International Journal of Software and Informatics, ISSN 1673-7288
http://www.ijsi.org, ijsi@iscas.ac.cn, +86-10-62661048
IJSI, 2023, 13(2): 143–155, doi: 10.21655/ijsi.1673-7288.00315

International Journal of Software and Informatics

Research Article

Multimodal Pre-training Method for Vision-language Understanding and Generation

Tianyi Liu (刘天义)[1,2,3], Zuxuan Wu (吴祖煊)[1,2,3], Jingjing Chen (陈静静)[1,2,3], Yugang Jiang (姜育刚)[1,2,3]

[1] (School of Computer Science, Fudan University, Shanghai 200438, China)
[2] (Shanghai Key Laboratory of IntelligentInformation Processing (Fudan University), Shanghai 200438, China)
[3] (Shanghai Collaborative Innovation Center of Intelligent Visual Computing (Fudan University), Shanghai 200438, China)
Corresponding author: Zuxuan Wu, zxwu@fudan.edu.cn; Yugang Jiang, ygj@fudan.edu.cn

Abstract Most existing vision-language pre-training methods focus on understanding tasks and use BERT-like loss functions (masked language modeling and image-text matching) during pre-training. Despite their good performance in the understanding of downstream tasks, such as visual question answering, image-text retrieval, and visual entailment, these methods cannot generate information. To tackle this problem, this study proposes Unified multimodal pre-training for Vision-Language understanding and generation (UniVL). The proposed UniVL is capable of handling both understanding tasks and generation tasks. It expands existing pre-training paradigms and uses random masks and causal masks simultaneously, where causal masks are triangular masks that mask future tokens, and such pre-trained models can have autoregressive generation abilities. Moreover, several vision-language understanding tasks are turned into text generation tasks according to specifications, and the prompt-based method is employed for fine-tuning of different downstream tasks. The experiments show that there is a trade-off between understanding tasks and generation tasks when the same model is used, and a feasible way to improve both tasks is to use more data. The proposed UniVL framework attains comparable performance to recent vision-language pre-training methods in both understanding tasks and generation tasks. Moreover, the prompt-based generation method is more effective and even outperforms discriminative methods in few-shot scenarios.

Keywords computer vision; multimodal learning; pre-training

Citation Liu TY, Wu ZX, Chen JJ, Jiang YG. Multimodal pre-training method for vision-language understanding and generation, *International Journal of Software and Informatics*, 2023, 13(2): 143–155. http://www.ijsi.org/1673-7288/315.htm

Inspired by the large-scale pre-trained models in natural language processing, people have proposed various vision-language pre-training methods to learn multimodal representations from large-scale image-text pairs. Once the pre-trained models are obtained, they can be fine-tuned to

This is the English version of the Chinese article “面向视觉语言理解与生成的多模态预训练方法. 软件学报, 2023, 34(5): 2024–2034. DOI: 10.13328/j.cnki.jos.006770”
Funding items: Major Program of Science and Technology Innovation—“New-Generation Artificial Intelligence” (2021ZD0112805); National Natural Science Foundation of China (62102092)
Received 2022-04-17; Revised 2022-05-29; Accepted 2022-08-24; IJSI published online 2023-06-29

execute the downstream tasks. This simple pre-training and fine-tuning paradigm has recently shown great potential in many challenging visual and linguistic tasks[1–5], such as visual question answering, image-text retrieval, image captioning, and visual implications.

The downstream vision-language tasks are divided into two categories: understanding tasks and generation tasks. Understanding tasks include visual question answering, visual implications, image classification, and image-text retrieval. Most existing vision-language pre-training methods standardize such tasks as discriminative tasks, which require models to select answers from pre-defined answer lists. For example, existing methods describe visual question answering as multiple-answer classification tasks and input the [CLS] tag into additional linear classifiers to obtain classification results[3, 4, 6–8]. These tasks usually require the model to roughly understand the semantic information of images and text. For instance, they need to judge whether the text describes the content of images and what the relationship (implication, neutrality, or contradiction) between image and text is. In contrast, generation tasks usually require the model to generate a complete sentence describing a specific image. A typical example is image captioning, which requires the model to output a sentence describing the content of the image. The existing vision-language methods use loss functions similar to BERT[9], such as Masked Language Modeling (MLM) and Image-Text Matching (ITM) for the learning of multimodal representations. They perform well in understanding tasks but cannot be directly applied to generation tasks.

Therefore, Unified multimodal pre-training for Vision-Language understanding and generation (UniVL) is proposed in this study. This pre-trained model can simultaneously process understanding tasks and generation tasks by sharing parameters. Specifically, an image encoder and a text encoder are used to encode images and texts, respectively. After that, a multimodal encoder is employed to fuse image and text features with cross-attention. Like the existing vision-language pre-training methods, this method uses two common training objectives, namely, MLM and ITM. However, unlike the previous methods, we use not only bidirectional visual masks but also causal masks in pre-training. Causal masks allow the model to carry out autoregressive decoding, which is essential for the generation of complete sentences. Unified pre-training for understanding and generation results in a unified model with shared parameters, thus reducing the need to train different models.

Typical vision-language pre-training methods always train multiple task-specific linear layers for different downstream tasks. This strategy is to enable a pre-trained model to adapt to different downstream tasks, and it needs to design task-specific objective functions. The prompt-based method has recently attracted people's attention, and it has been proven to be simple and effective. Downstream tasks can be re-normalized to the task pattern learned by the pre-trained model during pre-training. For instance, during theme classification, when the input is the sentence "He is playing basketball", the output includes multiple labels, such as health, politics, and sports. There is no need to add another linear classifier to fine-tune the pre-trained model; instead, a query statement can be constructed, "He is playing basketball. The theme is about __", and the pre-trained model is required to fill in the blank. As the blank-filling task is one of the pre-training objectives (MLM), the pre-trained model is familiar with the task. Compared with the widely used pre-training and fine-tuning paradigm, the prompt-based method converts the input and output formats of different downstream tasks into those processed by the model during pre-training, which can better unleash the potential of the pre-trained model. In this study, some of the previous classification tasks are standardized into text generation tasks, and language templates are used to fine-tune the pre-trained model.

Section 1 of this paper introduces the relevant methods and research status of large-scale pre-trained models. Section 2 presents the UniVL model built in this paper. Section 3 verifies

the effectiveness of the proposed model through comparative experiments. Finally, a summary is given.

1 Work Related to Large-scale Pre-trained Model

With the successful pre-training of large-scale language models, vision-language pre-training has recently attracted attention. Relevant research shows that vision-language pre-trained models perform well in multiple downstream tasks. Most of the existing methods are based on the Transformer architecture and use training objectives similar to BERT, namely multimodal MLM[2, 5, 7] and ITM[1, 7]. The former is to predict mask words or the visual features of masks according to input images and textual contexts, while the latter is to predict whether the input image matches the input text. They use a bidirectional visual mask in the self-attention module, which leads to differences between the form of the pre-training task and the form of the downstream task that requires auto-regressive generation. Inspired by UniLM[10, 11], we combine the causal mask and the bidirectional visual mask before training so that our pre-trained model can have the ability to understand and generate at the same time.

As the number of parameters in the pre-trained model increases, fine-tuning all parameters of the pre-trained model for a downstream task brings about great costs. The idea of the prompt-based method is to convert the data input and output formats of the downstream task into the formats that the model has learned during pre-training. For example, in natural language processing, the emotion classification task can be expressed as a natural language sentence with a mask mark, and the model needs to fill in the word "positive" or "negative." Due to the similarity of data input, the pre-trained model can be directly applied to downstream tasks without parameter adjustment.

Early prompt-based methods usually employ hand-made templates. For example, Petroni *et al*.[12] manually designed a cloze template for knowledge exploration tasks; GPT-3[13] designed prefixed templates for question-answering, translation, and exploration tasks. Although these manually designed templates are highly interpretable and generally effective, they need to be run in numerous trials. Different tasks require different fields of experience. In this paper, hand-made natural language templates and parametric templates are used for experiments.

2 Model Structure

The pre-trained model proposed in this paper consists of the visual encoder, the text encoder, and the multimodal encoder; their detailed introduction is given below.

2.1 Visual encoder

ViT[14] pre-trained on ImageNet-1k is taken as the visual encoder to extract image features. First, the input image $I \in R^{C\times H\times W}$ is expanded to $N = HW/P^2$ image blocks, where the resolution of the input image is $H \times W$; C is the number of channels; and the resolution of each image block is $P \times P$. Similar to the [CLS] tag used by BERT, ViT prepares a parametric learnable tag [CLS] for the image sequence. The visual encoder is composed of an alternating Multi-head Self-Attention (MSA) module and a Multi-Layer Perceptron (MLP) module, wherein the MLP module contains two linear layers and an activation layer. The visual encoder also uses layer normalization and residual connection in each layer.

$$
\begin{aligned}
z_0 &= [v_{\mathrm{CLS}}; v_p^1 V; v_p^2 V; \ldots; v_p^N V] + V_{\mathrm{pos}} \\
z_l' &= MSA(LN(z_{l-1})) + z_{l-1} && l = 1, \ldots, L_V \\
z_l &= MLP(LN(z_l')) + z_l' && l = 1, \ldots, L_V
\end{aligned}
$$

where $v_p^1, \ldots, v_p^N$ are the expanded two-dimensional (2D) image blocks; [CLS] is the head tag that can be learned, and z_l is the hidden state of the lth layer.

2.2 Text encoder

BERT is used as the text encoder. Similar to the visual encoder, the text encoder contains several layers of the MSA and MLP modules. The difference is that layer normalization is used after such modules. The input text $t \in R^{L\times O}$ uses the word embedding matrix $T \in R^{O\times H}$ and the position code $T_{\text{pos}} \in R^{(L+1)\times H}$ to be embedded as $t \in R^{L\times H}$.

$$
\begin{aligned}
p_0 &= [t_{\text{CLS}}; t^1T; t^2T; \ldots; t^NT] + T_{\text{pos}} \\
p_l' &= LN(MSA(p_{l-1})) + p_{l-1} && l = 1, \ldots, L_T \\
p_l &= LN(MLP(p_l')) + p_l' && l = 1, \ldots, L_T
\end{aligned}
$$

where $t^1, \ldots, t^N$ are the input words and p_l is the hidden state of the lth layer input sequence.

2.3 Multimodal encoder

The multimodal encoder is similar to the text encoder but requires an additional cross-attention calculation to fuse image features and text features.

$$
\begin{aligned}
m_0 &= p_{L_T} \\
m_l'' &= LN(MSA(m_{l-1})) + m_{l-1} && l = 1, \ldots, L_M \\
m_l' &= LN(MCA(m_l'', z_{L_V})) + m_l'' && l = 1, \ldots, L_M \\
m_l &= LN(MLP(m_l')) + m_l' && l = 1, \ldots, L_M
\end{aligned}
$$

where p_{L_T} is the output of the text encoder, z_{L_V} is the output of the visual encoder, and m_l is the hidden state of the lth layer sequence.

The attention mask used in the MSA module is bidirectional, and each tag can pay attention to all other tags. The bidirectional visual mask performs well in the discriminative task but is not suitable for the generation task. Generally, the generation task requires the model to generate tags in an autoregressive way, that is, from left to right. To solve this problem during pre-training, this study mixes two kinds of attention masks in different proportions in the self-attention module of the text encoder and of the multimodal encoder.

2.4 Multimodal prompt template

The previous vision-language pre-training methods always use the [CLS] tag as the multimodal image-text representations and add other linear layers to fine-tune the downstream tasks. For example, in Uniter[4], the visual question answering is described as a multiple-answer classification problem, which uses the [CLS] tag as the input of the linear layer and fine-tunes the linear layer. In contrast, this study uses the template to describe the visual question answering as a text generation problem. For instance, when answering "What is he doing?", we can continue to input the prompt "Answer:", and thus the complete input accepted by the model is "What is he doing? Answer:". The pre-trained model is required to fill such gaps through generation. In addition, as shown in Figure 1, the manually designed language template is replaced with parametric learnable tags. This is because it is difficult to design an appropriate natural language prompt template. It requires the expertise of domain experts and a significant amount of time for adjustment as subtle changes in each word of natural language prompts can have a significant impact on the performance of downstream tasks. The [UNUSED] tag of the marker is taken as a learnable tag as it is parametric and can be updated through backpropagation.

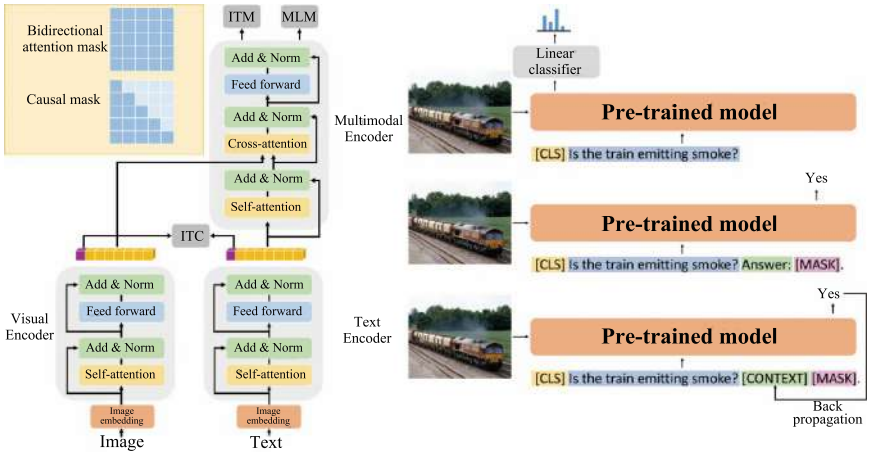

Figure 1 Model structure and multimodal prompt template proposed in this paper

2.5 Training objectives

Image-text contrastive loss has been proven to be effective for vision-language pre-training. Hence, it is applied to learn a common low-dimensional space to embed images and text. The matched image-text pairs are regarded as positive samples, and all other random image-text pairs in the training batch are regarded as negative samples. The sum of the two losses is minimized: one for image-to-text and the other for text-to-image.

$$\mathcal{L}_{i2t} = -\frac{1}{B}\sum_{i}^{B}\log\frac{\exp(x_i^{\mathrm{T}}y_i/\sigma)}{\sum_{j=1}^{B}\exp(x_i^{\mathrm{T}}y_i/\sigma)}$$

$$\mathcal{L}_{t2i} = -\frac{1}{B}\sum_{i}^{B}\log\frac{\exp(y_i^{\mathrm{T}}x_i/\sigma)}{\sum_{j=1}^{B}\exp(y_i^{\mathrm{T}}x_i/\sigma)}$$

$$\mathcal{L}_{itc} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i}$$

where x_i is the first-layer output of the visual encoder, y_i is the first-layer output of the text encoder, B is the size of the batch data, and σ is the temperature parameter of the scaled value.

Text tags are randomly masked with a probability of 15% and are replaced with a special [MASK] tag. The model needs to use images and contexts to predict the masked words.

$$\mathcal{L}_{\mathrm{mlm}} = \sum H(p_{\mathrm{mask}}, y_{\mathrm{mask}})$$

where H is the cross entropy, p_{mask} is the prediction probability of the model for mask tags, y_{mask} is the probability distribution of words, and $\mathcal{L}_{\mathrm{mlm}}$ is the sum of the cross entropy of each masked word.

The first tag output by the multimodal encoder is taken as the fused representation of the vision and language modes, and then Softmax is used to predict the matching probability p_{itm} of the two after the addition of a fully connected layer. The ITM loss function predicts whether the image and text pairs match or not. The image or text in the matched samples is replaced with the image or text randomly selected from other samples to create a negative sample.

$$\mathcal{L}_{\mathrm{itm}} = \sum H(p_{\mathrm{itm}}, y_{\mathrm{itm}})$$

where y_{itm} is a two-dimensional thermal vector, representing the truth value label, where 1 represents the matched image-text pair, and 0 represents the mismatched image-text pair; $\mathcal{L}_{\text{itm}}$ is the sum of the cross entropy of all positive and negative samples.

3 Experimental analysis

3.1 Experimental data

The experiment used GCC3M and COCO as the pre-training data sets and followed Karpathy's segmentation method for COCO[15]. The total number of different images in the final training set reached 2.84 millions.

3.2 Implementation details

A 12-layer ViT-B/16[14] was taken as the image encoder and was initialized using the weights pre-trained on ImageNet-1k. The text encoder was initialized by the first six layers of the BERT model, and the multimodal encoder was initialized by the last six layers of the BERT model. 30 cycles of pre-training were performed with 2,048 batches of data on 32 NVIDIA Tesla V100 32 GB GPUs. The learning rate of the AdamW optimizer used was 1E–4, and the weight attenuation was 0.02.

3.3 Downstream tasks

Image-text retrieval requires the model to select an image that meets a given description from a candidate image set or to select a description statement that meets the image content from a candidate description set. Therefore, it includes two subtasks: image-to-text retrieval and text-to-image retrieval. The study used the image-text contrastive loss and ITM loss to evaluate the similarity between images and texts. During the reasoning process, the visual encoder and text encoder were first employed to calculate the feature similarity of all image–text pairs. Then, the top K pairs with the highest score were selected. After that, the multimodal encoder was used to calculate the ITM scores and rank them. The proposed model was evaluated on Flickr30K[16] and COCO[17].

Image captioning aims to generate sentences describing image contents. Since causal masks are used in this study to pre-train the model, the proposed model can directly generate a sentence for the image. In the sentence generation process, the tags [CLS] and [MASK] were first used as input to encode the input of the image encoder. The special tag [CLS] was the beginning of the sentence, and the proposed model predicted the word at [MASK]. Then, another [MASK] was added to the generated tag sequence to predict the next word, and so on. When the model output [SEP], the generation process terminated. Beam search was used in the experiment, with a beam size of 5, and the experimental results on the COCO image caption data set were reported.

Visual question answering requires the model to answer a given question for a given image. The existing methods usually describe visual question answering as a multiple-answer classification problem. In this study, visual question answering was regarded as a text-generation task. Two prompt templates were designed for visual question answering: natural language prompt template and parametric prompt template for learnable contexts. The proposed method was evaluated on VQAv2[18].

Fine-grained image classification focuses on identifying image classes that are difficult to distinguish, such as the species of flowers or animals. In this study, the fine-grained image classification task was used to evaluate the multimodal understanding ability of the model. The fine-grained image classification task was standardized to text generation, and the natural language prompt template and the parametric prompt template of learnable contexts were designed. Compared with the discriminative method, the prompt-based method has a better

few-shot learning ability. The proposed method was evaluated on Food101[19], Flowers102[20], and DTD[21].

Visual implications[22] refer to a fine-grained visual reasoning task used to predict whether the relationship between images and text is inclusive, neutral, or contradictory. Natural language prompt templates and parametric prompt templates that could learn contexts were designed for visual implications, and it was compared with discriminant methods.

3.4 Experimental results

The image-text retrieval task was employed to evaluate the vision-language understanding ability of the pre-trained model. Table 1 reports the results of zero-shot and fine-tuned image-text retrieval on Flickr30K. For zero-shot retrieval, UniVL uses fewer data to achieve results similar to those of CLIP[23] and ALIGN[24]. For fine-tuned retrieval, the recall of UniVL is much higher than that of UNITER[4] and similar to that of ALIGN[24], though it is pre-trained on a larger data set (1.2B).

Table 1 Experimental results of multimodal retrieval

Method	Quantity of pre-trained images	Multimodal retrieval result (zero-shot/fine-tuned) on Flickr30K					
		Image-text retrieval			Text-image retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
UNITER	4M	83.6/87.3	95.7/98.0	97.7/99.2	68.7/75.6	89.2/94.1	93.9/96.8
CLIP	400M	88/–	98.7/–	99.4/–	68.7/–	90.6/–	95.2/–
ALIGN	1.2B	**88.6/95.3**	98.7/**99.8**	99.7/**100**	**75.7/84.9**	**93.8/97.4**	**96.8/98.6**
UniVL	3M	86.8/94.3	**98.7**/99.4	**99.7**/99.8	73.4/82.8	92.1/96.7	96.0/98.4

The image captioning task was used to evaluate the generation ability of the pre-trained model. Following Karpath's segmentation method, this study evaluated the performance of automatic image captioning on the COCO dataset. The algorithm re-split the training image and the verification image into 113,287, 5,000, and 5,000 for training, verification, and testing, respectively. Table 2 shows the results of four common indicators, i.e., BLEU4[25], CIDEr[26], METEOR[27], and SPICE[28]. In addition, the proposed pre-trained model was compared with other generative vision-language pre-training methods. The result indicates that the UniVL model has comparable performance to the recent generative pre-training methods.

Table 2 Experimental results of image captioning

Method	BLEU4	CIDEr	METEOR	SPICE
Unified VLP	36.5	117.7	28.4	21.3
XGPT	**37.2**	**120.1**	28.6	21.8
VL-T5	34.6	116.1	**28.8**	**21.9**
VL-BART	34.2	114.1	28.4	21.3
UniVL	35.6	116.8	28.6	21.4

Visual question answering was standardized as a text-generation task rather than a multi-answer classification task, and a natural language prompt template and a parametric prompt template that could learn contexts were designed for it. The natural language prompt template is "[QUESTION] Answer: [ANSWER]", where [QUESTION] represents the question text, and [ANSWER] represents the answer text. The word tag in [ANSWER] was masked, and the MLM loss was optimized in the fine-tuning process. In the reasoning process, the input text was "[QUESTION] Answer: [MASK]". The model predicted the word and added [MASK] repeatedly to the generated sequence until the tag [SEP] was obtained. The parametric prompt template of the learnable context replaced the natural language prompt with a learnable tag, while for the visual question answering, it was "[QUESTION] [CTX] [Answer]". [CTX] is the sequence of learnable tags, and this study uses the tag [UNUSED] of the BERT marker as

[CTX] with a length of 16. Compared with the natural language prompt, [CTX] is a parametric tag prompt that can be updated with other parameters.

The previous discriminative method standardizes the visual question answering as a multi-answer classification problem and requires the model to select answers from the pre-defined answer list. For a fine-grained comparison of the discriminative method and the generative method, this paper divided Karpathy's test data set into two categories: questions with answers in the pre-defined answer list (intra-domain samples) and questions without answers in the list (extra-domain samples). The size of the pre-defined answer list was 3,129; the number of intra-domain and extra-domain samples was 25,750 and 530, respectively. Typical discriminative methods cannot answer extra-domain questions because they have few answers, and the answers are not in the pre-defined answer list. To compare the generalization ability of discriminative methods and generative methods, we appended the answers of extra-domain samples to the pre-defined answer list and used the expanded answer list to fine-tune the discriminative method. For discriminative methods, we input the first hidden state output by the multimodal encoder into an additional linear classifier to predict the answer. As shown in Table 3 (LC indicates fine-tuning with linear layers, NLP implicates the use of a natural language template, and LCP denotes the learnable parametric template), compared with the discriminative method, the generative method based on prompts performs better in both categories, and in the comparison about the extra-domain samples, the improvement is more significant. Moreover, we used different amounts of Karpathy training data to evaluate the few-shot learning ability of the prompt-based method. As shown in Table 4, the performance of the natural language prompt template and the parametric prompt template of the learnable context is better than that of the discriminative method.

Table 3 Experimental results of visual question answering on COCO data set

Method	Intra-domain	Extra-domain	Mean
LC	70.8	3.7	69.4
NLP	68.4	13.9	67.3
LCP	**72.1**	**15.1**	**71.0**

Table 4 Experimental results of visual question answering under different numbers of training samples

Method	Number of training samples (intra-domain/extra-domain)			
	4k	22k	44k	88k
LC	0.5/0	5.6/0	10.9/0.5	15.4/0.9
NLP	0.9/**0.1**	11.9/**0.9**	14.8/1.1	18.3/1.6
LCP	**0.9**/0	**12.4**/0.7	**16.7/1.5**	**20.1/2.4**

For a fair comparison with the latest vision-language pre-training method, we referred to the previous method UNITER[4] and used the training set and verification set of VQAv2 to fine-tune the pre-trained model. As shown in Table 5, UniVL has achieved the performance equivalent to the state-of-the-art method.

Table 5 Experimental results of visual question answering and visual implication test sets

Method	Visual question answering		Visual implications	
	test-dev	test-std	val	test
VisualBERT	70.8	71	–	–
12-in-1	73.15	–	–	76.95
UNITER	72.7	72.91	78.59	78.28
VilT	70.94	–	–	–
VILLA	**73.59**	**73.67**	79.47	79.03
UniVL	72.31	72.53	**79.70**	**80.00**

Similar to visual question answering, we standardized image classification as a text-generation task and designed a natural language prompt template and a parametric prompt template of learnable contexts for image classification. The natural language prompt template was "a photo of [CATEGORY]", and the corresponding learnable context prompt was "[CTX] [CATEGORY]". [CATEGORY] is the class name of the image, which was masked during the fine-tuning process. As shown in Table 6, for this downstream task, the parametric prompt template of learnable contexts is an effective method because it is a form closer to the pre-training task.

Table 6 Experimental results of image classification using discriminative method and generative method

Method	Fine-tuning module	Food101	Flowers102	DTD
LC	VE	92.8	**93.8**	**65.4**
NLP	VE	88.4	90.1	53.3
LCP	VE	92.8	93.4	62.1
	TE	78.6	74.6	19.2
	ME	79.4	76.2	20.6
	VETE	92.8	93.5	63.3
	VEME	**93.3**	93.7	63.5

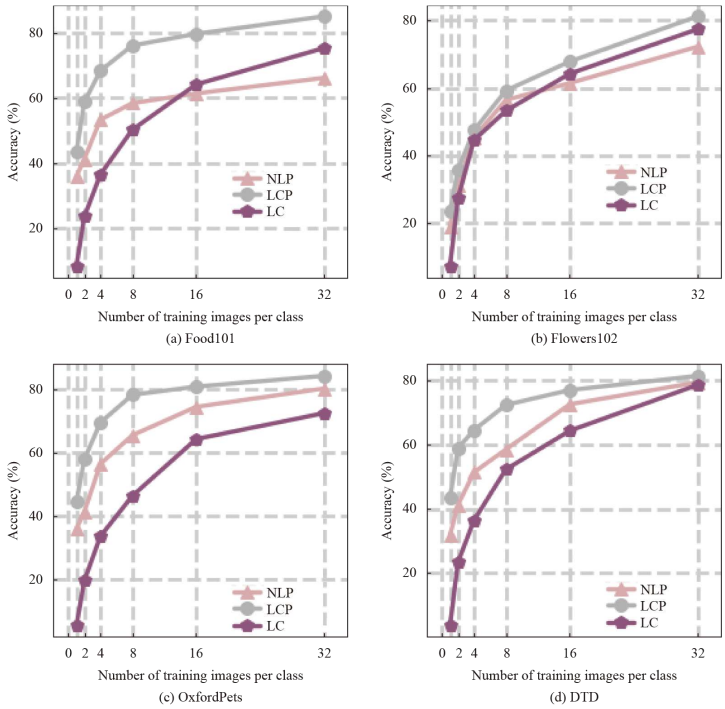

Figure 2 Few-shot learning results of different methods

Figure 2 shows that the parametric prompt template of learnable contexts has a better few-shot learning ability. Since the prompt template is a more familiar input data form of the pre-trained model, the prompt template of learnable contexts can better use the knowledge learned from the pre-trained model. Unlike visual question answering, the image classification task has simple text input, and the number of training samples is small. Therefore, it is unwise to update all parameters of the pre-trained model. As shown in Table 6, the visual encoder is

the key to the fine-grained image classification task in that the text is very simple and contains almost no semantic information except the class name.

Not all generative methods perform better than discriminative methods. Discriminative methods are more suitable for downstream tasks with fewer classes, and visual implications are one of them. Upon the design of the natural language prompt template and the parametric prompt template of learnable contexts, visual implications were standardized into a text generation task. The natural language prompt was "[SENTENCE] Relationship: [LABEL]", and the parametric prompt template of learnable contexts was "[SENTENCE] [CTX] [LABEL]". [LABEL] refers to the relationship between image and [SENTENCE], which could be implicated, neutral, and contradictory, which was masked in the fine-tuning process. Unlike predicting a word from the vocabulary according to the score of each word, we also sorted the score of each possible answer and returned the answer with the highest score in the reasoning process. This is a discriminative method, which is more suitable for the pre-trained model compared with the use of additional linear classifiers because the input is an image with text, with the goal of MLM. The results are shown in Table 7, where 1 in 3 means that we use three words to predict [MASK]: implication, neutrality, and contradiction. It should be noted that this should be realized in visual implications since each answer of visual implications is a word, and there is no common prefix. Compared with the generative method, the discriminative method is more suitable for visual implications because the set of candidate answers is too small.

Table 7 Experiment results of visual implications

Method	val	test
LC	78.4	78.1
NLP	65.4	65.7
NLP (1 in 3)	75.3	75.9
LCP	77.6	78.0
LCP (1 in 3)	**79.7**	**80.0**

3.5 Ablation experiment

We first evaluated the effectiveness of the causal mask matrix in pre-training. Table 8 shows the multimodal generation and understanding abilities of the pre-trained model. During pre-training, we used different mixing ratios of the bidirectional attention mask matrix and the causal mask matrix, as well as different numbers of image-text pairs. We employed the image captioning task to evaluate the generation ability of the model and used image classification, visual implications, and image-text retrieval to evaluate the understanding ability of the model. For image captioning, we appended the special tag [MASK] to the sequence and predicted the word iteratively until the model output the special tag [SEP]. For understanding tasks, we input

Table 8 Experimental results of different data volumes and different causal mask matrix proportions

Quantity of image-text pairs	Proportion of causal mask matrixes	Generation task				Understanding task			
		Image captioning				Visual implications	Image-text retrieval	Text-image retrieval	Image classification
		B1	B4	R	C	Acc	Acc	Acc	Acc
0.75M	0.0	15.7	3.5	10.1	11.7	50.9	56.4	43.3	64.3
	0.33	50.7	20.2	35.8	68.5	49.6	52.1	41.0	59.7
	0.66	58.7	23.4	37.5	78.4	47.5	51.8	39.8	66.9
	1.0	66.9	24.8	38.7	84.9	33.3	41.9	27.7	47.1
1.5M	0.0	24.9	5.3	16.8	18.2	73.5	82.6	70.4	85.4
	0.33	59.8	22.9	38.0	73.7	72.4	79.4	67.9	79.9
	0.66	65.1	24.8	38.9	82.4	72.2	80.8	69.4	85.1
	1.0	69.4	26.0	39.4	89.7	61.9	70.2	61.3	61.7
3.4M	0.5	**96.1**	**35.6**	**67.0**	**116.8**	**78.1**	**94.3**	**82.8**	**92.8**

the first hidden state output by the multimodal encoder to an additional linear classifier to predict the answer. Understanding tasks require the model to judge the relationship between the image and the sentence, which is a closed task. The model needs to select answers from the pre-defined set. It is more difficult for generation tasks because the model needs to generate open answers.

As shown in Table 8, as the proportion of the causal mask matrix in pre-training rises, the model performs better in generation tasks. However, with the increase in causal masks and the decrease in bidirectional attention masks, the model performs worse in understanding tasks. It is found that in general, causal masks are beneficial for generation tasks while bidirectional attention masks are conducive to understanding tasks. The increase in training data has greater benefits for both understanding and generation tasks.

We used the tag [UNUSED] of the BERT marker as a component of the parametric prompt template of learnable contexts. For visual question answering and visual implications, the input text contains a sentence, a prompt template, and a tag [MASK], and the prompt template can be at the beginning or in the middle of the input text. It is worth noting that the prompt should not appear at the end of the input text due to the causal masks used, and [MASK] cannot process the tag on the right. For image classification, the input text only contains [CTX] and [MASK], and the prompt template should be on the left side of [MASK]. As shown in Table 9, it can be seen that only one [CTX] cannot effectively prompt the model, and as the prompt length grows, the accuracy of different downstream tasks can be improved. However, when the prompt template is too long, the efficiency is low. Although the length of 32 is doubled compared with the length of 16, the accuracy of downstream tasks remains almost unchanged. For visual question answering and visual implications, the prompt template may as well be placed in the middle rather than at the beginning because the prompt template in the middle is closer to [MASK] and is a more effective signal for the subsequent text generation.

Table 9 Experimental results for different prompt template length and position

Task	Position	Length of template				
		1	4	8	16	32
VQA	begin	66.4	69.2	70.4	70.8	71.0
VQA	mid	67.9	69.5	70.4	71.0	71.1
VE	begin	77.3	78.2	78.8	79.4	79.9
VE	mid	77.5	78.5	78.6	80.1	80.1
IC (Food101)	begin	90.6	91.4	91.9	92.1	92.5
IC (Flowers102)	begin	93.0	93.6	94.2	94.4	94.0

4 Summary

This paper proposed UniVL, which can handle vision-language understanding and generation tasks. The experiments showed that the proposed method has achieved equivalent performance to the current vision-language method for understanding and generation tasks. Moreover, the study put forward a prompt-based method, which is simple and effective and can fine-tune different downstream tasks.

References

[1] Lu JS, Batra D, Parikh D, Lee S. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 2.

[2] Su WJ, Zhu XZ, Cao Y, Li B, Lu LW, Wei FR, Dai JF. VL-BERT: Pre-training of generic visual-linguistic representations. Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020. 1–16.

[3] Lu JS, Goswami V, Rohrbach M, Parikh D, Lee S. 12-in-1: Multi-task vision and language representation learning. Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10437 –10446. [doi: 10.1109/CVPR42600.2020.01045]

[4] Chen YC, Li LJ, Yu LC, El Kholy A, Ahmed F, Gan Z, Cheng Y, Liu JJ. UNITER: Universal image-text representation learning. arXiv:1909.11740, 2020.

[5] Li XJ, Yin X, Li CY, Zhang PC, Hu XW, Zhang L, Wang LJ, Hu HD, Dong L, Wei FR, Choi Y, Gao JF. OSCAR: Object-semantics aligned pre-training for vision-language tasks. Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 121–137. [doi: 10.1007/978-3-030-58577-8_8]

[6] Li LH, Yatskar M, Yin D, Hsieh CJ, Chang KW. VisualBERT: A simple and performant baseline for vision and language. arXiv:1908.03557, 2019.

[7] Kim W, Son B, Kim I. ViLT: Vision-and-language transformer without convolution or region supervision. Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 5583–5594.

[8] Gan Z, Chen YC, Li LJ, Zhu C, Cheng Y, Liu JJ. Large-scale adversarial training for vision-and-language representation learning. Proc. of the 34th Advances in Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 6616–6628.

[9] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[10] Bao HB, Dong L, Wei FR, Wang WH, Yang N, Liu XD, Wang Y, Gao JF, Piao SH, Zhou M, Hon HW. UniLMv2: Pseudo-masked language models for unified language model pre-training. Proc. of the 37th Int'l Conf. on Machine Learning. PMLR, 2020. 642–652.

[11] Dong L, Yang N, Wang WH, Wei FR, Liu XD, Wang Y, Gao JF, Zhou M, Hon HW. Unified language model pre-training for natural language understanding and generation. Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 1170.

[12] Petroni F, Rocktäschel T, Riedel S, Lewis PSH, Bakhtin A, Wu YX, Miller AH. Language models as knowledge bases? Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: ACL, 2019. 2463–2473.

[13] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Proc. of the 34th Advances in Neural Information Processing Systems. Curran Associates Inc., 2020. 1877–1901.

[14] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021. 1–21.

[15] Karpathy A, Li FF. Deep visual-semantic alignments for generating image descriptions. Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 3128–3137. [doi: 10.1109/CVPR.2015.7298932]

[16] Plummer BA, Wang LW, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 2641–2649. [doi: 10.1109/ICCV.2015.303]

[17] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. Proc. of the 13th European Conf. on Computer Vision. Zurich: Springer, 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]

[18] Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6904–6913.[doi: 10.1109/CVPR.2017.670]

[19] Bossard L, Guillaumin M, van Gool L. Food-101–mining discriminative components with random forests. Proc. of the 13th European Conf. on Computer Vision. Zurich: Springer, 2014. 446–461. [doi: 10.1007/978-3-319-10599-4_29]

[20] Nilsback ME, Zisserman A. Automated flower classification over a large number of classes. Proc. of the 6th Indian Conf. on Computer Vision, Graphics & Image Processing. Bhubaneswar: IEEE, 2008.

722–729. [doi: 10.1109/ICVGIP.2008.47]

[21] Cimpoi M, Maji S, Kokkinos I, Mohamed S, Vedaldi A. Describing textures in the wild. Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 3606–3613. [doi: 10.1109/CVPR.2014.461]

[22] Xie N, Lai F, Doran D, Kadav A. Visual entailment: A novel task for fine-grained image understanding. arXiv:1901.06706, 2019.

[23] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021.8748–8763.

[24] Jia C, Yang YF, Xia Y, Chen YT, Parekh Z, Pham H, Le QV, Sung YH, Li Z, Duerig T. Scaling up visual and vision-language representation learning with noisy text supervision. Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 4904–4916.

[25] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: A method for automatic evaluation of machine translation. Proc. of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia: ACL, 2002. 311–318. [doi: 10.3115/1073083.1073135]

[26] Vedantam R, Zitnick CL, Parikh D. CIDEr: Consensus-based image description evaluation. Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 4566–4575. [doi: 10.1109/CVPR.2015.7299087]

[27] Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language. Proc. of the 9th Workshop on Statistical Machine Translation. Baltimore: ACL, 2014. 376–380. [doi: 10.3115/v1/W14-3348]

[28] Anderson P, Fernando B, Johnson M, Gould S. SPICE: Semantic propositional image caption evaluation. Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 382–398. [doi: 10.1007/978-3-319-46454-1_24]

Tianyi Liu, master's degree candidate. His research interest is computer vision.

Jingjing Chen, Ph.D., associate researcher. Her research interests include multimedia content analysis, computer vision, robust and trustworthy artificial intelligence.

Zuxuan Wu, Ph.D., associate researcher. His research interests include computer vision and deep learning.

Yugang Jiang, Ph.D., professor, Ph.D. supervisor. His research interests include multimedia information processing, computer vision, robust and trustworthy artificial intelligence.