



## 저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

M.S. THESIS

A Multimodal, Multispeaker Abstractive  
Summarization Dataset of Discussion  
Threads

멀티모달 다중화자 토의 글타래의 추상적 요약을 위한  
데이터셋

August 2023

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING  
COLLEGE OF ENGINEERING  
SEOUL NATIONAL UNIVERSITY

Keighley Shea Overbay

A Multimodal, Multispeaker Abstractive  
Summarization Dataset of Discussion Threads

멀티모달 다중화자 토의 글타래의 추상적 요약을 위한  
데이터셋

지도교수 김 건 희

이 논문을 공학석사학위논문으로 제출함

2023 년 05 월

서울대학교 대학원

컴퓨터공학부

키일리 쉐이 오버베이

키일리 쉐이 오버베이의 공학석사 학위논문을 인준함

2023 년 07 월

위 원 장	신 영 길	(인)
부위원장	김 건 희	(인)
위 원	이 상 구	(인)

# Abstract

With recent advances in artificial intelligence and large language models, automatic summarization of documents such as news articles, dialogues, and online discussions has been improving rapidly. However, much of these improvements have been limited to text-only summarization, and have not addressed that many online discussions are increasingly multimodal, consisting of not only text but also videos and images. While the growing number of multimodal online discussions necessitates automatic summarization to save time and reduce content overload, existing summarization datasets do not sufficiently cover this domain. To address this, we present **mREDDITSUM**, the first multimodal discussion summarization dataset. It consists of 3,033 discussion threads where a post solicits advice regarding an issue described with an image and text, and respective comments express diverse opinions. We annotate each thread with a human-written summary that captures both the essential information from the text, as well as the details available only in the image. Experiments show that popular summarization models—GPT-3.5, BART, and T5—consistently improve in performance when visual information is incorporated. We also introduce a novel method, cluster-based multi-stage summarization, that outperforms existing baselines and serves as a competitive baseline for future work.

**Keywords:** Deep Learning, Natural Language Processing, Computer Vision, Abstractive Summarization, Multimodal Summarization, Dataset Annotation

**Student Number:** 2021-29898

# Contents

<b>Abstract</b>	<b>i</b>
<b>Chapter 1 Introduction</b>	<b>1</b>
1.1 Purpose of Research . . . . .	1
1.2 Related Work . . . . .	4
1.2.1 Discussion Thread Summarization . . . . .	5
1.2.2 Multimodal Summarization . . . . .	6
<b>Chapter 2 The mREDDITSUM Dataset</b>	<b>8</b>
2.1 Data Selection . . . . .	8
2.2 Data Annotation . . . . .	10
2.2.1 Step 1: Original Post Summarization . . . . .	11
2.2.2 Step 2: Comment Cluster Summarization . . . . .	11
2.2.3 Step 3: Summary Synthesis . . . . .	12
2.3 Dataset Analyses . . . . .	12
2.3.1 Statistics . . . . .	12
2.3.2 Abstractiveness . . . . .	13
2.3.3 Relatedness between Text and Images . . . . .	13

<b>Chapter 3</b>	<b>Models and Experiments</b>	<b>15</b>
3.1	Task Definitions . . . . .	15
3.2	Evaluation Metrics . . . . .	16
3.2.1	ROUGE . . . . .	16
3.2.2	BertScore . . . . .	16
3.3	Models . . . . .	16
3.3.1	Baseline Models . . . . .	17
3.3.2	Cluster-based Multi-stage Summarization . . . . .	19
3.4	Implementation Details . . . . .	19
<b>Chapter 4</b>	<b>Results and Analysis</b>	<b>22</b>
4.1	Experiment Results . . . . .	22
4.2	Qualitative Analysis . . . . .	24
4.3	Human Evaluation . . . . .	24
<b>Chapter 5</b>	<b>Conclusion</b>	<b>27</b>
<b>Appendix A</b>	<b>Annotation Interface</b>	<b>28</b>
<b>Appendix B</b>	<b>Additional Sample Data</b>	<b>31</b>
<b>Appendix C</b>	<b>Further Analyses</b>	<b>34</b>
C.0.1	Summarization based on the Length of Input Threads . .	34
C.0.2	Summarization per Subreddit . . . . .	37
<b>Acknowledgements</b>		<b>44</b>
<b>요약</b>		<b>45</b>

# List of Figures

Figure 3.1	An illustration of Cluster-based Multi-stage Summarization (CMS): (1) comments are first clustered by similarity, (2) each cluster is summarized in a sentence, and (3) the cluster-summaries are summarized. . . . .	20
Figure 4.1	Human evaluation results of randomly sampled summaries of CMS-T5-Imgcap and T5-ImgCap models. . . .	25
Figure A.1	An example of instructions given for Task 1: Original Post Summarization. . . . .	29
Figure A.2	An example of the Cluster Summarization task presented to workers on Amazon Mechanical Turk. . . . .	29
Figure A.3	An example of the Summary Editing task presented to workers on Amazon Mechanical Turk. . . . .	30
Figure C.1	The influence of the number of comments in the thread on summarization performance (ROUGE-1) on BART-based models measured on the test set. . . . .	35

Figure C.2	The influence of the number of comments in the thread on summarization performance (ROUGE-1) of T5-based models. The results are based on the test set. . . . .	35
Figure C.3	ROUGE scores obtained from our CMS-T5-ImgCap model on the test set, categorized by different subreddits. The number of input words is indicated in parentheses. . . .	36



# List of Tables

Table 1.1	An example from the MREDDITSUM dataset. Both the post, several viewpoints from the comments, and the overall thread are summarized along with important content only available the image (in green), or in both image and text (in blue). . . . .	2
Table 1.2	A comparison of MREDDITSUM and other summarization datasets. Among forum-based and multi-turn datasets, MREDDITSUM is the only multimodal dataset, and it has the highest summary length, number of turns, and number of speakers. Length is reported in the average number of words, and turns are the average number of each instance of a post, comment, or speaker change. Statistics are taken from the respective papers for AnswerSumm [1], ConvoSumm[2], SamSUM [3], CNN/DM [4], MSMO DailyMail [5], and How2 [6]. . . . .	4
Table 2.1	The subreddits used for data collection and the number of threads collected for each. . . . .	10

Table 2.2	Average statistics across the original post, comment clusters, and full thread structures of our dataset. . . . .	13
Table 2.3	A comparison of Extractive Oracle ROUGE scores on MREDDITSUM and related datasets. The lower the score, the more abstractive the summaries are. Results for related works are from the respective papers[1, 2]. . . . .	14
Table 4.1	Results for the summarization task on mRedditSum. Models with “-ImgCap” in the name incorporate image information via image caption, and “VG-”, via image embedding. Others are text-only models. Cluster-based multi-stage summarization (CMS) is our proposed method of processing discussions in three stages. . . . .	23
Table 4.2	Examples of summaries generated from various models. Across all models, hallucinations regarding the image (highlighted in red) are present; however, these are reduced with multimodal models that incorporate image-only information (highlighted in green). Our CMS models tended to include more relevant details (blue) while removing irrelevant comments (orange). . . . .	26
Table B.1	Another example from our dataset, from the <i>fashionadvice</i> subreddit. . . . .	32
Table B.2	Another example from our dataset, from the <i>designmyroom</i> subreddit. . . . .	33

# Chapter 1

## Introduction

### 1.1 Purpose of Research

With the increased popularity of online discussion forums like *Reddit*, discussion threads—each consisting of a post and comments—of various lengths have been quickly accumulating. It has thus become overwhelming for users to sift through the threads to find the information they seek, which in turn has led to the development of automated means for text-only discussion summarization [7, 2, 1].<sup>1</sup>

However, discussion threads are often multimodal, containing images in addition to text. This added modality cannot be ignored, as it plays a key role in the respective discussions. For example, in Figure 1.1, the image of the couch is essential for discussing which coffee table would go well with it. Yet, multimodal summarization has so far been limited to news and instructional domains [5, 8, 9, 10] that are not easily transferable to online discussions surrounding

---

<sup>1</sup>See Table 1.2 for an overview of summarization datasets.

<p><b>Post:</b></p>  <p>We got this couch for our living room and I need help finding the perfect coffee table to go with it. ...</p>	<p><b>Post Summary:</b></p> <p>The OP asked for help with finding the right coffee table shape to match their <b>brown leather sectional</b>.</p>
<p><b>Comments:</b></p> <p><b>User 1:</b> I would do a circular table.</p> <p><b>User 2:</b> Definitely round! There are a lot of sharp angles already ...</p> <p><b>User 3:</b> There's way too much furniture in this space, the ottoman has to go ...</p> <p><b>User 4:</b> I think you should look for a natural wood triangular shaped coffee table ...</p> <p><b>User 5:</b> You should get a rug. ...</p>	<p><b>Comment Summaries:</b></p> <p><b>C1,C2,C4:</b> Commenters suggest a differently shaped coffee table from the <b>square one in the picture</b>, round or triangular or hexagonal.</p> <p><b>C3:</b> A commenter suggests eliminating the ottoman as it takes up too much space.</p> <p><b>C5:</b> A commenter suggests adding a rug.</p>
<p><b>Full Summary:</b></p> <p>The OP asked for help with finding the right coffee table shape to match their <b>brown leather sectional</b>. Commenters suggested a differently shaped coffee table from the <b>square one</b> he has already, such as round, triangular, or hexagonal. A few commenters suggested eliminating the <b>ottoman</b>, as it is too big for the small space. Others suggested adding a rug.</p>	

Table 1.1: An example from the MREDDITSUM dataset. Both the post, several viewpoints from the comments, and the overall thread are summarized along with important content only available the image (in green), or in both image and text (in blue).

images. For many of these posts, the context provided by the image is crucial to understanding the ongoing discussion, and thus understanding the content of the image also becomes necessary for generating a high-quality summary.

To fill the gap, we tackle multimodal discussion summarization. In particular, we consider Reddit discussion threads in which the post solicits advice regarding an issue described with an image and text, and commenters offer opinions, as opposed to simple reactions or jokes. Here, the goal is to generate an abstractive summary faithfully capturing the information from the

post—both image and text—and comments. This task is especially challenging because along with the need to effectively process the multimodal input, a quality summary must provide good coverage of commenters’ varying perspectives and opinions without redundancy.

To facilitate research in this direction, we present the Multimodal Reddit Summarization (MREDDITSUM) dataset, consisting of 3,033 Reddit discussion threads—posts (text and image) and comments (text-only)—each accompanied by a human-written summary, as shown in Figure 1.1. To construct the dataset, we carefully selected subreddits with discussions surrounding an image, and collected summaries that not only summarize the text, but also make reference to relevant information present only in the image. (See Appendix B for additional examples.)

We also propose cluster-based multi-stage summarization (CMS), a novel method to summarize multimodal discussions. It processes discussions in three stages: (i) comments are first clustered by similarity, (ii) each cluster is summarized in a sentence, and (iii) the cluster-summaries are summarized.

Experiments show that CMS consistently outperforms popular large language models (LLMs) for summarization—GPT-3.5 [11], BART [12], and T5 [13]. Also, incorporating image information, either as a dense vector or in text caption, consistently boosts the performance. We plan to make the dataset and code public.

Our main contributions are as follows:

- We present MREDDITSUM, the first multimodal discussion summarization dataset with human-written summaries with essential information from both the text and the image.
- We propose cluster-based multi-stage summarization (CMS), a novel method

Dataset	Domain	# Docs	Doc Len	Sum Len	# Turns	# Speakers	Modality
mREDDITSUM(ours)	Forum	3,033	691.0	<b>91.0</b>	<b>22.6</b>	<b>15.59</b>	<i>t, i</i>
AnswerSumm	Forum	4,631	787.0	47.0	6.4	6.17	<i>t</i>
ConvoSumm <sub>reddit</sub>	Forum	500	641.0	65.0	7.88	*	<i>t</i>
SamSUM	Dialog	16,396	124.1	23.4	12.19	2.39	<i>t</i>
CNN/DM	News	286,817	766.0	53.0	1	1	<i>t</i>
MSMO Daily-Mail	News	314,581	722.7	55.0	1	1	<i>t, i</i>
How2	Video	79,114	291.0	33.0	1	1	<i>t, v</i>

\*: speaker info not provided / *t*: text / *i*: image / *v*: video

Table 1.2: A comparison of mREDDITSUM and other summarization datasets. Among forum-based and multi-turn datasets, mREDDITSUM is the only multi-modal dataset, and it has the highest summary length, number of turns, and number of speakers. Length is reported in the average number of words, and turns are the average number of each instance of a post, comment, or speaker change. Statistics are taken from the respective papers for AnswerSumm [1], ConvoSumm[2], SamSUM [3], CNN/DM [4], MSMO DailyMail [5], and How2 [6].

to summarize multimodal discussions outperforming competitive baselines like GPT-3.5, BART and T5, as well as their vision-guided variations.

## 1.2 Related Work

We highlight two main areas of related work in abstractive summarization: discussion thread summarization and multimodal summarization. We provide a comparison of related summarization datasets in Table 1.2, where statistics are taken from the respective papers for AnswerSumm [1], ConvoSumm[2], SamSUM [3], CNN/DM [4], MSMO DailyMail [5], and How2 [6]. Note that speaker info was not provided for ConvoSumm<sub>reddit</sub>, and CNN/DM, MSMO DailyMail, and How2 do not have any multi-turn or multi-speaker information.

### 1.2.1 Discussion Thread Summarization

Despite the prevalence of discussion threads online, it has traditionally been an understudied area for automatic summarization. This is likely due to the fact that until recently, most automatic summarization work has focused on extractive summarization. In extractive summarization, snippets of text are taken directly from the input and are used as a pseudo-summary of the document. One small extractive summarization dataset has been created [7]; however, it was not explored much further as extractive summarization is an unnatural choice for summarizing dialogues and discussions.

More recently, several abstractive summarization datasets have been proposed that offer a more fitting summary for a discussion. ConvoSumm [2] presented a dataset of 2000 summarized forum threads, 500 from each of 4 different domains including NYT articles, Reddit, StackExchange, and Email threads. AnswerSumm [1] is another dataset consisting of 4,631 question-answering discussion threads sourced from StackExchange. AnswerSumm shares the most similarities with our dataset, as they also summarize multi-speaker threads, and their annotation pipeline shares key similarities with ours. They also cluster the comments and summarize these groups before going through a final summary editing process, similar to our pipeline. The key differences between this dataset and ours is that AnswerSumm is only text-based with no images and operates in a different domain, as they are all question-answering threads curated from StackExchange. In contrast, our dataset includes both images and text, and focuses on Reddit threads where the images play a key role. Additionally, in our annotation pipeline we also summarize the original post and image as well, which to our knowledge has not been done in any other forum summarization dataset. This is useful because oftentimes the posts alone may

have unclear intent that may require context derived from the image or forum domain itself to understand.

Other related summarization datasets include other multi-turn datasets such as SamSUM [3], which consists of chat-like dialogues and human-annotated summaries, and EmailSum [14], which consists of work-related emails and both long and short reference summaries.

Overall, though there is a small variety of existing thread summarization datasets, they are all currently only text-based and none of these tackle both original post and thread summarization.

### **1.2.2 Multimodal Summarization**

Though other multimodal research areas such as VQA [15] and text-image pre-training [16, 17, 18] have been gaining attention in recent years, there only exist a small handful of works that address multimodal summarization. Generally speaking, multi-modal summarization aims to generate a summary that includes salient information from inputs with multiple modalities. Tasks such as Multimodal Summarization with Multimodal Outputs [5] take both a news article and image-caption pairs from that article and generates summaries that include both a textual summary as well as the most salient images from that article.

However, our task aims to generate a unimodal output—that is, a purely textual summary. This is similar to the multimodal summarization done on the How2 Dataset [8, 9], where a textual transcript of the video along with the video frames are generated into a text summary. [6] reported that incorporating the additional modality of the video frames into their summarization models showed improvement compared to text-only based models. Though this multimodal summarization task is the most similar to ours, there are some key



differences. The How2 dataset uses short video captions as pseudo-summaries, instead of detailed human-annotated summaries like we curate for MREDDIT-SUM. Additionally, our text is a rich multi-speaker discussion, rather than a transcript of audio. Finally, MREDDITSUM’s threads are specifically selected to include images where their information is necessarily included in the summary, whereas there is no such assurance for How2’s videos.

## Chapter 2

# The MREDDITSUM Dataset

In order to tackle multimodal discussion summarization, we curate a new dataset, the MREDDITSUM dataset. Here we discuss both the data selection process we used in determining the most useful target discussions, as well as the annotation process used for gathering high-quality human summaries.

### 2.1 Data Selection

To construct a meaningful multimodal discussion summarization dataset, we imposed three major criteria when selecting Reddit threads to be included in the dataset.

**Criterion 1** The discussion thread needs to contain an image. Since Reddit does not allow images embedded in comments in many subreddits, this means that the post itself needs to contain an image.

**Criterion 2** The discussion needs to be centered around an image in such a way that the information only available from the image plays a key role in the discussion. In some threads, an image may not provide any significant information, e.g. it is a favorite animal of the original poster. In such cases, simply summarizing the text is sufficient, and a multimodal model is unnecessary. On the other hand, threads with posts that present an image and a discussion topic regarding the image tend to result in discussions that can be sufficiently summarized only with the information available from the image. The latter type better suits our purpose.

**Criterion 3** The discussion needs to contain content that can be meaningfully summarized. Many Reddit threads that include images are meant to incite reactions from other users, or to be shared in a jocular manner that prompts commenters to make jokes. Though these are interesting in and of themselves, summarizing them proves difficult and not helpful: simply noting that jokes were told or that people were impressed does not create a meaningful summary. On the other hand, some threads clearly ask for advice or opinions, thereby eliciting diverse responses from a number of commenters. Summarizing these opinions along with the advice sought in the post would be helpful for readers to understand the gist of the threads. Again, this latter type better suits our purpose.

Given the aforementioned criteria, we identified 9 subreddits—presented in Table 2.1—that consist primarily of image-based posts where the original poster is soliciting advice or opinions about either clothing or interior design. We collected all threads from these subreddits with over 5 comments from years 2015-2022. Collection was done with RedCaps [19] API, modified to collect all comments from each thread. We additionally followed similar preprocessing

Subreddit	Category	# Threads
r/outfits	Clothes	161
r/fashionadvice	Clothes	529
r/plussizefashion	Clothes	19
r/handbags	Clothes	90
r/petitefashionadvice	Clothes	112
r/weddingdress	Clothes	108
r/designmyroom	Interior	1098
r/malelivingspace	Interior	642
r/femalelivingspace	Interior	258

Table 2.1: The subreddits used for data collection and the number of threads collected for each.

steps, removing all posts that contained NSFW content or images with faces. Additionally, we filtered the comments themselves to remove any comments with NSFW content, or comments posted by bots. All responses to these removed comments were also removed. We also replaced all URLs with [URL], and anonymized all authors.

## 2.2 Data Annotation

We then began annotating the data after selecting qualified workers from Amazon Mechanical Turk. We limited our workers to those from English-speaking countries with a HIT approval rate over 98%, with greater than 5000 HITs approved. For all tasks, workers were required to complete a qualification task where the results were manually checked for quality. After passing this task, workers were allowed to work on the main tasks where our data was collected. Any workers who were found to submit low-quality work had their qualification revoked. Additional detail on the annotation interface and instructions can be found in Appendix A. The annotation was conducted in a 3-step annotation pipeline as follows.

### 2.2.1 Step 1: Original Post Summarization

In the first step, we present annotators with the original post along with the image from that post. We ask the annotators to summarize in a single sentence the intent of the original poster, as well as the most relevant details from the image. We use this method because a post that simply reads “*blue or black?*” may only be comprehensible when paired with the image of blue and black heels next to a blue dress, and a true text-only summary should be comprehensible without the image. Our summary may then read “*The original poster asked if blue or black heels would match better with a strapless, knee-length blue dress.*”, thus eliminating the need to view the image to comprehend the question. In this way, all information necessary to understand the question should be self-contained within the summary, and annotators were instructed as such.

### 2.2.2 Step 2: Comment Cluster Summarization

For the second step, we first cluster the comments in order to identify groups of comments that share a similar opinion. We follow the method described in AnswerSUMM([1] in order to allow for clusters of varying sizes and number. We use a RoBERTa-based model fine-tuned for semantic similarity to get sentence embeddings of the top-level comments from each thread. We then use agglomerative clustering with average linkage, cosine distance, and a maximum distance of 0.5 to generate clusters of comments.

After clustering the comments, we then rank them according to their size and Reddit score. The Reddit score of a comment is defined as the number of upvotes minus the number of downvotes it has received. We take the sum of all Reddit scores of the top-level comments in a single cluster to assign a saliency-score to that cluster. We then take the top 5 clusters with the highest saliency-scores and use these for annotation. We do this to limit the size of the

summary and to help remove irrelevant comments, while encouraging larger clusters of comments with a similar sentiment.

We then take these groups of comments and present them to annotators along with the original post and image, and ask them to summarize within one or two sentences the main opinions present in that group of comments. We also encourage the annotators to reference objects or details from the image when necessary for summarization. For consistency, we also instruct the annotators to refer to the commenters as “*Commenters*” as opposed to people, users, or other words.

### **2.2.3 Step 3: Summary Synthesis**

For the final step, we concatenate the original post summary as well as the comment cluster summaries, in descending order of their saliency-scores, as defined above. We then present these summaries once more to annotators and ask them to edit them for fluency and readability. In particular, we encourage annotators to reduce repetitive wording when possible, add connectives between sentences, and to rearrange sentences so that related topics are next to each other and the overall summary reads as more natural. We also ensure all summaries are written entirely in the past-tense for consistency. After this step, the summary is complete.

## **2.3 Dataset Analyses**

### **2.3.1 Statistics**

The resulting dataset contains a total of 3,033 posts and summaries. We split these into a train, test, and validation set of sizes 2729, 152, and 152, respectively. We present further statistics in Table 1.2, where we compare with similar summarization datasets from a few different domains. The average summary

Structure	Document	Summary
Original Post	1.62 sents 18.87 words	1.07 sents 23.14 words
Comment Clusters	6.63 sents 85.05 words	1.34 sents 20.17 words
Full Thread	21.6 comments 37.41 sents 691 words	5.32 sents 91.0 words

Table 2.2: Average statistics across the original post, comment clusters, and full thread structures of our dataset.

length for MREDDITSUM is longer than other datasets; however, this is not surprising given the nature of summarizing varying opinions, of which there could be many. Additionally, we describe the structure-level statistics in Table 2.2; note that while the average length of the Original Post summary is longer than the document, this is due to the additional image description and context. For the full thread, the summary is 13.2% as long as the input on average, which is comparable to SamSUM’s 19% and How2’s 11.3%.

### 2.3.2 Abstractiveness

Extractive-Oracle ROUGE scores in Table 2.3 show that our dataset is similar in abstractiveness to other multi-turn datasets, and much more abstractive than DailyMail. Though scores are not available for MSMO, it is expected that the scores would be similar to DailyMail.

### 2.3.3 Relatedness between Text and Images

We also calculate the CLIPScore [20], a metric that measures the correlation between text and an image, to determine how grounded our summaries are to the images from each thread. Our summaries have an average CLIPScore of 74.62, the post summaries alone achieve 74.89, and the comment cluster summaries alone score 68.34. These suggest our summaries, especially the post

Dataset	Extractive Oracle ROUGE		
	R1	R2	RL
MREDDITSUM (ours)	37.43	13.33	33.57
AnswerSumm	40.05	18.45	35.70
ConvoSumm <sub>reddit</sub>	35.74	10.45	30.74
DailyMail	55.23	30.55	51.24

Table 2.3: A comparison of Extractive Oracle ROUGE scores on MREDDITSUM and related datasets. The lower the score, the more abstractive the summaries are. Results for related works are from the respective papers[1, 2].

summaries, are well-correlated with the images.



## Chapter 3

# Models and Experiments

### 3.1 Task Definitions

We consider the multimodal summarization task, where the input includes all original text and the image and the output is a text-only summary that describes both the document and image. The text includes the post and comments, and the goal is to accurately summarize both the original poster’s intent and commenters’ opinions. For this task, we format the text input as the following: "Original Post: Original Post", with "Image: Image Caption." appended for models that include image captions. We then additionally append the comments in the form "User 1: Comment 1. User 2: Comment 2. ...", where each username has been anonymized. Comments are listed in the order that they are scraped from Reddit in. The target output is the result of our final summary.

## 3.2 Evaluation Metrics

Following the standard metric for summarization evaluation, we use the ROUGE evaluation metrics for our baseline models, as well as BertScore. We describe them briefly as follows:

### 3.2.1 ROUGE

ROUGE [21] is widely used as an automatic evaluation metric for summarization. It measures the salience of system-generated summaries by comparing n-grams in the generated summary and reference summary. There are three common variants of the ROUGE score that we consider: ROUGE-1 (R1) measures unigram overlap, ROUGE-2 (R2) measures bigram overlap, and ROUGE-L (RL) determines the longest common subsequence between summaries. The `rouge`<sup>1</sup> package was used to compute the scores.

### 3.2.2 BertScore

BertScore [22] is another metric widely used to evaluate generated text. It computes a similarity score similar to other metrics, but instead of using exact n-gram matches, it computes a ‘soft’ token similarity score using contextual embeddings output from BERT. We use the `bert_score`<sup>2</sup> python package to compute BertScore, and use the default RoBERTa-large model and rescale with baseline.

## 3.3 Models

We evaluate several baseline models on MREDDITSUM, including those that use only text information as well as those that use image information, either in the

---

<sup>1</sup><https://github.com/pltrdy/rouge>

<sup>2</sup>[https://github.com/Tiiiger/bert\\_score](https://github.com/Tiiiger/bert_score)

form of image embeddings or captions.

### **3.3.1 Baseline Models**

We consider three text-only baseline models: GPT-3.5 (zero-shot), BART and T5 (fine-tuned), as well as their extensions to make use of image information, either as image captions or embeddings.

#### **Extractive Baselines (Lead-1, Lead-Comment, Ext-Oracle)**

We include several extractive baselines for comparison of extractive performance. Lead-1 uses the first sentence from the document as the summary, and Lead-Comment uses the leading top 5 comments from the thread. Ext-Oracle extracts passages from the document to achieve the maximum possible ROUGE score, and thus is the highest possible performance from an extractive model.

#### **Text-only Baselines (GPT-3.5, BART, T5)**

GPT-3.5 [23] is an LLM that has shown excellent zero-shot performance in summarization tasks [24, 25]. We use the largest model, text-davinci-003, through the OpenAI API, with the prompt "Summarize what the original post was asking about and the general opinions of the commenters.", which is determined empirically to perform well and closely mimic the instructions given to annotators. We also evaluate two finetuned models, BART-base [12] and T5-base [13], which are high-performing LLMs with good summarization abilities. We pre-train them on the CNN/DailyMail [4] summarization dataset before fine-tuning it for our task.

### **Extensions with Image Captioning (GPT-3.5-ImgCap, BART-ImgCap, T5-ImgCap)**

We extend the text-only baselines to incorporate visual information through the use of an image caption, denoted as GPT3.5-ImgCap, BART-ImgCap, and T5-ImgCap, respectively. They take advantage of powerful LLMs without large amounts of multimodal training to understand visual features. For image captions, we use the BLIP2 model [18] trained on COCO image captions [26] and generate multiple image captions for each image using nucleus sampling. Since a more detailed and grounded image caption that describes concrete objects is best for this task, we use an image-grounding model, GLIP [27], to score each caption by grounding it with the image, and calculate how many image-text grounded pairs are above a threshold of 0.7. We then select the image caption with the highest score and append the caption to the input after the original post. We then fine-tune BART-ImgCap and T5-ImgCap as described above; for GPT3.5-ImgCap, we use the caption-appended prompt.

### **Extensions with Vision-Guidance (VG-BART, VG-T5)**

Vision-Guided BART and T5 are presented in [6] for multimodal summarization. They include additional visual layers that receive video embeddings as input, and show state-of-the-art performance in multimodal summarization for the How2 dataset. We modify the original models by instead using 768-D ViT-base [28] image embeddings as input, as they have shown excellent performance as an image backbone. We use cross-modal dot product attention with a forget-gate and image transformer, as this version performed best in our experiments. We use the same T5-base and BART-base pretrained on CNN/DM to initialize the encoder and decoder. For VG-BART, we pretrain the visual layers using the COCO image captions before fine-tuning on our dataset; VG-T5 shows no

performance increase from visual pretraining, so we initialize its layers from scratch.

### 3.3.2 Cluster-based Multi-stage Summarization

One challenge in summarizing discussions is that they can be very long. To confirm that this is causing an issue, we conduct a preliminary experiment on the fine-tuned BART model by comparing the results of two different test subsets: the *long* subset with more than 22 turns and the *short* subset with less than or equal to 22 turns. The performance on the long subset is noticeably worse than that on the short subset, lower by 4.95 ROUGE-1 and 6.1 BertScore.

To effectively handle this challenge, we present a novel method named **cluster-based multi-stage summarization (CMS)**, consisting of three stages (See Figure 3.1):

1. **Comment Clustering.** Similar comments are clustered using RoBERTa sentence embedding and agglomerative clustering.
2. **Cluster Summarization.** Each cluster is summarized in about a sentence using an LLM with image captioning, or a vision-guided LLM, such as VG-BART or VG-T5.
3. **Cluster-summary Summarization.** The cluster summaries are concatenated and further reduced into a coherent summary using a separate model, which is either an LLM with image captioning or a vision-guided LLM.

## 3.4 Implementation Details

The fine-tuned models are trained for 50 epochs each on a single Titan X GPU for BART models, and a Titan RTX for the larger T5 models. We use a batch

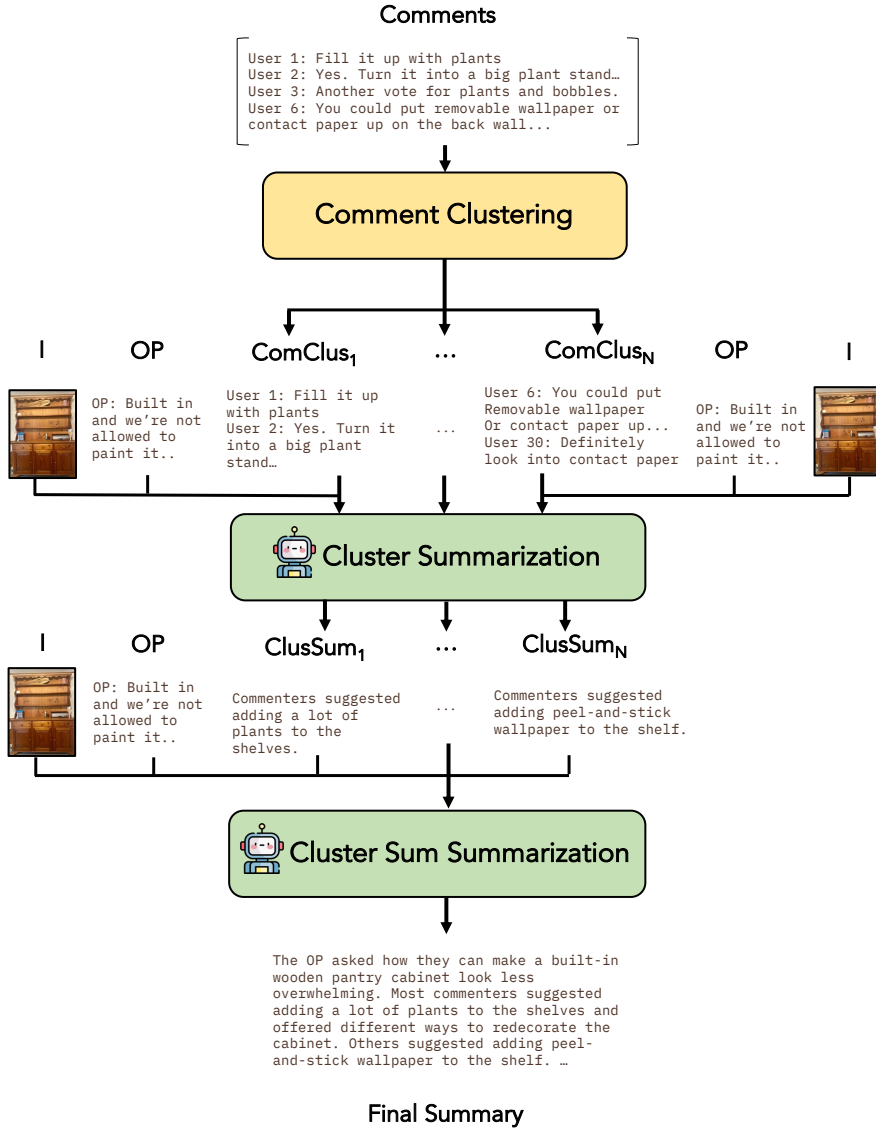


Figure 3.1: An illustration of Cluster-based Multi-stage Summarization (CMS): (1) comments are first clustered by similarity, (2) each cluster is summarized in a sentence, and (3) the cluster-summaries are summarized.

size of 4, and following [6, 13, 29], we use learning rates  $6\text{e-}4$  and  $3\text{e-}5$  to fine-tune the pre-trained parts of model weights, and a learning rate of  $1.5\text{e-}4$  to train the newly added visual layers in VG-BART and VG-T5. The decoding process uses beam-search with a beam size of 5. The average training time for BART, T5, BART-Cap, and T5-Cap was approximately 5 hours; the average training time for VG-BART and VG-T5 was approximately 8 hours, with the additional visual layers adding about 100 million extra parameters to each model. We use the same training epochs, batch size, learning rates, and beam-search size for cluster-based multi-stage summarization. All results shown are an average of two runs.

# Chapter 4

## Results and Analysis

### 4.1 Experiment Results

Table 4.1 shows the results of all models evaluated across the test set. We see that our model, Cluster-based Multi-stage Summarization (CMS), outperformed baseline models for all metrics across both T5 and BART-based models. We believe this is due to our models’ ability to better handle the long length of input threads; see § C.0.1 for more detailed analysis. In general across all model types, models that contain image information through an image caption outperform those that only have access to text-information. This supports that our dataset requires multimodal understanding in order to perform well on the summarization task. Vision-Guided models using text embeddings showed mixed results, with a marginal or no improvement over text-only models; we believe this to be due to a limitation of these models to effectively incorporate image information. Though they show strong performance on the How2 summarization task [6], mRedditSum has longer input and summary length, images,



Model	R1	R2	RL	BertS
<i>Extractive</i>				
Lead-1	15.23	3.46	13.24	11.89
Lead-Comment	22.86	5.55	20.43	7.16
Ext-Oracle	36.52	11.95	31.42	16.71
<i>Zero-shot Prompting</i>				
GPT-3.5	34.29	9.10	30.39	30.15
GPT-3.5-ImgCap	<b>34.59</b>	<b>9.41</b>	<b>30.59</b>	<b>31.07</b>
<i>Fine-tuned</i>				
BART	44.33	18.4	41.71	41.61
VG-BART	44.97	18.75	42.29	40.85
BART-ImgCap	44.91	18.54	42.12	41.34
CMS-VG-BART (ours)	45.13	18.81	42.56	42.13
CMS-BART-ImgCap (ours)	<b>45.55</b>	<b>19.28</b>	<b>42.87</b>	<b>43.89</b>
T5	45.29	18.97	42.4	42.32
VG-T5	45.58	18.94	42.75	42.3
T5-ImgCap	45.61	18.97	42.63	42.59
CMS-VG-T5 (ours)	45.71	19.21	42.97	42.72
CMS-T5-ImgCap (ours)	<b>47.29</b>	<b>19.86</b>	<b>44.13</b>	<b>44.74</b>

Table 4.1: Results for the summarization task on mRedditSum. Models with “-ImgCap” in the name incorporate image information via image caption, and “VG-”, via image embedding. Others are text-only models. Cluster-based multi-stage summarization (CMS) is our proposed method of processing discussions in three stages.

and fewer documents, likely contributing to the performance differences. Additionally, we note that T5 models show the best performance, followed by BART models and GPT3.5 models. For GPT3.5 models, we note that the low scores are likely due to inconsistencies in summary format, length, and detail, due to the zero-shot setting, but still receive relatively reasonable BertScore scores.

We provide further analyses on the effect of input length and subreddit category on performance in § C.

## 4.2 Qualitative Analysis

In addition to our automatic evaluation, we check the test results manually for qualitative analysis. Several results can be found in Table 4.2. The primary advantage of our method, CMS, is that it has a greater coverage of relevant opinions compared to the baseline models. It is better able to filter out irrelevant or strange comments, while keeping the important opinions and including ones that are presented late in the thread.

We also find that all models, even those incorporating image information, are still prone to hallucinations of what is in the image. These include incorrect descriptions of object color and style, as well as describing objects that are not present in the image at all. Though our multimodal models are generally better at incorporating visual details than text-only models, their power to interpret the image seems still limited; we believe this to be due to potential undertraining of the text-vision fusion layers in the VG models, and the limitations of image caption models.

Thus, while our CMS model can overcome one weakness of the baseline multimodal summarization models, we still believe there to be significant room for improvement in the field of multimodal models, and hope that MREDDITSUM can help facilitate such research.

## 4.3 Human Evaluation

We additionally perform human evaluation studies via AMT to compare the summaries generated from CMS-T5-ImgCap (ours) versus the baseline T5-ImgCap model. Based on similar works such as [30], we use three metrics to measure the summary quality: fluency, faithfulness, and overall quality. **fluency** measures which is more naturally written, **faithfulness** measures how truthful

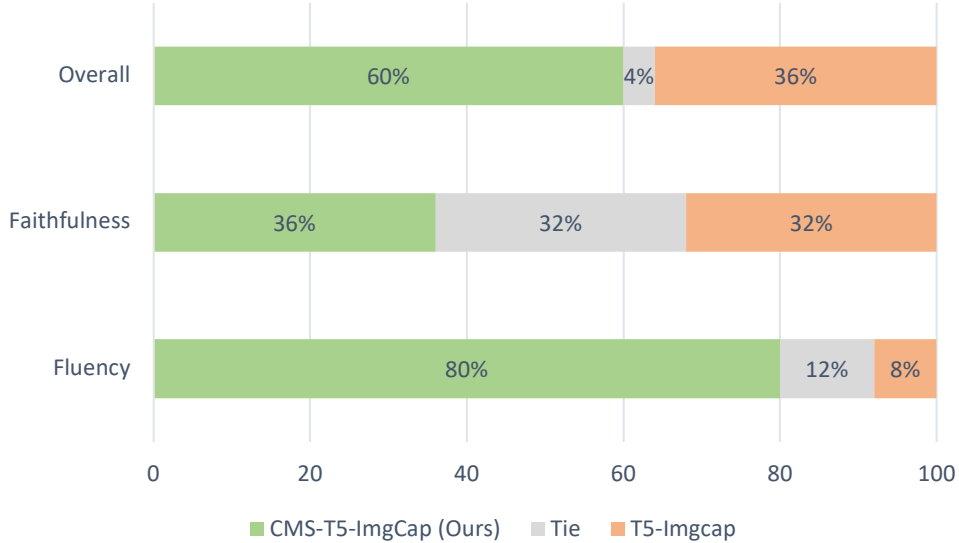


Figure 4.1: Human evaluation results of randomly sampled summaries of CMS-T5-Imgcap and T5-ImgCap models.

the summary is to the document, and the **overall quality** represents general user preference. We randomly sample 25 datapoints from the test set and receive 3 annotations per sample. We note that this limited number of datapoints is due to the fact that this evaluation task is highly challenging for human annotators, given that the input, including the original post, threads, and image, is long and complex.

Figure 4.1 shows the majority vote results that our summaries are overall more preferable in terms of fluency and overall quality, with similar performance for faithfulness. We believe this to be an indicator of our model’s better ability to effectively summarize the thread, offering more fluent and higher-quality summaries. The similar faithfulness scores are likely due to our method sharing the same base model and image caption as the baseline, T5-ImgCap, granting a similar ability to incorporate correct image and text information.



Image	Thread	T5	T5-ImgCap	CMS-T5-ImgCap
	Original Post: Looking for ideas on how to style this necklace User 1: This gives me Ms. Frizzle vibes. Maybe embrace that? User 2: Yes! Funky skirt that poofs out, fitted shirt! User 3: Fitted button up collard white shirt with high waisted shorts and flats OP: Ooh that's a great idea! Thank you! User 4: [URL] Link to necklace if anyone is interested...	The OP asked for ideas on how to style a <b>black and white floral</b> necklace. One commenter suggested wearing the necklace with a fitted shirt. Another suggested wearing a button up collared white shirt with high waisted shorts and flats.	The OP wanted advice on how to style their necklace, which has a <b>floral pattern</b> on it. One commenter suggested wearing the necklace with a fitted shirt. Another suggested wearing a button up collared white shirt with high waisted shorts and flats.	The OP asked for ideas on how to style their necklace, which is <b>white</b> with a <b>dinosaur</b> pattern on it. One commenter suggested wearing a <b>funky skirt</b> and fitted shirt with the necklace. Another commenter recommended a button up collared white shirt with high waisted shorts and flats. One user <b>shared a link</b> to OP's necklace.
	Original Post: [Update] Couch arrived and this is the rug I ended up choosing. Any coffee table and accent chair recommendations? User 1: Something with a warm wood color would look nice. Check out antique shops or local furniture shops to find something that has some character and is made of solid wood. User 2: Yes, round or oval would look nice....	The OP asked for advice on a coffee table and accent chair for their living room, which has a <b>brown leather armchair</b> and <b>tan leather</b> couch. Most commenters suggested a wood coffee table with a walnut finish and a solid white marble top. One commenter recommended a round or oval coffee table. Another suggested brown throw pillows and blankets to match the rug. <b>One user suggested OP get a non-shedding dog.</b>	The OP asked for advice on a coffee table and accent chair for their <b>blue couch</b> . Most commenters suggested a wood coffee table with a walnut finish and a solid white marble top. One commenter recommended a round or oval coffee table. <b>Another commenter suggested getting a non-shedding dog.</b>	The OP asked for help with picking out a coffee table and accent chair for their <b>blue couch</b> . Most commenters suggested getting a warm wood coffee table. <b>Others suggested a brown leather armchair or cream colored accent chair.</b> One commenter suggested getting throw pillows and blankets to match the rug. Another commenter asked where the rug was from, and the OP said it was from Apt2B.

Table 4.2: Examples of summaries generated from various models. Across all models, hallucinations regarding the image (highlighted in red) are present; however, these are reduced with multimodal models that incorporate image-only information (highlighted in green). Our CMS models tended to include more relevant details (blue) while removing irrelevant comments (orange).

## Chapter 5

# Conclusion

Online discussions are increasingly becoming multimodal, yet there are not sufficient resources for summarizing them. To this end, we presented MREDDITSUM, the first multimodal discussion summarization dataset containing 3,033 discussion threads and images with human-written summaries. Threads were carefully chosen so that the images play a key role in the respective threads, and summaries were written to capture this. Experiments showed that summarization models making use of visual information consistently outperform those that do not. Additionally, we introduced Cluster-based Multi-stage Summarization, which accounted for the structure of discussion thread data and outperformed baseline methods. We hope this dataset will help to facilitate active research on multimodal discussion summarization.

# Appendix A

## Annotation Interface

We listed a total of 3 tasks on Amazon Mechanical Turk for our data pipeline. We informed all annotators that this data would be used to help in summarizing Reddit threads, and asked them to agree with the Reddit Terms of Use before participating and notified them that participating in the HIT constituted acceptance of these terms of use.

We provided annotators with detailed instructions of the task and several acceptable and unacceptable examples to help them perform the task. In Figure 2, we show the instructions provided for Task 1; similar instructions were used in the other two tasks. Additionally, we show the annotation interface used for Tasks 2 and 3 in Figures 3 and 4.

INSTRUCTIONS


**Summarizing Posts + Images:**  
 You are to summarize what kind of advice or opinions a Reddit post is asking for, along with how it is related to the given image. Ideally, the summary should be a single sentence that contains all necessary information from both the post and image. That is, someone should be able to understand the summary completely even without seeing the image.

**Be sure to:**

- Start your sentence with "The OP...".
- Write in third-person (no "I" or "you") past tense.
- Summarize in your own words, rather than copy-pasting the original post.
- Include relevant information and details from the image.
- Do **not** include personal opinions, or irrelevant details.

Figure A.1: An example of instructions given for Task 1: Original Post Summarization.

Post:



OP: "what would you do to cover the ugly breaker box in my bedroom? obviously it still needs to be accessible, but i keep my door closed a lot and it's such an eyesore. the wall is 58 inches wide and 10 feet high."

Comments:

Group 1:

- **User 4:** you could hang a large framed painting. something liked stretched canvas over frame. no glass. that way to can move it and not worry about glass breaking. some will even install a hinge on one side so it opens like a cabinet door.
- **User 6:** i think a painting or frame would be nice- you can hinge it if you want to give easier access. i feel like a gallery style wall with multiple pieces would be best otherwise your art would look off center and weirdly close to the corner, if that makes sense.
  - **OP:** makes total sense. the closeness to the edge of the wall is really what's throwing me off about it.
- **User 7:** hang something over it. a painting or some kind of rug.
- **User 14:** i would paint it the same color as the wall and hang 2-3 large frames on the wall. you could easily


Write a summary sentence. You may write multiple sentences if necessary for this group.

Enter your summary for group 1...

No summary sentence is necessary - these comments are redundant or irrelevant.

Figure A.2: An example of the Cluster Summarization task presented to workers on Amazon Mechanical Turk.

Post 1:



OP: my room for about a year while flight instructing. what can i do to improve without being too permanent?

**Original Summary 1:**

The OP wanted temporary solutions to improve their sparse bedroom with just a bed, dresser and fake plant. One user commented that the room looks like a dressing room for a rub-n-tug. One user recommended getting a rug. Another user suggested curtains which easy to put and add a lot of character and soften the place and a floor lamp. Another user suggested moving the bed so it is not directly under the window and adding a small nightstand to the room. The OP said the recommendation are great so far and is now planning. Another user suggested adding a life-size statue to make the place less lonely.

**Edited Summary 1:**

The OP wanted temporary solutions to improve their sparse bedroom with just a bed, dresser and fake plant. One user commented that the room looks like a dressing room for a rub-n-tug. One user recommended getting a rug. Another user suggested curtains which easy to put and add a lot of character and soften the place and a floor lamp. Another user suggested moving the bed so it is not directly under the window and adding a small nightstand to the room. The OP said the recommendation are great so far

Figure A.3: An example of the Summary Editing task presented to workers on Amazon Mechanical Turk.



## Appendix B

### Additional Sample Data

We show a few additional datapoints from the MREDDITSUM dataset in Table 6 and 7. Table 6 shows a datapoint from the fashion category, whereas Table 7 shows a datapoint from the interior design category.


<p><b>Image:</b></p> 
<p><b>Post Caption:</b> what could you pair these with?</p>
<p><b>Comments:</b>  <b>User 1:</b> Dressy black pants, colorful blouse, and blazer....  <b>User 2:</b> You can pair this with shorts, slacks, or jeans—basically anything. Just make sure that the color of your top &amp; bottom matches.  <b>User 3:</b> If you are looking for women’s wear I would say a very wide leg pastel high waisted pant with a tight/fitted top in same color scheme or white.  <b>User 4:</b> This reminds me of the kind of shoes I see in anime with sailor style uniforms tbh  <b>User 5:</b> A Goodwill donation          ...       </p>
<p><b>Summary:</b>          The OP wanted to know what to wear with a pair of white leather loafers that have a thick black sole and low heel. One commenter thought pastel pants and a white top to match the shoes would work. Another commenter said that OP’s shoes would pair with any sort of bottoms, but cautioned that the top and bottom color should match. One user shared links for OP to use as inspiration. Another user thought that the shoes looked like anime sailor shoes. Two commenters didn’t like OP’s shoes, and suggested they be thrown away or donated.       </p>

Table B.1: Another example from our dataset, from the *fashionadvice* subreddit.

**Image:**



**Post Caption:**

Just moved into my first home and this space bothers me. Need some advice to make it look more cohesive.

**Comments:**

**User 1:** center feels empty. if it were me, i'd place one of those vintage wooden radio clocks in the middle. that's oddly specific i know...

**User 2:** Change nothing but add a vase of fresh white flowers in the center

**User 3:** Center large art piece and move it down. Lean the small art off center behind Candles, use the basket as a trinket tray on console.

**User 4:** I really like your art. I agree that the center needs something, maybe a plant or a stack of books.

**User 5:** I rather like it. The only thing missing is something sort of tallish in the center to fill that space. Like, it is the perfect spot for a vase filled with flowers. Some color and life! If a floral subscription isn't in your future lol maybe a full plant would fit the bill

...

**Summary:**

The OP asked what to do with a space in their home that presently has light blue walls and a brown sideboard with a lamp and candlesticks on it. Most commenters agreed the space looked good as-is, but recommended just adding something in the empty center of the table, such as a vase of white flowers or a large plant. Others thought a vintage wooden radio clock or traditional record player in the same green color as the candles would look perfect, while another suggested a stack of nice books. Others said to center the large wall art, and to check local thrift stores for a substantial but short statement piece to be the center accent decor. Others recommended using a basket as a trinket tray or just buying a marble tray for trinkets on the table. They also said to lean the smallest art pieces against the wall behind the candles, or get rid of the candles altogether.

Table B.2: Another example from our dataset, from the *designmyroom* subreddit.

# Appendix C

## Further Analyses

### C.0.1 Summarization based on the Length of Input Threads

To better understand whether CMS effectively handles long inputs, we run a further analysis using BART-based models (see Figure C.1). As the number of comments increases, the R1 score consistently decreases. This indicates that summarization indeed becomes more challenging when the input length is longer. However, the performance gap between the baseline models (i.e., BART, BART-ImgCap) and the CMS-BART-ImgCap generally increases as the number of comments grows, supporting the idea that CMS better handles longer threads. As our model generates cluster summaries in stage 1, it reduces the average input length by 82.8%, and thus achieving better performance even on relatively challenging long inputs. We also provide results from T5-based models in Figure C.2, showing similar trends; the gap between the baseline models and the CMS-T5-ImgCap is large when the number of comments falls within the range of  $[15,20)$  and  $[20,25)$ .

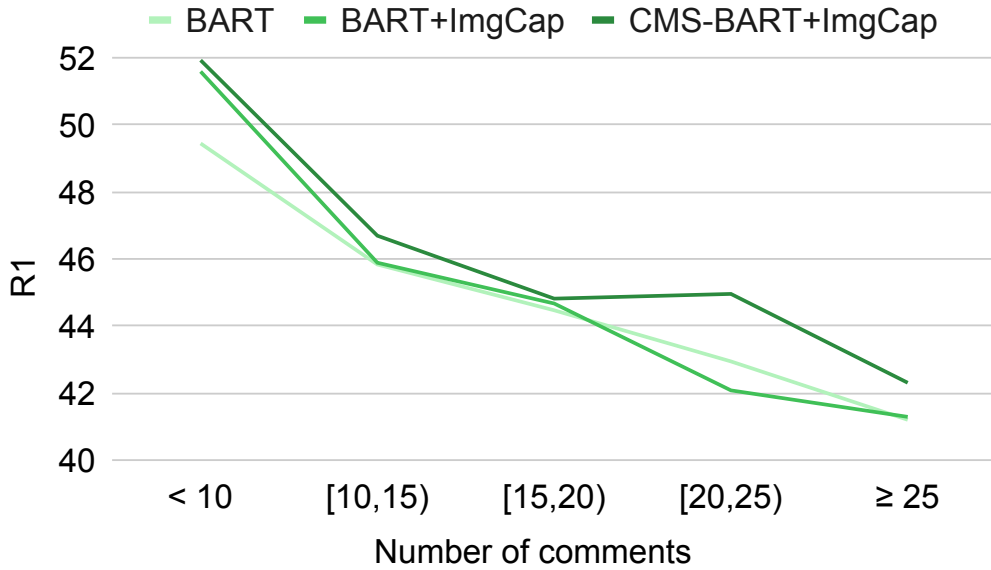


Figure C.1: The influence of the number of comments in the thread on summarization performance (ROUGE-1) on BART-based models measured on the test set.

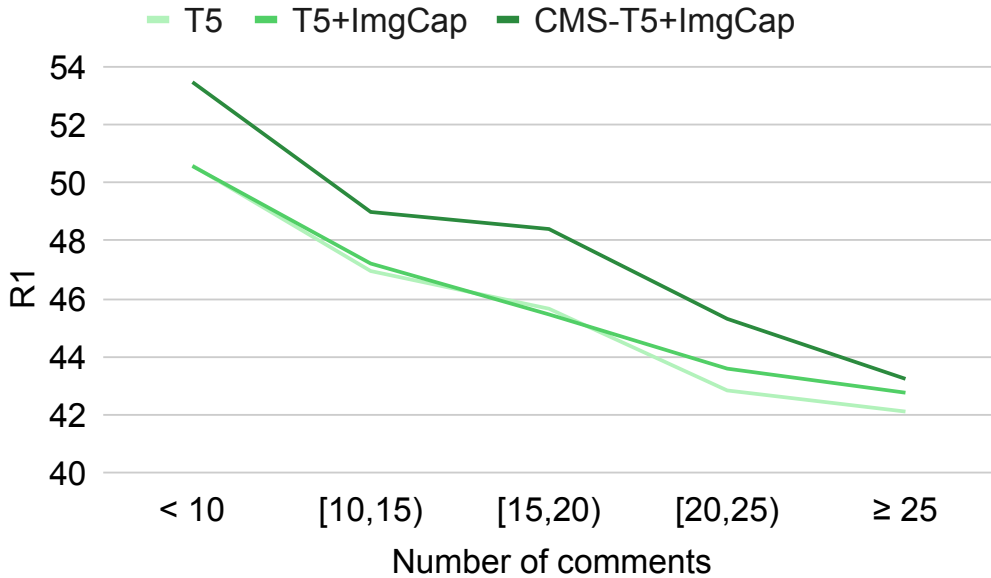


Figure C.2: The influence of the number of comments in the thread on summarization performance (ROUGE-1) of T5-based models. The results are based on the test set.

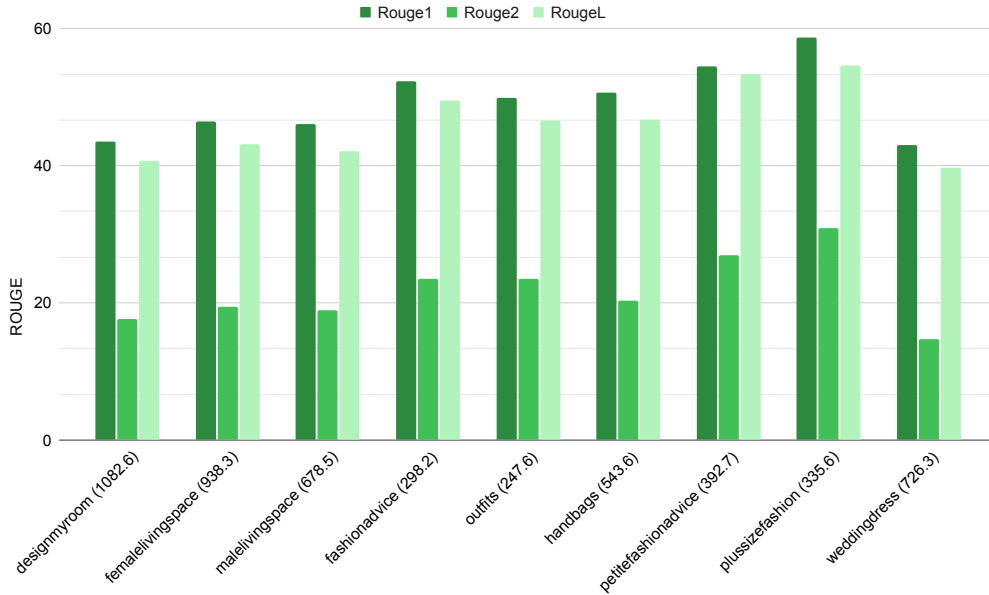


Figure C.3: ROUGE scores obtained from our CMS-T5-ImgCap model on the test set, categorized by different subreddits. The number of input words is indicated in parentheses.

### C.0.2 Summarization per Subreddit

We further explore the summarization across 9 different subreddits, as shown in Figure C.3.

The results reveal that subreddits within the ‘Interior’ category (i.e., the left three subreddits in Figure C.3) exhibit lower ROUGE scores in comparison to subreddits within the ‘Clothes’ category (i.e., the right six subreddits in Figure C.3). This discrepancy can be attributed to the difference in the input lengths across each subreddit. Given that the average input length of examples from the ‘Interior’ category exceeds that of examples from the ‘Clothes’ category, it is more difficult for our model to summarize the former. Additionally, we can also explain this gap by comparing the difference between domains. Specifically, while the model can easily comprehend clothing images by focusing on only salient objects, comprehending interior images is more challenging as it necessitates a broader range of information (e.g., wall color, spatial relationship between furniture, etc). Consequently, summarizing examples from the ‘Interior’ category proves to be more challenging for the models than summarizing examples from the ‘Clothes’ category.

# Bibliography

- [1] A. Fabbri, X. Wu, S. Iyer, H. Li, and M. Diab, “AnswerSumm: A manually-curated dataset and pipeline for answer summarization,” in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Seattle, United States), pp. 2508–2520, Association for Computational Linguistics, July 2022.
- [2] A. Fabbri, F. Rahman, I. Rizvi, B. Wang, H. Li, Y. Mehdad, and D. Radev, “ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Online), pp. 6866–6880, Association for Computational Linguistics, Aug. 2021.
- [3] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, “SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization,” in *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, (Hong Kong, China), pp. 70–79, Association for Computational Linguistics, Nov. 2019.



- [4] R. Nallapati, B. Zhou, C. dos Santos, Ç. Gulçehre, and B. Xiang, “Abstractive text summarization using sequence-to-sequence RNNs and beyond,” in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, (Berlin, Germany), pp. 280–290, Association for Computational Linguistics, Aug. 2016.
- [5] J. Zhu, H. Li, T. Liu, Y. Zhou, J. Zhang, and C. Zong, “MSMO: Multimodal summarization with multimodal output,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (Brussels, Belgium), pp. 4154–4164, Association for Computational Linguistics, Oct.-Nov. 2018.
- [6] T. Yu, W. Dai, Z. Liu, and P. Fung, “Vision guided generative pre-trained language models for multimodal abstractive summarization,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, (Online and Punta Cana, Dominican Republic), pp. 3995–4007, Association for Computational Linguistics, Nov. 2021.
- [7] S. Bhatia, P. Biyani, and P. Mitra, “Summarizing online forum discussions – can dialog acts of individual messages help?,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 2127–2131, Association for Computational Linguistics, Oct. 2014.
- [8] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, “How2: a large-scale dataset for multimodal language understanding,” *arXiv preprint arXiv:1811.00347*, 2018.
- [9] S. Palaskar, J. Libovický, S. Gella, and F. Metze, “Multimodal abstractive summarization for how2 videos,” in *Proceedings of the 57th Annual Meet-*

- ing of the Association for Computational Linguistics*, (Florence, Italy), pp. 6587–6596, Association for Computational Linguistics, July 2019.
- [10] N. Liu, X. Sun, H. Yu, W. Zhang, and G. Xu, “Multistage fusion with forget gate for multimodal summarization in open-domain videos,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Online), pp. 1834–1845, Association for Computational Linguistics, Nov. 2020.
- [11] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” 2020.
- [12] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “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*, (Online), pp. 7871–7880, Association for Computational Linguistics, July 2020.
- [13] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” 2020.
- [14] S. Zhang, A. Celikyilmaz, J. Gao, and M. Bansal, “EmailSum: Abstractive email thread summarization,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International*

- Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Online), pp. 6895–6909, Association for Computational Linguistics, Aug. 2021.
- [15] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, and D. Parikh, “Vqa: Visual question answering,” 2016.
  - [16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” 2021.
  - [17] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” 2022.
  - [18] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” 2023.
  - [19] K. Desai, G. Kaul, Z. Aysola, and J. Johnson, “Redcaps: Web-curated image-text data created by the people, for the people,” in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual* (J. Vanschoren and S. Yeung, eds.), 2021.
  - [20] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, “CLIPScore: A reference-free evaluation metric for image captioning,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, (Online and Punta Cana, Dominican Republic), pp. 7514–7528, Association for Computational Linguistics, Nov. 2021.

- [21] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*, (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, July 2004.
- [22] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with BERT,” *CoRR*, vol. abs/1904.09675, 2019.
- [23] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, “Training language models to follow instructions with human feedback,” 2022.
- [24] T. Goyal, J. J. Li, and G. Durrett, “News summarization and evaluation in the era of gpt-3,” 2022.
- [25] A. Bhaskar, A. R. Fabbri, and G. Durrett, “Zero-shot opinion summarization with gpt-3,” in *Findings of the Association for Computational Linguistics: ACL 2023*, Association for Computational Linguistics, 2023.
- [26] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, “Microsoft coco captions: Data collection and evaluation server,” 2015.
- [27] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, K.-W. Chang, and J. Gao, “Grounded language-image pre-training,” 2022.
- [28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image

recognition at scale,” in *International Conference on Learning Representations*, 2021.

- [29] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” 2019.
- [30] S. Zhang, A. Celikyilmaz, J. Gao, and M. Bansal, “Emailsum: Abstractive email thread summarization,” 2021.

# Acknowledgements

I would like to thank my advising professor, Gunhee Kim, for his support in this research. He has provided consistent counseling throughout the project and it would not have been possible without his guidance. I would also like to thank Professor Joonsuk Park who has additionally acted as a co-corresponding author in this research project on behalf of Naver. He has provided me with invaluable advice and direction during this research project, as well as personal encouragement that has helped us see this project to completion. I would also like to thank several members of the SNU Vision and Learning Lab who have either participated with me in this research or provided their own feedback, including but not limited to Fatima Pesaran, Jaewoo Ahn, Dayoon Ko, and Seokhee Hong.

## 요약

인공지능 기술과 대규모 언어 모델의 발전에 힘입어, 뉴스, 대화, 토의를 위한 자동 요약 기술 또한 빠르게 발전했다. 그러나, 대부분의 자동 요약 기술은 텍스트만 요약하는 것에 한정되어 있으며, 비디오와 이미지를 수반하여 이뤄지고 있는 온라인상 많은 토의를 위한 기술은 거의 다뤄지지 않았다. 현재 요약 데이터 세트들 또한 텍스트들로만 이뤄져 있으며, 이러한 멀티모달 (Multimodal) 영역을 다루는 요약 데이터 세트는 충분치 않다. 이를 해결하기 위하여, 우리는 첫 멀티모달 토의 요약 데이터 세트인 mRedditSum을 선보인다. Reddit의 서브 레딧(subreddits) 으로부터 모은 3,033개의 고품질의 토의 스레드(thread)들로 이루어진 본 데이터 세트는 이미지와 텍스트에 기반하여 조언을 구하는 글과 그 글에 다양한 의견으로 답하는 답변들로 구성돼 있다. 멀티모달의 특성에 맞게, 각 스레드에 해당하는 요약은 텍스트뿐만 아니라 이미지에서만 얻을 수 있는 정보들을 취합하여 사람이 작성하였다. 우리는 자동 요약에 자주 쓰이는 대규모 언어 모델들 - T5, BART, GPT-3 - 을 활용하여 실험을 진행하였고, 이미지 캡션(caption) 혹은 비전-텍스트 퓨전 계층(vision-text fusion layer)이 사용되었을 때, 자동 요약의 성능이 향상함을 보였다.

**주요어:** 딥러닝, 자연언어처리, 컴퓨터비전, 멀티모달 요약, 데이터세트

**학번:** 2021-29898