

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-XXXX-86077

Bc. Timotej Zaťko

**Uplatnenie interpretovateľnosti a
vysvetliteľnosti neurónových sietí pri
vyhodnocovaní medicínskych obrazových
dát**

Priebežná správa o riešení DP1

Študijný program: Inteligentné softvérové systémy

Študijný odbor: 18. Informatika

Miesto vypracovania: Ústav informatiky, informačných systémov a softvérového
inžinierstva (FIIT)

Vedúci práce: Ing. Martin Tamajka

máj 2020

Návrh zadania diplomovej práce

Finálna verzia do diplomovej práce¹

Študent:

Meno, priezvisko, tituly: Timotej Zaťko, Bc.

Študijný program: Inteligentné softvérové systémy

Kontakt: timi.zatko@gmail.com

Výskumník:

Meno, priezvisko, tituly: Martin Tamajka, Ing.

Projekt:

Názov: Uplatnenie interpretateľnosti a vysvetliteľnosti neurónových sietí pri vyhodnocovaní medicínskych obrazových dát

Názov v angličtine: Application of interpretability and explainability of neural networks in the evaluation of medical images

Miesto vypracovania: Ústav počítačového inžinierstva a aplikovanej informatiky, FIIT STU

Oblast problematiky: počítačové videnie, hlboké neurónové siete, analýza medicínskych obrazových dát, vysvetliteľnosť a interpretateľnosť

Text návrhu zadania²

Umelá inteligencia a špeciálne hlboké neurónové siete sa za posledných desať rokov stali jedným z dominantných výskumných problémov, pričom v mnohých úlohách významne prekonávajú doterajšie prístupy. Zatiaľ čo vo výskume je prípustná istá miera neistoty alebo nepresnosti, v oblastiach ako je medicína je žiaduce, aby algoritmy umelej inteligencie poskytovali účinné mechanizmy kontroly správnosti predikcie. V medicínskej oblasti sa už umelá inteligencia uplatnila pri výrobe liekov, monitorovaní zdravotného stavu, chirurgických zákrokov a aj pri odhalovaní chorôb. Práve pri odhalovaní chorôb, akými sú napríklad rakovina plúc, rakovina kože alebo Alzheimerova choroba, sa využívajú hlboké neurónové siete za účelom získania klinicky relevantných informácií z medicínskych obrazových dát.

Analyzujte doménu medicínskych obrazových dát a súčasný stav problematiky interpretateľnosti a vysvetliteľnosti predikcie neurónovej siete. Navrhnite metódu na detekciu nesprávnych rozhodnutí alebo odhadovanie miery správnosti modelu neurónovej siete pri vyhodnocovaní medicínskych obrazových dát. Navrhnutú metódu implementujte a dosiahnuté výsledky vyhodnote na dostatočne veľkej dátovej množine. Dosiahnuté výsledky porovnajte s inými súčasnými riešeniami.

¹ Vytlačiť obojstranne na jeden list papiera

² 150-200 slov (1200-1700 znakov), ktoré opisujú výskumný problém v kontexte súčasného stavu vrátane motivácie a smerov riešenia

Literatúra³

- MONTAVON, Grégoire, Wojciech SAMEK and Klaus-Robert MÜLLER, 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* [online]. 2018, vol. 73, pp. 1-15.
- STURM, Irene, Sebastian LAPUSCHKIN, Wojciech SAMEK and Klaus-Robert MÜLLER, 2016. Interpretable deep neural networks for single-trial EEG classification. *Journal of Neuroscience Methods* [online]. 2016, vol. 274, pp. 141-145.

Vyššie je uvedený návrh diplomového projektu, ktorý vypracoval(a) Bc. Timotej Zaťko, konzultoval(a) a osvojil(a) si ho Ing. Martin Tamajka a súhlasí, že bude takýto projekt viesť v prípade, že bude pridelený tomuto študentovi.

V Bratislave dňa 1.6.2020

Podpis študenta

Podpis výskumníka

Vyjadrenie garanta predmetov Diplomový projekt I, II, III

Návrh zadania schválený: áno / nie⁴

Dňa:

Podpis garanta predmetov

³ 2 vedecké zdroje, každý v samostatnej rubrike a s údajmi zodpovedajúcimi bibliografickým odkazom podľa normy STN ISO 690, ktoré sa viažu k téme zadania a preukazujú výskumnú povahu problému a jeho aktuálnosť (uveďte všetky potrebné údaje na identifikáciu zdroja, pričom uprednostnite vedecké príspevky v časopisoch a medzinárodných konferenciách)

⁴ Nehodiace sa prečiarknite

Čestne vyhlasujem, že som túto prácu vypracoval samostatne, na základe konzultácií a s použitím uvedenej literatúry.

V Bratislave, 6. máj 2020

Timotej Zatko

Anotácia

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií
Študijný program: Inteligentné softvérové systémy

Autor: Bc. Timotej Zaťko

Diplomová práca: Uplatnenie interpretovateľnosti a vysvetliteľnosti neurónových sietí pri vyhodnocovaní medicínskych obrazových dát

Vedúci diplomového projektu: Ing. Martin Tamajka

máj 2019

Súčasný vplyv umelej inteligencie na spoločnosť je nespochybniteľný. Využitie si už našla v rôznych oblastiach našich životov či už je to v smartfónoch pri odomykaní tvárou alebo najnovšie pri kontrole používania ochranného rúška pri vstupe do obchodov. Umelá inteligencia sa postupne dostáva do oblasti medicíny, kde má potenciál zachraňovať životy. Aby, teda mohla byť spoľahlivým pomocníkom doktorov pri diagnóze ochorení je nevyhnutné, aby jej rozhodnutie bolo možné vysvetliť.

V oblasti medicíny je možné použitie neurónových sietí, pretože dokážu veľmi dobre pracovať s obrazovými dátami, a tak sa dajú využiť napríklad pri diagnóze Alzheimerovej choroby z rádiologických snímkov. Ich problémom však je, že sa správajú ako "čierna skrinka" čo bráni v tom, aby sa dostali do bežnej praxe.

V tejto práci sme navrhli nový spôsob interpretovania neurónových sietí, navrhli sme spôsob porovnania s existujúcimi prístupmi a overenia pri vysvetľovaní roz hodnutí neurónovej siete deketujúcich Alzheimerovu chorobu z MRI snímkov.

Annotation

Slovak University of Technology Bratislava
Faculty of Informatics and Information Technologies
Degree Course: Intelligent Software Systems

Author: Bc. Timotej Zatko

Diploma's Thesis: Application of interpretability and explainability of neural networks in the evaluation of medical images

Supervisor: Ing. Martin Tamajka

2019, May

The current impact of artificial intelligence on society is undeniable. It has already been used in various areas of our lives, whether it is in smartphones for unlocking via face recognition or, most recently, for controlling the use of protective masks when entering shops or groceries. Artificial intelligence is entering the field of medicine, where it has the potential to save lives. Thus, in order to be a reliable assistant to doctors for example in the diagnosis of the disease, it is necessary that its decisions can be explained.

In the field of medicine, the usage of neural networks is possible, because they can work very well with image data, and so they can be used, for example, in the diagnosis of Alzheimer's disease from radiological images. However, their problem is that they behave like a "black box" which prevents them from getting into common practice.

In this work, we proposed a novel method of interpreting neural networks, we proposed a process of comparison with existing approaches and verification in explaining the neural network decisions detecting Alzheimer disease from MRI images.

Pod'akovanie

Ďakujem môjmu školiteľovi Ing. Martinovi Tamajkovi za odbornú pomoc a vedenie pri tvorbe tejto práce.

Obsah

1	Úvod	5
2	Analýza	7
2.1	Alzheimerova choroba	7
2.1.1	Diagnostika Alzheimerovej choroby	8
2.1.2	Biologické ukazovatele	8
2.1.3	Obrazové a rádiologické ukazovatele	9
2.2	Neurónové siete	10
2.2.1	Neurón	12
2.2.2	Dopredné neurónové siete	13
2.2.3	Konvolučné neurónové siete	13
2.2.4	Interpretovanie neurónovej siete	15
2.2.5	Vysvetľovanie predikcie neurónovej siete	16
2.2.5.1	Analýza senzitivity	18
2.2.5.2	LRP (angl. layer-wiser relevance propagation) . . .	19
2.2.5.3	RISE - Randomized Input Sampling for Explanation	21
2.3	Využitie neurónových sietí pri odhalovaní Alzheimerovej choroby .	22
2.3.1	Vysvetľovanie rozhodnutí neurónových sietí detegujúcich Alzheimerovu chorobu	24
2.4	Spracovanie obrazu	25
2.4.1	Inpainting	25
2.5	Zhrnutie	26
3	Ciele práce	29

Obsah

3.1	Vytvorenie novej, alebo vylepšenie existujúcej metódy pre vysvetľovanie rozhodnutí neurónových sietí	29
3.2	Využitie vytvorenej metódy na určenie miery správnosti modelu neurónovej siete detekujúcej Alzheimerovu chorobu	30
4	Návrh riešenia	31
4.1	RISEI - Randomized Input Sampling for Explanation with Inpainting	32
4.2	Overenie riešenia	34
4.2.1	Dátová sada	34
4.2.2	Experimenty	35
4.2.2.1	Určenie kvality metódy vysvetľovania rozhodnutí modelu	35
4.2.2.2	Určenie správnosti modelu	35
4.3	Záver	35
	Literatúra	37

Dodatok A Plán práce

A.1	Zimný semester
A.2	Letný semester

Zoznam použitých skratiek

AD angl. Alzheimier disease (Alzheimerova choroba) – používa sa na označenie pacientov trpiacich Alzheimerovou chorobou

AUC angl. area under curve (plocha pod krivkou)

CN angl. cognitive normal (kognitívne zdravý) – používa sa na označenie pacientov bez kognitívneho poškodenia (tj. zdravých jedincov)

MCI angl. mild cognitive impairment (mierne kognitívne poškodenie) – používa sa na označenie pacientov s miernym kognitívnym poškodením

MRI angl. magnetic resonance imaging (magnetická rezonancia) – je spôsob tvorby rádiologických snímkov ľudského tela, používa sa pri diagnostike Alzheimerovej choroby

PET angl. positron emission tomography (pozitrónová emisná tomografia) – je spôsob tvorby rádiologických snímkov ľudského tela, používa sa pri diagnostike Alzheimerovej choroby

Obsah

1. Úvod

Umelá inteligencia sa už dávno stala súčasťou nášho každodenného života. Prichádzame s ňou do kontaktu neustále, keď odomykáme telefón vlastnou tvárou alebo keď pomocou prekladača prekladáme text to iného jazyka. Jej využitie je tiež rozšírené v oblasti medicíny, kde má potenciál zachraňovať životy. Využíva sa pri výrobe liekov, monitorovaní zdravia, analýze zdravotných plánov, chirurgických zákrokov a aj pri odhaľovaní chorôb [1]. Práve pri odhaľovaní chorôb sa častokrát využívajú hlboké neurónové siete, a to napríklad pri detekcii rakoviny kože, rakoviny pľúc alebo Alzheimerovej choroby z obrazových dát.

Neurónovým sieťam sa už podarilo dosiahnuť také dobré výsledky, že sú porovnatelné s expertmi v medicínskej oblasti. Ich problémom však je, že sa správajú ako "čierna skrinka", čo v oblasti medicíny nie je žiadúce. Preto je nevyhnutné, aby boli rozhodnutia neurónovej siete interpretovateľné a pacient s lekárom vedeli, na základe čoho sa neúronová sieť rozhodla. Lekári by si mali svoje rozhodnutia vedieť obhájiť. Aby sa teda neurónové siete mohli stať bežným pomocníkom lekárov, je vysvetliteľnosť ich rozhodnutí dôležitá.

2. Analýza

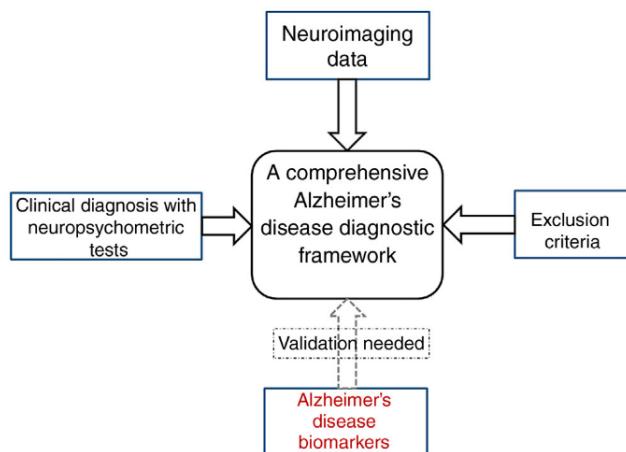
2.1 Alzheimerova choroba

Alzheimerova choroba je najčastejšou príčinou demencie. Prvotné príznaky tejto choroby sú zhoršenie pamäti, zabúdanie nedávnych udalostí, mien, neschopnosť rozoznávať známe miesta či orientovať sa v čase [2]. Jej priebeh sa vyznačuje postupným poklesom kognitívnych funkcií, postupným zhoršením pamäte, myslenia, rozprávania a schopnosti učenia sa [3]. Najčastejšie sa vyskytuje u ľudí starších ako 65 rokov, s pravdepodobnosťou výskytu až 50% po dovršení 85 rokov života [3]. S narastajúcim vekom človeka sa zvyšuje pravdepodobnosť ochorenia. Pravdepodobnosť ochorenia zvyšujú tak tiež úrazy hlavy, poruchy prekrvenia mozgu, pozitívna rodinná anamnéza či vzdelanie (protože ľudia s nižším vzdelaním majú väčšie riziko rozvoja tohto ochorenia) [2]. Toto ochorenie sa vyskytuje častejšie u žien ako u mužov, v pomere 2:1 [4].

Alzheimerova choroba nie je "iba" o strate pamäti, ale aj šiestou najčastejšou príčinou smrti v USA [5]. Medzi rokmi 2000 až 2017 sa počet úmrtí v USA viac ako zdvojnásobil [5]. Ľudia starší ako 65 rokov ktorým bola diagnostikovaná táto choroba sa v priemere dožívajú 4 až 8 rokov po jej diagnóze [5].

2.1.1 Diagnostika Alzheimerovej choroby

Alzheimerova býva diagnostikovaná kombináciou viacerých ukazovateľov. Pri určovaní diagnózy sa používajú neuropsychometrické (kognitívne) testy, rádiologické snímky (angl. neuroimaging data), biologické ukazovatele a špecifické kritériá, na základe ktorých je možné vylúčenie iných chorôb u pacienta z jeho história vývoja ochorenia [4]. T. Khan zadefinoval tieto ukazovatele do tzv. komplexného rámca pre diagnózu Alzheimerovej choroby (Obr. 2.1). V súčasnosti sa v tejto oblasti skúmajú biologické ukazovateľe (ich identifikácia a použitie), keďže používanie (a teda aj vytvorenie) rádiologických ukazovateľov je drahé [4] (vyžaduje si to zaškolený personál a vybavenie). Biologické ukazovateľe zatiaľ nie sú dostatočne spoľahlivé [4].



Obr. 2.1: **Komplexný rámec pre diagnózu alzheimerovej choroby.** Pozostáva z neuropsychometrických testov, rádiologických snímok (z PET, MRI...), biologických ukazovateľov (napr. úrovne hladín určitých proteínov v krvnej plazme) Alzheimerovej choroby a kritérií vylúčenia iných neurologických chorôb.[4]

2.1.2 Biologické ukazovatele

Biologické ukazovatele (angl. biomarkers) sú merateľné biologické ukazovatele slúžiace na detekciu prítomnosti choroby. National Institute of Health definguje bio-

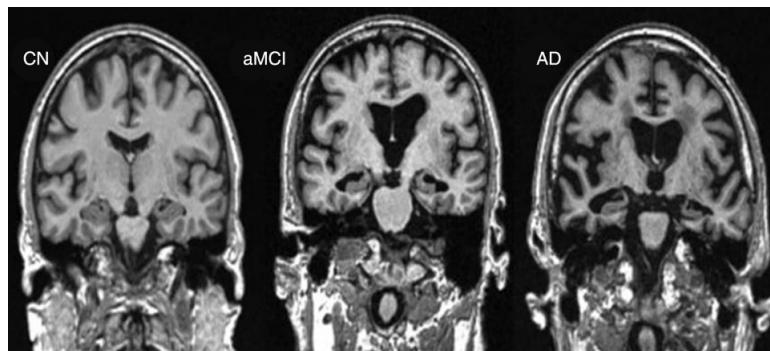
logický ukazovateľ ako indikátor určitého objektívneho merania a hodnotenia biologického procesu, patogénneho procesu alebo farmakologického hodnotenia terapeutickej účinnosti [6]. Alzheimerova choroba môže byť identifikovaná sledovaním týchto biologických ukazovateľov napríklad v krvnej plazme [4] alebo v mozgovo-miechovej tekutine (angl. cerebrospinal fluid) (ako úrovne hladín proteínov P-tau and A β 42) [4] (angl. cerebrospinal fluid).

2.1.3 Obrazové a rádiologické ukazovatele

Identifikovanie Alzheimerovej choroby je v súčasnosti možné aj z rádiologických snímkov. Tvorba rádiologických snímkov je v súčasnosti možná pomocou techník akými sú počítačová tomografia s jednou fotónovou emisiou (angl. single-photon emission computed tomography - SPECT), pozitrónová emisná tomografia (angl. positron emission tomography PET), počítačová tomografia (angl. computed tomography - CT), magnetická rezonancia (magnetic resonance imaging - MRI) a magnetická rezonančná spektroskopia (angl. magnetic resonance spectroscopy - MRS) [4].

Snímky z magnetickej rezonancie (MRI) dokážu zachytiť odumieranie tkaniva (na základe biologických procesov), ktoré sa odohráva v rôznych častiach mozgu [4]. Príklad takéhoto snímku sa nachádza na obrázku 2.2.

Snímky z pozitrónovej emisnej tomografie (PET) dokážu zachytiť pokles mozgovej aktivity, ktorá je u pacientov s Alzheimerovou chorobou nižšia. Mozgová aktivita odráža úroveň metabolizmu glukózy v mozgu. Na miestach v mozgu, ktoré sú touto chorobou postihnuté, je úroveň metabolizmu glukózy nižšia. Tento jav je znázornený na obrázku 2.3.



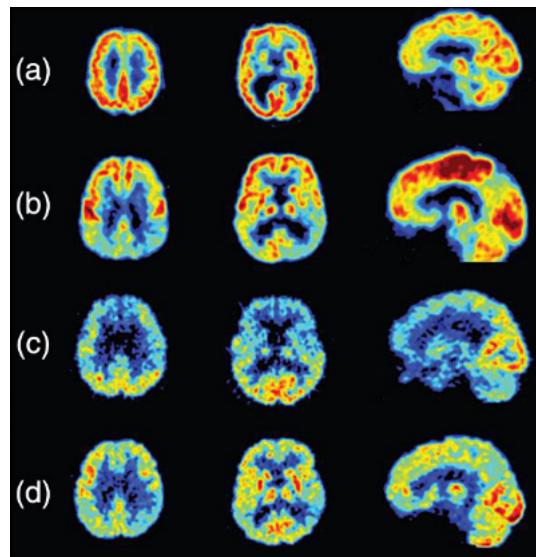
Obr. 2.2: **Typické odumieranie mozgového tkaniva zachytené magnetickou rezonanciou.** Obrázok zľava, označený ako CN (angl. cognitive normal), reprezentuje kognitívne normálneho jedinca. Obrázok v strede, označený ako aMCI (angl. amnestic mild cognitive impairment) reprezentuje jedinca s miernym kognitívnym poškodením - na obrázku je zreteľný úbytok mozgového tkaniva (šedá farba) najmä v strede mozgu (ale aj na jeho okrajoch) oproti kognitívne normálnemu jedincovi. Posledný obrázok označený ako AD (angl. Alzheimer's disease) reprezentuje jedinca s Alzheimerovou chorobou - na obrázku je zreteľný značný úbudok mozgového tkaniva. [4]

2.2 Neurónové siete

Neurónové siete patria medzi obľúbené techniky strojového učenia. Špeciálnou kategóriou sú hlboké neurónové siete (často označované skratkou DNN od angl. deep neural network), ktoré sa oproti obyčajným neurónovým sieťam odlišujú počtom vrstiev. Hlbokým neurónovým sieťam sa doteraz podarilo dosiahnuť v mnohých úlohách výnimočné výsledky, v ktorých častokrát už dokázali poraziť človeka. V našej oblasti obrazových rádiologických dát sa používajú najmä konvolučné neurónové siete.

Haykin et al. [7] definujú neurónovú sieť nasledovne:

Neurónová sieť je veľký paralelný distribuovaný procesor tvorený jednoduchými procesorovými jednotkami, ktorý má prirodzený sklon ukladať poznatky a sprístupňovať ich na použitie. Ľudskému mozgu sa podobá v dvoch aspektoch:



Obr. 2.3: **Snímky normálneho mozgu a mozgu postihnutého Alzheimerovou chorobou z pozitrónovej emisnej tomografie (PET).** [4] Na obrázkoch je viditeľná úroveň metabolizmu glukózy, u pacientov s Alzheimerovou chorobou je táto úroveň nižšia (žltá a modrá farba na obrázkoch). (a) Mozog kognitívne zdravého jedinca - vyznačuje sa vyššou mozgovou aktivitou. (b) Mozog vyznačujúci symptómy Alzheimerovej choroby - je vidieť nižšiu aktivitu v niektorých častiach mozgu oproti kognitívne zdravému jedincovi. (c) Mozog postihnutý frontotemporálnou demenciou (angl. frontotemporal dementia), tiež sa vyznačuje nižšou mozgovou aktivitou. (d) Mozog postihnutý Alzheimerovou chorobou.

1. Neurónová sieť získava vedomosti zo svojho prostredia prostredníctvom procesu učenia.
2. Na uchovanie získaných poznatkov sa používajú prepojenia medzi jednotlivými neurónami.

Neurónové siete sú teda inšpirované fungovaním mozgu človeka, keďže napodobňujú jeho fungovanie.

2.2.1 Neurón

Neurón (Obr. 2.4) je základnou stavebnou jednotkou neurónových sietí. Matematicky sa dá zapísat ako [7]:

$$y_k = \varphi(b_k + \sum_{j=1}^m w_{kj} \cdot x_j) \quad (2.1)$$

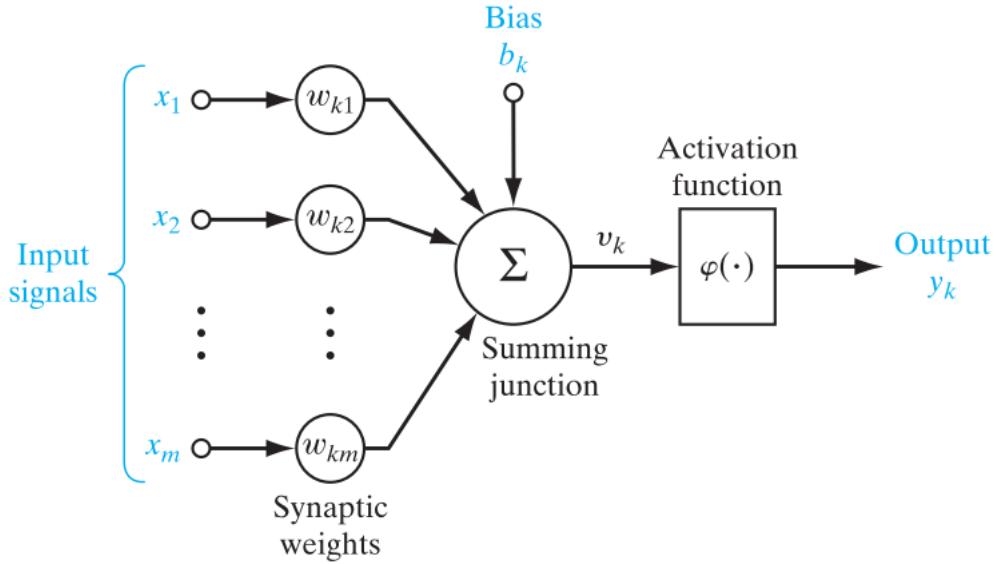
Kde:

- x_1, x_2, \dots, x_m sú vstupné signály
- $w_{k1}, w_{k2}, \dots, w_{km}$ sú váhy neurónu k
- b_k je sklon neurónu k
- $\varphi(\dots)$ je aktivačná funkcia
- y_k je výsupný signál neurónu k

Parametrami, ktoré sa počas trénoania neurónovej siete menia sú váhy w_{kj} a sklon b_k , tieto parametre sú takzvané trénovateľné parametre. Tieto parametre sa upravujú pri spätej propagácii (angl. backpropagation), kedy sa minimalizuje chybová funkcia (angl. loss function).

V neurónových sietiach s viac vrstvami sa stávajú výstupné signály y neurónov jednej vrstvy vstupom x do ďalšej.

Aktivačná funkcia zabezpečuje nelinearitu neurónu, medzi najpoužívanejšie aktivačné funkcie patria Sigmoid ($S(x) = \frac{1}{1+e^{-x}}$), Tanh alebo ReLU ($ReLU(x) = \max(0, x)$). Jednotlivé neuróny si môžeme predstaviť ako nelineárne funkcie, ktorých spojením do viac vrstiev dokážu skladať ešte zložitejšie a komplexnejšie funkcie.



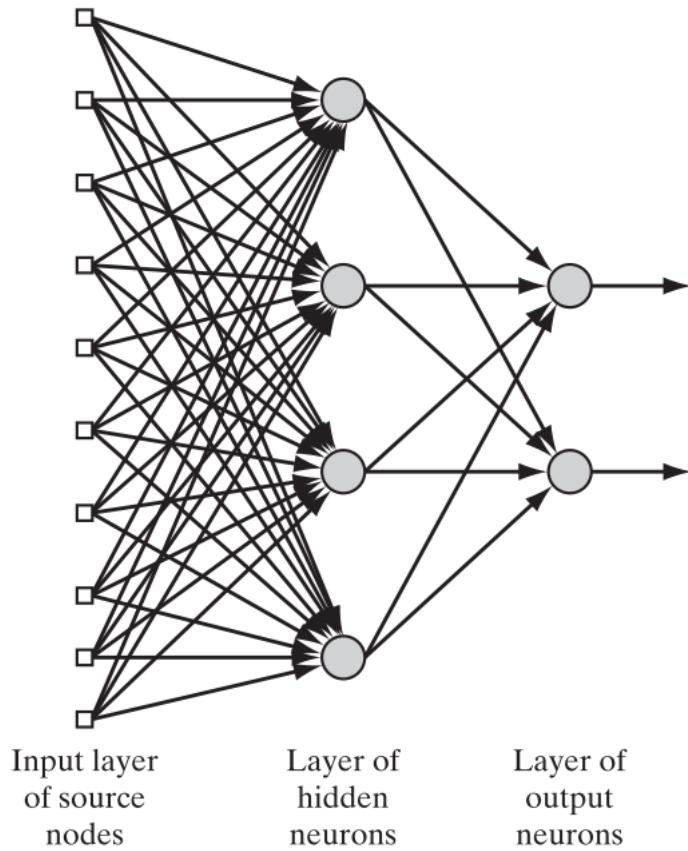
Obr. 2.4: **Model neurónu.** [7] Neurón sa skladá zo vstupných signálov a váh, ktoré sú na tieto signály aplikované, sklon (b_k - angl. bias) a aktivačnej funkcie, ktorá zabezpečuje nelinearitu. Vzorec 2.1 matematicky popisuje správanie neurónu.

2.2.2 Dopredné neurónové siete

Dopredné neurónové siete (Obr. 2.5) sú jednou z mnoha architektúr neurónových sietí. V dopredných neurónových sieťach výstupný signál z jednej vrstvy nemôže byť vstupným signálom do jej predošej vrstvy. Signál je prenášaný iba v jednom smere – dopredu. Dopredné neurónové siete sa môžu skladať z viacerých vrstiev. Základom je vstupná a výstupná vrstva a ľubovoľný počet skrytých vrstiev. Ich počet nie je limitovaný, avšak v hlbokých neurónových sieťach (tj. sieťach s veľkým početom skrytých vrstiev) môže nastáť problém miznúceho gradientu.

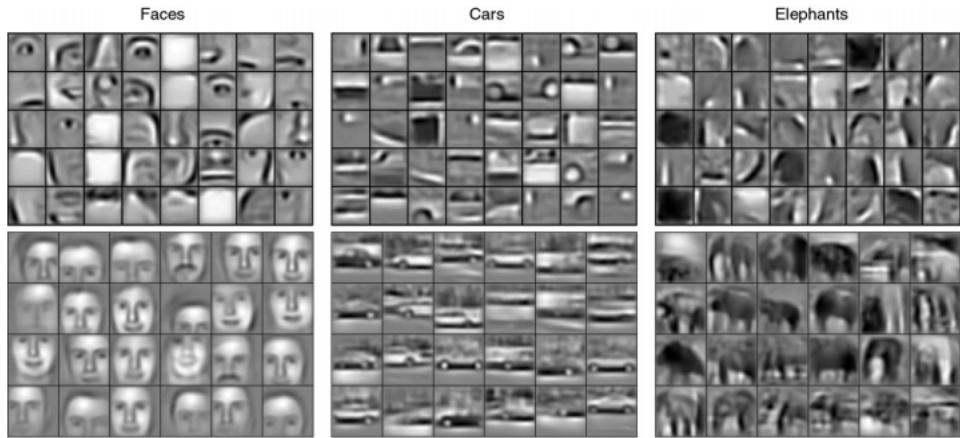
2.2.3 Konvolučné neurónové siete

Konvolučné neurónové siete sa používajú prevažne v doméne obrazových dát. Tieto siete majú schopnosť naučiť sa rozpoznávať špecifické štruktúry/tvary z obrázka. Toto dokážu pomocou takzvaných konvolučných filtrov, ktoré sa v nižších vrstvách

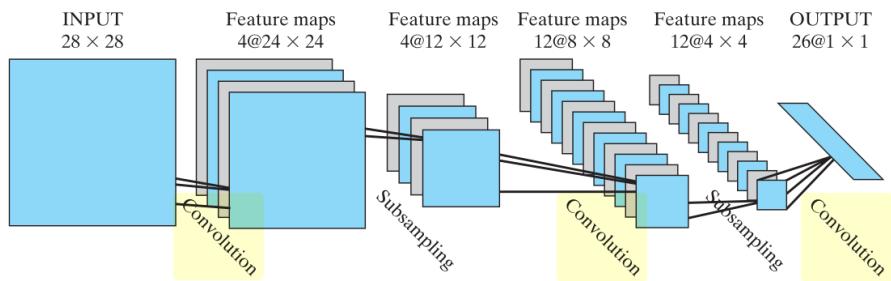


Obr. 2.5: **Model doprednej neurónovej siete.** [7] Dopredné neurónové siete sa skladajú zo vstupnej vrstvy, skrytých vrstiev a výstupnej vrstvy. Ked' hovoríme o počte vrstiev vstupnú vrstvu nepočítame. Neurónová sieť na obrázku má teda dve vrstvy.

naučia rozoznávať jednoduchšie tvary, akými sú napríklad obrysy alebo hrany (Obr. 2.6). V tých vyšších vrstvách sú to zložitejšie štruktúry akými môžu byť celé objekty v závislosti od typu úlohy na ktorú boli trénované. Ak bola neurónová sieť trénovaná napríklad na klasifikáciu zvierat, môže tým objektom byť pes alebo morča, v prípade ak je úlohou neurónovej siete detekcia Alzheimerovej choroby možu týmito objektami byť niektoré väčšie časti mozgu (napr. hippocampus).



Obr. 2.6: Vizualizácia druhej (hore) a tretej vrstvy (dole) konvolučných neurónových sietí naučených na špecifické kategórie objektov (tváre, autá a slony). [8] Nižšie vrstvy rozoznávajú jednoduchšie štruktúry zatiaľ čo vyššie už dokážu rozoznať aj tie zložitejšie.



Obr. 2.7

2.2.4 Interpretovanie neurónovej siete

Montavon; Samek; Müller (2018) definujú interpretovanie ako mapovanie abstraktného konceptu (napríklad predikovanej triedy) do domény, ktorej človek dokáže porozumieť. Ako príklad domény, ktorá je interpretovateľná uvádzajú obrázky (pole pixelov) alebo text (sekvencia slov) [9]. Medzi domény, ktoré nie sú interpretovateľné zaraďujú napríklad latentné vektorové reprezentácie slov (angl. word embeddings) alebo iné abstraktné vektorové reprezentácie [9]. Na rozdiel od vstupných dát do neurónovej siete, ktoré sú zvyčajne interpretovateľné, neuróny na výstupnej

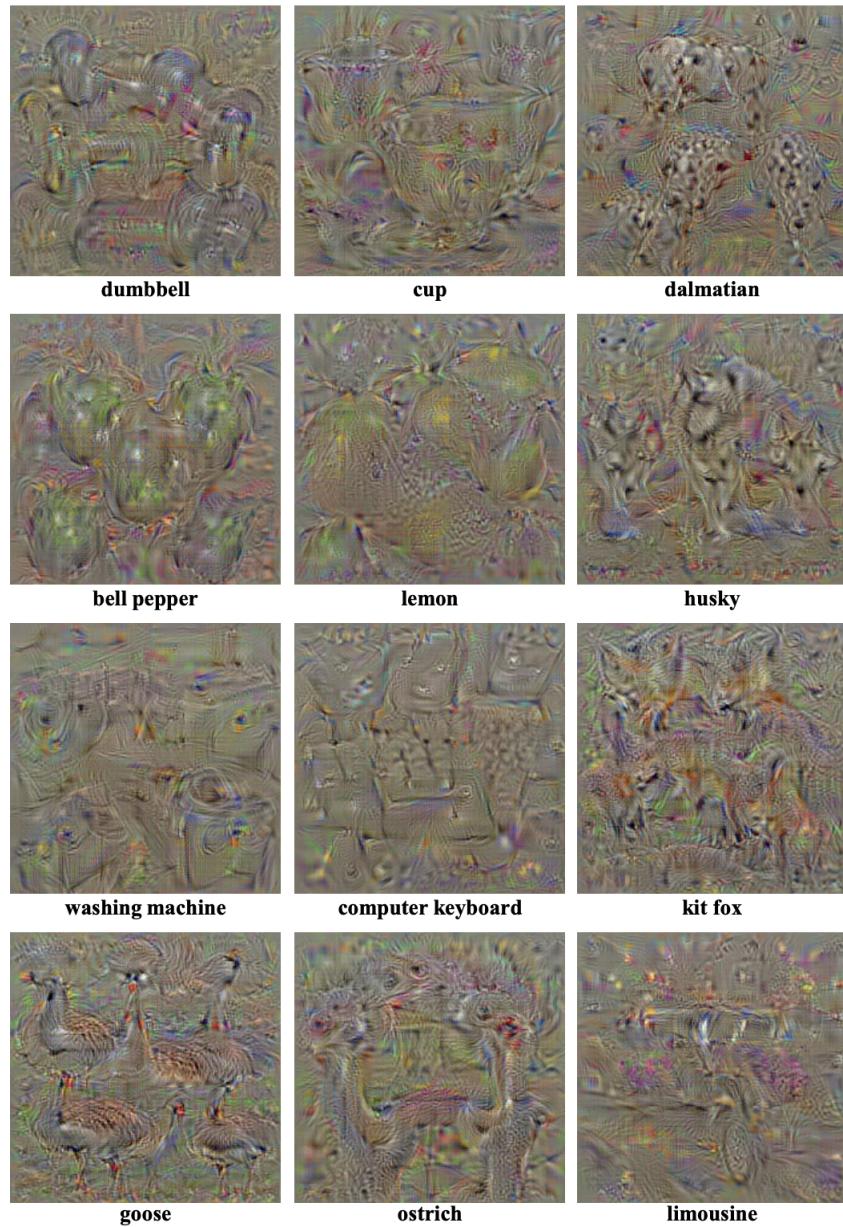
vrstve a v skrytých vrstvách sú abstraktné a vyžadujú dodatočné úsilie na ich interpretovanie. Jedným zo spôsobov interpretovania týchto neurónov je maximalizácia aktivácie (angl. activation maximization).

Maximalizácia aktivácie (angl. Activation maximization) Maximalizácia aktivácie je metóda na nájdenie takého vstupného prototypu, ktorý vyprodukuje najväčšiu mieru aktivácie pre zvolený neurón (zvyčajne je to neurón hľadanej triedy na najvyššej vrstve). Takýto vstupný prototyp je nájdený tak, že neurónovej sieti je daný na vstup neutrálny obrázok, ktorý v danej doméne nereprezentuje žiadnu triedu (zvyčajne sa jedná o šedý obrázok) a je optimalizovaná funkcia maximalizácie aktivácie pomocou poklesu gradientu [9] (angl. gradient descent). Pri aplikovaní tejto metódy na obrazové dátá výsledné prototypy vyzerajú tak ako na obrázku 2.8.

Maximalizácia aktivácie s expertom Na získanie realistickejších prototypov (prototypov, ktoré sa viac podobajú vstupným dátam) l_2 -regularizácia (používaná v maximalizácii aktivácie) je nahradená takzvaným “expertom”, ktorý sa snaží naučiť distribúciu hľadanej triedy [9]. Oproti l_2 -regularizácii, ktorá hľadá vstup maximalizujúci pravdepodobnosť triedy, expert hľadá taký vstup, ktorý je najpravdepodobnejší pre zvolenú triedu. Ako “expert” môže byť použitý napríklad Gaussian RBM (angl. Restricted Boltzmann machine) [9].

2.2.5 Vysvetľovanie predikcie neurónovej siete

Montavon; Samek; Müller (2018) definujú vysvetľovanie ako kolekciu vlastností dát, ktoré sú z interpretovateľnej domény, ktoré prispeli k výslednému rozhodnutiu (napr. zaradenie do určitej triedy - klasifikácia) pre určité pozorovanie [9]. Rozdiel oproti interpretovaniu teda je, že pri interpretovaní hľadáme vzorový prototyp (vzorové pozorovanie) pre zvolenú triedu, zatiaľ čo pri vysvetľovaní sa snažíme zistiť prečo, a teda ktoré z vlastností vstupu najviac prispeli (t.j. sú najviac



Obr. 2.8: Maximalizácia aktivácie aplikovaná na obrazové dátá. [10] Výsledné vzorové prototypy pre jednotlivé triedy nevyzerajú prirodzene, sú prevažne šedé s farebnými črtami objektov. Tieto vzorové prototypy nereprezentujú príklady vstupov "z reálneho sveta" ale ideálne vstupy pre jednotlivé triedy. Takéto vstupy nerónová sieť bežne nedostane.

relevantné) k výslednej predikcii neurónovej siete (napr. zaradenie pozorovania do určitej triedy).

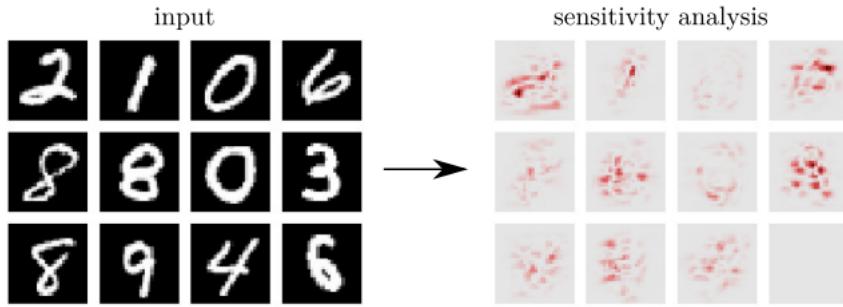
Niektoré metódy vysvetľovania fungujú na základe zakrývania častí obrázka a sledovaním zmeny predikcie predikovanej triedy – perturbačné metódy, iné zasa na základe spätného šírenia (angl. backpropagation) – napr. LRP, analýza senzitivity.

Každá z metód má svoje výhody a nevýhody, napríklad výhodou perturbačných metód je, že môžu byť použité na akýkoľvek model, keďže jediné čo potrebujú je výstup (predikciu) z modelu. Ich nevýhodou však je, že sú pomalé. Niektoré z metód vysvetľovania bližšie opíšeme v tejto sekcií.

2.2.5.1 Analýza senzitivity

Analýza senzitivity slúži na vysvetľovanie predikcie neurónovej siete. Táto metóda identifikuje, ktoré z vlastností vstupného pozorovania najviac prispievajú, či už pre alebo proti, výslednej predikcii. Najviac dôležité sú také vlastnosti, ktorých zmenou sa najvýraznejšie zmení výsledná predikcia. Na takéto vlastnosti je výsledná predikcia najviac citlivá [9].

Výsledok analýzy senzitivity znázornený v tepelnej mape (angl. heatmap) je zobrazený na obrázku 2.9. Analýza senzitivity zachytáva teda vlastnosti vstupného pozorovania, ktoré k výslednej predikcii prispievajú pozitívne aj negatívne (napr. zmenením určitej vlastnosti vstupu sa výrazne zníži zaradenie do danej triedy). Na výslednej tepelnej mape vlastnosti, ktoré k výslednej predikcii prispievajú pozitívne, a vlastnosti, ktoré k výslednej predikcii prispievajú negatívne (proti), nevieme rozlísiť. Vieme len, že zmenením danej vlastnosti výrazne ovplyvníme predikciu.



Obr. 2.9: **Analýza senzitivity aplikovaná na konvolučnú neurónovú sieť trénovanú na dátovej sade MNIST.** [9]

Červenou farbou sú zobrazené miesta ktoré najviac prispievajú, či už pre alebo proti, výslednej predikcii. Čím je červená farba výraznejšia, tým viac je výsledok senzitívny na zmenu daného pixela.

2.2.5.2 LRP (angl. layer-wiser relevance propagation)

Metóda vrstvami propagovanej relevancie, ďalej len LRP (angl. layer-wise relevance propagation), sa od analýzy senzitivity odlišuje tým, že vo výslednej tepelnej mape dokáže odlísiť vlastnosti, ktoré prispeli pozitívne alebo negatívne k výslednej predikcii (v závislosti od použitých parametrov α a β).

Táto technika funguje tak, že vstupný obrázok dopredným šírením "prejde" neurónovou sieťou, pričom sú zozbierané aktivácie neurónov v jednotlivých vrstvách. Následne je neurónovou sieťou spätným šírením propagované skôre z výstupu neurónovej siete v podobe relevancie až k vstupnému obrázku.

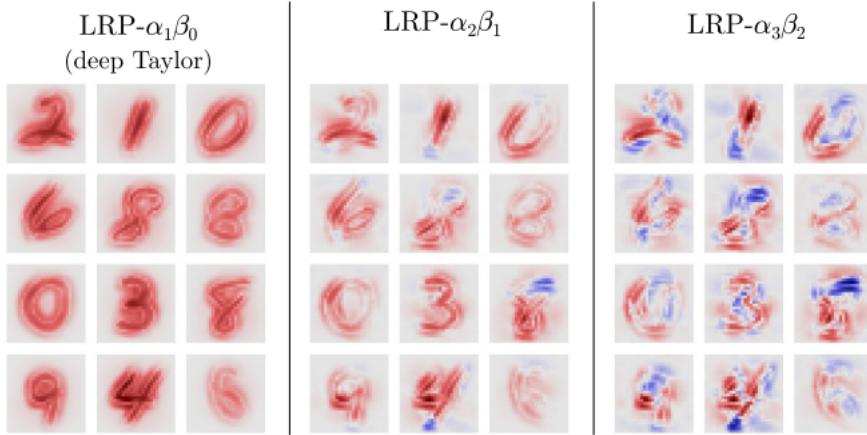
Nasledovné vzorce 2.2, 2.3, 2.4 [9] vyjadrujú spôsob výpočtu propagovanej relevancie medzi vrstvami. j a k sú jednotlivé vrstvy, pričom k je vrstva, z ktorej je relevancia R propagovaná. Parametre α a β upravujú, koľko pozitívnej (α) alebo negatívnej (β) relevancie je vytvorennej počas fázy spätného šírenia relevancie. Pri ich nastavovaní musí platiť, že $\alpha - \beta = 1$ a zároveň $\beta \geq 0$. Súčet pozitívnej a negatívnej relevancie je však medzi vrstvami vždy rovnaký [9], výsledok použitia rôznych hodnôt α a β je znázornený na obrázku 2.10. $R_{j \leftarrow k}^+$ (Obr. 2.2) a $R_{j \leftarrow k}^-$ (Obr. 2.4) vyjadrujú množstvo pozitívnej (+), resp. negatívnej (-) relevancie propagovanej z vrstvy k do vrstvy j . a_j je aktivácia neurónu, na ktorý je propagovaná

relevancia.

$$R_{j \leftarrow k}^+ = \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} \quad (2.2)$$

$$R_{j \leftarrow k}^- = \frac{a_j w_{jk}^-}{\sum_j a_j w_{jk}^-} \quad (2.3)$$

$$R_j = \sum_k (\alpha R_{j \leftarrow k}^+ - \beta R_{j \leftarrow k}^-) R_k \quad (2.4)$$



Obr. 2.10: Výsledné vysvetlenie (v podobe tepelnej mapy) vytvorené použitím LRP s rôznymi hodnotami α a β na dátovej sade MNIST. [9] Pozitívna relevancia je zobrazená červenou farbou [9]. Negatívna relevancia je zobrazená modrou farbou [9]. V prípade, že použijeme $\alpha = 1$ a $\beta = 0$ stráčame informáciu o tom, ktoré pixely negatívne (tj. sú proti výslednej predikcii) prispeli k výslednej predikcii (a opačne).

Výhodou LRP oproti analýze senzitivity je, že vysvetlenie (výsledná tepelná mapa) vytvorené technikou LRP je pre rôzne obrázky vždy rôzne [11]. Naopak, pri analýze senzitivity je vysvetlenie vždy rovnaké pokial v architektúre neurónovej siete neboli použité združovacie vrstvy (angl. pooling layers) [11]. Ďaľším rozdielom je, že vo výslednom vysvetlení LRP rozlišuje, ktoré vlastnosti pozitívne alebo negatívne prispeli k negatívnej predikcii.

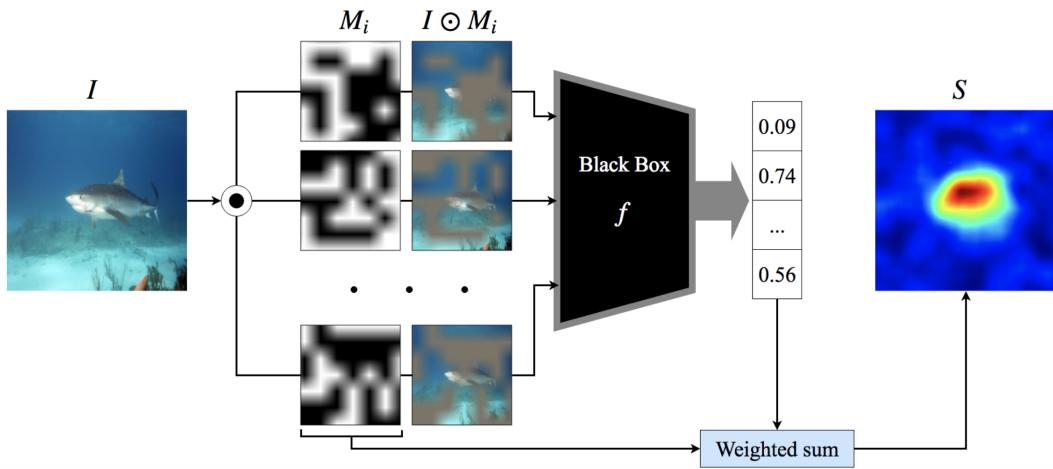
2.2.5.3 RISE - Randomized Input Sampling for Explanation

Túto metódu môžeme zaradiť medzi perturbačné metódy, keďže je tiež založená na zakrývaní jednotlivých častí obrazu a sledovaním zmeny výslednej predikcie modelu. Už z názvu modelu (*Randomized Input Sampling for Explanation*) je zrejmé, že táto metóda využíva náhodu na zakrývanie jednotlivých častí vstupného obrazu. Vstupný obraz je prekrytý náhodou maskou, ktorá je vytvorená nasledovne [12]:

- Je vytvorená náhodná binárna (tj. iba z bielej a čiernej farby) maska o malej veľkosti (napríklad 8px x 8px).
- Táto maska je zväčšená (angl. upsampled) pomocou bilineárnej interpolácie [12] (angl. bilinear interpolation) na veľkosť ktorá je mierne väčšia ako veľkosť obrázka s ktorým bude prekrytá (kvôli oreznávaniu). Tým sa zníži jej kvalita a ostré hrany medzi bielymi a čiernymi časťami sa zjemnia. Masky už teda nie sú binárne.
- Z masky je náhodne vyrezaná náhodná časť o veľkosť prekrývaného obrázka.

Toto sa opakuje N krát. Výsledná tepelná mapa je vypočítaná ako vážený priemer všetkých vygenerovaných masiek, kde váhy sú skóre (pravdepodobnosť predikovanej triedy) z modelu. Tento proces je zobrazený na obrázku 2.11.

Autori porovnali túto metódu s metódami *GradCAM* (Selvaraju et al. 2017) [13] a *LIME* (Ribeiro et al. 2016) [14]. Metóda *Rise* si oproti týmto dvom metódam počínala lepšie (Obr. 2.12). Vykonali niekoľko experimentov, v ktorých porovnali architektúry neurónových sietí *ResNet50* (He et al. 2016) [15] a *VGG16* (Simonyan; Zisserman 2014) [16] natrénované na dátových sadách PASCAL VOC07 (Everingham et al. 2010) [17] a MSCOCO2014 (Lin et al. 2014) [18]. Sledovali metriky *insertion* a *deletion* (Obr. 2.12). Metrika *insertion* je vyjadrená ako plocha pod krivkou (AUC) funkcie $y = f(x)$, kde y je istota predikcie a x je počet pridaných najdôležitejších pixelov, dôležitosť pixelov je určená metódou vysvetľovania predikcie neurónovej siete a môže byť zobrazené pomocou tepelnej mapy. Metrika *deletion* naopak odoberá najdôležitejšie pixely z obrázka.



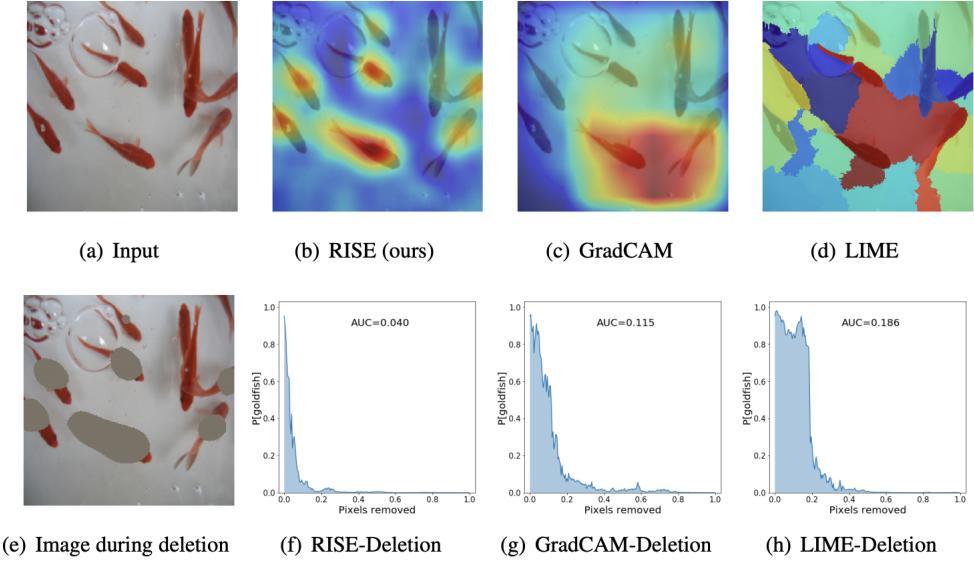
Obr. 2.11: Metóda *Rise*. [12] Vygenerované masky nahradzajú vstupný obrázok na, ktorý sú aplikované. Z výstupných predikcií jednotlivých masiek je nakoniec vypočítaná tepelná mapa.

Výhodou tejto metódy je, že oproti bežným perturbačným metódam je výrazne rýchlejšia.

2.3 Využitie neurónových sietí pri odhalovaní Alzheimerovej choroby

Neurónovým sietiam sa doposiaľ podarilo dosiahnuť veľmi dobré výsledky pri odhalovaní Alzheimerovej choroby. Medzi state-of-the-art riešenia patrí konvolučná neurónová sieť (Obr. 2.13) od Esmaeilzadeh et al. s presnosťou **94.1%** (a s F_2 skóre 0.93) na populárnej dátovej množine s názvom *ADNI-1*. Tento výsledok dosiahli v úlohe klasifikácie iba do CN a AD (bez MCI). Vstupom do tejto neurónovej siete boli snímky z magnetickej rezonancie (MRI) ale aj demografické informácie akými sú napríklad vek alebo pohlavie. Autor avšak nereportuje úspešnosť modelu, ktorý bol natrénovaný iba z obrazových dát, táto úspešnosť by bola pravdepodobne o niečo nižšia.

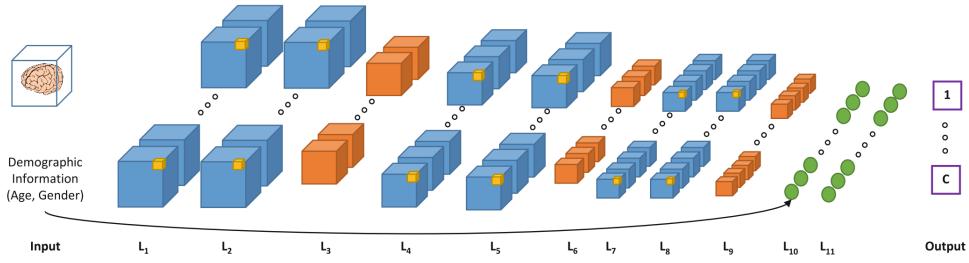
V prípade klasifikácie do všetkých troch tried - CN, MCI a AD autori tejto práce



Obr. 2.12: Porovnanie metódy *RISE* s *GradCAM* alebo *LIME*. [12] V prvom riadku sú tepelné mapy jednotlivých metód pre vstup. V druhom riadku je znázornená porovnávaná metrika *deletion*. Táto metrika sleduje vzťah medzi odobratím najdôležitejších pixelov a výslednou predikciou modelu. Je vyčíslená pomocou výpočtu plochy pod krivkou (AUC). Na grafoch si môžeme všimnúť, že metóda *RISE* potrebuje odobrať menej pixelov na to aby klesla pravdepodobnosť predikovanej triedy. To znamená, že tepelná mapa (metódy *RISE* oproti ostatným metódam) lepšie zaznamenáva dôležité pixely pre predikovanú triedu.

dosiahli horšie výsledky oproti binárnej klasifikácii. Ich model dokázal správne zaradiť pacienta s presnosťou **61.1%** (a s F_2 skóre 0.62) [19]. Pri dosiahnutí tohto výsledku použili tzv. učenie s prenosom (angl. transfer learning), ktoré im zlepšilo úspešnosť modelu až o 5.1% z pôvodných 54%. Model, z ktorého učili prenosom je už skôršie spomínaný model na binárnu klasifikáciu pacientov s Alzheimerovou chorobou.

Autori experimentovali trénovalím dvoch rôznych modelov, jedného jednoduchšieho a druhého zložitejšieho. Lepší bol jednoduchší model, pretože neboli tak náhľahlí na pretrénovanie. V týchto modeloch použili dropout, l_2 regularizáciu a augmentované dátá (obrázky otocili po osi x). Tieto "vylepšenia" pridávali po-



Obr. 2.13: Architektúra konvolučnej neurónovej siete použitej pri detekcii Alzheimerovej choroby. [19] Modré kocky sú konvolučné vrstvy, oranžové kocky sú *max-pooling* vrstvy, posledné dve (zelené) vrstvy sú plne prepojené vrstvy. Môžeme si všimnúť, že do posledných dvoch plne prepojených vrstiev okrem obrazových dát vstupujú aj informácie o veku a pohlaví.

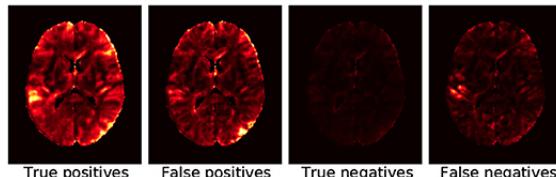
stupne a sledovali rozdiel v úspešnosti modelu, každé jedno z týchto vylepšení výrazne zlepšilo úspešnosť modelu. V kroku predspracovania dát odstránili z obrázkov také časti, ktoré nepredstavovali tkanoivo mozgu (napr. lebka) technikou s názvom BET (Smith 2002) [20], pretože z nich sa Alzheimerova choroba nedá diagnostikovať.

Niekteré práce (Suk et al. 2016) sa zaobrali dokonca klasifikáciou do štyroch tried: AD, CN, pMCI (angl. progressive MCI - pacienti ktorí pokročili k AD do 18 mesiacov), sMCI (angl. stable MC - - pacienti ktorí nepokročili k AD do 18 mesiacov). Táto úloha je samozrejme náročnejšia, najlepší model v tomto prípade dosahoval presnosť 53.72% [21]. V prípade binárnej klasifikácie (AD vs CN) sa autorom podarilo dosiahnuť presnosť až **95.09%**, oproti Esmaeilzadeh et al. však použili aj rádiologické snímky z PET. Táto práca sa ďalej vyznačuje adaptívou selekciou črt, vďaka ktorej sa autorom podarilo dosiahnuť tak dobré výsledky.

2.3.1 Vysvetľovanie rozhodnutí neurónových sietí detegujúcich Alzheimerovu chorobu

Existujúce práce sa už zaobrali metódami vysvetľovania rozhodnutí neurónových sietí detekujúcich Alzheimerovu chorobu. Böhle; Eitel; Weygandt; Ritter 2019

uviedli možnosti analýzy rozhodnutí za účelom ich vysvetľovania. Konkrétnie sa zaoberali metódami vrstvami propagovanej relevancie (LRP) a vedenou spätnou propagáciou (angl. guided backpropagation). Uvádzajú LRP ako metódu na vysvetľovanie inividuálnych rozhodnutí neurónovej siete kde naopak vedenú spätnú propagáciu ako metódu na zistenie oblastí, na ktoré je neurónová sieť senzitívna. Tieto metódy skúmali porovnávaním priemerov tepelných máp (angl. heatmaps) všetkých pozorovaní v predikovaných triedach (2 - AD, HC). Taktiež porovnávali priemerné tepelné mapy pozorovaní podľa spôsobu zaradenia výslednej predikcie (4 - true positive, true negative, false positive, false negative) (Obr. 2.14). Okrem iného porovnávali mieru relevancie pri metóde LRP v jednotlivých častiach mozgu u pozorovaní s Alzheimerovou chorobou a u pozorovaní bez nej. Možným vylepšením tejto práce je vyskúšanie metódy LRP aj na pacientoch s miernym kognitívnym poškodením (angl. mild-cognitive impairment), nie len na pacientoch s Alzheimerovoch chorobou a zdravých jedincoch.



Obr. 2.14: Priemerná relevancia (z metódy LRP ($\beta = 0$))
pozorovaní podľa spôsobu zaradenia výslednej predikcie
Najviac relevancie je na miestach so žltou farbou. [22]

2.4 Spracovanie obrazu

TODO:

2.4.1 Inpainting

TODO:

2.5 Zhrnutie

Alzhemierova choroba je bez pochyby veľmi nebezpečnou chorobou, keďže nie je "iba" o strate pamäti ale patrí k častím príčinám smrti (Sek. 2.1). Diagnostika tejto choroby pozostáva najmä z neuropsychometrických testov a analýzy rádiologických snímkov (napr. z PET, MRI). V súčasnosti tieto rádiologické snímky posudzujú doktori samotný. Práve tu je priestor pre umelú inteligenciu, aby im pri posudzovaní týchto snímkov pomohla.

V doméne obrazových dát sa používajú najmä konvolučné neurónové siete, pretože majú veľmi dobrú schopnosť naučiť sa rozoznávať špecifické objekty z obrázka. Konvolučné neurónové siete sa v nižších vrstvách naučia rozoznávať jednoduchšie tvary/hrany a vo vyšších zložitejšie šruktúry až celé objekty. Keďže jednou z možností diagnostiky Alzheimerovej choroby je diagnostika pomocou rádiologických snímkov, je možné použiť neurónové siete práve pri detekcii tohto ochorenia.

Neurónovým sieťam sa doteraz podarilo dosiahnuť veľmi dobré výsledky pri detekcii Alzheimerovej choroby, niektoré state-of-the art riešenia dosahujú presnosť až **95.09%** (Suk et al. 2016). S takto vysokou úspešnosťou môžu byť veľmi dobrým pomocníkom doktorov. Do úvahy však musíme zobrať, že tieto výsledky boli dosiahnuté bez klasifikácie MCI pacientov. V reálnom svete doktora navštívia všetky typy pacientov - CN, MCI a AD. V tomto prípade neurónové siete dosahujú rádovo nižšiu presnosť (**61.1%**, Böhle et al. 2019). Niektoré práce dosiahli tieto výsledky použitím informácií o veku a pohlaví pacienta. Keďže pravdepodobnosťu výskytu Alzheimerovej choroby po dovršení 85 rokov života je až 50% (Sek. 2.1), je možné, že sa pri vyššom veku pacienta model začne rozhodovať najmä na základe tejto informácie a nie na základe obrazových dát. Zároveň to však môže neurónovej sieti pomôcť, ak nebude brať tento atribút ako hlavný indikátor Alzheimerovej choroby, ale skôr ako pomocný atribút, ktorý bude meniť jej správanie u rôznych typov pacientov. Tu je však dôležité, takúto neurónovú sieť podrobieť dôkladnej analýze jej rozhodnutí. Osobne si ale myslím, že v produkčnom modeli by sa tento atribút mal vynechať.

Kapitola 2. Analýza

Ďalším problémom neurónových sietí je, že sa správajú ako čierne skrinky. Preto je potrebné ich rozhodnutia interpretovať, aby bolo pre doktora zrejmé na základe čoho neurónová sieť urobila svoju predikciu. V tomto práve môžu pomôcť metódy na vysvetľovanie rozhodnutí neurónovej siete (tzv. white-box metódy), alebo iné black-box metódy.

Bežným používaním neurónových sietí ako pomocníka pre doktorov, nebráni len ich vysvetliteľnosť, ale aj ich schopnosť detektie ochorenia, keďže aj tu je priestor na zlepšenie - napr. úspešnosti klasifikácie do CN, MCI a AD.

Pre pochopenie správania sa neurónových sietí poznáme metódy jej interpretovania a vysvetľovania jej rozhodnutí. Interpretovaním neurónovej siete zistujeme, ako si napríklad neurónová sieť predstavuje jednu z tried, ktorú klasifikuje. Vysvetľovaním jej rozhodnutí zas zistujeme na základe čoho neurónová sieť spravila svoje rozhodnutie, a teda ktoré zo vstupných vlastností pozorovania ju navideli k zaradeniu do určitej triedy. Niektoré z týchto metód (LRP a vedená spätná propagácia) už boli použité pri vysvetľovaní rozhodnutí neurónových sietí detekujúcich Alzheimerovu chorobu, avšak zatiaľ len pri binárnej klasifikácii pacientov.

3. Ciele práce

Vychádzajúc zo zadania projektu a na základe poznatkov nadobudnutých z analýzy domény a problému, sme si stanovili nasledovné ciele.

3.1 Vytvorenie novej, alebo vylepšenie existujúcej metódy pre vysvetľovanie rozhodnutí neurónových sietí

Existujú rôzne metódy pre vysvetľovanie rozhodnutí neurónových sietí. Či už sú to tzv. white-box metódy (ako napríklad LRP) alebo tzv. black-box metódy, ktoré je možné použiť na ľubovoľný typ modelu. Žiada z týchto metód nie je dokonalá (každá má svoje plusy a minusy v rôznych aspektoch) a je tu teda priestor na vytvorenie novej (lepšej) alebo vylepšenie existujúcej metódy. V prípade vylepšenia existujúcej metódy je nutné túto metódu porovnať najmä s vylepšovanou metódou a následne s inými metódami.

3.2 Využitie vytvorenej metódy na určenie miery správnosti modelu neurónovej siete detekujúcej Alzheimerovu chorobu

Pri neurónových sieťach detekujúcich Alzheimerovu chorobu je dôležité, aby sa naučili klasifikovať pacientov na základe relevatných črt z rádiologických snímkov. Práve, preto je potrebné určiť mieru správnosti modelu podľa toho či sa model rozhoduje práve na základe týchto črt a nie iných. Na to sa využívajú metódy na vysvetľovanie rozhodnutí neurónových sietí, v tomto prípade sa použije novovytvorená metóda.

4. Návrh riešenia

Pre použitie neurónových sietí v bežnej praxi doktorov pri diagnostike Alzheimerovej choroby je nevyhnutné, aby sa rozhodnutia neurónových sietí dali vysvetliť. Preto navrhujeme metódu na vyvsetľovanie rozhodnutí neurónových sietí, ktorú overíme na MRI snímkoch u pacientov (CN, MCI a AD).

Vychádzajúc cieľa práce *3.1 Vytvorenie novej alebo vylepšenie existujúcej metódy pre vysvetľovanie rozhodnutí neurónových sietí* navrhujeme metódu, ktorá vychádza z už existujúcej metódy *RISE* (Sek. 2.2.5.3). Táto metóda dosiahla veľmi dobré výsledky oproti metódam GradCAM a LIME a je, teda vhodným základom na možné vylepšenia. Táto metóda funguje na princípe zakrývania častí obrázka (tak ako iné perturbačné metódy). Po takomto prekrytí u iných perturbačných metódach vznikajú ostré hrany, čo môže neurónovú sieť myliť, *Rise* tento problém ale nemá. Avšak tento prekryv býva zvyčajne v čiernej farbe. Keďže v MRI snímky sú v odtieňoch čiernej (a u AD jedincov je na snímkoch oveľa viac ”čiernej” z dôvodu úbytku mozgového tkaniva) môže byť práve toto ďalším zdrojom zmätenia pre neurónovú sieť. Preto navrhujeme zakrývané miesta dokresliť určitou metódou spracovania obrazu (Sek. 2.4) alebo na zakrytie použiť inú hodnotu.

4.1 RISEI - Randomized Input Sampling for Explanation with Inpainting

Metódu sme pomenovali *Randomized Input Sampling for Explanation with Inpainting* (t.j. náhodné vzorkovanie vstupu pre vysvetlovanie s dokreslovaním) so skratkou RISEI.

Keďže metóda vychádza už z existujúcej metódy, časť našej metódy je samozrejme rovnaká. Proces vytvorenia vysvetlenia klasifikácie do triedy T pre obrázok O modelom je teda nasledovný:

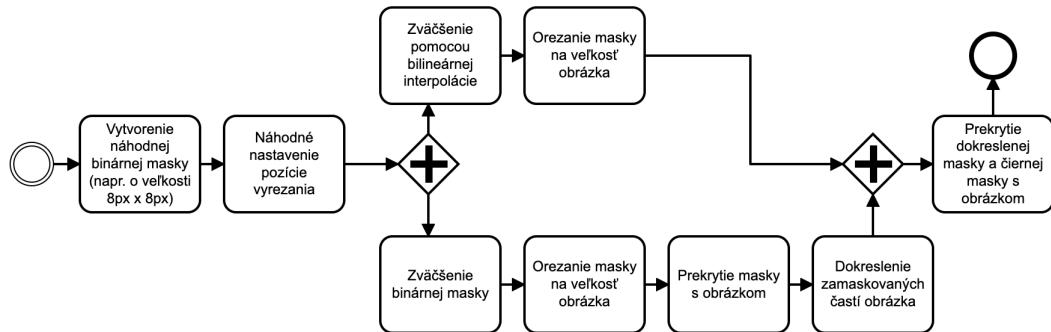
1. Vytvorenie N náhodne zamaskovaných obrázkov z obrázka O
2. Vloženie zamaskovaných obrázkov do modelu a následné získanie pravdepodobnosti pre triedu T
3. Vytvorenie a vizualizácia vysvetlenia pomocou tepelnej mapy

Toto sú 3 hlavné kroky z ktorých pozostáva táto metóda, ďalej bližšie popíšeme jednotlivé z nich.

Vytvorenie náhodne zamaskovaných obrázkov. Vytvorenie náhodne zamaskovaných obrázkov tiež pozostáva z niekoľkých krokov, pričom niektoré z nich môžu bežať paralelne. Tento sme znázornili diagramom (Obr. 4.1). Masky sa vytvárajú paralelne, pretože ”čierna” maska ma jemné hrany a na dokreslenie potrebujeme naopak masku s ostrými hranami.

Oproti metóde *RISE* vytvárame o jednu masku naviac, a teda je originálny obrázok prekrytý z viacerými maskami. Jednotlivé masky cez seba prekryjeme, pričom každej z nich nastavíme určité množstvo priehľadnosti. S týmto pomerom môžeme ďalej experimentovať a výsledky porovnávať. Môžeme porovnať použitie iba dokreslenej masky s iba čierou maskou a tiež s použitím oboch v rôznych pomeroch.

Vytvorenie ”čiernej” masky je rovnaké, ako pri metóde *RISE*. Dokreslená maska



Obr. 4.1: BPMN diagram generovania jedného obrázka prekrytého maskou

vznikne dokreslením zakrytých (zamaskovaných) častí obrázka pomocou jedného z algoritmov na dokreslovanie (angl. inpainting). Tieto algoritmy sme popísali v sekcií 2.4 Spracovanie obrazu. Obrázok 4.2 je príkladom dokreslenia častí vzorového obrázka na základe masky náhodne vygenerovanej masky. V našej metóde budeme experimentovať s rôznymi farbami prekrytia, nielen s čierной.



Obr. 4.2: Niektoré časti vzorového obrázka (vľavo) boli dokreslené podľa náhodne vygenerovanej binárnej masky (v strede). Výsledný obrázok (vpravo) môže byť ešte prekrytý "čierной" maskou s určitou príehľadnosťou.

Vytvorenie a vizualizácia vysvetlenia pomocou tepelnej mapy. Tento krok je identický s originálnou metódou *RISE*. Nasledovný vzorec 4.1 vyjadruje výpočet dôležitosti I pre každý pixel $[x, y]$ obrázka, kde n je počet všetkých zamás-

kovaných obrázkov. Funkcia $p(k, x, y)$ vracia 0 ak pixel $[x, y]$ bol v danom zamaskovanom obrázku k prekrytý, inak vracia predikciu (pravdepodobnosť) z modelu pre zamaskovaný obrázok k .

$$I_{x,y} = \frac{\sum_k^n p(k, x, y)}{n} \quad (4.1)$$

Táto metóda do originálnej pridáva niekoľko parametrov a najmä výpočtovo náročné dokreslovanie, preto bude nutné nájsť vhodné nastavenie parametrov, aby výpočet vysvetlenia neboli príliš časovo náročný. Práve výpočtová náročnosť môže byť slabinou tejto metódy.

4.2 Overenie riešenia

Našu metódu budeme najskôr porovnávať s originálnou metódou RISE (tj. či sa nám podarilo spraviť lepšiu metódu) a následne s metódou LRP. Tieto experimenty môžeme vykonávať na CN a AD vzorkách; a aj na CN, MCI a AD vzorkách. Budeme sledovať kvalitu navrhnutej metódy (oproti ostatným metódam) a na základe týchto tepelných máp budeme vyhodnocovať mieru správnosti modelu.

4.2.1 Dátová sada

Experimenty budeme vykonávať na dátovej sade ADNI, ktorá obsahuje MRI snímky AD pacientov. Táto dátová sada bola použitá aj na trénovanie state-of-the-art modelu na diagnostiku Alzheimerovej choroby [19], ale aj pri vysvetľovaní rozhodnutí neurónovej siete pomocou LRP [22]. Na tejto dátovej sade budeme musieť vykonať rovnaké predspracovanie ako Böhle et al., aby sme sa s ich výsledkami mohli porovnať. Prípadne môžeme vykonať vlastné predspracovanie, ale budeme musieť vykonať aj experimenty s metódou LRP.

4.2.2 Experimenty

Najskôr budeme vyhodnocovať nami navrhnutú metódu pomocou sledovania kvality tepelných máp. Následne budeme overovať správnosť modelu pomocou nami navrhнутej metódy.

4.2.2.1 Určenie kvality metódy vysvetľovania rozhodnutí modelu

Kvalitu metódy vysvetľovania rozhodnutí modelu budeme sledovať určovaním kvality tepelnej mapy. Tá v kontexte našej práce hovorí o tom, do akej miery tátu mapa odzrkadľuje to, na základe čoho sa model rozhoduje. Toto budeme merať metrikami *insertion (AUC)* a *deletion (AUC)*, ktoré sme bližšie popísali v sekcii 2.2.5.3. Táto metrika nám povie, ako dobrá je naša metóda na vysvetľovanie.

4.2.2.2 Určenie správnosti modelu

Správnosť modelu budeme určovať na základe tepelných máp vytvorených pomocou metódy na vysvetľovanie predikcií modelu. Budeme overovať do akej miery dávajú tepelné mapy zmysel v kontexte skutočnej anatómie mozgu. Sledujeme, že či tepelná mapa nehovorí o tom, že sa model rozhodol na základe takej oblasti mozgu, z ktorej sa Alzheimerova choroba nedá zistiť. Veľkú úlohu pri určovaní správnosti modelu zohráva aj kvalita natréновaného modelu. Táto metrika je súborom niekoľkých metrík z práce od Böhle et al.

4.3 Záver

V tejto kapitole sme navrhli metódu na vysvetľovanie rozhodnutí modelov strojového učenia a spôsob jej implementácie. Navrhnutú metódu budeme overovať na neurónových sieťach detekujúcich Alzheimerovu chorobu s cieľom odhaľovania nesprávnych rozhodnutí.

Literatúra

1. AMISHA, Paras Malik; PATHANIA, Monika; RATHAUR, Vyas Kumar. Overview of artificial intelligence in medicine. *Journal of family medicine and primary care*. 2019, roč. 8, č. 7, s. 2328.
2. 2019. Dostupné tiež z: <http://www.alzheimer.sk/informacie/alzheimerovachoroba.aspx>.
3. DUTHEY, Béatrice. Background paper 6.11: Alzheimer disease and other dementias. *A Public Health Approach to Innovation*. 2013, s. 1–74.
4. KHAN, Tapan. *Biomarkers in Alzheimer's Disease*. Academic Press, 2016.
5. 2017. Dostupné tiež z: <https://www.alz.org/alzheimers-dementia/facts-figures>.
6. WORKING, G Biomarkers Definitions. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther*. 2001, roč. 69, č. 3, s. 89–95.
7. HAYKIN, Simon S et al. *Neural networks and learning machines/Simon Haykin*. New York: Prentice Hall, 2009.
8. LEE, Honglak; GROSSE, Roger; RANGANATH, Rajesh; NG, Andrew. Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks. Dostupné z DOI: [10.1145/2001269](https://doi.org/10.1145/2001269).
9. MONTAVON, Grégoire; SAMEK, Wojciech; MÜLLER, Klaus-Robert. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*. 2018, roč. 73, s. 1–15.

10. SIMONYAN, Karen; VEDALDI, Andrea; ZISSERMAN, Andrew. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*. 2013.
11. MÜLLER, Klaus-Robert; SAMEK, Wojciech; MONTAVON, Gregoire; LAPUSCHKIN, Sebastian; ARRAS, Leila. *Explaining and Interpreting Deep Neural Networks*. Dostupné tiež z: http://iphome.hhi.de/samek/pdf/DTUSummerSchool2017_1.pdf.
12. PETSIUK, Vitali; DAS, Abir; SAENKO, Kate. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*. 2018.
13. SELVARAJU, Ramprasaath R; COGSWELL, Michael; DAS, Abhishek; VEDANTAM, Ramakrishna; PARIKH, Devi; BATRA, Dhruv. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision*. 2017, s. 618–626.
14. RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. "Why should i trust you? Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, s. 1135–1144.
15. HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing; SUN, Jian. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, s. 770–778.
16. SIMONYAN, Karen; ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. 2014.
17. EVERINGHAM, Mark; VAN GOOL, Luc; WILLIAMS, Christopher KI; WINN, John; ZISSERMAN, Andrew. The pascal visual object classes (voc) challenge. *International journal of computer vision*. 2010, roč. 88, č. 2, s. 303–338.
18. LIN, Tsung-Yi; MAIRE, Michael; BELONGIE, Serge; HAYS, James; PERONA, Pietro; RAMANAN, Deva; DOLLÁR, Piotr; ZITNICK, C Lawrence.

- Microsoft coco: Common objects in context. In: *European conference on computer vision*. 2014, s. 740–755.
- 19. ESMAEILZADEH, Soheil; BELIVANIS, Dimitrios Ioannis; POHL, Kilian M; ADELI, Ehsan. End-to-end Alzheimer's disease diagnosis and biomarker identification. In: *International Workshop on Machine Learning in Medical Imaging*. 2018, s. 337–345.
 - 20. SMITH, Stephen M. Fast robust automated brain extraction. *Human brain mapping*. 2002, roč. 17, č. 3, s. 143–155.
 - 21. SUK, Heung-Il; LEE, Seong-Whan; SHEN, Dinggang; INITIATIVE, Alzheimer's Disease Neuroimaging et al. Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis. *Brain Structure and Function*. 2016, roč. 221, č. 5, s. 2569–2587.
 - 22. BÖHLE, Moritz; EITEL, Fabian; WEYGANDT, Martin; RITTER, Kerstin. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Frontiers in aging neuroscience*. 2019, roč. 11, s. 194.

A. Plán práce

A.1 Zimný semester

V tomto semestri plánujem pracovať na implementácii navrhnutej metódy, ktorú budem overovať v experimentoch a postupne vylepšovať. V tomto semestri plánujem:

- Natrénovať model na detekciu Alzheimerovej choroby z MRI snímkov
- Implementovať navrhnutú metódu
- Experimentovať s hyper-parametrami navrhnutej metódy
- Skúmať dosiahnuté výsledky, hľadať príčiny a možné vylepšenia
- Priebežne písat' prácu – implementáciu a dosiahnuté výsledky

A.2 Letný semester

V tomto semestri budem pracovať na finalizácii tejto práce, navrhnutú metódu plánujem už iba vylepšovať a pracovať na záverečnom dokumente. V tomto semestri plánujem:

- Písat' prácu a jej jednotlivé časti - implementácia, technická dokumentácia, dosiahnuté výsledky, záver

Dodatok A. Plán práce

- Vykonáť úpravy v navrhnutej metóde na základe doterajších výsledkov experimentov
- Vyhodnotiť a porovnať vykonalé experimenty
- Porovnať navrhnutú metódy s existujúcimi metódami
- Odovzdať prácu

Dodatok A. Plán práce

Dodatok A. Plán práce

Dodatok A. Plán práce