ViL-Sum: Enhancing Vision and Language Representations via Multi-task Learning for Multi-modal Summarization

Anonymous ACL submission

Abstract

With the advance of multimedia on the Internet, multi-modal summarization has drawn much attention. Most current methods follow a pipeline strategy, where an off-the-shelf object detector is used to extract visual features which are then fused with language representations for decoder to generate. However, these methods suffer two issues 1) separate vision and language representations fail to capture the interrelations within the two modalities; 2) from the local view, the semantic alignments between images and paragraphs are missing. In order to address these problems, in this paper, we propose a novel Vision-Language Summarization (ViL-Sum) model with a multi-task learning framework. Specifically, we train our model with two auxiliary tasks in a multi-task manner, that are images selection and images reordering. In this way, the interrelations within image and text are well captured. Besides, to further enhance the vision-language representation, we employ a unified transformer-based encoder-decoder structure. The encoder simultaneously takes image and text as input and jointly learns the representations of both. Then the representations are used by the decoder to generate the summary. Experimental results show that ViL-Sum significantly outperforms current state-of-the-art methods. In further analysis, we find that the enhanced representations via multi-task training and joint modeling learn reasonable relations between image and text.

1 Introduction

The dramatic increase of multi-modal data (including text, image, audio, and video) on the Internet makes research on multi-modal summarization necessary. Multi-modal summarization is the task of automatically capturing salient information and generating summaries from one or more modalities inputs (Evangelopoulos et al., 2013; Li et al., 2017). Compared with traditional multi-modal summarization tasks, which only generate a text-modality summary, Zhu et al. (2018) proved that multi-modal summary with both text and images can effectively increase the satisfaction of users. Intuitively, people can grasp key information easier from multiple modalities than only from the text. In this paper, we mainly focus on the multi-modal summarization with multi-modal outputs (MSMO) task, which is the parent set of multi-modal summarization with text-only outputs. A simple example of multi-modal summarization with multi-modal outputs (MSMO) is shown in Fig. 1.

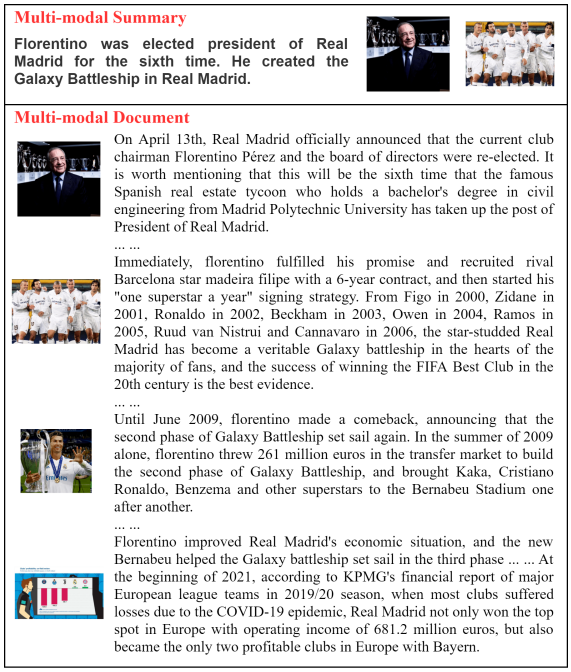

Figure 1: An example for explaining the semantic alignment between images and paragraphs in the document. "..." means some content was omitted.

Existing multi-modal summarization models (Li et al., 2017, 2018; Chen and Zhuge, 2018; Zhu et al., 2018; Khullar and Arora, 2020; Zhu et al., 2020; Im et al., 2021) simply added separate encoders for different modalities into single-modal encoder-decoder framework as shown in Fig. 2a and 2b. The representation of different modalities is obtained separately from single-modal encoders,

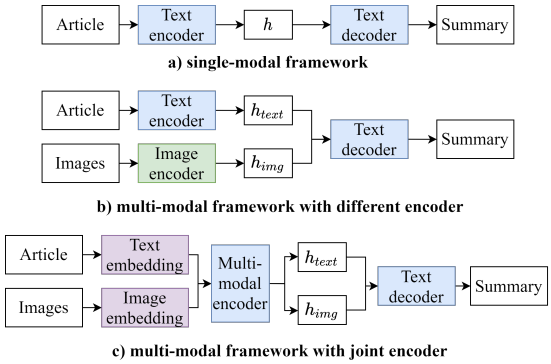

Figure 2: Multi-modal summarization frameworks with different encoder structure.

which leads to the model not effectively capturing the interaction between them. Recently, some works on vision-language representation learning (Li et al., 2020a; Xu et al., 2021; Zhou et al., 2020) have demonstrated that jointly encoding different modalities with the same encoder can improve the performance of natural language understanding tasks (e.g. classification). However, existing generation tasks (e.g. summarization) still ignored this method. In this paper, we proposed a Vision-Language Summarization (ViL-Sum) model which employs a unified transformer-based encoder-decoder structure. The encoder of ViL-Sum simultaneously encodes images and the document jointly for learning the interrelation of them. Specifically, we employ an image tokenizer to convert images into visual token embeddings and the concatenation of it and word embeddings is the input of the multi-modal encoder. The joint representations from the encoder are fed into the decoder to generate a summary. The multi-modal framework with the joint encoder of ViL-Sum is shown in Fig. 2c.

From the local view, existing works ignored another vital problem, that they can not align semantics between certain images and text paragraphs in the document. We give a semantic alignment example in Fig. 1. The semantic of paragraphs in the document is highly corresponding to the image on the left. However, the hard alignment of them is not available in existing datasets. To further enhance vision-language representation and align semantics, we proposed two simple but effective auxiliary tasks to train our ViL-Sum via multi-task learning. The first task is image selection, which selects several summary-related images as part of the multi-modal summary, which forces the model to learn the interrelation between images and text. The second task is images reordering, which aims to align semantics between images and text paragraphs via reordering shuffled images based on vision-language representations. Intuitively, the order of images is coincident with the order of paragraphs in the document. The reorder of images can force the model to learn the alignment of semantics between the two modalities. Finally, we train ViL-Sum with text summary generation, images selection, and image reordering tasks in a multi-task manner.

Our contributions can be summarized as follows: 1) We proposed a novel Vision-Language Summarization (ViL-Sum) model, which can jointly encode images and text to capture their interrelation. 2) We employ a multi-task learning framework to train our model with text summary generation and two elaborately designed simple tasks, which can effectively enhance vision-language representations and semantic alignment. 3) Our model outperforms all current state-of-the-art methods on automatic and manual evaluation metrics. 4) In further analysis, we find that the improvement is exactly from the enhanced representations via multi-task learning.

2 Methodology

2.1 Main Architecture

The overall framework of our ViL-Sum and the details of the image tokenizer are shown in Fig. 3. The main architecture is a unified Transformer encoder-decoder model, where encoder and decoder both consist of 12 standard Transformer blocks. Before input to the transformer encoder, we employ an image tokenizer to convert images into embeddings and concatenate them with document token embeddings. Then, we feed the hidden states from the transformer encoder into the transformer decoder to generate text summary, image selector layer to select several images for summary, and image reorder layer to reorder shuffled images. We will describe the details of each component in the next.

2.2 Vision-Language Joint Representation

We formalized the input and output of our ViT-Sum as (D, I) and (S, I_S), where $D = \{t_1, t_2, \ldots, t_T\}$ refers to the sequence of tokens from the input document, $I = \{img_0, img_1, \ldots, img_M\}$ refers to the sequence of input images from the input document, $S = \{t_1, t_2, \ldots\}$ refers to the se-

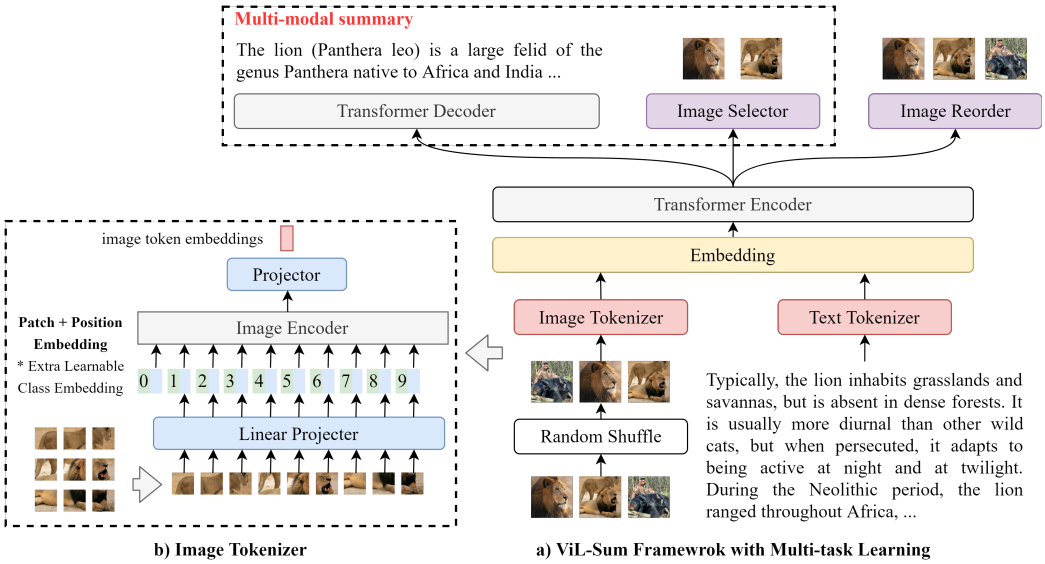

Figure 3: The overall framework of our ViL-Sum. Figure a) is the whole encoder-decoder architecture of ViL-Sum with multi-task learning. Figure b) is the detail of image tokenzier to project images into visual token embeddings.

quence of tokens from gold text summary, and $I_S = \{img_1, img_2, \ldots, img_K\}$ refers to K selected images for the summary.

2.2.1 Document Embeddings

Each document is firstly converted into the sequence of tokens $\{t_1, t_2, \ldots, t_T\}$ and then two special tokens "$\langle s \rangle$" and "$\langle \backslash s \rangle$" are added to represent the start and end of the document. After that, we map each token into vector representation $E_D = \{e_{start}, e_1, \ldots, e_T, e_{end}\}$.

2.2.2 Image Embeddings

Different from previous methods, which extract many image features via existing object detection models. We split each image into several patches, then encode them following ViT (Dosovitskiy et al., 2021). The details of image tokenizer are shown in Fig. 3b.

Firstly, we reshape image $img \in \mathbb{R}^{H \times W \times C}$ into a sequence of flattened 2D patches $\{img^p \in \mathbb{R}^{N \times (P^2 \cdot C)}\}_{p=1}^{N}$, where (H, W) is the resolution of the original image, C is the number of channels, (P, P) is the resolution of each image patch, and $N = HW/P^2$ is the resulting number of patches. Then, we can obtain a sequence of image patches $\{img^p\}_{p=1}^{N}$ as the input of image tokenzier.

Secondly, the patches are linearly projected to patch embeddings $e^p = E \times img_i^p$, where $E \in \mathbb{R}^{(P^2 \cdot D) \times C}$. We also add a special token "[class]" with learn-able embedding e^0. We add position embeddings and patch embeddings as input Z_0 of image encoder to retain positional information of images:

$$Z_0 = [e_i^0; e_i^1; \ldots; e_i^N] + E_{pos} \quad (1)$$

Where $Z_0, E_{pos} \in \mathbb{R}^{(N+1) \times D}$, E_{pos} is position embeddings.

Finally, we follow ViT to employ 12 transformer blocks to encode these patches of each image, which also can be replaced by any other encoders.

$$Z_{\ell+1} = \texttt{Transformer}(Z_\ell), \ell = 1, 2, ..., L \quad (2)$$

And the global max-pooling of output vectors is obtained as the visual token embedding $v_i \in \mathbb{R}^D$ of image img_i.

$$v_i = \texttt{Maxpooling}(Z_L) \quad (3)$$

Through the image tokenizer, we can convert the sequence of input images into a sequence of visual token embeddings $E_v = \{v_i\}_{i=1}^{M}$.

2.2.3 Multi-modal Encoder

The input of the multi-modal encoder is the concatenation of visual token embeddings E_v and token embeddings E_D. We can formalize the input as $H_0 = \{E_v; E_D\}$ and then encode visual and text embeddings with 12 transformer blocks. Finally, we can obtain vision-language representation $H_L = \{h_{v_1}, \ldots, h_{v_M}, h_{start}, h_1, \ldots, h_{end}\}$ from last layer output.

2.3 Enhanced by Multi-task Learning

We train our ViL-Sum with text summary generation task and two auxiliary tasks in a multi-task manner, which are used to enhance vision-language representation and semantic alignment.

2.3.1 Visual-enhanced Summary Generation

We feed the vision-language representation H_L from the multi-modal encoder as input of decoder with 12 transformer blocks. The target of the model is to minimize the negative log-likelihood of label text y tokens given input document D and images I via updating model parameters θ. The loss function of summary generation task is as follows:

$$\mathcal{L}_\theta^{GEN} = -\sum_{j=1}^{|y|} \log P_\theta(y_j | y_{<j}, D, I) \quad (4)$$

2.3.2 Images Selection

We also train our ViL-Sum with multi-modal output reference following (Zhu et al., 2020). To build pseudo image selection labels of training data, we employs similarity between image caption and gold summary to select top-K images as labels $\hat{y}$ (K is empirically set as 3). The similarity is the average of ROUGE-1, ROUGE-2 and ROUGE-L scores. The probability to select each image is $y_i = P(img_i) = \sigma(W \cdot h_{v_i} + b)$ and loss function of the image selection task is as follows:

$$\mathcal{L}_\theta^{IS} = \frac{1}{M} \sum_{i=1}^{M} -[\hat{y}_i \log y_i + (1 - \hat{y}_i) \log(1 - y_i)] \quad (5)$$

2.3.3 Images Reordering

To further enhance the vision-language representation and semantic alignment, we proposed the image reordering task to joint train ViL-Sum. We shuffle the order of input images and then employ vision representations to predict the position of each image through $y_i = P(pos_i) = \texttt{softmax}(W \cdot h_{v_i} + b)$ and minimize the objective function:

$$\mathcal{L}_\theta^{IR} = \frac{1}{M} \sum_{i=1}^{M} \sum_{c=1}^{C} -\hat{y}_{ic} \log y_{ic} \quad (6)$$

where C is the number of categories, depending on the number of input images.

	train	valid	test
#Documents	293,965	10,355	10,261
#AvgTokens(D)	721	766	731
#AvgTokens(S)	70	70	72
#Images	1,928,356	68,520	71,509
#AvgImgs	6.56	6.62	6.97

Table 1: Statistical information of MSMO. D refers to the input document. S refers to the summary.

2.3.4 Joint Training

We train our ViL-Sum with all three tasks (i.e., Summary Generation, Image Selection, Image Reordering) jointly by simultaneously minimizing three loss functions.

$$\mathcal{L}_\theta^{TOTAL} = L_\theta^{GEN} + L_\theta^{IS} + L_\theta^{IR} \quad (7)$$

However, the caption of the image is not always available. If we remove the images selection task, we can select images via measuring similarity between generated summary and vector representations of images. Our proposed multi-modal encoder and the image reordering task still help the model achieve excellent performance.

3 Experiments

3.1 Dataset

We employ the MSMO dataset (Zhu et al., 2018) to evaluate the effectiveness of our proposed methods. MSMO is a large-scale dataset for Multi-modal Summarization with Multi-modal output. Each example in the dataset is a triplet (document, images, summary). This dataset contains online news articles (723 tokens on average) paired with multiple image-caption pairs (6.58 images on average) and multi-sentence summaries (70 tokens on average). For test data, based on text reference, at most three images are annotated to produce a multi-modal reference by humans. The detailed statistical information of MSMO is shown in Tab. 1.

3.2 Settings

We train our model for 10 epoches on 8 V100 GPUs using Adam (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.99$, a batch size of 64. We also use linear learning rate warm-up with 1,000 steps. The weight-decay is set as 10^{-4}. We employ ViT-B/16 and BART-base to initialize our image tokenizer and the main encoder-decoder model. We set the max length of input images and tokens to be 10 and 512 respectively. For image tokenizer, we employ the same setting with ViT-b/16 in (Dosovitskiy

et al., 2021). When testing, we generate the summary with a beam size of 3, and the minimum and maximum decoding lengths are set as 15 and 150 separately.

3.3 Metrics

We evaluate the pictorial summary with the MMAE metric (Zhu et al., 2018). [1]

MMAE consists of three sub-metrics: ROUGE score (ROUGE-L), Image Precision (IP), and Image Text Relevance (MAX_{sim}). ROUGE (Lin, 2004) score can measure the salience of text in generated summary, which is widely used for measuring summarization systems. The image precision can measure the salience of selected images and is computed as Equ. (8).

$$\textbf{IP} = \frac{|ref_{img} \cap rec_{img}|}{|rec_{img}|} \tag{8}$$

where ref_{img} and rec_{img} denote reference images and recommended images by MSMO systems respectively. MAX_{sim} can measure the relevance between selected images and generated text summary, which trains an image-text retrieval (Faghri et al., 2018) model with max-margin loss to evaluate Image-Text relevance. Finally, Zhu et al. (2018) choose the linear regression results of 3 metrics as MMAE with human judgments and the weight for ROUGE-L, MAX_{sim}, and IP is 1.641, 0.854, 0.806 respectively, the intercept is 1.978.

We report the results of ROUGE-1/2/L, MAX_{sim}, IP, and MMAE of each model to comprehensively measure their performance. The results of our model are all the averages of three different checkpoints.

3.4 Comparison Models

To show the effectiveness of our models, we compare our model with the existing multimodal summarization methods (ATG, ATL, HAN, GR) (Zhu et al., 2018) and MOF^{RR}_{dec} (Zhu et al., 2020) using multiple metrics. We also report the result of PGC (See et al., 2017), which is a single-modal summarization model. To prove the effectiveness of our proposed joint representation and multi-task learning, we compared with BART-base (Lewis et al., 2020) model and a reproduced two-stream model BART-cross which has the same structure with MOF^{RR}_{dec} and replace GRU and VGG19 (Liu and Deng, 2015) with BART and ViT (Dosovitskiy et al., 2021) respectively. To be fair, we mainly compare our model with BART-base and BART-cross due to previous methods did not employ pre-trained models.

1) **PGC:** It is the widely used pointer-generator network that allows both copying words from the input text and generating words from a fixed vocabulary.

2) **ATG:** It refers to the multi-modal attention model which fuses images with static visual features from VGG19 (Liu and Deng, 2015) and selects images by measuring the visual attention distribution.

3) **ATL:** It replaces the image global features of ATG with local features (multiple pooling features), which select images by measuring the sum of visual attention distribution over the local patch features of each image.

4) **HAN:** It is based on ATL and a hierarchical attention mechanism is added, which first attends to the image patches to get the intermediate vectors to represent images and then attends to these vectors to get the visual context vector.

5) **GR:** It is an extractive method that employs LexRank (Erkan and Radev, 2004) to rank captions of images and select images based on the rank score of captions. And it also employs PGC to generate the text summary.

6) **MOF^{RR}_{dec}:** MOF is based on ATG and training with multi-modal optimization function, MOF^{RR}_{dec} is the best version of MOF which feed the last hidden state of decoder into image discriminator (dec) with ROUGE-ranking (RR) pseudo-labels.

7) **BART-base:** It is a pre-trained seq2seq generation model, which achieved promising results in many generation NLP tasks, especially on text summarization. We employ this model to confirm the contribution of visual features for a summary generation.

8) **BART-cross:** We build BART-cross with model structure from previous ATG, ATL, HAN, GR, and MOF^{RR}_{dec}. It encodes images and text with different encoders. We employ BART-base as the main encoder-decoder model and encode text input. Before feeding into the BART decoder, we fuse image and text representation with cross attention like ATG. BART-cross is a strong baseline with separate encoders for different modalities.

[1] Comment: Zhu et al. (2020) also proposed a MMAE+ to better evaluate MSMO task. However, the author did not release their MR model, which is the core component of their MMAE+. We find that the performance of MMAE and MMAE+ is very closer and consistent.

	Model	ROUGE-1	ROUGE-2	ROUGE-L	MAX$_{sim}$	IP	MMAE
baseline	PGC (See et al. 2017)	41.11	18.31	37.74	-	-	-
	ATG (Zhu et al., 2018)	40.63	18.12	37.53	25.82	59.28	3.35
	ATL (Zhu et al., 2018)	40.86	18.27	37.75	13.26	62.44	3.26
	HAN (Zhu et al., 2018)	40.82	18.30	37.70	12.22	61.83	3.25
	GR (Zhu et al., 2018)	37.13	15.03	30.21	**26.60**	61.70	3.20
	MOF$^{RR}_{dec}$ (Zhu et al., 2020)	**41.20**	**18.33**	**37.80**	26.38	**65.45**	**3.37**
ours	BART-base	43.75	20.70	40.66	-	-	-
	BART-cross	43.67	20.65	40.65	30.25	65.98	3.45
	ViL-Sum	**44.29**	**20.96**	**41.34**	32.17	66.27	3.48
	+ selection	44.20	20.90	41.22	34.47	68.18	3.51
	+ reordering	44.21	20.98	41.20	34.35	69.03	3.52
	+ selection, reordering	44.16	20.88	41.21	**34.52**	**71.73**	**3.55**

Table 2: The main results of all comparison models on different metrics. Models in baseline are based on the pointer network with Bi-GRU. Models in ours are based on the BART-base model. All reported results of ours are the average of three different checkpoints.

3.5 Main Results

3.5.1 Performance of Joint Representation

The main results of all models are shown in Tab. 2. We can see that compared with the baselines, our ViL-Sum gains significant improvement on all metrics, and the joint modeling of the two modalities does not hurt the performance of ROUGE scores. However, the performance of ATG, ATL, HAN, and GR all hurt ROUGE scores by simply introducing images as independent visual features. Through the multi-modal objective optimization, MOF$^{RR}_{dec}$ has a significant improvement on IP and does not decrease the quality of generated text summary. This situation proves that modeling vision and language information independently did not bring in the revenue for text summary generation. We can see that BART-cross, which also introduces images as independent features, also has lower ROUGE scores than BART-base, which also proves the previous conclusion. Our ViL-Sum obtains better ROUGE scores via encoding different modalities with the same encoder and the Image Precision (IP) and MAX$_sim$ both have a significant improvement. This demonstrated that using the joint multi-modal encoder to obtain vision-language representation is better than using separate encoders.

3.5.2 Performance of Multi-task Learning

The result of ViL-Sum without multi-task learning has achieved new state-of-the-art performance. In this section, we will analyze the influence of our proposed multi-task learning. From the results, we can see that the introduction of images selection and reordering bring a slight decrease in ROUGE scores. Meanwhile, the IP and MAX$_{sim}$ scores increase significantly, which makes the overall score MMAE is better than ViL-Sum without multi-task training.

We report the ablation study results of two auxiliary tasks in the second block of Tab. 2. From the results in the second block, we can see that images selection and reordering both can bring improvement on IP and MAX$_{sim}$ scores. The combination of two tasks can push the overall score of MMAE to a new state-of-the-art. The comparison of these models demonstrated that the introduction of multi-task learning exactly improved the vision-language representation and semantic alignment, which is reflected in the improvement of the multi-modal metrics: IP, MAX$_{sim}$ and MMAE.

Systems	Human Score
BART-base	3.29
BART-cross	3.46
ViL-Sum (best)	**3.78**
Gold	4.02

Table 3: Results evaluated by human annotators. Each summary is scored by three persons, and we take the average value.

4 Discussion

4.1 Human Evaluation

We randomly sample 100 examples from the MSMO test set to conduct the human evaluation. The multi-modal summary of gold reference, BART-base, BART-cross, and our ViL-Sum (best) are evaluated by three human annotators. Each annotator will give each example with a rating scale from 1 (worst) to 5 (best). Tab. 3 shows

K	ROUGE-L	MAX$_{sim}$	IP	MMAE
1	40.97	34.63	70.94	3.54
2	41.12	34.33	70.40	3.53
3	**41.21**	**34.52**	**71.73**	**3.55**
4	41.08	34.49	70.61	3.53

Table 4: Results of ViL-Sum under different hyper-parameters, where K is the selected image number.

	ROUGE-L	MAX$_{sim}$	IP	MMAE
ViT	**41.21**	**34.52**	**71.73**	**3.55**
Linear	40.18	33.89	70.44	3.51
Vision	41.10	34.28	71.04	3.54

Table 5: Results of ViL-Sum with different image tokenizer. Linear means image tokenizer which replace transformer blocks with linear layer. Vision is an image tokenizer from Vision Transformer.

the average scores from three annotators (t-test, $p < 0.05$). We can see that our ViL-Sum outperforms two strong baselines and BART-cross with multi-modal summary is better than BART-base with the single-modal summary.

4.2 Impact of Different K

Tab. 4 depicts the experimental results of our model performance varying with different K (the image number at summary). Since the gold reference in the test set contains three images, the consistency between training and test makes the model perform best when K is 3. We also can see that our model is not very sensitive with K.

4.3 Impact of Different Image Tokenizer

We employ different image tokenizers to prove the robustness of our framework. The results of them are shown in Tab. 5. Linear is the simple version of ViT which replaces the transformer image encoder with a simple linear layer to map the images into visual token embeddings. Vision is an image tokenizer from Vision Transformer (Wu et al., 2020), which can convert one image into several visual tokens embeddings. From the results, we can see that different tokenizers all can gain satisfactory performance.

4.4 Case Study and Relevance Visualization

To analyze the effectiveness of our proposed methods, we choose an example from the test set and visualized the relevance of 1) summary sentences and selected images; 2) selected paragraphs and images; 3) all tokens and images in Fig. 4 and 5. Each color block means cosine similarity between image and text object. The darker color refers to a higher similarity in heatmaps. From three different relevant visualizations, we can see that our model can effectively align semantic representation of summary sentences and selected images as shown in Fig. 4b. The input images can be aligned with paragraphs by training with image reordering as shown in Fig. 4c. We also report the heatmap of all input tokens and images in Fig. 5, which is consistent with Fig. 4b and 4c.

5 Related Work

5.1 Vision-Language Representation

Large-scale Transformers-based (Vaswani et al., 2017) models (Radford et al., 2019; Devlin et al., 2019; Lewis et al., 2020) have achieved the state-of-the-art results on many Natural Language Processing (NLP) tasks, which always pre-train on a large corpus with self-supervised tasks and then fine-tune on specific NLP tasks. With the success of transformers in NLP, many works begin to employ Transformers and pre-train with image-text pairs to joint represent vision and language semantic information for multi-modal downstream tasks, which can be called vision-language pre-training (VLP) model. Most existing VLP models (Tan and Bansal, 2019; Li et al., 2021) adopt two different encoders to model vision and language separately, which extracts visual features by an object detection model, and then combines the derived object-centric representation of the image and text embedding. Recently, many methods (Li et al., 2020b; Zhou et al., 2020; Li et al., 2020c; Zhang et al., 2021; Xu et al., 2021) employed an unified encoder to obtain vision-language representation and achieved better performance on downstream tasks.

5.2 Multi-modal Summarization

Recently, text summarization models have achieved remarkable performance on different type methods (Liu and Lapata, 2019; Zhong et al., 2020; Lewis et al., 2020; Zhang et al., 2019; Liang et al., 2021) with the development of pre-trained language models. Different from text summarization, multi-modal summarization is a task to generate a condensed summary to cover main information from multimedia data. One of the most significant characteristics of this task is it is not only based on text information, but it can also employ rich visual

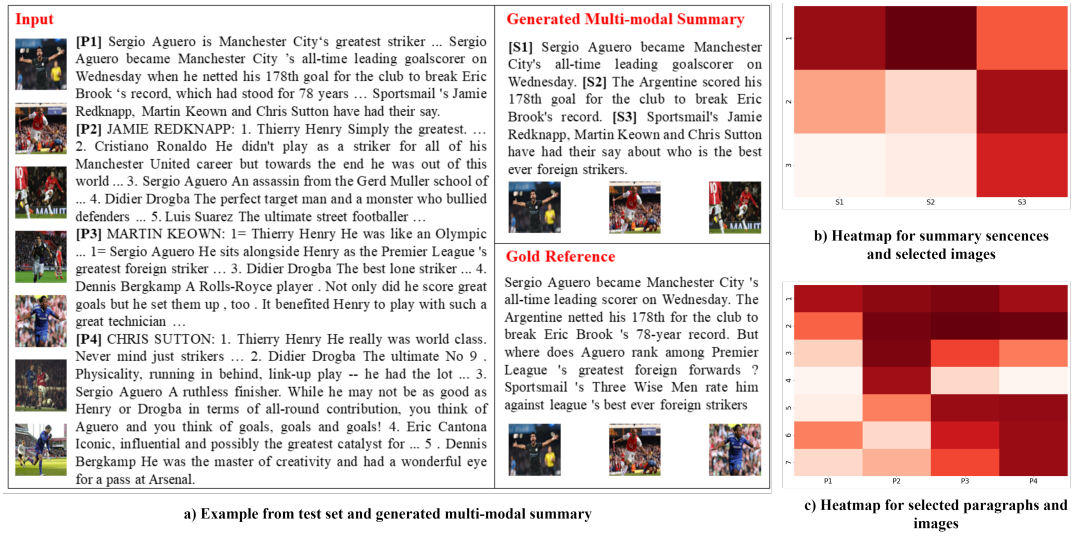

Figure 4: Example from the test set with the generated multi-modal summary. Fig. a) is the full example. Fig. b) is the heatmap that shows the relevance of the summary and selected images. Fig. c) is the heatmap that shows the relevance of selected paragraphs and images. Each color block means cosine similarity between image and text object. The darker color refers to higher similarity.

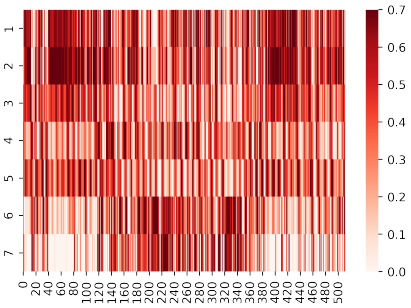

Figure 5: The heatmap shows the relevance of all input tokens and images. The darker color refers to higher similarity.

information from images, audio, and videos. Multi-modal summarization task can be divided into two types with different output: single-modal output (Evangelopoulos et al., 2013; Chen and Zhuge, 2018; Li et al., 2018) and multi-modal output (Bian et al., 2015; Zhu et al., 2018, 2020). Compared with single-modal output, multi-modal output summary can increase users' satisfaction (Zhu et al., 2018) and first proposed a large-scale Multi-modal Summarization with Multi-modal Output (MSMO) dataset. To tackle the gap between training and testing in MSMO task, Zhu et al. (2020) proposed two methods to obtain pseudo image labels and training the model with multi-modal optimization objectives.

However, previous works all obtain vision-language representation via separate encoders for different modalities, which has been proved weaker than joint representation in vision-language representation learning research (Zhou et al., 2020; Xu et al., 2021). Besides, they ignored the special semantic alignment between different modalities. In this paper, we proposed a novel Vision-Language Summarization (ViL-Sum) model with a multi-task learning framework to tackle these issues.

6 Conclusion

In this paper, we propose a novel Vision-Language Summarization (ViL-Sum) model with a multi-task learning framework, which can enhance the vision-language representation and align the semantics of different modalities. Our model achieved new state-of-the-art results on automatic and manual evaluation metrics.

Limitations: We only evaluate our model on one dataset due to the lack of MSMO datasets. However, we believe our model can obtain nice performance on other multi-modal tasks which need to align the semantic of paragraphs and images.

Broader Impact: The alignment of different modalities is an important problem in multi-modal tasks, in this paper, our proposed image reordering task is very simple yet effective for semantic alignment. We believe it can be employed in more scenarios (e.g. vision-language pre-training models). We also prove the joint modeling of images and texts is effective for the summary generation task.

References

Jingwen Bian, Yang Yang, Hanwang Zhang, and Tat-Seng Chua. 2015. Multimedia summarization for social events in microblog stream. *IEEE Transactions on Multimedia*, 17(2):216–228.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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*.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568.

Fartash Faghri, David J. Fleet, J. Kiros, and S. Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*.

Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. Self-supervised multimodal opinion summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 388–403, Online. Association for Computational Linguistics.

Aman Khullar and Udit Arora. 2020. MAST: Multimodal abstractive summarization with trimodal hierarchical attention. In *Proceedings of the First International Workshop on Natural Language Processing Beyond Text*, pages 60–69, Online. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR 2015*.

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.

Chenliang Li, Ming Yan, Haiyang Xu, Fuli Luo, Wei Wang, Bin Bi, and Songfang Huang. 2021. Semvlp: Vision-language pre-training by aligning semantics at multiple levels.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11336–11344.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020b. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11336–11344. AAAI Press.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018. Multi-modal sentence summarization with modality attention and image filtering. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4152–4158. International Joint Conferences on Artificial Intelligence Organization.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102, Copenhagen, Denmark. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020c. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*.

Xinnian Liang, Shuangzhi Wu, Mu Li, and Zhoujun Li. 2021. Improving unsupervised extractive summarization with facet-aware modeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1685–1697, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Shuying Liu and Weihong Deng. 2015. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734.

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.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. 2020. Visual transformers: Token-based image representation and processing for computer vision.

Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. 2021. E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 503–513, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13041–13049.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9749–9756.