 **Progress Report (Chanakya UG Fellows Project)**

|  |  |
| --- | --- |
| **Name: Raghvendra Kumar** | **Affiliation: IIT PATNA** |
| **Email Id and Contact No:** **raghvendra.kumar1004@gmail.com** | **Project Title:**  **Multimodal Abstractive Summarization with Multimodal Output** |
| **Name and Designation of Supervisor:**  **Dr. Sriparna Saha** **Associate Professor, CSE Dept.** | **Progress report Month and Year:**  **June, 2023** |
| **Broad area of project work:** **Natural Language Processing (NLP)** | **Sub area:**  **Summarization** |
| **Date of start: 22 July 2022** | **Date of Completion (Optional):** |
| **Achieved Objectives during the quarter period (Add extra sheet if required):**  I have been working on multiple problem statements, since the start of my PhD, which are correlated with each other and collectively contributes towards the overall completion of my designated project **“Multimodal Abstractive Summarization with Multimodal Output”**. The following section briefly introduces them all:   * **Large Scale Multi-Lingual Multi-Modal Summarization Dataset**   I have worked as capacity of third author for this paper.  **Work Abstract**: Significant developments in techniques such as encoder-decoder models have enabled us to represent information comprising multiple modalities. This information can further enhance many downstream tasks in the field of information retrieval and natural language processing; however, improvements in multi-modal techniques and their performance evaluation require large-scale multi-modal data which offers sufficient diversity. Multi-lingual modelling for a variety of tasks like multi-modal summarization, text generation, and translation leverages information derived from high-quality multi-lingual annotated data. In this work, we present the current largest multi-lingual multi-modal summarization dataset M3LS, and it consists of over a million instances of document-image pairs along with a professionally annotated multi-modal summary for each pair. It is derived from news articles published by British Broadcasting Corporation (BBC) over a decade and spans 20 languages, targeting diversity across five language roots, it is also the largest summarization dataset for 13 languages and consists of cross-lingual summarization data for 2 languages. We formally define the multi-lingual multi-modal summarization task utilizing our dataset and report baseline scores from various state-of-the-art summarization techniques in a multi-lingual setting. We also compare it with many similar datasets to analyse the uniqueness and difficulty of M3LS.  *Paper got accepted as oral presentation in EACL 2023: The 17th Conference of the European Chapter of the Association for Computational Linguistics, May 2-6, 2023*   * **Can Multimodal Pointer Generator Transformers Produce Topically Relevant Summaries?**   I have worked as capacity of second author for this paper.  **Work Abstract**: Due to the growth in demand for brief and pertinent multimedia material over the past few years, multimodal summarization has attracted a lot of study interest. Recently Transformers have been widely used for various sequence processing tasks due to their fast parallel processing ability compared to LSTMs. Although Multimodal Summarization (MS) has tractioned much research interest of late, a research gap exists in producing topic-relevant multimodal summaries. Since any summary deals with concise information, it should carry the essence of the topic from which it was derived. Further, due to the lack of alignment information among the images and the inter-modal segments, MS systems also face difficulty choosing appropriate pictorial summaries. To study these research questions, we propose a Multitask learning-based Multimodal Pointer Generator Transformer (MPGT), which utilizes the topic information of the samples to produce multimodal summaries. We also augment the popular MSMO dataset for this study with similar "On-Topic" and "Off-Topic" images. Our results show that inter-modal attention among images helps achieve better alignment in the visual modality and improves image precision scores. Our analysis also provides discussions on how we can further enhance topic-relevant MS systems.  *Paper got accepted as oral presentation in IJCNN 2023: International Joint Conference on Neural Networks*   * **Multimodal Rumour Detection: Catching news that never transpired!**   I have worked as capacity of first author for this paper.  **Work Abstract**: The growth of unverified multimodal content on microblogging sites has emerged as a challenging problem in recent times. One major roadblock to this problem is the unavailability of automated tools for rumour detection. Previous work in this field mainly involves rumour detection for textual content only. As per recent studies, the incorporation of multiple modalities (text and image) is provably useful in many tasks since it enhances the understanding of the context. This paper introduces a novel multimodal architecture for rumour detection. It consists of two attention-based BiLSTM neural networks for the generation of text and image feature representations, fused using a cross-modal fusion block and ultimately passing through the rumour detection module. To establish the efficiency of the proposed approach, we extend the existing PHEME-2016 data set by collecting available images and in case of non-availability, additionally downloading new images from the Web. Experiments show that our proposed architecture outperforms state-of-the-art results by a large margin.  *Paper got accepted as poster presentation in ICDAR 2023: The 17th International Conference on Document Analysis and Recognition* | |
| **Work which remains to be done under the project (Add extra sheet if required):**  All three of the above-mentioned papers have been finished to the best of their abilities and have been approved for presentation at conferences. We are now working on new concepts that also utilize/extend these works.   * For **Large Scale Multi-Lingual Multi-Modal Summarization Dataset**: Since this dataset contains multiple languages, we for the beginning are working on developing a dual summarizer model that can effectively summarize both English and Hindi text, while also being multimodal as it holds significant importance in today's globalized and diverse world. Such a model addresses the growing need for efficient and accurate summarization across different languages and modalities, providing numerous benefits in various domains. The idea behind this is that since English and Hindi are being widely spoken languages, particularly in regions such as India, having a model that can summarize both languages enable individuals to access information in their preferred language. This fosters inclusivity and ensures that important knowledge and insights are accessible to a broader audience, promoting cross-cultural exchange and understanding. * For **Multimodal Rumour Detection: Catching news that never transpired!** Since, this work helps us to find trustworthy tweets for crisis related events, we further try to develop a model which generates abstractive summaries of Twitter crisis events having multimodal output. Also, to further analyse our tweet summarization methodology, we have worked to create another dataset based on a collection of tweets associated with eight disaster/crisis events, which trended on Twitter from 2019-2023, namely, Bangalore Riots, Bangladesh Riots, Capitol Hill riots, George Floyd, Hong Kong riots, Israel Palestine, Northern Ireland Brexit, Peshawar Bombings, using a web scraper for social networking services. We used a maximum limit of 2000 tweets per event. Using this scraping tool, we download individual source tweets in English language. Reply tweets were not considered. As we know, users can also upload images, videos, GIFs, etc. on Twitter, but here we only download the image modality. Then, for dataset annotation, we had one graduate student write down gold-standard summaries for each topic, within 200 words. we used the Contrastive Language–Image Pre-training (CLIP) model to select the 10 most relevant images from each topic, by comparing them with the gold standard summaries obtained from the textual part. | |

**Brief Description of the project work done during this quarter period (Add extra sheet if required):**

* **Multimodal Abstractive Summarization of Crisis Events on Twitter**

The summarization of microblogs is an interesting problem as we are faced with some unique challenges, for instance, as documents increase in length, they tend to exhibit greater variability and fluctuations in the information they contain. This is especially true for longer texts such as the collection of microblogs based on an event, or interviews which often feature multiple shifts in dialogue or chain of thoughts. Consequently, it becomes more challenging to determine the contextual information that holds value for creating a summary. Models need to develop a more intricate understanding of the connections between keywords, topics, and phrases as the text expands. As documents increase in length, they tend to exhibit greater variability and fluctuations in the information they contain. This is especially true for longer texts such as the collection of microblogs based on an event, or interviews which often feature multiple shifts in dialogue or chain of thoughts. Consequently, it becomes more challenging to determine the contextual information that holds value for creating a summary. Models need to develop a more intricate understanding of the connections between keywords, topics, and phrases as the text expands. One of the approaches for handling long text summarization is to use BERTSUM, which is a fine-tuned BERT model, which performs extractive summarization. This novel architecture utilizes BERT as the encoder, which excels in handling lengthy inputs and capturing intricate word-to-word relationships. The pipeline involves employing BERT as a baseline model and subsequently fine-tuning it for the specific domain. This approach has gained significant popularity as leveraging the pre-trained baseline model can significantly enhance accuracy in the new domain and necessitates fewer data points.

A notable drawback of using BERTSUM for long text summarization is the maximum input token limit of 512 imposed by BERT. One of the solutions suggested for tackling the problem of summarizing long documents is chunking, where a document is split into multiple parts, and each part is used as a separate input to the model. Though the input length is no longer a concern with this method, the approach is still computationally expensive as each chunk needs to be encoded individually. Additionally, long-range attention is lost as there is no attention between the chunks. There have been multiple attempts to address this maximum input size problem, by increasing the limit to as high as 4096 tokens. Also, due to the quadratic dependency with respect to the input length, it is not feasible to use the model for time-critical scenarios. Longformer, a special type of Transformer, tried to tackle this by introducing an attention mechanism that gives rise to linear-time dependency.

Here, since we are having a huge input length, we have employed the method of extractive-abstractive summarization, wherein we select the important tweets through the extractive approach, thereby shortening the text content and reducing it below the maximum input token limit. After this, we will employ our methodology to get an abstractive summary through a Transformer-based model, named BIGBIRD, based on the extractive summary.

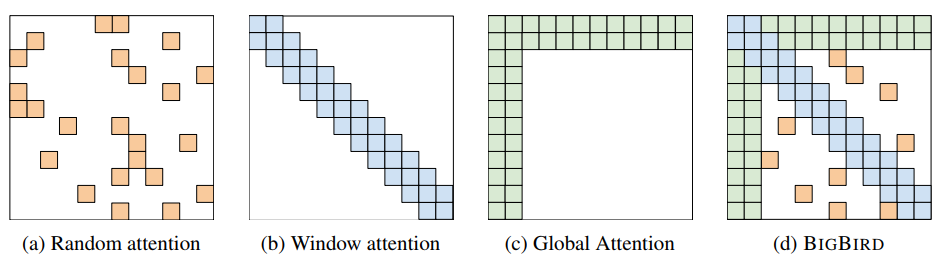
Now, to get a better understanding of the context, it is better to have another modality as part of our summary. Here, we are having images as part of the dataset, so we will be having an image summary as well. The images having the highest correspondence with the textual part and highly correlated with the other images in the dataset will be part of the final multimodal summary.

**Preprocessing of Data**:

First, the tweets are concatenated to get a single string. Then it is pre-processed by removing duplicate tweets, user mentions, hyperlinks, and stopwords. Then, the TF-IDF scores of each word in the tweet are calculated. Then, using the SpaCy library, we then create bigrams starting with that word, and fetch the tweets containing those bigrams. These tweets are then concatenated to form the extractive summary.

**Model Overview**

For the text modality, each of the extractive summaries obtained in the previous step is tokenized using the transformers tokenizer. Then the BigBird-Pegasus model, fine-tuned on the PubMed dataset, is built. BigBird-Pegasus, due to the combination of its random, window, and global attention mechanisms, has the capability to process sequences with a maximum length of 4096, with a linear dependency on input sequence length, unlike traditional Transformers. An abstractive text summary is generated by the model. For the image modality, the multimodal CLIP (Contrastive Language-Image Pretraining) model is used for selecting the best 10 representative images out of the available ones. The abstractive text summary obtained in the previous step is used for the text input, and all the images present in the dataset are used as image input.



Above picture denotes the attention mechanisms used in BIGBIRD

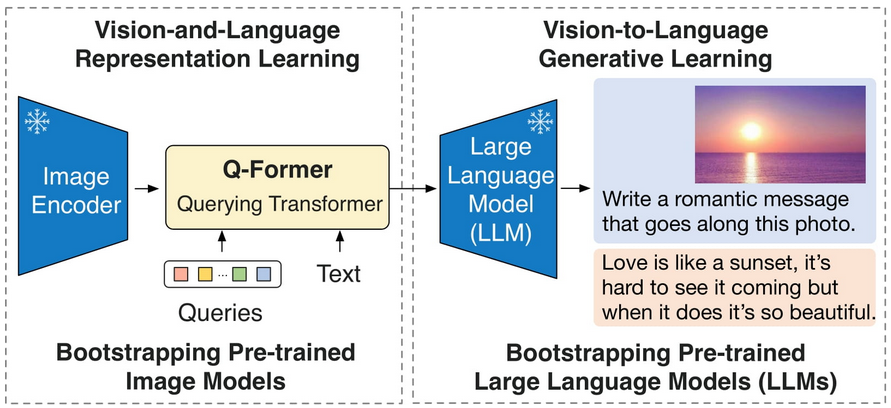
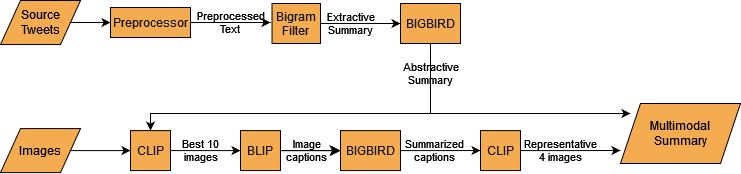


Image captioning using BLIP-2 Model

After this step, the BLIP-2 model, leveraging OPT-6.7b (a large language model with 6.7 billion parameters), was used for the generation of image captions for the top 10 images. Three models make up BLIP-2: an image encoder, a Querying Transformer (Q-Former), and a large language model. The weights of the image encoder and large language model are initialized from pretrained checkpoints and fixed while training the Querying Transformer, which maps a set of "query tokens" to query embeddings, from the embedding space of the image encoder to the LLM. The set of captions thus generated are concatenated and fed to the BigBird-Pegasus model for summarization. Next, the CLIP model uses this image caption summary and the 10 images to generate the top 4 images with the maximum similarity to the image caption summary. The abstractive text summary and these images together constitute the final multimodal summary. The architecture of the proposed model is represented in figure below.



**Experimental Setting and Evaluation Metrics**

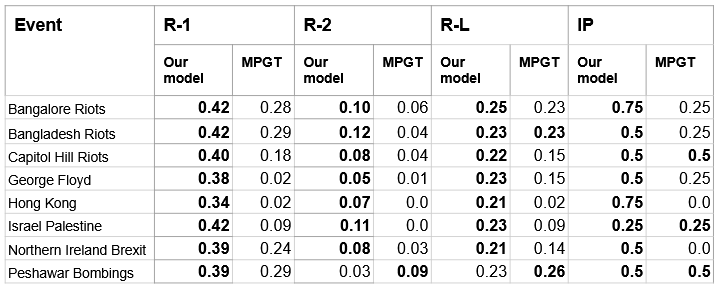
In this section, I explain the hyperparameters we used, and all the technical implementation details. Python libraries NumPy, Pandas, Keras, NLP, and Spacy were used on the PyTorch framework to execute the tests. We conducted our experiments using an NVIDIA A100-PCIE-40GB GPU with 40GB HBM2, connected via PCIe interface. The system's performance was assessed based on parameters like ROUGE-1, ROUGE-2, ROUGE-L, and Image Precision metrics. The BigBird-Pegasus model, fine-tuned on the PubMed dataset, with parameters block size=128, num random blocks=2, max length=4096, attention type='block sparse', was used for both the abstractive text summary generation task and image summary generation task. We set the max length as 4096, which is greater than the median input length of 3817. Nevertheless, some input samples need to be truncated as they are longer than 4096 tokens. The summaries have a median length of 162. Considering these statistics, setting the model's maximum generation length to 200 appears to be a reasonable choice. We can now create the BigBird-Pegasus evaluation function. First, we tokenize each article up to a maximum length of 4096 tokens. Then, we use beam search (with num beams=4 and max length=200) to generate the predicted abstract of the article. Finally, the predicted abstract tokens are decoded, and the resulting predicted abstract string is saved in the batch. For image caption generation, the model employs a masked self-attention Transformer as a text encoder and a ViT-B/32 Transformer architecture as an image encoder. By using a contrastive loss, these encoders are trained to optimize the similarity of (picture, text) pairs. This similarity score is used to generate captions with max new tokens=20.

**Output of the work done (Add extra sheet if required):**

For automatic evaluation of the textual summary, we have used the ROUGE score (R-1/R-2/R-L) metric. The image relevance is evaluated using the Image Precision (IP) metric. Here, we compare the results of our model on the new dataset that we created and is mentioned above with the state-of-the-art technique, Multimodal Pointer Generator Transformer (MPGT) (model from the 2nd paper, Can Multimodal Pointer Generator Transformers Produce Topically Relevant Summaries?), which was trained on the CNN-DailyMail dataset and provides multimodal summaries. Other unimodal models like Primera (a Longformer model) for text summary and GIT-large for image summary generation were also compared.

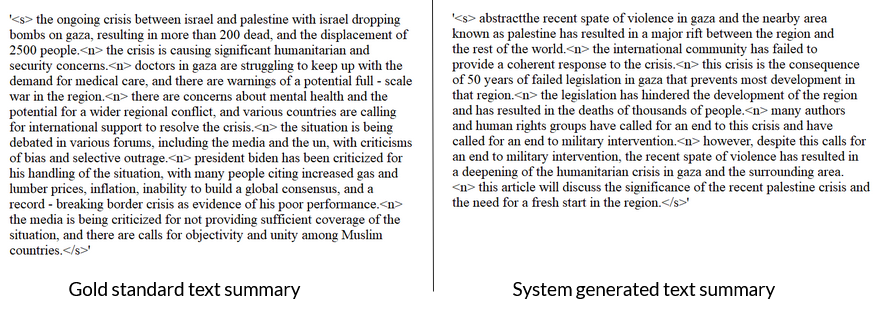
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Rouge-1** | **Rouge-2** | **Rouge-L** | **IP** |
| Primera (Text) | 19.93 | 3.39 | 16.05 | - |
| GIT-large (Image) |  |  | - |  |
| MPGT | 23.08 | 3.07 | 18.74 | 25.0 |
| BIGBIRD+BLIP | **39.53** | **8.17** | **22.62** | **56.25** |

As can be seen from the table above, our multimodal abstractive summarizer performs well in both the metrics ROUGE and Image Precision. The slight increase in IP value from 50.0 to 56.25 highlights the fact that the inter-modal image selection mechanism implemented in our model does help in the selection of better images. There is a jump of 17% in ROUGE-1 scores, which indicates the efficacy of our model. ROUGE-2 scores have also increased due to the selection of bigrams. Overall, we can say that a zero-shot transformer model like the one proposed, with BIGBIRD attention mechanism, without the chunking approach, and inter-modal image selection with the caption generation technique, is effective for the task of microblog summarization.



Above table shows event-wise results of MPGT and proposed model

Also, as seen in the figure above our proposed model performs well compared to the MPGT, with better results in almost all the events, except the Peshawar bombings event. The system-generated summaries of our proposed model are meaningful, with correct grammar and semantics too. Here in the next figure, we have the textual summary generated by our proposed model (right-side) vs the gold-standard summary written by the annotator (left-side) for the Israel Palestine event. We see that the model provides a more-or-less balanced result, except that it ignores important statistics like the number of people killed, etc. In the last figure, similarly, we have the pictorial summary generated by the proposed model (right-side) vs the gold-standard summary chosen by the annotator (left-side) for the Israel Palestine event. The model is mainly selecting images with text. It seems the model is a bit biased towards the text.



The above figure shows the gold standard vs Model-generated textual summary for the Israel-Palestine event.

The above figure shows the gold standard vs model-generated pictorial summary for the Israel-Palestine event.  
  
**References:**

1. Yang Liu. Fine-tune BERT for extractive summarization. CoRR, abs/1903.10318, 2019
2. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pretraining of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
3. Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020
4. Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences, 2021.
5. Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017