深度学习认知架构的反表征主义转向

刘 伟,符 征

(河南大学 马克思主义学院,河南 开封 475001)

摘要:当代认知研究发展出了符号主义和联结主义两种不同的范式。符号主义的计算—表征是以思想语言假说为基础的“句法图像”,具有内容与载体相分离、符号语境无关性等表征特征。深度学习是对联结主义技术的创新和深化,其认知架构是具有分布式加工和叠加存储、语境敏感和原型提取学习等特点的亚符号计算,表现出一系列的反表征特征,反映在深度网络中并不以明确的概念表征为对象的操作,推动了认知哲学中反表征主义的兴起。在充分理解符号主义和深度学习认知架构表征方式的基础上,探索二者在某种程度上的统一,也许是值得努力的目标。

关键词:深度学习;认知哲学;表征;反表征主义

[中图分类号]N02 **[文献标识码]**A **[文章编号]**1672-934X(2024)04-0054-07

DOI:10.16573/j.cnki.1672-934x.2024.04.007

The Anti-Representationism Turn in Deep Learning Cognitive Architectures

Liu Wei,Fu Zheng

(School of Marxism, Henan University, Kaifeng, Henan 475001, China)

Abstract: Two different paradigms, symbolism and connectionism, have evolved from contemporary cognitive research. The computational-representation of symbolism is a "syntactic image" based on the language-of-thought hypothesis, which is featured by the separation between content and its carriers and by the context-independence of symbols. Deep learning has been an innovation and depth to connectionism technique, whose architectures are characterized as a kind of sub-symbol computation with such features as distributed processing and superposition storage, context-sensitivity, and prototype extraction, has demonstrated a series of anti-representational features which have been reflected by operations in deep networks that do not target explicit conceptual representations, and has driven the rise of anti-representationism in cognitive philosophy. Exploring unity between symbolism and deep learning architectures in some degree based on fully understanding their representational approaches may be a worthwhile goal.

Key words: deep learning; cognitive philosophy; representation; anti-representationism

当前,深度学习是一种相当有影响力的认知架构。通过对联结主义技术进行一系列深化和创新,深度学习技术催生了一大批震惊世界的智能产品。作为一种认知架构,深度学习在

收稿日期:2024-03-27

基金项目:国家社会科学基金重大项目(22&ZD045)

作者简介:刘 伟(1991—),男,硕士研究生,研究方向为人工智能哲学、认知哲学;

符 征(1979—),男,副教授,主要从事认知哲学研究。

很大程度上表现出反表征主义的特征。这与以表征和计算为基础的符号主义产生了严重冲突。本文将分析表征的概念以及不同范式在表征问题上冲突的深层次原因。

一、认知哲学中的表征概念

在认知哲学中，表征（Representation）是指一种有关外部世界或内部心智状态的符号化形式或信息结构。表征是认知过程中对外部世界或内部思维的编码和存储，通过符号、符号系统、模型或图像等形式来表达和处理信息。

传统哲学认为，人的绝大部分心理活动都是以心理表征为对象的活动。这种以心理表征为基础对象来研究心灵或认知活动的观点，被称为“表征主义”。在认知科学中，表征特指心理表征或者内在表征。心理表征被看作是外部事物在心理活动中的内部显现，即通过表征外在事物与认知主体建立内在联系，进而实现对世界的认识。心灵如何表征事物？人的信念、欲望、希望、恐惧等心理状态何以表征世界？当“我希望/相信”的时候，“我”的每一个“希望/相信”都指向某种事物，即表达“我”对该事物的一种态度。同时，每一个“希望/相信”都有具体内容，如“我相信冰箱里有啤酒”的时候，相信的内容就是“冰箱里有啤酒”。在哲学上分别使用“命题态度”和“内容”这两个术语来描述表征的两个特征，不同的内容可以区分不同的态度。

在计算机加工中，表征指的是将信息以某种形式或结构进行编码和存储的过程。在计算机领域，表征主要用于描述和处理各种不同类型的数据和信息，为计算机系统提供处理、存储和传输信息的基础。表征在计算机加工中扮演着桥梁和媒介的角色，帮助计算机系统理解、处理和交互各种形式的数据和信息。简单来看，就是将一系列符号，即一些概念构造的表征，输入计算机之后，计算机通过预设的指令集来操作符号，并输出结果。正是基于这种理解，第一代人工智能和认知科学认为，人脑对信息的处理如同计算机的处理流程，是通过“信息输入—计算表征—输出”的计算—表征模型来实现的。因此，在人工智能领域，“大多数技术是通过使用明晰的表征知识和谨慎设计的搜索算法来实现智能的”[1]。

部分研究者通过计算机类比，将认知状态归属于计算的表征状态，计算机根据规则系统地处理符号表征。人的心智相对于大脑，正如计算机的软件相对于硬件，于是，产生了以符号加工为基础理论的符号主义进路。符号主义认为，表征在认知科学中扮演着重要角色。它允许我们将外部世界的信息转化为内部可处理的内容，帮助我们理解和描述外部世界的事物及其关系。通过表征，我们可以对复杂的信息进行处理、存储、检索和传递，进而实现认知活动和决策。

符号主义中最典型的理论是杰里·福多提出的“思想语言假说”。福多认为，表征必须是符号化和结构化的，“对思维最恰当的理解，是将其视为心智中的表征结构以及在这些结构上进行操作的计算程序”[2]。认知状态与过程是由心灵中某些承载信息的结构（表征）转化与储存的。对同一类型的例示（Token），心灵具有不同的存储方式。而同一信息可以存储于不同的载体，如语言表征和图像表征，信息储存方式的不同就是表征载体的不同。结构化的表征对应的是结构化的载体。如果命题态度通过内容在系统性、生成性与因果关系上有效的话，那么它们必须能够将内容的结构映射到具有结构特征的载体之上。思想内容可以通过结构化的命题态度来描述，故讨论命题

态度时需要对表征的内容与载体进行区分。

二、符号主义表征的特征

(一)内容与载体分离

福多认为,人类语言的基本框架对应于我们内在表征系统的结构,如果人类没有语言,就无从思考。当人们希望或者相信某件事情时,这个思维的表征就是一个句子。所以,当一个思考者具有一个内容为P的信念时,他的头脑中就写下了一个意义为P的句子。如果将思想语言的结构与机器语言的结构进行对比,那么可以发现思想语言具有与计算机语言一致的结构。如果将心理表征看作一个形式化系统,就可以将心灵与语言都置于符号操作装置之上。由此,借助计算机这一有效模型,就能够深入研究表征的变化机制。

语言之所以能够作为心理表征的载体,是因为在某一种语言体系中,词语和句子作为语言的元素都具有句法和语义结构,但词语和句子的特征与它们的形式并不是意义相关的。句法理论负责的是语言的基本表达形式或表达形式的组合的"合法性",即只保证其形式上符合语法或合法性。维特根斯坦说:"每个符号都是死的,是什么赋予了它生命?"[3]人类头脑中的符号何以具有句法性?福多认为,表征具有一种组合的句法和语义,"结构上复杂(分子)的表征,系统地建立在结构上简单(原子)的成分的基础上;并且一个分子表征的语义内容是它的原子成分的语义内容及其他的语法/形式结构的函数。"[4]一些符号被归为简单符号(原子符号),其实,它们是按照规则产生复杂符号(分子符号)的表征。符号具有了语义特征,才成为符号,而它们之所以具有语义特征,是因为它们表征了客观世界的某一事物。所以,福多认为,句法是在表征的因果作用与它们的内容之间起调节作用的,表征之间的语义关系可以由它们的句法关系得以模仿[5]。

句子的意义是由其构成成分的意义及其组合方式决定的。如"约翰喜欢玛丽"这个句子的意思,完全由其构成成分"约翰""喜欢""玛丽"的意义以及这些构成成分出现的顺序和句法角色来决定。符号主义认为,思想语言是通过符号与符号之间的纯形式规则运行,可计算的只有也只能是句法。因此,纯形式规则的推理可以确保语义属性通过句法属性反映出来。正如豪格兰德所说:"你只要管好句法,语义就会负责好自己的事情。"[6]克拉克称之为"句法图像"(Syntactic Image)[7](P6)。

(二)语境无关性

在"句法图像"中,信息是由符号串来表示的,就像数据在计算机内存里或纸张上被表征一样。载体符号串与心理内容形成一种表征关系,"这些内容被例示在内部符号串中,并借助内部符号串的因果而具有了因果效力",同时,"符号串是参与组成计算经济的结构化实体"[7](P6)。符号主义认为,对心灵进行表征的最好层次就是符号层次——使用计算机程序中的一个实体来指代现实世界中的某个实体。

福多和派利夏恩认为,表征潜力是由先天的表征基础,即符号系统所决定的,这些原子化的符号承载了固定的内容,并组合起来产生相对固定的语义内容。在句法图像中,表征是基于文本(符号)的,文本承载着固定的、语境不变的内容,同一符号可以在不同的语境中重复出现,并且能够保持语义不变。这种原子符号拥有固定内容,并且是完全独立于语境的,这样的表征在内部形成了一个符号系统。

句法图像表征因其句法结构而具有结构性。福多和派利夏恩认为,符号的组合产生新的结构,必须有"语境不变的组合原则"的参与,如果符号没有这种不变性,则无法保证符号系统推理的有效性。符号主义依赖于符号处理体

系结构，因此，可以很自然地直接表征数据结构，然后对这些结构进行处理，其计算的对象就是句法结构表征。符号主义并没有说明这种形式（规则）是什么，虽然其关涉逻辑规则，但并不能认为所有的规则都只是逻辑规则。

经典的句法图像依赖于串联编码形式，表征是按典型的串行方式而非并行方式处理的。串联编码是使用简单元素的信息空间的并置来创建复杂的结构。串联编码是在表达式本身保留表达式成分的例示，以及例示之间的顺序关系，即当一个表征被例示，表征的成分也会被例示。符号计算依赖于这种串联编码的空间形式，其中，符号表达式中的简单例示能够通过内部系统组合成复杂表达式的例示，在形成复杂表达式时，它们不会以任何方式被改变。例如，符号“P”，无论是单独出现还是出现在“P&Q”等表达式中，它的意义都是相同的。

总之，符号主义的句法图像表征是基于离散的原子符号和组合规则的结构化表征，涉及对无语境的内在内容承载者，即符号的操控，这对思想的生成性和系统性能够给予恰当的解释。因为，在真实的思维领域，可以通过少量的符号和逻辑规则组成无限的思维世界。句法图像虽然可以作为一个人类思维良好的解释模型，但这种静止的、被先天的表征基础所约束的表征变化是较弱的。对比经典符号主义模型下的人工智能，同样在计算主义纲领下的深度学习所依赖的深度网络肯定也存在计算，但在深度学习中计算与表征是分开的，计算处于概念层之下。因此，深度学习认知架构避免使用原子符号，而主张亚符号表征。

三、深度学习认知架构的特征

受神经系统多层联结架构的启发，深度学习认知架构是一个由大量神经元按照并行结构和层次结构组合而成的自适应、自组织的神经元网络系统。深度学习认知架构的运行基础是一个具有“深度”的学习网络，即“深度网络”。深度网络是由大量的单元组成的分层次网络，通常存在一个输入层、一个输出层以及多个起中介作用的隐藏层。每一个单元都与其他单元相关联，并且不同单元之间的联结具有不同的权重，信息是以非表征性的方式、以神经网络单元间的权重或连接强度来存储的。根据激活函数，每个单元的激活取决于它相邻单元的联结强度和活动。深度学习认知架构对计算和表征的理解都与符号主义不同。

（一）分布式加工和叠加存储

深度学习认知架构中的信息传递是分布并行的，也就是网络状态的变化在整个网络上同时发生，而非串联式地一步一步进行。与符号主义中存储在单独的、固定位置中的符号不同，深度网络以整体论的方式来存储和提取信息。分布式表征将意义分布到神经元的结构网络中，是对概念或命题更加复杂的表征。分布式表征的特点是信息传递通道多、神经元之间的联系多，且每一个神经元都处于由它周围神经元构成的语境信息中。因此，分布式表征体现出整体性、关联性、动态性的特点，表现出抗损坏、快速重新学习和自行恢复未排练项目的能力，还可以很好地处理符号主义的遗留问题，如“框架问题”。同时，系统具有多个联结来传递信息，当部分神经元遭到破坏时，信息仍然能够准确无误地传递下去。

分布式加工以及对所获得的信息施加语义度量的规则，导致了信息的叠加存储。所谓叠加，是指如果用于表征项目1的资源与用于表征项目2的资源具有相同的外延（Coextensive），就可以说这两种表征是完全叠加的。范·盖尔德定义了完全叠加表征：如果表征R是每个项目C_i的保守表征（R只用来表征C），则一系列项目C_i的表征R因参与表征C而被叠加[8]。由

于深度网络不是根据个别单元之间的逻辑关系获得,而是根据这些模式之间的相似性和差异性获得的,因此,表征的内部携带了有关内容的信息。表征内容之间的语义相似性,被认为是表征载体之间的相似性。也就是说,状态A与B叠加,状态B与C叠加,但A与C之间并不存在任何共同的成分。维特根斯坦所说的"家族相似性"正是如此,家族成员之间彼此相似却不一定具有共同特征,而是通过重叠的特征相互关联。基于深度网络的这种叠加内部资源(即采用叠加存储)用来表征语义的相似,是在深度学习中网络可以进行原型提取和学习训练的基础。

(二)语境敏感和原型提取学习

与符号主义认知加工中的语境无关性不同,深度学习认知加工是语境敏感的。符号主义认为,符号的语义是静态的,与语境无关的,在不同规则的句子中可以重复使用和替换符号。而在深度网络中,不存在语境无关的原子符号,因为整个系统通过非任意激活模式编码。例如,"杯中的咖啡洒到桌上"这一状态可以表征为一组微特征,这些微特征根据整体状态的语境而发生变化。咖啡在"洒出"的语境分布式模式中可能包括"接触桌布"等特征,而"在杯中"的语境中包括"接触瓷器"等特征。

在这种叠加存储系统中,一个或一组权值不表征任何固定语义的内容,因为每个权值都具备多重表征能力。权重产生的激活模式只在特定输入的意义上可以被认为表达了涉及语境的内容。这些模式受到语境的影响,无法显示出明确的系统性。

深度网络的另一个显著特征是可以通过"训练而习得",这是深度学习能够从数据中提取特征的主要能力。不同于符号主义认为的表征潜力是由符号系统所决定的,深度网络的深度学习是在没有任何原始资料的情况下获取某个领域的知识,它的基础并不是符号系统,而是初始的联结权重。但初始权重并不能被视作是一组有效的表征元素,因为这样的权重并不具有任何表征内容。

深度网络利用原型式表征来进行训练学习。原型概念是指围绕假定原型,通过与原型共同特征的对比,从而确定某事物是否属于该类别。例如,比较企鹅和麻雀,查找哪一个更符合"鸟"的典型原型。若找出麻雀属于"鸟"的典型原型越快速,则意味着麻雀表现出的原型特征越多。在深度网络训练时,首先,对内部一组样本进行分布式加工和叠加编码;其次,形成对样本共同特征的兴奋联结;然后,统计输入大量样本之间相互关联的语义特征;最后,提取统计中心的趋势并对特征集进行编码。

在深度学习中,环境和训练数据起着重要作用。深度网络是以系统性作为其加工的基本特征,但训练数据是从一开始就拥有其完整的表征信息,并被组织成一个系统性的符号系统。这种系统性也是在训练中逐步获得的,因此,知识基础和加工特性的变化是相联并行的。"知识的表征是以这样一种方式建立的,即知识必然影响加工过程。在加工过程中使用知识不再是在记忆中找到相关信息并加以利用的问题:它是加工本身的主要部分。"[9]通过对大量数据的叠加分布式的持续训练和对环境的适应性学习,深度网络可以发生质的变化。例如,一个开始只是机械地提取过去时态特征的深度网络,在经过持续训练之后,可以很突然地表现出对过去时态规则的掌握,同时获得不规则时态的知识。通过持续训练,深度学习在神经元计算基础上产生表征的"统计涌现"。这种表征涌现能力的根源在于知识和加工特性的深层次相互渗透,强调这种能力的整体性,就必须考虑环境和单个联结的作用。

这种方式似乎是将符号主义的符号与分布式的联结进行结合的产物,但是符号因其级别

太高而无法形成良好的心智模型[10]。联结单元也不同于符号主义的原子符号，神经元或者单个联结节点并不是表征，这种计算不是冯·诺依曼式的，而是从较低层次的计算中产生的。因此，联结主义排斥符号表征，采用了更低级别的亚符号表征。

四、反表征主义的兴起

在计算系统中，作为句法的实体是用来计算可操控的计算例示（Computational Token）。在符号主义系统中，这些计算例示是原子符号，同时也是语义的承担者，计算对象也是表征对象。而在亚符号系统中，计算例示是单个节点或者联结，表征则是分布式的权重和激活模式。联结主义者并不认可计算的对象就是语义解释的对象。可以看到，在这样的系统中，计算与表征是完全分开的，计算的对象比语义解释的对象更具有细粒度。查尔默斯认为，“在符号系统中，计算层次与表征层次一致。在亚符号系统中，计算层位于表征层之下。”[11]计算的载体可以不是表征的载体，表征的载体可以是整个网络的状态。因此，亚符号表征是非句法的，或者说其“句法”结构不在符号层面。

福多认为，思维语言的句子既是计算的载体，也是心理内容的载体，但联结主义显然并不需要如此。首先，在深度学习表征中，如“约翰喜欢玛丽”的表述，可以构建不包含任何显式表征的部分。我们可以轻易地从符号表征中提取相关成分的信息，但深度网络不需要显式地提取这些信息。这表明，联结主义模型与福多的命题“语言是人类认知的前提”是相反的，“在我们系统中流转的不是符号，而是刺激与抑制”[12]，换句话说，思维没有句法。斯莫伦斯基将深度网络描述为两级架构：“心理表征和心理加工不受同一形式实体的支持，没有‘符号’可以同时完成这两项工作。新的认知结构基本上是两级的：一方面，是处理机制的形式化、算法化说明，另一方面，是语义解释，必须在两个不同的描述层次上进行。”[13]其次，与符号主义的文本是语境无关的原子不同，联结主义主张的是内在语境敏感和结构敏感表征。深度网络将单元之间分布式的激活模式作为表征的中心，使其具有了丰富的内部结构，且这种结构不同于句法规则的直接的显式结构，而是低层级的间接操作。深度网络表征具有微观语义——系统地反映表征意义的内部模式。“通过比较简单和不太抽象的层次结构组成更复杂的特征表征。”[14]例如，在简单概念网络中，“白狗”与“黑狗”的激活模式具有相似性，因为它们都是“狗”这一概念下的子概念，这显示了激活模式与概念意义的统一性。

总之，深度网络采用了与符号主义不同的范式，表征的加工混合、原型提取学习、语境敏感性以及“语义度量”叠加的使用等构成联结主义表征的基础，强调表征的“并行性、分布性、动态性、涌现性”[15]。这些都是以深度学习认知为代表的第二代人工智能得以飞速发展的特性。尽管联结主义既有句法（计算例示），也有语义对象（表征），但是二者并不在一个层次。将认知过程的解释诉诸亚符号层面的计算，网络的整体语义通过分布式的表征承载。在由不同激活模式构成的分布式网络计算中，网络内部不但是语境敏感的，也是结构敏感的。

深度学习认知架构的兴起推动了认知哲学中反表征主义的兴起。深度网络中的分布式加工和叠加存储、语境敏感和原型提取学习等特性，反映了在深度网络中并不以明确的概念表征为对象的操作。虽然一些加工结果看上去具有表征特性，但是这些表征特性只是出现在部分加工结果中，而不是出现在加工过程中。这种认知架构的哲学启示是：也许心灵活动的底层逻辑并不是以概念表征为对象的，而那些亚概念符号的活动才是心灵活动的真正承担者。这一结论无疑对柏拉图以来的整个西方心灵理

论产生巨大的冲击。

从结果上来看,基于符号主义的人工智能侧重于模拟人类的抽象思维形式,深度网络则侧重于模拟人类的思维硬件构成形式。当然,这两种进路都存在一定的局限性,目前,还没有一种方法可以一致适用所有领域。当前,调和深度学习认知和符号人工智能的方法,如"表示对象—关系"方法[16]、"神经—向量—符号"方法[17]等已被提出。在新的条件下,探索符号主义和深度学习在某种程度上的统一,也许是值得努力的目标。

[参考文献]

[1] 魏屹东.表征概念的起源、理论演变及本质特征[J].哲学分析,2012(3):96-118,166,199.

[2] [加]P·萨伽德.心智 认识科学导论[M].朱菁,陈梦雅,译.上海:上海辞书出版社,2012:11.

[3] Wittgenstein L. Philosophical investigations: the English text of the third edition[M].New York:Macmillan,1953:432.

[4] Aydede M. The language of thought hypothesis[EB/OL]. Stanford Encyclopedia of Philosophy Archive,http://plato.stanford.edu/archives/fall2010/entries/language-thought.

[5] 符征.语义引擎的形成及其应用[J].自然辩证法研究,2013(11):21-25.

[6] Haugeland J. Semantic engines: an introduction to mind design[M]. Haugeland J. Mind design. Cambridge: The MIT Press,1981:44.

[7] Clark A. Associative engines: connectionism, concepts, and representational change[M]. Cambridge: The MIT Press, 1993:6.

[8] Gelder T V. What is the "D" in "PDP"? a survey of the concept of distribution[M].Ramsey W,Stich S P,Rumelhart D E. Philosophy and connectionist theory: a perspective from psychology and artificial intelligence. New Jersey: Lawrence Erlbaum Associates,1991:43.

[9] Mcclelland J L,Rumelhart D E,Hinton G E.The appeal of parallel distributed processing[M]. Collins A, Smith E E. Readings in cognitive science: a perspective from psychology and artificial intelligence.Amsterdam:Elsevier, 1988:52-72.

[10] Hinton G E, McClelland J L, Rumelhart D E. Distributed representations[M]. Rumelhart D E, McClelland J L, the PDP research group. Parallel distributed processing: explorations in the microstructure of cognition.Cambridge: The MIT Press,1986:77-109.

[11] Chalmers D J. Subsymbolic computation and the chinese room[M]. Dinsmore J. The symbolic and connectionist paradigms: closing the gap. London: Lawrence Erlbaum, 1992:25.

[12] [英]蒂姆·克兰.机械的心灵:心灵、机器与心理表征哲学导论[M].杨洋,译.北京:商务印书馆,2016:211.

[13] Smolensky P. Connectionism, constituency, and the language of thought[M].Lower B,Rey G.Meaning in mind: fodor and his critics.Oxford:Basil Blackwell,1991:203.

[14] Buckner C.Deep learning: a philosophical introduction[J]. Philosophy Compass,2019,14(10):e12625.

[15] 李光辉,陈刚.人工智能的内在表征何以可能[J].自然辩证法通讯,2019(3):41-47.

[16] Garnelo M, Shanahan M. Reconciling deep learning with symbolic artificial intelligence: representing objects and relations[J].Current Opinion in Behavioral Sciences,2019, 29:17-23.

[17] Hersche M, Zeqiri M, Benini L, et al. A neuro-vector-symbolic architecture for solving Raven′s progressive matrices[J]. Nature Machine Intelligence, 2023, 05(04): 363-375.