智能计算体系结构与系统

**一、研究背景**

近年来，人工智能“符号主义”、“连接主义”、“进化主义”和“统计主义”等四种流派呈现出交叉融合的趋势，并诞生了AlphaGo等突出成果。以AlphaGo为代表的人工智能技术在蓬勃发展的同时，也暴露出了传统冯·诺依曼计算机结构的不足之处。传统CPU以通用目的而设计，而且存在“存储墙”问题，在处理深度学习和超大规模脉冲神经网络（SNN）等专门问题时性能与能效性等方面无法达到最优，因此主流的人工智能系统框架大多使用CPU+GPU的混合结构，但是这种混合结构也同样面临着能效性的问题。使用专门集成电路（ASIC）可以取得较好的能效性，然而ASIC的设计周期可以长达数月至一年，远远长于算法更迭的周期，因而出现了基于FPGA或CGRA等可编程结构的设计方案。总体而言，基于深度学习等技术框架的应用和基于SNN的类脑研究对智能体系结构在高能效性、可扩展性、高容错性、实时性和支持神经网络模型的灵活性等方面提出了新的要求。

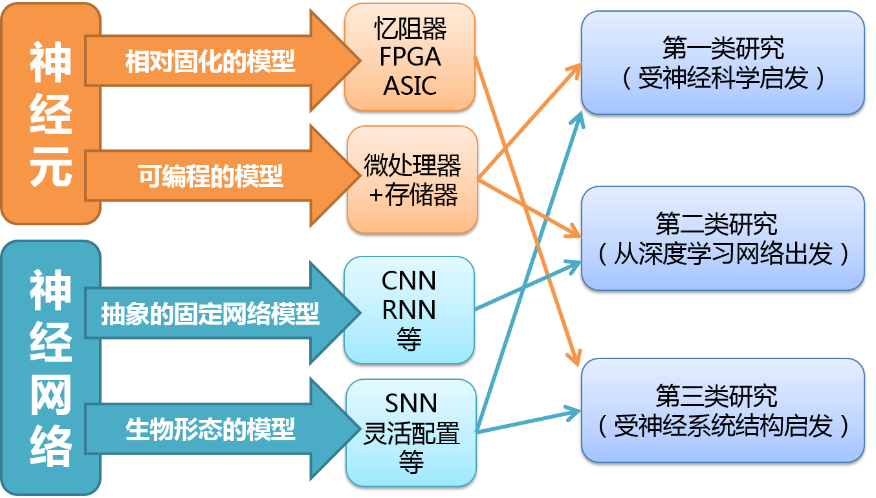
现如今相当多的人工智能系统在专项任务中取得了令人瞩目的成果，但是我们仍需注意到一点：这些人工智能系统都是针对专门任务设计的，而不是针对更广泛的智能行为而设计。然而早在1979年，Vernon Mountcastle就发现大脑用相同的原则来处理视觉、听觉、触觉等行为。他发现大脑新皮层的所有区域都执行相同的基础操作，会产生不同分工的皮层仅仅是由于它们从不同的来源接收到信号输入。这一事实说明，为了使人工智能有更进一步的发展我们仍有必要探索新的智能计算体系结构。

从计算机科学的角度来看，针对智能计算体系结构的研究可以分为三类：

第一类研究被神经科学所启发，主要目的是构建能够高精确度地模拟大规模生物神经系统结构的软硬件系统，从而更好地理解大脑的结构以及工作过程。尽管这类研究的一个重要目的是为神经科学的发展提供支持，但是其中的许多项目开拓出了与传统的超级计算机集群不同的体系结构。

第二类研究从现有的深度学习网络结构出发，针对特定类型的神经网络以及特定的训练算法开发专用的硬件来加速深度学习中的学习或训练过程。

第三类研究受神经系统结构启发，参考神经系统的结构来开发新形态的硬件系统。这类系统参考生物的神经系统来组织自身结构，并利用数字电路或模拟电路来实现神经系统中的组成成分，如神经元、突触和灰质等。这类研究与第一类研究不同，因为其目的并不是为了模拟神经系统，而是希望从神经科学中得到启发来构建低功耗高性能的新型体系结构；同时，这类研究也不同于第二类，因为这类研究并没有针对特定的一种或一类神经网络模型去进行设计。



**二、研究现状**

根据上一节介绍的研究分类，分别对国内外智能计算体系结构的发展成果进行简要的梳理和总结。

第一类研究中有代表性的包括欧盟2011年启动的BrainScaleS[Brüderle2011]，HP公司联合Boston大学开发的Cog Ex Machina[Snider2011]和曼彻斯特大学的SpiNNaker系统[Furber2013]。

* 2005年，欧盟启动了FACETS（Fast Analog Computing with Emergent Transient States）项目，由海德堡大学牵头研制基于模拟混合信号的神经形态芯片。BrainScaleS（Brain-inspired multiscale computation in neuromorphic hybrid systems）在2011年启动，作为FACETS的延续项目，目标是研发大规模并行类脑计算机。BrainScaleS中包括了4500万个可编程的动态突触，使用FACETS项目中开发的晶圆级别硬件系统作为组成单元。2016年3月，欧盟人类大脑计划宣布将BrainScaleS系统通过互联网对外开放使用，以以支持神经微回路模拟以及在机器学习和认知计算中应用类脑原理的相关研究.
* Cog Ex Machina是一款低功耗、全数字平台的硬件系统，利用多线程和张量数据表示实现了层级结构、并行处理等生物大脑的特征，建立了可以与模拟真实环境实时交互的大脑模型。Cog Ex Machina系统尚未达到生物规模的模拟能力，但是仍为研究者提供了一条成本相对较低地快速建模验证新算法的途径。
* 曼彻斯特大学开发的SpiNNaker系统基于ARM 968处理器研发，整个系统中共使用了1036800个处理器，峰值处理速度可以达到每秒228兆Dhrystone指令，而能量消耗可以控制在90kW内。SpiNNaker使用软件模拟神经元，每个核可以模拟1000个神经元，突触则使用SDRAM进行模拟。2016年3月，SpiNNaker与BrainScaleS系统一同通过互联网对外开放使用。

第二类研究由企业主导的情况较多，其中具有代表性的成果有谷歌公司开发的TPU [Jouppi2017]、微软公司的Branwave[Burger2017]以及中科院计算所开发的“寒武纪”芯片[Chen2014]。

* TPU是谷歌公司专门为TensorFlow深度学习框架所设计的一款处理器，主要加速深度学习中的矩阵运算部分。TPU使用脉动阵列结构对矩阵运算进行加速，核心思想是保持矩阵运算单元的忙碌状态。与同时期的CPU和GPU相比，TPU可以将每瓦特的计算性能提高30-80倍。
* 微软公司在2017年的Hot Chips会议上初次披露了自己的深度学习加速平台Brainwave。Brainwave平台主要面向实时AI应用，未来将会在微软的云平台Azure和Bing搜索服务中投入使用。Brainwave以深度神经网络DNN为主要加速对象，可以将DNN网络映射到一个FPGA池中，在Intel的14nm Stratix 10 FPGA上可以达到29.5 Teraflops的运算速度。
* 中科院计算所采取了不同的研究思路，从针对神经网络的指令系统出发，开发了“寒武纪”芯片。“寒武纪”芯片每秒能处理160亿个神经元和2.56万亿个突触运算，可达到每秒512G次浮点运算速度，可广泛适应各种智能处理应用。

第三类研究的成果包括IBM开发的TrueNorth神经网络芯片[Merolla2014]、斯坦福大学开发的Neurogrid[Benjamin2014]、浙江大学的“达尔文”芯片[Shen2016]以及清华大学的“天机”系列芯片[Shi2015]。

* TrueNorth基本结构由硬件神经元和神经元之间的脉冲连接组成，神经元模型使用注重速度的LIF模型。TrueNorth具有4096个处理器核，每个核中包含256个硬件神经元，总计可以模拟100万个神经元与2.56亿个突触。
* Neurogrid基于数模混合电路构建，利用晶体管的亚阈值工作区来实现突触和神经元的特殊功能，用以实时模拟大尺度神经网络。Neurogrid可以以极低的功耗实时模拟16×256×256个神经元与数十亿个突触构成的神经网络系统。
* “达尔文”使用LIF神经元模型，本身作为协处理器辅助完成计算任务，最大支持2048个神经元和约420万个突触，以及15种突触延迟。“达尔文”的工作频率为70MHz，虽然频率很低，但处理延迟并未显著增加，并在多种分类算法中取得了高准确率的结果。
* 清华大学的“天机”芯片可以支持多种神经网络模型。“天机”单个芯片中包含6个核，每个核中包含神经元、权值网络、路由和参数管理等模块，运行频率为100MHz，可达到每秒153.6G次定点运算速度。

**三、研究内容**

为了在以神经形态计算为代表的智能计算系统关键技术领域取得突破，应在神经科学、计算机系统、电子器件和材料等领域形成密切的相互协作。为了达到这一目的，需要加强如下几点的研究：

**（1）智能计算相关算法和基础理论研究。**基于神经科学、非线性动力学和计算科学等理论，深入研究智能计算相关基础理论及算法。了解系统基本特性及操作原理，在适应现有计算机训练学习的同时取得算法、理论等基础性突破，为智能计算系统的设计研究提供理论支撑。

**（2）面向智能计算的大规模模拟平台。**面向智能计算开展大规模模拟平台的研究，可在不同抽象层次支持不同神经系统结构的研究需要。提供器件级、电路级、材料级的大规模模拟，支持和小型器件的联合使用；高层模拟支持对计算模块、连接模式、可编程性要求的检查和测试；提供对系统性能、特性及能效的模拟估计功能。模拟平台可对理论结果进行检测并协助训练算法的开发，从而有效改进设计。利用大型模拟平台的开发，深入研究智能计算系统研发中的软件、结构、器件等问题，并进一步优化系统结构。

**（3）超大规模类脑计算机平台。**参考人类大脑神经网络的基本结构，研究设计新型超级神经形态计算机体系结构和系统，重点解决类脑模拟中的超高并行、超大规模、低功耗和复杂互连等问题，支持灵活神经元模型和神经网络配置，实现数百亿个神经元、数万亿个突触的实时模拟仿真。类脑计算机平台可以用于探索人脑的结构和运行机理，支持人工智能的理论创新。此外，平台还可以探索新的计算机体系结构，突破传统超级计算机的局限性，实现低功耗、高并行性的新型超算系统。

**（4）面向智能计算的芯片及系统架构。**从计算模块研发、芯片研制、系统架构设计等方面对神经形态硬件展开研究。结合神经科学、材料科学、器件工艺、电路设计及软件等多领域需求，改进材料和器件的性能参数，利用新材料、新器件、新电路和新体系结构构建完整的功能计算系统，并面向智能计算系统的集成和原型设计展开研究。

**（5）面向智能计算的配套基础软件。**研发面向智能计算的配套基础软件环境。该软件环境应对模拟仿真平台和神经形态硬件提供良好支持，基础软件环境须和大规模模拟仿真平台形成配合。软件应提供较好的编程开发环境，具备可视化功能、用户界面和便捷的调试分析工具，以支持软硬件开发人员的调试及系统交互。

**（6）基准套件和评测技术。**对评测指标和基准进行研究，对智能计算模型和系统进行量化评价。定义一套面向智能计算系统测试的基准测试套件，该评测标准及套件应准确反映智能计算系统的功能和性能特性，应包含分类、控制、异常检测等系统功能及神经科学相关的基准程序在内。基准套件中评测程序的选择应足够多样化以避免限制智能计算系统的创新，其程序和数据集应对社区开放。

**四、参考文献**

1. [Benjamin2014] Benjamin, Ben Varkey, et al. "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations." Proceedings of the IEEE 102.5 (2014): 699–716.
2. [Brüderle2011] Brüderle, Daniel, et al. "A comprehensive workflow for general-purpose neural modeling with highly configurable neuromorphic hardware systems." Biological cybernetics 104.4-5 (2011): 263–296.
3. [Burger2017] Doug Burger, "Microsoft unveils Project Brainwave for real-time AI", August 22, 2017. Accessed at http://www.microsoft.com/en-us/research/blog/microsoft-unveils-project-brainwave/
4. [Chen2014] Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu, Yunji Chen, and Olivier Temam, "DianNao: A Small-Footprint High-Throughput Accelerator for Ubiquitous Machine-Learning", in Proceedings of the 19th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'14), 2014.
5. [Furber2013] Furber, Steve B., David R. Lester, Luis A. Plana, Jim D. Garside, Eustace Painkras, Steve Temple, and Andrew D. Brown. "Overview of the SpiNNaker system architecture." IEEE Transactions on Computers 62.12 (2013): 2454¨C2467.
6. [Jouppi2017] Jouppi et al. "In-Datacenter Performance Analysis of a Tensor Processing Unit." In Proceedings of the 44th Annual International Symposium on Computer Architecture, 1-12. Toronto, ON, Canada: ACM. 2017.
7. [Merolla2014] Merolla, Paul A., et al. "A million spiking-neuron integrated circuit with a scalable communication network and interface." Science 345.6197 (2014): 668–673.
8. [Shen2016] Shen, Juncheng, et al. "Darwin: a neuromorphic hardware co-processor based on Spiking Neural Networks." Science China Information Sciences59.2 (2016): 1¨C5.
9. [Shi2015] L. Shi et al., "Development of a neuromorphic computing system," 2015 IEEE International Electron Devices Meeting (IEDM), Washington, DC, 2015, pp. 4.3.1-4.3.4.
10. [Snider2011] G. Snider, R. Amerson, D. Carter, H. Abdalla, M. S. Qureshi, J. Léveillé, et al., "From synapses to circuitry: Using memristive memory to explore the electronic brain," Computer, vol. 44, p. 21, 2011.