目录

概述总括

技术奇点

多模态大模型的技术奇点在于：
1、BERT等模型证明了Trans former在NLP领域性能好，并且对于数据量、模型大小而言未见上限；
2、ViT 模型将Trans former模型迁移到CV领域，让Transformer 能够处理图像；
3、BEiT 将生成式预训练从NLP迁移到CV， 图像大规模自监督学习成为可能。
4、扩散模型与多模态大模型结合，推动文生图领域发展。

应用催化

各式多模态场景下的应用持续推动多模态模型的演进

- **Transformer**
  - 处理图像
  - Vision Transformer
  - 生成式预训练
  - BEiT
- 扩散模型
- 多模态检索
- 虚拟人/智能助理

1、以BERT 为主的Transformer 模型取得很好的效果，但是 仅限于文本领域；
2、Transformer 中自注意力机制和前向传播网络权重共享适合于多模态模型。

1、将图片patch 化，解决了Transformer 不能应用于图像领域问题；
2、patch embedding 提取图像特征高效；
3、基于ViT模型衍生了视频Transformer 相关模型。

1、将生成式预训练MLM方法从NLP迁移至CV， 实现CV大规模自监督预训练；
2、统一多模态大模型BEiT -3前身。

与CLIP结合，衍生多个文图生成模型，文图生成领域火爆

智能家居

机器人技术

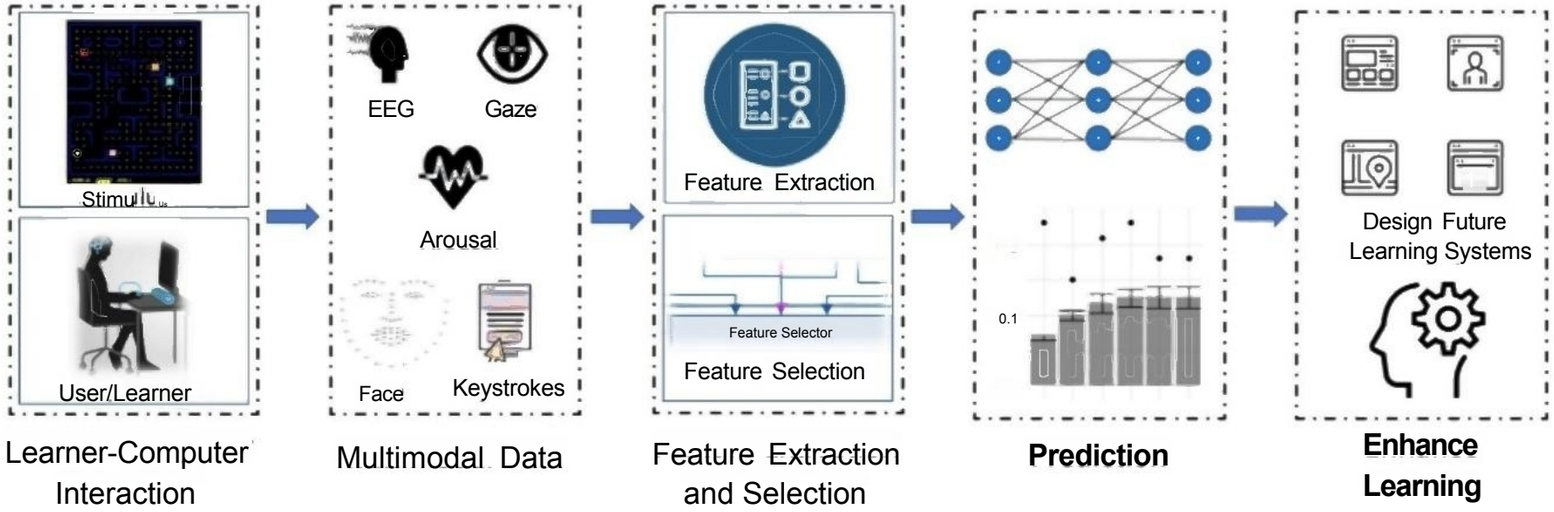
机器翻译

……

1.1 多模态表示包含两个或两个以上事物表现形式

- 模态是事物的一种表现形式，多模态通常包含两个或者两个以上的模态形式，是从多个视角出发对事物进行描述。生活中常见多模态表示，例如传感器的数据不仅仅包含文字、图像，还可以包括与之匹配的温度、深度信息等。
- 使用多模态数据能够使得事物呈现更加立体、全面，多模态研究成为当前研究重要方面，在情感分析、机器翻译、自然语言处理和生物医药前沿方向取得重大突破。

图表：利用多模数据能有助于学习



资料来源：Multimodal data as a means to understand the learning experience

1.2 多模态发展主要经历五个时代

图表：多模态模型发展的五个阶段

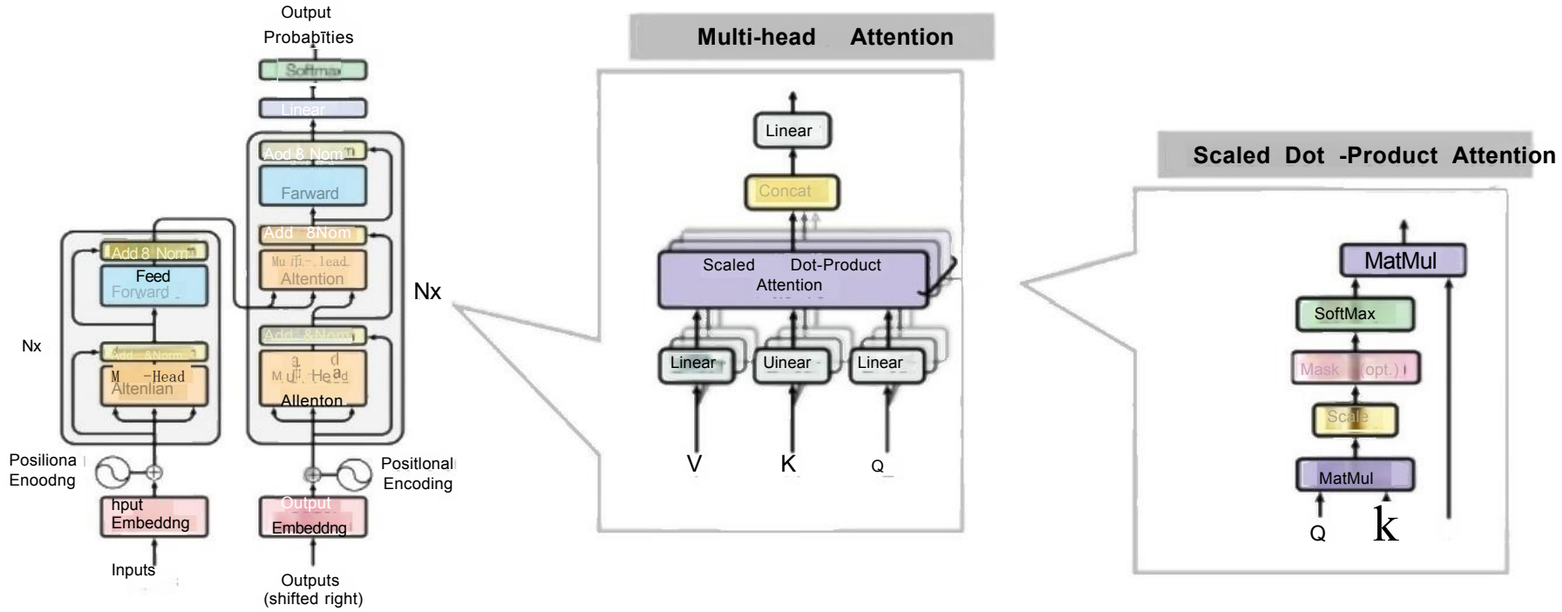
1970s-1980s末 行为时代	1980s末-2000 计算时代	2000-2010 交互时代	2010-2020 深度学习时代	2020至今 大模型开启新时代
1973 多模式行为疗法(Arnold Lazarus) 人格的七个维度 1980 多模态信号检测：独立决策与整合 1983 婴儿在多模态事件中的物质感知与时间同步性 1986 McGurk效应	视频音频语音识别(AVSR) 1986 第一个AVSR系统 多模态/多感知接口 1993 Glove-talk(Fels,Hinton) 多模态人机交互 多媒体计算 镜头边缘检测(1991-) 静态/动态视频摘要(1992-) 高级解析((1997-) 自动标注(1999-)	拟人类多模态交互 2001 AMI Project 记录、处理会议数据 2003 CALO Project Siri的前身 2008 SSP Project 社交信号处理网络 多媒体信息检索 2001 NIST TRECVID 视频检索竞赛 010 DICITAL VIDEO RETRiEVAL Ns	2011 多模态深度学习(Ngiam) AVEC情感竞赛  2012 基于深度玻尔兹曼机的多模态学习(Srivastava,Salakhutdinov) 2015 显示，参加和讲述：具有视觉注意的神经图像字幕生成(Xu等)	2021 CLIP模型诞生 2022 基于CLIP的DALL ·E 2模型发布 BEiT-3模型发布 2023 微软发布微软KOSMOS-1；谷歌发布PaLM-E,把图像和语言模型的能力拓展到对机器人的控制

资料来源：carnegie mellon university

1.3 Transformer 颠覆传统模型，但限于单模态领域

- 2 017年 **Transformer被提出，颠覆了传统的深度学习模型**，在机器翻译任务上实现了最好性能。 Transformer 在大规模语料库上进行自监督预训练，然后在下游任务进行微调受到人们的关注，许多预训练大模型都是遵守这一范式提出，例如BERT 、GPT等。
- 虽然基于Transformer 的大模型都取得了很好的效果，但还是限于单一模态(文本)上，无法将其self -attention 中良好的泛化能力迁移到其他模态(图像、视频等)中。**Transformer不能迁移图像领域的主要原因在于输入长度限制，**以 BERT为例，其输入数据的长度只能支持512,而对于像素为224*224的图片来讲，其输入远大于512。

图表：Transformer基本架构



资料来源：Attention Is All You Need

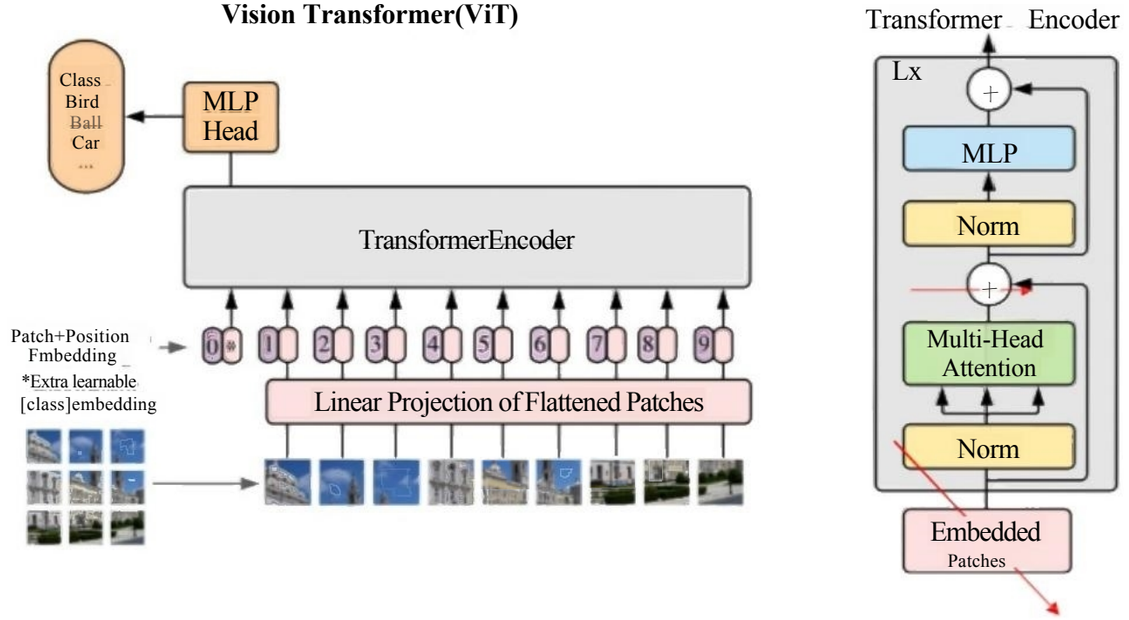
中信建投证券 CHINA SECURITIES

1.4 ViT的出现打通了CV和NLP之间壁垒，推动多模态演进

- Transformer (Self -attention) 在文本领域优秀的表现吸引着计算机视觉研究者，许多人开始将Transformer 机制引入到计算机视觉。

- Transformer **限制在于其输入数据大小，需要考虑** 输入策略。谷歌借鉴前人的思想，在强大的算力资源加持下，提出ViT模型。
- ViT模型通过将图片进行切割成一个个patch （原文将一张图片切割成16个 patch), 对 patch进行处理，通过线性映射，变成**图Tapme 构**接受的输入，打通了C 和NLP之间的壁垒。

将图片切割，解决输入大小问题



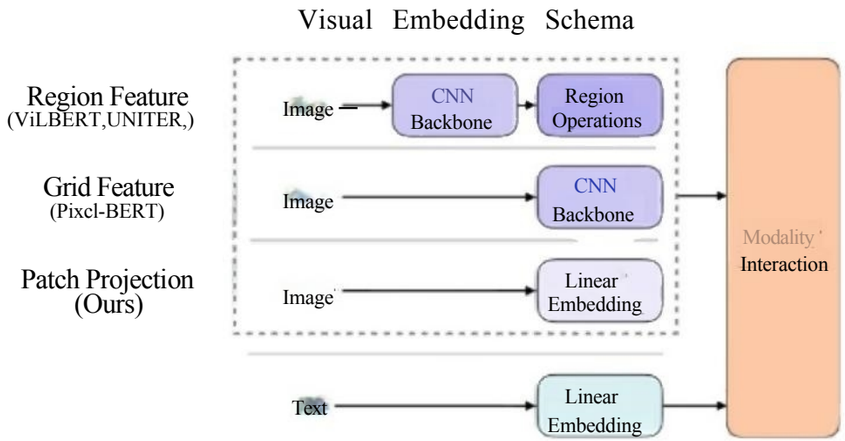
ViT将图片的2D信息，通过切割，转化为类似文本的1D信息。

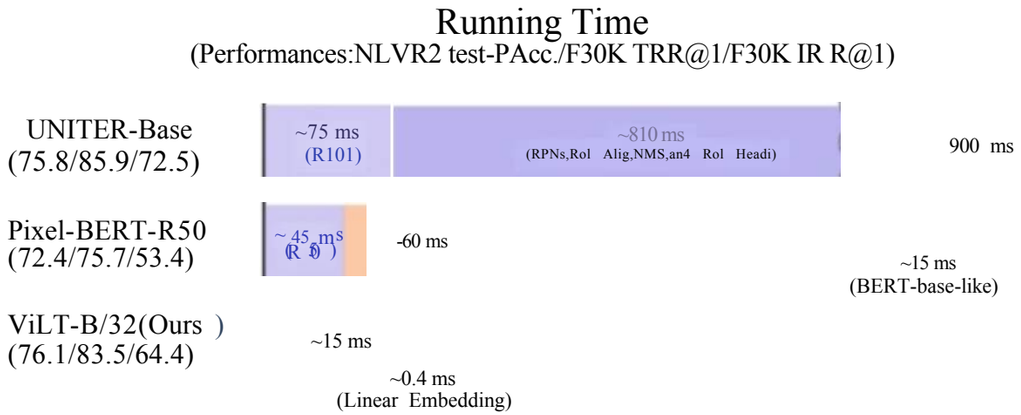
资料来源：An image is worth 16x16 words:Transformers for image recognition at scale

1.5 ViT 中 的Patch embedding 在提取视觉特征方面效率优势明显

- **ViT不仅能够让Transformer能够对图像进行处理，而且ViT 图像特征提取策略相较于之前的方式效率更高。**
- 如左图，虚线框内是三种视觉提取方式，分别为基于Region 、基于Grid 和ViT中线性映射方法进行视觉特征提取。在ViT之前，视觉算法中的视觉特征多基于Region 提取，大多会存在一个目标检测器，使用目标检测方法提取视觉特征。ViT 在预训练阶段舍弃了目标检测器，使用基于Patch 的视觉特征，几乎只相当于一个线性 embedding ，降低了运算复杂度。
- 如右图所示，ViLT多模态模型中在视觉特征提取方面使用了Patch embedding方法，实现了运行效率的大大提升，在特征提取阶段 ViLT -B/32的视觉特征提取阶段仅用0.4ms, 远高于Region(885ms) 和 Grid(45ms) 方法。

图表：ViLT 模型使用 Patch embedding提取视觉特征并取得很好效率



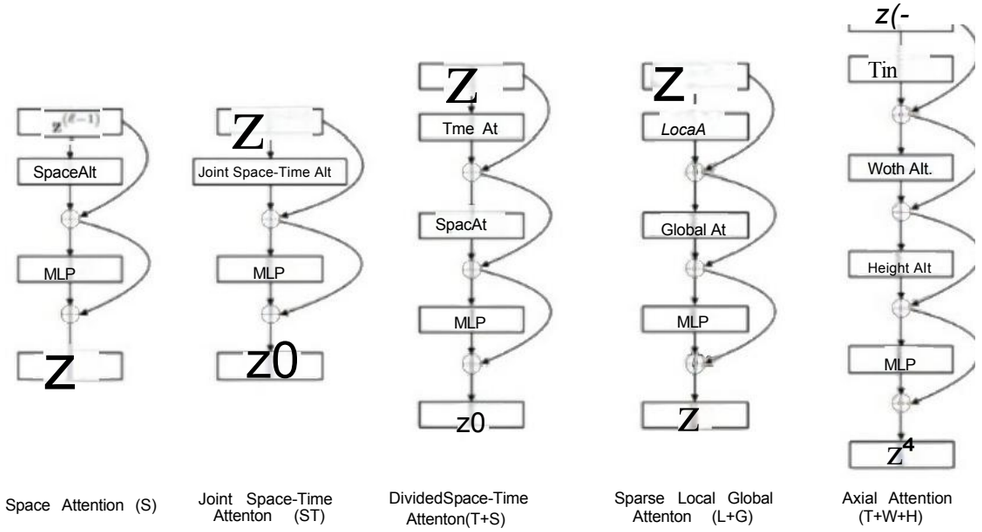


资料来源：ViLT:Vision -and -Language Transformer Without Convolution or Region Supervision

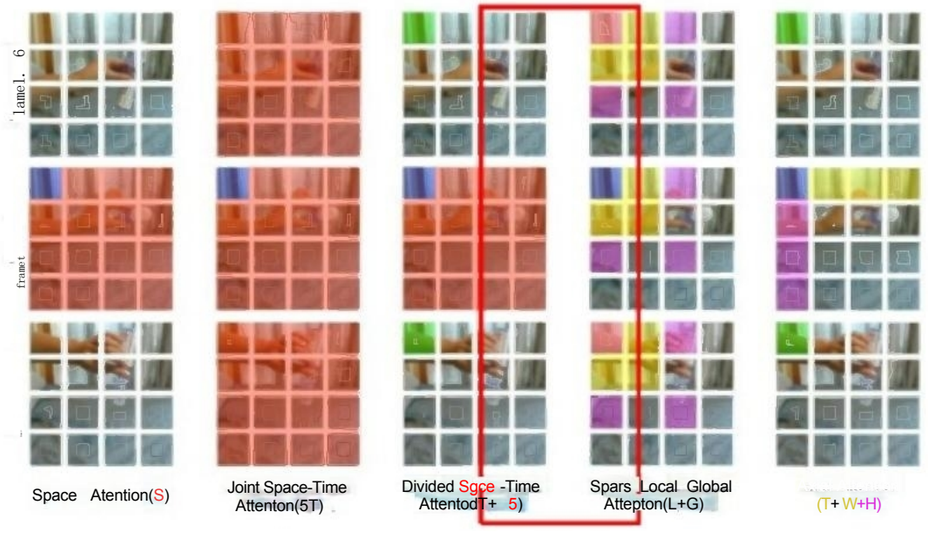
1.6 基 于 Vision Transformer ,Video Transformer 模型出现

- 1、**视频领域基于ViT模型推出各类Video Transformer。**视频是一个典型的多模态形式，里面包含图像、声音、文字等。
- 2、在ViT之前，视频方面的任务，如视频理解等，基本是通过3D卷积网络展开的。借鉴ViT思想，许多Video Transformer 被提出来，其中包括TimeSformer ,TimeSformer 将每一帧视频抽象成图像，并与其前一帧和后一帧相结合进行运算。与3D卷积神经网络 (CNN) 相 比 ，TimeSformer 的训练速度大约是其4倍，而推断所需的计算量不足其十分之一。TimeSformer 的高效让在高空间分辨率（例如高达560x560 像素的帧）和长视频（包括高达96 帧）上训练模型成为可能。

图表：自注意力机制在视频领域应用机制



图表：自注意力机制在视频领域应用详情



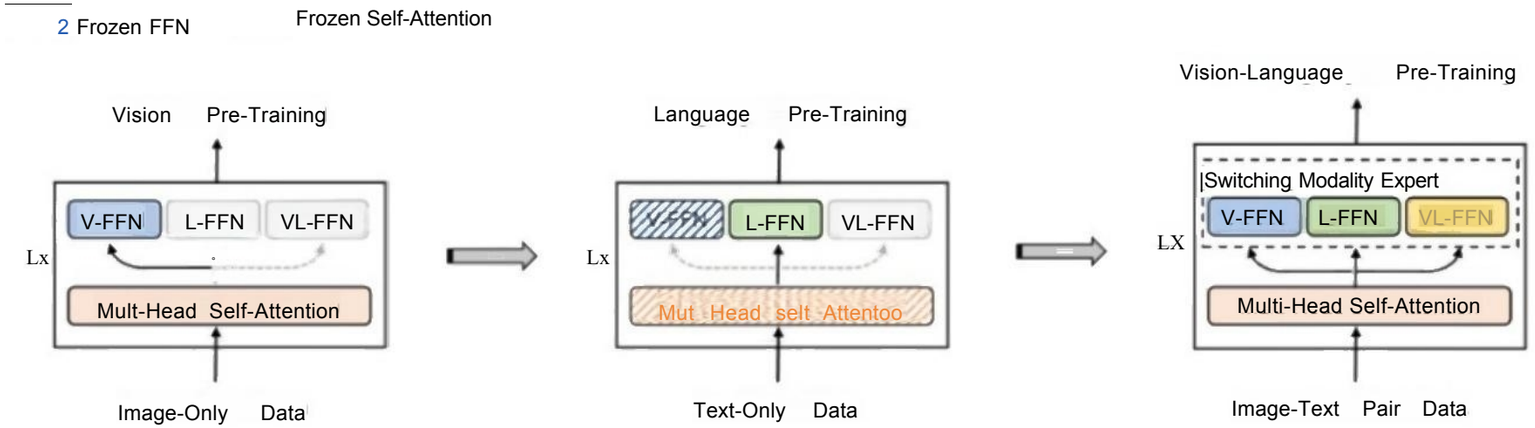
TimeSformer 将视频的每一帧看作一个图像，采取五种策略对图像中的像素进行处理，发现第三种处理方式最好。

资料来源：Is Space -Time Attention All You Need for Video Understand

1.7 Transformer 权重共享决定其适合多模态

- Transformer 存在权重共享，模型内部的某些模块可以共享权重参数。 Transformer 的权重共享主要是由于其自注意力模块和前向传播网络都和输入序列长度无关。
- **这种权重共享理念同样适合用于多模态模型中。**例如，图文多模态中，图像训练得到的权重参数可以用于训练文本，结果依然有效，甚至不用fine -tune。
- 许多多模态模型都借鉴了Transformer 里面的权重共享理念，典型的案例包括VLM。模型，该模型首先在BEiT中使用大规模纯图像数据预训练视觉网络和自注意力模块，然后冻结视觉网络和自注意力模块，通过对大量纯文本数据进行建模训练语言网络，最后使用视觉-语言预训练整个模型。

图表：VMo 预训练阶段



冻住的前向传播和自注意力共享视觉与文本参数

资料来源：LMO:Unified Vision -Language Pre-Training with Mixture -of -Modality -Experts

1.8 BEiT 模型的出现将生成式预训练从NLP迁移到CV上

- **生成式预训练是自监督学习重要方法和训练目标，**生成式预训练核心是在没有标签或者人工标注的情况下，学习如何产生数据。生成式预训练在自然语言处理中取得较大成功。**BEiT 模型的出现，将生成式预训练从NLP迁移到QV上，就是将BERT中的掩码语言学习 (MIM)方法应用到图像领域。**之后的MAE 模型也是基于 BEiT的工作展开的。**如 果 说ViT 将Transformer迁移到CV中，那么 BEiT 就是将 BERT迁移到CV中 。**BEiT 解决了CV上生成式预训练的两个问题：
  - 1、如何将图像信息转化为NLP中离散的token ，BEiT 使用到了dVEA 方法将图像离散化；
  - 2、使用成熟的ViT结构将处理图像信息。
- 通过以上两点，BEiT成功将MM/MIM方法应用图像领域，将生成式预训练迁移到cv 上，实现CV领域中大规模自监督预训练。

图 表 ：BEiT 模型预训练架构

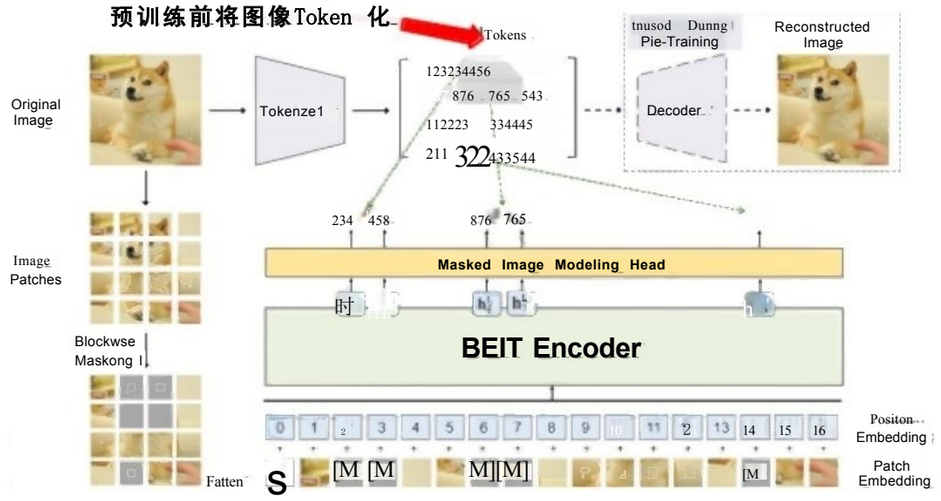
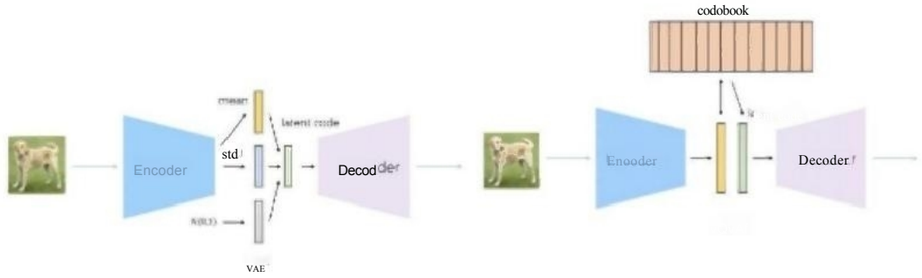


图 表 ：dVAE架构



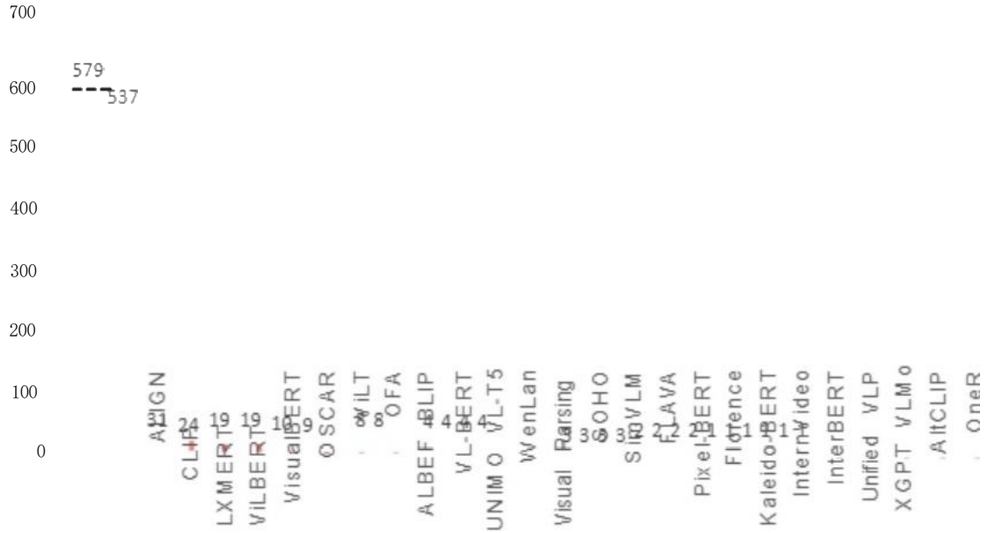
资料来源：Beit:Bert pre-training of image transformers

1.9 图文多模态是多模态模型中目前重要方向

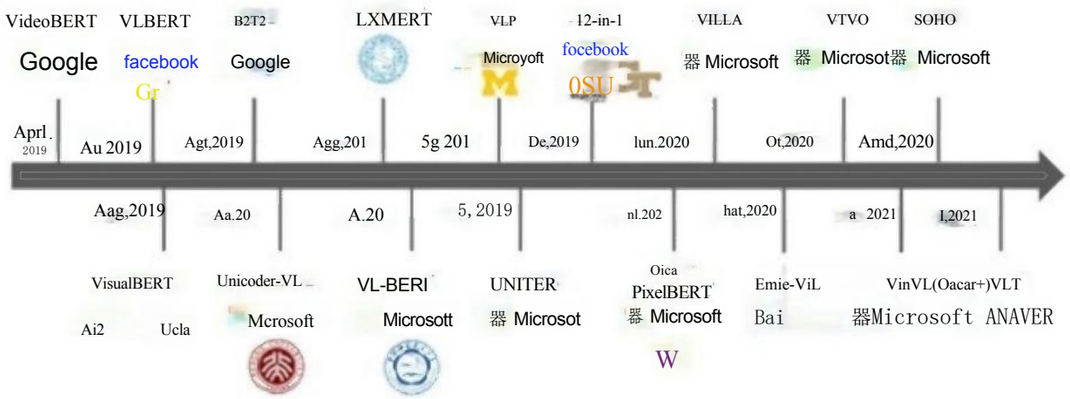
- 图文多模态任务是目前视觉语言预训练模型(VLP) 中最重要的任务之一。图文任务包括图文检测、图文分类、图文分割等。根据Paper with code网站上VLP领域中模型相关论文数量来看，ALIGN 和CLIP 模型相关论文数量最多，均超过500篇，这两个模型均是以图像-文本为对象展开的研究。

- 其中ALIGN 是谷歌2021年6月提出，CLIP 是OpenAI2021 年2月提出。

图表：VLP模型的相关论文数量(Paper with code数据)



图表：多模态模型主要情况

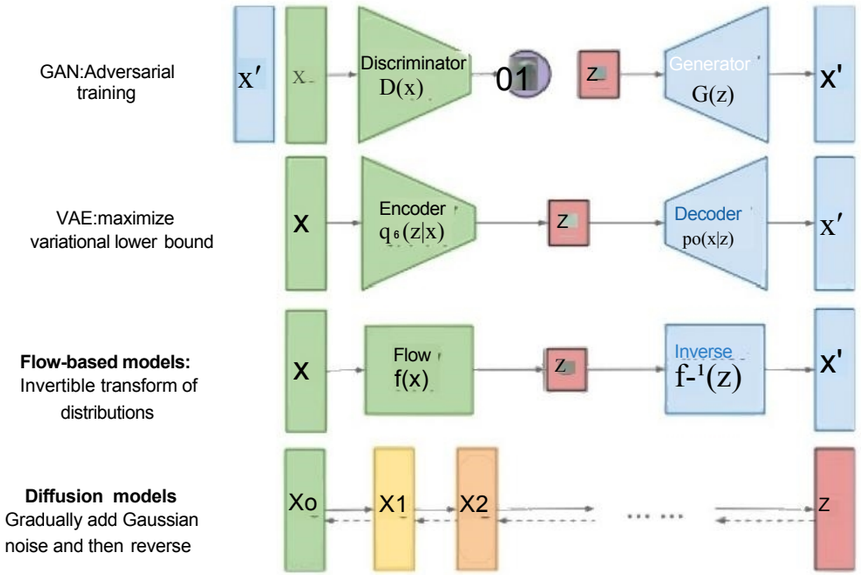


资料来源：Paper with code ,Vision -Language Intelligence :Tasks ,Representation Learning ,and Large Models

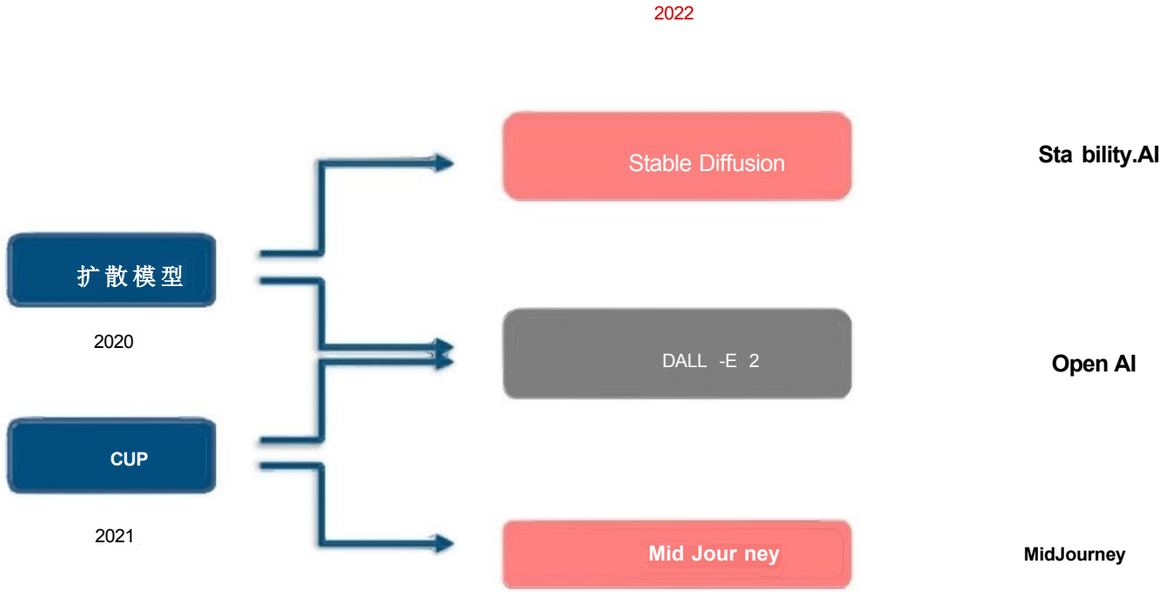
1.10 扩散模型推动多模态中文本图像生成发展

- 扩散模型**是一种继GN.VAE、Flow-based 模型之后最新的生成模型，从气体扩散的物理过程中获得灵感，通过正向扩散和反向扩**散两个过程进行生成。在OpenAI、英伟达、谷歌推出大模型后，扩散模型受到了较多的关注。
- 扩散模型与多模态预训练大模型的结合主要应用在文本图像生成领域。以扩散模型和多模态预训练大模型CLIP为基础模型，2022年4月OpenAI 发布文本生成图像模型DALL·E 2,之后谷歌推出Imagen, 直接对标DALL·E 2。

图表：几种生成式模型



图表：扩散模型与CLIP融合



资料来源：Github

1.11 多模态模型有包括COCO 在内的多个预训练数据集

- 和文本大模型或者视觉大模型类似，多模态预训练大模型也需要大量数据提前进行预训练，然后针对下游场景进行微调。
- 多模态模型目前用于许多预训练数据集，包括Flickr30k 、coc0、LAION -400M、RedCaps 在内的多项英文图像/文本数据集，也包括Wukong、WuDaoM、WSCD在内的多项中文数据集。这些数据集一般都是以图像文本对的形式存在，例如，LAION -400M包含CLIP模型过滤的4亿个图像文本对数据集； Wikong 包含1亿个中文图像文本对；Flickr30K 包含31000张图片，每张都与5个句子相关。

- LAION 是多模态模型数据集领域重要组织，他们是公益/非营利性组织，推出了LAION-400M、LAION -5B、Clip H/14等数据集，并且完全开源。

图表：多模态模型常见数据集

数据集	年份	规模(图文对数量)	语言	是否可获取
SBU Captions	2011	1M	English	是
Flickr30k	2014	145K	English	是
CoCo	2014	567K	English	是
FashionGen	2.018	300k	English	是
VQA v2.0	2017	1.1M	English	是
CC3M	2018	3M	English	是
GQA	2019	1M	English	是
LAIT	2.020	10M	English	否
CC12M	2021	12M	English	是
AltT ext	2.021	1.8B	English	否
TVQ A	2018	21,793	Eng lish	是
HT100M	2019	136M	English	是
WebVid2M	2.021	2.5M	English	是
YFCC -100M	2015	100M	English	是

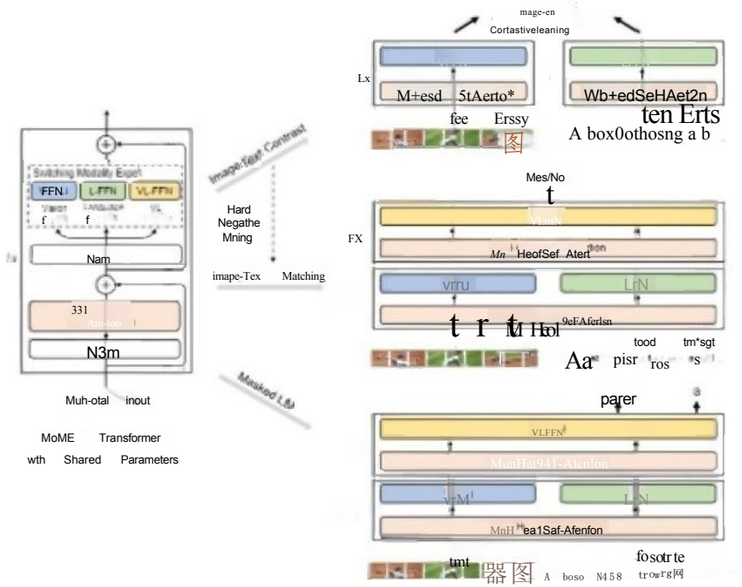
数据集	年份	规模	语言	是否可获取
LAION -400M	2021	400M	English	是
RedCap s	2021	12M	English	是
Wukong	2022	100M	Chinese	是
CxC	2021	24K	English	是
Product1M	2021	1M	Chinese	是
WIT	2021	37.5M	Multi -lingual	是
JFT-300M	2017	30M	English	否
JFT-3B	2021	3000M	English	否
IG-3.5B -17k	2018	350M	English	否
M6-Corpus	2021	60M	Chinese	否
M5Product	2021	6M	English	是
LocalizedNarratives	2020	849k	English	是
RUC -CAS-WenLan	2021	30M	Chinese	否
WuD aoMM	2022	600M	Chinese	是

资料来源：Vision -Language Intelligence :Tasks ,Representation Learning ,and Large Models

1.12 多模态模型大一统成趋势

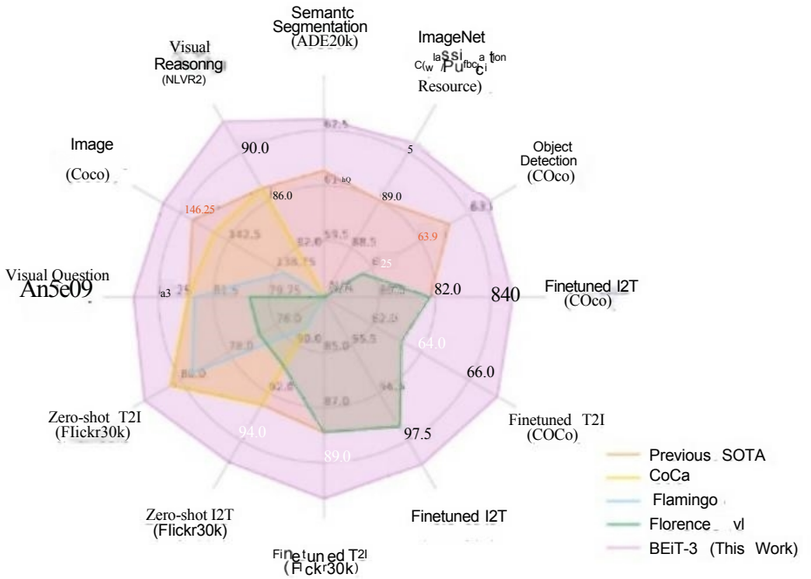
- 2022年8月，微软推出 BEiT -3模型，引领图像、文本、多模态迈向大一统。
- BEiT -3提出了掩码图像建模，将masked data modeling 引入到图像预训练任务，将图像和文本同等看待，以统一的方式对图像、文本、图像-文本对进行建模和学习。**实际上，微软在2021年11月就推出了统一模型VMO** ，使用混合模态专家（MOME） 的方式来进行不同模态中进行预训练，训练出不同的编码器，用于不同的下游任务。 BEiT-3 在其基础上简化模型并增大预训练数据量，最终在多项下游任务上表现亮眼。2023年3月15日，微软旗下OpenAI 推出多模态大模型GPT-4。

图表：WMO预训练框架



VLMO 在前向传播层，使用三个“专家”处理不同预训练任务

图表：BEiT -3在多项任务上表现亮眼



资料来源：Image as a Foreign Language :BEII Pretraining for All Vision and Vision -Language Task ,LMO:Unified Vision -Language Pre-Training withMixture -of-Modality -Experts

1.13 视频/音频等领域模态融合进展也较快

- 在视频、音频领域，多模态融合同样是一种趋势。
- **图文多模态模型逐渐迁移至视频-文本/音频-文本多模态领域**，典型的代表是CoCa模型，图文领域中推出后，在视频领域就推出了VideoCoCa ,CLIP 模型推出后，在视频领域就推出了VideoCLIP 模型。

- **一些统一多模态大模型的出现也在推动该领域的发展。**例如，阿里达摩院推出的mPLUG -2多模态大模型，不仅在图-文下游任务中取得很好的效果，也能进行视频领域的任务，例如在视频问答、视频字幕等领域相关工作上均取得了不错的成绩。

- 在音频多模态领域中比较著名的模型是谷歌推出的MusicLM 模型，能通过文字生成音乐。

图表：视频多模态领域中的数据集及模型

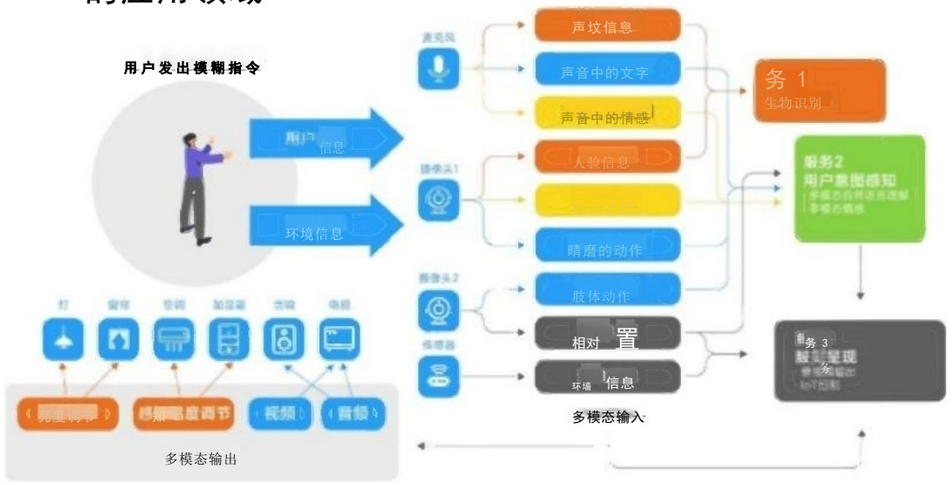
领域	数据 集	Best Model	评价	评价标准
Video Question Answering	ActivityNet -QA	VideoCoCa	56.10 %	Accuracy
	MSRVTT -QA	mPLUG -2	48%	
	iVQA	Text +Text (no Multimodal Pretext Training)	40.20 %	
	MSRVTT -MC	VIOLETv2	97.60 %	
	TVQA	FrozenBiLM	82%	
	NExT -QA	HiTeA	63.10 %	
	Howto100M -QA	Hero w/pre -training	77.75 %	
Video Captioning	MSR -VTT	mPLUG -2	57.8	BLUE -4
	YouCook2	UniVL	17.35	
	ActivityNet Captions	VideoCoCa	14.5	
	Hindi MSR -VTT	SBD_Keyframe	41.01	
Video Retrieval	MSR-VTT-1kA	HunYuan_tvr (huge)	62.9	text -to-video R@1
	MSR -VTT	InternVideo	55.2	
	MSVD	HunYuan_tvr (huge)	59	
	YouCook2	VideoCLIP	32.2	
	TVR	Hero w/pre -training	4.34	
	TGIF	MDMMT -2	25.5	

资料来源：Paper with code

1.14 多模态广泛存在于机器人、数字人、智能家居等领域

- **多模态在交互、感知、内容分发等众多领域都有较为重要的地位。**
- 多模态交互在家庭与办公场景下应用广泛，多模态交互可以进一步提升用户与智能家居设备的交互体验，提升了用户完成相同意图的效率与成功率。
- 多模态感知包括车场景和语音助手下的用户意图感知，例如，在驾车场景中，随着多屏主控等智能座舱技术进步，各种智能终端可以通过多模态交互实现意图识别准确率更高的用户体验。
- 多模态内容分发场景下，虚拟人结合动作、表情、情感、文本等信息，输出给用户。

图表：在家里通过多模态方式发出指令的应用领域



图表：多模态技术能够合成虚拟形象，给予用户多模态的信息

图表：多模态技术的应用领域

应用	公司/市场领域
文本生成	阿里商品推荐
机器翻译	有道AR翻译
	搜狗同传3.0
多模态检索	谷歌图像检索
	爱奇艺人脸识别
智能个人助理	阿里小蜜
	小爱同学
数字人	虎牙直播
	小爱虚拟形象
传感器智能	智能车舱

资料来源：多模态技术白皮书，中信建投

目录

多模态预训练关键要素总括

模型训练要素

1. 对图文进行tokenization ，转化为模型能处理的形式

文字使用成熟的BERT等模型进行处理

图像特征提取包括Grid、Region、Patch based方式

需要重要视觉特征；基于patch 的方式更高效

2. 设置学习目标

图文对比（ITC）

掩码语言模型（ MLM)

图文匹配（ITM）

使用不同的学习目标会带来不同的结果

3. 模型结构

Encoder -only

Encoder -decoder

常见的是Encoder-only 结构，用于图文检索等任务，encoder-decoder 结构适合相关生成任务

可以通过叠加多个模型结构/模态融合方式改变模型性能

4. 模态融合方式

Fusion Encoder

Dual Encoder

Fusion Encoder通过融合方式对模态进行处理；Dual Encoder分别对各模态进行处理

5. 提升数据质量

ALBEF 动量蒸馏生成伪标签

BLIP生成图像描述并和原来的进行比较过滤

6.Prom pt

CLIP

Visual ChatGPT

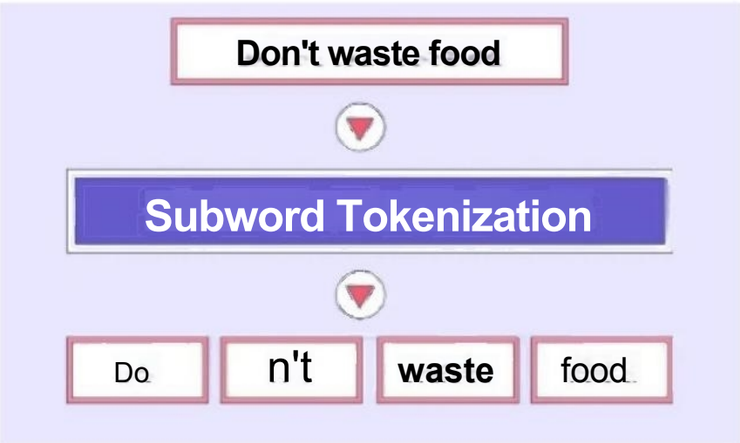
Prompt工程在多模态中更加重要，例如以上两个模型采取Prompt方式提建

2.1 图文需要Tokenization 和Embedding

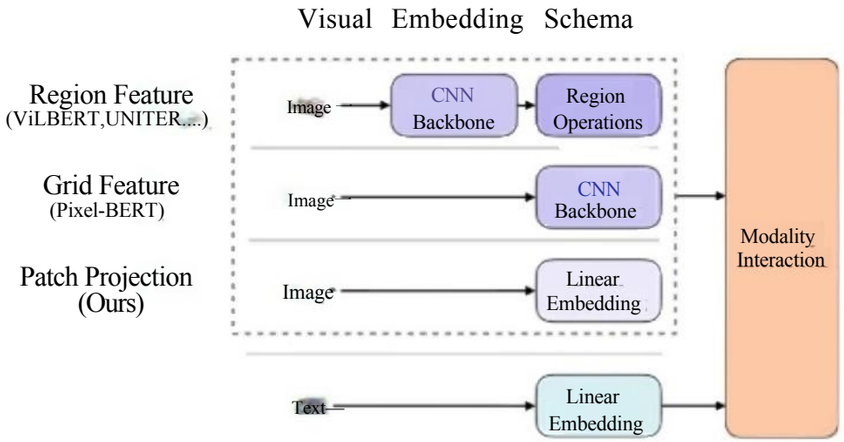
- **Token是模型输入的基本单元， Embedding是Token映射后的向量，用于计算。**
- 文字方面早期一般使用Word2Vec 进行Tokenization ，包 括CBOW和skip-gram, 虽 然Word2Vee 计算效率高，但是存在着词汇量不足的问题，因此子词分词法 (subword tokenization ） 被提出，使用字节对编码 (BPE) 将词分割成更小的单元，该方法已被应用于 BERT 等众多Transformer 模型中。

- 图像的Tokenization 要比文本更加复杂，可以分为基于region, 基于 grid和基于 patch三类方式。基于 grid的方式直接使用CN 进行图像网格信息提取，基于region 的方式由预训练的目标检测器进行特征提取，基于 patch 的方式将图像切割成小块，提取小块上的线性投影。

图表：子词分词法示例



图表：图像编码的三种方式

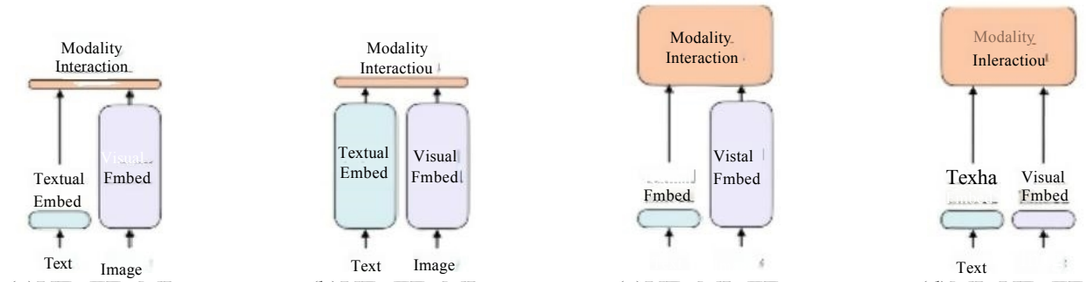


资料来源：Vision -Laguage Intelligence:Tasks,Representation Learning ,amd Large Moels ,Umicoder -π:A Universal Eacoder forvision and Language by Cross -Modal Pre -Training

2.2 多模态模型中要重视视觉特征

- **相较于文本特征而言，多模态模型中视觉特征更为重要。**
  - 当前多模态预训练大模型中，不论CLIP、UNITER 还是ViLT,在模型构造方面，视觉特征的embedding 层数或者复杂度要超过文本特征,体现出视觉特征更重要，多模态需要从视觉特征中学习到更多知识。
  - 根据METER模型中的数据显示，在视觉特征端进行优化对结果产生的影响要远大于对文本端进行的优化。

图表：多模态融合的四种形式



(a)VE>TE>MI (b)VE=TE>MI (c)VE>MI>TE (d)MI>VE=TE

多模态的主要形式中，无一例外
视觉特征要大于等于文本特征

图表：文字/视觉特征改变对结果影响

Text Fne.	vQAv2 Acc.	VE Acc.	IR R@1	TR R@1	SQuAD EM	MNLI Acc.
Emb-only	67.13	74.85	49.06	68.20		
ELECTRA	69.22	76.57	41.80	58.30	86.8	**88.8**
CLIP	69.31	75.37	**54.96**	**73.80**		
DeBERTa	69.40	**76.74**	51.50	67.70	**87.2**	**88.8**
BERT	69.56	76.27	49.60	66.60	76.3	84.3
RoBERTa	69.69	76.53	49.86	68.90	84.6	87.6
ALBERT	69.94	76.20	52.20	68.70	86.4	87.9

无论文本特征如何改变，对结果影响不大

Vision Encoder	VQAv2	VE	IR	TR	ImageNet
Dis.DeiT B-384/16	67.84	76.17	34.84	52.10	85.2
BEiT B-224/16	68.45	75.28	32.24	59.80	85.2
DeiT B-384/16	68.92	75.97	33.38	50.90	82.9
ViT B-384/16	69.09	76.35	40.30	59.80	83.97
CLIP B-224/32	69.69	76.53	49.86	68.90	
vOL04-448/32	71.44	76.42	40.90	61.40	**86.8**
CaiT M-384/32	71.52	76.62	38.96	61.30	86.1
CLIP B-224/16	71.75	77.54	57.64	**76.90**	
Swin B-384/32	**72.38**	77.65	52.30	69.50	86.4

视觉特征改变对结果影响比较大

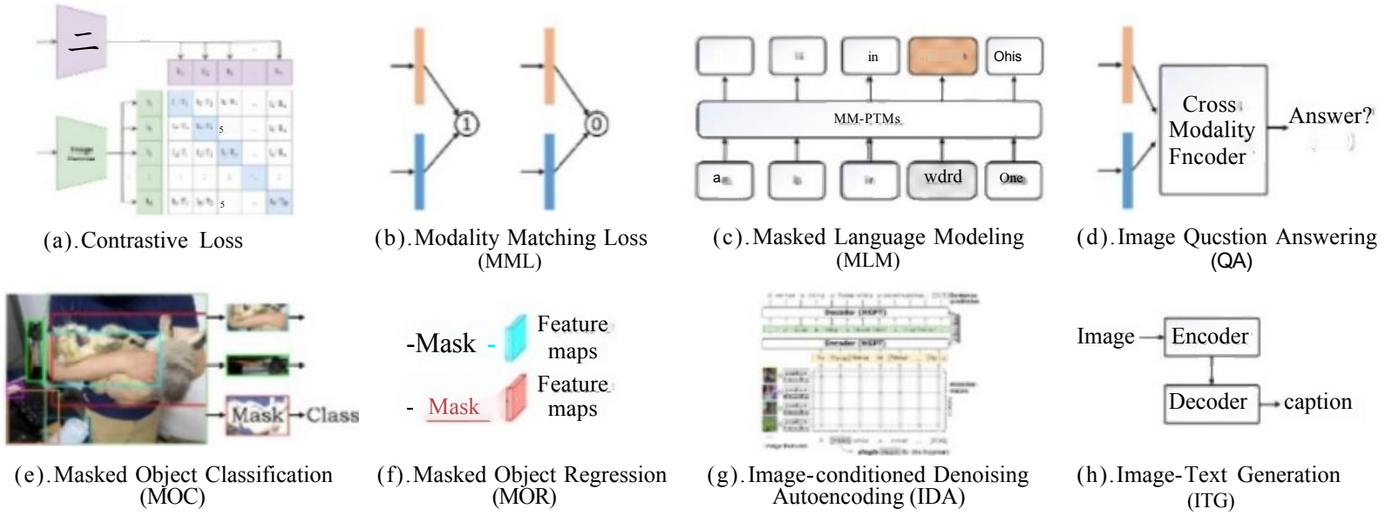
资料来源：ViLI:Vision -and -Language Transformer Without Convolution or Region Supervision ,An Empirical study of Training End -to-End Vision and Language Transformers

2.3 如何设计学习目标是多模态训练的重要一步

- 学习目标是多模态预训练非常重要的一步，**目前的多模态的预训练学习任务主要包括图文对比（ITC）、掩码语言学习（MIM）掩码视觉学习（MMM）、图文匹配（ITM）等。**

- ITC是通常构造正负样本对，通过对比学习方式，对齐图像和文本；
- ITM可以看作是一个二分类任务，目标是预测一对图像和文本是否匹配；
- MLM是让模型学习语言和视觉内容之间的隐式关系，目标是从已知的语言和视觉内容中重建掩码语言标记；
- 此外还包括掩码物体分类（MOC）、掩码物体回归（MOR）、行为预测（AP）、图文生成（ITG）等。

图表：多模态中学习目标



(a).Contrastive Loss (b).Modality Matching Loss (MML) (c).Masked Language Modeling (MLM) (d).Image Qucstion Answering (QA)

(e).Masked Object Classification (MOC) (f).Masked Object Regression (MOR) (g).Image-conditioned Denoising Autoencoding (IDA) (h).Image-Text Generation (ITG)

资料来源：Large -scale Multi -Modal Pre-trained Models:A Comprehensive Survey

2.4不同的多模态预训练学习目标可能带来不一样的结果

- **同时使用不同的预训练学习目标可能会增强多模态模型的效果，** 例如UNITER模型中，使用更多的学习目标效果一般要更好，UNITER使用MLM+ITM+MRC-kl+MRFR+WRA等多个学习目标在在多个细分场景下表现要更好。
- **使用过多的学习目标可能效果并不好。** 例如，METER模型中，在MM和ITM上再加入MIM学习模型，效果比使用单个学习目标要好，但不如仅仅使用两个学习目标，这一方面可能是学习目标之间的冲突导致的，另外一方面可能是图像中存在噪声，MIM重建图像噪声的监督学习没有任何意义导致的。

图表：UNITER模型在使用不同学习目标得到不同结果

Pre-training Data		Pre-training Tasks	Meta-Sum	VQA	IR (Flickr)		TR (Flickr)		Ref-COCO+
			NLVR²		test-dev	val	test-dev	val	val
None	1	Nom	314.34	67.03	61.74	65.55	51.02		68.73
Wikipodia+ BookCorpus	2	MLM(text ouly)	346.24	69.39	73.92	83.27	50.86		68.80
In domain (CoCO+VG)	3	MRFR	344.66	69.02	72.10	82.91	52.16		68.47
	4	ITM	385.29	70.04	78.93	89.91	74.08		72.33
	5	MLM	386.10	71.29	77.88	89.25	74.79		72.89
	6	MLM+ITM	393.04	71.55	81.64	91.12	75.98		72.75
	7	MLM+ITM+MRC	393.97	71.46	81.39	91.45	76.18		73.49
	8	MLM+1M+MRFR	396.24	71.73	81.76	92.31	76.21		74.23
	9	MLM+ITM+MRC-kl	397.09	71.63	82.10	92.57	76.28		74.51
	10	MIM+ITM+MRC-kl+MBFR	399.97	71.92	83.73	92.87	76.93		74.52
	11	MLM+rM+MRC-k+MRFR+WRA	400.93	72.47	83.72	93.03	76.91		74.80
	2	MLM+ITM+MNC-kl+MIFT (wlognd.mk	396.51	71.68	82.31	92.08	76.15		74.29
Out-of-domain (SBU+OX)	13	MIM+ITM+MRC-kl+MBFR+WRA	396.91	71.56	84.34	92.57	75.66		72.78
In-domain + Out-of domain	14	MLM+mM+MRC-kl+MRFB+WRA	405.24	72.70	85.77	94.28	77.18		75.31

图表：METER模型在使用不同学习目标得到不同结果

Pre-training Objectives	VQAv2	Flickr-ZS IR	Flickr-ZS TR
MLM	74.19		
ITM	72.63	53.74	71.00
MLM+ITM	74.98	66.08	78.10
MLM+ITM+MIM (In-batch Negatives)	74.01	62.12	76.90
MLM+ITM+MIM(Discrete Code)	74.21	59.80	76.30

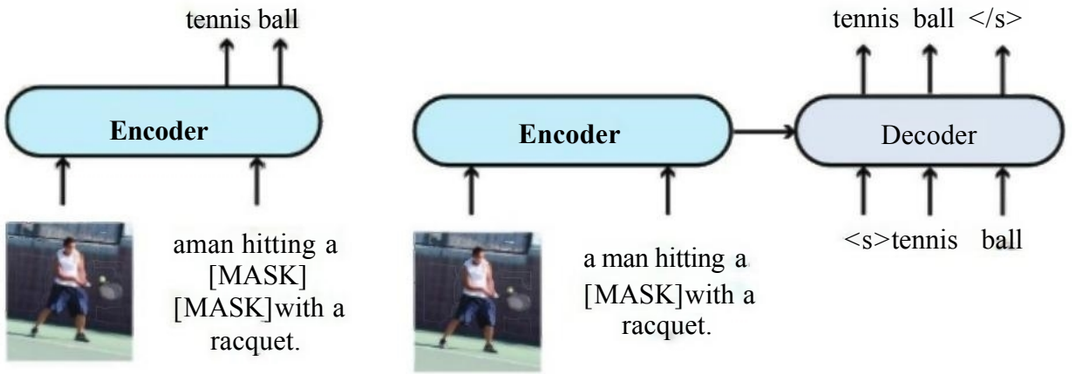
资料来源：UNITER:UNiversal Image -TExtRepresentation Learning ,An Empirical Study of Training End -to-End Vision -and -Language Transformers

2.5 多模态模型结构包括encoder -only 和encoder -decoder 两类

- **根据模型的结构不同，多模态可以分为encoder-only和encoder-decoder 两类。**
- 顾名思义， encoder -only指模型只用了transformer 的编码器部分，多模态的输入直接通过 encoder 进行输出，而encoder-decoder 则是使用了transformer 中的编码器和解码器两部分，解码器同时获得解码器的输出结果以及之前生成的 token ,使 用 自回归产生输出。

- 常见的多模态模型是encoder -only,包 括CLIP、ALBEF 等，适合图文检索，但是不适合生成任务，例如image captioning 等；
- Encoder -Deocder 模型包括 VL-T5、SimVLM 等，利用了decoder 的特性，适合生成任务，但是不太适合图文检索等。

图表：多模态中的Encoder-only和Encoder-Decoder架构



(a)Encoder-Only **(b)** Encoder-Decoder

图表：Encoder-only和Encoder-Decoder基本情况小结

架构	基本情况	代表
Encoder -only	常见；适合图文检索，但不适合生成任务	CLIP、ALBEF
Encoder-Decoder	不适合图文检索，适合生成任务	VL-T5、SimVLM

资料来源：An epirical study of training end-to-end visiom -and-lamguage transforners ,LP:Botstraping Language -Image Pre-training fotOnified Vision -Language Understanding