Machine Intelligence Research
www.mi-research.net

20(2), April 2023, 289-298
DOI: 10.1007/s11633-022-1372-x

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] 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 that 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 using 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? 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 (MMSS), generative pre-trained language model (GPLM), natural language generation, deep learning, artificial intelligence.

Citation: L. Jing, Y. Li, J. Xu, Y. Yu, P. Shen, X. Song. Vision enhanced generative pre-trained language model for multimodal sentence summarization. *Machine Intelligence Research*, vol.20, no.2, pp.289–298, 2023. http://doi.org/10.1007/s11633-022-1372-x

1 Introduction

Sentence summarization is a task that aims to generate a short summarization of 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 employ the encoder-decoder framework for semantic understanding and text generation. For example, Li et al.[3] utilized 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

Research Article
Special Issue on Large-scale Pre-training: Data, Models, and Fine-tuning
Manuscript received June 25, 2022; accepted August 30, 2022; published online January 11, 2023
Recommended by Associate Editor Hao Dong

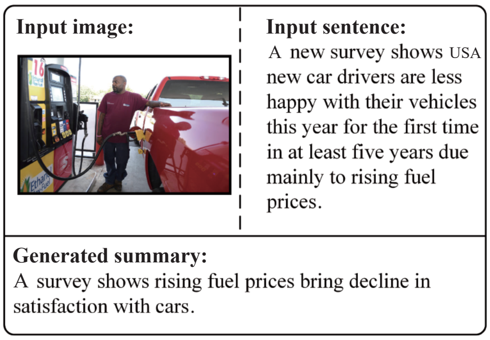

Fig. 1 Illustration of the task of multimodal sentence summarization

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 has 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:

1) 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.

2) 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.

3) 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 by-product, 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.

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 in a long text by rephrasing or using 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 out inconsequential information from the source sentence.

Although these methods have achieved remarkable success, they overlook the benefit of pre-training and training 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-

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

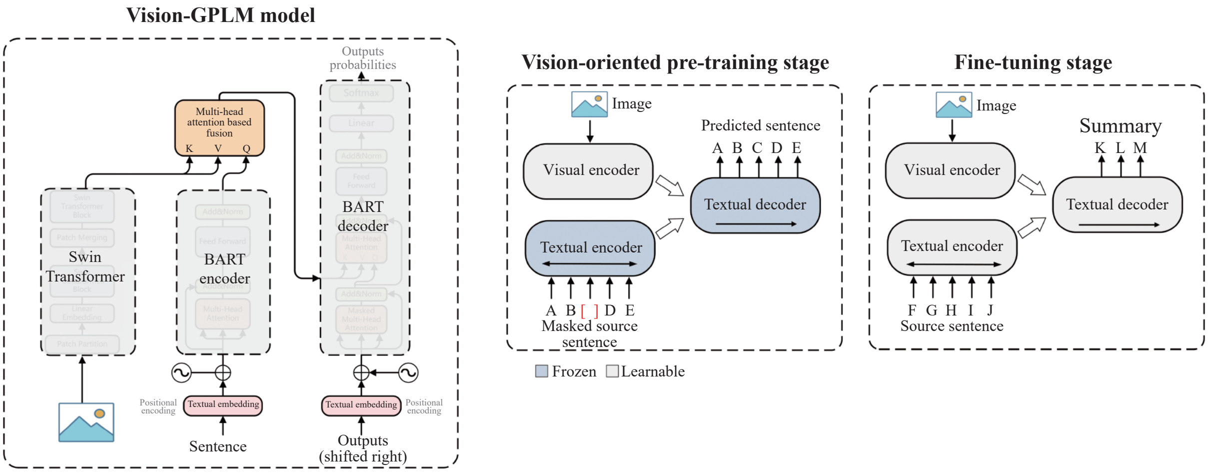

Fig. 2 Illustration of our proposed model and two training stages. In the vision-oriented pre-training stage, the parameters of the textual encoder and textual decoder are frozen while the visual encoder is trained to predict mask tokens. In the fine-tuning stage, all components are learnable and trained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

trained models that introduced a 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) with 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 utilized the full Transformer architecture for natural language generation. Meanwhile, Raffel et al.[6] proposed T5, which 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, Song et al.[17] adapted the generative pre-trained language model BART for a multimodal product summarization task which summarizes the image of the product and its textual description into a short text. 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 a natural language description for an image. Early studies[18, 19] 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 structures. Meanwhile, the search-oriented methods[20, 21] directly adopted the sentence of the similar image or selected a semantic similar sentence from a sentence set to get the target sentence. Obviously, these methods are limited by the size of the human-generated sentence set and cannot generate a new sentence. Recently, with the development of deep learning, many works[22–27] utilized neural networks to learn the probability distribution in the common semantic space of visual content and textual content, and generate a new sentence, achieving state-of-the-art performance.

Despite the success of the image captioning methods mentioned above, they are not suitable for the multimodal sentence summarization task because they cannot tackle the textual input.

3 Methodology

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

3.1 Task formulation

Suppose that 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:

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

where $\boldsymbol{\Theta}$ 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 power 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 extraction. 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:

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

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

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

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

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

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

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

where $\boldsymbol{Z} \in \mathbf{R}^{M \times d_2}$ is the extracted textual feature, and 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[28], we chose the Swin Transformer[29] as 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 the Swin Transformer to extract the visual features by feeding the split patches as follows:

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

where $v_i \in \mathbf{R}^{H_{in} \times W_{in} \times 3}$ is the i-th split patch. H_{in} and W_{in} are the height and width of the RGB patch image. 3 refers to the number of RGB channels. $\boldsymbol{I'} \in \mathbf{R}^{D_0}$ is the output feature vector of Swin. D_0 is the dimension of the output of the Swin Transformer. $\boldsymbol{W}_I \in \mathbf{R}^{D_0 \times D_1}$ is a linear transformation matrix, and $\boldsymbol{b}_I \in \mathbf{R}^{D_1}$ is the bias vector. $\boldsymbol{I} \in \mathbf{R}^{D_1}$ 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 information into the GPLM (i.e., BART), we resort to a 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[30], and multimodal abstractive summarization[8]. Suppose we have H attention heads, and the attention function of the H-th attention head can be formulated as follows:

$$\begin{cases} \boldsymbol{Q}_i = \boldsymbol{Z} \boldsymbol{W}_i^q \\ \boldsymbol{K}_i = \boldsymbol{I} \boldsymbol{W}_i^k \\ \boldsymbol{V}_i = \boldsymbol{I} \boldsymbol{W}_i^v \\ \boldsymbol{O}_i = softmax\left(\dfrac{\boldsymbol{Q}_i \boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{d_1}}\right) V_i \end{cases} \tag{6}$$

where $\boldsymbol{W}_i^q \in \mathbf{R}^{d_2 \times \frac{d_2}{H}}$, $\boldsymbol{W}_i^k \in \mathbf{R}^{D_1 \times \frac{d_2}{H}}$, and $\boldsymbol{W}_i^v \in \mathbf{R}^{D_1 \times \frac{d_2}{H}}$ are the learnable matrices in the i-th attention head, which aim to project the text feature and the image feature into the same semantic space, and $\boldsymbol{V}_i$. $softmax(\cdot)$ is the softmax activation function. $\boldsymbol{O}_i \in \mathbf{R}^{M \times \frac{d_2}{H}}$ 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:

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

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

Finally, due to the superiority of residual connection[31] in many computer vision tasks[29, 32] and natural language processing tasks[5, 9], we apply an element-wise addition between textual features $\boldsymbol{Z}$ and multimodal representation $\boldsymbol{O}$ as follows:

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

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

3.2.3 Text generation

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

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

where $\hat{\boldsymbol{p}}_j \in \mathbf{R}^{|\mathcal{V}|}$ 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 $\hat{\boldsymbol{p}}_j$.

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: the 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.

Algorithm 1. Training procedure of Vision-GPLM
Input: Training set $\mathcal{D}$.
Output: Parameters $\boldsymbol{\Theta}$.
1) Initialization parameters $\boldsymbol{\Theta}$.
2) **repeat**
3) Randomly sample a batch of (X, V, Y) from $\mathcal{D}$.
4) Update $\boldsymbol{\Theta}_V$ by optimizing the loss function in (10)
5) **until** Swin Transformer converges.
6) **repeat**
7) Randomly sample a batch of (X, V, Y) from $\mathcal{D}$.
8) Update $\boldsymbol{\Theta}$ by optimizing the loss function in (11)
9) **until** $\mathcal{M}$ converges.

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[33] with GPLMs, and adapt better to GPLMs.

Inspired by the masked language model objective presented in previous works[5, 9, 34], we mask certain input tokens randomly 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, and 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:

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

where $\hat{\boldsymbol{p}}_j^{Mask}[t*]$ denotes the element of $\hat{\boldsymbol{p}}_j^{Mask}$ 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. $\boldsymbol{\Theta}_V$ are the parameters of the 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. Toward the optimization of our model, we adopt the standard cross-entropy loss to fulfill the output supervision as follows:

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

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

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

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

tences 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 split patches, W_{in} and H_{in}, are both 4. The dimensions of the token embedding d_1 and that of the encoded representation d_2 are both 768. The dimension of the output representation of the Swin Transformer D_0 is 1 024. The number of attention heads is set to 8. The size of vocabulary $\mathcal{V}$ is 50 265. We utilized three widely-used summarization metrics, ROUGE-1, ROUGE-2, and ROUGE-L[35], for comparison. Note that all the experiments were conducted five times, and the average performance is reported.

4.2 On model comparison (RQ1)

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

Lead.[4] It is a simple baseline that takes the first eight words of the source sentence as the summary.

Compress.[36] This method summarizes a 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.[37] This is a textual summarization model which incorporates textual selective encoding.

Multi-source.[38] This is a multimodal hierarchical attention model for text summarization.

Doubly-attentive.[39] This is a multimodal machine translation model equipped with a doubly-attentive mechanism.

PGNet.[40] 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.

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

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 Table 2, we can acquire the following observations. 1) Vision-GPLM achieves state-of-the-art performance compared to all baselines on 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 BART on all metrics. This verifies that Vision-GPLM can further improve the generation ability of GPLMs by injecting visual information.

Table 2 Performance (%) comparison among different methods. The best results are in bold, and the second best results 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%

4.3 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 the vision-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 visual geometry group (VGG)[41] and ResNet[31], respectively.

Table 3 shows the ablation study results of our proposed model. From Table 3, we have the following observations. 1) Vision-GPLM consistently surpasses w/o-Image on 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 on all metrics. This suggests the powerful visual feature extraction capacity of the Swin Transformer and its knowledge learned from the vision-oriented pre-training stage is valuable.

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

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 test 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 can not. Therefore, "railway town" is not 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 Fig. 4, the multi-head attention based fusion mechanism does assign different levels of confidence to different tokens in the source sentence. This verifies that the multi-head attention based fusion does contribute to the multimodal sentence summarization task. Notably, the multi-head attention based fusion mechanism assigns high confidence to 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 boosts 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 that the visual modality can enhance the quality of generated summaries. Moreover, we also show the benefit of using the Swin Transformer instead of VGG or ResNet for the visual feature extraction. In the future, we plan to adopt

Source sentence: Jiangxi province in east china will spend one billion yuan on the development of small towns along the beijing-kowloon railway, which runs ### km across the province.

Reference summary: Jiangxi to develop railway towns

Summary from w/o-Image: Jiangxi to spend one billion yuan on development of small towns
ROUGE-1: 37.49%, ROUGE-2: 14.28%, ROUGE-L: 37.49%

Summary from Vision-GPLM: Jiangxi to develop small towns along railway
ROUGE-1: 83.33%, ROUGE-2: 39.99%, ROUGE-L: 66.66%

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. The ROUGE-1, ROUGE-2, and ROUGE-L scores for each sentence are given.

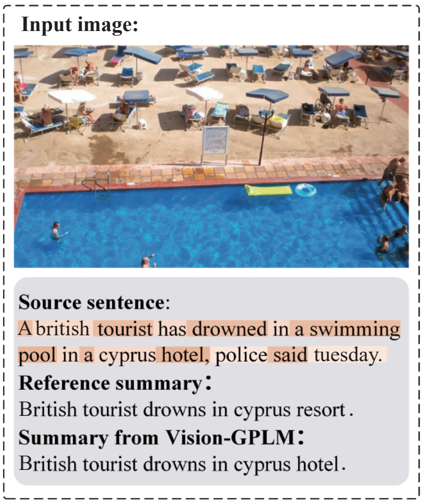

Fig. 4 Illustration of the multi-head attention based fusion mechanism. The color depth of the orange bar stands for the confidence of the word learned by the attention mechanism. The darker color refers to the larger attention weight.

more advanced generative pre-trained language models (e.g., T5) to solve the multimodal sentence summarization task.

References

[1] A. M. Rush, S. Chopra, J. Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp. 379–389, 2015. DOI: 10.18653/v1/D15-1044.

[2] S. Chopra, M. Auli, A. M. Rush. Abstractive sentence sum-marization with attentive recurrent neural networks. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, USA, pp. 93–98, 2016. DOI: 10.18653/v1/N16-1012.

[3] H. R. Li, J. N. Zhu, T. S. Liu, J. J. Zhang, C. Q. Zong. Multi-modal sentence summarization with modality attention and image filtering. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, pp. 4152–4158, 2018.

[4] H. R. Li, J. N. Zhu, J. J. Zhang, X. D. He, C. Q. Zong. Multimodal sentence summarization via multimodal selective encoding. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain, pp. 5655–5667, 2020. DOI: 10.18653/v1/2020.coling-main.496.

[5] M. Lewis, Y. H. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020. DOI: 10.18653/v1/2020.acl-main.703.

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Q. Zhou, W. Li, P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, vol. 21, no. 1, Article number 140, 2020.

[7] Y. H. H. Tsai, S. J. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency, R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 6558–6569, 2019. DOI: 10.18653/v1/P19-1656.

[8] T. Z. Yu, W. L. Dai, Z. H. Liu, P. Fung. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, pp. 3995–4007, 2021. DOI: 10.18653/v1/2021.emnlp-main.326.

[9] J. Devlin, M. W. Chang, K. Lee, K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, USA, Minnesota, pp. 4171–4186, 2019. DOI: 10.18653/v1/N19-1423.

[10] J. T. Gu, Z. D. Lu, H. Li, V. O. K. Li. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, pp. 1631–1640, 2016. DOI: 10.18653/v1/P16-1154.

[11] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, Lake Tahoe, USA, pp. 3111–3119, 2013.

[12] J. Pennington, R. Socher, C. Manning. GloVe: Global vectors for word representation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, pp. 1532–1543, 2014. DOI: 10.3115/v1/D14-1162.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, USA, pp. 6000–6010, 2017.

[14] X. Song, J. J. Chen, Z. X. Wu, Y. G. Jiang. Spatial-temporal graphs for cross-modal text2video retrieval. *IEEE Transactions on Multimedia*, vol. 24, pp. 2914–2923, 2022. DOI: 10.1109/TMM.2021.3090595.

[15] T. Hasan, A. Bhattacharjee, M. S. Islam, K. Mubasshir, Y. F. Li, Y. B. Kang, M. S. Rahman, R. Shahriyar. Xl-Sum: Large-scale multilingual abstractive summarization for 44 languages. In *Proceedings of Findings of the Association for Computational Linguistics*, pp. 4693–4703, 2021. DOI: 10.18653/v1/2021.findings-acl.413.

[16] A. Nighojkar, J. Licato. Improving paraphrase detection with the adversarial paraphrasing task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 7106–7116, 2021. DOI: 10.18653/v1/2021.acl-long.552.

[17] X. M. Song, L. Q. Jing, D. T. Lin, Z. Z. Zhao, H. Q. Chen, L. Q. Nie. V2P: Vision-to-prompt based multi- modal product summary generation. In *Proceedings of the 45th*

International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, pp. 992–1001, 2022. DOI: 10.1145/3477495.3532076.

[18] T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations*, Scottsdale, USA, 2013. DOI: doi.org/10.48550/arXiv.1301.3781.

[19] G. Kulkarni, V. Premraj, S. Dhar, S. M. Li, Y. Choi, A. C. Berg, T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, IEEE, Colorado Springs, USA, pp. 1601–1608, 2011. DOI: 10.1109/CVPR.2011.5995466.

[20] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. Forsyth. Every picture tells a story: Generating sentences from images. In *Proceedings of the 11th European Conference on Computer Vision*, Springer, Heraklion, Greece, pp. 15–29, 2010. DOI: 10.1007/978-3-642-15561-1_2.

[21] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. F. Han, A. Mensch, A. Berg, T. Berg, H. DauméIII Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, pp. 747–756, 2012.

[22] Q. Z. You, H. L. Jin, Z. W. Wang, C. Fang, J. B. Luo. Image captioning with semantic attention. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 4651–4659, 2016. DOI: 10.1109/CVPR.2016.503.

[23] T. Yao, Y. W. Pan, Y. H. Li, Z. F. Qiu, T. Mei. Boosting image captioning with attributes. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 4904–4912, 2017. DOI: 10.1109/ICCV.2017.524.

[24] T. Yao, Y. W. Pan, Y. H. Li, T. Mei. Exploring visual relationship for image captioning. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 711–727, 2018. DOI: 10.1007/978-3-030-01264-9_42.

[25] L. Ke, W. J. Pei, R. Y. Li, X. Y. Shen, Y. W. Tai. Reflective decoding network for image captioning. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Republic of Korea, pp. 8887–8896, 2019. DOI: 10.1109/ICCV.2019.00898.

[26] M. Cornia, M. Stefanini, L. Baraldi, R. Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 10575–10584, 2020. DOI: 10.1109/CVPR42600.2020.01059.

[27] Y. W. Pan, T. Yao, Y. H. Li, T. Mei. X-linear attention networks for image captioning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 10968–10977, 2020. DOI: 10.1109/CVPR42600.2020.01098.

[28] B. W. Cheng, A. G. Schwing, A. Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Proceedings of the 35th Conference on Neural Information Processing Systems*, pp. 17864–17875, 2021.

[29] Z. Liu, Y. T. Lin, Y. Cao, H. Hu, Y. X. Wei, Z. Zhang, S. Lin, B. N. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, USA, pp. 9992–10002, 2021. DOI: 10.1109/ICCV48922.2021.00986.

[30] J. Cho, J. Lei, H. Tan, M. Bansal. Unifying vision-and-language tasks via text generation. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 1931–1942, 2021.

[31] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. H. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations*, pp. 1–21, 2021.

[33] J. Yosinski, J. Clune, Y. Bengio, H. Lipson. How transferable are features in deep neural networks? In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 3320–3328, 2014.

[34] W. L. Taylor. "Cloze procedure": A new tool for measuring readability. *Journalism & Mass Communication Quarterly*, vol. 30, no. 4, pp. 415–433, 1953. DOI: 10.1177/107769905303000401.

[35] C. Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of Text Summarization Branches Out*, Barcelona, Spain, pp. 74–81, 2004.

[36] J. Clarke, M. Lapata. Global inference for sentence compression an integer linear programming approach. *Journal of Artificial Intelligence Research*, vol. 31, no. 1, pp. 399–429, 2008.

[37] Q. Y. Zhou, N. Yang, F. R. Wei, M. Zhou. Selective encoding for abstractive sentence summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, pp. 1095–1104, 2017. DOI: 10.18653/v1/P17-1101.

[38] J. Libovický, J. Helcl. Attention strategies for multi-source sequence-to- sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, pp. 196–202, 2017. DOI: 10.18653/v1/P17-2031.

[39] I. Calixto, Q. Liu, N. Campbell. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, pp. 1913–1924, 2017. DOI: 10.18653/v1/P17-1175.

[40] A. See, P. J. Liu, C. D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, pp. 1073–1083, 2017. DOI: 10.18653/v1/P17-1099.

[41] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, USA, 2015.

Liqiang Jing received the B. Eng. degree in computer science and technology from School of Computer Science and Technology, Hefei University of Technology, China in 2020. He is now a master student in computer technology at Department of Computer Science and Technology, Shandong University, China.

His research interests include multimodal learning and natural language processing.

E-mail: jingliqiang6@gmail.com

ORCID iD: 0000-0001-9827-5835

Yiren Li received the B. Eng. degree in finance from Hebei University of Economics and Business, China in 2004, and the degree in industry and business administration from Tianjin University, China in 2007. He is currently the deputy general manager of HBIS Group and the chairman of HBIS Digital Technology Co., Ltd., China. 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, China. He has published more than 20 papers.

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

E-mail: liyiren@hbisco.com (Corresponding author)

Junhao Xu is an undergraduate student in data science and big data technology at Department of Computer Science and Technology, Shandong University, China.

His research interests include information retrieval and natural language processing.

E-mail: xujunhao.cn@gmail.com

Yongcan Yu is an undergraduate student in data science and big data technology at Department of Computer Science and Technology, Shandong University, China.

His research interests include computer vision and recommendation system.

E-mail: yuyongcan0223@gmail.com

Pei Shen received the B. Eng. degree in computer and application from Hebei University of Science and Technology, China in 2010. He is currently the general manager of HBIS Digital Technology Co., Ltd, China. 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, China.

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

E-mail: shenpei@hbisco.com

Xuemeng Song received the B. Eng. degree in in electronic information engineering from University of Science and Technology of China, China in 2012, and the Ph. D. degree in computer science from School of Computing, National University of Singapore, Singapore in 2016. She is currently an associate professor of Shandong University, 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)

ORCID iD: 0000-0002-5274-4197