信号的语义刻画与度量

石光明[1,2],高大化[1,2],杨旻曦[1],谢雪梅[1],董明皓[3],李雷达[1],于 凯[1]
(1. 西安电子科技大学人工智能学院,陕西西安 710071; 2. 鹏城实验室,广东深圳 518055;
3. 西安电子科技大学生命科学技术学院,陕西西安 710071)

摘 要: 相比于基于比特数据的信息处理及通信技术,人类通过语义处理和传递信息的方式,在面对智能体间传递处理海量信息这一问题时显得更为高效和自然. 然而由于目前缺乏关于语义度量和刻画的数学描述,涉及语义的应用无法兼顾可解释性和泛化性,无法发挥语义的高效自然的优势. 本文围绕语义的度量和刻画,首先依据信息科学和神经科学相关结论,讨论了语义的内涵,并指出语义具有模块化、多模态、层级化的特点;接着提出了一种多模态信号的语义刻画和度量的数学描述;然后为了验证所提信号语义的刻画和度量的可行性和有效性,在MNIST(Mixed National Institute of Standards and Technology database)手写数字识别和水声目标识别两个应用中进行了实验,获得比传统深度学习更好的性能;最后将语义用于视频编码,实现了远超传统方法的压缩比,展现了语义在通信领域的实用价值. 这为未来建立以语义为基础的新型信息处理与通信技术奠定了理论和实践基础.

关键词: 语义; 语义刻画; 语义度量; 语义基元; 语义计算; 语义识别
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2022)09-2068-11
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210238

Semantic Characterization and Measurement of Signals

SHI Guang-ming[1,2], GAO Da-hua[1,2], YANG Min-xi[1], XIE Xue-mei[1], DONG Ming-hao[3], LI Lei-da[1], YU Kai[1]
(1. *School of Artificial Intelligence*, *Xidian University*, *Xi'an*, *Shaanxi* 710071, *China*;
2. *Peng Cheng Laboratory*, *Shenzhen*, *Guangdong* 518055, *China*;
3. *School of Life Science and Technology*, *Xidian University*, *Xi'an*, *Shaanxi* 710071, *China*)

Abstract: Compared with the modern information processing and communication technology based on bits, the semantic-based way of processing and transmitting information used by humans is more efficient and natural in the face of the massive information that needs to be sent between agents. However, due to the lack of mathematical description of semantic measurement and characterization, semantics' applications cannot consider both interpretability and generalization. Therefore, they cannot give play to the advantages of efficient and natural semantics. This paper focuses on the measurement and characterization of semantics. Firstly, we discuss the connotation of semantics based on the relevant conclusions of information science and neuroscience, and concludes that semantics has the characteristics of modularity, multi-mode, and hierarchy. Then, a semantic description of multimodal signals and a mathematical description of their measurement are proposed. Next, to verify the feasibility and effectiveness of the characterization and measurement of the proposed signal semantics, experiments are carried out in two applications: MNIST (Mixed National Institute of Standards and Technology database) handwritten digital recognition and underwater acoustic target recognition, and the results are better than those of the traditional deep learning. Finally, the semantics is used for video coding, and a compression ratio far exceeding that of traditional methods is achieved. This lays a theoretical and practical foundation for establishing new information processing and communication technology based on semantics.

Key words: semantic; semantic characterization; semantic measurement; semantic primitive; semantic computing; semantic recognition

收稿日期:2021-02-09;修回日期:2022-02-02;责任编辑:梅志强
基金项目:国家重点研发计划课题(No.2019YFA0706604); 国家自然科学基金(No.61871304, No.61976169, No.61875157); 鹏城实验室重大攻关项目(No.PCL2021A12)

1 引言

凭借着丰富的感官和发达的大脑,人类能对多模态信号进行处理分析,从而形成生存优势,并在沟通、交流和通信基础上形成了现代社会,以比特数据为基础的信息处理与通信技术更是将人类带入了信息社会. 但随着人工智能技术的发展,大量智能体(本文中的智能体是包括人在内的智慧性系统,如会决策的机器人)之间(如人-机、机-机之间)产生了巨大而频繁的信息传递需求,这给以比特数据为基础的信息处理与通信技术带来了巨大挑战. 以视频通信为例,使用目前最新的视频编码方法 H265 传输单路 4K/30 帧的视频就需要 40 Mbps 的带宽,使用 5G 的终端设备仍然无法同时传输几十路视频以满足大型线上会议场景下的交互需求. 因此,探寻并建立一套突破比特数据局限的新型信息处理技术具有时代意义. 其中,如何找到一种更高效、更自然的信息表达方法是开展新型信息处理技术研究的基础.

通常,人类处理和传递信息时并不是以精准传递数据信号为主,而是以传递语义(语言中包含的意思)为首要目的,这对智能体而言相当高效和自然. 事实上,从远古至今,人类近距离交互一直使用的是声波. 尽管声波带宽有限,但相比远古人类之间交流、处理的信息量,现代人类之间交流的信息量多了很多,而交流所用的声波带宽资源并没有增加,增加的是人类对语义的表达与理解能力. 近年来,国内外学者已经对语义在信息处理和通信中的应用展开了大量的研究. 在人工智能领域,直接从信号中提取所需信息的过程被称为语义识别. 对以语言文字为代表的离散符号所承载的语义的研究主要分为两类:(1)借鉴自然语言的语法规则,构建描述语义的离散符号系统. 该思路下的早期工作是由文献[1]首先提出的框架逻辑. 随后 Baader 等人提出了更为完善且更适合计算机处理的描述逻辑[2]. 近年来,随着数据和算力的发展,基于知识图谱[3]、事理图谱[4]等灵活的离散图数据结构的语义存储方法和基于图神经网络[5,6]的数据挖掘方法被广泛应用于学界和业界;(2)基于语料数据,构建反映语言规则的语言模型. 在早期,学界主要使用条件随机场[7]、贝叶斯网络[8]等概率模型对语言进行建模. 随着深度学习的发展,从早期的句子顺序建模[9,10]到词语嵌入式表示[11],到基于注意力的大型跨任务模型[12,13],越来越多的基于深度模型的语言模型被提出,并逐渐成为离散语义研究的主流方法. 而对以图像为代表的连续信号所承载的语义研究主要分为三类:(1)建立信号和语义概念的直接映射. Lowe 提出一种设计语义不变性的模板,并通过模板匹配进行图像识别方法[14]. Pedro 等人在文献[15]中提出利用深度模型识别出的图像中语义概念明确的部件,进而综合得到整体目标的检测结果. Koh 等人在文献[16]中提出一种先使用深度模型预测图像中包含的语义概念,再由分类器做出判断的可解释图像分类方法. (2)融合离散符号表示的语义先验. Lu 等人提出了一种结合人类语言先验的视觉关系检测方法[17]. 该方法用深度神经网络将特征提取网络得到的图像特征和词向量表示的人类语言先验知识融合,实现了对图像中物体之间的关系的检测. Wang 等人通过语义嵌入和知识图谱实现了零样本识别[18]. 该方法利用图卷积神经网络,将知识图谱表示的图像类别知识映射到语义空间. 然后融合图像特征和语义向量,实现对训练时没有见过的类别的图像进行识别. (3)分析数据在表征空间中的分布. Caron 等人在文献[19]中提出了一种通过在表征空间上进行数据聚类实现无监督图像分类的方法. Li 等人在文献[20]中基于数据在表征空间上的聚类中心,为每个类别构建多个原型,再将待测数据和原型匹配得到分类结果.

然而,不论是传统机器学习方法还是深度学习方法,由于语义似乎只可意会不可言传,虽然学者们在研究中大量涉及了语义及其应用,但并没有从物理和数学等方面对语义进行刻画、表达、度量和计算,导致其无法兼顾可解释性和泛化性,这是造成当今的信息处理与通信技术是非语义模式的重要原因之一. 在本文中,我们提出了一种多模态信号的语义刻画和度量的数学描述. 我们首先依据信息科学和神经科学相关结论,给出了具有模块化、多模态、层级化特点的语义刻画方法,包括基于语义基元的表达方法和语义计算模型;接着,在语义刻画的基础上,给出了语义空间、语义相似度、语义距离和语义度量的数学描述;最后,为了验证所提信号语义的刻画和度量的可行性和有效性,我们在 MNIST (Mixed National Institute of Standards and Technology database) 手写数字识别和水声目标识别两个应用中进行了实验,获得了比传统深度学习更好的性能.

2 语义的内涵

“语义(semantic)”一词在人工智能领域被广为使用,被用于指代信号中的可理解含义的表征,如语义分割[21]、语义分析[22]、语义理解[23],甚至语义计算[24]. 然而,此类表征都是高维张量或者文本,存在着可解释性差,泛化能力差的问题,制约了通用模型的产生. 目前为止,对信号中语义的直观且通用的数学描述仍然是一个极具挑战性的难题,还没有有效的解决思路. 其原因之一是语义基本内涵不易定义从而难以度量;原因之二是人们对语义的产生机理和过程不了解. 而信息科学和神经科学的一些工作对语义基本内涵和语义产

生的机理过程的探寻有着重要借鉴价值. 因此本节将分别介绍信息科学和神经科学对语义的相关研究,并以此总结出语义的特点,为第3节中的信号语义的刻画和度量的数学描述奠定基础.

2.1 信息科学中的语义

信息论的创始人Shannon在其奠定现代信息论基础的论文文献[25]中率先提到语义层面的信息交互问题. Shannon在其之后出版的《通信的数学理论》[26]一书中指出,语义问题关心的是收信者对信息的理解是否与发信者想表达的含义一致或接近. 并将通信问题归为三个层面:(1)技术问题:通信符号如何准确地进行传输?(2)语义问题:传输的符号如何精确地传达含义?(3)效用问题:收到的含义如何以期望的方式有效地影响行为? 不同于符号层面只关注经过符号编码调制的信号载波是否正确传输,语义层面的信息交互是需要交互双方能够理解信号中的内容或含义,从而提取其中的信息. 语义层面的信息也不再是由符号的熵简单定义,而是通过接受信号前后的语义差异性定义,即,先从对方的信号中感知出语义,然后与自己的已知语义对比,如果存在差异,这个差异就是信息. 在Shannon之后的学者在语义信息理论框架下的语义刻画与度量展开讨论,并率先开展了基于语义而非比特数据的通信方法的探究. Guler等人[27]提出了一种语义误差,作为语义信息准确性的衡量标准,用于计算交互双方语义的偏差距离. Bao等人[28]进一步指出在进行语义信息交互过程中,交互双方需要具有共有知识储备,才能进行顺畅的语义交流. Basu等人[29]提出了语义容量的概念,并指出语义容量等于信息源的平均语义熵,确立了语义压缩的下界. 此外Willems等人[30]研究了语义编码,使用语义相似性指导机器学习算法的优化,实现了数据间关系的更紧凑表示. 目前信息科学对语义的研究主要是以信息论中关于不确定性的论述为基础,将香农信息论对比特的理论迁移至语义,形成了以语义符号为基础的语义信息论. 由此,我们认为信号语义具有模块性. 在节3中,我们将使用有限个预定义了语义的信号作为语义符号构成信号语义刻画的基础.

2.2 神经科学中的语义

Hubel和Wiesel[31]发现大脑视觉皮层中存在相同图像特征选择性和相同感受野位置的众多神经细胞,以垂直于大脑表面的方式排列成柱状结构,称为神经元功能柱(functional column). 同一个功能柱内所有的神经细胞都编码了相同的视觉信息,它们只对某一种视觉特征发生反应,从而形成该种视觉特征的基本单位. 类似神经元功能柱的模块化结构在大脑中有着不同尺度的体现. 以视觉神经信号传输过程为例(如图1).

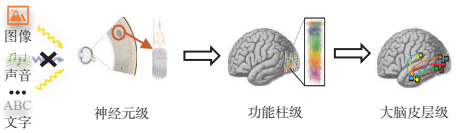


图1 视觉语义形成过程示意图

视觉语义的产生源自视网膜,并终止于大脑[32],其中语义以稀疏的皮质区域和连接模式的形式呈现. 该过程分层次地涉及基于低级感官的处理,并且可以分为三个级别[33]. 处理的初始级别发生在视网膜上,该过程将视网膜上的光的模式传输为编码的电信号,然后传输到大脑中. 中级处理通过将视觉世界分为属于对象的轮廓和表面并将这些元素与背景隔离,从而参与视觉图元的识别. 在此阶段,视觉信息处理依赖于功能柱[34],它表示一个单元,其中包含大脑皮层中一组神经元中任何给定的感受野的完整神经元集合[35]. 柱状组织假说是目前最广泛用于解释信息皮质处理的方法[36]. 上层涉及对象识别,其核心是语义的分类,表现为连续语义空间向大脑皮层表面区域的一种映射[32]. 因此,一个语义由大脑中的多个子系统表示,这一现象可以被称为大脑区域网络[33,37]. Huth和Gallant等人使用了功能性磁共振成像(functional Magnetic Resonance Imaging, fMRI),利用有声读物故事以及来自计算语言学的机器学习算法,以探讨如何将语义映射到相应激活的大脑皮质区域[38]. 由此产生的图谱表明,语义系统是复杂的层级表现模式,在主体内相对稳定,并且在个体之间分布基本一致[38]. 由此,我们认为信号语义具有模块化、多模态、层级化的特点. 在第3节中,我们将使用语义计算模拟神经信号随着尺度提升而逐渐抽象的过程,对基础语义进行抽象、调整、拓展等延伸操作,使其具有模块化、多模态、层级化的特点.

3 语义的刻画和度量

3.1 语义的刻画

鉴于信号语义具有模块化、多模态、层级化的特点,我们将从基于基元的基础语义定义和基于语义计算的语义延伸两个方面对语义进行刻画.

3.1.1 基础语义

设δ表示一个基础的语义,如语义“黑色”. 那么语义“黑色”存在不同模态下的信号语义基元,如声音模态$f_\delta(t)$为“Hēi”这一声音信号,图色模态$F_\delta(x,y)$为图片“■”,文字模态s_δ为中文词“黑”. 此时,可定义语义$\delta:=\{f_\delta(t),F_\delta(x,y),s_\delta,\cdots\}$,其中$f_\delta,F_\delta,s_\delta$等分别对应一维、二维、离散符号等模态空间中的某个由基元表示的语义特征函数. δ集合内的元素在语义层面是相互等价的,

即f_δ的语义是δ,F_δ的语义是δ,s_δ的语义也是δ,换句话说它们之间可以相互表达或标注. 例如,“黑”(文本),“Hēi”(声音),“■”(图色)这三者是不同的信号模态,它们之间建立了语义等价概念. 看到“黑”文字符号,就能发音“Hēi”,脑中也能产生黑色影像. 对应在英语语系中,同样的语义可以用“Black”(文本)、“[blæk]”(声音)、“■”(图色)来约定相互等价的语义.

更一般地,如果语义有N个模态,则可进一步定义其语义为:

$$\delta=\{F_{\Theta_1},F_{\Theta_2},\cdots,F_{\Theta_N}\} \tag{1}$$

其中在模态Θ_i上定义的特征函数的集合为$F_{\Theta_i}=\{F_{\Theta_i}^{(1)},F_{\Theta_i}^{(2)},\cdots,F_{\Theta_i}^{(n_i)}\}$.

选择合适的语义基元对于定义基础语义而言十分关键. 在实际应用中,可根据具体应用场景,通过专家手工设计或数据学习得到. 当人类知识可靠时,通过专家手工设计基元. 比如对语义“人的躯体”,可以根据人类知识手工设计头、手、腿、躯干等人体部件作为语义基元. 当人类知识不可靠时,则需要收集足够的表达某个语义的数据,利用机器学习算法获取基元. 例如建立语义“木质纹理”,则可收集足够的木质材料的图像数据,然后利用主成分分析等算法提取基元.

3.1.2　语义计算

众所周知,智能体会对语义进行扩充、融合、提升和凝练,为此我们定义了满足这些操作的语义计算方法. 人在学习的过程中,会不断地对现有的语义进行扩充,从计算角度看,这种语义内涵的扩充称为语义加法. 语义加法有两种情况:一是语义中拓展若干新的特征函数,即语义与特征函数集合相加;二是两个语义融合为新的、更全面的语义,即语义与语义相加.

3.1.2.1　语义加法

(1) 语义与特征函数集合相加

语义与特征函数集合相加主要用于对某一语义具有属性的细化或扩充. 用$S_1=\{F_{1\Theta_1},F_{1\Theta_2},\cdots,F_{1\Theta_N}\}$表示一个语义,用$F=\{F_{\Theta_1},F_{\Theta_2},\cdots,F_{\Theta_N}\}$表示一个特征函数的集合. 那么语义和特征函数集合的加法可以定义为:

$$\begin{aligned}S_1'&=S_1+F\\&=\{F_{1\Theta_1}\cup F_{\Theta_1},F_{1\Theta_2}\cup F_{\Theta_2},\cdots,F_{1\Theta_N}\cup F_{\Theta_N}\}\end{aligned} \tag{2}$$

其中,$F_{1\Theta_1}\cup F_{\Theta_1}$表示两个集合之间的并集操作. 我们通过上式将语义和特征函数的集合表示为所有子空间内集合的并集. 因此,这种加法是不会产生新的特征函数的. 需要指出的是,特征函数集合$F=\{F_{\Theta_1},F_{\Theta_2},\cdots,F_{\Theta_N}\}$和语义的定义形式相同,因此该操作也可以用来实现两个语义概念的合并,而不会产生新的特征函数.

通过语义和特征函数集合之间的加法可以描述人在学习外语时语义的变化过程. 对于3.1.1小节中定义的语义“黑”可以表示为$S=\{F_{\Theta_s},F_{\Theta_v},F_{\Theta_c}\}$. 其中$F_{\Theta_s},F_{\Theta_v},F_{\Theta_c}$分别表示声音、图像、符号语义模态的特征函数集合. 我们假设智能体只懂汉语,即声音模态的特征函数集合$F_{\Theta_s}=\{F_{\text{Chn}}\}$只记录了汉语读音“Hēi”. 而通过学习英文中单词black的发音“[blæk]”$F=\{F_{\Theta_s}|F_{\Theta_s}=\{F_{\text{Eng}}\}\}$之后,语义“黑”就可以通过式(2)所述的语义和特征函数集合之间的加法进行扩展,得到$S=\{F_{\Theta_s},F_{\Theta_v},F_{\Theta_c}\}$,　$F_{\Theta_s}=\{F_{\text{Chn}},F_{\text{Eng}}\}$.

(2) 语义与语义相加

语义和语义相加用于融合语义特征,产生新的特征函数,从而形成新的语义,其定义为:

$$\begin{aligned}S'=S_1\oplus S_2=\{&\text{fusion}(F_{1\Theta_1},F_{2\Theta_1}|\Theta_1),\\&\text{fusion}(F_{1\Theta_2},F_{2\Theta_2}|\Theta_2),\cdots,\text{fusion}(F_{1\Theta_N},F_{2\Theta_N}|\Theta_N)\}\end{aligned} \tag{3}$$

其中$\text{fusion}(F_{1\Theta_1},F_{2\Theta_1}|\Theta_1)$表示同一个模态内的两个特征函数集合之间进行特征融合:

$$\begin{aligned}\text{fusion}(F_{1\Theta_k},F_{2\Theta_k}|\Theta_k)=\{&f(F_{1\Theta_k}^{(a)},F_{2\Theta_k}^{(b)})|\forall F_{1\Theta_k}^{(a)}\in F_{1\Theta_k},\\&F_{2\Theta_k}^{(b)}\in F_{2\Theta_k}\}\end{aligned} \tag{4}$$

其中$\text{fusion}(F_{1\Theta_k}^{(a)},F_{2\Theta_k}^{(b)})$表示两个特征函数之间的特征融合. 我们使用上式将集合之间的特征融合表示为两个集合中所有特征函数两两进行特征融合,从而产生一个新的特征函数的集合. 特征函数之间的特征融合的方式有很多,甚至可以使用神经网络实现复杂的非线性特征融合,这里我们定义为最简单最通用的方法:

$$\text{fusion}(F_{1\Theta_k}^{(a)},F_{2\Theta_k}^{(b)})=(F_{1\Theta_k}^{(a)}+F_{2\Theta_k}^{(b)})/2 \tag{5}$$

通过语义间加法可以描述光线或者颜料的颜色组合过程. 若仿照3.1.1小节分别定义光的三原色红、绿、蓝的语义. 则可以根据式(5)对三原色的语义中的图像模态的特征函数进行融合,得到新的颜色的图像特征函数. 而若将式(5)中的特征融合函数设定为三项加权求和,就可以得到任意颜色.

由于语义与语义相加需要在两个集合间组合计算,因此计算复杂度较高. 在实际应用中,如果连续使用语义与语义相加,将会导致语义特征函数集合的规模快速扩大,进而引发组合爆炸问题. 语义与语义相加产生的大量同一模态的语义特征函数构成了一个语义子空间的一组过完备原子,其中一些语义特征函数可以由其他特征函数近似线性表示,即存在冗余. 因此,在完成语义与语义相加的操作后,可以采用聚类等方式对新的特征函数集合进行去冗余,从而缩减语义特征函数集合的规模,从而避免组合爆炸.

3.1.2.2　语义乘法

(1) 语义的数乘

人在学习的过程中,会根据新总结的经验在已有概念的基础上进行调整. 在语义计算的框架中,我们定

义语义的数乘来描述这种语义概念的权重调整：

$$S'=c\cdot S=\{c_{11}\cdot F_{11},c_{12}\cdot F_{12},\cdots,c_{Nn}\cdot F_{Nn}\} \tag{6}$$

其中，$c=\{c_{11},c_{12},\cdots,c_{Nn}\}$ 表示每个特征函数对应的数乘的常数，F_{ki} 和 c_{ki} 表示模态 Θ_k 下的第 i 个特征函数及其对应的常数.

通过语义的数乘可以描述语言环境发生变化时，人对语义做出的调整. 一个懂得中英双语的人在国内生活时间长了之后，对中文发音更加敏感. 当语言环境再次改变的时候，则根据式(6)再做出相应的调整.

(2) 语义的直积

联想、抽象能力是人类智慧重要的组成部分. 面对不同的概念，将其关联起来，组合成更高一级的语义，便是联想、抽象能力的本质. 通过之前对语义的定义，可以把联想抽象能力理解为从低级语义生成更高一级语义的过程. 语义本质上是特征函数的集合，因此使用直积来表示这种过程：

$$S'=S_1\otimes_R S_2=\{F_{1\Theta_1},F_{1\Theta_2},\cdots,F_{1\Theta_N}\}\otimes_R\{F_{2\Theta_1},F_{2\Theta_2},\cdots,F_{2\Theta_N}\} \tag{7}$$

其中，最右项表示规则 R 指导下的两个集合之间的笛卡尔乘积(Cartesian product). 此处引入规则 R 是为了减少没有意义的模态之间的组合，从而减少笛卡尔乘积运算后集合的规模，以避免发生组合爆炸问题. 比如，我们可以定义声音模态 Θ_1 和图像模态 Θ_2 进行组合，其他的模态将不参与组合，此时笛卡尔乘积的结果便只有 $\{(F_{1\Theta_1},F_{2\Theta_1}),(F_{1\Theta_1},F_{2\Theta_2}),(F_{1\Theta_2},F_{2\Theta_1}),(F_{1\Theta_2},F_{2\Theta_2})\}$ 四种组合结果. 需要注意的是，此处的组合是有向组合，即一般情况下 $(F_{1\Theta_1},F_{2\Theta_1})\neq(F_{2\Theta_1},F_{1\Theta_1})$，则：

$$S_1\otimes_R S_2\neq S_2\otimes_R S_1 \tag{8}$$

这样的规定可以表示一定的因果、先后次序，更加丰富语义的表达能力. 需要进一步指出的是，模态组合的结果是生成新的模态，例如 $\{(F_{1\Theta_1},F_{2\Theta_1}),(F_{1\Theta_1},F_{2\Theta_2}),(F_{1\Theta_2},F_{2\Theta_1}),(F_{1\Theta_2},F_{2\Theta_2})\}$ 就生成了四种新的高维模态，即 $\{(\Theta_1,\Theta_1),(\Theta_1,\Theta_2),(\Theta_2,\Theta_1),(\Theta_2,\Theta_2)\}$. 在新的高维模态下，将同样按照笛卡尔乘积的形式生成高维特征函数：

$$F_{(\Theta_i,\Theta_j)}=F_{1\Theta_i}\otimes F_{2\Theta_j}=\{(F_{1\Theta_i}^{(a)},F_{2\Theta_j}^{(b)})|\forall F_{1\Theta_i}^{(a)}\in F_{1\Theta_i},F_{2\Theta_j}^{(b)}\in F_{2\Theta_j}\} \tag{9}$$

用语义直积可以描述我们根据所学声母韵母组合成汉语拼音的过程. 例如，我们在学习汉语拼音时，会区分声母和韵母，分开学习，然后再将声母韵母按一定规则组合起来，就能形成所有的汉语拼音. 从语义角度理解，我们可以首先定义所有声母和韵母的语义分别为 S_1,S_2. 然后将式(7)中的规则定义为仅限声音模态直积，对 S_1,S_2 进行组合. 接着依照式(8)，根据拼音的组合规则，只保留声母在前韵母在后的组合结果. 最后依照式(9)，将声母和韵母的特征函数直积成拼音的特征函数.

由于语义与语义相加需要在两个集合间组合计算，因此计算复杂度较高. 在实际应用中，需要去除新产生的高维特征函数集合中的冗余. 由于新产生的高维特征函数是由两个属于不同语义子空间的特征函数组合而成，不便直接使用聚类等方法去除冗余. 因此，可以先通过主成分分析等方法先对高维特征函数进行降维，再在低维特征空间上通过聚类去除冗余.

3.2 语义的度量

在本小节中，我们将以语义为元素，给出了语义空间的数学描述，并基于语义空间提出了语义相似度、语义距离和语义度量的数学描述.

3.2.1 语义空间

若 $\delta(i)$ 为一个具体的语义，把所有的语义用集合 $S:=\{\delta(i)\}$ 表示，就构成了语义空间，其中每一种模态特征信号的集合为对应的模态子空间. 同一个模态特征信号组成一个语义子空间. 语义空间是由多个不同模态信号的子空间组成. 在任何类型模态信号中，定义那些不再细分的基本特征函数为语义基元. 语义基元能够支撑这类模态信号的语义子空间，当然这些语义基元可以按不同的时空关系组合或融合再次形成高层次含义或概念的语义符号，这种融合可以逐级提升[39]. 一个语义的特征函数是由相应的子空间的多个语义基元结构化组合而成. 例如，在视觉空间，一些基本的点、线、面、曲线、圆、三角形、四边形等是视觉空间的语义基元，组成视觉语义子空间的基函数. 圆、三角和线可再次融合形成某一类物体的语义符号，不同复杂程度的基函数代表不同层级的语义特征函数. 这个语义也可能用听觉模态子空间的基函数，例如拼音的声母和韵母的发音. 整个语义空间由相互表示等价含义的多个模态信号子空间组成，如图2所示.

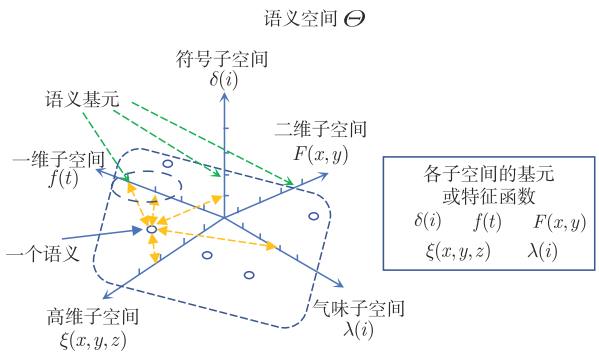


图2 语义空间示意图

3.2.2 语义特征谱

在某个模态对应的语义子空间中,假设人工设计或学习到的特征函数用一个点表示,则在此子空间的众多特征函数就是一个点阵图,它们呈现有序结构. 一维特征函数点阵呈现的是时间先后的一维序列结构,可以用矢量表示;二维特征函数点阵呈现二维图像结构,可以用矩阵表示;高维特征函数点阵呈现高维图结构. 这些结构都称为语义特征序.

如在一维声波语义子空间,预先定义一维基元特征函数集合$\{f_1(t),f_2(t),f_3(t),\cdots,f_n(t)\}$,集合中的每一个元素为一个基元,它们是一个个特定连续的基本声波. 某个语义在声波子空间可以由一维声波信号s_{v}表示:

$$s_{\mathrm{v}}=\sum_{i=1}^{n}\omega_i f_i(t-t_i) \tag{10}$$

其中,ω_i是对应特征的强度系数,即为s_{v}的特征谱;t_i是对应特征的时延,即为s_{v}的特征序. 又如在二维图像模态的语义子空间,其语义基元的特征函数集合为$\{F_1(x,y),F_2(x,y),F_3(x,y),\cdots,F_n(x,y)\}$,每个基元的特征函数都对应一个特定的二维图像. 某个语义可以在图像子空间用二维图像信号s_{p}表示:

$$s_{\mathrm{p}}=\sum_{i=1}^{n}\omega_i F_i(x-x_i,y-y_i) \tag{11}$$

其中,ω_i是对应特征的强度系数,即为s_{p}的特征谱;(x_i,y_i)是对应特征的空间结构点集,即为s_{p}的特征序.

3.2.3 语义之间的距离

前面已经给出了语义子空间的概念,并指出语义是由不同语义子空间中定义的若干特征函数描述的. 在同一个子空间中的特征函数很容易定义距离;而属于不同子空间的特征函数由于物理意义不同,无法定义距离. 因此语义之间的距离可以定义为所有子空间内特征函数距离的集合. 如果它们之间不存在相同的语义子空间,则表明这两者语义距离无穷大;如果存在部分相同的语义子空间,则它们之间的语义距离定义为语义子空间距离的集合. 对于语义子空间集合$\Theta=\{\Theta_1,\Theta_2,\cdots,\Theta_N\}$,其中在子空间$\Theta_i$上定义的特征函数的集合为$F_{\Theta_i}=\{F_{\Theta_i}^{(1)},F_{\Theta_i}^{(2)},\cdots,F_{\Theta_i}^{(n_i)}\}$. 于是任意两个语义可以记为$S_1=\{F_{1\Theta_1},F_{1\Theta_2},\cdots,F_{1\Theta_N}\}$和语义$S_2=\{F_{2\Theta_1},F_{2\Theta_2},\cdots,F_{2\Theta_N}\}$,其中$F_{1\Theta_i},F_{2\Theta_i}\subseteq F_{\Theta_i},\quad i=1,2,\cdots,N.$由此,语义之间的距离定义为一个集合:

$$D_s=\{d_1,d_2,\cdots,d_N|d_i=\mathrm{dis}(F_{1\Theta_i},F_{2\Theta_i})\} \tag{12}$$

其中每个元素表示两个语义在语义子空间Θ_i下的特征函数集合之间的距离. 需要指出的是,当两个语义特征函数集合是同一模态时,他们之间的距离便具有意义,而与其所属的语义无关. 距离可定义为:

$$\mathrm{dis}(F_{1\Theta_i},F_{2\Theta_i})=\frac{1}{N_a\cdot N_b}\cdot\sum_{F_{\Theta_i}^{(a)}\in F_{1\Theta_i},F_{\Theta_i}^{(b)}\in F_{2\Theta_i}}\mathrm{dis}(F_{\Theta_i}^{(a)},F_{\Theta_i}^{(b)}),\quad F_{1\Theta_i},F_{2\Theta_i}\neq\varnothing \tag{13}$$

其中,N_a和N_b分别为$F_{1\Theta_i}$和$F_{2\Theta_i}$中特征函数的个数,两个特征函数之间的距离度量dis(·)如下:

$$\mathrm{dis}(a,b|\Theta_i)=\frac{\sqrt[p]{\sum_{k=1}^{\dim(\Theta_i)}(F_{\Theta_i}^{(a)}(k)-F_{\Theta_i}^{(b)}(k))^p}}{\dim(\Theta_i)} \tag{14}$$

其中,$\dim(\Theta_i)$为语义子空间Θ_i的维数,p为闵可夫斯基距离(Minkowski distance)的参数,其取值应根据具体应用而定.

需要说明的是,当任意一个集合为空集时,距离定义为无穷大:

$$\mathrm{dis}(F_{1\Theta_i},F_{2\Theta_i})=\infty,\ F_{1\Theta_i}\ \text{or}\ F_{2\Theta_i}=\varnothing \tag{15}$$

这里,语义的距离是一个集合$D_s=\{d_1,d_2,\cdots,d_N|d_i=\mathrm{dis}(F_{1\Theta_i},F_{2\Theta_i})\}$.

3.2.4 语义之间的相似度

与语义之间的距离定义类似,可以定义语义之间的相似度集合:

$$D_s=\{d_1,d_2,\cdots,d_N|d_i=\mathrm{sim}(F_{1\Theta_i},F_{2\Theta_i})\} \tag{16}$$

其中每个元素表示两个语义在语义子空间Θ_i下的特征函数集合之间的相似度. 要定义特征函数集合之间的相似度,就需要先明确两个特征函数之间的相似度$\mathrm{sim}(a,b|\Theta_i)=\mathrm{sim}(F_{\Theta_i}^{(a)},F_{\Theta_i}^{(b)})$. 特征函数之间的相似度可以有很多定义方式,例如:采用人工标注的方式,对所有特征函数两两之间的相似度进行一个预设;或者采用神经网络模型预测相似度. 此处,我们给出最通用的定义方法,采用特征函数之间的闵可夫斯基距离的倒数作为特征函数之间的相似度:

$$\mathrm{sim}(a,b|\Theta_i)=\frac{\dim(\Theta_i)}{\sqrt[p]{\sum_{k=1}^{\dim(\Theta_i)}(F_{\Theta_i}^{(a)}(k)-F_{\Theta_i}^{(b)}(k))^2}+\varepsilon} \tag{17}$$

其中,$\dim(\Theta_i)$为语义子空间Θ_i的维数,p为闵可夫斯基距离(Minkowski distance)的参数,其取值应根据具体应用而定;$F_{\Theta_i}^{(\cdot)}(k)$表示特征函数的第$k$维;$\varepsilon$为一个很小的数,避免除法错误. 基于特征函数之间的相似度,我们给出集合之间的相似度的计算公式:

$$\mathrm{sim}(F_{1\Theta_i},F_{2\Theta_i})=\frac{\sum_{F_{\Theta_i}^{(a)}\in F_{1\Theta_i},F_{\Theta_i}^{(b)}\in F_{2\Theta_i}}\mathrm{sim}(F_{\Theta_i}^{(a)},F_{\Theta_i}^{(b)})}{|F_{1\Theta_i}|\cdot|F_{2\Theta_i}|},\quad F_{1\Theta_i},F_{2\Theta_i}\neq\varnothing \tag{18}$$

其中,$|\cdot|$表示集合的长度. 通过该式,我们将集合间的相似度定义为两个集合中所有特征函数两两之间相

似度的平均值. 同样,当任意一个集合为空集时,相似度定义为0:

$$\mathrm{sim}(F_{1\Theta_i},F_{2\Theta_i})=0,\ F_{1\Theta_i}\ \mathrm{or}\ F_{2\Theta_i}=\varnothing \tag{19}$$

3.2.5 信号的语义度量

人在理解一种新事物的时候,往往使用我们熟知的各个属性对其进行衡量. 有了前述语义定义之后,我们可以把这个过程看作是求一个信号在各个语义上的投影,从而实现信号的语义度量. 设语义 $S=\{F_{\Theta_1},F_{\Theta_2},\cdots,F_{\Theta_N}\}$,其中在子空间 Θ_i 特征函数的集合为 $F_{\Theta_i}=\{F_{\Theta_i}^{(1)},F_{\Theta_i}^{(2)},\cdots,F_{\Theta_i}^{(n_i)}\}$,对应的特征谱为 $\{\omega_{\Theta_i}^{(1)},\omega_{\Theta_i}^{(2)},\cdots,\omega_{\Theta_i}^{(n_i)}\}$. 定义在该子空间的度量 $\left\|S_{\Theta_i}\right\|=\sqrt{\sum_{j=1}^{n_i}(\omega_{\Theta_i}^{(j)})^2}$,则其在整个语义空间的度量定义为:

$$\|S\|=\sqrt{\sum_{j=1}^{n_i}(S_{\Theta_i})^2} \tag{20}$$

4 基于语义度量和计算的应用实例

为了说明语义这一核心概念的有效性和可行性,本节基于本文提出的语义度量和计算方法,分别在MNIST手写数字图像分类和水声目标识别任务上进行了仿真验证.

4.1 MNIST手写数字识别

目前,大多数机器学习方法都是基于数据驱动的,需要使用大量样本数据,消耗大量算力对模型进行训练后,才能用于图像分类与识别. 针对此问题,本文提出了:(1)基于人类知识和语义计算的语义符号库构建方法;(2)基于语义度量的识别网络(总体框图如图3所示). 其主要思路是首先利用人类知识从样本中抽取语义,然后通过语义计算构建语义符号库;在识别过程中,结合语义符号库,通过识别网络对待识别图像进行语义度量,完成识别过程.

该方法的优势在于只需要使用少量图像样本构建语义符号库,不需要对网络进行训练或仅需少量训练,即可用于图像识别. 在MNIST数据集上的对比实验结果表明本文方法远优于数据驱动的卷积神经网络[40];并且训练数据量越小,效果差异越大.

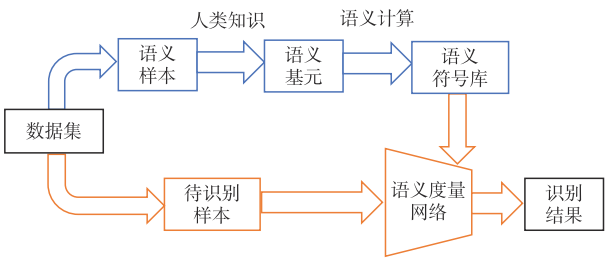


图3 基于语义度量和计算的图像识别框架

4.1.1 语义符号库的构建

基于人类知识和语义计算的手写数字语义符号库构建方法如图4所示.

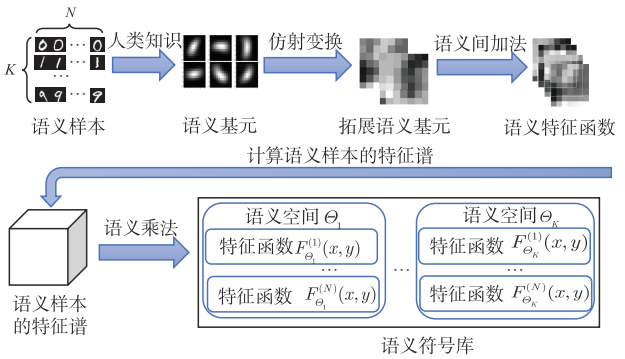


图4 基于人类知识和语义计算的手写数字语义符号库构建方法流程图

第1步,根据人类知识设计出手写数字的基础语义基元. 通过分析MNIST图像总结出手写数字的笔画可以拆解成各个方向上的短弧线,据此设计出如图4中所示的基础语义基元集合 $P=\{p_i(x,y)\}_{i=1}^{N}$,其中每个基元都可以用11×11的矩阵来表示.

第2步,通过仿射变换和语义间加法对基础语义基元进行组合和扩充,进而得到特征函数. 具体过程是,先通过仿射变换对基础语义基元集合进行扩充,这种经过仿射变换之后的基元可以在几何形变后保持相同的语义,提升了泛化性. 然后再将变换后的基元利用式(3)所示进行语义相加,得到候选特征函数 $F=\frac{1}{N}\sum_{i=1}^{N}T_i(p_i(x,y)))$.

第3步,使用语义乘法(如式(9)所示)计算出语义样本对应的图像语义空间下的特征函数. 首先以筛选出的特征函数作为卷积核,对语义样本进行卷积得到式(11)定义的特征谱集合. 接着根据像素点空间近邻关系作为式(7)中的组合规则 R,按照式(9)进行语义乘法,得到图结构的高维特征函数. 然后对该图进行图傅里叶变换,将高维特征函数转换为图像语义空间下的特征函数 $F_I=\{\boldsymbol{f}_{i,j}|\boldsymbol{f}_{i,j}\in\mathbf{R}^{(W_gH_gC)}\}_{i=1,j=1}^{K,N}$. 最后将特征函数按照式(1)构成集合,便得到了每张图像的语义.

第4步,根据MNIST数据集中对所选语义样本的类别标注,将同一类别的所有图像语义特征函数定义为一个语义空间,同样根据式(1)便完成了描述手写数字类别的语义符号库的构建,即 $S=\{F_{\Theta_i}|F_{\Theta_i}=\{F_{\Theta_i}^{j}|F_{\Theta_i}^{j}=\boldsymbol{f}_{i,j}\}_{j=1}^{N}\}_{i=1}^{K}$.

4.1.2 基于语义度量的图像识别

在完成语义符号库的构建之后,便可基于语义度量方法对图像进行识别. 首先,将待识别图像通过语义

符号库中的特征函数卷积得到式(3)中的特征谱,再经过语义乘法和图傅里叶变换方法计算出图像特征向量 $\boldsymbol{f}_I \in \mathbf{R}^{(W_g H_g C)}$. 然后,按照3.2.5节介绍的方法,根据符号语义库 S 中的 K 个子语义,对图像特征向量按照式(20)进行语义度量得到 K 个标量,组成语义向量 $\boldsymbol{f}_F \in \mathbf{R}^K$. 最后通过计算语义向量的softmax得到图像类别的预测概率向量 $\boldsymbol{p}=\left(\frac{\exp(f_i)}{\sum_{i=1}^{K}\exp(f_i)}\right)_{i=1}^{K}$,完成手写体数字的识别.

4.1.3　实验结果

本实验在MNIST数据集上将传统的卷积神经网络(Convolutional Neural Network, CNN)和本文方法进行了对比. 用于对比的CNN由4个3×3的卷积层和一个全连接层组成. 对比实验针对不同训练数据量计算了两种方法的测试准确率,对比结果如表1所示. 其中 n-shot代表每一类使用 n 张图片进行训练.

表1　本文方法和卷积神经网络在不同数据量下的对比

数据集大小	卷积神经网络	本文方法
1-shot	45.35	65.93
10-shot	62.61	83.13
100-shot	72.08	92.85
1000-shot	87.53	97.41
6000-shot	96.19	98.08

根据表1实验结果,我们可以得出如下结论:(1)在使用相同训练数据量条件下,本文的识别方法均优于传统卷积神经网络. 更进一步,本文方法比使用10倍数据量的传统卷积神经网络的准确率更高. 这表明了本文所提图像识别方法的有效性,从而验证了语义定义、度量与计算的可行性;(2)在1-shot到100-shot时,本文方法的准确率比传统卷积神经网络的准确率高20%以上,随着数据量增加,两者之间的性能差异逐渐缩短. 这是因为当训练数据较少时,语义知识发挥主要作用;当训练数据逐步增加时,数据驱动模型将逐步接近知识驱动模型的效果. 从这个角度讲,基于语义的知识驱动模型更适合用于训练数据缺乏的场景.

4.2　水声目标的识别

为了进一步验证本文所提出的语义度量方法的有效性,我们又将语义的概念用于水声目标的识别. 水下声音信号受到海洋背景噪声大、海况复杂、季节变化等多方面因素的影响,可用于有效识别的特征少,识别难度大. 现有的水声信号识别方法主要基于谱分析法,识别过程没有明确的语义,因此识别的精度有限. 基于本文所提出的语义概念,我们首先定义几种具有语义属性的水声信号基元表达,在此基础上构建用于水声信号识别的语义知识图谱,然后借助于图卷积神经网络(Graph Convolutional Network, GCN)进行语义基元间的关联推理,进而获得更高层次的语义表达,实现基于语义推理的水声信号识别.

本实验的目标针对三类舰船的水声信号进行识别. 实验中,我们首先定义了6种水声的语义基元特征,具体如表2所示.

表2　水声信号识别的语义基元定义

语义基元	描述
叶频	对应于轴频与叶片数的乘积
叶片数	螺旋桨的固有属性,根据轴频和叶频计算获得
音调	声音的固有属性,与声音信号的周期密切相关
音色	声音的固有属性
LOFAR(Low Frequency Array)谱基元	反映水声信号在时、频维度上的性质
通用声音基元	通过自然声音识别任务训练的VGG (Visual Geometry Group)网络,描述声音的一般属性

利用语义基元间的先验知识,构建图5所示的知识图谱,利用图卷积网络进行语义基元间的关系推理. 在三类水声信号的分类问题上进行了实验验证,三类水声信号的样本数分别为112、136和153. 实验过程中,采用80%的数据进行模型训练,剩下20%的数据用于测试. 为了验证基于语义推理的水声识别算法的有效性,将算法与传统基于支持向量机(Support Vector Machine, SVM)分类的方法进行了对比实验,实验结果如表3所示.

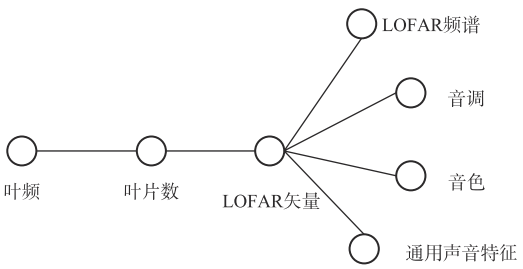


图5　基于语义基元的水声识别知识图谱

表3　基于语义推理的水声识别方法与传统基于SVM分类方法的性能对比

方法	传统特征SVM分类	语义度量与推理网络
分类精度	56.87	93.50

从表3的实验结果可以看出,传统基于特征SVM分类的方法识别精度较低,采用基于语义推理的网络识别的性能得到了大幅度的提升. 这是因为水声信号样本数量较少,且传统特征的语义层次低,因此基于SVM的方法无法获理想的识别性能;而语义基元可以从不同的层面描述水声信号的属性,图卷积网络又可以进一步对语义基元进行抽象获得更高层次的语义,形成更为深刻的信号表达. 因此,结合语义基元和图卷

积网络推理的方法能够有效捕捉水声信号的高层次语义特征,进而实现更准确的识别.

4.3 面向达意通信的视频信号语义编码

我们进一步将本文所提语义度量方法付诸实践,用于视频信号编码. 常用的视频编码技术以尽可能完整的传递视频信号为目的,其编码数据量随着视频清晰度提升而迅速增长,已经无法满足智能物联网等智能时代背景下的视频通信场景的需求. 然而,在大多数应用场景中,视频通信并不需要始终完整地传输视频信号,而只需传输其中的语义信息,实现达意通信即可. 例如,在视频会议场景中,通信双方需要的是面部表情和肢体动作所传达的意义,而不需要对方所处的环境、衣物纹理等信息. 因此,通过面向达意通信的视频语义编码能够有效地节省通信带宽,满足大规模视频通信需求.

本实验以大规模视频会议为背景,对视频中的人体姿态语义进行层级编码,其流程如图6所示. 我们将人体姿态语义分解为关节点和动作姿态这两级语义. 先根据人类知识定义了人体上14个关节点(如图6初级语义符号所示),组成初级特征函数库. 再对由14个关键点组成的人体骨骼图进行聚类,得到10种标准动作,组成高级特征函数库. 根据特征函数库,我们依次提取视频信号中的人体关节点和动作姿态这两级语义,作为语义编码结果. 具体流程为:首先,将输入的视频信号以初级特征函数库为标准语义进行信号语义度量,根据人体骨骼连接关系先验,便可将度量结果记录为图结构的初级语义符号;随后,以高级特征函数库,对初级符号进行语义度量,得到高级语义符号;最后,根据场景所需语义的层级,选择初级语义符号或高级语义符号进行传输.

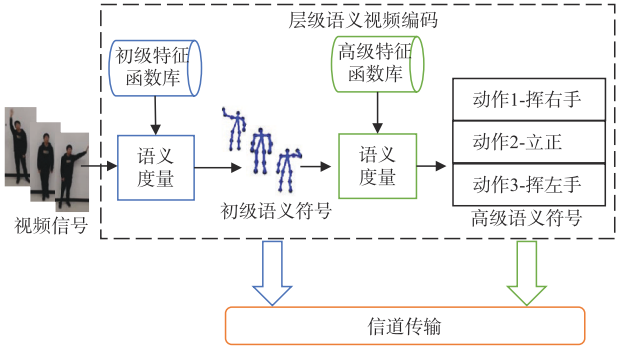


图6 视频信号层级语义编码

由于语义符号为图结构,在形象表达视频信号中语义的同时,还能极大减小信道传输的数据量. 我们对时长42 s、每秒25帧、分辨率512×512的原始视频进行了MPEG(Moving Picture Experts Group)编码和层级语义编码,其性能对比如表4所示. 从表中对比结果可以看出,面向达意通信的语义编码能够显著降低信道传输的压力,甚至可以在3 Kbps的带宽下传输视频语义. 而传统编码使用如此低的码率时,完全无法分辨视频中的语义. 因此本文所提语义度量方法在视频的达意通信中具有应用意义.

表4 层级语义编码和MPEG编码的性能对比

压缩方法	原始视频	MPEG编码	初级语义编码	高级语义编码
视频大小/Kb	268800	5990.4	34.1	0.5
压缩率	1.00	0.0223	1.27e−4	1.86e−6

5 结论

本文从基于语义的新型信息处理与通信技术引入,针对目前缺乏语义刻画和度量的数学描述这一问题,依据信息科学和神经科学相关结论,讨论了语义的内涵,并指出语义具有模块化、多模态、层级化的特点,由此提出了一种多模态信号的语义刻画和度量的数学描述. 为了验证所提信号语义的刻画和度量方法的可行性和有效性,分别在MNIST手写数字识别和水声目标识别两个应用中进行了实验,实验结果表明,基于语义的分类识别网络能达到比传统深度学习更好的效果. 本文还将语义用于视频编码,实现了远超传统方法的压缩比,展现了语义在通信领域的实用价值. 这为未来建立以语义为基础的新型信息处理与通信技术奠定了理论和实践基础.

参考文献

[1] KIFER M, LAUSEN G. F-logic: a higher-order language for reasoning about objects, inheritance, and scheme[J]. ACM SIGMOD Record, 1989, 18(2): 134-146.

[2] BAADER F. The Description Logic Handbook: Theory, Implementation, and Applications[M]. Cambridge, UK: Cambridge University Press, 2003

[3] BOLLACKER K, COOK R, TUFTS P. Freebase: a shared database of structured general human knowledge[C]// AAAI'07: Proceedings of the 22nd national conference on Artificial intelligence - Volume 2. Menlo Park, CA: AAAI Press, 2007: 1962-1963.

[4] LI Z Y, DING X, LIU T. Constructing narrative event evolutionary graph for script event prediction[C]//IJCAI'18: Proceedings of the 27th International Joint Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2018: 4201-4207.

[5] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. (2016-09-09). https://arxiv.org/abs/1609.02907.

[6] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al.

Graph attention networks[EB/OL]. (2017-10-30). https://arxiv.org/abs/1710.10903.

[7] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//ICML'01: Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc, 2001: 282-289.

[8] TERENT'YEV A N, BIDYUK P I. Method of probabilistic inference from learning data in Bayesian networks[J]. Cybernetics and Systems Analysis, 2007, 43(3): 391-396.

[9] MIKOLOV T, KARAFIÁT M, BURGET L, et al. Recurrent neural network based language model[C]//Interspeech 2010. Chiba: ISCA, 2010: 1045-1048.

[10] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[11] GOLDBERG Y, LEVY O. word2vec Explained: deriving Mikolov et al.′s negative-sampling word-embedding method[EB/OL].(2014-02-15). http://arxiv.org/abs/1402.3722.

[12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all You need[C]//NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2017: 6000-6010.

[13] DEVLIN J, CHANG M-W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018-10-11). https://arxiv. org/abs/1810.04805v2.

[14] LOWE D G. Object recognition from local scale-invariant features[C]//Proceedings of the Seventh IEEE International Conference on Computer Vision. Piscataway: IEEE, 1999: 1150-1157.

[15] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645.

[16] KOH P W, NGUYEN T, TANG Y S, et al. Concept bottleneck models[EB/OL]. (2020-07-07). https://arxiv. org/abs/2007.04612

[17] LU C W, KRISHNA R, BERNSTEIN M, et al. Visual relationship detection with language priors[C]//Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016: 852-869.

[18] WANG X L, YE Y F, GUPTA A. Zero-shot recognition via semantic embeddings and knowledge graphs[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6857-6866.

[19] CARON M, BOJANOWSKI P, JOULIN A, et al. Deep Clustering for Unsupervised Learning of Visual Features [C]//Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 139-156.

[20] LI O, LIU H, CHEN C, et al. Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions[C] // Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2018: 3530-3537.

[21] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014: 580-587.

[22] CHUTE C G. Classification and retrieval of patient records using natural language: An experimental application of latent semantic analysis[C]//Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society Volume 13. Piscataway: IEEE, 1991: 1162-1163.

[23] TURNEY P D, PANTEL P. From frequency to meaning: Vector space models of semantics[J]. Journal of Artificial Intelligence Research, 2010, 37: 141-188.

[24] MATSUNO K. Semantic commitments as a mode of non-programmable computation in the brain[J]. Bio Systems, 1992, 27(4): 235-239.

[25] SHANNON C E. A mathematical theory of communication[J]. The Bell System Technical Journal, 1948, 27(3): 379-423.

[26] SHANNON C E, WEAVER W. The Mathematical Theory of Communication[M]. Urbana: University of Illinois Press, 1949

[27] GÜLER B, YENER A, SWAMI A. The semantic communication game[J]. IEEE Transactions on Cognitive Communications and Networking, 2018, 4(4): 787-802.

[28] BAO J, BASU P, DEAN M K, et al. Towards a theory of semantic communication[C]//2011 IEEE Network Science Workshop. Piscataway: IEEE, 2011: 110-117.

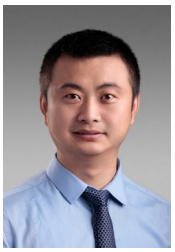
[29] BASU P, BAO J, DEAN M, et al. Preserving quality of information by using semantic relationships[J]. Pervasive and Mobile Computing, 2014, 11: 188-202.

[30] WILLEMS F M J, KALKER T. Semantic compaction, transmission, and compression codes[C]//Proceedings of International Symposium on Information Theory, 2005

ISIT. Piscataway: IEEE, 2005: 214-218.
[31] HUBEL D H, WIESEL T N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex[J]. The Journal of Physiology, 1962, 160(1): 106-154.
[32] HUTH A G, NISHIMOTO S, VU A T, et al. A continuous semantic space describes the representation of thousands of object and action categories across the human brain[J]. Neuron, 2012, 76(6): 1210-1224.
[33] HANDJARAS G, RICCIARDI E, LEO A, et al. How concepts are encoded in the human brain: A modality independent, category-based cortical organization of semantic knowledge[J]. NeuroImage, 2016, 135: 232-242.
[34] DEFELIPE J, MARKRAM H, ROCKLAND K S. The neocortical column[J]. Frontiers in Neuroanatomy, 2012, 6: 22.
[35] MOUNTCASTLE V B. The columnar organization of the neocortex[J]. Brain: a Journal of Neurology, 1997, 120 (Pt 4): 701-722.
[36] ROCKLAND K S. Five points on columns[J]. Frontiers in Neuroanatomy, 2010, 4: 22.
[37] BI Y C, WANG X Y, CARAMAZZA A. Object domain and modality in the ventral visual pathway[J]. Trends in Cognitive Sciences, 2016, 20(4): 282-290.
[38] HUTH A G, DE HEER W A, GRIFFITHS T L, et al. Natural speech reveals the semantic maps that tile human cerebral cortex[J]. Nature, 2016, 532(7600): 453-458.
[39] SHI G M, ZHANG Z Q, GAO D H, et al. Knowledge-guided semantic computing network[J]. Neurocomputing, 2021, 426: 70-84.
[40] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

作者简介

石光明 男，1965年06月出生于江西省南昌市．长江学者特聘教授，IEEE Fellow，IET Fellow．现任西安电子科技大学人工智能学院教授．主要研究方向为人工智能、语义通信．
E-mail: gmshi@xidian.edu.cn



高大化 男，1979年08月出生于河南省开封市．现任西安电子科技大学人工智能学院教授．主要研究方向为智能信息处理、智能感知．
E-mail: dhgao@xidian.edu.cn

杨旻曦 男，1996年10月出生于四川省成都市．现为西安电子科技大学人工智能学院博士研究生．主要研究方向为表征学习、语义通信．
E-mail: mxyang@stu.xidian.edu.cn

谢雪梅 女，1967年01月出生于陕西省西安市．现任西安电子科技大学人工智能学院教授．主要研究方向为场景理解与视频分析、多模态融合．
E-mail: xmxie@mail.xidian.edu.cn

董明皓 男，1984年05月出生于陕西省西安市．现为西安电子科技大学分子与神经影像教育部工程研究中心副教授．主要研究方向为脑机混合智能、人体效能增强．
E-mail: dminghao@xidian.edu.cn

李雷达 男，1982年10月出生于江苏省徐州市．现任西安电子科技大学人工智能学院教授．主要研究方向为图像感知质量评价．
E-mail: ldli@xidian.edu.cn

于 凯 男，1996年11月出生于山东省日照市莒县．2018至2021年于西安电子科技大学人工智能学院攻读硕士学位．主要研究方向为图像识别、增量学习．
E-mail: yukai_nathan@163.com