# [1]Multimodal Summarization: Integrating Text and Image Inputs for Enhanced Text and Visual Summaries

林嫈容
112423006
ingridlin0303@gmail.com

方子樽
112423008
jmjmjim1101@gmail.com

吳若瑜
112423024
jjjoyuwu@gmail.com

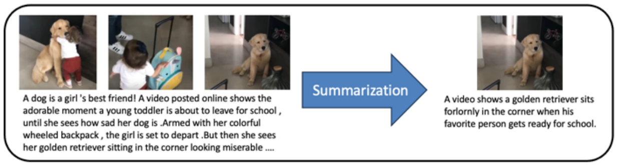
## ABSTRACT

Multimodal summarization, which integrates textual and visual data to generate comprehensive summaries, has garnered significant attention due to its potential to enhance user understanding and information retention. This paper explores a novel approach to multimodal summarization by dynamically integrating features extracted from text and images using pretrained models like ResNet50, BART, and VisualBERT. Our method aims to generate coherent text summaries accompanied by representative images, thereby improving the overall quality and informativeness of the summaries. Through empirical evaluation using ROUGE metrics, cosine similarity, and Euclidean distance, we demonstrate that our approach significantly outperforms traditional text-only summarization methods. The results indicate that incorporating visual information not only improves the semantic alignment but also enhances the structural coherence of the summaries. Despite the promising outcomes, our research is limited by the size of the dataset, which is constrained by our hardware capabilities. Future work will focus on optimizing the models for larger datasets and further refining the feature extraction and integration processes to achieve even better performance.

## Keywords

Multimodal Summarization; Cross-modality Learning; BART Model ;ResNet50 ; VisualBERT ;MSMO

## 1. Introduction


A dog is a girl 's best friend! A video posted online shows the adorable moment a young toddler is about to leave for school , until she sees how sad her dog is .Armed with her colorful wheeled backpack , the girl is set to depart .But then she sees her golden retriever sitting in the corner looking miserable ....

A video shows a golden retriever sits forlornly in the corner when his favorite person gets ready for school.

**Figure 1. Multimodal Input Multimodal Output(MSMO)**

Multimodal summarization has garnered significant attention in natural language processing and computer vision. This field involves integrating text and images to create comprehensive summaries, leveraging the strengths of each modality. Such summaries provide concise, informative representations of content, enhancing user understanding and information retention.

Traditional summarization techniques have focused primarily on text, but the inclusion of images adds context and appeal, improving the informativeness of summaries. With rapid advancements in multimedia technologies, combining visual and textual data has become increasingly effective. Multimodal summarization, which merges images and text, is gaining interest for its potential to deliver richer, more informative content. The growing importance of multimodal summarization is further underscored by Apple's plans to introduce this feature in their new iOS system, highlighting the recognized value of integrating visual and textual elements to enhance user experience.

Our research focuses on using multimodal inputs (both images and text) to generate text summaries and then selecting the most appropriate image based on the generated summary, thus creating a comprehensive multimodal output (see Figure 1).However, current approaches often prioritize text generation over image selection quality. To address this, we propose a novel method that dynamically integrates visual and textual information using pretrained models like ResNet50, BART, and VisualBERT. Our approach aims to fuse these features effectively, generating coherent summaries with representative images to enhance overall quality and informativeness.

Through empirical evaluation using ROUGE scores, our method strives to demonstrate improved performance over existing techniques, advancing the field of multimodal summarization and offering more user-friendly solutions for information consumption.

## 2. Related Work

### 2.1 Traditional Summarization

Text summarization has seen significant advancements in recent years. Numerous studies have explored various approaches and techniques to improve the accuracy and effectiveness of summarization models. We have reviewed and summarized some of the key papers in the field of text summarization, focusing on the methodologies and outcomes. Table 1 provides an overview of these papers.

**Table 1. Overview of the paper of text summarization**

| Author(s) | Brief Description | Method |
|---|---|---|
| Yadav, D. & Desai, J.(2023)[1] | Provides a detailed state-of-the-art analysis of text summarization concepts, approaches, techniques, datasets, and evaluation metrics. Discusses future research opportunities. | |
| La Quatra, M. & Cagliero, L.(2023)[2] | Introduces BART-IT, a sequence-to-sequence model based on BART for Italian text summarization, showing superior performance in ROUGE scores. | BART |
| Akshaya A. & Kambli M.(2022)[3] | Explores deep learning techniques for text summarization, highlighting methods based on neural networks. | Seq2Seq , Transformer |
| Md. Motiur Rahman, | Presents an abstractive text summarization model using multi- | MAPCoL |

| | | |
|---|---|---|
| Fazlul Hasan Siddiqui (2019)[4] | layered attentional peephole convolutional LSTM (MAPCoL) optimized with central composite design and response surface methodology for generating summaries. | |

## 2.2 Multimodal Summarization

Multimodal summarization has recently gained significant attention due to the rapid growth of multimedia data. Unlike traditional summarization methods that focus solely on text, multimodal summarization aims to integrate and summarize information from multiple modalities, such as text and images, to create a more comprehensive and informative summary. This approach not only enhances the informativeness of the summaries but also improves user engagement and satisfaction by providing a richer and more contextual representation of the information.

In recent years, several studies have explored various approaches and techniques to improve the accuracy and effectiveness of multimodal summarization models. Table 2 provides an overview of some of the key papers in this field, highlighting their methodologies and outcomes. These papers demonstrate the innovative strategies and models developed to handle the complexities of multimodal data, paving the way for future research and applications in this area.

**Table 2. Overview of the paper of multimodal summarization**

| Author(s) | Brief Description | Method |
|---|---|---|
| Zhu, J., Zhou, Y., Zhang, J., Li, H., Zong, C. (2020)[5] | Proposes Multimodal Summarization with Multimodal Output (MSMO) that combines textual and visual modalities to generate a pictorial summary. | MSMO |
| Zhang, Z., Meng, X., Wang, Y., Jiang, X., Liu, Q., Yang, Z. (2021)[6] | Introduces UniMS, a unified framework for multimodal summarization that integrates extractive and abstractive objectives with image selection. | UniMS |
| Atri, Y. K., Pramanick, S., Goyal, V., & Chakraborty, T. (2021)[7] | The authors introduce a novel approach to multimodal summarization, which combines extractive and abstractive techniques to select the most relevant images and text for summaries. This method leverages both pre-trained models and custom algorithms for improved accuracy. | UniMS |
| Liu, S., Li, H., Liu, T., Zhou, Y., Zhang, J., Zong, C. (2018)[8] | Proposes MSMO task, collecting a large-scale dataset and developing a multimodal attention model to generate text and select the most relevant image. | MSMO |

In summary, the research on multimodal summarization has made significant progress in recent years, showcasing various innovative approaches to handle and integrate multimodal data. These studies have not only improved the accuracy and richness of summaries but also significantly enhanced user satisfaction and engagement. Future research can build on this foundation to explore more advanced technologies and applications, driving the widespread adoption of multimodal summarization in different fields.

# 3. Method

## 3.1 Dataset

Our dataset originates from the paper "MSMO: Multimodal Summarization with Multimodal Output". This paper provides a downloadable dataset from a cloud storage, with the original data sourced from the Daily Mail website. The Daily Mail is a British newspaper website renowned for its extensive news coverage and multimedia content, spanning topics from current events to entertainment.

There are a total of 12 data combinations available for access. Due to the large size of each file and the limitations of our computer performance, we selected the smallest of these files—data5, which is 6.1 GB—as our dataset, and further compressed it for easier use.

The training set includes 14,507 texts and 94,580 images, while the test set comprises 10,262 texts and 68,521 images. This selection and processing method enable us to conduct efficient research within the constraints of limited hardware resources.

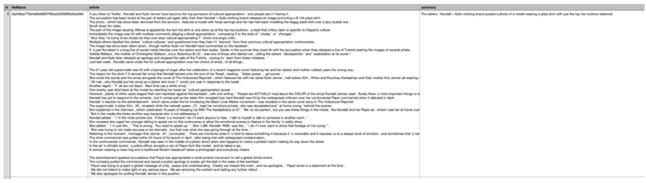
| | Texts | Images |
|---|---|---|
| Training set | 14,507 | 94,580 |
| Testing set | 10,262 | 68,521 |

### *3.1.1 Data Extraction Process*

The downloaded dataset contains numerous text files (txt) and image files (jpg). Each text file includes the following information:

- Filename：Identifies the text file and is used to locate the corresponding set of images.
- Title：The headline or title of the article.
- Body：The main content of the article, providing detailed information.
- Summary：A concise overview of the article, summarizing the key points.

We extracted these elements from each text file and organized them into a CSV file, as illustrated in the example table (see Fig. 2). The CSV file includes columns for the filename, article body, and summary. This structured format allows us to easily access and process the data for our summarization model.



| id | fileName | article | summary |
|---|---|---|---|
| 1 | [illegible] | If you listen to Twitter , Kendall and Kylie Jenner have become the top purveyors of cultural appropriation - and people are n't having it .<br>The accusation has been levied at the pair of sisters yet again after their their Kendall + Kylie clothing brand released an image promoting a $ 145 plaid shirt .<br>The photo , which has since been removed from the account , features a model with hoop earrings and her hair tied back modeling the baggy plaid shirt over a lacy bustier bra .<br>Scroll down for video<br>The part of the image causing offense is apparently the fact the shirt is only done up at the top two buttons – a style that critics claim is specific to Hispanic culture .<br>Immediately the image was hit with multiple comments alleging cultural appropriation , comparing it to the style of ` cholas ' or ` chongas ' .<br>` Now they 're trying to be cholas do they ever stop cultural appropriating ? ' wrote one angry critic .<br>Multiple others labelled the sisters ` culture vultures ' and questioned how they had n't ` learned ' from their previous cultural appropriation controversies .<br>The image has since been taken down , though neither Kylie nor Kendall have commented on the backlash .<br>It 's just the latest in a long line of social media frenzies over the sisters and their styles . Earlier in the summer they were hit with the accusation when they released a line of T-shirts bearing the images of several artists .<br>Voletta Wallace , the mother of Christopher Wallace , a.k.a. Notorious B.I.G. , was one of those who lashed out , calling the sisters ` disrespectful ' and ` exploitation at its worst ! '<br>Kendall and Kylie later released an apology and stopped the sale of the T-shirts , vowing to ` learn from these mistakes . '<br>Just last week , Kendall came under fire for cultural appropriation over her choice of emoji , of all things .<br>The 21-year-old supermodel was hit with a barrage of anger after her celebration of a recent magazine cover featuring her and her sisters and mother rubbed users the wrong way .<br>The reason for the stink ? A tanned flat emoji that Kendall tacked onto the end of her Tweet , reading : ` Sister power ... girl power . '<br>She wrote the words and the emoji alongside the cover of The Hollywood Reporter , which features her with her sister Kylie Jenner , half sisters Kim , Khloe and Kourtney Kardashian and their mother Kris Jenner all wearing<br>` Oh nah , why Kendall put her emoji as a darker skin tone ? ' wrote one user in response to the tweet .<br>Another raged : ` Y' all are not black . Next time use a white emoji . '<br>One snarky user shot back at the model by rewriting her tweet as ` cultural appropriation power . '<br>However , plenty of other users staged their own backlash against the backlash , with one writing : ` People are ACTUALLY mad about the COLOR of the emoji Kendall Jenner used . Surely there 's more important things to b<br>Kendall has yet to respond to the remarks , but it comes just as her sister Kim revealed how hard Kendall was hit by the widespread criticism over her controversial Pepsi commercial when it debuted in April .<br>Kendall 's reaction to the advertisement - which came under fire for trivializing the Black Lives Matter movement - was revealed in the same cover story in The Hollywood Reporter .<br>The supermodel 's sister Kim , 36 , revealed while the catwalk queen , 21 , kept her emotions private , she was devastated and ` at home crying ' behind the scenes .<br>Kim explained in the interview , which celebrated 10 years of Keeping Up With The Kardashians on E! : ` We 're not perfect , but you see these things in the media , like Kendall and her Pepsi ad , where I see her at home cryi<br>` But in the media she looks another way because she 's not addressing it . '<br>Kendall added : ` I 'm the most private one . If there 's a moment I do n't want anyone to hear , I talk to myself or talk to someone in another room . '<br>Kim revealed she urged her younger sibling to speak out on the controversy or allow the emotional scenes to feature in the family 's reality show .<br>She added : ` I 'm just like , `` This is wrong . You need to speak up '' . She -LSB- Kendall -RSB- was like , '' I do n't ever want to show that footage of me crying '' .<br>` She was trying to not make excuses or be dramatic , but that was what she was going through at the time . '<br>Referring to the moment , momager Kris Jenner , 61 , concluded : ` There are moments when it 's hard to leave something in because it 's vulnerable and it exposes us to a deeper level of emotion , and sometimes that 's har<br>The drink commercial was pulled within 24 hours of its launch in April , after being met with widespread condemnation .<br>In the controversial commercial , Kendall was seen in the middle of a photo shoot when she happens to notice a protest march making its way down the street .<br>In the ad 's climatic scene , a police officer accepts a can of Pepsi from the model , and he takes a sip .<br>A woman wearing a nose ring and a traditional Muslim headscarf takes a photograph and everybody cheers .<br>The advertisement sparked accusations that Pepsi has appropriated a racial protest movement to sell a global drinks brand .<br>The company pulled the commercial and issued a public apology to poster girl Kendall in the wake of the backlash .<br>` Pepsi was trying to project a global message of unity , peace and understanding . Clearly we missed the mark , and we apologize , ' Pepsi wrote in a statement at the time .<br>` We did not intend to make light of any serious issue . We are removing the content and halting any further rollout .<br>` We also apologize for putting Kendall Jenner in this position . ' | The sisters ' Kendall + Kylie clothing brand posted a photo of a model wearing a plaid shirt with just the top two buttons fastened |

,

**Figure 2. Example of CSV File Format for Dataset**

### *3.1.2 Summary of Dataset Processing*

a. Downloading and Selecting Data：We downloaded the dataset from the provided cloud storage. Due to hardware limitations, we selected the smallest data file (data5, 6.1 GB) and further compressed it for ease of use.

b. Organizing Data：The dataset includes text files with filenames, article bodies, and summaries. We extracted these elements and organized them into a CSV file.

c. CSV File Structure:

- fileName：Used to locate the corresponding set of images.
- artical：Contains the detailed content of the article.
- summary：Provides a brief overview of the article.

By organizing the data in this manner, we facilitated efficient processing and analysis, enabling our model to generate text summaries and select appropriate images based on the text content.

## 3.2 Baseline

We chose the paper "Exploiting Pseudo Image Captions for Multimodal Summarization" as our baseline due to its innovative approach to bridging the semantic gap between visual and textual modalities in multimodal summarization tasks. This paper effectively addresses the challenge of integrating images and text to create coherent and informative summaries, making it a relevant and robust benchmark for our work. Additionally, the method's use of image captions to enhance cross-modal alignment provides a solid foundation for comparing and improving upon existing techniques.

Figure 3 illustrates the architecture of our baseline method. It integrates the ResNet and Transformer models for feature extraction from images and text, respectively. The extracted features are then fused and processed by an LSTM-based summary model to generate the final output. This architecture ensures effective handling and integration of multimodal data, leading to high-quality text summaries.

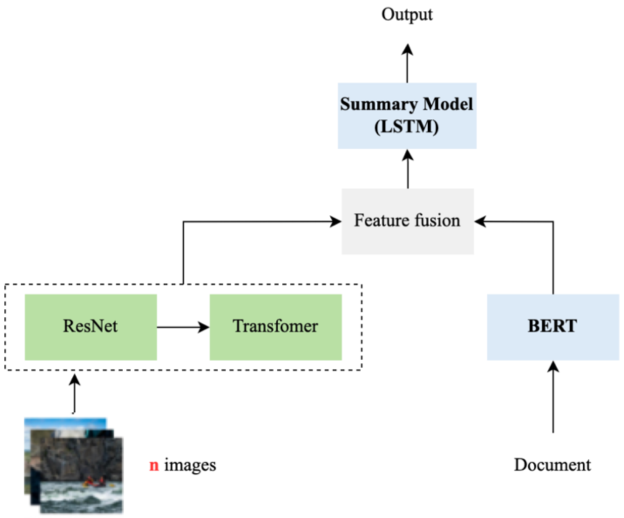


**Figure 3. Baseline Method Architecture**

### *3.2.1 Models Used in Baseline Method*

a. DistilBERT

- Purpose：Used for text preprocessing and feature extraction.
- Description：DistilBERT is a streamlined version of BERT (Bidirectional Encoder Representations from Transformers), smaller, faster, and lighter. It retains 97% of BERT's language understanding capabilities but is 60% faster, achieving a balance between efficiency and performance. Using knowledge distillation, it inherits and compresses BERT's model size, reducing computational resource needs while maintaining robust language comprehension. In natural language processing, DistilBERT efficiently tokenizes and extracts features from text, reducing training and inference time. This method is particularly beneficial for projects handling large datasets, enhancing overall efficiency and effectiveness.

b. ResNet152

- Purpose：Used for image feature extraction.
- Description：ResNet152 is a deep residual network with 152 layers. It is highly effective in image recognition tasks due to its deep architecture, which allows it to learn intricate patterns in visual data. For our baseline method, ResNet152 is used to extract high-dimensional feature vectors from images, providing a rich representation of the visual content.

c. LSTM (Long Short-Term Memory)

- Purpose：Used for decoding and generating text summaries.
- Description：LSTM is a type of recurrent neural network (RNN) capable of learning long-term dependencies, making it particularly effective for sequence prediction tasks. In our baseline method, an LSTM-based decoder processes features combined from text and images to generate coherent and contextually relevant text summaries. The structural design of LSTM allows it to remember past information and forget details that are no longer relevant, a crucial feature when dealing with lengthy articles. This capability makes LSTM highly powerful in handling long sequence data that requires understanding and generating content based on complex contexts. Therefore, when creating text summaries closely related to image content, LSTM effectively integrates multimodal features to ensure the summaries are not only accurate but also semantically coherent.

### *3.2.2 Data Preprocessing*

a. Text Data Processing：We utilize the DistilBERT tokenizer (DistilBertTokenizer) for processing text data. This involves converting the raw text into a format that the model can understand. Additionally, DistilBERT is employed to extract textual features, transforming the text into a representation that can be further processed.

b. Image Data Processing：For image processing, we use a pre-trained ResNet152 model to extract image features. The raw images are converted into high-dimensional feature vectors through the ResNet152 model. These extracted features are then averaged to create a unified feature representation for each image.

### *3.2.3 Feature Fusion*

In your model, feature fusion is achieved by directly concatenating text and image features. This fusion method involves merging features from two different sources along the feature dimension to form a unified feature vector. This vector is then used as the input for further processing in the neural network. Specifically, this is done using PyTorch's torch.cat function, which concatenates two feature arrays along the designated dimension. This method is not only straightforward and intuitive but also effective, as it allows the model to consider information from both text and images simultaneously during learning, enhancing its ability to process multimodal data.

### *3.2.4 Summary Generation*

The combined features are fed into a Long Short-Term Memory (LSTM) based decoder, designed to generate a sequence of outputs. The LSTM model processes the fused features and generates a coherent text summary. This involves decoding the combined features into meaningful text, effectively summarizing the input content.

### *3.2.5 Model Architecture*

a. Decoder：An LSTM-based decoder that receives the fused features and generates the text summary.

b. Summary Model：An integration model that combines the decoding results to produce the final text summary. This model is responsible for the coherent fusion of features and the generation of concise textual output.

By using these models, our baseline method effectively handles the integration of multimodal data, leading to high-quality text summaries. This comprehensive approach ensures that the generated summaries are both textually and visually informative, providing a more holistic representation of the original content.

Although the Baseline method has achieved some success in handling multimodal data, there are still some limitations. Firstly, the Baseline method has limited capability in integrating image and text features, making it difficult to fully capture the complex relationships between images and text. Secondly, the models used in the Baseline method have room for improvement in feature extraction and representation when dealing with multimodal data. Additionally, the data preprocessing steps are not comprehensive enough, which may result in generated summaries that lack accuracy and relevance.

To address these issues, our approach introduces the VisualBERT model, which can better understand and integrate image and text information, thereby generating more meaningful and coherent multimodal summaries. We also utilize the BART model for text feature extraction and summary generation, enhancing the quality and efficiency of text processing. Finally, we implement comprehensive improvements in data preprocessing and feature extraction to ensure high-quality input data, further enhancing the accuracy and relevance of the final summaries.

These improvements effectively address the limitations of the Baseline method, enabling our approach to perform better in multimodal data processing and summary generation.

## 3.3 Our Method

### *3.3.1 Models Used in Our Method*

a. ResNet50

- Purpose：Used for image feature extraction.
- Description：ResNet50 is a deep residual network that helps in extracting high-dimensional feature vectors from images. It is known for its excellent performance in visual recognition tasks and is widely used for processing images to capture relevant visual features.

b. BART (Bidirectional and Auto-Regressive Transformers)

- Purpose: Used for text feature extraction and text summarization.
- Description: BART is an advanced denoising autoencoder designed for pretraining sequence-to-sequence models specifically for natural language processing tasks. It integrates the strengths of bidirectional encoders and the characteristics of autoregressive decoders. This hybrid model structure allows BART to excel in tasks that require a deep understanding of context and the generation of coherent text. It is particularly adept at generating detailed and contextually relevant text summaries, making it highly suitable for applications such as summarization, translation, and text completion. The pretraining of BART involves corrupting text with a noising function and learning to reconstruct the original text, which enhances its ability to handle various noise patterns and textual inconsistencies, thereby improving the robustness and effectiveness of generating high-quality text outputs.

c. VisualBERT

- Purpose: Used for refining the combined visual and textual features.
- Description: VisualBERT is a transformative model that is adept at handling both visual and textual information, making it particularly effective for tasks that require the integration of these two types of data. It operates by enhancing the fused features, providing a deep understanding of the interplay between text and images. This capability is crucial for generating meaningful multimodal summaries, where the coherence between visual content and textual description is essential. VisualBERT uses the attention mechanism common to BERT architectures to focus on relevant aspects of both text and images, facilitating a more nuanced interpretation and richer representation of multimodal data.

Our proposed method leverages multimodal inputs (text and images) to generate text summaries and select the most appropriate image based on the text summary. Figure 4 illustrates the architecture of our method (see Fig. 4), which integrates textual and visual information to produce comprehensive multimodal summaries. The process involves the following key steps:

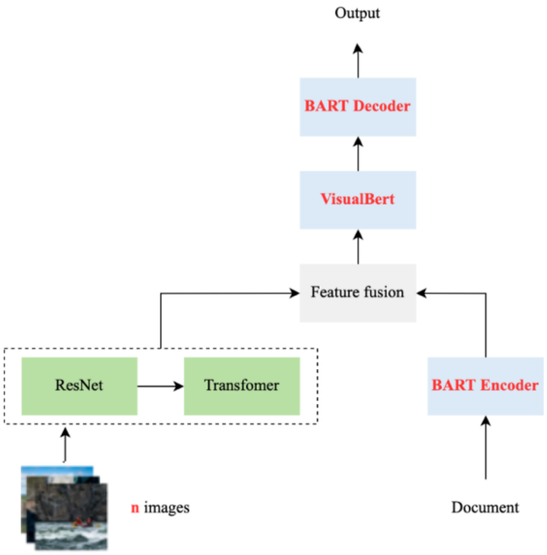


**Figure 4. Our Method Architecture**

### 3.3.2 Data Preprocessing

a. Text Data Processing：We preprocess the textual data by removing whitespace, converting to lowercase, and normalizing the input data. This step ensures that the text is in a suitable format for model processing.

b. Image Data Processing：We use a pre-trained ResNet50 model to process image data. The images are first resized and normalized before being fed into the ResNet50 model, which extracts relevant features representing the visual content of the images.

### 3.3.3 Feature Extraction

a. Text Feature Extraction：Text feature extraction is performed using a pretrained BART model, which is specifically designed for language understanding tasks. First, the BART tokenizer is used to preprocess the text, converting it into a format that the model can recognize, and applying truncation and padding as necessary to meet the input requirements of the model. Next, the processed data is fed into the BART model, and the encoder's last hidden state is obtained without performing gradient calculations. Finally, these hidden states are averaged to produce a composite feature vector. This vector serves as an input feature for subsequent tasks, effectively capturing the deep semantic information of the text.

b. Image Feature Extraction：The image feature extraction method utilizes a pretrained ResNet-50 model, which excels in image classification tasks. First, each image undergoes preprocessing, including resizing, center cropping, conversion to tensor, and normalization. Then, the processed image is input into the ResNet-50 model to extract its feature vector. For each set of images corresponding to a file name, all image features are extracted sequentially and averaged to obtain a composite feature vector. This method effectively captures the common features of multiple images, providing high-quality image representations for subsequent tasks.

### 3.3.4 Feature Fusion

VisualBERT integrates features from text and images. First, text and image features are extracted separately and then input into VisualBERT. VisualBERT, through its bidirectional encoder structure and attention mechanism, effectively captures and understands the interplay between text and images. This method enables the model to generate comprehensive multimodal feature representations, enhancing contextual understanding and semantic coherence. As a result, it provides more accurate and meaningful outcomes in subsequent tasks, such as text summarization. The strength of VisualBERT lies in its ability to simultaneously process visual and textual information while preserving and enhancing the interactions between them.

### 3.3.5 Summary Generation

The fused features are generated by the VisualBERT model, which further refines these combined features. The VisualBERT model, through its bidirectional encoder structure and attention mechanism, effectively captures and understands the interplay between text and images, enhancing feature representation and preparing it for the summarization task. This fusion method enables the model to generate feature representations that are more semantically coherent and contextually relevant. Subsequently, the output from the VisualBERT model, along with the original text input, is fed into the BART decoder. The BART decoder leverages BART's powerful text generation capabilities, combined with the visual context provided by the images, to generate the final text summary.

By adopting this multimodal approach, our goal is to produce summaries that are not only textually coherent but also visually informative. This method fully utilizes the complementary nature of image and text data, allowing the generated summaries to more comprehensively reflect the original content. By combining the visual information from images and the semantic information from text, the final summaries are significantly improved in both quality and informativeness. This not only enhances the accuracy and relevance of the summaries but also provides a richer understanding of the content.Evaluation indicators

# 4. Evaluation indicators

## 4.1 text-based evaluation metric

### 4.1.1 ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics specifically designed to evaluate the quality of automatic summarization and machine translation. It achieves this by comparing the generated summary against reference summaries created by human annotators. Widely used in the field of natural language processing (NLP), ROUGE helps in assessing the performance of various text summarization models.

#### 4.1.1.1 ROUGE Metrics

a. ROUGE-N：ROUGE-N is an automated metric used to evaluate summary quality. It mainly measures the number of n-gram co-occurrences between candidate summaries and reference summaries.[9] [10]

- ROUGE-1：ROUGE-1 measures the overlap of unigrams (single words) between the system-generated summary and the reference summaries. It evaluates the match of individual words, where n=1. ROUGE-1 is used to evaluate the basic content overlap and the presence of essential keywords in the generated summary. A higher ROUGE-1 score indicates a higher similarity between the generated summary and the reference summary at the word level.

- ROUGE-2：ROUGE-2 measures the overlap of bigrams (two consecutive words) between the system-generated summary and the reference summaries. It assesses the match of word pairs, where n=2. ROUGE-2 is used to evaluate the fluency and coherence of the summary by considering the sequence of words. A higher ROUGE-2 score indicates a higher similarity between the generated summary and the reference summary at the bigram level.

b. ROUGE-L：ROUGE-L measures the longest common subsequence (LCS) between the system-generated summary and the reference summaries. It evaluates the match of the longest sequence of words that appear in the same order in both summaries. ROUGE-L captures the sentence-level structure similarity and allows for an evaluation of the summary's ability to preserve the order of the content. A higher ROUGE-L score indicates a higher similarity between the generated summary and the reference summary in terms of overall structure and sequence.

The ROUGE metric provides a robust method for assessing the quality of automated summarization. By measuring the overlap of words, phrases, and sentence structures between candidate and reference summaries, these metrics help us comprehensively evaluate generated summaries. These scores are calculated using the rouge-score suite. They ensure that the abstract not only captures key content but also maintains coherence and flow, providing a comprehensive assessment of the abstract's quality. In summary, ROUGE-1 measures word-level matching, ROUGE-2 measures bigram-level matching, and ROUGE-L measures sentence structure and word order.

## 4.2 The correlation between images and text

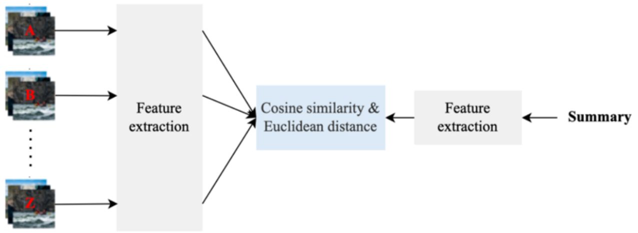


**Figure 5. Process of Correlating Images and Text Using Feature Extraction and Similarity Metrics**

This Figure depicts the process of correlating images and text using feature extraction and similarity metrics. The figure shows a series of images labeled A to Z on the left side, which undergo feature extraction to capture their essential characteristics. Similarly, a text summary also undergoes feature extraction. These extracted features are then compared using cosine similarity and Euclidean distance, which are techniques to measure the similarity between the image features and the text summary features. The comparison results help in determining how closely the images relate to the text summary. This figure serves as a visual guide to understanding how the correlation between images and text is calculated, with cosine similarity and Euclidean distance being explained individually in sections 4.2.1 and 4.2.2.

### *4.2.1 Cosine similarity*

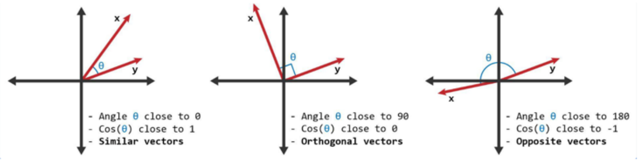


**Figure 6. Cosine similarity**[11]

Cosine similarity is a measure used to evaluate the similarity between two vectors by calculating the cosine of the angle between them. As illustrated in Figure 6, cosine similarity measures the orientation of the vectors, not their magnitude. When the angle between two vectors is small, their cosine similarity is close to 1, indicating high similarity. Conversely, when the angle is large, the cosine similarity is close to 0, indicating low similarity (see Fig. 6). The formula is as follows:

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2}\sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

where $A$ A and $B$ B are two vectors, $\cdot$ · denotes the dot product of the vectors, ‖ A ‖ and ‖ B ‖ are the magnitudes of the vectors. The cosine similarity ranges from -1 to 1, where values closer to 1 indicate higher similarity, and values closer to 0 indicate lower similarity. Cosine similarity is widely used in high-dimensional data similarity calculations, especially in text summarization and image similarity research.[12] [13]

### *4.2.2 Euclidean distance*

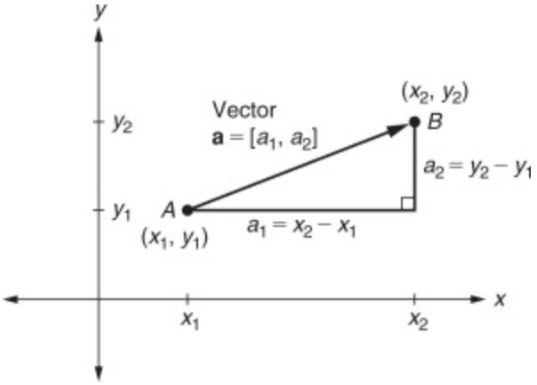


**Figure 7. Euclidean distance between 2 points.**[14]

Euclidean distance is a measure of the straight-line distance between two points in a multidimensional space. As shown in Figure 7, Euclidean distance calculates the direct distance between points regardless of their dimensions. This metric is useful for determining the absolute difference in features between data points (see Fig. 7). The formula is as follows:

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$$

where A and B are two vectors. A smaller Euclidean distance indicates a higher similarity between the vectors' features, while a larger distance indicates greater dissimilarity. Although Euclidean distance has limitations in high-dimensional data, it remains a commonly used metric in many machine learning applications. [15]

By combining cosine similarity and Euclidean distance, we can more comprehensively assess the correlation between images and summaries, thereby improving the performance of summary generation models.

## 5. Result

In this section, we present and analyze the evaluation results of our multimodal summarization method using ROUGE metrics, cosine similarity, and Euclidean distance.

### 5.1 Example of output

The following example illustrates the input text, the generated output summary, and the corresponding image selected by our multimodal summarization model.

Input1:

*‘A security alert has closed Liverpool John Lennon airport , with around 1,000 passengers being held on their flights or evacuated from the terminal .A suspect package was reported to airport staff around 6pm and in consultation with Merseyside Police it was decided to evacuate the area and call in the army bomb disposal unit .The area was declared safe and the terminal was reopened shortly before 9pm .According to an airport spokesman , passengers on around a dozen aircraft were inconvenienced by the incident .Passenger Pauline Cox complained about the lack of information . She tweeted : Security alert at Liverpool airport so we ca n't get off our plane . Anyone know anything ? 'An airport statement said : ` Earlier this evening a suspicious package was found at the airport . As a precautionary measure the terminal was evacuated and the Army Bomb Disposal Team were called to undertake further examination . ` Following this , the area was declared safe and passengers are now returning to the terminal . ` Operations are resuming as normal , though there may be some disruption as a result .’*

Ouput Summary1:

*‘A suspect package was spotted at Liverpool John Lennon airport, The airport along with Police decided to evacuate the area, An army bomb disposal unit attended the scene and declared it safe before 9pm.’*

Selected Image:

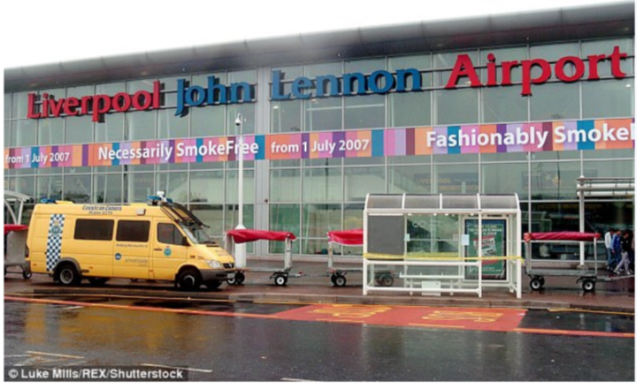


**Figure 8: Example1 of Output**

Input2:

*‘Chelsea defender David Luiz is targeting more success after winning the Premier League in his first season since returning to England . The Brazil centre-back was often cast as a figure of ridicule in his first spell at Stamford Bridge , but he silenced his critics with stellar performances at the heart of a three-man defence under Antonio Conte . Luiz , who re-joined the Blues for # 34million from Paris Saint-Germain last summer , will team up with new signing Tiemoue Bakayoko next week after the midfielder completed his switch from Monaco . And the centre-back believes his side can build on the success of last season with yet more silverware in the coming 10 months . When you play for a big club you have to have big ambitions , he told the club 's official website . Chelsea play all the competitions and we try to win all the competitions , so it 's my ambition too . We will try to win everything . Luiz is determined for the club not to slip backwards in the manner they did during their last title defence , and having set the bar high , the 30-year-old expects a rigorous pre-season training period ahead . He added : It 's difficult to be back in training . It 's a bit hard in pre-season , like always . But we have to prepare ourselves in the best way because now is the time to get the oxygen for when we need it during the season , so now is the time to work hard and rest well . We work hard at Cobham -LSB- the club 's training ground -RSB- and then at home we need to be careful with the food , with the discipline to sleep . I have stuff I do at home to recover well also , so it 's part of our job .It 's not just the five or six hours we are here at Cobham , it 's important to act like a professional every day of the year . ’*

Ouput Summary2:

*‘David Luiz believes Chelsea players should match the ambitions of the club, Brazil centre-half believes the Blues can build on the success of last season, Luiz expects a rigorous pre-season training period ahead of the big kick-off, 30-year-old will team up with new # 40m signing Tiemoue Bakayoko next week.’*

Selected Image:

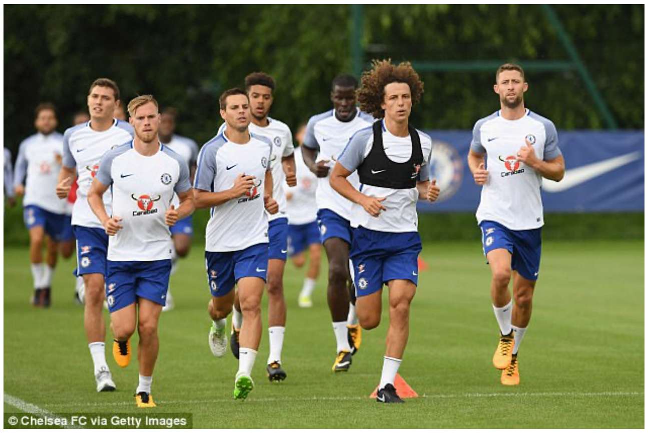


**Figure 9: Example2 of Output**

This example demonstrates how the input text is processed by the model to generate a concise summary and select a relevant image. The output summary captures the key details of the incident, ensuring that essential information is conveyed succinctly. The selected image of Liverpool John Lennon Airport visually complements the textual summary, providing a holistic representation of the event.

The effectiveness of this approach is evident in how well the model integrates textual and visual data to produce comprehensive and informative summaries. This example highlights the model's capability to handle real-world scenarios, offering a practical solution for multimodal summarization tasks.

## 5.2 Rouge metrics

The ROUGE metrics provide a comprehensive evaluation of the quality of the generated summaries by measuring the overlap with reference summaries at different granularities. Table 3 and Table 4 summarize the results of our experiments for the baseline and the BART models.

**Table 3. ROUGE Scores for Baseline and BART Models**

| | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| Baseline | 28.77 | 10.51 | 27.43 |
| BART | 32.53 | 21.19 | 32.42 |

For the Baseline model：

- ROUGE-1：The score of 28.77 for the baseline indicates that approximately 28.77% of the words in the generated summaries match the words in the reference summaries, suggesting a relatively high word-level match.
- ROUGE-2：The score of 10.51 for the baseline indicates that approximately 10.51% of the bigrams (two consecutive words) in the generated summaries match those in the reference summaries, suggesting a lower bigram-level match.
- ROUGE-L：The score of 27.43 for the baseline indicates that approximately 27.43% of the longest common subsequences in the generated summaries match those in the reference summaries, indicating a good match in terms of sentence structure and sequence.

For the BART model：

- ROUGE-1：The score of 32.53 indicates an improvement, with a higher percentage of word-level matches compared to the baseline.
- ROUGE-2：The score of 21.19 shows a significant improvement in bigram-level matching, suggesting enhanced fluency and coherence.
- ROUGE-L：The score of 32.42 reflects better preservation of sentence structure and sequence, indicating a more effective summarization process.

The BART model with multimodal input (text + image) shows significant improvement over the baseline model in all ROUGE metrics.

**Table 4. ROUGE Scores for BART (Text Only) and BART (Text + Image) Models**

| | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| BART(text) | 31.47 | 19.80 | 31.33 |
| BART(text+image) | 32.53 | 21.19 | 32.42 |

The BART model with only text input shows notable performance, with improvements in all three ROUGE metrics compared to the baseline. However, the addition of images further enhances the performance, indicating the benefits of incorporating visual information into the summarization process.

For the BART (Text) model：

- ROUGE-1：The score of 31.47 indicates a strong word-level match, showing the model's ability to capture key terms effectively from the text alone.
- ROUGE-2：The score of 19.80 demonstrates improved bigram-level coherence, suggesting that the text-only model can understand and reproduce sequences of words more accurately.
- ROUGE-L：The score of 31.33 reflects better preservation of sentence structure, indicating that the text-only model maintains a good overall structure of the original content.

For the BART (Text + Image) model：

- ROUGE-1：The score of 32.53 shows a significant improvement over the text-only model, indicating a higher word-level match. This suggests that visual context helps in identifying and including more relevant keywords.
- ROUGE-2：The score of 21.19 reflects a substantial increase in bigram-level matching, indicating that visual information aids in capturing the flow and coherence of the original content more effectively.
- ROUGE-L：The score of 32.42 indicates improved sentence structure and sequence matching, highlighting the model's enhanced ability to preserve the overall structure and order of the original content.

These results demonstrate that integrating visual information into the summarization process enhances the ability to capture and convey the content more accurately and coherently. The substantial improvements in ROUGE-2 and ROUGE-L scores highlight the importance of considering both textual and visual contexts to generate more comprehensive and contextually rich summaries.

By leveraging multimodal inputs, our approach effectively addresses the limitations of traditional text-only summarization methods, offering a more holistic understanding of the content and improving user experience with richer, more informative summaries. The BART model with multimodal input provides a balanced enhancement across all levels of matching, from individual words to the overall structure, confirming the benefits of a multimodal approach in summarization tasks.

## 5.3 Cosine Similarity

Cosine similarity measures the similarity between the feature vectors of the generated summaries and the reference summaries. Table 5 summarizes the results:

**Table 5. Cosine Similarity Scores**

| | Cosine Similarity |
|---|---|
| Baseline | 0.039 |
| BART | 0.04 |

The slightly higher cosine similarity score for the BART model indicates that the generated summaries are more similar to the reference summaries in terms of vector orientation. This suggests that the BART model, especially when incorporating multimodal inputs, can better capture the semantic orientation of the summaries, resulting in more semantically aligned outputs.

- Baseline: The cosine similarity score of 0.039 shows a moderate level of similarity between the generated and reference summaries in terms of their directional alignment in the vector space.
- BART: The score of 0.04, although slightly higher, indicates that the multimodal approach further enhances the semantic similarity.This improvement, though incremental, highlights the added value of integrating visual features into the text summarization process, resulting in more coherent and contextually relevant summaries.

The average score of BART model(0.04) is relatively low, indicating that the similarity between the text descriptions and the corresponding images is low in most cases. This could mean:(a)Feature Extraction Issues: The current methods for extracting features from text or images may not be capturing the key information effectively enough to link the text with the images.(b)Model Selection: The models currently in use (such as ResNet and BERT) might not be fully suited for this specific task or may require further fine-tuning and optimization.(c)Data Discrepancy: There may be inherent differences between the content of the text and the images, leading to a naturally low similarity score even for the most relevant text-image pairs.

The improvement in cosine similarity underscores the importance of considering both text and image data to capture the nuanced meanings and contextual relationships within the content, thereby generating summaries that are not only accurate but also semantically rich.

## 5.4 Euclidean Distance

Euclidean distance measures the difference between the feature vectors of the generated summaries and the reference summaries. Table 6 summarizes the results:

**Table 6. Euclidean Distance Scores**

| | Euclidean Distance |
|---|---|
| Baseline | 53.33 |
| BART | 50.94 |

The lower Euclidean distance for the BART model suggests that the generated summaries are closer to the reference summaries in feature space, indicating better performance.

- Baseline: The Euclidean distance of 53.33 indicates a higher degree of difference between the generated and reference summaries, reflecting the limitations of the baseline model in accurately capturing the essential features of the content.
- BART: The reduced Euclidean distance of 50.94 for the multimodal approach signifies a closer alignment with the reference summaries. This reduction demonstrates that incorporating visual data helps in producing summaries that are not only contextually relevant but also structurally similar to the reference content.

The average value of about BART model(50.943546),which is a high distance, suggesting a significant distance between the feature spaces of the text descriptions and the corresponding images. This could be due to several factors: (a)Feature Mismatch: The current methods for feature extraction might not effectively capture common or relevant features between the text and images.(b)Data Discrepancy: If there are natural mismatches between the text and image content within the dataset, even the most relevant text-image pairs might have large Euclidean distances.(c)Normalization and Alignment: Differences in the scale of the features extracted from text and images might cause the distances to appear larger. Appropriate normalization or scaling of features might reduce this distance, indicating a better match.

The combination of improved cosine similarity and reduced Euclidean distance highlights the effectiveness of our multimodal approach in producing high-quality summaries. These summaries closely resemble the reference summaries in both semantic and structural aspects, confirming the advantages of leveraging multimodal inputs in summarization tasks. The analysis of these metrics underscores the potential of multimodal summarization techniques in creating more accurate, coherent, and contextually enriched summaries.

By understanding these potential causes, we can take steps to improve the feature extraction process, choose more suitable models, and address any inherent discrepancies in the data to achieve better alignment and similarity scores.

# 6. Conclusion

## 6.1 Conclusion

In this study, we have presented a novel approach to multimodal summarization that integrates textual and visual data to generate comprehensive and coherent summaries. Our method leverages pretrained models like ResNet50, BART, and VisualBERT to dynamically fuse features from both modalities, resulting in improved performance over traditional text-only summarization methods. The empirical evaluations using ROUGE metrics, cosine similarity, and Euclidean distance demonstrate the effectiveness of our approach in producing high-quality summaries that closely resemble reference summaries in both semantic and structural aspects. The integration of visual information significantly enhances the overall quality and informativeness of the summaries, providing a more holistic understanding of the content. Future research will focus on optimizing the models for larger datasets and refining the feature extraction and integration processes to further improve the performance and applicability of multimodal summarization techniques.

## 6.2 Limitation

The primary limitation of our study is the relatively small size of the dataset used for training and evaluation. This constraint is mainly due to the limited computational resources available, which restricts our ability to process larger datasets. As a result, the findings and improvements demonstrated in this study may not fully generalize to larger and more diverse datasets.

Specific limitations include:

- Hardware Limitations：The limited computational resources prevented us from using larger datasets, which could have provided more comprehensive training and evaluation, potentially leading to better performance.
- Feature Extraction Issues：Current methods for extracting features from text and images might not capture the key information effectively, which could affect the alignment and integration of multimodal data.
- Model Selection：The pretrained models used (such as ResNet and BERT) might not be fully optimized for this specific task, indicating a need for further fine-tuning and optimization.
- Data Discrepancy：There may be inherent differences between the text and images, leading to lower similarity scores even for the most relevant text-image pairs.
- Dataset Quality：Some images in the dataset may not be relevant or accurately representative of the textual content, suggesting that a higher quality dataset could yield better results.

## 6.3 Future Work

Future work will focus on addressing these limitations to further improve the performance and applicability of multimodal summarization techniques.

Future directions include:

- Expanding the Dataset: Using larger and more diverse datasets to train and evaluate the models. This will help in understanding how well the model generalizes to different types of content and improves performance. Improving
- Feature Extraction: Refining the methods for extracting and integrating features from both text and images to enhance alignment and coherence. This includes exploring more advanced feature extraction techniques that can better capture the essential information from both modalities.
- Exploring More Models: Running the dataset through various frameworks to compare performance and determine the most effective approach. This could involve testing newer models or different combinations of existing models to find the optimal configuration.
- Domain-Specific Analysis: Evaluating the models on specific domains to assess their effectiveness across different types of content. This will help in understanding how the model performs in specialized areas and whether domain-specific training can further enhance performance.
- Optimizing Computational Resources: Leveraging advanced hardware and optimized algorithms to handle larger datasets more efficiently. This includes using more powerful GPUs and optimizing the code to reduce computational overhead and processing time.

By addressing these limitations and refining our approach, we aim to advance the field of multimodal summarization and develop more robust and scalable solutions for real-world applications. In summarizing our research findings and future work directions, expanding datasets, improving feature extraction, exploring more models, conducting domain-specific analysis, and optimizing computational resources will enhance the performance and applicability of multimodal summarization techniques. Through these improvements, we hope to overcome current limitations, enhance the model's generalization ability and performance in various application scenarios, and ultimately achieve higher-quality and more informative multimodal summaries. These future directions will drive further development in multimodal technology, bringing more innovation and applications to the fields of natural language processing and computer vision.

## 7. Reference

1. Yadav D, Desai J, Yadav AK. Automatic Text Summarization Methods: A Comprehensive Review. Published online March 3, 2022. doi:10.48550/arXiv.2204.01849

2. La Quatra M, Cagliero L. BART-IT: An Efficient Sequence-to-Sequence Model for Italian Text Summarization. *Future Internet*. 2023;15(1):15. doi:10.3390/fi15010015

3. Akshaya A, Kambli M. Text Summarization Using Deep Learning : An Analysis of sequence-to-sequence models and transformers. Published online January 1, 2022:176-187. doi:10.46501/IJMTST0807026

4. Rahman MM, Siddiqui FH. An Optimized Abstractive Text Summarization Model Using Peephole Convolutional LSTM. *Symmetry*. 2019;11(10):1290. doi:10.3390/sym11101290

5. Zhu J, Zhou Y, Zhang J, Li H, Zong C, Li C. Multimodal Summarization with Guidance of Multimodal Reference. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020;34(05):9749-9756. doi:10.1609/aaai.v34i05.6525

6. Zhang Z, Meng X, Wang Y, Jiang X, Liu Q, Yang Z. UniMS: A Unified Framework for Multimodal Summarization with Knowledge Distillation. Published online February 15, 2022. Accessed June 4, 2024. http://arxiv.org/abs/2109.05812

7. Atri YK, Pramanick S, Goyal V, Chakraborty T. See, hear, read: Leveraging multimodality with guided attention for abstractive text summarization. *Knowledge-Based Systems*. 2021;227:107152. doi:10.1016/j.knosys.2021.107152

8. Zhu J, Li H, Liu T, Zhou Y, Zhang J, Zong C. MSMO: Multimodal Summarization with Multimodal Output.

9. Lin CY. ROUGE: A Package for Automatic Evaluation of Summaries. In: ; 2004:74-81. Accessed June 5, 2024. https://aclanthology.org/W04-1013

10. Cao M. A Survey on Neural Abstractive Summarization Methods and Factual Consistency of Summarization. Published online April 20, 2022. doi:10.48550/arXiv.2204.09519

11. Cosine Similarity. Accessed June 5, 2024. https://www.learndatasci.com/glossary/cosine-similarity/

12. Sohangir S, Wang D. Improved sqrt-cosine similarity measurement. *Journal of Big Data*. 2017;4(1):25. doi:10.1186/s40537-017-0083-6

13. Is Cosine-Similarity of Embeddings Really About Similarity? ar5iv. Accessed June 5, 2024. https://ar5iv.labs.arxiv.org/html/2403.05440

14. Distance formula - an overview | ScienceDirect Topics. Accessed June 5, 2024. https://www.sciencedirect.com/topics/mathematics/distance-formula

15. Ifeanyi-Reuben N, Ugwu C, Nwachukwu E. Comparative Analysis of N-gram Text Representation on Igbo Text Document Similarity. *International Journal of Applied Information Systems*. 2017;12:1-7. doi:10.5120/ijais2017451724

## 8. Work Division

The authors confirm and agree that their work description and contribution is correct.

| **Name** | **Work description** | **Contribution** |
|---|---|---|
| 林嫈容<br>112423006 | 1. Experiment Design<br>2. New Method Coding | 33% |
| 方子樽<br>112423008 | 1. Experiment Design<br>2. Baseline Method Coding | 33% |
| 吳若瑜 | 1. Experiment Design | 33% |

|  |  |  |
|---|---|---|
| 112423024 | 2. Drafting Document |  |