MixBAI: Knjižnica za razlaganje biometričnih modelov z uporabo metod interpretacije mehanizmov

Robert Jutreša[1], Peter Rot[1,2], Peter Peer[1], Vitomir Štruc[2], Klemen Grm[2]

[1]*Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Večna pot 113*
[2]*Fakulteta za elektrotehniko, Univerza v Ljubljani, Tržaška cesta 25*
E-pošta: rj7149@student.uni-lj.si

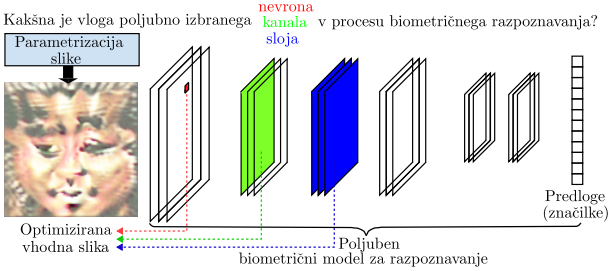
Abstract

In an effort to enhance the recognition of individuals using biometric models, researchers are developing increasingly complex models, making it challenging to interpret their decision making. To address this, we have developed the MixBAI library, offering biometric researchers a convenient and straightforward way to inspect such models using recently proposed mechanistic interpretability. Our aim is to implement recently proposed techniques for the interpretation of existing and newly developed recognition models in a plug-and-play manner. In this paper, we primarily focus on the functionality offered by our library that allows for the inspection of either individual components of the model, such as neurons, channels, or layers, or their combinations (e.g., multiplication or sum of neurons, etc.), by optimizing an input image to maximally activate a predefined selected component(s). We show that mechanistic interpretability can be a useful tool for understanding the inner workings of face recognition models.

1 Uvod

Za globoke nevronske mreže (NN) je značilno, da je njihov proces odločanja netransparenten, saj človek njihove odločitve lahko razume le delno [3]. Modeli za biometrično razpoznavanje, ki temeljijo na nevronskih mrežah, sicer dosegajo najbolj natančne rezultate, a je njihova praktična uporabnost omejena na aplikacije, kjer podrobna razložljivost odločitvega procesa ni strogo zahtevana. Razložljiva umetna inteligenca (ang. explainable artificial intelligence, XAI) človeku omogoča vpogled v odločitveni proces inteligentnih sistemov [10].

Biometrično razpoznavanja je tipičen primer, kjer je razložljivost odločitve bistvenega pomena, saj ima napačna odločitev modela lahko resne posledice za vpletene osebe (npr. zavrnitev prehoda čez mejo, obtoževanje nedolžnega za kaznivo dejanje, itd.). Raziskovalci so predlagali mnogo pristopov za razlago biometričnih modelov, vendar so javno dostopne rešitve tipično razvite za razlaganje specifičnega modela (tj. niso splošne). To pomeni, da uporaba obstoječih pristopov za razlaganje drugih modelov, ki še niso podprti, tipično zahteva veliko prilagoditev programskega okolja, kar raziskovalcem vzame veliko časa. V sklopu tega dela zato razvijemo



Slika 1: Ilustracija uporabe knjižnice MixBAI, s katero interpretiramo mehanizme v biometričnih modelih. Z njo lahko raziskovalci na enostaven način analizirajo, kakšne vhodne slike povzročijo v naprej določeno aktivacijo na poljubnem nevronu ali na njihovi poljubni kombinaciji (npr. aktivacija kanala, sloja, itd.), kar omogoča interpretacijo.

MixBAI knjižnico, ki podpira sodobne metode za interpretacijo, pri čemer razvijalci lahko na preprost način (i) zamenjajo obravnavan model, (ii) zamenjajo metodo za interpretacijo in (iii) spremenijo hiperparametre posamezne metode. S knjižnico lahko analiziramo posamezne nevrone, kanale in sloje, kot je ilustrirano na sliki 1.

Raziskovalci so predlagali mnogo parametrizacij in regularizacijskih postopkov, katerih cilj je izboljšati kvaliteto vhodnih slik in prikaz semantične informacije na njih [12, 14]. Izkaže se, da različni postopki dajejo bolj ali manj interpretabilne slike, pri čemer so v obstoječi literaturi te pristopi ovrednoteni na modelih za bolj splošno razpoznavanje (npr. za zbirko ImageNet). Ker pa imajo biometrični modeli drugačne lastnosti od tovrstnih modelov, je potrebno interpretabilnost oz. kvaliteto pridobljenih slik posebej oceniti tudi v konteksu biometričnega konteksta, kar je tudi eden izmed primarnih ciljev tega dela.

2 Sorodna dela

Razložljiva biometrija. Osnovni cilj razložljive biometrije je omogočiti razumevanje osnovnih gradnikov biometričnega sistema in interpretirati odločitve, ki le-ta sprejme [7]. Izčrpen pregled literature je na voljo v delih [5] in [13], v tem delu pa se osredotočamo predvsem na opis metod, ki razlagajo, kakšna je vloga posameznih gradnikov v modelu. Na visokem nivoju lahko metode za interpretacijo odločitev biometričnih modelov delimo v dve različni skupini. Cilj prve skupine je razložiti, katere

obrazne lastnosti na dveh različnih slikah obraza so povzročile odločitev, da (ne) gre za isto osebo [13]. Cilj druge skupine pa je interpretirati vlogo posameznih komponent (npr. nevrona, kanala ali sloja) v nevronski mreži [6]. Kljub temu, da MixBAI delno že podpira metode iz prve skupine, se v tem delu osredotočamo predvsem na opis podprtih metod iz druge skupine, katerih cilj je interpretirati vlogo posamezne komponente v omrežju. Ideja t.i. *pristopov za interpretacijo mehanizmov* (ang. mechanistic interpretability) je, da z analizo in razumevanjem posameznih komponent lahko sklepamo kako deluje celoten sistem [1]. Poudariti je potrebno, da se tovrstni mehanizmi razlikujejo od metod za interpretacijo, ki temeljijo na optimizaciji map intenzivnosti (angl. heatmap), kot sta na primer Grad-CAM [14] in IG [8].

Interpretacija preko optimizacije vhodne slike. Za mnoge biometrične sisteme velja predpostavka, da na vhod sprejmejo sliko biometričnega vzorca (npr. sliko obraza, očesa, ipd.) [2]. Zato lahko preko analize vhodnih slik in aktivacij v nevronskem omrežju, ki jih le-te povzročijo, sklepamo na vlogo posameznih delov omrežja pri procesu odločanja. Ker so odvodi v tovrstnih omrežjih tipično na voljo, lahko vhodno sliko optimiziramo glede na določen kriterij (npr. želimo, da vhodna slika maksimizira aktivacijo izbranega nevrona ali sloja). Pozitivna lastnost tovrstnega pristopa je, da so rezultati v obliki slik (tj. v formatu, ki ga ljudje lahko interpretiramo), za katere po optimizaciji sklepamo na njihovo vlogo v nevronskem omrežju. V podjetju Google so za tovrstno raziskovanje modelov, ki so definirani v okolju Tensorflow razvili knjižnico Lucid [1], katere delni prepis obstaja tudi za modele v okolju PyTorch [2]. V našem delu smo omenjeni knjižnici vzeli kot osnovo za razvoj knjižnice v okolju PyTorch, pri čemer smo funkcionalnost razširili in priredili za prikladno uporabo na področju biometrije.

3 Knjižnica MixBAI

V tej sekciji najprej natančneje opišemo postopek optimizacije vhodnih slik in nato predstavimo posamezne funkcionalnosti, ki so na voljo v knjižnjici MixBAI. Funkcionalnosti so povzete v tabeli 1 in natančneje opisane v naslednjih odstavkih.

Optimizacija vhodnih slik. Pri izbranem biometričnem modelu s prednaučenimi utežmi je cilj z optimizacijo vhodnih slik $\mathbf{x}$ ustvariti optimizirano sliko $\mathbf{x}^*$, ki v modelu vzbudi maksimalno aktivacijo določene komponente a_i, npr. nevrona, sloja, kanala, ali kombinacije le-teh. Aktivacijo ureja regularizacijska funkcija R, ki je parametrizirana z θ. Matematično to lahko zapišemo kot:

$$\mathbf{x}^* = \arg\max(a_i(\mathbf{x}) - R_\theta(\mathbf{x})).$$

Dodana vrednost knjižnice MixBAI je, da lahko razpoznavalni model in izbiro komponent, ki jih želimo analizirati, spreminjamo na preprost, modularen način.

Parametrizacije vhodnih slik. Biometrični modeli so odvedljivi v odvisnosti od svojega vhoda in parametrizacija slike mora biti prav tako odvedljiva. Preko gradientnega spusta lahko spreminjamo njene parametre (vrednosti slikovnih elementov, uteži, itd.), da zagotovimo čim boljšo vizualizacijo. Zato smo implementirali več parametrizacij slik, ki so definirane spodaj.

Parametrizacija v RGB prostoru. Najbolj preprost način parametrizacije vhodne slike je, da kot osnovo vzamemo tri kanalno sliko, kjer vrednosti za vsak RGB kanal naključno vzorčimo iz normalne porazdelitve (vrstica 5, tabela 1).

Parametrizacija v frekvenčnem prostoru (FFT). Vzorčimo lahko tudi vrednosti RGB kanalov iz frekvenčnega prostora. To storimo tako, da za x in y koordinate vzorčimo iz frekvence diskretne Fourierjeve transformacije, ki jih nato skaliramo, da dobimo enotno transformacijo. Nato z inverzom N-dimenzionalne diskretne Fourierjeve transformacije za realni vnos pridobimo vrednosti, ki jih petvorimo v željeno velikost slike [6]. Prednost tega je, da s tem zagotovimo prostorsko dekorelacijo vrednosti znotraj posameznega kanala. Za zagotovitev dekorelacije med barvnimi kanali se izvede razcep Choleskega na matriki kovarianc RGB kanalov slik, na katerih je bil model učen (vrstica 6, tabela 1).

Optimizacija α kanala. Optimizacija slik z RGB ali FFT parametrizacijo ima slabost, da z njo ne dobimo vpogleda v pomembnost posameznih značilk. S tem, ko parametrizaciji dodamo četrti kanal (α oziroma prosojnost), lahko v vhodni sliki optimiziramo tudi globino (kaj je v ospredju in ozadju slike). To dosežemo tako, da štiri kanalno sliko združimo nazaj v tri kanalno RGB sliko pred tem, ko jo damo v model, pri čemer jo zmešamo z naključnim ozadnjem glede na vrednosti kanala α (vrstica 9, tabela 1).

Parametrizacija CPPN [11]. Mreža za ustvarjanje kompozicijskih vzorcev (angl. Compositional Pattern Producing Network, CPPN) dodatno omeji optimizacijo s pravilom, da morajo biti sosednji slikovni elementi podobnih barv. Parametrizacija spreminja naključno inicializirane parametre več-nivojske nevronske mreže, ki preslika (x, y) koorinate slike v RGB vrednosti (vrstica 7, tabela 1).

Parametrizacija DIP [12]. Globok predpogoj o sliki (angl. Deep Image Prior, DIP) predpostavlja, da pravilno zasnovana arhitektura modela brez predhodnega učenja zajame nizko-nivojske značilke slik. DIP se veliko uporablja v literaturi za obnavljanje slik. V tem delu smo DIP implementirali z uporabo arhitekture UNet [9], z naključno inicializiranimi utežmi. UNet združuje informacije iz različnih nivojev, kar omogoča generiranje bolj razložljivih slik (vrstica 8, tabela 1).

Regularizacije L_1, L_2 [6] in popolna varianca (angl. total variation) [4]. Optimizirane slike zaradi naključne inicializacije vsebujejo veliko šuma. Prisotnost šuma lahko bodisi omejimo z regularizacijo gradienta (parametrizacija DIP ali FFT), ali pa regulariziramo samo sliko (vrstice 10, 11 in 13, tabela 1). Poleg zmanjšanega šuma, regularizacija povzroči hitrejšo konvergenco optimizacije.

Raznolikost (angl. diversity) [6]. Za boljšo utemeljivtev interpretacije tipično potrebujemo večje število slik. Na

[1] Lucid: `https://github.com/tensorflow/lucid`
[2] Lucent: `https://github.com/greentfrapp/lucent`

	Funkcionalnost	Enačba	Referenčna slika
Kriterijske funkcije	Aktivacija nevrona	$a_{i,c,x,y}(\mathbf{x}) = a_i[c,x,y]$	2c
	Aktivacija sloja	$a_i(\mathbf{x}) = \frac{1}{H \cdot W \cdot C} \sum_{l=0}^{H-1} \sum_{k=0}^{W-1} \sum_{h=0}^{C-1} a_i[h,l,k]$	2a
	Aktivacija kanala	$a_{i,c}(\mathbf{x}) = \frac{1}{H \cdot W} \sum_{l=0}^{H-1} \sum_{k=0}^{W-1} a_i[c,l,k]$	2b
	Poljubne interakcije	*n/a*	2d
Paramet-rizacije	RGB prostor	$\mathbf{x}_{k,x,y} \sim \mathcal{N}(\mu, \sigma^2)$	2e
	Frekvenčni prostor [6]	*n/a*	2f
	CPPN [11]	*n/a*	2g
	DIP [12]	*n/a*	2h
Regularizacije	α kanal	$\mathbf{x}'_{\text{RGB}} = \mathbf{x}_{\text{RGB}} \cdot \mathbf{x}_\alpha + \text{BG}_{\text{RGB}} \cdot (1 - \mathbf{x}_\alpha)$ $a_i(\mathbf{x}) = a_i(\mathbf{x}) \cdot (1 - \bar{\mathbf{x}_\alpha})$	2p
	L_1	$R_{L_1}(\mathbf{x}) = \sum_{l=0}^{H-1} \sum_{k=0}^{W-1} \sum_{h=0}^{C-1} \lvert\lvert\mathbf{x}[h,l,k]\rvert\rvert - \varepsilon$	2j
	L_2	$R_{L_2}(\mathbf{x}) = \sum_{l=0}^{H-1} \sum_{k=0}^{W-1} \sum_{h=0}^{C-1} \sqrt{(\mathbf{x}[h,l,k] - \varepsilon)^2}$	2k
	Raznolikost	$G_{i,j} = \sum_{x,y} a_i[k,x,y] \cdot a_i[l,x,y]$ $R_C(\mathbf{x}) = \sum_a \sum_{a \neq b} \frac{\text{vec}(G_a) \cdot \text{vec}(G_b)}{\lvert\lvert\text{vec}(G_a)\rvert\rvert \cdot \lvert\lvert\text{vec}(G_b)\rvert\rvert}$	2m, 2n,2o
	Popolna Varianca	$R_{TV}(\mathbf{x}) = \sum_{i,j} \left((x_{i,j+1} - x_{i,j})^2 + (x_{i+1,j} - x_{i,j})^2 \right)^{\frac{\beta}{2}}$	2i

$\mathbf{x}$ — parametrizacija vhodne slike, $\mathbf{x}'$ — nova vhodna slika, α — alfa kanal, a — aktivacija, i — indeks sloja, c — indeks kanala, x, y — koordinati nevrona, H — višina konv. sloja, W — širina konv. sloja, C — št. kanalov, $a_i[c,x,y]$ — indeks aktivacije, $x_{k,w,h}$ — vhodna slika s k kanali, višino w in širino h, $\mathbf{x}[h,l,k]$ — indeks slikovnega elementa v vhodni sliki, ε — manjša konstantna vrednost, BG — ozadje, R — regularizacijska funkcija, G — Gramova matrika, $\bar{\mathbf{x}_\alpha}$ — povprečna vrednost α kanala.

Table 1: Tabela podprtih funkcionalnosti v knjižnici MixBAI.

podlagi podobnosti / razlik med njimi, lahko sklepamo o tem ali neka komponenta res kodira določen faktor variacije (spreminjajujoč se faktor med slikami). Zagotovimo jo tako, da minimiziramo kosinusno podobnost med slikami, ki jih optimiziramo hkrati (vrstica 12, tabela 1).

4 Eksperimenti in rezultati

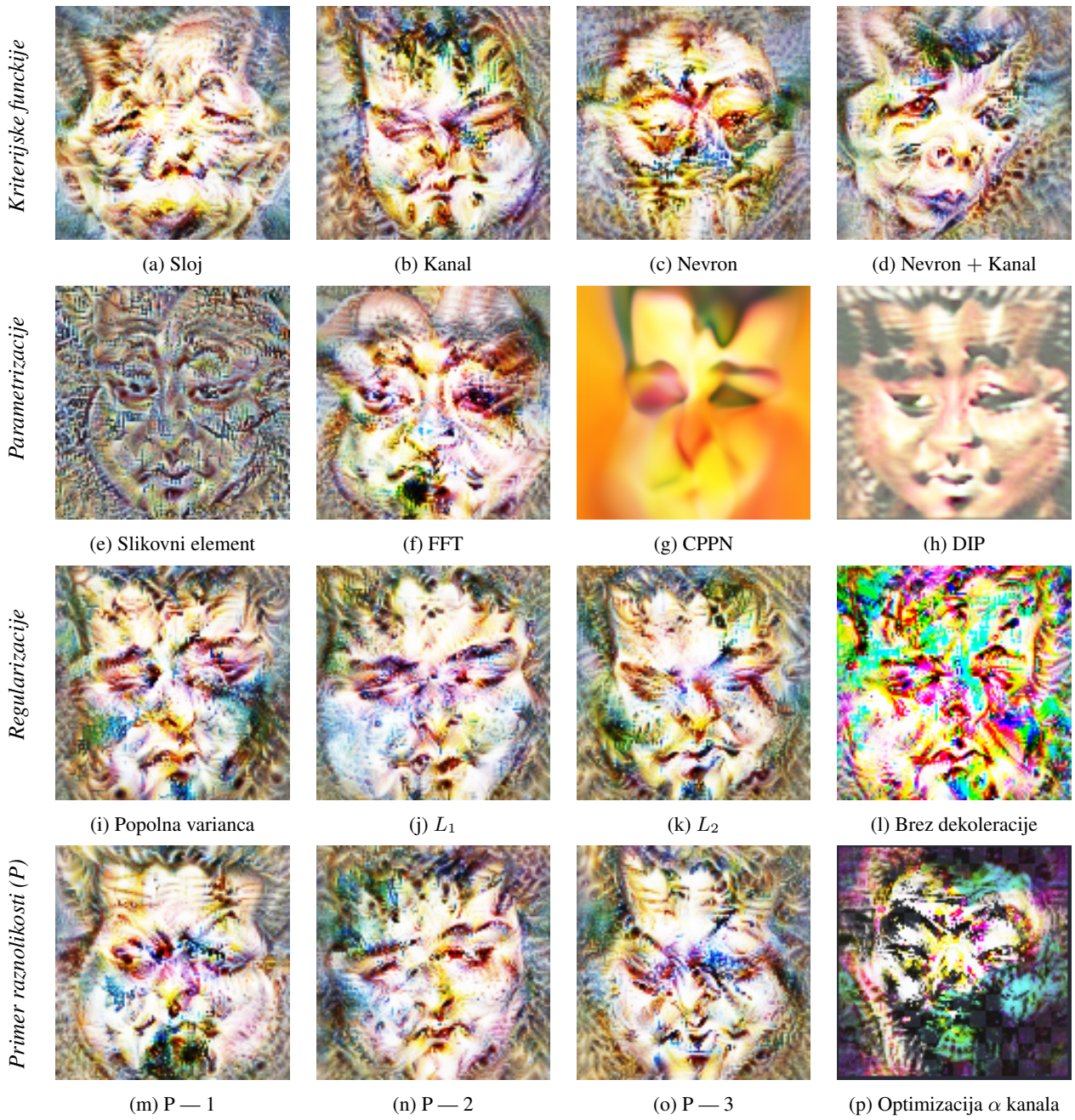
Biometrični model. Knjižnico MixBAI smo testirali z uporabo modela AdaFace [2], ki je trenutno eden izmed najbolj natančnih modelov za razpoznavanje obrazov.

Optimizacija. Za optimizacijo slike smo uporabili optimizator AdamW s 512 koraki optimizacije. Ugotovili smo, da je interpretabilnost rezultata močno odvisna od nastavitve stopnje učenja, ki smo jo nastavili na vrednosti $5\text{e}{-2}$, $2\text{e}{-3}$ (pri uporabi CPPN parametrizacije) ali $5\text{e}{-4}$ (pri uporabi DIP parametrizacije). Za vse vizualizacije na sliki 2 smo (če pod sliko ni navedeno drugače) uporablili tri kanalno FFT parametrizacijo vhodne slike. Ciljna komponenta za aktivacijo (če pod sliko ni navedeno drugače) je bila kanal. Gre za prvi kanal sloja z ReLU funkcijo zadnjega konvolucijskega bloka od modela AdaFace. V primeru, ko optimiziramo kombinacijo nevron + kanal, je kanal omenjen, medtem ko je centralni nevron prvega kanala vhod v ReLU funkcijo. Optimizacijo smo izvajali na grafičnih karticah NVidia RTX 4090 s 24GB VRAM-a, pri čemer optimizacija posamezne slike traja manj kot 25 sekund, ne glede na izbiro parametrizacije.

Ocenjevanje interpretabilnosti. V kontekstu našega dela kvalitativno ovrednotimo koliko informacije o obrazih lastnostih je razvidno iz slik, ki jih pridobimo z različnimi optimizacijskimi postopki. Na optimiziranih slikah želimo videti jasne obrazne atribute, kot so oči, nos, usta oz. ustnice, gube, ušesa, ipd., kar omogoča interpretacijo. Prav tako si želimo slike s čim manj šuma. Ker v sledečih primerih optimiziramo komponente, ki so v zadnjih slojih modela, pričakujemo, da se posamezni atributi sestavljajo v predstavitev celotnega obraza.

Kvalitativna analiza interpretabilnosti. Uporaba različnih parametrizacij, optimizacijskih ciljev in regularizacijskih funkcij daje zelo različne rezultate, ki so močno občutljivi od izbire hiperparametrov. Za AdaFace se je izkazalo, da s FFT (slika 2f) in DIP (slika 2h) dobimo najbolj interpretabilne vhodne slike. To trdimo na podlagi najbolj jasnih obraznih linij, med katerimi najbolj izstopata oblika obraza ter pozicija in oblika oči. Parametrizacija slikovnih elementov (slika 2e) pogosto vrne manj kvalitetne slike, ki jih je težje interpretirati. Z uporabo CPPN (slika 2g) nismo pridobili razložljivih rezultatov (odsotnost ust in zamegljene obrazne linije).

Z vidika regularizacije, metode vračajo podobne rezultate (slike 2i, 2j, 2k), je pa pomembno, da izvedemo dekorelacijo barvnega prostora, saj so v nasprotnem primeru optimizirane slike nereprezentativne. Optimizacija α kanala (slika 2p) pogosto vrne bolj jasne obrazne atribute, saj le-te loči od ostalega obraza (npr. opazimo bolj definirane čelne gube, oči in obrvna loka). Za kvaliteto raznolikosti (slike 2m, 2n, 2o) se izkaže, da je odvisna od zas-



(a) Sloj (b) Kanal (c) Nevron (d) Nevron + Kanal

(e) Slikovni element (f) FFT (g) CPPN (h) DIP

(i) Popolna varianca (j) L_1 (k) L_2 (l) Brez dekoleracije

(m) P — 1 (n) P — 2 (o) P — 3 (p) Optimizacija α kanala

Slika 2: Primeri optimiziranih vhodnih slik z uporabo različnih komponent modela za razpoznavanje (nevrona, sloja, kanala), različnih parametrizacij in regularizacijskih funkcij.

tavljenega cilja (slike 2b, 2c, 2d, 2l). Bolj kot je ta kompleksen, bolj raznoliki bodo rezultati in več informacij bomo pridobili o faktorjih variacije.

5 Sklepi in nadaljnje delo

Predstavili smo knjižnico MixBAI, ki razvijalcem na področju biometrije omogoča preprosto testiranje metod za interpretacijo biometričnih modelov, pri čemer smo se osredotočali na implementacijo optimizacijskih postopkov, ki maksimizirajo aktivacije posameznih komponent v izbranem modelu (AdaFace). V nadaljevanju bomo implementirali metode za interpretacijo obraznih lastnosti, na podlagi katerih se je model odločil, da sta si dva obraza podobna oz. ne. Raziskali bomo, na katerem nivoju model zaznava posamezne obrazne atribute, in kako se le-te sestavljajo v prestavitev celotenga obraza. Interpretirati bomo poskušali tudi druge sodobne modele za biometrično razpoznavanje, pri čemer ne bomo omejeni le na obraze, ampak bomo poskusili tudi na drugih biometričnih modalnostih (šarenica, periokularni del, itd.).

Zahvala

Raziskave so bile financirane preko ARIS programa P2-0250 “Metrologija in biometrični sistemi” ter ARIS razikovalnega projekta J2-50069 “Interpretacija mehanizmov za razložljivo biometrično umetno inteligenco (MIXBAI)”.

Viri in Literatura

[1] Leonard Bereska in Efstratios Gavves. »Mechanistic Interpretability for AI Safety–A Review«. V: *arXiv preprint arXiv:2404.14082* (2024).

[2] Minchul Kim, Anil K. Jain in Xiaoming Liu. *AdaFace: Quality Adaptive Margin for Face Recognition*. 2023. arXiv: 2204.00964 [cs.CV].

[3] Octavio Loyola-González. »Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View«. V: *IEEE Access* 7 (okt. 2019), str. 154096–154113.

[4] Aravindh Mahendran in Andrea Vedaldi. *Understanding Deep Image Representations by Inverting Them*. 2014. arXiv: 1412.0035 [cs.CV].

[5] Pedro C. Neto in sod. *Causality-Inspired Taxonomy for Explainable Artificial Intelligence*. 2024. arXiv: 2208.09500 [cs.CV].

[6] Chris Olah, Alexander Mordvintsev in Ludwig Schubert. »Feature Visualization«. V: *Distill* (2017).

[7] P. Jonathon Phillips in Mark A. Przybocki. »Four Principles of Explainable AI as Applied to Biometrics and Facial Forensic Algorithms«. V: *CoRR* abs/2002.01014 (2020). arXiv: 2002.01014.

[8] Zhongang Qi, Saeed Khorram in Fuxin Li. »Visualizing Deep Networks by Optimizing with Integrated Gradients.« V: *CVPR workshops*. Zv. 2. 2019, str. 1–4.

[9] Olaf Ronneberger, Philipp Fischer in Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].

[10] Waddah Saeed in Christian Omlin. »Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities«. V: *Knowledge-Based Systems* 263 (2023), str. 110273.

[11] Kenneth O Stanley. »Compositional pattern producing networks: A novel abstraction of development«. V: *Genetic programming and evolvable machines* 8 (2007), str. 131–162.

[12] Dmitry Ulyanov, Andrea Vedaldi in Victor Lempitsky. »Deep Image Prior«. V: *International Journal of Computer Vision* 128.7 (2020), str. 1867–1888.

[13] Jonathan R Williford, Brandon B May in Jeffrey Byrne. »Explainable face recognition«. V: *European conference on computer vision*. Springer. 2020, str. 248–263.

[14] Bolei Zhou in sod. »Learning deep features for discriminative localization«. V: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, str. 2921–2929.