A Survey of Vision-Language Pre-Trained Models

Yifan Du[1,3†], **Zikang Liu**[1†], **Junyi Li**[1,2] and **Wayne Xin Zhao**[1,3*]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]DIRO, Université de Montréal
[3]Beijing Key Laboratory of Big Data Management and Analysis Methods
{yifandu1999, jasonlaw8121, batmanfly}@gmail.com, junyi.li@umontreal.ca

Abstract

As transformer evolves, pre-trained models have advanced at a breakneck pace in recent years. They have dominated the mainstream techniques in natural language processing (NLP) and computer vision (CV). How to adapt pre-training to the field of Vision-and-Language (V-L) learning and improve downstream task performance becomes a focus of multimodal learning. In this paper, we review the recent progress in Vision-Language Pre-Trained Models (VL-PTMs). As the core content, we first briefly introduce several ways to encode raw images and texts to single-modal embeddings before pre-training. Then, we dive into the mainstream architectures of VL-PTMs in modeling the interaction between text and image representations. We further present widely-used pre-training tasks, and then we introduce some common downstream tasks. We finally conclude this paper and present some promising research directions. Our survey aims to provide researchers with synthesis and pointer to related research.

1 Introduction

We now live in a world with various modalities (voice, vision, odors, *etc.*), among which vision and language are two critical ones. In academia, there exist large amounts of works focusing on V-L tasks. These tasks require the agent to jointly process information from these two modalities and utilize them to answer complex questions. For example, visual question answering (VQA) [Antol *et al.*, 2015] takes an image and the corresponding question as input and gives the correct answer; image captioning [Lin *et al.*, 2014] generates a description for a given image.

Deep learning has revolutionized the fields of artificial intelligence. Various deep models have been applied to solve V-L tasks, such as recurrent neural network (RNN) [Arevalo *et al.*, 2017], convolutional neural network (CNN) [Huang *et al.*, 2020] and transformer [Vaswani *et al.*, 2017]. Despite the success of deep learning, most of these models are designed for specific tasks, which leads to poor transferability. Pre-training a huge model on large-scale general datasets and then fine-tuning it on specific downstream tasks is one technique to increase transferability. Pre-training is first discovered to be effective in the field of CV [Simonyan and Zisserman, 2014]. After the proposal of transformer [Vaswani *et al.*, 2017] and BERT [Devlin *et al.*, 2018], the paradigm of pre-training and fine-tuning becomes prevalent in the field of NLP. With its powerful capability to model long-range dependency, transformer has become the backbone of most Pretrained Language Models (PLMs). BERT and GPT-3 [Brown *et al.*, 2020] are typical PLMs that significantly outperform previous methods and achieve new state-of-the-art results on various downstream tasks.

Due to the success of pre-trained models in the field of CV and NLP, many works have tried to pre-train large-scale models on both vision and language modalities, called Vision-Language Pre-Trained Models (VL-PTMs). By pre-training on large-scale image-text corpora, VL-PTMs can learn universal cross-modal representations, which are beneficial for achieving strong performance in downstream V-L tasks [Zellers *et al.*, 2019; Tan and Bansal, 2019]. For example, LXMERT [Tan and Bansal, 2019] employs a dual-stream fusion encoder to learn V-L representations, significantly outperforming traditional models on VQA [Antol *et al.*, 2015], NLVR^2 [Suhr *et al.*, 2018] tasks via pre-training on 9.18M image-text pairs. Besides, VL-PTMs achieve strong results on many other V-L tasks like visual commonsense reasoning [Zellers *et al.*, 2019] and image captioning [Lin *et al.*, 2014]. These VL-PTMs utilize different single-modal encoders, elaborate V-L interaction schemes, and devise various pre-training tasks.

However, there lacks a comprehensive survey to summarize the recent progress in this field. Mogadala *et al.* [2021] mainly reviews existing V-L tasks, datasets, and traditional solutions but rarely introduces V-L pre-training methods. Ruan and Jin [2022] focus on Video-Language PTMs instead of Vision-Language PTMs. Different from them, our survey aims to present a thorough review of VL-PTMs, which summarizes recent research progress and provides pointers to related research. We present the recent mainstream VL-PTMs in Table 1.

Basically, there are three steps to pre-train a VL-PTM: 1)

[†] Equal Contribution.
[*] Corresponding Author.

encode images and texts into latent representations preserving their semantics (Section 2); 2) design a performant architecture to model the interaction between two modalities (Section 3); and 3) devise effective pre-training tasks to train the VL-PTMs (Section 4). After learning universal vision and language features, VL-PTMs can be fine-tuned on various downstream V-L tasks (Section 5). Finally, we conclude this survey and point out some promising research directions in Section 6.

2 Learning Vision-Language Representation

As discussed in Section 1, encoding images and texts as embeddings preserving input semantics is the first step in pre-training a VL-PTM. The ways of encoding images and texts are quite different because of the discrepancy between the two modalities. Almost all VL-PTMs utilize a transformer-based PTM as a text encoder, but how to learn visual representations based on visual contents is still an open problem. In what follows, we introduce several methods to encode the images and texts into single-modal embeddings before feeding them into a cross-modal transformer.

Pre-training Dataset. The initial step in pre-training VL-PTMs is to construct large-scale image-text pairs. We formally define the pre-training dataset as $\mathcal{D} = \{(W, V)\}_{i=1}^{N}$, where W and V denote the text and image, respectively, and N is the number of image-text pairs. Specifically, each text will be tokenized as a sequence of tokens $W = \langle w_1, ..., w_n \rangle$. Similarly, each image will be also transformed into a sequence of object features (or grid features, or patch features), denoted as $V = \langle v_1, ..., v_m \rangle$. In Table 2 we list several widely-used or recently proposed pre-training datasets.

Text Representation. Most of existing studies on VL-PTMs follow BERT [Devlin *et al.*, 2018] to preprocess the raw text. The text sequence is first split into tokens and concatenated with "[CLS]" and "[SEP]" tokens, denoted as $W = \langle \text{[CLS]}, w_1, ..., w_n, \text{[SEP]} \rangle$. Each token w_j will be mapped to a word embedding. Besides, a positional embedding indicating the position and a segment embedding indicating the modality type are added with the word embedding to obtain the final embedding of w_j.

Image Representation. To align with the sequence of embeddings of the paired text, the image V will also be represented as a sequence of embedding vectors. In this way, we can unify input representation as a sequence of embeddings for both modalities. Unlike relationships among words in text, relationships among visual concepts in images are critical for V-L tasks but difficult to capture. For example, in order to generate a description of an image, the model is expected to infer the complex relationships among various objects in the image. Therefore, many works elaborate different vision encoders to model these relationships and the attributes of objects. Early works like ViLBERT [Lu *et al.*, 2019] and LXMERT [Tan and Bansal, 2019] first utilize Faster R-CNN [Ren *et al.*, 2015] to detect a sequence of object regions from an image and then encode them as a sequence of Region-Of-Interest (ROI) features. Besides, some VL-PTMs get rid of the bounding box and encode an image into pixel-level grid features. For example, pixel-BERT [Huang *et al.*, 2020] and SOHO [Huang *et al.*, 2021] abandon Faster R-CNN in favor of ResNet so that the visual encoder could view an image as a whole, avoiding the risk of neglecting some critical regions. Apart from these methods, many works try to follow the success of ViT [Dosovitskiy *et al.*, 2020] to utilize a transformer to extract vision features. In this scenario, the transformer in VL-PTMs is tasked with the objective of modeling object relationships within an image. An image is firstly split into several flattened 2D patches. Then the embeddings of image patches are arranged in a sequence to represent the original image. ALBEF [Li *et al.*, 2021a] and SimVLM [Wang *et al.*, 2021b] feed patches to an ViT encoder to extract vision features, which lead the way to a full-transformer VL-PTM.

3 Modeling Vision-Language Interaction

After encoding images and texts into single-modal embeddings, the next step is to design an encoder to integrate information from both vision and language modalities. For example, to answer a question about an image, the model needs to combine the linguistic information from both questions and answers, then localize the corresponding region in the paired images, and lastly align linguistic meanings with visual clues. Based on the way of aggregating information from different modalities, we categorize the encoder into *fusion encoder*, *dual encoder* and the combination of both.

3.1 Fusion Encoder

The fusion encoder takes text embeddings and image features as input and designs several fusion approaches to model V-L interaction. After self-attention or cross-attention operation, the hidden states of the last layer will be treated as the fused representation of different modalities. There are mainly two types of fusion schemes for modeling the cross-modal interaction: single stream and dual stream.

Single-stream Architecture. The single-stream architecture assumes that the potential correlation and alignment between two modalities are simple, which can be learned by a single transformer encoder. Therefore, the text embeddings and image features are concatenated together, adding some special embeddings to indicate position and modalities, and fed into a transformer-based encoder.

Although different V-L tasks require different input formats (*e.g.*, ⟨*caption*, *image*⟩ for image captioning, ⟨*question*, *answer*, *image*⟩ for VQA), single-stream architecture can handle them in a unified framework due to the unordered representation nature of transformer attention. VisualBERT [Li *et al.*, 2019] and V-L BERT [Su *et al.*, 2019] utilize segment embedding to indicate input elements from different sources. Instead of simply using image-text pair, OSCAR [Li *et al.*, 2020c] adds object tags detected from the image and represents the image-text pair as a ⟨*word*, *tag*, *image*⟩ triple to help the fusion encoder better align different modalities. Since the single-stream architecture performs self-attention directly on two modalities, they may neglect intra-modality interaction. Thus some works propose to employ dual-stream architecture to model V-L interaction.

VL-PTM	Text encoder	Vision encoder	Fusion scheme	Pre-training tasks	Multimodal datasets for pre-training
Fusion Encoder					
VisualBERT [2019]	BERT	Faster R-CNN	Single stream	MLM+ITM	COCO
Uniter [2020]	BERT	Faster R-CNN	Single stream	MLM+ITM+WRA+MRFR+MRC	CC+COCO+VG+SBU
OSCAR [2020c]	BERT	Faster R-CNN	Single stream	MLM+ITM	CC+COCO+SBU+Flickr30k+VQA
InterBert [2020]	BERT	Faster R-CNN	Single stream	MLM+MRC+ITM	CC+COCO+SBU
ViLBERT [2019]	BERT	Faster R-CNN	Dual stream	MLM+MRC+ITM	CC
LXMERT [2019]	BERT	Faster R-CNN	Dual stream	MLM+ITM+MRC+MRFR+VQA	COCO+VG+VQA
VL-BERT [2019]	BERT	Faster R-CNN+ ResNet	Single stream	MLM+MRC	CC
Pixel-BERT [2020]	BERT	ResNet	Single stream	MLM+ITM	COCO+VG
Unified VLP [2020]	UniLM	Faster R-CNN	Single stream	MLM+seq2seq LM	CC
UNIMO [2020b]	BERT, RoBERTa	Faster R-CNN	Single stream	MLM+seq2seq LM+MRC+MRFR+CMCL	COCO+CC+VG+SBU
SOHO [2021]	BERT	ResNet + Visual Dictionary	Single stream	MLM+MVM+ITM	COCO+VG
VL-T5 [2021]	T5, BART	Faster R-CNN	Single stream	MLM+VQA+ITM+VG+GC	COCO+VG
XGPT [2021]	transformer	Faster R-CNN	Single stream	IC+MLM+DAE+MRFR	CC
Visual Parsing [2021]	BERT	Faster R-CNN + Swin transformer	Dual stream	MLM+ITM+MFR	COCO+VG
ALBEF [2021a]	BERT	ViT	Dual stream	MLM+ITM+CMCL	CC+COCO+VG+SBU
SimVLM [2021b]	ViT	ViT	Single stream	PrefixLM	C4+ALIGN
WenLan [2021]	RoBERTa	Faster R-CNN + EffcientNet	Dual stream	CMCL	RUC-CAS-WenLan
ViLT [2021]	ViT	Linear Projection	Single stream	MLM+ITM	CC+COCO+VG+SBU
Dual Encoder					
CLIP [2021]	GPT2	ViT, ResNet		CMCL	self-collected
ALIGN [2021]	BERT	EffcientNet		CMCL	self-collected
DeCLIP [2021b]	GPT2, BERT	ViT, ResNet, RegNetY-64GF		CMCL+MLM+CL	CC+self-collected
Fusion Encoder+ Dual Encoder					
VLMo [2021a]	BERT	ViT	Single stream	MLM+ITM+CMCL	CC+COCO+VG+SBU
FLAVA [2021]	ViT	ViT	Single stream	MMM+ITM+CMCL	CC+COCO+VG+SBU+RedCaps

Table 1: Glossary of Representative VL-PTMs. MLM/MVM: (Cross-Modal) Masked Language/Vision Modeling. ITM: Image-Text Matching. MRC: Masked Region Classification. MRFR: Masked Region Feature Regression. VG: Visual Grounding. GC: Grounded Captioning. WRA: Word-Region Alignment. CMCL: Cross-Modal Contrastive Learning. DAE: Denoising AutoEncoding

Dataset	Size	Reference
COCO	328,124	[Lin *et al.*, 2014]
VG	108,077	[Krishna *et al.*, 2017]
CC	3.1M	[Sharma *et al.*, 2018]
SBU	1M	[Ordonez *et al.*, 2011]
LAION	400M	https://laion.ai/laion-400-open-dataset/
RedCaps	12M	[Desai *et al.*, 2021]

Table 2: Widely Used Pre-training Datasets

Dual-stream Architecture. Different from self-attention operation in single-stream architectures, dual-stream architectures adopt a cross-attention mechanism to model V-L interaction, where the query vectors are from one modality while the key and value vectors are from the other. A cross-attention layer usually contains two unidirectional cross-attention sub-layers: one from language to vision and another from vision to language. They are responsible for exchanging information and aligning the semantics between the two modalities.

Dual-stream architectures assume that the intra-modal interaction and cross-modal interaction need to be separated to obtain better multimodal representations. ViLBERT [Lu *et al.*, 2019] utilizes two transformers to further model intra-modality interaction after the cross-modal module. LXMERT [Tan and Bansal, 2019] does not use extra transformers but appends a self-attention sub-layer after cross-attention sub-layer to further build internal connections. In the cross-modal sub-layers, the parameters of the attention module are shared between the two streams. In this case, the model learns a single function to contextualize image and text embeddings. ALBEF [Li *et al.*, 2021a] employs two separate transformers before cross-attention for images and texts, which better decouples intra-modal interaction and cross-modal interaction. This type of architecture helps to encode the input in a more comprehensive way. However, the extra feature encoder makes it parameter-inefficient.

3.2 Dual Encoder

Although the fusion encoder could model cross-modal interaction at different levels and achieves state-of-the-art results on many V-L tasks, it relies on a heavy transformer network to model V-L interaction. When performing cross-modal matching tasks like Image-Text Retrieval, the fusion encoder has to jointly encode all possible image text pairs, which leads to a quite slow inference speed.

In contrast, a dual encoder utilizes two single-modal encoders to encode two modalities separately. Then, it adopts straightforward methods such as shallow attention layer [Lee *et al.*, 2018] or dot product [Radford *et al.*, 2021; Jia *et al.*, 2021] to project the image embedding and text embedding to the same semantic space for computing V-L similarity scores. Without the complex cross-attention in transformer, the V-L interaction modeling strategy in the dual encoder is much more efficient. Thus, the feature vectors of images and text can be pre-computed and stored, which is more effective for retrieval tasks than the fusion encoder. Although dual encoder models like CLIP [Radford *et al.*, 2021] have shown surprising performance on image-text retrieval tasks, they fail in some hard V-L understanding tasks such as NLVR [Suhr *et al.*, 2018]. This is attributed to the shallow interaction between the two modalities.

3.3 Combination of Fusion Encoder and Dual Encoder

Based on the observation that fusion encoder performs better on V-L understanding tasks while dual encoder performs better on retrieval tasks, it is natural to combine the benefits of the two types of architectures. FLAVA [Singh *et al.*, 2021] first adopts a dual encoder to obtain single-modal representations. Then the single-modal embeddings are sent to a fusion encoder to obtain cross-modal representation. Apart from its model design, FLAVA conducts several unimodal pre-training tasks to improve the quality of single-modal representations. VLMo [Wang *et al.*, 2021a] introduces **M**ixture-**o**f-**M**odality-**E**xpert (**MoME**) and unifies a dual encoder and a fusion encoder into a single framework. After pre-training on images, texts, and image-text pairs by stage, VLMo can not only be fine-tuned on V-L understanding tasks, but also be applied to efficient image-text retrieval.

4 Cross-Modal Pre-training Tasks

According to Section 1, after the input images and texts are encoded as vectors and fully interacted, the next step is to design pre-training tasks for VL-PTMs. The designed pre-training tasks have a great impact on what VL-PTM can learn from the data. In this section, we introduce some widely-used pre-training tasks.

4.1 Cross-Modal Masked Language Modeling (MLM)

Cross-modal MLM is similar to MLM in the BERT model. In cross-modal MLM, VL-PTMs predict masked tokens not only based on unmasked tokens, but also by taking vision features into account. The dependency on vision modality differentiates cross-modal MLM from MLM in NLP. This task has been proven to be quite effective for pre-training VL-PTMs because it helps the model to align vision and text by considering the relationship between image and text. Formally, the objective can be defined as:

$$L_{\mathrm{MLM}} = -\mathbb{E}_{(W,V)\in\mathcal{D}} \log P_\theta\left(w_m | w_{\setminus m}, V\right), \tag{1}$$

where $w_m, w_{\setminus m}$ represent the masked tokens and unmasked tokens respectively, and $(W,V) \in \mathcal{D}$ represents a text W and an image V sampled from dataset $\mathcal{D}$.

Due to the distinction between cross-modal MLM and MLM in NLP, an effective masking strategy is necessary for cross-modal MLM. If the method is too simple, the model may be able to predict the masked tokens purely on the basis of their surrounding tokens. By masking some tokens that rely on the image, VL-PTMs will take into account the image features, thus aligning tokens and their corresponding objects in the image. ViLT [Kim *et al.*, 2021] utilizes the Whole Word Masking strategy, which prevents the model from predicting tokens solely by words co-occurrence; InterBERT [Lin *et al.*, 2020] masks several consecutive segments of text to make this pre-training task more difficult and improves its performance on downstream tasks further.

4.2 Cross-Modal Masked Region Prediction (MRP)

Similar to the cross-modal MLM, cross-modal MRP masks some RoI features with zeros and predicts them based on other image features. The model learns object relationships by inferring from other unmasked regions and learns V-L alignments by inferring from the text. There are two kinds of learning objectives: Masked Region Classification (MRC) and Masked Region Feature Regression (MRFR).

Masked Region Classification (MRC). MRC learns to predict the semantic class of each masked region. This task is motivated by the observation that VL-PTMs just learn high-level semantics of images instead of raw pixels from the language side. To predict the region class, the hidden state $\boldsymbol{h}_{v_i}$ of the masked region v_i from VL-PTMs is fed into a fully-connected (FC) layer, followed by a softmax function to form a predicted distribution on K object classes. The final objective is to minimize the cross-entropy (CE) loss between the predicted distribution and the detected object category, which can be formally defined as:

$$\mathcal{L}_{\mathrm{MRC}} = \mathbb{E}_{(W,V)\in\mathcal{D}} \sum_{i=1}^{l} \mathrm{CE}\big(\mathrm{softmax}(\mathrm{FC}(\boldsymbol{h}_{v_i})), c(v_i)\big), \tag{2}$$

where l is the amount of masked regions, and $c\left(v_i\right)$ represents the true label of the masked region, such as the object detection output or the (pre-defined) visual tokens.

Through the cross-modal attention in VL-PTMs, the hidden state $\boldsymbol{h}_{v_i}$ contains information from both vision and language modality, which makes it possible to predict visual semantic class from the text. As for the ground-truth label $c\left(v_i\right)$, an intuitive method is to regard object tags (with the highest confidence score) detected from object dector as the true labels[Tan and Bansal, 2019; Su *et al.*, 2019]. However, these labels are pseudo, which highly relies on the quality of the pre-trained object detectors, thus there are some variants of this task. ViLBERT [Lu *et al.*, 2019] and UNITER [Chen *et al.*, 2020] propose to consider the raw output of the detector as soft labels, which is a distribution of object classes. In this scenario, the objective becomes the KL-divergence between two distributions. SOHO [Huang *et al.*, 2021] first maps the CNN-based grid features to visual tokens, and then predicts the masked visual tokens based on their surrounding tokens.

Masked Region Feature Regression (MRFR). MRFR learns to regress the masked region feature $\boldsymbol{h}_{v_i}$ to its corresponding original region feature $\hat{E}_V(v_i)$, which can be written as:

$$\mathcal{L}_{\mathrm{MRFR}} = \mathbb{E}_{(W,V)\in\mathcal{D}} \sum_{i=1}^{l} \|\mathrm{FC}(\boldsymbol{h}_{v_i}) - \hat{E}_V(v_i)\|^2. \tag{3}$$

In this formula, the region feature $\hat{E}_V(v_i)$ of v_i is computed based on an unmasked image and l represents the amount of masked regions. MRFR requires the model to reconstruct the high-dimensional vectors instead of semantic class. When images are represented as a sequence of region features by faster R-CNN, simple masking strategies like random masking can give satisfying performances [Tan and Bansal, 2019; Chen *et al.*, 2020; Li *et al.*, 2020b]. However, random masking will not be so effective when images are converted into grid features or patch features, because the model will directly duplicate neighbor features as the predicted features. Visual parsing [Xue *et al.*, 2021] uses patch features to represent an image and assumes that visual tokens (region features) of high attention weights have similar semantics. It first randomly masks a visual token as a pivot token, and continues to mask k tokens with top-k attention weights. SOHO[Huang *et al.*, 2021] pre-trains a vision dictionary and masks all the features sharing the same visual index to avoid information leakage.

4.3 Image-Text Matching (ITM)

Cross-modal MLM and MRP help VL-PTMs learn the fine-grained correlation between images and texts, while ITM empowers VL-PTMs with the ability to align them at a coarse-grained level. ITM is similar to the Next Sentence Prediction (NSP) task in NLP, which requires the model to determine whether an image and a text are matched. Given an image-text pair, a score function s_θ measures the alignment probability between the image and text. The objective function is:

$$\begin{aligned}\mathcal{L}_{\text{ITM}} = &-\mathbb{E}_{(W,V)\in\mathcal{D}}\Big[y \log s_\theta\left(\boldsymbol{h}_{w_{\text{[CLS]}}}, \boldsymbol{h}_{v_{\text{[IMG]}}}\right) \\ &+(1-y)\log\left(1 - s_\theta\left(\boldsymbol{h}_{w_{\text{[CLS]}}}, \boldsymbol{h}_{v_{\text{[IMG]}}}\right)\right)\Big],\end{aligned} \tag{4}$$

where $y \in \{0, 1\}$ represents whether W and V are matched with each other or not, and $\boldsymbol{h}_{w_{\text{[CLS]}}}$ and $\boldsymbol{h}_{v_{\text{[IMG]}}}$ are the representations of $w_{\text{[CLS]}}$ and $v_{\text{[IMG]}}$, respectively.

The key to this task is how to represent an image-text pair in a single vector so that the score function s_θ could output a probability. UNITER [Chen *et al.*, 2020], Unicoder [Li *et al.*, 2020a] and SOHO [Huang *et al.*, 2021] concatenate the word sequence W and the object sequence V and take the final hidden state of the "[CLS]" token as the fused representation. By feeding it into a fully-connected layer layer, they can reduce the dimension to predict the alignment probability. While, ViLBERT [Lu *et al.*, 2019] adopts the representation of the "[IMG]" and "[CLS]" tokens to represent image and text respectively, and the fused representation is computed by element-wise product between them.

4.4 Cross-Modal Contrastive Learning (CMCL)

CMCL aims to learn universal vision and language representation under the same semantic space by pushing the embeddings of matched image-text pairs together while pushing the non-matched ones apart. The image-to-text contrastive loss can be formulated as:

$$\mathcal{L}_{\text{i2t}} = -\mathbb{E}_{(W,V)\in\mathcal{D}}\left[\log \frac{s_\theta\left(\boldsymbol{h}_{v_{\text{[IMG]}}}, \boldsymbol{h}_{w_{\text{[CLS]}}}\right)}{\sum_{W'} s_\theta\left(\boldsymbol{h}_{v_{\text{[IMG]}}}, \boldsymbol{h}_{w'_{\text{[CLS]}}}\right)}\right], \tag{5}$$

where W' belongs to the negative samples set of V, $\boldsymbol{h}_{w_{\text{[CLS]}}}$, $\boldsymbol{h}_{v_{\text{[IMG]}}}$ and $\boldsymbol{h}_{w'_{\text{[CLS]}}}$ are the representations of $w_{\text{[CLS]}}$, $v_{\text{[IMG]}}$ and $w'_{\text{[CLS]}}$, respectively, and s_θ is a score function to justify how similar a given image-text pair is. It is worth noting that the contrastive loss in CMCL is symmetrical, and the text-to-image contrastive loss is formulated similarly.

CLIP [Radford *et al.*, 2021] and ALIGN [Jia *et al.*, 2021] leverage large-scale image-text pairs to learn transferable visual representations and exhibit surprising zero-shot transfer to image classification tasks. ALBEF [Li *et al.*, 2021a] proposes to adopt momentum distillation to facilitate contrastive learning on massive noisy image-text pairs. WenLan [Huo *et al.*, 2021] employs MoCo [He *et al.*, 2020] and maintains a queue to store negative samples, which has been proven to be effective for contrastive learning. UNIMO [Li *et al.*, 2020b] incorporates a large volume of unimodal data during contrastive learning, allowing vision and language to enhance each other. It outperforms previous works on both multimodal and unimodal downstream tasks. [Yang *et al.*, 2022] claims that CMCL does not guarantee similar inputs from the same modality stay close by, so they introduce intra-modal contrastive learning to benefit representation learning.

5 Adapting VL-PTMs to Vision-Language Downstream Tasks

Pre-training tasks are able to help VL-PTMs to learn general visual and linguistic features, which can be applied to various downstream tasks. In this section, we introduce several common vision-language integration tasks and how VL-PTMs are adapted to them. Basically, we categorised these downstream tasks into cross-modal matching, cross-modal reasoning and vision and language generation.

5.1 Cross-Modal Matching

Cross-modal matching requires VL-PTMs to learn cross-modal correspondences between different modalities. We introduce two commonly-used cross-modal matching tasks: image text retrieval and visual referring expression.

Image Text Retrieval (ITR). ITR is a typical cross-modal matching task. This task requires retrieving an image that matches a given sentence most and vice versa. Early VL-PTMs that utilize a fusion-encoder architecture obtain a fused vector representation which is later projected to a similarity score [Lu *et al.*, 2019; Li *et al.*, 2019; Li *et al.*, 2020c]. Dual-encoder architectures such as CLIP [Radford *et al.*, 2021] and ALBEF [Li *et al.*, 2021a] are more efficient for ITR as they can pre-compute and store the embeddings of images and texts before retrieval.

Visual Referring Expression (VRE). VRE is an extension of the referring expression task in NLP. The goal is to localize the region in an image that corresponds to a specific textual description. Most VL-PTMs (*e.g.* [Lu *et al.*, 2019]) take the final representation of the extracted region proposals as input and learn a linear projection to predict a matching score, which is the same strategy as ITR during fine-tuning.

5.2 Cross-Modal Reasoning

Cross-modal reasoning requires VL-PTMs to perform language reasoning based on visual information. Ignoring any modality gives poor performance. Here we present two commonly-used cross-modal reasoning tasks.

Visual Question Answering (VQA). VQA is a widely-used cross-modal reasoning task. Different from text-based QA, VQA requires answering questions about images. Most researchers consider VQA as a classification task and require the model to select a correct answer from an answer pool. VL-PTMs with a fusion-encoder architecture usually map the final cross-modal representation (usually corresponds to the input [CLS] token) to the distribution of answer labels. However, VL-PTMs with a dual-encoder architecture are not so effective for VQA tasks because the interaction between the two modalities is too shallow to conduct cross-modal reasoning. There are also some works modeling VQA as a generation task [Cho *et al.*, 2021; Wang *et al.*, 2021b], which can generalize better to real-world open-ended scenarios.

Natural Language for Visual Reasoning (NLVR). NLVR provides an image pair and a textual statement as input and requires the model to decide whether the statement is true about the image pair, thus can be considered as a binary classification task. Most VL-PTMs first encode the given two image-text pairs separately, then a classifier is trained over the concatenation of the two embeddings to make a prediction [Tan and Bansal, 2019; Chen *et al.*, 2020].

Visual Commonsense Reasoning (VCR). VCR is considered to be another kind of VQA task. The main difference between VCR and VQA is that VCR's questions pay more attention to visual common sense. Different from VQA, the VCR task can be decomposed into two multi-choice subtasks: question answering (Q $\rightarrow$ A) and answer justification (Q + A $\rightarrow$ R). Most VL-PTMs utilize the same approach in VQA to solve these two sub-tasks. For the question answering sub-task, the procedure is the same as VQA. For the answer justification sub-task, the concatenations of question and answer are treated as the new questions and the rationales become the options. A linear layer is trained to predict a score for each possible option [Lu *et al.*, 2019].

5.3 Vision and Language Generation

Based on the source modal and target modal, the generation task can be divided into text-to-image generation and image-to-text generation (multimodal text generation).

Text-to-Image Generation. Text-to-Image generation is the task of generating a corresponding image from a descriptive text. X-LXMERT [Cho *et al.*, 2020] first converts continuous visual representations to discrete cluster centroids and then ask the model to predict the cluster ids of masked regions. DALL-E [Ramesh *et al.*, 2021] trains a codebook to tokenize images and formulates text-to-image generation task as an autoregressive generative task. It achieves new state-of-the-art results on MS-COCO [Lin *et al.*, 2014] in zero-shot setting.

Multimodal Text Generation. Multimodal text generation can be regarded as a special type of conditional text generation, where the condition includes not only texts but also images. Usually, a decoder is needed for the generation process. Image captioning is a typical image-to-text generation task that requires the model to generate a description of an image. XGPT [Xia *et al.*, 2021] and VL-T5 [Cho *et al.*, 2021] encode the images first and then employ a decoder to generate the captions autoregressively. Multimodal machine translation is another generation task that aims to introduce images to improve translation quality. VL-T5 [Cho *et al.*, 2021] tackles this task using the same strategy as in image captioning.

As for the connection between VL-PTMs architecture and downstream tasks, fusion encoder is more suitable than dual encoder on cross-modal reasoning tasks for its powerful ability to model interaction. The dual encoder is more suitable for cross-modal retrieval tasks since it keeps a similar performance as fusion encoder does while being more efficient.

6 Conclusion and Future Directions

In this paper, we present an overview of VL-PTMs. We review the commonly-used architectures and discuss their advantages and disadvantages. We also introduce several mainstream approaches to pre-training a VL-PTM and adapt it to downstream tasks. Though VL-PTMs have made significant progress on V-L tasks compared to traditional methods, there are still several challenges that could be the directions of future research.

Unified Model Architecture. Transformer-based models have shown surprising performance on NLP, CV, and multimodal tasks. The success of transformer-based models in various domains indicates the possibility of using a single transformer model to learn a representation of different modalities and building a general agent to handle tasks in different domains. UNIMO [Li *et al.*, 2020b] and FLAVA [Singh *et al.*, 2021] make some inspiring attempts in this direction, but their performance on some tasks is much worse than the task-specific baselines. Data2vec [Baevski *et al.*, 2022] adopts self-supervised learning to unify vision, speech , and language. This model achieves competitive results to predominant methods on several tasks, which paves the way for a powerful unified model.

Model Compression and Acceleration. Despite the great success achieved by VL-PTMs in various fields, it is difficult to deploy such a huge model in real-life scenarios, thus leading to a direction of VL-PTM compression and acceleration. Knowledge distillation has been used to compress VL-PTM [Fang *et al.*, 2021], but some other traditional compression methods such as quantization and pruning for VL-PTMs are yet to be explored. As for model acceleration, Li *et al.* [2021b] construct a data-efficient paradigm for V-L pre-training. Despite all these achievements, only a few efforts focus on improving VL-PTM's inference speed.

Advanced Pre-training Methods. Though the current pre-training method seems quite effective, the potential of advanced pre-training methods is yet to be explored. Using adversarial samples to enhance pre-training has been shown to be effective [Gan *et al.*, 2020], which helps VL-PTMs to overcome the overfitting issue. Stage-wise pre-training [Wang *et al.*, 2021a] has been proposed for better single-modal representation. With that ahead, the potential of pre-training methods are not fully developed, which is worth further studies.

Reaching the Limit of VL-PTMs. Nowadays, with the success of large-scale PLMs in NLP, many researchers have also tried to build a deeper model or use a larger dataset for V-L pre-training. ALIGN [Jia *et al.*, 2021] has a number of 675.4 million parameters and collects a huge dataset consisting of 1.8 billion image-text pairs for pre-training. It achieves state-of-the-art results on almost all downstream tasks. Wenlan [Fei *et al.*, 2021] expands the dataset to 650 million image-text pairs and shows astonishing performance on both vision-language understanding and generation tasks. In the future, VL-PTMs will need more high-quality data and more parameters to reach a higher recognition level.

References

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015.

[Arevalo *et al.*, 2017] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

[Baevski *et al.*, 2022] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. *URL https://ai. facebook. com/research/data2veca-general-framework-for-self-supervi sed-learning-in-speech-vision-and-language/. Accessed*, 2022.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, pages 1877–1901, 2020.

[Chen *et al.*, 2020] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120. Springer, 2020.

[Cho *et al.*, 2020] Jaemin Cho, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi. X-lxmert: Paint, caption and answer questions with multimodal transformers. *arXiv preprint arXiv:2009.11278*, 2020.

[Cho *et al.*, 2021] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. *arXiv preprint arXiv:2102.02779*, 2021.

[Desai *et al.*, 2021] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Fang *et al.*, 2021] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Compressing visual-linguistic model via knowledge distillation. In *CVPR*, pages 1428–1438, 2021.

[Fei *et al.*, 2021] Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. Wenlan 2.0: Make ai imagine via a multimodal foundation model. *arXiv preprint arXiv:2110.14378*, 2021.

[Gan *et al.*, 2020] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *NeurIPS*, pages 6616–6628, 2020.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.

[Huang *et al.*, 2020] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

[Huang *et al.*, 2021] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *CVPR*, pages 12976–12985, 2021.

[Huo *et al.*, 2021] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*, 2021.

[Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

[Kim *et al.*, 2021] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*, 2021.

[Krishna *et al.*, 2017] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, pages 32–73, 2017.

[Lee *et al.*, 2018] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018.

[Li *et al.*, 2019] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[Li *et al.*, 2020a] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020.

[Li *et al.*, 2020b] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

[Li *et al.*, 2020c] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, pages 121–137. Springer, 2020.

[Li *et al.*, 2021a] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 2021.

[Li *et al.*, 2021b] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[Lin *et al.*, 2020] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. Interbert: Vision-and-language interaction for multi-modal pretraining. *arXiv preprint arXiv:2003.13198*, 2020.

[Lu *et al.*, 2019] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

[Mogadala *et al.*, 2021] Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *JAIR*, pages 1183–1317, 2021.

[Ordonez *et al.*, 2011] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *NeurIPS*, 2011.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[Ramesh *et al.*, 2021] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831, 2021.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015.

[Ruan and Jin, 2022] Ludan Ruan and Qin Jin. Survey: Transformer based video-language pre-training. *AI Open*, 2022.

[Sharma *et al.*, 2018] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Singh *et al.*, 2021] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. *arXiv preprint arXiv:2112.04482*, 2021.

[Su *et al.*, 2019] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[Suhr *et al.*, 2018] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

[Tan and Bansal, 2019] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[Wang *et al.*, 2021a] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021.

[Wang *et al.*, 2021b] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

[Xia *et al.*, 2021] Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, and Ming Zhou. Xgpt: Cross-modal generative pre-training for image captioning. In *NLPCC*, pages 786–797, 2021.

[Xue *et al.*, 2021] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training. *NeurIPS*, 2021.

[Yang *et al.*, 2022] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. *arXiv preprint arXiv:2202.10401*, 2022.

[Zellers *et al.*, 2019] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731, 2019.

[Zhou *et al.*, 2020] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, pages 13041–13049, 2020.