Intro to Vision-Language Pretrain (VLP)

also called Multi-modal Learning

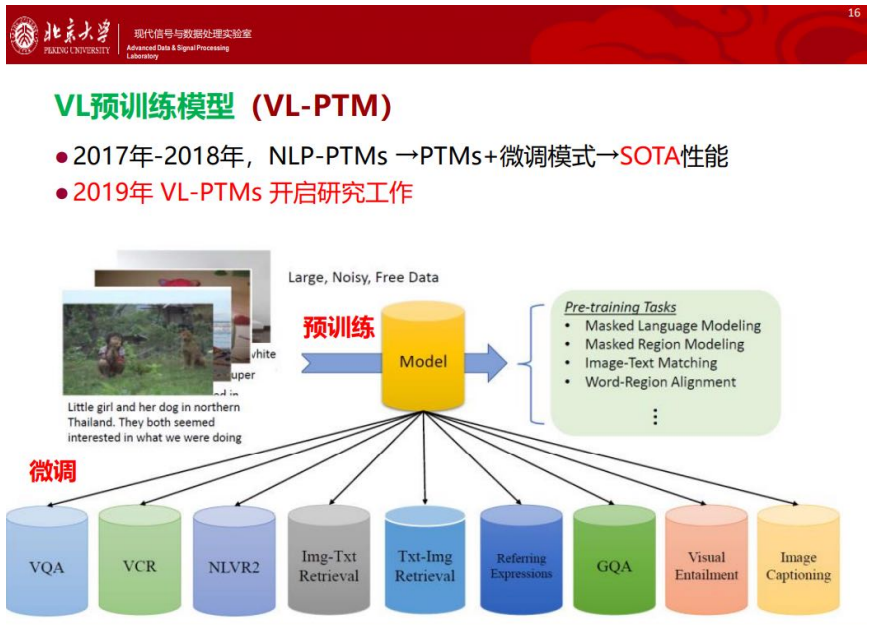
Breezedeus, 2022.07 →

Contents

Vision-Language (VL) 历史

VL学习的发展分为三个阶段 [1]

1. 2014-2018年：专门的模型被设计用于不同的任务
2. 2019-2021年（**BERT**诞生）：通过使用**高质量标签的VL数据集**进行预训练，模型能够学习视觉和语言的联合表征
   - ViLBERT
   - UNITER

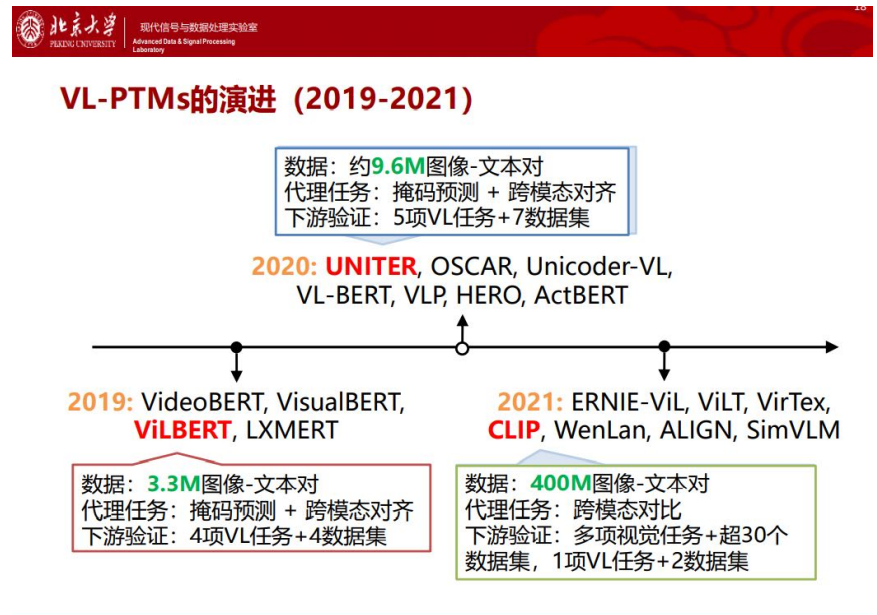


1. 北大邹月娴：视觉-语言预训练模型演进及应用 ↩

Vision-Language (VL) 历史

VL学习的发展分为三个阶段 [1]

3. 2021年-：随着CLIP的出现，期望在**更大的弱标签数据集**上预训练VL模型，并通过VL预训练获得性能强大的基于零样本或少样本的视觉模型
   - CLIP



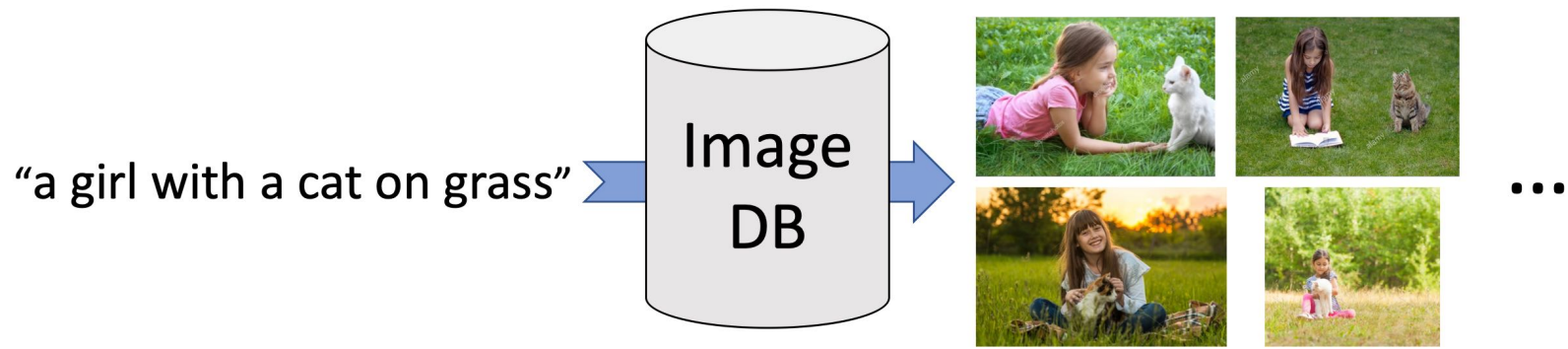
1. 北大邹月娴：视觉-语言预训练模型演进及应用

多模态任务

下游任务（Downstream Tasks）[1]

一些常见任务：

- Text-Image / Image-Text Retrieval



1. VQA2VLN2021 part1, part2 ↩️

多模态任务

下游任务（Downstream Tasks）[1]

一些常见任务：

- Visual Question Answering

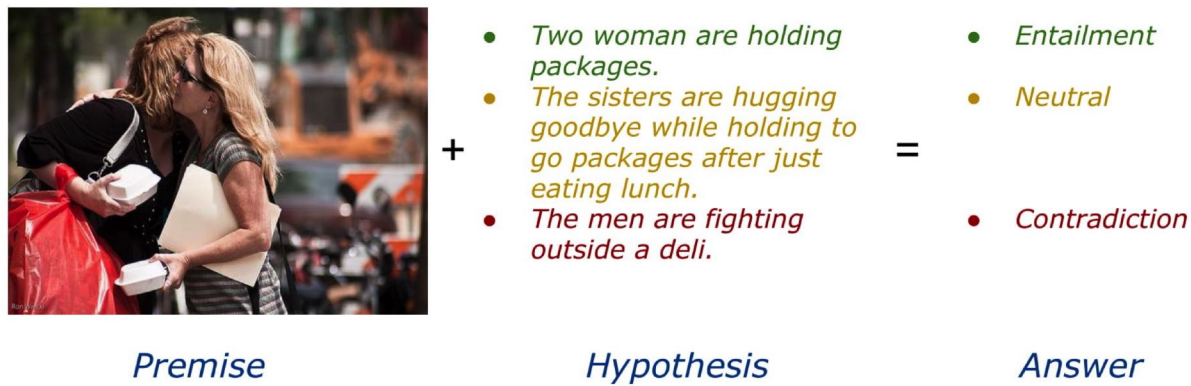
1. VQA2VLN2021 part1, part2 ↩️

多模态任务

下游任务（Downstream Tasks）[1]

一些常见任务：

- Visual Entailment



---

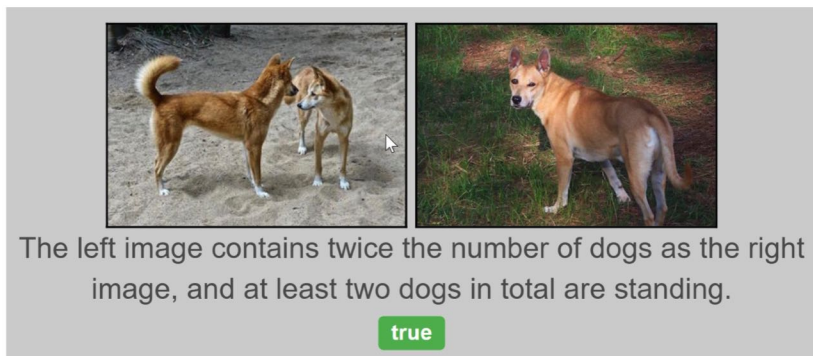
1. VQA2VLN2021 part1, part2 ↩

多模态任务

下游任务（Downstream Tasks）[1]

一些常见任务：

- Natural Language for Visual Reasoning



---

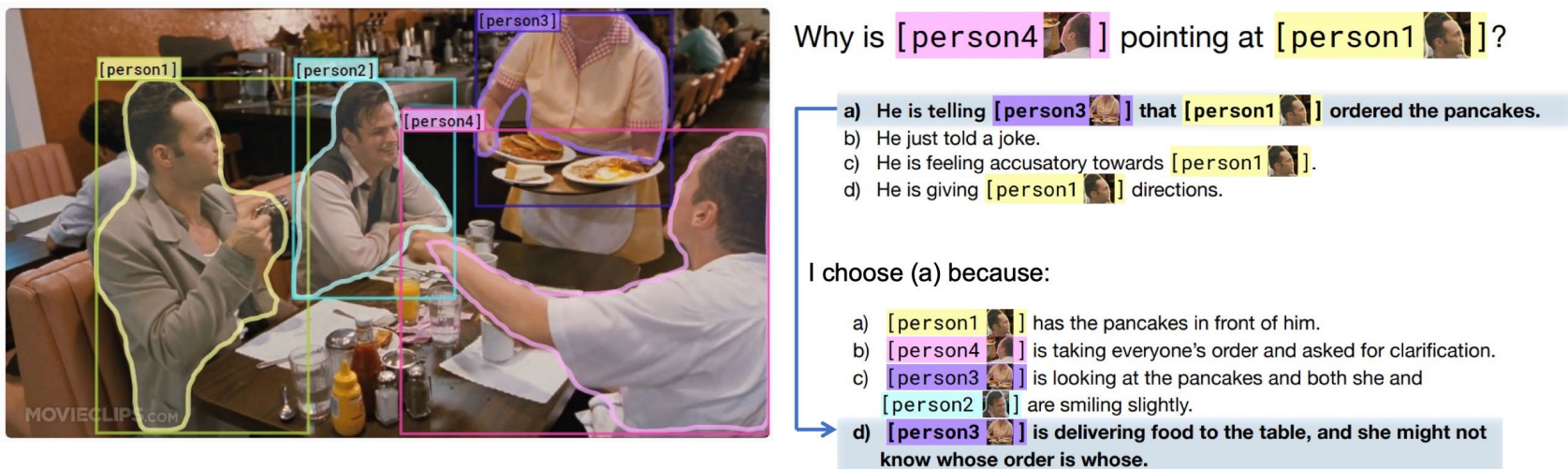
1. VQA2VLN2021 part1, part2 ↩

多模态任务

下游任务（Downstream Tasks）[1]

一些常见任务：

- Visual Commonsense Reasoning



---

1. VQA2VLN2021 part1, part2 ↩

多模态任务

下游任务（Downstream Tasks）[1]

一些常见任务：

- Referring Expression Comprehension

woman washing dishes

---

1. VQA2VLN2021 part1, part2 ↩

多模态任务

下游任务（Downstream Tasks）

More:

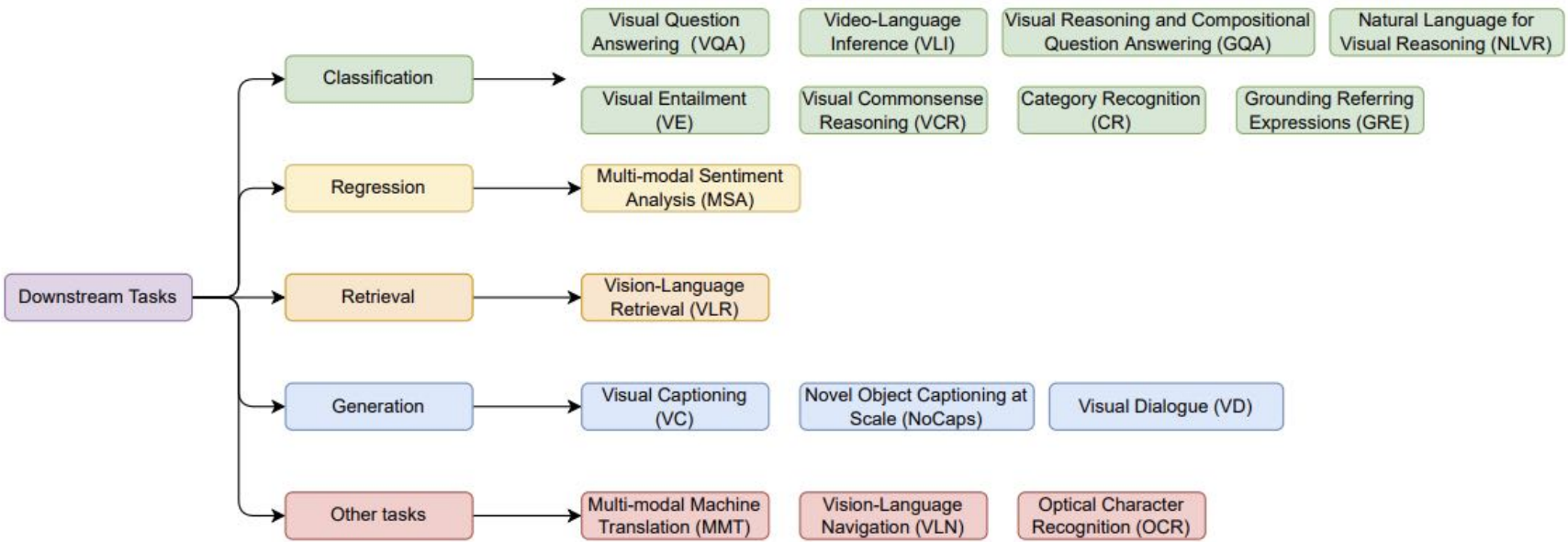


Fig. 2 Illustration of downstream tasks in VLP.

模型总结

Feature Extraction

Text Features

- LSTM
- BERT

Image Features

- CNN: feature maps
- RoI: detected objects
- Patch: `16 x 16` patches
- Pixel
- 也可以把目标检测出的标签（Object Tags）作为额外的信息拼接到输入

Video Features

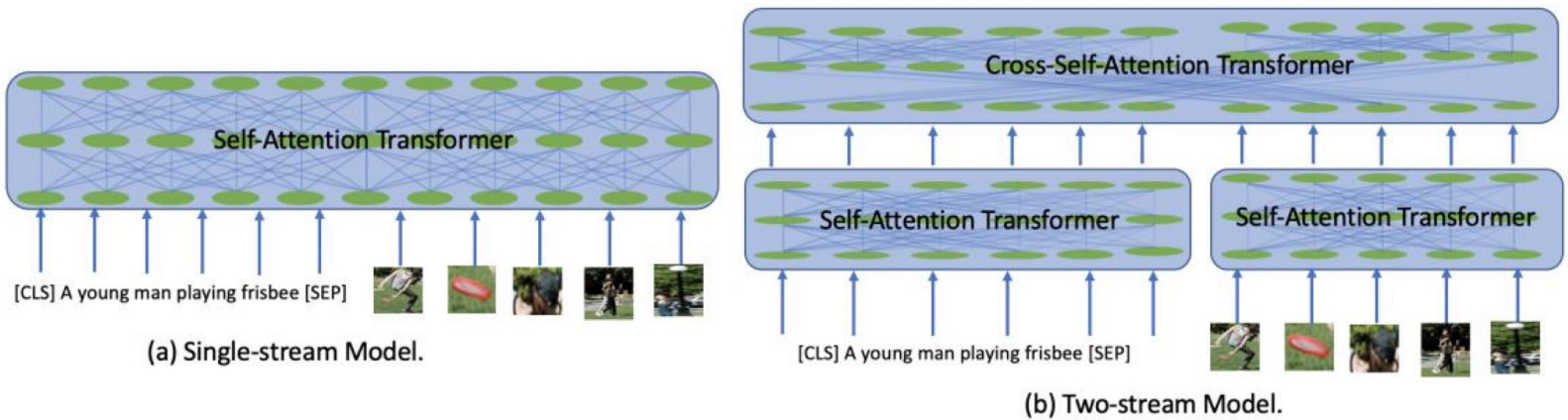
- 先抽取帧图片特征，再把时序图片特征变成定长特征
  - temporal maxpooling
  - ...

模型总结

Model Architecture

两大类别：

1. 单流模型（Single-Stream）
2. 双流模型（Dual-Stream）



- **Encoder-only** vs. **Encoder-decoder**

模型总结

预训练任务

Masked Language Modeling (MLM)

$$\mathcal{L}_{\mathrm{MLM}}(\theta) = -E_{(\mathbf{w},\mathbf{v})\sim D} \log P_\theta \left(\mathbf{w}_\mathbf{m} \mid \mathbf{w}_{\backslash \mathbf{m}}, \mathbf{v}\right)$$

Masked Language Modeling (MLM)

术语：

- Image Regions: $\mathbf{v} = \{v_1, \ldots, v_K\}$
- Sentence Tokens: $\mathbf{w} = \{w_1, \ldots, w_T\}$
- Masking Indices: $\mathbf{m} \in \mathbb{N}^M$

模型总结

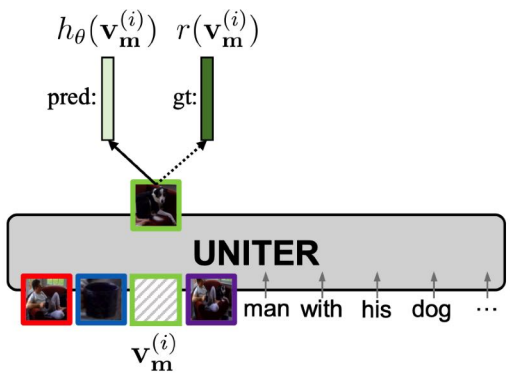
预训练任务

Masked Region Modeling (MRM)

模仿MLM，只不过是对图片 token 进行随机mask

- **Masked Region Feature Regression (MRFR)**

$$f_\theta\left(\mathbf{v_m} \mid \mathbf{v}_{\backslash \mathbf{m}}, \mathbf{w}\right)=\sum_{i=1}^{M}\left\|h_\theta\left(\mathbf{v_m^{(i)}}\right)-r\left(\mathbf{v_m^{(i)}}\right)\right\|_2^2$$



术语：

- Image Regions: $\mathbf{v}=\{v_1,\ldots,v_K\}$
- Sentence Tokens: $\mathbf{w}=\{w_1,\ldots,w_T\}$
- Masking Indices: $\mathbf{m}\in\mathbb{N}^M$

模型总结

预训练任务

Masked Region Modeling (MRM)

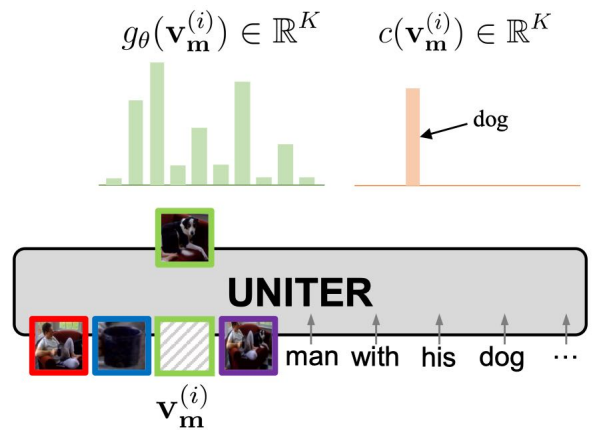
模仿MLM，只不过是对图片 token 进行随机mask

- **Masked Region Classification (MRC)**
  - hard label: predicts the **semantic class** for the corresponding image region

$$f_\theta\left(\mathbf{v}_\mathbf{m} \mid \mathbf{v}_{\backslash \mathbf{m}}, \mathbf{w}\right)=\sum_{i=1}^{M} \mathrm{CE}\left(c\left(\mathbf{v}_\mathbf{m}^{(i)}\right), g_\theta\left(\mathbf{v}_\mathbf{m}^{(i)}\right)\right)$$

术语：

- Image Regions: $\mathbf{v}=\{v_1, \ldots, v_K\}$
- Sentence Tokens: $\mathbf{w}=\{w_1, \ldots, w_T\}$
- Masking Indices: $\mathbf{m} \in \mathbb{N}^M$



模型总结

预训练任务

Masked Region Modeling (MRM)

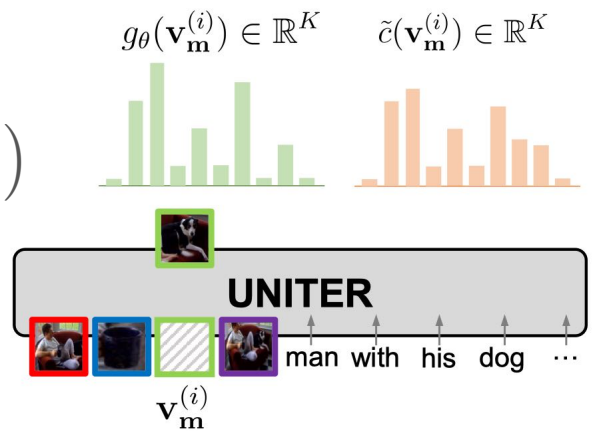
模仿MLM，只不过是对图片 token 进行随机mask

- **Masked Region Classification – KL Divergence (MRC-kl)**
  - soft label: predicts a **distribution over semantic classes** for the corresponding image region

$$f_\theta\left(\mathbf{v}_\mathbf{m} \mid \mathbf{v}_{\backslash \mathbf{m}}, \mathbf{w}\right)=\sum_{i=1}^{M} D_{KL}\left(\tilde{c}\left(\mathbf{v}_\mathbf{m}^{(i)}\right) \| g_\theta\left(\mathbf{v}_\mathbf{m}^{(i)}\right)\right)$$

术语：

- Image Regions: $\mathbf{v}=\{v_1, \ldots, v_K\}$
- Sentence Tokens: $\mathbf{w}=\{w_1, \ldots, w_T\}$
- Masking Indices: $\mathbf{m} \in \mathbb{N}^M$



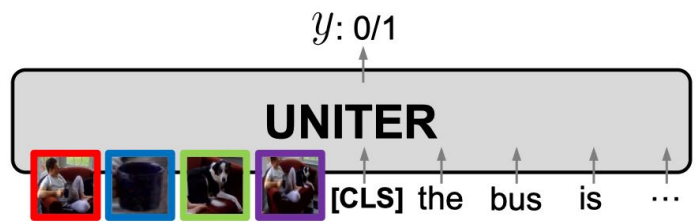
模型总结

预训练任务

Image-Text Matching (ITM)

- 判断当前 `text-image` pair是不是匹配：将图片的`[IMG]` token和文本的`[CLS]` token 进行匹配

$$\mathcal{L}_{\mathrm{ITM}}(\theta) = -E_{(\mathbf{w},\mathbf{v})\sim D}\left[y \log s_\theta(\mathbf{w},\mathbf{v}) + (1-y)\log\left(1 - s_\theta(\mathbf{w},\mathbf{v})\right)\right])$$



模型总结

预训练任务

Vision-Language Contrastive Learning (VLC)

- Vision-Language paired **InfoNCE**

$$p_m^{v2t}(I) = \frac{\exp\left(s\left(I, T_m\right)/\tau\right)}{\sum_{m=1}^{M}\exp\left(s\left(I, T_m\right)/\tau\right)}$$

$$p_m^{t2v}(T) = \frac{\exp\left(s\left(T, I_m\right)/\tau\right)}{\sum_{m=1}^{M}\exp\left(s\left(T, I_m\right)/\tau\right)}$$

$$\mathcal{L}_{\mathrm{VLC}} = \frac{1}{2}\mathrm{E}_{(I,T)\sim D}\left[\mathrm{CE}\left(y^{v2t}, p^{v2t}(I)\right) + \mathrm{CE}\left(y^{t2v}, p^{t2v}(T)\right)\right]$$

模型总结

预训练任务

Word-Region Alignment (WRA)

- **Loss of WRA:**

$$\mathcal{L}_{\mathrm{WRA}}=\min_{\mathrm{T}\in II(\mathbf{a},\mathbf{b})}\sum_{i=1}^{T}\sum_{j=1}^{K}\mathrm{T}_{ij}\cdot c\left(\mathbf{w}_i,\mathbf{v}_j\right)$$

where $c\left(\mathbf{w}_i,\mathbf{v}_j\right)$ is the cost function evaluating the distance between $\mathbf{w}_i$ and $\mathbf{v}_j$, $\mathrm{T}\in II(\mathbf{a},\mathbf{b})=\left\{\mathrm{T}\in\mathbb{R}^{T\times K}\mid \mathbf{T1}_m=\mathbf{a},\ \mathrm{T}^{\top}\mathbf{1}_n=\mathbf{b}\right\}$, $\mathbf{a}$ and $\mathbf{b}$ Dirac function coefficients centered on $\mathbf{w}_i$ and $\mathbf{v}_j$

- Optimal Transport Problem

任务： 存储在不同地区的 N 个仓库（位置 $\{x_i\}_{i=1}^N$ ，每个仓库有物资（ $\{G_n\}_{n=1}^N$ ），需要将这些物资分发到 M 个不同的地方 (位置 $\{y_i\}_{i=1}^M$ ，货物数量需求 $\{h_j\}_{j=1}^M$ ）。各个仓库及分发地点之间距离为 $\{c(x_i,y_j)\}_{i,j=1}^{M,N}$ 。

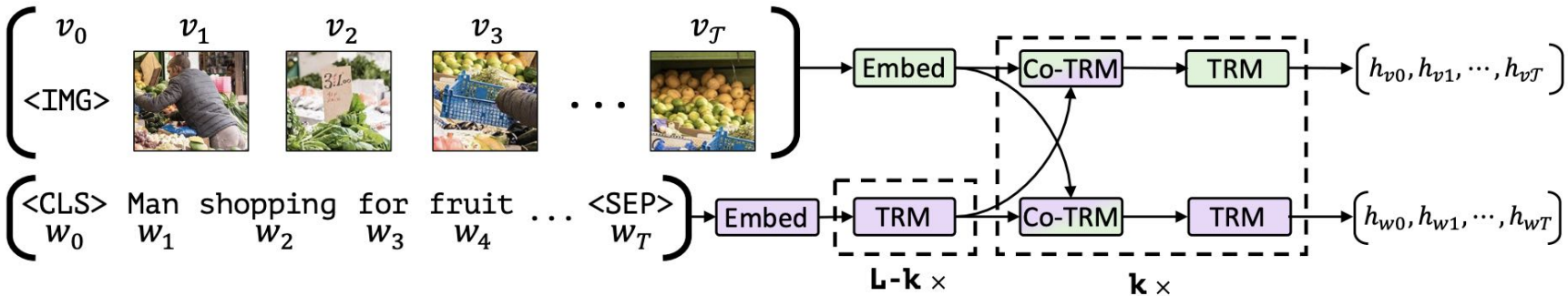
目标： 怎样的运输策略才能最有效的完成物资的分配。为简化问题，假设只考虑距离及运输货物数量为主要因素，且运 k 个一定比运1个成本更高，可得目标优化问题为

$$L=\arg\min_{\Gamma}\sum_{i,j=1}^{M,N}\Gamma_{ij}c\left(x_i,y_j\right)\tag{1}$$

其中 Γ 为运输矩阵，矩阵的元素 Γ_{ij} 表示为从仓库 i 发往地点 j 的物资数。且满足 $\sum_i\Gamma_{ij}=G_i$ ， $\sum_j\Gamma_{ij}=h_i$

代表性模型

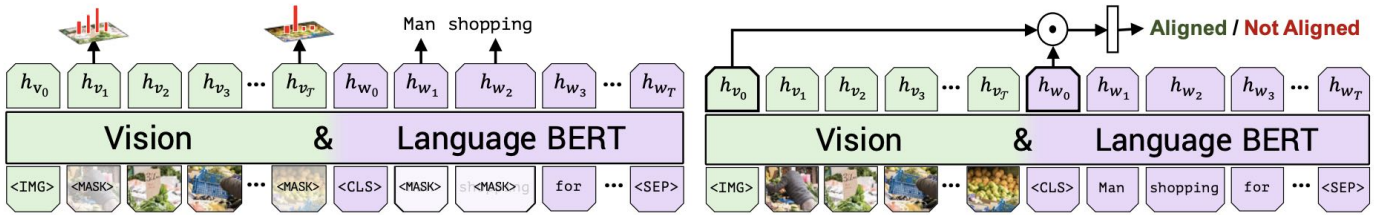
ViLBERT



- dual-stream
- Image Features: detected objects (RoI) + 5-d spatial location vector
  - 5-d vector: region position (normalized top-left and bottom-right coordinates) and the fraction of image area covered

代表性模型

ViLBERT



(a) Masked multi-modal learning (b) Multi-modal alignment prediction

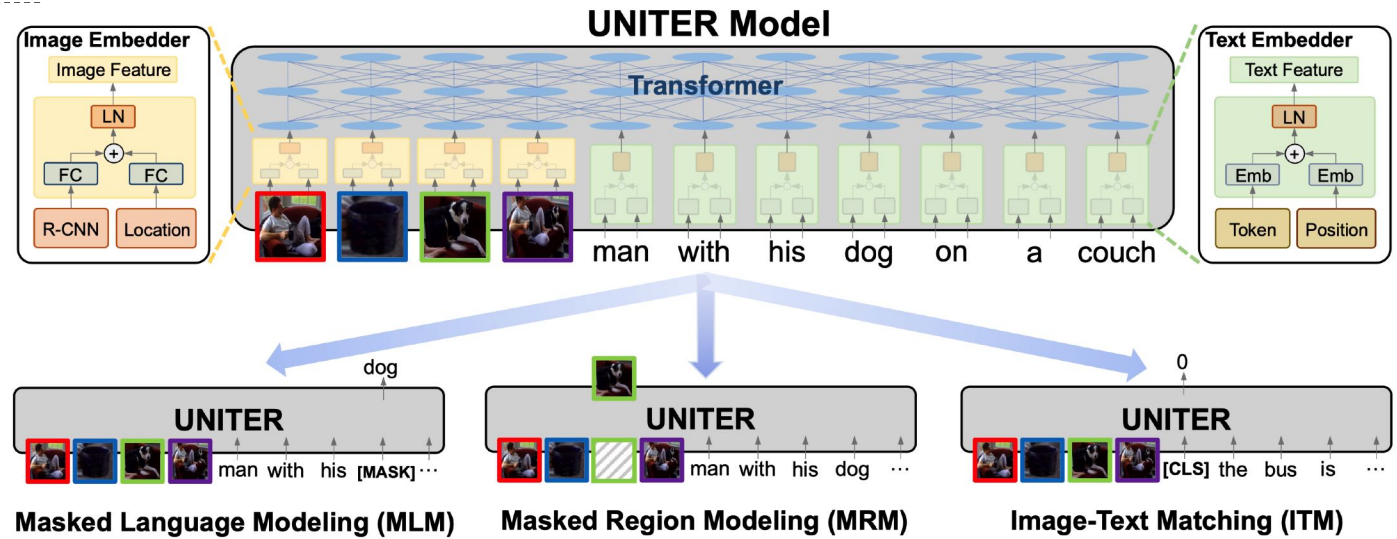
Figure 3: We train ViLBERT on the Conceptual Captions [24] dataset under two training tasks to learn visual grounding. In masked multi-modal learning, the model must reconstruct image region categories or words for masked inputs given the observed inputs. In multi-modal alignment prediction, the model must predict whether or not the caption describes the image content.

Training Tasks:

- **MRM**: the model predicts a **distribution over semantic classes** for the corresponding image region; KL divergence
- **ITM**

代表性模型

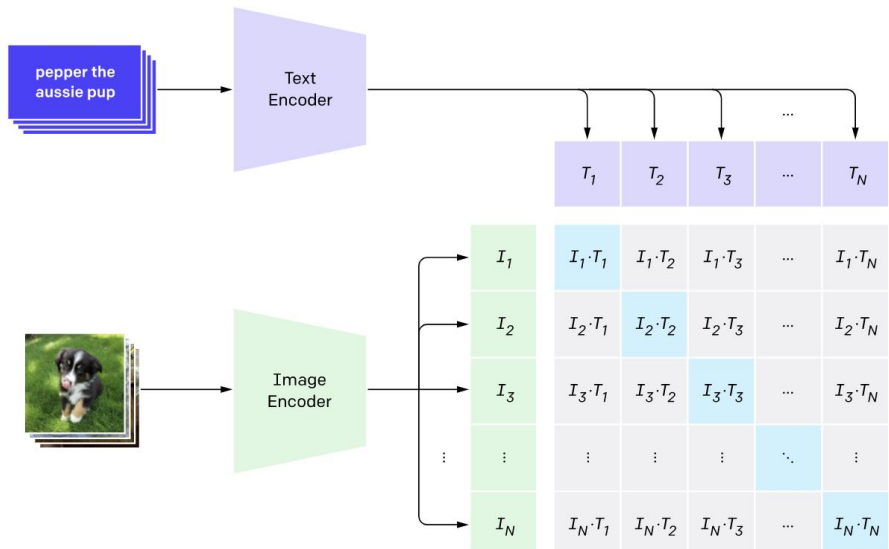
UNITER



- single-stream
- Image Input: detected objects (RoI)
- Text Input: tokens

代表性模型

CLIP

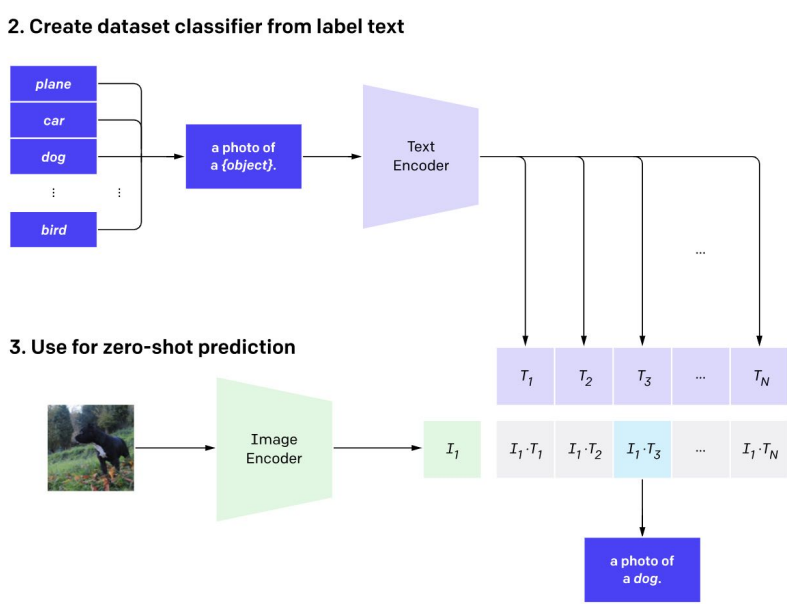


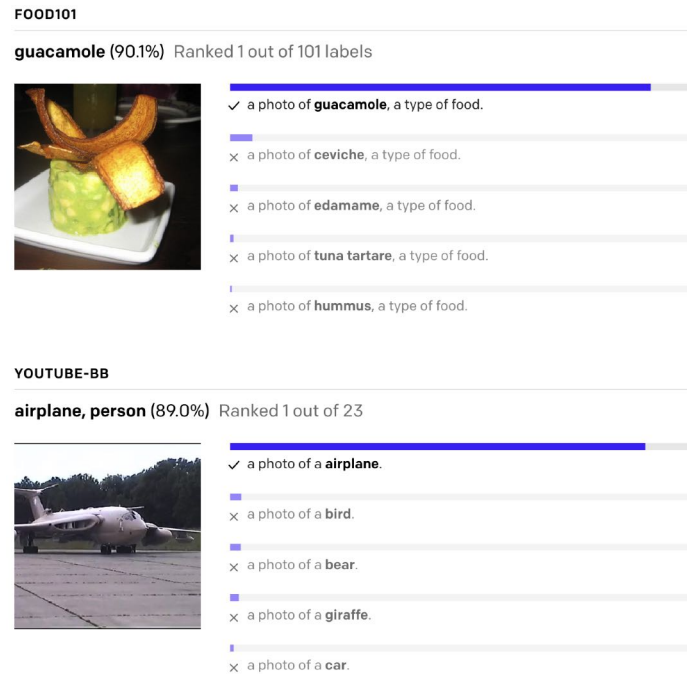
- dual-stream
- train on **large weakly** paired text-image examples
- VLC: InfoNCE Loss
- Image Encoder: ResNet or Vision Transformer (ViT)；ViT 效果更好，所以论文中的实验来自ViT encoder

代表性模型

CLIP 的多才多艺

- zero-shot classifcation

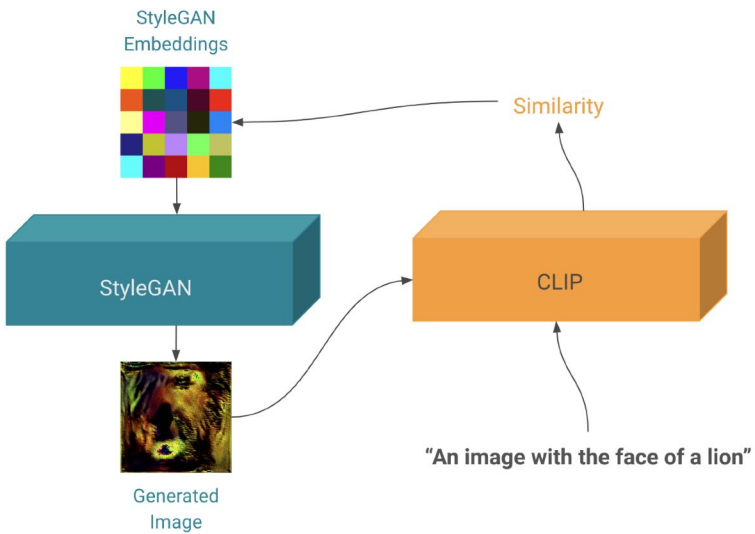




代表性模型

CLIP 的多才多艺

- Image Generation: CLIP + StyleGAN

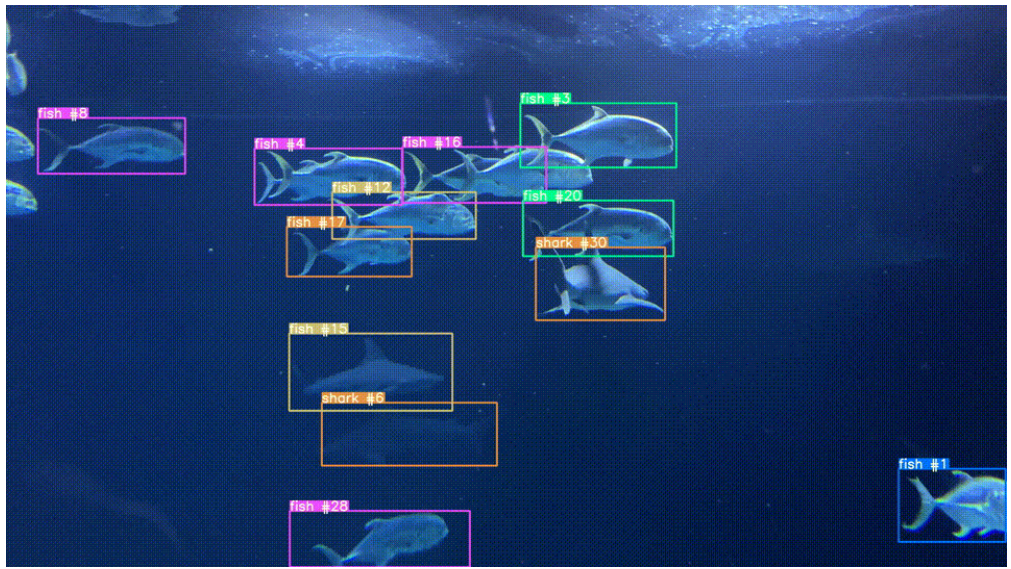


"An image with the face of Elon Musk with blonde hair"

代表性模型

CLIP 的多才多艺

- Object Tracking



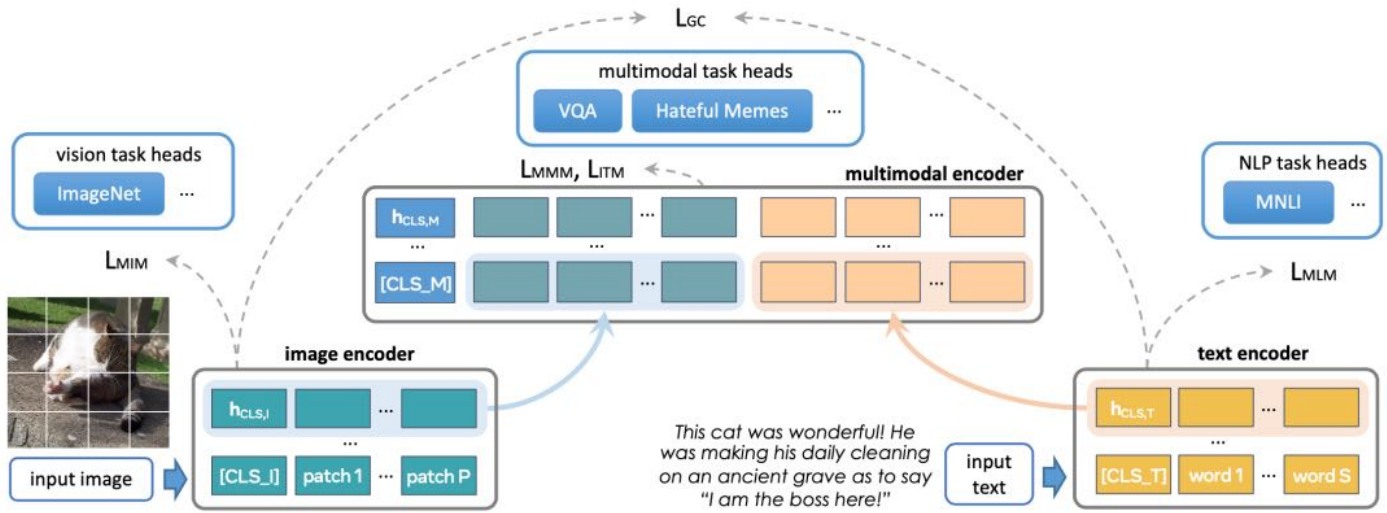
代表性模型

CLIP 的多才多艺

- Image Captioning: CLIP + GPT2
- Image Search / Similarity / Ranking
- ...

代表性模型

FLAVA

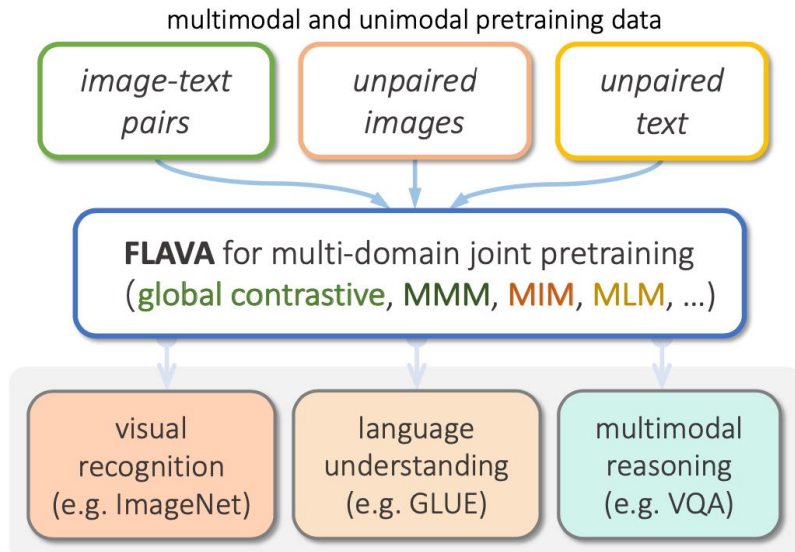


- single-stream
- train data: 单独文本、单独图片、匹配的图片-文本
- **Image Encoder**: 和 ViT 一样，使用 patches

代表性模型

FLAVA

- MIM: 利用dVAE将图片转化为类似词向量的token；再参照BEiT，对masked隐状态进行分类
- 使用了三种多模态预训练任务：
  - VLC
  - ITM
  - Masked Multimodal Modeling (MMM)：在上层的多模态模块进行，会同时随机mask图像的`patches`和文本的`tokens`；修复任务和MLM、MIM相同



代表性模型

悟空

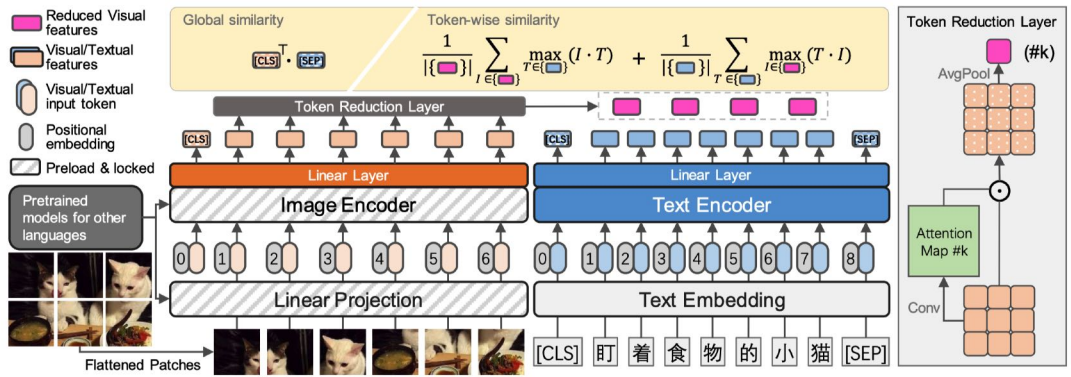


Figure 1: Overviews of our released models. Our Chinese pre-trained models consist of an image encoder and a text encoder with visual tokens and textual tokens as inputs. We have three variations of pretrained models: global similarity (**CLIP**-style); token-wise similarity (**FILIP**-style) and token-wise similarity with token reduction layer (**Wukong**-style).

- dual-stream
- Image input: patch
- 融合时有3种相似度计算方式
  - Token Reduction Layer：降低视觉token的数量，从 `16*16+1` 降到 `12` 或者 `24` （`12` 或者 `24` 个带权重的 **AvgPooling** ）

代表性模型

文澜 BriVL

- 双流模型
- CLIP + MoCo, 提升训练效率
- Image encoder: RoI

R2D2

- 来自奇虎360
- 算是单流模型，不适合检索，适合排序
- 使用 VLC + ITM 做跨模态训练，MLM for text
- 使用了**蒸馏技术**

CogView

- Text-to-Image generation
- GPT 结构
- **Image tokenizer**: a discrete Auto-Encoder, which is similar to the stage 1 of VQ-VAE or d-VAE

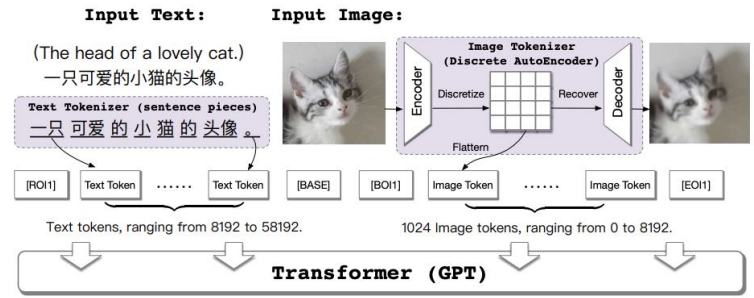


Figure 3: The framework of CogView. `[ROI1]`, `[BASE1]`, etc., are seperator tokens.

- CogView2
  - Hugging Face Demo

代表性模型

More (From VLP)

Table 2 The summary of mainstream image-text VLP models. The number of downstream tasks determines whether the model is generic or domain-specific VLP. FE: Feature Extraction. PT: Pre-training. Emb: Embedding. SC in Datatsets column: self-constructed or self-collected. MTL in Datatsets column: all datasets for multi-task learning in corresponding work. See other abbreviations in Datatsets column in Table 1.

Model	Domain	Vision FE	Language FE	Multimodal Fusion	Decoder	PT Objectives	PT Datasets	Downstream Tasks
VisualBERT [44]	Image	OD-RFs	Emb	Single-stream	No	MLM+VLM	COCO	GRE+NLVR+VCR+VQA
ViLBERT [45]	Image	OD-RFs	Emb	Dual-stream	No	MLM+VLM+MVM	COCO+VG	VLR+NLVR+VE+VQA
LXMERT [46]	Image	OD-RFs+Xformer	Xformer	Dual-stream	No	MLM+VLM+MVM+VQA	COCO+VG+VQA+GQA+VGQA	GQA+NLVR+VQA
B2T2 [47]	Image	CNN-GFs	Emb	Single-stream	No	MLM+VLM	CC3M	VCR
Unicoder-VL [48]	Image	OD-RFs	Emb	Single-stream	No	MLM+VLM+MVM	CC3M+SBU	VLR+VCR
VL-BERT [49]	Image	OD-RFs	Emb	Single-stream	No	MLM+MVM	CC3M	GRE+VCR+VQA
VLP [50]	Image	OD-RFs	Emb	Dual-stream	Yes	MLM+LM	CC3M	VC+VQA
UNITER [8]	Image	OD-RFs	Emb	Single-stream	No	MLM+VLM+MVM+WRA	COCO+VG+SBU+CC3M	GRE+VLR+NLVR+VCR+VE+VQA
12-IN-1 [51]	Image	OD-RFs	Emb	Single-stream	No	MLM+MVM	MTL	GQA+GRE+VC+NLVR+VE+VQA
VisDial-BERT [52]	Image	OD-RFs	Emb	Dual-stream	No	MLM+VLM+MVM	CC3M+VQA	VD
ImageBERT [22]	Image	OD-RFs	Emb	Single-stream	No	MLM+VLM+MVM	LAIT+CC3M+SBU	VLR
PREVALENT [53]	Image	CNN-GFs+Xformer	Xformer	Single-stream	No	MLM+MVM	Matterport3D	VLN
XGPT [10]	Image	OD-RFs	Emb	Dual-stream	Yes	MLM+IDA+VC+TIFG	CC3M	VC+VLR
InterBER [54]	Image	OD-RFs	Emb	Single-stream	No	MLM+VLM+MVM	COCO+CC3M+SBU	VLR+VCR
PixelBERT [55]	Image	CNN-GFs	Emb	Single-stream	No	MLM+VLM	COCO+VG	VLR+NLVR+VQA
OSCAR [56]	Image	OD-RFs	Emb	Single-stream	No	MLM+VLM	COCO+SBU+CC3M+FLKR+VQA+GQA+VGQA	GQA+VC+VLR+NLVR+NoCaps+VQA
VLN-BERT [57]	Image	OD-RFs	Emb	Dual-stream	No	MLM+VLM+MVM	CC3M	VLN
FashionBERT [58]	Image	Xformer	Emb	Single-stream	No	MLM+VLM+MVM	FashionGen	VLR
VILLA [59]	Image	OD-RFs+Xformer	Xformer	Single-stream	No	MLM+VLM+MVM	COCO+VG+CC3M+SBU	GRE+VLR+NLVR+VCR+VE+VQA
ERNIE-ViL [60]	Image	OD-RFs	Emb	Single-stream	No	MLM+MVM	CC3M+SBU	GRE+VLR+VCR+VQA
RVL-BERT [61]	Image	OD-RFs	Emb	Single-stream	No	MLM+VLM+MVM	CC3M	VC+VQA
VinVL [62]	Image	OD-RFs	Emb	Single-stream	No	MLM+VLM	COCO+CC3M+SBU+FLKR+VQA+GQA+VGQA	GQA+VC+VLR+NLVR+NoCaps+VQA
VL-T5 [63]	Image	OD-RFs	Emb	Single-stream	Yes	MLM+VLM+VQA+GRE+VC	COCO+VG+VQA+GQA+VGQA	GQA+GRE+VC+MMT+NLVR+VCR+VQA
ViLT [64]	Image	ViT-PFs	Emb	Single-stream	No	MLM+VLM	COCO+VG+SBU+CC3M	VLR+NLVR+VQA
ALIGN [24]	Image	CNN-GFs	Xformer	Dual-stream	No	VLC	ALIGN	VLR
Kaleido-BERT [12]	Image	CNN-GFs	Emb	Single-stream	No	MLM+VLM+AKPM	FashionGen	CR+VC+VLR
MDETR [11]	Image	Xformer	Xformer	Single-stream	Yes	OD+MLM+VLC	COCO+VG+FLKR+GQA	GQA+VQA
SOHO [65]	Image	CNN-GFs	Emb	Single-stream	No	MLM+VLM+MVM	COCO+VG	VLR+NLVR+VE+VQA
E2E-VLP [66]	Image	CNN-GFs	Emb	Single-stream	Yes	OD+MLM+VLM	COCO+VG	VC+VLR+NLVR+VQA
Visual Parsing [67]	Image	Xformer	Emb	Single-stream	No	MLM+VLM+MVM	COCO+VG	VLR+VCR+VE+VQA
CLIP-ViL [68]	Image	CNN-GFs	Emb	Single-stream	Yes	MLM+VLM+VQA	COCO+VG+VQA+GQA+VGQA	VE+VLN+VQA
ALBEF [69]	Image	Xformer	Xformer	Dual-stream	No	MLM+VLM+VLC	COCO+VG+CC3M+SBU	VLR+NLVR+VQA
SimVLM [70]	Image	CNN-GFs	Emb	Single-stream	Yes	PrefixLM	ALIGN	VC+NLVR+VE+VQA
MURAL [71]	Image	CNN-GFs	Xformer	Dual-stream	No	VLC	CC12M+ALIGN	VC+VLR
VLMO [72]	Image	ViT-PFs	Emb	Single-stream	No	MLM+VLC+VLM	COCO+VG+CC3M+SBU	VQA+NLVR+VLR
METER [73]	Image	Xformer	Xformer	Dual-stream	No	MLM+VLM	COCO+VG+CC3M+SBU	VLR+NLVR+VE+VQA
X-VLM [74]	Image	Xformer	Xformer	Single-stream	No	MLM+VLM+VG	COCO+VG+CC3M+SBU	VLR+NLVR+VE+VQA
TCL [75]	Image	Xformer	Xformer	Single-stream	No	MLM+VLM+TCL	COCO+VG+CC3M+SBU	VLR+NLVR+VE+VQA

代表性模型

Video-Language Models (From VLP)

Table 3 The summary of mainstream video-text VLP models. The number of downstream tasks determines whether the model is generic or domain-specific VLP. FE: Feature Extraction. PT: Pre-training. Emb: Embedding. SC in Datatsets column: self-constructed or self-collected. MTL in Datatsets column: all datasets for multi-task learning in corresponding work. See other abbreviations in Datatsets column in Table 1.

Model	Domain	Vision FE	Language FE	Multimodal Fusion	Decoder	PT Objectives	PT Datasets	Downstream Tasks
VideoBERT [29]	Video	CNN-GFs	Emb	Single-stream	No	MLM+VLM+MVM	SC	AC+VC
CBT [76]	Video	CNN-GFs+Xformer	Xformer	Single-stream	No	VLC	Kinetics	AC+AS+VC
UniVL [77]	Video	CNN-GFs	Xformer	Dual-stream	Yes	MLM+VLM+VC	HT100M	AS+ASL+MSA+VC+VLR
HERO [9]	Video	CNN-GFs+Xformer	Xformer	Single-stream	No	MLM+VLM+MVM+FOM	HT100M+TV	VC+VLI+VQA+VLR
MMFT-BERT [78]	Video	OD-RFs+Xformer	Xformer	Single-stream	No	VQA	TV	VQA
ActBERT [79]	Video	OD-RFs+CNN	Emb	Single-stream	No	MLM+VLM+MVM	HT100M	AS+ASL+VC+VQA+VLR
CLIP [30]	Image / Video	CNN/Xformer	Xformer	Dual-stream	No	VLC	SC	OCR +AC etc.
Frozen [28]	Video	ViT-PFs	Emb	Dual-Stream	No	VLC	WebVid2M+CC3M	VLR
Region-Learner [80]	Video	ViT-PFs	Emb	Dual-Stream	No	VLC	WebVid2M+CC3M	VLR
CLIP4Clip [81]	Video	ViT-PFs	Emb	Dual-Stream	No	VLC	WebVid2M+CC3M	VLR
CLIP2Video [82]	Video	ViT-PFs	Emb	Dual-Stream	No	VLC	WebVid2M+CC3M	VLR

Datasets

中文数据集：

- 诺亚-悟空（介绍文章）：包括 1亿条图文对
- 智源-悟道: WuDaoMM Base 全量数据集约有6.5亿图文对，包含强相关数据5千万对和弱相关数据6亿对
  - 支持了包含文澜、Cogview等大规模中文多模态预训练模型的训练
- Zero: 23 million image-text pairs, collected from the search engine and contains images and corresponding textual descriptions, which is filtered from 5 billion image-text pairs by user click-through rate

More: (From VLP)

Table 1 Details of some popular pre-training datasets for VLP. Names of some datasets are abbreviated for the convenience of subsequent description. FLKR represents Flickr30k, and HT100M represents HowTo100M.

Dataset	# Images	# Image-text Pairs	Duration (hrs)	# Clips	# Videos
SBU [13]	875K	875K	-	-	-
FLKR [14]	29K	145K	-	-	-
COCO [15]	113K	567K	-	-	-
VG [16]	108K	5.4M	-	-	-
VGQA [16]	108K	1.8M	-	-	-
VQA [17]	83K	444K	-	-	-
Matterport3D [18]	104K	104K	-	-	-
FashionGen [19]	260K	260K	-	-	-
CC3M [20]	3M	3M	-	-	-
GQA [21]	82K	1M	-	-	-
LAIT [22]	10M	10M	-	-	-
CC12M [23]	12M	12M	-	-	-
ALIGN [24]	1.8B	1.8B	-	-	-
Kinetics400 [25]	-	-	817	306K	306K
TVQA [26]	-	-	461	22K	925
HT100M [27]	-	-	134K	136M	1.2M
WebVid2M [28]	-	-	13K	2.5M	2.5M

开源项目

中文VLP项目

- 悟空，Gitee
- CogView
- R2D2
- PaddleMM
- 太初/紫东太初：**图文音**多模态预训练模型

References

- [resources] https://github.com/lonePatient/awesome-pretrained-chinese-nlp-models#Multi-Modal
- [resources] https://github.com/yuewang-cuhk/awesome-vision-language-pretraining-papers
- VLP: A Survey on Vision-Language Pre-training, 2022; 综述 | 最新视觉-语言预训练综述 - 腾讯云开发者社区
- 北大邹月娴：视觉-语言预训练模型演进及应用
- Vision-Language Intelligence: Tasks, Representation Learning, and Large Models, 2022; 万字深度好文！视觉-语言（VL）智能：任务、表征学习和大型模型
- [slides] Tutorial on Multimodal Machine Learning
- 多模态预训练模型综述 - 知乎
- Recent Advances in Vision-and-Language Research, CVPR 2020 Tutorial
- VQA2VLN Tutorial 2021

Thanks

Learn More

Bilibili · Zhihu · Weibo · Github