软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.006125]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

多模态视觉语言表征学习研究综述*

杜鹏飞[1,2], 李小勇[1,2], 高雅丽[1,2]

[1](可信分布式计算与服务教育部重点实验室(北京邮电大学), 北京 100876)

[2](北京邮电大学网络空间安全学院, 北京 100876)

通讯作者: 李小勇 E-mail:lxyxjtu@163.com

摘要: 我们生活在一个由大量不同模态内容构建而成的多媒体世界中,不同模态信息之间具有高度的相关性和互补性,多模态表征学习的主要目的就是挖掘出不同模态之间的共性和特性,产生出可以表示多模态信息的隐含向量.该文章主要介绍了目前应用较广的视觉语言表征的相应研究工作,包括传统的基于相似性模型的研究方法和目前主流的基于语言模型的预训练的方法.目前比较好的思路和解决方案是将视觉特征语义化然后与文本特征通过一个强大的特征抽取器产生出表征,其中 Transformer[1] 作为主要的特征抽取器被应用表征学习的各类任务中.文章分别从研究背景、不同研究方法的划分、测评方法、未来发展趋势等几个不同角度进行阐述.

关键词: 多模态表征学习;表征学习;多模态机器学习;深度学习

中图法分类号: TP311

Survey of Multimodal Visual Language Representation Learning

DU Peng-Fei[1,2], LI Xiao-Yong[1,2], GAO Ya-Li[1,2]

[1](Key Laboratory of Trustworthy Distributed Computing and Service of Ministry of Education(Beijing University of Posts and Telecommunications), Beijing 100876, China)

[2](School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: We live in a multimedia world built from a large number of different modal contents. . The information between different modalities is highly correlated and complementary. The main purpose of multi-modal representation learning is to mine the different modalities. Commonness and characteristics produce implicit vectors that can represent multimodal information. This article mainly introduces the corresponding research work of the currently widely used visual language representation, including traditional research methods based on similarity models and current mainstream pre-training methods based on language models. The current better ideas and solutions are to semanticize visual features and then generate representations with textual features through a powerful feature extractor. Transformer[1] is currently used in various tasks of representation learning as the mainstream network architecture. This article elaborated from several different angles of research background, division of different studies, evaluation methods, future development trends, etc.

Key words: Multimodal Representation Learning; Representation Learning; Multimodal Machine Learning; Deep Learning

模态是事情经历和发生的方式,我们生活在一个由多种模态信息构成的世界,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.对于多模态问题我们需要充分利用多种模态间的互补性和冗余性,充分挖掘模态之间的信息从而消除数据的异构问题带来的挑战.多模态机器学习

* 基金项目:国家自然科学基金(61370069,61672111);国家自然科学基金-通用技术基础研究联合基金(U1836215);北京市自然科学基金(4162043); 国家重点研发计划(2016QY03D0605)

Foundation item: National Natural Science Foundation of China (61370069, 61672111); Natural Science Foundation of Beijing, China (4162043); National Key Research and Development Program of China (2016QY03D0605)

收稿时间: 2020-05-11; 修改时间: 2020-06-26; 采用时间: 2020-08-07; jos 在线出版时间: 2020-09-10

的应用很广泛,比较早期的应用可以追溯到1989年提出的一个视听语音任务[2],借助隐马尔可夫模型[3]通过视觉模态补充听觉模态信息.另外就是情感识别研究领域目前已经从单模态识别逐步转向多模态识别的研究,多模态情感的研究的主要是借助视觉、语音、文本、脑电等模态信息对情感状态进行识别,从输出上来看可分为分类问题(输出愤怒、高兴、悲伤等不同情感)和回归问题(输出一个到情感空间的映射值),相应的研究数据集有 Busso 等人[4] 通过诱导方式录制的基于情感分类的 IEMOCAP 数据集,Mckeown 等人录制的基于连续值的 SEMAINE 数据库[5] .另外比较常见的应用包括媒体描述、事件识别、多媒体检索、视觉推理、视觉问答等等.

Baltrušaitis 在多模态机器学习综述[6] 一文中将多模态机器学习研究分为几个方向:多模态表征学习、多模态翻译、多模态对齐、多模态融合、多模态联合学习.在解决多模态问题时多模态表征学习是一个关键的研究点,一般来说机器学习模型的好坏严重依赖于数据特征的选择,传统的机器学习中很大一部分工作都在于特征的挖掘以及特征的抽取和选择方面,这些工作的结果可以支持有效的机器学习数据表征,但是这样的特征工作比较耗费时间,尤其是一些基于手工特征的方法没有能力从原始数据抽炼出有用的知识,特征工程的目的是将人的先验知识转化为可以被机器学习识别的特征,从而弥补自身的缺点.利用表征学习的方法可以从数据中学习出有用的表征以减少对特征工程的依赖从而在一些具体任务中能取得更好的应用.首先一个好的表征要尽可能的包含更多数据的本质信息,相比于单个模态,多模态的表征学习面临很多的挑战比如噪音处理,模态之间的融合方式,丢失的模态信息处理,不同模态处理的差异化,实时性和效率等等.Bengio[7] 指出好的表征主要有几个特点:数据平滑、时空相关、数据稀疏、自然聚类等等.多模态表征空间相似的数据在实际意义或者实体概念上要存在相似性,在单一模态信息丢失的情况下可以通过另外一种模态的信息进行补充.

在过去的一段时间内单模态的表征学习取得了很大的发展,在图像领域,过去很长时间内盛行的一些手工特征比如 SIFT（尺度不变特征变化）特征[8] 和 HOG（方向梯度直方图）[9] 特征,逐渐被卷积神经网络[10] 代替,通过卷积神经网络可以充分挖掘视觉的二维和三维信息的表征含义,目前很多视觉任务都采用在一个充分预训练的卷积模型上进行微调的方式.语音领域中的一些手工特征比如梅尔频率倒谱系数（MFCC）也逐渐被一些基于数据驱动的深度学习的方法所代替[11] .另外就是在自然语言处理领域表征学习的发展尤其迅速,过去文本领域一直效果很好的基于词频统计的 TF-IDF 特征[12] 逐渐被 word2vec[13] 等隐式表征向量所代替,这些隐式表征充分挖掘了文本信息的潜在含义可以对文本进行更丰富的信息表达,另外像卷积神经网络、递归神经网络等也常被用来作为文本表征的挖掘工具,另外近年来基于预训练技术的表征学习模型逐渐兴起,并逐渐霸榜 NLP 的各类任务,其基本模式为通过在海量无标注数据集中进行自监督学习,然后再接一个具体的下游任务比如文本分类知识问题等,其中最有代表性的为谷歌提出的 BERT[14] ,BERT 在 GLUE[15] 的各项任务中都取得显著提升,BERT 的成功充分证明了对数据表征进行充分学习重要性.借鉴于单模态表征学习的一些方法,多模态的表征学习也取得了一定的进展,视觉语言表征学习是多模态表征学习中最有代表性的,而且视觉语言结合的任务也是多模态任务中最常见和占比最大的,这篇文章中我们主要介绍基于视觉语言统一表征学习的一些方法、应用、数据集以及面临的难点.

这篇文章的核心部分组织如下,首先第一节介绍相应的背景知识包括多模态表征学习的一些基本定义和划分,常用预训练技术,第二节分别比较了视觉语言表征学习的两种研究框架,第三节开始介绍基于相似性的视觉语言表征学习的方法,第四节为核心部分,主要介绍基于预训练架构的视觉语言表征模型,第五节介绍视觉语言统一表征的质量评估方法,第六节讲了视觉语言表征学习的发展趋势.

1 背景知识

1.1 表征学习

表征学习作为机器学的一个专门领域吸引了越来越多的学者的研究,很多机器学习的专门会议比如 NIPS 和 ICML 都会定期举办专门的研讨会,另外还有专门针对表征学习的会议 ICLR.表征学习本质上是特征

工程的一种延伸,传统传统特征工程挖掘的一些特征都是在对数据进行一些分析后在一些经验基础上结合一些数学分析得到的,目前典型的表征学习的方法是通过深度学习的方法从数据中自动化的挖掘出有效的隐性特征,以降低人工挖掘特征成本,更方便高效的挖掘出与具体任务无关但是可以在下游任务中有较好的应用的隐含向量.

Bengio[7] 指出表征学习有两条主线:一是概率图模型,二是神经网络模型,这两个根本区别是对每一层描述为概率图还是计算图,或者说隐层的节点是潜在的随机变量还是计算节点.从概率图模型角度来研究,表征学习的问题可以解释为试图恢复一组描述观测数据分布的潜在随机变量.我们可以将观测数据表示成为x,将潜在变量联合空间上的概率模型表示成h,表征学习的概率图模型可以表示成$p(x,h)$,表征值被认为是一个推理过程的结果,以确定给定数据的潜在变量的概率分布即$p(h|x)$,也就是后验概率,估计过程就是最大化训练数据正则化的可能性.概率图模型又可以分为有向图模型和无向图模型.有向图模型又称为贝叶斯网络,有向图模型的图节点之间有前后依赖关系,后面节点的概率依赖于前面节点的概率输出.其联合分布的构建方式表示为$p(x,h)=p(x|h)p(h)$.目前基于有向图模型进行表征学习的例子有:主成分分析(PCA), 稀疏编码,Sigmod 信念网络等等.无向图模型又被称为马尔可夫网络,其前后节点之间没有明显的依赖关系,其公式为:

$$p(x,h)=\frac{1}{Z_\theta}\prod_i \psi_i(x)\prod_j \eta_j(h)\prod_k \nu_k(x,h) \tag{1-1}$$

其中$\psi_i(x)$代表可见变量之间的连接,$\eta_j(h)$代表隐含变量之间的连接,$V_k(x,h)$代表隐含变量和可见变量之间的连接,其中分配函数Z_θ保证分布的归一化.无向图模型用于表征学习的一个典型代表是波尔兹曼机(RBM).概率图模型总是学习与潜在变量相关的尤其是后验分布给出的一个观察输入,如果模型有超过两个关联层时,其计算会变的非常复杂,而且潜在变量的后验分布还不是一个简单的可用特征向量,为了最后提取出稳定的确定性的数值特征值,通常还需要借助自动编码器.基于神经网络的自动编码器的表征学习方法与基于概率图的表征学习模型的方法的区别是概率图模型是由显式概率函数定义的,然后经过训练以最大化数据可能性,而自动编码器框架通过编码器和解码器进行参数化,自动编码框架允许在编码器和解码器中使用不同的矩阵.自动编码器训练的一个实际优点是定义了一个简单的可跟踪优化目标,可以来监视进程.为了将重构误差最小化以捕获数据生成分布的结构,在训练准则或者参数化过程中一定要防止自动编码器学习自身函数,从而在任何地方产生零重建错误.基础的自动编码器在于找到一个值的参数向量θ,从而将重建误差最小化.自动编码器的定义如下:

$$\mathcal{J}_{\mathrm{DAE}}(\theta)=\sum_t L\left(x^{(t)},g_\theta\left(f_\theta\left(x^{(t)}\right)\right)\right) \tag{1-2}$$

其中$x^{(t)}$是训练数据,L是进行优化的目标函数,其主要训练框架采用神经网络,主要训练方法采用随机梯度下降法等.f_θ主要用于编码,g_θ主要用于解码.编码维数小于输入维数的欠完备自编码器可以学习数据分布的最显著特征,如果赋予这类编码器过大的容量或者隐藏编码维数大于输入时也会发生类似情况,针对这一情况提出正则自编码器,正则自编码器根据要建模的数据分布的复杂性选择合适的编码维数,选择编码维数和编码器、解码器容量等,根据选择就可以成功训练任意架构的自编码器.去噪自动编码器是在自动编码器的基础上在输入中加入随机噪声.去噪自动编码器的表示方程如下:

$$\mathcal{J}_{\mathrm{DAE}}=\sum_t \mathbb{E}_{q\left(\tilde{x}|x^{(t)}\right)}\left[L\left(x^{(t)},g_\theta\left(f_\theta(\tilde{x})\right)\right)\right] \tag{1-3}$$

其中$\mathbb{E}_{q\left(\tilde{x}|x^{(t)}\right)}[\cdot]$是对样本进行随机过程的一个平均,随机噪声编码的方式很多,比如加性各向同性高斯噪声、用于灰度图像的加盐噪声、还有近来被广泛应用的掩码噪声,其中 NLP 各项任务中取得较大提升的 BERT 模型就是借鉴去噪自动编码方法对输入字符进行随机掩码的方法.由 Rifai 等人提出的基于对比损失的自动编码[16]

的方法通过加入一个解析收缩惩罚项,将学习特征的灵敏度与输入的无穷小变化相结合从而具备更好的性能.

1.2 多模态表征学习的定义及划分

过去的十年内通过神经网络或者概率图模型对自然语言处理、语音、图像进行表征的方法层出不穷,而同一时间内多模态表征学习的早期研究主要通过对单模态表征进行简单连接的方式进行,后来借鉴单模态表征尤其是自然语言处理领域的一些成功经验,多模态表征尤其是视觉语言的统一表征开始逐渐兴起. Baltrušaitis[6] 等人汇聚了到 19 年为止多模态表征的一些研究进展,根据输出的表征是否在一个统一的表征空间内将多模态表征分为统一表征和协同表征.统一表征融合多个单模态信号并将他们映射到一个统一表征空间内,协同表征分别处理每一个模态的信息,但是在不同模态之间增加相似性的约束.协同表征和统一表征的构造如图 1 所示:

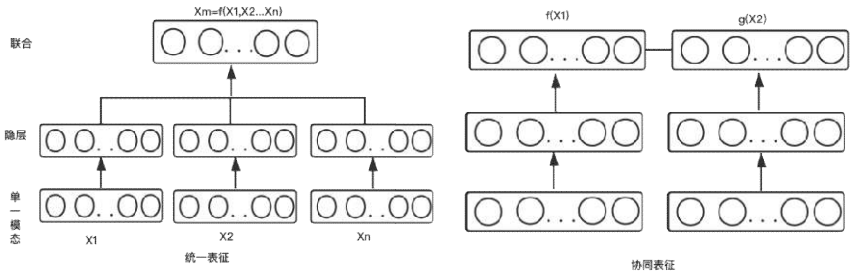

Fig.1 Structure of joint and coordinated representations

图 1 统一表征与协同表征的基本结构

统一表征将所有的多模态数据映射到一个公共空间,适用于在做推断时所有模态都存在的情况,其被广泛应用于一些视觉语言匹配任务、语音识别辅助、情感识别和多模态姿态估计等.协同表征将每一个模态映射到单独的空间,其中的每一个模态都是相互独立的,但是不同模态之间存在关联关系,协同表征可以将其中的一个模态单独拿来用,适用于像跨模态检索等场景.

从处理模态的形式上来区分,多模态表征的研究涉及图像加音频[17] [18] [19] ,视频加音频[20] [21] ,图像加文本[22] [23] [24] [25] [26] [27] [28] [29] [1] [30] [32] [33] [34] [34] [35] [36] [37] [38] [39] [40] ,视频加文本[41] [42] [43] 等等,其中视觉语言表征的研究比较多,而且其研究框架较为通用,其详细的信息在后面介绍.

1.3 预训练技术

随着深度学习的兴起,预训练技术逐步被广泛应用,其大概框架是预先训练一个模型,然后利用已经训练好的底层网络参数在目前的网络结构基础上再增加一个下游任务,其中底层网络参数在下游任务训练过程中可以不做改变的方式叫冷冻(Frozen),另外一种是底层网络参数在下游任务训练过程中随着训练进程一起改变,这种方式叫微调(Fine-Tunning).这种预训练加下游任务的方法在图像和视频领域取得了较好的效果,比如一般在做目前检测或者图像分类等任务时一般都会使用一个基于 ImageNet 数据集的预先训练好的网络然后再进行微调.虽然目前图像领域对预训练技术有一些质疑,并且认为基于 ImageNet 的预训练网络不能明显改善准确率[46] ,但是其明显改善了模型的鲁棒性和不确定性的估计[47] .而且一些基于无监督学习的预训练技术可以充分的利用海量的无标注样本[48] 从而为下游任务提供更加丰富的特征.

在 NLP 领域表征学习可以追溯到 2003 年的 NNLM[45] ,但是由于各种原因其应用效果不佳,至到 Word2Vec[13] 的诞生基于 NLP 领域的表征学习才开始逐步兴起,Word2Vec 与 NNLM 架构类似其充分利用文本表达的语序关系,通过句子中词的上下位词来进行训练（通过上下文预测词或者通过词预测上下文）从而产出词的向量,但是由于这种方法产生的词向量是静态,所以其无法较好的解决多义词的问题.从 2018 年开始诞

生的 ELMO[49] 、GPT[50] 以及目前广泛应用的 BERT 通过利用预训练技术并采用 LSTM、Transformer 等特征提取器有效的解决了多义词问题.尤其是 BERT 其通过借鉴随机噪声编码器的思想,通过对文本进行随机掩码的方式从而有效的提升了文本表征的质量,并且在 NLP 的各项任务中得到了显著的提升.BERT 的一些设计思想和架构也随即被应用到了视觉文本表征领域.

2 视觉语言表征学习的研究框架

基于视觉语言的多模态表征是多模态表征中的一个重要研究方向,其在内容消费、医疗影像等领域有着广泛应用.视觉语言表征学习的本质是学习到视觉模态和语言模态到一个空间的映射,其可以充分利用视觉模态和语言模态之间的互补性,剔除模态间的冗余性,从而学习到更好的特征表示.目前较主流且性能较好的研究框架主要分为两种:一种是基于对比学习或者称为相似性学习的,其主要是在相似性的约束条件下优化每一种模态的表征.另外一种是基于自回归或者自编码的预训练架构的,其借助于 Transformer[1] 等高效神经网络对各种数据模态的样本编码成特征然后再进行重构.两种研究框架表现形态如图 2 所示.

基于相似性学习的方法通过一个度量函数衡量视觉模态信息和语言模态信息的差异,相似性学习的目标是学习一个编码器 f 使得:

$$\text{score}\left(f(m), f\left(m^{+}\right)\right) >> \text{score}\left(f(m), f\left(m^{-}\right)\right) \tag{2-1}$$

其中 m^{+} 是和 m 相似的正样本,m^{-} 是和 m 不相似的负样本,score 为相似性的度量函数,相似性度量又可以建模为回归问题、分类问题、排序问题,其根据输入数据的不同格式和不同的目标损失函数来建模模态之间的关系,其输入模态被限制为两种.

相似性学习的方法只需要在各自特征空间上学习到区分性,而基于预训练架构的方法需要对每个模态元素之间的细节进行重构,其构建模型表示如下:

$$\begin{aligned} &VisionRegions: \mathbf{v}=\left\{v_1,\ldots,v_K\right\} \\ &SentenceTokens: \mathbf{w}=\left\{w_1,\ldots,w_T\right\} \\ &f_{emb}=f\left(v,w\right) \end{aligned} \tag{2-2}$$

其中 $\mathbf{v}$ 为视觉区域单元,w 表示为文本模态信息,f 为深度神经网络,一般为 Transformer 神经网络结构,其堆叠了多个多头自注意力层和前馈神经网络,自注意力子层的设计使其在处理多模态序列编码时相比其他结构具备更好的性能.其中一种典型的框架是采用类似于 BERT 这种基于语言模型掩码的自编码架构的,比如 VisualBERT[39] 、ImageBERT[40] 等,视觉输入通过预处理的方法转化成与文本单元类似的一个个视觉单元,然后视觉单元和语言单元通过掩码任务实现语言模型,这种架构从本质上是将视觉模态和语言模态处理成语言序列任务,通过自监督的方式从海量数据中学习出两种模态的联合编码.针对联合表征的具体用途(用于理解或者用于生成),其可以分别采用自编码模型或者自回归模型来进行编码.更好的损失函数定义和更海量的数据都有助于提升这种表征的质量,还有就是视觉单元的描述方式,视觉维度信息的刻画方式也会对最终表征好坏产生影响.预训练架构可以让视觉语言表征在一阶段通过自编码或者自回归的方式进行充分的模态融合产生高质量的视觉语言表征,然后在二阶段或者三阶段应用于具体任务.目前在一些视觉语言具体的应用任务中,这种方式取得准确率最高.

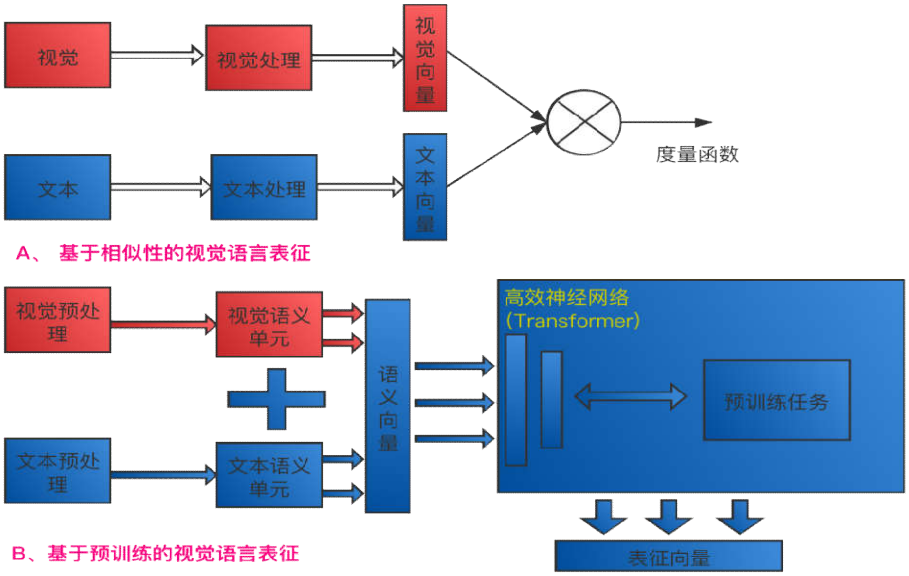

Fig.2 Two architectures for visual language representation learning

图 2 视觉语言表征学习的两种架构

3 基于相似性的视觉语言表征学习

3.1 总体架构

基于相似性的表征学习是在一个协同的空间内最小化不同模态之间的距离,其输入数据主要为具有排序或者正负关系的视觉文本信息对,通过不同的建模方法实现视觉语言表征的学习.其建模的损失函数可以为铰链损失或者三元损失.输入为具备排序关系的正负对或者三塔结构.比如在基于堆叠注意力网络的图文联合表征（SCAN）[51]中其采用损失函数如下:

$$l_{\text{hard}}(I,T)=\left[\alpha-S(I,T)+S\left(I,\hat{T}_h\right)\right]_+ +\left[\alpha-S(I,T)+S\left(\hat{I}_h,T\right)\right]_+ \tag{3-1}$$

其中 S 为相似函数方程(主要基于 cosin 函数进行改造),(I,T)为正例的图像文本对,$(I,\hat{T}_h)$为负例的文本图像对,$(\hat{I}_h,T)$为负例的图像文本对,该损失函数融合了铰链损失和三元损失函数,其优化目标是在一定间距内使得图文配对的正例的取值大于图文不配对和文图不配对的负例的取值.基于相似性的视觉语言表征学习主要受限于度量损失函数、相似性计算方法和进行相似度量的粒度,其中从相似性度量的粒度来看主要分为基于粗粒度的匹配和精细粒度的匹配,从发展来看越精细粒度的相似性计算所产生的表征越能产生更好的效果.下文分别从粗粒度相似度匹配模型和细粒度的相似度性匹配模型两个方面进行阐述,同时介绍不同模型的特性.

3.2 基于粗粒度的相似度匹配模型

最早的一个工作是由 Weston 等人在 WSABIE[52] 中提出的,其主要通过计算图像模态和图像的标注文本之间的相似性.WSABIE 中使用排序损失来度量标注数据与图像之间的相似性.

$$L(k)=\sum_{j=1}^{k}\alpha_j,\ \text{with}\ \alpha_1\geq\alpha_2\geq\cdots\geq 0 \tag{3-2}$$

L 可以选择不同的优化方法,其中α为对一张图片的不同标注的排序.WSABIE 同时引入了在线学习排序的方法来实时优化参数,但是由于 WSABIE 只研究了从图像特征到嵌入空间的线性映射,可用的标签仅仅是图像训练集中提供的标签无法扩展到新的类别,DeViSE[25] 基于深度零样本学习的理念在不同模态的预训练向量之间建立了一个线性映射,首先采用 skip-gram 的方法对文本部分产生文本向量,另外采用一个卷积神经网络对图片进行基于目标检测的预训练,视觉部分的最终的投影层是一个线性变换,将视觉部分的 4096 维的表征映射成语言模型的 500 维或者 1000 维.最终的损失主要基于相似性,融合点积运算和铰链损失,我们定义最终的损失函数为:

$$\text{loss(image,label)}=\sum_{j\neq label}\max\left[0,\text{margin}-\vec{t}_{\text{label}}M\vec{v}(\text{image})+\vec{t}_j M\vec{v}(\text{image})\right] \tag{3-3}$$

其中$\vec{v}(image)$是视觉层的输出向量,$\vec{t}_{\text{label}}$是为所提供的文本标签学习嵌入向量的行向量,$\vec{t}_j$是其他文本的向量,M 是线性映射层的训练参数.

DeViSE 在对文本部分进行预训练时利用了基于 skip-gram 的语言模型,Lazaridou[53] 等人进行了扩展将视觉部分加入了进去,构成了多模态的 skip-gram 模型,视觉损失部分将词汇表示的视觉信息考虑在内,其中损失的计算为:

$$-\sum_{w'\sim P_n(w)}\max\left(0,\gamma-\cos\left(u_{w_t},v_{w_t}\right)+\cos\left(u_{w_t},v_{w'}\right)\right) \tag{3-4}$$

u_{w_t}是我们最终想学习的多模态增强的单词表示,v_{w_t}是与文本部分匹配的视觉模态的向量表示,$v_{w'}$是从视觉词典中负采样的视觉单词向量,其通过最大化匹配的图文向量和不相匹配的图文向量的差异来进行相似性度量.

针对损失函数的优化可以有效提升产出的表征的质量,VSE++[54] 提出了一种新的损失函数计算方案,其主要针对疑难的负例,加大样本与疑难负例之间的距离,其损失函数采用三元损失:

$$\ell_{MH}(i,c)=\max_{c'}\left[\alpha+s\left(i,c'\right)-s(i,c)\right]_{+}+\max_{i'}\left[\alpha+s\left(i',c\right)-s(i,c)\right]_{+} \tag{3-5}$$

针对每一给定的正例(i,c),负例的选择为$i'=\arg\max_{j\neq i}s(j,c)$ 和$c'=\arg\max_{d\neq c}s(i,d)$,其中 s 为距离函数, 也就是选择距离正样本距离较远的负样本进行训练.

借鉴于深度语义相似性度量模型 DSSM[55] 这种基于无监督方式度量查询向量和返回的匹配文档相似性的方法,DMSM[56] 将视觉模态作为查询而文本模态作为返回的匹配文档,采用余弦度量函数度量两种模态的距离$R(Q,D)=\text{cosine}\left(y_Q,y_D\right)=\left(y_Q^{T}y_D\right)/\left\|y_Q\right\|\left\|y_D\right\|$,对于每一个输入的图像文本对,我们计算和文本相关的图像的后验概率为:

$$P(D|\ Q)=\frac{\exp(\gamma R(Q,D))}{\Sigma_{D'\in\mathbb{D}}\exp\left(\gamma R\left(Q,D'\right)\right)} \tag{3-6}$$

其中γ是验证集上的平滑因子,$\mathbb{D}$为与查询图像匹配的所有候选文档的集合,对每一个查询图像我们选择一个相关的文本片段和 N 个不相关文本片段来计算后验概率.最终的损失函数采用负 log 损失:

$$L(\Lambda)=-\log\prod_{Q,D^{+}}P\left(D^{+}|\ Q\right) \tag{3-7}$$

ReviSE[57] 采用最大平均差异（MMD）的方法,度量视觉模态和文本模态分布的差异:

$$\mathcal{L}_{MMD}=\left\|\mathbf{E}_{p}\left[\phi\left(\tilde{v}_h\right)\right]-\mathbf{E}_{q}\left[\phi\left(t_h\right)\right]\right\|_{\mathcal{H}_k}^{2} \tag{3-8}$$

其中 p 和 q 是视觉向量和文本向量的分布,优化 MMD 函数的过程可以看作是缩小两种模态的相似差距.
最终损失函数融合图文配对二分类损失、无监督的图像和属性二分类损失以及 MMD 损失三种损失函数.

以上的相似度模型对视觉信息和文本信息的提取都分别采用神经网络结构输出的隐含向量,尤其是视觉信息大部分都是基于一个全局的卷积神经网络提取特征向量,没有对每一种模态的特征进行细粒度语义级别学习,为了更精细的进行不同模态下相似性度量,下面提出了基于一些细粒度的模态提取的方法.

3.3 基于细粒度的相似性匹配模型

为了对每种模态的高层次语义信息（尤其视觉模态）进行捕获,从而实现细粒度匹配,一般采用全局特征与局部特征融合和增加自注意力机制等方案.

You 等人[31] 提出了基于卷积神经网络融合局部特征和全局特征进行相似度量计算的表征输出模型.其使用卷积神经网络分别进行文本和视觉部分的特征提取,融合局部特征和全句特征进行相似度的计算,其中全局特征的损失函数为:

$$\sum_{i}\sum_{j}\left(\alpha-f\left(v_i,s_i\right)+f\left(v_i,s_j\right)\right)+\sum_{i}\sum_{j}\left(\alpha-f\left(v_i,s_i\right)+f\left(v_j,s_i\right)\right) \tag{3-9}$$

v_i和s_i分别代表第 i 张图像和其相对应文本的全局表征,$f(\mathrm{v,s})$是计算两个向量的相似性的函数,除了全局特征比较外,还针对中间特征也就是局部特征进行了比较,中间层的二维卷积通常都包括三维特征（卷积核数,长度和宽度）,为此设计了一个从局部特征到全局特征的线形映射:

$$l_v'=T\left(l_v\right)W_v+b_v \tag{3-10}$$

其中$T\left(l_v\right)\in\mathbb{R}^{(h_v\times w_v)\times f_v}$是一个变形操作,我们可以通过一个 Softmax 函数计算局部特征和全局特征的相关性:

$$\alpha_v = \mathrm{softmax}\left(l_v' s\right) \tag{3-11}$$

然后我们通过局部视觉特征计算全局视觉特征.

$$c_v = \sum_i \alpha_{vi} l'_{v(i,:)} \tag{3-12}$$

文本特征的计算方式类似,局部特征的相似性损失函数如下:

$$\sum_i I(i) \sum_j \left(\gamma - f\left(c_{vi}, s_i\right) + f\left(c_{vi}, s_j\right)\right)$$
$$\sum_i I(i) \sum_j \left(\gamma - f\left(v_i, c_{si}\right) + f\left(v_j, c_{si}\right)\right) \tag{3-13}$$

如果第 i 对全局损失 I(i)为 1,局部损失才计算,否则 I(i)为 0.

SCO 模型[58] 提取了图像的多个候选区域,然后采用多标签的卷积神经网络对每一个候选区域进行分类得到分类的向量,然后再利用逐元素最大池化的方法得到一个得分向量作为局部特征,通过 VGG 模型抽取全局特征然后通过门控机制将全局特征和局部特征进行融合得到视觉融合向量,再与 LSTM 输出的文本向量进行相似度匹配.

Wu[59] 等人提出的融合方法同样考虑了全局对齐和局部对齐,同时对文本句子进行解析分别提取出实体对象、属性、实体关系的三元组.对于局部对齐,其中文本部分是通过提取实体对象与对应的图像做排序损失,全局对齐的损失主要包括实体关系与图像特征的排序损失、句子与图像特征的排序损失、文本融合与图像特征的排序损失,其表示如下:

$$\mathbf{u}_{comp} = \mathrm{Norm}\left(\Phi\left(\left\{\mathbf{u}_{obj}\right\} \cup \left\{\mathbf{u}_{attr}\right\} \cup \left\{\mathbf{u}_{rel}\right\}\right)\right) \tag{3-14}$$

另外一种细粒度特征提取的方案是通过注意力机制进行特征权重计算,SCAN[51] 模型采用目标检测的方法提取图像的不同特征区域,然后对文本切分为一个个的文本单词,首先用对应每个图像区域与文本中每个单词做注意力运算,然后再用每个图像区域与句子向量进行注意力运算,从而确定图像区域重要性.PFAN[60] 模型通过对图像进行分块然后针对不同的块输出不同的隐含位置向量,同时将每个块的位置向量与原始图像进行注意力运算,从而产生带有位置权重信息的视觉向量,针对文本模态采用 GRU 提取向量,最后采用一个三元损失作为度量损失函数.

3.4 总结

基于相似性的视觉语言表征学习模型以相似性为度量标准优化每种模态的隐含向量,首先在使用上其不能作为一个统一表征输出,需要采用一定拼接方式将两种向量连接起来,同时在训练过程中由于存在大量样本如何高效计算损失也需要解决的问题.

4 预训练架构的视觉语言统一表征学习

Transformer 凭借强大的特征学习能力、预训练加下游任务的多阶段架构、基于随机掩码构建的自动编码机制在 NLP 领域取得巨大成功,从 2019 年开始多模态领域开始借鉴 BERT 在 NLP 领域的一些成功经验,由此诞生了像 VideoBERT[44] 、ViLBert[33] 、ImageBERT[40] 、LXMERT[37] 、UNITER[35] 等一系列基于预训练架构和 Transformer[1] 特征抽取的多模态模型,并取得了较好的效果.表 1 中展示了在视觉推理任务中近些年评测的结果.

Table 1 NLVR2 presents the task of determining whether a natural language sentence is true about a pair of photographs.

表 1 NVLR2 任务用于判断自然语言处理中句子对是否正确

排名	模型	开发集(准确率)	公共测试集(准确率)	未开放测试集(准确率)
	人工评测	96.2	96.3	96.1
1	UNITER	78.4	79.5	80.4
2	LXMERT[37]	74.9	74.5	76.2
3	VisualBERT[39]	67.4	67.0	67.3
4	MaxEnt	54.1	54.8	53.5
5	CNN+RNN	53.4	52.4	53.2
6	FiLM	51.0	52.1	53.0

如上表所示,像 UNINTER、LXMERT、VisualBERT 等采用类 BERT 预训练架构的多模态表征模型相比其他架构的模型有显著提升.

4.1 总体架构

如图 3 所示 VisualBERT[39] 展示了类 BERT 视觉文本统一表征预训练架构的一个典型结构,通过 Transformer 中的 self-attention 机制隐式的对齐输入文本元素和输入图像中的区域,复用了 BERT 的加掩码操作的编码方式,整个架构上采用预训练加下游任务微调的模式.

文本输入部分的处理与原始 BERT 类似,对原始输出文本产生字词向量、段落向量、位置向量等三个输入向量.对视觉部分的输入进行隐式表达,通过目标检测的方法提取图像关键区域,类似于文本中的词组,图像的输入同样产生三个与文本输入类似的向量,分别是图像目标区域的向量,图像文本段落向量(是图像还是文本),图像目标区域位置坐标进行平均加权的位置向量.视觉部分的三个隐含向量与文本部分的三个隐含向量进行拼接然后作为 Transformer-encoder 编码的输入,预训练目标函数包括两个（1）预测文本加图像组成的输入向量的随机掩码（2）图像文本匹配任务,其中每一个图像有多个描述,从中选择一个作为正例,从其他图像的描述中随机选择一个作为负例,进行二分类预测.目标函数的选择对多模态表征作用很大,像 ViLBERT[33] 、ImageBERT[40] 等很多都是通过优化预训练的目标函数从而提高了输出表征的质量.通过自监督的预训练任务后产出了一个较高维度的多模态表征.对于一些具体的视觉文本类下游任务比如图像描述、视觉问答等在目前已经训练好的网络结构上进行微调,具体实现为在输出的隐层后面再接一个面向具体任务的损失函数.比如视觉问答任务是针对一张图像提出问题然后选出匹配的答案,其本质上属于一个多分类的任务,因而一般后面接一个交叉熵损失.

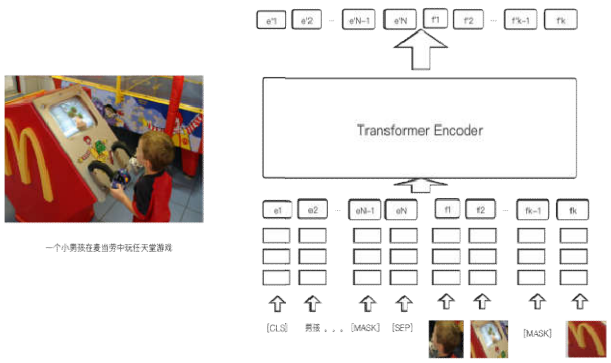

Fig.3 Main architecture of VisualBERT. Image regions and language regions are combined with a Transformer to allow the self-attention to discover implicit alignments .

图 3 VisualBERT 的主干结构,图像检测区域和文本区域进行输入组合,Transformer 通过自注意机制发现隐含对齐

4.2 几种不同划分

4.2.1 基于内容理解与内容生成的划分

一个典型的 Transformer 架构由编码器和解码器两部分组成,其中编码器部分主要应用于内容的理解,比如 BERT,解码器部分则侧重于内容的生成和回归,典型有 GPT 这种模型.目前产出的视觉语言统一表征框架多是基于 Transformer 自动编码架构的,侧重于内容理解部分,另外的架构就是融合自回归和自编码两种模型的架构,其可以支持内容理解和内容生成的通用任务,如表 2 所示.

Table 2 Unified representation of visual language based on encoder and decoder

表 2 基于编码和解码架构区分的视觉语言统一表征

类型	方法	适用领域
基于自编码的	ViLBERT	图文
	VL-BERT	图文
	VisualBERT	图文
	LXMERT	图文
	Unicoder-VL	图文
基于自回归和自编码的	VideoBERT	视频文本
	CBT[61]	视频文本
	VLP	图文
	UniViLM	视频文本

VLP[62] 是一个典型的混合编码解码结构的网络框架.从结构上讲自编码和自回归结构的一个主要区别在于进行掩码遮罩操作时,自编码方式的掩码可以是随机掩码的,而在自回归的方式中由于考虑到序列关系,所以其掩码操作必须是按顺序进行掩码的.

4.2.2 单流结构与双流结构

对于输入的文本特征向量和视觉特征向量有两种方式进行融合,一种是文本特征和视觉特征拼接然后接一个自动编码器进行融合,另外一种就是分别对文本特征和视觉特征进行独立编码,然后通过交叉注意力机制实现不同模态信息的融合,具体如表 3 所示.

Table 3 Two streams and Single stream

表 3 双流结构和单流结构

类型	方法	适用领域
单流结构	Unicoder-VL	图文
	B2T2	图文
	VisualBERT	图文
	ImageBERT	图文
	VLBERT	图文
	UNITER	图文
双流结构	LXMERT	图文
	ViLBERT	图文

双流结构通过对视觉部分和文本部分进行分别编码,然后再通过交叉编码的方式充分学习了每种模态的特征,相比单流结构双流结构的特征学习更加充分,类似于对不同模态的特征进行了一次特征提取之后又进行了交叉的特征提取,其典型结构如图 4 所示,ViLBERT 中引入了联合注意力机制进行不同模态之间的学习,联合注意力机制最早见于 Faster-RCNN[47]结构中,每一种模态的查询向量和键值向量同时作为另外一种模态的查询向量和键值向量,注意力模块为每种模态产生了基于注意力的池化特征,视觉流中有了基于文本注意力的先验条件,文本流中有了基于视觉注意力的先验条件.ViLBERT 分别输出视觉模态和文本模态的表征,然后通过线性加权融合的方式产生联合表征.另外类似的模型 LXMERT 中每一路包含两个自注意力子层、一个交叉自注意力层和两个前向编码,与 ViLBERT 操作查询向量和键值向量的方式不同,其第 K 层的自注意力交叉层的输入为前 k-1 层的视觉向量和文本向量,具体如下:

$$\hat{h}_i^k = \text{Cross Att}_{\text{L}\to\text{R}}\left(h_i^{k-1}, \left\{v_1^{k-1}, \ldots, \mid v_m^{k-1}\right\}\right) \tag{4-1}$$

$$\hat{v}_j^k = \text{Cross Att}_{\text{R}\to\text{L}}\left(v_j^{k-1}, \left\{h_1^{k-1}, \ldots, h_n^{k-1}\right\}\right) \tag{4-2}$$

其中 $\hat{h}_i^k$ 代表第 k 层的文本输入向量,$\hat{v}_j^k$ 代表第 k 层的视觉输入向量.交叉自注意力模型分别作用于前 k-1 层的视觉输入向量和文本输入向量.LXMERT 的输出为三个向量分别为视觉向量、文本向量和对文本向量进行池化操作后的交叉模态向量.

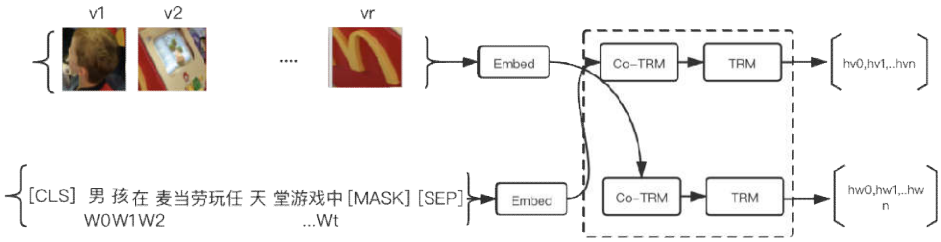

Fig.4 . Two-Stream Deep Architecture

图 4 双流结构

4.3 特征预处理

4.3.1 文本特征处理

在神经网络训练的过程中,首先需要构建一个词典,然后对词典中每个词做向量表,对于新来的词首先需要加入词典中,这些会导致词典越来越大,过大的词典主要会带来两个问题:1）稀疏问题:某些词汇出现的频率很低,得不到充分训练.2）计算量问题:词典过大,也就导致隐含向量的计算量变大.单纯基于词典的方式不能解决袋外词集的问题（出现不在词表中的词）,解决这个问题主要是通过建立字符级别的模型,字符级别的模型试图使用 26 个字母加上一些符号表示所有词汇,这种处理方式虽然可以较好的解决袋外词集的问题,但是模型的粒度变小,输入长度变长,使得数据更加稀疏并且难以学习长远程的依赖关系.词级别模型导致袋外词集问题,而字级别模型粒度小,所以就诞生了子词级别（subword-level）的处理方式.比如训练集的词汇:"old older oldest smart smarter smartest" 采用词级别的词典表示为" old older oldest smart smarter smartest"长度为 6,而采用子词级别的处理方式表示为"old smart erest",其长度为 4.目前预训练模型中常用子词算法包括:BPE 算法[64] 和 WordPiece[65] 算法.

BPE(字节对)编码或二元编码属于数据压缩算法,其中最常见的为一对连续字节数据被替换为该数据中不存在的字节,其后期使用时需要一个替换表来重建原始数据,其算法描述如下:

1. 准备足够大的训练语料.
2. 确定期望的子词词表大小.
3. 将单词拆分成字符序列并在末尾添加后缀"</w>",统计单词频率.
4. 统计每一个连续字节对的出现频率,选择最高频合并成新的子词.
5. 重复第四步至到达到第二步设定的子词词表大小或者下一个最高频字节对出现频率为 1.

WordPiece 算法是 BPE 的变种,不同点在于 WordPiece 基于概率生成新的子词而不是下一最高频字节对.算法描述如下:

1. 训练语料数据准备.
2. 确定期望的子词词表大小.

3. 将单词变成字符序列.
4. 基于第三步数据训练语言模型.
5. 从所有的子词单元中选择加入语言模型后能最大程度地增加训练数据概率的单元作为新的单元.
6. 重复第五步直到达到第二步设定的子词词表大小或概率增量低于某一阈值.

4.3.2 图像特征处理

卷积神经网络是目前比较通用的图像特征提取的方法,目前大部分的图像任务大多基于一个效果较好的卷积网络比如 ResNet-101[66] 提取图像表征然后在一个具体任务上进行应用.BERT 在处理文本任务时其输入的信息都是词或者字,是一个小的语义单元,将整张图片向量作为输入将无法很好的学习视觉语义单元信息,所以一般对图片进行目标检测操作,然后将检测后的结果进行处理然后作为一个语义单元作为输入.下表中展示 Unicode-VL 模型在句子检索和图像检索任务中使用 ResNeXt[67] 模型和使用 FasterR-CNN[68] 提取检测框的差别.

Table 4 Performance between FasterR-CNN and ResNeXt

表 4 使用 FasterR-CNN 和 ResNeXt 的差距

处理方法	句子检索			图像检索		
	R@1	R@5	R@10	R@1	R@5	R@10
无微调 ResNeXt	4.0	14.1	22.9	4.3	15.5	26.1
ResNext	4.7	18.5	28.5	5.6	18.8	30.9
FasterRCNN 36 个框	66.1	93.2	96.9	57.8	86.7	93..8
FasterRCNN 100 个框	82.6	96.6	99.3	68.5	92.7	96.9

使用目标检测提取图像的 36 个框或者 100 个框然后将其作为视觉语义单元进行输入,其输出的表征的质量远远高于单纯的使用 ResNeXt[67] 进行像素级特征提取所产生的表征的质量.将目标检测方法应用多模态任务最早由 Anderson[69] 等人在 Bottom-Up and Top-Down Attention for Image Captioningand Visual Question Answering 一文中提出,文章提出了一种自下而上和自顶向下的方法用于多模态的视觉问答和图像描述的任务,该架构与目前视觉文本预训练架构类似,只是其主要应用于具体的下游任务.其中自下而上的模块采用 FasterR-CNN[68] 作为特征提取器,FasterR-CNN 的骨干网络主要采用了基于 ImageNet 数据集进行预训练的 RestNet-101 网络,然后其在 Visual Genome[70] 数据集上进行训练,Visual Genome 数据集是由斯坦福人工智能实验室的李菲菲教授提出,其目的是构建一个包含丰富语义信息的视觉数据集,整个数据集大约包含 10.8 万张图片,平均每张图片含有 21 个物体、16 个属性,同时其还标示了两个物体之间的关系,也就是该数据集同时包含实体、实体属性和实体之间的关系等三种.Faster-RCNN 最终检测输出的目标是从 2000 类实体和 500 类属性中选出的 1600 类实体和 400 类属性,通过这种方式训练的目标检测模型作为图片的特征提取器可以有效的提取图片中的视觉语义特征.

目前每一张图片输出的检测框一般包含 36 个框或者 100 个框,如上表 4 所示输出 100 个框的效果相对更好一些.在 Unicoder-VL[38] 中将每个检测到的对象的位置坐标表示为由归一化的左下和右上坐标组成四维向量,其包含了位置和大小信息,最后我们将位置向量和检测出的目标框中的特征向量进行相加,然后通过线性映射的方式将其映射成另外的向量且其维度与文本输入向量的维度相同从而可以进行拼接.VL-BERT 同样将每个目标区域坐标表示成一个四维的向量,不同的是其不是简单的直接与目标区域特征向量拼接,而是利用与 BERT 位置向量相同的处理方式通过一个正弦和余弦方程映射成一个 2048 维的向量[71],这样相当于产生了图像维度的位置向量.ViLBERT 也是将输出的位置向量和检测框中的特征向量进行求和,区别是其位置向量为一个五维的向量,除了归一化的左上和右下坐标外还增加了一个区域占比(目标区域面积占整张图片面积的比重).UNINTER 中则采用一个七维的位置向量（归一化的四个位置坐标,长度,高度,区域面积）.

4.3.3 视频特征的预处理

VideoBERT[44] 、UniViLM[43] 是目前典型的视频文本预训练的架构,预处理时首先需要将视频特征向量化.VideoBERT 通过每秒 20 帧的速度进行采样,以 30 帧为一个单元,通过在 Kinetics[72] 视频动作数据集上预训练 S3D[73] 模型对视频帧进行特征抽取.通过分层聚类的方式对视频特征进行处理,设置分层聚类的层数为 4,聚类簇为 12,总共可以聚类产生 12 的 4 次方一共 20736 个类,相当于两万个最小的语义单元.UniViLM 则对视频切帧处理后使用 ResNet-152 提取 2 维特征,使用 ResNeXt-101 为骨干网络的三维卷积网络提取 3 维特征,将三维特征和二维特征拼接成一个 4096 维的向量,再后接一个 Transformer 结构进行视频特征抽取.

4.4 预训练任务

4.4.1 预训练数据集

对于多模态表征的预训练任务而言,一个好的数据集直接影响了最终产生的表征的质量的好坏.现在收集的用于视觉语言统一表征预训练的数据集如下:

- MS-COCO[74] :第一个版本由微软在 2014 年发布,最开始数据集由 20G 左右的图片和 500M 左右的文本文件.COCO 通过在 Flickr 上搜索 80 个对象的类别和各种场景来收集图像,有 33 万张图片,其中 20 万张有标注.其标注类型主要包括目标实例、目标关键点和图像描述,其本质是一个可以支持多个任务的数据集,并不是专门为视觉语言描述任务设计,且其数据量相对而言不是很大,可以作为视觉语言表征预训练任务的基线版本.
- Conceptual Captions[75] 数据集:这个数据集由谷歌发布,其一共包含 330 万张配对的图像以及该对象对应的描述.数据集通过一个流式处理框架从互联网中的上亿个网页中构建,首先基于图像进行过滤,其只保存 JPG 格式图片,同时过滤了涉及色情的内容.然后抽取网页中 Alt-text 标签之间的文本,过滤掉搜索引擎优化词和推特的标签词.使用一个图文配对分类器将没有与之配对文本的图片过滤掉.该数据集发布后成为视觉语言预训练任务的标配数据,目前很大一部分视觉语言预训练任务类似 ViLBERT、VLBERT、Unicoder-VL 等都采用该数据集进行预训练.
- SBUCaptions[76] 数据集:在从网络中筛选而成的图像描述系统中输入查询图像,根据查询结果筛选候选的匹配图像,基于抽取出的一些高维度信息比如对象、场景进行重排序,返回最相关的图像中的文本描述,同时过滤一些带噪音的描述,最终结果中包含了 1 百万的图像和其相对应的文本描述,该数据集采用两种图像描述生成的方法,一种为查询结果的描述迁移,一种为利用全局表示和图像内容的直接估计生成描述,图像描述的相关性总体较高,但也存在一些错误,数据量规模中等,进行预训练任务时一般和其他数据集融合使用.
- LAIT(Large-scale weak-supervised Image-Text)[40] 数据集:该数据集由微软收集目前最大的一个图文配对数据集,一共包含 1000 万的图像文本数据,每一张图片的平均描述为 13 个字符.收集方法与 Conceptual Caption 类似,首先从互联网中收集网页信息过滤掉非英文部分,对图片进行过滤,保留长度和高度大于 300 像素的,使用二元分类器丢弃了一些不可学习的图片.使用用户定义的元数据信息作为图像文本描述,同时制定了一系列的过滤规则用于文本过滤.训练了一个弱分类器进行图文匹配的判断,对于一张图片有个配对描述的情况只选择得分最高的配对.该数据集属于目前规模最大的预训练任务数据集,ImageBERT 也利用该数据集取得了当时为止的最好效果.
- HowTo100M[77] 数据集:从海量教学视频中进行数据收集,视频的内容主要为教授一些复杂的任务,其中包括来自 122 万段人类表演和活动的教学网络视频,描述了超过 2 万 3 千个不同的视觉任务.该数据集的规模非常大,涵盖的种类较多.
- Youcook2[78] :是两个下游任务的域内数据集.它包含了 2000 个烹饪视频,89 个食谱,14K 的视频剪辑.总时长为 176 小时(平均 5.26 分钟).每个视频片段都有一个注释句子.该数据集主要是和烹饪相关的视频,领域受限.
- MSR-VTT[79] :包含针对 10000 个视频的 200,000 个描述,覆盖类别为 257 种,平均句子长度为 9.28.视

频描述重复,66%的视频具备同样描述.

- VATEX[80] 中英文视频描述数据集:该数据集一共包含 41250 个视频,82500 个视频描述,600 个类别,82500 个描述都是唯一的,每一个视频都有 20 个描述,其中 10 个中文,10 个英文,其中 5 对是中英文相互对应的翻译,英文不少于 10 个单词,非翻译的中文句子不少于 15 个字,相比 MSR-VTT 数据集其支持多语言,而且视频描述的多样性较高,规模更大.

下表中列出了不同数据集的一些特点和差异:

- Table 5 The difference between different pretrained datasets
- 表 5 不同预训练数据集比较

类型	数据集	数据规模	获取方式	优点	缺点
图像加文本	MS-COCO	33 万+	基于 Flickr 搜索	支持多种任务	数据量相对较少
	Conceptual Captions	330 万+	从互联网上亿个网页中获取,然后进行过滤	规模较大	缺少图片描述的校验,所以存在一定误差
	SBUCaptions	100 万+	基于网络中爬取得到的图文配对库	多种图像描述生成方式,图像的文本描述准确率较高	数据规模一般,且收集过程复杂
	LAIT	1000 万+	互联网爬取,基于弱分类器判断	规模巨大,用于预训练任务效果明显	数据量超大,对预训练任务的机器性能是一种挑战
视频加文本	HowTo100M	122 万	Youtube 视频获取	规模较大,海量视频	规模海量,筛选规则不强
	Youcook2	2000	烹饪视频获取	领域性较强	只涵盖烹饪领域数据
	MSR-VTT	10000	互联网获取	类别涵盖全,视觉内容多样化	视频描述重复,60%的视频具备同样描述
	VATEX	41250	互联网获取	多样性强,中英双语描述,数据规模较大	部分视频缺少双语描述

4.4.2 预训练损失函数

对于多模态的预训练任务而言预训练损失函数的选择和设计至关重要,目前主要的预训练任务整理如下:

(1) **图像文本掩码**:其基本模式与 BERT 掩码语言建模的任务类似.VisualBERT 中只对文本向量进行掩码,对图像部分不进行掩码操作.ViLBERT 对 15%的视觉和文本输入都进行随机掩码.图像掩码对图像区域的 90%的图像特征进行归零,区域的 10%保持不变,对于图像掩码的处理通过最小化掩码区域分布和非掩码区域分布的 KL 散度实现.

(2) **视觉文本匹配**:本质上是一个二分类任务,VisualBERT 基于 COCO 数据集,正样本是一张图片和该图片匹配的描述,负样本是一张图片以及随机选择的其他图片的描述.该方法同样被 VL-BERT 使用.

(3) 掩码视觉区域:UINTER 中对视觉区域做掩码有三种方式,一种是掩码区域特征回归,每一个视觉区域都是一个高维度的向量,让输出的向量尽可能接近被掩码掉的区域特征向量,使用 L2 损失让两个**向量**的距离尽可能的小.第二种方法是掩码区域特征分类,每一个视觉区域都对应一个分类标签,我们的目的就是采用交叉熵等损失函数使掩码区域的分类和真实的分类类似.第三种方法是掩码区

域 KL 散度,一般 KL 散度主要用来衡量数据分布之间的差异,我们采用 KL 散度损失来度量视觉掩码区域和真实视觉区域的之间的分布式差异.

(4) **序列到序列目标损失:**微软 VLP[62] 中为了构建即满足内容理解任务又满足内容生成任务的联合表征引入了序列到序列的损失函数,这种方式保证自注意力掩码操作是顺序的.我们首先定义自注意力掩码为 M

$$M_{jk}=\begin{cases}0, & \text{allow to attend}\\ -\infty, & \text{prevent from attending}\end{cases}\qquad j,k=1,\ldots,U \tag{3-3}$$

然后在自注意力编码操作中引入 M

$$\begin{aligned}&V=W_V^l H^{l-1}, Q=W_Q^l H^{l-1}, K=W_K^l H^{l-1}\\ &A^l=\text{softmax}\left(\frac{Q^\top K}{\sqrt{d}}+M\right)V^\top\end{aligned} \tag{3-4}$$

H^{l-1}为上一层的输出, A^l 为l层的输出, A^l 计算过程中通过引入 M 操作保证自注意力操作的顺序.

(5) 场景图预测:简单的视觉语言匹配任务会丢失很多文本和视觉模态中的细粒度信息,ERNIE-ViL[63] 提出了基于场景图的目标预测任务,包括物体目标预测,属性预测以及物体关系预测,物体目标预测为随机选取图中一部分物体,掩码其在句子中词,然后根据上下文和图片对掩码预测,属性预测和物体关系预测方式类似,其掩码句子中的属性词和关系词然后基于相应上下文和图像进行掩码预测.这一预测任务属于从粗粒度的视觉语言匹配到精细粒度的视觉语言语义匹配的迈进.

4.4.3 多阶段预训练

ImageBERT 中采用了一种多阶段预训练的方法,首先在 LAIT[40] 数据集上进行第一阶段的预训练,然后再在 Conceptual Caption 和 SBUCaptions 数据集上进行第二阶段的预训练,最后再接一个具体的下游任务.

4.5 下游任务

基于 BERT 自监督预训练的框架,通常会在训练好的预训练参数的基础上接一些具体的下游任务,多模态视觉语言预训练的下游应用任务很多,从内容理解类到内容生成类的,下游任务的性能和效果的好坏一定程度上反映了训练出来的表征质量的好坏,下面从理解和生成两个角度选取一些有代表性的下游任务

4.5.1 内容理解类

典型的下游任务包括视觉问答、视觉推理、视觉联合推理、图像检索、视频检索.视觉问答是指根据给定的图片提问,从候选中选择出正确的答案,VQA2.0[81] 中从 COCO 图片中筛选了超过 100 万的问题,我们训练模型来预测最常见的 3129 个回答,其本质上可以转化成一个分类问题.视觉推理相对问答更为困难,其可以分解为两个子任务视觉问答(Q->A)和选出答案的原因（QA->R）,除了回答用自然语言表达的具有挑战性的视觉问题外,模型还需要解释为什么作出这样的回答,其最开始由华盛顿大学提出,同时发布的 VCR[82] 数据集包含 11 万的电影场景和 29 万的多项选择问题.NLVR2[83] 是一个关于自然语言和图像联合推理的数据集,重点关注语义多样性、组合性和视觉推理挑战,该任务是确定一个自然语言的标题对给出的一对图像是否正确,数据集由超过 10 万的英语句子和网络图片组成.图像检索任务是给定一个句子,选择它们对应的边界区域,Flickr30K[84] 数据集包括 3 万张图片和 25 万条注释.视频检索是根据给定的一个查询语句从视频中查询出相关的片段,在推断过程中,模型根据文本输入和候选视频计算片段相似性,从而选择最合适的视频片段.Youcook2[84] 是一个视频检索的数据集,它包含 2000 条烹饪的视频和 89 个食谱,每一个视频片段都有一个相应的视频描述,可以适用于视频检索的任务.

4.5.2 内容生成类

典型的内容生成类应用主要包括图像描述和视频描述,图像描述是根据输入的一张图片自动生成其对应的文字性描述,类似于看图说话,图像描述是一个很典型的多模态生成的任务,其可以被看作是动态的目标检测,最早的做法主要是利用图像处理的一些算子提取出图像的特征,然后利用一些浅层分类器得到可能的目标[85],后来谷歌提出 show and tell 仿照机器翻译的架构通过卷积神经网络提取图像特征做为递归神经网络的输入,通过编码-解码的结构来生成目标语言文字[1] ,后来又引入注意力机制以关注局部特征.利用图像目标检测的结果做为输入等,基于 Transformer 预训练然后微调的一个典型架构是 VLP[62] ,其他在预训练损失函数中定义了序列到序列的任务,其可以直接用于图像描述的任务.典型的图像描述数据集包括 MSCOCO[74] 等,视频描述与图像描述类似是针对视频生成文本描述,其本质上也是一个序列生成任务,用于视频描述的数据集比较多,常见的比如 MSR-VTT[1] 视频描述数据集,其每个视频片段包含 20 个人工标注的句子数据,其总共有来源于 1 万条视频的 20 个分类的 20 万条视频片段,可以用于生成任务.

4.6 总结

基于预训练架构的视觉语言表征模型可以灵活应用于各类下游任务,但是由于其模型大、结构的灵活性较差、参数较多,所以其计算量大,应用场景被限制,另外由于视觉信息的多样化,所以进行视觉单元提取时很难涵盖全面.

5 视觉语言统一表征质量评估

对于一般的分类任务而言准确率、召回率等指标就可以很好的衡量分类算法的好坏,同样对于嵌入式表征我们也需要有一套评估标准来衡量其输出的表征的质量的好坏.对于 Word2Vec 等字词向量算法在被提出时也指定了一系列的质量评估的方法,比如相似度评价:通过标注好的词汇相似性数据集(WS-353、SimLex-999)进行的相关性度量.类比任务:比如中国+北京=法国+巴黎.分类任务:根据词向量计算文本向量然后进行文本分类,根据文本分类的准确率评估向量质量.聚类可视化:比如 t-SNE[87] 通过 t 分布对数据点进行相似性的建模.视觉语言表征由于涉及到跨模态的表示所以其质量评估的方法更加复杂一些.综合目前视觉语言表征的一些产出总结的一些表征质量评估的方法如下几种.

5.1 零样本学习评估

零样本学习就是识别过去从未见过的数据类别,即产出的表征在不经过微调的情况下不仅能识别出已知的数据类别还能识别出未知的数据类别.其中用于质量评估的任务主要包括句子检索和图像检索.目前被用来进行验证的数据集包括:

MSCOCO 数据集[74] :包含 33 万张图片,每张图片包含 5 个文本描述,他被分割成训练集、验证集和测试集.

Flickr 数据集:Flickr8k 数据集包含来自 Flickr 数据集的 8000 张图片,Flickr30k 数据集包含三万张图片.其中每一张图片包含 5 个描述,每个描述都用相同意思但是不同的方式描述同一张图片.

句子检索是基于句子查询相匹配的图片,图像检索主要是基于图像查询相匹配的句子.这里会定义几个指标 R@1,R@5,R@10,分别表示召回的前一条,前五条和前十条数据中,正确的数据占的百分比.下表中为列出了在 Flickr30K 数据集上进行零样本评估的一些主流模型的比对:

Table 6 Performance of different model on Flickr30k datasets with zero-short learning method

表 6 几种不同模型在 Flickr30k 数据集上的零样本评估性能比对

主要模型	Flickr30k					
	图像到文本检索			文本到图像检索		
	R@1	R@5	R@10	R@1	R@5	R@10
ViLBERT	31.86	61.12	72.8			
Unicoder-VL	43.7	75.6	81.5	64.3	85.8	92.3
ImageBERT	54.3	79.6	87.5	70.7	90.2	94
UNITER	62.34	85.62	91.48	75.1	93.7	95.5

可见 UNITER 模型由于在预训练任务中对视觉语言模态分布差异的充分学习,从而在零样本评估中取得较好的效果.

5.2 面向具体任务的评估

面向具体任务的评估是直接在具体的任务上进行训练,本质上属于有监督的训练,如表 5 所示面向具体任务的训练效果明显好于零样本学习评估的效果.视觉语言表征模型可以针对 MSCOCO 和 Flickr 数据集的句子检索和图像检索任务进行有监督训练,从而评估模型好坏.另外一些有监督的任务包括多标签分类等, NUS-WIDE 是一个多标签分类的数据集[88] ,黄等人将基于社交网络图片学习的视觉与文本联合表征可以在 NUS-WIDE 数据集上进行评估[89] .一般基于相似性的表征学习架构会采用这种评估方式,同样采用 Flickr30k 数据集,下表中列出了一些模型直接基于跨模态检索任务进行训练得到的性能评估:

Table 7 Performance of different model on Flickr30k datasets with task-specific method

表 7 几种不同模型在 Flickr30k 数据集上的面向具体任务的性能评估比对

主要模型	Flickr30k					
	图像到文本检索			文本到图像检索		
	R@1	R@5	R@10	R@1	R@5	R@10
DeVise	4.5	18.1	29.2	6.7	21.9	32.7
VSE++	52.0	-	92.0	64.7	-	95.9
SCO	56.7	87.5	94.8	69.9	92.9	97.5
SCAN	58.8	88.4	94.8	72.7	94.8	98.4
PFAN	61.6	89.6	95.2	76.5	96.3	99
Unicoder-VL（面向任务）	63.9	91.6	96.5	75.1	94.3	97.8

由此可见在这种评估方式下,一些细粒度的视觉语言相似度模型比如 PFAN 通过对模态特征的学习可以达到与 transformer 这种自编码结构接近或者更优的效果,对一些可能没有那么预训练数据的专有领域,采用相似性学习的方法也不失为一种较好的方案.

5.3 预训练加下游任务评估

针对下游任务的评估是在产生的统一表征的基础上针对具体的任务进行微调,从表 8 中可以看出这种方式的准确率最高,这也表明这种方法的先进性.进行评估的下游任务囊括了上一章中介绍的各类任务包括内容生成类和内容理解类的.比如视觉问答、视觉推理、视觉联合推理、图像检索、视频检索等,以及图像描述等.目前为止视觉问答、推理和联合推理任务中表现较好的模型 UNITER、ImageBERT 以及 ERNIE-ViL,其提升的关键是引用了更大的数据集和设计了更合理的预训练损失函数.在表 9 中针对不同模型的性能进行了

比对.

Table 8 Performance of Unicoder-VL under different evaluation method

表 8 几种不同评估方式下 Unicoder-VL 的表现

处理方法	Flickr30k 句子检索（1000 条测试）		
	R@1	R@5	R@10
Unicoder-VL(零样本)	43.7	75.6	85.3
Unicoder-VL(具体任务)	75.1	94.3	97.8
Unicoder-VL(预训练+微调)	82.6	96.6	99.3

Table 9 Performance of different model on Flickr30k datasets with pretrain+fine-tune method

表 9 几种不同模型在 Flickr30k 数据集上采用预训练加微调方式的性能评估比对

主要模型	Flickr30k					
	图像到文本检索			文本到图像检索		
	R@1	R@5	R@10	R@1	R@5	R@10
PFAN	50.4	78.7	86.1	70.0	91.8	95.0
ViLBERT	58.2	84.9	91.5	-	-	-
UNITER	71.5	91.2	95.2	84.7	97.1	99.0
Unicoder-VL	71.5	90.9	94.9	86.2	96.3	99.0
ImageBERT	73.1	92.6	96.0	87.0	97.6	99.2
ERNIE-ViL	76.7	93.58	96.44	88.10	98.0	99.2

预训练架构模型凭借预训练时利用海量数据对模态间信息的充分学习,在进行下游任务微调时取得了较好的效果,且其相应指标明显高于基于相似性学习的模型,这也充分证明了在采用 transformer 进行编码的预训练阶段针对模态间的互补性和冗余性进行了很好的学习,而针对具体的下游任务就是在已经学习到的参数基础上进行优化.

6 视觉语言表征学习的发展趋势

从目前发展趋势及表征质量测评效果来看基于预训练架构的视觉语言表征学习方法相较于基于相似性的表征学习有一定的优势,但同时其产出表征质量的好坏对海量的预训练数据依赖也比较大,所以基于相似性的表征学习在一些数据相对匮乏的专有领域会有一定优势,综合不同表征学习框架的优缺点和多模态表征的一些特点,未来有以下几点值得深度研究:

1) 支持内容理解与内容生成的通用表征框架:目前基于预训练的统一表征框架大多偏向内容理解方向,比如 ViLBERT、VisualBERT、ImageBERT 等,针对图像描述等生成类任务的预训练框架以及理解与生成通用的预训练框架也是未来研究方向,XGPT[90] 在预训练阶段采用图像描述任务作为预训练任务,其引入三类跨模态生成类预训练任务,包括图像为条件的语言掩码任务,以图像为条件的降噪自编码任务,以文本为条件的图像特征生成任务.生成类预训练任务与理解类预训练任务的一个很大不同是生成类预训练任务既引入编码架构又引入解码结构,同时生成类预训练任务中也增加了序列到序列的预测任务,未来如何更好的构建更加通用的预训练框架是一个值得研究的问题.

2) 训练及推断性能提升:目前基于预训练架构的视觉语言统一表征虽然在视觉问答、跨模态检索等任务中相比较原来的架构有较大的提升,但是在进行实际推断任务时其速度较慢,纠其原因主要分为几个方面:(1) 大部分框架中的图像特征主要采用基于 Faster-RCNN[68] 两阶段目标检测的方式提取,虽然精度有一定保证但是速度很慢.这方面优化可以采用效率更高的单阶段检测框架,或者更换骨干网络,比如用 ResNeXt 替

换原有骨干网络.（2）基于 Transformer 架构的模型计算量大,参数较多,所以可以采用蒸馏、量化、压缩等手段进行提升,比如 TinyBERT[91] 通过两阶段蒸馏的方式,同时对预训练任务和下游任务进行蒸馏,教师模型和学生模型优化的损失函数分别为隐含层损失和注意力矩阵损失,其分别对预训练任务和下游任务同时进行蒸馏操作,TinyBERT 模型大小比 BERT-BASE 小 7.5 倍,推断速度为其 9 倍,在实际应用中可以结合具体的多模态预训练任务利用蒸馏的方法进行提速.另外 transformer 也有一些实现速度提升的变种,比如基于因式分解的稀疏 Transformer[92] 和利用局部敏感哈希替换点积运算的 Reformer[93] 等,可以利用这些模型改造后替换 Transformer 的原始模型.

3) 细粒度特征挖掘:无论是基于相似性的还是基于预训练的框架,更精细粒度的特征提取是提升表征质量的一个很好的方向,目前有一些视觉语言统一表征的预训练模型是基于图像的像素级输入的,比如 MMBT 模型[94] 就是通过 ResNet 算法提取图像特征,然后通过一个池化卷积操作输出不同特征映射单元作为视觉 token 输入,再将视觉词组与文本词组作为 Transformer 结构的联合输入,但是这种结构相比于单纯的将文本的输出向量和图像输出向量融合的方式提升精度并不高.另外就是 Pixel-BERT[1] ,为了解决基于特征提取方式提取视觉特征导致的分类数目有限的问题,其采用像素级特征表示视觉模态,通过采用随机采样像素点的方式避免过拟合,在视觉问答等下游任务中表现较好,超越了 ViLBERT 和 UNITER 等模型.ERNIE-ViL 是采用场景图预测的方式,将句子分割成物体、属性、关系的三元组,然后与图像信息进行联合预测,还有一种思路就是可以引入知识图谱作为实体信息的补充,从而进行知识增强.针对视觉模态可以挖掘更多的高层语义信息,比如人脸特征、文字识别特征等.

7 总结

文章首先介绍了一些相应的背景知识包括表征学习的主要研究思路包括基于概率图的模型和神经网络的模型,同时介绍了多模态统一表征的划分和预训练技术.然后介绍了视觉语言表征的几种研究方向,包括基于相似性的视觉语言表征学习以及基于预训练架构的视觉语言统一表征学习,其中基于预训练架构的模型为近年来研究重点,相关领域产生的一些成果较多,文章从模型结构、预处理方案、预训练任务、下游任务等角度进行了分别阐述,针对多模态表征的质量评估文章介绍了零样本学习评估、面向具体任务的评估、预训练加下游任务评估等几种方式.最后文章结合目前视觉语言表征的一些待解决问题和一些新兴的研究思路介绍了视觉语言表征学习的未来发展趋势.多模态视觉语言表征学习目前被越来越多的研究者所重视,并且成为了目前极其火热的一个研究方向,相信该领域未来可以更好的推动多模态学习和人工智能的发展.

8 致谢

在此,向对本文在组织撰写过程中提供帮助的老师和同学们表示感谢.

References:

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems. 2017: 5998-6008.

[2] Yuhas B P, Goldstein M H, Sejnowski T J. Integration of acoustic and visual speech signals using neural networks. IEEE Communications Magazine, 1989, 27(11): 65-71.

[3] Juang B H, Rabiner L R. Hidden Markov models for speech recognition. Technometrics, 1991, 33(3): 251-272.

[4] Busso C, Bulut M, Lee C C, *et al*. IEMOCAP: interactive emotional dyadic motion capture database . Language Resources & Evaluation, 2008, 42 (4): 335.

[5] McKeown G, Valstar M, Cowie R, et al. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE transactions on affective computing, 2011, 3(1): 5-17.

[6] Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(2): 423-443.

[7] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8): 1798-1828.

[8] Lowe D G. Distinctive image features from scale-invariant keypoints. International journal of computer vision, 2004, 60(2): 91-110.

[9] Dalal N, Triggs B. Histograms of oriented gradients for human detection. 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). IEEE, 2005, 1: 886-893.

[10] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012: 1097-1105.

[11] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine, 2012, 29(6): 82-97.

[12] Chowdhury G G. Introduction to modern information retrieval[M]. Facet publishing, 2010.

[13] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[14] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[15] Wang A, Singh A, Michael J, et al. Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461, 2018.

[16] Rifai S, Vincent P, Muller X, et al. Contractive Auto-Encoders: Explicit Invariance During Feature Extraction. international conference on machine learning, 2011: 833-840.

[17] Mroueh Y, Marcheret E, Goel V, et al. Deep multimodal learning for Audio-Visual Speech Recognition. international conference on acoustics, speech, and signal processing, 2015: 2130-2134.

[18] Ngiam J, Khosla A, Kim M, et al. Multimodal Deep Learning. international conference on machine learning, 2011: 689-696.

[19] Kim Y, Lee H, Provost E M, et al. Deep learning for robust feature generation in audiovisual emotion recognition. international conference on acoustics, speech, and signal processing, 2013: 3687-3691.

[20] Kahou S E, Bouthillier X, Lamblin P, et al. Emonets: Multimodal deep learning approaches for emotion recognition in video. Journal on Multimodal User Interfaces, 2016, 10(2): 99-111.

[21] Nicolaou M A, Gunes H, Pantic M. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Transactions on Affective Computing, 2011, 2(2): 92-105.

[22] Silberer C, Lapata M. Learning Grounded Meaning Representations with Autoencoders. meeting of the association for computational linguistics, 2014: 721-732.

[23] Srivastava N, Salakhutdinov R. Multimodal Learning with Deep Boltzmann Machines. neural information processing systems, 2012: 2222-2230.

[24] Rajagopalan S S, Morency L P, Baltrusaitis T, et al. Extending long short-term memory for multi-view structured learning. European Conference on Computer Vision. Springer, Cham, 2016: 338-353.

[25] Frome A, Corrado G S, Shlens J, et al. Devise: A deep visual-semantic embedding model. Advances in neural information processing systems, 2013: 2121-2129.

[26] Kiros R, Salakhutdinov R, Zemel R S. Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539, 2014.

[27] Vendrov I, Kiros R, Fidler S, et al. Order-embeddings of images and language. arXiv preprint arXiv:1511.06361, 2015.

[28] Zhang D, Li W. Large-scale supervised multimodal hashing with semantic correlation maximization. national conference on artificial intelligence, 2014: 2177-2183.

[29] Cao Y, Long M, Wang J, et al. Deep visual-semantic hashing for cross-modal retrieval. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 1445-1454.

[30] Yang G, Miao H, Tang J, et al. Multi-kernel Hashing with Semantic Correlation Maximization for Cross-Modal Retrieval. international conference on image and graphics, 2017: 23-34.

[31] You Q, Zhang Z, Luo J. End-to-end convolutional semantic embeddings. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 5735-5744.

[32] Alberti C, Ling J, Collins M, et al. Fusion of detected objects in text for visual question answering. arXiv preprint arXiv:1908.05054, 2019.

[33] Lu J, Batra D, Parikh D, et al. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. neural information processing systems, 2019: 13-23.

[34] Shi B, Ji L, Lu P, et al. Knowledge Aware Semantic Concept Expansion for Image-Text Matching. international joint conference on artificial intelligence, 2019: 5182-5189.

[35] Chen Y,Li L, Yu L, et al. UNITER: Learning Universal Image-Text Representations. arXiv: Computer Vision and Pattern Recognition, 2019.

[36] Su W, Zhu X, Cao Y, et al. Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530, 2019.

[37] Tan H, Bansal M. Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490, 2019.

[38] Li G, Duan N, Fang Y, et al. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training. arXiv: Computer Vision and Pattern Recognition, 2019.

[39] Li L H, Yatskar M, Yin D, et al. VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv: Computer Vision and Pattern Recognition, 2019.

[40] Qi D, Su L, Song J, et al. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. arXiv preprint arXiv:2001.07966, 2020.

[41] Pan Y, Mei T, Yao T, et al. Jointly Modeling Embedding and Translation to Bridge Video and Language. computer vision and pattern recognition, 2016: 4594-4602.

[42] Xu R, Xiong C, Chen W, et al. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. national conference on artificial intelligence, 2015: 2346-2352.

[43] Luo H, Ji L, Shi B, et al. Univilm: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353, 2020.

[44] Sun C,Myers A, Vondrick C, et al. Videobert: A joint model for video and language representation learning. Proceedings of the IEEE International Conference on Computer Vision. 2019: 7464-7473.

[45] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model. Journal of machine learning research, 2003, 3(Feb): 1137-1155.

[46] He K, Girshick R, Dollár P. Rethinking imagenet pre-training. Proceedings of the IEEE international conference on computer vision. 2019: 4918-4927.

[47] Hendrycks D, Lee K, Mazeika M. Using pre-training can improve model robustness and uncertainty. arXiv preprint arXiv:1901.09960, 2019.

[48] He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9729-9738.

[49] Peters M E, Neumann M, Iyyer M, et al. DEEP CONTEXTUALIZED WORD REPRESENTATIONS. north american chapter of the association for computational linguistics, 2018: 2227-2237.

[50] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. 2018.

[51] Lee K H, Chen X, Hua G, et al. Stacked cross attention for image-text matching. Proceedings of the European Conference on Computer Vision (ECCV). 2018: 201-216.

[52] Weston J, Bengio S, Usunier N. Large scale image annotation: learning to rank with joint word-image embeddings. Machine Learning, 2010, 81(1):21-35.

[53] Lazaridou A, Pham N T, Baroni M. Combining language and vision with a multimodal skip-gram model. arXiv preprint arXiv:1501.02598, 2015.

[54] Faghri F, Fleet D J, Kiros J R, et al. Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612, 2017.

[55] Huang P S, He X, Gao J, et al. Learning deep structured semantic models for web search using clickthrough data. Proceedings of the 22nd ACM international conference on Information & Knowledge Management. 2013: 2333-2338.

[56] Fang H, Gupta S, Iandola F, et al. From captions to visual concepts and back. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1473-1482.

[57] Hubert Tsai Y H, Huang L K, Salakhutdinov R. Learning robust visual-semantic embeddings. Proceedings of the IEEE International Conference on Computer Vision. 2017: 3571-3580.

[58] Huang Y, Wu Q, Song C, et al. Learning semantic concepts and order for image and sentence matching. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6163-6171.

[59] Wu H, Mao J, Zhang Y, et al. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6609-6618.

[60] Wang Y, Yang H, Qian X, et al. Position focused attention network for image-text matching. arXiv preprint arXiv:1907.09748, 2019.

[61] Sun C, Baradel F, Murphy K, et al. Learning Video Representations using Contrastive Bidirectional Transformer. arXiv preprint arXiv:1906.05743, 2019.

[62] Zhou L, Palangi H, Zhang L, et al. Unified Vision-Language Pre-Training for Image Captioning and VQA.AAAI. 2020: 13041-13049.

[63] Yu F, Tang J, Yin W, et al. ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph. arXiv preprint arXiv:2006.16934, 2020.

[64] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909, 2015.

[65] Schuster M, Nakajima K. Japanese and korean voice search.2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2012: 5149-5152.

[66] He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks. ECCV. Springer, Cham, 2016: 630-645.

[67] Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition.2017: 1492-1500.

[68] Ren S, He K, Girshick R, et al. Faster-Rcnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems.2015:91-99.

[69] Anderson P, He X, Buehler C,et al. Bottom-up and top-down attention for image captioning and visual question answering. Proceedings of the IEEE conference on computer vision and pattern recognition.2018: 6077-6086.

[70] Krishna R, Zhu Y,Groth O,et al.Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision, 2017, 123(1): 32-73.

[71] Hu H, Gu J, Zhang Z, et al. Relation networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018: 3588-3597.

[72] Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017.

[73] Xie S, Sun C, Huang J, et al. Rethinking spatiotemporal feature learning for video understanding. arXiv preprint arXiv:1712.04851, 2017, 1(2): 5.

[74] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context. European conference on computer vision. Springer, Cham, 2014: 740-755.

[75] Sharma P, Ding N, Goodman S, et al. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 2556-2565.

[76] Ordonez V, Kulkarni G, Berg T L. Im2text: Describing images using 1 million captioned photographs.Advances in neural information processing systems. 2011: 1143-1151.

[77] Miech A, Zhukov D, Alayrac J B, et al. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. Proceedings of the IEEE international conference on computer vision. 2019: 2630-2640.

[78] Zhou L, Xu C, Corso J J. Towards automatic learning of procedures from web instructional videos. Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[79] Xu J, Mei T, Yao T, et al. Msr-vtt: A large video description dataset for bridging video and language. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 5288-5296.

[80] Wang X, Wu J, Chen J, et al. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. Proceedings of the IEEE International Conference on Computer Vision. 2019: 4581-4591.

[81] Goyal Y, Khot T, Summers-Stay D, et al. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering.Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6904-6913.

[82] Zellers R, Bisk Y, Farhadi A, et al. From recognition to cognition: Visual commonsense reasoning.Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6720-6731.

[83] Suhr A, Zhou S, Zhang A, et al. A corpus for reasoning about natural language grounded in photographs.arXiv preprint arXiv:1811.00491, 2018.

[84] Plummer B A,Wang L,Cervantes C M,et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models.Proceedings of the IEEE international conference on computer vision. 2015: 2641-2649.

[85] Farhadi A, Hejrati M, Sadeghi M A, et al. Every picture tells a story: Generating sentences from images. European conference on computer vision. Springer, Berlin, Heidelberg, 2010: 15-29.

[86] Vinyals O, Toshev A, Bengio S, et al. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(4): 652-663.

[87] Maaten L, Hinton G. Visualizing data using t-SNE. Journal of machine learning research, 2008, 9(Nov): 2579-2605.

[88] Chua T S, Tang J, Hong R, et al. NUS-WIDE: a real-world web image database from National University of Singapore. Proceedings of the ACM international conference on image and video retrieval. 2009: 1-9.

[89] Huang F, Zhang X, Li Z, et al. Learning social image embedding with deep multimodal attention networks. Proceedings of the on Thematic Workshops of ACM Multimedia 2017. 2017: 460-468.

[90] Xia Q, Huang H, Duan N, et al. Xgpt: Cross-modal generative pre-training for image captioning. arXiv preprint, 2020

[91] Jiao X, Yin Y, Shang L, et al. Tinybert: Distilling bert for natural language understanding. arXiv preprint arXiv:1909.10351, 2019.

[92] Child R, Gray S, Radford A, et al. Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509, 2019.

[93] Kitaev N, Kaiser Ł, Levskaya A. Reformer: The efficient transformer. arXiv preprint arXiv:2001.04451, 2020.

[94] Kiela D, Bhooshan S, Firooz H, et al. Supervised multimodal bitransformers for classifying images and text. arXiv preprint arXiv:1909.02950, 2019.

[95] Huang Z, Zeng Z, Liu B, et al. Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers. arXiv preprint arXiv:2004.00849, 2020.