

多模态学习综述(MultiModal Learning)

极市平台

2023-03-01 22:00:08

发表于广东

手机阅读

𐄞

以下文章来源于张浩在路上，作者Hao Zhang



张浩在路上

开发者，关注机器学习和商业化变现，这里分享自己的学习笔记和日常思考。

↑ 点击蓝字 关注极市平台



作者 | 张浩@知乎（已授权）
来源 | <https://zhuanlan.zhihu.com/p/582878508>
编辑 | 极市平台

极市导读

Jeff Dean 谈2020年机器学习趋势：多任务和多模式学习将成为突破口。也正如他预言的一样，多模态学习在行业内越来越火爆。本文对多模态的发展历史以及经典的多模态任务、面临的挑战等都进行了详细的阐述，希望能帮助大家多模态方向有更加清晰的认知。>>加入极市CV技术交流群，走在计算机视觉的最前沿

原文首发于我的博客：

多模态学习(MultiModal Learning)：

<https://imzhanghao.com/2022/10/27/multimodal-learning/>

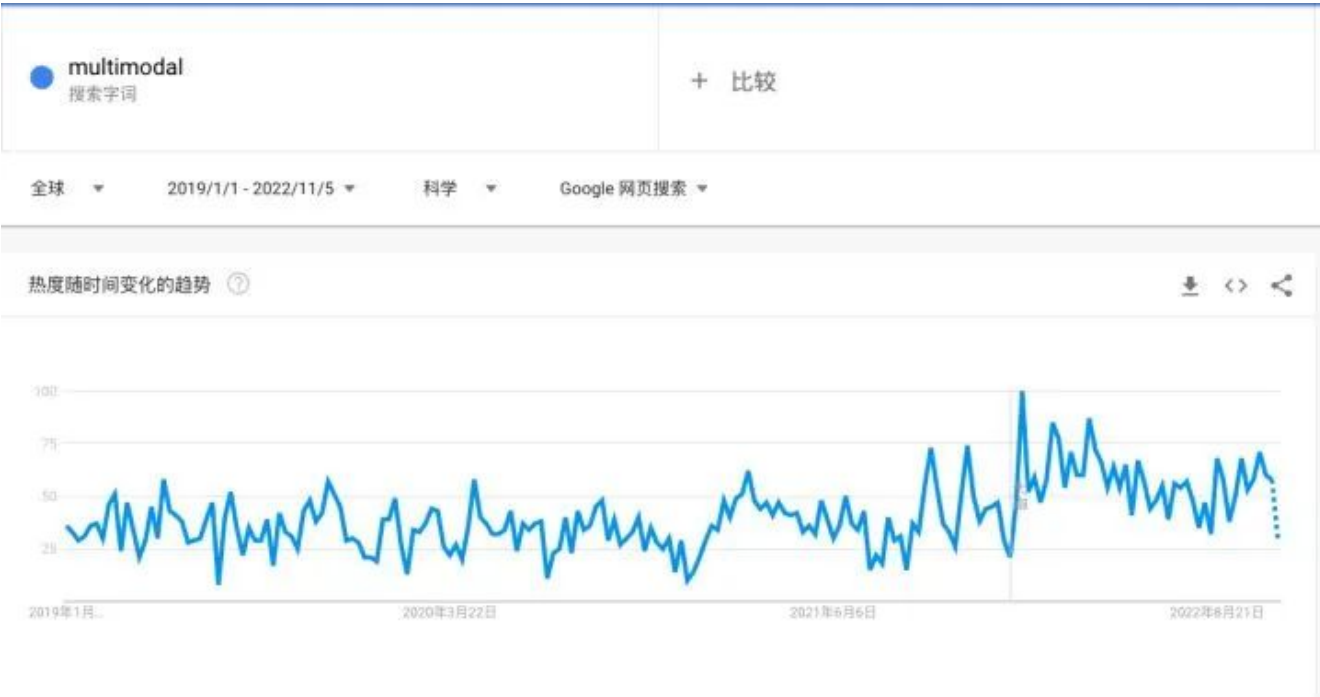
最早开始关注到多模态机器学习是看到Jeff Dean在2019年年底NeurIPS大会上 的一个采访报道，讲到了2020年机器学习趋势：多任务和多模态学习将成为突破口。

VentureBeat: What are some of the trends you expect to emerge, or milestones you think may be surpassed in 2020 in AI?

Dean: I think we'll see much more multitask learning and multimodal learning, of sort of larger scales than has been previously tackled. I think that'll be pretty interesting.

Jeff Dean 谈2020年机器学习趋势：多任务和多模式学习将成为突破口

站在2022年，也正如他预言的一样，多模态学习在行业内越来越火爆。

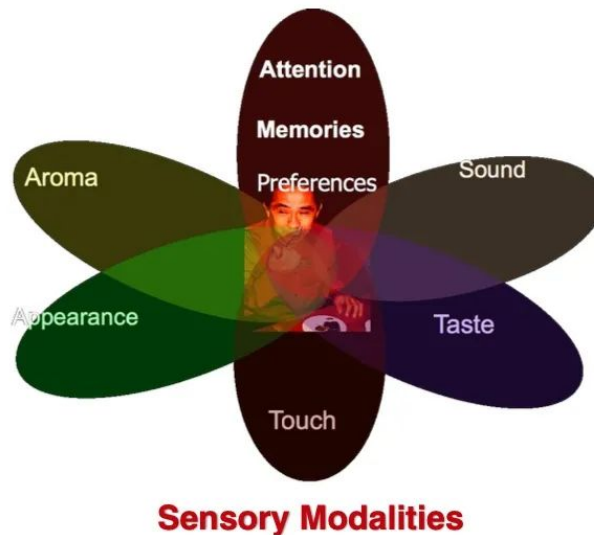


多模态机器学习在Google Trends上的表现

一、定义

多模态机器学习，英文全称 MultiModal Machine Learning (MMML)

模态（modal）是事情经历和发生的方式，我们生活在一个由多种模态（Multimodal）信息构成的世界，包括视觉信息、听觉信息、文本信息、嗅觉信息等等，当研究的问题或者数据集包含多种这样的模态信息时我们称之为多模态问题，研究多模态问题是推动人工智能更好的了解和认知我们周围世界的关键。



1.1 模态

模态是指一些表达或感知事物的方式，每一种信息的来源或者形式，都可以称为一种模态。例如，人有触觉，听觉，视觉，嗅觉；信息的媒介，有语音、视频、文字等；多种多样的传感器，如雷达、红外、加速度计等。以上的每一种都可以称为一种模态。

相较于图像、语音、文本等多媒体(Multi-media)数据划分形式，“模态”是一个更为细粒度的概念，同一媒介下可存在不同的模态。比如我们可以把两种不同的语言当做是两种模态，甚至在两种不同情况下采集到的数据集，亦可认为是两种模态。

1.2 多模态

多模态即是从多个模态表达或感知事物。多模态可归类为同质性的模态，例如从两台相机中分别拍摄的图片，异质性的模态，例如图片与文本语言的关系。

多模态可能有以下三种形式：

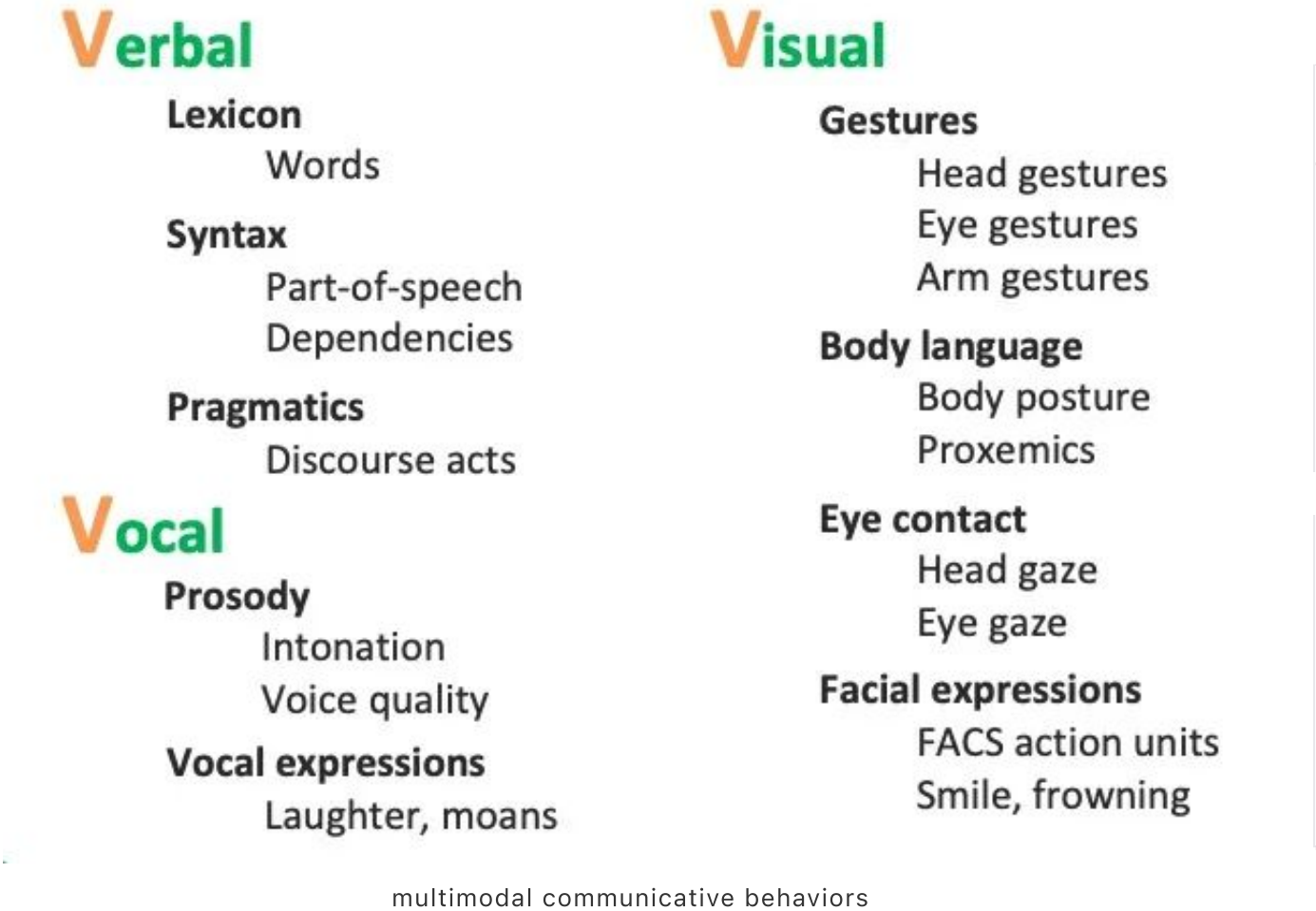
- 描述同一对象的多媒体数据。如互联网环境下描述某一特定对象的视频、图片、语音、文本等信息。下图即为典型的多模态信息形式。



"下雪"场景的多模态数据(图像、音频与文本)

- **来自不同传感器的同一类媒体数据**。如医学影像学中不同的检查设备所产生的图像数据，包括B超(B-Scan ultrasonography)、计算机断层扫描(CT)、核磁共振等；物联网背景下不同传感器所检测到的同一对象数据等。
- **具有不同的数据结构特点、表示形式的表意符号与信息**。如描述同一对象的结构化、非结构化的数据单元；描述同一数学概念的公式、逻辑符号、函数图及解释性文本；描述同一语义的向量、词袋、知识图谱以及其它语义符号单元等。

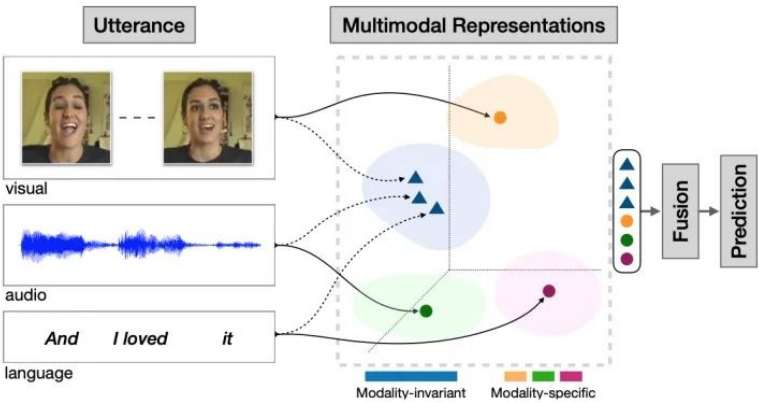
通常主要研究模态包括"**3V**"：即**Verbal(文本)**、**Vocal(语音)**、**Visual(视觉)**。人跟人交流时的多模态：



1.3 多模态学习

多模态机器学习是从多种模态的数据中学习并且提升自身的算法，它不是某一个具体的算法，它是一类算法的总称。

从语义感知的角度切入，多模态数据涉及不同的感知通道如视觉、听觉、触觉、嗅觉所接收到的信息;在数据层面理解，多模态数据则可被看作多种数据类型的组合，如图片、数值、文本、符号、音频、时间序列，或者集合、树、图等不同数据结构所组成的复合数据形式，乃至来自不同数据库、不同知识库的各种信息资源的组合。对多源异构数据的挖掘分析可被理解为多模态学习。



多模态学习举例

二、发展历史

Four eras of multimodal research

- The “behavioral” era (1970s until late 1980s)
- The “computational” era (late 1980s until 2000)
- The “interaction” era (2000 - 2010)
- The “deep learning” era (2010s until ...)

多模态发展的四个时期

2.1 行为时代

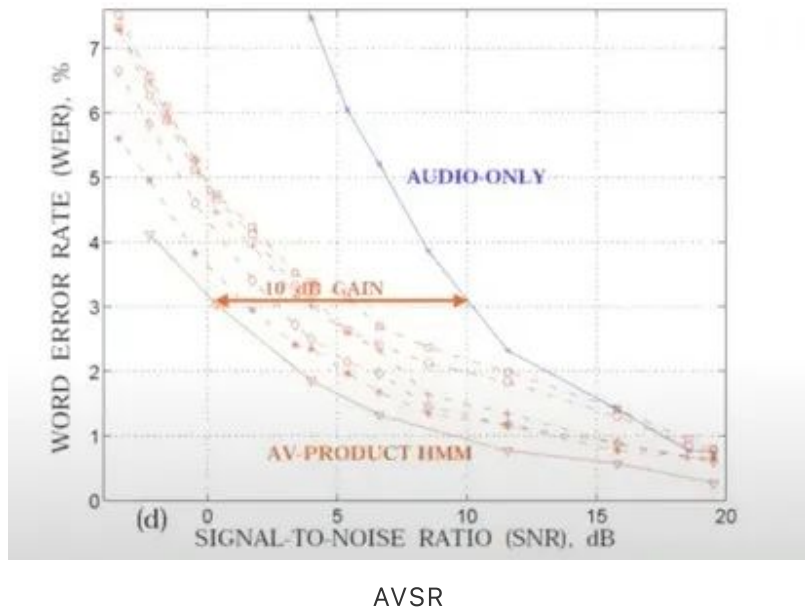
The “behavioral” era (1970s until late 1980s), 这一阶段主要从心理学的角度对多模态这一现象进行剖析。

- Chicago 的 McNeill 认为手势是说话人的思考行为，是言语表达的重要组成部分，而不仅仅是补足。
- 1976年的McGurk效应：当语音与唇形不符合时，大脑会脑补出中和的声音MCGURK, H., MACDONALD, J. Hearing lips and seeing voices. Nature 264, 746–748 (1976). The McGurk Effect Video

2.2 计算时代

The “computational” era (late 1980s until 2000), 这一阶段主要利用一些浅层的模型对多模态问题进行研究，其中代表性的应用包括视觉语音联合识别，多模态情感计算等等。

- 视频音频语音识别(AVSR)，在声音的低信噪比下，引入视觉信号能够极大提升识别准确率



- 多模态/多感知接口：情感计算：与情感或其他情感现象有关、源于情感或有意影响情感的计算[Rosalind Picard]
- 多媒体计算：CMU曾有过信息媒体数字视频库项目[1994-2010],

2.3 交互时代

The "interaction" era (2000 - 2010), 这一阶段主要主要从交互的角度入手，研究多模态识别问题，其中主要的代表作品包括苹果的语音助手Siri等。

拟人类多模态交互过程

- IDIAP实验室的AMI项目[2001-2006]，记录会议录音、同步音频视频、转录与注释；
- Alex Waibel的CHIL项目，将计算机置于人类交互圈中，多传感器多模态信号处理，面对面交互



AMI Project [2001-2006, IDIAP]

- 100+ hours of meeting recordings
- Fully synchronized audio-video
- Transcribed and annotated



CHIL Project [Alex Waibel]

- Computers in the Human Interaction Loop
- Multi-sensor multimodal processing
- Face-to-face interactions

IMI Project & CHIL Project

- 2003-2008 SRI的学习和组织认知助手，个性化助手，Siri就是这个项目的衍生产品
- 2008-2011 IDIAP的社交信号处理网络，数据库<http://sspnet.eu>。



CALO Project [2003-2008, SRI]

- Cognitive Assistant that Learns and Organizes
- Personalized Assistant that Learns (PAL)
- Siri was a spinoff from this project



SSP Project [2008-2011, IDIAP]

- Social Signal Processing
- First coined by Sandy Pentland in 2007
- Great dataset repository: <http://sspnet.eu/>

CALO Project & SSP Project

2.4 深度学习时代

The “deep learning” era (2010s until ...), 促使多模态研究发展的关键促成因素有4个，
1) 新的大规模多模态数据集，2) GPU快速计算，3) 强大的视觉特征抽取能力，4) 强大的语言特征抽取能力。

表示学习三篇参考文献

- Multimodal Deep Learning [ICML 2011]

- Multimodal Learning with Deep Boltzmann Machines [NIPS 2012]
- Visual attention: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention [ICML 2015]

三、典型任务

3.1 跨模态预训练

- 图像/视频与语言预训练。
- 跨任务预训练

3.2 Language-Audio

- Text-to-Speech Synthesis: 给定文本，生成一段对应的声音。
- Audio Captioning: 给定一段语音，生成一句话总结并描述主要内容。(不是语音识别)

3.3 Vision-Audio

- Audio-Visual Speech Recognition(视听语音识别): 给定某人的视频及语音进行语音识别。
- Video Sound Separation(视频声源分离): 给定视频和声音信号(包含多个声源), 进行声源定位与分离。
- Image Generation from Audio: 给定声音，生成与其相关的图像。
- Speech-conditioned Face generation: 给定一段话，生成说话人的视频。
- Audio-Driven 3D Facial Animation: 给定一段话与3D人脸模版，生成说话的人脸3D动画。

3.4 Vision-Language

- Image/Video-Text Retrieval (图(视频)文检索): 图像/视频 \leftrightarrow 文本的相互检索。

- Image/Video Captioning(图像/视频描述): 给定一个图像/视频, 生成文本描述其主要内容。
- Visual Question Answering(视觉问答): 给定一个图像/视频与一个问题, 预测答案。
- Image/Video Generation from Text: 给定文本, 生成相应的图像或视频。
- Multimodal Machine Translation: 给定一种语言的文本与该文本对应的图像, 翻译为另外一种语言。
- Vision-and-Language Navigation(视觉-语言导航): 给定自然语言进行指导, 使得智能体根据视觉传感器导航到特定的目标。
- Multimodal Dialog(多模态对话): 给定图像, 历史对话, 以及与图像相关的问题, 预测该问题的回答。

3.5 定位相关的任务

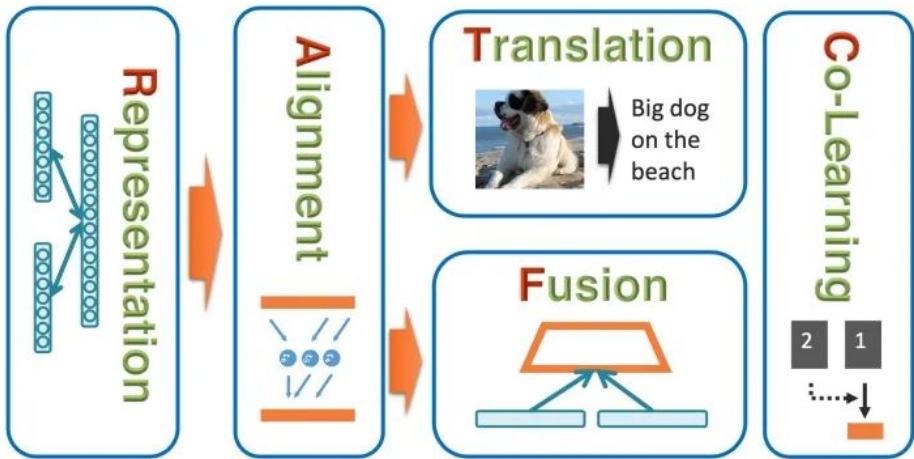
- Visual Grounding: 给定一个图像与一段文本, 定位到文本所描述的物体。
- Temporal Language Localization: 给定一个视频即一段文本, 定位到文本所描述的动作(预测起止时间)。
- Video Summarization from text query: 给定一段话(query)与一个视频, 根据这段话的内容进行视频摘要, 预测视频关键帧(或关键片段)组合为一个短的摘要视频。
- Video Segmentation from Natural Language Query: 给定一段话(query)与一个视频, 分割得到query所指示的物体。
- Video-Language Inference: 给定视频(包括视频的一些字幕信息), 还有一段文本假设(hypothesis), 判断二者是否存在语义蕴含(二分类), 即判断视频内容是否包含这段文本的语义。
- Object Tracking from Natural Language Query: 给定一段视频和一些文本, 追踪视频中文本所描述的对象。

- Language-guided Image/Video Editing: 一句话自动修图。给定一段指令(文本), 自动进行图像/视频的编辑。

3.6 更多模态

- Affect Computing (情感计算): 使用语音、视觉(人脸表情)、文本信息、心电、脑电等模态进行情感识别。
- Medical Image: 不同医疗图像模态如CT、MRI、PETRGB-D模态: RGB图与深度图

四、技术挑战



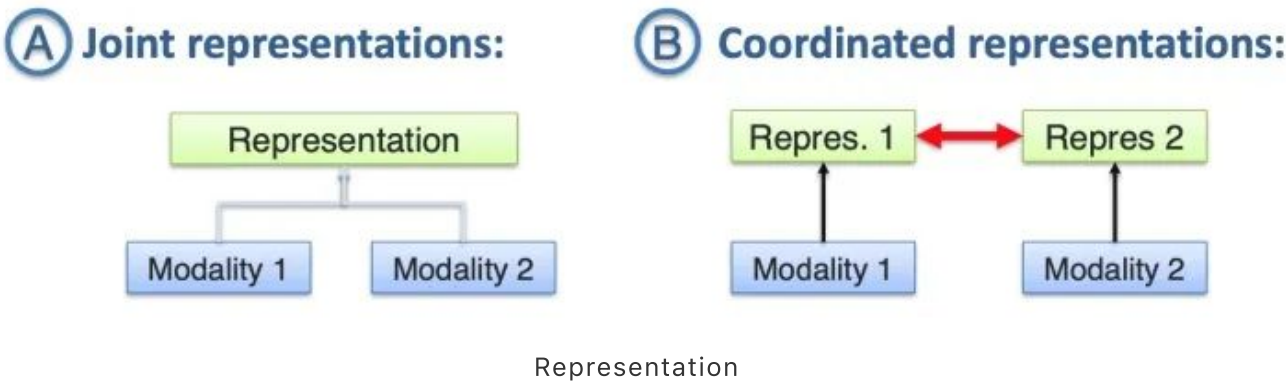
多模态学习的技术挑战

4.1 表征Representation

第一个基本挑战是学习如何以**利用多种模态的互补性和冗余性的方式表示和总结多模态数据**。多模态数据的异质性使得构建这样的表示具有挑战性。例如，语言通常是象征性的，而音频和视觉形式将被表示为信号。

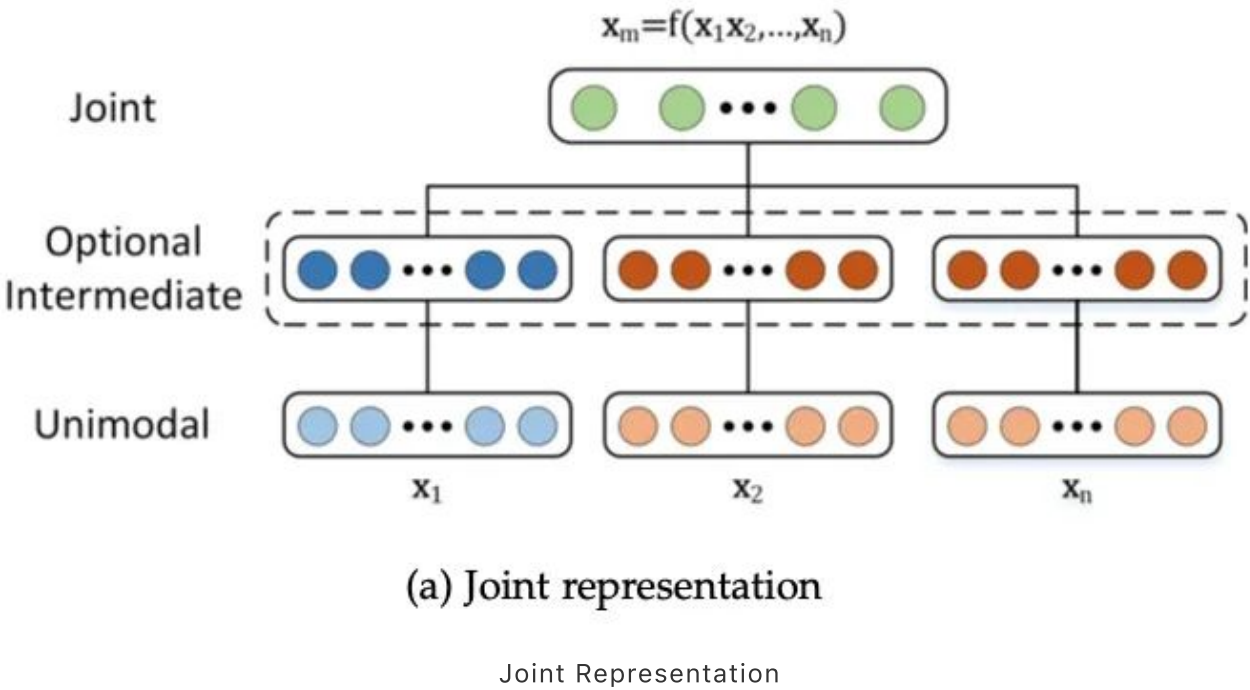
单模态的表征负责将信息表示为计算机可以处理的数值向量或者进一步抽象为更高层的特征向量，而多模态表征是指通过利用多模态之间的互补性，剔除模态间的冗余性，从而学习到更好的特征表示。

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

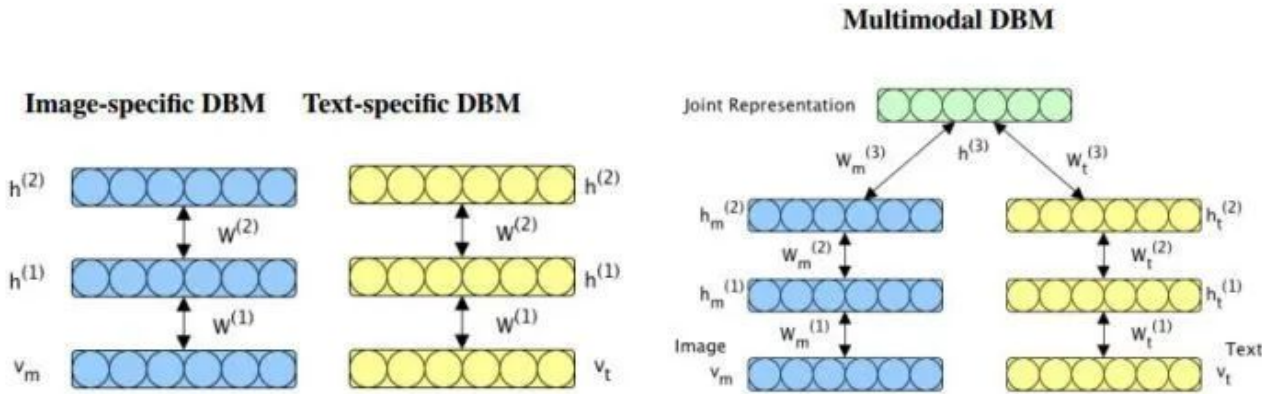


4.1.1 联合表征

联合表征（Joint Representation）将多个模态的信息一起映射到一个统一的多模态向量空间，Joint结构注重捕捉多模态的互补性，融合多个输入模态 x_1, x_2 获得多模态表征 $x_m = f(x_1, \dots, x_n)$ ，进而使 x_m 完成某种预测任务。



Multimodal learning with deep boltzmann machines (NIPS 2012) 提出将 deep boltzmann machines（DBM）结构扩充到多模态领域，通过 Multimodal DBM，可以学习到多模态的联合概率分布。



Multimodal DBM 模型

在获得图像与文本间的联合概率分布后，我们在应用阶段，输入图片，利用条件概率 $P(\text{文本}|\text{图片})$ ，生成文本特征，可以得到图片相应的文本描述；而输入文本，利用条件概率 $P(\text{图片}|\text{文本})$ ，可以生成图片特征，通过检索出最靠近该特征向量的两个图片实例，可以得到符合文本描述的图片。













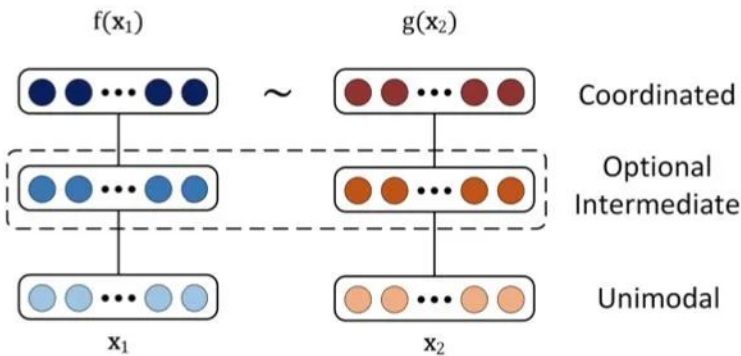
Image	Given Tags	Generated Tags	Input Tags	Nearest neighbors to generated image features	
	pentax, k10d, kangarooisland, southaustralia, sa, 300mm, australia, australiansealion	beach, sea, surf, strand, shore, wave, seascape, sand, ocean, waves	nature, hill, scenery, green, clouds		
	< no text >	night, lights, christmas, nightshot, nacht, nuit, notte, longexposure, noche, nocturna	flower, nature, green, flowers, petal, petals, bud		
	aheram, 0505, sarahc, moo	portrait, bw, balckandwhite, people, faces, girl, blackwhite, person, man	blue, red, art, artwork, painted, paint, artistic, surreal, gallery, bleu		
	unseulpixel, naturey crap	fall, autumn, trees, leaves, foliage, forest, woods, branches, path	bw, blackandwhite, noiretblanc, bianconero, blancoynegro		

Figure 1: **Left:** Examples of text generated from a Deep Boltzmann Machine by sampling from $P(\mathbf{v}_{txt}|\mathbf{v}_{img};\theta)$. **Right:** Examples of images retrieved using features generated from a Deep Boltzmann Machine by sampling from $P(\mathbf{v}_{img}|\mathbf{v}_{txt};\theta)$.

Multimodal DBM 应用

4.1.2 协同表征

协同表征 (Coordinated Representation) 将多模态中的每个模态分别映射到各自的表示空间，但映射后的向量之间满足一定的相关性约束（例如线性相关）。Coordinated结构并不寻求融合而是建模多种模态数据间的**相关性**，它将多个(通常是两个)模态映射到协作空间，表示为： $f(x_1) \sim g(x_2)$ ，其中 \sim 表示一种协作关系。网络的优化目标是这种协作关系(通常是相似性，即最小化cosine距离等度量)。



(b) Coordinated representations

Coordinated Representation

Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models

(NIPS 2014) 利用协同学习到的特征向量之间满足加减算数运算这一特性，可以搜索出与给定图片满足“指定的转换语义”的图片。例如：狗的图片特征向量 - 狗的文本特征向量 + 猫的文本特征向量 = 猫的图片特征向量 -> 在特征向量空间，根据最近邻距离，检索得到猫的图片。

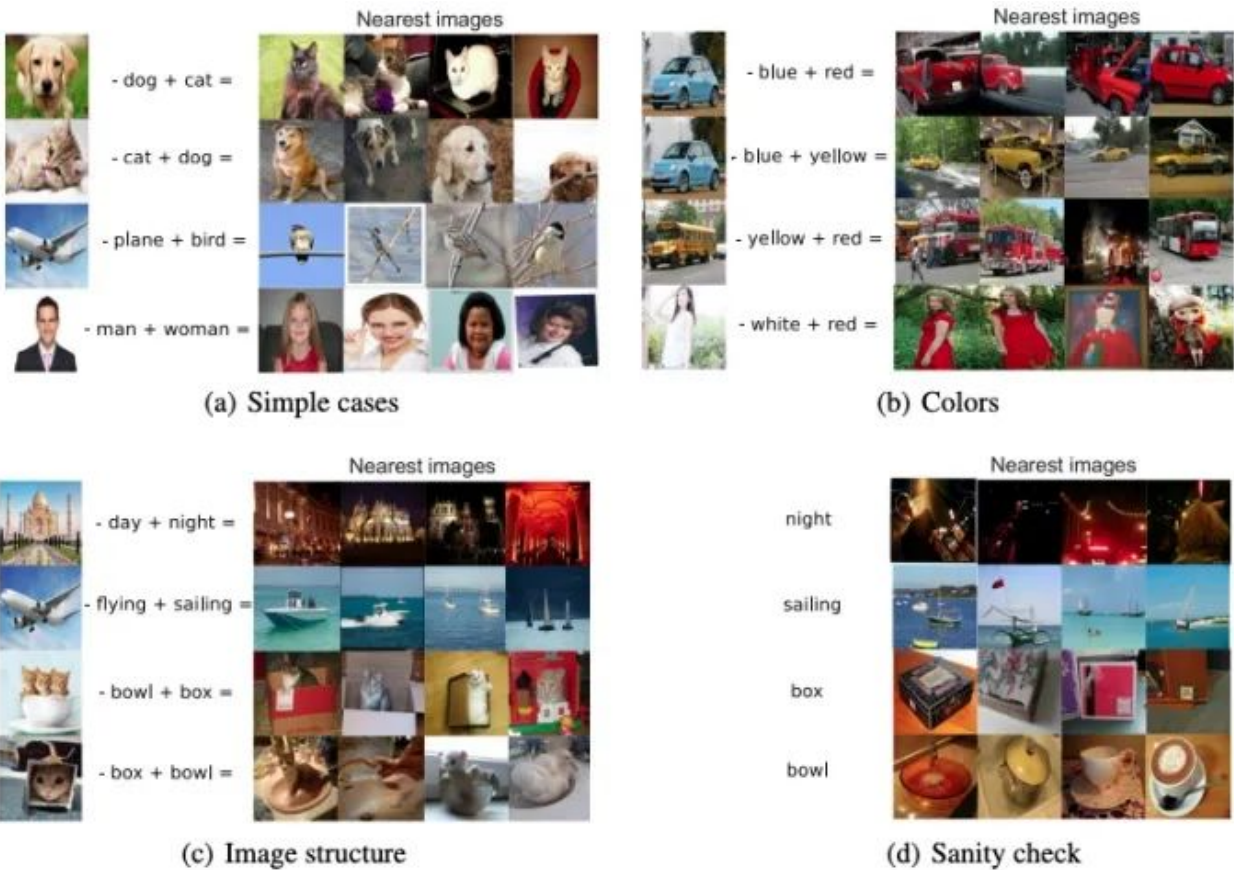


Figure 4: Multimodal vector space arithmetic. Query images were downloaded online and retrieved images are from the SBU dataset.

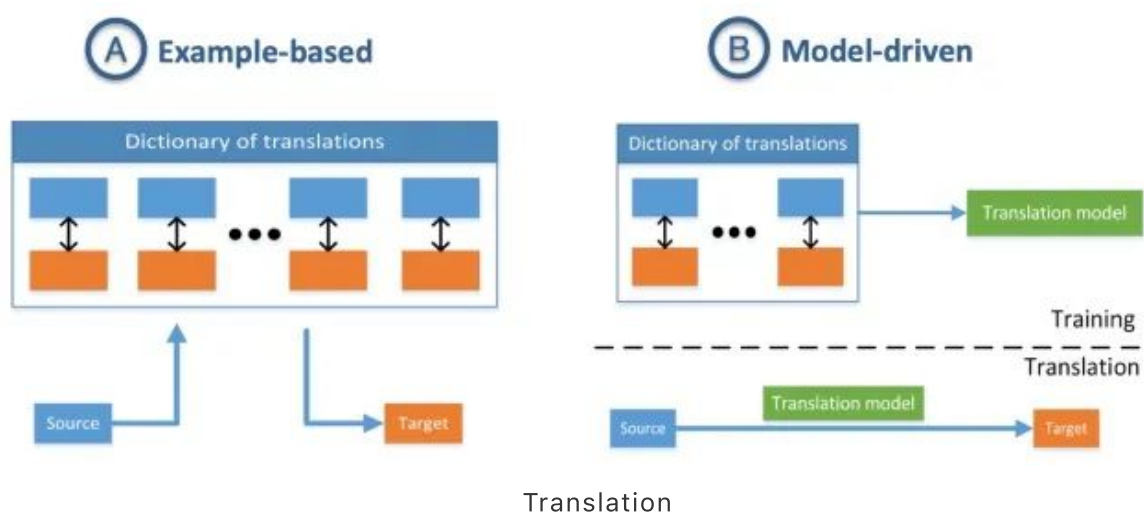
4.2 翻译Translation

第二个挑战涉及如何将数据从一种模式转化（映射）到另一种模式。不仅数据是异构的，而且模态之间的关系通常是开放式的或主观的。例如，存在多种描述图像的正确方法，并且可能不存在一种完美的翻译。

4.2.1 常见应用

- **机器翻译（Machine Translation）**：将输入的语言A（即时）翻译为另一种语言B。类似的还有唇读（Lip Reading）和语音翻译（Speech Translation），分别将唇部视觉和语音信息转换为文本信息。
- **图片描述（Image captioning）或者视频描述（Video captioning）****：对给定的图片/视频形成一段文字描述，以表达图片/视频的内容。
- **语音合成（Speech Synthesis）**：根据输入的文本信息，自动合成一段语音信号。

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



4.2.2 基于实例的方法

基于实例的方法从词典中**检索**最佳翻译，词典一般指训练集中的数据对 $\{(x_1, y_1), \dots, (x_N, y_N)\}$ 。给定测试样本 \hat{x} ，模版法直接检索在词典中找到最匹配的翻译结果 y_i

，并将其作为最终输出。

检索可分为单模态检索或跨模态检索：

- **单模态检索**首先找到与 \hat{x} 最相似的 x_i ，然后获得 x_i 对应的 y_i ；
- **多模态检索**直接在 $\{y_1, \dots, y_N\}$ 集合中检索到与 \hat{x} 最相似的 y_i ，性能通常优于单模态检索。

为进一步增强检索结果的准确性，可选择top-K的检索结果 $\{y_{i_1}, y_{i_2}, \dots, y_{i_k}\}$ ，再融合K个结果作为最终输出。

4.2.3 模型驱动的方法

基于模型的首先在字典上训练一个翻译模型，然后使用该模型进行翻译。

- **基于语法的模型 (Grammar-based models)** 即人为设定多个针对目标模态的语法模版，将模型的预测结果插入模版中作为翻译结果。以图像描述为例，模版定义为who did what to whom in a place，其中有四个待替换的插槽。通过不同类型的目标/属性/场景检测器可以获得who, what, whom, place等具体单词，进而完成翻译。
- **编码-解码器模型 (Encoder-decoder models)** 首先将源模态的数据编码为隐特征 z ，后续被解码器用于生成目标模态。以图像描述为例，编码器(一般为CNN+spatial pooling)将图像编码为一个或多个特征向量，进而输入到RNN中以自回归的方式生成单词序列。
- **连续型生成模型 (Continuous generation models)** 它针对源模态与目标模态都为流数据且在时间上严格对齐的任务。以文本合成语音为例，它与图像描述不同，语音数据与文本数据在时间上严格对齐。WaveNet采用了CNN并行预测解决该类问题，当然，编码-解码器理论上也可完成该任务，但需处理数据对齐问题。

4.2.4 翻译的评估困境

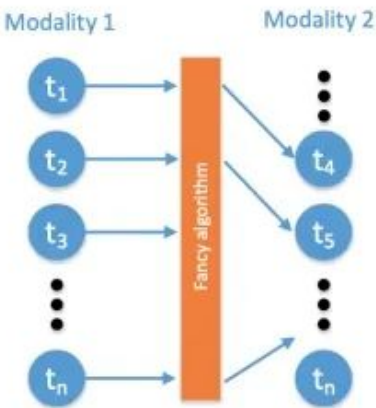
多模态翻译方法面临的一个主要挑战是它们很难评估。语音识别等任务只有一个正确的翻译，而语音合成和媒体描述等任务则没有。有时，就像在语言翻译中，多重答案是正确的，决定哪个翻译更好往往是主观的。

- 人工评价是最理想的评估，但是耗时耗钱，且需要多样化打分人群的背景以避免偏见。
- 自动化指标是视觉描述领域常用的替代方法，包括BLEU，Meteor，CIDEr，ROUGE等，但它们被证实与人的评价相关性较弱。
- 基于检索的评估和弱化任务(例如：将图像描述中一对多映射简化为VQA中一对一的映射)也是解决评估困境的手段。

4.3 对齐Alignment

第三个挑战是从两种或多种不同的模态中识别（子）元素之间的直接关系。例如，我们可能希望将食谱中的步骤与显示正在制作的菜肴的视频对齐。为了应对这一挑战，我们需要测量不同模式之间的相似性并处理可能的长期依赖和歧义。

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



A Explicit Alignment

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

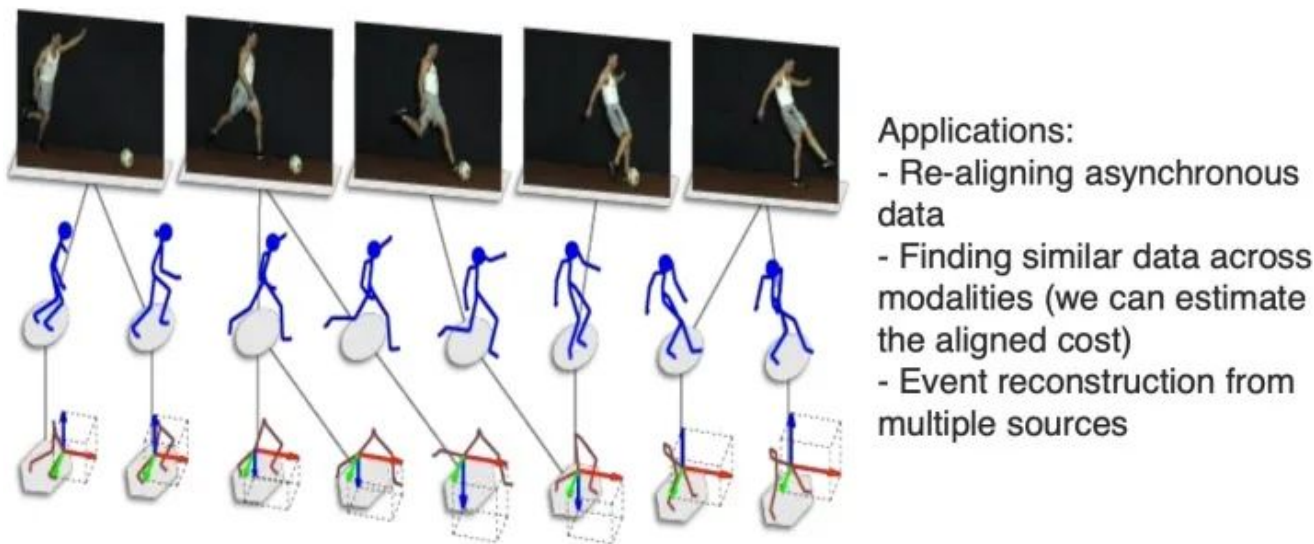
B Implicit Alignment

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

Alignment

4.3.1 显式对齐

如果模型的主要目标是对齐来自两个或多个模态的子元素，那么我们将其分类为执行显式对齐。显式对齐的一个重要工作是相似性度量。大多数方法都依赖于度量不同模态的子组件之间的相似性作为基本构建块。



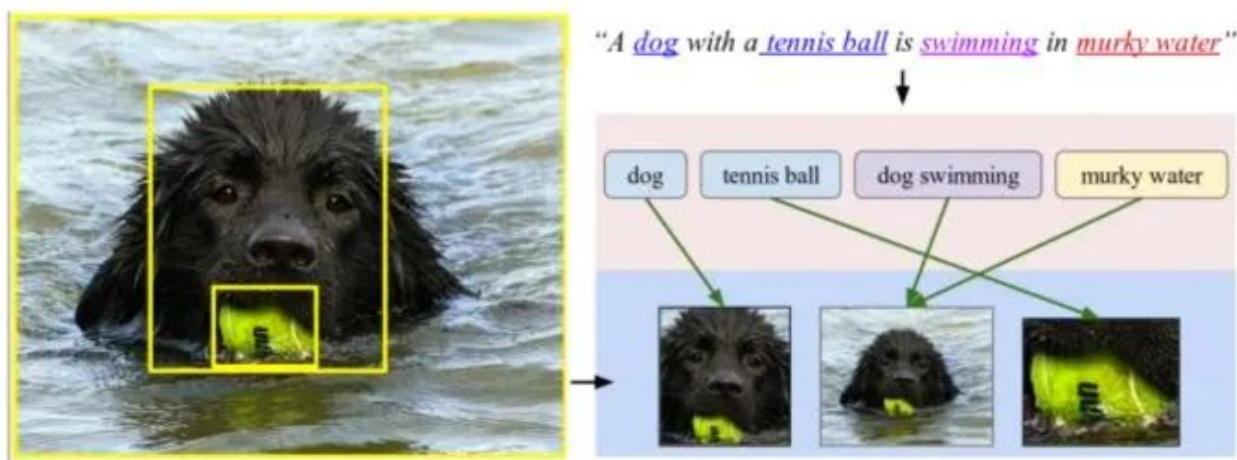
显式对齐

包括无监督和弱监督的方法：

- **无监督对齐**：给定两个模态的数据作为输入，希望模型实现子元素的对齐，但是训练数据没有“对齐结果”的标注，模型需要同时学习相似度度量和对齐方式。
- **有监督对齐**：有监督方法存在标注，可训练模型学习相似度度量。

4.3.2 隐式对齐

隐式对齐用作另一个任务的中间(通常是潜在的)步骤。这允许在许多任务中有更好的表现，包括语音识别、机器翻译、媒体描述和视觉问题回答。这些模型不显式地对齐数据，也不依赖于监督对齐示例，而是学习如何在模型训练期间潜在地对齐数据。



Karpathy et al., Deep Fragment Embeddings for Bidirectional Image Sentence Mapping,
<https://arxiv.org/pdf/1406.5679.pdf>

隐式对齐

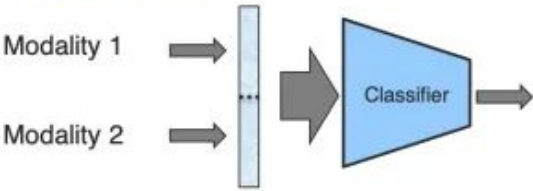
4.4 融合Fusion

第四个挑战是结合来自两个或多个模态的信息来执行预测。例如，对于视听语音识别，将嘴唇运动的视觉描述与语音信号融合以预测口语。来自不同模态的信息可能具有不同的预测能力和噪声拓扑，并且可能在至少一种模态中丢失数据。

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

A Model-Agnostic Approaches

1) Early Fusion



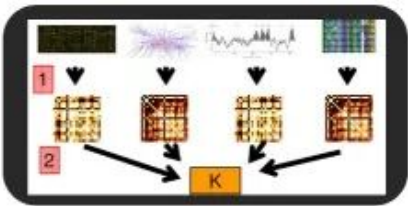
2) Late Fusion



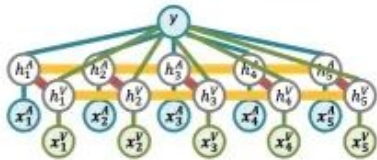
Fusion

B Model-Based (Intermediate) Approaches

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



Multiple kernel learning



Multi-View Hidden CRF

Fusion

4.4.1 模型无关的方法

- **早期融合（Early Fusion）**：指在模型的浅层(或输入层)将多个模态的特征拼接起来，然后再级联深度网络结构，最后接上分类器或其他模型。Early Fusion 是学者对多模态融合的早期尝试，通过将各模态的底层特征进行融合学习相关性，由于只需要训练一个共同的模型，复杂度可控。但是，由于多个模态的数据来源不一致，会给拼接造成很大的难度，并且直接对原始数据进行拼接会引起较大的特征维度，对数据预处理也非常敏感。

- **晚期融合 (Late Fusion)**：独立训练多个模型，在预测层(最后一层)进行融合，可以理解为集成方法 Ensemble Methods 的一种。Late Fusion 方式的各模态单独处理，特征独立互不影响，即使某个模态信息丢失也可以正常训练，具有很强的灵活性。但是，该方式没有充分利用模态间底层特征的相关性，并且由于涉及多个模态的分别训练，也会带来较大的计算复杂度。
- **混合融合 (Hybird Fusion)**：同时结合前融合和后融合，以及在模型中间层进行特征交互。Hybird Fusion是一种**逐级融合方式**，在不同层级上依次对不同模态进行融合，综合了上述两种方式的优点，既利用了模态间信息的相关性，也具有一定的灵活性，目前大部分多模态融合都是采用这种方法。

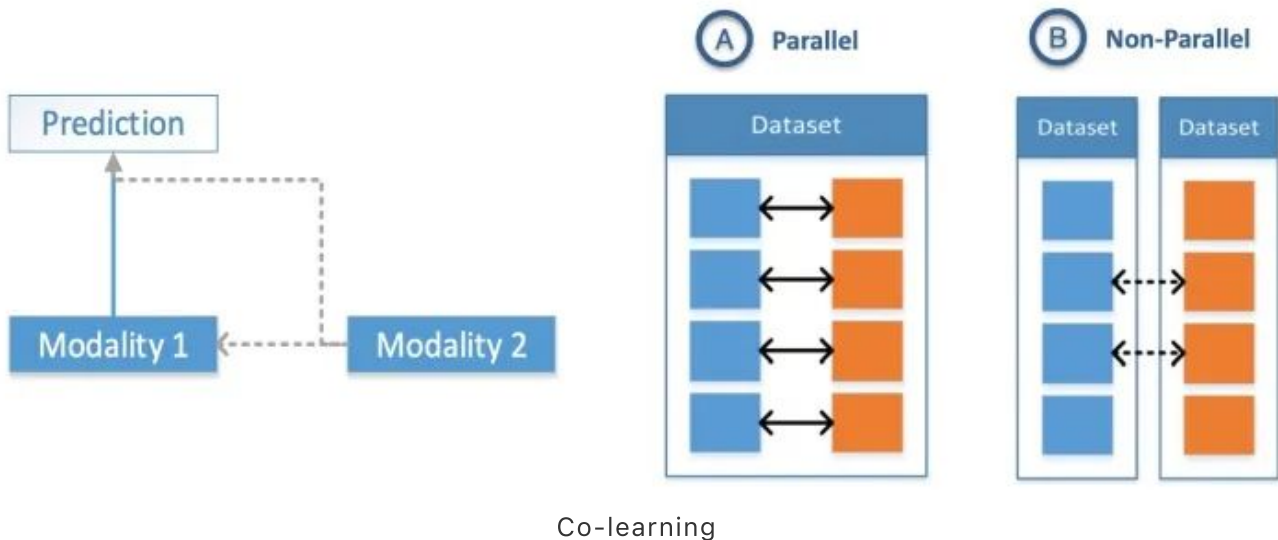
4.4.2 基于模型的方法

- **Deep Neural Networks**：神经网络进行端到端的训练，使用LSTM、卷积层、注意力层、门机制、双线性融合等设计序列数据或图像数据的复杂交互。
- **Multiple Kernel learning**：多核学习（将不同的核用于不同的数据模态/视图）
- **Graphical models**：利用隐马尔可夫模型或贝叶斯网络建模数据的联合概率分布(生成式)或条件概率(判别式)

4.5 协同学习Co-learning

第五个挑战是在模态的表示和它们的预测模型之间转移知识。协同学习探索了**从一种模态中学习的知识如何帮助在不同模态上训练的计算模型**。当其中一种模式的资源有限（例如，带注释的数据）时，这一挑战尤其重要。辅助模态（helper modality）通常只参与模型的训练过程，并不参与模型的测试使用过程

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



4.5.1 并行

需要训练数据集，其中来自一种模态的观察结果与来自其他模态的观察结果直接相关，例如在一个视听语音数据集中，视频和语音样本来自同一个说话者。

4.5.2 非并行

不需要来自不同模式的观察结果之间的直接联系，通常通过使用类别重叠来实现共同学习，例如，在零样本学习中，使用来自Wikipedia的纯文本数据集扩展传统的视觉对象识别数据集以改进视觉对象识别的泛化能力。

4.5.3 混合

通过共享模式或数据集桥接

五、SOTA模型 - CLIP

CLIP全称Contrastive Language-Image Pre-training，是OpenAI最新的一篇NLP和CV结合的多模态的工作，在多模态领域迈出了重要的一步。**CLIP在无需利用ImageNet的数据和标签进行训练的情况下，就可以达到ResNet50在ImageNet数据集上有监督训练的结果。**

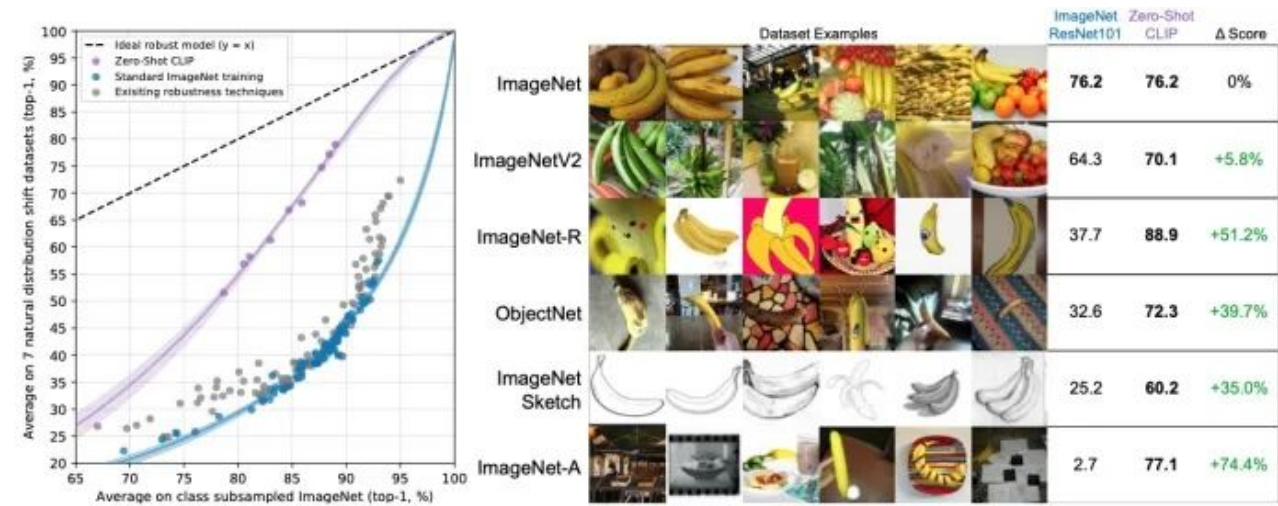


Figure 13. Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models. (Left) An ideal robust model (dashed line) performs equally well on the ImageNet distribution and on other natural image distributions. Zero-shot CLIP models shrink this “robustness gap” by up to 75%. Linear fits on logit transformed values are shown with bootstrap estimated 95% confidence intervals. (Right) Visualizing distribution shift for bananas, a class shared across 5 of the 7 natural distribution shift datasets. The performance of the best zero-shot CLIP model, ViT-L/14@336px, is compared with a model that has the same performance on the ImageNet validation set, ResNet-101.

CLIP Zero shot

CLIP主要的贡献就是利用无监督的文本信息，作为监督信号来学习视觉特征。

5.1 原理

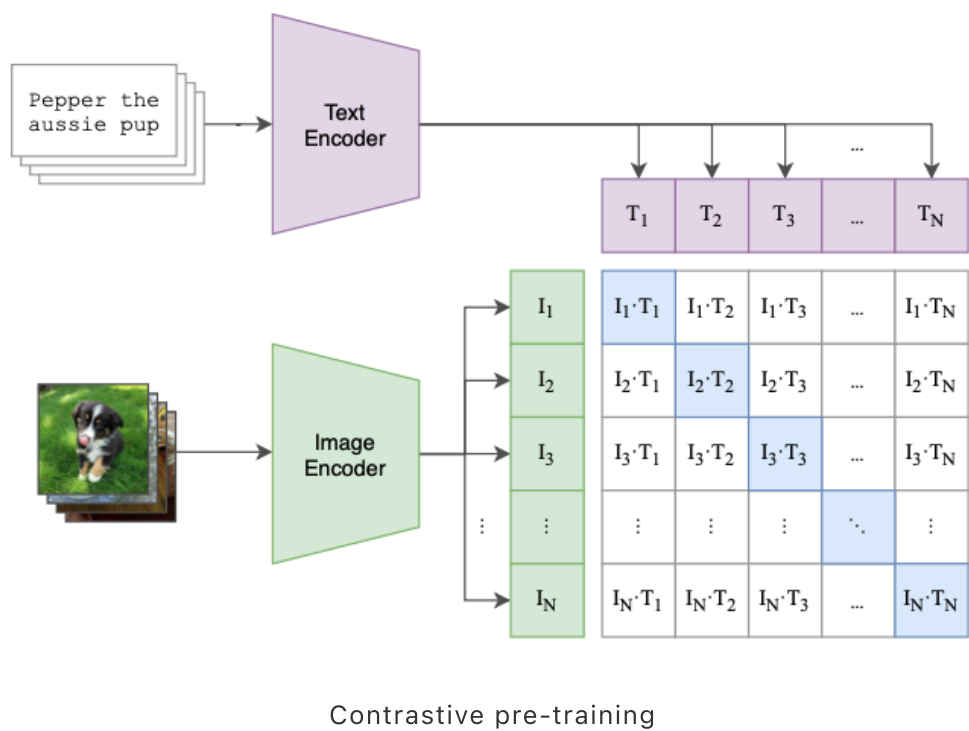
CLIP不预先定义图像和文本标签类别，直接利用从互联网爬取的 400 million 个image-text pair 进行图文匹配任务的训练，并将其成功迁移应用于30个现存的计算机视觉分类。



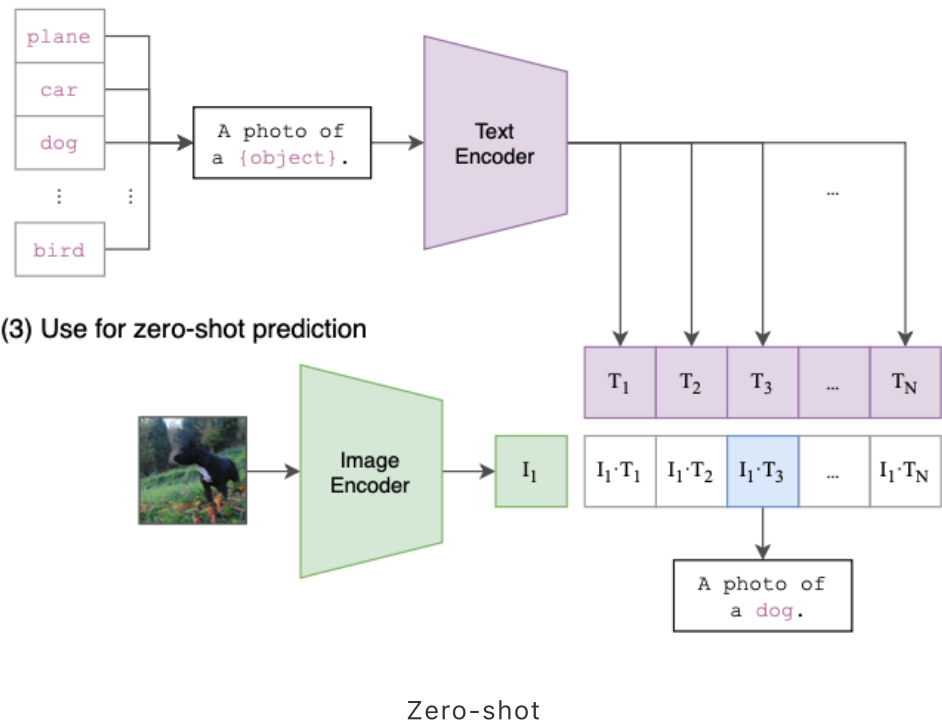
5.2 流程

- **Contrastive pre-training**: 预训练阶段，使用图片 - 文本对进行对比学习训练；
- **Create dataset classifier from label text**: 提取预测类别文本特征；
- **Use for zero-shot prediction**: 进行 Zero-Shoot 推理预测；

(1) Contrastive pre-training



(2) Create dataset classifier from label text



阶段1： Contrastive pre-training 在预训练阶段，对比学习十分灵活，只需要定义好 正样本对 和 负样本对 就行了，其中能够配对的 image-text 对即为正样本。具体来说，先分别对图像和文本提特征，这时图像对应生成 $I_1, I_2 \dots I_n$ 的特征向量（Image Feature），文本对应生成 $T_1, T_2 \dots T_n$ 的特征向量（Text Feature），中间对角线为正样本，其余均为负样本。

阶段2： Create dataset classifier from label text 基于400M数据上学得的先验，仅用数据集的标签文本，就可以得到很强的图像分类性能。现在训练好了，然后进入前向预测阶段，通过 prompt label text 来创建待分类的文本特征向量。

阶段3: Use for zero-shot prediction 最后就是推理见证效果的时候，对于测试图片，选择相似度最大的那个类别输出。在推理阶段，无论来了张什么样的图片，只要扔给 Image Encoder 进行特征提取，会生成一个一维的图片特征向量，然后拿这个图片特征和 N 个文本特征做余弦相似度对比，最相似的即为想要的那个结果，比如这里应该会得到 “A photo of a guacamole.”，

5.3 实现

```
# image_encoder - ResNet or Vision Transformer
# text_encoder - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l] - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T) #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss = (loss_i + loss_t)/2
```

Numpy-like pseudocode for the core of an implementation of CLIP.

5.4 后续

StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery (ICCV 2021 Oral) StyleCLIP是结合CLIP和StyleGAN的一个工作，通过文字上的改变，从而去引导图像的生成。<https://github.com/orpatashnik/StyleCLIP>

![StyleCLIP 例子]([https://cdn.jsdelivr.net/gh/cynthia-yawian/upic@main/uPic/tutieshi_640x360_14s \(1\).gif](https://cdn.jsdelivr.net/gh/cynthia-yawian/upic@main/uPic/tutieshi_640x360_14s%20(1).gif))



CLIPDraw: Exploring Text-to-Drawing Synthesis through Language-Image Encoders CLIPDraw也是利用文字来指导图像生成的一个工作，只是想法更加简单，不需要进行模型训练，而是使用预训练的 CLIP 语言图像编码器作为度量，以最大化给定描述和生成的绘图之间的相似性，最后就可以生成很多简笔画的图像。



ViLD: Open-vocabulary Object Detection via Vision and Language Knowledge Distillation / Google 用CLIP来做物体检测和分割的任务，在CLIP出来一个月半月以后，就Google就出了这篇文章。作者指出，如果你用传统的物体检测方法，算法只能告诉你这些只是玩具，也就是下图蓝色的base categories，但是当你利用了这种自然语言之后，你就拜托了基础类的这个限制，就可以检测出来新的类，也就是红色的noval categories。

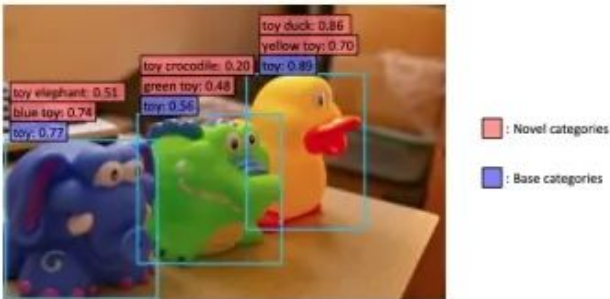
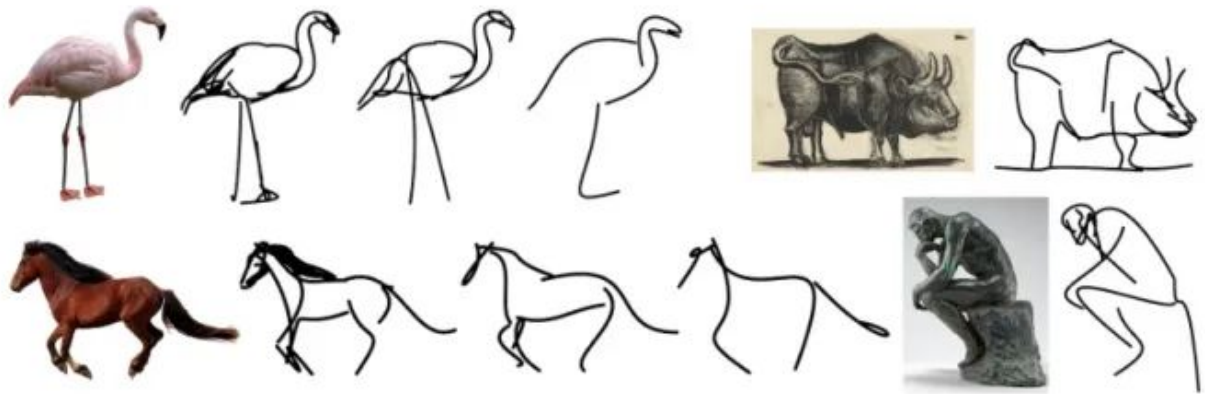


Figure 1: An example of our open-vocabulary detector with arbitrary texts. After training on base categories (purple), we can detect novel categories (pink) that are not present in the training data.

ViLD

CLIPasso: Semantically-Aware Object Sketching (SIGGRAPH 2022 Best Paper Award) 用CLIP提炼语义概念，生成图片目标的高度抽象线条画(速写)

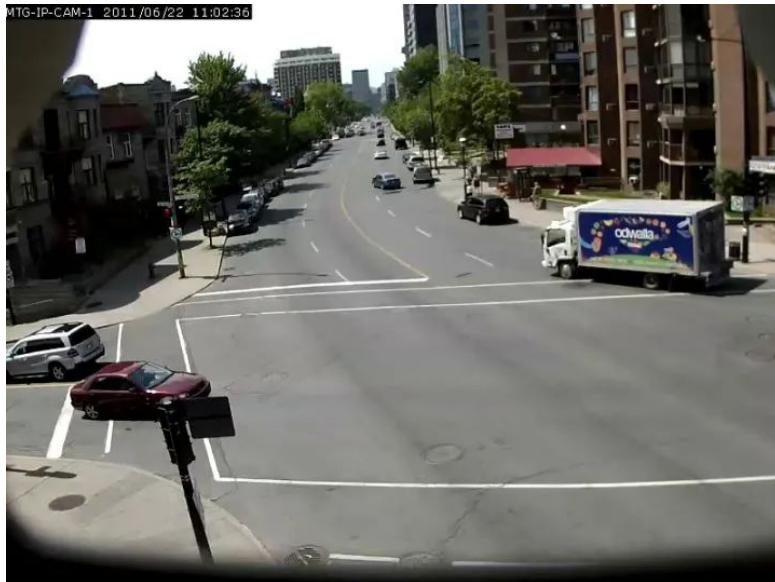


Our work converts an image of an object to a sketch, allowing for varying levels of abstraction, while preserving its key visual features. Even with a very minimal representation (the rightmost flamingo and horse are drawn with only a few strokes), one can recognize both the semantics and the structure of the subject depicted.

CLIPasso

应用：Contrastive Language-Image Forensic Search <https://github.com/johanmodin/clifs> 使用CLIP完成视频检索，看一个视频里面有没有出现过一个人或者一些场景，通过直接输入文本的这种形式进行检索。

A truck with the text "odwalla"



A truck with the text "odwalla"

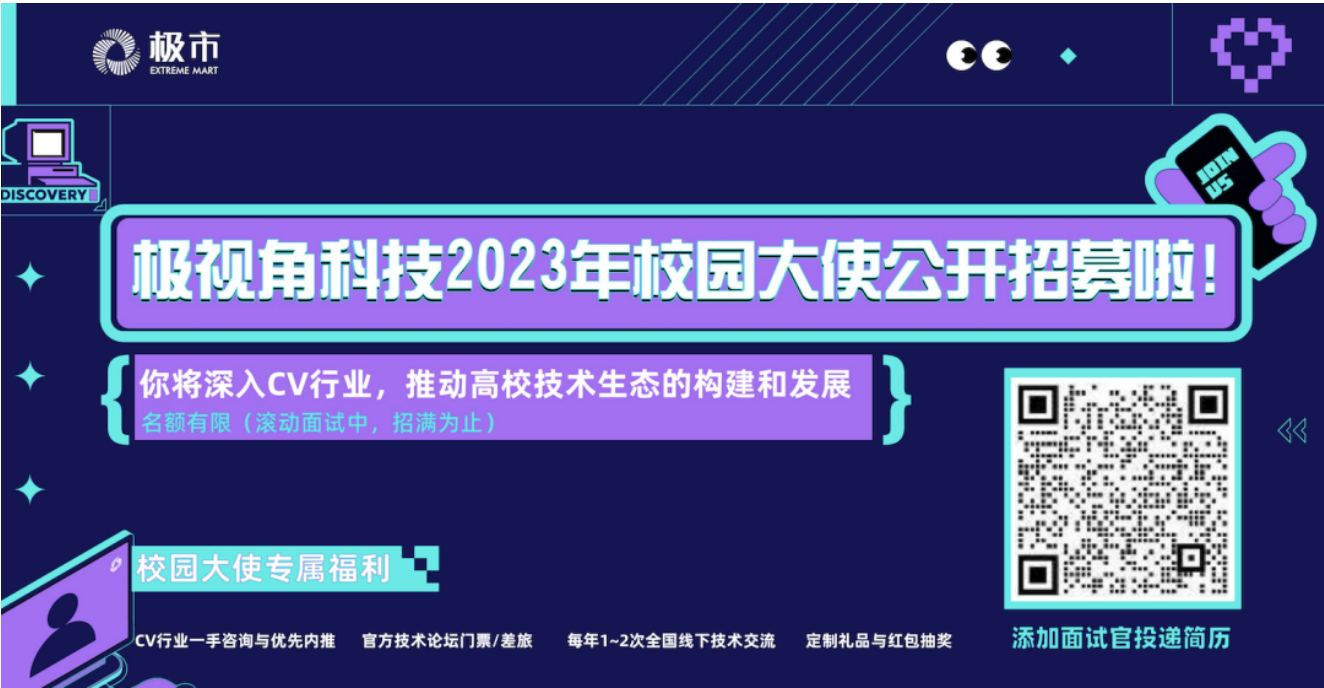
A white BMW car



A white BMW car

参考资料

- [1]多模态学习方法综述 / 陈鹏 / 工程科学学报 / 2019
- [2]Multimodal Machine Learning: A Survey and Taxonomy / Tadas Baltrusaitis / 2017
- [3]MultiModal Machine Learning / Louis-Philippe Morency / CMU
- [4]CMU-10707 第二十一讲 多模态机器学习 / 华年ss / 知乎
- [5]Multimodal Learning with Deep Boltzmann Machines / Nitish Srivastava / 2012
- [6]多模态学习综述及最新方向 / yougeii / 知乎
- [7]Learning transferable visual models from natural language supervision / 2021
- [8]CLIP: Connecting Text and Images / openai / blog
- [9]Awesome-CLIP / yzhuoning



公众号后台回复“极市直播”获取100+期极市技术直播回放+PPT



极市平台

为计算机视觉开发者提供全流程算法开发训练平台，以及大咖技术分享、社区交流、竞...
848篇原创内容

公众号

极市干货

技术干货：损失函数技术总结及Pytorch使用示例 | 深度学习有哪些trick？ | 目标检测正负样本区分策略和平衡策略总结

实操教程：GPU多卡并行训练总结（以pytorch为例） | CUDA WarpReduce 学习笔记 | 卷积神经网络压缩方法总结



极市CVPR2023交流群已成立



群聊: CVPR2023 论文交流群



该二维码7天内(3月7日前)有效, 重新进入将更新

△长按扫码进群

添加极市小助手微信 (ID : cvmart4)

备注: 姓名-学校/公司-研究方向-城市 (如: 小极-北大-目标检测-深圳)

即可申请加入极市 [目标检测](#)/[图像分割](#)/[工业检测](#)/[人脸](#)/[医学影像](#)/[3D/SLAM](#)/[自动驾驶](#)/[超分辨率](#)/[姿态估计](#)/[ReID](#)/[GAN](#)/[图像增强](#)/[OCR](#)/[视频理解](#)等技术交流群

每月大咖直播分享、真实项目需求对接、求职内推、算法竞赛、干货资讯汇总、与 **10000+**来自港科大、北大、清华、中科院、CMU、腾讯、百度等名校名企视觉开发者互动交流~



极市平台

为计算机视觉开发者提供全流程算法开发训练平台, 以及大咖技术分享、社区交...
848篇原创内容

公众号

△点击卡片关注极市平台, 获取最新CV干货

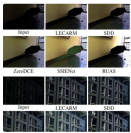
[点击阅读原文进入CV社区](#)

[收获更多技术干货](#)

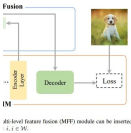
[阅读原文](#)

喜欢此内容的人还喜欢

ICCV23 | 将隐式神经表征用于低光增强，北大张健团队提出NeRCo
极市平台



ICCV 2023 | Pixel-based MIM: 简单高效的多级特征融合自监督方法
极市平台



ICCV 2023 | 南开程明明团队提出适用于SR任务的新颖注意力机制（已开源）
极市平台

