

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 DP2

Študijný program: Inteligentné softvérové systémy

Študijný odbor: 18. Informatika

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

Vedúci práce: Ing. Martin Tamajka

január 2021

Návrh zadania diplomovej práce

Finálna verzia do diplomovej práce ¹

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. Zatial' čo vo výskume je prípustná istá miera neistoty alebo nepresnosti, v oblastiach ako je medicína je žiadúce, 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 vyhodnoťte 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 súhlasi, že bude takýto projekt viesť v prípade, že bude pridelený tomuto študentovi.

V Bratislave dňa 29.1.2020



Pôpis študenta



Pôpis výskumníka

Vyjadrenie garanta predmetov Diplomový projekt I, II, III

Návrh zadania schválený: áno / nie⁴

Dňa: 17. 2. 2020



Pôpis 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 (uvedte 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, 3. január 2021

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
január 2021

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 rozhodnutí neurónovej siete detekujúcich 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	7
2	Analýza	9
2.1	Alzheimerova choroba	9
2.1.1	Diagnostika Alzheimerovej choroby	10
2.1.2	Biologické ukazovatele	10
2.1.3	Obrazové a rádiologické ukazovatele	11
2.2	Neurónové siete	12
2.2.1	Neurón	14
2.2.2	Dopredné neurónové siete	15
2.2.3	Konvolučné neurónové siete	15
2.2.4	Architektúry konvolučných neurónových sietí	18
2.2.5	Interpretovanie neurónovej siete	19
2.2.6	Vysvetľovanie predikcie neurónovej siete	20
2.2.6.1	Analýza senzitivity	22
2.2.6.2	LRP (angl. layer-wiser relevance propagation) . . .	23
2.2.6.3	RISE - Randomized Input Sampling for Explanation	24
2.3	Využitie neurónových sietí pri odhalovaní Alzheimerovej choroby .	27
2.3.1	Vysvetľovanie rozhodnutí neurónových sietí detegujúcich Alzheimerovu chorobu	29
2.4	Spracovanie obrazu	30
2.4.1	Rekonštrukcia obrazu	31
2.5	Zhrnutie	32

Obsah

3 Ciele práce	35
3.1 Vytvorenie novej, alebo vylepšenie existujúcej metódy pre vysvetľovanie rozhodnutí neurónových sietí	35
3.2 Využitie vytvorenej metódy na určenie miery správnosti modelu neurónovej siete detegujúcej Alzheimerovu chorobu	36
4 Návrh riešenia	37
4.1 RISEI - Randomized Input Sampling for Explanation with Inpainting	38
4.2 Overenie riešenia	41
4.2.1 Dátová sada	41
4.2.2 Experimenty	41
4.2.2.1 Určenie kvality metódy vysvetľovania rozhodnutí modelu	41
4.2.2.2 Určenie správnosti modelu	42
4.3 Záver	43
5 Implementácia	45
5.1 Metóda RISEI	45
5.1.1 Generovanie masiek	45
5.1.2 Vytvorenie tepelných máp	53
5.1.3 Vyhodnotenie tepelných máp	54
5.1.3.1 Metriky insertion & deletion	54
5.2 Model na detekciu Alzheimerovej choroby na základe MRI snímkov	56
6 Overenie riešenia	59
6.1 Experimenty	59
6.2 Záver	59
Literatúra	61
Dodatok A Plán práce	
A.1 Letný semester - DP1	

Obsah

A.2 Zimný semester - DP2

 A.2.1 Vyjadrenie k plneniu plánu

A.3 Letný semester - DP3

Obsah

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 viedieť obhájiť. Aby sa teda neurónové siete mohli stať bežným pomocníkom lekárov, je vysvetliteľnosť ich rozhodnutí dôležitá. Avšak toto nie je jedinou motiváciou pre vysvetliteľnosť rozhodnutí neurónových sietí. Novovznikajúce regulácie, ako napríklad pripravovaná regulácia s názvom "Right to Explanation" od Európskej Únie [2] vyžadujú vysvetliteľnosť systémov umelej inteligencie. Motivácia je teda aj legislatívna.

2. Analýza

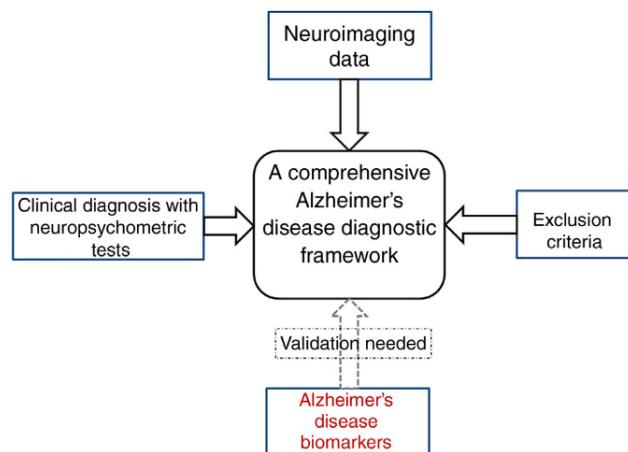
2.1 Alzheimerova choroba

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 [3]. Jej priebeh sa vyznačuje postupným poklesom kognitívnych funkcií, postupným zhoršením pamäte, myslenia, rozprávania a schopnosti učenia sa [4]. Najčastejšie sa vyskytuje u ľudí starších ako 65 rokov, s pravdepodobnosťou výskytu až 50% po dovršení 85 rokov života [4]. S narastajúcim vekom človeka sa zvyšuje pravdepodobnosť ochorenia. Pravdepodobnosť ochorenia zvyšujú taktiež úrazy hlavy, poruchy prekrvenia mozgu, pozitívna rodinná anamnéza či vzdelanie (protože ľudia s nižším vzdelaním majú väčšie riziko rozvoja tohto ochorenia) [3]. Toto ochorenie sa vyskytuje častejšie u žien ako u mužov, v pomere 2:1 [5].

Alzheimerova choroba nie je “iba” o strate pamäti, ale aj šiestou najčastejšou príčinou smrti v USA [6]. Medzi rokmi 2000 až 2017 sa počet úmrtí v USA viac ako zdvojnásobil [6]. Ľudia starší ako 65 rokov ktorým bola diagnostikovaná táto choroba sa v priemere dožívajú 4 až 8 rokov po jej diagnóze [6].

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 [5]. 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 [5] (vyžaduje si to zaškolený personál a vybavenie). Biologické ukazovateľe zatiaľ nie sú dostatočne spoľahlivé [5].



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.[5]

2.1.2 Biologické ukazovatele

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

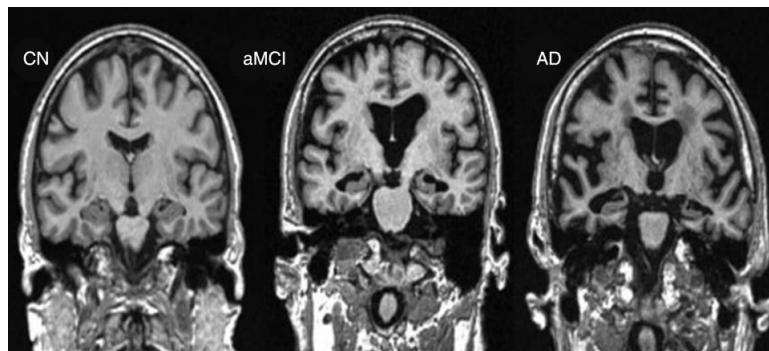
logický ukazovateľ ako indikátor určitého objektívneho merania a hodnotenia biologického procesu, patogénneho procesu alebo farmakologického hodnotenia terapeutickej účinnosti [7]. Alzheimerova choroba môže byť identifikovaná sledovaním týchto biologických ukazovateľov napríklad v krvnej plazme [5] alebo v mozgovo-miechovej tekutine (angl. cerebrospinal fluid) (ako úrovne hladín proteínov P-tau and A β 42) [5] (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) [5].

Snímky z magnetickej rezonancie (MRI) dokážu zachytiť odumieranie tkaniva (na základe biologických procesov), ktoré sa odohráva v rôznych častiach mozgu [5]. Príklad takého snímku sa nachádza na obrázku 2.2.

Snímky z pozitrólovej 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.



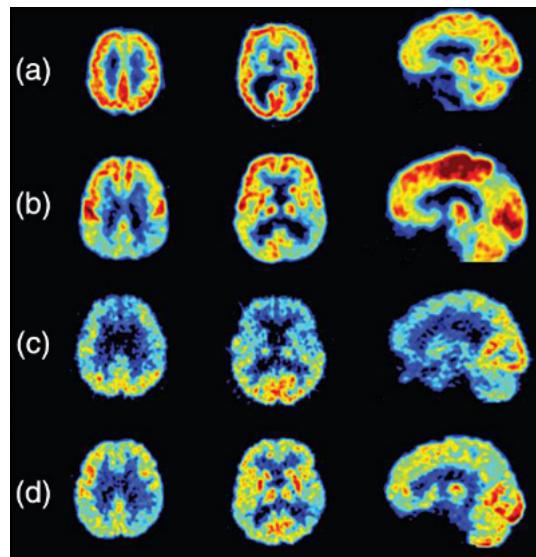
Obr. 2.2: **Typické odumieranie mozgového tkaniva zachytené magnetickou rezonanciou.** Obrázok zľava, označený ako CN (angl. cognitive normal), reprezentuje kognitívne normálneho jedinca. Obrázok v strede, označený ako aMCI (angl. amnestic mild cognitive impairment) reprezentuje jedinca s miernym kognitívnym 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. [5]

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 prekonať človeka. V našej oblasti obrazových rádiologických dát sa používajú najmä konvolučné neurónové siete.

Haykin et al. [8] 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).** [5] 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 [8]:

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

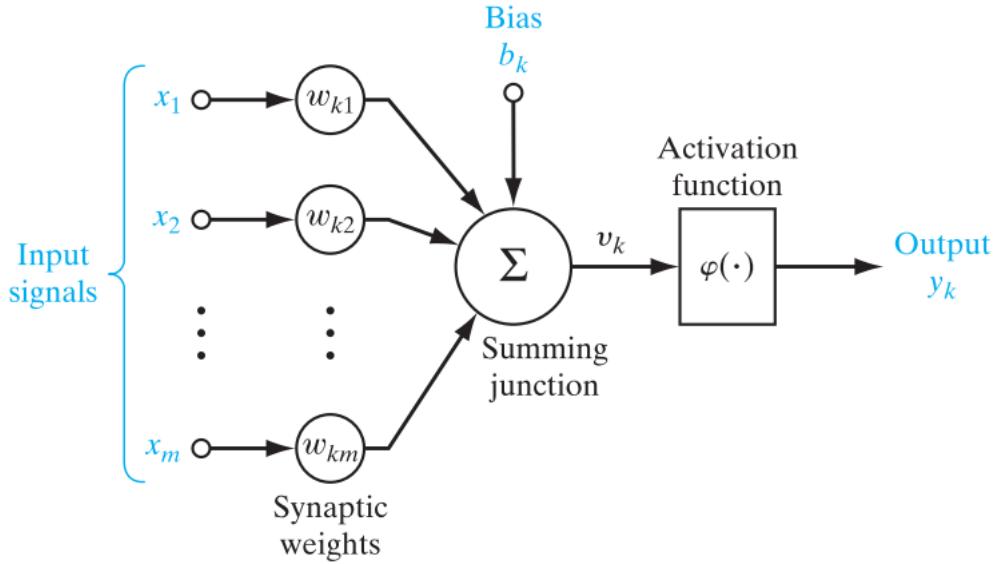
Kde:

- 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.



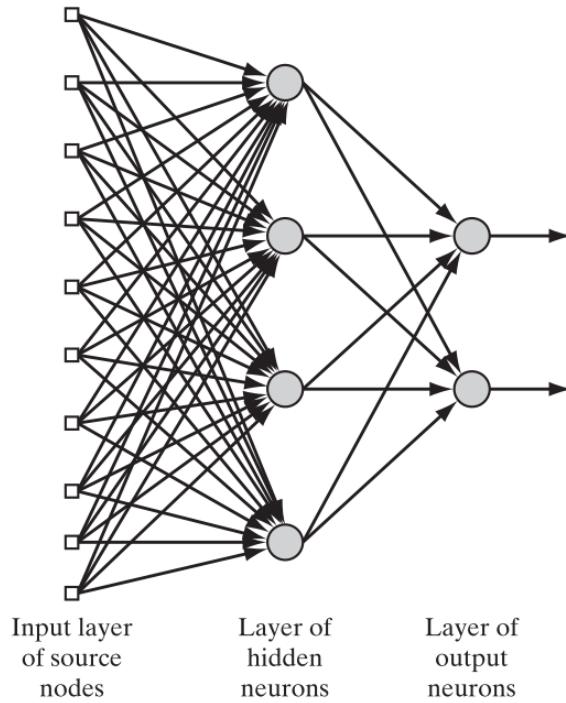
Obr. 2.4: **Model neurónu.** [8] Neurón sa skladá zo vstupných signálov a váh, ktoré sú na tieto signály aplikované, sklon (b_k - angl. bias) a aktivačnej funkcie, ktorá zabezpečuje nelinearitu. Vzorec 2.1 matematicky popisuje správanie neurónu.

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

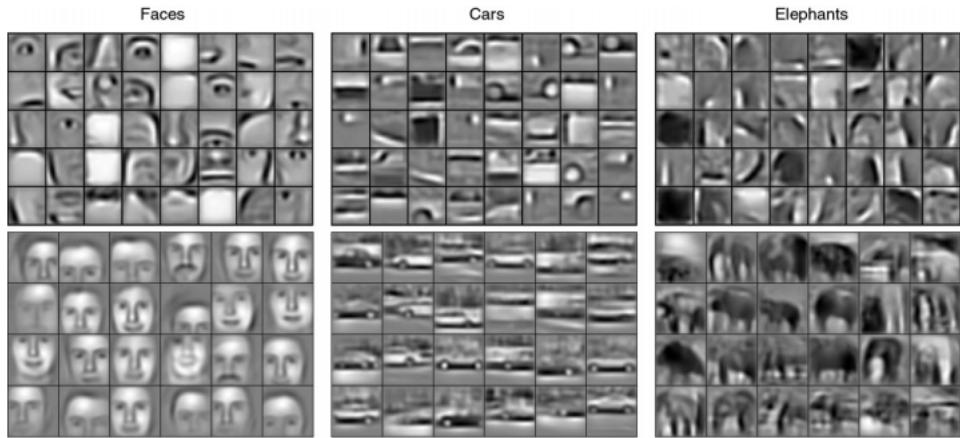
Konvolučné neurónové siete sa používajú prevažne v doméne obrazových dát. Tieto siete majú schopnosť naučiť sa rozpoznávať špecifické štruktúry/tvary z obrázka. Toto dokážu pomocou takzvaných konvolučných filtrov, ktoré sa v nižších vrstvách



Obr. 2.5: **Model doprednej neurónovej siete.** [8] Dopredné neurónové siete sa skladajú zo vstupnej vrstvy, skrytých vrstiev a výstupnej vrstvy. Keď 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).

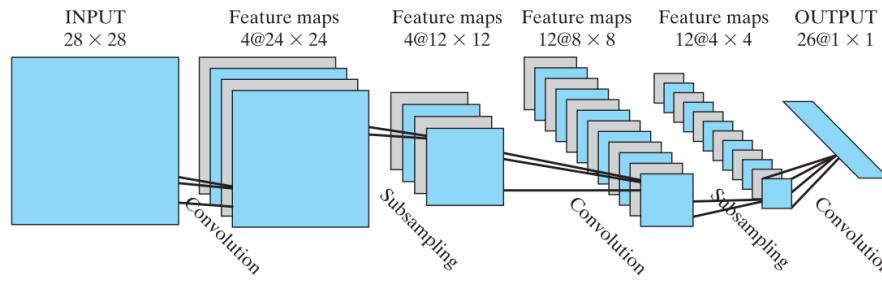
Základnými stavebnými blokmi konvolučných neurónových sietí sú konvolučné vrstvy (angl. convolutional layers) a združovacie vrstvy (angl. pooling layers).



Obr. 2.6: Vizualizácia druhej (hore) a tretej vrstvy (dole) konvolučných neurónových sietí naučených na špecifické kategórie objektov (tváre, autá a slony). [9] Nižšie vrstvy rozoznávajú jednoduchšie štruktúry zatiaľ čo vyššie už dokážu rozoznávať aj tie zložitejšie.

Konvolučné vrstvy Pomocou konvolučných vrstiev sa neurónová sieť učí extrahovať črty z obrázka [8]. Konvolúcia prebieha tak, že tzv. jadro (angl. kernel) sa posúva po tzv. mape vlastností (angl. feature map) a matematickými operáciami z pôvodnej mapy vlastností a svojich parametrov vytvára novú mapu vlastností. Tieto parametre sú trénovateľné, čo umožňuje sa každému jadru naučiť určitú črtu - napr. hranu. Konvolučná vrstva tiež dokáže znižovať komplexitu modelu (a teda aj počet jeho parametrov) jej hyper parametrami (angl: stride, padding, depth).

Združovacie vrstvy Cieľom združovacích vrstiev je postupne znižovať dimenzionalitu dát, tým znižovať počet parametrov modelu, a teda aj jeho komplexitu [10]. Najčastejšie sa používajú vrstvy združujúce maximom (angl. max-pooling), ale existujú aj vrstvy združujúce priemerom či súčtom.



Obr. 2.7: Príklad architektúry konvolučnej neurónovej siete.

[8] V tejto architektúre neurónovej siete sa používajú tri konvolučné vrstvy (označené ako *convolution*) a dve združovacie vrstvy (označené ako *subsampling*). Môžeme si všimnúť, že konvolučné vrstvy postupne pridávajú mapy vlastností (tiež označované ako: angl. "volumes") a tiež mierne znižujú ich veľkosť. Združovacie vrstvy zasa výrazne znižujú ich veľkosť (až o polovicu) a tým aj počet parametrov v neurónovej sieti.

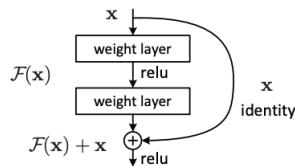
2.2.4 Architektúry konvolučných neurónových sietí

Architektúra neurónovej siete hovorí o tom, ako neurónová sieť vyzerá - koľko má vrstiev, z akých vrstiev sa skladá (konvolučné, združovacie, husté), koľko filtrov je v jednotlivých vrstvách a pod. Nie každá architektúra je vhodná na každý problém. Ak je problém jednoduchý, môže byť použitie veľmi hlbokej neurónovej siete zbytočné. Taktiež, jednoduchšia architektúra potrebuje menej výpočtových zdrojov na natrénovanie a je odolnejšia voči pretrénovaniu. Spomeniem niekoľko najznámejších architektúr, ktoré sú používané najmä pri klasifikácii obrazových dát.

- VGG [11] - hlboká neurónová sieť, so 16 alebo s 19 vrstvami. Skladá sa s konvolučných a združovacích vrstiev.
- ResNET [12] - hlboká neurónová sieť skladajúca sa z reziduálnych blokov. Reziduálne bloky obsahujú skracovacie spojenia, "skratky" (angl. shortcut connections) ako nástroj na zabránenie miznúcemu a explodujúcemu gradientu. Táto architektúra bola navrhnutá s 20, 32, 44, 56, 110 a 1202 vrstvami.
- Inception (GoogLeNet) [szegedy20s15going] - hlboká neurónová sieť skla-

dajúca sa s inception blokov. Každý blok robí niekoľko rôznych konvolúcii zo vstupu daného bloku, ktoré sú následne spojené v združovacom bloku.

Taktiež existuje niekoľko ďalších vylepšení Inception architektúry (Inception v1 až v4), dokonca aj kombinácia s architektúrou ResNet.



Obr. 2.8: Reziduálny blok v architektúre ResNET. Informácia z predhádzajúceho bloku je súčasťou výstupu aktuálneho bloku pomocou skracovacieho spojenia. [12]

2.2.5 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) [13]. Medzi domény, ktoré nie sú interpretovateľné zaraďujú napríklad latentné vektorové reprezentácie slov (angl. word embeddings) alebo iné abstraktné vektorové reprezentácie [13]. Na rozdiel od vstupných dát do neurónovej siete, ktoré sú zvyčajne interpretovatelné, 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 [13] (angl. gradient descent). Pri aplikovaní tejto metódy na obrazové dátá výsledné prototypy vyzerajú tak ako na obrázku 2.9.

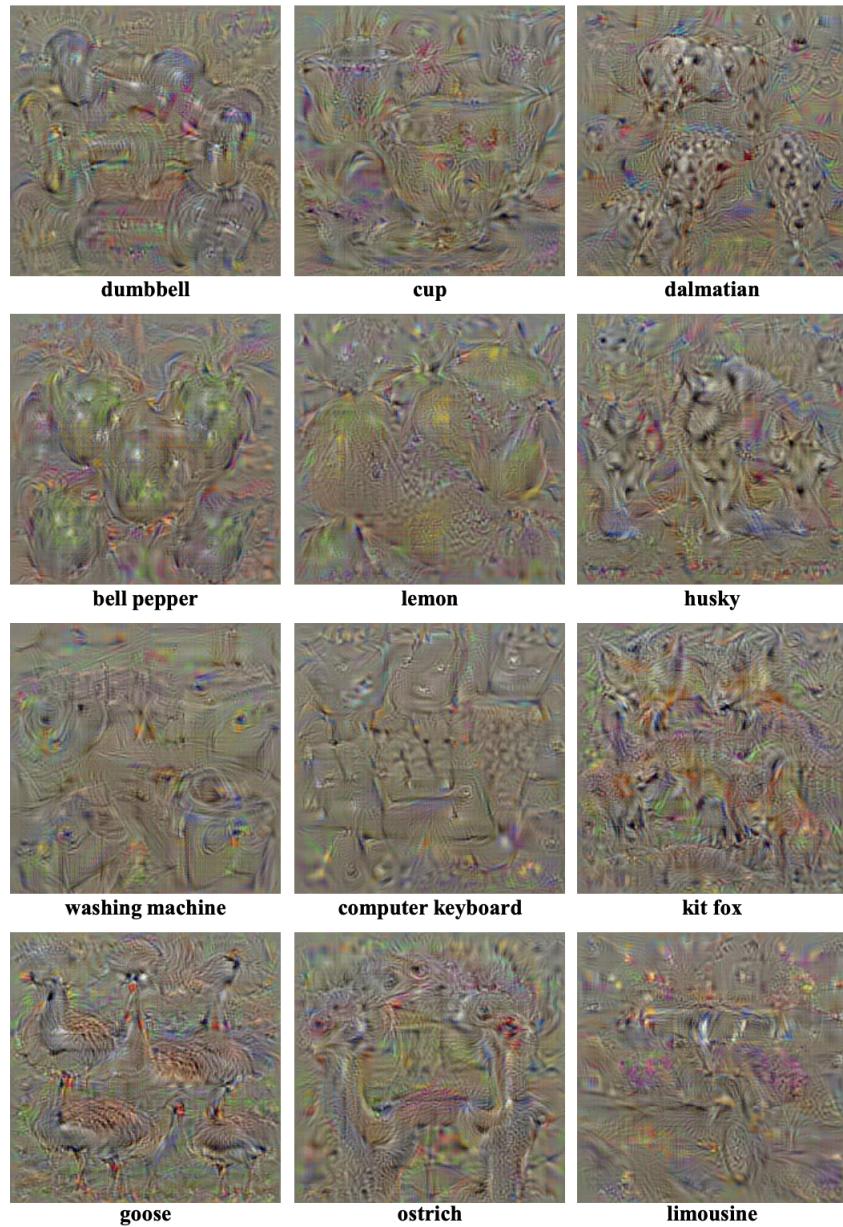
Maximalizácia aktivácie s expertom Na získanie realistickejších prototypov (prototypov, ktoré sa viac podobajú vstupným dátam) l_2 -regularizácia (používaná v maximalizácii aktivácie) je nahradená takzvaným “expertom”, ktorý sa snaží naučiť distribúciu hľadanej triedy [13]. Oproti l_2 -regularizácii, 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) [13].

2.2.6 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 [13]. 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 (tj. sú najviac relevantné) k výslednej predikcii neurónovej siete (napr. zaradenie pozorovania do určitej triedy).

Niekteré 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



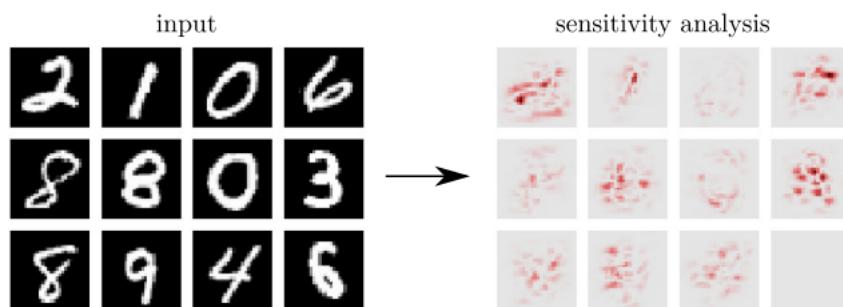
Obr. 2.9: Maximalizácia aktivácie aplikovaná na obrazové dátá. [14] Výsledné vzorové prototypy pre jednotlivé triedy nevyzerajú prirodzene, sú prevažne šedé s farebnými črtami objektov. Tieto vzorové prototypy nereprezentujú príklady vstupov "z reálneho sveta" ale ideálne vstupy pre jednotlivé triedy. Takéto vstupy nerónová sieť bežne nedostane.

metód vysvetľovania bližšie opíšeme v tejto sekcií.

2.2.6.1 Analýza senzitivity

Analýza senzitivity slúži na vysvetľovanie predikcie neurónovej siete. Táto metóda identifikuje, ktoré z vlastností vstupného pozorovania najviac prispievajú výslednej predikcii. Najviac dôležité sú také vlastnosti, ktorých zmenou sa najvýraznejšie zmení výsledná predikcia. Na takéto vlastnosti je výsledná predikcia najviac citlivá [13].

Výsledok analýzy senzitivity znázornený v tepelnej mape (angl. heatmap) je zobrazený na obrázku 2.10. 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. 2.10: **Analýza senzitivity** aplikovaná na konvolučnú neurónovú sieť trénovanú na dátovej sade MNIST. [13]

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.

2.2.6.2 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 odlišiť 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 skóre z výstupu neurónovej siete v podobe relevancie až k vstupnému obrázku.

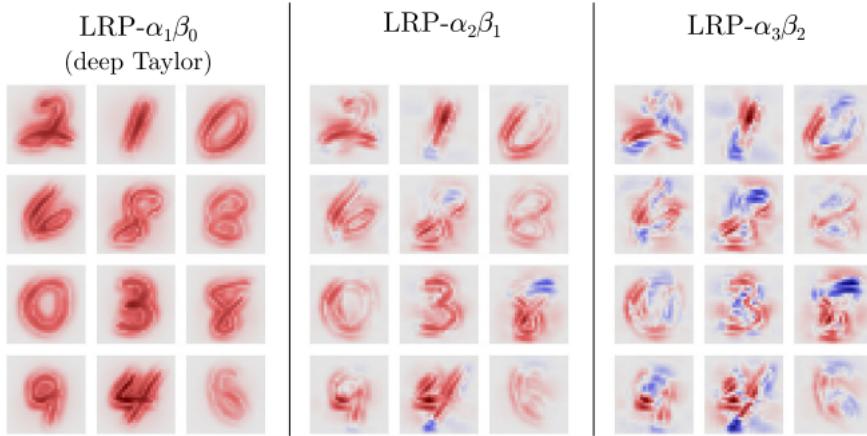
Nasledovné vzorce 2.2, 2.3, 2.4 [13] vyjadrujú spôsob výpočtu propagovanej relevancie medzi vrstvami. j a k sú jednotlivé vrstvy, pričom k je vrstva, z ktorej je relevancia R propagovaná. Parametre α a β upravujú, koľko pozitívnej (α) alebo negatívnej (β) relevancie je vytvorené počas fázy spätného šírenia 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 [13], výsledok použitia rôznych hodnôt α a β je znázornený na obrázku 2.11. $R_{j \leftarrow k}^+$ (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.

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

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

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

Výhodou LRP oproti iným metódam, ako napríklad dekonvolúciu je, že vysvetlenie (výsledná tepelná mapa) vytvorené technikou LRP je pre rôzne obrázky vždy



Obr. 2.11: Výsledné vysvetlenie (v podobe tepelnej mapy) vytvorené použitím LRP s rôznymi hodnotami α a β na dátovej sade MNIST. [13] Pozitívna relevancia je zobrazená červenou farbou [13]. Negatívna relevancia je zobrazená modrou farbou [13]. V prípade, že použijeme $\alpha = 1$ a $\beta = 0$ strácamo informáciu o tom, ktoré pixely negatívne (tj. sú proti výslednej predikcii) prispeli k výslednej predikcii (a opačne).

rôzne [15]. Naopak, pri dekonvolúcii je vysvetlenie vždy rovnaké pokiaľ v architektúre neurónovej siete neboli použité združovacie vrstvy (angl. pooling layers) [15]. Ďaľším rozdielom je (aj oproti analýze senzitivity), že vo výslednom vysvetlení LRP rozlišuje, ktoré vlastnosti pozitívne alebo negatívne prispeli k negatívnej predikcii.

2.2.6.3 RISE - Randomized Input Sampling for Explanation

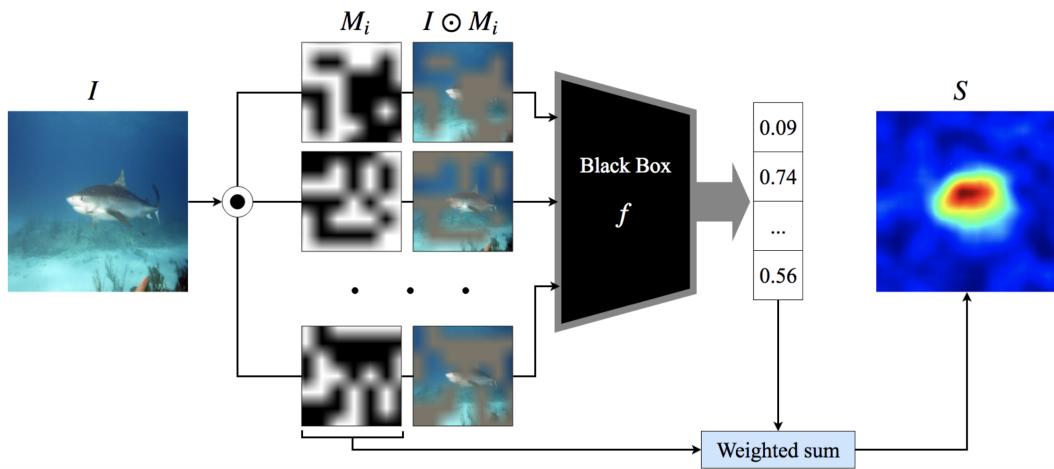
Túto metódu môžeme zaradiť medzi perturbačné metódy, keďže je tiež založená na zakrývaní jednotlivých častí obrazu a sledovaním zmeny výslednej predikcie modelu. Už z názvu modelu (*Randomized Input Sampling for Explanation*) je zrejmé, že táto metóda využíva náhodu na zakrývanie jednotlivých častí vstupného obrazu. Vstupný obraz je prekrytý náhodou maskou, ktorá je vytvorená nasledovne [16]:

- Je vytvorená náhodná binárna (tj. iba z bielej a čiernej farby) maska o malej

veľkosti (napríklad 8px x 8px).

- Táto maska je zväčšená (angl. upscaled) pomocou bilineárnej interpolácie [16] (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 2.12.

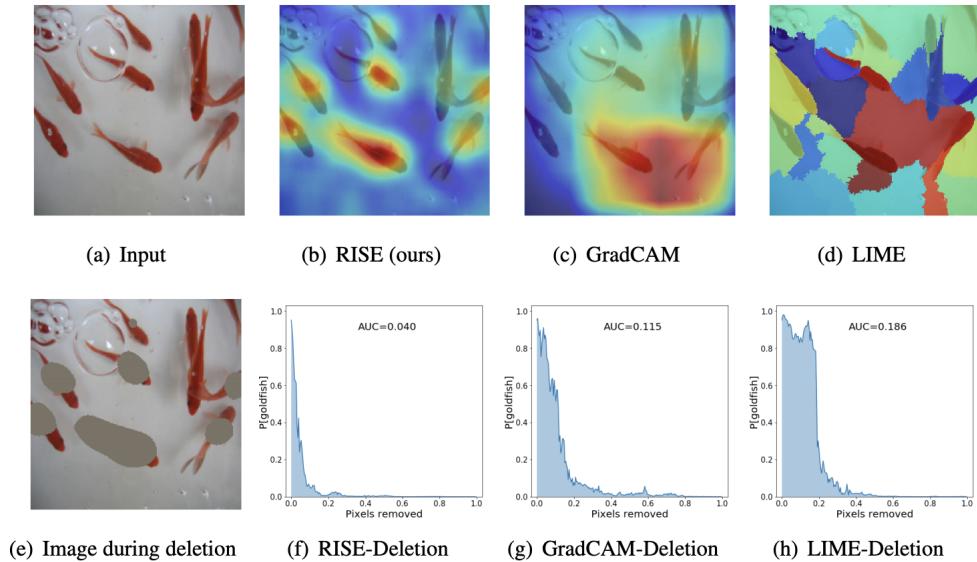


Obr. 2.12: Metóda *Rise*. [16] Vygenerované masky nahradzajú vstupný obrázok na, ktorý sú aplikované. Z výstupných predikcií jednotlivých masiek je nakoniec vypočítaná tepelná mapa.

Autori porovnali túto metódu s metódami *GradCAM* (Selvaraju et al. 2017) [17] a *LIME* (Ribeiro et al. 2016) [18]. Metóda *Rise* si oproti týmto dvom metódam počínala lepšie (Obr. 2.13). Vykonali niekoľko experimentov, v ktorých porovnali architektúry neurónových sietí *ResNet50* (He et al. 2016) [12] a *VGG16* (Simonyan; Zisserman 2014) [11] natrénované na dátových sadách PASCAL VOC07 (Everingham et al. 2010) [19] a MSCOCO2014 (Lin et al. 2014) [20]. Sledovali metriky *insertion* a *deletion* (Obr. 2.13). 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.

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



Obr. 2.13: Porovnanie metódy *Rise* s *GradCAM* alebo *LIME*. [16] V prvom riadku sú tepelné mapy jednotlivých metód pre vstup. V druhom riadku je znázornená porovávaná 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.

2.3 Využitie neurónových sietí pri odhalovaní Alzheimerovej choroby

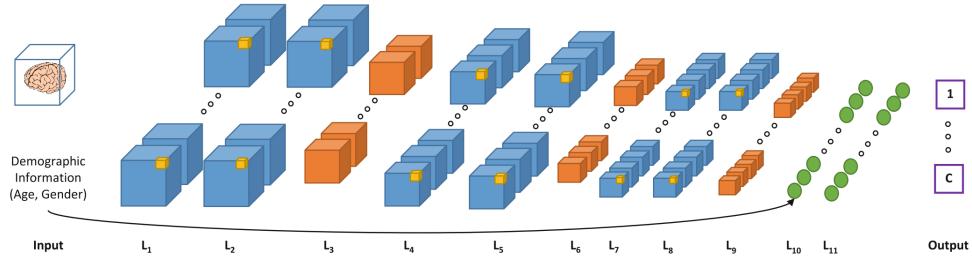
Neurónovým sieťam sa doposiaľ podarilo dosiahnuť veľmi dobré výsledky pri odhalovaní Alzhemiemerovej choroby. Ako vstup používajú rádiologické snímky ako sú z MRI či PET. Tieto rádiologické ukazovateľe sme bližšie popísali v sekciu 2.1.3. Okrem rádiolgických snímkov môžu byť vstupom do neurónovej siete demografické údaje o pacientovi, či výstupy z rôznych klinických alebo kongitívnych testov. Tačéto údaje o pacientoch obsahuje populárna dátová množina *ADNI-1* [21].

Neurónové siete natrénované na predikciu Alzheimerovej choroby sa líšia najmä v:

- **predspracovaní** - vstupné dátá sú zmenšené/zväčšené rôznymi algoritmiami na rôzne veľkosti, častokrát sa z rádiologických snímkov odstraňuje lebka
- **type vstupných dát** - môžu to byť rádiologické snímky (MRI, PET), vlastné črty extrahované z rádiologických snímkov (MRI, PET), alebo kombinácia takého snímkov/črt, s inými, napríklad demografickými údajmi
- **architektúre** - môžu to byť konvolučné siete s 2D konvolúciami (v prípade, že sa používa iba časť rádiologickej snímky, alebo vlastné črty) alebo 3D konvolúciami (ak je vstup celý rádiologický snímok, angl. "full volume"), alebo iné architektúry ako ResNET (reziduálne neurónové suete) alebo VGG
- **ako boli natrénované** - pri niektorých neurónových sieťach autori využili učenie prenosom (angl. transfer learning) a rôzne spôsoby augmentácie vstupov

Ako príklad 3D konvolučnej neurónovej siete uvediem neurónovú sieť od Esmaeilzadeh et al. s presnosťou **94.1%** (a s F_2 skóre 0.93) na populárnej dátovej množine s názvom *ADNI-1* (Obr. 2.14). 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.



Obr. 2.14: Architektúra konvolučnej neurónovej siete použitej pri detekcii Alzheimerovej choroby. [22] 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.

V prípade klasifikácie do všetkých troch tried - CN, MCI a AD autori tejto práce dosiahli horsie výsledky oproti binárnej klasifikácii. Ich model dokázal správne zaradiť pacienta s presnosťou **61.1%** (a s F_2 skóre 0.62) [22]. 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 trénovaním dvoch rôznych modelov, jedného jednoduchšieho a druhého zložitejšieho. Lepší bol jednoduchší model, pretože neboli tak náchylní na pretrénovanie. V týchto modeloch použili dropout, l_2 regularizáciu a augmentované dátá (obrázky otočili po osi x). Tieto "vylepšenia" pridávali postupne 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 tkivo mozgu (napr. lebka) technikou s názvom BET (Smith 2002) [23], pretože z nich sa Alzheimerova choroba nedá diagnostikovať.

Niekteré práce (Suk et al. 2016) sa zaoberali 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% [24]. 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. V tejto práci autori taktiež vykonali odstránenie lebky zo vstupných snímkov počas fázy predspracovania.

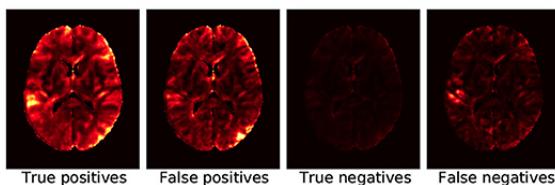
Učenia prenosom (angl. transfer learning) je veľmi dobrým spôsobom na zrýchlenie trénovania a zlepšenie úspešnosti modelu. Hosseini-Asl et al. využili učenie prenosom a to tak, že najskôr netrénovali 3D konvolučný autoenkdér, ktorý mal za úlohy rekonštruovať vstup - tj. vstupný radiologický snímok. Z tohto autoenkdéra zobraťi jednu jeho časť - enkdér za ktorý dali konvolučné vrstvy, ktoré dotrénovali na detekciu Alzheimerovej choroby. Enkdér teda slúžil na ektrakciu črt.

Neurónové siete sa v niektorých prácach používajú v kombinácii s inými algoritmami strojového učenia. Suk et al. použili kombináciu riedkych regresných modelov (angl. sparse regression models) a 2D konvolučnej neurónovej siete, kde výstupy z týchto regresiných modelov slúžili ako vstup do neurónovej siete.

2.3.1 Vysvetľovanie rozhodnutí neurónových sietí detegujúcich Alzheimerovu chorobu

Existujúce práce sa už zaoberali 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 siet 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. 2.15). 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 Alzheimerovou chorobou a zdravých jedincoch.



Obr. 2.15: Priemerná relevancia (z metódy LRP - $\beta = 0$)
pozorovaní podľa spôsobu zaradenia výslednej predikcie

Najviac relevancie je na miestach so žltou farbou. [27]

2.4 Spracovanie obrazu

Keďže pri diagnostike Alzheimerovej choroby sa pracuje s rádiologickými snímkami, čo sú trojrozmerné obrazové dátá, pri jej detekcii neurónovými sieťami je potrebné tieto dátá spracovať technikami spracovania obrazu.

Metódy spracovania obrazu podľa Chen [28] rozdeľujeme do nasledovných kategórií:

- vylepšovanie obrazu (angl. image enhancement)
- rekonštrukcia obrazu (angl. image restoration)
- analýza obrazu (angl. image analysis)

- kompresia obrazu (angl. image compression)

Pri **vylepšovaní obrazu** je obraz upravovaný predovšetkým heuristickými technikami [28], môže sa napríkald jednať o upravenie jasu, kontrastu alebo farieb. Cieľom **rekonštrukcie obrazu** je zrekonštruovať poškodené časti obrazu, napr. pri fotografiách to môžu byť ich vyblednuté časti. Metódy **analýzy obrazu** umožňujú obraz spracovať tak, že je možné z neho automaticky získať (extrahovať) informácie [28]. Príkladmi analýzy obrazu je segmentácia obrazu, extrakcia hrán alebo analýza textúry. **Kompresia obrazu** umožňuje zmenšenie veľkosti obrazu znižovaním počtom potrebných bitov na jeho reprezentáciu [28]. Môže sa jednať o zmenšenie rozmerov obrazu, alebo počtu farieb potrebných na jeho reprezentáciu.

V našej doméne budeme pracovať so všetkými týmito technikami. Ako príklad môžem uviesť odstránenie takých častí obrazu, ktoré nepredstavujú mozgové tkanivo (BET - Smith 2002). Táto technika je kombináciou analýzy obrazu - identifikácia častí na odstránenie a vylepšenia obrazu - samotné odstránenie tých častí. Kompresia obrazu sa používa, v časti predspracovania pred tým ako je samotný snímok použitý ako vstup do neurónovej siete. Metódy rekonštrukcie obrazu sa bežne v tejto oblasti nepoužívajú, avšak my by sme ich chceli v našej práci použiť pri vytváraní novej metódy, preto sa im budeme bližšie venovať.

2.4.1 Rekonštrukcia obrazu

Metódy rekonštrukcie obrazu, alebo inak nazývané aj dokreslenia obrazu (angl. inpainting), podľa Ravi; Pasupathi; Muthukumar.; Krishnan [29] môžeme zaraďať do nasledovných kategórií:

- dokresľovanie založené na syntéze textúr
- poloautomatické a rýchle digitálne dokresľovanie
- dokresľovanie založené na parciálnej diferenciálnej rovnici
- dokresľovanie na základe predlohy a vyhľadávania

- hybridné dokresľovanie

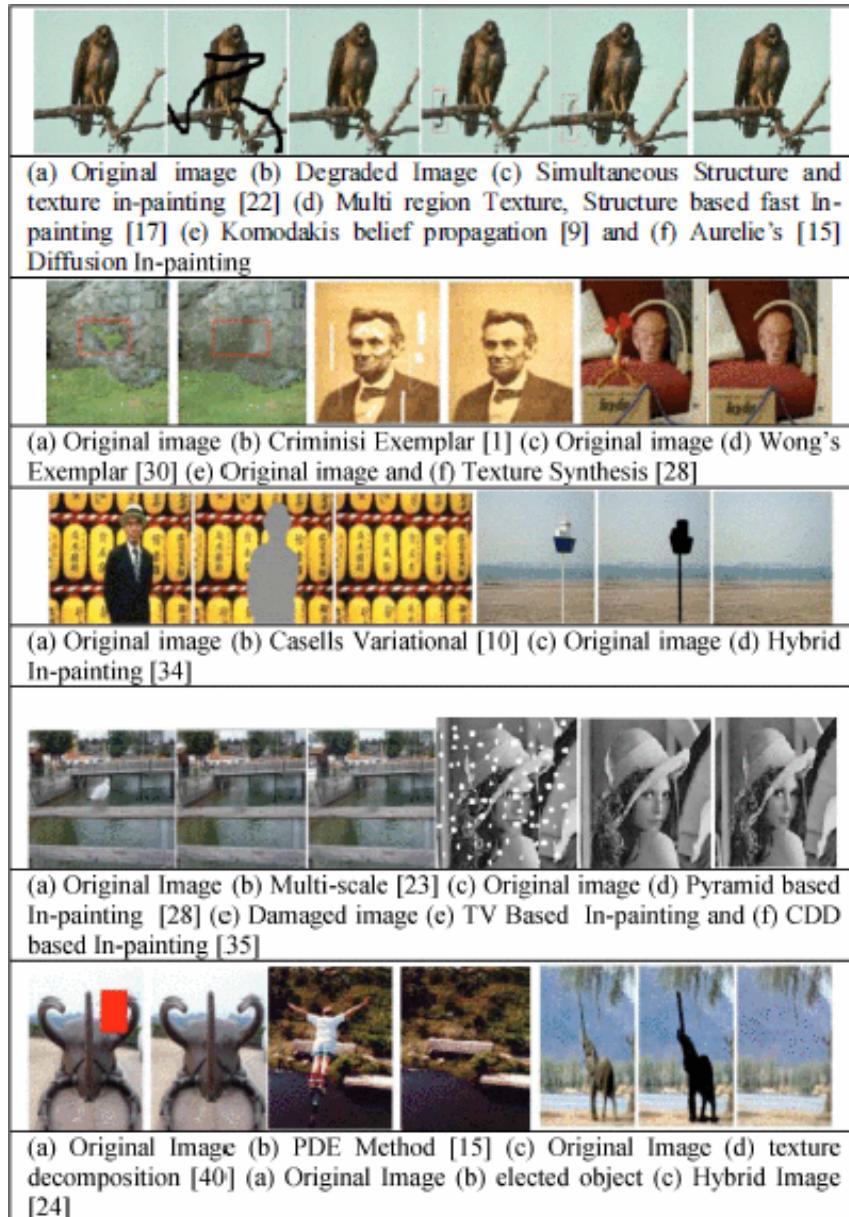
Tieto metódy sa líšia rýchlosťou dokresľovania, schopnosti dokreslovať veľké/malé plochy a predovšetkým kvalitou dokreslenia. Metódy dokresľovania založené na syntéze textúr fungujú dobre pre väčšie chýbajúce oblasti, avšak v ich výsledku môžu vzniknúť nežiadúce hrany [29]. Dokresľovanie na základe predlohy má zas problémy so zakrivenými štruktúrami [29]. Obr. 2.16 zobrazuje príklady použitia niektorých techník dokreslenia obrazu.

2.5 Zhrnutie

Alzheimierova choroba je bez pochyby veľmi nebezpečnou chorobou, keďže nie je "iba" o strate pamäti ale patrí k častým 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 štruktúry 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 Alzheimerevej 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



Obr. 2.16: Príklady dokreslenia obrázkov rôznymi metódami [29].

použitím informácií o veku a pohlaví pacienta. Keďže pravdepodobnosťou 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

Kapitola 2. Analýza

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 podrobiť dôkladnej analýze jej rozhodnutí. Osobne si ale myslím, že v produkčnom modeli by sa tento atribút mal vynechať.

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 vysvetľovania rozhodnutí modelov (napr. RISE, LIME...).

Bežnému používaniu 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 zisťujeme, ako vyzerá vzorové pozorovanie pre jednu z tried, ktorú klasifikuje. Vysvetľovaním jej rozhodnutí zas zisťujeme 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. Žiadna 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. Cieľom je teda vytvoriť novú metódu, ktorá vytvára presnejšie vysvetlenia ako iné metódy, alebo vylepsiť existujúcu metódu, ktorá vytvára presnejšie vysvetlenia ako metóda, z ktorej vychádza.

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

Pri neurónových sieťach detekujúcich Alzheimerovu chorobu je dôležité, aby sa naučili klasifikovať pacientov na základe relevantných 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. Cieľom je teda určiť správnosť modelu detegujúceho Alzheimerovu chorobu pomocou vytvorenej metódy pre vysvetľovanie rozhodnutí neurónovej siete.

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. 2.2.6.3). 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 jednej hodnote. Keďže metóda *Rise* bola pôvodne použitá na obrázky vo farebnej schéme RGB, tento prekryv sa zvyčajne robí v čiernej farbe - tj. v ($r = 0, g = 0, b = 0$). MRI snímky nepoužívajú žiadnu farebnú schému, ale zachytávajú intenzitu (hodnoty sú zväčša reálne čísla). V tomto prípade môžeme zakrývať maximálnou alebo minimálnou hodnotou (minimálna hodnota je ekvivalentná RGB v prípade šedej). Toto zakrutie môže byť práve ďalším zdrojom zmätenia pre neurónovú sieť, keďže úbytky tkaniva sú vyjadrené nízkymi hodnotami na snímkoch. Preto navrhujeme zakrývané miesta dokresliť určitou metódou spracovania obrazu (Sek. 2.4) alebo na zakrytie použiť inú hodnotu. Pôvodná metóda bola ale narvhnutá pre obrázky (tj. 2D) a nie 3D volumetrické dátá, preto budeme musieť metódu *Rise* upraviť aby vedela pracovať

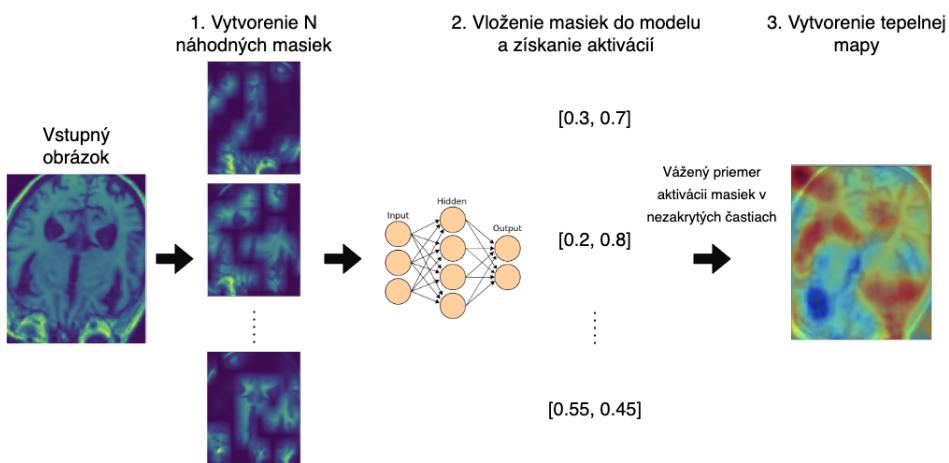
s 3D dátami - tj. budeme generovať 3D masky a podobne.

4.1 RISEI - Randomized Input Sampling for Explanation with Inpainting

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

Kedžže metóda vychádza už z existujúcej metódy, časť našej metódy je samozrejme rovnaká. Proces vytvorenia vysvetlenia klasifikácie (Obr. 4.1) do triedy T pre obrázok O modelom je teda nasledovný:

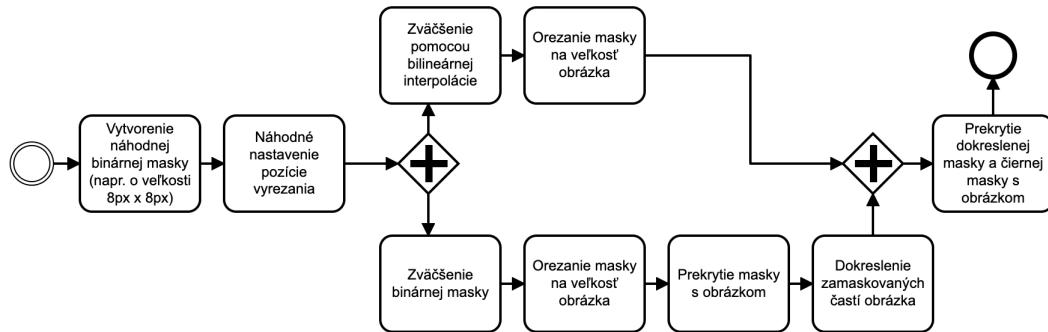
1. Vytvorenie N náhodne zamaskovaných obrázkov z obrázka O .
2. Vloženie zamaskovaných obrázkov do modelu a následné získanie pravdepodobností pre triedu T .
3. Vytvorenie a vizualizácia vysvetlenia pomocou tepelnej mapy.



Obr. 4.1: Proces vysvetlenia klasifikácie - vytvorenia tepelnej mapy.

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

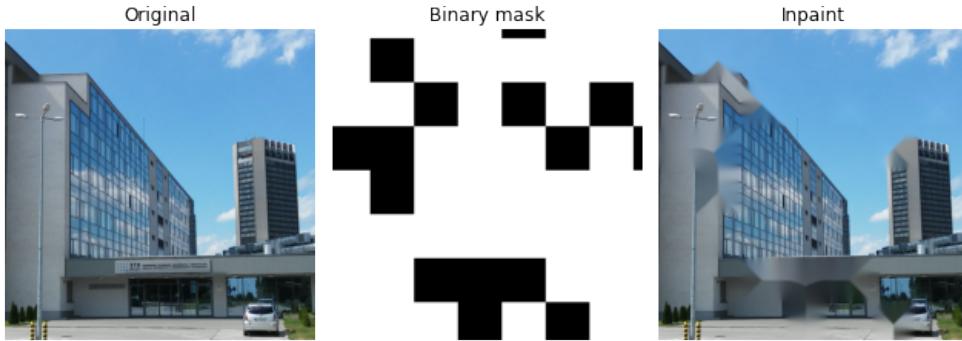
Vytvorenie náhodne zamaskovaných obrázkov. Vytvorenie náhodne zamaskovaných obrázkov tiež pozostáva z niekoľkých krokov, pričom niektoré z nich môžu bežať paralelne. Tento sme znázornili diagramom (Obr. 4.2). Masky sa vytvárajú paralelne, pretože ”čierna” maska ma jemné hrany a na dokreslenie potrebujeme naopak masku s ostrými hranami.



Obr. 4.2: BPMN diagram generovania jedného obrázka prekrytého maskou

Oproti metóde *Rise* vytvárame o jednu masku naviac, a teda je originálny obrázok prekrytý s 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 ”čierной” maskou a tiež s použitím oboch v rôznych pomeroch.

Vytvorenie ”čiernej” masky je rovnaké, ako pri metóde *Rise*. Dokreslená maska vznikne dokreslením zakrytych (zamaskovaných) častí obrázka pomocou jedného z algoritmov na dokreslovanie (angl. inpainting). Tieto algoritmy sme popísali v sekcií 2.4 Spracovanie obrazu. Obrázok 5.1.1 je príkladom dokreslenia častí vzorového obrázka na základe masky náhodne vygenerovanej masky (tentotýkľad je v 2D, naša metóda bude pracovať s 3D). V našej metóde budeme experimentovať s rôznymi hodnotami prekrycia (priemer, maximum, minimum, medián).



Obr. 4.3: 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 prichľadnosťou.

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 voxel $[x, y, z]$ snímku, kde n je počet všetkých zamaskovaných snímkov. Funkcia $p(k, x, y, z)$ vracia vracia predikciu (tj. aktiváciu v kontexte neurónových sietí) pre predikovanú triedu (v prípade binárnej klasifikácie) z modelu pre zamaskovaný snímok k . Funkcia $c(k, x, y, z)$ vracia mieru zakrytie/dokreslenia maskou, pričom $H(c) = \langle 0, 1 \rangle$, kde 1 znamená úplné prekrytie/dokreslenie a 0 žiadne prekrytie/dokreslenie. Rovnako, ako metóde *Rise*, počítame vážený priemer.

$$I_{x,y,z} = \frac{\sum_k^n p(k, x, y, z) * (1 - c(k, x, y, z))}{\sum_k^n p(k, x, y, z)} \quad (4.1)$$

Navrhovaná metóda do originálnej metódy 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 jednou zo slabín tejto metódy. Takisto aj samotná dokreslená časť obrázka môže byť príčinou zmätenia neurónovej siete.

4.2 Overenie riešenia

Našu metódu budeme najskôr porovnávať s originálnou metódou RISE (tj. či sa nám podarilo vytvoriť 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 [22], ale aj pri vysvetľovaní rozhodnutí neurónovej siete pomocou LRP [27]. 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 navrhnutej metódy, avšak je nutné aby metóda generovala kavlitné tepelné mapy.

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 táto mapa odzrkadľuje to, na základe čoho sa model rozhoduje. Toto budeme merať metrikami *insertion (AUC)* a *deletion (AUC)*, ktoré sme bližšie popísali v sekcii 2.2.6.3. Táto metrika nám povie, ako dobrá je naša metóda na vysvetľovanie.

Kedžže naša metóda generuje tepelné mapy pomocou vygenerovania veľkého množstva náhodných masiek, je preto vhodné skúmať ako sú tieto tepelné mapy konzistentné pri niekoľkých požitiach metódy na tom istom MRI snímku. Konzistentnosť máp môžeme merať pomocou podobnosti medzi jednotlivými tepelnými mapami (napr. ako súčet absolútnej hodnôt rozdielov medzi voxelmi v oboch tepelných mapách). čím je táto podobnosť väčšia, tým je metóda pri generovaní máp viac konzistentná.

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 akéj 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 natrénovaného modelu, tút budeme meráť pomocou metrík z práce od Böhle et al. v ktorej sa autori zaoberali vyhodnocovaním tepelných máp z metódy LRP. Tieto metriky sú nasledovné (relevancia je v našom prípade teplota na tepelnej mape):

- súčet relevancie v jednotlivých častiach mozgu (podľa segmentačných másk) pre AD a CN
- hustota relevancie v jednotlivých častiach mozgu (podľa segmentačných másk) pre AD a CN, berie ohľad na veľkosť danej časti mozgu
- prírastok relevancie v jednotlivých častiach mozgu (podľa segmentačných másk) vypočítaný ako pomer priemernej relevancie každej triedy v danej časti mozgu

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 sietach detegujúcich Alzheimerovu chorobu s cieľom odhaľovania nesprávnych rozhodnutí.

5. Implementácia

5.1 Metóda RISEI

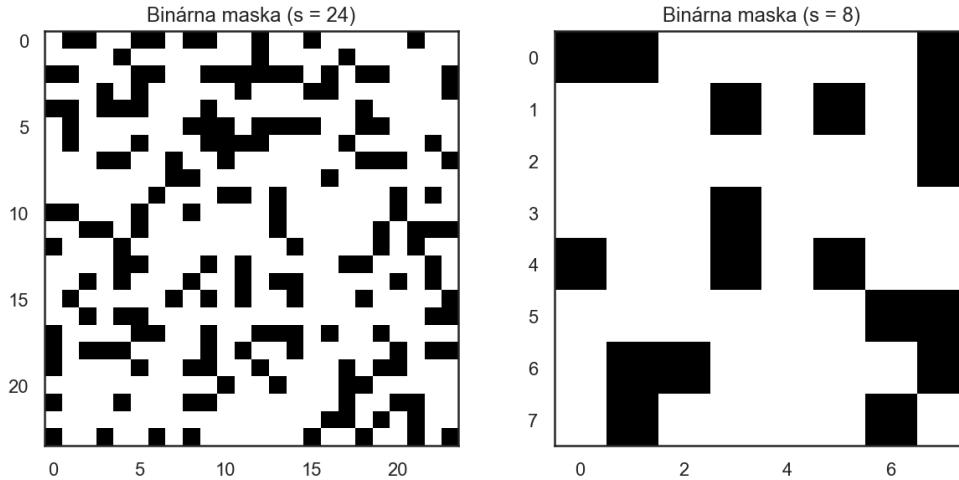
Metódu RISEI sme sa rozhodli implementovať v jazyku Python, keďže plánujeme používať knižnice pre strojové učenie akými sú *tensorflow* či *scikit-learn*.

5.1.1 Generovanie masiek

Na základe BPMN diagramu (Obr. 4.2) sme implementovali proces generovania masiek. Generovanie masiek prebieha paralelne vo viacerých procesoch použitím knižnice Python *multiprocessing*. Metóda RISE pracuje s trojrozmernými dátami, avšak diagramy v tejto sekcii zobrazujú snímky a masky v 2D (konkrétnie určitú vrstvu z 3D snímku) kvôli jednoduchšej vizualizácii. V tejto sekcii popíšeme jednotlivé kroky generovania masiek.

Vytvorenie náhodnej binárnej masky Náhodné binárne masky generujeme pomocou knižnice *numpy*. Pomocou nasledovného kódu vygenerujeme N náhodných masiek 3D binárnych matice. Obr. zobrazuje takúto binárnu maticu, ale v 2D. *size* (veľkosť) a *probability* (pravdepodobnosť) sú hyper-parametrami RISEI metódy. *size* hovorí o veľkosti generovanej masky, čím je toto číslo väčšie tým bude výsledná maska viac fragmentovaná na malé plochy. *probability* hovorí o tom, s akou pravdepodobnosťou daná plocha neprekrytá maskou. RISE používa predvolenú hodnotu *size* = 8.

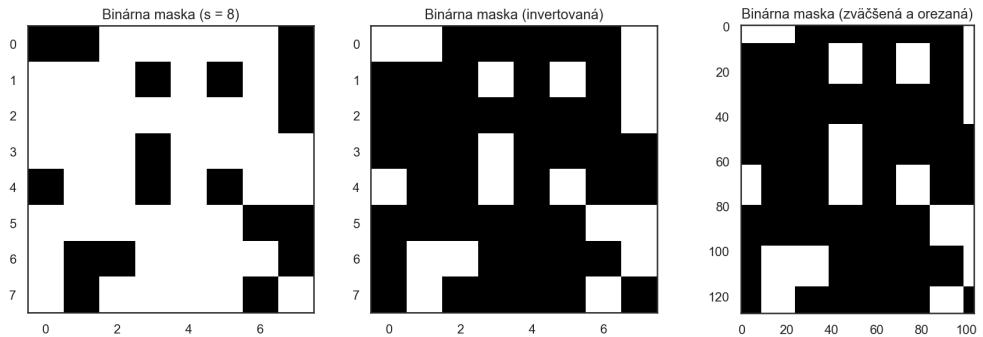
```
binary_masks = np.random.rand(N, size, size, size) < probability
```



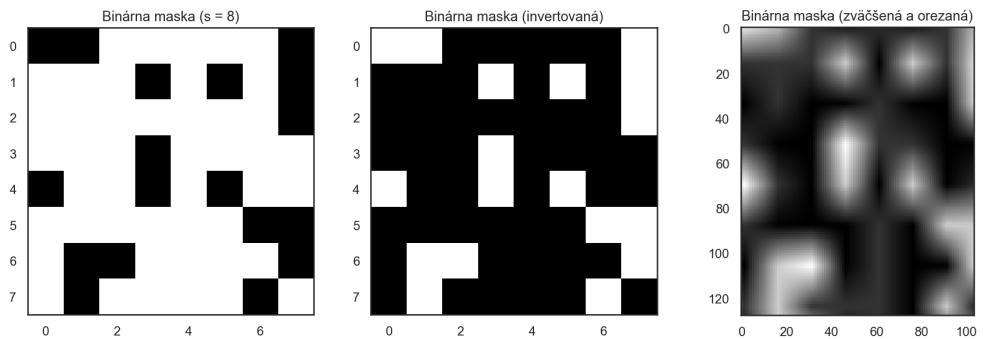
Obr. 5.1: Porovnanie dvoch binárnych masiek s rôznou veľkosťou (*size*), čím väčšia veľkosť, tým je obrázok viac fragmentovaný.

Náhodné nastavenie pozície vyrezania, zväčšenie binárnej masky a orezenie na veľkosť obrázka Binárnu masku zväčšíme na veľkosť vstupného snímku plus menší offset (o veľkosti *size*). Následne zo zväčšenej masky na náhodnej pozícii vyrežeme masku o veľkosti vstupného snímku (Obr. 5.2). Táto maska určuje, ktoré miesta na snímku bude treba dokresliť - biele miesta, čiže jednotky. Tento krok v pôvodnej implementácii RISE nie je.

Zväčšenie pomocou bilineárnej interpolácie a orezanie masky na veľkosť obrázka Tak ako v poôvodnej implementácii RISE, vytvoríme "čiernu" masku na zakrytie častí obrázku. Pôvodnú binárnu masku pomocou bilineárnej interpolácie (funkcia *resize* z knižnice *scikit-learn*) zväčšíme na veľkosť vstupného snímku plus menší offset, následne vyrežeme na náhodnej pozícii masku o veľkosti vstupného snímku (táto náhodná pozícia je rovnaká ako pri orezávaní binárnej masky bez interpolácie, preto je v BPMN diagrame v samostatnom kroku).



Obr. 5.2: Vygenerovaná maska je zväčšená a orezaná na veľkosť vstupného snímku (ten je o veľkosti [104, 128, 104] pričim na obrázkoch je vizualizovaná druhá a tretia dimenzia). Úplne vľavo je binárna maska o veľkosti 8. V strede je invertovaná binárna maska (kvôli ďalšiemu pracovanou s ňou) a vpravo je orezaná binárna maska o veľkosti vstupného snímku.



Obr. 5.3: Vygenerovaná maska je zväčšená pomocou bilineárnej interpolácie a orezaná na veľkosť vstupného snímku (ten je o veľkosti [104, 128, 104] pričim na obrázkoch je vizualizovaná druhá a tretia dimenzia). Úplne vľavo je binárna maska o veľkosti 8. V strede je invertovaná binárna maska (kvôli ďalšiemu pracovanou s ňou) a vpravo je orezaná interpolovaná ”čierna” maska o veľkosti vstupného snímku.

Prekrytie masky s obrázkom a dokreslenie zamaskovaných častí obrázka
 Kedže pracujeme nad trojrozmernými dátami, pokúsili sme sa použiť dokreslovanie obrázka v 3D. Na to sme sa pokúsili použiť funkciu *inpaint* s knižnicou *scikit-image*, avšak dokreslenie jednej masky bolo veľmi časovo náročné (trvanie bolo až v minútach kde dokreslenie v 2D je v sekundách) a my ich potrebujeme generovať

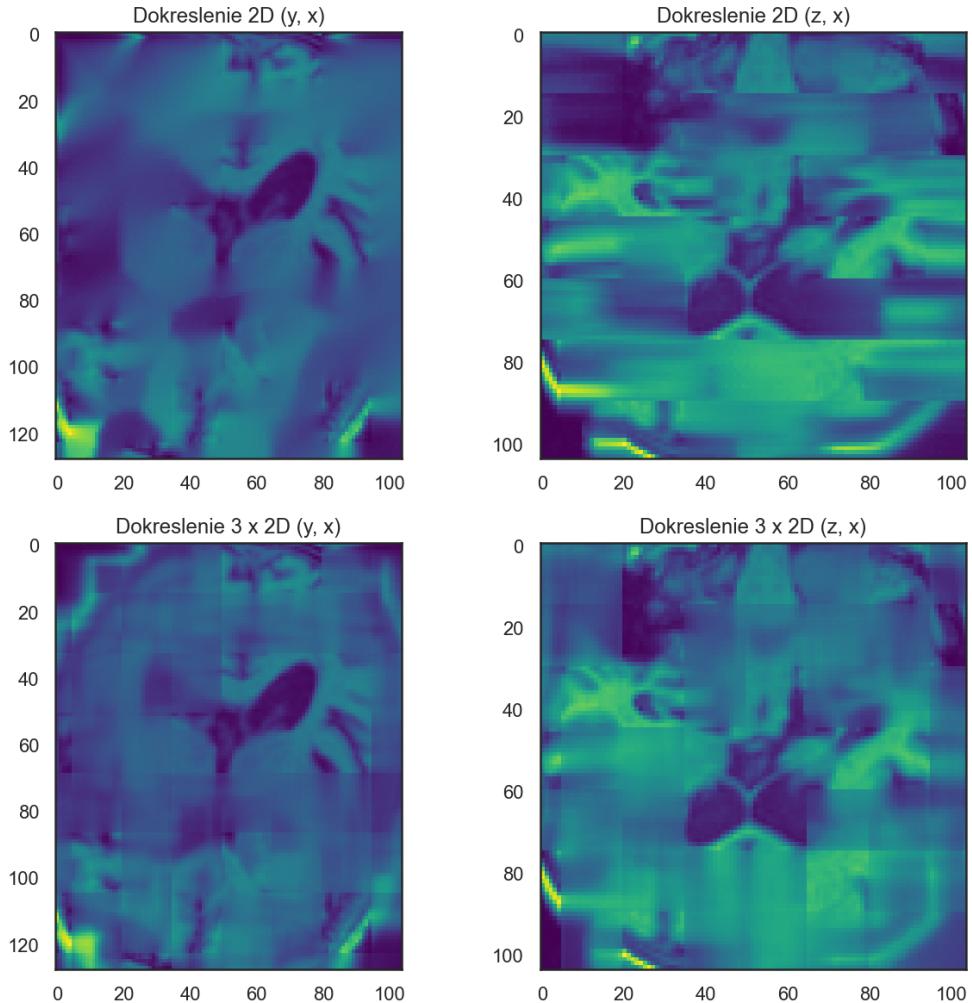
tisíce, preto sme trojrozmerného dokreslovania upustili.

Dokreslovanie dvojrozmerných snímkov z 3D snímku má avšak svoje nevýhody. Nech máme snímky o veľkosti $[z, y, x]$, pri 2D dokreslení musíme dokreslovať z snímkov o veľkosti $[y, x]$ (alebo y snímkov o veľkosti $[y, x]$, alebo x snímkov o veľkosti $[y, z]$). Pri takomto dokreslovaní, dokreslenie z pohľadu $[y, x]$ vyzerajá byť správne, avšak z iného pohľadu, napr. $[z, x]$ sa javí byť dokreslenie nesprávne, najmä kvôli vzniknutým ostrím hranám (Obr. 5.4). Toto sme sa pokúsili obýť tak, že dokreslujeme zo všetkých troch pohľadov a robíme priemer pre každý voxel zo všetkých troch dokreslení. Takto je výsledok o niečo lepší, tj. z každej strany je dokreslenie lepšie ako nesprávne dokreslenie z 2D ale o niečo horšie ako správne dokreslenie z 2D. Na označenie miest, ktoré treba dokresliť sme použili zväčšenú binárnu masku (Obr. 5.2). Dokreslenie vykonávame funkciou *inpaint* z knižnice *cv2* (*Open CV*). Používame dokreslovací algoritmus *cv2.INPAINT_TELEA*, keďže pomocou neho sme dosahovali vizuálne najlepšie výsledky. Funkcia *cv2.inpaint* vyzaduje ako parameter *inpaint_radius* (Obr. 5.6), čo je jedným z hyper parametrov našej metódy.

Keďže sa pôvodná implementácia RISE prekrýva miesta tak, aby nevznikali ostré hrany medzi zakrytím miestom a pôvodným obrázkom, a teda vznikol plynulý prechod, aj pri dokreslení vytvárame plynulý prechod medzi dokreslením a pôvodným obrázkom (Obr. 5.5). Tento prechod je implementovaný nasledovne.

```
# binary_mask int[z, x, y] - upsized binary mask
# image float[z, x, y] - original image
# mask float[z, x, i] - upsized and interpolated binary mask
# inpaint_radius int
inpainted = cv.inpaint(image, binary_mask, inpaint_radius,
                       cv2.INPAINT_TELEA)
inpainted_blend = image * mask + inpainted * (1 - mask)
```

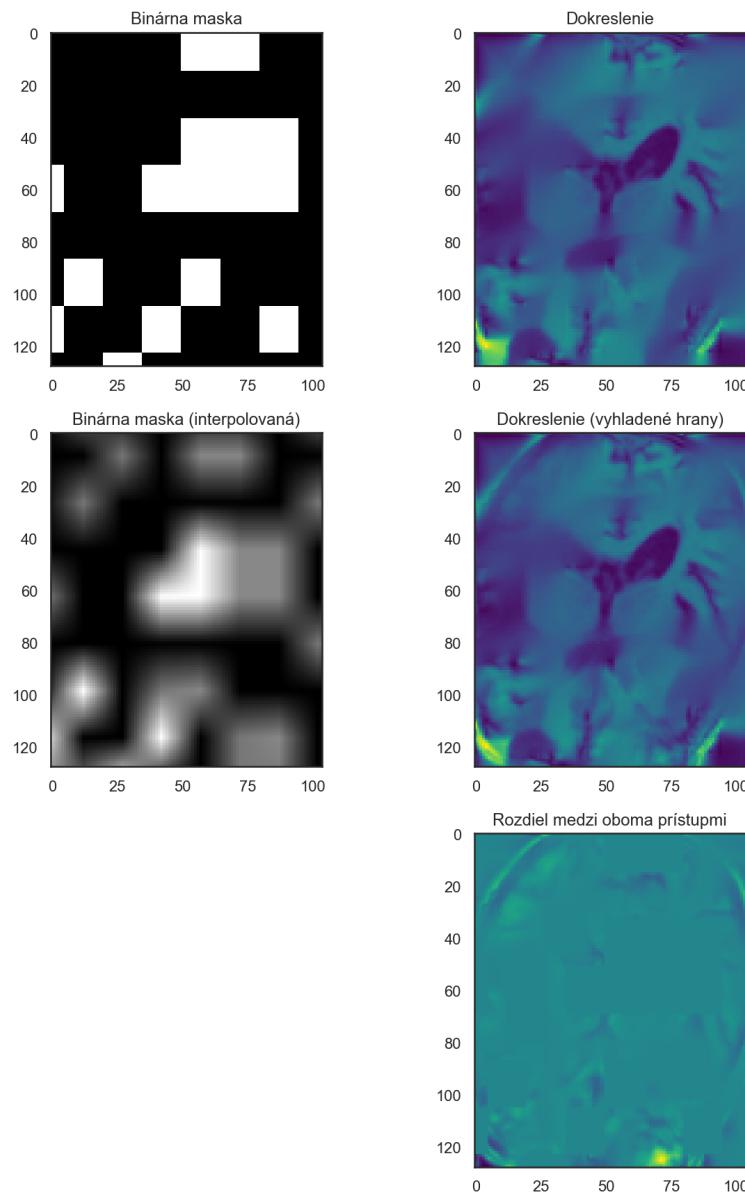
Prekrytie dokreslenej masky a čiernej masky s obrázkom Keďže prekrývam tri rôzne vrstvy - originálny snímok, čiernu masku a dokreslený snímok môžem



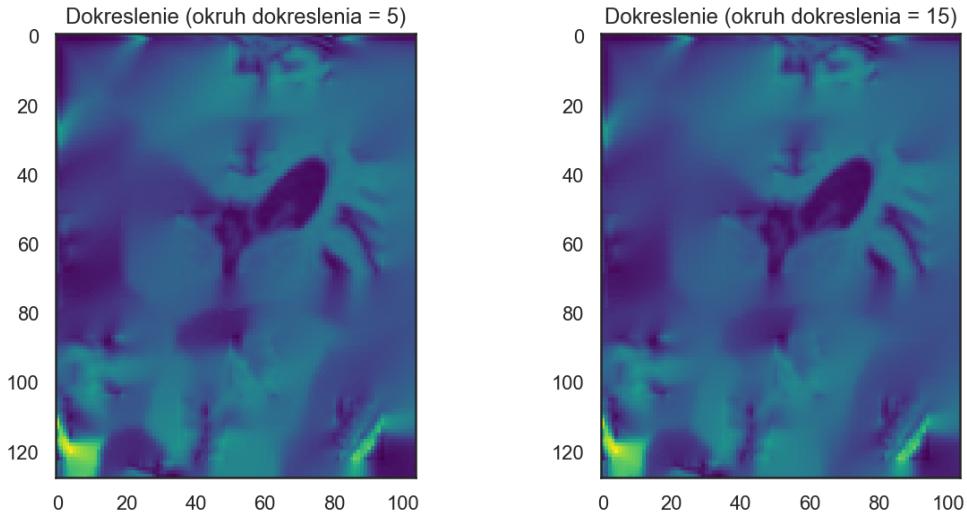
Obr. 5.4: Porovnanie 2D dokreslenia (iba v jednej dimenzii) a spriemerovaného 3x 2D dokreslenia (v každej dimenzii). Použitie iba 2D dokreslenia je kvalitné iba v jednej dimenzii a v ostatných je deštruktívne - vytvára ostré hrany. Použitie 3x 2D dokreslenia a spriemerovanie pre každý voxel produkuje celkom dobré dokreslenia po všetkých dimenziách.

tieto vrstvy skombinovať v rôznom pomere a tým vytvoriť nový obrázok.

Toto som implementoval zavedením parametrov $b1$ a $b2$ (skratka od slova prechod, angl. blend), ktoré hovoria o pomere medzi originálnym snímkom a dokresleným snímkom, a originálnym snímkom spojeným s dokreslením a čierrou maskou (Obr.



Obr. 5.5: Príklad vyhladzovania hrán dokreslenia - splynutie dokreslenia s pôvodným snímkom (štvrty snímok). Druhý snímok zobrazuje ostré hrany po dokreslení - bez splývania s obrázkom. Piaty snímok zobrazuje rozdiel medzi oboma prístupmi. Môžeme si všimnúť, že na obrázku sú viditeľné miesta, kde sa nachádza prechod na interpolovanej binárnej maske. O tieto miesta (informácie) je dokreslenie s vyhladenými hranami "bohatšie".

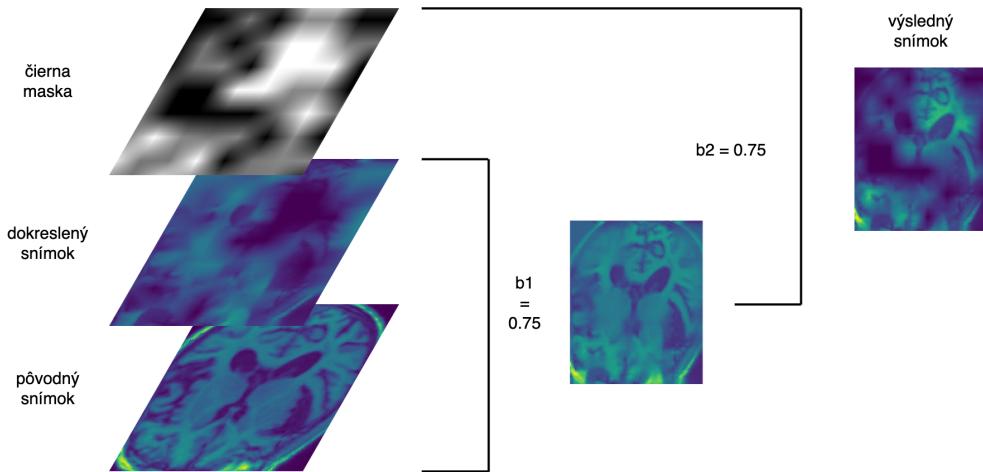


Obr. 5.6: Porovnanie okruhov dokreslenia (parameter *inpaint_radius*), rozdiel vo výsledku nie je veľmi viditeľný, avšak s väčším okruhom dokreslenia je generovanie rádovo pomalšie. (pri generovaní bolo vypnuté splynutie dokreslenia so snímkom aby bol rozdiel aspoň trochu viditeľný)

5.7). Pri týchto parametroch platí, že $0 \leq b1, b2 \leq 1$. Takto zadefinované parametre mi umožňujú vytvoriť zakaskovaný snímok iba s čierrou maskou ($b1 = 0$, $b2 = 1$) či iba s dokreslením ($b1 = 1$, $b2 = 0$).

Názov "čierna" maska pochádza z pôvodnej implementácie RISE, kde sa obrázok prekrýval čierrou maskou. V našej implementácii neprekrývame farbou, ale hodnotou, tj. "čierna" je hodnota 0 (minimum). Okrem použitia hodnoty 0, môžeme použiť aj 1, *priemer* či *medián* (toto je ďalším hyper-parametrom našej metódy). Zjednodušená (a menej efektívna, v produkčnej implementácii sa niektoré inštrukcie nevykonávajú keď $b1$ je 0 alebo $b2$ je 0) implementácia spojenia jednotlivých vrstiev vyzerá nasledovne.

```
# image float[z, x, y] - original image
# inpainted_blend float[z, x, y] - inpainted image
# mask float[z, x, i] - upsized and interpolated binary mask
# b1 float <0, 1>
# b2 float <0, 1>
```



Obr. 5.7: Príklad, ako vyzerá spojenie originálneho snímku, dokresleného snímku a čiernej masky. V diagrame je zobrazený aj výsledok medzikroku spojenia dokresleného snímku a pôvodného snímku. Parametre boli nastavené na $b1 = 0.75$ a $b2 = 0.75$.

```
# b2_value string - what value use in "black" mask
(min/max/mean/median)

# merge with inpainted image
new_image = (1 - b1) * original_image + b1 * inpainted_blend

value = 0 # black
if b2_value == 'max':
    value = 1 # white
elif b2_value == 'mean':
    value = np.mean(original_image)
elif b2_value == 'median':
    value = np.median(original_image)
# merge with "black" mask
new_image = b2 * mask * new_image + (b2 * (1 - mask) * value)
```

Kompletný zoznam parametrov metódy RISEI sa nachádza v tabuľke 5.1.

Názov	Dátový typ	Popis
s	int	Veľkosť strany binárnej 3D matice.
p	float	Pravdepodobnosť, že plocha nebude prekrytá maskou.
b1	float	Miera prekrytia medzi originálnym snímkom a dokresleným snímkom.
b2	float	Miera prekrytia s "čierrou" maskou.
b2_value	string	Hodnota "čiernej" masky, môže to byť minimum, maximum, medián, priemer.
in_paint_radius	float	Polomer dokreslenia algoritmom z knižnice OpenCV.

Tabuľka 5.1: Zoznam parametrov metódy RISEI.

5.1.2 Vytvorenie tepelných máp

Na základe návrhu (Sekcia 4.1) sme implementovali vytváranie tepelných máp. Keďže generovanie tepelnej mapy si vyžaduje vygenerovať veľký počet zamaskovaných snímkov, ktoré v istom momente musia byť všetky uložené v pamäti, generujeme a vyhodnocujeme zamaskované snímky v dávkach (angl. batch). Zdrojový kód nižšie, implementuje vytvorenie jednej tepelnej mapy.

```
# image_x float[z, x, y, 1] - original image
# masks_count int - how many masks are generated to create a heatmap
# batch_size - how many masks to evaluate on model
# risei_batch_size int - how many masks to generate in one batch
# seed int int - seed for mask generation
# cls_idx int - index of target class in model output vector
# model tf.keras.Model - instance of tensorflow model

risei = RISEI(s=8, p=0.5, b1=0.5, b2=0.5, b2_value='median',
              in_paint_radius=5)
heatmap = np.zeros(shape=image_x.shape[:3])
batch_count = math.ceil(masks_count / risei_batch_size)
weights = 0
```

```
for batch_idx in range(batch_count):
    batch_masks_count = min(risei_batch_size, masks_count - batch_idx *
                           risei_batch_size)

    # reshape input for RISEI since it works with [z, y, x] shape
    # batch_x float[z, x, y] - images to evaluate with masks already
    # applied
    # masks float[z, x, y] - interpolated binary masks (so we know which
    # places we inpainted or masked)
    batch_x, masks = risei.generate_masks(batch_masks_count,
                                           image_x.reshape(image_x.shape[:3]), seed=seed)
    y_pred_batch_x = model.predict(batch_x.reshape((-1, *image_x.shape)),
                                   batch_size=batch_size)

    for mask, y_pred in zip(masks, y_pred_batch_x):
        # invert the mask, since 1 is for no masking
        # y_pred is the activation for the input masked image on last
        # layer (softmax)
        heatmap = heatmap + y_pred[cls_idx] * (1 - mask)
        weights += y_pred[cls_idx]

heatmap = heatmap / weights
```

5.1.3 Vyhodnotenie tepelných máp

Zatiaľ sme implementovali, podľa návrhu riešenia (Sekcia 4.2.2.1), iba metriky *insertion* a *deletion*.

5.1.3.1 Metriky insertion & deletion

Tieto metriky fungujú tak, že postupne odstraňujeme/pridávame pixely z obrázku a tieto obrázky vkladáme do modelu a zaznamenávame si aktiváciu na poslednej

vrstve pre predikovanú triedu. V prípade obrázkov, a teda dvojrozmerných dát je to ešte výpočtovo zvládnuteľné, avšak v prípade trojdimenzionálnych rádiologických simkov to už môže byť problém. Naše vstupné snímky majú po zmenšení rozmer $[104, 128, 104]$, čiže ak aby sme odstraňovali zo snímku po jednom voxelovi, museli by sme vykonať 1 384 448 evaluácií pomocou nášho modelu (čo trvá niekoľko hodín, aj pri evaluovaní v maximálnych možných dávkach vzhľadom na pamäť grafickej karty). Preto sme sa rozhodli, že budeme pridávať po n (100) voxeloch v každom kroku. V prípade metódy insertion vkladáme do snímku plného núl (môžeme prípadne aj jednotiek). Kedže kód je rozsiahlejší, uvedieme len pseudokód.

```
method = 'insertion'

step_size = 150 # how many voxels to insert/delete in one evaluation
image_x, image_y = get_image()
image_y_pred = model.predict(image_x)
heatmap = get_heatmap()
voxels = get_ordered_voxels_by_heat(heatmaps)
sequence = get_images_sequence(voxels, step_size) # create a sequence
    from images where each next image has n inserted/deleted voxels
y_pred = []

for batch_x, batch_y in sequence:
    batch_y_pred = model.predict(batch_x)
    for y in batch_y_pred:
        y_pred.append(y)

auc = metrics.auc([i * step_size for i in range(len(y_pred))], y_pred) /
    get_voxels_count(image_x)
```

5.2 Model na detekciu Alzheimerovej choroby na základe MRI snímkov

6. Overenie riešenia

6.1 Experimenty

TODO: DP2

6.2 Záver

TODO: we need to compare our method with other methods like LRP or other occlusion methods
TODO: evaluate the consistency of heatmaps, the method several times for the same image and find out if heatmaps are consistent - this way we can find also optimal number of masks
TODO: compare with segmentation masks, if more heat is in important areas

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. GILPIN, Leilani H; BAU, David; YUAN, Ben Z; BAJWA, Ayesha; SPECTER, Michael; KAGAL, Lalana. Explaining explanations: An overview of interpretability of machine learning. In: *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. 2018, s. 80–89.
3. 2019. Dostupné tiež z: <http://www.alzheimer.sk/informacie/alzheimerovachoroba.aspx>.
4. DUTHEY, Béatrice. Background paper 6.11: Alzheimer disease and other dementias. *A Public Health Approach to Innovation*. 2013, s. 1–74.
5. KHAN, Tapan. *Biomarkers in Alzheimer's Disease*. Academic Press, 2016.
6. 2017. Dostupné tiež z: <https://www.alz.org/alzheimers-dementia/facts-figures>.
7. WORKING, G Biomarkers Definitions. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther*. 2001, roč. 69, č. 3, s. 89–95.
8. HAYKIN, Simon S et al. *Neural networks and learning machines/Simon Haykin*. New York: Prentice Hall, 2009.
9. 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).

10. O'SHEA, Keiron; NASH, Ryan. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*. 2015.
11. SIMONYAN, Karen; ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. 2014.
12. 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.
13. MONTAVON, Grégoire; SAMEK, Wojciech; MÜLLER, Klaus-Robert. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*. 2018, roč. 73, s. 1–15.
14. SIMONYAN, Karen; VEDALDI, Andrea; ZISSERMAN, Andrew. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*. 2013.
15. 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.
16. PETSIUK, Vitali; DAS, Abir; SAENKO, Kate. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*. 2018.
17. 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.
18. 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.

Literatúra

19. 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.
20. 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.
21. 2017. Dostupné tiež z: <http://adni.loni.usc.edu/>.
22. 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.
23. SMITH, Stephen M. Fast robust automated brain extraction. *Human brain mapping*. 2002, roč. 17, č. 3, s. 143–155.
24. 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.
25. HOSSEINI-ASL, Ehsan; KEYNTON, Robert; EL-BAZ, Ayman. Alzheimer's disease diagnostics by adaptation of 3D convolutional network. In: *2016 IEEE International Conference on Image Processing (ICIP)*. 2016, s. 126–130.
26. SUK, Heung-Il; LEE, Seong-Whan; SHEN, Dinggang; INITIATIVE, Alzheimer's Disease Neuroimaging et al. Deep ensemble learning of sparse regression models for brain disease diagnosis. *Medical image analysis*. 2017, roč. 37, s. 101–113.
27. 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.
28. CHEN, Wai Kai. *The electrical engineering handbook*. Elsevier, 2004.

Literatúra

29. RAVI, S.; PASUPATHI, P.; MUTHUKUMAR., S.; KRISHNAN, N. Image in-painting techniques - A survey and analysis. In: *2013 9th International Conference on Innovations in Information Technology (IIT)*. 2013, s. 36–41.

Literatúra

A. Plán práce

A.1 Letný semester - DP1

V tomto semestri plánujem pracovať na analýze domény, návrhu metódy a jej implementácií.

A.2 Zimný semester - DP2

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.2.1 Vyjadrenie k plneniu plánu

V tomto semestri sa nám podarilo splniť všetky stanovené ciele. Natrénovali sme niekoľko modelov detekujúcich Alzheimerovu chorobu z MRI snímkov. Čo sa týka úspešnosti týchto modelov, bohužiaľ sa nám nepodarilo dosiahnuť tak dobré výsledky ako u iných prác. Avšak, naším cieľom nie je natrénovať najlepší model, takže táto úspešnosť vyzerá byť zatial pre nás postačujúca.

Metódu sme implementovali, tak, ako sme ju navrhli, pričom sme pridali vylepšenia ako multiprocessing - paralelné generovanie masiek.

S hyper-parametrami navrhnutej metódy sme experimentovali (ale nie so všetkými, pretože ich je veľa), avšak sme nerobili žiadne prehľadávanie optimálnych parametrov.

Dosiahnuté výsledky sme skúmali a diskutovali ich v závere overenia riešenia pričom sme navrhli ďalšie kroky.

A.3 Letný semester - DP3

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
- Vykonáť úpravy v navrhnutej metóde na základe doterajších výsledkov experimentov
- Vyhodnotiť konzistenciu tepelných máp
- Optimalizovať vstupné parametre do RISEI metódy
- Porovnať navrhnutú metódu s existujúcimi metódami

Dodatok A. Plán práce

- Vyhodnotiť a porovnať vykonané experimenty
- Odovzdať prácu