Multimodal Machine Learning

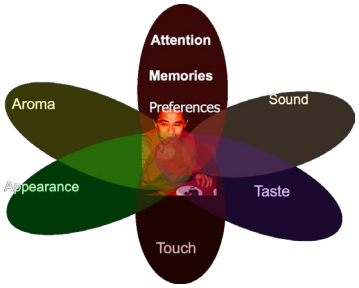
Aimer

20220812

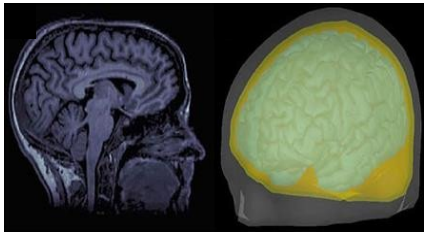
提　纲

一 多模态机器学习现状

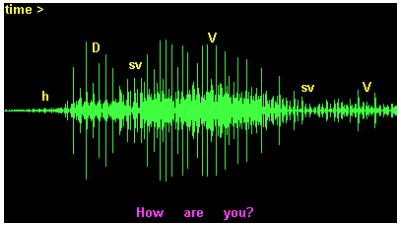
多元社区和多模态（Multiple Communities and Modalities）



Psychology

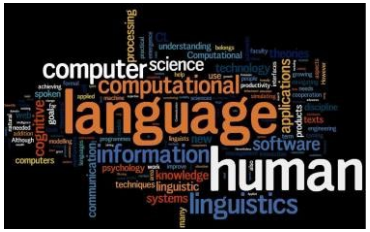


Medical



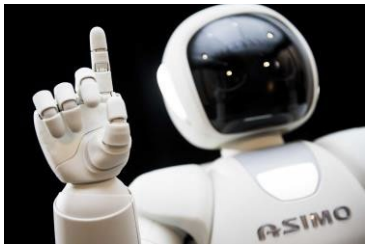
Speech

Vision

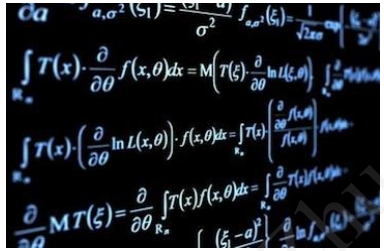


Language

Multimedia



Robotics



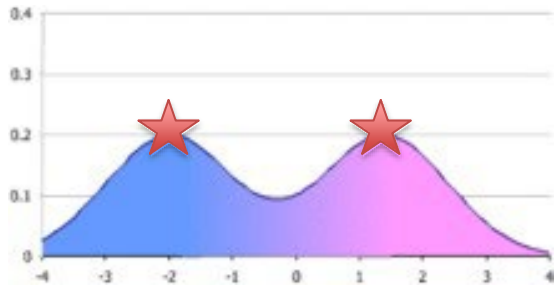
Learning

一 多模态机器学习现状

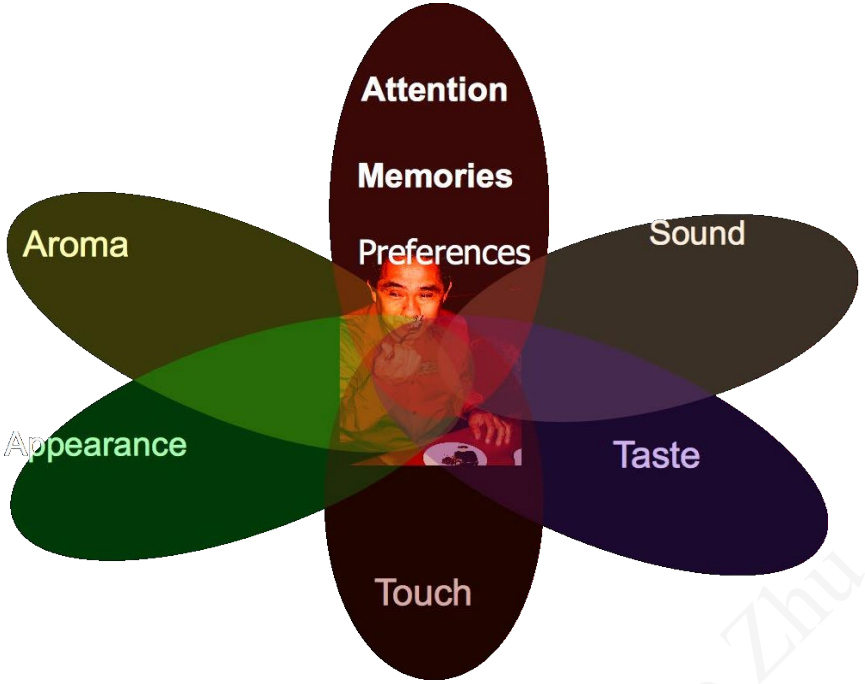
Multimodal的定义：

Modality（模态）

一种做某事或体验某事的特定方式。它指的是某种类型的信息和/或存储信息的表示格式。



Multimodal distribution



Sensory Modalities
（感官模式）

- 多模式，即概率密度函数中明显的 "峰值"（局部最大值）

一 多模态机器学习现状

模态（Modality）

某事发生或经历的方式

- 模式指的是某种类型的信息和/或 储存信息的表现形；
- 感觉方式：感觉的主要形式之一，如视觉或触觉；交流的渠道。

媒介（ Medium ）

储存或传递信息的手段或工具；通信/传输系统

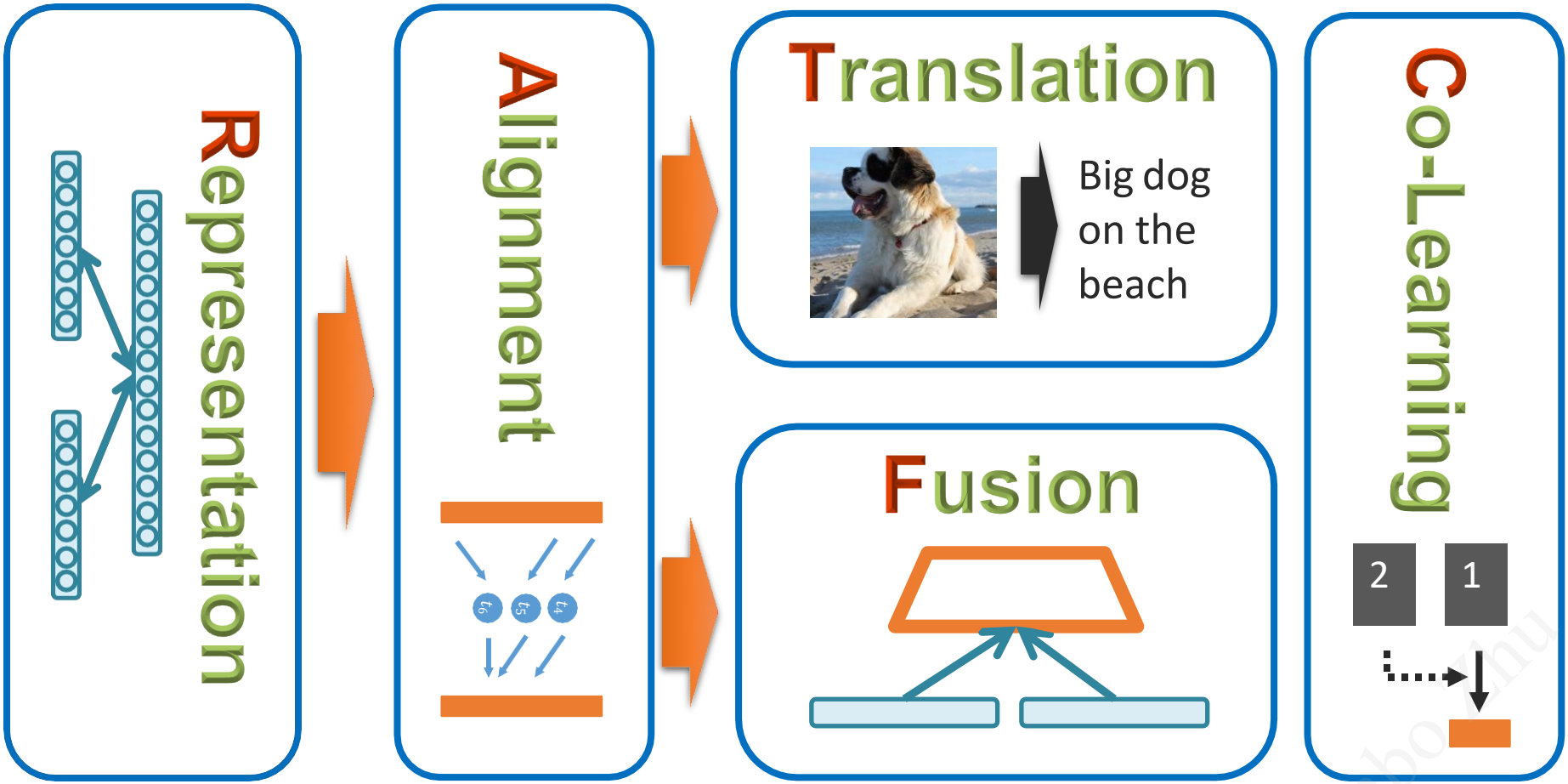
- 媒介是将这种信息传递给解释者的感官的手段。

各种模态例子：

- 自然语言（包括口语或书面）
- 视觉（来自图像或视频）
- 听觉（包括语音、声音和音乐）
- 触觉/触摸
- 嗅觉、味觉和自我运动
- 生理信号
- 心电图（ECG）、皮肤电导率
- 其他模态：
- 红外线图像、深度图像、fMRI

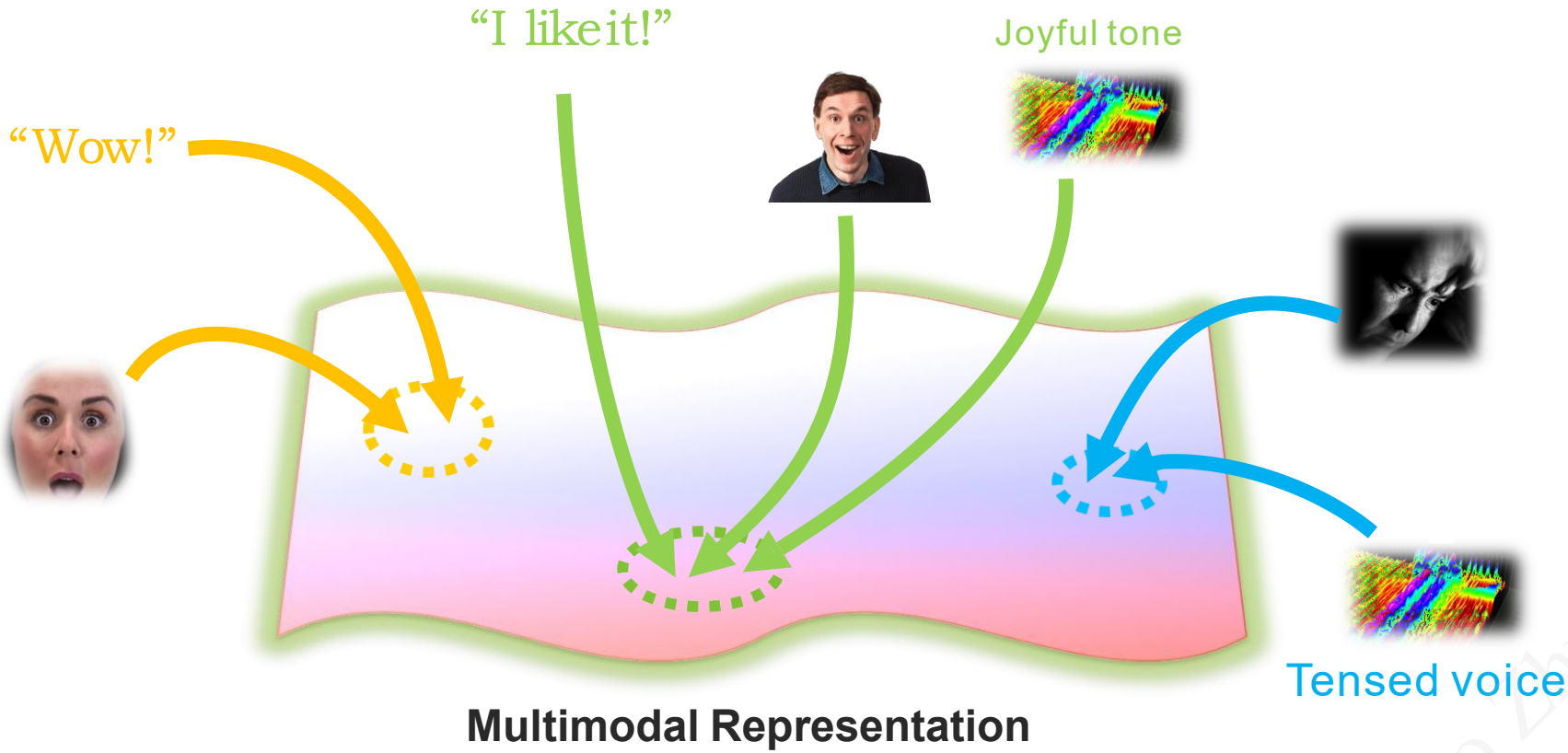
一 多模态机器学习现状

多模态机器学习的五大类（挑战）：



Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: A survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(2): 423-443.

一 多模态机器学习现状



一 多模态机器学习现状

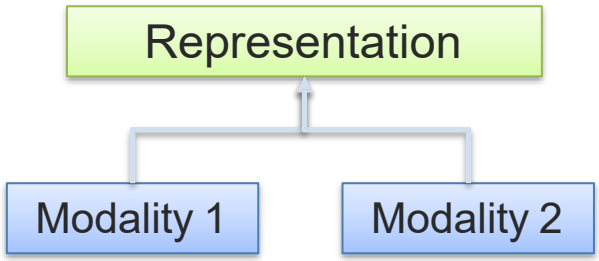
第一类多模态表示（Representation）

Definition: Learning how to represent and summarize multimodal data in away that exploits the complementarity and redundancy.

学习如何表示和总结多模态数据，以利用其互补性和冗余性。

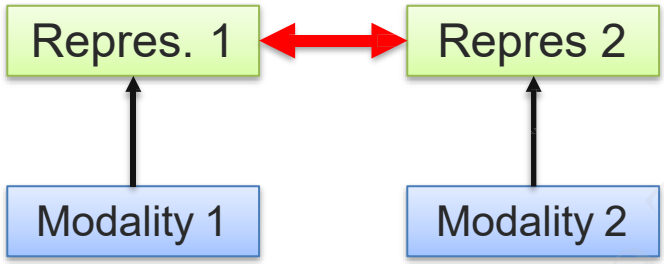
联合表示

A Joint representations:

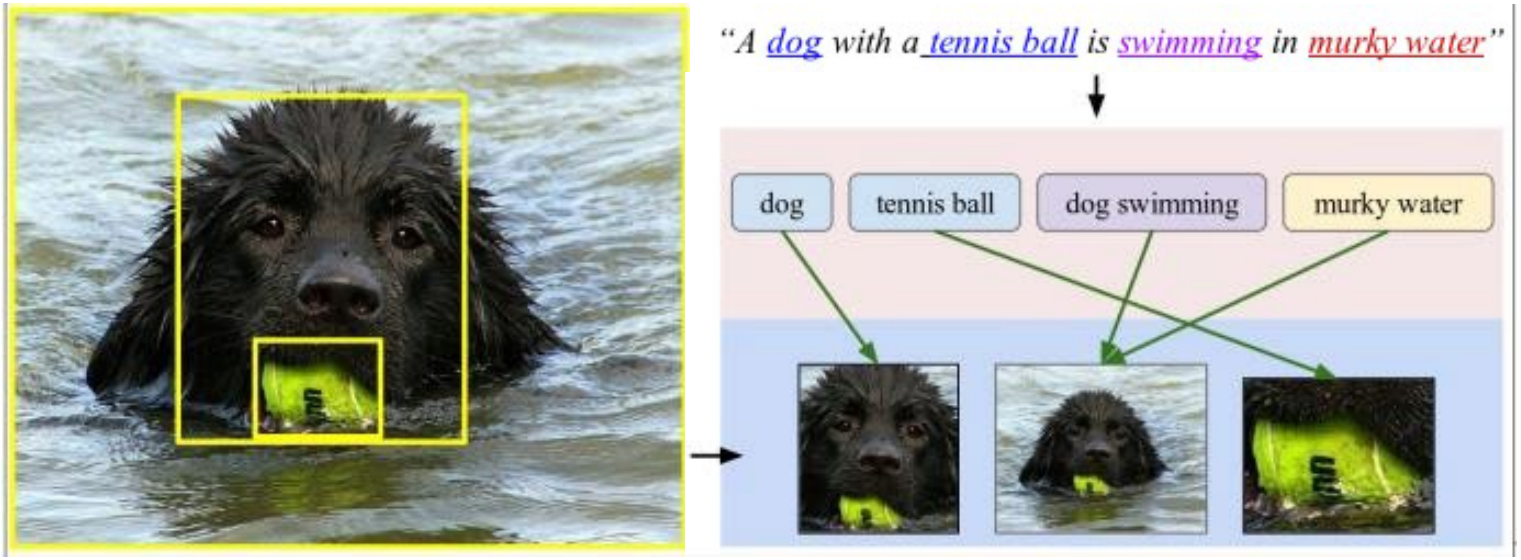


协同表示

B Coordinated representations:



一 多模态机器学习现状



Multimodal Alignment

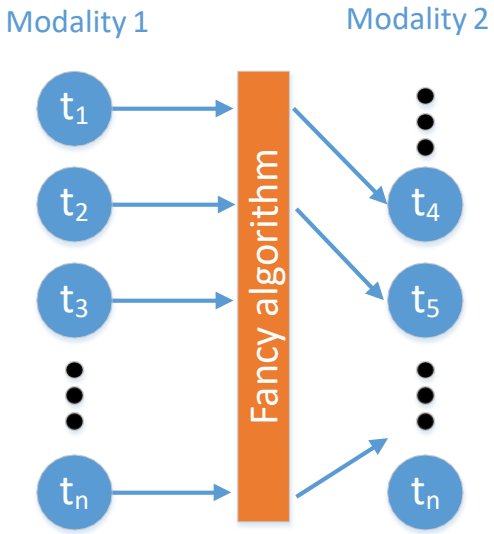
Karpathy A, Joulin A, Fei-Fei L F. Deep fragment embeddings for bidirectional image sentence mapping[J]. Advances in Neural Information Processing Systems, 2014, 27.

一 多模态机器学习现状

第二类多模态对齐（Alignment）

Definition: Identify the direct relations between (sub)elements from two or more different modalities.

识别来自两个或多个不同模式的（子）元素之间的直接关系。



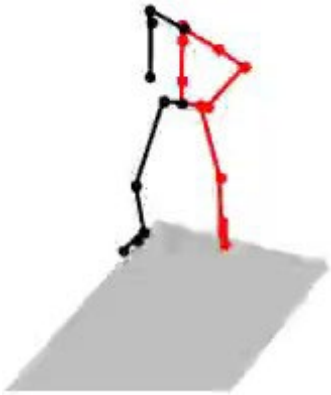
A **Explicit Alignment** 显示对齐

The goal is to directly find correspondences between elements of different modalities

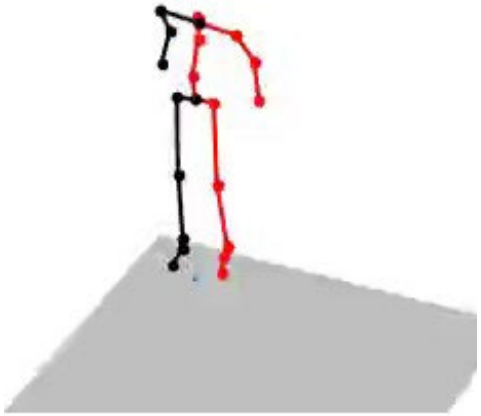
B **Implicit Alignment** 隐示对齐

Uses internally latent alignment of modalities in order to better solve a different problem

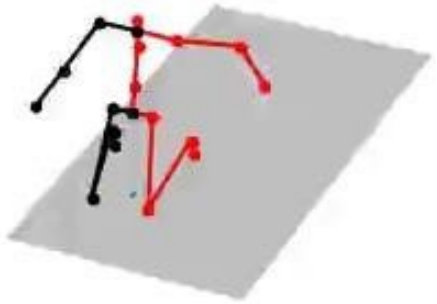
一 多模态机器学习现状



a person jogs a few steps



A person steps forward then turns around and steps forwards again.



A kneeling person raises their arms to the sides and stand up.

Multimodal Translation

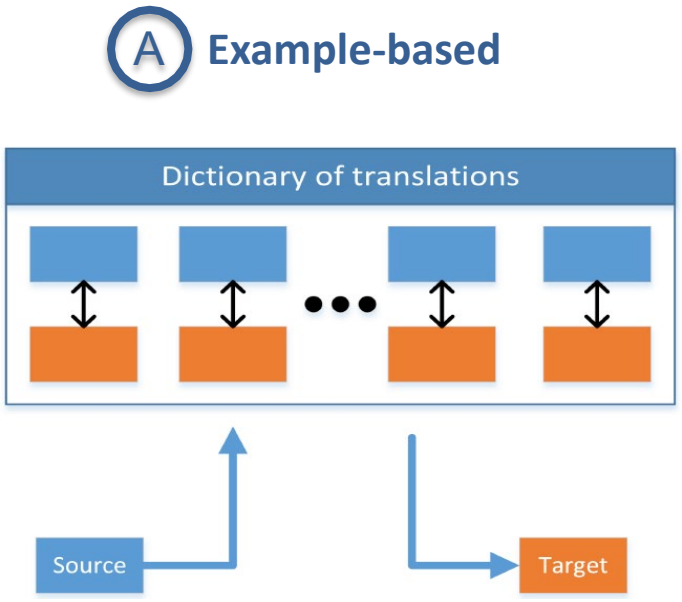
Ahuja C, Morency L P. Language2pose: Natural language grounded pose forecasting[C]//2019 International Conference on 3D Vision (3DV). IEEE, 2019: 719-728.

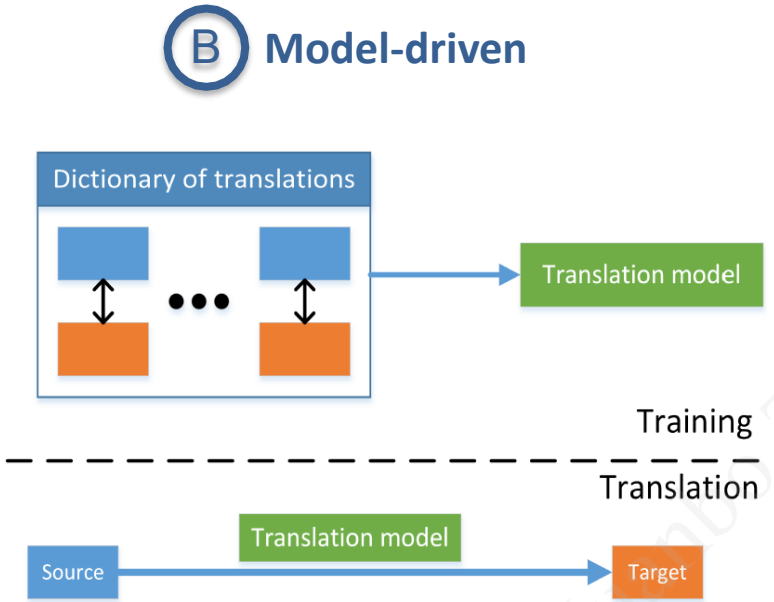
一 多模态机器学习现状

第三类多模态转换（Translation）

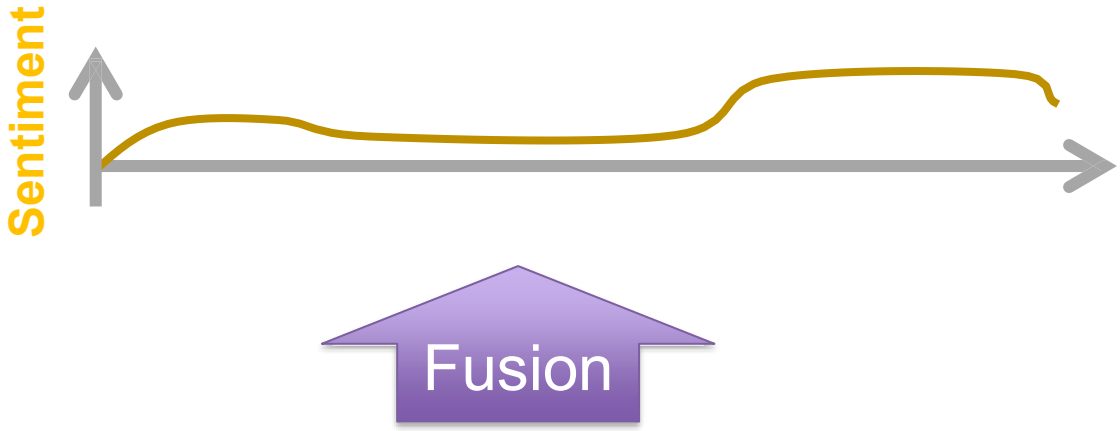
Definition: Process of changing data from one modality to another, where the translation relationship can often be open-ended or subjective.

将数据从一种模式转变为另一种模式的过程，其中的翻译关系往往可以是开放式的或主观的。





一 多模态机器学习现状



Multimodal Fusion

一 多模态机器学习现状

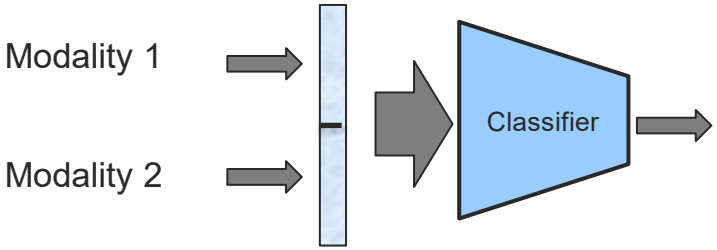
第四类多模态融合（Fusion）

Definition: To join information from two or more modalities to perform a prediction task.

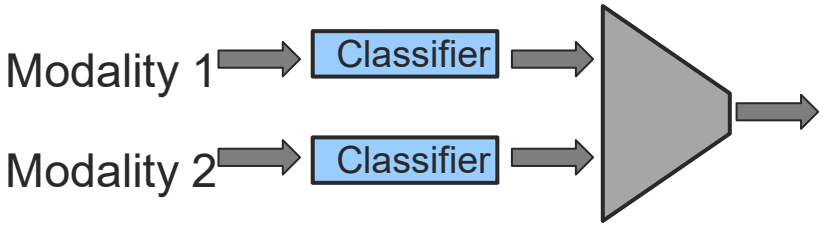
将两种或多种模式的信息结合起来，以执行预测任务。

A Model-Agnostic Approaches

1) Early Fusion

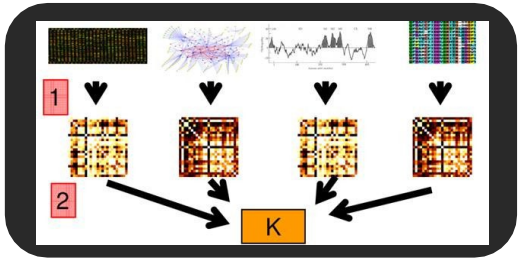


2) Late Fusion

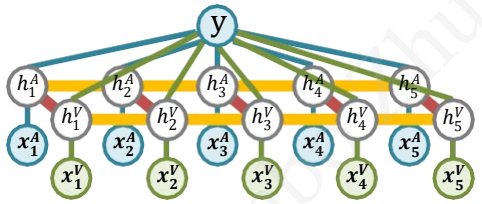


B Model-Based (Intermediate) Approaches

1) Deep neural networks
2) Kernel-based methods
3) Graphical models

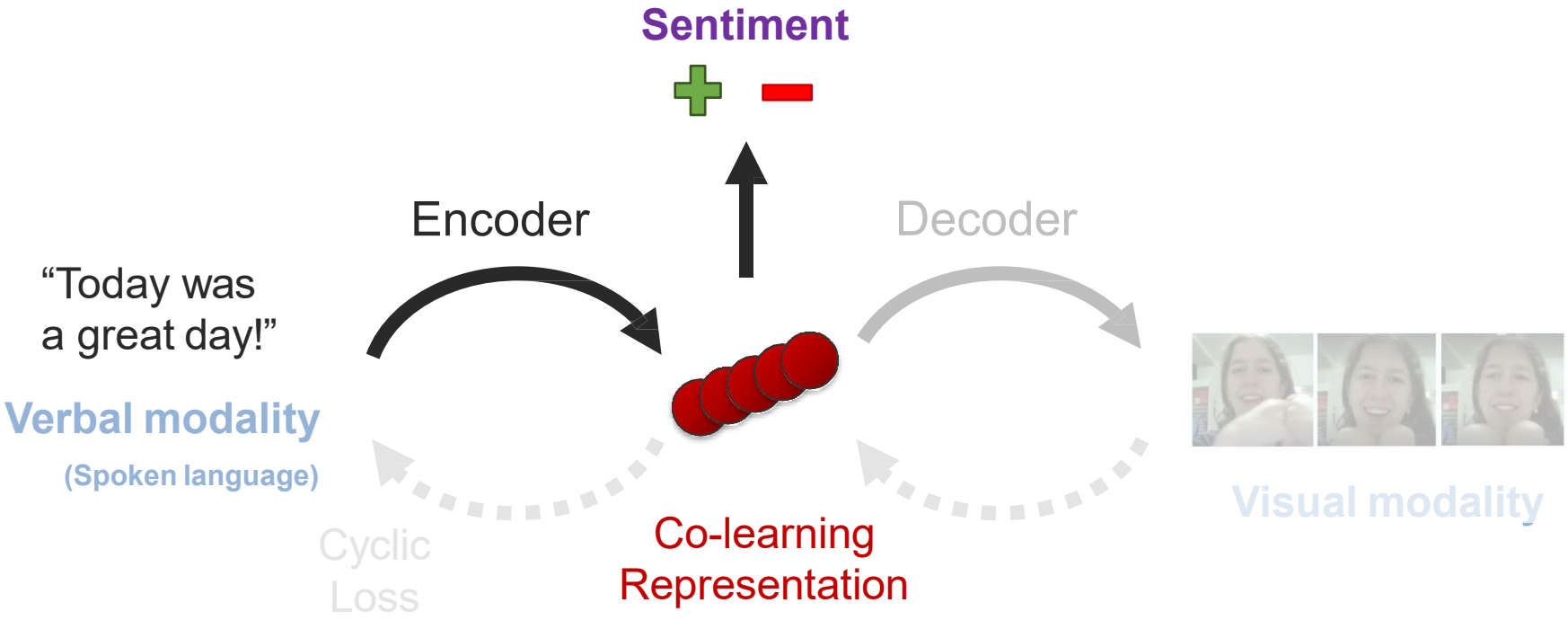


Multiple kernel learning



Multi-View Hidden CRF

一 多模态机器学习现状



Multimodal Co-Learning

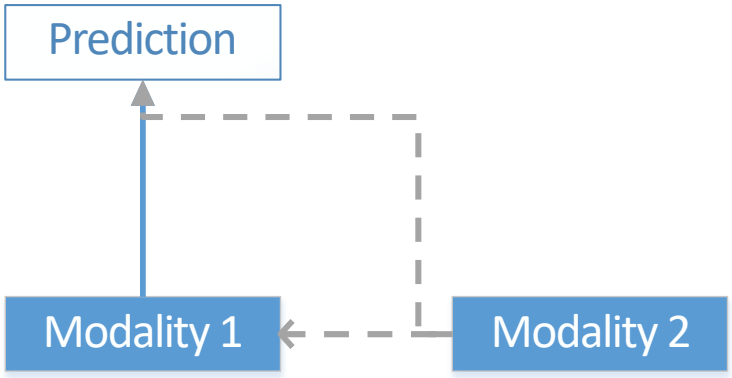
Pham H, Liang P P, Manzini T, et al. Found in translation: Learning robust joint representations by cyclic translations between modalities[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 6892-6899..

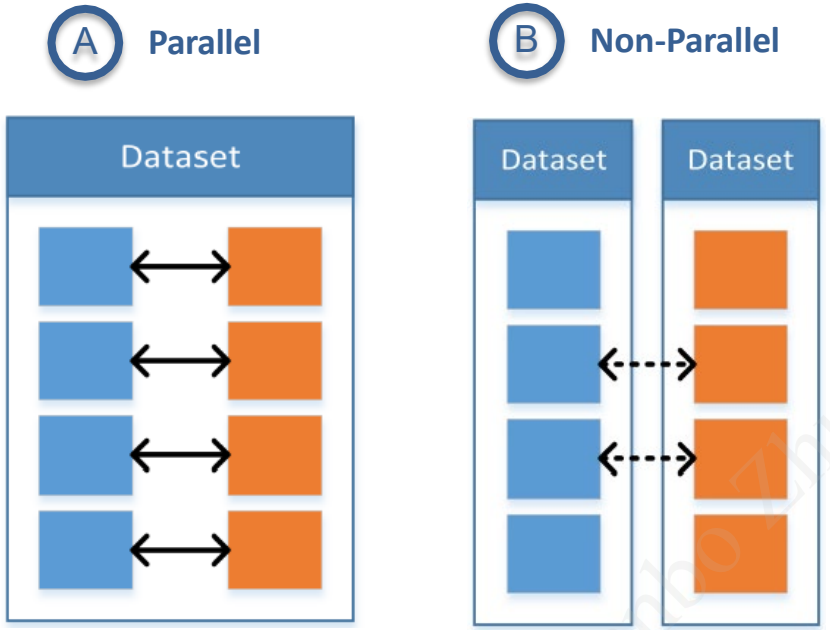
一 多模态机器学习现状

第五类多模态协同学习（Co-Learning）

Definition: Transfer knowledge between modalities, including their representations and predictive models.

在不同的模式之间转移知识，包括其代表和预测模型。





一 多模态机器学习现状

多模态研究的分类法例子

Representation

- Joint
  - *Neural networks*
  - *Graphical models*
  - *Sequential*
- Coordinated
  - *Similarity*
  - *Structured*

Translation

- Example-based
  - *Retrieval*
  - *Combination*
- Model-based
  - *Grammar-based*
  - *Encoder-decoder*
  - *Online prediction*

Alignment

- Explicit
  - *Unsupervised*
  - *Supervised*
- Implicit
  - *Graphical models*
  - *Neural networks*

Fusion

- Model agnostic
  - *Early fusion*
  - *Late fusion*
  - *Hybrid fusion*
- Model-based
  - *Kernel-based*
  - *Graphical models*
  - *Neural networks*

Co-learning

- Parallel data
  - *Co-training*
  - *Transfer learning*
- Non-parallel data
  - *Zero-shot learning*
  - *Concept grounding*
  - *Transfer learning*
- *Hybrid data*
  - *Bridging*

一 多模态机器学习现状

多模态研究的任务

- 情感识别(Affect recognition)
  - 情感(Emotion)
  - 劝服(Persuasion)
  - 性格特征(Personality traits)
- 媒体描述 (Media description)
  - 图片说明(Image captioning)
  - 视频说明(Video captioning)
  - VQA(Visual Question Answering)
- 事件识别(Event recognition)
  - 行动识别(Action recognition)
  - 分割(Segmentation)
- 多媒体信息检索

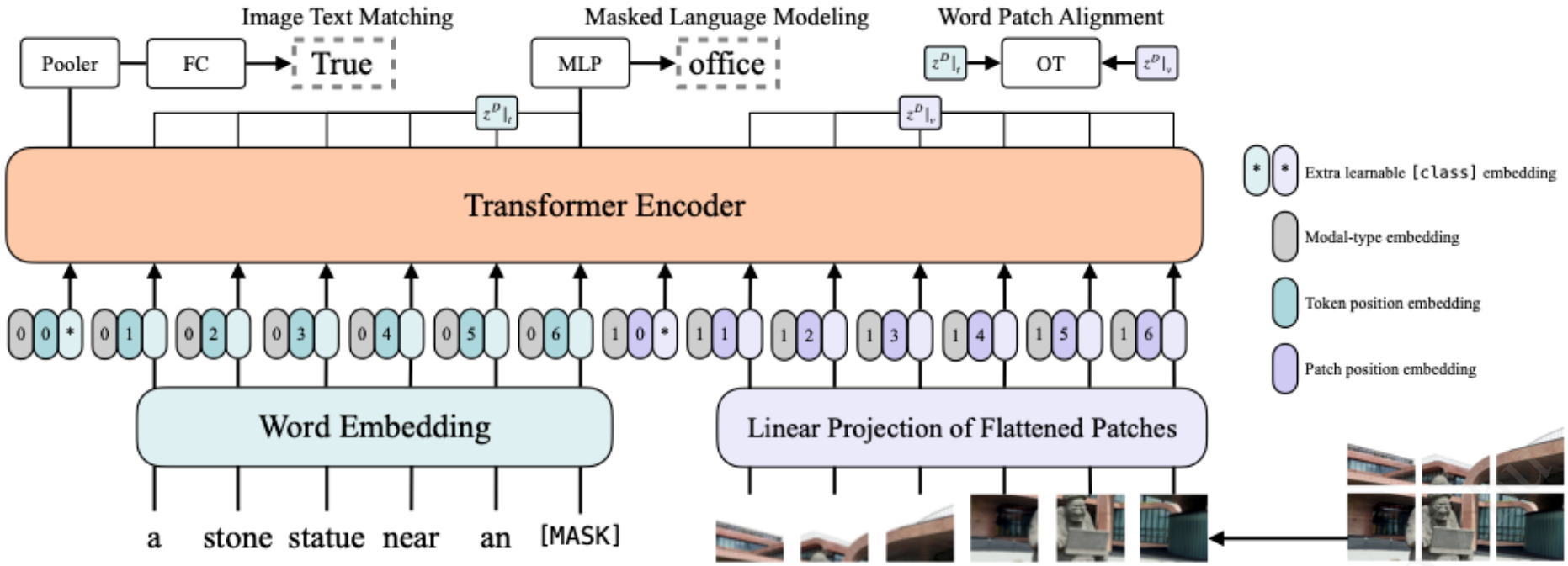
  Multimedia information retrieval
  - 基于内容的/跨媒体的

    Content based/Cross-media

二 MMML论文例子

ViLT：Vision-and-Language Transformer

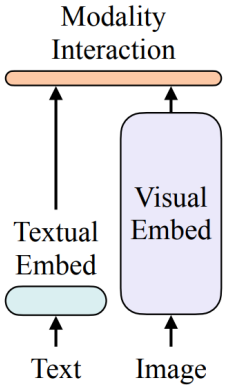
-----最简单的多模态Transformer



Kim W, Son B, Kim I. Vilt: Vision-and-language transformer without convolution or region supervision[C]//International Conference on Machine Learning. PMLR, 2021: 5583-5594.

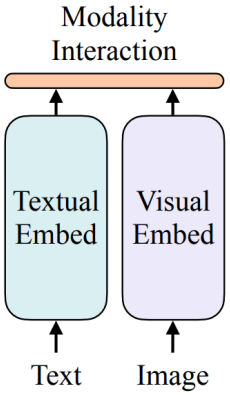
二 MMML论文例子

视觉语言模型的四种结构类别：



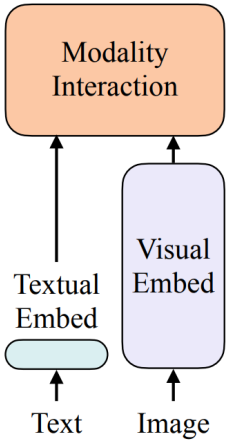
(a) VE > TE > MI

视觉语义嵌入(VSE)模型，如VSE++、SCAN



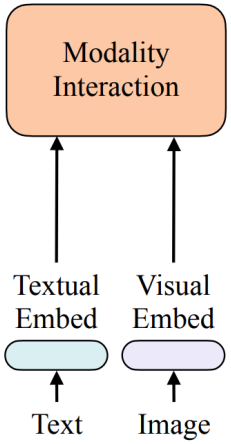
(b) VE = TE > MI

CLIP（Contrastive Language-Image Pre-training）模型



(c) VE > MI > TE

视觉语言预训练VLP（Vision-Language Pre-training）

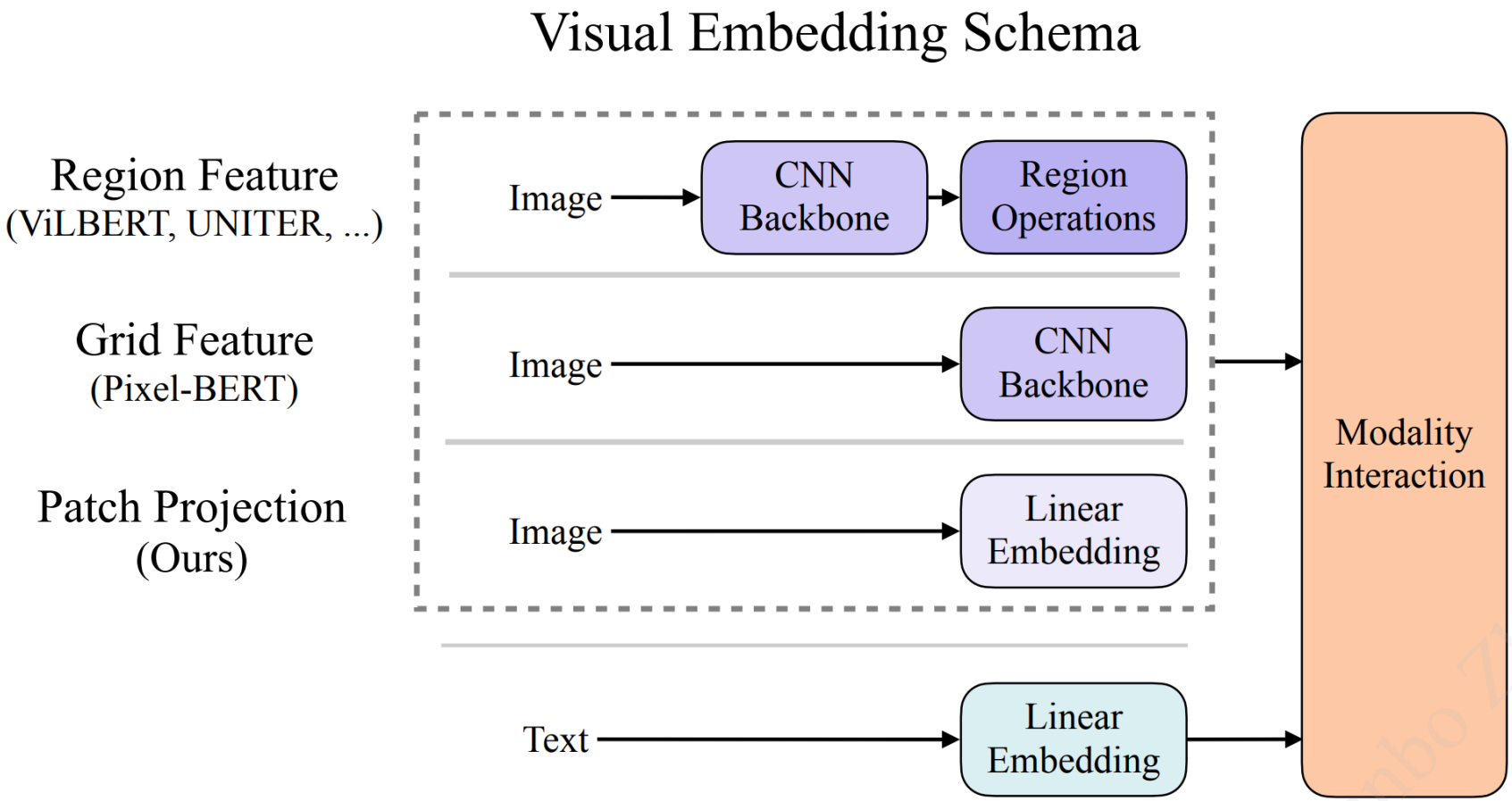


(d) MI > VE = TE

Vision-and-Language Transformer

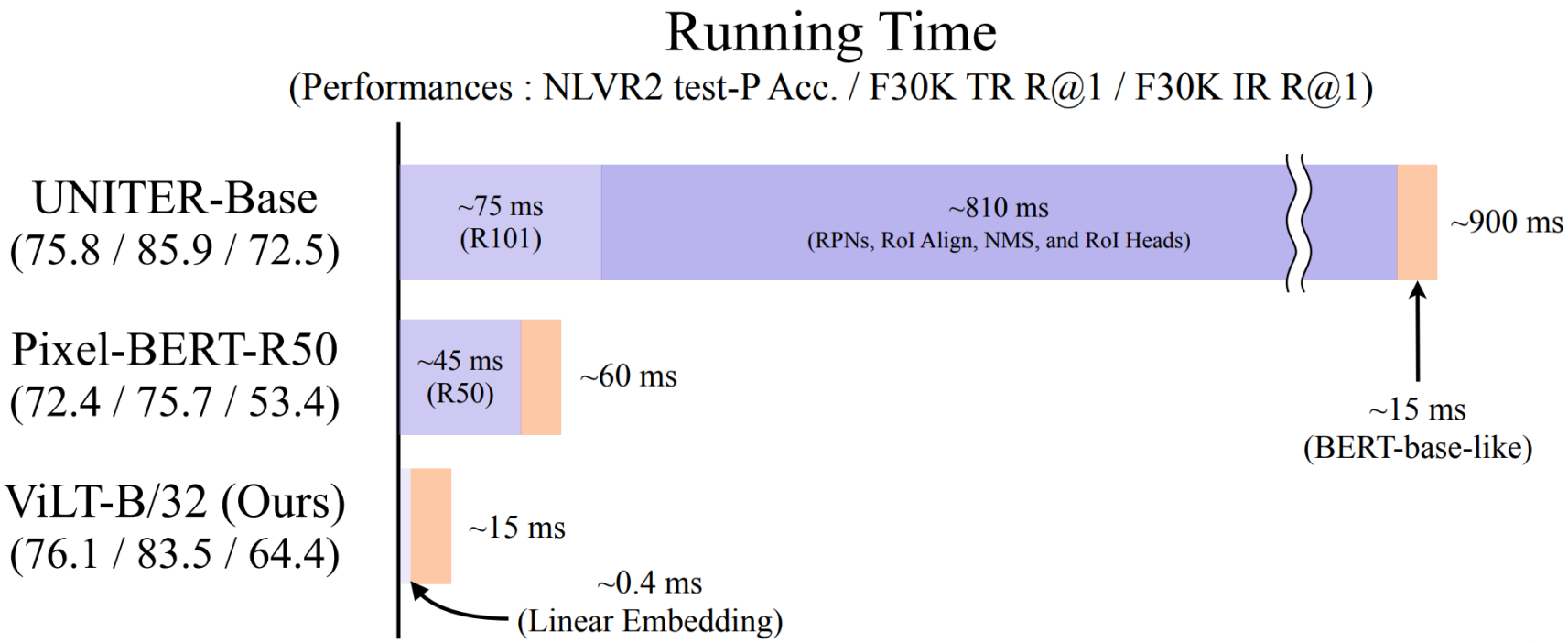
二 MMML论文例子

ViLT与其他主流做法的对比：



二 MMML论文例子

ViLT与其他主流做法的对比:



在效率上大大提升且表现出相似的性能，相比于region feature的方法速度快了60倍，相比于grid feature的方法快了4倍，而且下游任务表现出相似甚至更好的性能。

二 MMML论文例子

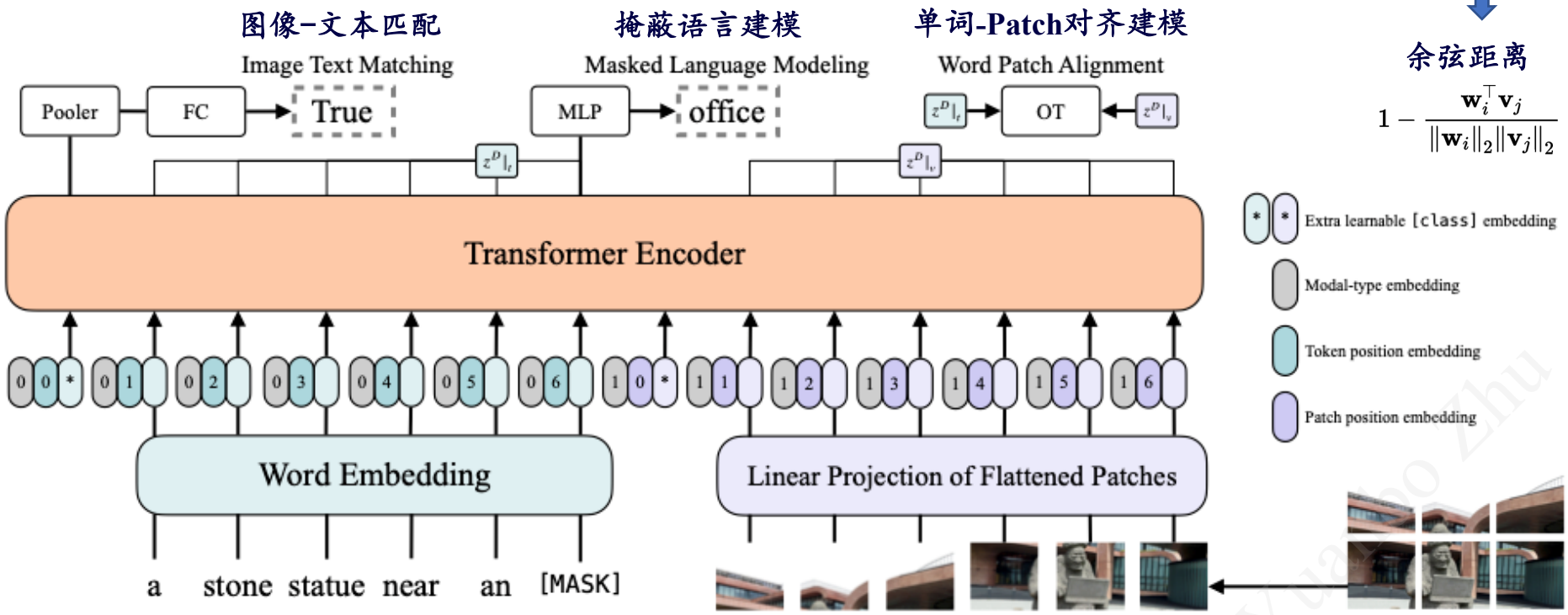
ViLT模型结构：

二分类交叉熵损失 $\mathcal{L}_{\mathrm{ITM}}(\theta) = -E_{(\mathbf{w},\mathbf{v})\sim D}\left[y \log s_\theta(\mathbf{v},\mathbf{w}) + (1-y)\log(1-s_\theta(\mathbf{v},\mathbf{w}))\right]$

BERT MLM损失 $\mathcal{L}_{\mathrm{MLM}}(\theta) = -\mathbb{E}_{(\mathbf{w},\mathbf{v})\sim D} \log P_\theta(\mathbf{w_m} \mid \mathbf{w}\setminus\mathbf{m}, \mathbf{v})$ 负对数似然(Negative log-likelihood, NLL)

改UNITER对齐损失 $\mathcal{L}_{\mathrm{WPA}}(\theta) = \mathcal{D}_{ot}(\boldsymbol{\mu},\boldsymbol{\nu}) = \min_{\mathbf{T}\in\Pi(\mathbf{a},\mathbf{b})} \sum_{i=1}^{T}\sum_{j=1}^{K} \mathbf{T}_{ij}\cdot c(\mathbf{w}_i,\mathbf{v}_j)$ $\boldsymbol{\mu} = \sum_{i=1}^{T} \mathbf{a}_i \delta_{\mathbf{w}_i}$ $\boldsymbol{\nu} = \sum_{j=1}^{K} \mathbf{b}_j \delta_{\mathbf{v}_j}$ $c(\mathbf{w}_i,\mathbf{v}_j)$

余弦距离 $1 - \frac{\mathbf{w}_i^\top \mathbf{v}_j}{\|\mathbf{w}_i\|_2 \|\mathbf{v}_j\|_2}$



二 MMML论文例子

ViLT数据处理方法

Whole Word Masking

[Original Sentence]
使用语言模型来预测下一个词的probability。
[Original Sentence with CWS]
使用 语言 **模型** 来 **预测** 下 一个 词 的 **probability** 。

probability这个词被切分成”pro”、”#babi”和”#lity”3个WordPiece

[Original BERT Input]
使 用 语 言 **[MASK]** 型 来 **[MASK]** 测 下 一 个 词 的 **pro [MASK] ##lity** 。
[Whold Word Masking Input]
使 用 语 言 **[MASK] [MASK]** 来 **[MASK] [MASK]** 下 一 个 词 的 **[MASK] [MASK] [MASK]** 。

Image Augmentation 魔改版RandAugment

颜色反转，因为文本通常也包含颜色信息；

```
transforms = [
"Identity", "AutoContrast", "Equalize", "Rotate", "Solarize",
"Color", "Posterize", "Contrast", "Brightness", "Sharpness",
"ShearX", "ShearY", "TranslateX", "TranslateY"]
def randaugment(N, M):
sampled_ops = np.random.choice(transforms, N)
return [(op, M) for op in sampled_ops]
```

N：将被选择出的操作数，M：失真级数

以及切除，因为它可能会清除分散在整个图像中的小而重要的物体

二 MMML论文例子

实验结果

Visual Embed	Model	Time (ms)	VQAv2 test-dev	NLVR2 dev	NLVR2 test-P
Region	w/o VLP SOTA	~900	70.63	54.80	53.50
	ViLBERT	~920	70.55	-	-
	VisualBERT	~925	70.80	67.40	67.00
	LXMERT	~900	72.42	74.90	74.50
	UNITER-Base	~900	72.70	75.85	75.80
	OSCAR-Base†	~900	73.16	78.07	78.36
	VinVL-Base†‡	~650	75.95	82.05	83.08
Grid	Pixel-BERT-X152	~160	74.45	76.50	77.20
	Pixel-BERT-R50	~60	71.35	71.70	72.40
Linear	ViLT-B/32	~15	70.33	74.41	74.57
	ViLT-B/32ⓐ	~15	70.85	74.91	75.57
	ViLT-B/32ⓐ⊕	~15	71.26	75.70	76.13

- Classification Tasks Visual Question Answering
- Natural Language for Visual Reasoning

Visual Embed	Model	#Params (M)	#FLOPs (G)	Time (ms)
Region	ViLBERT^{36+36}	274.3	958.1	~900
	VisualBERT^{36+128}	170.3	425.0	~925
	LXMERT^{36+20}	239.8	952.0	~900
	UNITER-Base^{36+60}	154.7	949.9	~900
	OSCAR-Base^{50+35}	154.7	956.4	~900
	VinVL-Base^{50+35}	157.3	1023.3	~650
	Unicoder-VL$^{100+?}$	170.3	419.7	~925
	ImageBERT^{100+44}	170.3	420.6	~925
Grid	Pixel-BERT-X152$^{146+?}$	144.3	185.8	~160
	Pixel-BERT-R50$^{260+?}$	94.9	136.8	~60
Linear	ViLT-B/32^{200+40}	87.4	55.9	~15

参数和处理速度对比

Visual Embed	Model	CNN Backbone	RoI Head	NMS	Trans. Layers
Region	ViLBERT	R101	C4	PC	~15
	VisualBERT	X152	FPN	PC	12
	LXMERT	R101	C4	PC	~12
	UNITER-Base	R101	C4	PC	12
	OSCAR-Base	R101	C4	PC	12
	VinVL-Base	X152	C4	CA	12
	Unicoder-VL	X152	FPN	PC	12
	ImageBERT	X152	FPN	PC	12
Grid	Pixel-BERT-X152	X152	-	-	12
	Pixel-BERT-R50	R50	-	-	12
Linear	ViLT-B/32	-	-	-	12

VLP模型的组成部分

二 MMML论文例子

实验结果

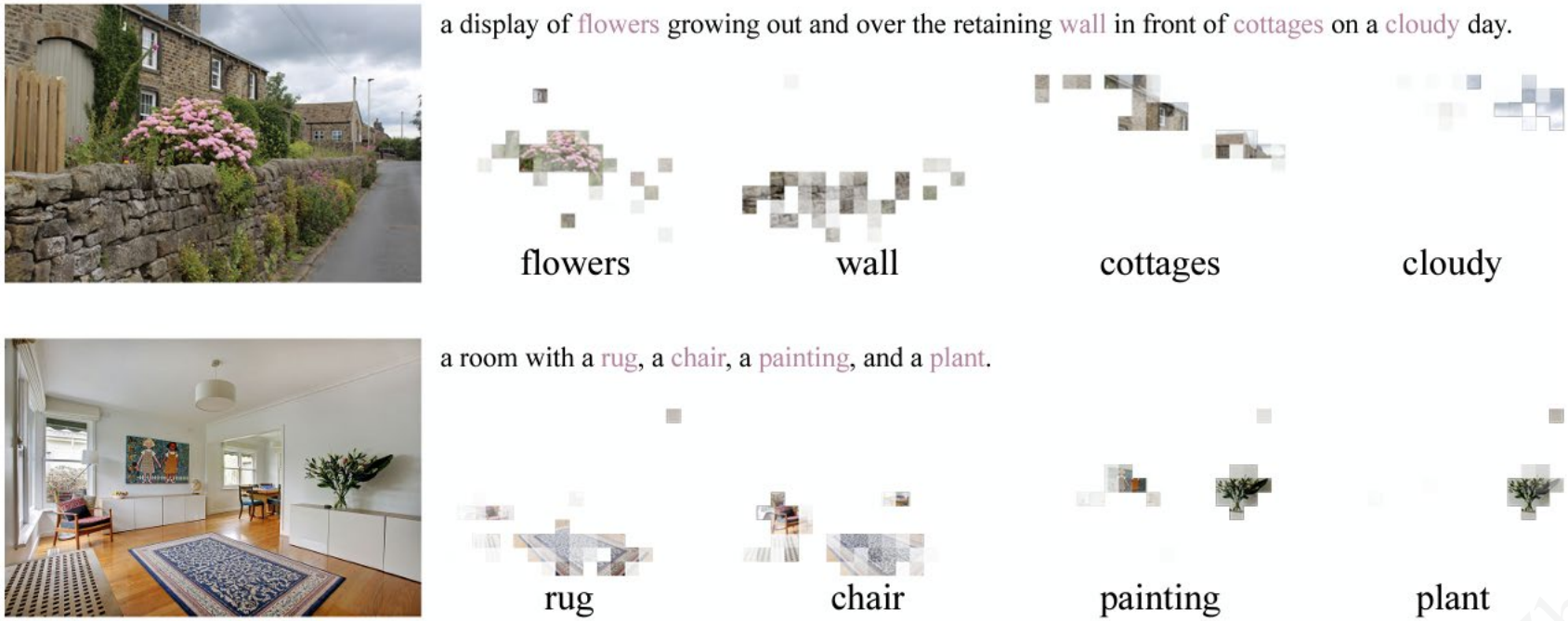
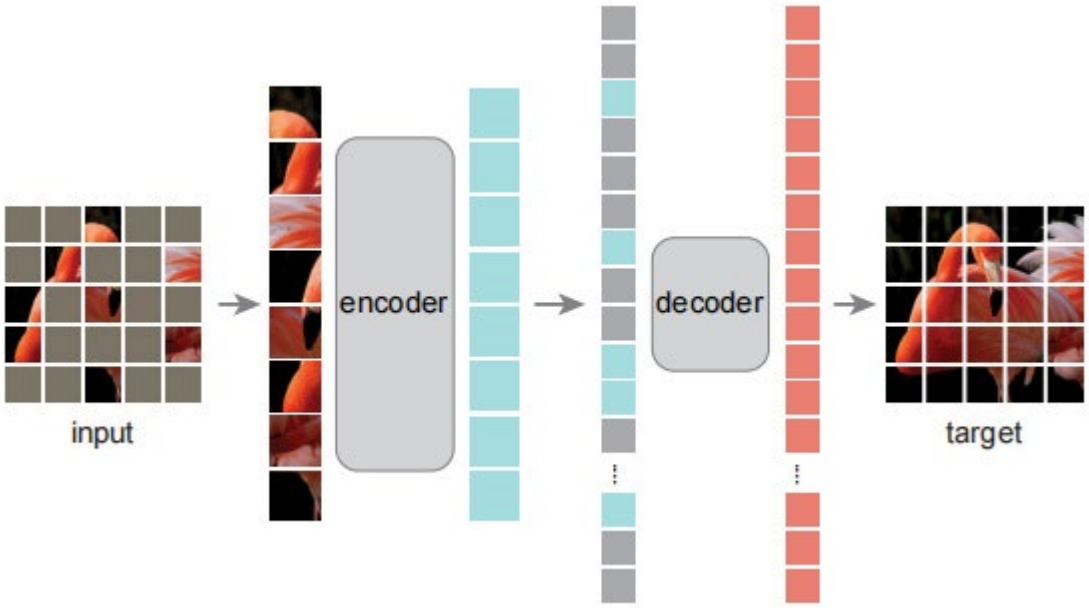


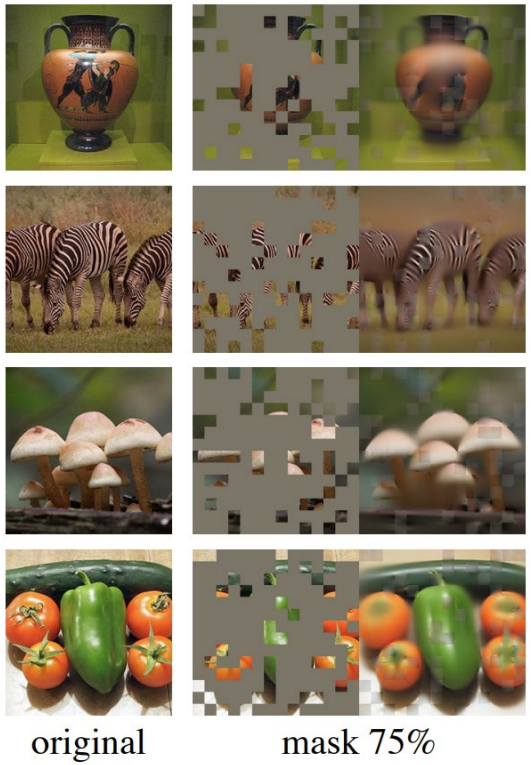
Figure 4. Visualizations of transportation plan of word patch alignment. Best viewed zoomed in.

三 未来研究展望

ViLT相关的后续研究



Masked AutoEncoders(MAE)

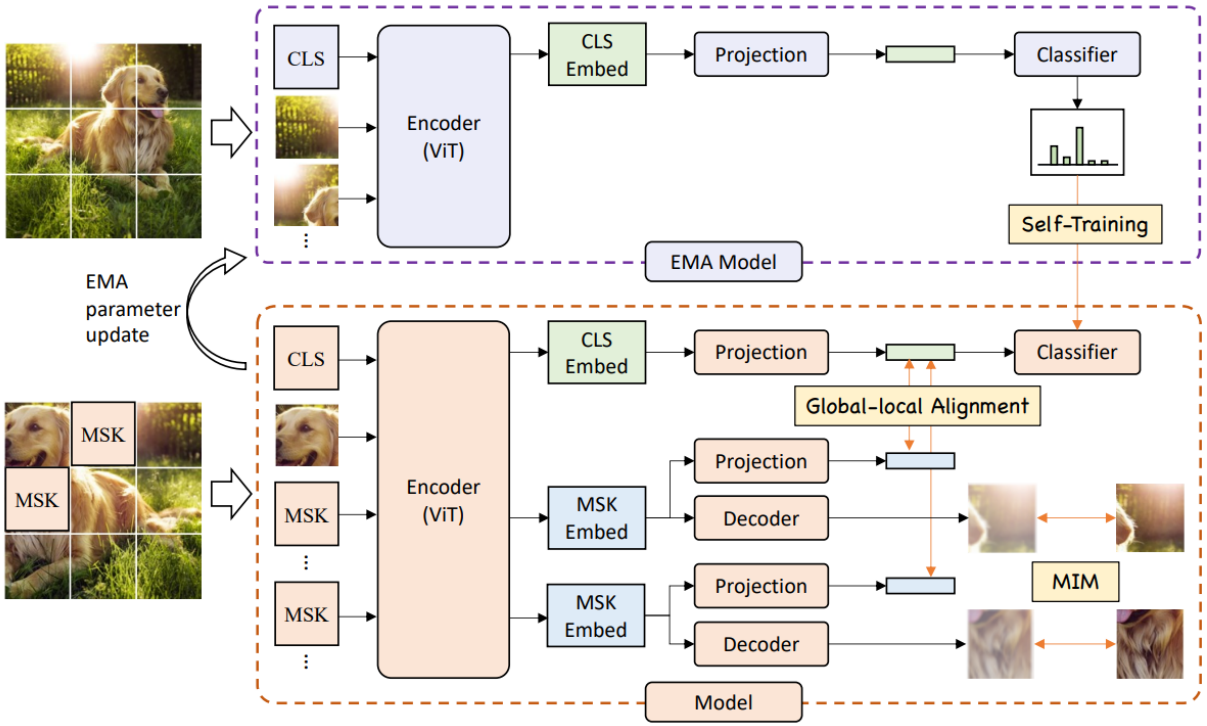


MAE重建验证图像

He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 16000-16009.

三 未来研究展望

ViLT相关的后续研究



Masked Unsupervised Self-Training (MUST)

Li J, Savarese S, Hoi S C H. Masked Unsupervised Self-training for Zero-shot Image Classification[J]. arXiv preprint arXiv:2206.02967, 2022.

Merci！ 谢谢！