工程科学学报
Chinese Journal of Engineering

多模态学习方法综述

陈鹏 李擎 张德政 杨宇航 蔡铮 陆子怡

A survey of multimodal machine learning

CHEN Peng, LI Qing, ZHANG De-zheng, YANG Yu-hang, CAI Zheng, LU Zi-yi

引用本文:
陈鹏, 李擎, 张德政, 杨宇航, 蔡铮, 陆子怡. 多模态学习方法综述[J]. 工程科学学报, 2020, 42(5): 557-569. doi: 10.13374/j.issn2095-9389.2019.03.21.003
CHEN Peng, LI Qing, ZHANG De-zheng, YANG Yu-hang, CAI Zheng, LU Zi-yi. A survey of multimodal machine learning[J]. *Chinese Journal of Engineering*, 2020, 42(5): 557-569. doi: 10.13374/j.issn2095-9389.2019.03.21.003

在线阅读 View online: https://doi.org/10.13374/j.issn2095-9389.2019.03.21.003

工程科学学报, 第 42 卷, 第 5 期: 557−569, 2020 年 5 月
Chinese Journal of Engineering, Vol. 42, No. 5: 557−569, May 2020
https://doi.org/10.13374/j.issn2095-9389.2019.03.21.003; http://cje.ustb.edu.cn

多模态学习方法综述

陈　鹏[1,2)], 李　擎[1,2)✉], 张德政[3,4)], 杨宇航[1)], 蔡　铮[1)], 陆子怡[1)]

1) 北京科技大学自动化学院, 北京 100083　2) 工业过程知识自动化教育部重点实验室, 北京 100083　3) 北京科技大学计算机与通信工程学院, 北京 100083　4) 材料领域知识工程北京市重点实验室, 北京 100083

✉通信作者, E-mail: liqing@ies.ustb.edu.cn

摘　要　大数据是多源异构的. 在信息技术飞速发展的今天, 多模态数据已成为近来数据资源的主要形式. 研究多模态学习方法, 赋予计算机理解多源异构海量数据的能力具有重要价值. 本文归纳了多模态的定义与多模态学习的基本任务, 介绍了多模态学习的认知机理与发展过程. 在此基础上, 重点综述了多模态统计学习方法与深度学习方法. 此外, 本文系统归纳了近两年较为新颖的基于对抗学习的跨模态匹配与生成技术. 本文总结了多模态学习的主要形式, 并对未来可能的研究方向进行思考与展望.

关键词　多模态学习; 统计学习; 深度学习; 对抗学习; 特征表示

分类号　TP18

A survey of multimodal machine learning

CHEN Peng[1,2)], *LI Qing*[1,2)✉], *ZHANG De-zheng*[3,4)], *YANG Yu-hang*[1)], *CAI Zheng*[1)], *LU Zi-yi*[1)]

1) School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China

2) Key Laboratory of Knowledge Automation for Industrial Processes, Ministry of Education, Beijing 100083, China

3) School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

4) Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China

✉ Corresponding author, E-mail: liqing@ies.ustb.edu.cn

ABSTRACT　"Big data" is always collected from different resources that have different data structures. With the rapid development of information technologies, current precious data resources are characteristic of multimodes. As a result, based on classical machine learning strategies, multi-modal learning has become a valuable research topic, enabling computers to process and understand "big data". The cognitive processes of humans involve perception through different sense organs. Signals from eyes, ears, the nose, and hands (tactile sense) constitute a person's understanding of a special scene or the world as a whole. It reasonable to believe that multi-modal methods involving a higher ability to process complex heterogeneous data can further promote the progress of information technologies. The concepts of multimodality stemmed from psychology and pedagogy from hundreds of years ago and have been popular in computer science during the past decade. In contrast to the concept of "media", a "mode" is a more fine-grained concept that is associated with a typical data source or data form. The effective utilization of multi-modal data can aid a computer understand a specific environment in a more holistic way. In this context, we first introduced the definition and main tasks of multi-modal learning. Based on this information, the mechanism and origin of multi-modal machine learning were then briefly introduced. Subsequently, statistical learning methods and deep learning methods for multi-modal tasks were comprehensively summarized. We also introduced the main styles of data fusion in multi-modal perception tasks, including feature representation, shared mapping, and co-training. Additionally, novel adversarial learning strategies for cross-modal matching or generation were reviewed. The main methods for multi-modal learning were outlined in this paper

收稿日期: 2019−03−21

基金项目: 国家重点研发计划(云计算和大数据专项)资助项目(2017YFB1002304)

with a focus on future research issues in this field.

KEY WORDS multi-modal learning; statistical learning; deep learning; adversarial learning; feature representation

早在公元前 4 世纪, 多模态的相关概念和理论即被哲学家和艺术家所提出, 用以定义融合不同内容的表达形式与修辞方法[1−2]. 20 世纪以来, 这一概念被语言学家更为广泛地应用于教育学和认知科学领域[3]. 近年来, 描述相同、相关对象的多源数据在互联网场景中呈指数级增长, 多模态已成为新时期信息资源的主要形式.

人类的认知过程是多模态的. 个体对场景进行感知时往往能快速地接受视觉、听觉乃至嗅觉、触觉的信号, 进而对其进行融合处理和语义理解. 多模态机器学习方法更贴近人类认识世界的形式. 本文首先介绍了多模态的概念与基本任务, 分析了多模态认知学习的起源与发展. 结合互联网大数据形态, 本文重点综述了多模态统计学习方法、深度学习方法与对抗学习方法.

1 多模态学习的定义、基本任务与发展过程

1.1 多模态学习的定义

本文主要采用了新加坡国立大学 O′Halloran 对“模态”的定义, 即相较于图像、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[4]. 概括来说, “多模态”可能有以下三种形式.

(1)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. 图 1 即为典型的多模态信息形式.

(2)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.

(3)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; 描述同一语义的词向量、词袋、知识图谱以及其它语义符号单元等[5].

因此, 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(Multimodal machine learning)”, 其相关概念有“多视角学习”和“多传感器信息融合”. 来自不同数据源或由不同特征子集构成的数据被称作多视角数据, 每个数据源、每种数据类型均可被看作一个视角. 卡内基梅隆大学的 Morency 在 ACL2017(The 55th Annual Meeting of the Association for Comput-

图 1 “下雪”场景的多模态数据(图像、音频与文本)

Fig.1 Multimodal data for a “snow” scene (images, sound and text)

ational Linguistics, CCF A 类会议）的 Tutorial 报告[6]中，将大量的多视角学习方法归类为多模态机器学习算法．笔者认为，“多视角学习”强调对数据“视角”的归纳和分析，“多模态学习”则侧重“模态”感知和通道．“视角”和“模态”的概念是相通的，一个模态即可被视作一个视角．“多传感器信息融合（Multi-sensor information fusion）”为在物理层面与“多模态学习”相关的术语，即对不同传感器采集的数据进行综合利用，其典型应用场景有物联网、自动驾驶等．

1.2　多模态机器学习的基本任务

多模态学习的基本任务可包括以下几个方面．

多源数据分类：单模态的分类问题只关注对一类特定数据的分析和处理，相较于单一通道，多模态数据更接近大数据背景下信息流真实的形态，具有全面性和复杂性．

多模态情感分析：情感分析问题的本质也是分类问题，与常规分类问题不同，情感分类问题所提取的特征往往带有明确的情绪信号；从多模态的角度分析，网络社交场景中所衍生的大量图片、文本、表情符号及音频信息均带有情感倾向．

多模态语义计算：语义分析是对数据更为高层次的处理，理想状态下，计算机能够处理一个特定场景下不同数据的概念关系、逻辑结构，进而理解不同数据中隐含的高层语义；对这种高层语义的理解是有效进行推理决策的前提．

跨模态样本匹配：现阶段，最常见的跨模态信息匹配即为图像、文本的匹配，如 Flickr30k[7] 数据集中的实例；图像文本匹配任务为较为复杂的机器学习任务，这一任务的核心在于分别对图像、文本的特征进行合理表示、编码，进而准确度量其相似性．

跨模态检索：在检索任务中，除了实现匹配外，还要求快速的响应速度以及正确的排序；多模态信息检索通过对异构数据进行加工，如直接对图片进行语义分析，在有效特征匹配的情况下对图片采用基于内容的自动检索形式；为适应快速检索的需要，哈希方法被引入多模态信息检索任务中，跨模态哈希方法将不同模态的高维数据映射到低维的海明空间，有效减小了数据存储空间，提高了计算速度．

跨模态样本生成：跨模态生成任务可以有效构造多模态训练数据，同时有助于提高跨模态匹配与翻译的效果，目前由图像到文本（如图像语义自动标注）、图像到图像（如图片风格迁移）的生成任务发展较为成熟，由文本到图像的生成任务则较为新颖．

多模态人机对话：即在基本对话（文本模态）生成任务的基础上，进一步对人的表情、语调、姿势等多模态信息进行采集，采用模态融合的方法对多模态信号进行分析处理．多模态人机对话的理想状态是在有效感知多模态信号的前提下给出拟人化的多模态输出，构建更为智能、沟通更加顺畅的人机交互形式．

多模态信息融合：多模态融合要求对多源数据进行综合有效地筛选和利用，实现集成化感知与决策的目的，常见的信息融合方式有物理层融合、特征层融合、决策层融合几个类型．物理层融合指在感知的第一阶段，在传感器层级对采集到的数据进行融合处理，这种处理方式可被概括为多传感器信息融合（Multi-sensor information fusion），是工业生产场景中极为常见的信息融合方法；特征层融合指在特征抽取和表达的层级对信息进行融合，如对同一场景中不容摄像头采集到的图像采用相同的特征表达形式，进而进行相应的叠加计算；决策层融合指对不同模态的感知模型所输出的结果进行融合，这种融合方式具有较好的抗干扰性能，对于传感器性能和种类要求相对不高，但具有较大的信息损耗．

1.3　多模态机器学习的发展——从符号计算到深度学习

随着计算机技术的发展，多模态认知的概念从传统的教育学、心理学、语言学的范畴拓展至信息科学领域．20 世纪 60 ~ 70 年代，科学家利用符号和逻辑结构模拟人类的思维逻辑，如利用语法树分析文本信息[8]，利用规则库构建专家决策系统[9]．由于人类认知过程的复杂性与流动性，有效、实时地制定逻辑结构和规则形式成为制约“符号主义”认知智能的主要因素．

20 世纪 80 年代至 21 世纪初，统计机器学习方法在智能信息处理的各个领域取得了令人瞩目的成就．Cortes 和 Vapnik 提出的支持向量机模型可以快速、准确地处理高维、非线性的模式识别问题[10]；Pearl 所构建的概率图模型赋予了计算机依据概率推理的能力[11]；进一步地，Jelinek 将信息论与隐马尔科夫模型引入语音识别与自然语言处理领域，奠定了近代统计自然语言处理学派的根基，使自然语言处理的工程化应用成为可能[12]．

在这一阶段，受麦格克效应的启发[13]，许多计算机科学家致力于构建基于视觉信号和声音信号

的多模态语音识别系统, 如唇语–声音语音识别系统[14], 有效提高了识别准确率. 这一时期的多模态信息系统还被应用于人机交互场景, 如 Fels 等提出的 Glove-talk 框架(1992 年)采用 5 个多层神经网络实现对手势、声音、语义的机器感知[15]. 这一神经网络模型的结构还比较简单, 其采用的后向传播训练方法易出现过拟合现象, 因而无法对复杂的大规模数据进行处理.

2010 年至今, 随着 Dropout 训练模式[16] 的提出、Relu 激活函数[17] 的引入乃至深度残差结构[18] 对网络的调整, 深度神经网络在许多单一模态的感知型机器学习任务中取得了优于传统方法的效果. 以 AlexNet[19]、ResNet[18]、GoogleNet[20] 为代表的改进卷积神经网络(Convolutional neural network, CNN)模型在 ImageNet[21] 图像分类任务中甚至取得了超过人类的表现; 长短记忆模型(Long short term memory, LSTM)和条件随机场(Conditional random field, CRF)的组合结构在自然语言序列标注特别是命名实体识别任务中实现了极为成功的商业化、工程化应用[22]. 多模态深度学习已成为人工智能领域的热点问题. Ngiam 等在 ICML2011 (28th International Conference on Machine Learning) 的大会论文中对多模态深度学习进行了前瞻性的综述, 而这一阶段的深度学习主要网络结构为深度玻尔兹曼机(Deep boltzmann machines)[23]. 卡内基梅隆大学的 Baltrusaitis 等也开展了大量的多模态深度学习研究[24].

在国内, 北京交通大学的 Zhang 等[25], 北京邮电大学的 Wang 等在跨模态信息匹配和检索领域开展了许多卓有成效的工作[26]; 清华大学的 Liu 等对视觉模态、触觉模态的数据展开研究, 并将其应用于机器人综合感知场景[27]; 清华大学的 Fu 等则在图像语义标注领域取得了若干突破[28].

在人工智能技术突飞猛进的今天, 开展数据驱动的多模态学习方法研究, 能够取得更为全面有效的解决方案. 对多模态数据的分析处理可采用机器学习手段来完成, 处理多模态数据的机器学习方法即可被视为多模态学习方法. 机器学习是利用数据优化算法的一种人工智能手段, 它涵盖统计学习与深度学习等方法. 近几年, 对抗学习技术被广泛地应用于跨模态匹配和生成任务中, 并取得了令人瞩目的效果. 后文将分别对多模态统计学习方法、多模态深度学习方法、多模态对抗学习方法进行综述与分析.

2　多模态统计学习方法

广义的统计学习(Statistical learning)即采用统计学的相关理论, 赋予计算机处理数据能力的机器学习方法. 如统计学家和数学家 Breiman 提出的随机森林(Random forest)算法[29], Breiman 和 Friedman 等一同提出的分类回归树(Classification and regression trees, CART)算法[30], Cortes 和 Vapnik 提出的支持向量机(Support vector machine, SVM)算法[10] 等. 统计学习方法和经典机器学习方法在概念上是基本重合的. 上述统计学习界的领军学者分别在不同角度完善了该领域的基本概念和理论体系. 如 Breiman 在数据建模和算法建模两个角度重新解读了机器学习的建模方式, 即数据建模方式往往预设数据符合某种分布形式, 如线性回归、逻辑回归等, 进而进行参数估计和假设推断; 而算法建模则试图通过算法去直接寻找映射函数以达到由输入预测输出的目的, 如决策树与神经网络结构[31]. Vapnik 和 Cervonenkis 归纳了他的 VC(Vapnik–Chervonenkis dimension)维理论, 不仅对典型的分类器模型与这些模型所能区分的集合大小进行系统总结, 还给出了对模型最大分类能力进行分析的有效方法[32].

受计算资源等因素的制约, 统计学习方法的处理样本往往是中小规模的数据集, 在许多任务(如图像处理和自然语言处理任务)的处理过程中, 需要人参与的特征处理过程. 多模态机器学习技术是伴随着统计学习理论的完备、大量新颖有效的统计学习方法的提出逐渐发展的. 本节将结合多模态数据的特点, 对相应的统计学习方法进行介绍.

2.1　核学习方法与多核学习

核学习(Kernel learning)方法是一种将低维不可分样本通过核映射的方式映射到高维非线性空间, 实现对样本有效分类的方法[33], 如图 2 所示. 核学习方法是支持向量机(SVM)算法的有力理论支撑, 也随着支持向量机的广泛应用被研究者和工程技术人员所关注. 事实上, 早在 1909 年, 英国数学家 Mercer 即提出了其重要的 Mercer 定理, 即任何半正定的函数都可作为核函数, 奠定了核学习方法的理论基础[34]. 在 Mercer 定理的基础上, 波兰裔美国数学家 Aronszajn 进一步发展了再生核希尔伯特空间理论, 使其能够被引入到模式识别任务中[35].

多核学习方法为不同模态的数据和属性选取

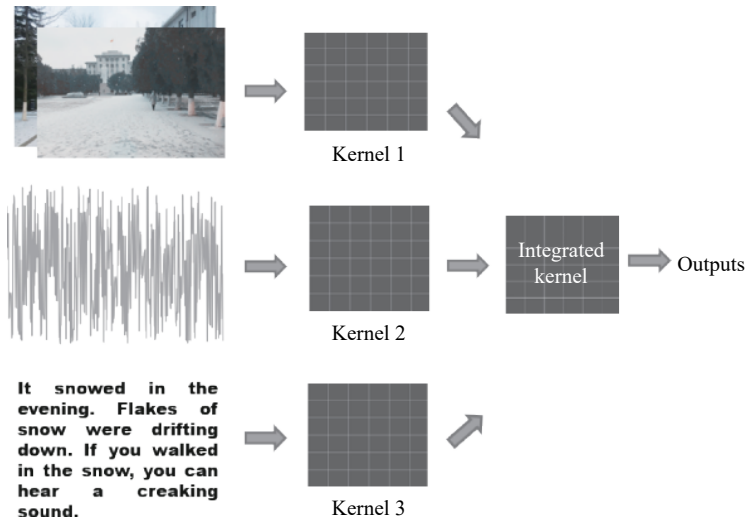

图 2　多核学习

Fig.2　Multi-kernel learning

不同的核函数，进而采用特定方法对不同核函数进行融合．目前，随着多核学习方法被深入研究并应用于不同的场景，不同形式的核函数及其改进形式被提出．如对于数值型数据的分类问题，高斯核具有较好的处理效果[36]；字符串核对序列型问题的分类处理（如文本、音频、基因表达等）具有较大的优势[37]；对于人脸识别问题和行人识别问题，则可以采用直方图交叉核[38]．

多核学习方法可以较好地处理异构数据的分类和识别问题．早期的多核数据融合方法多采用对不同核进行线性叠加组合的形式，为生物医学工程领域许多问题的求解（如基因功能分析、蛋白质功能预测与定位等）提供了有力的解决方案[39]．线性叠加的核融合方式具有机理简单、可解释性强、计算速度快等优势，但其叠加系数往往较难确定，在叠加的同时可能造成一定的信息损失．文献 [40] 提出采用“核组合”的方式解决该问题，即将不同的核矩阵组合，构成一个更高维的矩阵作为新的核矩阵完成映射与分类的任务．文献 [41] 提出了一种改进的判别函数，并采用梯度下降法优化该表达式中的核参数．文献 [42] 则采用粒子群优化算法对核参数进行优化选择．

2.2　典型相关性分析

典型相关性分析（Canonical correlation analysis, CCA）是一种用途广泛的统计学分析算法，由 Hotelling 于 1935 年提出[43]，并由 Cooley 和 Lohnes 推动其发展[44]．在多模态领域，CCA 被广泛地应用于度量两种模态信息之间的相关特征，并在计算中尽可能保持这种相关性．

CCA 算法的本质是一种线性映射，采用 CCA 对复杂的非线性多模态信息进行拟合可能造成信息的损耗．在 CCA 的基础上，Akaho 提出了与核方法结合的非线性的 Kernel CCA 算法[45]．CCA 的其他改进形式还有判别典型相关分析（Discriminant canonical correlation analysis, DCCA）[46]、稀疏典型相关分析（Sparse discriminant canonical correlation analysis, SCCA）等[47]．

2.3　共享子空间学习

在高层语义空间中，多源数据具有较强的相关性．对于底层的特征表示，不同来源的数据往往具有较大差别．共享子空间学习对多源数据的相关关系进行挖掘，得到多模态特征的一致性表示，如图 3 所示．

共享子空间学习可通过投影的方式实现，最常见的投影方法即 2.2 节中给出的 CCA 方法及其改进形式．SVM-2K 算法是投影型共享子空间学习的典型算法，该算法结合 SVM 与 Kernel CCA[45] 对两个模态的特征进行有效映射、表示和整合[48]．张量分析及因子分解也是典型的共享子空间学习方法，这种方法的主要思想是将一个模态的信息看作一阶张量，通过因子分解、判别式分析等形式实现降维并对特征进行相关表示，其典型方法为联合共享非负矩阵分解（Joint shared nnnegative matrix factorization, JSNMF）算法[49]．从任务驱动的角度来分类，典型的共享子空间学习方法还有基于多任务学习的共享子空间学习方法[50]、基于多标签学习的共享子空间学习方法等[51]．

基于统计学习的子空间投影的形式相对简单，难以处理较为复杂的语义感知任务，对于相似模态的数据（如不同传感器的图像数据）优势明

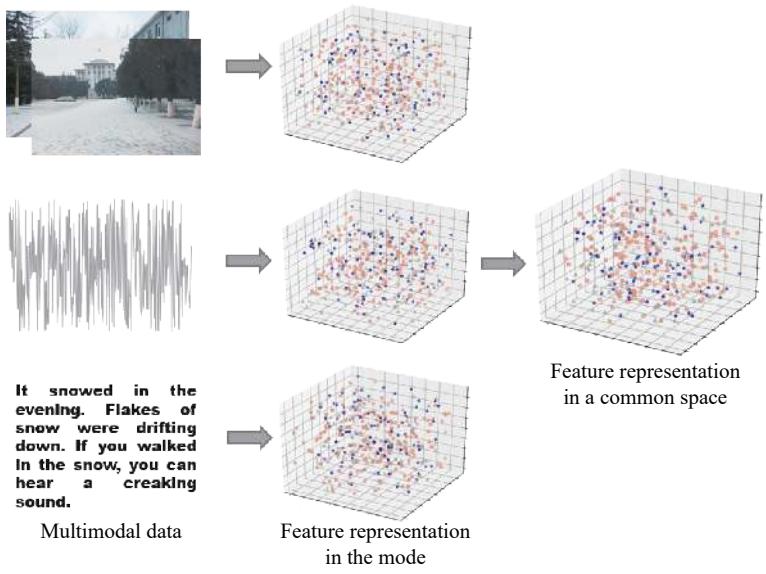

图 3　共享子空间学习

Fig.3　Common subspace learning

显, 但在跨度较大的模态上表现不佳. 近年来, 随着深度学习的兴起, 许多研究者将深度学习模型应用于多源信息处理领域. 从结果上来看, 绝大多数的深度学习多源信息处理方法将不同模态的数据通过深度神经网络特征学习映射到了同一个共享子空间, 因此深度学习方法也可被视为共享子空间学习. 对该方法将在第三部分中作进一步的介绍.

2.4　协同训练方法

协同训练(Co-training)是一种典型的弱监督学习方法, 该方法由 Blum 和 Mitchel 于 1998 年提出[52]. 在多模态数据处理领域, 它的大致思想是分别采用两个模态的有标签数据 X_1、X_2 训练两个分类器, 进而用这两个分类器对各自模态内的无标签数据进行处理. 在此基础上, 将分类结果中达到一定置信度的样本作为训练集的补充, 扩大训练集规模, 进一步对分类器进行训练. 在满足一定停止条件, 如达到一定迭代代数后, 将两个分类器的训练数据进行交换, 即采用 X_1 模态中的数据对分类器 2(Classifier$_2$)进行训练, 同时采用 X_2 中的数据对分类器 1(Classifier$_1$)进行训练. 协同训练的原理图如图 4 所示. 这种联合训练方法使分类器学习到不同数据源中尽可能多的知识, 同时具备了较好的泛化性能. 协同训练假定数据集满足三个条件: 1)数据之间相互独立; 2)单一模态内的数据均能完整地对对象进行描述; 3)存在充分的样本对

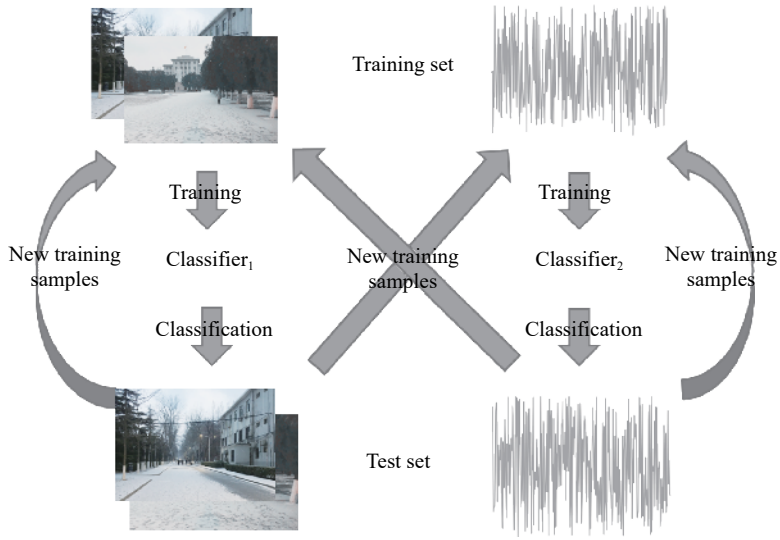

图 4　协同训练

Fig.4　Co-training

分类器进行训练. 然而在实际的应用场景中, 往往很难满足上述的条件. 研究者提出了多种改进手段以提升协同训练的性能.

文献 [53] 在协同训练中改进了多模态优化函数, 从而更为精确地处理拼写与上下文间的一致性信息; 文献 [54] 将支持向量机和期望最大算法 (Expectation maximization, EM) 相结合提出 Co-EM 算法, 提升多模态分析性能; 文献 [55] 在 Co-EM 的基础上进一步引入主动学习 (Active learning) 策略, 提高了算法的鲁棒性.

3　多模态深度学习方法

基本的神经网络模型 (浅层结构) 可被归纳为一种特殊的统计学习方法. 不同于支持向量机的核技巧采用核映射转化问题, 神经网络结构直接采用非线性映射 (激活函数) 的形式拟合数据分布规律. 神经网络是深度学习的起源, 后者是对采用深度神经网络完成机器学习任务的各种机器学习方法的概括. 近年来, 深度学习方法已成为推动人工智能技术的主要力量. 隐层大于 1 的神经网络即可被看作深度神经网络, 常见的深度神经网络模型有卷积神经网络 (Convolutional neural networks, CNN)[56]、循环神经网络 (Recurrent neural networks, RNN)[57]、深度信念网络 (Deep belief networks, DBN)[58] 等. 深度学习的发展建立在统计学习的高度繁荣之上, 得益于不断发展的互联网技术积累了大量的数据资源, 以及更为普及的高性能计算硬件. 有别于统计学习依赖于专家知识来确定特征的限制, 深度学习模型可以自动地在数据中学习特征表示, 从而能够对海量数据进行处理, 在一定程度上实现端到端的机器学习系统.

3.1　卷积神经网络与图像处理

Lécun 于 1998 年提出了经典卷积神经网络的雏形 LeNet, 并将其应用于手写字符识别[56]. 针对 CNN 训练过程中的过拟合问题, Srivastava 等提出了 Dropout 方法, 即在网络结构中以一定概率将某些神经元暂时丢弃[16]. 这种方法被应用于 AlexNet[19] 中. 在 AlexNet 之后, 改进了的 CNN 结构不断刷新 ImageNet 图像分类的记录. 如牛津大学的 VGG (Visual geometry group)[59] 模型和 Google 公司的 Inception[20] 系列模型, 在增加 CNN 网络层数的同时设计了精巧丰富的卷积核结构, 从而降低参数数量, 提高训练速度. 微软公司的 ResNet[40] 模型引入残差结构, 有效解决了梯度消失问题. 在图像分类之外的计算机视觉任务中, CNN 同样取得了优于经典图像处理方法的效果. 如目标检测 (Object detection) 领域的 Yolo (You only look once) 模型[60], 语义分割 (Semantic segmentation) 领域的 FCNN (Fully convolutional networks) 模型[61] 等. 有理由认为, CNN 及其改进形式能够较好地对视觉模态特征进行表示和处理.

此外, 对于文本数据, CNN 也体现出卓越的性能. 文献 [62] 采用 CNN 对短文本进行分类, 在保证可靠精度的同时提高分类速度. 文献 [63] 提出基于序列的深度卷积语义分析模型, 采用卷积结构生成句子的向量化表示, 进而进行深层分析. 文献 [64] 中也采用 CNN 对句子进行建模, 并将这种建模方法应用于句子匹配.

3.2　循环神经网络与自然语言理解

近年来, 自然语言处理域的研究热点正在从经典的统计学习方法向深度学习方法转变. 典型的深度文本处理模型即循环神经网络 (Recurrent neural network, RNN) 结构[57]. 该结构源于蒙特利尔大学 Bengio 等于 2003 年提出的神经语言模型[65]. 神经语言模型实现了语言最基本的单元——词的向量化表示. 受文献 [65] 启发, C&W 词向量[66]、Word2Vec 词向量[67] 等文本表示模型相继被提出.

神经语言模型的提出使文本转化为稠密的向量成为可能, 已成为目前处理自然语言任务的主流算法. 值得一提的是, 文献 [65] 至 [67] 中的文本表示及学习方法均为较为浅层的结构, 其价值在于通过弱监督、无监督的手段得到文本的表示形式, 进而供较为深层的神经网络机器学习模型进行挖掘分析.

在神经语言模型的基础上, 大量的深度神经网络结构被改良并进一步应用于自然语言处理任务, 如 RNN[57]、LTSM[68] 被广泛地应用于文本分类[69]、实体识别[22] 等任务. 由于 RNN 能够出色地学习序列样本中不同时刻的信息及其相互关系, RNN 结构在机器翻译、对话生成等序列分析及序列生成任务中的优势极为突出[70]. RNN 的主要改进形式为 LSTM[68] 和 GRU (Gated recurrent unit)[71]. 这些变体在 RNN 中添加了特殊的“门”结构来判断信息的价值, 进而模拟人类大脑的记忆和遗忘过程. 在 LSTM 的基础上, 其双向形式 BiLSTM[72]、基于 Attention 的 BiLSTM[73] 相继被提出. 相较于经典的 RNN[57], LSTM[68] 和 GRU[71] 可以更有效地对序列进行建模, 建立更为精确的语义依赖关系. 在合理标注的前提下, RNN 结构在自然语言实体识别任务中已实现了极为出色的工程应用, 其典

型算法为 LSTM+CRF, 即通过 LSTM 提取深度特征, 用条件随机场(Conditional random field, CRF)模型进行文本序列标注[22].

此外, RNN 还能很好地处理时间序列数据, 即对数值模态进行分析预测[74]. 在语音识别领域, RNN 是最为出色的算法之一[75]. 该模型还能够出色地处理图像标注[76]、视频解析[77] 任务.

3.3 面向多模态数据的深度学习

通过上文分析, 可以发现深度学习模型具有更好的跨模态适应性. 多模态深度学习始于 Ngiam 等发表于 ICML 2011 的《Multimodal Deep Learning》, 文中的数据来源为视觉模态(唇语)和音频模态, 其构建的深度学习模型以玻尔兹曼机(Restricted boltzmann machine, RBM)为基本单元, 通过对视频、音频数据进行编码、联合表示、学习和重构, 实现对字母、数字的识别[23].

近年来, 已有很多卓有成效的多模态深度学习方法被提出. 如文献 [76] 在学习机制上进行改良, 即在对训练集进行学习时, 不再构建图片-句子标签之间的映射关系, 而是将图片中的对象和句子中的实体匹配起来, 首先对图片采取目标检测的任务, 进而学习单词和细粒度图像区域之间的关系, 在此基础上生成标注句子. 这一方式简化了对 Image-Caption 任务的训练集标注需求, 即从句子简化为单词. 文献 [77] 结合 LSTM 的特性, 构建了能够对多幅图像或视频内容进行理解和描述的深度神经网络框架, 实现对视觉序列的文本描述. 文献 [78] 设计了 CNN-LSTM 混合编码器对数据进行编码, 进而采用排序损失(Pairwise ranking loss)函数对数据进行训练. 文献 [79] 借鉴了在基于 RNN 的机器翻译任务中的研究进展, 用 CNN 替代 RNN 作为图片的编码器. 在设计模型框架的同时, 该文还提出了得到相关细节描述的概率公式. 文献 [80] 设计了基于图片的问答模型, 该模型能够根据 CNN 编码的图片和问题句子, 生成正确的问题答案. 文献 [81] 重点研究了采用 CNN 模型的基于内容的图片检索问题, 并分析了深度卷积神经网络对高维语义特征的有效表达能力. 文献 [82] 则采用多模态深度学习框架, 通过构建多个 LSTM 结构处理情感分类问题. 文献 [83] 提出一种多模态无监督机器翻译方法, 采用描述同一内容的图片链接跨语种语料, 实现语义对应与融合. 文献 [84] 采用强化学习的手段对文本和视觉场景进行匹配, 进而对自动驾驶决策进行推理.

4 多模态对抗学习方法

跨模态迁移与跨模态生成是多模态学习的常见任务. 针对多源异构的复杂数据, 迁移学习可以在不同模态间转化知识. 近年来, 基于对抗学习策略的迁移学习方法取得了优于经典迁移学习方法的性能. 跨模态生成任务有助于构造完整的多模态认知场景, 同时能够提高在不同模态间进行迁移、匹配与翻译的能力.

生成对抗网络(Generative adversarial networks, GAN)的基本框架由 Goodfellow 等于 2014 年提出[85]. 该框架主要由两个互为博弈的结构——生成器 G(Generator)和判别器 D(Discriminator)构成. 对 GAN 进行对抗式训练的主要目标在于得到一组高性能的 G 与 D, 使 G 能够生成足够真实的样本, 而 D 则能够对以假乱真的样本进行区分. GAN 的性能是在交互式的对抗学习中提高的.

文献 [86] 中提出的 DCGAN 方法将 CNN 结构和 GAN 结合, 赋予对抗学习强大的图片生成能力. 在文献 [87] 中, Wasserstein 距离被引入来替代经典的 KL 散度(Kullback–Leibler divergence), 该方式可有效避免 GAN 训练过程中的“模式崩溃”, 即只能生成有限模式图片的问题. 文献 [88] 则提出 CGAN 模型, 在 GAN 结构中结合条件变量, 这一“条件”可以是类别标签, 也可以是跨模态样本的向量化表示.

4.1 基于对抗学习的跨模态迁移与域适应

迁移学习是跨模态学习的有效方法. 在迁移学习中, 常采用源域、目标域的概念表述迁移对象. 源域涉及已学习到的数据源或问题, 目标域则包含需要采用迁移学习方法进行处理的数据或新问题. 在跨模态问题中, 可将数据全面、结果较好的模态作为源域, 将数据资源较为有限的模态作为目标域.

采用 GAN 的对抗学习域适应(跨模态分类、匹配)方法在近几年取得了令人瞩目的成绩. 文献 [89] 给出了采用 GAN 结构处理跨模态域适应问题的基本模型 ADDA(Adversarial discriminative domain adaptation). 在 ADDA 中, 两个不同模态的数据分别经由 CNN 编码. 判别器 D 对源域和目标域进行判别, 该对抗学习的过程能够对齐目标域、源域的特征, 从而能够将源域(模态 A)的分类器应用于目标域(模态 B). 在此基础上, 文献 [90] 设计了双向 GAN 结构进一步优化域适应性能. 文献 [91]、[92] 针对目标域的类别, 设计了多个生成-判别单元, 具有针对性地进行跨模态迁移. 文献 [93] 则采

用质心对齐的手段，强化对抗学习中跨模态特征对齐的效果.

4.2　基于对抗学习的跨模态生成

根据 O'Halloran 所给出的细粒度模态划分[4]，跨模态生成涉及“图像—图像”生成、“图像—文本”生成及“文本—图像”生成三个典型任务.

在由图像到图像的样本生成任务（如图像风格迁移、图像高分辨率重构）中，GAN 是最为成功的方法之一. 文献 [94] 中的 LAPGAN 算法采用拉普拉斯金字塔结构，以串联的形式在多个尺度采用生成—对抗的学习方法生成高质量图片. 文献 [95] 中的 SAGAN 将自然语言处理领域的 Attention 机制[96] 引入 GAN 模型，有效利用了图片中的全局信息和局部信息. 文献 [97] 提出 SNGAN，采用谱范数对网络参数进行归一化，从而能够有效调整梯度，提高 GAN 的优化性能. 文献 [98] 中提出的 BigGAN 采用 ResNet 为特征提取器，以图片类别标签作为条件输入，经过在 ImageNet 上的大量训练，能够取得极为逼真的高质量图片. 基于对抗学习的图片风格迁移方法有 pix2pix[99]、CycleGAN[100]、StarGAN[101]、MUNIT[102] 等. pix2pix[99] 以 CGAN 为基础，将目标样本作为条件变量，输入给 GAN 模型，同时采用了改进的 CNN 特征表示模型（U-Net[103]）. CycleGAN[100] 采用循环训练方法，首先采用对抗学习在目标域生成具有源域内容、目标域风格的图片，接着将该图片进一步变换至源域，构成一个循环. 这种循环训练方式不依赖于大量的训练样本，能够实现有效的弱监督图片生成. StarGAN[101] 在 CycleGAN 的基础上针对多个不同的域进行编码，通过互异的域标签和图片内容的叠加，实现多个域（模态）的切换. MUNIT[102] 则引入 ResNet 中的残差模块（Residual blocks），设计了更为巧妙的编码器和解码器，对图片内容和风格分别进行编码和训练，实现无监督跨模态样本生成.

在由图像生成文本的任务（如图像语义标注）中，CGAN 也是基本的方法. 该方法将图片向量作为 GAN 的条件，指导对图片标签的向量生成[88]. 由于文本模态自身的序列特点，在目前常见的以生成描述性句子为目标的图像语义标注任务、基于视觉的问答任务中，RNN、GRU、LSTM 等结构常被用作文本编码/解码器，并能够取得优于 GAN 的效果[104−106]. 部分典型的方法在 3.3 节中进行了简要的介绍.

根据文本合成图片是较为新颖的跨模态生成问题，也是最近几年中对抗学习领域的研究热点. 文献 [107] 中的 GAN-CLS 模型是具有开创性的工作之一，该文利用细粒度的标签信息训练图像编码器和文本编码器，提高跨模态编码的相关性，同时采用流形差值优化等策略，生成与描述内容较为契合的图片样本. 文献 [108]、[109] 中提出的 StackGAN 系列模型则采用两阶段的生成方法生成具有更高像素的图片，先生成与文本描述相一致的包含轮廓、颜色等基本信息的低分辨率图片，在该图基础上进一步生成高像素、细粒度的图片样本，两阶段的生成过程均包含文本描述作为条件输入. 文献 [110] 则采用层次化的生成方法，首先根据文本描述生成对象的边界框，进而填充图像细节内容. 文献 [111] 中的 AttnGAN 进一步采用注意力机制（Attention）选取文本模态中的细节信息，经由多步的 Attention 和对抗学习，依次生成低像素、高像素的图片.

5　结论与展望

大数据背景下，多模态数据对同一对象的描述存在形式多源异构、内在语义一致的特点. 不同的模态形式分别描述对象在某一特定角度下的特征. 随着机器学习技术的发展，多模态学习领域的研究热点逐渐从经典的统计学习方法转移到深度学习方法. 对于视觉模态，CNN 逐渐成为最有效的特征表示方法；对于文本模态及相关、类似的序列预测任务，LSTM 也逐渐取代概率图模型，取得主导地位. 而对抗学习的兴起使得跨模态任务更为多样化.

对于多模态学习方法的研究可以从以下几个方向进一步展开：（1）对不同模态的样本进行更为精细化的特征表示，实现有效的跨模态匹配，利用模态互补构建更为完整的特征描述体系；（2）克服学习样本数量的限制，研究弱监督、无监督的多模态学习方法；针对该问题，对抗学习方法是可行的解决方案之一；（3）研究有效的模型融合框架，一方面是组合不同的算法以取得高质量的数据分析结果，另一方面是用模型融合指导对多模态数据的融合；（4）研究效果更为真实、性能更加稳定的跨模态生成方法；（5）应用背景从通用领域向垂直领域拓展，针对特定的应用场景（如医疗场景）实现可行的解决方案.

参　考　文　献

[1]　Rhianna K. Pedwell J A. Hardy S L, et al. Effective visual design and communication practices for research posters: Exemplars

based on the theory and practice of multimedia learning and rhetoric. *Biochem Mol Biol Educ*, 2017, 45(3): 249

[2] Welch K E. *Electric Rhetoric: Classical Rhetoric, Oralism, and A New Literacy*. Cambridge: MIT Press, 1999

[3] Berlin James A. Contemporary composition: the major pedagogical theories. *College English*, 1982, 44(8): 765

[4] O'Halloran K L. Interdependence, interaction and metaphor in multi-semiotic texts. *Social Semiotics*, 1999, 9(3): 317

[5] O'Halloran K L. Classroom discourse in mathematics: a multi-semiotic analysis. *Linguistics Educ*, 1998, 10(3): 359

[6] Morency L P, Baltrusaitis T. Tutorial on multimodal machine learning [R/OL]. *Language Technologies Institute* (2016-6-26) [2019-03-05]. https://www.cs.cmu.edu/~morency/MMML-Tutorial-ACL2017.pdf

[7] Plummer B A, Wang L W, Cervantes C M, et al. Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models // *Proceedings of IEEE International Conference on Computer Vision* (*ICCV* 2015). Santiago, 2015: 2641

[8] von Glasersfeld E, Pisani P P. The multistore parser for hierarchical syntactic structures. *Commun ACM*, 1970, 13(2): 74

[9] Jackson P. *Introduction to Expert Systems*. 3rd Ed. Boston: Addison Wesley, 1998

[10] Cortes C, Vapnik V. Support-vector networks. *Machine Learning*, 1995, 20(3): 273

[11] Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann Publishers, 1988

[12] Jelinek F. *Statistical Methods for Speech Recognition*. Cambridge: MIT Press, 1997

[13] McGurk H, MacDonald J. Hearing lips and seeing voices. *Nature*, 1976, 264(5588): 746

[14] Petajan E D. *Automatic Lipreading to Enhance Speech Recognition* (*Speech Reading*) [Dissertation]. University of Illinois at Urbana-Champaign, 1984

[15] Fels S S, Hinton G E. Glove-Talk: a neural network interface between a data-glove and a speech synthesizer. *IEEE Trans Neural Networks*, 1993, 4(1): 2

[16] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. *J Machine Learning Res*, 2014, 15: 1929

[17] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks // *Proceedings of the* 14*th International Conference on Artificial Intelligence and Statistics*. Fort Lauderdale, 2011: 315

[18] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition // *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR* 2016). Las Vegas, 2016: 770

[19] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks // *Advances in Neural Information Processing Systems*. Lake Tahoe, 2012: 1097

[20] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the impact of residual connections on learning // *Proceedings of Thirty-First AAAI Conference on Artificial Intelligence*. San Franciso, 2017: 4278

[21] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database // *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR* 2009). Miami, 2009: 248

[22] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition // *Proceedings of the* 2016 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, 2016: 260

[23] Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning // *Proceedings of the* 28*th International Conference on Machine Learning*. Bellevue, 2011: 689

[24] Baltrusaitis T, Ahuja C, Morency L P. Multimodal machine learning: a survey and taxonomy. *IEEE Trans Pattern Anal Machine Intelligence*, 2019, 41(2): 423

[25] Zhang L, Zhao Y, Zhu Z F, et al. Multi-view missing data completion. *IEEE Trans Knowledge Data Eng*, 2018, 30(7): 1296

[26] Wang L Q, Sun W C, Zhao Z C, et al. Modeling intra- and inter-pair correlation via heterogeneous high-order preserving for cross-modal retrieval. *Signal Process*, 2017, 131: 249

[27] Liu H P, Li F X, Xu X Y, et al. Multi-modal local receptive field extreme learning machine for object recognition. *Neurocomputing*, 2018, 277: 4

[28] Fu K, Jin J Q, Cui R P, et al. Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts. *IEEE Trans Pattern Anal Machine Intelligence*, 2017, 39(12): 2321

[29] Breiman L. Random forests. *Machine Learning*, 2001, 45(1): 5

[30] Breiman L, Friedman J H, Olshen R A, et al. *Classification and Regression Trees*. Florida: Chapman and Hall/CRC, 1998

[31] Breiman L. Statistical modeling: the two cultures. *Statist Sci*, 2001, 16(3): 199

[32] Vapnik V N, Cervonenkis, A. J. *Empirical Inference.* Berlin: Springer, 2013

[33] Schölkopf B, Smola A J, Bach F. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge: MIT Press, 2002

[34] Mercer J. Functions of positive and negative type, and their connection with the theory of integral equations. *Philos Trans R Soc London Ser A*, 1909, 209(441-458): 415

[35] Aronszajn N. Theory of reproducing kernels. *Trans Am Math Soc*, 1950, 68(3): 337

[36] Steinwart I, Hush D, Scovel C. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Trans Inf Theory*, 2006, 52(10): 4635

[37] Lodhi H, Saunders C, Shawe-Taylor J, et al. Text classification

using string kernels. *J Machine Learning Res*, 2002, 2(3): 419

[38] Wu J X, Rehg J M. Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel // 2009 *IEEE* 12*th International Conference on Computer Vision*. Kyoto, 2009: 630

[39] Lanckriet G R, Deng M, Cristianini N, et al. Kernel-based data fusion and its application to protein function prediction in yeast. // *Proceedings of Pacific Symposium on Biocomputing*. Hawaii, 2004: 300

[40] Lee W J, Verzakov S, Duin R P W. Kernel combination versus classifier combination // *Proceedings of International Workshop on Multiple Classifier Systems, MCS* 2007. Prague, 2007: 22

[41] Gönen M, Alpaydin E. Localized multiple kernel learning // *Proceedings of the* 25*th International Conference on Machine learning*. Helsinki, 2008: 352

[42] Jiang T J, Wang S Z, Wei R X. Support vector machine with composite kernels for time series prediction // *Proceedings of International Symposium on Neural Networks*. Nanjing, 2007: 350

[43] Hotelling H. Relations between 2 sets of variants. *Biometrika*, 1935, 28(3-4): 312

[44] Cooley W W, Lohnes P R. *Multivariate Procedures for the Behavioral Sciences*. New York: John Wiley & Sons, 1962

[45] Akaho S. A kernel method for canonical correlation analysis // *Proceedings of the International Meeting of the Psychometric Society* (*IMPS*2001). Osaka, 2001: 1

[46] Wang S, Lu J F, Gu X J, et al. Unsupervised discriminant canonical correlation analysis based on spectral clustering. *Neurocomputing*, 2016, 171: 425

[47] Hu H F. Multiview gait recognition based on patch distribution features and uncorrelated multilinear sparse local discriminant canonical correlation analysis. *IEEE Trans Circuits Syst Video Technol*, 2014, 24(4): 617

[48] Farquhar J D R, Hardoon D R, Meng H, et al. Two view learning: SVM-2K, theory and practice // *Proceedings of the* 18*th International Conference on Neural Information Processing*. Vancouver, 2005: 355.

[49] Ozerov A, Fevotte C. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Trans Audio Speech Language Process*, 2010, 18(3): 550

[50] Zhang J, Huan J. Inductive multi-task learning with multiple view data // *Proceedings of the* 18*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Beijing, 2012: 543

[51] Kong X N, Ng M K, Zhou Z H. Transductive multilabel learning via label set propagation. *IEEE Trans Knowledge Data Eng*, 2013, 25(3): 704

[52] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training // *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*. Madison, 1998: 92

[53] Collins M. Unsupervised models for named entity classification. // *Proceedings the* 1999 *of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. College Park, 1999: 100

[54] Brefeld U, Scheffer T. Co-EM support vector learning // *Proceedings of the Twenty-first International Conference on Machine Learning*. Banff, 2004: 16

[55] Muslea I, Minton S, Knoblock C A. Active + semi-supervised learning = robust multi-view learning // *Proceedings of the* 19*th International Conference on Machine Learning*. Sydney, 2002: 435

[56] Lécun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. *Proc IEEE*, 1998, 86(11): 2278

[57] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model // *Eleventh Annual Conference of the International Speech Communication Association*. Makuhari, 2010: 1045

[58] Hinton G E. Deep belief networks[J/OL]. *Scholarpedia* (2009-04-11) [2019-03-05]. http://www.scholarpedia.org/article/Deep_belief_networks

[59] Simonyan K, Zisserman A. Very Deep convolutional networks for large-scale image recognition. // *Proceedings of International Conference on Learning Representations* 2015. San Diego 2015: 1

[60] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas 2016: 779

[61] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Trans Pattern Anal Machine Intelligence*, 2017, 39(4): 640

[62] Kim Y. Convolutional neural networks for sentence classification // *Proceedings of* 2014 *Conference on Empirical Methods in Natural Language Processing*. Doha, 2014: 1746

[63] Shen Y L, He X D, Gao J F, et al. A latent semantic model with convolutional-pooling structure for information retrieval // *Proceedings of the* 23*rd ACM International Conference on Conference on Information and Knowledge Management*. Shanghai, 2014: 101

[64] Hu B T, Lu Z D, Li H, et al. Convolutional neural network architectures for matching natural language sentences // *Advances in Neural Information Processing Systems*. Montreal, 2014: 2042

[65] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model. *J Machine Learning Res*, 2003, 3(6): 1137

[66] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning // *Proceedings of the* 25*th International Conference on Machine Learning*. Helsinki, 2008: 160

[67] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J/OL]. *arXiv* (2013-09-07) [2019-03-05]. https://arxiv.org/pdf/1301.3781.pdf

[68] Graves A, Schmidhuber J. Frame-wise phoneme classification with bidirectional LSTM and other neural network architectures.

Neural Networks, 2005, 18(5-6): 602

[69] Liu P F, Qiu X P, Huang X J. Recurrent neural network for text classification with multi-task learning // *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. New York, 2016: 2873

[70] Sundermeyer M, Alkhouli T, Wuebker J, et al. Translation modeling with bidirectional recurrent neural networks // *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*). Stroudsburg, 2014: 14

[71] Cho K, van Merrienboer B, Bahdanau D, et al. On the properties of neural machine translation: encoder-decoder approaches. // *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Doha, 2014: 103

[72] Wollmer M, Eyben F, Graves A, et al. Bidirectional LSTM networks for context-sensitive keyword detection in a cognitive virtual agent framework. *Cognitive Comput*, 2010, 2(3): 180

[73] Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification // *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics*. Berlin, 2016: 207

[74] Zhang J, Man K F. Time series prediction using RNN in multi-dimension embedding phase space // *SMC*'98 *Conference Proceedings*. 1998 *IEEE International Conference on Systems, Man, and Cybernetics*. San Diego, 1998: 1868

[75] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks // 2013 *IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, 2013: 6645

[76] Karpathy A, Joulin A, Fei-Fei L F. Deep fragment embeddings for bidirectional image sentence mapping // *Advances in Neural Information Processing Systems*. Montreal, 2014: 1889

[77] Donahue J, Hendricks L A, Rohrbach M, et al. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Trans Pattern Anal Machine Intelligence*, 2014, 39(4): 677

[78] Kiros R, Salakhutdinov R, Zemel R. Unifying visual-semantic embeddings with multimodal neural language models. // *Deep Learning and Representation Learning Workshop: NIPS* 2014. Montreal, 2014: 1

[79] Mitchell M, Han X F, Dodge J, et al. Midge: Generating image descriptions from computer vision detections // *Proceedings of the* 13*th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, 2012: 747

[80] Ma L, Lu Z D, Li H. Learning to answer questions from image using convolutional neural network // *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. Phoenix, 2016: 3567

[81] Wan J, Wang D Y, Hoi S C H, et al. Deep learning for content-based image retrieval: a comprehensive study // *Proceedings of the* 22*nd ACM International Conference on Multimedia*. Orlando, 2014: 157

[82] Wöllmer M, Metallinou A, Eyben F, et al. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling // *Proceedings of the* 11*th Annual Conference of the International Speech Communication Association* (*INTERSPEECH*2010). Makuhari, 2010: 2362

[83] Su Y H, Fan K, Bach N, et al. Unsupervised multi-modal neural machine translation // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Seattle, 2019: 10482

[84] Wang X, Huang Q Y, Celikyilmaz A, et al. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Seattle, 2019: 6629

[85] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets // *Advances in Neural Information Processing Systems*. Monteral, 2014: 2672

[86] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[J/OL]. *arXiv* (2016-01-07) [2019-03-05]. https://arxiv.org/pdf/1511.06434.pdf

[87] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks // *Proceedings of the* 34*th International Conference on Machine Learning*. Sydney, 2017: 214

[88] Mirza M, Osindero S. Conditional generative adversarial nets[J/OL]. *arXiv* (2014-11-06) [2019-03-05]. https://arxiv.org/pdf/1411.1784.pdf

[89] Tzeng E, Hoffman J, Saenko K, et al. Adversarial discriminative domain adaptation // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Hawaii, 2017: 7167

[90] Liu M Y, Tuzel O. Coupled generative adversarial networks // *Advances in Neural Information Processing Systems*. Barcelona, 2016: 469

[91] Pei Z Y, Cao Z J, Long M S, et al. Multi-adversarial domain adaptation // *Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence*. New Orleans, 2018: 3934

[92] Cao Z J, Long M S, Wang J M, et al. Partial transfer learning with selective adversarial networks // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt lake city, 2018: 2724

[93] Xie S A, Zheng Z B, Chen L, et al. Learning semantic representations for unsupervised domain adaptation // *Proceedings of the* 35*th International Conference on Machine Learning*. Long Beach, 2018: 5423

[94] Denton E L, Chintala S, Szlam A, et al. Deep generative image models using a laplacian pyramid of adversarial networks // *Advances in Neural Information Processing Systems*. Montreal, 2015: 1486

[95] Zhang H, Goodfellow I, Metaxas D, et al. Self-attention generative adversarial networks [J/OL]. *arXiv* (2018-05-21)[2019-03-05]. https://arxiv.org/pdf/1805.08318.pdf

[96] Rush A M, Chopra S, Weston J. A neural attention model for

abstractive sentence summarization // *Proceedings of the* 2015 *Conference on Empirical Methods in Natural Language Processing*. Lisbon, 2015: 379

[97] Miyato T, Kataoka T, Koyama M, et al. Spectral normalization for generative adversarial networks[J/OL]. *arXiv* (2018-02-16) [2019-03-05]. https://arxiv.org/pdf/1802.05957.pdf

[98] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis[J/OL]. *arXiv* (2019-02-25) [2019-03-05]. https://arxiv.org/pdf/1809.11096.pdf

[99] Isola P, Zhu J Y, Zhou T H, et al. Image-to-image translation with conditional adversarial networks // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, 2017: 1125

[100] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks // *Proceedings of the IEEE International Conference on Computer Vision*. Venice, 2017: 2223

[101] Choi Y, Choi M, Kim M, et al. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, 2018: 8789

[102] Huang X, Liu M Y, Belongie S, et al. Multimodal unsupervised image-to-image translation // *Proceedings of the European Conference on Computer Vision* (*ECCV*). Munich, 2018: 172

[103] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation // *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Munich, 2015: 234

[104] Anderson P, He X D, 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*. Salt Lake City, 2018: 6077

[105] Chen X P, Ma L, Jiang W H, et al. Regularizing RNNs for caption generation by reconstructing the past with the present // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, 2018: 7995

[106] Chen F H, Ji R R, Sun X S, et al. Groupcap: Group-based image captioning with structured relevance and diversity constraints // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, 2018: 1345

[107] Reed S, Akata Z, Yan X, et al. Generative adversarial text to image synthesis. // *Proceedings of The* 33*rd International Conference on Machine Learning*. New York, 2016: 1060

[108] Zhang H, Xu T, Li H S, et al. Stackgan: text to photo-realistic image synthesis with stacked generative adversarial networks // *Proceedings of the IEEE International Conference on Computer Vision*. Venice, 2017: 5907

[109] Zhang H, Xu T, Li H S, et al. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Trans Pattern Anal Machine Intelligence*, 2019, 41(8): 1947

[110] Hong S, Yang D D, Choi J, et al. Inferring semantic layout for hierarchical text-to-image synthesis // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, 2018: 7986

[111] Xu T, Zhang P C, Huang Q Y, et al. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, 2018: 1316