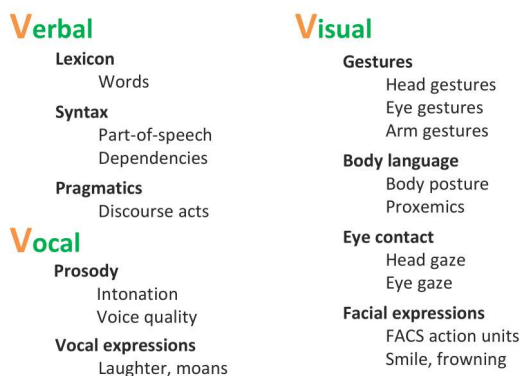


# 1 什么是多模态

模态是指人接受信息的特定方式，如文字、声音、图片视频、K线图。来自单一模态的信息无法全面地理解事物，关联来自多种模型的信息能做出更好的判断。

模态包括但不限于下列形式：



## 2 技术演进路径

多模态学习从 1970 年代起步，经历了几个发展阶段，在 2010 后全面步入 Deep Learning 阶段：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 ...)。

在深度学习阶段的主要研究方向为

### Representation learning (a.k.a. deep learning)

- 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]

### Key enablers for multimodal research:

- New large-scale multimodal datasets
- Faster computer and GPUS
- High-level visual features
- “Dimensional” linguistic features

主要挑战是情绪识别、图片打标题、视频描述、问答。

### Audio-Visual Emotion Challenge (AVEC, 2011- )



- Emotional dimension estimation
- Standardized training and test sets
- Based on the SEMAINE dataset

### Emotion Recognition in the Wild Challenge (EmotiW 2013- )



- Emotional dimension estimation
- Standardized training and test sets
- Based on the SEMAINE dataset

实际应用中的任务有情绪识别、给多媒体添加描述、事件识别。

#### Affect recognition

- Emotion
- Persuasion
- Personality traits

#### Media description

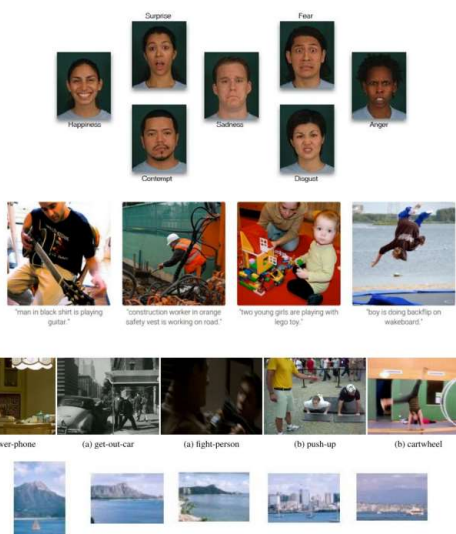
- Image captioning
- Video captioning
- Visual Question Answering

#### Event recognition

- Action recognition
- Segmentation

#### Multimedia information retrieval

- Content based/Cross-media



## 3 多模态机器学习的核心任务

经过广泛调研，有人总结出了多模态机器学习的 5 个核心任务（论文：Multimodal Machine Learning: A Survey and Taxonomy）。即

- 多模态表示学习（Representation）：如何将多个模态数据所蕴含的语义信息数值化为实值向量。
- 模态间映射（mapping）：如何将某一特定模态数据中的信息映射至另一模态。
- 对齐(Alignment)：如何识别不同模态之间的部件、元素的对应关系。
- 融合(fusion)：主要研究如何整合不同模态间的模型与特征。
- 协同学习(co-learning)：主要研究如何将信息富集的模态上学习的知识迁移到信息匮乏的模态，使各个模态的学习互相辅助。典型的方法包括多模态的零样本学习、领域自适应等

### Multimodal Machine Learning: A Survey and Taxonomy

By Tadas Baltrušaitis, Chaitanya Ahuja,  
and Louis-Philippe Morency

<https://arxiv.org/abs/1705.09406>

- ✓ 5 core challenges
- ✓ 37 taxonomic classes
- ✓ 253 referenced citations

举例解释上述 5 个任务。

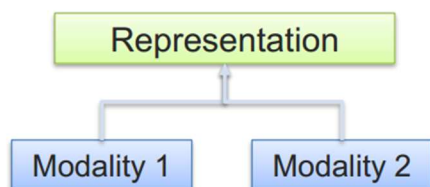
### 3.1 多模态表示学习

单模态的表示学习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，而多模态表示学习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。主要包括两大研究方向：联合表示（Joint Representations）和协同表示（Coordinated Representations）。

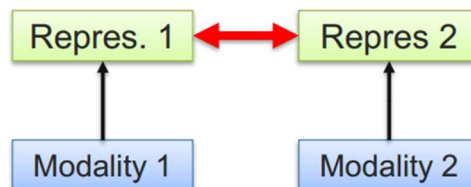
联合表示将多个模态的信息一起映射到一个统一的多模态向量空间；

协同表示负责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）。

### Ⓐ Joint representations:

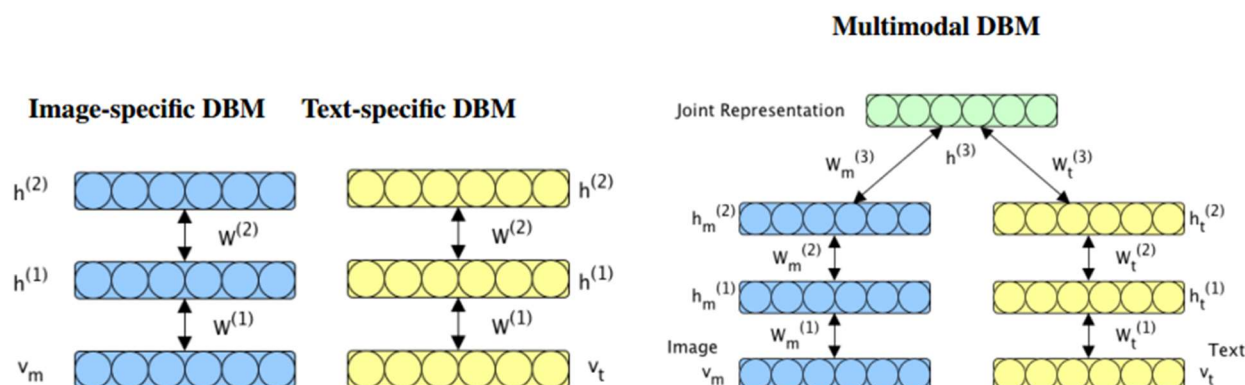


### Ⓑ Coordinated representations:















利用多模态表示学习到的特征可以用来做信息检索，也可以用于的分类/回归任务。下面列举几个经典的应用。

在来自 NIPS 2012 的《Multimodal learning with deep boltzmann machines》一文中提出将 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。

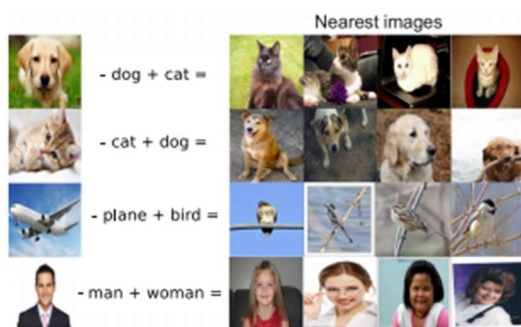


论文中的实验通过 Bimodal DBM，学习图片和文本的联合概率分布  $P(\text{图片}, \text{文本})$ 。在应用阶段，输入图片，利用条件概率  $P(\text{文本}|\text{图片})$ ，生成文本特征，可以得到图片相应的文本描述；而输入文本，利用条件概率  $P(\text{图片}|\text{文本})$ ，可以生成图片特征，通过检索出最靠近该特征向量的两个图片实例，可以得到符合文本描述的图片。如下图所示：

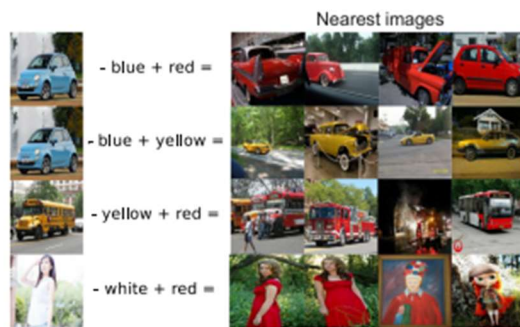
| Image   | Given Tags   | Generated Tags   | Input Text  | 2 nearest neighbours to generated image features                                   |   |
|---|--|--|---|--|---|
|  | pentax, k10d, kangaroosland, southaustralia, sa, australia, australiansealion, 300mm | beach, sea, surf, strand, shore, wave, 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, blackandwhite, woman, 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, noiret blanc, biancoenero, biancoynegro              |  |  |

协同表示学习一个比较经典且有趣的应用是来自于《Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models》这篇文章。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



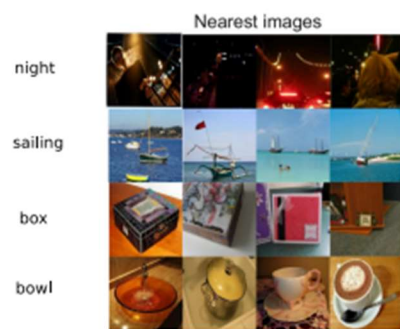
(a) Simple cases



(b) Colors



(c) Image structure



(d) Sanity check



## 3.2 映射 Mapping

转化也称为映射，负责将一个模态的信息转换为另一个模态的信息。常见的应用包括：

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。



"man in black shirt is playing guitar."



"construction worker in orange safety vest is working on road."



"two young girls are playing with  
lego toy."

**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。

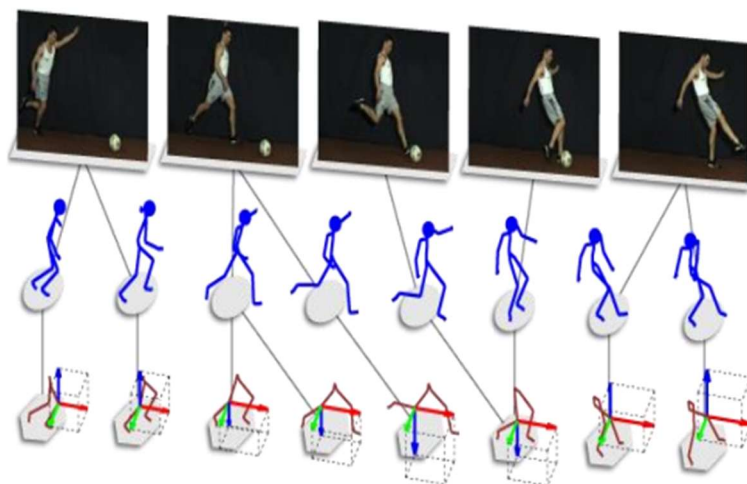


模态间的转换主要有两个难点，一个是 open-ended，即未知结束位，例如实时翻译中，在还未得到句尾

的情况下，必须实时的对句子进行翻译；另一个是 **subjective**，即主观评判性，是指很多模态转换问题的效果没有一个比较客观的评判标准，也就是说目标函数的确定是非常主观的。例如，在图片描述中，形成怎样的一段话才算是图片好的诠释？也许一千个人心中有一千个哈姆雷特吧。

### 3.3 对齐 Alignment

多模态的对齐负责对来自同一个实例的不同模态信息的子分支/元素寻找对应关系。这个对应关系可以是时间维度的，比如下图所示的 **Temporal sequence alignment**，将一组动作对应的视频流同骨骼图片对齐。类似的还有电影画面-语音-字幕的自动对齐。



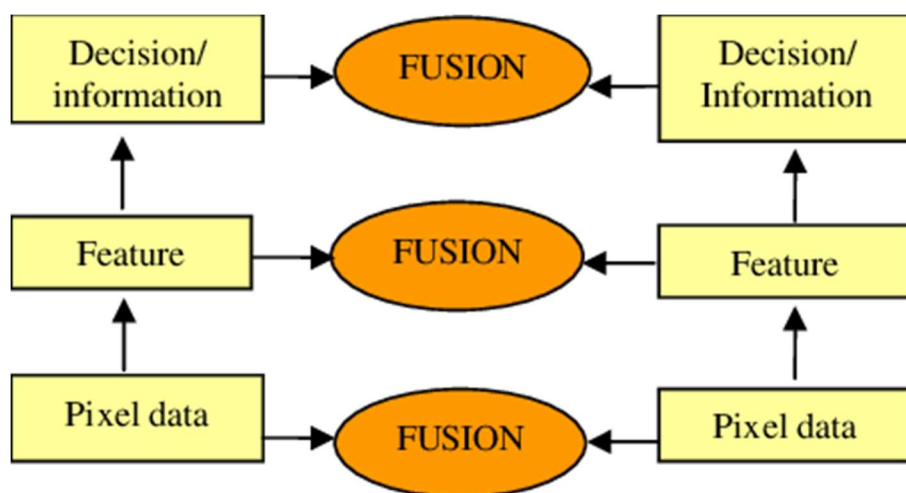
对齐又可以是空间维度的，比如图片语义分割（**Image Semantic Segmentation**）：尝试将图片的每个像素对应到某一种类型标签，实现视觉-词汇对齐。



### 3.4 多模态融合

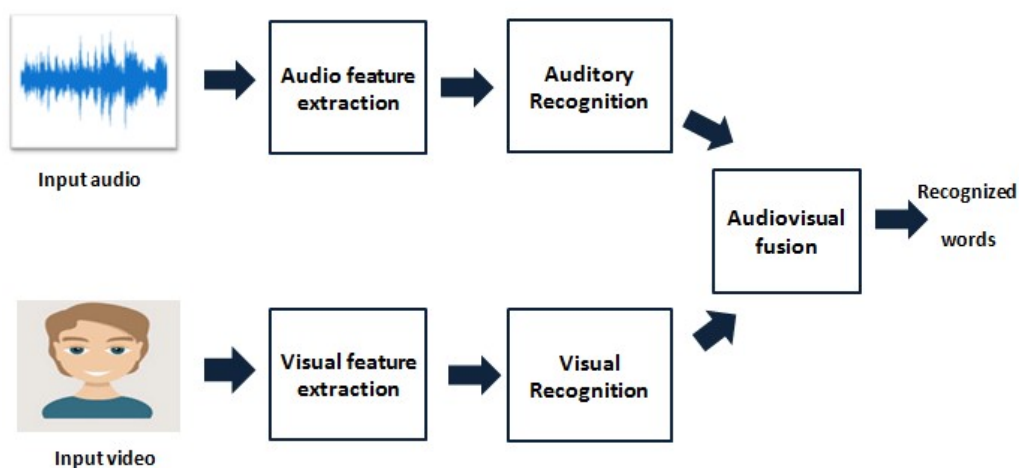
多模态融合（**Multimodal Fusion**）负责联合多个模态的信息，进行目标预测（分类或者回归），属于 **M** 多模态学习的最早的研究方向之一，也是目前应用最广的方向，它还存在其他常见的别名，例如多源信息融合（**Multi-source Information Fusion**）、多传感器融合（**Multi-sensor Fusion**）。

按照融合的层次，可以将多模态融合分为 **pixel level**，**feature level** 和 **decision level** 三类，分别对应对原始数据进行融合、对抽象的特征进行融合和对决策结果进行融合。而 **feature level** 又可以分为 **early** 和 **late** 两个大类，代表了融合发生在特征抽取的早期和晚期。当然还有将多种融合层次混合的 **hybrid** 方法。

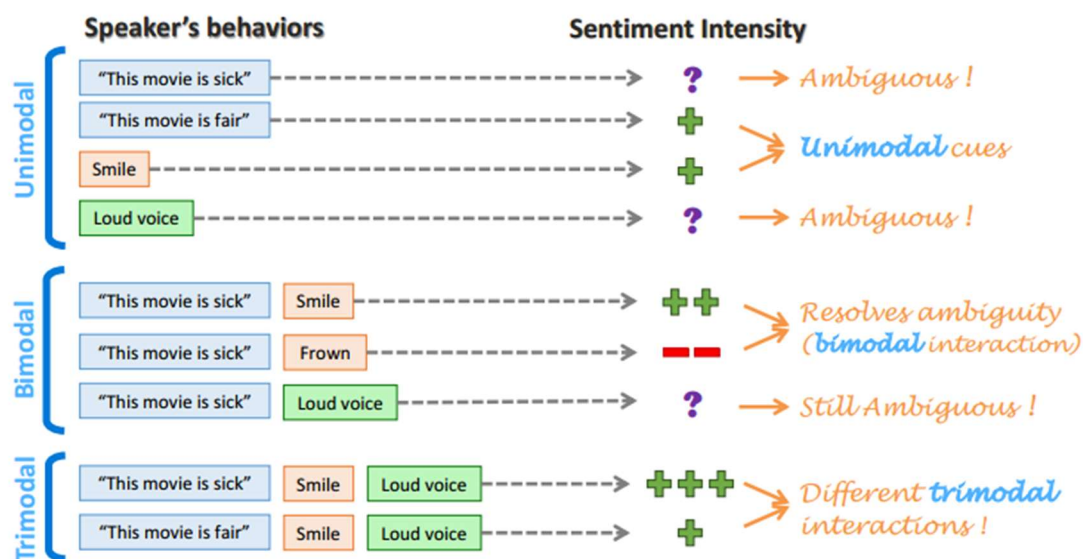


常见的机器学习方法都可以应用于多模态融合，下面列举几个比较热门的研究方向。

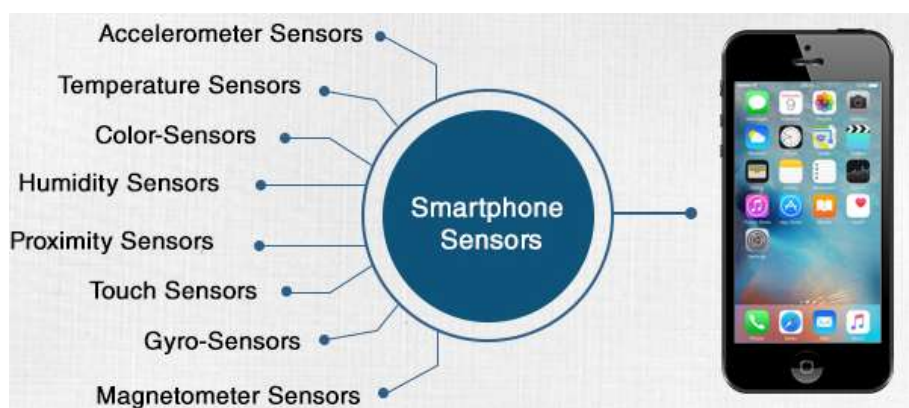
**视觉-音频识别 (Visual-Audio Recognition):** 综合源自同一个实例的视频信息和音频信息，进行识别工作。



**多模态情感分析 (Multimodal sentiment analysis):** 综合利用多个模态的数据（例如下图中的文字、面部表情、声音），通过互补，消除歧义和不确定性，得到更加准确的情感类型判断结果。



手机身份认证（Mobile Identity Authentication）：综合利用手机的多传感器信息，认证手机使用者是否是注册用户。



多模态融合研究的难点主要包括如何判断每个模态的置信水平、如何判断模态间的相关性、如何对多模态的特征信息进行降维以及如何对非同步采集的多模态数据进行配准等。

若想了解传统的机器学习方法在此领域的应用，推荐学习清华大学出版的《多源信息融合》（韩崇昭等著）一书。

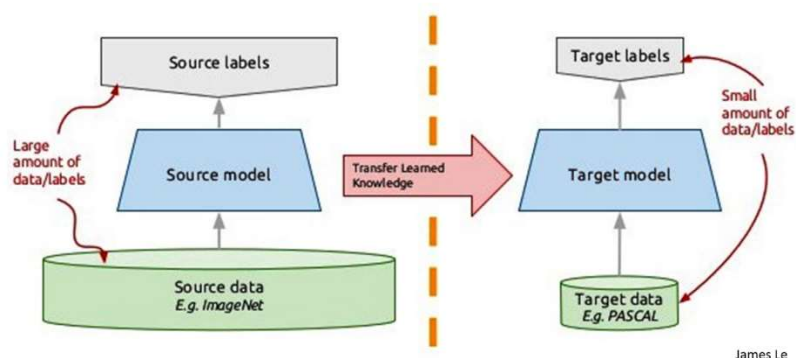
### 3.5 协同学习 Co-learning

协同学习是指使用一个资源丰富的模态信息来辅助另一个资源相对贫瘠的模态进行学习。

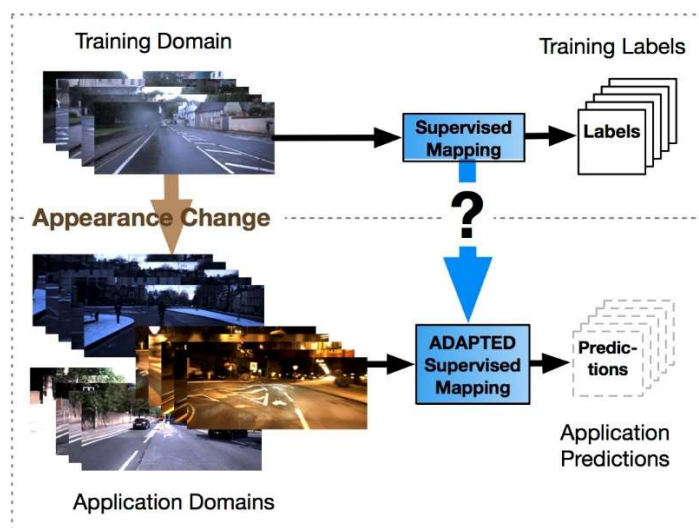
比如迁移学习（Transfer Learning）就是属于这个范畴，绝大多数迈入深度学习的初学者尝试做的一项工作就是将 ImageNet 数据集上学习到的权重，在自己的目标数据集上进行微调。



## Transfer learning: idea

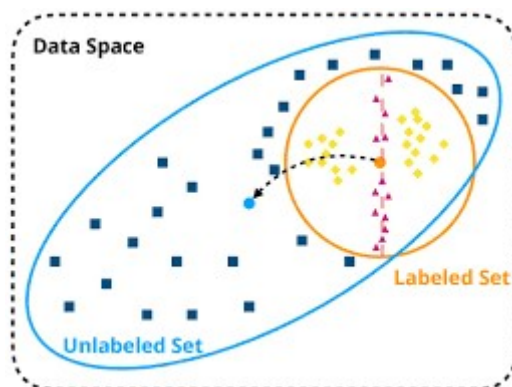


迁移学习比较常探讨的方面目前集中在领域适应性（Domain Adaptation）问题上，即如何将 train domain 上学习到的模型应用到 application domain。



迁移学习领域著名的还有零样本学习（Zero-Shot Learning）和一样本学习（One-Shot Learning），很多相关的方法也会用到领域适应性的相关知识。

Co-learning 中还有一类工作叫做协同训练（Co-training），它负责研究如何在多模态数据中将少量的标注进行扩充，得到更多的标注信息。



通过以上应用我们可以发现，协同学习是与需要解决的任务无关的，因此它可以用于辅助多模态映射、融合及对齐等问题的研究。

## 3.6 综述文献

更多细节详见

许倩倩, 黄庆明. 多模态学习研究进展综述。 <https://zhuanlan.zhihu.com/p/39878607>

Ramachandram D, Taylor G W. Deep multimodal learning: A survey on recent advances and trends[J]. IEEE Signal Processing Magazine, 2017, 34(6): 96-108.

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



MMML-Tutorial-  
ACL2017.pdf

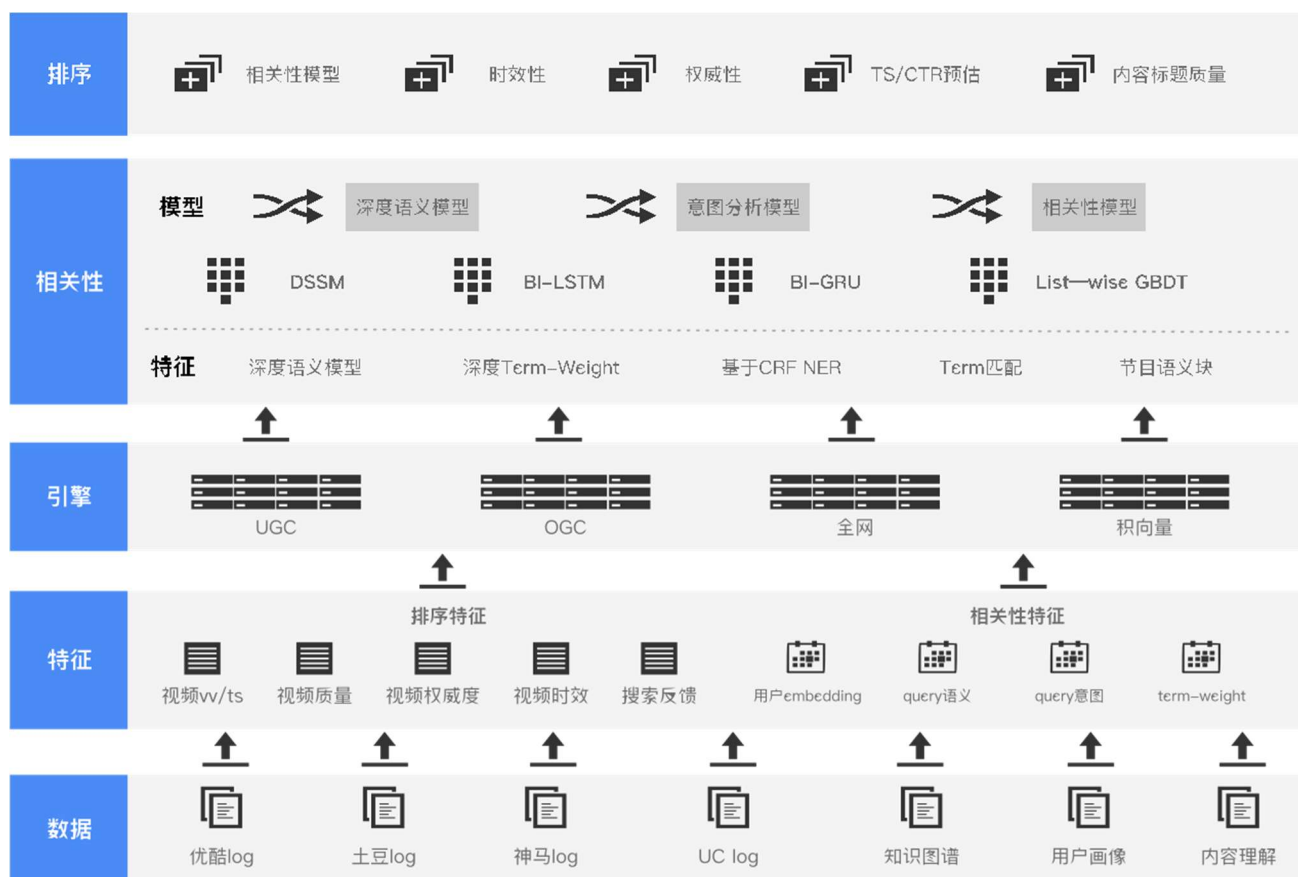
## 4 应用案例

### 4.1 优酷在多模态内容理解上的研究及应用

[https://www.infoq.cn/article/xgP\\_eyfidAA2l5ShcCPp](https://www.infoq.cn/article/xgP_eyfidAA2l5ShcCPp)

应用场景：用户搜索

- **视频理解与分析**, 对视频内容做细颗粒度拆解, 将图像、视频动作、人物、声音、背景音乐等信息通过检测和识别等手段做标签化, 通过上述手段完成对视频内容降维到文字模态的转换
- **视频内容逐帧向量化**, 为保证召回兜底, 采用 query、视频向量化处理, 作为文字模态召回的有益补充
- **搜索查询意图识别**, 用户在使用搜索时是有状态的, 不同上下文环境下同一个查询词表达的意图不尽相同
- **搜索排序**, 排序对于搜索引擎是个至关重要的模块, 既有算法技术的一面, 更有业务属性的一面, 这里要兼顾平台视角和用户视角, 单纯的 CTR 优先或者业务干预优先都是不可取的, 需要排序的设计者能够从机制设计的视角来思考



难题：不同用户喜欢的因素不一样、  
方案：视频、音频、文字 同时推荐，计算量大、

### 应用场景：视频数字资产化

- 将视频抽象为一个向量。
- 元素级解构：将视频拆解为人物、情绪、场景、动作、服化道，例如接吻和同时使用图像和声音判断。
- 自动生成电影电视海报，个性化的海报。

## 5 图文融合模型

微博的推荐系统已经在使用多模态信息：用户信息、物料信息、博主信息、浏览时长、转评赞，等等。这些信息的可以用文本和数字表示，使用方便。

微博的形式通常是文字+图片或者文字+视频，单一模态（文字、图片或视频）可能无法理解博主的完整意思。下左图中，文字“爱了爱了”没有实体信息，只从红唇图片也不好确定要表达的信息，如果将红唇与“知名美妆博主”身份联系起来，就知道这条微博中的实体信息是口红。下右图，文字中的实体词毛戈平是著名化妆师，但他的中秋礼盒可能是各种各样的化妆品，无法确定这条微博对应的具体物品；4张图中有3张的主体是盒子1张是口红，仅从图片判断，微博对应的具体物品应该是盒子；如果将文字中的毛戈平和图片中的口红联系起来，就知道这条微博对应的物品是口红。



理解图片是理解视频的基础，因此，微博中的多模态机器学习可以从文本+图片开始。目前，图片分析相当成熟，工具众多；文本分析的最强通用工具是 BERT。因此，本文重点调研以 BERT 为基础的图文融合方法。

## 5.1 模型摘要与详述

这里给出 7 篇论文的摘要，全文见插入的 PDF 文件，内容精华见插入的 PPT 文件。

### 5.1.1 VL-BERT

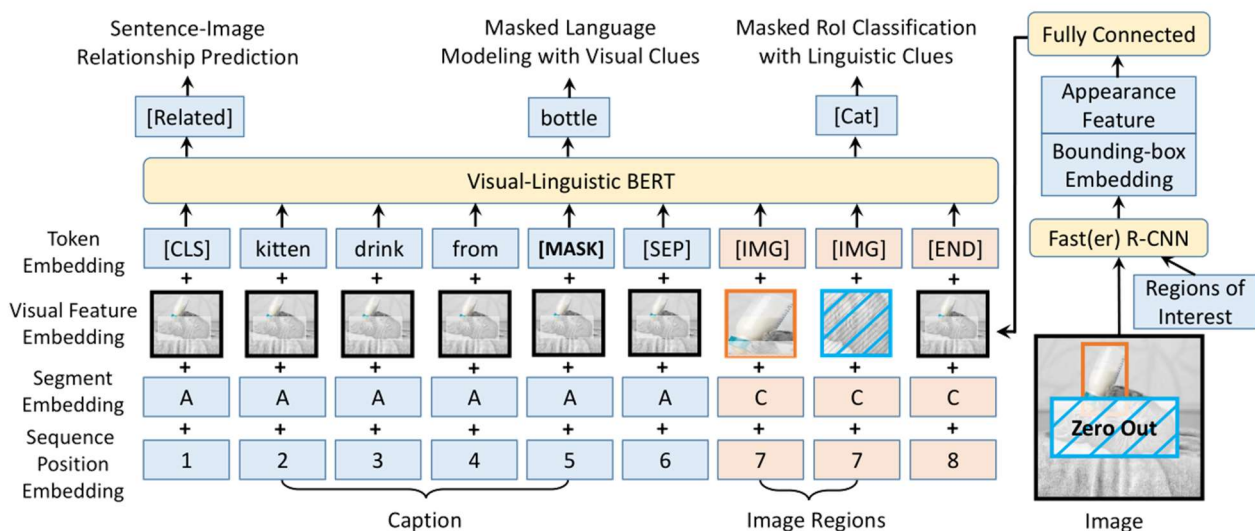
VL-BERT: Pre-training of Generic Visual-Linguistic Representations

VL-BERT：通用视觉-语言表征预训练

论文地址：<https://arxiv.org/abs/1908.08530>

论文摘要：作者们设计了一种新的用于视觉-语言任务的可预训练的通用表征，名为 VL-BERT。VL-BERT 把简单有效的 Transformer 模型作为主干并进行拓展，视觉和语言嵌入特征可以同时作为输入。输入中的每个元素可以是来自句子的一个单词，也可以是输入图像中的一个感兴趣区域。模型的设计也为了能够和所有视觉-语言的下游任务兼容。作者们在大规模的 Conceptual Captions 上对模型进行预训练，三个预训练任务为：带有视觉线索的掩蔽文字建模、带有语言线索的感兴趣区域分类、句子-图像关系预测。作者们通过大量的实证分析表明预训练阶段可以更好地对齐视觉-语言线索，并为视觉问答、视觉常识推理、代指词汇理解等下游任务带来收益。值得一提的是 VL-BERT 在 VCR 排行榜上取得了单一模型的最好成绩。





模型架构：用图像兴趣区域组成的图像句子代替 BERT 中的下一句句子，兴趣区域向量是 Faster R-CNN 特征向量+边框坐标向量；目标任务有 3 个：图文关系二分类、遮挡语言模型、遮挡兴趣区域模型。



VL-BERT论文-英.pdf



VL-BERT.pptx

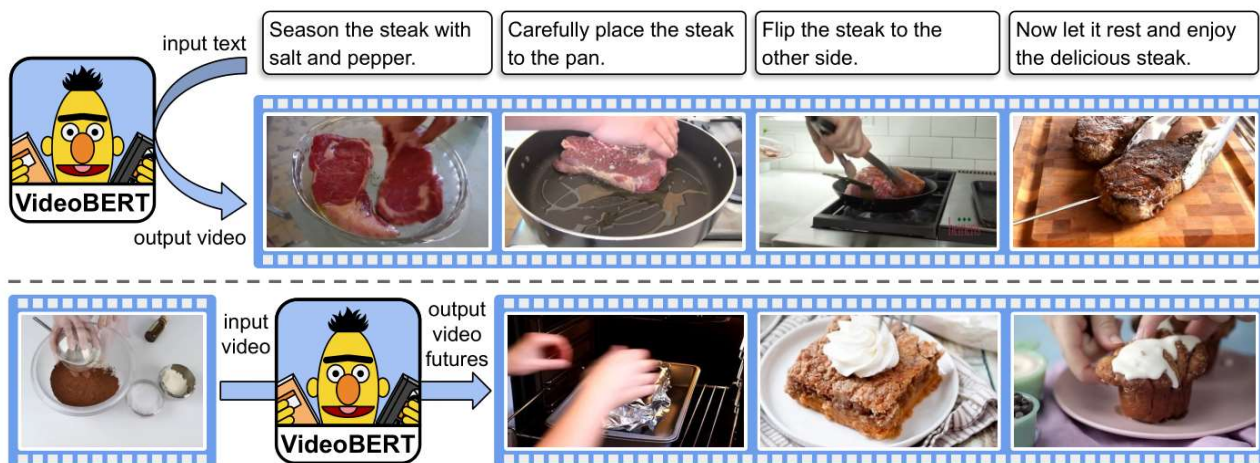
## 5.1.2 VideoBERT

VideoBERT: A Joint Model for Video and Language Representation Learning

VideoBERT：一个视频和语言表征的联合学习模型

论文地址：<https://arxiv.org/abs/1904.01766>

论文摘要：为了利用 YouTube 之类的公众媒体平台上的大规模无标签数据，自监督学习如今变得越来越重要。目前的大多数方法都是学习一些低阶表征，而这篇论文中作者们提出了一个视觉和语意的联合模型，在没有额外显式监督的条件下学习高阶特征。具体来说，作者们借鉴了语言建模中十分成功的 BERT 模型，在它的基础上进行改进，从视频数据的向量量化和现有的语音识别输出结果上分别导出视觉 token 和语言学 token，然后在这些 token 的序列上学习双向联合分布。作者们在多项任务中测试了这个模型，包括动作分类和视频描述。作者们表明了这个模型可以直接用于开放词汇库的分类任务，也确认了大规模训练数据以及跨模态信息都对模型的表现有重大影响。除此之外，这个模型的表现超过了最优秀的视频描述模型，作者们也通过量化结果验证了这个模型确实学习到了高阶语义特征。



论文目标：根据菜谱文本匹配烹饪视频片段，根据视频片段预测后续几步的操作视频。

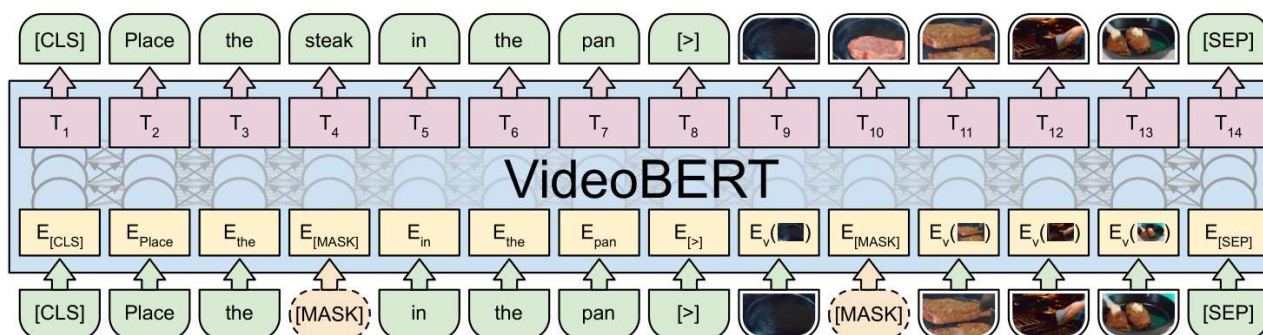


Figure 3: Illustration of VideoBERT in the context of a video and text masked token prediction, or *cloze*, task. This task also allows for training with text-only and video-only data, and VideoBERT can furthermore be trained using a linguistic-visual alignment classification objective (not shown here, see text for details).

模型架构：BERT 架构不变，重点在视频句子构造方法。将视频切分成 1.5 秒的片段，再识别动作，按动作聚类，用类中心那帧代表本类动作。



videoBERT.pdf



videoBERT.pptx

### 5.1.3 ViLBERT

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

ViLBERT：为视觉-语言任务训练非任务专用的视觉语言表征

论文地址：<https://arxiv.org/abs/1908.02265>

论文摘要：这篇论文中作者们提出了 ViLBERT（视觉和语言 BERT），一个学习任务无关的图像内容与自然语言联合表征的模型。作者们把热门的 BERT 架构拓展为一个支持两个流输入的多模态模型，它在这两个流中分别预处理视觉和文本输入，并在联合注意力 transformer 层中进行交互。作者们先在大规模自动采集数据集 Conceptual Captions 上通过两个代理任务预训练模型，然后把它迁移到多个现有的视觉-语言任务上，包括视觉问答、视觉常识推理、代指词、基于说明的

图像检索，过程中也只对基础架构做很小的调整。相比于目前的任务专用模型，作者们的做法带来了巨大的表现提升，在所有 4 个任务上都得到了最好的成绩。作者们的成果也代表了学习视觉和语言之间联系的一种新思路，不再局限于某个具体任务训练过程中的学习，而是把视觉-语言联系作为一个可预训练、可转移的模型能力。

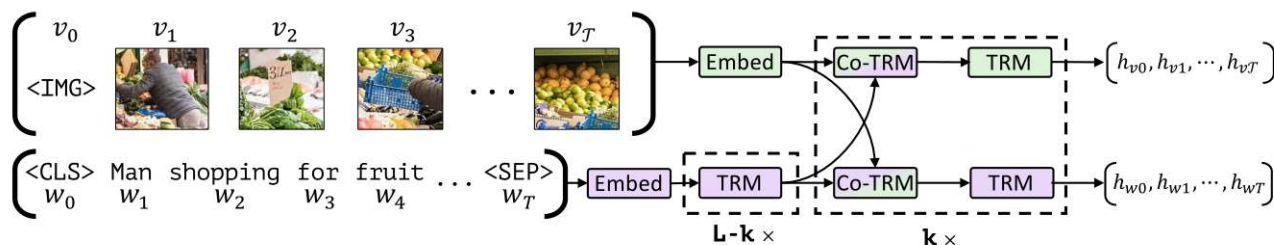
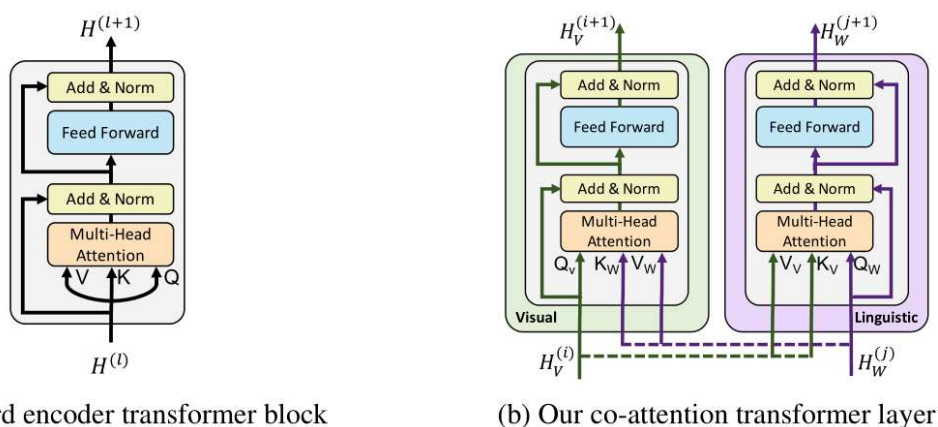


Figure 1: Our ViLBERT model consists of two parallel streams for visual (green) and linguistic (purple) processing that interact through novel co-attentional transformer layers. This structure allows for variable depths for each modality and enables sparse interaction through co-attention. Dashed boxes with multiplier subscripts denote repeated blocks of layers.

模型架构：构造了一个跨模态变换器，用于融合文本和图像向量。

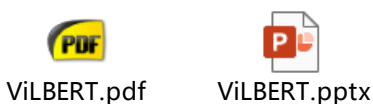


(a) Standard encoder transformer block

(b) Our co-attention transformer layer

Figure 2: We introduce a novel co-attention mechanism based on the transformer architecture. By exchanging key-value pairs in multi-headed attention, this structure enables vision-attended language features to be incorporated into visual representations (and vice versa).

跨模态变换器是 2 个并列的变换器，相互交换输入的 Key 向量和 Value 向量。



## 5.1.4 VisualBERT

VisualBERT: A Simple and Performant Baseline for Vision and Language

VisualBERT：一个简单有效的视觉和语言基准线模型

论文地址：<https://arxiv.org/abs/1908.03557>



论文摘要：这篇论文里作者们提出了 VisualBERT，这是一个可以对一系列不同的视觉-语言任务进行建模的框架，而且简单灵活。VisualBERT 包含了一组层叠的 Transformer 层，借助自我注意力把输入一段文本中的元素和一张相关的输入图像中的区域隐式地对齐起来。除此之外，作者们还提出了两个在图像描述数据上的视觉-语言关联学习目标，用于 VisualBERT 的预训练。作者们在 VQA、VCR、NLVR2 以及 Flickr30K 这四个视觉-语言任务上进行了实验，结果表明 VisualBERT 以明显更简单的架构在所有任务中都达到了做好的表现或者和竞争者相当的表现。作者们的进一步分析表明 VisualBERT 可以在没有任何显式监督的情况下建立语言元素和图像中区域之间的联系，而且也对句法关系和追踪（根据描述建立动词和图像区域之间的关系）有一定的敏感性。

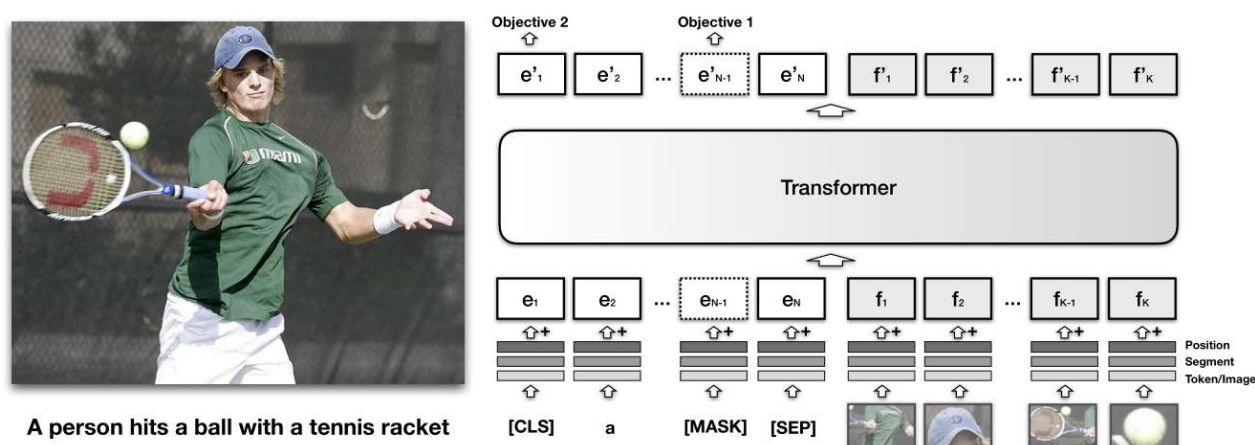


Figure 2: The architecture of VisualBERT. Image regions and language are combined with a Transformer to allow the self-attention to discover implicit alignments between language and vision. It is pre-trained with a masked language modeling (Objective 1), and sentence-image prediction task (Objective 2), on caption data and then fine-tuned for different tasks. See §3.3 for more details.

模型架构：图像向量是 3 个向量的和，兴趣区域向量，由 Faster R-CNN 或其它卷积网络提供（随数据集而变）。



VisualBERT.pdf



visualBERT.pptx

## 5.1.5 B2T2

Fusion of Detected Objects in Text for Visual Question Answering

视觉问答中的检测到物体的文本融合

论文地址：<https://arxiv.org/abs/1908.05054>

论文摘要：论文作者们开发了一种简单但强有力的神经网络，它可以合并处理视觉和自然语言数据，作为多模态模型的持续改进。模型的名字是 B2T2（Bounding Boxes in Text Transformer，文本 Transformer 中的边界框），它也在同一个统一架构中利用了把单词指向图像中的一部分的参考信息。B2T2 在视觉常识推理（<http://visualcommonsense.com/>）数据集上有优秀的表现，相比此



前公开的基准模型降低了 25% 错误率，也是公共排行榜上目前表现最好的模型。作者们进行了详细的对照试验，表明在早期就把视觉特征和文本分析相结合是这个新架构发挥出好效果的重要原因。

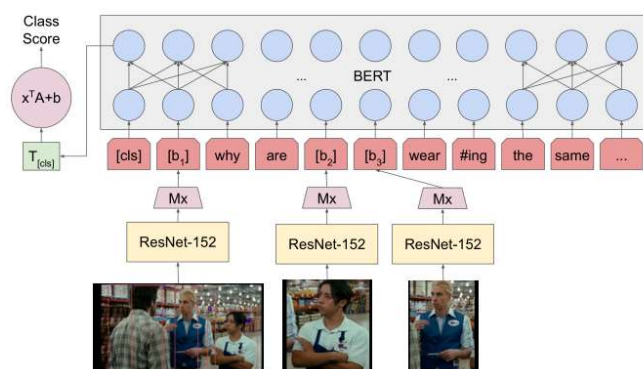


Figure 3: B2T2 architecture with early fusion. Bounding boxes are inserted where they are mentioned in the text and at the end of the input, as described in Sec. 4.

模型架构：目标任务视觉常理推理，把图像向量填到空位里，用真假图像构造负样本。



B2T2.pdf



b2t2.pptx

## 5.1.6 Unicoder-VL

Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training

Unicoder-VL：一个通过跨模态预训练生成的语言和视觉通用编码器

论文地址：<https://arxiv.org/abs/1908.06066>

论文摘要：作者们提出了 Unicoder-VL，这是一个以预训练的方式学习视觉和语言的联合表征的通用编码器。这个模型借鉴了 XLM 和 Unicoder 等跨语言、预训练模型的设计思路，视觉和语言内容都会被传入一个多层 transformer 中，作为跨模态预训练阶段；预训练阶段使用三个任务，包括掩蔽语言建模、掩蔽对象标签预测以及视觉-语言匹配。前两个任务会让模型学习从基于语言和视觉内容输入的联合 token 学习到内容相关的表征；后一个任务尝试预测一张图像和一段文本描述之间是否相符。在大量的图像-描述对上预训练之后，作者们把 Unicoder-VL 迁移到了图像-文本检索任务上，只添加了一个额外的输出层，就在 MSCOCO 和 Flickr30K 两个数据集上都取得了目前最佳的表现。

## 5.1.7 LXMERT

LXMERT: Learning Cross-Modality Encoder Representations from Transformers

LXMERT：从 Transformers 中学习跨模态编码器表征

论文地址：<https://arxiv.org/abs/1908.07490>

论文摘要：视觉-语言推理需要对视觉概念、语言语义有一定的理解，尤其还需要能在这两个模态之间进行对齐、找到关系。作者们提出了 LXMERT 框架来学习这些语言和视觉的联系。在 LXMERT 中，作者们构建了一个大规模 Transformer 模型，它含有三个编码器：一个对象关系编码器、一个语言编码器和一个跨模态编码器。接着，为了让模型具备联系视觉和语言语义的能力，作者们用大量的图像和句子对进行了模型预训练，使用了 5 个不同的、有代表性的预训练任务：掩蔽语言建模、掩蔽对象预测（特征回归和标签检测）、跨模态对齐以及图像问答。这些任务既可以帮助学习同一个模态内的联系，也可以帮助学习跨模态的联系。在预训练的参数基础上进行精细调节之后，模型在 VQG 和 GQA 两个视觉问答数据集上都取得了最好成绩。作者们还把这个预训练跨模态模型适配到了一个有挑战的视觉推理任务 NLVR2 上，把最好成绩从此前的 54% 正确率一举提升到了 76%，表明了模型有良好的泛化性。最后，作者们通过对照试验证明了他们新设计的模型部件和预训练策略都对结果有很大的帮助。代码和预训练模型可以参见 <https://github.com/airsplay/lxmert>

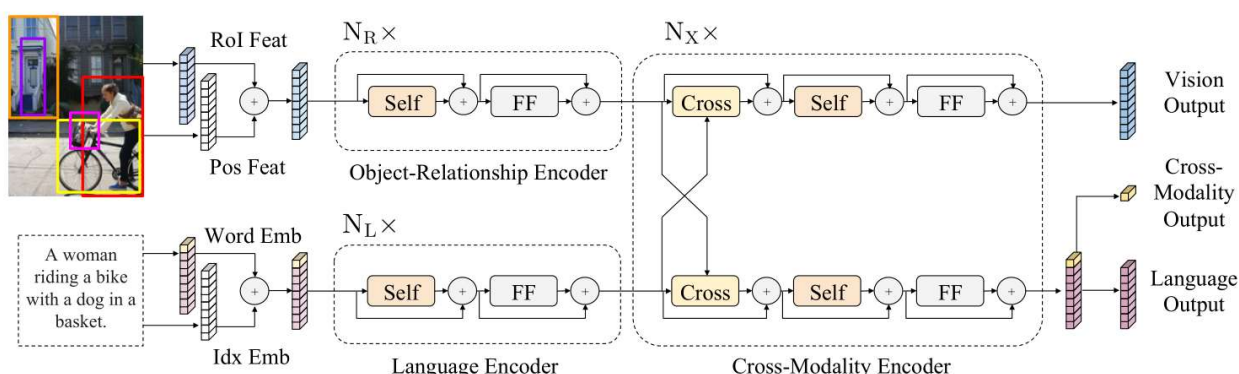


Figure 1: The LXMERT model for learning vision-and-language cross-modality representations. ‘Self’ and ‘Cross’ are abbreviations for self-attention sub-layers and cross-attention sub-layers, respectively. ‘FF’ denotes a feed-forward sub-layer.

模型架构：跨模态的方法是交换 self-attention 中的向量。兴趣区域特征向量用 Faster R-CNN 提取。



LXMERT.pdf



LXMERT.pptx

## 5.2 模型评价与展望

上一节的 7 个模型中，图像融入文本的方式都将图像向量化，形成一个图像句子，像单词句子一样输入解码器。期待的效果是将图片中的物品、动作与单词对应起来。提取图像向量的工具是各种卷积网络，例如 Faster R-CNN 和 ResNet。

这里有一个严重的限制，这些卷积网络在 ImageNet 上训练而来，能够识别的目标种类十分有限，通常只有 1000 类物品。在这 1000 类物品中，相当一部分还是微博中不常见的。微博中大量出现的物品，都不在这 1000 类中，导致这些模型难以直接应用到微博业务中。

中文数据集难题。这些模型的训练，用到几个图像描述、图像问答、图像常识推理数据集，这些数据集中文本均是英文。若要应用到微博业务中，首先需要建立中文数据集，新建或将已经数据集汉化，是一项巨大的工作。

如果强行使用卷积网络来计算所有图像的特征向量，那么对 1000 类以外的物品很可能不可靠，这是国为训练的时候只用到 1000 类物品，这个网络只对这 1000 类物品有区分度，对其它的物品很可能没有区分度，至少区分度不高。

若想将这图文融合模型应用到微博业务中，又不想建立带标签的数据集，那么可以考虑无监督的图像分类、目标检测，不再受 1000 类的限制。只是需要进一步调研无监督学习，就目前了解的情况，图像无监督分类正确率还在 50% 以下，不可用。

## 6 微博中的应用场景

### 6.1 打标签

有同事在 2017 年融合图文信息来给微博打标签，性能比 fastText 高 2-3%，但没有上线，原因是耗时比 fastText 多 20 倍以上。

打标签是个分类问题，需要准备的训练样本。目前标签存在问题，影响分类的准确度：

- 标签之间相关性较强。例如：化妆品、美妆爱用物、口红，这 3 个标签的范围大小不一致，还有包含关系。
- 人工打标签。二级标签 850 余个，三级标签约 50 万个。人工判断，标签可能不准确、不全面，标签的使用频率不均衡。

考虑当前推荐系统中同时使用标签、关键字作为特征，关键字粒度比标签小，可以将图片中的信息映射到关键字上，这样能在不改动推荐系统的情况下使用图片特征。

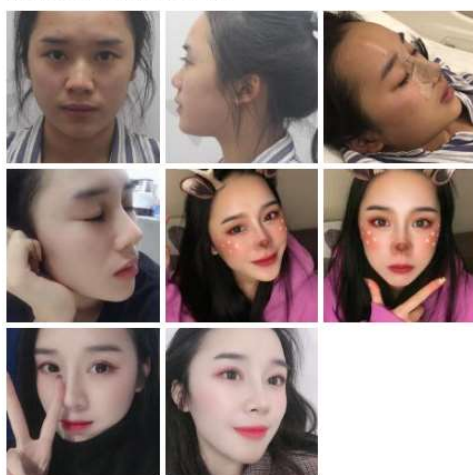
具体来说，对图片中的人脸，识别出来对应到明星人名上。对非人脸图片，按频道映射关键词，接着对关键词聚类（用以减少关键词数量）。图片中的有效信息多为物品，频道中的物品关键词通常不太多。

图片中物品与关键映射也有不小的挑战，例如下面的微博，文本中提到了鼻综合、鼻子、鼻梁、鼻头，而图片中显示的整张脸，并不是鼻子特写。再者，鼻综合是个抽象词，与图中的实体对应起来不太容易。

北京鼻综合案例反馈

妹子看着身边不少朋友都做了鼻综合，加上自己鼻子确实不好看，鼻梁低的像是没有似的，鼻头和鼻翼也很肥大，大家都变好看了，自己难免不动心的

妹子用的是硅胶，之前还一直担心出现透光的问题，不过现在都一年了，除了变好看了，别的问题都没有，也太好看了吧！



## 6.2 图文向量

推荐系统在排序使用的是大规模 FM 模型，据了解，还没有用上文本向量或者图像向量。根据 FM 的算法原理，FM 中也难以使用向量特征。

计算图文相似度。论文 B2T2 中考虑过个任务，仍然依赖于图像特征抽取器。

## 6.3 下一步工作

调研无监督目标检测，希望能识别无限数量类别的物品，以便对齐微博中的图片与文本。