ŽILINSKÁ UNIVERZITA V ŽILINE

AUTOREFERÁT
DIZERTAČNEJ PRÁCE

Žilina, Máj 2023

Ing. Jaroslav Kopčan

Žilinská univerzita v Žiline
Fakulta riadenia a informatiky

Ing. Jaroslav Kopčan

Autoreferát dizertačnej práce

VYSVETLITEĽNÉ ROZPOZNÁVANIE VZOROV

na získanie akademického titulu "**philosophiae doctor**" (**PhD.**)
v štúdijnom programe doktorandského štúdia
aplikovaná informatika
v štúdijnom odbore
informatika

Žilina, Máj 2023

Dizertačná práca bola vypracovaná v dennej forme doktorandského štúdia na Katedre informačných sietí, Fakulte riadenia a informatiky Žilinskej univerzity v Žiline.

Predkladateľ: *Ing.* ***Jaroslav Kopčan***
Katedra informačných sietí
Fakulta riadenia a informatiky
Žilinská univerzita v Žiline

Školiteľ: *prof. Ing.* ***Martin Klimo****, PhD.*
Katedra informačných sietí
Fakulta riadenia a informatiky
Žilinská univerzita v Žiline

Oponenti: *prof. Ing.* ***Peter Sinčák****, CSc.*
Katedra kybernetiky a umelej inteligencie
Fakulta elektrotechniky a informatiky
Technická univerzita Košice,
doc. Ing. ***Michal Gregor****, PhD.*
Ústav konkurencieschopnosti a inovácií
Žilinská univerzita v Žiline

Autoreferát bol rozoslaný dňa: **4.7.2023**
Obhajoba dizertačnej práce sa koná dňa **22.8.2023** o **11** hod. pred komisiou pre obhajobu dizertačnej práce schválenou odborovou komisiou v študijnom odbore **informatika**, v študijnom programe **aplikovaná informatika**, vymenovanou dekanom Fakulty riadenia a informatiky Žilinskej univerzity
v Žiline dňa **3.7.2023**

prof. Ing. ***Karol Matiaško****, PhD.*
predseda odborovej komisie
študijného programu **aplikovaná informatika**
v študijnom odbore **informatika**

Fakulta riadenia a informatiky
Žilinská univerzita
Univerzitná 8215/1
010 26 Žilina

Anotácia

Táto dizertačná práca sa zameriava na vysvetliteľnosť predikcií klasifikátorov hlbokého učenia. Cieľom je vytvoriť systém, ktorý poskytuje vysvetlenia predikcií a pomáha používateľom pochopiť súvisiace prvky. Pre zachovanie výkonu využívame post-hoc prístup. Presná interpretácia prvkov je náročná a má charakter vedeckej práce, preto predstavujeme nástroj, ktorý pomáha výskumníkom porozumieť novo extrahovaným prvkom, ktoré môžu byť v konkrétnej doméne neznáme.

Po zhodnotení súčasného stavu sme zistili, že súčasné metódy vysvetliteľnosti nezvažujú existenciu anomálií. Preto navrhujeme integrovať detektor anomálií do systému vysvetliteľnosti. Naším inovatívnym prístupom k detekcii anomálií je porovnanie mechanizmov použitých pri vzniku testovaných obrazov s tými ktoré boli použité pri tvorbe obrazov v trénovacej sade, namiesto merania podobnosti medzi nimi.

Kľúčové slová: vizuálne trasovanie objektov, hlboké učenie, Siamské neurónové siete, latentné priestory, mechanizmus pozornosti, homografia, analýza dopravy.

Počet strán:	*162*	*Počet použitej literatúry:*	*189*
Počet obrázkov:	*48*	*Počet tabuliek:*	*16*

Annotation

This dissertation researches the explainability of predictions made by deep learning classifiers. We aim to create a system that provides explanations of predictions and helps end-users understand the involved features. To maintain high performance, we use a post-hoc explanation strategy. The exact interpretation of features is a tough task-dependent discipline that has the nature of scientific work. Because of this, we introduce a tool to assist researchers in comprehending newly extracted features that may not be familiar in the specific application domain.

We found that current explainability methods do not consider anomalies. To address this, we suggest integrating an anomaly detector into the explainability system. Our innovative approach to anomaly detection compares the underlying mechanisms used to form the tested image with those used for forming images in the training set, rather than measuring the similarity between them.

Key words: explainability, interpretability, anomaly detection, deep generative modeling, deep learning, feature extraction, pattern recognition, fuzzy logic.

Number of pages:	*162*	*Number of references:*	*189*
Number of figures:	*48*	*Number of tables:*	*16*

1 Úvod

Modely neurónových sietí sú výrazne ťažko vysvetliteľné a podávajú väčšinou žiadne alebo veľmi slabé vysvetlenia pre koncového užívateľa. Tento stav nie je ideálny ani pre používanie v praxi ako aj budúci výskum a vývoj, keďže neuronové siete vo veľkej miere predstavujú modely čiernych skriniek. Tento problém však nie je jednoduchý keďže s narastajúcim výkonom modelu rastie aj jeho robustnosť čo znamená že model má obrovské počty parametrov a informácie v modeloch sú ťažko sledovateľné pričom chýba ich jasná reprezentácia. Druhý dôvod predstavuje nelinearita týchto systémov kedy aktivačné funkcie dokážu transformovať priestor spôsobom, ktorý je pre človeka nepochopiteľný.
Ďalší významný problém predstavuje aj skutočnosť, že dáta, ktoré nepatria do modelu, sa často nezohľadňujú pri otázke vysvetliteľnosti. To môže spôsobiť nejasnosti pri interpretácii, keď sa model snaží vysvetliť nevysvetliteľné údaje. Preto by detekcia anomálií mala byť neoddeliteľnou súčasťou danej problematiky.

Výskum, ktorý predkladá táto dizertačná práca, má za cieľ riešiť problém vysvetliteľnosti v hlbokom učení a zdôrazňuje výzvy, ktorým sa čelí pri aplikácii hlbokého učenia na detekciu anomálií. Experimenty si kladú za cieľ demonštrovať potenciál hlbokého učenia, dosiahnuť efektivitu a robustnosť systémov detekcie anomálií a taktiež poskytnúť lepšie pochopenie toho, ako sa nový návrh môže aplikovať v reálnych situáciách. Ďalším cieľom je vyvinúť metódu, ktorá môže zlepšiť vysvetliteľnosť modelov hlbokého učenia a zároveň zachovať ich presnosť. Hlavný prínos práce je v oblasti vysvetliteľnosti, ktorý poskytuje lepšie pochopenie toho, ako modely hlbokého učenia robia svoje predikcie. Prínosom pre oblasť detekcie anomálií je poskytnutie poznatkov o možnostiach a obmedzeniach modelov hlbokého učenia pre túto úlohu.
Vzhľadom na vyššie spomenuté aspekty, v rámci tejto dizertačnej práce boli zadefinované nasledovné ciele:

- **Primárnym cieľom tejto práce je vyvinúť metódu post-hoc vysvetlení pre rozhodnutia vykonané systémom hlbokého učenia, konkrétne pre rozpoznávanie obrázkov.**
- **Sekundárnym cieľom je vyvinúť techniku detekcie anomálií, ktorá sa vyhýba vysvetleniam anomálnych vstupov a spolieha sa na hlboké generatívne modelovanie. Takýto systém by mohol slúžiť ako rozšírenie metód vysvetľovania, čím by sa zvýšila robustnosť a spoľahlivosť systémov. Alternatívne by mohol byť implementovaný aj ako samostatná aplikácia.**

Nasledujúca kapitola poskytuje súčasný stav v oblasti vysvetliteľnosti metód umelej inteligencie a detekcie anomálií.

2 Teoretické východiská a súčasný stav

2.1 Vysvetliteľnosť v strojovom učení

Všeobecná kvantitatívna definícia vysvetliteľnosti umelej inteligencie (AI) neexistuje, avšak kvalitatívne ju môžeme definovať takto:

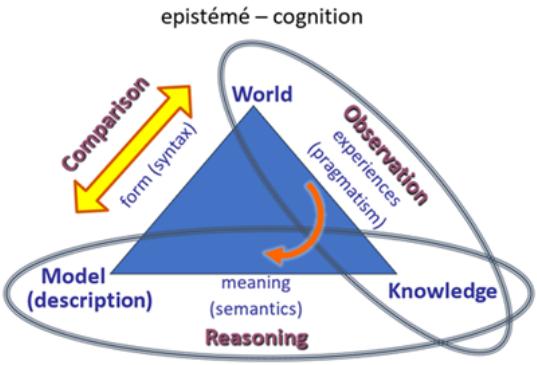
Vysvetliteľnosť je úroveň na ktorej dokáže človek porozumieť príčine rozhodnutia systému.[11]

Čím je vyššia úroveň vysvetliteľnosti modelu, tým jednoduchšie je pre človeka pochopiť, prečo boli vykonané určité rozhodnutia alebo predpovede. Taktiež by sme mohli definovať vysvetliteľnosť ako úroveň na ktorej dokáže človek konzistentne predpovedať výstupy modelu. Takáto definícia však platí iba do určitej komplexnosti modelu. S pribúdajúcim množstvom dát (alebo narastajúcou zložitosťou dát) a zväčšujúcou sa zložitosťou modelu, nie je pre človeka možné, aby vedel predikovať výsledok ešte pred tým ako ho určí samotný model. Potom je však dôležité že vie porozumieť príčine rozhodnutia modelu. Jednoduchá metrika kvality vysvetliteľnosti môže byť určená porovnaním, napríklad tak, že jeden model má lepšiu vysvetliteľnosť ako druhý ak sú jeho rozhodnutia pochopiteľnejšie pre koncového užívateľa ako rozhodnutia z iného modelu [7].

2.2 Všeobecný popis a motivácia

Vysvetliteľnosť je základnou požiadavkou na vedecký prístup k riešeniu problémov. Jeho základný princíp vyjadruje epistemologický trojuholník (Fig.2). alebo aj „semiotický trojuholník“ [5], v ktorom sú uvedené rôzne interpretácie vrcholov trojuholníka. Iné varianty reťazca „world – knowledge – model“ sú napr. „thing – logos – states of mind“ (Aristotle), „object – sign – interpretant“ (Pierce).

Zjednodušene môžeme povedať, že z okolitého sveta pozorovaním abstrahujeme vedomosti vo forme vedomostí, z ktorých logickou úvahou zostavujeme model pozorovanej reality. Použitím modelu robíme rozhodnutia a pokiaľ je model správny, tieto rozhodnutia budú správne aj v realite. Strojové učenie tento postup zmenilo na priamu identifikáciu parametrov modelu z reálnych dát a pokiaľ je táto transformácia nelineárna (hlboké neurónové siete), stáva sa pre požívateľa čiernou skrinkou. Preto sa domnievame, že pre vysvetliteľnosť rozhodnutia získaného neurónovou sieťou je potrebné napodobniť proces ľudského myslenia a prispôsobiť proces tvorby rozhodnutí epistemologickému trojuholníku. Podvedome sa tento koncept uplatnil aj vo vývoji hlbokých neurónových sietí, keď hlboká neurónová sieť pre rozpoznávanie vzorov bola rozdelená na časť extrakcie príznakov a časť generovania rozhodnutí z príznakov (napr. klasifikácia).



Obr. 1: Epistemologický trojuholník.

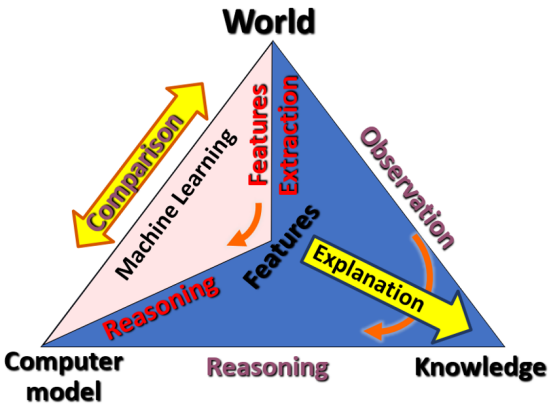
Preto aj vysvetliteľnosť systému rozpoznávania vzorov je potrebné rozdeliť na dve časti: vysvetliteľnosť extrahovaných príznakov a vysvetliteľnosť odvodenia rozhodnutia z extrahovaných príznakov. Príznaky môžu byť priamo vysvetliteľné expertom v danej oblasti a takéto sú niekedy získavané priamo predspracovaním dát a použité ako vstup do neurónovej siete. Túto transformáciu vedomostí na príznaky nazývame príznakovým inžinierstvom.

Príkladom je spracovanie zvuku, kedy vieme že vďaka tvaru baziliárnej membrány vo vnútornom uchu počuje človek v spektrálnej oblasti, preto ako vstup neurónovej siete sa používal priamo spektrogram zvuku. Ak príznak objektov poznáme aspoň približne (napr. vek zobrazovanej osoby – face aging [10]), potom aj neurónovú sieť pre extrakciu príznakov môžeme natrénovať tak aby vzory s určitou hodnotou zvoleného príznaku boli umiestnené v určitej časti príznakového (latentného) priestoru. Ak takto extrahované príznaky použijeme na rekonštrukciu obrazu (variational autoencoder [8]), potom zmenou polohy príznakového vektora vieme odpovedajúcu vlastnosť vtláčať rekonštruovanému obrazu [15]. Dokonca pre vytvorenie vzoru nemusíme ani použiť príznaky extrahované zo vstupných dát, ale môžeme natrénovať generatívnu neurónovú sieť (generative adversarial network - GAN) tak, aby vzor generovala z náhodných súradníc príznakového vektora z definovaného podpriestoru latentného priestoru.

Na extrakciu príznakov známych používateľovi a teda pre neho vysvetliteľných, sú k dispozícii účinné nástroje. Ak však chceme využiť výhody získané používaním nelineárnych systémov na rozpoznávanie vzorov, musíme pripustiť že získané príznaky sú pre používateľa nové a teda nie sú priamo vysvetliteľné. V práci zastávame názor, že vysvetliteľnosť príznakov spočíva priamo v získavaní vedomostí z extrahovaných príznakov. Tento proces je podobný získavaniu vedomostí z dát a

preto má rovnaký vedecký postup. Výhoda náhrady vstupných dát extrahovanými príznakmi spočíva v tom, že extraktor príznakov extrahuje významné príznaky z pohľadu aplikácie t.j. dáta predspracuje a môže vydolovať príznaky, ktoré by človeku ostali skryté kvôli ľudskej zaujatosti minulými skúsenosťami. Toto oprostenie sa týka hľadania parametrov modelu strojovým učením, avšak architektúra modelu a voľba jeho hyperparametrov je doposiaľ stále na človekovi. Určité skúsenosti máme aj s interpretáciou zložitejších príznakov získaných lineárnymi extraktormi. Príkladom toho sú príznaky získané Principal Component Analysis.

Vysvetliteľnosť teda nechápeme ako automatizovaný postup ktorý dá vysvetlenie aj nezainteresovanému používateľovi. Naopak, chápeme ho ako proces objavovania, získavania vedomostí zo štúdia príznakov požívaných konkrétnym systémom rozpoznávania vzorov v konkrétnej aplikácii (Fig. 3). To síce môže viesť k úzkej špecializácii, ale to je dnes typické aj pre čoraz užšiu špecializáciu vedných odborov. Príkladom môže byť analýza ťahu v hre Go programom AlphaGo, ktorý bol pre hráčov dovtedy neznámy, ale bol rozhodujúci pre víťazstvo nad hráčom Lee Sedol. Pochopenie transformácie príznakov na vedomosti je učením sa používateľa od počítača. Navyše analýzou príznakov môžeme dospieť aj k tomu, že síce príznak významne prispel k rozpoznaniu obrazu, ale príznak nie je významný pre rozpoznávaný objekt v obraze (Clever Hans effect) [14] [1].



Obr. 2: Vysvetliteľnosť strojového učenia v koncepte epistemologického trojuholníka

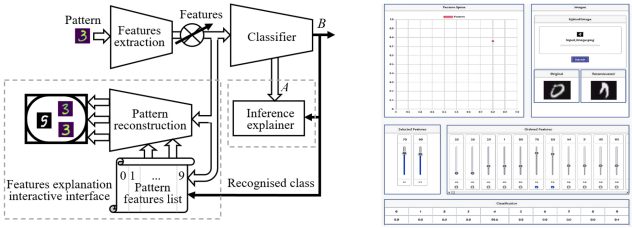
Úlohou používateľa teda nie je vysvetliť ktoré príznaky boli extrahované zo vstupných dát (nerobí vysvetliteľnosť extrakcie príznakov), ale vysvetliť prečo bolo urobené práve toto rozhodnutie zo získaných príznakov. Preto budeme trvať na tom,

aby bežne používaný nelineárny klasifikátor bol nahradený resp. podporený lineárnym systémom alebo logickými pravidlami. Táto práca sa zameriava na implementáciu post-hoc vysvetliteľnej podpory.

3 Navrhovaný prístup k vysvetliteľnosti klasifikácie

Táto práca predstavuje prístup určený pre zlepšenie používateľského pochopenia vzorov a ich vplyvu na rozpoznávanie tried. Model predstavuje vysvetliteľný klasifikátor vo forme fuzzy logiky, ktorá poskytuje konzistentné vysvetlenia pre podmnožinu vzorov patriacich do jednej triedy. V prístupe založenom na príkladoch skúmame vzorky s podobnými znakmi buď z rovnakej triedy, alebo z inej triedy (kontrafaktuálne) v porovnaní s vysvetľovaným vzorom. Tento prístup sa používa počas fázy rozpoznávania a vysvetľovania nového vzoru. Pri vysvetľovaní funkcií môže používateľ experimentovať s variáciami hodnôt príznakov a sledovať vplyv na rozhodnutie a zodpovedajúci vzor. Toto poskytuje vizuálnu reprezentáciu logickej funkcie použitej na vysvetlenie klasifikácie, ktorá poskytuje rovnaký výsledok ako nelineárny klasifikátor.

Ako názorný príklad sme vybrali vysvetliteľnosť rozpoznávania rukou písaných číslic z databázy MNIST. Ako znázorňuje bloková schéma na Fig.3 a), systém rozpoznávania rukou písaných číslic pozostáva z extraktora príznakov, užívateľského prostredia a fuzzy logického modelu ktorý je hlavnou súčasťou post-hoc vysvetliteľnosti nelineárneho klasifikátora. Fig. 3 b).



Obr. 3: Návrh vysvetliteľného systému

Ako už bolo uvedené, tvrdíme, že je nevyhnutné zachovať nelinearitu extraktora príznakov, aj keď príznaky samotné nie sú okamžite vysvetliteľné a proces vysvetľovania je analogický s objavovaním prostredníctvom pozorovania udalostí. Naša perspektíva je v súlade s experimentálnym prístupom Galilea Galileiho k získavaniu vedomostí, na rozdiel od Platónovej myšlienky pasívneho pozorovania. Pre tento

post-hoc prístup sme vyvinuli nástroj, ktorého bloková schéma je znázornená na obrázku 3a) (úprava [3]). Testuje sa nelineárny klasifikátor pozostávajúci z extraktora prvkov a klasifikátora. Experiment zahŕňa variácie znakov a pozorovanie ich vplyvu na rekonštruovaný obraz, výsledok klasifikácie a relevantnosť znakov. Model dekodéra, ktorý rekonštruuje obraz pomocou extrahovaných prvkov, sa trénuje, zatiaľ čo kompresor (súčasť klasifikátora - extraktor príznakov) zostáva nemenný. Grafické rozhranie predstavuje 2D podpriestor príznakového priestoru, pričom sa zobrazuje originálny obraz a obraz po zmene príznakov.

V práci zastávame názor, že ľudským vedomostiam v procese poznania zodpovedajú extrahované príznaky v procese strojového učenia. Doteraz sme popísali spôsob ako naučiť používateľa vysvetliť extrahované príznaky, ďalej budeme popisovať spôsob ako použiť príznaky na vysvetlenie klasifikácie vzoru.

Skúmame dva prístupy na využitie funkcií $f_i, i \in \{1,...,M\}$ pri určovaní klasifikovanej triedy pomocou inferencie:

1. Ich priame použitie, ako to robí nelineárny klasifikátor, kde $y_i = f_i$.
2. Využitie miery relevantnosti funkcií, kde $y_i = \rho(f_i) = R_i$.

Čo sa týka **2. bodu** - Na tému atribučných metód, teda využitie miery relevantnosti príznakov, existuje mnoho literatúry, ktorá hodnotí význam premenných v neurónových sieťach, ktoré aplikujeme na príznaky. Za účelom testovania navrhovanej metódy, budeme brať do úvahy dve vzorové techniky na výpočet relevancie prvkov: postup spätného šírenia založený na prepočte gradientu a prístup relevantnosti príznakov [13],[12].

Ako predstaviteľa prístupu k relevantnosti funkcií používame Layer-Wise Relevance Propagation [2]. Spätný výpočet relevantnosti LRP metódy je definovaný ako:

$$R_i^l = \sum_{k=1}^{B_{l+1}} R_{i\leftarrow k}^{l,l+1} = \sum_{k=1}^{B_l+1} \frac{a_i^l w_{i,k}^{l,l+1}}{\varepsilon + \sum_{j=1}^{B_l} a_i^l w_{j,k}^{l,l+1}} R_k^{l+1} \tag{1}$$

$$i \in \{1,\dots B_l\}, l \in \{1,\dots,L-1\}$$

Ďalej, ako predstaviteľa gradientového prístupu sme zvolili metódu Vanilla Gradient pre zobrazenie teplotných máp (Saliency Maps) [6]. Táto metóda taktiež veľmi pekne ukazuje všeobecný princíp, ktorý nasledujú aj ostatné gradientovo založené metódy.

Skóre triedy dôležitosti S_i^j príznaku f_i pre konkrétnu triedu j je definované ako paricialna derivácia:

$$S_i^j = \frac{\delta a_j^L}{\delta f_i}\Big|_{f_i=a_i^0}, i \in \{1,\dots,M\}, j \in \{1,\dots,N\} \tag{2}$$

V oboch prípadoch, teda buď surový extrahovaný príznak alebo jeho vypočítaná relevantnosť sa musia normalizovať na jednotkový interval, aby sa jeho hodnota mohla interpretovať ako pravdivostná hodnota výroku fuzzy logiky [9]. Normalizujeme pre hodnoty $y_i \in \mathbb{R}, i \in \{1,...,M\}$ a použijeme min-max lineárnu normalizáciu:

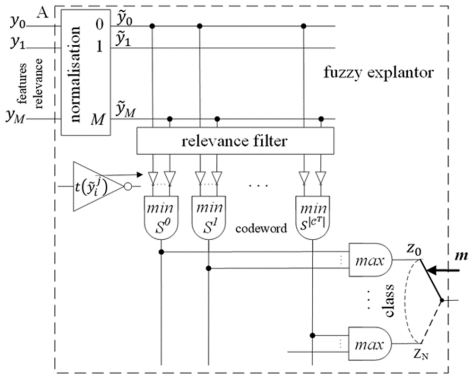
$$\begin{aligned} \tilde{y_i} &= \frac{y_i - y_{min}}{y_{max} - y_{min}}, \\ y_{min} &= min\{y_1, \ldots, y_M\}, \\ y_{max} &= max\{y_1, \ldots, y_M\} \end{aligned} \tag{3}$$

Existuje tiež možnosť normalizácie funkcií pomocou funkcie Sigmoid:

$$\tilde{y_i} = S(y_i) = \frac{1}{1 + e^{-y_i}} \tag{4}$$

4 Fuzzy model ako vysvetliteľný klasifikátor

V predchádzajúcej časti boli spomenuté spôsoby ako možno merať významnosť príznakov za účelom ich použitia pre fuzzy logiku a ich následnú normalizáciu pre Zadehovu fuzzy funkciu. Takéto predspracovanie dát je prvým krokom v našej navrhovanej architektúre, ako môžeme vidieť na obrázku 4 - Normalizačný modul.



Obr. 4: Architektúra Fuzzy logického klasifikátora

Ďalšia fáza zahŕňa modul relevančného filtra, ktorý transformuje príznaky do zodpovedajúcich stavov relevantnosti. Pozitívna relevancia predstavuje vyjadrenie,

do akej miery je konkrétny vstupný príznak alebo skupina príznakov dôležitá na vytváranie presných predikcií modelom hlbokého učenia.

Takže v relevančnom filtri interpretujeme príznak $c_i, i \in \{1,\ldots,M\}$ ako pozitívne relevantný $c_i = 1$, ak je hodnota alebo váha jeho relevantnosti $\tilde{y_i} \in [0,1]$ vyššia ako daná hranica $\tilde{y_i} > \frac{1}{2} + \Delta$ a ako negatívne relevantný $c_i = 0$, ak je jeho hodnota nižšia ako špecifický hranica $\tilde{y_i} < \frac{1}{2} - \Delta$.

Tretí prípad je pre irelevantné príznaky $c_i = X$ if $\frac{1}{2} - \Delta \leq \tilde{y_i} \leq \frac{1}{2} + \Delta$:

$$c_i = \begin{cases} 1, & \tilde{y_i} > \frac{1}{2} + \Delta \\ X, & \frac{1}{2} - \Delta \leq \tilde{y_i} \leq \frac{1}{2} + \Delta \\ 0, & \tilde{y_i} < \frac{1}{2} - \Delta \end{cases} \tag{5}$$

Ďalej odvodíme pravdivostnú hodnotu pre príslušné príznaky:

$$t(\tilde{y_i} = c_i) = \begin{cases} 1 - \tilde{y_i}, & c_i = 0 \\ \tilde{y_i}, & c_i = 1 \end{cases}, c_i \neq X, i \in \{1,\ldots,M\}. \tag{6}$$

Cieľom navrhovaného fuzzy klasifikátora je priradiť vstupnému vzoru $\mathbf{x}$ s relevančným vektorom $\tilde{\mathbf{y}}$ kódové slovo relevancie $\tilde{\mathbf{c}}$ (vysvetlenie) s najvyššou mierou pravdivosti. Pravdivostná hodnota tvrdenia, že vyhodnotené kódové slovo $\tilde{\mathbf{c}}$ sa rovná kódovému slovu $\mathbf{c^j}$, je:

$$t(\tilde{\mathbf{c}} = \mathbf{c^j}) = \min_{\substack{i=1,\ldots,M \\ c_i^j \neq X}} t(\tilde{c_i} = c_i^j), j = 1,\ldots,3^M \tag{7}$$

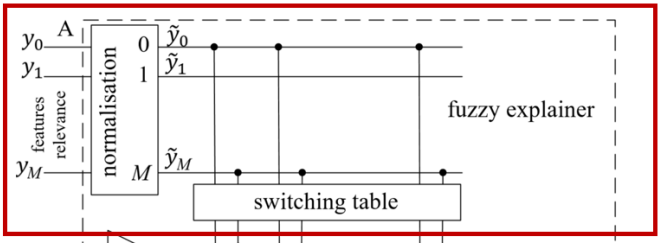
Následne fuzzy model priradí k rozpoznanému vzoru kódové slovo $\mathbf{c}^{\tilde{m}}$ s maximálnou pravdivostnou hodnotou (klasifikácia vykonaná fuzzy modelom):

$$\tilde{m} = \arg\max_{c^j \in \mathbb{C}} t(\tilde{\mathbf{c}} = \mathbf{c}^j) \tag{8}$$

4.1 Trénovanie Fuzzy Klasifikátora

Hlavnou úlohou trénovania fuzzy klasifikátora je nájsť relevantné (kódové) slová $\mathbf{c} = \{(c_1,\ldots,c_i,\ldots,c_M)\} c_i \in \{0,X,1\}, i \in \{1,\ldots,M\}$, ktoré maximalizujú pravdepodobnosť, že fuzzy model zaradí kódové slovo do rovnakej triedy ako nelineárny klasifikátor. Keď sa zameriavame na interpretáciu nelineárneho klasifikátora, je rozumné očakávať, že navrhovaný fuzzy model by sa mal správať čo najbližšie k klasifikátoru.

Na trénovanie modelu sú potrebné dva kroky pri použití trénovacej množiny údajov:



Obr. 5: Tréningová fáza (tvorba kódových slov) s navrhovaným Fuzzy modelom

1. V počiatočnej fáze pridelíme každý vzor v trénovacej množine zodpovedajúcemu kódovému slovu na základe vzorca (5). Vytvorené kódové slová a ich príslušnú triedu zhromažďujeme v štatistickej tabuľke priradených tried ku kódovému slovu. Počas tohto 'tréningu' zvyšujeme počet výstupných kódových slov pre špecifickú triedu rozpoznanú fuzzy modelom. Informácie sa potom zhromažďujú v štatistickej tabuľke pre triedy priradené konkrétnemu kódovému slovu (ako je znázornené v tabuľke 1).

Codeword $\mathbf{c}^1$			...	Codeword $\mathbf{c}^{\lvert\mathbb{C}\rvert}$		
Class1	...	Class N		Class1	...	Class N
n_1^1	...	n_N^1	...	$n_1^{\lvert\mathbb{C}\rvert}$	...	$n_N^{\lvert\mathbb{C}\rvert}$

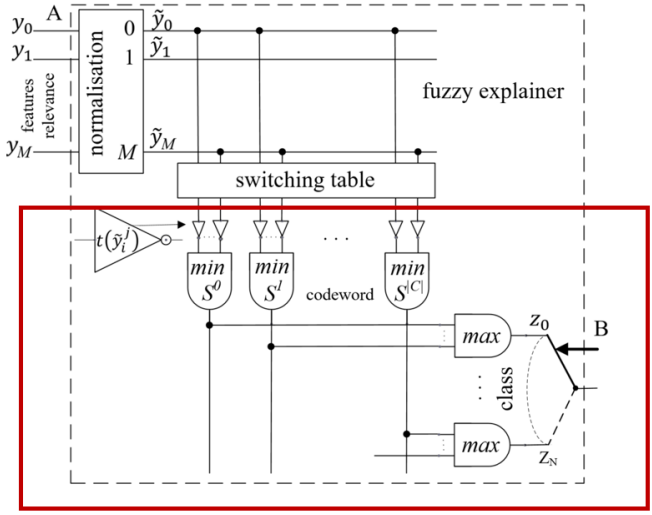
Tabuľka 1: Príznakové mapovanie do kódových slov

2. Počas druhého kroku sa vyberú príslušné kódové slová (pozri podmienku dominancie 9) a priradia sa ich príslušným triedam. Tento proces sa vykonáva nasledujúcim spôsobom. Hodnota n_k^j spojená so vzorom triedy $k \in \{1,\ldots,N\}$ priradená kódovému slovu $\mathbf{c}^j \in \{1,\ldots,3^M\}$ označuje pomer presných a nepresných klasifikácií vytvorených fuzzy modelom.
Inými slovami, hodnota n_k^j označuje, koľkokrát klasifikátor správne identifikoval danú triedu $n_k^{j^+}$ v porovnaní s počtom nesprávnych klasifikácií $n_k^{j^-}$, t.j. $n_k^j = n_k^{j^+} - n_k^{j^-}$ Kódové slovo $\mathbf{c}^j \in \{1,\ldots,3^M\}$ sa považuje za vhodné pre konkrétnu triedu $k \in \{1,\ldots,N\}$, ak platí nasledovná podmienka:

$$n_k^j > \alpha \sum_{i \neq k} n_i^j \tag{9}$$

4.2 Inferencia Fuzzy Klasifikátora

Na získanie predikcie modelu pre predkladaný vzor **x** patriaci do triedy $m \in \{1,\ldots,N\}$ sa privádza ako vstup do nelineárneho klasifikátora (DNN), a ten identifikuje pre daný vzor najvýraznejšie **f**.



Obr. 6: Fuzzy model - inferencia

Následne klasifikačná časť DNN využíva extrahované znaky **f** pre predikciu triedy vzoru. Z rozhodnutia klasifikátora čiernej skrinky odvodíme relevantnosť príznakov **y** a privedieme ich na vstup vysvetliteľného klasifikátora. Vo fáze testovania teda získame zodpovedajúce kódové slová z inferencie testovacích vzoriek. Podľa vytvorenej Tabuľky kódových slov (1) nájdeme najbližšie kódové slovo (diskrétny vektor), ktoré nebolo počas tréningovej fázy vyradené.

Pre evaluáciu Fuzzy klasifikátora budeme používať nasledovné metriky:

$$explanation \quad \tilde{\mathbf{c}} = \underset{j:\mathbf{c}^j \in C^T_{\tilde{m}}}{\operatorname{argmax}} \; t(\tilde{\mathbf{c}} = \mathbf{c}^{\mathbf{j}}) \tag{10}$$

Pre presnosť predikcií black-boxu:

$$acc_{BBox} \quad = \frac{\sum_{i=1}^{n} \delta(\tilde{\tilde{m}}_i = m_i)}{n} \tag{11}$$

kde

$$\delta(\tilde{\tilde{m}}_i = m_i) = \begin{cases} 1, & \tilde{\tilde{m}}_i = m_i \\ 0, & \tilde{\tilde{m}}_i \neq m_i \end{cases}$$

Presnosť predikcií fuzzy modelu môžeme hodnotiť podobne:

$$acc_{fuzzy} \quad \rho = \frac{\sum_{i=1}^{n} \delta(\tilde{m}_i = m_i)}{n} \tag{12}$$

Vierohodnosť vysvetlení je jeden z najdoležitejších atribútov pri hodnotení návrhu z hladiska vysvetliteľnosti:

$$fidelity \quad \rho_r = \frac{\sum_{i=1}^{n} \delta(\tilde{m}_i = \tilde{\tilde{m}}_i)}{n} \tag{13}$$

a nakoniec stabilita vysvetlení ktorá vyjadruje mieru, do akej vysvetlenie zostáva konzistentné **v rámci triedy**:

$$stability \quad H(C_t, C_m) = \sum_{i=1}^{n} \delta(b_i^t \neq b_i^m) \tag{14}$$

4.3 Experimentálna činnosť

Výsledky uvedené v tabuľke 2 ukazujú výkonnosť navrhovaného post-hoc modelu na databáze ImageNet. Celková presnosť architektúr nelineárnych klasifikátorov je však taktiež výrazne nižšia ako 90%. V prípade modelov DenseNet121 a VGG16 sme boli schopní dosiahnuť porovnateľnú úroveň presnosti s post-hoc vysvetliteľným modelom. Použili sme tiež metriku stability (pozri 14) na posúdenie podobnosti medzi vysvetleniami inštancií v rámci každej triedy pre konkrétnu architektúru. Je dôležité mať stabilné vysvetlenia, pretože náhla zmena vo vysvetlení vhľadom na malé zmeny na vstupe môže viesť k nedôvere alebo nejednoznačnosti, pokiaľ ide o správanie modelu.
Keď sa zaoberáme komplexnejšími dátami, ako je ImageNet, metóda redukcie príznakov (zväčšovanie pásma irelevantnosti) prináša veľké výhody. V tomto konkrétnom experimente sme sa rozhodli použiť architektúru DenseNet121. Tabuľka 5 zobrazuje priemerný počet pozitívnych a negatívnych príznakov spolu s celkovým výkonom modelu pod rôznymi prahovými úrovňami. Na základe výsledkov môžeme konštatovať, že táto príznaková redukcia (celková suma $N_{POSITIVE}$ a $N_{NEGATIVE}$) bola úspešná, pričom miera vernosti vysvetlenia bola zachovaná nad 99%. v experimentoch figuruje spodná a horná hranica symetricky centrovaná okolo stredu jed-

Relevance Method	ResNet50			VGG 16		
	Fidelity p_r	**Accuracy Fuzzy p**	**Accuracy B_Box**	**Fidelity p_r**	**Accuracy Fuzzy p**	**Accuracy B_Box**
Vanilla Grad	5.09	4.58	67.46	47.89	26.82	64.44
Raw normalised features	**34.09**	**31.04**	**67.46**	41.01	36.84	64.44
LRP relevance	-	-	67.46	21.74	17.8	64.44
Guided Backprop	5.09	4.58	67.46	15.66	12.66	64.44
DeconvNet	5.09	4.58	67.46	**100.0**	**64.44**	**64.44**

Relevance Method	InceptionV3			DenseNet121		
	Fidelity p_r	**Accuracy Fuzzy p**	**Accuracy B_Box**	**Fidelity p_r**	**Accuracy Fuzzy p**	**Accuracy B_Box**
Vanilla Grad	11.35	10.42	75.84	100	71.2	71.2
Raw normalised features	**65.87**	**60.49**	**75.84**	40.24	36.99	71.2
LRP relevance	-	-	75.84	-	-	71.2
Guided Backprop	11.35	10.42	75.84	**100**	**71.2**	**71.2**
DeconvNet	11.35	10.42	75.84	**100**	**71.2**	**71.2**

Tabuľka 2: Výsledky pre databázu ImageNet

Architecture / Codeword Length	Relevance Method	Bit Match	Stability (%)
VGG16/4096 bits	Raw	2971.28	72.54
	Vanilla Grad	3789.96	92.53
	Guided Backprop	3937.23	96.12
	DeconvNet	**4096**	**100**
ResNet50/2048 bits	Raw	2004.03	72.54
	Vanilla Grad	**2046.44**	**99.92**
	Guided Backprop	**2046.44**	**99.92**
	DeconvNet	**2046.44**	**99.92**

Tabuľka 3: Priemetná stability predikčných vysvetlení v rámci tried pre architektúry VGG16 a ResNet50

notkového intervalu, začínajúc od jednej tretiny:

$$LowerLimit = \frac{1}{3} - \Delta, \quad UpperLimit = \frac{2}{3} + \Delta \tag{15}$$

Architecture / Codeword Length	Relevance Method	Bit Match	Stability (%)
DenseNet121/1024 bits	Raw	984.12	96.12
	Vanilla Grad	**1024**	**100**
	Guided Backprop	1022.51	99.85
	DeconvNet	**1024**	**100**
InceptionV3/2048 bits	Raw	1937.14	94.58
	Vanilla Grad	**2046.67**	**99.93**
	Guided Backprop	**2046.67**	**99.93**
	DeconvNet	**2046.67**	**99.93**

Tabuľka 4: Priemetná stability predikčných vysvetlení v rámci tried pre architektúry Inception V3 a DenseNet121

DenseNet121/ImageNet	Δ	0	0.1	0.2	0.3
VanillaGrad/ GuidedBackprop/DeconvNet	$N_{POSITIVE}$	10.59	5.19	2.5	2.04
	$N_{NEGATIVE}$	774.12	275.15	44.28	21.75
	Fidelity $\rho_r(\%)$	100	100	99.78	99.12
Raw normalized features	$N_{POSITIVE}$	4.12	2.44	1.5	1.05
	$N_{NEGATIVE}$	989.03	949.35	856.96	627.17
	Fidelity $\rho_r(\%)$	40.24	40.24	38.61	27.57

Tabuľka 5: Metoda redukcie príznakov aplikovaná pre DenseNet-121/ImageNet

4.4 Vyhodnotenie

V tejto práci sme navrli metódu, ktorej experimentálne výsledky ukazujú, že vysvetliteľný fuzzy klasifikátor môže zodpovedať klasifikáciou takmer identicky s nelineárnymi klasifikátormi. Parita výkonu sa dosiahne vtedy, keď klasifikátory striktne oddeľujú proces extrakcie prvkov od klasifikácie a dokážu úspešne extrahovať zmysluplné a efektívne príznaky zo vstupných údajov (závisí od architektúry).

Primárnym krokom je použitie relevantnosti príznakov ako vstupných hodnôt pre fuzzy klasifikátor. Najlepšie výsledky dosiahli metódy, ktoré určujú relevanciu funkcie pomocou metód založených na gradiente a celkovým víťazom medzi testovanými metódami sa stal DeconvNet. Pri zvažovaní implementácie tohto post-hoc vysvetlenia je však kľúčové premyslieť si celý proces vrátane dostupných údajov a architektúry. Hoci navrhovaná všeobecná koncepcia zostáva nemenná, skutočná implementácia musí byť prispôsobená konkrétnej úlohe.

Príspevok fuzzy logiky k vysvetliteľnosti v rámci klasifikačnej domény:

- Post-hoc klasifikátor je no forme fuzzy logického výrazu. Sada pravidiel (IF, THEN) je základom logiky (modus ponens: $(P \to Q) \wedge P \to Q$), pre neurčité dáta rozšírenej na fuzzy logiku (miera pravdivosti výroku z jednotkového intervalu).

- Dostávame mieru pravdivosti klasifikácie. Spätne môžeme odstrániť irelevantné príznaky a zistiť mieru pravdivosti kritických príznakov.

- Keďže ide o post-hoc vysvetliteľnosť, môžeme určiť mieru dôveryhodnosti vysvetlenia (fidelity).

- Ako mieru pravdivosti príznakov môžeme použiť samotné hodnoty príznakov, alebo ich relevancie. Potom vieme posúdiť vhodnosť rôznych výpočtov relevancie porovnaním dôveryhodnosti, prípadne priemernej pravdivosti klasifikácie vysvetliteľného klasifikátora.

- Môžeme hľadať optimálnu transformáciu príznakov na ich pravdivosti $t = \mu(f)$. Čím sa zbavujeme náhrady pravdivosti priamych príznakov ich relevanciou - predmet ďalšieho výskumu.

- Nájdená fuzzy logická funkcia je v tvare úplnej normálnej disjunktnej formy. Počet logických operácií je možné minimalizovať, podobne ako sa minimalizujú Booleove logické funkcie. Logickú funkciu môžeme napr. prepísať do tvaru terciárneho rozhodovacieho stromu.

5 Navrhované prístupy k detekcii anomálií

Motto - Anomália je pozorovanie, ktoré sa tak výrazne odlišuje od iných pozorovaní až budí podozrenie že dané pozorovanie bolo vytvorené úplne iným mechanizmom.

– Hawkins,[4].

Predmetná citácia tvrdí, že nezrovnalosť môže byť dostatočne významná na to, aby oprávňovala k záveru, že vzory neboli vytvorené rovnakým mechanizmom. Preto nie je nevyhnutné, aby sme mali znalosti o presnom mechanizme tvorby daného vzoru, ale postačí, aby sme pochopili mechanizmus zodpovedný za generovanie jeho obrazu.

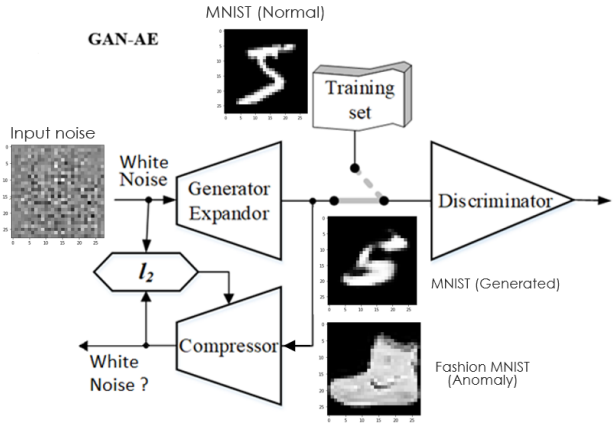
Podľa toho očakávame, že normálne vzorky v rovnakej triede budú vytvorené rovnakým mechanizmom. Bohužiaľ, tento mechanizmus nie je známy pre tréningové/pôvodné dáta. Máme však mechanizmus, ako generovať vzorky, ktoré sa nedajú odlíšiť od trénovacích dát počas Turingovho testu vykonávaného trénovanou

neurónovou sieťou – Diskriminátor. Ak sú vygenerované dáta podobné trénovacím dátam, potom očakávame rovnaký mechanizmus generovania pre obe, pretože úspešne zachytíme pôvodnú distribúciu dát.

5.1 Charakteristiky dát ako indikácie anomálií

V počiatočných fázach výskumu pre detekciu anomálií sme sa zaoberali hlavne praktickým overením využitia generatívnych modelov pre detekciu anomálií konvenčnými spôsobmi(vrodené vlastnosti týchto modelov).

Ďalšie experimenty pre výskum v oblasti detekcie anomálií sme prispôsobili overeným predpokladom z predošlých experimentov a navrhli sme architektúru detekcie pri ktorej by účinnosť detegovania nebola závislá od zložitosti dát. Zamerali sme sa na porovnávanie charakteristík jednotlivých dát, pričom sme za metriku pre detekciu anomálií uvažovali výraznú odlišnosť charakteristík anomálnych dát od charakteristík originálnych dát. Experiment pozostával z dvoch krokov, kde v prvom sme najprv natrénovali GAN sieť na trénovacích dátach MNIST-U a po úspešnom tréningu sme generátor GAN siete využili v ďalšom zapojení k jeho inverznej transformácii. Tréning parametrov inverznej transformácie ku generátoru bol hodnotený pomocou priemernej štvorcovej chyby (MSE), medzi zdrojovým šumom a rekonštruovaným šumom 7.



Obr. 7: Architektúra pre detekciu anomálií a jej princíp

Po natrénovaní bola inverzná transformácia schopná takmer presne rozkladať falzifikáty na výstupný šum podobný tomu, ktorý bol na vstupe pre daný falzifikát. Pre originálne obrazy nemôže byť rekonštruovaný výstupný vektor šumu porovnaný

s generujúcim šumom, pretože ten nepoznáme. Pracovnou hypotézou je, že ak poznáme mechanizmus tvorby falzifikátov z rovnomerného šumu a poznáme proces rekonštrukcie rovnomerného šumu z falzifikátov, potom aj pre originály, ktoré sú podobné falzifikátom bude rovnaký proces degenerácie viesť tiež k rovnomernému šumu. Kľúčom k tomu sú štatistické testy s otázkou, či rekonštruovaný šum tvoria, nezávislé náhodné veličiny s rovnomerným rozdelením. Preto nás zaujímali hlavne charakteristiky výstupného šumu ako bolo spomenuté vyššie, a to či sú rovnaké ako pre vstupný šum.

Jadrom veci sú teda štatistické testy, ktorých cieľom je zistiť, či rekonštruovaný šum pozostáva z nezávislých náhodných premenných s rovnomerným rozdelením.

Na testovanie uniformity vektorov šumu sme použili Chí-kvadrát test. Tento štatistický test je špeciálne navrhnutý tak, aby určil, či je daná vzorka údajov odvodená zo špecifikovaného rozdelenia, ako je napríklad rovnomerné rozdelenie. Pri testovaní uniformity vektorov šumu by nulová hypotéza predpokladala, že vektory šumu sú rovnomerne rozdelené, zatiaľ čo alternatívna hypotéza by naznačovala opak. Vypočíta sa testovacia štatistika H a potom sa porovná s kritickou hodnotou rozdelenia chí-kvadrát. Ak je testovacia štatistika väčšia ako kritická hodnota, nulová hypotéza H_0 sa zamietne, čo naznačuje, že vektory šumu nie sú rovnomerne rozdelené, čo naznačuje odľahlé hodnoty.

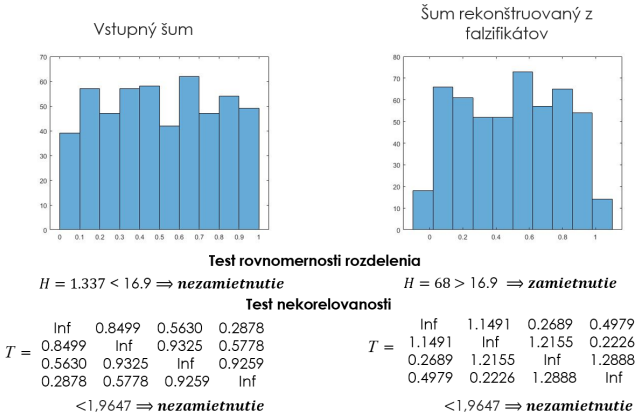
$$H = \sum_{j=1}^{k} \frac{(X_j - np_j)^2}{np_j} \sim \chi^2(k-1) \tag{16}$$

kde k je počet tried, p_j je teoretická pravdepodobnosť 0,1, n je počet prvkov v testovanej vzorke údajov a H testovacia štatistika.
Uvažovanú nulovú hypotézu možno formulovať nasledovne:

H_0 - rekonštruované šumové vektory (dáta) pochádzajú z rovnomerného rozdelenia, takže údaje nie sú vo svojej podstate anomálne.

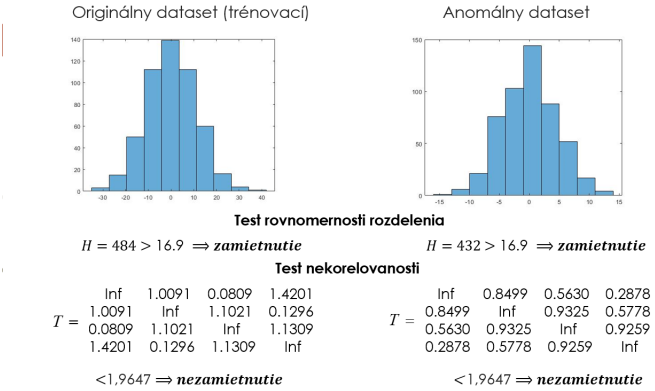
Kritická hodnota rozdelenia chí-kvadrát χ^2 je 16,9.

Štatistické testy odhalili že test rovnomernosti nebol prijatý pre rekonštruované šumy z generovaných falzifikátov, a teda degenerátor nebol efektívne natrénovaný tak aby sa naučil zachovávať charakteristiky vstupného šumu ako môžeme vidieť na obrázku 8. Taktiež takto natrénovaná inverzná transformácia nedokázala správne rozložiť originálne obrazy na výstupný šum, ktorý by prešiel testom uniformnosti, ako je možné vidieť na 9. Z týchto výsledkov vyplýva že súčasný návrh samotný nestačí trénovať iba na falzifikátoch vytváraných generátorom ale taktiež aj na originálnych dátach. Taktiež účelová funkcia MSE nie je schopná rovnako presne zachytiť rozdelenie dát pri trénovaní - hlavne na chvostoch rozdelení. Tento fakt je spôsobený aj tým že chyba sa priemeruje a tréning sa sústredí viac na priemerné



Obr. 8: Výsledky testov pre vstupný šum generátora(vľavo) a rekonštruovaný šum(vpravo)

hodnoty ako na individuálne.



Obr. 9: Výsledky testov pre vstupný šum generátora(vľavo) a rekonštruovaný šum(vpravo)

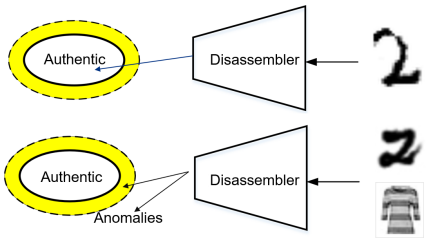
Na základe zhrnutia uvedeného v tabuľke 6 možno pozorovať, že testovacie štatistiky pre scenáre 2, 3 a 4 sú vyššie ako prahová hodnota H. Toto zistenie je významné, pretože naznačuje, že rekonštruované vektory šumu pre tieto scenáre vykazujú anomálie. Avšak iba 4. scenár bol skutočným anomálnym vstupom. Scenár č.1 napokon prešiel testom, ktorý potvrdil, že pri trénovaní inverznej transformácie sme použili skutočne uniformný šum.

Scenarios	Test Statistic H	Decision
1. Input Noise	1.337	Accept null hypothesis
2. Authentic FMNIST fakes	68	Reject null hypothesis
3. Original FMNIST images	484	Reject null hypothesis
4. Anomalous MNIST images	432	Reject null hypothesis

Tabuľka 6: The statistical analysis for hypothesis H_0 - inference of FMNIST trained Inverse Transformation

5.2 Experimenty s detekciou anomálií na základe vzdialenosti

Vzhľadom na náš predchádzajúci predpoklad, že falošné vzorky majú podobnosť s originálnymi obrázkami a že máme znalosti o procese použitom na vytvorenie falošných obrazov, je jasné, že anomálie nemohli pochádzať z rovnakého procesu. Opäť platí, že trénovaním modelu inverznej transformácie, teda inverznej funkcie generátora, môžeme odhaliť tieto odľahlé hodnoty, ktoré boli vytvorené rôznymi procesmi. Nasledovné experimenty sme spracovali za účelom jasne priestorovo oddeliť anomálie od originálnych dát a generovaných autentických falzifikátov. Jedno možné predspracovanie je možné vykonať tak že originálne dáta sa budú nachádzať vnútri jednotkovej hypergule zatiaľ čo anomálie budú mimo. Na základe vzdialenosti je potom možné trénovať inverznu transformáciu spôsobom že originálne dáta a generované autentické falzifikáty budú rekonštruované do vnútra hypergule, zatiaľ čo iné dáta - anomálie budú rekonštruované mimo objektu.



Obr. 10: Princíp detekcie anomálií

Je dôležité poznamenať, že ak majú rekonštruované dáta pre trénovacie vzorky svoje sférické súradnice v rámci jednotkovej hypersféry, akékoľvek počiatočné dáta

získané z anomálnych vzoriek musia byť umiestnené mimo nej. S týmto cieľom sme teda trénovali model inverznej transformácie.

Metrics	Inverse Transformation
Detection Accuracy (%)	41.73
Reconstruction MSE metric	1.0859

Tabuľka 7: Vyhodnotenie detekcie anomálií pre dataset MNIST

Najnovšie zistenia tréningu inverznej transformácie (pozri tabuľku 7)) vykazujú určité pokroky v porovnaní s koncepciou uniformity. Výsledky však nie sú spoľahlivé pri zisťovaní anomálií, čo možno pripísať nedostatočnosti MSE chyby ako jedinej účelovej funkcii v tejto situácii. Vhodnou náhradou za stratovú funkciu by mohla byť Čebyševova vzdialenosť. Okrem toho je možné implementovať stratégiu rekurzívnej rekonštrukcie na zlepšenie tréningu inverznej transformácie. Dalo by sa to urobiť začlenením modelu diskriminátora pôsobiaceho ako kritik na základe klasifikácie obrázkov vytvorených z rekonštruovaných vektorov šumu generovaných modelom inverznej transformácie.

5.3 Vyhodnotenie

Oblasť detekcie anomálií bola podrobená rozsiahlemu výskumu. Napriek tomu nepovažujeme dosiahnuté výsledky za dostatočné.

Počiatočné experimenty na detekciu anomálií boli uskutočnené pomocou DGM-modelov ako Autoenkóder a GAN sieť. Pri detekcii anomálií založenej na Autoenkóderi sme hodnotili chybu rekonštrukcie testovaného vzoru ako detekčnú metriku. Chyba rekonštrukcie porovnáva vstupné a rekonštruované vzorky za predpokladu, že pôvodné obrázky poskytujú menšiu chybu v dôsledku trénovania AE na týchto údajoch.
Podobne v sieťach GAN sme predpokladali, že dobre natrénovaný generátor vytvorí autentické falzifikáty, ktoré diskriminátor klasifikuje podobne ako pôvodné dáta. Akékoľvek vzorky, ktoré sa nezobrazia počas tréningu, budú diskriminátorom automaticky klasifikované ako falošné alebo anomálie. Náš predpoklad sa čiastočne potvrdil. Účinnosť detekcie anomálií pomocou štandardných modelov DGM (AE a GAN) však závisí od zložitosti pôvodných údajov a anomálií. Rozlišovanie medzi takýmito údajmi je náročnejšie, keď sú pôvodné údaje zložité a anomálie jednoduché.

Pri experimentoch s uniformitou dát sme sa rozhodli zamerať na porovnanie mechanizmov tvorby individuálnych obrazov na detekciu anomálií tým, že zohľadníme

významný rozdiel v šume použitom na vytvorenie anomálnych údajov z charakteristík pôvodných údajov ako metriky. Tento prístup bol zameraný na predchádzanie problémom s nestabilnou detekciou anomálií, ktoré môžu vzniknúť v dôsledku zložitosti údajov. Experimenty nepriniesli uspokojivé výsledky.

V poslednom koncepte sme zmenili tréningový prístup pre navrhovanú inverznú transformáciu, namiesto testu uniformity sme použili koncept jednotkovej hypergule, kde pôvodné dáta ležia vo vnútri sféry, zatiaľ čo odľahlé hodnoty sa nachádzajú ďalej. Model bol trénovaný na rekonštrukciu vytvorených obrázkov čo najbližšie k ich pôvodným náprotivkom (vektorom vstupného šumu). Trénovaním inverznej transformácie sa naučí základný koncept vzdialenosti medzi rôznymi typmi obrázkov. Počas testovacej fázy by teda mali byť pôvodné obrázky a autentické falzifikáty zrekonštruované blízko stredu hypersféry a naopak. Tento prístup vykazuje zatiaľ najlepšie výsledky avšak ani tie nie sú dostatočné a táto oblasť vyžaduje ďalšie skúmanie.

6 Záver a zhrnutie prínosov

Hlavným cieľom tejto dizertačnej práce bolo prispieť do oblasti vysvetliteľnosti s použitím hlbokého strojového učenia. Náš vecný vedecký príspevok, či už po stránke teoretickej alebo praktickej, sa skladá z dvoch hlavných častí - vysvetliteľnosť predikcií a detekcia anomálnych vzorov.

Vo všeobecnosti hlavné poznatky tejto práce sú:

1. Vysvetliteľné rozhodnutia/predikcie musia byť analogické ku ľudskému spôsobu rozhodovania, ktorý je definovaný epistemologickým trojuholníkom.

2. Podľa 1. bodu musí byť rozpoznávací systém konceptuálne rozdelený na extrakciu príznakov a klasifikáčnú časť.

3. Interpretácia extrahovaných príznakov má charakter vedeckej práce, v ktorom ľudia získavajú znalosti o príznakoch tak, ako veda získava znalosti o svete. Tento proces musí byť podporovaný vhodnými nástrojmi.

4. Vysvetlenie klasifikácie musí byť založené na logike (binárnej alebo fuzzy).

5. Zadehova fuzzy logika s rozdelením príznakov na pozitívne relevantné, negatívne relevantné a irelevantné hodnoty poskytuje vhodný rámec pre vysvetlenie získanej klasifikácie - fuzifikácia.

6. Fuzzifikácia príznakov ovplyvňuje aproximáciu vysvetliteľných tried na predikcie nelineárneho systému. Relevantnosť príznakov sa zdá byť dobrým východiskom pre nelineárnu transformáciu hodnôt príznakov na ich pravdivostné

hodnoty. Ich aplikáciou (DecovnNet) možno dosiahnuť 100% vernosť (fidelity) vysvetlení na databáze Image Net.

7. Prvé vykonané experimenty naznačujú, že klasifikátor vo forme fuzzy logiky funkcií môže prekonať rozhodovacie stromy.

8. Pre korektnú vysvetliteľnosť je potrebný anomálny detektor ako filter pre zabránenie o pokus vysvetlovania anomálií. Definícia [4] anomálnych dát popisuje anomálie ako vzorky s iným mechanizmom vzniku, a dáva nám návrh ako aplikovať hlboké generatívne modelovanie pre detekciu. Túto myšlienku sme aplikovali dvoma spôsobmi:

    - Ak sú vstupné dáta generátora nezávislé pseudonáhodné premenné s rovnomerným rozdelením, inverzná transformácia testovaného obrazu musí poskytnúť rovnaké vlastnosti pre výstupy. Prvé výsledky tohto overovania nápadov nie sú uspokojivé. Dôvodom sa zdá byť generovanie dát z hyperkocky.
    - Ak sú základom generátora pseudonáhodné premenné z hypergule s rovnomerným rozložením polomeru, inverzná transformácia testovaného obrazu musí poskytnúť súradnice v rámci hypersféry. Vzhľadom na zložitosť tréningu inverznej transformácie bol tento prístup ponechaný na ďalší výskum.

Literatúra

[1] Anders, C.J., Marinc, T., et al.: 'Analyzing imagenet with spectral relevance analysis: Towards imagenet un-hans'ed', *CoRR*, vol. abs/1912.11425, 2019

[2] Bach, S., Binder, A., et al.: 'On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation', *PLOS ONE*, vol. 10, no. 7, pp. 1–46, 2015, doi:10.1371/journal.pone.0130140

[3] Buzgo, M.: 'Image classifier with explainable features', 2022

[4] Hawkins, D.: *Identification of outliers*, Monographs on applied probability and statistics, Chapman and Hall, London [u.a.], 1980, ISBN 041221900X

[5] Hébert Louis, J.M.K.: 'Sign structures/signo - applied semiotics theories',

[6] Karen Simonyan, Andrea Vedaldi, A.Z.: 'Deep inside convolutional networks: Visualising image classification models and saliency maps', pp. 1–8, 2013, doi:arXiv:1312.6034

[7] Kim, B., Khanna, R., Koyejo, O.O.: *Examples are not enough, learn to criticize! Criticism for Interpretability*, vol. 29, Curran Associates, Inc., 2016

[8] Kingma, D.P., Welling, M.: 'Auto-encoding variational bayes', 2013, doi:10.48550/ARXIV.1312.6114

[9] Klimo, M., Lukáč, P., Tarábek, P.: 'Deep neural networks classification via binary error-detecting output codes', *Applied Sciences*, vol. 11, no. 8, 2021, ISSN 2076-3417, doi:10.3390/app11083563

[10] LeCun, Y., Bengio, Y., Hinton, G.: 'Deep learning', *nature*, vol. 521, no. 7553, p. 436, 2015

[11] Miller, T.: 'Explanation in artificial intelligence: Insights from the social sciences', *ArXiv.org*, pp. 1–66, 2017, doi:arXiv:1706.07269

[12] Montavon, G.: 'Gradient-based vs. propagation-based explanations: An axiomatic comparison', *Explainable AI*, 2019

[13] Samek, W., Montavon, G., et al.: 'Toward interpretable machine learning: Transparent deep neural networks and beyond', *CoRR*, vol. abs/2003.07631, 2020

[14] Samek, W., Müller, K.: *Towards Explainable Artificial Intelligence*, pp. 5–22, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2019, doi:10.1007/978-3-030-28954-6_1

[15] Zhi-Han, Y.: 'Training latent variable models with auto-encoding variational bayes: A tutorial', 2022, doi:10.48550/ARXIV.2208.07818

7 Zoznam vlastných publikácií autora

- Martin Klimo, Jaroslav Kopčan, and Ľubomír Králik. "*Explainability as a method for learning from computers*" IEEE Access 11 (2023): 2169 – 353.
- Jaroslav Kopčan, Martin Klimo, and Ondrej Škvarek. "*Do neural networks recognize patterns as well as students?*" 2022 International Conference on Emerging eLearning Technologies and Applications: proceeding (ICETA). IEEE, 2022.
- Jaroslav Kopčan, Martin Klimo, and Ondrej Škvarek. "*Anomaly detection using Autoencoders and Deep Convolution Generative Adversarial Networks*" 2021 International scientific conference on sustainable, modern and safe transport: proceeding (TRANSCOM). IEEE, 2021.
- Jaroslav Kopčan, Martin Klimo, and Ondrej Škvarek. "*PCA Tail as the Anomaly Indicator*" 2020 International Conference on Emerging eLearning Technologies and Applications: proceeding (ICETA). IEEE, 2020.
- Jaroslav Kopčan. "*Explainable AI: a brief introduction*" 2021 Mathematics in science and technologies, proceedings of the MIST conference 2020.