Interpretacija mehanizmov obraznih biometričnih modelov s kontrastnim multimodalnim učenjem

Anastasija Manojlovska[1], **Vitomir Štruc**[1], **Klemen Grm**[1]

[1]*Fakuilteta za Elektrotehniko UL, Tržaška 25, Ljubljana*
E-pošta: am6417@student.uni-lj.si

Povzetek

Razložljiva umetna inteligenca (XAI) povečuje transparentnost sistemov umetne inteligence. Ta študija uporablja model CLIP (Contrastive Language-Image Pretraining) podjetja OpenAI za prepoznavanje obraznih atributov v podatkovni zbirki VGGFace2 z uporabo anotacij atributov iz podatkovne zbirke MAADFace. Z poravnavo slik in opisov v naravnem jeziku prepoznamo atribute, kot so starost, spol in pričeska, ter ustvarimo razlage v naravnem jeziku. Raziskujemo tudi integracijo predhodno naučenih modelov za prepoznavanje obrazov in dodajanje razvrščevalnih plasti za izboljšanje razvrščanja atributov. Prednaučen model CLIP, se je izkazal najboljše pri prepoznavanju atributov Moški in Črn, saj je dosegel vrednosti AUC 0,9891 oz. 0,9829.

1 Uvod

Razložljiva umetna inteligenca (XAI) postaja ključna v sodobnih sistemih umetne inteligence zaradi svoje preglednosti in vpogleda v odločitvene procese. Na hitro razvijajočem se področju računalniškega vida je presečišče med jezikom in slike ključno za napredek globokega učenja. V naši raziskavi uporabljamo model CLIP (Contrastive Language-Image Pretraining) [11], razvit pri OpenAI, za interpretacijo slik z opisi v naravnem jeziku. S tem modelom prepoznavamo obrazne atribute v zbirkah slik, kot je VGGFace [2], ki vključuje slike slavnih osebnosti z različnimi obraznimi značilnostmi.

Naš cilj je prepoznati atribute obraza, kot so starost, spol, pričeska in druge značilnosti, ter generirati naravnojezikovne razlage za zaznane značilnosti, kar povečuje razumevanje in zaupanje uporabnikov ter razložljivost sistemov umetne inteligence. Raziskujemo tudi integracijo predhodno naučenih modelov za prepoznavanje obrazov s klasifikacijskimi plastmi, kar povečuje natančnost in učinkovitost klasifikacije obraznih atributov ob ohranitvi interpretacijskih sposobnosti modela CLIP. Naš cilj je premostiti vrzel med tehnično zmogljivostjo in uporabniško razložljivostjo ter zagotoviti natančnost in razumljivost spoznanj, pridobljenih z umetno inteligenco.

2 Kratek pregled področja

Razložljiva umetna inteligenca (XAI) v biometriji. Razložljiva umetna inteligenca (XAI) je ključna za preglednost v sistemih umetne inteligence, zlasti v računalniškem vidu in biometriji. Arrieta et al. [1] predstavijo tehnike XAI in njihovo vlogo pri prepoznavanju slik. Williford et al. [13] prispevajo k razložljivemu prepoznavanju obrazov z merili in protokoli za ocenjevanje. Phillips in Przybocki [10] raziskujeta integracijo XAI v biometrične in obrazne forenzične algoritme.

Multimodalno učenje. Multimodalno učenje, ki združuje različne vrste podatkov, izboljšuje učinkovitost in razlago modelov umetne inteligence. Kiela et al. [7] dokazujejo učinkovitost povezovanja vizualnih in besedilnih podatkov, medtem ko Ho in Nvasconcelos [6] predstavita kontrastivno učenje z nasprotnimi primeri (CLAE) za bolj zahtevne pozitivne in negativne pare.

Prepoznavanje obraznih atributov. Liu et al. [9] predlagajo kaskadne konvolucijske nevronske mreže (LNet in ANet) za izboljšanje prepoznavanja obraznih atributov. Zhang et al. [16] razvijajo algoritem za prepoznavanje spola in ocenjevanje starosti z večopravilnim učenjem in konvolucijskimi nevronskimi mrežami za učinkovito ekstrakcijo značilnosti.

3 Metodologija

Naš pristop razlage obraznih predlog preko razpoznavanja atributov sestoji iz korakov:

1. Pridobivanja obraznih predlog iz slike.
2. Preizkusa razpoznavanja obraznih atributov iz predlog.

Preizkušena sta bila dva pristopa. V prvem se značilke slike pridobijo s slikovnim kodirnikom modela CLIP, tekstovni opisi pa se zakodirajo s tekstovnim transformerjem CLIP. V drugem pristopu se uporabijo vektorji značilk iz prednaučenega modela za prepoznavanje obrazov AdaFace [8], nato pa se dodajo linearni sloji za večznačno ali binarno razvrščanje atributov.

3.1 CLIP model

CLIP [11] ima dve glavni komponenti: slikovni in besedilni kodirnik. Slikovni kodirnik lahko temelji na arhitekturah ResNet(50/101) [5] ali Vision Transformer (ViT) [4], medtem ko je besedilni kodirnik zasnovan na arhitekturi Transformer.

Pri eksperimentih uporabljamo slikovni transformer ViT-B/32 in kodiramo pripadajoče tekstovne opise slik z besedilnim transformerjem CLIP. Prednaučen model CLIP dodatno naučimo na pare slik in generiranih opisov ter primerjamo uspešnosti.

Kodirnika preslikata slikovne in besedilne podatke v skupni prostor značilk, CLIP pa povezuje slike z besedilnimi opisi s kontrastno funkcijo izgube. Model CLIP smo dotrenirali na zbirki VGGFace2, kjer smo ustvarili besedilne opise slik na podlagi označenih atributov ter tvorili pare (slika, tekst). Za učenje uporabljamo funkcijo izgube prečne entropije, ki maksimizira kosinusne podobnosti $sim(.)$ med vektorji značilk I_{emb} in T_{emb}. Enačba 1 prikazuje izračun kosinusne podobnosti.

$$sim(I_{emb}, T_{emb}) = \cos(\theta(I_{emb}, T_{emb})) = \frac{I_{emb} \cdot T_{emb}}{\|I_{emb}\|\|T_{emb}|} = \frac{\sum_i I_{emb_i} T_{emb_i}}{\sqrt{\sum_i I^2_{emb_i}} \sqrt{\sum_i T^2_{emb_i}}}, \quad (1)$$

kjer sta I_{emb} in T_{emb} vektorja slikovnih oz. tekstovnih značilk, Θ pa je kot med njima v skupnem projekcijskem prostoru.

Funkcija izgube prečne entropije za pare (slika-tekst ali tekst-slika) je podana z naslednjo enačbo:

$$H(P, Q) = \sum_{i=1}^{N} -P_i \log Q_i, \quad (2)$$

kjer je P_i verjetnost $i - tega$ elementa v pravi porazdelitvi P (1 za pravilen par ali 0 za druge), Q_i pa verjetnost $i - tega$ elementa v napovedani porazdelitvi P, ki se izračuna iz logitov z uporabo funkcije softmax, N pa število vseh možnih parov.

Končna funkcija simetrične izgube prečne entropije (CE) je srednja vrednost izgub prečne entropije med slikami in besedili ter med besedili in slikami, kot prikazuje enačba 3.

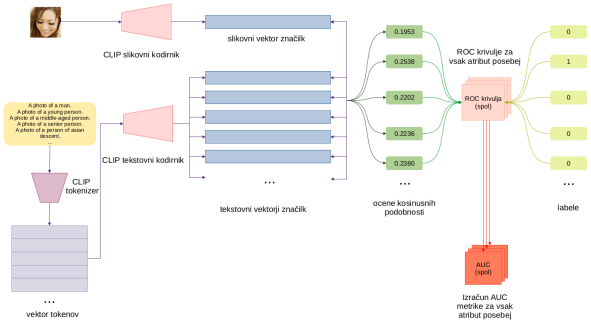
$$CE = \frac{H_I(P_I, Q_I) + H_T(P_T, Q_T)}{2}, \quad (3)$$

kjer $H_I(P_I, Q_I)$ predstavlja izgubo prečne entropije med sliko in besedilom, izpeljano iz logitov slike, $H_T(P_T, Q_T)$ pa izgubo prečne entropije med sliko in besedilom, iz logitov besedila.

Natrenirane modele uporabljamo za prepoznavo atributov, pri čemer izračunamo podobnost med slikami in besedili atributov v prostoru značilk. Na sliki 1 je grafično predstavljen ta postopek.

3.2 Večznačni razvrščevalnik in več binarnih razvrščevalnikov

AdaFace je model za prepoznavanje obrazov, ki izboljša kakovost značilk s prilagodljivo funkcijo izgube, temelječi na meji. Temelji na globoki konvolucijski nevronski mreži (npr. ResNet, Inception) in dinamično prilagaja mejo glede na težavnost vzorca, kar izboljša razločevanje v zahtevnih pogojih. Spremenjena funkcija izgube ArcFace [3] omogoča modelu, da se osredotoči na zahtevne vzorce, kar povečuje robustnost in zanesljivost sistema.



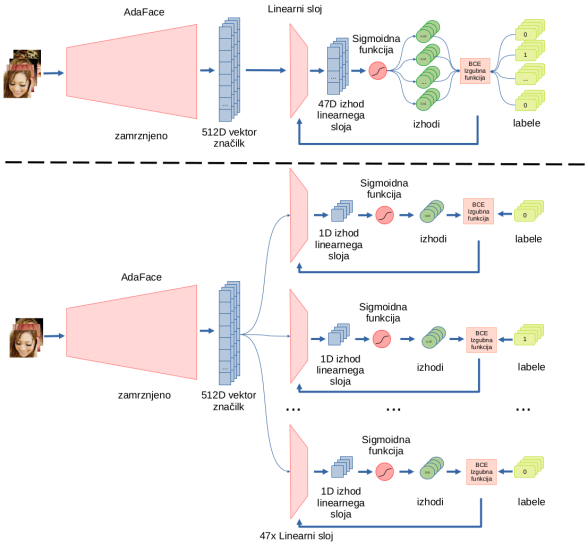
Slika 1: Postopek prepoznavanja atributov s pomočjo modela CLIP.

Modelu dodamo linearne klasifikacijske plasti na dva načina. Prvi način je večznačni klasifikacijski sloj, ki zaznava vse atribute na sliki. Vhod v ta sloj je 512-dimenzionalen vektor značilk iz AdaFace, izhod pa je 47-dimenzionalen, kar predstavlja število atributov. Tik pred generiranjem izhoda dodamo sigmoidno plast, ki zagotavlja, da se vsak atribut obravnava ločeno. Uporabimo funkcijo izgube Binarne Prečne Entropije (BCE), kot prikazuje enačba 4. AdaFace ostane zamrznjen in služi le za generiranje značilk.

$$BCE = -\frac{1}{N} \sum_{i=1}^{N} (y_i \log(\hat{y_i}) + (1 - y_i) \log(1 - \hat{y_i})) \quad (4)$$

V drugem pristopu modelu dodamo linearno klasifikacijsko plast s 512-dimenzionalnim vhodom in enim izhodom. Za vsak atribut posebej zgradimo svoj razvrščevalnik, ki ga učimo z uporabo funkcije izgube Binarne Prečne Entropije, pri čemer zamrznemo AdaFace.

Oba pristopa sta grafično predstavljena na sliki 2.



Slika 2: Postopek učenja večznačnega (zgornji del slike) in več binarnih razvrščevalnikov (spodnji del slike).

Uspešnost prepoznavanja posameznih atributov pred-

stavimo z metriko AUC (Area Under the Curve) pod ROC krivuljo (Receiver Operating Characteristic).

4 Eksperimenti

4.1 Podatkovne zbirke

Za ta projekt sta bili uporabljeni podatkovni zbirki VGGFace2 [2] in MAADFace [12] za anotacije atributov. VGGFace2 vsebuje 3,31 milijona slik 9131 oseb, povprečno 362,6 slike na osebo, z raznolikostjo v pozi, starosti, osvetlitvi, narodnosti in poklicu. MAADFace je zbirka anotacij atributov za VGGFace2 slike. Vsaka slika ima 47 binarnih atributov obraza, kot so mlad, privlačen, barva las, brada, nasmeh, klobuk itd. Primeri slik z različnimi pozami, starostjo, spolom in atributi so vidni na sliki 3.

VGGFace2 je razdeljena na učni del s 3 milijoni slik, na katerih izvedemo uglaševanje (*angl fine-tuning*) modelov, in testni del s 160 tisoč slikami, ki jih uporabljamo za validacijo modelov. Poleg slike, uporabljamo tudi pripadajoče anotirane atribute iz MAADFace zbirko.

Slika 3: Primeri slik iz VGGFace2 zbirke, ki prikazujejo različne atribute.

4.2 Podrobnosti eksperimentov

Slike najprej poravnamo s pomočjo algoritma MTCNN (Multi-task Cascaded Convolutional Networks) [15], ki predstavlja ogrodje globokega učenja, zasnovano za zaznavanje in poravnavo obrazov. Slike, pri katerih algoritem MTCNN ne zazna obraza, ne upoštevamo, posledično se število slik zmanjša za približno 10 %.

Tvorjenje besedilnih opisov. Za učenje modela CLIP je bilo treba generirati besedilne opise slik na podlagi anotiranih atributov. Glavna omejitev modela je nastavljena maksimalna dolžina konteksta 77 tokenov pri tokenizaciji, kar omejuje dolžino opisov in število atributov.

Na začetku smo upoštevali 32 atributov, pri čemer smo generirali približno 77,000 različnih besedilnih opisov za celotno zbirko, zaradi česar je med paketnim učenjem znotraj enega paketa bilo preveč ponovitev istega opisa.

Preprosta strategija za sprostitev omejitve je interpolacija pozicijskega vložka in učenje CLIP modela s pari (slika-tekst), ki vključujejo dolge opise slik, podobno kot je opisano v članku [14]. Na ta način smo povečali dolžino konteksta na 305 ($\lambda = 4$) ali 153 ($\lambda = 2$) tokenov, kar omogoča generiranje daljših opisov s 44 atributi.

4.3 Metrike vrednotenja

Za ocenjevanje naših modelov smo uporabili krivulje ROC (Receiver Operating Characteristic) in površino pod krivuljo (AUC) kot metrike vrednotenja. Krivulja ROC prikazuje razmerje med stopnjo resničnih pozitivnih rezultatov in stopnjo lažnih pozitivnih rezultatov pri različnih pragovih, krivulja AUC pa predstavlja eno vrednost, ki povzema uspešnost modela.

5 Rezultati

ROC krivulje (slika 4) in vrednosti AUC (tabela 1) primerjajo uspešnost različnih modelov pri prepoznavanju obraznih atributov.

Večznačni proti več binarnih razvrščevalnikov. Binarni razvrščevalniki so na splošno boljši od večznačnih razvrščevalnikov. To nakazuje, da se pri nalogah, kjer so razlike med dvema razredoma jasne in dobro ločene, lahko doseže višje AUC vrednosti.

CLIP-prednaučen. Ta model je najboljši pri atributih, kot so *Male* (0,9891) in *Black* (0,9829), ter je boljši od večznačnega model in pripadajočimi binarnimi modeli. Visoka učinkovitost predhodno naučenega modela kaže, da uporaba obsežnih in raznolikih zbirk podatkov med predhodnim učenjem pomaga učinkovito zajeti splošne lastnosti obrazov.

Donaučeni CLIP modeli. Donaučeni modeli kažejo mešane rezultate. Medtem ko se model *CLIP-fine-tuned-77* dobro pokaže pri atributih *White* (0,9651) in *Brown_Hair* (0,9556), so uglašeni modeli pri večih kontekstih še vedno boljši pri večini atributov. Model *CLIP-fine-tuned-153* je na splošno najboljši med uglašenimi različicami. Daljše dolžine konteksta (153 in 305) izboljšajo uspešnosti, ker omogočajo, da model ujame podrobnejše semantične odnose.

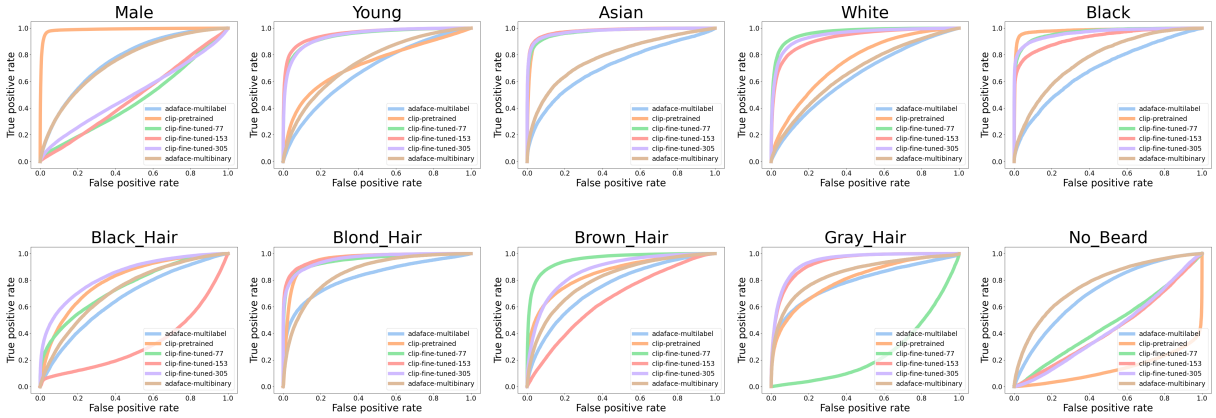
6 Zaključek

Ta študija opisuje uspešnost modela CLIP pri prepoznavanju obraznih atributov v podatkovni zbirki VGGFace. Model CLIP je najboljši pri prepoznavanju atributov *Male* in *Black*, saj je dosegel AUC 0,9891 in 0,9829, kar presega modele z več oznakami in binarne modele. Dodatno učenje je pokazalo nekaj izboljšav, vendar so se nekateri rezultati pri daljših kontekstih poslabšali zaradi kompleksnosti in morebitnega pretiranega prilagajanja. Uspeh modela CLIP poudarja učinkovitost obsežnih in raznolikih zbirk podatkov pri zajemanju posplošljivih značilnosti. Prednost večbinarnega modela pred modelom z več oznakami kaže prednosti učenja jasne ločitve med pozitivnim in negativnim razredom.

Model AdaFace za razpoznavanje obrazov ni sistematično boljši od prednaučenega CLIP modela. Iz tega sklepamo, da razpoznava obraze brez eksplicitnega kodiranja mehkih biometričnih atributov.

Zahvala

Raziskave v okviru pričujočega prispevka so bile financirane preko ARIS programa P2-0250 “Metrologija in biometrični sistemi” ter ARIS razikovalnega projekta J2-



Slika 4: ROC krivulje za izbrane atribute in posamezne modele.

Tabela 1: Vrednosti AUC za izbrane atribute posameznih modelov.

Model \ Attribute	*Male*	*Young*	*Asian*	*White*	*Black*	*Black_Hair*	*Blond_Hair*	*Brown_Hair*	*Gray_Hair*	*No_Beard*
AdaFace MultiLabel	0.7583	0.6797	0.7109	0.6394	0.6918	0.6640	0.8243	0.7474	0.7857	0.7070
AdaFace MultBinary	0.7467	0.7419	0.7829	0.6727	0.7704	0.7223	0.8548	0.8051	0.8491	**0.7686**
CLIP-pretrained	**0.9891**	0.7288	0.9723	0.7454	**0.9829**	0.7982	0.9507	0.8520	0.8048	0.1350
CLIP-fine-tuned-77	0.4620	0.9454	0.9722	**0.9651**	0.9670	0.7475	0.9513	**0.9556**	0.2555	0.4685
CLIP-fine-tuned-153	0.4848	**0.9580**	**0.9785**	0.9270	0.9293	0.3253	**0.9672**	0.6528	0.9341	0.4333
CLIP-fine-tuned-305	0.5073	0.9460	0.9756	0.9479	0.9642	**0.8381**	0.9585	0.8738	**0.9425**	0.4344

50069 “Interpretacija mehanizmov za razložljivo biometrično umetno inteligenco (MIXBAI)”.

Literatura

[1] Alejandro Barredo Arrieta, Francisco Herrera, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[2] Qiong Cao, Andrew Zisserman, et al. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[3] Jiankang Deng, Stefanos Zafeiriou, et al. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, October 2022.

[4] Alexey Dosovitskiy, Neil Houlsby, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Chih-Hui Ho and Nuno Nvasconcelos. Contrastive learning with adversarial examples. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17081–17093. Curran Associates, Inc., 2020.

[7] Douwe Kiela, Davide Testuggine, et al. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.

[8] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759, 2022.

[9] Ziwei Liu, Xiaoou Tang, et al. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[10] P Jonathon Phillips and Mark Przybocki. Four principles of explainable ai as applied to biometrics and facial forensic algorithms. *arXiv preprint arXiv:2002.01014*, 2020.

[11] Alec Radford, Ilya Sutskever, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[12] Philipp Terhörst, Arjan Kuijper, et al. Maad-face: A massively annotated attribute dataset for face images. *IEEE Transactions on Information Forensics and Security*, 16:3942–3957, 2021.

[13] Jonathan R Williford, Jeffrey Byrne, et al. Explainable face recognition. In *European conference on computer vision*, pages 248–263. Springer, 2020.

[14] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. *arXiv preprint arXiv:2403.15378*, 2024.

[15] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, October 2016.

[16] Wendong Zhang, Xianjing Zhou, et al. Research on face detection and face attribute recognition based on deep learning. In *2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, volume 1, pages 18–22. IEEE, 2021.