Leveraging Entity Information for Cross-Modality Correlation Learning: The Entity-Guided Multimodal Summarization

Anonymous ACL submission

Abstract

The rapid increase in multimedia data has spurred advancements in Multimodal Summarization with Multimodal Output (MSMO), which aims to produce a multimodal summary that that integrates both text and relevant images. The inherent heterogeneity of content within multimodal inputs and outputs presents a significant challenge to the execution of MSMO. Traditional approaches typically adopt a holistic perspective on coarse image-text data or individual visual objects, overlooking the essential connections between objects and the entities they represent. To integrate the fine-grained entity knowledge, we propose an Entity-Guided Multimodal Summarization model (EGMS). Our model, building on BART, utilizes dual multimodal encoders with shared weights to process text-image and entity-image information concurrently. A gating mechanism then combines visual data for enhanced textual summary generation, while image selection is refined through knowledge distillation from a pre-trained vision-language model. Extensive experiments on public MSMO dataset validate the superiority of the EGMS method, which also prove the necessity to incorporate entity information into MSMO problem.

1 Introduction

With the rapid development of multimedia content across the Internet, the task of Multimodal Summarization with Multimodal Output (MSMO) has emerged as a research direction of considerable significance (Zhu et al., 2018, 2020; Mukherjee et al., 2022; Zhang et al., 2022b,a), especially for news content summary (Zhu et al., 2018). Specifically, as shown in Figure 1, given the source text and corresponding images, MSMO aims to produce a multimodal summary with a textual abstract alongside a pertinent image. Instead of providing exclusively text-based summaries, MSMO considers and generates more diverse multimodal information, which constitutes a significant research but also puts high challenges for the interaction between text and images (Zhu et al., 2020).

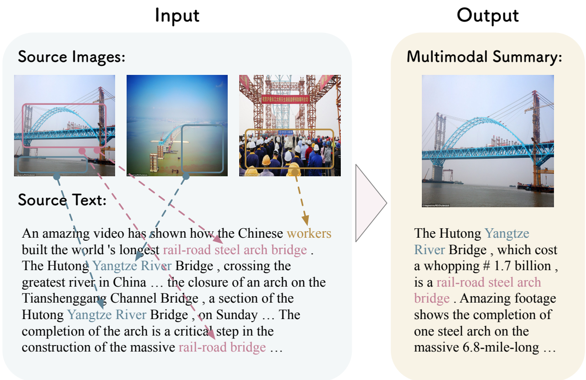

Figure 1: An example of entity-object correlations in multimodal data from MSMO problem. Entities *rail-road steel arch bridge* and *Yangtze River* correspond with elements in the associated images, suggesting inherent cross-modality correlations.

Since Zhu et al. (2018) proposed the MSMO task and collected the first large-scale English corpus, there has been a surge of research in academia exploring this area. However, most of the existing methodologies (Zhu et al., 2018, 2020; Mukherjee et al., 2022; Zhang et al., 2022b) integrated comprehensive image and text data without allocating explicit attention to discrete constituents within these modalities. Zhang et al. (2022a) have made strides in enhancing the domain by facilitating interactions between textual components at the granular word level and discrete objects in visual content. Nonetheless, these visual objects tend to relate to entity-level content in text rather than individual words. For example, from Figure 1, we can see that multi-word entities *rail-road steel arch bridge* and *Yangtze River* within the textual corpus exhibit correspondence with elements depicted in the accompanying images. The explicit extraction of these entities is posited to enhance comprehension of the image content. However, to the best of our knowledge, there are few works focusing on incorporating entity information into MSMO problem.

Indeed, there are many technical challenges inherent in designing effective solutions to incorporate entity information into MSMO process. The first of these pertains to the heterogeneity of the data involved, which can be textual, pictorial, or entity-based. This diversity imposes significant hurdles in attaining efficient cross-modality interaction. Second, the traditional frameworks employed for text decoding are predominantly designed to process purely textual inputs, thus creating a conundrum when the need arises to incorporating multimodal data into the decoding procedure. Third, the task of image selection, which tends to operate independently, frequently suffers from an absence of adequate labeling information, as there are no golden labels in the training set.

To tackle the above challenges, we propose an Entity-Guided Multimodal Summarization model (EGMS). Similar to UniMS (Zhang et al., 2022b), our study employs the BART framework as the foundational architecture for our model development. Specifically, we reconfigure the architecture of BART's text-centric encoder to establish a Shared Multimodal Encoder. It incorporates a pair of multimodal encoders with shared parameter weights, designed to model textual and visual data alongside entity-specific visual information. For the decoding process, we design a Multimodal Guided Text Decoder. It first employs a gated image fusion module to effectively merge the image representations that have been enriched with disparate modal information, and further utilizes the multimodal information for text generation. Subsequently, we introduce a Gated Knowledge Distillation module, which serves to harness the expertise of a pre-trained vision-language model, functioning as an auxilary guide for the learning process of image selection. Finally, we conduct extensive experiments on public MSMO datasets, where the experimental results demonstrate the effectiveness of our proposed EGMS method. Our code is available via `https://github.com/AnonymousEGMS/EGMS`.

2 Related Work

2.1 Multimodal Summarization

Multimodal summarization (UzZaman et al., 2011) is defined as a task that aims at distilling concise and precise syntheses from heterogeneous data sources, encompassing textual, visual, and audio content, etc. Research endeavors (Chen and Zhuge, 2018; Li et al., 2018; Zhang et al., 2021) have predominantly concentrated on the incorporation of supplementary and ancillary modal information to augment the depiction of a solitary modality. For example, Li et al. (2018) design image filters with the intent to selectively harness visual information, thereby augmenting the semantic richness of the input sentence.

Recently, there has been a burgeoning interest in the domain of multimodal summarization with multimodal output (MSMO). Zhu et al. (2018) construct the first large-scale corpus MSMO for this novel summarization task, which integrates textual and visual inputs to produce a comprehensive pictorial summary. They also propose a multimodal attention framework to jointly synthesize textual summary and select the most relevant image. Then Zhu et al. (2020) introduce a novel evaluation metric that integrates multimodal data to better combine visual and textual content during both the training and assessment stages. Mukherjee et al. (2022) and Zhang et al. (2022b) propose to solve the multimodal summarization task in a multitask training manner. And Zhang et al. (2022a) adopt a graph network and a hierarchical fusion framework to learn the intra-modal and inter-modal correlations inherent in the multimodal data respectively.

2.2 Knowledge Graph Augmented Models

Knowledge Graphs (KGs) store and organize information about different things and how they relate to each other in a structual way. World knowledge is commonly expressed using fact triplets, which consist of three elements: the subject entity, the relation, and the object entity denoted as (h, r, t). Since the introduction of TransE (Bordes et al., 2013), a multitude of knowledge graph embedding techniques (Ji et al., 2015; Zhong et al., 2015; Shi and Weninger, 2017) have emerged, aiming to translate the entities and relationships within these graphs into numerical vectors so that they can be easily applied to various downstream tasks.

Zhang et al. (2019) and Chen et al. (2019) leverage external knowledge graphs to enhance the textual content for improved performance in text classification tasks. Moreover, Hu et al. (2022) concentrate on the integration of external knowledge into the verbalizer mechanism to enhance the effectiveness and stability of prompt tuning for zero and few-shot text classification tasks. Yu et al. (2022) improve Fusion-in-Decoder (Izacard and

Grave, 2021) by employing a knowledge graph to establish the structural interconnections among the retrieved passages in Open-Domain Question Answering (ODQA) problem, achieving comparable or better performance with a much lower computation cost. And Kale et al. (2023) construct a Chest X-Ray knowledge graph then use it for radiology report generation.

3 Preliminary

3.1 Problem Formulation

Given a multimodal input $D = \{T, P\}$, where $T = \{t_1, t_2, ..., t_L\}$ is a sequence of L tokens of the article text and $P = \{p_1, p_2, ..., p_M\}$ is the collection of the M in-article images, our proposed model first extracts all the entities $K = \{k_1, k_2, ..., k_N\}$ in the article text and then summarizes $\{D, E\}$ into a multimodal summary $S = \{S_t, S_i\}$. $S_t = \{s_1, s_2, ..., s_l\}$ denotes the textual summary limited by a max length of l. The pictorial summary S_i is an extracted subset of the image input P.

3.2 BART Architecture

BART (Bidirectional and Auto-Regressive Transformers) (Lewis et al., 2020) functions as a denoising autoencoder, designed to reconstruct an original document from its corrupted counterpart.

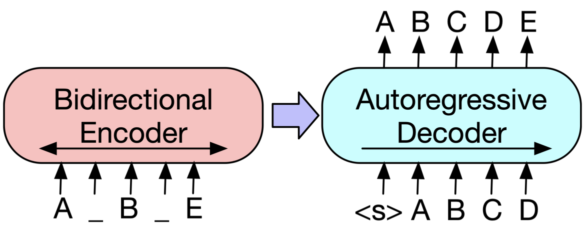

Figure 2: BART architecture from Lewis et al. (2020).

As shown in Figure 2, it uses a standard Transformer-based neural machine translation architecture, incorporating a bidirectional encoder, coupled with a left-to-right autoregressive decoder. In the process of optimizing BART for text generation applications, the source text is initially fed into the encoder module. Following this, the desired output text, which is prepended with the decoder's designated initial token, is introduced to the decoder module.

4 Model

4.1 Model Overview

We propose a novel multimodal summarization framework enhanced by an external knowledge graph, as shown in Figure 3. Building upon the BART architecture, our model has been adapted to accommodate multimodal inputs, specifically textual and visual data. Recognizing that images often depict objects which correspond to real-world entities, our approach seeks to leverage this multimodal data more effectively. To this end, we utilize an external knowledge graph to extract entities from the textual content, which in turn facilitates a better interpretation of the visual information. This integration aims to improve the coherence and richness of the generated summaries by bridging the semantic gap between the textual and visual modalities.

4.2 Shared Multimodal Encoder

Text-Image Encoder Given the inherent restriction of BART's context length, capped at 1024 tokens, it is imperative to deliberate on the regulation of image input dimensions to ensure compatibility with the model's processing capabilities. Following Li et al. (2023), we use a frozen Q-Former to transform image features $r_i^{|IE| \times d_{IE}}$, which are derived from a frozen image encoder, into a fixed number of output features $v_i^{|Q| \times d_Q}$, each corresponding to a predefined learned query q:

$$\begin{aligned} r_i &= [r_{i,1}, r_{i,2}, ..., r_{i,|IE|}] = f_{img-enc}(p_i), \\ v_i &= [v_{i,1}, v_{i,2}, ..., v_{i,|Q|}] \\ &= f_{Q-Former}(q_1, q_2, ..., q_{|Q|}; r_i), \end{aligned} \tag{1}$$

Then, we enhance the textual encoding capabilities of BART by transitioning to a multimodal encoding framework. For text-image encoder, this involves the integration of textual embeddings, denoted as e_t, with corresponding visual embeddings e_v. The concatenated embeddings serve as input to the encoder f_{ti-enc}, which then yields contextualized representations:

$$\begin{aligned} e_t &= W_t \cdot [t_{CLS}, t_1, t_2, ..., t_L, t_{SEP}], \\ e_{v_i} &= [v_{CLS}, W_v \cdot v_i] + e_{intra-pos}, \\ e_{ti} &= [e_t, e_v] + e_{multi-pos} \\ &= [e_t, e_{v_1}, ..., e_{v_M}] + e_{multi-pos}, \\ h_{ti} &= [h_{T_{ti}}, h_{V_{ti}}] = f_{ti-enc}(e_{ti}), \end{aligned} \tag{2}$$

where special tokens t_{CLS} and t_{SEP} serve as delimiters to denote the start and end of each sentence respectively. The embeddings $e_{intra-pos}$ and $e_{multi-pos}$ represent the intra-image positional information and the multimodal positional context within the framework. The matrices W_t and W_v

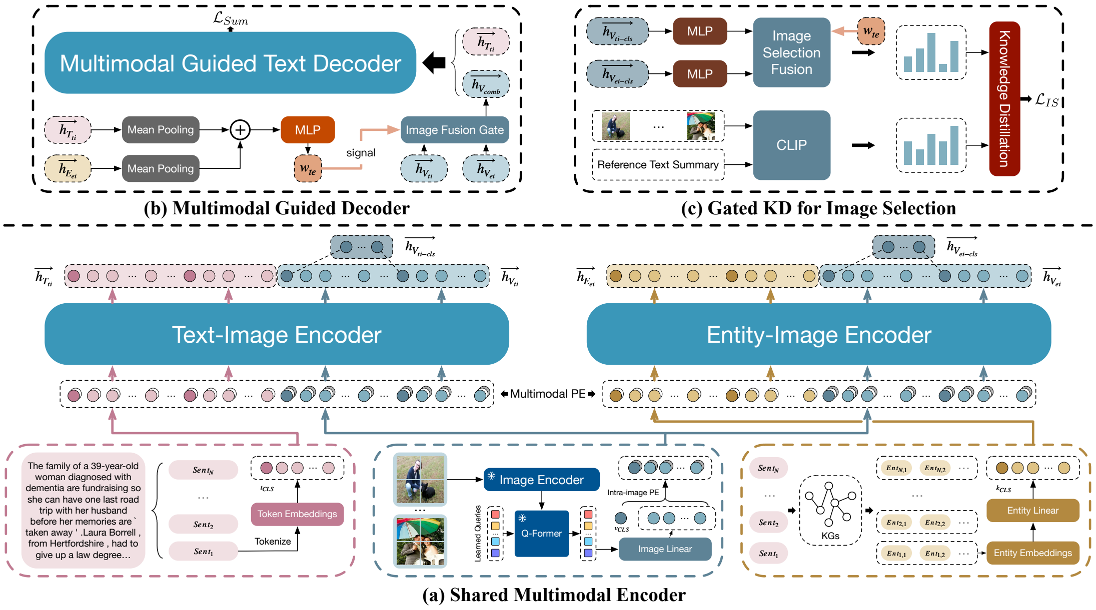

Figure 3: The architecture of our proposed EGMS model. It consists of three parts: (a) Shared Multimodal Encoder; (b) Multimodal Guided Decoder; (c) Gated Knowledge Distillation for Image Selection.

are employed for embedding linguistic tokens and projecting image features into a shared multimodal space respectively. Following Dosovitskiy et al. (2021) and Zhang et al. (2022b), we add a learnable special token, represented by the embedding vector v_{CLS}, to signify the initiation of an image sequence. The corresponding encoded state at the output of the encoder is then utilized as a holistic representation of the image.

Entity-Image Encoder For the reasons already explained in the introduction, we propose to incorporate entity-level information to enhance the exploitation of multimodal data.

First, we extract entities from the text utilizing an external knowledge graph. For the clarity and simplicity, we adopt the classical TransE model (Bordes et al., 2013) to obtain a representation of the entities in the knowledge graph, which contains intricate structural relationships among the entities. Similar to the text-image encoder, the entity embeddings e_e concatenated with visual embeddings e_v are subsequently processed by the entity-image encoder f_{ei-enc}, yielding an enriched image representation that encapsulates augmented entity-specific information:

$$\begin{aligned} e_e &= W_{e_2} \cdot W_{e_1} \cdot [k_{CLS}, k_1, k_2, ..., k_M], \\ e_{ei} &= [e_e, e_v] + e_{multi-pos}, \\ h_{ei} &= [h_{E_{ei}}, h_{V_{ei}}] = f_{ei-enc}(e_{ei}), \end{aligned} \tag{3}$$

where k_{CLS} is used to demarcate sequences of entities contained in discrete sentences. The matrix W_{e_1} represents the embedding matrix for entities, which is initialized utilizing embeddings derived from the pre-trained TransE model. Concurrently, the matrix W_{e_2} is employed to project entity features into a unified multimodal space for further integration of modalities. Notably, this encoder shares its parameter weights with the aforementioned text-image encoder.

4.3 Multimodal Guided Decoder

Gated Image Fusion To integrate the visual representations derived from dual encoders, each amalgamated with textual and entity-based information respectively, we introduce a gated image fusion module. Visual information integrated with textual and entity representations from the respective encoders will be merged together:

$$h_{te} = Mean(h_{T_{ti}}) \oplus Mean(h_{E_{ei}}), \tag{4}$$

where $\oplus$ is the concatenation operation.

Then h_{te} will serve as the input for a weight computation module, which is designed to quantitatively assess the salience of the visual representations in conjunction with corresponding multimodal inputs:

$$w_{te} = \sigma_w^2(W_w^2 \cdot \sigma_w^1(W_w^1 \cdot h_{te} + b_w^1) + b_w^2), \tag{5}$$

where σ_w^2 is $Sigmoid$ activation function, making the value of this weight between 0 and 1.

Subsequently, the derived weight w_{te} serves as the signal to control the fusion of dual image representations that encapsulate different modal information, yielding an augmented image representation that is enriched with both textual and entity information:

$$h_{V_{comb}} = w_{te} \cdot h_{V_{ti}} + (1 - w_{te}) \cdot h_{V_{ei}}. \quad (6)$$

Multimodal Guided Text Decoder Similar to BART, the architecture of our model incorporates a conventional autoregressive transformer decoder within its decoding module. In contrast to relying exclusively on textual representations during the encoding phase, our proposed model also utilizes the aforementioned augmented image representations. These representations serve as encoder hidden states that are subsequently fed into the decoder:

$$h_{enc-hid} = [h_{T_{ti}}, h_{V_{comb}}]. \quad (7)$$

The decoder attends to the sequence of previously generated tokens, denoted as $s_{<j}$, as well as the encoder output hidden states $h_{enc-hid}$, and predicts the conditional probability distribution of subsequent text tokens. So for the abstractive summarization task, our model is trained by minimizing the negative log-likelihood:

$$\mathcal{L}_{Sum} = -\sum_{j=1}^{|S|} \log p(s_j | s_{<j}, h_{enc-hid}, \phi), \quad (8)$$

where ϕ denotes all the parameters of the model.

4.4 Gated Knowledge Distillation for Image Selection

In the current multimodal summarization dataset, only the test set has visual references, which could be instrumental in guiding the selection of salient images during the training phase.

Zhang et al. (2022b) propose to adopt Knowledge Distillation (KD) technique (Hinton et al., 2015) to distill the inherent relevance between textual and visual information, which can get image references without any image captions or visual references. Rather than using only the text-integrated image representations as Zhang et al. (2022b), we incorporate entity information as well. Specifically, we use the output hidden states of v_{CLS} derived from both encoders as comprehensive image representations and feed them to two distinct multi-layer perceptrons to obtain scores:

$$\begin{aligned} g_{ti}(p) &= W_t^2 \cdot \sigma_t^1(W_t^1 \cdot h_{v_{ti-cls}} + b_t^1) + b_t^2, \\ g_{ei}(p) &= W_e^2 \cdot \sigma_e^1(W_e^1 \cdot h_{v_{ei-cls}} + b_e^1) + b_e^2. \end{aligned} \quad (9)$$

And we combine them with the weight calculated in Eq.(5) to futher utilize multimodal information:

$$g(p) = w_{te} \cdot g_{ti}(p) + (1 - w_{te}) \cdot g_{ei}(p). \quad (10)$$

We employ CLIP (Radford et al., 2021) as the teacher model to calculate the similarity scores between each image p and the textual summary S_t:

$$l(S_t, p) = sim(\mathcal{T}(S_t), \mathcal{V}(p)), \quad (11)$$

where $\mathcal{T}$ and $\mathcal{V}$ are its textual and visual encoder respectively, and sim is the cosine similarity function.

Through knowledge distillation, our model is intended to emulate the score distribution of the teacher model. By using Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951), this approach can be modeled as minimizing the following objective function with temperature τ:

$$\mathcal{P}_p(p, \tau) = \frac{\exp(\frac{g(p)}{\tau})}{\sum_{p \in P} \exp(\frac{g(p)}{\tau})}, \quad (12)$$

$$\mathcal{Q}_p(S_t, p, \tau) = \frac{\exp(\frac{l(S_t,p)}{\tau})}{\sum_{p \in P} \exp(\frac{l(S_t,p)}{\tau})}, \quad (13)$$

$$\mathcal{L}_{IS} = KL(\mathcal{P} || \mathcal{Q}) = -\sum_{p \in P} \mathcal{P}_p \cdot \ln \frac{\mathcal{Q}_p}{\mathcal{P}_p}. \quad (14)$$

4.5 Training

Inspired by Li et al. (2023), we divide the training process of our proposed model into two main stages: an initial phase dedicated to aligning the modalities of images and text, followed by a subsequent phase focusing on fine-tuning.

Modal Matching In the modal matching phase, parameter optimization is confined to the weights of the image feature projection matrix W_v, and the embedding v_{CLS} of the visual initiation token. This targeted approach leverages the text-image encoder and the decoder exclusively, thereby enhancing the model's focus on the pertinent multimodal information while alleviating the impact of other information. The training process is governed by minimizing the negative log-likelihood:

$$\begin{aligned} \mathcal{L} &= -\sum_{j=1}^{|S|} \log p(s_j | s_{<j}, h_{ti}, \phi), \\ &= -\sum_{j=1}^{|S|} \log p(s_j | s_{<j}, [h_{T_{ti}}, h_{V_{ti}}], \phi). \end{aligned} \quad (15)$$

Statistics	Train	Valid	Test
#Samples	293,965	10,355	10,261
#AvgTokens(A)	720.87	766.08	730.80
#AvgTokens(S)	70.12	70.02	72.16
#AvgImgs	6.56	6.62	6.97

Table 1: The data statistics of MSMO dataset. **#AvgTokens(A)** and **#AvgTokens(S)** denote the average number of tokens in articles and reference summaries respectively.

Fine-tuning In the fine-tuning phase, the model parameters are initially set using the weights obtained from the modal matching stage. Subsequently, the entire proposed framework is employed, with adjustments made to all learnable parameter weights. The training loss of our model is a sum of the objectives of image selection and abstractive text summarization:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{IS} + \mathcal{L}_{Sum}, \tag{16}$$

where α is a hyper-parameter that modulates the salience of the image selection loss within the total training loss.

5 Experiment

5.1 Experiment Setup

Datasets For multimodal summarization with multimodal output, we use the MSMO dataset, which is introduced by Zhu et al. (2018). This is the first and only large-scale English corpus specifically curated for this task. It comprises a collection of online news articles sourced from $DailyMail$ website[1], each accompanied by several images and corresponding manually-written highlights that serve as the reference summary. More statistics about the dataset are illustrated in Table 1. Within the test set, a maximum of three images are annotated to provide a pictorial reference.

Evaluation Metrics In text summarization tasks, the evaluation of summary quality usually employs the **ROUGE** metric (Lin, 2004), which quantifies the degree of lexical correspondence between the produced sentences and the reference summaries. All the ROUGE scores in this paper refer to the F-1 ROUGE scores calculated by official script. In addition, Zhu et al. (2018) introduce the metric of image precision (**IP**) to assess the quality of the output image, delineating the methodology as follows:

$$\text{IP} = \frac{|\{ref_{img}\} \cap \{rec_{img}\}|}{|\{rec_{img}\}|}, \tag{17}$$

where ref_{img} and rec_{img} denote the reference images and the recommended ones.

Implementation Details Our model utilizes the released checkpoint[2] of a BART-like model, BRIO (Liu et al., 2022), to initialize corresponding parameters. And we take released CLIP model (Radford et al., 2021)[3] as the teacher model for image selection knowledge distillation. For the image processing, we employ the vision feature extractor of BLIP-2 (Li et al., 2023)[4] to get visual features. The number of the learned queries is set to 32, resulting in an allocation of 33 token positions within the encoder for each image. And we set the upper limit of image number to 8. Noting that we concatenate multimodal tokens together as the input for two dual-modal encoders, the maximum number of textual and entity tokens is constrained by the encoder's maximum context length as well as the length of the image sequence.

The train set of MSMO dataset is partitioned into 20 discrete subsets. Therefore, we employ a cumulative training strategy, wherein the model undergoes iterative training on each subset in succession. After training on each subset, the model's parameters are saved as checkpoints and evaluated on validation set. We identify the top-3 checkpoints as determined by the minimal validation loss. Subsequently, we compute and present the mean results derived from these checkpoints on test set.

In the process of image selection, we choose the image with greatest score as computed in Eq.(10). And for text summarization, we use beam search with a beam size of 5 in decoding.

Baseline Models To demonstrate the efficacy of the proposed model, we conduct comparative analyses with extant methodologies in both text-based and multimodal summarization domains:

- **BertSum** (Liu and Lapata, 2019) uses a general framework for both extractive and

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

[2] https://huggingface.co/Yale-LILY/brio-cnndm-uncased

[3] https://huggingface.co/openai/clip-vit-base-patch32

[4] https://github.com/salesforce/LAVIS/tree/main/projects/blip2

abstractive text summarization, with its encoder based on BERT (Kenton and Toutanova, 2019). It has several raviants, out of which **BertAbs** and **BertExtAbs** can be used for abstractive text summarization.

- **BART** (Lewis et al., 2020), constructed as a denoising autoencoder, employs a sequence-to-sequence framework with significant applicability in the domain of text summarization.

- **ATG/ATL/HAN** utilizes a pointer-generator network (See et al., 2017) and a multimodal attention mechanism, with variants reflecting different image representation approaches for attention operations.

- **MOF**^{RR} (Zhu et al., 2020) ranks images via ROUGE score comparison of captions to textual reference, forming a visual reference. Variants of incorporating different hidden states into image discriminator are denoted as **MOF**$^{RR}_{enc}$ and **MOF**$^{RR}_{dec}$.

- **UniMS** (Zhang et al., 2022b) proposes to merge textual and visual data to BART (Lewis et al., 2020) encoder to construct a multimodal representation. Subsequently, it employs a visually guided decoder to integrate textual and visual modalities in guiding abstractive text generation.

5.2 Experimental Result

As shown in Table 2, our proposed EGMS method outperforms all baselines in all metrics, which proves the effectiveness of our method and the necessity to incorporate knowledge graphs.

The outcomes of this study reveal a number of intriguing phenomena: (1) By fine-tuning BART for summarization task, it can achieve competitive results with models that introduce visual information. This proves that BART exhibits robust language modeling proficiencies, thereby indicating its substantial potential for applications in multimodal information modeling. The findings herein reinforce the rationale for its deployment in our modeling endeavors. (2) The UniMS framework, also based on BART model, has shown great improvements, especially in ROUGE-2 and ROUGE-L scores. This advancement suggests that the integration of visual data facilitates the model's capacity to process and interpret extended text sequences, surpassing the merely word-level analyses. Such findings are consistent with our initial hypothesis, which postulates that the incorporation of entity-level information rather than word-level would yield a more robust understanding of the multimodal data.

Model	R-1	R-2	R-L	IP
Text Abstractive				
BertAbs*	39.02	18.17	33.20	-
BertExtAbs*	39.88	18.77	38.36	-
BART	42.93	19.95	39.97	-
Multimodal Abstractive				
ATG*	40.63	18.12	37.53	59.28
ATL*	40.86	18.27	37.75	62.44
HAN*	40.82	18.30	37.70	61.83
MOF$^{RR}_{enc}$*	41.05	18.29	37.74	62.63
MOF$^{RR}_{dec}$*	41.20	18.33	37.80	65.45
UniMS*	42.94	20.50	40.96	69.38
EGMS	**44.47**	**21.20**	**41.43**	**75.81**

Table 2: Experimental results for multimodal summarization on MSMO dataset. Results marked by * are taken from respective papers and Zhang et al. (2022b).

Model	R-1	R-2	R-L	IP
EGMS	**44.47**	**21.20**	**41.43**	**75.81**
-w/o IS	44.25	21.05	41.21	-
-w/o EI	44.29	21.10	41.22	75.65
-w/o TI	44.35	21.07	41.35	62.88

Table 3: Ablation experiments on MSMO dataset. 'IS' stands for Image Selection module. 'EI' and 'TI' refer to the encoded visual representations derived from Entity-Image Encoder and Text-Image Encoder respectively.

5.3 Ablation Study

In this subsection, we conduct ablation experiments to prove the effectiveness of different components of EGMS model. We remove Image Selection (IS) module, image representations derived from Entity-Image Encoder (EI) and Text-Image Encoder (TI) respectively. More specifically, by removing Image Selection module, we reduce MSMO problem to a multimodal summarization task with only textual output. Removing 'EI' means that we only use the encoded visual representations from Text-Image Encoder for summary generation and iamge selection. To elaborate, the weight w_{te} from Eq.(5) is fixed to 1. Likewise, when removing 'TI', reliance

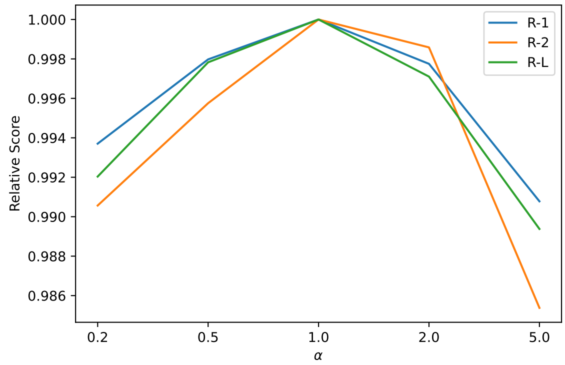

Figure 4: Hyperparameter study on MSMO dataset. The results in the graph are normalized by the result of the corresponding metric with $\alpha = 1.0$.

is exclusively placed on the visual representations from Entity-Image Encoder, with the corresponding weight being constrained to 0.

The results are listed in Table 3. Analysis of the data reveals a consistent decline across all ablation variants, thereby demonstrating the validity and non-redundancy of our proposed EGMS method. Besides, we can find that the entity information predominantly enhances the capacity of the model to generate concise summaries, while the improvement of the model's image selection accuracy is smaller. This differential impact suggests that comprehensive textual data may suffice for the selection of pertinent images. However, the integration of additional entity information can have an advantage in the precise identification of salient components, aligning well with the core requirements of the summarization task.

5.4 Parameter Sensitivity

To study the impact of the loss hyperparameter α in EGMS, a series of parameter sensitivity analyses were performed on the MSMO dataset. The results are reported in Figure 4. $\alpha = 1.0$ is the best hyperparameter of our model. From the results, we can see that larger or smaller α will lead to decrease on the summarization performance. This is reasonable as the hyperparameter controls the weight of the Image Selection loss in the total loss. A large weight will affect the Abstractive Summarization loss, while a small weight reduces the usefulness of the text-image multimodal knowledge aids learned from the teacher model in modeling multimodal information.

Model	Text		Image
	Coherence	Relevance	Relevance
BART	3.47	3.22	-
EGMS	**4.20**	**4.02**	**3.66**
-w/o IS	3.75	3.64	-
-w/o EI	3.84	3.64	3.53
-w/o TI	3.84	3.67	3.45

Table 4: Human evaluation of different model outputs.

5.5 Human Evaluation

To further evaluate our models performance, we randomly select 120 data samples from test set for human evaluation. Subsequently, three graduate students are enlisted to evaluate them on a scale ranging from one to five, addressing various qualitative aspects. For abstractive text summarization, *coherence* measures whether the summary is smooth and fluent. And *relevance* assesses the extent to which the summary content corresponds with the information presented in the original document. For image selection, *relevance* indicates the text-image relevance of the multimodal summary. Table 4 indicates that our method can generate more coherent and relevant summaries compared to other variants and baselines.

6 Conclusions

In this paper, we propose an Entity-Guided Multimodal Summarization model (EGMS), that incorporates entity-specific information into solving MSMO problem. Based on BART, our model introduces a pair of multimodal encoders with shared weights to concurrently process text-image and entity-image information. Subsequently, a gating mechanism is used to fuse the visual representations, which will further be utilized in the generation of textual summaries. As for image selection, we also use a gating mechanism and distill knowledge from a pre-trained vision-language model. Extensive experiments on public MSMO dataset demonstrat the effectiveness of our proposed method. We hope our work could lead to more future studies in this field.

7 Limitations

In our proposed EGMS method, incorporating the knowledge graph requires the entity recognition process, which will consume additional time compared with other MSMO methods. And if we need to use other domains' knowledge graphs, it will be requisite to undertake retraining of the entity representations and the model. However, by utilizing a general-purpose knowledge graph, our model can be applied in most scenarios. Another limitation is that since the MSMO dataset is labeled with pictorial references only on the test set, we adopt a method that utilizes knowledge distillation for image selection learning. And the results of such an approach can be affected by the performance of the teacher vision-language model.

References

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Jindong Chen, Yizhou Hu, Jingping Liu, Yanghua Xiao, and Haiyun Jiang. 2019. Deep short text classification with knowledge powered attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6252–6259.

Jingqiang Chen and Hai Zhuge. 2018. Abstractive text-image summarization using multi-modal attentional hierarchical RNN. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4046–4056, Brussels, Belgium. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. 2022. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2225–2240, Dublin, Ireland. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 687–696.

Kaveri Kale, Pushpak Bhattacharyya, Milind Gune, Aditya Shetty, and Rustom Lawyer. 2023. KGVL-BART: Knowledge graph augmented visual language BART for radiology report generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3401–3411, Dubrovnik, Croatia. Association for Computational Linguistics.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

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

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

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, Chengqing Zong, et al. 2018. Multi-modal sentence summarization with modality attention and image filtering. In *IJCAI*, pages 4152–4158.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

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

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Sourajit Mukherjee, Anubhav Jangra, Sriparna Saha, and Adam Jatowt. 2022. Topic-aware multimodal summarization. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 387–398, Online only. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Baoxu Shi and Tim Weninger. 2017. Proje: Embedding projection for knowledge graph completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Naushad UzZaman, Jeffrey P Bigham, and James F Allen. 2011. Multimodal summarization of complex sentences. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pages 43–52.

Donghan Yu, Chenguang Zhu, Yuwei Fang, Wenhao Yu, Shuohang Wang, Yichong Xu, Xiang Ren, Yiming Yang, and Michael Zeng. 2022. KG-FiD: Infusing knowledge graph in fusion-in-decoder for open-domain question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4961–4974, Dublin, Ireland. Association for Computational Linguistics.

Chenxi Zhang, Zijian Zhang, Jiangfeng Li, Qin Liu, and Hongming Zhu. 2021. Ctnr: Compress-then-reconstruct approach for multimodal abstractive summarization. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. 2019. Integrating semantic knowledge to tackle zero-shot text classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1031–1040, Minneapolis, Minnesota. Association for Computational Linguistics.

Litian Zhang, Xiaoming Zhang, and Junshu Pan. 2022a. Hierarchical cross-modality semantic correlation learning model for multimodal summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11676–11684.

Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2022b. Unims: A unified framework for multimodal summarization with knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11757–11764.

Huaping Zhong, Jianwen Zhang, Zhen Wang, Hai Wan, and Zheng Chen. 2015. Aligning knowledge and text embeddings by entity descriptions. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 267–272.

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

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Multimodal summarization with guidance of multimodal reference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9749–9756.

A Multimodal Summary Sample

Figure 5: An example of multimodal summary.

To better show the effectiveness of our proposed EGMS method, we illustrate a case study in Figure 5. From this figure, we can find that our model accurately recognizes the entity *smelling salts*. And each image in the source input contains information about it. When considering a word-level approach, the isolated word *salts* is not able to get the corresponding meaning accurately. However, the incorporation of entity-level information allows for an enhanced understanding of the correlations between textual data and visual elements, thereby improving the model's capacity for multimodal learning.