Pozornosť ako biologicky inšpirovaný koncept pre vysvetliteľné, robustné a efektívne strojové učenie

Igor Farkaš, Barbora Cimrová, Štefan Pócoš, Iveta Bečková

Fakulta matematiky, fyziky a informatiky
Univerzita Komenského v Bratislave
{farkas,cimrova,pocos,beckova}@fmph.uniba.sk

Abstrakt

Strojové učenie zožalo vďaka hlbokým neurónovým sieťam významné úspechy, čo sa týka riešenia rozmanitých úloh ako sú klasifikácia obrázkov, jazykové úlohy, alebo rozhodovanie v hrách. Na druhej strane, známe sú nedostatky týchto metód umelej inteligencie ako napríklad nízka efektivita trénovania, netransparentnosť alebo absencia robustnosti natrénovaných modelov. V príspevku predstavíme koncept pozornosti z pohľadu psychológie a neurovedy, kde zahŕňa širšie spektrum schopností s rozmanitými mechanizmami v mozgu, ako aj z pohľadu strojového učenia, kde zavedenie pozornosti prispelo k zlepšeniu presnosti a vysvetliteľnosti modelov neurónových sietí, avšak nie efektivity trénovania a robustnosti. Dá sa teda predpokladať, že potenciál v tomto smere nebol ešte vyčerpaný.

1 Úvod

Strojové učenie, a s tým súvisiaca umelá inteligencia, sa teší v ostatnej dekáde vysokej popularite vďaka hlbokým neurónovým sieťam, pretože tie umožnili nachádzať úspešné riešenia rôznych úloh ako sú klasifikácia obrázkov, úlohy v prirodzenom jazyku či hranie hier (Schmidhuber, 2015). Výsledky týchto výpočtových modelov v mnohých prípadoch dosahujú úroveň človeka, a niekedy ho aj prekonávajú (Mnih a spol., 2015). O umelých neurónových sieťach je známe, že sú architektonicky inšpirované sieťami v mozgu človeka a ich procesy učenia pripomínajú učenie u ľudí (na príkladoch). Učenie na príkladoch sa javí ako najlepší spôsob, ako dosiahnuť zložité správanie, ktoré formálne predstavuje matematické zobrazenie vstupov na výstupy (napr. obrázkov na predikovanú kategóriu). Tento konekcionistický prístup stojí vo fundamentálnom kontraste so symbolovými modelmi na báze logiky a ontológií, kde ťažisko spočíva v expertíze dizajnéra, ktorý de facto vytvorí hotový znalostný systém. Hlboké neurónové siete učiace sa s učiteľom využívajú čisto empirický prístup (tzv. end-to-end), pri ktorom sa sieť učí úlohy priamo z pôvodných vstupov (a nerieši sa explicitne extrakcia príznakov). Neurónové siete majú veľa pozitív a je zjavné, že v ostatných rokoch hrajú prvé husle v strojovom učení, pričom významnú úlohu zohrávajú aj vo výpočtovej kognitívnej vede (Farkaš, 2011). Cieľom tohto príspevku je však poukázať na súčasné nedostatky neurónových sietí a možnosť ich odstránenia alebo aspoň zmiernenia, pomocou mechanizmov pozornosti.

2 Nedostatky umelých neurónových sietí

Najmä v súvislosti s hlbokými modelmi umelých neurónových sietí, ktoré mávajú veľa skrytých vrstiev, a teda aj astronomický počet voľných (trénovateľných) parametrov (t.j. váh medzi neurónmi), vznikli tri hlavné problémy: (1) nízka efektivita trénovania, (2) nízka transparentnosť, a (3) absencia robustnosti. Stručne si vysvetlíme každý z týchto nedostatkov.

2.1 Nízka efektivita trénovania

Neurónové siete bežne potrebujú veľa opakovaní príkladov, na ktorých sa učia. Počet potrebných opakovaní obyčajne závisí od veľkosti siete, zložitosti úlohy, ako aj ďalších faktorov (napr. hyperparametre siete). Dĺžka trénovania výrazne narastá u hlbokých modelov, ktoré súčasne potrebujú obrovské množstvo príkladov, a tie sú našťastie v súčasnosti už dostupné. Ukazuje sa, že to pomáha tomu, aby sieť predišla preučeniu (t.j. zameraniu sa na detaily v trénovacích dátach), a tým pádom slabej generalizácii (t.j. predikcii na testovacích dátach).

Existujú aj snahy, ako znížiť časovú náročnosť trénovania, lebo tá už sa stáva aj ekologickým problémom (trénovanie neurónovej siete je dosť energeticky náročné). Na druhej strane, sú známe aj metódy učenia na pár príkladoch (few-shot learning), alebo len jednom (one-shot learning), ale tie majú tiež svoje obmedzenia a predstavujú len malú časť použiteľných prístupov.

Dlhé trvanie učenia má dva hlavné dôvody. Po prvé, sieť v podstate začína pri trénovaní od nuly (čisto empirický prístup), zatiaľ čo u človeka sa predpokladajú už nejaké predispozície alebo znalosti získané z predchádzajúcich skúseností. Známe sú rôzne heuristiky ako sieť správne inicializovať, aby sa „dobre učila", ale toto problém efektívnosti nerieši. Po druhé, učenie zložitejších klasifikačných úloh predstavuje tzv. ne-

konvexný problém, ktorého dostatočne dobré riešenie (lokálne minimum chybovej funkcie) hľadáme iteratívnym spôsobom, čo je v podstate pohyb dosť naslepo vo vysokorozmernom priestore trénovateľných parametrov.

2.2 Nízka transparentnosť

Nízka transparentnosť neurónových sietí priamo vyplýva z ich architektúry a reprezentácie znalostí (pomocou reálnych čísel), ktoré sú ukryté vo váhach medzi neurónmi a aktivitách neurónov. Vzhľadom na úspešnosť týchto modelov je dôležité hľadať spôsoby, ako neurónovým sieťam lepšie porozumieť. Vysvetliteľná umelá inteligencia sa stala dôležitou vetvou výskumu, zameranou na pochopenie rôznych metód umelej inteligencie (Barredo Arrieta and others, 2020). Súčasťou tejto agendy je aj vysvetlenie toho, prečo neurónová sieť dáva na výstupe to, čo dáva (Montavon a spol., 2018). Vysvetlenia sú dôležité pre rôzne cieľové skupiny, či už expertov, užívateľov alebo pacientov, s čím súvisia aj rôzne úrovne vysvetlenia (expert rozumie aj matematickým formulám, zatiaľ čo bežný človek uprednostní vysvetlenie v prirodzenom jazyku alebo obrázkoch).

2.3 Absencia robustnosti

Absencia robustnosti natrénovaných neurónových sietí je najnovšie identifikovaný problém, ktorý bráni v nasadzovaní týchto modelov do rôznych kľúčových aplikácií. Tento problém prakticky znamená, že sieť sa dá ľahko oklamať. Samozrejme, nie hocijako, ale oveľa triviálnejšie, než človek. Geniálna myšlienka autorov (Szegedy a spol., 2014) tejto idey spočívala v návrhu takých špeciálnych vstupov pre úspešne natrénovanú sieť, pre ktoré dáva úplne zlé predikcie, častokrát s vysokou mierou presvedčenia.

Najbežnejším príkladom je klasifikácia obrázkov do tried (pričom na počte tried nezáleží). Natrénovaná sieť s vysokou presnosťou predikuje správne triedy na testovacích dátach, no napriek tomu sa dá ľahko oklamať obrázkami, ktoré boli len málo, no veľmi špecificky, pozmenené. To naznačuje, tieto vstupy sú dosť zriedkavé na to, aby sa neprejavili na testovacej chybe, no zároveň dosť bežné, aby sa dali vhodnými metódami nájsť. Skúmanie robustnosti neurónových sietí patrí medzi aktívne oblasti výskumu (Bečková a spol., 2020; Pócoš a spol., 2022).

Absencia robustnosti je asi najväčší problém, pretože otázka bezpečnosti je v modernom technologickom svete kľúčová. Nutná dĺžka trénovania sa dá zvládnuť, a v prospech toho hrá aj zrýchľujúci sa hardvér. Nízka transparentnosť sa možno nikdy nebude dať úplne prekonať, a možno riešenie bude spočívať v získaní dôveryhodnosti systému umelej inteligencie, ak bude správne fungovať (ani človek nedokáže vždy jasne zdôvodniť svoje rozhodnutie). Avšak absencia robustnosti nie je tolerovateľná, aj preto, že človek ponúka spoľahlivejšie, robustnejšie riešenie, z čoho vyplýva ďalšia potreba inšpirovať sa biologickými systémami. Pozornosť je jednou z ciest.

3 Mechanizmy pozornosti

Pozornosť je pojem známy v bežnom jazyku ale aj vo vedeckom skúmaní, najmä v psychológii a neurovede. Počiatky jeho skúmania siahajú na koniec 19. storočia, keď svetovo známy americký filozof William James ju opísal nasledovne: „Každý vie, čo je pozornosť. Je to ovládnutie mysle, v jasnej a živej forme, jedným zo zdanlivo niekoľkých súčasne možných objektov alebo myšlienkových pochodov." (James, 1890).

Odvtedy sa však chápanie pozornosti výrazne posunulo, až do takej miery, že súčasnú perspektívu niektorí autori, napríklad Hommel a spol. (2019), opisujú veľmi pesimisticky: „Nikto nevie, čo pozornosť je." V článku argumentujú, že existujú tri hlavné problémy v chápaní konceptu pozornosti u ľudí: Po prvé, koncept pozornosti vyvoláva mylné predstavy o jednom koherentnom súbore kognitívnych alebo nervových operácií, v závislosti od úrovne analýzy, ktoré všetky prispievajú k tomu, čo nazývame „pozornosť". Ako druhý problém uvádzajú to, že pozornosť sa uvádza ako problém, ktorý sa snažíme vysvetliť, ale aj ako samotné vysvetlenie (napr. pozornosť ako výsledok kapacitných obmedzení mozgu na jednej strane, verzus pozornosť ako schopnosť vysporiadať sa s týmito obmedzeniami). A po tretie, predpokladá sa, že pozornosť predstavuje konkrétny súbor kognitívnych alebo nervových operácií od iných, zdanlivo odlišných operácií, ako sú tie, ktoré súvisia s rozhodnutiami, zámermi, motiváciou, emóciami ale najmä plánovaním a vykonávaním akcií.

Je teda zložité nájsť jednotiaci konceptuálny rámec, ktorý by zastrešil všetky významy pozornosti, no niektorí autori sa o to snažia (Lindsay, 2020).

3.1 Koncept pozornosti v psychológii a neurovede

Vedecké skúmanie pozornosti má svoj pôvod v psychológii, kde dôsledné experimentovanie so správaním môže viesť k presným prejavom tendencií a vlastností pozornosti pri rôznych podmienkach. Cieľom kognitívnej vedy a kognitívnej psychológie je premeniť tieto pozorovania na modely mentálnych procesov, ktoré by mohli vytvárať takéto vzorce správania. Takýchto teoretických a výpočtových modelov bolo vytvorených veľa, s rôznymi predpokladanými základnými mechanizmami (Driver, 2001; Borji a Itti, 2013).

V oblasti neurovedy zase dostupnosť dát z neurofyziologických meraní aktivít neurónov v mozgu u zvierat, spolu s neinvazívnymi metódami merana

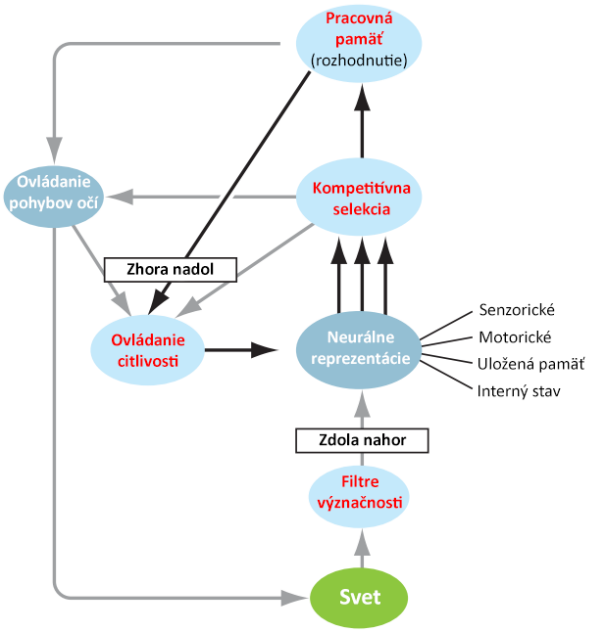
ľudskej mozgovej aktivity (ako napr. EEG, fMRI a MEG), umožnili priamo pozorovať základné neurálne koreláty kognitívnych procesov. To priamo umožňuje návrh výpočtových neurálnych modelov, ktoré dokážu replikovať empirické dáta a ponúkať tak mechanistické vysvetlenie rôznych prejavov pozornosti.

Spektrum toho, čo označujeme ako pozornosť je naozaj rozmanité. Pri nahliadnutí do učebníc kognitívnej psychológie (napr. Eysenck a Keane 2000) nachádzame rôzne príklady schopností: (1) vybrať vonkajšie udalosti pre ďalšie interné spracovanie (sústredená pozornosť); (2) ignorovať zavádzajúce informácie a/alebo irelevantné lokácie (selektívna pozornosť); (3) automaticky spracovávať nepodstatné informácie (mimovoľná pozornosť); (4) selektívne integrovať informácie patriace k jednej udalosti v rámci zmyslových modalít a medzi nimi (integrácia informácií); (5) uprednostniť spracovanie udalostí z konkrétnej lokácie (priestorová pozornosť); (6) systematicky vyhľadávať cieľovú udalosť (vizuálne vyhľadávanie); (7) vykonávať viacero úloh súčasne (rozdelená pozornosť); (8) ovládať priestorové parametre pohybov očí (selektívna pozornosť na akciu); (9) uprednostniť jeden cieľ pred ostatnými (pozornosť zameraná na cieľ); (10) uprednostniť jeden objekt, pamäťovú položku alebo vedomú reprezentáciu pred ostatnými (objektovo sústredená pozornosť); a (11) konsolidovať informácie pre neskoršie použitie a sústrediť sa na očakávanie možnej udalosti počas určitého času (trvalá pozornosť).

Každopádne, tieto rôzne typy pozornosti je možné kategorizovať, čo trochu zvyšuje ich pochopenie. Lindsay (2020) uvádza nasledovné typy pozornosti: (1) Pozornosť ako nabudenie alebo ako bdelosť, (2) senzorická pozornosť v rôznych modalitách (s dominanciou vizuálnej), zameraná na príznaky, alebo na lokáciu, (3) pozornosť pri exekutívnom riadení, a (4) interakcie pozornosti s pamäťou. S typmi pozornosti sa spájajú rôzne aspekty, ako napr. skrytá/otvorená (angl. covert/overt) pozornosť, procesy zdola nahor verzus zhora nadol.

Zrozumiteľnú schému funkčných mechanizmov pozornosti vyjadruje obr. 1. Ide o permanentnú interakciu s prostredím, v rámci ktorej sa uplatňujú štyri komponenty: (1) pracovná pamäť, (2) ovládanie senzitivity, (3) kompetitívna selekcia a (4) automatické filtrovanie význačných stimulov. Každý proces výrazne a zásadne svojím spôsobom prispieva k pozornosti, pričom vôľou riadené zameranie pozornosti zahŕňa prvé tri procesy, fungujúce zhora nadol, ktoré fungujú v rekurentnej slučke. Opačným smerom pôsobí automatická detekcia význačných stimulov.

Pracovná pamäť je špecifická forma pamäti, cez ktorú prechádza spracovanie akejkoľvek senzorickej informácie nielen z okolitého sveta, ale aj vnútorného sveta. Obsah pracovnej pamäte (s kapacitnými obmedzeniami) je tak okamžite spracovateľný v danom kontexte a stáva sa predmetom pozornosti. To, ktorá informácia získa prístup do pracovnej pamäti, je



Obr. 1. Funkčné mechanizmy pozornosti (podľa Knudsen 2007).

výsledkom kompetitívnej selekcie (súťaženia o miesto v pracovnej pamäti). Pracovná pamäť sa týka všetkých modalít a má v nich svoje špecifiká. Je široko distribuovaná v mozgu, s centrom riadenia v prefrontálnej kôre.

Ovládanie citlivosti modulované zhora nadol hrá kľúčovú úlohu pri optimalizácii informácie, ktorá je v centre pozornosti. To sa dosahuje buď zameraním pohľadu na objekt, čím sa zvyšuje priestorové rozlíšenie, alebo zvýšením pomeru signál-šum. Toto sa týka všetkých senzorických modalít, pamäti, aj vnútorných stavov. Neurálna evidencia týchto modulačných procesov pochádza najmä z (invazívnych) elektrofyziologických meraní najmä na mozgoch opíc, kde vidieť zmeny citlivosti neurónov, pričom ich zníženie, resp. zvýšenie (t.j. miery aktivity neurónu) sa potom dá interpretovať ako neurálna implementácia miery pozornosti. Informácia sa môže dostať do mozgu aj zdola nahor, bez zasahovania mechanizmov zhora nadol. Príkladom sú podnety s vysokou význačnosťou, ktoré skrátka automaticky upútajú pozornosť človeka alebo zvieraťa. Takýto prístup k pracovnej pamäti riadený vonkajšími podnetmi, bežne označovaný ako pozornosť zdola nahor, odráža účinky filtrov význačnosti (salientnosti) na mnohých úrovniach v centrálnom nervovom systéme, ktoré selektujú vlastnosti tých podnetov, ktoré budú pravdepodobne dôležité. Tieto filtre sú realizované rôznymi neurálnymi mechanizmami.

Prezentovaný pohľad navodzuje konceptualizáciu pozornosti ako inherentnej súčasti všetkých perceptuálnych a kognitívnych procesov človeka či zvieraťa, ktorá funguje v permanentnej slučke, v rámci ktorej dochádza k adaptívnemu a flexibilnému riade-

Tab. 1. Typické prístupy k mechanizmom pozornosti v strojovom učení (prevzaté z Niu a spol. (2021)).

Kritérium	Pozornosť
jemnosť pozornosti	spojtá/diskrétna globálna/lokálna
forma vstupných príznakov	na položku na lokáciu
vstupné reprezentácie	vzájomná, na seba, spoločná, hierarchická
výstupné reprezentácie	jeden výstup, viac hláv, viacrozmerná

niu obsahu práve spracovávanej informácie. Táto informácia môže pritom pochádzať z vonkajšieho prostredia (vlastnosť objektu, lokácia v priestore) alebo môže byť interne generovaná (obsah dlhodobej pamäti, pravidlo relevantné pre rozhodovanie). Chun a spol. (2011) ponúkajú taxonómiu mechanizmov pozornosti práve z tejto perspektívy. Súčasne argumentujú proti existencii jednotiaceho modelu pozornosti vzhľadom na rozmanitosť mechanizmov, ktoré stoja za jej prejavmi.

3.2 Koncept pozornosti v strojovom učení

Mechanizmy pozornosti v umelých systémoch nie sú novou záležitosťou, no napriek trom dekádam výskumu v tejto oblasti, najmä v umelých neurónových sieťach, tieto stále vo väčšine prípadov nedosahujú úroveň človeka. Mechanizmy pozornosti boli aplikované v mnohých úlohách (pozri napr. prehľad v Niu a spol. (2021)), no najvýznamnejšie dve oblasti predstavuje prirodzený jazyk (Galassi a spol., 2021), ktorý zahŕňa rôzne úlohy a počítačové videnie (Guo a spol., 2022), kde najčastejšou úlohou je klasifikácia obrazových dát.

Typická implementácia pozornostného mechanizmu spočíva v tom, že sieť sa trénuje s učiteľom (najčastejšie pomocou algoritmu spätného šírenia chyby) tak, aby dokázala správne riešiť úlohy vďaka zameraniu pozornosti na časť vstupu. Za touto naučenou schopnosťou sa skrýva iteratívne nastavenie matíc parametrov, ktoré určujú algebraické transformácie vektorov aktivít na rôznych vrstvách siete (spolu s nelinearitami neurónov). Mechanizmy pozornosti pritom počas testovania pôsobia typicky zdola nahor.

V oblasti strojového učenia tiež existujú snahy o unifikáciu mechanizmov pozornosti, možno aj preto, že spektrum existujúcich mechanizmov a použitých reprezentácií je oveľa užšie ako v mozgu. Niu a spol. (2021) vo svojom prehľade výskumu prezentovali jednotiaci model pozornosti (obr. 4 v článku), ktorý sa týka hlbokých neurónových sietí. Mechanizmy pozornosti rozdelili podľa štyroch kritérií, ako znázorňuje tab. 1.

Pri klasifikácii obrázkov model zameria pozornosť na časť obrázka, ktorá výrazne prispeje k predikcii správnej kategórie, alebo pri jazykovom preklade na relevantné slovo zdrojovej vety. Pri rozlišovaní spôsobu váhovania jednotlivých komponentov hovoríme o spojitej (alebo globálnej) pozornosti, ak ku každému komponentu prislúcha kladná váha, hoci aj veľmi malá. V prípade „ostrého" výberu komponentov hovoríme o diskrétnej pozornosti.

Ak vstupné príznaky sú vektory, hovoríme o pozornosti na položku, pričom pozornosť je smerovaná na jednotlivé vstupné vektory. Menej častým prípadom je pozornosť na lokáciu, ktorá sa využíva ak nemáme viac vstupov, no chceme nájsť nejakú informatívnu oblasť vstupu (používa sa najmä na obrázky).

Výpočet pozornostných váh závisí od zdroja informácií. Ak počítame pozornosť jedného vektora vzhľadom na druhý, ide o vzájomnú pozornosť (napr. pri preklade vety do iného jazyka). V prípade, že v tomto procese figuruje iba jeden vektor, ide o pozornosť na seba (klasifikácia obrázka). V niektorých aplikáciách je možné miešať pozornosť reprezentácií z rôznych príznakov. Tu hovoríme o spoločnej pozornosti. Napokon, pozornosť môže byť aj hierarchická, napríklad keď máme viac úrovní reprezentácie a v každej použijeme nejakú formu pozornosti.

Čo sa týka výstupu pozornostného mechanizmu, ten môžeme reprezentovať rôzne. Bežný spôsob je použiť jeden výstup (zvyčajne vektor) ktorý sumarizuje niekoľko vstupných vektorov. Rozšírením tejto myšlienky je použitie viacerých pozornostných hláv, ktoré podporujú bohatšiu reprezentáciu informácie v modeli. Zaujímavou alternatívou je aplikovanie viacrozmernej pozornosti, ktorá pri vhodnej aplikácii ponúka možnosť zmysluplnej reprezentácie vstupu viacerými možnými spôsobmi.

V aktuálnom prehľadovom článku Guo a spol. (2022) ponúkajú alternatívnu taxonómiu prístupov s využitím mechanizmov pozornosti v oblasti počítačého videnia. Tieto mechanizmy boli využité v rôznych úlohách, t.j. okrem klasifikácie obrazu pri detekcii objektov, sémantickej segmentácii, porozumenia videu, generovania obrazu, 3D videnia, ako aj multimodálnych úlohách. Pozornosť v týchto prístupoch možno algoritmicky zamerať na rôzne aspekty (lokácia, čas, farebný kanál, alebo aj vetvu v spracovaní, napr. lokálnu/globálnu), ako aj ich kombinácie. Z článku je zrejmé, že rozmanitosť modelov v ostatných rokoch narástla významne.

4 Záver

Je zrejmé, že mechanizmy pozornosti sa stali kľúčovou súčasťou modelov neurónových sietí s cieľom vylepšiť ich vlastnosti. Toto sa čiastočne podarilo, pretože pozornosť pomáha zvyšovať presnosť modelov a prispieva ich k ich vysvetliteľnosti (aj keď na pomerne nízkej úrovni). Napriek rozmanitosti prístupov pretrvávajúcim

problémom ostáva najmä absencia robustnosti. V tomto smere by mohla pomôcť ďalšia inšpirácia z biológie. Je možné, že vyriešenie tohto problému si bude vyžadovať aj iné koncepty než pozornosť.

Pri snahe o porovnanie mechanizmov pozornosti v psychológii a v strojovom učení môžeme pozorovať, že existuje čiastočný prekryv medzi oboma oblasťami, keď niektoré koncepty majú aj svoje náprotivky. Napr. Lindsay (2020) uvádza len dva príklady: otvorená vizuálna pozornosť u človeka pripomína diskrétnu priestorovú pozornosť v umelom systéme, a skrytá vizuálna pozornosť odpovedá spojitej vizuálnej pozornosti zameranej na lokáciu alebo na nejakú črtu. Jedným zo základných rozdielov je to, že v živých systémoch dominuje mechanizmus pozornosti zhora nadol, a to v rámci permanentnej slučky s prostredím. Toto absentuje pri klasifikácii obrázkov, ale aj pri iných úlohách. V každom prípade, zakomponovanie mechanizmu pozornosti sa zdá byť nutnou, a možno nepostačujúcou zložkou pri dosiahnutí vysvetliteľného a robustného umelého systému s efektívnym učením.

Poďakovanie

Tento výskum bol podporený Slovenskou spoločnosťou pre kognitívnu vedu.

Literatúra

Barredo Arrieta and others, A. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

Bečková, I., Pócoš, Š. a Farkaš, I. (2020). Computational analysis of robustness in neural network classifiers. Farkaš, I., Masulli, P. a Wermter, S. (zost.), V *Artificial Neural Networks and Machine Learning – ICANN 2020*, str. 65–76. Springer.

Borji, A. a Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207.

Chun, M., Golomb, J. a Turk-Browne, N. (2011). A taxonomy of external and internal attention. *Annual Reviews of Psychology*, 62:73–101.

Driver, J. (2001). A selective review of selective attention research from the past century. *British Journal of Psychology*, 92:53–78.

Eysenck, M. a Keane, M. (2000). *Cognitive Psychology: A Student's Handbook*. Psychology Press, Philadelphia, 7. vyd.

Farkaš, I. (2011). Konekcionizmus v náručí výpočtovej kognitívnej vedy. Kvasnička, V. a spol. (zost.), V *Umelá inteligencia a kognitívna veda III*, str. 19–62. Vydavateľstvo STU v Bratislave.

Galassi, A., Lippi, M. a Torroni, P. (2021). Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):4291–4308.

Guo, M., Xu, T., Liu, J. a spol. (2022). Attention mechanisms in computer vision: A survey. *Computational Visual Media*, 8:331–368.

Hommel, B., Chapman, C. S., Cisek, P., Neyedli, H. F., Song, J.-H., a Welsh, T. N. (2019). No one knows what attention is. *Attention, Perception & Psychophysics*, 81:2288–2303.

James, W. (1890). *Principles of Psychology*. New York: Holt.

Knudsen, E. (2007). Fundamental components of attention. *Annual Review of Neuroscience*, 30(1):57–78.

Lindsay, G. (2020). Attention in psychology, neuroscience, and machine learning. *Frontiers in Computational Neuroscience*, 14.

Mnih, V. a spol. (2015). Human-level control through deep reinforcement learning. *Nature*, 518:529–542.

Montavon, G., Samek, W. a Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15.

Niu, Z., Zhong, G. a Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62.

Pócoš, Š., Bečková, I. a Farkaš, I. (2022). Examining the proximity of adversarial examples to class manifolds in deep networks. arXiv: 2204.05764 [cs.LG].

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. a Fergus, R. (2014). Intriguing properties of neural networks. V *International Conference on Learning Representations*.