Znanstveni empiričnoraziskovalni prispevek

Eksperimentator, 4, 58–66 (2020)
Društvo študentov psihologije Slovenije, ISSN 2630-2934

Modeliranje simptomov semantične demence z globokimi nevronskimi mrežami

Lev Kiar Avberšek[1]
magistrski študent Oddelka za psihologijo, Filozofska fakulteta, Univerza v Ljubljani

Povzetek

Globoko učenje (ang. deep learning) je skupina metod strojnega učenja, pri katerih jim njihova hierarhična struktura in veliko število procesnih slojev omogočata kompleksne transformacije vhodnih podatkov in posledično učenje reprezentacij na različnih nivojih abstraktnosti. Uspeh globokih nevronskih mrež pri nalogah kategoriziranja objektov na fotografijah je povzročil velik razmah njihove uporabe za modeliranje kognitivnih procesov, ki temeljijo na vidnih reprezentacijah. Nekatere značilnosti napak globokih nevronskih mrež pri kategoriziranju objektov so podobne simptomom semantične demence – motnja, za katero je značilno zelo rigidno učenje konceptov. V raziskavi smo preučili, kako je mogoče z globokimi nevronskimi mrežami simulirati omenjene simptome. Sodelovalo je 45 zdravih udeležencev, ki so preko spletnega vprašalnika ocenili tipičnost in psevdotipičnost primerov objektov na fotografijah v 5 nadrednih kategorijah s po 10 podrednimi kategorijami, na katerih smo urili in evalvirali model. Evalvirane so bile tri končne različice modela – brez izpada nevronov, s 70 % izpadom in z 90 % izpadom. Rezultati so pokazali, da natančnost globokih nevronskih mrež pri kategoriziranju variira kot funkcija tipičnosti in izpada nevronov iz sistema. Vzorec napak pri psevdotipičnih primerih se ni ujemal z vzorcem, ki ga izkazujejo ljudje s semantično demenco, kljub temu je bilo pri različici s 70 % izpadom nevronov zaznano vedenje, ki nakazuje analogije z vzorcem napak omenjene motnje. Raziskava predstavlja eksploratoren uvod v preučevanje odnosa med simptomi semantične demence in vedenjem globokih nevronskih mrež. Dobro bi bilo, da nadaljnje raziskave preučijo dejanske (vedenjske kot tudi možganske) reprezentacije oseb s semantično demenco.

Ključne besede: globoke nevronske mreže, globoko učenje, semantična demenca, kategoriziranje, računski model

Modelling the symptoms of semantic dementia with deep neural networks

Abstract

Deep learning is a set of machine learning techniques. Hierarchical structure and many processing layers allow deep neural networks to make complex transformations of input data and learn representations on multiple levels of abstraction. The recent success of deep neural networks in the categorization of objects on images triggered a boom in their usage for modelling cognitive processes based on visual representations. Some characteristic errors that the deep neural networks make in categorization resemble symptoms of semantic dementia – a disorder in which learning of concepts is aggravated and rigid. In the present study, we have examined the potential of simulating symptoms of semantic dementia with deep neural networks. 45 healthy participants assessed typicality and pseudotypicallity of objects on images. These objects belonged to 5 superordinate categories, each of them containing 10 subordinate categories, used for training of the model and its evaluation. Three final versions of the model were evaluated; one with no dropout of neurons, one with 70% dropout and one with 90% dropout. The results showed that the accuracy of deep neural networks in categorization varies as a function of typicality and to as a function of the neuronal dropout. The pattern of errors that the model made with pseudotypical cases did not match the pattern typical for the patients with semantic dementia. However, the version with 70% dropout showed behavior that indicates some analogies with symptoms of semantic dementia. The study brought forth an exploratory introduction into studying the relationship between symptoms of semantic dementia and behavior of deep neural networks. Future research should focus on the examination of (behavioral and brain) representations of patients with semantic dementia.

Keywords: deep neural networks, deep learning, semantic dementia, categorization, computational model

[1]lev.avbersek@gmail.com

Uvod

Globoko učenje (ang. deep learning) je skupina metod strojnega učenja, ki s svojo hierarhično strukturo omogoča računskim modelom, sestavljenim iz številnih procesnih slojev, da se naučijo reprezentacij podatkov na različnih nivojih abstraktnosti. Procesni sloji izvajajo ne-linearne transformacije vhodnih podatkov; s pomočjo tega se je model zmožen naučiti zelo kompleksnih funkcij. Abstraktnost reprezentacij slojev narašča z globino modela, s tem pa omogoča odkrivanje kompleksnih struktur podatkov (Lecun, Bengio in Hinton, 2015).

Nedavni uspeh globokih nevronskih mrež (ang. deep neural networks – DNNs) pri nalogah kot so prepoznava vidnih objektov in pisanega ali branega jezika (He idr., 2016; Krizhevsky, Sutskever in Hinton, 2012; Simonyan in Zisserman, 2015; glej tudi LeCun, Bengio in Hinton, 2015 za pregled) je povzročil množično uporabo teh računskih modelov v nevroznanosti, predvsem za simulacijo raznih kognitivnih funkcij, ki temeljijo na vidnih reprezentacijah (Jozwik idr., 2017). Zaradi široke uporabe in podobnosti s procesi, ki potekajo v človeških možganih, smo v pričujoči študiji preučili ali je z globokimi nevronskimi mrežami mogoče simulirati simptome semantične demence.

Raziskave, ki se ukvarjajo z računsko nevroznanostjo ugotavljajo, da globoke nevronske mreže v svojih reprezentacijah »zaobjamejo« določen del semantičnih informacij. Več raziskovalcev je namreč ugotovilo, da globoke nevronske mreže, podobno kot ljudje, ne kategorizirajo objektov zgolj glede na njihovo fizično podobnost, temveč tudi glede na kategorično pripadnost. A. Zeman in sodelavci (2020) so npr. ugotovili, da globoke nevronske mreže v svojih reprezentacijah informacije o kategoriji objekta kodirajo neodvisno od informacij o obliki objekta. Rezultati njihove raziskave so pokazali, da so informacije o obliki objekta v večji meri prisotne v zgodnjih slojih modela. S pomočjo analize podobnosti reprezentacij (ang. Representational similarity analysis) pa so pokazali, da reprezentacije zgodnjih slojev modela najvišje korelirajo z reprezentacijami zgodnjega vidnega korteksa (V1). Nasprotno je največ informacij o kategoriji objekta kodiranih v zadnjem polno-vezanem sloju, katerega reprezentacije najvišje korelirajo z reprezentacijami anteriornega ventrotemporalnega korteksa; s predelom, kjer človeški možgani kodirajo informacije o kategorijah objektov.

Do ugotovitev o prisotnosti semantičnih informacij v globokih nevronskih mrežah so prišli tudi S. Bracci in sodelavci (2019). V svoji raziskavi so pri živih stvareh razločili obliko in kategorijo. Njihovi testni dražljaji so spadali v tri kategorije; nežive objekte, živali in nežive objekte, ki so izgledali kot živali. Ugotovili so, da so tako ljudje kot globoke nevronske mreže bolj kot na podlagi fizične podobnosti, nagnjeni h kategoriziranju glede na kategorično pripadnost objektov. Analiza podobnosti reprezentacij je pokazala, da so reprezentacije neživih objektov, ki so izgledali kot živali, skladne s kategoričnim modelom, ki predvideva, da so reprezentacije neživih objektov bolj podobne reprezentacijam objektov kot reprezentacijam živali. Korelacija reprezentacij globoke nevronske mreže s kategoričnim modelom se je višala s tem, ko so se sloji modela približevali zadnjemu, polno-vezanemu sloju. Prav v tem sloji je tudi dosegla najvišjo vrednost korelacije.

Kljub temu pa pri delovanju ljudi in globokih nevronskih mrež prihaja do nekaterih nedoslednosti, zaradi katerih lahko sklepamo, da si človeške in računalniške reprezentacije morda niso povsem podobne. Jozwik in kolegi (2017) so npr. pokazali, da globoke nevronske mreže pojasnijo več variance v človeškem presojanju podobnosti objektov, merjenim z analizo podobnosti reprezentacij, kot modeli, ki temeljijo na lastnostih objektov (torej fizični podobnosti), vendar manj variance kot kategorični modeli, ki objekte razvrščajo glede na semantično kategorijo. To nakazuje, da globokim nevronskim mrežam, kljub njihovi odličnosti pri nalogah kategoriziranja, primanjkujejo nekatere semantične informacije. Skladno s tem so Szegedy in sodelavci (2013) ugotovili, da v posameznih enotah globokih nevronskih mrež ni nikakršnih semantičnih informacij, saj med temi enotami in njihovimi naključnimi linearnimi kombinacijami ni bilo mogoče razlikovati. Avtorji (Szegedy idr., 2013) so zato sklepali, da so semantične informacije razpršene preko mnogih enot modela. Peterson in sodelavci (2016) pa so s pomočjo metode večrazsežnostnega lestvičenja (ang. multidimensional scaling) odkrili, da vzorec sodb o podobnosti objektov na fotografijah, po katerem se ravnajo globoke nevronske mreže, ni podoben vzorcu sodb o podobnosti, po katerem se ravnajo ljudje.

Tako kot velja za mnoge koncepte v računalništvu, izvirata ideja o strukturi posamičnih slojev ter ideja o celotni arhitekturi globokih nevronskih mrež iz delovanja človeških možganov. Arhitektura globokih nevronskih mrež spominja na pot vidnega sistema v možganih: lateralni genikulatni nukleus – V1 – V2 – V4 – inferotemporalni korteks (LeCun, Bengio in Hinton, 2015). Globlje razumevanje specifičnih (dis)funkcij človeških možganov in računskih modelov, ki pojasnjujejo delovanje možganov, lahko tako dosežemo s preučevanjem pojavov, pri katerih je delovanje omenjenih kompleksnih sistemov v določenem vidiku podobno. Eden izmed takšnih je semantična demenca.

Semantična demenca je selektivna in progresivna motnja semantičnega spomina, ki vpliva na vse receptivne in ekspresivne modalitete (Lambon, Ralph & Patterson, 2008; Bozeat, Lambon Ralph, Patterson, Garrard, & Hodges, 2000; Warrington, 1975). Za ljudi s semantično demenco je značilno zelo rigidno učenje konceptov. Sicer so se zmožni naučiti imena (kategorije) objektov, vendar teh, razen ob očitni vizualni podobnosti, niso zmožni posplošiti na druge primere iste kategorije (Mayberry idr., 2011). Po »Hub-and-spokes« modelu, ki razlaga strukturo in delovanje človeškega semantičnega sistema, (Patterson idr., 2007; Pobric idr., 2010b) so gradniki konceptov modalno specifične informacije (spokes) in naučena, modalno neodvisna jedrna reprezentacija (hub). Jedrna reprezentacija ustvarja multidimenzionalni semantični prostor, kamor so lahko projecirane specifične modalne informacije (Mayberry idr., 2011). Množina dimenzij je pomembna, saj omogoča kompleksne ne-linearne povezave med koncepti in mo-

dalnimi informacijami (Lambon Ralph idr. 2010; Rogers idr., 2004; Rogers in McClelland, 2004). Modalno-specifične informacije kodirajo različne regije možganskega korteksa (glede na modalnost), nevralni substrat jedra pa naj bi bilo ventrolateralno anterotemporalno območje (Lambon Ralph idr., 2010; Pobric idr., 2010b; Landon Ralph in Patterson, 2008; Patterson idr., 2007). Volumetrične analize atrofije in hipometabolizma so pokazale, da je ravno to območje pri pacientih s semantično demenco najbolj okrnjeno (Nestor, Fryer in Hodges, 2006; Galton idr., 2001; Mummery idr., 2000).

Rezultati študije Lambon Ralpha in sodelavcev (2010) so pokazali, da zaradi degradacije semantičnega sistema in posledično zmanjšane računske zmožnosti jedrne reprezentacije, pacienti s semantično demenco težje zaznajo globlje konceptualne kot površinske lastnosti. E. Mayberry in sodelavci (2011) so v svoji raziskavi preučili vpliv tipičnosti primerov kategorije na uspešnost kategoriziranja pri pacientih s semantično demenco. Primere, ki si znotraj kategorije delijo mnogo lastnosti, zaznavamo kot tipične. Obstajajo pa tudi primeri, ki površinsko (po fizičnem izgledu) niso podobni večini primerov dane kategorije. Prav tako med mnogimi kategorijami obstaja določena mera prekrivanja. Semantični sistem mora variabilnost znotraj in med kategorijami preseči in postaviti meje v multidimenzionalnem semantičnem prostoru na pravo mesto. Z degradacijo multidimenzionalnega semantičnega prostora jedrne reprezentacije ta zmožnost postaja okrnjena. To ima še posebej močan vpliv na primere, ki se nahajajo na robovih meja multidimenzionalnega semantičnega prostora. Mednje štejemo atipične primere, ki se površinsko razlikujejo od tipičnih primerov svoje kategorije (npr. noj, ki spada pod kategorijo ptice) in psevdotipične primere, ki so površinsko podobni tipičnim primerom druge kategorije (npr. metulj, ki je površinsko podoben pticam; Mayberry idr., 2011). Rezultati raziskave E. Mayberry in sodelavcev (2011) so pokazali, da premajhno posploševanje (napačna kategorizacija atipičnih primerov) in pretirano posploševanje (nepravilna kategorizacija psevdotipičnih primerov) potekata simultano in variirata kot funkcija tipičnosti, ter okrnjenosti semantičnega sistema.

Poskusi izdelave računskih modelov semantične demence v nevroznanosti že obstajajo. Model Lambon Ralpha in sodelavcev (2007), ki je konceptualno in arhitekturno podoben sodobnim globokim nevronskim mrežam, je vseboval tri sloje. Eden je kodiral strukturo vizualne podobnosti, drugi strukturo verbalnih propozicij in imen posameznih objektov, tretji, semantični sloj, pa je omogočal pretok podatkov med vizualnim in verbalnim slojem. Model je bil izurjen na 30 primerih kategorije »živečih stvari« in 30 primerih kategorije »stvari, ki jih je ustvaril človek«, tako da je asociiral vizualne reprezentacije, verbalne opise in posamična imena. Končni model je bil podvržen dvema vrstama napak; (I) odstranitvi naključno izbranih uteži in (ii) spreminjanjem vrednosti posamičnih uteži. Prva vrsta poškodbe je imela dve funkcionalni posledici za signale, ki prihajajo in odhajajo iz semantičnih enot. S tem, ko so bile odstranjene pomembne uteži, je signale izkrivila, poleg tega pa jih je »zameglila«, saj je bilo na voljo manj povezav, preko katerih bi model pošiljal informacije. Posledično so bile informacije manj učinkovite pri aktivaciji uteži v semantični enoti. Simptomatika modela z omenjeno napako je bila podobna simptomatiki oseb s semantično demenco. Prihajalo je do premajhnega in pretiranega posploševanja, ki ni bilo omejeno na specifične kategorije.

V primerjavi z omenjenim modelom, so globoke nevronske mreže veliko globje (8–150 slojev) in urjene na mnogo več podatkih (približno 1,2 milijona fotografij). Prav tako njihova primarna naloga ni semantična destilacija, temveč kategoriziranje objektov na fotografijah. Prisotnost semantičnih informacij v globokih nevronskih mrežah je tako verjetno posledica njihove velike kompleksnosti.

Kljub njihovi odločnosti pri kategoriziranju objektov, globoke nevronske mreže izkazujejo nekatere pomanjkljivosti, predvsem v primerjavi s kategoričnimi modeli. Do tega naj bi prihajalo zaradi primanjkljaja specifičnih semantičnih informacij (Jozwik idr., 2017). Poleg tega je eden izmed glavnih ciljev računalniških inženirjev, da z raznimi tehnikami omogočijo, da je model sposoben posploševati izven primerov, na katerih je bil izurjen, hkrati pa preprečevati pretirano posploševanje (Ng, 2011). Drugače rečeno, model mora biti zmožen postaviti meje v multidimenzionalnem semantičnem prostoru na ustrezno mesto. Glede na to, da je to eden izmed primarnih problemov globokih nevronskih mrež, lahko sklepamo, da bodo imele le-te težave pri primerih, ki se nahajajo na robovih multidimenzionalnega semantičnega prostora; atipčnih in psevdotipičnih primerih.

Ena izmed najpogosteje uporabljenih tehnik regularizacije globokih nevronskih mrež, tj. omogočanja posploševanja, je t.i. »dropout« metoda (izpad), ki je v svojem bistvu podobna napaki, ki je bila inducirana v modelu Lambon Ralpha in sodelavcev (2007). Globoke nevronske mreže se zaradi velikega števila uteži naučijo preveč kompleksnih reprezentacij, posledica česar je nezmožnost posploševanja. Z naključno odstranitvijo deleža uteži lahko to preprečimo (Srivastava idr., 2014). Zaradi majhnega števila uteži je imel izpad problematične posledice za model Lambon Ralpha in sodelavcev (2007), ker pa v globokih nevronskih mrežah tudi po tej intervenciji ostane veliko uteži, to omogoča ravno pravšno razmerje med premajhnim in pretiranim posploševanjem.

Na podlagi ugotovitev raziskav, ki kažejo na analogije med vedenjem ljudi in globokih nevronskih mrež, smo se odločili podrobno analizirati kje in kako se te podobnosti pojavljajo na primeru semantične demence. Če povzamemo, je simptomatika podobna v tem, da tako ljudje s semantično demenco kot globoke nevronske mreže izkazujejo določen primanjkljaj semantičnih informacij, posledica česar so značilne napake v kategorizaciji objektov. Te se kažejo v napačni kategorizaciji primerov, ki površinsko niso podobni tipičnim predstavnikom svoje kategorije (t.i. atipični primeri) in napačni kategorizaciji primerov, ki so površinsko podobni tipičnim primerom druge kategorije (psevdotipični primeri).

Vprašanje je tudi, ali ima obsežno izločevanje posamičnih uteži iz modela (izpad) podobne učinke kot jih je imelo

pri modelu Lambon Ralpha in sodelavcev (2007). Če ima odstranitev določene količine uteži učinek generalizacije (zmanjšanje kompleksnosti modela) je verjetno, da bo imela odstranitev večine uteži učinek pretiranega znižanja kompleksnosti in posledično nezmožnosti generalizacije. Predvidevamo, da bodo globoke nevronske mreže pred odstranitvijo večine uteži izkazovale premajhno posploševanje, torej bo prihajalo do sistematičnih napak pri klasifikaciji atipičnih primerov. Po odstranitvi večine uteži pa bodo izkazovale pretirano posploševanje, torej bo prihajalo do sistematičnih napak pri klasifikaciji psevdotipičnih primerov.

Metoda

Udeleženci

V raziskavi je sodelovalo 45 zdravih udeležencev, od tega 38 žensk, šest moških, en udeleženec je izbral možnost »drugo«. Povprečna starost udeležencev je bila 20,31 let (*SD* = 2,27). Sedem udeležencev je zaključilo osnovno šolo ali manj, eden poklicno ali strokovno šolo, 26 srednjo šolo in 11 višjo šolo ali univerzo.

Postopek

Za vprašalnik smo izbrali 5 kategorij živih in neživih stvari (ptice, ribe, instrumenti, orodje in vozila). Vsako izmed petih kategorij je sestavljalo 10 podrednih kategorij (10 vrst ptic, 10 vrst rib, itd.). Podredne kategorije smo izbrali tako, da je po oceni avtorja članka tipičnost kategorij variirala. Nekateri primeri podrednih kategorij so bili torej bolj, drugi pa manj tipični. Za vsako izmed petih nadrednih kategorij smo izbrali tudi dva »psevdotipična« primera. Psevdotipičen primer je tisti, ki spada v eno nadredno kategorijo, vendar površinsko spominja na drugo nadredno kategorijo (npr. leteča riba površinsko spominja na ptiča). Pri izbiri smo upoštevali, da je bil eden izmed primerov bolj, drugegi pa manj psevdo tipični. Udeležencem je bila preko spletnega vprašalnika predstavljena fotografija vsakega primera podredne kategorije, nato pa so ti morali na 7-stopenjski lestvici oceniti tipičnost vsakega izmed predstavnikov (1 – zelo netipičen, 7 – zelo tipičen).

Ocena tipičnosti je predstavljala povprečje ocen tipičnosti vseh udeležencev za dano podredno kategorijo. Glede na oceno tipičnosti so bile podredne kategorije pri vsaki izmed nadrednih kategorij razvrščene na kontinuumu od najmanj do najbolj tipične. Iz vsake od nadrednih kategorij je bilo izbranih šest najbolj tipičnih, ki so bile uporabljene pri urjenju globoke nevronske mreže. Ostale štiri so bile namenjene preverjanju natančnosti (evalvaciji) računskega modela.

Za preverjanje natančnosti računskega modela smo uporabili globoko nevronsk mrežo VGG-16 (Simonyan in Zisserman, 2013), ki vsebuje 16 slojev, od tega dva polno-vezana globoka sloja in enega namenjenega kategoriziranju. Uteži zgodnjih slojev (prvih 13) so bile pridobljene s platforme Keras (https://keras.io/) in so bile rezultat urjenja globoke nevronske mreže s pomočjo slikovne baze Imagenet (http://www.image-net.org/). Ta vsebuje približno 1,2 milijona fotografij, ki spadajo v 1000 različnih kategorij. Uteži globokih (tj. zadnjih treh slojev) pa so bile uglašene z urjenjem na petih nadrednih kategorijah (ptice, ribe, instrumenti, orodje in vozila), skozi 160 epochov, pri čemer 1 epoch pomeni, da model pregleda 19.200 fotografij iz podatkovne baze za urjenje. Vsaka nadredna kategorija je bila sestavljena iz šestih podrednih, vsaka izmed podrednih kategorij pa je vsebovala 1.300 fotografij. Med učenjem modela je bila merjena »učna natančnost« (tj. natančnost kategoriziranja slik, na katerih se je model uril) in »validacijska natančnost« (tj. natančnost, pri kategoriziranju iz istih podrednih kategorij, vendar primerov, na katerih se model ni uril). Urjene so bile tri različice modela, in sicer (1) brez odstranjevanja nevronov, (2) z odstranitvijo 70 % nevronov v zadnjih treh slojih in (3) z odstranitvijo 90 % nevronov v zadnjih treh slojih. Po urjenju smo vsako različico VGG-16 evalvirali na ostalih štirih podrednih kategorijah, ki so predstavljale primere na kontinuumu tipičnosti, ter dveh kategorijah, ki sta predstavljali kontinuum psevdotipičnosti. Vsaka izmed kategorij je vsebovala 20 primerov, merili pa smo natančnost kategoriziranja (tj. odstotek pravilnih odgovorov).

Po urjenju smo pridobili aktivacije zadnjega polno-vezanega sloja za skupino fotografij, ki so jo sestavljali vsi psevdotipični primeri in šest predstavnikov podrednih kategorij vsake nadredne kategorije. Aktivacije predstavljajo vrednosti uteži, ki se aktivirajo v določenem sloju ob določeni fotografiji. Pridobljene aktivacije smo pretvorili v korelacijsko matriko, s pomočjo katere smo izvedli večrazsežnostno lestvičenje. Vrednosti šestih podrednih kategorij, pridobljenih z večrazsežnosnim lestvičenjem so bile povprečne, tako da je bilo končno število vrednosti po 10 za vsako koordinatno os (5 psevdotipičnih in 5 tipičnih nadrednih kategorij). Namen tega je bil dobiti kvalitativen vpogled v reprezentacije modela pri različnih primerih, oziroma odnose med reprezentacijami. Da bi odnose preučili kvantitativno, smo v dvodimenzionalnem prostoru izračunali evklidsko razdaljo med položajem psevdotipičnih primerov in primerov kategorij, ki so jim ti navidezno podobni (psevdo kategorija), ter evklidsko razdaljo med položajem psevdotipičnih primerov in primerov kategorij, ki jim v resnici pripadajo (resnična kategorija). Vse statistične analize so bile izveden v programu RStudio.

Rezultati

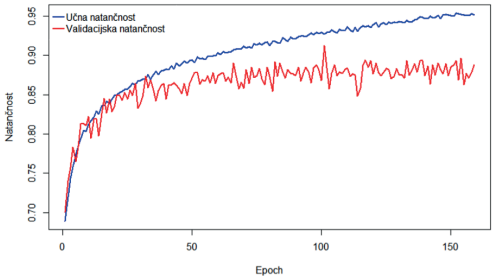
Najvišjo učno (približno 95%) in validacijsko natančnost (približno 90 %) je dosegla različica VGG-16 brez odstranjevanja nevronov (slika 1). Opazimo lahko razliko med razmerji učne in validacijske natančnosti. Medtem ko učna natančnost pri različici brez odstranjevanja nevronov (slika 1) vedno presegala validacijsko, je pri ostalih dveh različicah ravno obratno (sliki 2 in 3). Kljub temu, da je bilo v zadnjih treh slojih odstranjenih 70 % uteži (slika 2), je model dosegal visoko natančnost (približno 80% pri validacijski natančnosti). Nadaljnje odstranjevanje 20 %

nevronov pa je povzročilo velik upad v natančnosti. Validacijska natančnost pri modelu z 90 % izpadom je znašala le okoli 40% (slika 3).

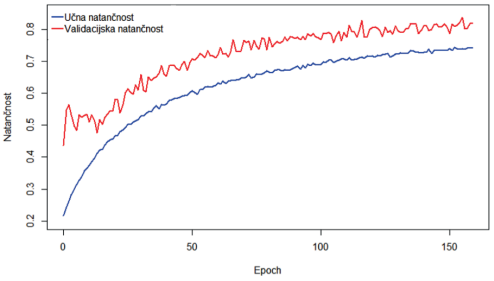
Pri različici brez odstranjevanja nevronov lahko opazimo dokaj linearen odnos med natančnostjo in oceno tipičnosti (slika 4) – z višanjem povprečne ocene tipičnosti se je večala tudi natančnost modela pri kategoriziranju. Obe različici, tako brez izpada, kot s 70 % izpadom sta za kategorije s srednjimi ocenami tipičnosti dosegla podobno natančnost. Medtem ko pri različici s 70 % izpadom še opazimo naraščanje natančnosti s tipičnostjo, je pri modelu z 90 % izpadom prihajalo do zgodnjega doseganja platoja natančnost. Zgolj primeri z najnižjo oceno tipičnosti so vplivali na njegovo natančnost.

Ker so udeleženci na kontinuumu psevdotipičnosti ocenjevali zgolj po dve kategoriji, smo te glede na povprečno oceno razdelili binarno. Primeri z višjo oceno tipičnosti so bili kategorija »psevdotipični«, primeri z nižjo oceno tipičnosti pa »psevdoatipični« (slika 5). Nepričakovano so bile vse tri različice modela bolj natančne pri kategoriziranju psevdotipičnih kot psevdoatipičnih primerov. Opazimo lahko, da je bil upad v natančnosti med VGG-16 brez izpada in VGG-16 s 70 % izpadom večji pri psevdotipičnih kot pri psevdoatipičnih primerih. Razlika med modelom brez izpada in modelom s 70 % izpadom pri natančnosti kategoriziranja psevdotipičnih primerov je približno 10 %, pri psevdoatipičnih primerih pa le približno 2 %. Izpad nevronov je pri tem modelu torej bolj vplival na zmožnost kategoriziranja psevdotipičnih primerov.

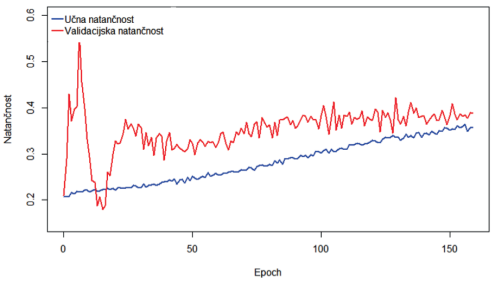
Slike 6, 7 in 8 prikazujejo rezultate večrazsežnostnega lestvičenja. Ti predstavljajo vzorec aktivacij v zadnjem polno-vezanem sloju treh modelov za vsako izmed kategorij. Bližje kot sta si v 2-dimenzionalnem prostoru dve kategoriji, bolj podoben je njun vzorec aktivacij v zadnjem polno-vezanem sloju modela. Tabela 1 prikazuje te razdalje v kvantitativni obliki. Povprečne vrednosti evklidskih razdalj kažejo, da so vsi trije modeli psevdotipične primere reprezentirali bližje njihovi resnični kot psevdo kategoriji.



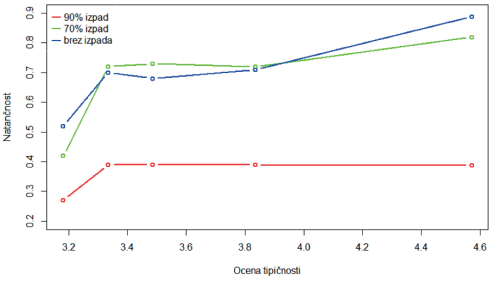
Slika 1. Urjenje VGG-16 brez izpada nevronov v zadnjih treh slojih.



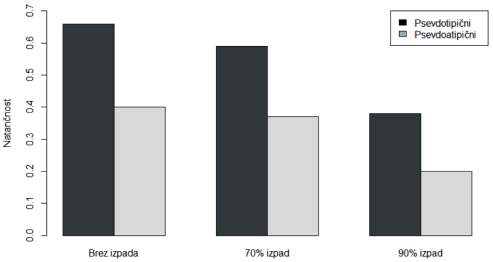
Slika 2. Urjenje VGG-16 s 70 % izpadom nevronov v zadnjih treh slojih.



Slika 3. Urjenje VGG-16 z 90 % izpadom nevronov v zadnjih treh slojih.



Slika 4. Razmerje med oceno tipičnosti in natančnostjo pri različici VGG-16 brez izpada, s 70 % in 90 % izpadom nevronov v zadnjih treh slojih.



Slika 5. Natančnost VGG-16 brez izpada, s 70% in 90% izpadom nevronov v zadnjih treh slojih pri psevdotipičnih in psevdoatipičnih primerih.

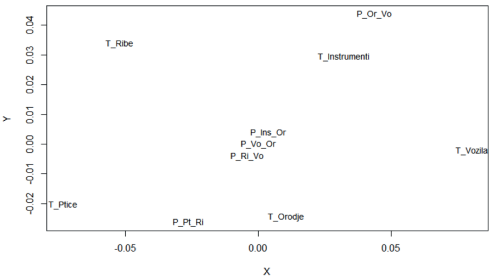
To pomeni, da je bil vzorec aktivacij v zadnjem polno-vezanem sloju modela ob prikazu psevdotipičnega primera bolj podoben vzorcu aktivacij njegove resnične kategorije (tisti, ki zares pripada) kot vzorcu aktivacij njegove psevdo kategorije (kateri tipičnim primerom je površinsko podoben). Razlika med povprečnima evklidskima razdaljama med psevdo in resnično skupino pa je bila najnižja pri modelu s 70 % izpadom.

Razprava

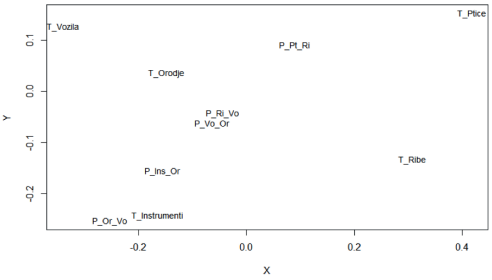
Pri razvrščanju objektov v razne kategorije imajo človeški semantični sistem in globoke nevronske mreže podobno nalogo; najti prave meje med kategorijami v multidimenzionalnemu semantičnemu prostoru, tako da ne pride do pretiranega ali premajhnega posploševanja. V primeru, da sistem tega ni zmožen storiti, je verjetnost napak pri kategoriziranju največja za primere, ki se nahajajo na robu semantičnih meja. To so atipični primeri, ki so površinsko manj podobni tipičnemu predstavniku dane kategorije kot tipični primeri in psevdotipični primeri, ki so površinsko podobni tipičnemu predstavniku druge kategorije (Mayberry idr., 2011). Napake pri kategoriziranju atipičnih in psevdotipičnih primerov so značilne za ljudi s semantično demenco. Po drugi strani se tudi pri urjenju globokih nevronskih mrež pojavljajo podobne napake. Modeli velikokrat posplošujejo pretirano ali premalo (Ng, 2011). Te podobnosti nas napeljujejo k sklepu, da so si računski procesi globokih nevronskih mrež in človeškega semantičnega sistema v določeni meri sorodni.

Rezultati urjenja VGG-16 brez odstranjevanja nevronov (slika 1) so pokazali, da je bil model nekoliko nagnjen k premajhnemu posploševanju. Učna natančnost je bila ves čas treninga višja od validacijske, kar pomeni, da je model slabše kategoriziral primere, na katerih ni bil urjen. Po drugi strani rezultati treninga drugih dveh različic (slika 2 in 3), kjer je bila odstranjena večina nevronov v globokih slojih, kažejo višjo validacijsko kot učno natančnost. Globoki nevronski mreži bolj natančno kategorizirata primere objektov na fotografijah, ki jih nista nikoli »videli«. To bi lahko bil indikator pretiranega posploševanja, vendar moramo biti pri interpretaciji takšnih rezultatov pazljivi. Kot rečeno spadajo validacijski primeri v iste kategorijo kot učni, vendar jih model za razliko od slednjih ni nikoli procesiral. Ne moremo predpostavljati, da bi se podobne napake kot pri validacijskih pojavljale tudi pri atipičnih in psevdotipičnih primerih. Poudariti je potrebno, da pretirano in premajhno posploševanje ne poteka znotraj enega modela. Za semantično demenco je namreč značilno, da simptomi pretiranega in premajhnega posploševanja konceptov potekajo simultano (Mayberry idr., 2011). V primeru globokih nevronskih mrež pa pretirano in premajhno posploševanje poteka znotraj dveh različic modela.

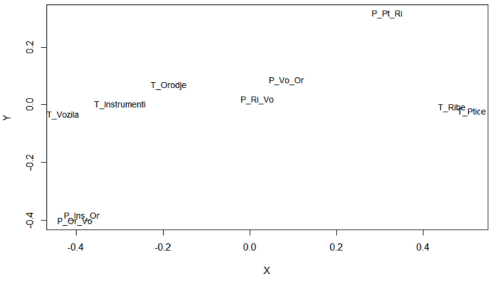
Analiza razmerja med oceno tipičnosti in natančnostjo kategoriziranja je potrdila, da podobno kot pri ljudeh s semantično demenco (Mayberry idr., 2011), tudi pri globokih nevronskih mrežah zmožnost kategoriziranja variira kot funkcija tipičnosti in funkcija okrnjenosti semaničnega



Slika 6. Multirazsežnostno lestvičenje psevdotipičnih in tipičnih primerov za zadnji polno-vezani sloj VGG-16 brez izpada. Opombe: T = tipični primeri, P = psevdotipični primeri. Pri psevdotipičnih primerih črki za prvim podčrtajem označujeta, kateri kategoriji je primer površinsko podoben (Pt = ptice, Ri = ribe, Ins = instrumenti, Or = orodje, Vo = vozila), črki za drugim podčrtajem pa označujeta, kateri kategoriji primer v resnici pripada.



Slika 7. Multirazsežnostno lestvičenje psevdotipičnih in tipičnih primerov za zadnji polno-vezani sloj VGG-16 s 70 % izpadom nevronov v zadnjih treh slojih.



Slika 8. Multirazsežnostno lestvičenje psevdotipičnih in tipičnih primerov za zadnji polno-vezani sloj VGG-16 z 90 % izpadom nevronov v zadnjih treh slojih. Opombe: T = tipični primeri, P = psevdotipični primeri. Pri psevdotipičnih primerih črki za prvim podčrtajem označujeta, kateri kategoriji je primer površinsko podoben (Pt = ptice, Ri = ribe, Ins = instrumenti, Or = orodje, Vo = vozila), črki za drugim podčrtajem pa označujeta, kateri kategoriji primer v resnici pripada.

sistema. Okrnjenost je bila do določene mere simulirana z izpadom večinskega deleža nevronov v globokih slojih. Razvidno je (slika 4), da je funkcija pri prvi različici modela, ki simulira neokrnjen sistem, dokaj linearna, medtem ko z okrnjenostjo narašča hitro doseganje platoja natančnosti. Do razlike v natančnosti je pri tretji različici modela z 90 % odstranitvijo nevronov v globokih slojih prišlo le med najmanj tipično skupino primerov in drugimi skupinami bolj tipičnih primerov (slika 4). Natančnost je hitro dosegla plato in ostala relativno nizka (< 40 %). To pomeni, da se zaradi primanjkljaja nevronov model ni bil zmožen naučiti dovolj kompleksnih reprezentacij, niti do tolikšne mere, da bi bil uspešen pri kategoriziranju tipičnih primerov. Rezultati druge različice modela, s 70 % odstranitvijo nevronov kažejo, da se je bil model zmožen naučiti se dovolj kompleksnih reprezentacij za bolj tipične primere, vendar pa je njegova natančnost eksponentno upadla pri skupini najmanj tipičnih, torej najbolj atipičnih primerov.

Razmerje med natančnostjo pri psevdotipičnih in manj psevdotipičnih (psevdoatipičnih) primerih kaže, da ni prihajalo do pretiranega posploševanja, ki je značilno za ljudi s semantično demenco. Pretirano posploševanje bi se pokazalo v napakah pri kategorizaciji psevdotipičnih primerov. Natančnost za te bi morala biti nižja kot natančnost za psevdoatipične primere. Rezultati globokih nevronskih mrež pa so bili ravno obratni. Do manj napak je prihajalo npr. pri kategoriziranju primera metulja (psevdotipičen primer), ki bi glede na površinske lastnosti lahko spadal v kategorijo ptice, kot pri primeru gosenice (psevdoatipičen primer), ki pticam ni površinsko podobna (Mayberry idr., 2011). Možna razlaga, zakaj do tega v primeru globokih nevronskih mrež ni prišlo, je izbor primerov psevdotipične in psevdoatipične skupine. Če ponazorimo s primerom; ko so morali udeleženci oceniti, koliko je tipičnemu predstavniku skupine orodje podoben traktor in koliko triangel, so višje ocene pripisali traktorju. Traktor pravzaprav površinsko (zaznavno) ni bolj podoben orodju kot je podoben vozilom, zato lahko sklepamo, da so se udeleženci pri ocenjevanju naslanjali na bolj globinske lastnosti, npr. funkcijo traktorja. Traktor je bil v tem primeru uvrščen v skupino psevdotipičnih primerov, ker pa na kontinuumu psevdotipičnosti, torej površinske podobnosti, ni bolj podoben orodju kot vozilom, globoka nevronska mreža ni imela težav s kategorizacijo fotografije v kategorijo »vozila«.

Medtem ko razmerje med natančnostjo pri kategoriziranju psevdotipičnih in psevdoatipičnih primerov ni kazalo simptomatike semantične demence, lahko opazimo, da sta imela modela brez izpada in s 70 % izpadom (slika 5) več težav pri kategoriziranju psevdotipičnih primerov kot primerov, ki so na kontinuumu tipičnosti uvrščeni najvišje (sliki 1 in 2). Najbolj tipične primere model brez izpada kategorizira s skoraj 90 % natančnostjo, model z 70 % izpadom pa s približno 80 % natančnostjo. Psevdotipične primere model brez izpada kategorizira s 66 % natančnostjo, kar je podobna natančnost kot jo model dosega pri srednjih treh primerih na kontinuumu tipičnosti. Model s 70 % izpadom pa pri psevdotipičnih primerih dosega 59 % natančnost, kar je nižje od natančnosti za srednje tri primere na kontinuumu tipičnosti, še zmeraj pa nekoliko višje kot je dosežena natančnost za najmanj tipične primere. Pri različici modela z 90 % izpadom, pri katerem gre za najvišji izpad nevronov, lahko opazimo kontradiktoren rezultat. Glede na to, da je sistem najbolj okrnjen, bi lahko sklepali, da bo razlika pri natančnosti med tipičnimi primeri na eni strani, ter atipičnimi in psevdotipičnimi primeri na drugi, največja. Rezultati kažejo ravno obratno; psevdotipični primeri so bili kategorizirani s podobno natančnostjo kot najbolj tipični primeri. Med atipičnimi in tipičnimi primeri pa je bila razlika manjša kot pri drugih dveh različicah modela. Ko smo presegli določeno mejo okrnjenosti sistema, je očitno model postal premalo kompleksen za učinkovito kategoriziranje, ne glede ne tipičnost ali psevdotipičnost.

Večrazsežnostno lestvičenje nam ponuja možnost za kvalitativen vpogled v reprezentacije globokih nevronskih mrež. S pomočjo te metode lahko ugotovimo, katere izmed primerov model reprezentira bolj podobno. Izkazalo se je, da so bile reprezentacije psevdotipičnih primerov pri vseh treh različicah modela, glede na izmerjeno evklidsko razdaljo, bližje njihovi resnični kot psevdokategoriji. Aktivacija nevronov v zadnjem polno-vezanem sloju je bila npr. pri leteči ribi bolj podobna aktivaciji nevronov pri tipični predstavnici rib kot pri tipični predstavnici ptic. Vendar lahko opazimo, da je bila najmanjša razlika v evklidski razdalji med psevdo in resnično skupino pri modelu s 70 % izpadom nevronov. Pri tej različici modela so se aktivacije nevronov, ki jih sproži psevdotipičen primer, v večji meri prekrivale s psevdo in resnično kategorijo. Razlika v aktivacij nevronov med omenjenima kategorijama je bila manjša. Posledično je prihajalo do večjega števila napak pri psevdotipičnih primerih. To potrjuje tudi rezultat zad-

Tabela 1. *Normalizirane evklidske razdalje med psevdotipičnimi primeri in njihovimi psevdo kategorijami, ter psevdotipičnimi primeri in njihovimi resničnimi kategorijami*

	VGG-16 brez izpada		VGG-16 70% izpad		VGG-16 90% izpad	
Kategorija	Psevdo	Resnična	Psevdo	Resnična	Psevdo	Resnična
Ptice	0,36	0,67	0,82	0,75	0,47	0,34
Ribe	0,60	1	0,91	0,84	0,69	0,70
Instrumenti	0,20	0,04	0	0,35	0,49	0,87
Orodje	0,85	0,55	0,73	1	0,96	0,38
Vozila	0,93	0,00	0,82	0,14	1	0,00
M	0,59	0,45	0,66	0,62	0,72	0,48

nje različice modela, kjer je bila razlika med aktivacijami nevronov med psevdo in resnično kategorijo relativno velika. Kot smo videli, temu modelu psevdotipični primeri v odnosu do tipičnih primerov ne predstavljajo težav, saj jih kategorizira z enako natančnostjo.

Rezultati vodijo k sklepu, da smo se z modelom s 70 % izpadom nevronov v globokih slojih najbolj približali nekaterim ključnim simptomom semantične demence. Simptomi semantične demence se kažejo v značilnih napakah pri kategoriziranju atipičnih in psevdotipičnih primerov. Vzporednice z globoko nevronsko mrežo s 70 % upadom lahko potegnemo zaradi več razlogov. Prvič, ta različica modela je pri urjenju kazala višjo validacijsko kot učno natančnost, kar pomeni, da je prihajalo do pretiranega posploševanja. Drugič, natančnost se je višala z oceno tipičnosti. Tretjič, model je dosegal nižjo natančnost pri psevdotipičnih primerih kot pri vseh primerih na kontinuum tipičnosti, razen najmanj tipičnih. Četrtič, evklidska razdalja med resnično in psevdo kategorijo je manjša kot pri ostalih dveh različicah, kar pomeni, da se reprezentacije omenjenih kategorij pri tej različici bolj prekrivajo.

Kljub temu ne smemo prezreti nekaterih resnih pomanjkljivosti raziskave. Prvič, model je bil treniran za kategoriziranje zgolj petih nadrednih kategorij, ki so vsebovale po šest podrednih kategorij. Globoke nevronske mreže so običajno izurjene za kategoriziranje veliko večjega števila kategorij. Avtorji VGG-16 (Simonyan in Zisserman, 2013) so model urili s pomočjo slikovne baze Imagenet, ki je vsebovala 1000 kategorij. Nevroznanstvene raziskave (npr. Bracci idr., 2019; Zeman idr., 2020), ki so ugotovile, da so globoke nevronske mreže zmožne kodirati semantične informacije, so bile izvedene na teh modelih. Prav tako se ljudje vsakodnevno soočamo z ogromnim številom kategorij, med katerimi je včasih težko postaviti ločnice. Veliko število kategorij, med katerimi lahko izbiramo, in njihova kontekstualna pogojenost, predstavljata drugo omejitev raziskave. Izbor tipičnih, atipičnih in psevdotipičnih primerov je namreč temeljil na a priori intuiciji. To pomeni, da smo izbrali kategorije, ki so se po naši oceni razlikovale glede na tipičnost ali psevdotipičnost. Sicer smo našo oceno skušali objektivizirati s tem, da smo pridobili kvantitativne ocene 45 udeležencev, vendar je bilo tako število podrednih kategorij (10 za vsako nadredno kategorijo) kot število kategorij psevdotipičnih kategorij (2 za vsako nadredno kategorijo) nizko.

Raziskava predstavlja eksploratoren uvod v raziskovanje odnosa med vedenjem globokih nevronskih mrež in simptomi semantične demence. Potrjene so bile nekatere predvidene analogije z značilnostjo napak kategoriziranja. Prihodnje raziskave se bodo morale osredotočiti na rigorozno definiranje kategorij, ter njihovih tipičnih in psevdotipičnih primerov. Poleg tega bi bilo v nadaljne študije dobro vključiti ljudi s semantično demenco in preveriti, do kolikšne mere se njihovi simptomi prekrivajo z značilnostmi globokih nevronskih mrež. Zanimivo bi bilo preučiti, kakšne so reprezentacije ljudi s semantično demenco. Potencialna metoda za to je analiza podobnosti reprezentacij (Representational Similarity Analysis – RSA; Kriegeskorte , Mur in Bandettini, 2008). S to metodo bi lahko preučili značilnosti reprezentacij atipičnih in psevdotipičnih objektov, tako na vedenjskem nivoju, kot nivoju aktivacije posameznih možganskih predelov, še posebej tistih, ki so vključeni v naloge kategoriziranja (ventrolateralno anterotemporalno območje), in jih primerjali z reprezentacijami v posameznih slojih globokih nevronskih mrež.

Reference

Bozeat, S., Lambon Ralph, M. A., Patterson, K., Garrard, P. in Hodges, J. R. (2000). Non-verbal semantic impairment in semantic dementia. *Neuropsychologia, 38*, 1207–1215.

Bracci, S., Ritchie, B. J., Kalfas, I. in Op de Beeck, H. (2019). The Ventral Visual Pathway represents animal appearance over animacy, unlike human behavior and deep neural networks. *The Journal of Neuroscience, 39*(33), 6513–6525.

Galton, C. J., Patterson, K., Graham, K., Lambon-Ralph, M. A., Williams, G., Antoun, N. idr. (2001). Differing patterns of temporal atrophy in Alzheimer's disease and semantic dementia. *Neurology, 57*, 216–225.

He, K., Zheng, X., Shaoqing, R. in Suri, J. (2016). Deep Residual Learning for Image Recognition. *Arxiv*. https://arxiv.org/abs/1512.03385

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R. in Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, *8*, 1726.

Kriegeskorte, N., Mur, M. in Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*, 4.

Krizhevsky, A., Sutskever, I. in Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. *NIPS*. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

Lambon Ralph, M. A. in Patterson, K. (2008). Generalization and differentiation in semantic memory: Insights from semantic dementia. V A. Kingstone in M. Miller (ur.), T*he year in cognitive neuroscience* (str. 61–76). New York: Annals of the NY Academy of Sciences.

Lambon Ralph, M. A., Lowe, C. in Rogers, T. T. (2007). Neural basis of category-specific semantic deficits for living things: evidence from semantic dementia, HSVE and a neural network model. *Brain, 130*(4), 1127–1137.

Lambon Ralph, M. A., Sage, K., Jones, R. W. in Mayberry, E. J. (2010). Coherent concepts are computed in the anterior temporal lobes. *Proceedings of the National Academy of Sciences, U.S.A., 107*, 2717–2722.

LeCun, Y., Bengio, Y. in Hinton, G. (2015). Deep Learning. *Nature, 521*, 436-444.

Mayberry, E. J., Sage, K. in Lambon Ralph, M. A. (2011). At the edge of semantic space: The breakdown of coherent concepts in semantic dementia is constrained by typicality and severity but not modality. *Journal of*

Cognitive Neuroscience, 23(9), 2240–2251.
Mummery, C. J., Patterson, K., Price, C. J., Ashburner, J., Frackowiak, R. S. J. in Hodges, J. R. (2000). A voxel based morphometry study of semantic dementia: The relation of temporal lobe atrophy to cognitive deficit. *Annals of Neurology, 47*, 36–45.
Nestor, P. J., Fryer, T. D. in Hodges, J. R. (2006). Declarative memory impairments in Alzheimer's disease and semantic dementia. *Neuroimage, 30*, 1010–1020.
Ng, A. (2011). Machine Learning. *Coursera*. https://www.coursera.org/learn/machine-learning
Patterson, K., Nestor, P. J. in Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience, 8*, 976–987.
Peterson, J. C., Abbott, J. T. in Griffiths, T. L. (2016). Adapting deep network features to capture psychological representations. *Arxiv*. https://arxiv.org/abs/1608.02164
Pobric, G., Jefferies, E. in Lambon Ralph, M. A. (2010). Category-specific versus category-general semantic impairment induced by transcranial magnetic stimulation. *Current Biology, 20*, 964–968.
Simonyan, K. in Zisserman, A. (2013). Very deep convolutional networks for large-scale image recognition. *Arxiv*. https://arxiv.org/abs/1409.1556
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. in Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*, 1929-1958.
Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. idr. (2013). Intriguing properties of neural networks. *Arxiv*. https://arxiv.org/abs/1312.6199
Warrington, E. K. (1975). The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology, 27*, 635–657.
Zeman, A. A., Ritchie, J. B., Bracci, S. in Op de Beeck, H. (2020). Orthogonal representations of object shape and category in deep convolutional neural networks and human visual cortex. *Scientific Reports, 10*(1).

Prispelo/Recieved: 5. 6. 2020
Sprejeto/Accepted: 29. 9. 2020