Az egységes számítási modell

The unified computing model

VÉGH János, MTA doktora, egyetemi tanár[1], Dr.BERKI Ádám József[2]

[1]Kalimános BT, Debrecen, 4032, Magyarország, Vegh.Janos@gmail.com
[2]Marosvásárhelyi „George Emil Palade" Orvosi, Gyógyszerészeti, Tudomány és Technológiai Egyetem, 540142, Marosvásárhely, Románia, berki.adam@yahoo.com

Abstract

Although the commonly used classic computing model was a biology-inspired one, von Neumann underlined that his simplified paradigm is based on strong omissions. Those strong simplifications make unsound applying the classic paradigm for both modern electronic technologies and biological computing systems. However, von Neumann did not provide another procedure (without critical omissions) to describe those systems. To describe the operation of those latter systems, a generalization of the computing model is required. The introduced general computing model considers computing as a chain of constrained processes, where the transfer and processing processes are aligned by events. Those events are generated and used in different ways in the different implementations, resulting in different behavior of the computing systems. However, the general principles of computing remain the same. Scrutinizing the general principles of computing reveals not only why various implementations make hard to imitate one implementation with another and why large-scale computing systems show the experienced performance issues, but also reveal the mechanism how biological computing systems can represent, store and process information; furthermore that why learning and machine learning are entirely different.

Keywords: computing model; information storing; computing paradigm; biomorph computing; machine learning

Kivonat

Bár a jelenleg használt számítási modellünket a biológia inspirálta, Neumann János hangsúlyozta, hogy az általa javasolt klasszikus számítási paradigma erős elhanyagolásokat használt. Ezek az egyszerűsítések értelmetlenné teszik, hogy az egyszerűsített klasszikus paradigmát a modern technológiai rendszerekben megvalósított vagy a biológiai rendszerekben megvalósuló számítások leírására használjuk. Neumann János azonban nem közölt olyan (elhanyagolások nélküli) eljárást, amelyet ilyen rendszerek leírására használhatnánk. Ilyen célra a számítási modell általános tárgyalása szükséges. Ilyen célra a számítási folyamatot adat átviteli és feldolgozási folyamatok megfelelően egymáshoz illesztett láncolataként kezeljük, ahol az illesztést események valósítják meg. Ezeket az eseményeket a különböző módokon megvalósított számítási rendszerek különböző módokon állítják elő. A számítási folyamat általános elvei azonban azonosak maradnak. Az általános számítási elvek figyelmes analízise nem csak azt fedi fel, hogy miért nehéz egyik féle számítási megvalósítást egy másikkal imitálni és miért tapasztaljuk a nagy-skálájú számítási rendszerekben az ismert teljesítőképesség problémákat, hanem megismertet azokkal a mechanizmusokkal is, amelyek használatával a biológia számítási rendszerek ábrázolják, tárolják és feldolgozzák az információt; továbbá, hogy miért alapvetően más a biológiai és a gépi tanulás eszköztára és technikája.

Kulcsszavak: számítási modell; informaáció tárolás; számítási paradigma; biomorph számítás; gépi tanulás

1. BEVEZETÉS

Gyakran halljuk, hogy nem csak technikai számítási rendszereink, hanem agyunk is „számol". Az utóbbiról nem igazán tudjuk, hogyan is működik [1], bár sok ismerettel rendelkezünk róla [2]; de az

előbbire vonatkozó ismereteink pontosságával kapcsolatban is merülnek fel kétségek [3]. Agyunkról a működés tanulmányozásának megkezdése óta tudjuk, hogy (a neurális folyamatok lassúsága miatt) működése ún. spatiotemporális jellegű, azaz ugyanazt az információt az agyon belül más helyütt más időben (és a környezet által okozott kis torzulásokkal) látjuk. Mivel a jel továbbítási idő (akár sokkal) hosszabb, mint a jel feldolgozási idő, tudjuk, hogy az agy működését nem írhatjuk le a technikai számításokra általánosan használt klasszikus számítási paradigma alapján, mivel az abban használt elhanyagolások miatt az ilyen esetre nem érvényes. A technikai számítási rendszerekben csak a nagyon nagy (vagy nagyon intenzív számítási kommunikációt végző) rendszerek esetén válik nyilvánvalóvá, hogy a jelenségek a mérési hibán túl eltérnek attól, amit a klasszikus paradigma alapján várnánk. A két esetben közös, hogy a klasszikus, egyszerűsített számítási paradigma hibás leírást eredményez. A helyes leírás megvalósításához vissza kell térnünk az általános, elhanyagolás nélküli modellhez [4].

A műszaki technológiában használatos információ közvetítési folyamatok – éppen biológiai működésünk lassúsága miatt – érzékelhetetlenül gyorsan zajlanak. Az intenzívebb használat és a nagyon nagy rendszerek iránti igény azonban felfedte, hogy a „nagyon gyors" azért nem végtelenül gyors, és – éppen azért, mert véges a terjedési és a feldolgozási sebesség – a nagy technológiai rendszerek viselkedése éppúgy eltér a mindennapos használatú „játék" rendszerek viselkedésétől, mint ahogyan csupán a neuronok működésének ismeretében nem tudjuk levezetni agyunk működésének tapasztalt viselkedését, pl. hogy van tudatunk. A biológiai jel terjedési sebessége 10 milliószor lassúbb, mint az elektromágneses hullámok esetén, továbbá a biológiai jel továbbítás kevert kémiai/elektromos továbbítása eltérő mechanizmusok használatát teszi lehetővé és szükségessé, a folyamatoknak az általános modellben feltételezett összefűzése is másként valósul meg. A két mechanizmus egyidejű tanulmányozása kölcsönösen segíti a működés jobb megértését, és rávilágít arra, hogy azok egységes modell alapján működnek, bár a folyamatok és események megvalósítási módja teljesen eltérő és ezért a kétféle megvalósítás tulajdonságai is erősen különböznek. Tudjuk, hogy a biológiai rendszerek élethosszig tanulnak, rövid és hosszú távon egyaránt; amit a technológia rendszerek csak nagyon rossz hatásfokkal és különféle (tanító és normál) üzemmódokkal tudnak elérni.

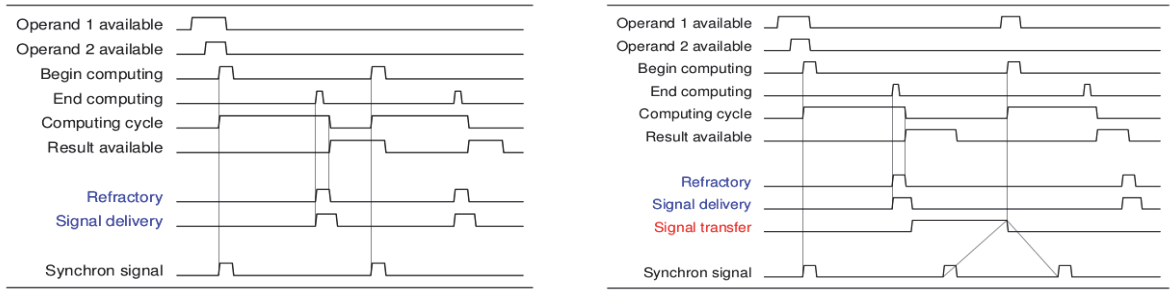
A technológia oldaláról sikerült felfedni, hogy – bár elvileg tudtuk, hogy az elektromágneses hullámok terjedési sebessége véges és a nagyon gyors elektronikus állapot változások követését már az anyagi tulajdonságok korlátozzák [5] – a fizikailag nagy méretű, nagyon sok, és a feladat megoldása során egymással adatokat cserélő számító elemet tartalmazó, nagyon sok adattal dolgozó nagy számítógépek teljesítőképessége már elérte a fizikailag lehetséges határt [6]; ezért bizonyos típusú feladatok megoldására a jelenlegi technológiai kivitel már elvileg sem alkalmas. A biológia oldaláról pedig, bár szinte szó szerint atomjaira szedték a neurobiológusok az agyat, nem sikerült megtalálni, hogy valójában hogyan tárolja a biológiai számítási rendszer az információt [7]; ezért annak pontos részleteit sem, hogy hogyan dolgozza azt fel. Mindennek ellenére ambiciózus tervek születtek és világszerte folynak jól támogatott kutatások jelenleg is az agy működésének számítógéppel történő modellezésére; inkább kevesebb, mint több sikerrel [8].

2. A SZÁMÍTÁSI MODELL

Háromnegyed évszázada Neumann János a neurális működés ismeretének akkori szintjén alkotta meg azt a számítási modellt, amelyen mai számítógépeink működése alapszik. Már akkor figyelmeztetett, hogy sokkal gyorsabb processzorok esetén az általa használt elhanyagolások nem érvényesek; továbbá, hogy az általa javasolt „klasszikus paradigma" nem használható neurális működés leírására. Mára a technológia fejlődése következtében számítási szerkezeteink működése szinte nem is hasonlít arra, amit a számítás tudomány alapját képező egyszerűsített paradigma feltételez [4].

Neumann alapgondolata az volt, hogy a számításokat lánc-szerűen végezzük, azaz az egyik elemi számító egység eredményét egy másik számító egység bemenő adatként használja, és hogy a számító egység működése ideje mellett az egységek közötti átviteli időt is figyelembe kell venni. A számítás és az átvitel kölcsönösen akadályozzák egymást: a műveletet nem is tudjuk elkezdeni, amíg az operandusok meg nem érkeztek; fordítva pedig addig nem lehet elszállítani az eredményt, amíg a számítás el nem készült. Ez még akkor is igaz, ha a technológiai kivitelben fizikailag ugyanaz a processzor végzi az egymás utáni számításokat: az eredményt az „output section"-ből át kell szállítani az „input section"-be, mielőtt a következő műveletet végeznénk. Ezek a tárolók hosszú hozzáférési idővel is rendelkezhetnek. Neumann korában azonban a művelet végzési idő néhány millisecundum nagyságú volt, az átviteli

idő meg néhány mikrosecundum nagyságú; ezért Neumann teljes joggal mondotta, hogy az akkori technológiát nagyon jó közelítéssel leírja a ma klasszikusnak nevezett, egyszerűsített modell. Mára viszont az időzítési viszonyok megfordultak. A lényegi elhanyagolást mutatja az 1. ábra. A teljes modell (jobb oldalt) tartalmazza az átviteli időt is, az egyszerűsített model (bal oldalt) pedig nem (azaz, feltételezi az azonnali kölcsönhatást). A teljes modell mind technológiai, mind biológiai számításra alkalmazható, bár a folyamatok sorba rendezését más módon előállított jelek végzik el.



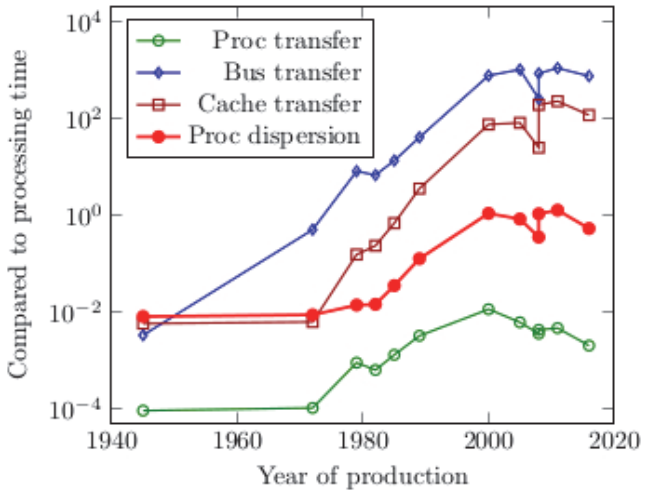
1. ábra: *A számítási és adat átviteli idők viszonya Neumann egyszerűsített (bal oldali) és teljes (jobb oldali) számítási modelljében. Láncolt műveletek esetén nyilván figyelembe kell venni az átviteli időt; ez rövid számítási idők és/vagy nagy fizikai méretű számítási rendszerek esetén alapvető különbséget jelent.*

3. SZINKRONIZÁLÁS

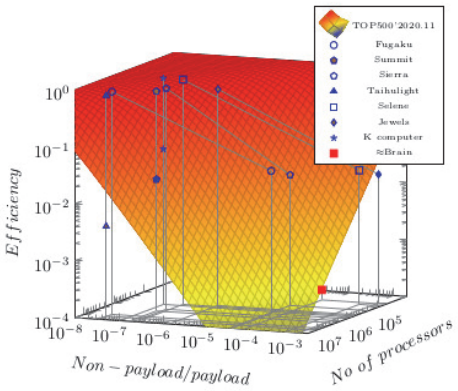
Összetett rendszerek esetén külön nehézséget jelent az elemi számító egységek működésének összehangolása, a szinkronizálás: mind a műveleti idő, mind a szállítási idő esetről esetre változik. A technológia és a biológia különböző szinkronizálási módot használ. Neumann az említett időzítési viszonyokra javasolta a központi órajel használatát szinkronizálásra technológiai számító rendszerekben. Ez a megoldás kétségtelenül egyszerűsíti kis rendszerek esetén a kivitelezést, de a szinkronizálás egyúttal valamelyik művelet várakoztatását (a hasznos teljesítőképesség csökkenését) is jelenti. Az alapfeltétel, hogy minden művelet befejeződjön egy órajel alatt, továbbá hogy a gazdaságos működtetéshez a műveletek időbeli hossza nagyon egyforma (diszperziója nagyon kicsi) legyen. A technológia fejlődés során óriási mértékben megnőtt a processzorok (elemi kapuk számában kifejezett) mérete és egyre kevésbé teljesül, hogy kicsi legyen a diszperzió (2. ábra). A diszperzió mérete még aránytalanabbul nő, ha a számítógép összetevőinek fizikai méretét is figyelembe vesszük (a Moore-megfigyelés csak a processzszoron belüli alkatrész sűrűségre teljesül, úgy, ahogy). Az ábrából az is megállapítható, hogy az átmeneti tároló (cache) bevezetése akkor vált szükségessé, amikor a buszon át történő memória kezelés súlya jelentősen megnőtt a mikroelektronikai sűrűség jelentős növekedése miatt. Mai processzorainkban a diszperzió nagyságrendje a feldolgozási idő nagyságrendjébe esik, azaz nem elhanyagolható. A következmény, hogy modern processzoraink az energia nagy részét hőtermelésre fordítják, számolás helyett.

További hatások miatt, de lényegében az idő elhanyagolása következtében, nagy rendszereknél óriási méretű hatásfok csökkenés következik be (3. ábra). Mivel a számítógép processzorok működését egy-szálú végrehajtásra optimalizálják, nagy rendszerek és a neurális működést imitáló alkalmazások esetén már jól látható, hogy a jelenleg használatos módokon nem lehetséges megoldani a nagy igényű rendszerre szabott feladatot. Nem növekszik a magányos számítógép processzor teljesítőképessége, nincs újabb millió-processzoros nagy szuperszámítógép, korlátokba ütközött a mesterséges neurális hálózatok fejlesztése is [9,10]. Az újonnan megjelent, főleg GPU-alapú szuperszámítógépek valódi feladatok esetén már processzoraiknak csak töredékét tudják használni [6].

A technikai megvalósításban a „Begin Computing” és az „End computing” események egyaránt a központi órajelhez kötődnek. Emiatt a feldolgozási folyamat hosszúsága rögzített idejű; és a paradigma az átviteli idő hosszát nullának tételezi fel. Mint az ábrán látható, ha az átviteli idővel együtt a teljes idő hosszabb, mint a szinkron jel periódus ideje, az vagy kisebb használható frekvenciát követel (azaz teljesítőképesség veszteséget eredményez) vagy szinkronizálás nélküli működéshez vezet.



2. ábra: *A processzor diszperziójának fejlődése a technológia változásával. Viszonyítási alapul a mindenkori működési sebesség szolgál.*



3. ábra*: A párhuzamosítással működő (elosztott) rendszerek hatásfokának változása a processzorok száma és a nem-számítási hasznos teljesítőképesség arány függvényében*

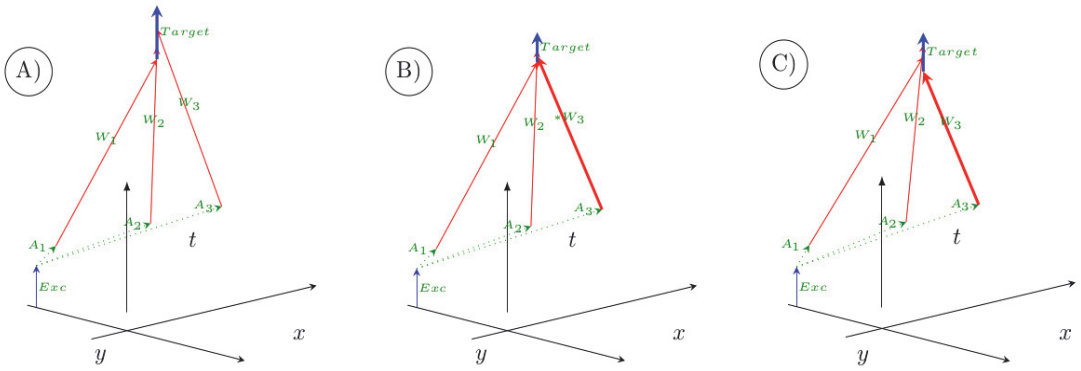
A biológiai rendszerek lényegesen nagyobbak lehetnek, mint a technológiaiak és természetes módon a végrehajtó egységek együttműködésére fektetik a hangsúlyt, nem pedig az egyes neuron képességeit fejlesztik az elképzelhető legjobbra (a neuronok „eldobhatók”: helyettesítik egymást). A számítási hatékonysághoz a biológiának másféle szinkronizálási módszert (aszinkron ön-szinkronizálás) kell használnia: a neurális impulzusok (a töltések) megérkezése szolgáltatja a „Begin Computing” jelet, amelyek közül az első ilyen impulzus indítja el a „Computing cycle” folyamatot. Hasonló módon, a neuron maga állítja elő az „End computing” jelet is, amikor elegendő töltést gyűjtött össze membránján, és a potenciál elérte a kisüléshez szükséges küszöbértéket. Ez azt is jelenti, hogy a feldolgozás hossza időben változó (a kapott jelektől és a feldolgozó egység belső állapotától függ). Azaz, a biológiai számítási rendszerek (technológiai fogalmakkal) a digitális és analóg technikák egy sajátos keverékét használják, ráadásul anatómiai mechanizmusok lehetővé teszik az „architektúra” paramétereinek menet közbeni megváltoztatását is. Túl azon, hogy a neuronok autonóm módon tudják változtatni a neurális impulzus elküldésének időpontját, az anatómiai paraméterek változtatásával a neuronok hálózata az impulzus terjedési (vezetési) sebességét is képes változtatni. Az együttműködés keretében az egyes neuronok kísérletezik ki a tanuláshoz szükséges időzítést, de azt a neuronhálózat véglegesíti anatómiai változások létrehozásával [11].

4. TANULÁS ÉS GÉPI TANULÁS

A fenti folyamat részletes vizsgálata feltárta, hogy az agyban az információ tárolása a neuronok működési idő paramétereinek változtatásával történik. A tanulás során a neuron saját hatáskörébe eső paramétereket módosít: az információt fogadó szinapszis erősségét és az azt a neuronhoz vezető axon (egy szakaszán) az információ továbbítás sebességét. A modellből következő mechanizmus működéséhez nem szükséges a neuronok közötti együttműködés, és a korábban feltételezett mechanizmusok szükségtelenné válnak; megérthető a neuron együttesek (assembly) működése is.

A 4. ábra mutatja azt a két lépésből álló módot, ahogyan neuronjaink tanulnak. Az A) ábrán látható az alapállapot, amikor egy neuron assembly három tagja küld információt ugyanazon cél-neuronnak. Az assembly tagjai függetlenek, különböző helyeken vannak, a jel indulási/érkezési ideje és sebessége is eltérhet. Amikor a beérkező jelek integrálja meghaladja a küszöb értéket (a töltés gyűjtés folyamata maga a „számítás”), a cél neuron (egy belső „Signal delivery” idő után) elküldi az eredményt (spike), ami aztán majd a „Signal transfer” idő elteltével jut el céljához. A neurális információ elküldésének időpontja attól függ, mennyi idő alatt éri el a membrán a küszöb feszültséget: a beérkezett töltéseket a membrán egy „szinaptikus erősség” faktorral szorozva integrálja, azaz a nagyobb szorzó faktorral rendelkező jelek hamarabb vezetnek kisüléshez. A B) ábra mutatja, mi történik, ha az A_3 neuron jelét fogadó szinapszis erőssége 50%-kal megnő, mivel a cél-neuron úgy találja, hogy az A_3 neuron jelét

érdemesebb nagyobb súllyal kezelni. Ekkor ez az utolsóként beérkező adalék az integrálás korábbi befejezését jelenti, azaz a cél-neuron hamarabb fog tüzelni. Ez a mechanizmus viszonylag gyors, csupán a neurotranszmitterek helyi gradiensének megváltoztatását igényli. A gradiens fenntartása viszont csak töltések pumpálásával lehetséges, ami jelentős energia befektetését igényli; ez a módszer azonban jól használható rövid távú tanulásra. Ha az A_3 neuron jelének hasznossága tartósnak bizonyul, a neuron hálózat megpróbálja ugyanezt a hatást más módon elérni. Amint azt a C) ábra mutatja, ha az A_3 tagtól származó jel sebessége megnő 10%-kal, akkor ugyanezt a hatást érhetjük el. A sebesség növelést a biológia úgy tudja megoldani, hogy növeli az axon szigetelő rétegének vastagságát, ami ugyan napokig/hetekig tartó folyamat, viszont hatása csaknem élethosszig tart. Miután azonban megnő a továbbítási sebesség, már nincs szükség a gradiens fenntartására, azaz kisebb energia befektetéssel érhető el ugyanaz a hatás. A biológia előbb megkeresi az optimális időzítést, majd minimalizálja az energia felhasználást. A biológia látja a tanulással növekvő szigetelő vastagságot (lásd [11] hivatkozásait), de a mechanizmust az időt független változóként kezelve, nem sikerült megértenie. Ennek következtében térnek el sajátságaik a technológiai rendszerekétől, bármennyire is „biomorphic” a technológiai rendszer.



4. ábra: *A neuron rövid és hosszú távú tanulásának folyamata. a) egy neuron assembly impulzusai érik el a cél-neuront. B) rövid távon, a cél-neuron módosítja szinaptikus súlyait C) hosszú távon, a rendszer megnöveli a kedvezményezett impulzus szállítási sebeségét.*

A biológia tehát két olyan módszert is használ, amelyet a jelenleg használatos technológia nem tud utánozni. Egyrészt kevert digitális/analóg számítási módszert használ, másrészt változtatni tudja az információ hordozó terjedési sebességét. Másrészt olyan szinkronizálási módszert alkalmaz, amelyik alkalmazkodik a változó sebességekhez és nagyon kis energia fogyasztást tesz lehetővé. A biológia előnyt kovácsol az lassúságból: a lassú folyamat maga tárolja az információt (érdekes párhuzam a kezdeti számítógépek késleltető művonalas tárolóival), továbbá a számítás során is kihasználja a működés lassúságát. A biológiai számító rendszerek működését csak az idő figyelembe vételével lehet megérteni. A működési modelljükben már a kiindulási ponttól eltérő technológiai és biológiai megvalósítások így csak nagyon durva közelítésként és csak nagyon egyszerű számítási rendszerekben valósítják meg ugyanazt a jelenséget. Kezdeti hasonlóságuk miatt próbálják a technológiai számítási rendszereket nagy méretű és nagy bonyolultságú biológiai számítási rendszerek imitálására is használni. Az eltérő számítási modell azonban egyrészt természetes módon korlátozza a technológiai számító rendszer teljesítőképességét, másrészt nagyobb méretű rendszerek és mélyebb szintű tanulmányozás esetén kifejezetten félrevezető eredményeket adhat. Az itt javasolt általánosított működési modell mindkét számítási rendszer esetén jól írja le a tapasztalt jelenségeket.
Az egyre jobban terjedő mesterséges intelligencia (MI) alkalmazások egyik meghökkentő sajátsága, hogy még viszonylag egyszerű feladatok megoldásának tanulása is akár hetekig tart a nagy teljesítményű számítógépes rendszereken. Ráadásul a konvergencia igen bizonytalan, a minta felismerés szinte csak

a betanításhoz használt minta elemeire ad jó eredményt. A MI esetében a számítógépes teljesítőképesség a játék szinten túl alig tud emelkedni: a teljesítőképesség pár tucat processzor után telítésbe megy [12]. Az eltérő mechanizmusok összehasonlítása alapján megállapítható az is [13], hogy a biológiai tanulás és a gépi tanulás alap feltevései és módszerei, és ezért eredményei is, is gyökeresen eltérnek.

5. ÖSSZEFOGLALÓ

Az általánosított számítási modell sikeresen írja le mind a technikai, mind a biológiai implementáció működését. A modell sikeresen megmagyarázta a nagy-skálájú technológiai implementációk tapasztalt teljesítőképesség korlátait, és értelmezni tudta a biológiai információ tárolás és feldolgozás módszerét. A modell alapján sikerült kimutatni, miért csak nevében rokon a tanulás és a gépi tanulás, ebből következően az intelligencia és a mesterséges intelligencia.

KÖSZÖNETNYILVÁNÍTÁS

A kutatás a Nemzeti Kutatási, Fejlesztési és Innovációs Alap K-136496 sz projektjének támogatásával készült.

IRODALMI HIVATKOZÁSOK

[1] Nature 571, S9 (2019): The four biggest challenges in brain simulation.
[2] Buzsáki Gy, The Brain from Inside Out, 2019, Oxford University Press, 2019, 978-0-19-090538-5
[3] J. Végh és Á. J. Berki, Do we know the operating principles of our computers better than those of our brain? CSCI'20 CSCI-ISAI: Artificial Intelligence (2020), pp 668-674, 10.1109/CSCI51800.2020.0012
[4] Végh J.: von Neumann's missing "Second Draft": what it should contain, 2020, Proc. 2020 Internat. Conf. on Computational Science and Computational Intelligence https://american-cse.org/sites/csci2020proc/pdfs/CSCI2020-6SccvdzjqC7bKupZxFmCoA/762400b260/762400b260.pdf
[5] Markov I. L.: Limits on fundamental limits to computation. Nature, 2014, 512(7513), 147-154
[6] Végh J.: Finally, how many efficiencies the supercomputers have? The J. Supercomputing, 2020, 12(76), 9430-9455
[7] Sterling P., Laughlin S.: Principles of Neural Design, 2017, MIT Press, 978-0-262-53468-0
[8] Végh J.: How Amdahl's Law limits performance of large artificial neural networks, Brain Informatics, 2019, 4(6), 1-11.
[9] Hutson M.: Core progress in AI has stalled in some fields, Science, 2020, 6494(368), 927, 10.1126/science.368.6494.927
[10] Végh J.: Which scaling rule applies to Artificial Neural Networks, Neural Computing and Applications, (2021) https://link.springer.com/article/10.1007%2Fs00521-021-06456-y
[11] Végh J., Berki Á.: Storing and processing information in technological and biological computing systems. 2021, Proc. 2021 Intern. Conf. on Computational Science and Computational Intelligence, in print. https://www.researchsquare.com/article/rs-88297/v1
[12] J. Keuper and F-J Pfreundt, Distributed Training of Deep Neural Networks: Theoretical and Practical Limits of Parallel Scalability, 2016, 2nd Workshop on Machine Learning in HPC Environments (MLHPC), 1469 – 1476, 10.1109/MLHPC.2016.006
[13] J. Végh and Á. J. Berki, Why learning and machine learning are different, Advances in Artificial Intelligence and Machine Learning, 1/2, (2021) 9.