Paper Reading

A Brief Survey of Representative Multimodal Pre-training Models

2023.06.08

Pandeng Li

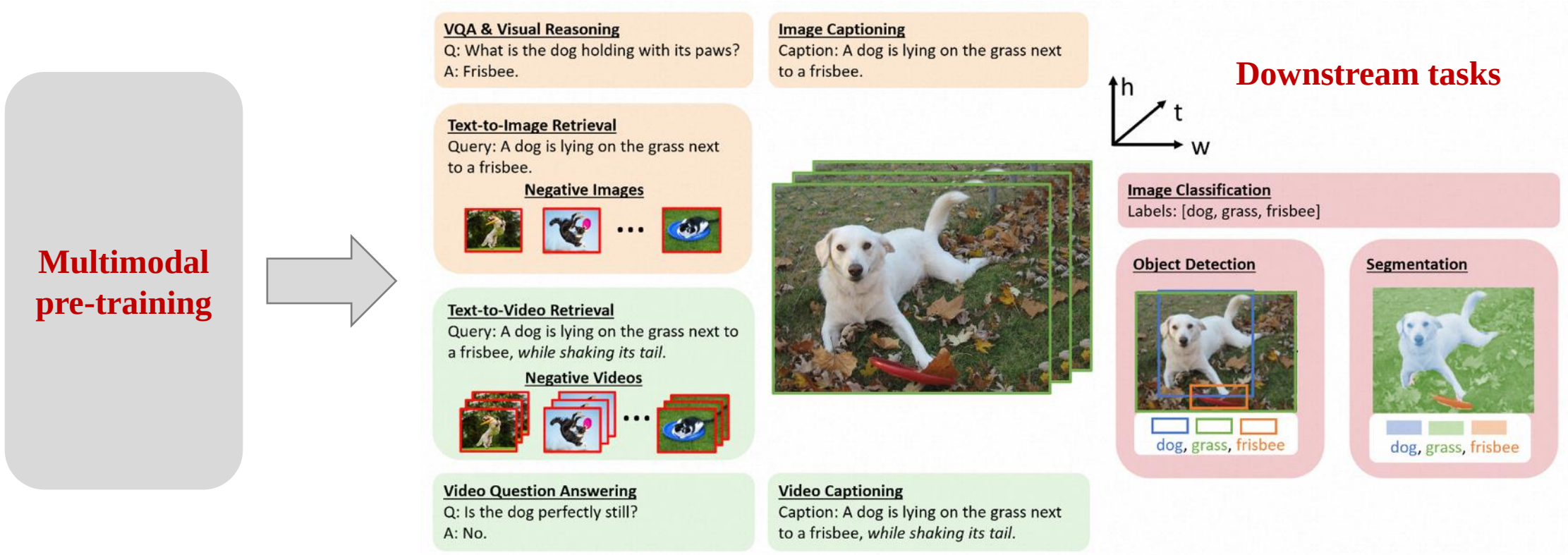
Contents

Task introduction

What is multimodal pre-training ?

- Multimodal pre-training refers to the process of training machine learning models on **multiple modalities**, such as text, images, and audio, before fine-tuning them for specific tasks.

- This pre-training allows the model to learn **general representations** from various data types, which can improve its performance when applied to specific downstream tasks.



Task introduction

Why do multimodal pre-training ?

- ✓ **Improved performance**: By learning from multiple data sources, models can gain a more comprehensive understanding of the data, leading to better performance on specific tasks.

- ✓ **Transfer learning**: Pre-trained models can be fine-tuned for various tasks, reducing the time and resources required for training from scratch.

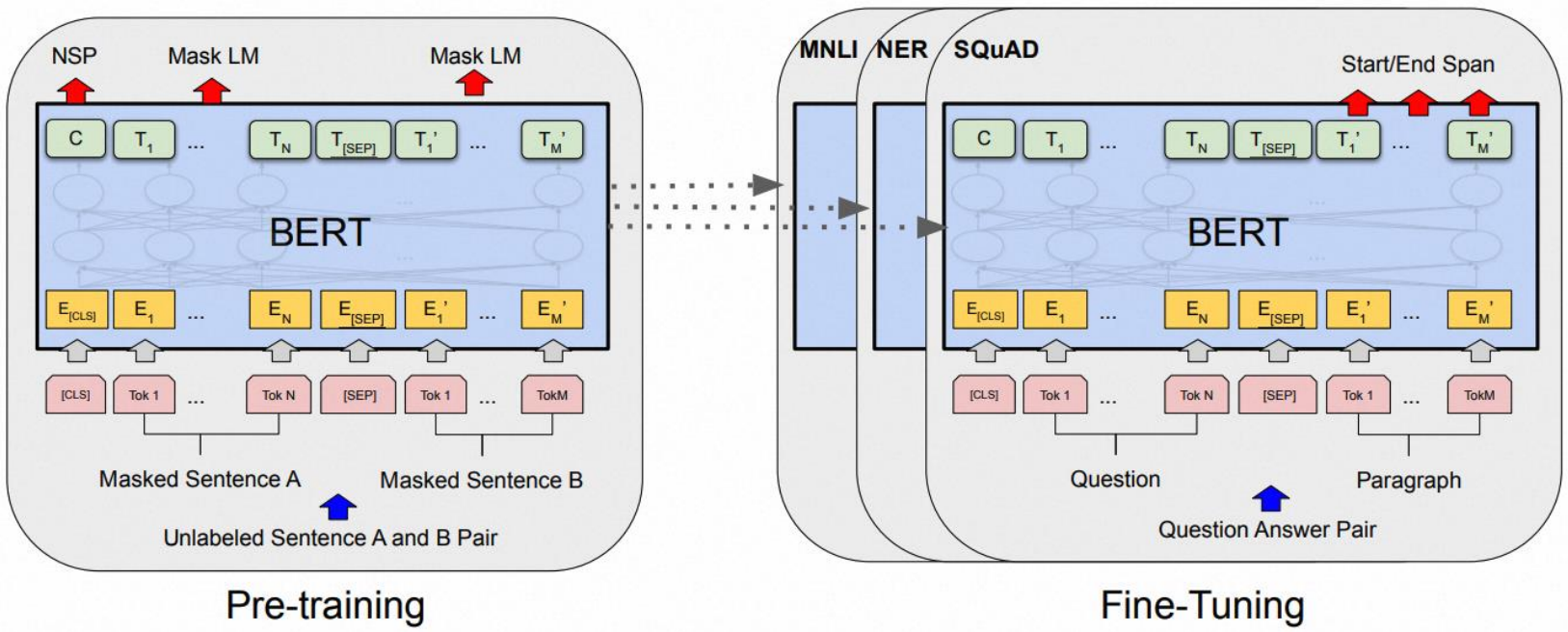
- ✓ **Leveraging complementary information**: Different modalities provide complementary information, which can help the model make more accurate predictions and improve generalization.

Method	#Pairs	MSRVTT			DiDeMo			ActivityNet			LSMDC			MSVD		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ClipBERT [42]	5.4M	22.0	46.8	59.9	20.4	48.0	60.8	21.3	49.0	63.5		-			-	
Frozen [5]	5M	31.0	59.5	70.5	34.6	65.0	74.7		-		15.0	30.8	39.8	33.7	64.7	76.3
VIOLET [27]	138M	34.5	63.0	73.4	32.6	62.8	74.7		-		16.1	36.6	41.2		-	
All-in-one [71]	138M	37.9	68.1	77.1	32.7	61.4	73.5	22.4	53.7	67.7		-			-	
LAVENDER [47]	30M	40.7	66.9	77.6	53.4	78.6	85.3		-		26.1	46.4	57.3	50.1	79.6	87.2
Singularity [41]	17M	42.7	69.5	78.1	53.1	79.9	88.1	48.9	77.0	86.3		-			-	
OmniVL [72]	17M	47.8	74.2	83.8	52.4	79.5	85.4		-			-			-	
VINDLU [15]	25M	46.5	71.5	80.4	61.2	85.8	91.0	55.0	81.4	89.7		-			-	
CLIP4Clip [55]	400M	44.5	71.4	81.6	42.8	68.5	79.2	40.5	72.4	83.4	21.6	41.8	49.8	46.2	76.1	84.6
CLIP-ViP [85]	500M	54.2	77.2	84.8	50.5	78.4	87.1	53.4	81.4	90.0	29.4	50.6	59.0		-	
InternVideo [76]	646M	55.2	79.6	87.5	57.9	82.4	88.9	62.2	85.9	93.2	34.0	53.7	62.9	58.4	84.5	90.4
	5M	46.3	72.7	82.0	54.8	83.0	89.0	52.1	80.5	89.6	30.3	51.8	61.4	67.0	92.7	96.7
UMT-B	17M	50.6	75.4	83.5	60.8	85.1	91.0	56.1	82.5	91.2	32.3	54.5	61.9	70.8	93.7	96.6
	25M	51.0	76.5	84.2	61.6	86.8	91.5	58.3	83.9	91.5	32.7	54.7	63.4	71.9	94. 5	97.8
	5M	53.3	76.6	83.9	59.7	84.9	90.8	58.1	85.5	92.9	37.7	60.6	67.3	76.9	96.7	98.7
UMT-L	17M	56.5	80.1	**87.4**	66.6	89.9	**93.7**	66.6	88.6	94.7	41.4	63.8	72.3	78.8	97.3	98.8
	25M	**58.8**	**81.0**	87.1	**70.4**	**90.1**	93.5	**66.8**	**89.1**	**94.9**	**43.0**	**65.5**	**73.0**	**80.3**	**98.1**	**99.0**

Contents

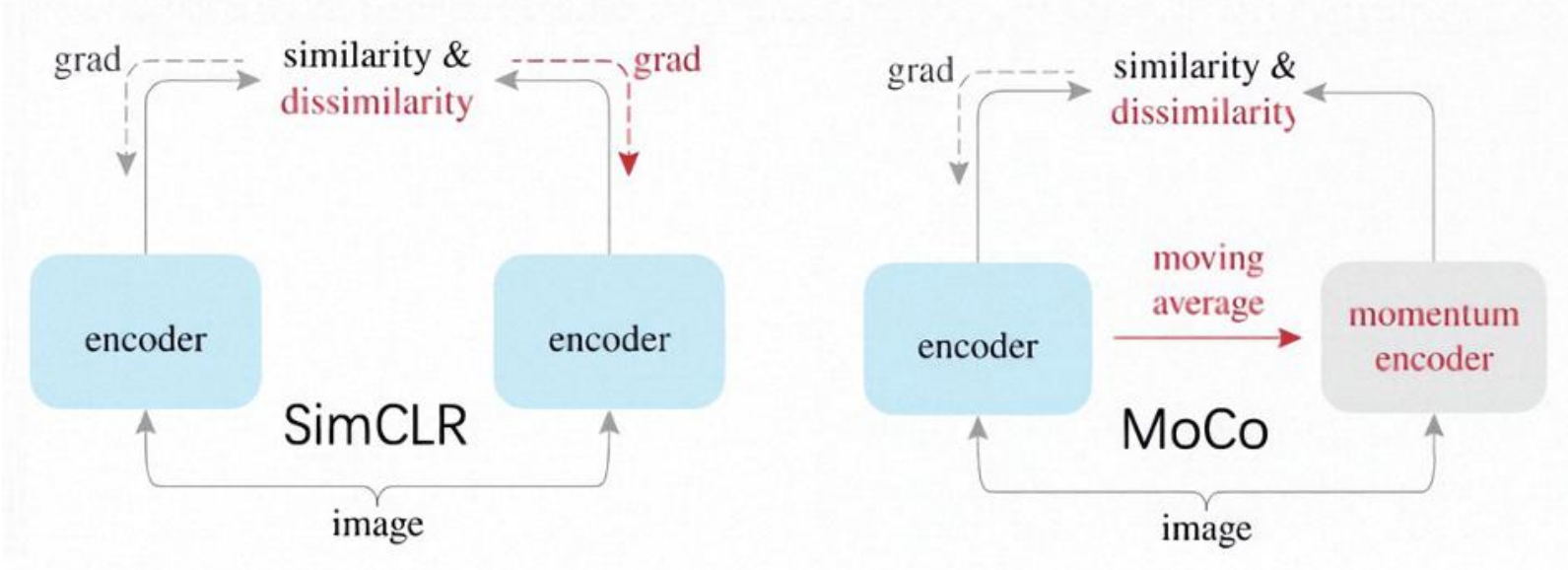
Background: Masked Language Modeling (Generation Perspective)

Bidirectional Encoder Representation from Transformers



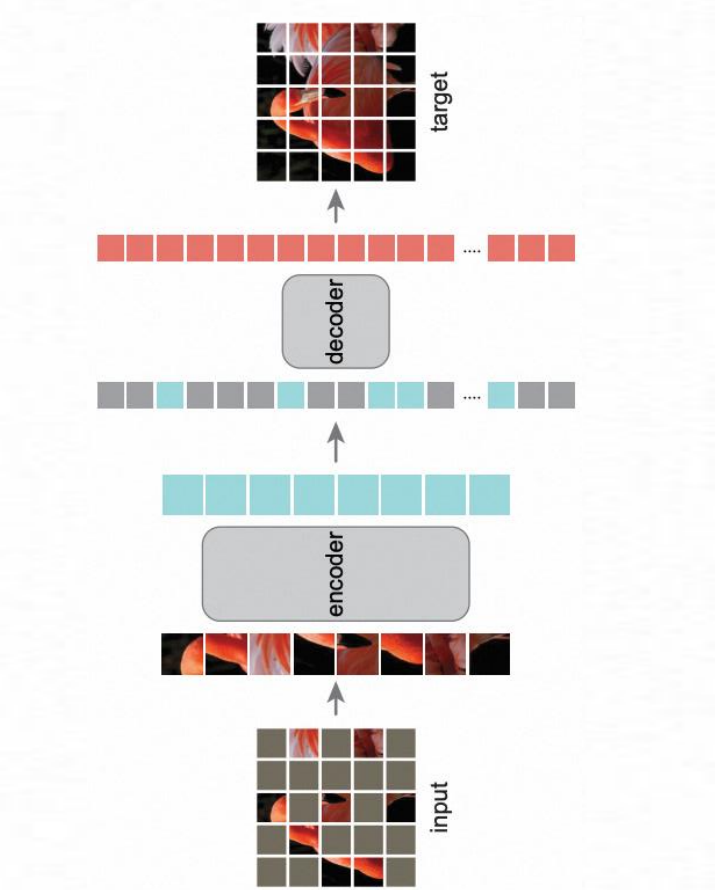
- ✓ **Masked Language Model**
- ✓ **Next Sentence Prediction**

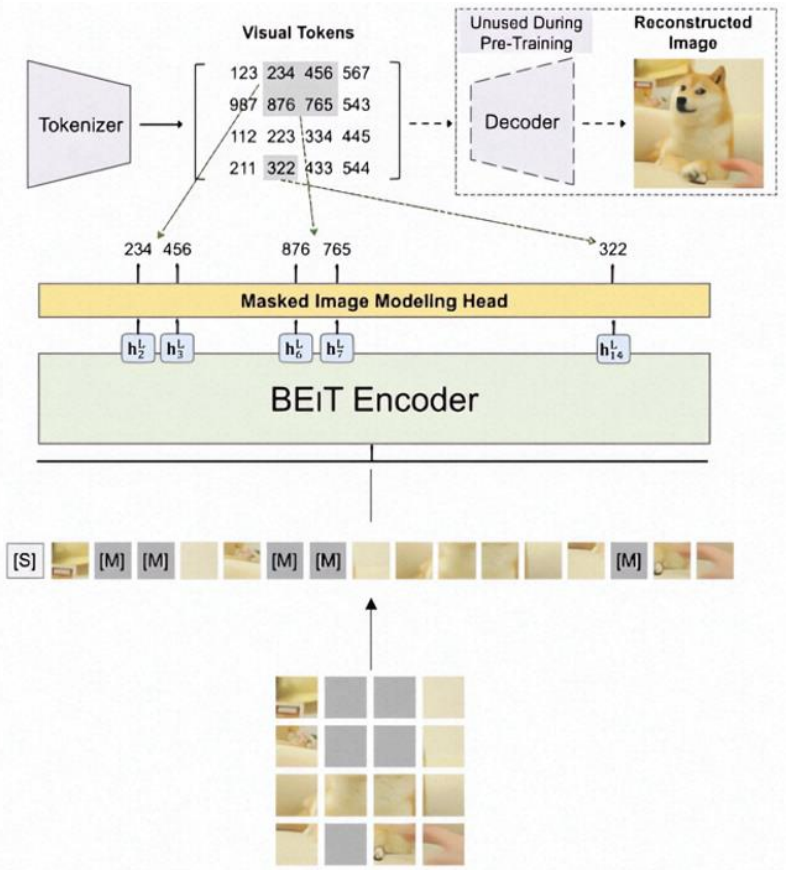
Background: Contrastive Learning (Discrimination Perspective)

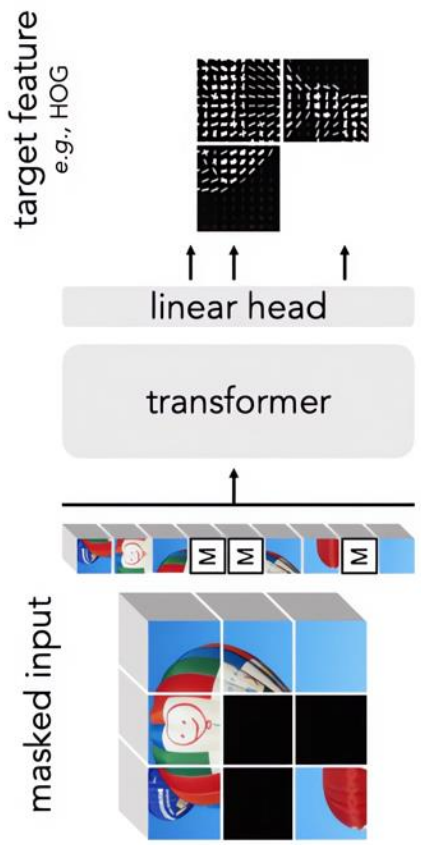


$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i / \tau)}$$

Background: Masked Image Modeling (Generation Perspective)







Masked Pixel Prediction (MAE)

Masked Token Prediction (BEiT)

Masked Feature Prediction (MaskFeat)

Contents

Pre-training scale

https://huggingface.co/blog/vision_language_pretraining

- **Medium-scale pre-training (2019/8-2021/8).** ViLBERT ,UNITER, OSCAR…
  - using image-text datasets up to 4M images (roughly 10M image-text pairs)
  - model sizes ranging from 110M (BERT-base) to 340M (BERT-large).

Conceptual Captions (CC3M)
SBU Captions (SBU)
COCO
Visual Genome (VG)

- **Large-scale pre-training (2021/8-now).** CLIP, BLIP, BEiT-3 …
  - pre-trained over roughly 1B-10B image-text pairs.
  - current models typically contain roughly 1B parameters

Conceptual 12M (CC12M)
LAION-400M, 2B

Model	Model Size				PT dataset size	PT Tasks
	Image Enc.	Text Enc.†	Fusion†	Total		
CLIP ViT-L/14 (Radford et al., 2021)	302M	123M	0	425M	400M	ITC
ALIGN (Jia et al., 2021)	480M	340M	0	820M	1.8B	ITC
Florence (Yuan et al., 2021)	637M	256M	0	893M	900M	ITC
SimVLM-huge (Wang et al., 2022k)	300M	39M	600M	939M	1.8B	PrefixLM
METER-huge (Dou et al., 2022b)	637M	125M	220M	982M	900M+20M[1]	MLM+ITM
LEMON (Hu et al., 2022)	147M[2]	39M	636M	822M	200M	MLM
Flamingo (Alayrac et al., 2022)	200M	70B	10B	80.2B	2.1B+27M[3]	LM
GIT (Wang et al., 2022d)	637M	40M	70M	747M	800M	LM
GIT2 (Wang et al., 2022d)	4.8B	40M	260M	5.1B	12.9B	LM
CoCa (Yu et al., 2022a)	1B	477M	623M	2.1B	1.8B+3B[4]	ITC+LM
BEiT-3 (Wang et al., 2022g)	692M[5]	692M[5]	52M[5]	1.9B	21M+14M[6]	MIM+MLM +MVLM
PaLI (Chen et al., 2022e)	3.9B	40M	13B	16.9B	1.6B	LM+VQA[7] +OCR+OD

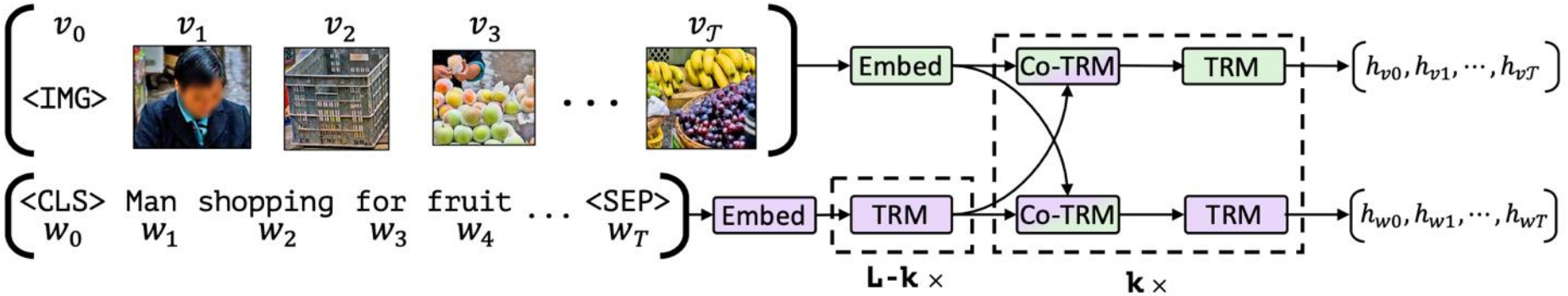
Contents

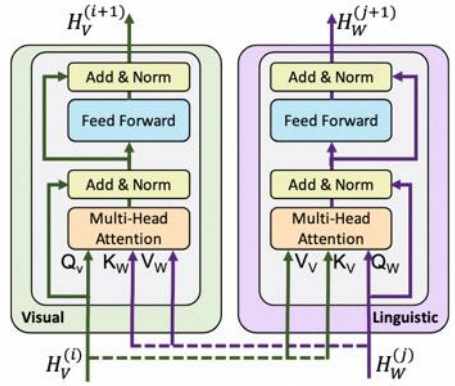
Medium-scale pre-training

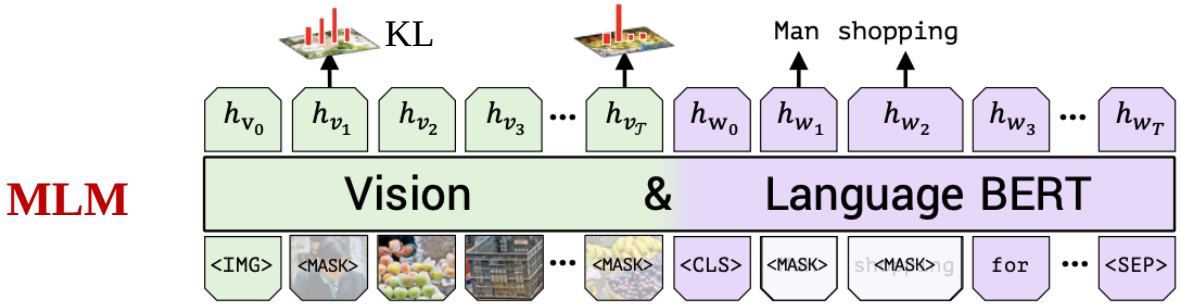
- **ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks**
- **UNITER: UNiversal Image-TExt Representation Learning**
- **Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks**
- **VisualBERT , LXMERT,VL-BERT, VILLA, VinVL, UNIMO…**

ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

(NeurIPS19, Facebook AI Research, cite 2347, https://github.com/facebookresearch/vilbert-multi-task)







MLM

(a) Masked multi-modal learning

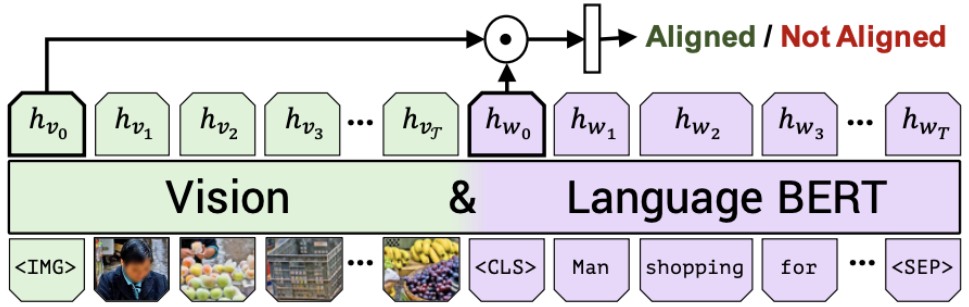


Image-Text Matching

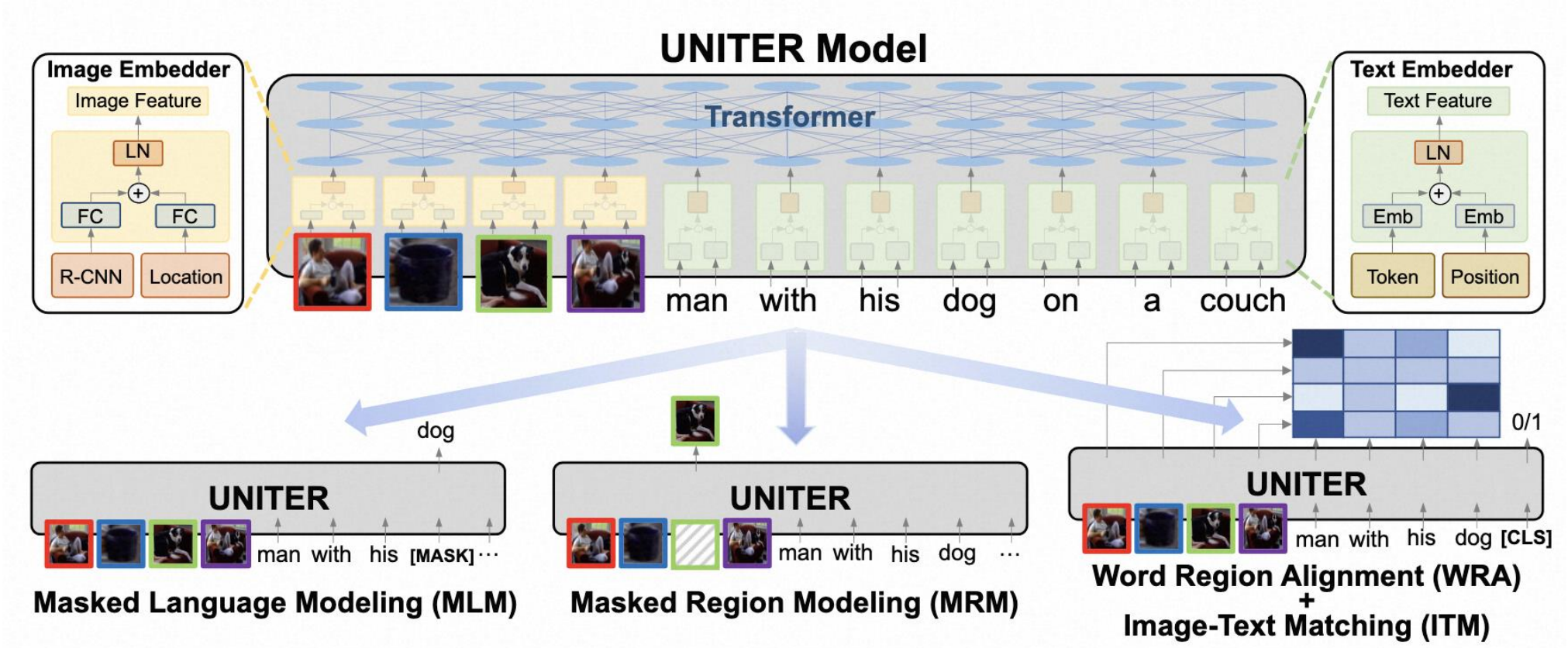
(b) Multi-modal alignment prediction

Faster R-CNN, BERT, 8 TitanX GPUs, CC3M

VQA, visual grounding, image-text retrieval

UNITER: UNiversal Image-TExt Representation Learning

(ECCV20, Microsoft Dynamics 365 AI Research, cite 1128, https://github.com/ChenRocks/UNITER)

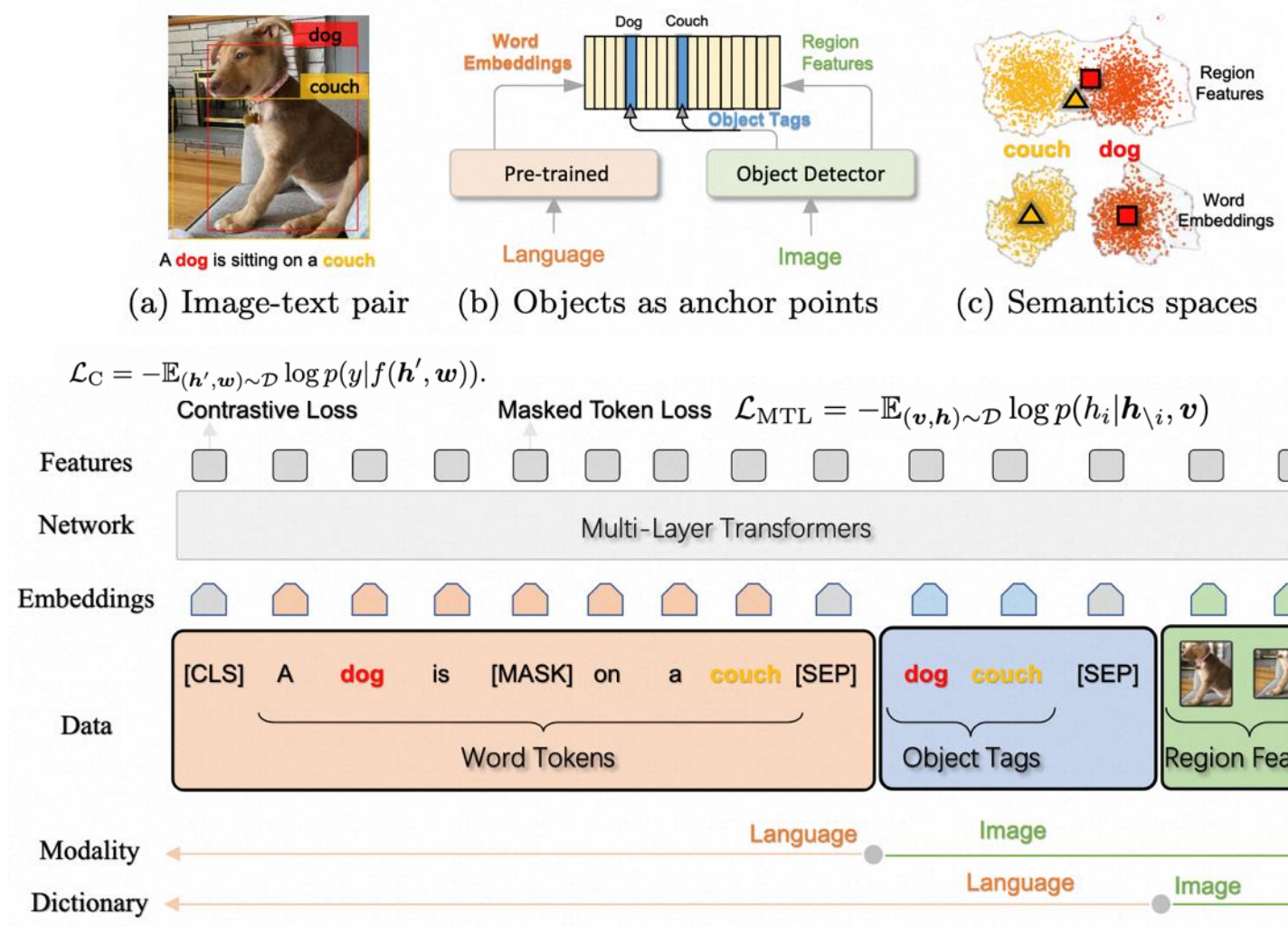


Faster R-CNN, BERT, 882 V100 GPU hours, [COCO, Visual Genome, CC3M, and SBU Captions]

a novel Word-Region Alignment pre-training task $\mathcal{L}_{\mathrm{WRA}}(\theta) = \mathcal{D}_{ot}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\mathbf{T} \in \Pi(\mathbf{a}, \mathbf{b})} \sum_{i=1}^{T} \sum_{j=1}^{K} \mathbf{T}_{ij} \cdot c(\mathbf{w}_i, \mathbf{v}_j)$

Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks

(ECCV20, Microsoft Corporation , cite 1185, https://github.com/microsoft/Oscar)



Faster R-CNN, BERT, 6.5M text-tag-image triples

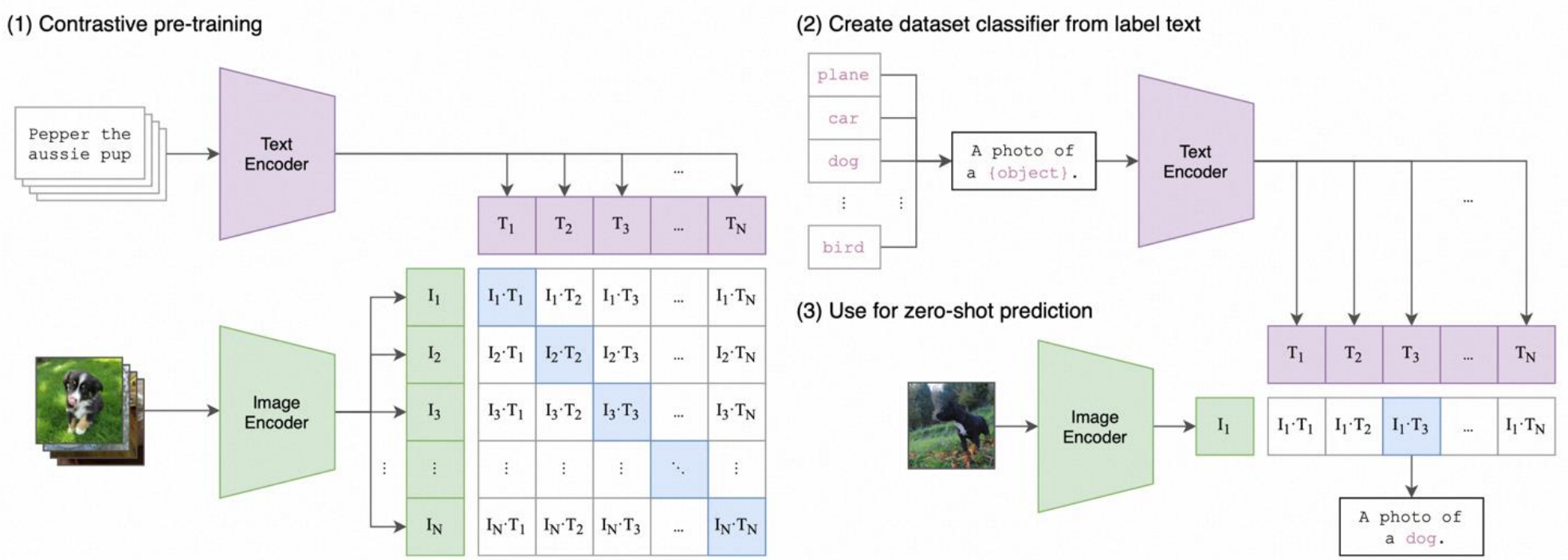
Contents

Large-scale pre-training

- **CLIP, ALIGN -- very simple (Open AI, Google Research)**
- **ALBEF, BLIP, BLIP-2 -- multimodal interaction, understanding and generation (Salesforce Research)**
- **BEIT V3-- MIM (Microsoft)**
- **EVA -- scaling up, best open-source model (Beijing Academy of Artificial Intelligence)**
- **SIMVLM , Flamingo, Florence, CoCa, OFA, GIT, PALI…**

Learning Transferable Visual Models From Natural Language Supervision
(ICML21, Open AI , cite 5679, https://github.com/OpenAI/CLIP)

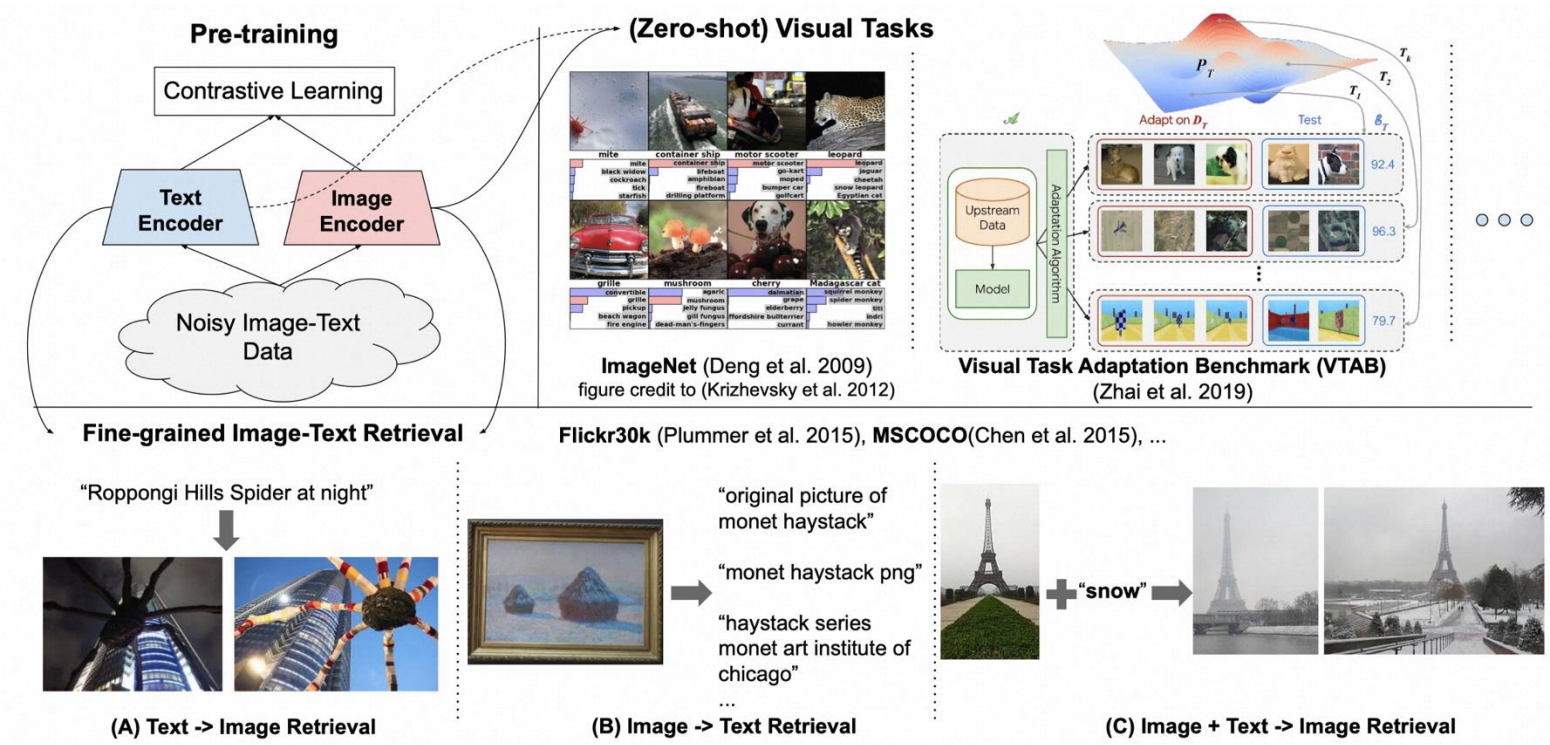
- modern pre-training methods within web-scale collections of text surpasses that of high-quality crowd-labeled NLP datasets.



5 ResNets & 3 ViTs, GPT-2, 73728 V100 GPU hours, 400M web data pairs

Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision

(ICML21, Google Research, cite 1175)



EfficientNet, BERT, 1024 Cloud TPUv3 cores, 1.8B pairs

Image-based filtering.

- remove pornographic images
- Image dimension >200
- Images with more than 1000 associated alt-texts are discarded
- Remove near-duplicates images in downstream evaluation datasets

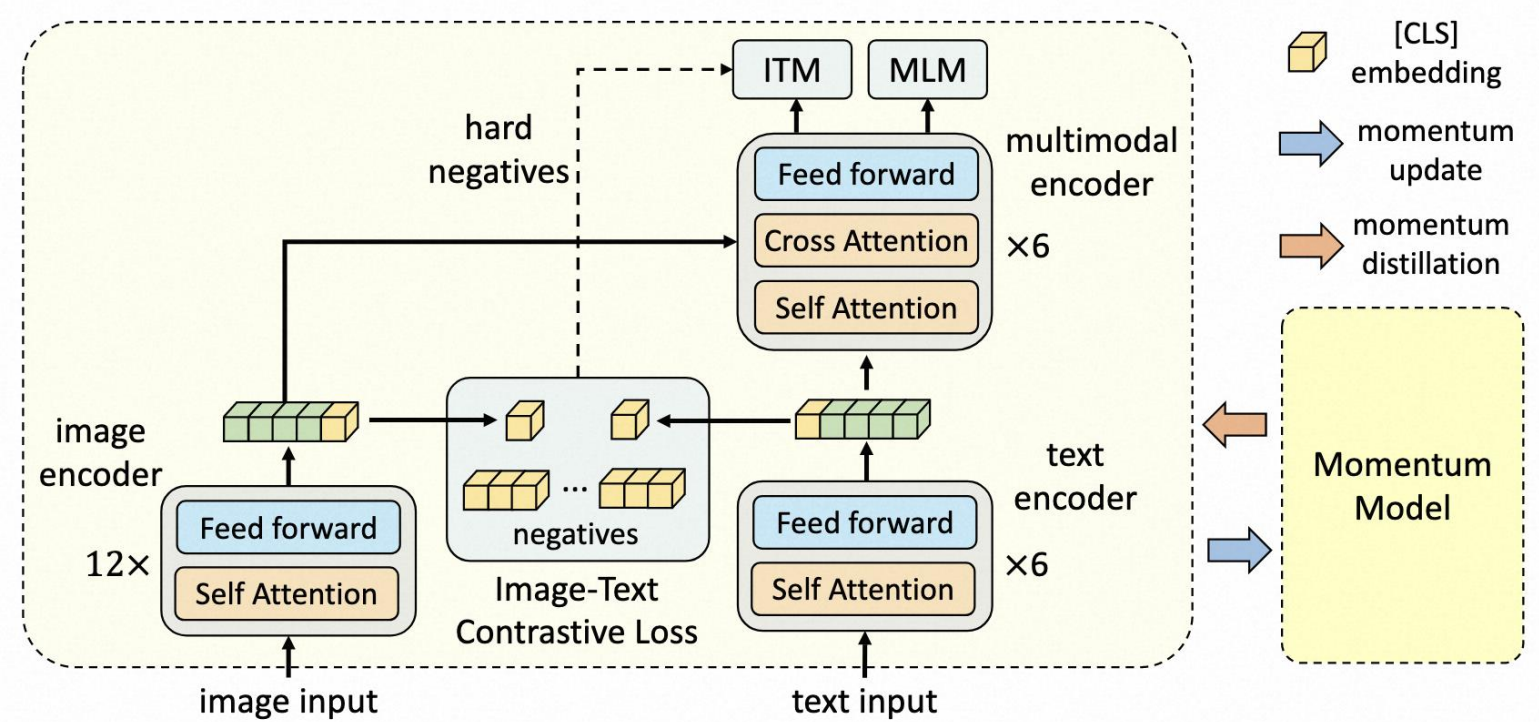
Text-based filtering.

- exclude alt-texts that are shared by more than 10 images.
- contain any rare token, irrelevant
- 3-20 words
- Noisy characters

		Flickr30K (1K test set)					
		image → text			text → image		
		R@1	R@5	R@10	R@1	R@5	R@10
Zero-shot	ImageBERT	70.7	90.2	94.0	54.3	79.6	87.5
	UNITER	83.6	95.7	97.7	68.7	89.2	93.9
	CLIP	88.0	98.7	99.4	68.7	90.6	95.2
	ALIGN	**88.6**	**98.7**	**99.7**	**75.7**	**93.8**	**96.8**
Fine-tuned	GPO	88.7	98.9	99.8	76.1	94.5	97.1
	UNITER	87.3	98.0	99.2	75.6	94.1	96.8
	ERNIE-ViL	88.1	98.0	99.2	76.7	93.6	96.4
	VILLA	87.9	97.5	98.8	76.3	94.2	96.8
	Oscar	-	-	-	-	-	-
	ALIGN	**95.3**	**99.8**	**100.0**	**84.9**	**97.4**	**98.6**

Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

(NeurIPS21, Salesforce Research , cite 569, https://github.com/salesforce/ALBEF)



- ViT-B/16, BERT, 8 A100, [CC3M/Conceptual 12M, SBU, COCO, Visual Genome]
- Image-text Retrieval, Visual Entailment, VQA, Visual Reasoning, Visual Grounding

Method	# Pre-train Images	Flickr30K (1K test set)					
		TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10
UNITER [2]	4M	83.6	95.7	97.7	68.7	89.2	93.9
CLIP [6]	400M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN [7]	1.2B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF	4M	90.5	98.8	99.7	76.8	93.7	96.7
ALBEF	14M	**94.1**	**99.5**	**99.7**	**82.8**	**96.3**	**98.1**

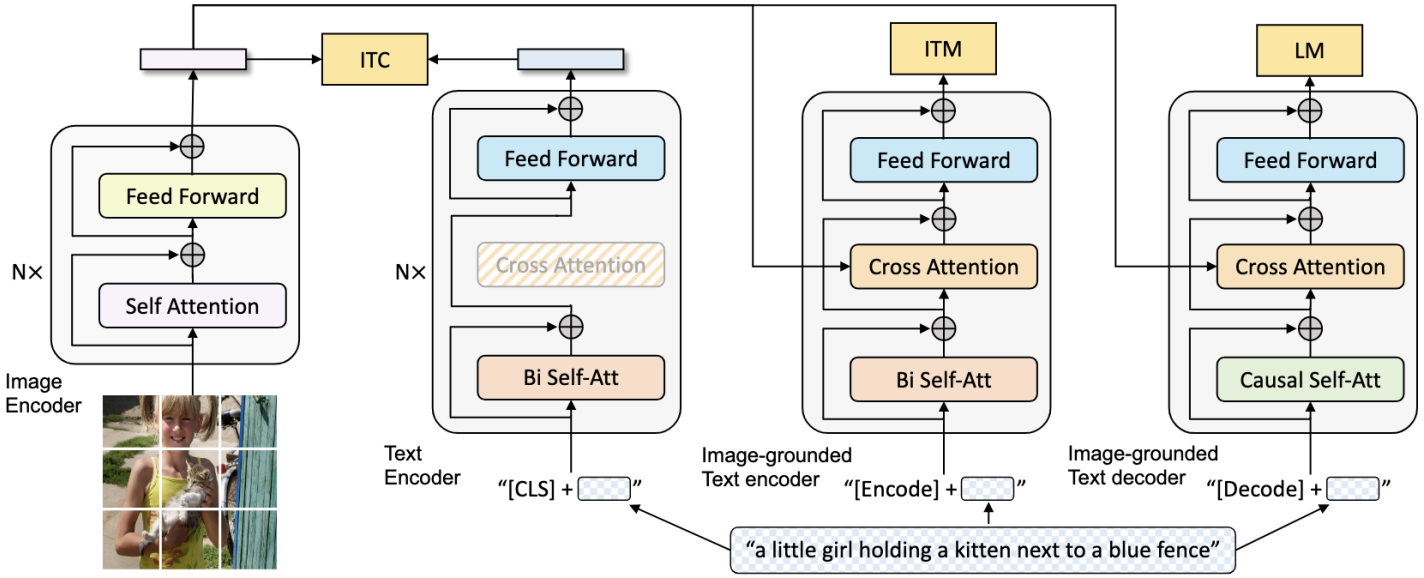
Table 3: Zero-shot image-text retrieval results on Flickr30K.

$$\mathcal{L}_{\text{itc}}^{\text{mod}} = (1-\alpha)\mathcal{L}_{\text{itc}} + \frac{\alpha}{2}\mathbb{E}_{(I,T)\sim D}\left[\text{KL}(\boldsymbol{q}^{\text{i2t}}(I) \parallel \boldsymbol{p}^{\text{i2t}}(I)) + \text{KL}(\boldsymbol{q}^{\text{t2i}}(T) \parallel \boldsymbol{p}^{\text{t2i}}(T))\right]$$

$$\mathcal{L}_{\text{mlm}}^{\text{mod}} = (1-\alpha)\mathcal{L}_{\text{mlm}} + \alpha\mathbb{E}_{(I,\hat{T})\sim D}\text{KL}(\boldsymbol{q}^{\text{msk}}(I,\hat{T}) \parallel \boldsymbol{p}^{\text{msk}}(I,\hat{T}))$$

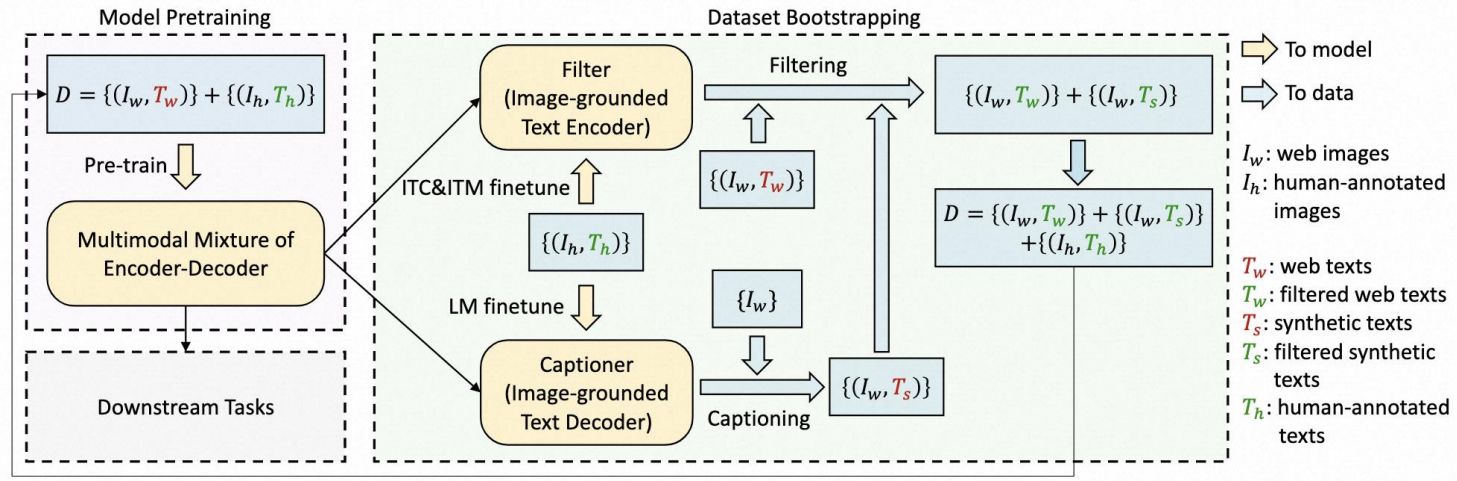
BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

(ICML22, Salesforce Research, cite 415, https://github.com/salesforce/BLIP)



- most pre-trained models only excel in either understanding-based tasks or generation-based tasks.
- performance improvement has been largely achieved by scaling up the dataset with noisy image-text pairs collected from the web

- ViT-B/16, BERT, 32 GPU cores
- 129M: [Conceptual 12M, SBU, COCO, Visual Genome, LAION-115M]

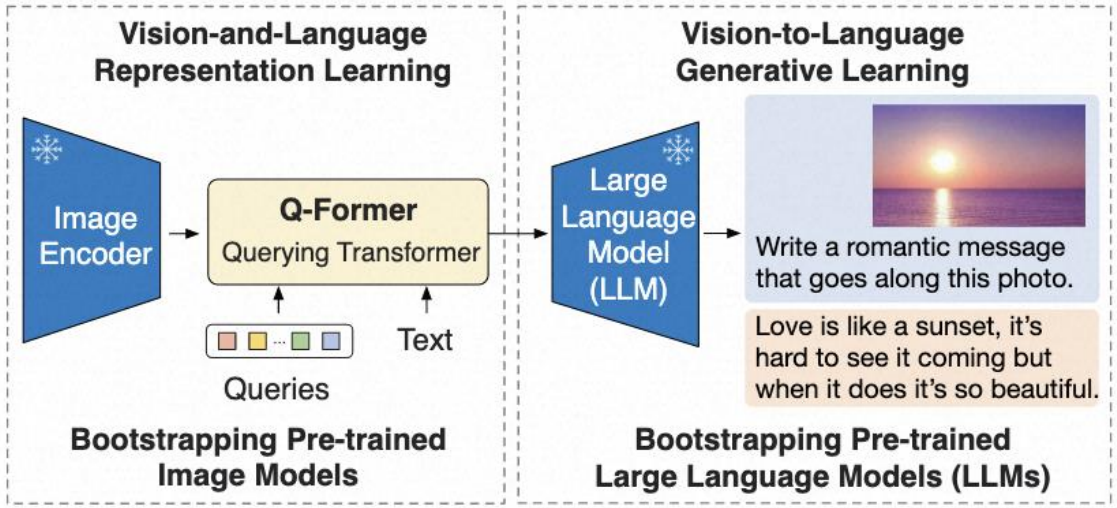


Method	Pre-train # Images	Flickr30K (1K test set) TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10
CLIP	400M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN	1.8B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF	14M	94.1	99.5	99.7	82.8	96.3	98.1
BLIP	14M	94.8	99.7	**100.0**	84.9	96.7	98.3
BLIP	129M	**96.0**	**99.9**	**100.0**	85.0	**96.8**	98.6
$\text{BLIP}_{\text{CapFilt-L}}$	129M	**96.0**	**99.9**	**100.0**	**85.5**	**96.8**	**98.7**
$\text{BLIP}_{\text{ViT-L}}$	129M	96.7	100.0	100.0	86.7	97.3	98.7

Table 6. Zero-shot image-text retrieval results on Flickr30K.

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

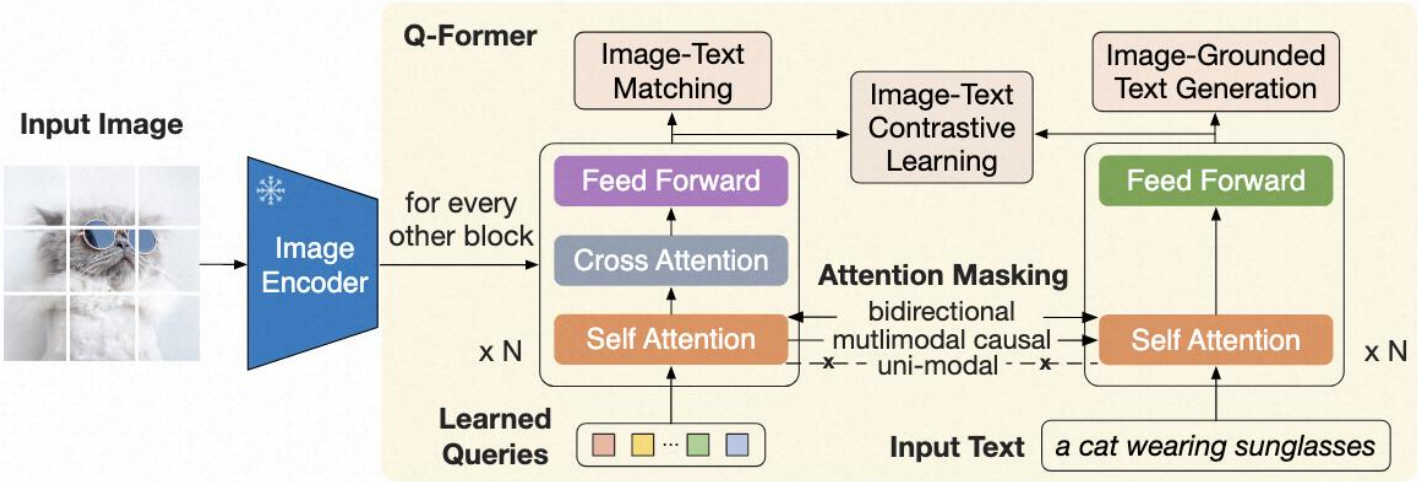
(arXiv:2301.12597, Salesforce Research, cite 112, https://github.com/salesforce/LAVIS)

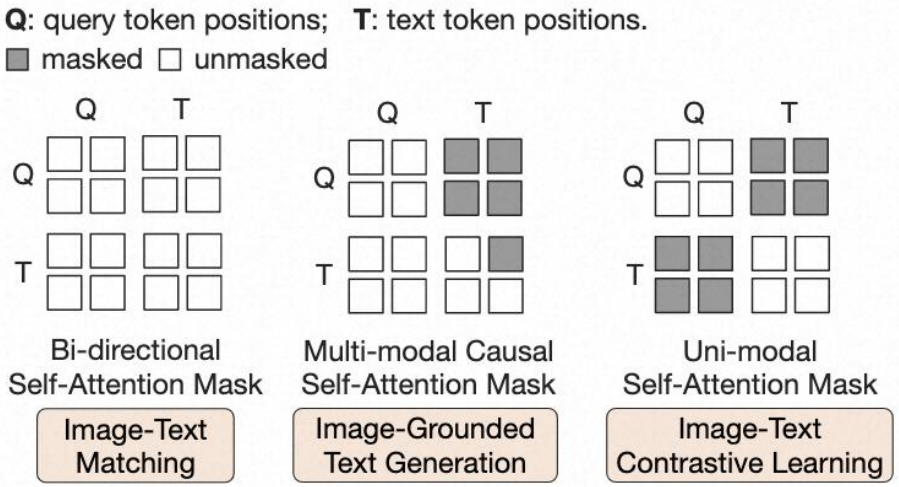


- Most sota models incur a high computation cost during pre-training, due to end-to-end training using large-scale models and datasets.

Frozen -> Flamingo -> BLIP-2

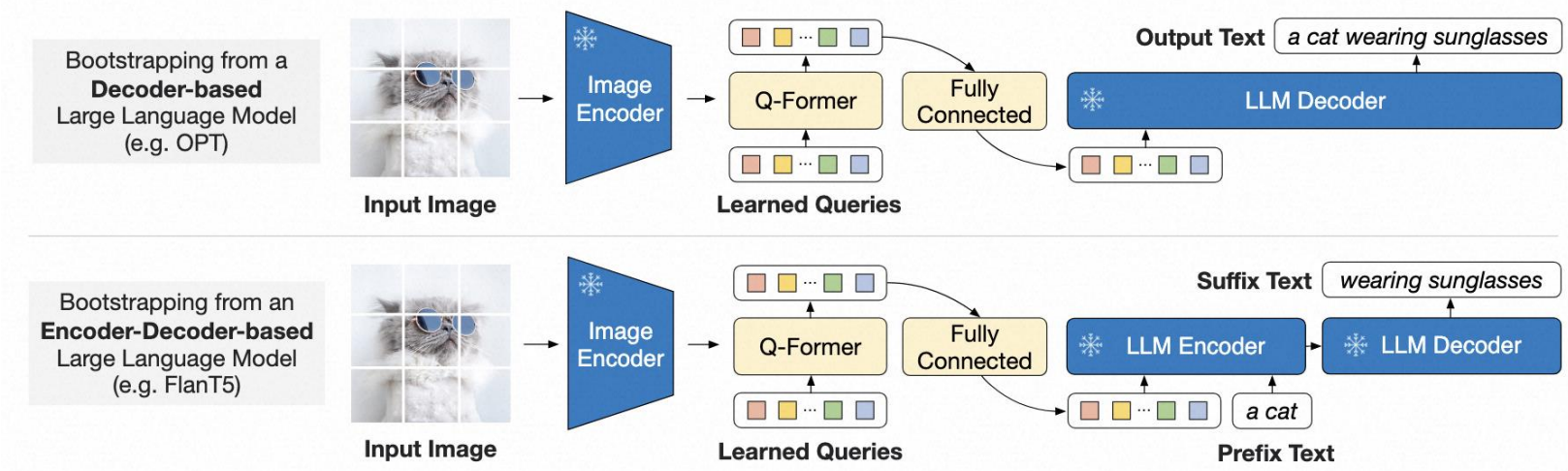
- ViT-L/14, BERT, OPT/FlanT5, **16 A100(40G), 9 days**
- 129M: [Conceptual 12M, CC3M, SBU, COCO, Visual Genome, LAION-115M] + synthetic captions from BLIP





BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

(arXiv:2301.12597, Salesforce Research, cite 112, https://github.com/salesforce/LAVIS)



Model	#Trainable Params	Flickr30K Zero-shot (1K test set) Image → Text R@1	R@5	R@10	Text → Image R@1	R@5	R@10	COCO Fine-tuned (5K test set) Image → Text R@1	R@5	R@10	Text → Image R@1	R@5	R@10
Dual-encoder models													
CLIP (Radford et al., 2021)	428M	88.0	98.7	99.4	68.7	90.6	95.2	-	-	-	-	-	-
ALIGN (Jia et al., 2021)	820M	88.6	98.7	99.7	75.7	93.8	96.8	77.0	93.5	96.9	59.9	83.3	89.8
FILIP (Yao et al., 2022)	417M	89.8	99.2	99.8	75.0	93.4	96.3	78.9	94.4	97.4	61.2	84.3	90.6
Florence (Yuan et al., 2021)	893M	90.9	99.1	-	76.7	93.6	-	81.8	95.2	-	63.2	85.7	-
BEIT-3(Wang et al., 2022b)	1.9B	94.9	99.9	**100.0**	81.5	95.6	97.8	84.8	96.5	98.3	67.2	**87.7**	**92.8**
Fusion-encoder models													
UNITER (Chen et al., 2020)	303M	83.6	95.7	97.7	68.7	89.2	93.9	65.7	88.6	93.8	52.9	79.9	88.0
OSCAR (Li et al., 2020)	345M	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5
VinVL (Zhang et al., 2021)	345M	-	-	-	-	-	-	75.4	92.9	96.2	58.8	83.5	90.3
Dual encoder + Fusion encoder reranking													
ALBEF (Li et al., 2021)	233M	94.1	99.5	99.7	82.8	96.3	98.1	77.6	94.3	97.2	60.7	84.3	90.5
BLIP (Li et al., 2022)	446M	96.7	**100.0**	**100.0**	86.7	97.3	98.7	82.4	95.4	97.9	65.1	86.3	91.8
BLIP-2 ViT-L	474M	96.9	**100.0**	**100.0**	88.6	97.6	**98.9**	83.5	96.0	98.0	66.3	86.5	91.8
BLIP-2 ViT-g	1.2B	**97.6**	**100.0**	**100.0**	**89.7**	**98.1**	**98.9**	**85.4**	**97.0**	**98.5**	**68.3**	**87.7**	92.6

zero-shot image-to-text generation

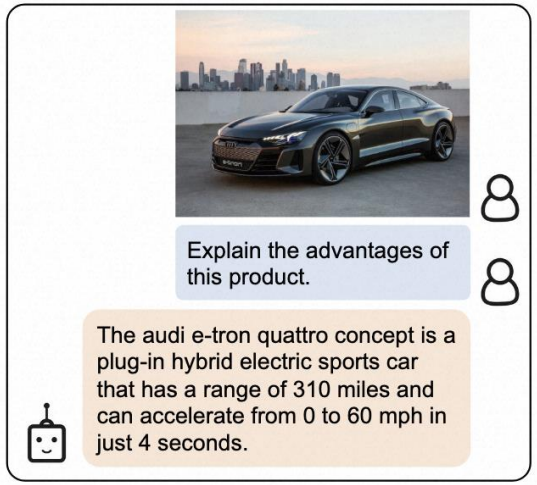
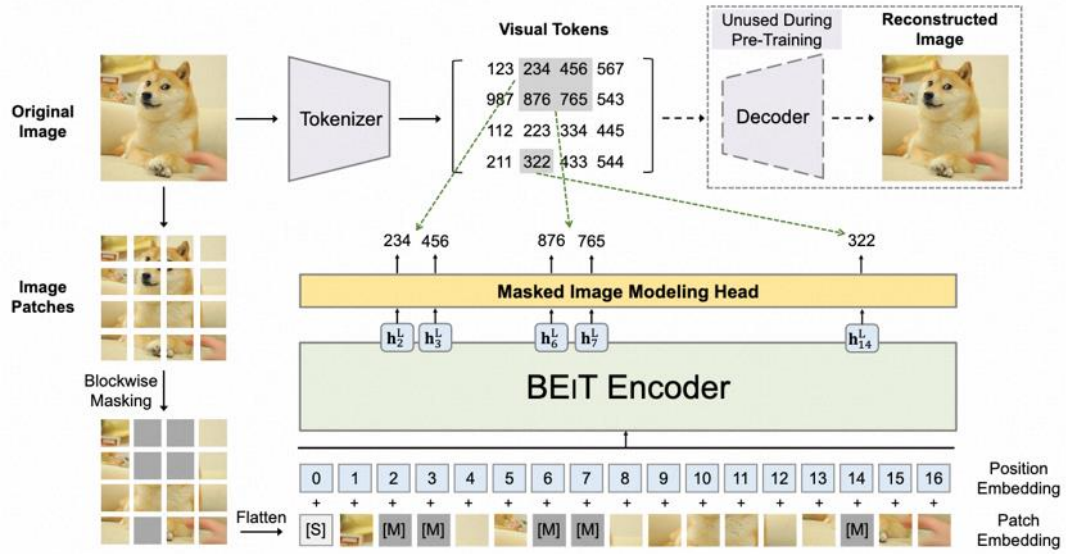
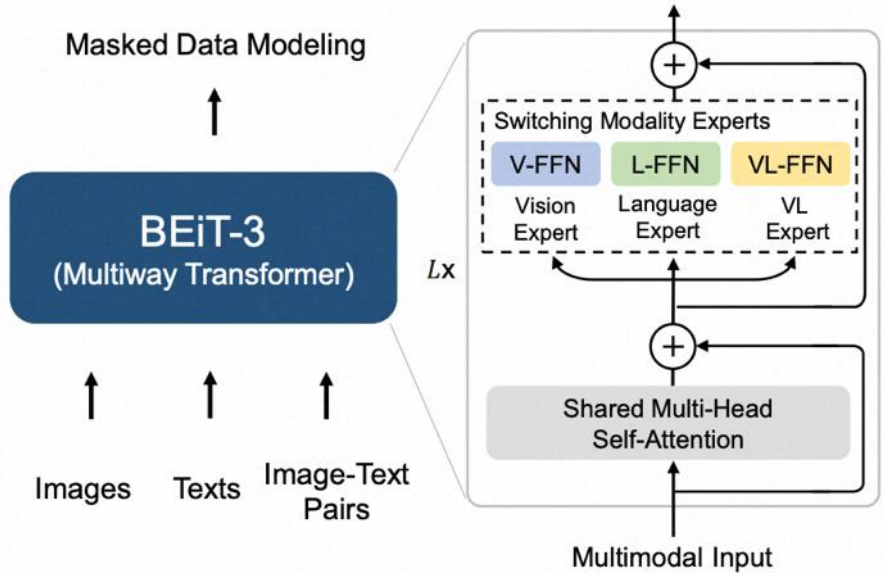


Image as a Foreign Language: BEIT Pretraining for All Vision and Vision-Language Tasks (CVPR23, Microsoft Corporation, cite 145, https://aka.ms/beit-3)

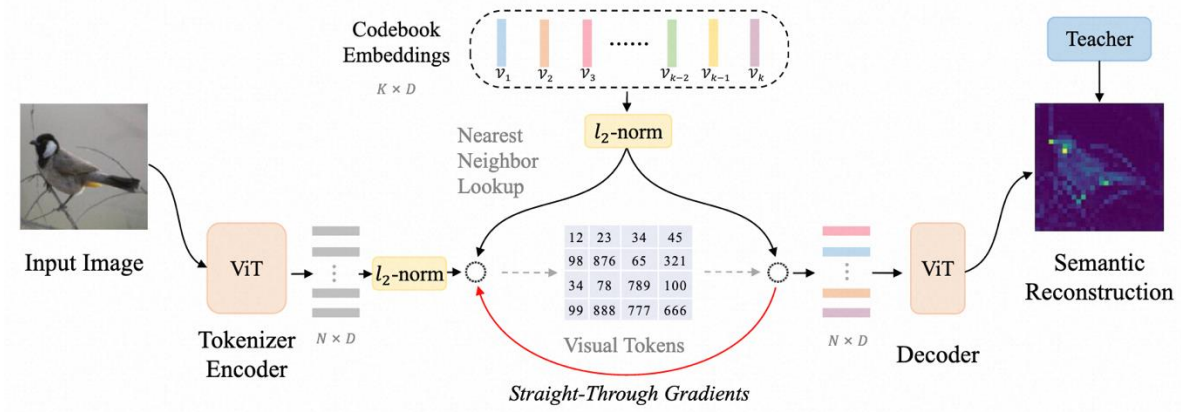
BEIT V1:



BEIT V3:



BEIT V2:



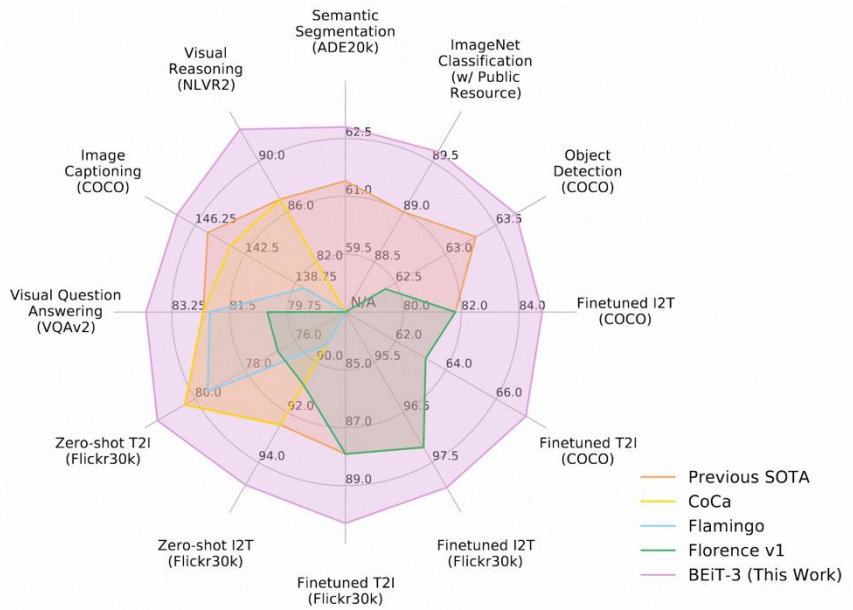
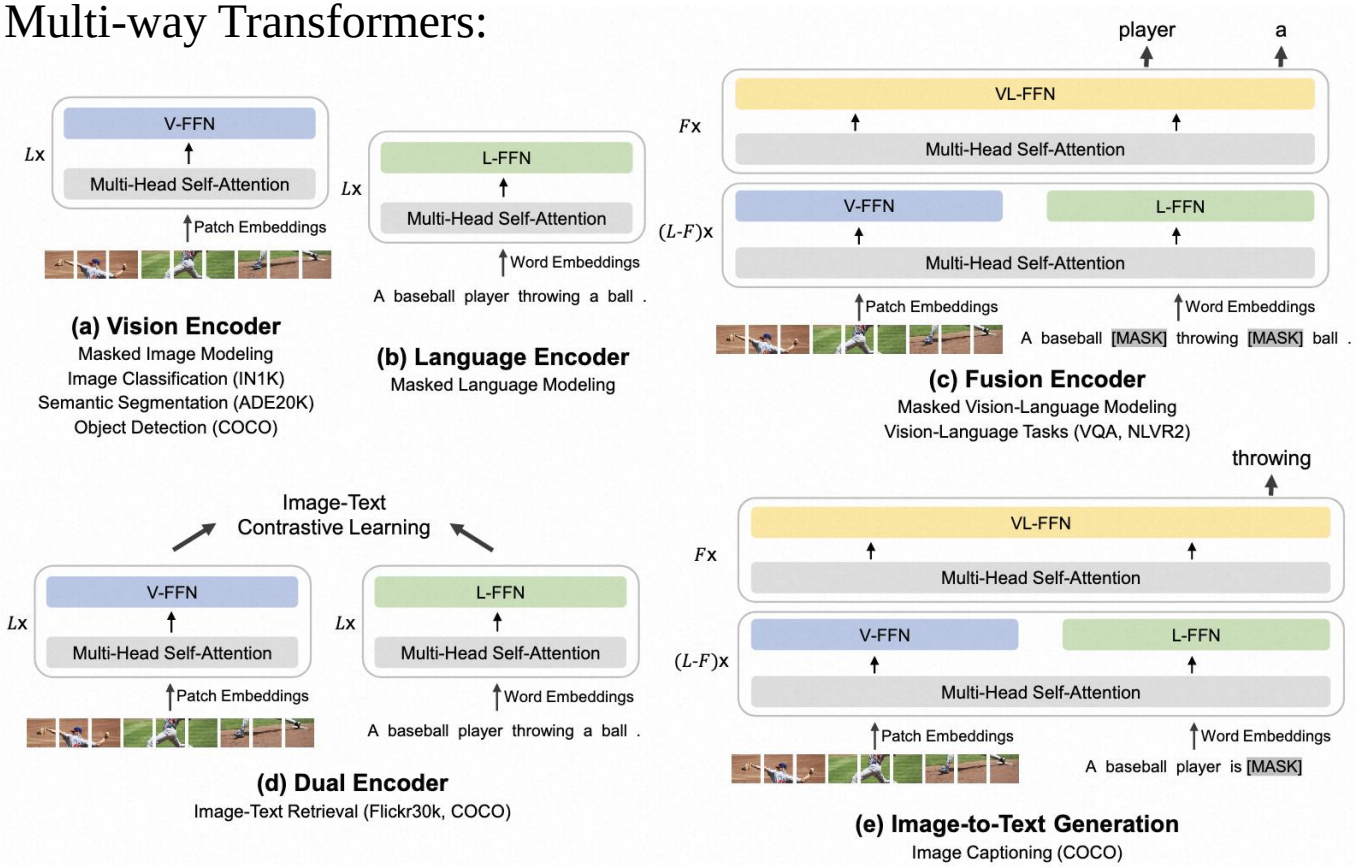


Image as a Foreign Language: BEIT Pretraining for All Vision and Vision-Language Tasks (CVPR23, Microsoft Corporation, cite 145, https://aka.ms/beit-3)

Multi-way Transformers:



Model	#Layers	Hidden Size	MLP Size	#Parameters				
				V-FFN	L-FFN	VL-FFN	Shared Attention	Total
BEIT-3	40	1408	6144	692M	692M	52M	317M	1.9B

Table 2: Model configuration of BEIT-3. The architecture layout follows ViT-giant [ZKHB21].

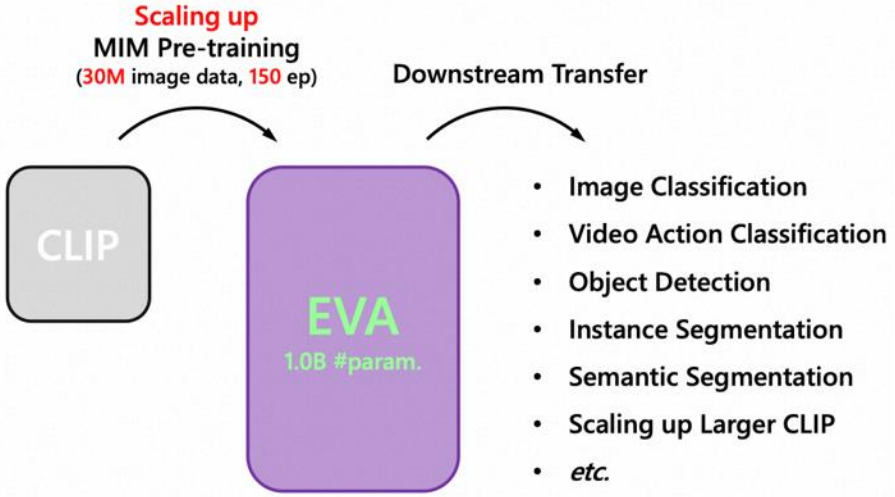
Data	Source	Size
Image-Text Pair	CC12M, CC3M, SBU, COCO, VG	21M pairs
Image	ImageNet-21K	14M images
Text	English Wikipedia, BookCorpus, OpenWebText, CC-News, Stories	160GB documents

Table 3: Pretraining data of BEIT-3. All the data are academically accessible.

Model	MSCOCO (5K test set)						Flickr30K (1K test set)					
	Image → Text			Text → Image			Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Fusion-encoder models												
UNITER [CLY+20]	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
VILLA [GCL+20]	-	-	-	-	-	-	87.9	97.5	98.8	76.3	94.2	96.8
Oscar [LYL+20]	73.5	92.2	96.0	57.5	82.8	89.8	-	-	-	-	-	-
VinVL [ZLH+21]	75.4	92.9	96.2	58.8	83.5	90.3	-	-	-	-	-	-
Dual encoder + Fusion encoder reranking												
ALBEF [LSG+21]	77.6	94.3	97.2	60.7	84.3	90.5	95.9	99.8	**100.0**	85.6	97.5	98.9
BLIP [LLXH22]	82.4	95.4	97.9	65.1	86.3	91.8	97.4	99.8	99.9	87.6	97.7	99.0
Dual-encoder models												
ALIGN [JYX+21]	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	**100.0**	84.9	97.4	98.6
FILIP [YHH+21]	78.9	94.4	97.4	61.2	84.3	90.6	96.6	**100.0**	**100.0**	87.1	97.7	99.1
Florence [YCC+21]	81.8	95.2	-	63.2	85.7	-	97.2	99.9	-	87.9	98.1	-
BEIT-3	**84.8**	**96.5**	**98.3**	**67.2**	**87.7**	**92.8**	**98.0**	**100.0**	**100.0**	**90.3**	**98.7**	**99.5**

EVA: Exploring the Limits of Masked Visual Representation Learning at Scale

(CVPR23, Beijing Academy of Artificial Intelligence, cite 10, https://github.com/baaivision/EVA)



patch size	#layers	hidden dim	mlp dim	attn heads	#param.
14×14	40	1408	6144	16	1011M

(a) EVA architecture configurations.

dataset	total size
ImageNet-21K, CC12M, CC3M, Object365, COCO, ADE	29.6M images

(b) datasets for pre-training EVA.

image size	batch size	optimizer	peak lr	(β_1, β_2)	pt epochs
224^2	4096	AdamW	1e-3	(0.9, 0.98)	150

(c) some pre-training settings and hyper-parameters.

precision	ZeRO	#gpus	samples / sec.	max mem.	pt days
fp16	stage-1	128	~3150	~26.5GB	~14.5

(d) basic statistics of EVA pre-training.

EVA= (最强**CLIP** 语义学习 + 最强**MIM** 几何结构学习) X 10亿参数 **ViT-g**

- Only open source image data
- Standard ViT-g, very simple!

model	precision	total #param.	image #param.	text #param.	clip training data	samples seen	image size	patch size	batch size	gpus for training
OpenAI CLIP-L	fp16	430M	304M	124M	CLIP-400M [73]	12B	224^2	14×14	32k	256×V100 (32GB)
ALIGN	bfloat16	834M	480M	354M	ALIGN-1.8B [73]	22B	289^2	-	16k	1024×TPUv3
Open CLIP-H	bfloat16	1.0B	632M	354M	LAION-2B [85]	32B	224^2	14×14	79k	824×A100 (40GB)
Open CLIP-g	bfloat16	1.3B	1.0B	354M	LAION-2B [85]	12B	224^2	14×14	64k	800×A100 (40GB)
EVA CLIP-g	fp16	1.1B	1.0B	124M	LAION-400M [86]	11B	224^2	14×14	41k	256×A100 (40GB)

(a) **CLIP model configurations.** EVA CLIP-g can be stably trained via fp16 precision with fewer image-text pairs (7B *v.s.* 12B / 32B) sampled from a smaller data pool (LAION-400M *v.s.* LAION-2B) on ~1/3×GPUs compared with other open-sourced billion-scale competitors.

EVA-CLIP: Improved Training Techniques for CLIP at Scale

(arXiv:2303.15389, Beijing Academy of Artificial Intelligence, cite 8, https://github.com/baaivision/EVA)

	zero-shot **text** retrieval						zero-shot **image** retrieval					
	Flickr30K			COCO			Flickr30K			COCO		
method	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
(a) comparisons with CLIP-**Base** baselines												
OpenAI CLIP-B/16	81.9	96.2	98.8	52.4	76.8	84.7	62.1	85.6	91.8	33.1	58.4	69.0
Open CLIP-B/16	86.3	97.9	99.4	59.4	81.8	88.6	69.8	90.4	94.6	42.3	66.7	77.1
EVA-02-CLIP-B/16	**85.7**	**96.7**	**98.9**	**58.7**	**80.7**	**88.2**	**71.2**	**91.0**	**94.7**	**42.2**	**66.9**	**76.3**
(b) comparisons with CLIP-**Large** baselines												
OpenAI CLIP-L/14	85.2	97.3	99.0	56.3	79.3	86.7	65.2	87.3	92.0	36.5	61.0	71.1
Open CLIP-L/14	88.7	98.4	99.2	62.1	83.4	90.3	75.0	92.5	95.6	46.1	70.7	79.4
EVA-02-CLIP-L/14	89.7	98.6	99.2	63.7	84.3	90.4	77.3	93.6	96.8	47.5	71.2	79.7
(c) comparisons with **larger** CLIPs trained with **more samples**												
OpenAI CLIP-L/14+	87.4	98.3	99.3	57.9	81.2	87.9	67.3	89.0	93.3	37.1	61.6	71.5
Open CLIP-H/14	90.8	99.3	99.7	66.0	86.1	91.9	77.8	94.1	96.6	49.5	73.4	81.5
Open CLIP-g/14	91.4	99.2	99.6	66.4	86.0	91.8	77.7	94.1	96.9	48.8	73.3	81.5
Open CLIP-G/14	92.9	99.3	99.8	67.3	86.9	92.6	**79.5**	**95.0**	**97.1**	**51.4**	74.9	**83.0**
EVA-01-CLIP-g/14	88.3	98.3	99.3	61.8	83.3	90.0	72.6	91.6	95.1	44.1	68.5	77.3
EVA-01-CLIP-g/14+	91.6	99.3	99.8	68.2	87.5	92.5	78.9	94.5	96.9	50.3	74.0	82.1
EVA-02-CLIP-L/14+	89.2	98.9	99.6	64.1	85.2	90.8	77.9	94.2	96.8	47.9	71.7	80.0
EVA-02-CLIP-E/14	92.4	99.3	**99.9**	68.1	87.7	**92.8**	78.8	94.6	97.0	50.8	74.7	82.5
EVA-02-CLIP-E/14+	**93.9**	**99.4**	99.8	**68.8**	**87.8**	**92.8**	78.8	94.2	96.8	51.1	**75.0**	82.7

Table 4: **Summary of zero-shot retrival performance on Flickr30K [53] and COCO [34]**

- CLIP->EVA
- EVA->EVA CLIP
- EVA CLIP -> EVA 2
- EVA 2 ->EVA 2 CLIP

Contents

Conclusion

- Transformer is the cornerstone of multimodal pre-training. Text and images are unified.
- This report introduces the upper bound of multimodal pre-training. Bigger Models, More Data ---- Scaling up !
- Bottleneck: computing power + data quality
- Multimodal-adapter

Contents

Future trend – GPT 3.5

- ✓ More modal information interaction - image, video, text, audio...
- ✓ **Efficient** multimodal pre-training framework.
- ✓ Conversational LLM can inspire multimodal pre-training: KOSMOS, MiniGPT-4, VisionLLM
  User feedback and logical reasoning ability? **In-context Few-shot Learning**?
- ✓ **Unified** interface, unified task form and the whole is unified. all modalities are placed in a unified generation model. “what I cannot create, I do not understand”- Richard Feynman. Chatgpt + Midjourney.
- ✓ Improve model interpretability.
  - Does multimodal model use visual knowledge? Which part of the knowledge of the image will the model use?
  - The existing LLM does not open the model, how to dynamically correct the wrong knowledge?
  - Confidence in model predictions, adding relevant citations/sources? Result traceability/inspection.
  - **Chain-of-thoughts** is very important, telling the human solution process.
- ✓ Visual reasoning benchmarks that are closer to the real world. How to comprehensively evaluate the performance and value of multimodal pre-training?