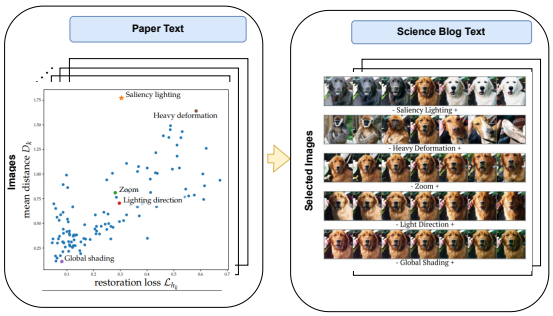

Figure 1: The illustration of our proposed task Longform Multimodal lay summarization. The image and the generated text can help reader quickly digest a paper in science blog format.

Our summarization addresses the challenges of both ”laysum” and ”longsum”. It is comprehensive, containing detailed information, and is also written in layman's terms. We introduce a novel type of summary. The generated summary is both extensive and phrased, easily comprehensible to laypeople. When presenting a paper, the author emphasizes the most crucial points, elucidating them more clearly. They ensure an informal explanatory tone, considering listeners may come from various domains. Conversely, in the written paper, the focus is primarily on those with domain knowledge, using a formal tone and technical terminology. This generated summary facilitates easier comprehension for readers. We make the resource and codes publicly available [1][2].

We summarize our contributions as follows :-

- 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.

2. Related Work

In this section, we discuss the related works in blog generation using AI, text summarization, multimodal summarization.

2.1. AI-based Blog Generation

Chen et al. (2013) developed an Automatic Travel Blog Generator, integrating mobile and desktop applications with a web platform. The system facilitates capturing, sharing, and organizing travel photographs and employs search engine and web mining techniques for efficient travel blog generation. Haiyan et al. (2014) proposed automatic generation of micro-blog user tags based on cluster analysis. Since science blog generation is similar to a lay and long summary, we discuss the related works about these below.

[1] https://github.com/sandeep82945/ScienceBlogGeneration.git

[2] https://www.iitp.ac.in/~ai-nlp-ml/resources.html

2.2. Text Summarization

Our proposed task bases on text summarization, the methods of which can be divided into extractive and abstractive methods (Widyassari et al., 2022). Extractive models (Zhang et al., 2018; Narayan et al., 2018; Xie et al., 2022; Narayan et al., 2018) directly pick sentences from article and regard the aggregate of them as the summary. In contrast, abstractive models (Gehrmann et al., 2018; Gerani et al., 2014; Oya et al., 2014; Jagan et al., 2016) generate a summary from scratch and the abstractive summaries are typically less redundant.

Two types of summaries have been introduced, *viz.* LongSumm (Long Scientific Document Summarization) and LaySumm (Lay Summarization) (Roy et al., 2021; Chandrasekaran et al., 2020). LongSumm task focuses on generating long summaries of scientific text. It is fundamentally different than generating short summaries that mostly aim at teasing the reader. The LongSumm task strives to learn how to cover the salient information conveyed in a given scientific document, taking into account the characteristics and the structure of the text. LaySumm addresses the issue of making research results available to a larger audience by automatically generating ‘Lay Summaries’, or summaries that explain the science contained within the paper in laymen's terms. A few works focused on improving the readability of biomedical document. (Luo et al., 2022) proposed adjustable readability level to cater to different user expertise levels. However, (Guo et al., 2024) proposed integration of external knowledge via retrieval-augmented methods produce lay language. In order to improve summary of scientific articles, a new way of prompting using content plans has been proposed (Creo et al., 2023).

2.3. Multimodal Summarization

A series of works (Wang and Bai, 2014; Greenbacker, 2011; Ahmad et al., 2004; Kumar et al., 2022) focused on generating better textual summaries with the help of multimodal input. Some of the prior research (Zhu et al., 2020; He et al., 2023; Zhu et al., 2018) have jointly generated text and select the most relevant image for these. A dataset, LoRaLay (Nguyen et al., 2023) incorporated visual/layout information alongside text for summarization. It addresses the challenges of long texts and the complex layouts of real-world documents.

In contrast to prior research, our work generates a science blog which adopts a multimodal and multi-output strategy, producing summaries that are simultaneously accessible to general audiences (lay) and sufficiently detailed for specialized readers (long). Additionally, to complement the visual elements, we employ an image retrieval method

to select figures from academic papers that are related to, and essential for, the generated blog post. As far as we know, generation of AI based science blog has never been explored.

3. Dataset

3.1. Dataset Collection

We collect papers, video presentation and the corresponding slide decks from openly available academic proceedings from Papertalk website[3]. Papertalk is a platform where scientists share a short video presentation (slides with narration) about a paper they have written. The papers are from several virtual conferences, especially in machine learning. In total, we collected information related to approximately 3k papers.

3.2. Extraction of Text and Figures from Paper

To extract text from the paper (in PDF form), we utilized a tool named Science Parse[4]. Science Parse processes scientific papers and returns them in a structured format. For extracting figures[5] and their corresponding captions from the paper, we employ PDFFigures 2.0[6]. This tool extracts figures, captions, tables, and section titles from scholarly documents, with a notable focus on documents within the field of computer science.

3.3. Video to Transcription Generation

In order to obtain the audio transcription of the video, we utilize the OpenAI Whisper model (Radford et al., 2022). Whisper is a general-purpose speech recognition model, trained on a vast dataset of diverse audio, and is capable of multitasking. It can perform multilingual speech recognition, speech translation, and language identification. We experiment with various versions of the model, including small, medium, large, and large-v2, and chose the model that produced the fewest errors in the generated transcription. Specifically, we employ the Whisper large model, named large-v2.

We found that the generated transcript sometimes does not recognize uncommon words or acronyms. To enhance reliability, we conducted post-processing of the audio transcript using the language model GPT-3.5-Turbo. We provided the following prompt instruction:

> **Prompt:** Correct the spellings in the generated audio transcript based on the paper [PAPER CONTENT] and **Transcript:** [Transcript].

Subsequently, to further improve reliability, we undertook manual tuning of the transcription.

3.4. Transcription to Science Blog Generation

We analyzed the presentation transcript and identified three major modifications required to transform it into a blog format.

Firstly, academic paper presentations can be delivered from various perspectives, but they are most often given in the first person, especially when the presenter is an author of the paper. In blogs, authors typically discuss others' research, which provides a more objective or distant perspective. Example: "The researchers aimed to investigate the impact of A on B...".

Secondly, the presentation includes the author's introduction, such as 'Hello everyone, I am Mishima. I present our paper...'. In contrast, blogs adopt a neutral tone, devoid of direct author references.

Thirdly, the presentation references its own format, indicating that it's a presentation. Blogs, however, strategically avoid such meta-referential elements, focusing on delivering content in a direct and immersive way. By prompting ChatGPT 3.5 with specific evaluation instructions, it achieves competitive correlation with human judgments compared to existing automatic metrics (Lai et al., 2023).

Hence, in order to convert transcription into blog format, we utilize the following prompt:

> **Prompt:** Given the following text from a presentation, please: 1. Convert the content to third person perspective. 2. Remove any references to the author or speaker. 3. Eliminate any mention of it being a presentation. **Presentation:** [Transcription]

Subsequently, to enhance the blog's reliability, we conduct manual tuning.

3.5. Figure Labelling

We hire an annotator to determine whether an image was present on a presentation slide. Given the task's straightforward nature, we enlisted an undergraduate student with four years of experience in scientific research publishing. We provide numerous examples from various papers to support and guide the annotator. The annotated data

[3] papertalk.org

[4] https://github.com/allenai/science-parse#science-parse

[5] We also classified the tables from the papers as figures for our purpose.

[6] https://github.com/allenai/pdffigures2

underwent regular checks, emphasizing the identification and correction of inconsistencies or ambiguities. We compensated the annotator at a rate of 4 USD per paper.

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.

4. Methodology

The proposed LMLS model is divided into two main parts, *viz.* Blog Generator(c.f. Figure 2) and Figure Selector. In this section, we outline the structure and intricacies of our model.

First, we discuss our proposed multimodal blog generation. The primary objective is to seamlessly integrate multimodal knowledge into the Longformer architecture. To achieve this, we introduce the Multimodal Contextual Fusion (MCF), an adapter-based module. Given a paper's textual information and its associated visual data, MCF effectively integrates multimodal information into textual representations. This adapter module can be effortlessly incorporated into multiple layers of the Longformer, enabling various levels of multimodal interactions. Figure 2 depicts our model's architecture.

4.1. Feature Extraction

Similar to prior works (Zhang et al., 2023; Zhu et al., 2023), we use the pre-trained feature extraction model, BLIP 2 (Li et al., 2023) to extract deep neural features for each figure of the paper. Specifically, we denote the generated figure feature as $F \in \mathbb{R}^{N \times C}$. We employ a Transformer encoder (Chen, 2023) to capture the sequential image context in the representations.

4.2. Multimodal Context Aware Attention

The conventional dot-product-based cross-modal attention mechanism facilitates direct interactions between textual representations and other modalities. In this setup, the textual representations act as the query, while the multimodal representations function as both the key and the value. Since each modality originates from a distinct embedding subspace, directly fusing multimodal information might not preserve the maximum contextual information. This can also introduce significant noise into the final representations. Drawing from the findings of (Yang et al., 2019), we advocate for multimodal fusion using Context Aware Attention. Initially, we produce key and value vectors conditioned on multimodal information, followed by the application of the traditional scaled dot-product attention. We provide a detailed explanation of this process below.

Given the intermediate representation H generated by the Blog Generator at a specific layer, we calculate the query, key, and value vectors Q, K, and $V \in \mathbb{R}^{n \times d}$, respectively, as given in Equation 1,

where W_Q, W_K, and $W_V \in \mathbb{R}^{d \times d}$ are the learnable parameters. Here, n denotes the maximum sequence length of the text, and d denotes the dimensionality of the generated vector.

$$[QKV] = H[W_Q W_K W_V] \tag{1}$$

Let $C \in \mathbb{R}^{n \times dc}$ denotes the vector obtained from visual representation. We generate multimodal information informed key and value vectors $\hat{K}$ and $\hat{V}$, respectively, as given by (Yang et al., 2019). To decide how much information to integrate from the multimodal source and how much information to retain from the textual modality, we learn matrix $\lambda \in \mathbb{R}^{n \times 1}$ (Equation 2). Note that U_k and $U_v \in \mathbb{R}^{dc \times d}$ are learnable matrices.

$$\begin{bmatrix} \hat{K} \\ \hat{V} \end{bmatrix} = (1 - \begin{bmatrix} \lambda_k \\ \lambda_v \end{bmatrix}) \begin{bmatrix} K \\ V \end{bmatrix} + \begin{bmatrix} \lambda_k \\ \lambda_v \end{bmatrix} \left(C \begin{bmatrix} U_k \\ U_v \end{bmatrix} \right) \tag{2}$$

Using the context-aware attention mechanism, we obtain the visual information infused vector H_v Finally, the multimodal information infused vectors $\hat{K}$ and $\hat{V}$ are used to compute the traditional scaled dot-product attention.

$$H_v = \text{Softmax}\left(\frac{Q\hat{K}_v^T}{\sqrt{\frac{d_k}{v}}} \right) \hat{V}_v \tag{3}$$

$$\hat{H} = H + H_v \tag{4}$$

The final multimodal information fused representation $\hat{H}$ is given by Equation 4. This vector $\hat{H}$ is inserted back into Blog Generator for further processing.

4.3. Relevance Figure Selection

To support the textual blog generated by our proposed model we set up a pipeline to retrieve the top k images from the pool of all figures in the input academic paper. In our corpus we had a relevance label for each image of the paper which signified if the paper was suitable for the presentation or not. We propose the problem as a image-text retrieval task since it will help us in selecting suitable images guided by the principle of relevant selection. The Blip 2 model is capable of handling an input of 768 Max tokens. For the first step, we split the summary text TS into paragraphs of 768 tokens. This ensures that a sentence does not break in between; if it does, we discard it.[7]

[7] In case a sentence breaks in between, we discard that sentence.

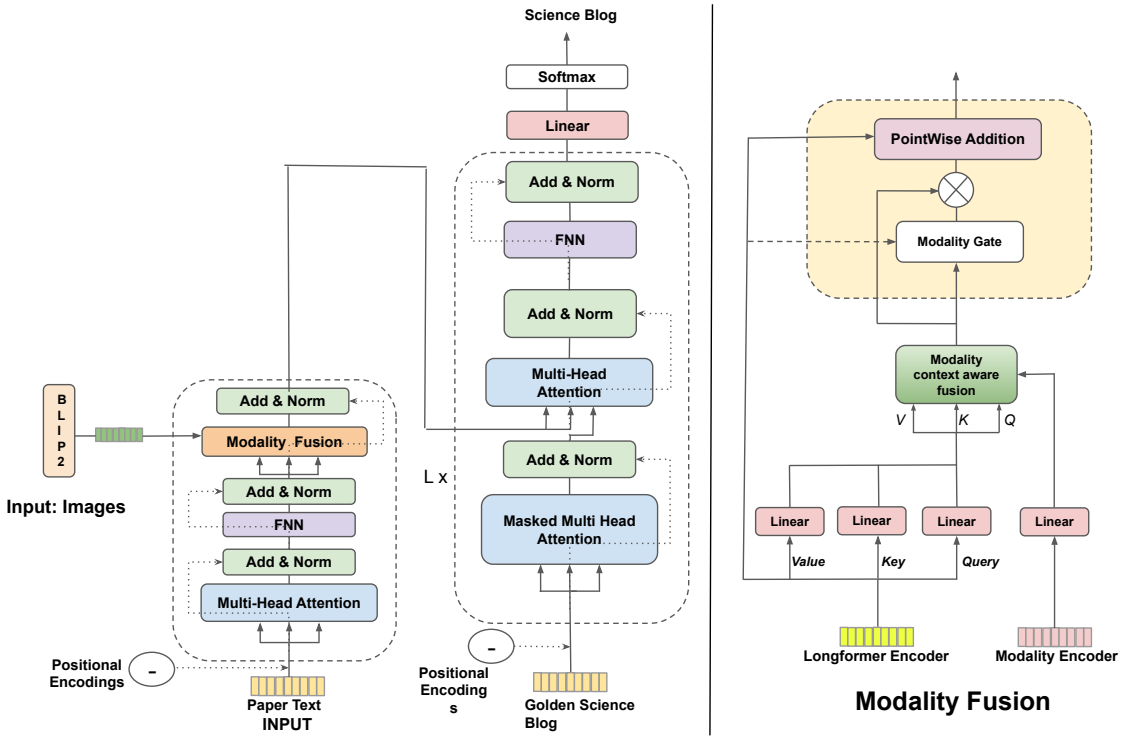

Figure 2: The architecture diagram of our proposed science blog generation (textual part).

$$TS = \{TS_1, TS_2, TS_3, \ldots, TS_n\} \quad (5)$$

where n is the number of paragraphs into which TS is broken.
Suppose the figures in the paper are represented by:

$$I = \{I_1, I_2, I_3, \ldots, I_p\} \quad (6)$$

where the document contains p figures.
Then, for an image I_j, we calculate the ITM score of that image against each text paragraph TS_i:

$$ITM_j = \text{mean}\Big(\text{ITMScore}(I_j, TS_1), \ldots, \text{ITMScore}(I_j, TS_n)\Big) \quad (7)$$

Image-Text Matching (ITM) score aims to learn fine-grained alignment between image and text representation. Similarly, we calculate the textual semantic similarity of the caption of the image with the paragraph text:

$$SS_j = \text{mean}\left(\text{Cosine}(E_j, E_1), \ldots, \text{Cosine}(E_j, E_n)\right) \quad (8)$$

Here, E represents the textual embedding obtained by the sentence transformer model[8].
Relevance score of a figure of a paper based on the science blog.

$$R_j = w_1 \cdot ITM_j + w_2 \cdot SS_j \quad (9)$$

Here, we propose a weighted sum[9] to determine how much weight should be given to the image-to-text similarity (ITM_j) and the text-to-text similarity (SS_j).

5. Experiments and Results

In this section, we illustrate our experimental settings and the comparative systems, followed by the results and its analysis. For a quantitative analysis of the science blog, we use the standard metrics for generative tasks – ROUGE-1/2/L (Lin, 2004). To capture the semantic similarity, we use the BERTScore (Zhang et al., 2020).

5.1. Experimental Setup

The number of images varies in a paper. So, while creating the image feature embedding, we padded with 0 in case the number of images is fewer than N and truncated in case it is more than N. We set the value of N to 12[10]. We implement our experiments in PyTorch and use the bart large variant provided via the Transformers (Lewis et al., 2020) package. Following (Yang et al., 2019), we use two separate 4-layer encoders with 8 attention heads to contextualize the figures.

[8] https://www.sbert.net/docs/usage/semantic_textual_similarity.html

[9] We set the weights as w_1= 0.7 and w_2= 0.3 empirically

[10] This was set empirically

Mode	Model	R1	R2	RL	BS
Textual	Transformers (Vaswani et al., 2017)	42.07	5.78	26.77	73.84
	DANCER (Gidiotis and Tsoumakas, 2020)	43.17	6.35	27.97	74.21
	BigBird (Zaheer et al., 2020)	45.37	4.83	32.46	75.90
	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
	MCF-TV_A	48.42	15.02	37.03	77.06
	MCF-TV_C	48.69	15.17	37.38	77.47

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

	R@1	P@1	R@2	P@2	R@3	P@3	R@4	P@4	R@5	P@5
Text to Text	0.2438	0.5691	0.4344	0.5233	0.6078	0.5032	0.7284	0.4651	0.815	0.4282
Text to Image	0.2111	0.5022	0.3977	0.4854	0.5719	0.4765	0.6946	0.4472	0.7867	0.4143
Text to text + image [0.5,0.5]	0.2427	0.5575	0.4268	0.5189	0.5919	0.4915	0.7213	0.4603	0.8093	0.4265
Text to text + image [0.7,0.3]	**0.2567**	**0.5852**	**0.4352**	0.5226	0.6069	**0.5041**	**0.7387**	**0.4698**	**0.8263**	**0.4311**
Text to text + image [0.3,0.7]	0.2304	0.5342	0.4171	0.5044	0.5956	0.4896	0.7062	0.452	0.801	0.421

Table 2: Figure Relevance scores based on various combinations

5.2. Results and Analysis

5.2.1. Text Based

As evident from Table 1, Longformer Encoder Decoder (seqlen: 16384) performs the best across all the metrics for the textual modality. We observed improvement of 1.44 points in the BERTScore, 3.89 points in the ROUGE-L score, 0.97 points in the ROUGE-1 score, and 0.2 points in the ROUGE-2 score when compared to the next best baseline. Pegasaus, Bigbird, Dancer and Transformer demonstrate admissible performance, considering that they have been trained from scratch.

5.2.2. Multimodality

Visual elements help authors present detailed results and complex relationships, patterns, and trends clearly and concisely (Schriger et al., 2006); reduce the length of the manuscript (Durbin, 2004) (Esteramorperez and Esteramorperez, 2020). Table 1 also shows the improvement in fusing caption information with the images. Thus, we gradually merge visual modalities using MCF module and obtain MCF-TVA and MCF-TVC for Longformer. We observe that the inclusion of figures leads to noticeable gains of 2-3% across the ROUGE, BERT scores. The rise in BERTScore also suggests that the multimodal variant generates a more coherent summary. Here, 'MCF-TVA' pertains to the model exclusively utilize visual elements, such as figures and tables to generate visual embeddings. On the other hand, 'MCF-TVC' is the model that also incorporates caption information and the visual elements during the training process.

Tables and figures are an integral part of a well-written scientific paper. We surmise that our model, to some extent, is able to gain more information from these visual cues and establish a relationship between text and figures of the paper while generating a blog.

5.2.3. Figure Relevance

For image retrieval, precision (P@k) and recall (R@k) metrics serve as pivotal indicators of a model's performance. We report the results in Table 2. In the 'Text to image' approach, we consider only the ITM score between figures and text as described in Equation 7. In the "Text to Text" approach, we take only the ITM score between text pairs as described in Equation 8. In the "Text to Text and image" approach, we determine the relevance score as the weighted sum of both "Text to Text" and "Text to image" scores as described in Equation 9. From the results, it is evident that combining text and figure to determine relevance scores offers superior performance compared to using either text-to-text or text-to-image alone. This suggests that incorporating both textual and visual information can provide a more comprehensive and accurate assessment of relevance. Among the various weight combinations tested for the text-to-text and image-to-text methods, the pairing of 0.7 and 0.3, respectively, yielded the best scores in most columns. This indicates that giving a higher emphasis on textual information (with a weight of 0.7) combined with a moderate consideration of visual content (with a weight of 0.3) produces the most optimal results.

5.2.4. Human Evaluation

Since the proposed LMLS task is a generative task, it is imperative to manually inspect the generated results. Consequently, we perform a human evaluation for a sample of 100 instances from our test set with the help of 5 evaluators. We ask the evaluators to judge the generated explanation, given the paper and the generated science blog. Each evaluator has to read the paper and then rate the generated science blog. We asked the responders to evaluate the summaries by rating

Model	RI	Readability	Diversity	Informativeness
LED-Large	3.0	4.0	4.25	3.0
MCF-TV$_C$	**3.5**	4.0	4.25	**3.75**

Table 3: Human evaluation results. Here, RI denotes relatedness to images

them between 1 to 5 on Likert Scale (Taherdoost, 2019) based on the following four questions:

- 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?

Table 3 displays the human evaluation analysis, providing average scores for each of the mentioned categories. Our analysis indicates that MCF-TVC bimodal produces blogs that are more syntactically informative than its textual counterpart. Additionally, it generates blogs that relate more closely to the figures referenced in the paper (showing an increase of 0.5 points compared to its textual counterpart). This reaffirms that these models can integrate information not explicitly present in the summary, such as graphical diagrams, architectural visuals, or tabular results.

5.3. Error Analysis

- **Error Propagation from Textual Science Blogs**: Our image retrieval framework relies on the textual content of science blogs to fetch relevant figures. Consequently, any generation error within the textual content may sometimes influence the image retrieval process leading to the retrieval of incorrect images.
- **Varying Length of Science Blogs**: As we have created the 'golden' science blog for training using transcriptions from presentations, which can inherently vary in length. Consequently, there are instances where the length of the generated science blogs does not align with that of the reference or 'golden' science blogs.
- **Irrelevant Words**: Given that scientific papers often contain equations and mathematical symbols, our proposed model occasionally introduces irrelevant words or produces sentences that are grammatically incorrect.

6. Conclusion and Future Work

In today's digital era, marked by abundant information and brief attention spans, reimagining the presentation of scientific knowledge is essential. Science blogs have emerged as effective platforms, translating intricate subjects into accessible narratives. To amplify their impact, the integration of multimodal summarization and image retrieval is vital. By combining concise summaries with appropriate imagery, we elevate both the clarity and appeal of scientific content. As the demand for effective science communication intensifies, the synergy of expertly constructed blogs with multimodal elements is indispensable, ensuring scientific insights are both understood and vividly depicted.

In the future, we would like to explore how this dataset can be utilized for various tasks, such as contextualizing scientific figures and tables. Specifically, we aim to automatically retrieve and rank snippets from the paper that are essential for interpreting their results, with the goal of making figures and tables more self-contained.

7. Limitations

We have proposed figure labeling system to determine the relevance of these images to be included in the generated blogs, however we have not investigated the locations of the images. This can be an interesting future work. Also, in some science blogs, the writer creates visual figures on their own. This paper limits itself to retrieving images contained within the paper solely for blog generation. This motivates working towards image generation for science blogs in the future.

8. Ethics Statement

While pre-trained language models have recently shown promising results in generating blog content, their effectiveness largely hinges on access to extensive annotated data. This dependency raises concerns due to the laborious and time-consuming nature of data annotation, which may be impractical for smaller research teams or institutions with limited resources. Nonetheless, our introduction of the LMLS is a preliminary step towards alleviating this challenge. Beyond this, our research contributes to a broader framework for addressing data scarcity, thus enhancing the viability of science blog generation systems in practical settings where annotated data is scarce. How-

ever, it is crucial to acknowledge that, like all machine learning models, the framework we propose is not foolproof and should be approached with caution when deployed in real-world applications. Writers should not depend on AI tools completely for science blog generation instead use this framework for a first draft generation. We have obtained permission from Papertalk to utilize the resources for our experiments.

9. Acknowledgements

Sandeep Kumar acknowledges the Prime Minister Research Fellowship (PMRF) program of the Govt of India for its support.

10. Bibliographical References

Saif Ahmad, Paulo C. F. de Oliveira, and Khurshid Ahmad. 2004. Summarization of multimodal information. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*. European Language Resources Association.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Dominique Brossard and Dietram Scheufele. 2013. Science, new media, and the public. *Science (New York, N.Y.)*, 339:40–1.

Paige Brown Jarreau. 2015. Science bloggers' self-perceived communication roles. *Journal of Science Communication*, 14(4):A02.

Muthu Kumar Chandrasekaran, Guy Feigenblat, Eduard H. Hovy, Abhilasha Ravichander, Michal Shmueli-Scheuer, and Anita de Waard. 2020. Overview and insights from the shared tasks at scholarly document processing 2020: Cl-scisumm, laysumm and longsumm. In *Proceedings of the First Workshop on Scholarly Document Processing, SDP@EMNLP 2020, Online, November 19, 2020*, pages 214–224. Association for Computational Linguistics.

Yi-Jiu Chen, Wei-Sheng Zeng, and Shian-Hua Lin. 2013. Automatic travel blog generator based on intelligent web platform and mobile applications. In *Information Technology Convergence - Security, Robotics, Automations and Communication, 5th International Conference on Information Technology Convergence and Services, ITCS 2013, Fukuoka, Japan, July 8-10, 2013*, volume 253 of *Lecture Notes in Electrical Engineering*, pages 355–364. Springer.

Zhe Chen. 2023. Attention is not all you need anymore. *CoRR*, abs/2308.07661.

Aldan Creo, Manuel Lama, and Juan Carlos Vidal. 2023. Prompting llms with content plans to enhance the summarization of scientific articles. *ArXiv*, abs/2312.08282.

Charles G. Durbin. 2004. Effective use of tables and figures in abstracts, presentations, and papers. *Respiratory care*, 49 10:1233–7.

Esteramorperez and Esteramorperez. 2020. How to manage complexity and realize the value of big data.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1602–1613. ACL.

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3029–3040.

Charles F. Greenbacker. 2011. Towards a framework for abstractive summarization of multimodal documents. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA - Student Session*, pages 75–80. The Association for Computer Linguistics.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2024. Retrieval augmentation of large language models for lay language generation. *Journal of Biomedical Informatics*, 149:104580.

Lv Haiyan, Che Xiaowei, and Ren Ying. 2014. The research on the automatic generation of microblog user tags based on clustering analysis. In *2014 IEEE 5th International Conference on Software Engineering and Service Science*, pages 633–636.

Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. 2023. Align and attend: Multimodal summarization with dual contrastive losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14867–14878.

Balaji Jagan, T. V. Geetha, and Ranjani Parthasarathi. 2016. Abstractive summarization: A hybrid approach for the compression of semantic graphs. *Int. J. Semantic Web Inf. Syst.*, 12(2):76–99.

Chaoya Jiang, Rui Xie, Wei Ye, Jinan Sun, and Shikun Zhang. 2023. Exploiting pseudo image captions for multimodal summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 161–175, Toronto, Canada. Association for Computational Linguistics.

Gargi Joshi, Ananya Srivastava, Bhargav Yagnik, Mohammed Hasan, Zainuddin Saiyed, Lubna A. Gabralla, Ajith Abraham, Rahee Walambe, and Ketan Kotecha. 2023. Explainable misinformation detection across multiple social media platforms. *IEEE Access*, 11:23634–23646.

Nicole M Krause, Dietram A Scheufele, Isabelle Freiling, and Dominique Brossard. 2021. The trust fallacy: Scientists' search for public pathologies is unhealthy, unhelpful, and ultimately unscientific. *American Scientist*, 109(4):226–232.

Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5956–5968, Dublin, Ireland. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multidimensional evaluation for text style transfer using chatgpt. *arXiv preprint arXiv:2304.13462*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Laura Nguyen, Thomas Scialom, Benjamin Piwowarski, and Jacopo Staiano. 2023. LoRaLay: A multilingual and multimodal dataset for long range and layout-aware summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 636–651, Dubrovnik, Croatia. Association for Computational Linguistics.

Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, and Raymond T. Ng. 2014. A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *INLG 2014 - Proceedings of the Eighth International Natural Language Generation Conference, Including Proceedings of the INLG and SIGDIAL 2014 Joint Session, 19-21 June 2014, Philadelphia, PA, USA*, pages 45–53. The Association for Computer Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Sayar Ghosh Roy, Nikhil Pinnaparaju, Risubh Jain, Manish Gupta, and Vasudeva Varma. 2021. Summaformers @ laysumm 20, longsumm 20. *CoRR*, abs/2101.03553.

David L. Schriger, R. Sinha, Sara Schroter, Pamela Y Liu, and Douglas G. Altman. 2006. From submission to publication: a retrospective review of the tables and figures in a cohort of randomized controlled trials submitted to the british medical journal. *Annals of emergency medicine*, 48 6:750–6, 756.e1–21.

Leona Yi-Fan Su, Dietram A. Scheufele, Larry Bell, Dominique Brossard, and Michael A. Xenos. 2017. Information-sharing and community-building: Exploring the use of twitter in science public relations. *Science Communication*, 39(5):569–597.

Hamed Taherdoost. 2019. What is the best response scale for survey and questionnaire design; review of different lengths of rating scale / attitude scale / likert scale.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Ting Wang and Changqing Bai. 2014. Understand the city better: Multimodal aspect-opinion summarization for travel. In *Web Information Systems Engineering - WISE 2014 - 15th International Conference, Thessaloniki, Greece, October 12-14, 2014, Proceedings, Part II*, volume 8787 of *Lecture Notes in Computer Science*, pages 381–394. Springer.

Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, and De Rosal Ignatius Moses Setiadi. 2022. Review of automatic text summarization techniques & methods. *Journal of King Saud University - Computer and Information Sciences*, 34(4):1029–1046.

John S Wilkins. 2008. The roles, reasons and restrictions of science blogs. *Trends in ecology & evolution*, 23(8):411–413.

Qianqian Xie, Jimin Huang, Tulika Saha, and Sophia Ananiadou. 2022. GRETEL: Graph contrastive topic enhanced language model for long document extractive summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6259–6269, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Baosong Yang, Jian Li, Derek F. Wong, Lidia S. Chao, Xing Wang, and Zhaopeng Tu. 2019. Context-aware self-attention networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 387–394. AAAI Press.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. *CoRR*, abs/2007.14062.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018. Neural latent extractive document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 779–784, Brussels, Belgium. Association for Computational Linguistics.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. MSMO: multimodal summarization with multimodal output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4154–4164. Association for Computational Linguistics.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Mul-

timodal summarization with guidance of multimodal reference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9749–9756.