UniMS: A Unified Framework for Multimodal Summarization with Knowledge Distillation

Zhengkun Zhang[1]*, **Xiaojun Meng**[2], **Yasheng Wang**[2], **Xin Jiang**[2], **Qun Liu**[2], **Zhenglu Yang**[1]†

[1]TKLNDST, CS, Nankai University, China, [2] Noah's Ark Lab, Huawei Technologies
zhangzk2017@mail.nankai.edu.cn,
{xiaojun.meng, wangyasheng, Jiang.Xin, qun.liu}@huawei.com,
yangzl@nankai.edu.cn

Abstract

With the rapid increase of multimedia data, a large body of literature has emerged to work on multimodal summarization, the majority of which target at refining salient information from textual and visual modalities to output a pictorial summary with the most relevant images. Existing methods mostly focus on either extractive or abstractive summarization and rely on qualified image captions to build image references. We are the first to propose a **U**nified framework for **M**ultimodal **S**ummarization grounding on BART, **UniMS**, that integrates extractive and abstractive objectives, as well as selecting the image output. Specially, we adopt knowledge distillation from a vision-language pretrained model to improve image selection, which avoids any requirement on the existence and quality of image captions. Besides, we introduce a visual guided decoder to better integrate textual and visual modalities in guiding abstractive text generation. Results show that our best model achieves a new state-of-the-art result on a large-scale benchmark dataset. The newly involved extractive objective as well as the knowledge distillation technique are proven to bring a noticeable improvement to the multimodal summarization task.

Introduction

Existing researchers (Srihari 1994; He and Deng 2017; Baltrusaitis, Ahuja, and Morency 2019) have evidenced that in order to understand the world around us, it needs to be able to interpret and reason from multimodal information. Recently proposed Multimodal Summarization with Multimodal Output (Zhu et al. 2018) (MSMO) that condenses long multimodal news to a short pictorial version, as shown in Fig. 1, is a new fashion of summarization. This method improves users' satisfaction with the ability to quickly grasp news highlights, and brings both research and commercial value in the face of information explosion.

A large amount of methods (Zhu et al. 2018, 2020; Zhang et al. 2021) are proposed for improving multimodal summarization. They serve as baselines for the large-scale benchmark dataset collected by Zhu et al. (2018), and provide insights for recent work on multimodal summarization. In order to obtain the pictorial summary including a piece of condensed text with images as shown in Fig. 1, existing methods typically consist of two objectives.

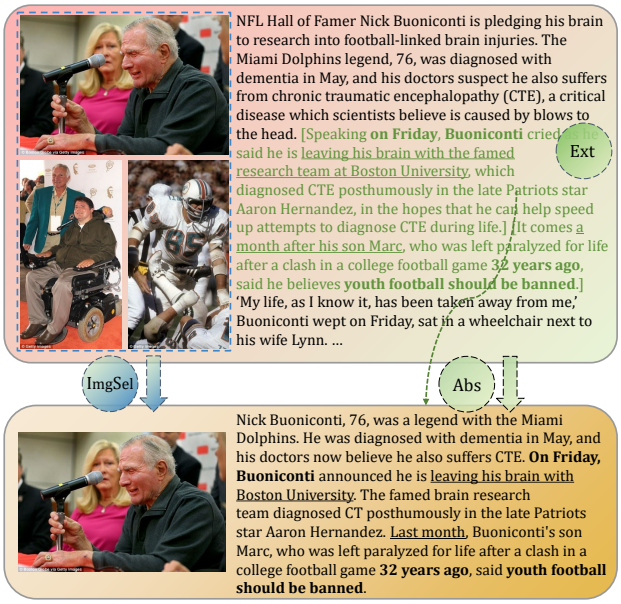

Figure 1: The illustration of multimodal summarization task. Our unified framework is able to achieve extractive and abstractive summarization objectives, as well as the image selection objective with no need of using image captions.

Firstly, as a critical modality of the pictorial summary, the objective of abstractive text summarization plays an important role in improving multimodal summarization. Recent advances in denoising autoencoder for pretraining sequence-to-sequence models such as BART (Lewis et al. 2020) have been shown to capture many language relevant facets for downstream tasks, which leads to a large improvement on text summarization metrics. We believe the existing high-capacity language model benefits summarization remarkably even in a multimodal setting. However, the original BART model has no support for multimodal input and output. Therefore, we are motivated to integrate textual and visual modalities into the BART model and extend its architecture to further improve the multimodal summarization task.

*Work is done at the internship of Noah's Ark Lab, Huawei Technologies.

†Corresponding authors.

Next, to leverage the visual modality and improve the quality of model-selected image, Zhu et al. (2020) proposes to incorporate image references into the training process, and jointly consider summary generation and image selection as training targets. Due to the lack of image references, Zhu et al. (2020) proposes a ROUGE-ranking method using the similarity of image caption and text reference to rank images, which hence builds the pseudo image reference for training. It assumes the image caption has already contained condensed textual information of visual modalities, which semantically matches with the corresponding image. However, this assumption largely depends on the quality of the image caption instead of the image content itself. To sum up, the effectiveness of this ROUGE-ranking method relies on the presence of high-quality image captions. Unfortunately, the image captions are not often qualified or even not present in a multitude of irregular multimedia data.

Thus, we propose to distill knowledge from a pretrained vision-language teacher model (*e.g.,* CLIP) to help our model on learning the relevance rank of images in the given textual context without any requirement on the existence and quality of image captions. Specially, given the recent process in contrastive vision-language pre-training (Radford et al. 2021), we are motivated to use the cosine similarity of text references and images to represent text-image content relevance, and use this relevance to build pseudo image references. As far as we know, distilling knowledge from vision-language models to mentor multimodal summarization remains unexplored.

Considering the above mentioned abstractive and image selection objectives, we design **UniMS**: a **Uni**fied framework for **M**ultimodal **S**ummarization grounding on BART, which integrates inputs of both textual and visual modalities along with multiple multimodal objective functions in a unified multitask framework. We modify the BART architecture separately on its encoder and decoder. In the encoder, we distill the knowledge from a pretrained vision-language teacher model to guide our encoder on selecting images. We additionally introduce a new extractive objective, together with the image selection objective, as multimodal supervised signals for our encoder. It could potentially reduce the modality-bias problem (Zhu et al. 2020) and improve the understanding of multi-modalities for our encoder, which eventually benefits the multitask abilities. In the decoder, we adopt visual guided attention to better integrate both modalities to achieve the abstractive objective.

Overall, our unified multitask framework extends the BART model by enabling both extractive and abstractive summarization, as well as image selection as output. To our knowledge, **UniMS** is the first unified framework including three objectives of multimodal summarization. Experimental results show that our framework achieves new state-of-the-art results in all tasks. Our contributions in this paper are three-fold:

1. We propose a unified multimodal summarization framework with an encoder-decoder multitask architecture on top of BART, which simultaneously outputs extractive and abstractive summaries, and image selection results.

2. Our framework adopts knowledge distillation to improve image selection without any requirement on the existence and quality of image captions. We further introduce the extractive objective in the encoder and visual guided attention in the decoder to better integrate both textual and visual modalities in the conditional text generation.

3. Our unified method achieves a new state-of-the-art result of multimodal summarization in all the subtasks, i.e., extractive, abstrastive, and image selection.

Related Work

Multimodal summarization takes two or more modalities of data as input, and outputs either single-modal or multimodal summary. The input modality could contain documents, images, audios and videos from the rich multimedia information (Li et al. 2017), which enables it to make use of more information than traditional text summarization. Zhu et al. (2018) construct a large-scale corpus for a novel multimodal summarization task, which takes the news with images as input, and outputs a pictorial summary. They also present a multimodal attention method serving as a baseline for this task.

To further improve the quality of the model-selected images, Zhu et al. (2020) propose to incorporate the multimodal reference into the training process, thus taking account of both summary generation and image selection to guide the abstractive model. LAMS (Zhang et al. 2021) considers image location as a critical factor for news summarization, and proposes a location-aware extractive approach for multimodal summarization.

Existing methods are only targeting either extractive or abstractive multimodal summarization. The tradeoff in the use of extractive and abstractive approaches has been discussed a lot in existing research (Cao et al. 2017; Jia et al. 2021): the former often achieve better factuality and efficiency, while the latter generates flexible and less redundant summaries. They both have merits in summarizing multimedia data.

We aim to propose a unified framework producing both extractive and abstractive summaries with a selected image preview that serves a multi-functional use. Our framework extends BART model (Lewis et al. 2020) to a unified multitask architecture. The model comparison with related work is presented in Table 1, and more details are discussed in Section 4.3. In addition, we distill the knowledge from state-of-the-art vision-language model CLIP (Radford et al. 2021) to guide our training process, which avoids any requirement on the existence and quality of image captions.

Methodology

Model Overview

We propose a new multimodal summarization framework that unifies multiple subtasks of multimodal summarization. We use the BART model (Lewis et al. 2020) as our backbone architecture which consists of an encoder and decoder component. The overall architecture of our framework is shown in Fig. 2.

Given a multimodal document $D = \{T, V\}$, where $T = \{t_1, t_2, ..., t_M\}$ is a sequence of M tokens and $V =$

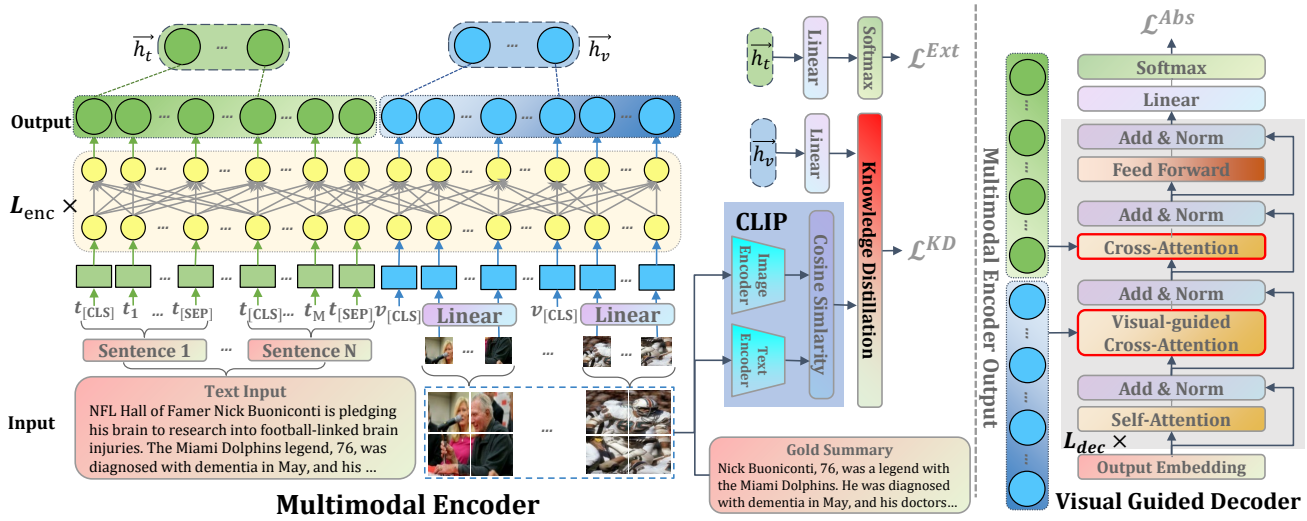

Figure 2: Overview of our proposed UniMS framework. At first, we extend BART encoder to multimodal encoder by embedding visual features with linear projection as additional input. To enhance the multimodal encoder's ability to understand multimodal input, we further introduce extractive text reference and image reference via knowledge distillation to guide our encoder. Besides, the proposed visual guided decoder separately attends to the visual and the textual output of the multimodal encoder.

$\{v_1, v_2, ..., v_N\}$ is a collection of N images, the proposed framework summarizes D into a multimodal summary $Y = \{Y_t, Y_v\}$, where $Y_t = \{y_1, y_2, ..., y_l\}$ denotes the textual summary limited by length l and Y_v is a subset of images extracted from the image collection V.

Multimodal Reference Enhanced Encoder

Multimodal Encoder. Inspired by Kim, Son, and Kim (2021), we adopt a simple scheme, linear projection operating on image patches for obtaining visual embeddings. Specifically, each input image v_i is sliced into patches and flattened as $\{v_{ik}\}$. Followed by linear projection and visual position embedding v_{pos}, v_i is embedded into e_{v_i}. It is worth noting that, unlike existing methods (Zhu et al. 2018, 2020), we take the first attempt to only use patch projection rather than any visual backbone.

Then, we extend the text encoder of BART to multimodal encoder, which takes the concatenation of text embeddings e_t and visual embeddings e_v as input and outputs their contextualized joint representations:

$$\begin{aligned} e_t &= [t_{[\text{CLS}]}; t_1; ...; t_M; t_{[\text{SEP}]}]W_t, \\ e_{v_i} &= [v_{[\text{CLS}]}; v_{i1}; ...; v_{ik}]W_v + v_{pos}, \\ e &= [e_t; e_v] = [e_t; e_{v_i}; ...; e_{v_N}] + e_{pos}, \\ h &= [h_t; h_v] = f_{\text{enc}}(e), \end{aligned} \tag{1}$$

where $t_{[\text{CLS}]}$ and $t_{[\text{SEP}]}$ are the special token embeddings introduced to mark the start and the end of the text input, e_{pos} is the multimodal sequence position embedding, f_{enc} denotes the encoder function, W_t and W_v are learnable parameters. Following Dosovitskiy et al. (2021), we add $v_{[\text{CLS}]}$ as the beginning token of each image, whose hidden state at the output of the encoder serves as the whole image representation.

Knowledge Distillation for Image Reference. The multimodal summarization training set has only plain summary text reference, which lacks image reference for helping image selection in the training process. Thus, the existing research (Zhu et al. 2020) considers **ROUGE-ranking** (RR) strategy to rank images and choose the top-k as image references, which thus extends text-only to multimodal reference. In detail, RR uses the similarity of image caption and text reference to rank images. However, this strategy relies on not only the presence of image captions but also the high relevance between images and their captions, which can lead to a very limited application scenario in reality.

Therefore, we adopt the Knowledge Distillation (KD) technique (Hinton, Vinyals, and Dean 2015) for distilling the text-image content relevance knowledge, which can get image reference without any image caption. Our proposed method aims to distill the knowledge from a teacher network, i.e., CLIP (Radford et al. 2021), to a student network, i.e., our encoder. The student network is trained to mimic the behaviors of the teacher network, on calculating text-image content relevance score as well as ranking images according to the precedent score.

Specifically, we use the hidden state $h_{v_{[\text{CLS}]}}$ of visual token `[CLS]` as the image representation $\overrightarrow{h_v}$, and feed it into a fully connected layer to predict a score:

$$g_v(\overrightarrow{h_v}) = W_v \overrightarrow{h_v} + b_v, \tag{2}$$

where W_v and b_v are learnable parameters.

Then, we employ CLIP (Radford et al. 2021) as a teacher model to calculate the cosine similarity scores between text embeddings generated by feeding textual summary Y_t into its text encoder $\mathcal{T}$, and image embeddings generated by feeding image collection V into its image encoder $\mathcal{V}$:

$$f(v, Y_t) = \text{sim}(\mathcal{V}(v), \mathcal{T}(Y_t)). \tag{3}$$

To quantify the variation of ranking score distribution for our encoder and the CLIP model, we use the Kullback-Leibler (KL) divergence (Kullback and Leibler 1951). Via distilling knowledge, our encoder intends to directly mimic the teacher model's score distribution. Formally, CLIP-based KD can be modeled as minimizing the following objective functions with temperature τ. Unless otherwise specified, all results reported in this paper use temperature $\tau = 10$ which we find to perform best.

$$p_v(\overrightarrow{h_v}, \tau) = \frac{\exp(g_v(\overrightarrow{h_v})/\tau)}{\sum_{v \in V} \exp(g_v(\overrightarrow{h_v})/\tau)}, \tag{4}$$

$$q_v(v, Y_t, \tau) = \frac{\exp(f(v, Y_t)/\tau)}{\sum_{v \in V} \exp(f(v, Y_t)/\tau)}, \tag{5}$$

$$\mathcal{L}^{\mathrm{KD}} = \mathcal{L}_{\mathrm{KL}}(\mathbf{p}||\mathbf{q}) = -\sum_{v \in V} p_v \ln \frac{q_v}{p_v}. \tag{6}$$

Extractive Text Reference Due to the fact that if the encoder only obtains the reference from one single modality, the system may suffer from the modality-bias problem (Zhu et al. 2020). Therefore, we introduce the extractive text reference to supervise our encoder. In other words, we regard extractive text summarization as one of the subtasks in our framework. We use a greedy algorithm similar to Nallapati, Zhai, and Zhou (2017) to obtain an oracle summary within each document as the extractive text reference. In detail, our algorithm generates an oracle consisting of multiple sentences, which are selected greedily to maximize the ROUGE-L score against the gold summary.

In our encoder, for representing individual sentences, we insert extra `[CLS]` and `[SEP]` tokens at the start and end of each sentence, expecting each `[CLS]` token to collect features for the sentence following it, i.e., $\overrightarrow{h_t} = h_{t_{[\mathrm{CLS}]}}$. Thus, on top of which, a fully connected layer can be used to predict an extractive score for each sentence:

$$g_t(\overrightarrow{h_t}) = W_t \overrightarrow{h_t} + b_t, \tag{7}$$

where W_t and b_t are learnable parameters.

Given the extractive score per sentence, the regular binary cross-entropy loss is employed as the objective function for extractive text reference:

$$p_t(\overrightarrow{h_t}) = \frac{\exp(g_t(\overrightarrow{h_t}))}{\sum_t \exp(g_t(\overrightarrow{h_t}))}, \tag{8}$$

$$\mathcal{L}^{\mathrm{Ext}} = -\sum_t \log p_t(\overrightarrow{h_t}). \tag{9}$$

Visual Guided Decoder

Different from the original BART, our decoder has to attend to both the textual and visual content of the multimodal document instead of single modality input. Inspired by the guidance-aware mechanism proposed by Dou et al. (2021), which can introduce multiple textual guidance signals into the transformer decoder, we design a visual guided text generation decoder for better utilizing our encoded visual signals. As shown in Fig. 2, after self-attention, the sequence in the decoder first attends to the encoded visual hidden states and generates the visual-guided cross-modal representation. After that, this sequence continuously attends to the textual hidden states to produce the final representation:

$$\begin{aligned} y &= \mathrm{LN}(y + \textsc{CrossAttn}(y; h_v)), \\ y &= \mathrm{LN}(y + \textsc{CrossAttn}(y; h_t)), \end{aligned} \tag{10}$$

Ideally, modeling the multimodal signals separately with two cross-attention blocks allows the decoder to be explicitly guided by signals from different modalities. Overall, the decoder iteratively attends to previously generated tokens $y_{<j}$ and the encoder outputs h, and subsequently predicts the probability of future text tokens $p_y(y_j|y_{<j}, h)$. For multimodal conditional text generation task, i.e., multimodal abstractive summarization subtask, we train our model by minimizing the negative log-likelihood:

$$\mathcal{L}^{\mathrm{Abs}} = -\sum_{j=1}^{|y|} \log p_y(y_j|y_{<j}, h). \tag{11}$$

Finally, the training loss $\mathcal{L}$ of our framework is a sum of the objectives of all these subtasks: image selection, extractive, and abstractive text summarization.

$$\mathcal{L} = \mathcal{L}^{\mathrm{KD}} + \mathcal{L}^{\mathrm{Ext}} + \mathcal{L}^{\mathrm{Abs}}. \tag{12}$$

Experiments

Datasets

We use the MSMO dataset, which is collected by (Zhu et al. 2018) for multimodal summarization. It contains online news articles from the $DailyMail$ website[1] paired with multiple images (6.58 images on average) and multi-sentence ground-truth summaries. The dataset includes 293,965 training pairs, 10,355 validation pairs, and 10,261 test pairs. In the test set, at most three images are annotated as a multimodal reference.

Implementation Details

Our framwork is built on 'bart-base'[2] version of BART (Lewis et al. 2020) with its initialized parameters and tokenizer. We use the released 'ViT-B-32'[3] version of CLIP (Radford et al. 2021) as the teacher network for knowledge distillation. Meanwhile, following the image pre-processing used in CLIP, we resize the resolution of each image into 224×224 with a patch projection yielding 49 patches. We limit the number of images to 10 and pad the visual token `[CLS]` and image patch tokens together in a total length of 500 for batch training. Besides, we truncate each article to 512 tokens, and the oracle summary in this paper is calculated on the truncated articles. All models are trained for 30,000 steps with 750 steps for warm-up. Model checkpoints are saved and evaluated on the validation set every 2,000 steps. We select the top-3

[1] http://www.dailymail.co.uk

[2] https://huggingface.co/facebook/bart-base

[3] https://github.com/openai/CLIP

Model	Multimodal	Ext.	Abs.	ImgSel
BertExt (2019)		✓		
BertAbs (2019)			✓	
BertExtAbs (2019)		✓	✓	
BART (2020)			✓	
ATG/ATL/HAN (2018)	✓		✓	✓
GR (2018)	✓	✓		✓
MOF (Zhu et al. 2020)	✓		✓	✓
LAMS (Zhang et al. 2021)	✓	✓		
UniMS	✓	✓	✓	✓

Table 1: Model comparison with baseline models. Multimodal, Ext., Abs. and ImgSel denote whether the model could support multimodal input, perform extractive text summarization, abstractive text summarization, and image selection subtasks, respectively.

checkpoints according to the validation loss and report the averaged results on the test set.

For abstractive summarization, we use beam search (size 5) in decoding, and tune the α for the length penalty (Wu et al. 2016) between 1.6 and 2.0 on the validation set; we decode until an end-of-sequence token is emitted. For extractive summarization, we first use the model to obtain the score for each sentence. We then rank these sentences descendingly according to their scores and select the top-3 sentences as the summary.

We report the F1 **ROUGE** score via ROUGE-1.5.5.pl (Lin 2004), which calculates the overlap lexical units between generated and ground-truth sentences. For image selection, we report image precision (**IP**) score, which is defined in (Zhu et al. 2018) to represent whether an image is correctly selected as output. Besides, $\mathbf{M_{sim}}$ is an image-text relevance metric which measures the maximum similarity between the output image and text summary via the cross-modal retrieval model, as stated in Zhu et al. (2018, 2020).

Baselines

To show the effectiveness of our unified framework, we compare our model with existing text and multimodal summarization methods:

BertSum (Liu and Lapata 2019) is an unified model for extractive and abstractive text summarization whose parameters are initialized with BERT (Devlin et al. 2019). Its variants include **BertExt**, **BertAbs** and **BertExtAbs**.

BART (Lewis et al. 2020) is the state-of-the-art abstractive pure text summarization model pretrained with a denoising autoencoding objective.

ATG/ATL/HAN/GR are proposed by Zhu et al. (2018), the former three of which leverage the pointer generator network (See, Liu, and Manning 2017) for multimodal summarization via adopting the visual attention on global features, local features and hierarchically local features of images. **GR** is an extractive method that ranks sentences and captions via LexRank (Erkan and Radev 2004) and guidance strategies.

Model	R-1	R-2	R-L	IP	M_{sim}
Text Abstractive					
BertAbs (2019)	39.02	18.17	33.20	-	-
BertExtAbs (2019)	39.88	18.77	38.36	-	-
BART (2020)	<u>41.83</u>	<u>19.83</u>	<u>39.74</u>	-	-
Multimodal Abstractive					
ATG (2018)*	40.63	18.12	37.53	59.28	25.82
ATL (2018)*	40.86	18.27	37.75	62.44	13.26
HAN (2018)*	40.82	18.30	37.70	61.83	12.22
$\text{MOF}^{\text{RR}}_{\text{enc}}$ (2020)*	41.05	18.29	37.74	62.63	26.23
$\text{MOF}^{\text{RR}}_{\text{dec}}$ (2020)*	<u>41.20</u>	18.33	37.80	<u>65.45</u>	<u>26.38</u>
UniMS	**42.94**	**20.50**	**40.96**	**69.38**	**29.72**
w/o Visual Guide	42.71	20.26	40.76	69.14	29.64
w/o $\mathcal{L}^{\text{Ext}}$	42.63	20.21	40.61	69.25	29.61
w/o Both	42.55	19.88	40.14	69.22	29.57
UniMS-VL	40.81	18.83	38.83	68.26	29.45

Table 2: Experimental results for multimodal summarization on MSMO test set, where "w/o Visual Guide" denotes to remove the visual-guided cross attention from the decoder and keep only one cross attention on all hidden states from two modalities, "w/o $\mathcal{L}^{\text{Ext}}$" denotes to the model without applying $\mathcal{L}^{\text{Ext}}$ to the encoder and "w/o Both" denotes to remove the former two. Results marked by * are taken from the authors' respective papers.

MOF variants (Zhu et al. 2020) introduce the multimodal objective function with image reference obtained by ROUGE-ranking (**RR**) as described in Section 3.2. Besides, they incorporate the last hidden states of the text encoder or the summary decoder into their proposed image discriminator and denote it as MOF_{enc} and MOF_{dec}.

LAMS (Zhang et al. 2021) proposes a location-aware extractive approach to utilize the image locations for multimodal summarization.

UniMS-VL: Cho et al. (2021) proposes VL-BART which extends BART text encoders to multimodal encoders by incoporating image region embeddings extracted by Fast-R-CNN (Ren et al. 2017) as additional input. VL-BART has shown good performances on vision-language tasks. Therefore, we also aim to train our model with their released checkpoint as a starting point.

The comparison among our **UniMS** and different baseline models is shown in Table 1, which demonstrates that our unified framework is able to cover all the subtasks of multimodal summarization.

Automatic Evaluation

Table 2 summarizes the results on abstractive summarization and image selection subtasks. The first block in the table includes abstractive summarization methods with text-only input, while the second block includes abstractive methods with multimodal input. We find that, by fine-tuning the pretrained BART model (Lewis et al. 2020), the results of abstractive text summarization is able to achieve competitive performances as state-of-the-art models that additionally use visual signals, which indicates the powerful language mod-

Model	R-1	R-2	R-L
ORACLE	50.15	28.56	47.91
LEAD-3	39.94	18.56	38.38
GR (2018)*	37.13	15.03	30.21
BertExt (2019)	39.02	18.17	33.20
LAMS-ATL (2021)*	42.48	19.75	38.78
LAMS-MFB (2021)*	**43.07**	20.28	39.34
UniMS	42.58	**20.29**	**40.91**
UniMS-VL	41.29	19.01	39.47

Table 3: Experimental results for extractive summarization on MSMO test set. Results marked by * are taken from the respective papers.

Model	R-L	IP	M_{sim}
ROUGE-ranking	40.90	69.25	29.56
UniMS	**40.96**	**69.38**	**29.72**

Table 4: Experimental results for multimodal summarization with different methods of building image references.

eling capabilities of BART and motivates us to extend BART to handle multi-modalities. Our proposed **UniMS** outperforms state-of-the-art methods in both abstractive summarization and image selection tasks, demonstrating the superiority of our multimodal summarization model.

In addition, our model can achieve extractive goals with $\mathcal{L}^{\text{Ext}}$. We display the results of our extractive summarization in Table 3. As an upper bound, the table begins with an extractive ORACLE result as described in Section 3.2. We also present the LEAD-3 baseline, which simply selects the first three sentences from a document as extractive output. The second block includes various extractive summarization methods with text-only input, i.e., BertExt, and multimodal input, i.e., LAMS. Our extractive model can achieve comparable performance to the state-of-the-art extractive method (Zhang et al. 2021) that additionally models human-labeled image locations as input.

Overall, our proposed unified multitask framework provides superior performances in all subtasks of multimodal summarization.

Model Analysis

Ablation study. Table 2 also shows the ablation test of our model when some components are removed. It shows the degradation in performances of abstractive summarization and image selection tasks when removing the corresponding components. Meanwhile, we find that training from the BART pretraining parameters is better than the VL-BART. It may be because the pre-training task of VL-BART is primarily designed for vision-language tasks, such as VQA, etc. Therefore, the original BART's capability, such as summarization, is corrupted during their training process. We believe using a visual-languaged pretrained model particularly designed for multimodal summarization tasks could further improve our framework, which we leave as future work.

Model	R-L	IP	M_{sim}	Para.↑%
LinearProjection	**40.96**	**69.38**	**29.72**	**10.07**
ResNet50	40.79	68.92	29.66	30.22
CLIP-ResNet50	40.83	69.14	23.73	28.78
CLIP-ViT-B-32	40.86	69.16	29.62	74.10
Fast-RCNN	40.76	68.33	29.62	41.73
UniMS-VL	38.83	68.26	29.45	41.73

Table 5: Experimental results for multimodal summarization with different visual backbones.

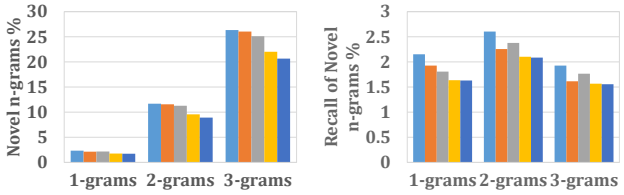

Figure 3: Novel n-grams and the recall of novel n-grams appearing in the ground-truth reference. Introducing the visual guided decoder and extractive objective can enhance the ability of our framework to generate more novel n-grams.

Image Reference. Table 4 depicts the model performances varying with different strategies of building image references. We adopt the ROUGE-ranking (**RR**) strategy from MOF (Zhu et al. 2020) to our framework. Results show that RR is a strong competitor since the high-quality image caption of MSMO dataset has already contained condensed textual information of visual modalities. However, RR relies on the presence of high-quality image captions, while our KD method is more flexible with no additional dependency on captions, leading to wide use in reality.

Visual Backbone. To study the impact of visual backbones in our framework, we conduct an experiment on evaluating the model performances vary with employing different visual backbones, as shown in Table 5. In terms of grid features, we use pretrained visual backbones, i.e., ResNet50, CLIP-ResNet50, and CLIP-ViT-B-32, to extract the 49 image patch features. In terms of region features, we represent an input image v with $n = 49$ object regions with Faster R-CNN (Ren et al. 2017) pretrained on Visual Genome (Anderson et al. 2018).

Results show that the pretrained visual backbones cannot significantly improve the performance of our model. Meanwhile, the use of pretrained visual backbones increases the overhead of obtaining visual embeddings. Using the linear projection only brings an additional 10% gain in parameters compared to BART. As a result, we choose linear projection to obtain visual embeddings in our framework.

Novel n-grams. Following Dou et al. (2021), we count the number of novel n-grams in the output summaries, namely the newly generated n-grams that do not appear in the source document. As shown in Fig. 3, all the variants of our model generate more novel n-grams and cover more novel n-grams that exist in the ground-truth references than the baseline

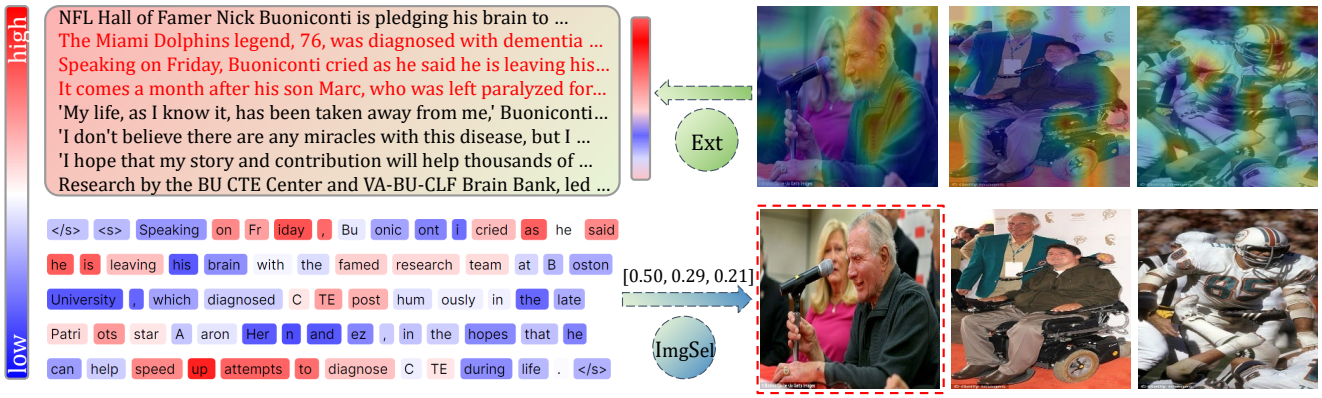

Figure 4: Visualizations for indicating how one modality affects the other in our multimodal encoder. Note that the importance scores decrease from red to blue. The red three sentences are extractive results and the first image with a red border owns the highest score 0.5 in the image selection. Our encoder is able to recognize summary reference related tokens as important, such as the person regions from images, and verb tokens like "said" and "attempt".

Models	Abs. Consist.	Abs. Relev.	Ext. Relev.	ImgSel Relev.
BART	2.18	2.40	-	-
UniMS	**2.32**	**2.45**	2.37	**2.44**
w/o Visual Guide	2.26	2.42	**2.41**	**2.44**
w/o $\mathcal{L}^{\text{Ext}}$	2.22	2.40	-	2.42
w/o Both	2.19	2.39	-	2.38

Table 6: Human evaluation of different model outputs. All the improvements are significant ($p < 0.001$).

BART. It indicates that our modifications on top of BART can indeed generate more novel expressions in the abstractive summary. Additionally, it is clear from Fig. 3 that introducing the extractive loss and visual guided decoder brings a significant improvement on enabling models to generate more novel n-grams.

Human Evaluation. We further conduct a manual evaluation to assess the quality of the generated multimodal summarization. We randomly sample 200 data points from the test set and recruit three people from Amazon Mechanical Turk to rate them between 1 to 3 points on multiple qualitative aspects. For extractive summarization, we ask the annotators to evaluate the relevance (i.e., Relev.), which measures whether the summary captures the key points of the source document. For abstractive summarization, we ask the annotators to evaluate the consistency (i.e., Consist.), which measures the factual alignment between the summary and the source document. In particular for image selection, we ask the annotators to measure whether the summary captures the key points of the selected image as the relevance (i.e., Relev.) score, which indicates the image-text relevance of multimodal output. Table 6 shows that our framework can generate more faithful and relevant summaries compared to other variants and baselines.

Visualizations It is of interest to visualize how one modality affects the other in our multimodal encoder. As shown in Fig. 4, we adopt the class activation mapping (CAM) (Zhou et al. 2016) technique to visualize the importance scores by feeding the last hidden states of our encoder to the fully connected layers used separately in extractive summarization and image selection tasks. Specifically, for extractive summarization, we employ its fully connected layer on the last hidden states of visual embeddings to indicate which region of images is regarded as important by this layer. Similarly for image selection, we employ its fully connected layer on the last hidden states of textual tokens to indicate which set of tokens is regarded as important by this image selection layer. As shown in Fig. 4, we find that our framework surprisingly attends to the person and player region from images, and the verb tokens such as "said" and "attempt", which is quite close to the key points of summary reference.

Conclusion

We propose a unified framework for multimodal summarization that is able to produce extractive and abstractive text output, and jointly select image output for the final pictorial summary. Grounding on BART, our method uses an encoder-decoder architecture consisting of three training objectives, i.e., extractive & abstractive text summarization, and image selection. In particular, extractive and image selection objectives cooperatively supervise the encoder. To improve the image selection, we distill the knowledge from existing pre-trained vision-language models. We also propose a visual guided decoder that separately attends to the textual and visual modalities while performing abstractive text generation. Overall, our best model achieves a new state-of-the-art result on the MSMO dataset. We believe such a unified framework with multi-functional use can serve as a stepping stone to further improve multimodal summarization, as well as baselines against which future methods are tested.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*, 6077–6086.

Baltrusaitis, T.; Ahuja, C.; and Morency, L. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2): 423–443.

Cao, Z.; Wei, F.; Li, W.; and Li, S. 2017. Faithful to the Original: Fact Aware Neural Abstractive Summarization. *CoRR*, abs/1711.04434.

Cho, J.; Lei, J.; Tan, H.; and Bansal, M. 2021. Unifying Vision-and-Language Tasks via Text Generation. In *ICML*, volume 139, 1931–1942.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 4171–4186.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Dou, Z.; Liu, P.; Hayashi, H.; Jiang, Z.; and Neubig, G. 2021. GSum: A General Framework for Guided Neural Abstractive Summarization. In *NAACL-HLT*, 4830–4842.

Erkan, G.; and Radev, D. R. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *J. Artif. Intell. Res.*, 22: 457–479.

He, X.; and Deng, L. 2017. Deep Learning for Image-to-Text Generation: A Technical Overview. *IEEE Signal Process. Mag.*, 34(6): 109–116.

Hinton, G. E.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *CoRR*, abs/1503.02531.

Jia, R.; Cao, Y.; Shi, H.; Fang, F.; Yin, P.; and Wang, S. 2021. Flexible Non-Autoregressive Extractive Summarization with Threshold: How to Extract a Non-Fixed Number of Summary Sentences. In *AAAI*, 13134–13142.

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *ICML*, volume 139, 5583–5594.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*, 7871–7880.

Li, H.; Zhu, J.; Ma, C.; Zhang, J.; and Zong, C. 2017. Multimodal Summarization for Asynchronous Collection of Text, Image, Audio and Video. In *EMNLP*, 1092–1102.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, Y.; and Lapata, M. 2019. Text Summarization with Pretrained Encoders. In *EMNLP-IJCNLP*, 3728–3738.

Nallapati, R.; Zhai, F.; and Zhou, B. 2017. SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. In *AAAI*, 3075–3081.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139, 8748–8763.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6): 1137–1149.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *ACL*, 1073–1083.

Srihari, R. K. 1994. Computational Models for Integrating Linguistic and Visual Information: A Survey. *Artif. Intell. Rev.*, 8(5-6): 349–369.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; Klingner, J.; Shah, A.; Johnson, M.; Liu, X.; Kaiser, L.; Gouws, S.; Kato, Y.; Kudo, T.; Kazawa, H.; Stevens, K.; Kurian, G.; Patil, N.; Wang, W.; Young, C.; Smith, J.; Riesa, J.; Rudnick, A.; Vinyals, O.; Corrado, G.; Hughes, M.; and Dean, J. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR*, abs/1609.08144.

Zhang, Z.; Wang, J.; Sun, Z.; and Yang, Z. 2021. LAMS: A Location-aware Approach for Multimodal Summarization (Student Abstract). In *AAAI*, 15949–15950.

Zhou, B.; Khosla, A.; Lapedriza, À.; Oliva, A.; and Torralba, A. 2016. Learning Deep Features for Discriminative Localization. In *CVPR*, 2921–2929.

Zhu, J.; Li, H.; Liu, T.; Zhou, Y.; Zhang, J.; and Zong, C. 2018. MSMO: Multimodal Summarization with Multimodal Output. In *EMNLP*, 4154–4164.

Zhu, J.; Zhou, Y.; Zhang, J.; Li, H.; Zong, C.; and Li, C. 2020. Multimodal Summarization with Guidance of Multimodal Reference. In *AAAI*, 9749–9756.