Vision Enhanced Generative Pre-trained Language Model for Multimodal Sentence Summarization

Liqiang Jing[1], Yiren Li[2], Junhao Xu[1], Yongcan Yu[1], Pei Shen[2] and Xuemeng Song[1]

[1]School of Science and Technology, Shandong University, Qingdao 266237, China.
[2]HBIS Digital Technology Co., Ltd, Shijiazhuang 050035, China.

Abstract

Multimodal sentence summarization (MMSS) is a new yet challenging task, which aims to generate a concise summary of a long sentence and its corresponding image. Although existing methods have gained promising success in MMSS, they overlook the powerful generation ability of generative pre-trained language models (GPLMs), which have shown to be effective in many text generation tasks. To fill this research gap, we propose to make use of GPLMs to promote the performance of MMSS. Notably, adopting GPLMs to solve MMSS inevitably faces two challenges: (1) what fusion strategy should we use to inject visual information into GPLMs properly? And (2) how to keep the GPLM's generation ability intact to the utmost extent when the visual feature is injected into the GPLM. To address these two challenges, we propose a vision enhanced generative pre-trained language model for MMSS, dubbed as Vision-GPLM. In Vision-GPLM, we obtain features of visual and textual modalities with two separate encoders and utilize a text decoder to produce a summary. In particular, we utilize multi-head attention to fuse the features extracted from visual and textual modalities to inject the visual feature into the GPLM. Meanwhile, we train Vision-GPLM in two stages: the vision-oriented pre-training stage and fine-tuning stage. In the vision-oriented pre-training stage, we particularly train the visual encoder by the masked language model task while the other components are frozen, aiming to obtain homogeneous representations of text and image. In the fine-tuning stage, we train all the components of Vision-GPLM by the MMSS task. Extensive experiments on a public MMSS dataset verify the superiority of our model over existing baselines.

Keywords: Multimodal Sentence Summarization, Gnerative Pre-trained Language Model, Natural Language Generation

1 Introduction

Sentence summarization is a task that aims to generate a short summarization for a long sentence. Because of its wide applications, *e.g.,* news summarization and product summarization, this task has attracted much research attention. The early studies focus on the pure sentence summarization task, namely, producing a condensed summary from an input long sentence [1, 2]. Despite their promising performance, these efforts overlook visual modality information(i.e., the image). Visual modality allows readers to grasp the key information at a glance, conveying important cues regarding the core events. Therefore, a few pioneer studies [3, 4] resorted to multimodal sentence summarization (MMSS). As shown in Fig 1, the MMSS aims to generate a textual summary based on its multimodal contents, *e.g.,* the text content and image. Most existing works on MMSS

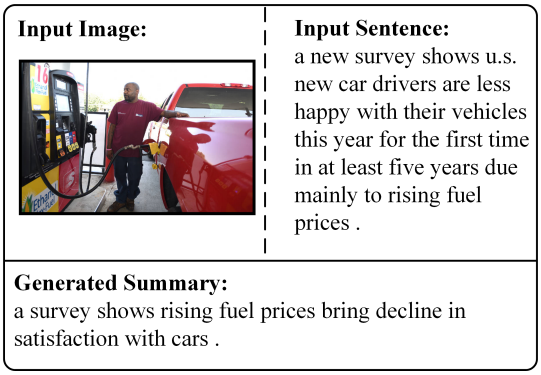

Fig. 1: Illustration of the task of multimodal sentence summarization.

employ the encoder-decoder framework for semantic understanding and text generation. For example, Li *et al.* [3] utilized the recurrent neural networks (RNNs) and convolutional neural networks (CNNs) as the textual encoder and visual encoder, respectively, and employed a textual decoder for multimodal sentence summarization.

Previous methods, however, follow the conventional train-from-scratch paradigm, overlooking the benefit of pre-training. In fact, the pre-training technique has shown its advance in a series of natural language processing (NLP) tasks. Several generative pre-trained language models (GPLMs) have shown excellent capability on language generation tasks, such as denoising autoencoder for pre-training sequence-to-sequence models [5] (BART) and Transfer Text-to-Text Transformer [6] (T5). Therefore, in this work, we aim to adapt GPLMs to promote the MMSS research line. Notably, we face two key challenges:

- **C1.** What fusion strategy should we use to inject visual information into GPLMs properly? GPLMs are trained on a text-to-text paradigm, and we need an effective fusion strategy to fuse visual and textual features.
- **C2.** How to keep GPLMs' generation ability intact to the utmost extent when the visual feature is injected into GPLMs? The input of multimodal data is heterogeneous, which may hurt the performance of GPLMs which are pre-trained on the pure textual modality.

To address these two challenges, we propose a vision enhanced generative pre-trained language model for multimodal sentence summarization: Vision-GPLM for short. As shown in Fig. 2, Vision-GPLM mainly consists of three components: multimodal feature extraction, multi-head attention based fusion, and text generation. Specifically, we first introduce a multi-head attention mechanism to fuse the visual representation to the GPLM to address the first challenge. The multi-head attention mechanism have shown its advance in many multimodal tasks [7, 8]. We then train the whole model in two stages: the vision-oriented pre-training stage and fine-tuning stage. In the vision-oriented pre-training stage, only the visual encoder is trained on the masked language model objective [9] while other components are fixed, aiming to obtain homogeneous representations of text and image. The fine-tuning stage is utilized to learn the task-aware knowledge to solve the MMSS task. To verify the effectiveness of our proposed model, we conduct extensive experiments on a publicly released dataset. The experimental results demonstrate that our model outperforms the state-of-the-art baselines.

Overall, our contributions can be concluded into three points:

- To the best of our knowledge, we are the first to adopt the GPLM to MMSS task. Furthermore, we incorporate the encoded visual feature into the GPLM through an advanced multi-head attention fusion strategy.
- To keep the GPLM's generation ability to the maximum extent, we train the model in two stages: the vision-oriented pre-training stage and fine-tuning stage.
- To justify the proposed model, we conduct extensive experiments on a widely used benchmark. The experimental results show that our model significantly outperforms the state-of-the-art baselines. As a byproduct, we release our source code to benefit the research community[1].

2 Related Work

Our work is related to sentence summarization, pre-trained language models, and image captioning.

[1] https://github.com/LiqiangJing/Vision-GPLM.

2.1 Sentence Summarization

Sentence summarization is one of the most common NLP tasks, and there are mainly two ways to summarize texts: *extraction sentence summarization* and *abstraction sentence summarization*. Extractive sentence summarization is extracting a subset of words from a sentence to represent the most significant aspects and combining them into a shorter sentence. Abstractive sentence summarization aims to generate a concise summary of the most important information of the long text by rephrasing or using the new words.

As abstractive sentence summarization can assist in overcoming the extraction techniques' grammatical inaccuracies and therefore produces better-quality summaries, recent works focus on abstractive sentence summarization. Early research mainly focused on generating the sentence summary based on the sequence-to-sequence (seq2seq) model. For example, Rush *et al.* [1] first presented a seq2seq model based on RNNs to generate a short summary for a long sentence. Based on this, Chopra *et al.* [2] further developed the seq2seq model equipped with a novel convolutional-attention based encoder for sentence summarization. In addition, Gu *et al.* [10] incorporated a copying mechanism into the seq2seq model to improve the fluency and accuracy of the generated summary. Despite their promising success, these methods overlook the visual modality, which also provides essential semantic cues and aids in sentence summary. To tackle this issue, some studies resorted to multimodal sentence summarization. For example, Li *et al.* [3] proposed a multimodal sentence summarization model which contained a modality-based attention mechanism for paying different attention to the input image and sentence. To grasp the highlights of the source sentence by the image, Li *et al.* [4] presented a multimodal selective gate network to filter away inconsequential information in the source sentence.

Although these methods have achieved remarkable success, they overlook the benefit of pre-training and train the model from scratch.

2.2 Pre-trained Language Models

Pre-training recently has shown its powerful ability for diverse NLP tasks, improving the model's performance for downstream tasks and reducing training costs. Word2vec [11] and GloVe [12] are examples of early pre-trained models which introduced shallow architecture to provide pre-trained word embeddings for downstream NLP tasks. Although the pre-trained word embeddings learned the semantic meaning of the word, they are context-free and hard to capture the semantic meaning of the whole sentence or document. With the advance of Transformer [13], increased research efforts have been committed to developing Transformer-based pre-trained models to capture context semantics. For example, Devlin et al. [9] pretrained the deep bidirectional encoder in Transformer (BERT) by two pre-training tasks: masked language model and next-sentence prediction. Despite its success in textual representation learning [14], BERT cannot be fine-tuned directly for language generation. Later, Lewis *et al.* [5] developed BART, which utlized the full Transformer architecture for natural language generation. Meanwhile, Raffel *et al.* [6] proposed T5 that transfers all NLP tasks to a "text-to-text" format and can be utilized for a variety of downstream NLP tasks such as document summarization [15] and paraphrase detection [16].

Due to the pre-trained language models having absorbed rich knowledge from large-scale corpus, many researchers have resorted to GPLMs to solve their specific tasks. For example, Yu *et al.* [8] developed vision guided generative pre-trained language models based on BART and T5 for multimodal video summarization task which summarizes videos into short texts based on their visual modalities and textual transcripts. Inspired by this, we also resorted to publicly released pre-trained language models to summarize sentence-image pairs into a short sentence.

2.3 Image Captioning

Image captioning aims to produce natural language description for a image. Early studies [17, 18] on image captioning firstly detected words from the image and then utilized predefined templates to convert detected words into a natural language sentence. These methods rely on templates and always generate similar sentence structure. Meanwhile, the search-oriented methods [19, 20] directly adopted the sentence of the similar image or selected a semantic similar

sentence from a sentences set to get the target sentence. Obviously, these methods are limited by the size of human-generated sentences set and can not generate a new sentence. Recently, with the development of deep learning, many works [21–26] utlized neural networks to learn the probability distribution in the common semantic space of visual content and textual content and generate a new sentence, and achieved state-of-the-art performance.

Despite the success of image cpationing methods mentioned above, they are not sitiable for the multimodal sentence summarization task because they can not tackle the texutal input.

3 Methodology

In this section, we first introdice the task formulation. Then, we detail the proposed Vision-GPLM.

3.1 Task Formulation

Suppose we have a set of N training triplets $\mathcal{D} = (X_1, V_1, Y_1), (X_2, V_2, Y_2), \cdots, (X_N, V_N, Y_N)$. $X_i = \{x_1^i, x_1^i, \cdots, x_{M_i}^i\}$ is the source sentence (*e.g.,* long news sentences), where x_j^i denotes the j-th token in the source text X_i. M_i refers to the total number of tokens which is a variable for different triplets. V_i is the image in the i-th triplet. $Y_i = \{y_1^i, y_2^i, \cdots, y_{O_i}^i\}$ stands for the target summary in the i-th triplet, where O_i denotes its total number of tokens. Based on these training triplets, our goal is to learn a multimodal sentence summarization model $\mathcal{M}$ which can generate a concise summary for the source sentence and image as follows,

$$Y = \mathcal{M}(X, V \mid \mathbf{\Theta}), \tag{1}$$

where $\mathbf{\Theta}$ stands for the parameters to be learned. For simplicity, we temporarily omit the index (i.e., the subscript i) of each training triplet.

3.2 Model Architecture

As shown in Fig. 2, the model architecture mainly consists of three components: multimodal feature extraction, multi-head attention based fusion, and text generation. As aforementioned, to utilize the powerful generation ability of the generative pre-trained language model, we resort to BART as our backbone for textual feature extraction and summary generation.

3.2.1 Multimodal Feature Extraction

We introduce the feature extraction of the multimodal input, *i.e.,* text feature extraction and vision feature extraction.

Text Feature Exraction. We utilize the embedding layer of BART to get the embedding of the source text. In particular, each token can be embedded with a linear transformation as follows,

$$\mathbf{e}_i = \mathbf{W}^T \mathbf{x}_i, i = 1, 2, \cdots, M, \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{|\mathcal{V}| \times d_1}$ is the token embedding matrix which can be optimized. $|\mathcal{V}|$ refers to the size of whole token vocabulary. d_1 is the dimension of the token embedding matrix. $\mathbf{x}_i \in \mathbb{R}^{|\mathcal{V}|}$ is the one-hot vector that indicates the index of the x_i in the token vocabulary. $\mathbf{e}_i$ is the embedding of the token x_i in the source sentence X.

To make the model aware of the positional order information of inputs, we introduce the positional embedding [13] to get the final embedding of the source text X as follows,

$$\mathbf{E} = [\mathbf{e}_1;\ \mathbf{e}_2;\ \cdots;\ \mathbf{e}_M]^T + \mathbf{E}_p, \tag{3}$$

where $\mathbf{E}_p \in \mathbb{R}^{M \times d_1}$ is the positional embedding and $\mathbf{E} \in \mathbb{R}^{M \times d_1}$ is the final embedding which encode the postional information of the source sentence X. [;] denotes the concatenation operation.

Then, we employ BART encoder to extract the textual feature. In particular, we feed the text embedding $\mathbf{E}$ into the encoder $\mathcal{E}$ of the pre-trained BART as follows,

$$\mathbf{Z} = \mathcal{E}(\mathbf{E}), \tag{4}$$

where $\mathbf{Z} \in \mathbb{R}^{M \times d_2}$ is the extracted textual feature and the d_2 is the dimension of the textual feature.

Vision Feature Extraction. Since the transformer models have achieved excellent performance in many computer vision tasks [27], we choose Swin Transformer [28] the visual encoder. In particular, we firstly split an input RGB image into K non-overlapping patches by a patch splitting module. Then, we employ Swin Transformer to extract the visual features by feeding the split patches as follows,

$$\begin{cases} \mathbf{I}' = Swin(v_1, v_2, \cdots, v_K) \\ \mathbf{I} = \mathbf{I}'\mathbf{W}_I + \mathbf{b}_I, \end{cases}, \tag{5}$$

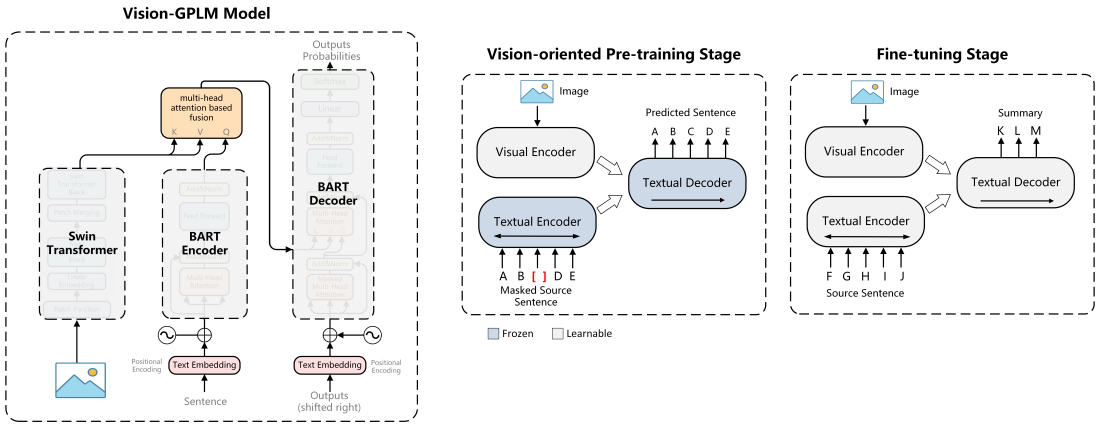

Fig. 2: Illustration of our proposed model and two training stages. In the vision-oriented pre-training stage, the parameters of the textual encoder and textual decoder are frozen while the visual encoder is trained to predict mask tokens. In the fine-tuning stage, all components are learnable and are trained to summarize sentences.

where $v_i \in \mathbb{R}^{H_{in} \times W_{in} \times 3}$ is the i-th splited patch. H_{in} and W_{in} are the height and width of the RGB patch image. 3 refers to the number of RGB channels. $\mathbf{I}' \in \mathbb{R}^{D_0}$ is the output feature vector of Swin. D_0 is the dimension of the output of the Swin Transformer. $\mathbf{W}_I \in \mathbb{R}^{D_0 \times D_1}$ is a linear transformation metrix, and $\mathbf{b}_I \in \mathbb{R}^{D_1}$ is the bias vector. $\mathbf{I} \in \mathbb{R}^{D_1}$ is the extracted visual features and D_1 is the dimension of the visual feature.

3.2.2 Multi-head Attention Based Fusion

In order to inject the visual informatin into the GPLM (*i.e.,* BART), we resort to multi-head attention based fusion strategy [13], which has achieved compelling success in many multimodal tasks, such as multimodal sentiment analysis [7], visual question answering [29], and multimodal abstractive summarization [8]. Suppose we have H attention heads, and the attention function of the i-th attention head can be formulated as follows,

$$\begin{cases} \mathbf{Q}_i = \mathbf{Z}\mathbf{W}_i^q \\ \mathbf{K}_i = \mathbf{I}\mathbf{W}_i^k \\ \mathbf{V}_i = \mathbf{I}\mathbf{W}_i^v \\ \mathbf{O}_i = softmax(\frac{\mathbf{Q}_i\mathbf{K}_i^T}{\sqrt{d_1}})\mathbf{V_i} \end{cases}, \tag{6}$$

where $\mathbf{W}_i^q \in \mathbb{R}^{d_2 \times \frac{d_2}{H}}$, $\mathbf{W}_i^k \in \mathbb{R}^{D_1 \times \frac{d_2}{H}}$, and $\mathbf{W}_\mathbf{i}^v \in \mathbb{R}^{D_1 \times \frac{d_2}{H}}$ are the learnable matrices in the i-th attention head, which aim to project the text feature and image feature into the same semantic space, and $\mathbf{V}_i$). $softmax(\cdot)$ is the softmax activation function. $\mathbf{O}_i \in \mathbb{R}^{M \times \frac{d_2}{H}}$ is the representation of the multimodal input (*i.e.,* the source sentence and the image) derived by the i-th head.

Next, we aggregate all heads from different subspaces to obtain the final multimodal representation as follows,

$$\mathbf{O} = [\mathbf{O}_1;\ \mathbf{O}_2;\ \cdots;\ \mathbf{O}_H]W_O \tag{7}$$

where $W_O \in \mathbb{R}^{d_2 \times d_2}$ is a trainable matrix. $\mathbf{O} \in \mathbb{R}^{M \times d_2}$ is the multimodal representation.

Finally, due to the superiority of residual connection [30] in many computer vision tasks [28, 31] and natural language processing tasks [5, 9], we apply an element-wise addition between textual features $\mathbf{Z}$ and mutlimodal representation $\mathbf{O}$ as follows,

$$\mathbf{Z}' = \mathbf{Z} + \mathbf{O}. \tag{8}$$

where $\mathbf{Z}' \in \mathbb{R}^{M \times d_2}$ is the final multimodal representation.

3.2.3 Text Generation

To generate the target text, we feed the multimodal representation $\mathbf{Z}'$ to the decoder $\mathcal{D}$ as

Algorithm 1 The Training Procedure of Vision-GPLM.

Input: training set $\mathcal{D}$.
Output: Parameters $\mathbf{\Theta}$.

1: Initialization parameters $\mathbf{\Theta}$.
2: **repeat**
3: Randomly sample a batch of (X, V, Y) from $\mathcal{D}$.
4: Update $\mathbf{\Theta}_V$ by optimizing the loss function in Eqn. (10)
5: **until** Swin Transformer converges.
6: **repeat**
7: Randomly sample a batch of (X, V, Y) from $\mathcal{D}$.
8: Update $\mathbf{\Theta}$ by optimizing the loss function in Eqn. (11)
9: **until** $\mathcal{M}$ converges.

follows,

$$\hat{\mathbf{p}}_\mathbf{j} = \mathcal{D}\left(\mathbf{Z}', \hat{y}_1, \hat{y}_2, \cdots, \hat{y}_{j-1}\right), \tag{9}$$

where $\hat{\mathbf{p}}_\mathbf{j} \in \mathbb{R}^{|\mathcal{V}|}$ is the predicted token distribution for the j-th token of the generated sentence. $\hat{y}_j$ is the derived token according to the largest element of $\hat{\mathbf{p}}_\mathbf{j}$.

3.3 Training Paradigm

Considering that the heterogeneity between the input sentence and image may hurt the text generation capability of BART, which is pre-trained simply on large-scale text corpus, we design our training paradigm with two stages: vision-oriented pre-training stage and fine-tuning stage. The former works on forcing the visual encoder (*i.e.,* Swin Transformer) to output homogeneous textual representations, narrowing the gap between textual and visual representations, while the latter targets fine-tuning the whole model in an end-to-end manner. The overall procedure of the optimization is briefly summarized in Algorithm 1.

3.3.1 Vision-oriented Pre-training

In the vision-oriented pre-training stage, we particularly train the visual encoder (*i.e.,* Swin Transformer) while keeping the textual encoder and decoder (*i.e.,* BART) fixed. In this way, the visual encoder can gain co-adapted features [32] with GPLMs, and adapt better to GPLMs.

Inspired by the masked language model objective presented in previous works [5, 9, 33], we mask certain input tokens at random and then train the model to predict those masked tokens. In particular, we randomly mask 5% tokens for every sentence, which is similar to BERT [9]. For tokens chosen to be masked, we replace tokens in the strategy that (1) 80% of the time with $[MASK]$ tokens, (2) 10% of the time with a random token, (3) 10% of the time with the unchanged input tokens. Considering that the object and event information delivered by the given image plays an important role in the summarization, we increase the masking probability of nouns by 10%, since objects and events are more likely to be described as nouns.

To force the visual encoder can learn the homogeneous feature of textual modality, we choose to mask the source sentence by the aforementioned mask strategy and then reconstruct the original source sentence as follows,

$$\mathcal{L}_{S1} = \min_{\mathbf{\Theta}_V} \frac{1}{M} \sum_{j=1}^{M} log(\hat{\mathbf{p}}_j^{Mask}[t*]), \tag{10}$$

where $\hat{\mathbf{p}}_j^{Mask}[t*]$ denotes the element of $\hat{\mathbf{p}}_j^{Mask}$ that corresponds to the j-th token of the source sentence X, and the j-th token is masked in the input sentence. M is the total number of masked tokens in the source sentence X. $\mathbf{\Theta}_V$ is the parameters of Swin Transformer. Notably, this loss is defined for a single sample.

3.3.2 Fine-tuning

To adapt the visual encoder trained in the vision-oriented pre-training stage, we train the entire model in an end-to-end manner. Towards the optimization of our model, we adopt the standard cross-entropy loss to fulfill the output supervision as follows,

$$\mathcal{L}_{S2} = \min_{\mathbf{\Theta}} \frac{1}{L} \sum_{j=1}^{L} log(\hat{\mathbf{p}}_j[t*]), \tag{11}$$

where $\hat{\mathbf{p}}_j[t*]$ denotes the element of $\hat{\mathbf{p}}_j$ that corresponds to the j-th token of the ground truth summary Y. L is the total number of tokens in

Table 1: The statistics of the MMSS dataset. #Train, #Valid and #Test denote the number of samples in the training set, validation set and testing set, respectively. #AvgSourceLength and #AvgSummaryLength are the average number of tokens for source sentences and summaries, respectively.

#Train	62,000
#Valid	2,000
#Test	2,000
#AvgSourceLength	22
#AvgSummaryLength	8

the ground truth summary Y. Notably, this loss is also defined for a single sample.

4 Experiment

To verify the effectivity of the proposed model Vision-GPLM, we conducted extensive experiments on a multimodal sentence summarization dataset to answer these research questions:

RQ1. Does Vision-GPLM outperform state-of-the-art methods?

RQ2. How does each component of Vision-GPLM affect its performance?

RQ3. What is the qualitative performance of Vision-GPLM?

4.1 Experimental Setting

Dataset. To verify the effectiveness of our model, we conducted extensive experiments on a widely-used multimodal sentence summarization dataset [3]. Each sample in this MMSS dataset is a triplet (*i.e.,* sentence, image, summary). The MMSS dataset contains 66,000 triplets. As shown in Table 1, the training set, validation set and test set consist of 62,000, 2,000 and 2,000 triplets, respectively. The average number of tokens in source sentences is 22, whereas the average number of tokens in summaries is 8.

Implementation Details. We trained our model on a Tesla T4 GPU, and the batch size is set to 16. We used the BART provided by Hugging Face[2] as our text encoder and decoder backbone. The height and width of input image's splited patches W_{in}, H_{in} are both 4. The dimension of the token embedding d_1 and that of the encoded representation d_2 are both 768. The dimension of the output representation of Swin Transformer D_0 is 1024. The number of attention heads is set to 8. The size of vocabulary $\mathcal{V}$ is 50265. We utilized three widely-used summarization metrics ROUGE-1, ROUGE-2 and ROUGE-L [34] for comparison. Note that all the experiments were conducted five times, and the average performance is reported.

[2] https://huggingface.co/docs/transformers/index.

Table 2: Performance (%) comparison among different methods. The best results are in bold case and the second best are underlined. R-1, R-2, R-L represent ROUGE-1, ROUGE-2, ROUGE-L, respectively. "Improvement . ↑" denotes the relative improvement of Vision-GPLM over the best baseline.

Model	**R-1**	**R-2**	**R-L**
Lead	33.6	13.4	31.8
Compress	31.6	11.0	28.9
ABS	36.0	18.2	31.9
Multi-Source	39.7	19.1	38.0
Doubly-Attentive	41.1	21.8	39.9
SEASS	44.9	23.0	42.0
PGNet	46.1	24.2	44.2
MAtt	47.3	24.9	44.5
TGSMR	48.2	25.6	45.3
BART	<u>51.4</u>	<u>29.1</u>	<u>48.6</u>
Vision-GPLM	**53.2**	**30.7**	**50.5**
Improvement. ↑	3.5%	5.5%	3.9%

4.2 On Model Comparision (RQ1)

To justify our model Vision-GPLM, we introduced several baselines for comparison.

- **Lead** [4]. It is a simple baseline which takes the first eight words of the source sentence as the summary.
- **Compress** [35]. This method summarizes sentence based on the syntactic structure of the source sentence.
- **ABS** [1]. This method summarizes the source sentence with a convolutional neural network (CNN) encoder and a neural network language model decoder.
- **SEASS** [36]. This is a textual summarization model which incorporates the textual selective encoding.
- **Multi-Source** [37]. This is a multimodal hierarchical attention model for text summarization.

Table 3: Ablation study results (%). The best results are in bold case. R-1, R-2, R-L represent ROUGE-1, ROUGE-2, ROUGE-L respectively.

Model	R-1	R-2	R-L
w/o-Image	51.4	29.1	48.6
w-Concate	52.3	29.6	49.5
w/o-Pre-training	52.4	30.0	49.7
w-VGG	50.2	27.8	47.6
w-Res	51.3	28.6	48.7
Vision-GPLM	**53.2**	**30.7**	**50.5**

- **Doubly-Attentive** [38]. This is a multimodal machine translation model equipped with a doubly-attentive mechanism.
- **PGNet** [39]. This is a textual sequence-to-sequence neural network model containing the copying mechanism.
- **MAtt** [3]. This is a hierarchical seq2seq model with a modality-based attention mechanism.
- **BART**. This is a denoising autoencoder model with transformer architecture which is pre-trained by reconstructing the original text of corrupted text with five noising functions.
- **TGSMR** [4]. This is a multimodal selective gate network for multimodal sentence summarization.

We report the performance comparison between our model and all the baselines in Table 2. From this table, we can acquire the following observations. (1) Vision-GPLM achieves the state-of-the-art performance compared to all baselines over all metrics. This demonstrates the superiority of Vision-GPLM. (2) It is worth noting that BART is already far ahead of other baselines by only utilizing textual information. The reason may be that BART has been well pre-trained on a vast corpus and learned transferable knowledge, which is overlooked by previous work. (3) Vision-GPLM surpasses the BART over all metrics. This verifies that Vision-GPLM can further improve the generation ability of GPLMs by injecting visual information.

4.3 On Ablation Study (RQ2)

To verify the importance of each module of Vision-GPLM, we introduce the following variant methods for ablation study.

- **w/o-Image**. To show the benefit of the image in MMSS, we design this method that only utilizes the source text to generate the summary. Actually, it is BART.
- **w-Concate**. To demonstrate the effect of the multi-head attention based fusion strategy, we directly utilize concatenation operation for multimodal fusion rather than the original multi-head attention based fusion in our model.
- **w/o-Pre-training**. To show the necessity of thevision-oriented pre-training, we remove the vision-oriented pre-training stage and directly apply fine-tuning.
- **w-VGG** and **w-Res**. In order to show the influence of different image encoders in our model, we replace Swin Transformer in our model with the VGG [40] and ResNet [30], respectively.

Table 3 shows the ablation study results of our proposed model. From this table, we have the following observations. (1) Vision-GPLM consistently surpasses w/o-Image over all metrics. This illustrates the importance of using visual information for sentence summarization. (2) Vision-GPLM exceeds w-Concate. This shows the superiority of the multi-head attention based fusion strategy. (3) The performance of Vision-GPLM drops when the vision-oriented pre-training stage is removed. The reason may be that directly injecting visual information into GPLM confuses the GPLM and hurts its generation ability. (4) Our model exceeds w-VGG and w-Res over all metrics. This suggests the powerful visual feature extraction capacity of Swin Transformer and its knowledge learned from the vision-oriented pre-training stage is valuable.

4.4 On Case Study (RQ3)

As shown in Fig. 3, to get an intuitive understanding of the multimodal sentence summarization ability of our model, we show a testing result of Vision-GPLM and its variant w/o-Image. As can be seen, the performance (*i.e.,* ROUGE-1, ROUGE-2 and ROUGE-L) of Vision-GPLM exceeds its variant w/o-Image. Looking into the generated summaries, we can learn that by incorporating the product's image, Vision-GPLM can capture the vital information (*i.e.,* railway) which appears in both the image and text, while w/o-Image cannot. Therefore, "railway town" is not

Fig. 3: Comparison between the summaries generated by Vison-GPLM and w/o-image for a testing sentence-image pair. The reference summary is the ground truth in this case. ROUGE-1, ROUGE-2, and ROUGE-L scores of each sentence are given.

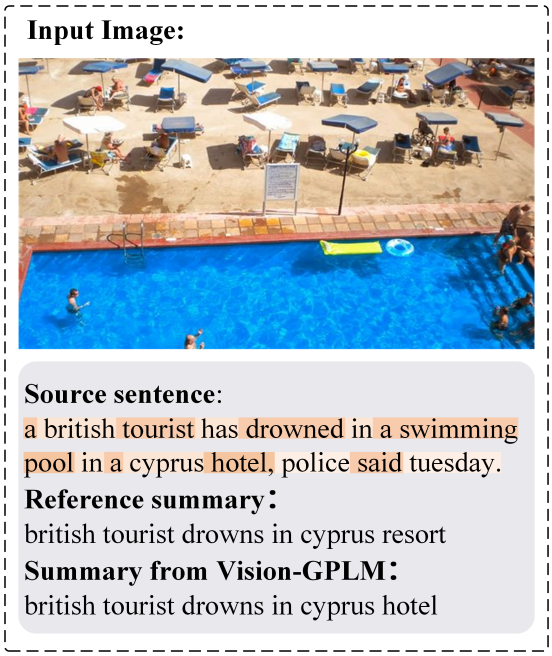

Fig. 4: Illustration of the multi-head attention based fusion mechanism. The color depth of the orange bar stands for the confidence of the word learned by the attention mechanism. The darker color refers to the larger attention weight.

mentioned in the summary produced by w/o-Image, which instead incorrectly focuses on how much it spends. This intuitively verifies the necessity of injecting the visual modality into the GPLMs for multimodal sentence summarization.

In addition, we studied the multi-head attention based fusion mechanism and we showed a testing sample on the confidence assignment with tokens in the source sentence in Fig. 4. From this figure, the multi-head attention based fusion mechanism does assign different level confidence into different tokens in the source sentence. This verifies that the multi-head attention based fusion does contribute to the multimodal sentence summarization task. Nobtaly, the multi-head attention based fusion mechanism assign the high confidence into the semantically identical parts of the image and source sentence (*e.g.,* "tourist", "swimming pool", and "hotel"), which is the significant semantic information in multimodality and hence boost the MMSS task.

5 Conclusions and Future Work

In this work, we present a vision enhanced generative pre-trained language model, which seamlessly unifies the heterogeneous multimodal data (*i.e.,* the source sentence and image) of the product into the common semantic space of the GPLM (*i.e.,* BART). Extensive experiments on a public multimodal sentence summarization dataset demonstrate the superiority of our model over existing cutting-edge methods. The ablation study verifies that each component of our model is effective and the visual modality can enhance the quality of generated summaries. Moreover, we also show the benefit of using Swin Transformer instead of VGG or ResNet for the visual feature extraction. In the future, we plan to adopt more advanced generative pre-trained language models (*e.g.,* T5) to solve the multimodal sentence summarization task.

References

[1] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *EMNLP*, pages 379–389, 2015.

[2] Sumit Chopra, Michael Auli, and Alexander M. Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *HLT-NAACL*, pages 93–98, 2016.

[3] Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. Multimodal sentence summarization with modality attention and image filtering. In *IJCAI*, pages 4152–4158, 2018.

[4] Haoran Li, Junnan Zhu, Jiajun Zhang, Xiaodong He, and Chengqing Zong. Multimodal sentence summarization via multimodal selective encoding. In *COLING*, pages 5655–5667, 2020.

[5] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880, 2020.

[6] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020.

[7] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, pages 6558–6569, 2019.

[8] Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *EMNLP*, pages 3995–4007, 2021.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019.

[10] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. Incorporating copying mechanism in sequence-to-sequence learning. In *ACL*, 2016.

[11] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[12] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[14] Xue Song, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Spatial-temporal graphs for cross-modal text2video retrieval. *IEEE Trans. Multim.*, 24:2914–2923, 2022.

[15] Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of ACL/IJCNLP*, pages 4693–4703, 2021.

[16] Animesh Nighojkar and John Licato. Improving paraphrase detection with the adversarial paraphrasing task. In *ACL/IJCNLP*, pages 7106–7116, 2021.

[17] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations, Workshop Track Proceedings*, 2013.

[18] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. Baby talk: Understanding and generating simple image descriptions. In *Conference on Computer Vision and Pattern Recognition*, pages 1601–1608. IEEE, 2011.

[19] Ali Farhadi, Seyyed Mohammad Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David A. Forsyth. Every picture tells a story: Generating sentences from

images. In *European Conference on Computer Vision*, volume 6314 of *Lecture Notes in Computer Science*, pages 15–29. Springer, 2010.

[20] Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa C. Mensch, Alexander C. Berg, Tamara L. Berg, and Hal Daumé III. Midge: Generating image descriptions from computer vision detections. In *Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. ACL, 2012.

[21] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Conference on Computer Vision and Pattern Recognition*, pages 4651–4659. IEEE, 2016.

[22] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *International Conference on Computer Vision*, pages 4904–4912. IEEE, 2017.

[23] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *European Conference on Computer Vision*, volume 11218 of *Lecture Notes in Computer Science*, pages 711–727. Springer, 2018.

[24] Lei Ke, Wenjie Pei, Ruiyu Li, Xiaoyong Shen, and Yu-Wing Tai. Reflective decoding network for image captioning. In *International Conference on Computer Vision*, pages 8887–8896. IEEE, 2019.

[25] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10575–10584. Computer Vision Foundation / IEEE, 2020.

[26] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Conference on Computer Vision and Pattern Recognition*, pages 10968–10977. IEEE, 2020.

[27] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, pages 17864–17875, 2021.

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 9992–10002, 2021.

[29] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, volume 139, pages 1931–1942, 2021.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[31] 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. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, pages 1–21, 2021.

[32] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NIPS*, pages 3320–3328, 2014.

[33] Wilson L Taylor. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953.

[34] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *ACL*, 2004.

[35] James Clarke and Mirella Lapata. Global inference for sentence compression: An integer linear programming approach. *J. Artif. Intell. Res.*, 31:399–429, 2008.

[36] Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. Selective encoding for abstractive sentence summarization. In *ACL*, pages 1095–1104, 2017.

[37] Jindrich Libovický and Jindrich Helcl. Attention strategies for multi-source sequence-to-sequence learning. In *ACL*, pages 196–202, 2017.

[38] Iacer Calixto, Qun Liu, and Nick Campbell. Doubly-attentive decoder for multi-modal neural machine translation. In *ACL*, pages 1913–1924, 2017.

[39] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *ACL*, pages 1073–1083, 2017.

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, pages 1–14, 2015.

Liqiang Jing is a graduate student with the Department of Computer Science and Technology, Shandong University. He received the B.E. degree in School of Computer Science and Technology from Hefei University of Technology, Anhui, in 2020.

His research interests include multimodal learning and natural language processing.

E-mail: jingliqiang6@gmail.com
ORCID iD: 0000-0001-9827-5835

Yiren Li, born in 1974, is currently the deputy general manager of HBIS Group and the chairman of HBIS Digital Technology Co.,Ltd. Previously, he successively served as the deputy director of Integrated Management Department of HBIS Group, director of Management Innovation Department of HBIS Group, and strategy director of HBIS Group. He has published more than 20 papers.

His research interests include intelligent applications in the iron and steel industry.

E-mail: liyiren@hbisco.com

Junhao Xu is an undergraduate student with the Department of Computer Science and Technology, Shandong University.

His research interests include information retrieval and natural language processing.

E-mail: xujunhao.cn@gmail.com

Yongcan Yu is an undergraduate student with the Department of Computer Science and Technology, Shandong University.

His research interests include computer vision and recommendation system.

E-mail: yuyongcan0223@gmail.com

Shen Pei, born in 1982, is currently the general manager of HBIS Digital Technology Co., Ltd. He is a member of Steel of Standardization Administration of China, vice chairman of the smart Enterprise Promotion Committee of the China Enterprise Federation, and director of the intelligent manufacturing alliance of the iron and steel industry.

His research interests include intelligent applications in the iron and steel industry.

E-mail: shenpei@hbisco.com

Xuemeng Song received the B.E. degree from University of Science and Technology of China in 2012, and the Ph.D. degree from the School of Computing, National University of Singapore in 2016. She is currently an associate professor of Shandong University, Jinan, China. She has published several papers in the top venues, such as ACM SIGIR, MM and TOIS. In addition, she has served as reviewers for many top conferences and journals.

Her research interests include the information retrieval and social network analysis.

E-mail: sxmustc@gmail.com (Corresponding author)