UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

Vladan Kovačević

PRIMENA METODE LRP U ANALIZI MODELA ZA PREPOZNAVANJE MODIFIKOVANIH LICA

master rad

Beograd, 2024.

Mentor:

dr Mladen Nikolić, vanredni profesor
Univerzitet u Beogradu, Matematički fakultet

Članovi komisije:

dr Jovana Kovačević, vanredni profesor
Univerzitet u Beogradu, Matematički fakultet

dr Aleksandar Kartelj, vanredni profesor
Univerzitet u Beogradu, Matematički fakultet

Datum odbrane: _____

Stricu Miroslavu

Naslov master rada: Primena metode LRP u analizi modela za prepoznavanje modifikovanih lica

Rezime: Duboke neuronske mreže preovladale su u mašinskom učenju u poslednjih nekoliko godina. U nekim zadacima koji su tradicionalno bili teški za računare uspele su da ostvare bolje rezultate čak i od čoveka. Ipak, njihova mana je što način na koji funkcionišu može biti vrlo težak ljudima za razumevanje. Zbog toga se uglavnom posmatraju kao crne kutije, trenirane da ostvare što bolje rezultate. Usled nedovoljnog razumevanja razloga zbog kojih model donosi odluke, često se dešava da model iz pogrešnih razloga daje dobre izlaze ili da ne koristi sve dostupne informacije, što može biti ozbiljan nedostatak, pogotovo kod modela koji se koriste u okviru bezbednosnih sistema. Tehnika preobražavanja lica predstavlja ozbiljnu pretnju sistemima za prepoznavanje lica zbog svoje jednostavnosti i efikasnosti. Prirodni izbor za prepoznavanje preobraženih lica su konvolutivne neuronske mreže, trenirane na velikom broju originalnih i preobraženih slika lica. Iako ove mreže ostvaruju visoku tačnost, njihova robusnost se dovodi u pitanje. LRP metoda je tehnika iz oblasti objašnjive veštačke inteligencije koja je primenu našla u mnogim zadacima u kojima se koriste slike. Korišćenjem LRP metode ustanovljeno je da se model pri prepoznavanju preobraženih lica najviše oslanja na oči, što nije poželjno, pogotovo kod modela koji se koriste u bezbednosti, jer ih takvo ponašanje čini podložnim raznim vrstama napada. Kako bi se prevazišli ovi nedostaci, podaci za trening su modifikovani tako da model bude nateran da u obzir uzme sve relevantne regije lica. Pomoću LRP metode utvrđeno je da se modeli trenirani na modifikovanim podacima pored očiju oslanjaju i na ostale bitne regije lica. Modeli trenirani na modifikovanim podacima imaju manju tačnost od klasično treniranih, ali ostvaruju bolje rezultate na ostalim metrikama i otporniji su na semantičke i suparničke napade, što ih čini robusnijim i pogodnijim za primenu u bezbednosti.

Ključne reči: mašinsko učenje, neuronske mreže, XAI, LRP, prepoznavanje lica, preobražavanje lica

Sadržaj

Glava 1

Uvod

Poslednjih nekoliko godina duboke neuronske mreže postale su najpopularniji skup modela u mašinskom učenju, ostvarivši značajne uspehe u mnogim oblastima. Za razliku od nekih jednostavnijih modela, dublje razumevanje procesa donošenja odluke kod dubokih neuronskih mreža je izuzetno teško zbog njihove kompleksnosti, te se obično posmatraju kao crne kutije, trenirane na velikoj količini podataka da za dati ulaz daju što tačnije predviđanje. Kod ovakvog pristupa često izostaje uvid u to zašto je za dati ulaz vraćen određeni izlaz. Može se desiti da modeli ostvare odlične rezultate, ali da ne koriste sve dostupne informacije na ulazu ili donose ispravne odluke iz pogrešnih razloga, što značajno umanjuje njihovu moć generalizacije i robusnost, te mogu biti podložni raznim vrstama napada koji uglavnom podrazumevaju zlonamerne modifikacije ulaza. Modeli koji se primenjuju u bezbednosti posebno su ugroženi kao česte mete napada.

Preobražavanje lica (eng. *face morphing*) je tehnika spajanja dve slike u kojoj se pomoću dva originalna lica generiše treće koje u istoj meri liči na originale. Ova tehnika predstavlja ozbiljnu pretnju po sisteme za prepoznavanje lica zbog svoje jednostavnosti i efikasnosti [9]. Sistemi za verifikaciju identiteta primarno koriste sistem za prepoznavanje lica koji često greškom može da lica obe osobe koje učestvuju u prevari identifikuje sa preobraženim licem. U nekim zemljama dozvoljeno je da osoba pri vađenju elektronskog dokumenta priloži svoju sliku. Ukoliko preobražena slika lica prođe sve kontrole uključujući ljudsku inspekciju i uđe u dokument, dve osobe mogle bi da koriste isti dokument. Problem prepoznavanja preobraženih lica privukao je pažnju zajednice mašinskog učenja prethodnih par godina. Za prepoznavanje preobraženih lica isprobani su razni modeli, ali i razne varijante preobražavanja, kao i dodatne modifikacije na preobraženim licima koje služe da uklone vizuelne tragove

koje ostavlja proces preobražavanja [25, 34]. Najbolje rezultate u rešavanju ovog problema uglavnom ostvaruju konvolutivne neuronske mreže, trenirane na velikom broju originalnih i lažnih (preobraženih) slika. Iako ove mreže ostvaruju visoku tačnost, njihova robusnost se dovodi u pitanje i potrebno ju je dodatno ispitati. Bitno je napomenuti da postoje i drugi pristupi za prepoznavanje preobraženih lica koji koriste dodatnu sliku jednog originalnog lica pored lica za koje se određuje da li je preobraženo ili ne, kao i pristupi kod kojih se na osnovu preobraženog i jednog originalnog lica vrši obrnuti proces preobražavanja - dobijanje drugog originalnog lica. U ovom radu neće biti razmatrani navedeni pristupi, već će fokus biti na glavnom problemu - na osnovu date slike lica odrediti da li je ona nastala preobražavanjem ili nije.

Objašnjiva veštačka inteligencija ili XAI (eng. *explainable artificial intelligence*) je oblast veštačke inteligencije čiji je cilj da sisteme koji koriste veštačku inteligenciju učini razumljivijim za čoveka. Najviše se primenjuje u mašinskom učenju, te možemo reći da je cilj XAI da na neki način objasni odluke modela mašinskog učenja i učini ih razumljivijim za ljude. LRP (eng. *layer-wise relevance propagation*) metoda je tehnika iz oblasti objašnjive veštačke inteligencije koja je primenu našla u mnogim problemima mašinskog učenja [18]. Za datu sliku na ulazu LRP metoda vraća toplotnu mapu (mapu relevantnosti) čije vrednosti predstavljaju relevantnost piksela, odnosno u kojoj meri je svaki piksel doprineo odluci modela. Korišćenjem LRP metode ustanovljeno je da se model treniran za detekciju preobraženih slika lica pri prepoznavanju preobraženih lica najviše oslanja na oči. Takvo ponašanje modela koji se primenjuju u bezbednosti nije poželjno, jer ih čini podložnim raznim vrstama napada. Da bi se taj nedostatak prevazišao, preobražene slike za trening su modifikovane kako bi se model naterao da u obzir uzme sve relevantne regije lica. Modifikacije podrazumevaju isecanje nekih glavnih regija preobraženog lica - očiju, usta, nosa i umetanje tih regija u jedno od dva originalna lica [27]. Pretpostavka je da će ciljanim pristustvom lažnih atributa na samo određenim regijama model biti nateran da se fokusira na sve bitne regije, te bi treniranje modela na modifikovanim podacima trebalo da učini modele robusnijim, što je ispitano u ovom radu. U cilju provere ove pretpostavke ispitana je otpornost modela na alternativne tehnike preobražavanja, semantičke napade i suparničke napade. Ovi napadi su česti kod modela koji se primenjuju u bezbednosti, te bi sposobnost modela da im se odupre trebalo uzeti kao bitnu meru kvaliteta [27]. U glavi 3 dat je kratak pregled objašnjive veštačke inteligencije i LRP metode. U glavi 4 opisana je tehnika preobražavanja slika

lica, kao i problemi koje ona predstavlja u bezbednosti. Detaljno su opisane dve različite tehnike preobražavanja. U glavi 5 navedeni su skupovi podataka korišćeni u ovom radu i opisane su modifikacije na podacima. U glavi 6 detaljno je opisana postavka problema, kao i različite varijante treninga. Ispitana je EER ocena kao glavna mera tačnosti na test skupu za različito trenirane modele, kao i otpornost na razne napade. Za dodatnu analizu rezultata i ponašanja različitih modela korišćena je LRP metoda.

Glava 2

Neuronske mreže i suparnički napadi

Neuronske mreže predstavljaju veoma širok spektar modela mašinskog učenja. Zbog svog uspeha u raznim poljima računarstva kao što su računarski vid i obrada prirodnih jezika, a od nedavno i generisanje slika i teksta, neuronske mreže često se (pogrešno) uzimaju kao sinonim za mašinsko učenje, pa čak i veštačku inteligenciju uopšte. Iako se kao ideja prvi put javljaju krajem pedesetih godina prošlog veka, nagli skok popularnosti doživele su tek početkom druge decenije 21. veka zahvaljujući razvoju hardverskih komponenti kao što su GPU (eng. *graphics processing unit*) i TPU (eng. *tensor processing unit*) potrebnih za njihovo efikasno treniranje, te ne čudi podatak da je najveći proizvođač ovih komponenti - NVIDIA, postao najvrednija kompanija na svetu u jednom trenutku [4].

Neuronske mreže možemo posmatrati kao usmereni težinski graf gde su čvorovi neuroni, a težine parametri koji se menjaju u procesu treniranja. Neuroni su funkcije koje (uglavnom) predstavljaju sumu ulaznih atributa pomnoženih težinama na koju se primenjuje aktivaciona funkcija. Cilj aktivacione funkcije je da uvede nelinearnost u ove modele i time proširi skup funkcija koje oni mogu da aproksimiraju. Neke od poznatih aktivacionih funkcija su $ReLU$, *sigmoidna* i $softmax$. U zavisnosti od strukture grafa dobijamo različite arhitekture neuronskih mreža. Jedna od najstarijih arhitektura su potpuno-povezane mreže (eng. *multi-layer perceptron*), koje karakteriše grupisanje neurona u slojeve tako da su neuroni jednog sloja povezani sa svim neuronima narednog sloja. Duboko učenje (eng. *deep learning*) karakteriše učenje složenih atributa i njihovih veza iz podataka, zbog čega su duboke neuronske mreže (u nastavku rada samo neuronske mreže ili mreže) veoma uspešne u radu sa sirovim podacima kao što su slike i zvuk. Svaki skriveni sloj mreže uči kompleksnije reprezentacije na osnovu onih koje dobija iz prethodnog sloja. Na ovaj način mreže

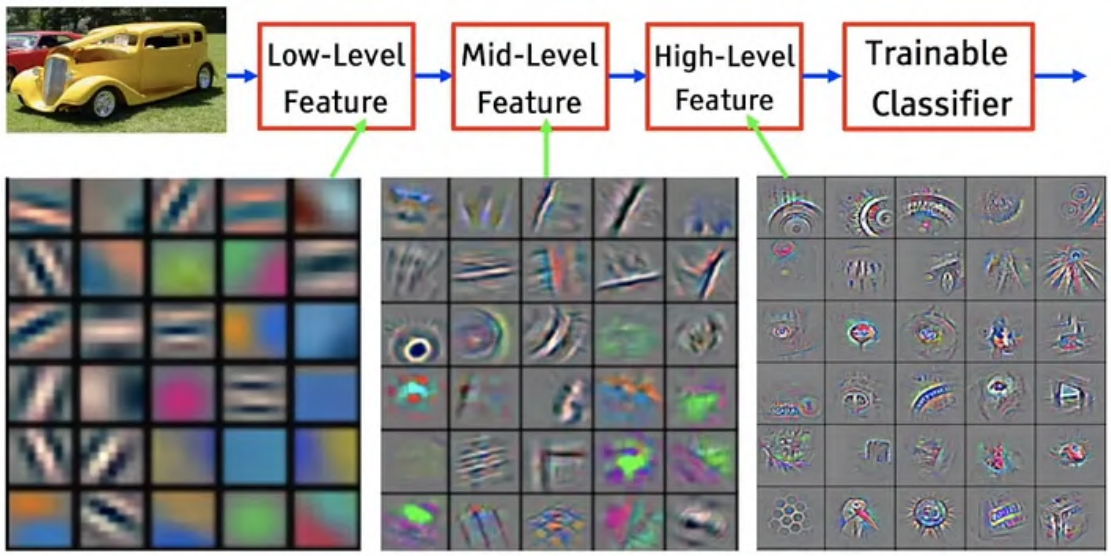
uče obrasce koje ljudi prirodno opažaju, te su zahvaljujući tome računari postali jednako uspešni (nekad čak i uspešniji) u rešavanju nekih problema koji su ljudima intuitivni, kao što su klasifikacija slika, odgovaranje na vizuelna pitanja, detekcija objekata na slici i slično.

Treniranje mreže podrazumeva minimizaciju greške, koja predstavlja odstupanje odluke mreže od ciljne promenljive za datu instancu, mereno funkcijom greškeF. U slučajevima klasifikacije, često korišćene funkcije greške su binarna i kategorička unakrsna entropija (eng. *categorical cross-entropy*). Minimizacija se uglavnom vrši metodama koje koriste gradijent, kao što su gradijentni spust i *adam*. Propagacija unazad (eng. *backpropagation*) je algoritam koji efikasno računa gradijente na osnovu pravila za kompoziciju funkcija i omogućava optimizaciju parametara sloj po sloj, od izlaza ka ulazu. Treniranje se zbog efikasnosti retko vrši na celom trening skupu odjednom, već se bira određeni broj instanci koji zovemo veličina podskupa (eng. *batch size*). Neke od mera kvaliteta modela za binarnu klasifikaciju su tačnost (eng. *accuracy*) udeo stvarno pozitivnih ili TPR (eng. *true positive rate*), udeo stvarno negativnih ili TNR (eng. *true negative rate*), udeo lažno pozitivnih ili FPR (eng. *false positive rate*) i udeo lažno negativnih ili FNR (eng. *false negative rate*). Kod modela koji se koriste u bezbednosti FPR i FNR se još zovu i udeo lažno prihvaćenih i udeo lažno odbijenih ili FAR i FRR (eng. *false accepted rate, false rejected rate*). FAR i FRR zavise od praga klasifikatora, iz tog razloga, češće korišćena mera je EER (eng. *equal error rate*). EER predstavlja FAR, odnosno FRR, kad je prag podešen tako da je FAR=FNR.

Konvolutivne neuronske mreže

Ljudski vid složene oblike registruje na osnovu jednostavnijih oblika i njihovog relativnog položaja. Na taj način čovek je u stanju da razazna različite figure na sceni koju posmatra. Filteri za prepoznavanje nekih jednostavnih oblika koristili su se u obradi slika i pre pojave neuronskih mreža, međutim, njih su kreirali ljudi i njihova primena bila je ograničena. Konvolutivne neuronske mreže inspirisane su načinom na koji ljudski vid procesira slike. Primenom operacije konvolucije za različite filtere izdvajaju se različiti atributi, po čemu su ove mreže i dobile ime. Ono što ih karakteriše je učenje filtera, bez potrebe čoveka da ih kreira. Pomoću naučenih filtera u stanju su da prepoznaju različite oblike na slici, a struktura neuronskih mreža omogućava im da na osnovu naučenih veza između jednostavnijih atributa grade složenije, od elementarnih oblika kao što su ivice, ćoškovi i kružići u početnim

slojevima, do veoma složenih (na primer oči, usta, uši) u kasnijim slojevima. Filteri koje ove mreže uče su možda najbolji vizuelni primer kako učenje reprezentacija kroz slojeve funkcioniše (slika 2.1). Njihova primena danas prevazilazi domen računarskog vida, ali ipak su ostale dominantne u onome za šta su primarno kreirane, a to je rad sa slikama.



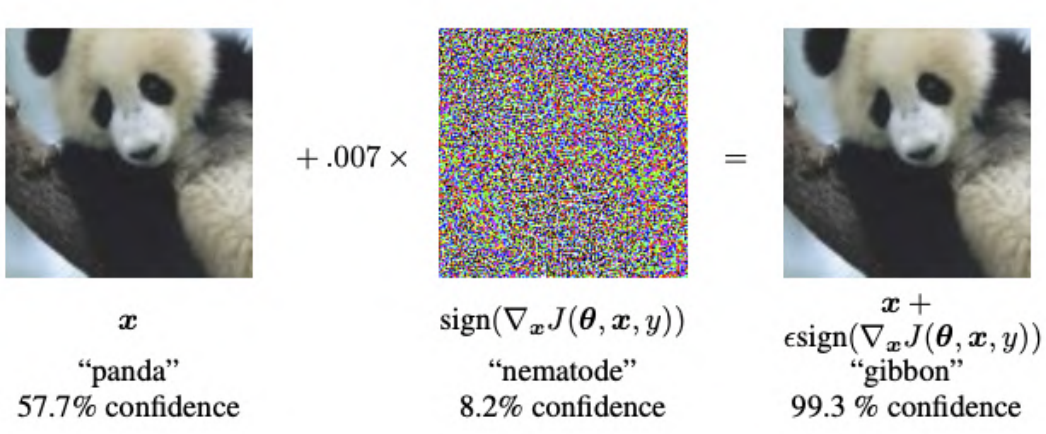
Slika 2.1: Vizuelni prikaz filtera koje konvolutivne mreže uče u različitim slojevima [36]

Konvolutivne neuronske mreže sastoje se od niza konvolutivnih slojeva, između kojih se često vrši agregacija kako bi se smanjila dimenzija izlaza. Agregacija obično podrazumeva zamenu dela izlaza maksimumom (eng. *max pooling*) ili prosekom (eng. *average pooling*) tog dela. Kako bi trening bio stabilniji i brži, nekad se dodaju i slojevi unutrašnje standardizacije (eng. *batch normalization*), a za regularizaciju se koristi izostavljanje (eng. *dropout*). Veoma poznata i dosta jednostavna arhitektura koja je u vreme nastanka ostvarila najbolje rezultate na ILVSRC [23] takmičenju (*ImageNet* skup podataka) je VGG19 [30]. Nešto novija i naprednija mreža, sa dosta komplikovanijom arhitekturom je Inception v3 [31]. Pri korišćenju ovih arhitektura u kreiranju novih, uglavnom se koriste težine već istrenirane mreže na velikim skupovima podataka (eng. *transfer learning*). Gornji (potpuno-povezani) slojevi koji u originalnoj mreži služe za klasifikaciju u zadatku za koji su trenirane menjaju se slojevima koji više odgovaraju novom zadatku. Konvolutivni deo mreže koji se zadržava (eng. *backbone*) koristi se za izvlačenje atributa (eng. *feature extraction*). Nekad se određeni broj slojeva (uglavnom onih bližih ulazu) zamrzava kako bi se

smanjio broj parametara i trening ubrzao, imajući u vidu da su neki univerzalni atributi već naučeni pri korišćenju težina ranije trenirane mreže na velikom skupu podataka. Fino podešavanje (eng. *fine-tuning*) je dodatan trening, nekad izvršen samo na određenim slojevima mreže u cilju daljeg prilagođavanja modela problemu koji rešava.

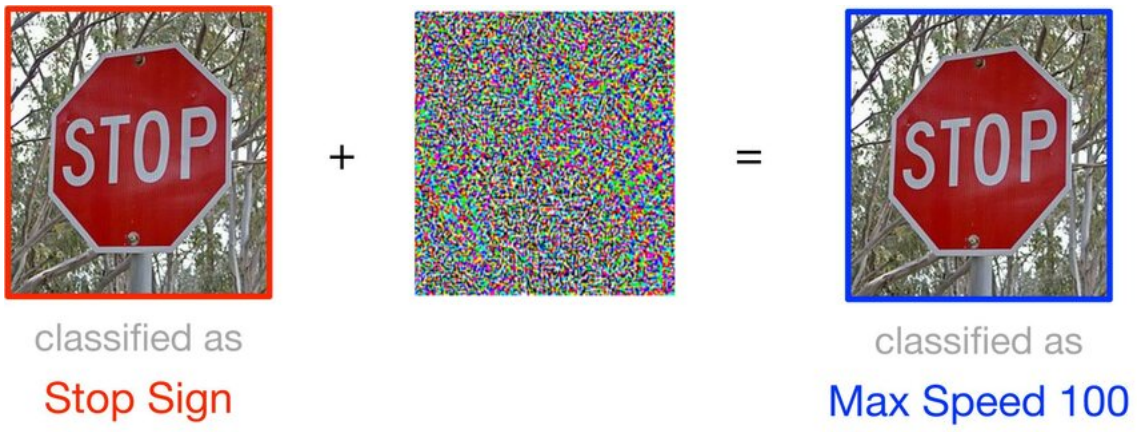
Suparnički napadi

Suparnički napad (eng. *adversarial attack*) je vrsta napada na model mašinskog učenja u kojem se eksploatišu njegove slabosti tako što se suptilnim modifikacijama ulazne instance model natera da donese pogrešnu odluku (slika 2.2). Modifikovane ulazne instance zovu se suparnički primeri (eng. *adversarial examples*) [10]. U idealnom slučaju, čovek ne može da razlikuje originalnu od modifikovane instance, što ove napade može učiniti opasnim i dovesti u pitanje bezbednost korišćenja modela mašinskog učenja u nekim poljima kao što je autonomna vožnja (slika 2.3). Zanimljive primene ovi napadi našli su na uličnim protestima, gde ljudi dizajniraju suparničku odeću čiji je cilj da prevari kamere za nadzor koje koriste sisteme za detekciju ljudi i prepoznavanje lica [12].



Slika 2.2: Primer suparničkog napada na model za klasifikaciju slika [10]

Suparnički primeri dobijaju se dodavanjem specijalno generisanog šuma kontrolisanog intenziteta na originalnu instancu. U idealnom scenariju, ako naš klasifikator f za ulaz x daje kao izlaz klasu $f(x)$, onda je šum koji tražimo najmanje s za koje važi $f(x) \neq f(x+s)$. Želimo da šum što slabijeg intenziteta utiče u što većoj meri na odluku modela. Jedan od načina dobijanja šuma za datu instancu je primenom metode FGSM (eng. *fast gradient sign method*) [10] na model. U scenariju napada



Slika 2.3: Primer suparničkog napada na model za prepoznavanje saobraćajnih znakova [39]

crnom kutijom (eng. *black-box attack*) napadač nema uvid u detalje modela kao što su tačna arhitektura i težine, jedino što ima na raspolaganju je korišćenje modela kao crne kutije - za dati ulaz može da dobije odgovarajući izlaz. Na ovakav model nije moguće primeniti FGSM. Međutim, uspešno generisanje šuma moguće je čak i sa mnogo slabijim pretpostavkama [20] - napadač nema nikakav uvid u arhitekturu modela i ne zna na kojim skupovima podataka je model treniran. Sa vrlo malo informacija, napadač je u stanju da definiše zamenski, suparnički model (eng. *adversary*) i istrenira ga, tako da pomoću njega dobije šum za željene instance. U ovom radu korišćene su nešto jače pretpostavke: pretpostavićemo da ne znamo arhitekturu modela koji napadamo, ali imaćemo uvid u podatke na kojima je model treniran. Suparnički model biće treniran na skupu koji je služio kao validacioni skup u treniranju modela koji napadamo.

Glava 3

XAI

Neuronske mreže ostvarile su ogromne uspehe u raznim poljima računarstva, a i šire. Ipak, način na koji one funkcionišu previše je kompleksan za ljude da bi ga u potpunosti razumeli. Uglavnom se posmatraju kao crne kutije (eng. *black-box models*) koje se treniraju da ostvare što bolje rezultate na datim metrikama, bez dodatnog objašnjenja zašto za dati ulaz vraćaju neki izlaz. U mnogim situacijama javlja se potreba za dubljom analizom modela, van samih rezultata koje oni daju.

Objašnjiva veštačka inteligencija ili XAI (eng. *explainable artificial intelligence*) je skup metoda u mašinskom učenju koje teže da odluke modela učine razumljivijim za ljude, sa ciljem postizanja dubljeg razumevanja modela, veće transparentnosti, stepena poverenja, robusnosti itd. Neki od razloga zbog kojih je ovo poželjno su [24]:

- verifikacija i poverenje,
- pravni aspekti,
- učenje novih koncepata od modela,
- poboljšanje modela.

Kada govorimo o verifikaciji i poverenju, želimo da utvrdimo da li model zaista uči ono što treba ili možda donosi ispravne odluke iz pogrešnih razloga [15]. Ovo je posebno bitno kod modela koji se koriste u bezbednosne, administrativne i medicinske svrhe. Često se dešava da se model pri donošenju odluke fokusira na pogrešne delove ulaza ili koristi vrlo malo informacija, što umanjuje njegovu moć generalizacije. U primenama u bezbednosti ovo može učiniti model podložnim raznim vrstama napada. U primenama u medicini želimo da znamo na primer na osnovu čega je

tkivo na slici klasifikovano kao tumor. Pravni aspekti odnose se na pravo na objašnjenje, izbegavanje diskriminacije, dodeljivanje odgovornosti, poštovanje ljudskih prava itd. Jedan od prvih zakona koji reguliše korišćenje sistema veštačke inteligencije tako što ih razvrstava u različite kategorije rizika je Akt Evropske unije o veštačkoj inteligenciji [33]. Jedan od primera učenja novih koncepata od modela su nove strategije u društvenim igrama kao što je go. Dubljim razumevanjem modela možemo utvrditi koje su njegove slabosti i kako možemo da ga poboljšamo, kako po pitanju performansi, tako i po pitanju bezbednosti, što je i fokus ovog rada.

Jedan od načina objašnjavanja odluke modela su toplotne mape (*eng. heatmaps*) kreirane od ulaznih instanci. Toplotne mape predstavljaju *lokalna* objašnjenja - daju odgovor na pitanje gde i u kojoj meri leži fokus modela za datu instancu. Od njih je moguće kreirati i *globalna* objašnjenja koja nezavisno od pojedinačnih ulaza daju uvid u ponašanje modela. Kod toplotnih mapa koje objašnjavaju ponašanje modela vrednosti atributa treba da budu srazmerne značaju odgovarajućeg atributa ulazne instance. Kod modela koji na ulazu dobijaju sliku, vrednosti piksela toplotne mape označavaju *relevantnost* tog piksela u donošenju odluke modela, odnosno u kojoj meri je taj piksel doprineo odluci modela.

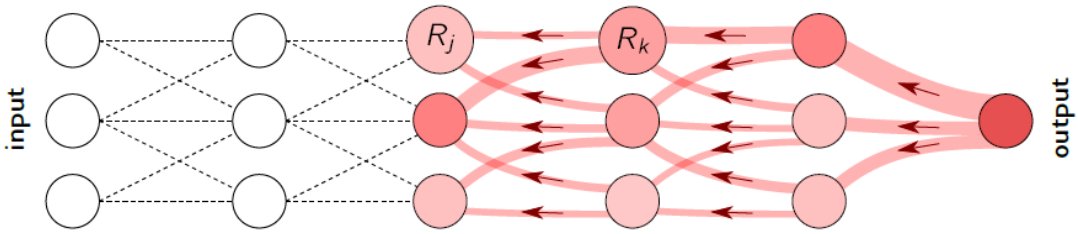
Ne postoji tačna definicija toplotnih mapa koje XAI metode vraćaju, ali postoje smernice koje nas upućuju na kvalitet metode. *Verodostojnost* nam govori da li metoda zaista oslikava nešto smisleno (npr. kod klasifikacije slika, da li se pri izlazu za klasu mačka može uočiti da se model na ulazu zaista fokusirao na atribute mačke), *razumljivost* podrazumeva da čovek može da razume objašnjenje koje metoda vraća, *efikasnost* garantuje da je problem koji metoda rešava dovoljno lak za rešavanja u smislu složenosti izračunavanja [3, 24]. Neformalno, najbolja ocena kvaliteta metode glasi - *metoda je dobra ukoliko su objašnjenja koja daje korisna na neki način.*

LRP

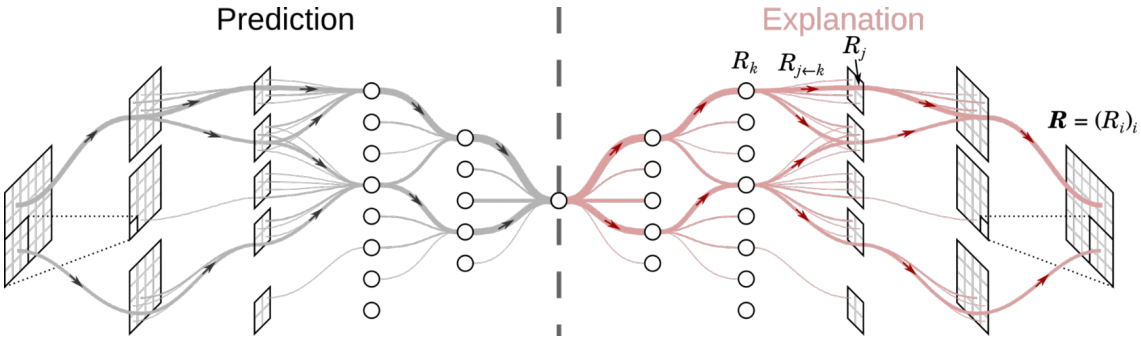
LRP (eng. *Layer-wise relevance propagation*) je XAI metoda koja se zasniva na dekompoziciji izlaza modela $f(x)$ za dati ulaz x, oslanjajući se na strukturu neuronske mreže kao acikličnog usmerenog grafa čiji su čvorovi (neuroni) grupisani u slojeve [18]. Dekompozicijom se relevantnost propagira sloj po sloj, od izlaza do ulaza, vodeći računa o zakonu održanja - *ukupna količina relevantnosti ne menja se kroz slojeve, samo se preraspoređuje po neuronima* (slike 3.1 i 3.2).
Neka su $N(l)$ indeksi neurona l-tog sloja, R_i^l relevantnost neurona l-tog sloja sa indeksom i, $f(x)$ izlaz modela za ulaz x. Tada je zakon održanja relevantnosti:

$$\sum_{i \in N(0)} R_i^0 = \sum_{i \in N(1)} R_i^1 = ... = \sum_{i \in N(l)} R_i^l = ... = f(x),$$

gde su R_i relevantnosti pojedinačnih neurona i-tog sloja.



Slika 3.1: LRP - propagiranje relevantnosti [18]



Slika 3.2: LRP - propagiranje relevantnosti za konvolutivne mreže [38]

LRP metoda ne daje eksplicitno pravila propagacije, već samo definiše ograničenja koja ona moraju da zadovoljavaju. Zbog toga imamo različite varijante LRP pravila u zavisnosti od slojeva na koje se primenjuju. Generalni oblik LRP pravila glasi:

$$R_j^l = \sum_{k \in N(l+1)} \frac{z_{jk}}{\sum_{j \in N(l)} z_{jk}} R_k^{l+1},$$

gde z_{jk} označava u kojoj meri je neuron j doprineo relevantnosti neurona k u narednom sloju.

Neka je $a_k^{l+1} = max(0, \sum_j a_j^l w_{jk})$ ReLU aktivacija k-tog neurona $l+1$ sloja. Neka često korišćena LRP pravila propagacije koja su primenjena u ovom radu su:

- LRP - 0 (osnovno pravilo)

$$R_j^l = \sum_{k \in N(l+1)} \frac{a_j^l w_{jk}}{\sum_{j \in N(l)} a_j^l w_{jk}} R_k^{l+1}$$

koje je pogodno za potpuno-povezane slojeve.

- LRP - ϵ

$$R_j^l = \sum_{k \in N(l+1)} \frac{a_j^l w_{jk}}{\epsilon + \sum_{j \in N(l)} a_j^l w_{jk}} R_k^{l+1}$$

koje je pogodno za srednje slojeve, do potpuno-povezanih. Hiperparametar ϵ „upija“ slabe relevantnosti. Iako strogo govoreći ne prati zakon održanja relevantnosti, za male vrednosti ϵ u praksi se dosta dobro pokazuje u generisanju jasnijih toplotnih mapa.

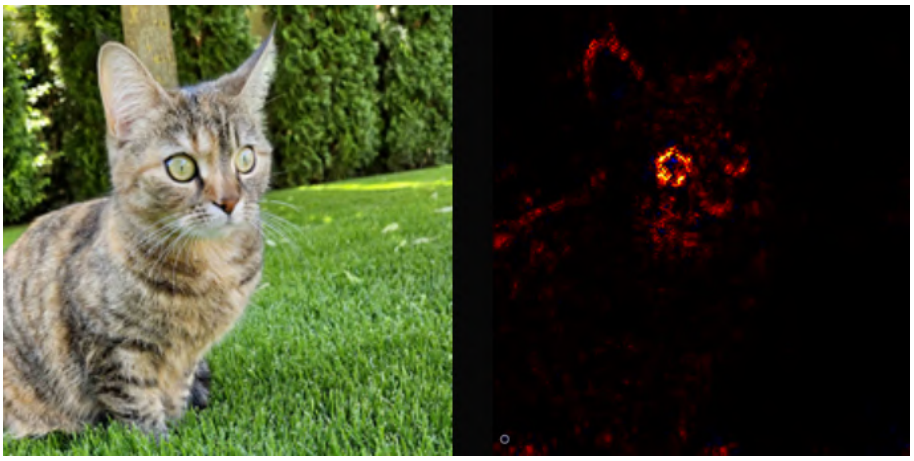
- LRP - $\alpha\beta$

$$R_j^l = \sum_{k \in N(l+1)} \left(\alpha \frac{(a_j^l w_{jk})^+}{\sum_{j \in N(l)} (a_j^l w_{jk})^+} - \beta \frac{(a_j^l w_{jk})^-}{\sum_{j \in N(l)} (a_j^l w_{jk})^-} \right) R_k^{l+1},$$

gde je $(x)^+ = max(0, x)$, a $(x)^- = min(0, x)$. Hiperparametri α i β kontrolišu u kojoj meri favorizujemo pozitivne, odnosno negativne doprinose, što može dati stabilnija objašnjenja koja su ljudima razumljivija. Kako bi važio zakon održanja relevatnosti postoji dodatan uslov $\alpha - \beta = 1$. Ovo pravilo je pogodno za početne slojeve.

Za slojeve kao što su slojevi agregacije i unutrašnje standardizacije postoje posebna pravila. Na primer, kod agregacije maksimuma, sva relevantnost se prosleđuje „pobedniku“ (eng. *winner-take-all*), neuronu koji je imao maksimalnu vrednost.

LRP metoda našla je primenu u mnogim oblastima mašinskog učenja. Neki od primera su klasifikacija slika (slika 3.3), klasifikacija teksta (slika 3.4), odgovaranje na vizuelna pitanja (eng. *visual question answering*) (slika 3.5) itd.



Slika 3.3: Primeri primene LRP metode - klasifikacija slika [8]

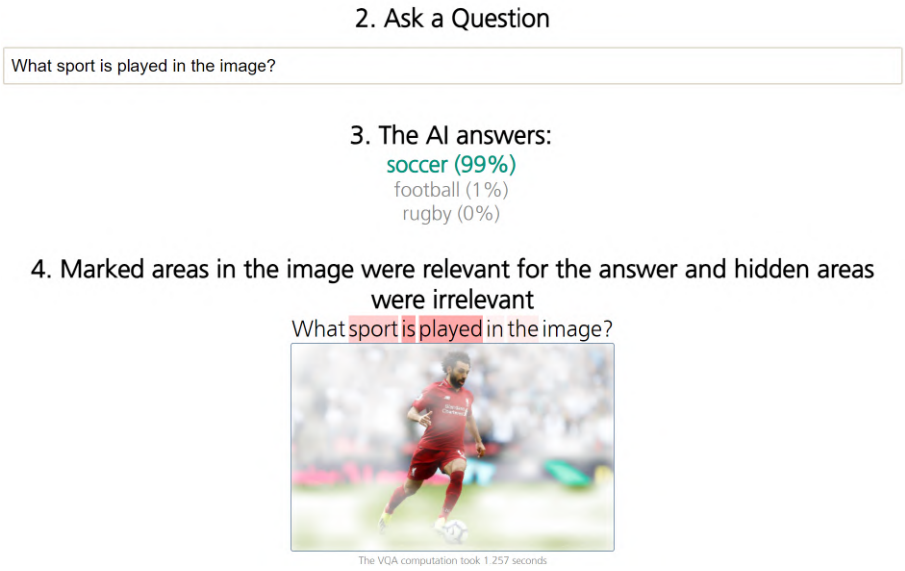
Science → Medicine

From: dsew@troi.cc.rochester.edu (David Sewell)
Subject: Theophylline/ephedrine and water bio-availability
Organization: University of Rochester - Rochester, New York
Lines: 19
Nntp-Posting-Host: troi.cc.rochester.edu

Does anyone know if either theophylline or ephedrine, or the two in combination, can reduce the body's ability to make use of available water? I had kind of an odd experience on a group hike recently, becoming dehyrated after about 9 hours of rigorous hiking despite having brought 1 1/2 gallons of water (c. 6 liters). I drank close to twice as much as anyone else, and no one else was dehydrated. I don't think general physical condition was an issue, since I was in at least the middle of the pack in terms of general stamina, so far as I could tell.

It may be that I just plain need more water than most people. But I am wondering if theophylline and/or ephedrine might be aggravating things. I took a couple of Primatene tablets during the hike to control asthma (24 mg. ephedrine, 100 mg. theophylline). I gather that both those drugs are diuretics. So now I'm wondering: does that mean they can reduce the body's ability to utilize available water? Would it be a particularly stupid thing to take that medication during hot-weather exercise? (I always assumed diuresis just meant you urinated a lot, but that wasn't the case yesterday.)

Slika 3.4: Primeri primene LRP metode - klasifikacija teksta [8]



Slika 3.5: Primeri primene LRP metode - odgovaranje na vizuelna pitanja [8]

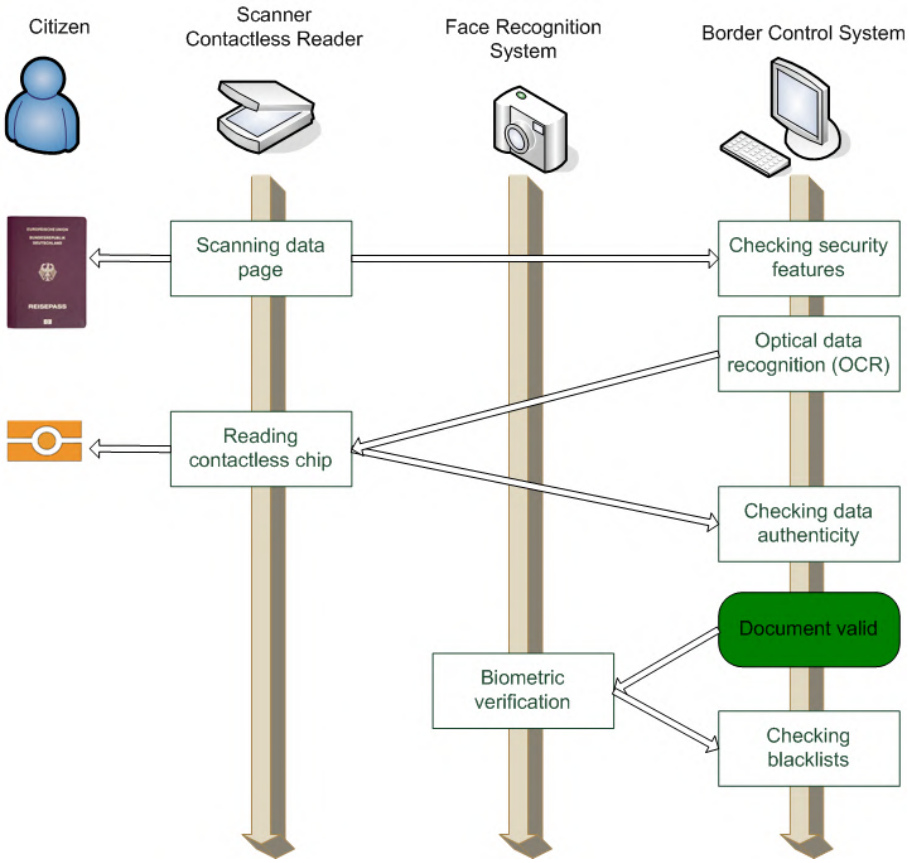
Glava 4

Tehnike preobražavanja lica

Elektronska putna dokumenta ili eMRTD (eng. *electronic machine readable travel documents*) koja sadrže biometrijske podatke nosioca postaju sve zastupljenija. Na osnovu biometrijskih podataka vrši se verifikacija identiteta, a samim tim moguće je automatizovati prolaz kontrolne tačke (eng. *automated border control*, *eGate*). Kao glavni biometrijski atribut za verifikaciju identiteta pomoću eMRTD izabrano je lice 2002. godine od strane ICAO [9] (eng. *International civil aviation organization*). Ova dokumenta poseduju čip koji sadrži sliku lica, a može sadržati i sekundarne atribute kao što su otisak prsta (eng. *fingerprint recognition*) ili dužica oka (eng. *iris recognition*) za dodatnu kontrolu.

Čovek pri dolasku na kontrolnu tačku prilaže elektronski dokument sa kojeg se čitaju biometrijski podaci, zatim se skenira lice i poredi sa podacima iz dokumenta pomoću sistema za prepoznavanje lica (eng. *face recognition system*). Ako je došlo do poklapanja i ako je dokument prošao ostale bezbednosne kontrole (potvrđena je validnost samog dokumenta itd), osoba uspešno prolazi kontrolnu tačku, dok je u slučaju odbijanja uglavnom potrebna intervencija prisutnog nadležnog službenika (slika 4.1).

Slika lica koja se nalazi u elektronskom dokumentu treba da ispunjava određene geometrijske i fotometrijske karakteristike propisane ISO standardom. Iako je uglavnom praksa da se slika lica koja se koristi u ovom dokumentu kreira pri samom izdavanju dokumenta u kontrolisanim uslovima, neke države dozvoljavaju da osoba kojoj se izdaje dokument priloži sliku, pri čemu je dodatno potrebno da se ispita da li ta slika zadovoljava propisane karakteristike (lice se jasno vidi, ujednačeno osvetljenje, osoba gleda pravo, osoba nije preterano našminkana itd) i da je naravno jasno da se na slici nalazi ta osoba, što vizuelno utvrđuje čovek.



Slika 4.1: eGate [37]

Ako je dozvoljeno da osoba pri vađenju dokumenta priloži svoju sliku, postoji mogućnost da su na toj slici izvršene modifikacije koje prolaze sve propise, uključujući i (manje ili više pažljivu) ljudsku inspekciju, a ipak posle mogu značajno uticati na sistem za prepoznavanje lica. Slučajne modifikacije (šum, distorzija) kao i neke namerne (filteri za ulepšavanje) mogu prouzrokovati nepoklapanje slike u dokumentu sa licem nosioca dokumenta od strane sistema za prepoznavanje lica. Druge modifikacije kao što je *preobražavanje lica* ako prođu nezapaženo mogu dovesti do poklapanja lica dve različite osobe sa slikom lica u dokumentu i predstavljaju ozbiljan bezbednosni propust u sistemu za prepoznavanje lica [9].

Preobražavanje slika

Preobražavanje slika, *stapanje slika* ili *morf slika* (eng. *image morphing* - od grčke reči za metamorfozu) je vizuelni efekat koji omogućava postepeni prelaz sa izvorne slike (eng. *source image*) na ciljnu sliku (eng. *target image*) nizom međuslika. Prva

slika u nizu je izvorna, svaka naredna slika sve manje liči na izvornu, a sve više na ciljnu, dok je poslednja slika ciljna. U slučaju dve slike lica, ova tehnika se zove još i *preobražavanje lica* (eng. *face morphing*) (slika 4.2).

Tehnika preobražavanja ima primenu u kinematografiji i dizajnu kao specijalni efekat. Prvi film koji je koristio ovu tehniku je *„Indiana Jones and The Last Crusade“* iz 1989. godine, a prvi muzički video je spot za pesmu *„Black and White“* Majkla Džeksona iz 1991. godine. U falsifikovanju slike lica za eMRTD koristi se slika u sredini niza jer ona u istoj meri liči i na izvornu i na ciljnu sliku. Ovu sliku zvaćemo *preobražena slika*.

Slika 4.2: Efekat preobražavanja lica

Najjednostavniji način da se preobražavanje postigne jeste *alfa mešanjem* (eng. *alpha blending*). Ako su S i T redom izvorna i ciljna slika predstavljene matrično sa tri kanala, onda je $\alpha * S + (1 - \alpha) * T$ slika nastala alfa mešanjem, za α između 0 i 1. U slučaju preobražavanja lica uzimamo $\alpha = 0.5$. Problem primene ove tehnike na licima je što se upečatljivi regioni ne poklapaju, zbog čega se kao rezultat pojavljuju dva nosa, četiri oka itd. Potrebno je pre mešanja izobličiti slike (*eng. warping*) tako da se odgovarajuće regije poklope. U ovom radu korišćena su dva algoritma za izobličavanje - *izobličavanje pomoću trouglova* (eng. *triangle warp*) i *izobličavanje zasnovano na atributima* koje ćemo zvati *Bajer-Nili algoritam* po autorima [5] (eng. *Beier-Neely warp*), te i odgovarajuće procedure preobražavanja zovemo *preobražavanje pomoću trouglova* i *Bajer-Nili preobražavanje*.

Preobražavanje pomoću trouglova

Koraci za dobijanje preobražene slike pomoću trouglova su (slika 4.3):

1. Poravnanje (eng. *aligning*) - obe slike se poravnaju tako da su oba oka na svakoj slici na istoj visini i da se centar lica nalazi u centru slike.

2. Ključne tačke lica (eng. *face landmarks*) - odrede se ključne tačke za oba lica. Od ta dva skupa dobije se treći koji predstavlja središta odgovarajućih parova ključnih tačaka dva lica.

3. Triangulacija (eng. *triangulation*) - izvrši se Delunijeva triangulacija (eng. *Delaunay triangulation*) za sva tri skupa ponaosob.

4. Izobličavanje - kako bi izvršili poklapanje ključnih regija svaki trougao prve slike preslika se u međusliku pomoću afine transformacije određene sa tri temena tog trougla i tri temena odgovarajućeg trougla dobijenog od tačaka središta. Isto se uradi i za drugu sliku.

5. Alfa-mešanje - dve međuslike dobijene u prethodnom koraku mešaju se sa koeficijentom $\alpha = 0.5$. U opštem slučaju, $\alpha \in [0, 1]$ što određuje i koliko je za par ključnih tačaka dve slike središnja blizu izvornoj, odnosno ciljnoj, te samim tim i koliko je preobraženo lice blizu izvornom, odnosno ciljnom licu.

Bajer-Nili preobražavanje

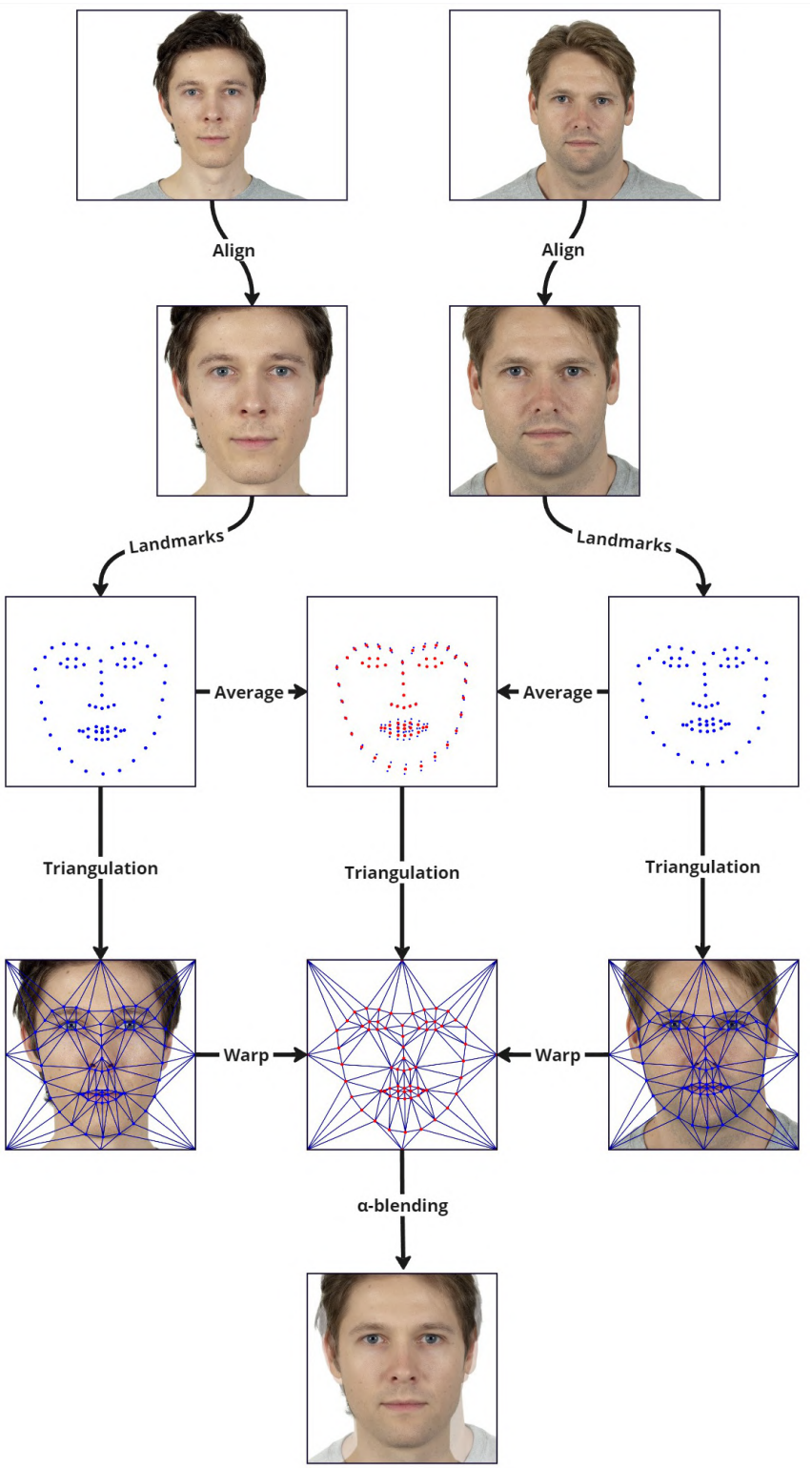
Za razliku od izobličavanja pomoću trouglova gde definišemo afina preslikavanja iz originalne u međusliku, u Bajer-Nili algoritmu traže se inverzna preslikavanja - ako je S originalna slika, a M međuslika u kojoj su regije izobličene i poklopljene, onda tražimo preslikavanje koje svakom pikselu iz M dodeljuje piksel iz S. Vrednost tog piksela se zatim kopira iz S u M. Umesto parova ključnih tačaka lica, u Bajer-Nili algoritmu koristimo parove duži (slika 4.6). Prvo se odrede odgovarajući parovi duži, zatim se za svaki par odredi „međuduž“ (duž dobijene od središta tačaka početka i središta tačaka kraja para duži). Neka je $\overrightarrow{P'Q'}$ usmerena duž na slici S, a $\overrightarrow{PQ}$ odgovarajuća međuduž na slici M. Neka je X tačka na slici M. Tada je njena inverzna slika X' određena na sledeći način:

$$u = \frac{(X - P) \cdot (Q - P)}{\|Q - P\|^2}$$

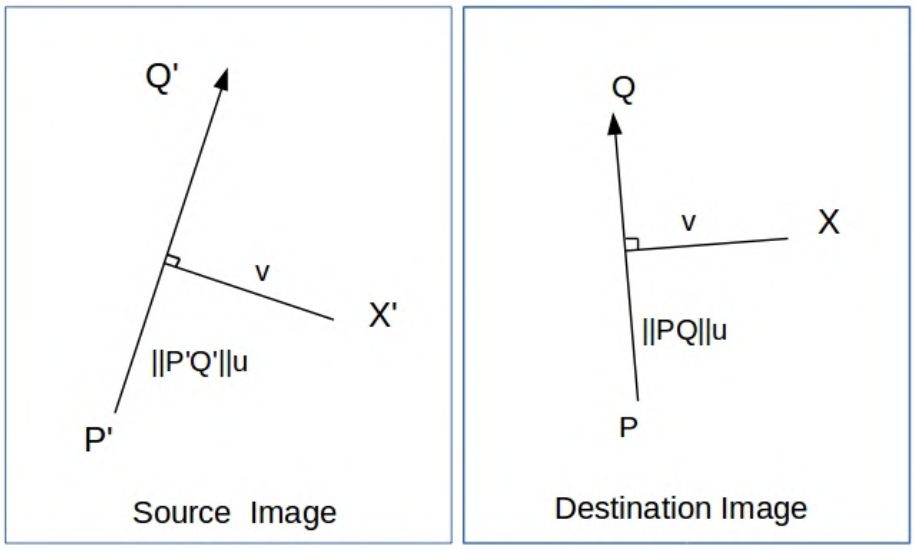
$$v = \frac{(X - P) \cdot Perpendicular(Q - P)}{\|Q - P\|}$$

$$X' = P' + u \cdot (Q' - P') + \frac{v \cdot Perpendicular(Q' - P')}{\|Q' - P'\|},$$

gde je *Perpendicular* normalan vektor na dati vektor, jednake dužine, svejedno kog smera, bitno je da se isti smer koristi u celom algoritmu (slika 4.4).



Slika 4.3: Preobražavanje lica pomoću trouglova



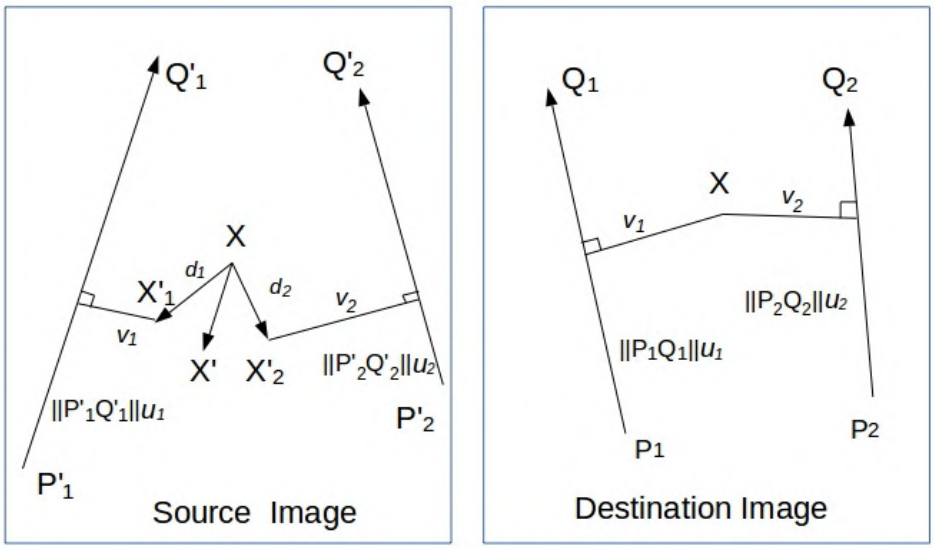
Slika 4.4: Inverzno preslikavanje u Bajer-Nili algoritmu za jednu duž [40]

U slučaju više duži, za tačku X odredi se odgovarajuća inverzna slika za svaku duž, zatim se kao konačna tačka uzme težinska suma tih tačaka (slika 4.5). Težina za svaku tačku je:

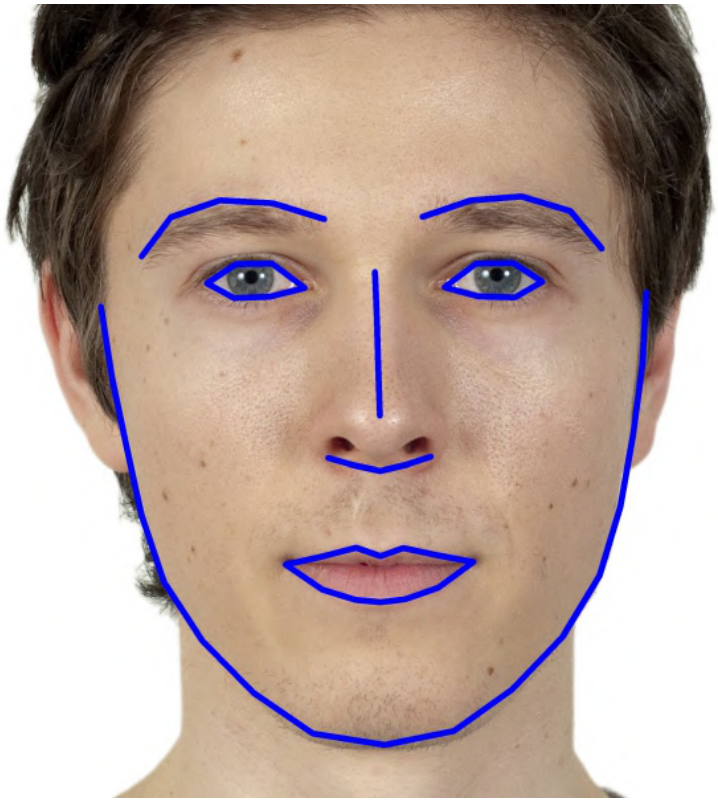
$$w = \left(\frac{length^p}{a + distance} \right)^b,$$

gde je *length* dužina duži, *distance* udaljenost inverzne slike tačke od duži, a a, b i p hiperparametri koji utiču na ponašanje algoritma. U ostatku rada za ove hiperparametre važiće $a = 0.01, b = 2, p = 0$.

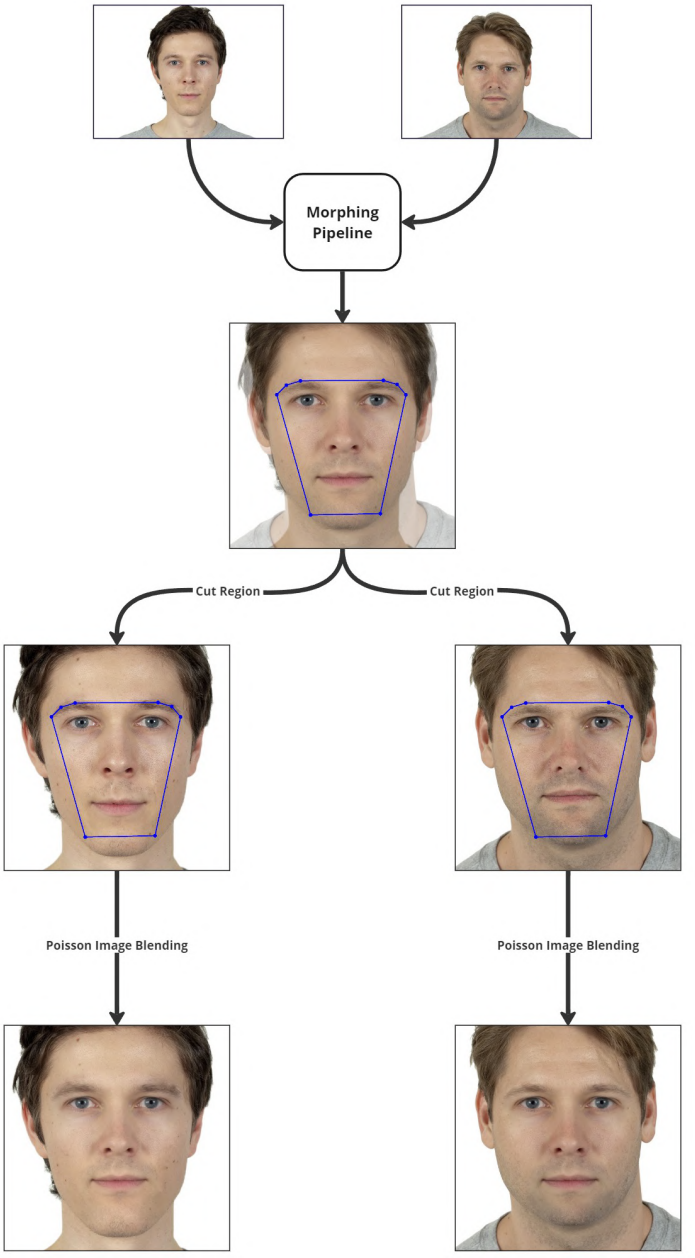
Uprkos tome što preobražena slika sadrži karakteristike oba lica i uspeva da prevari sistem za prepoznavanje lica, kao posledica procedure preobražavanja ostaju vidljivi tragovi mešanja kose, ušiju itd. (eng. *ghost artifacts*). Kako bi se ovo izbeglo, preobraženo lice se iseče, zatim se pomoću tehnike Poasonovog mešanja slika [21] (eng. *Poisson image editing*) umetne u jedno od dva originalna lica (slika 4.7), obično osobe koja prilaže sliku pri apliciranju za dokument.



Slika 4.5: Inverzno preslikavanje u Bajer-Nili algoritmu za više duži [40]



Slika 4.6: Duži u Bajer-Nili algoritmu određene na osnovu ključnih tačaka lica



Slika 4.7: Isecanje i umetanje preobraženog lica u originalno lice Poasonovim mešanjem slika

Glava 5

Priprema podataka

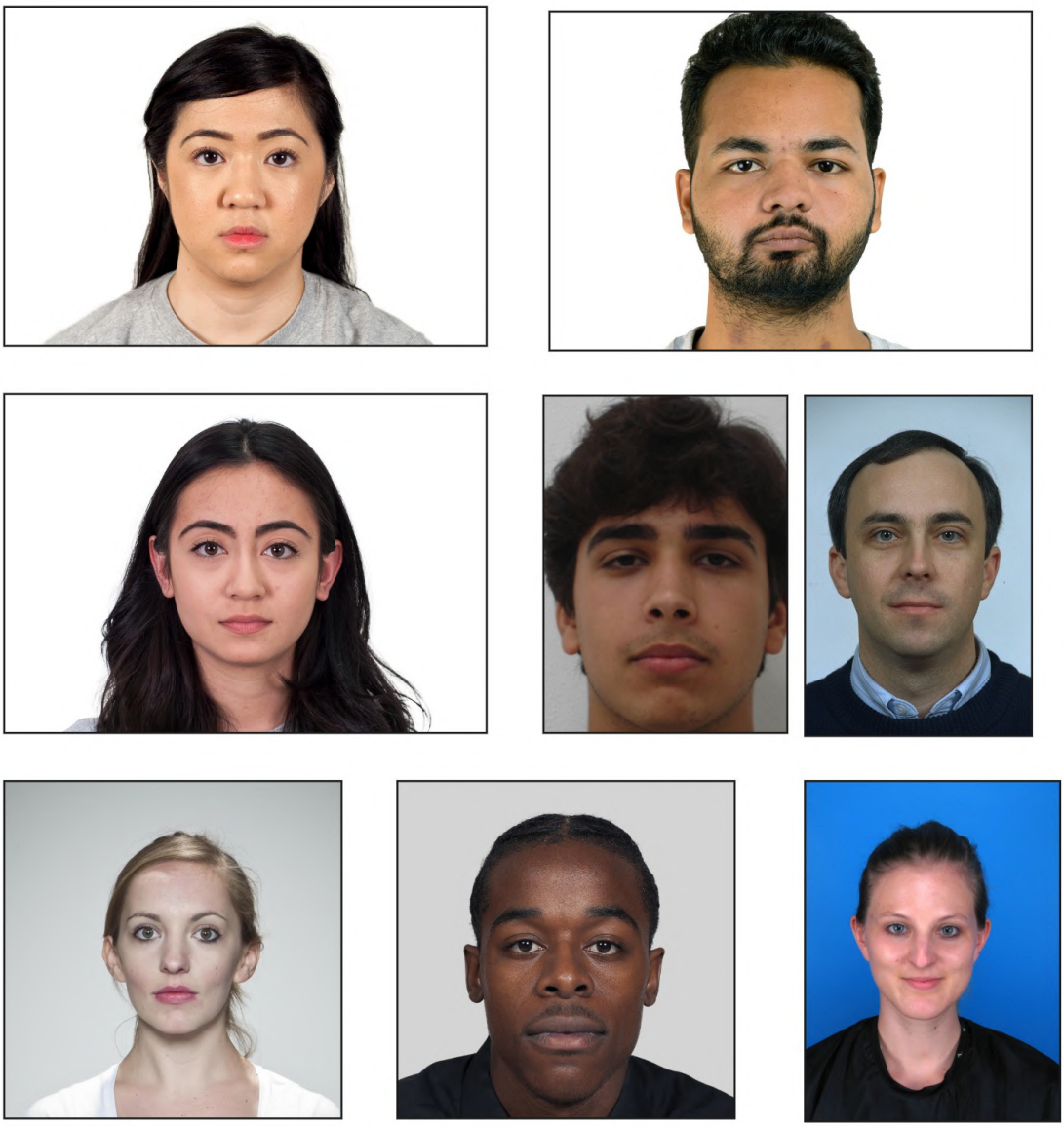
U ovom radu korišćeni su javno dostupni skupovi podataka ljudskih lica (slika 5.1):

- CFD [16] (sa dodacima CFD-MR [17] i CFD-INDIA [14]),
- FEI [32],
- FERET [22],
- London [6],
- OSF [35],
- Utrecht [11].

Iz svakog skupa odabrane su slike koje zadovoljavaju standarde slika za eMRTD. Sve osobe gledaju pravo, imaju neutralan izraz lica, ne nose naočare i lice im je ravnomerno osvetljeno. Postoji naravno nekolicina slika koje u manjoj meri odstupaju od ovih pravila. Broj slika i dimenzije slika po skupovima dati su u tabeli 5.1. Svaki od navedenih skupova podeljen je na particije za treniranje, validaciju i testiranje u razmeri 75% : 10% : 15%.

Kako bi se izbeglo prilagođavanje mreže određenim osobama i ceo proces približio realnom scenariju, za kreiranje preobraženih lica birani su parovi na sledeći način:

- Obe osobe su iz istog skupa.
- Preobražene slike za jednu particiju (treniranje, validacija ili testiranje) dobijene su isključivo od parova slika iz te particije.



Slika 5.1: Primeri slika iz skupova podataka redom: CFD, CFD-INDIA, CFD-MR, FEI, FERET, London, OSF, Utrecht

- Obe osobe su istog pola.
- Lica dve osobe su slična - prednost imaju parovi čiji vektori reprezentacije lica (eng. *face embeddings*) imaju manje kosinusno rastojanje.
- Svaka osoba se javlja približno jednak broj puta.
- Broj preobraženih slika za jedan skup približno je jednak broju slika tog skupa.

	CFD	FEI	FERET	London	OSF	Utrecht
Broj slika	823	168	360	102	74	67
Dimenzija slika	2444x1718	260x360	512x768	1350x1350	3120x3120	900x1200

Tabela 5.1: Broj originalnih slika po skupovima

Lažne (preobražene) slike generišu se procedurom preobražavanja lica pomoću trouglova koja je detaljno opisana u glavi 4. Bajer-Nili procedura koristi se samo pri ispitivanju otpornosti modela na alternativne tehnike preobražavanja opisane u glavi 6. Za izabrane dve slike vrši se poravnanje tako da se oči nalaze horizontalno, na istoj visini i da je lice uvećano i pomereno u centar slike. Određuje se 68 ključnih tačaka lica pomoću *Python* biblioteke *dlib* i dodaje se još 9 tačaka na ivice slike. Nakon toga se slike izobličavaju tako da se ključni atributi lica poklope i vrši se mešanje. Kako bi se izbegli očigledni tragovi koje ove procedure ostavljaju na kosi, ušima, vratu i slično, preobraženo lice se iseče i pomoću tehnike Poasonovog mešanja slika umetne u jedno (nasumično izabrano) od dva originalna lica.

Kvalitet preobraženih slika

Kvalitet preobraženih slika predstavlja njihovu sposobnost da prevare sistem za prepoznavanje lica. Kao mere kvaliteta uglavnom se koriste MAR (eng. *morph acceptance rate*) i rMAR (eng. *realistic morph acceptance rate*). MAR je relativan broj osoba koje sistem za prepoznavanje lica identifikuje sa njihovom preobraženom slikom. rMAR je relativan broj preobraženih slika koje se identifikuju sa oba originalna lica. rMAR je uvek manji ili jednak od MAR i predstavlja pogodniju meru za realan scenario u kojem preobraženo lice treba da se identifikuje sa obe osobe koje učestvuju u prevari. Kvalitet preobraženih lica ispitan je pomoću dva javno dostupna alata za kodiranje lica - *deepface* [28, 29] i *InsightFace* [7], koji nude istrenirane modele za dobijanje vektora reprezentacije lica. Dva lica se prepoznaju kao isto ako je kosinusno rastojanje između njihovih vektora reprezentacije manje od određenog praga. Prag je podešen na osnovu test particije javno dostupnog skupa lica LFW [13], tako da FAR bude manji od 0.1% (preporuka *FRONTEX*-a).

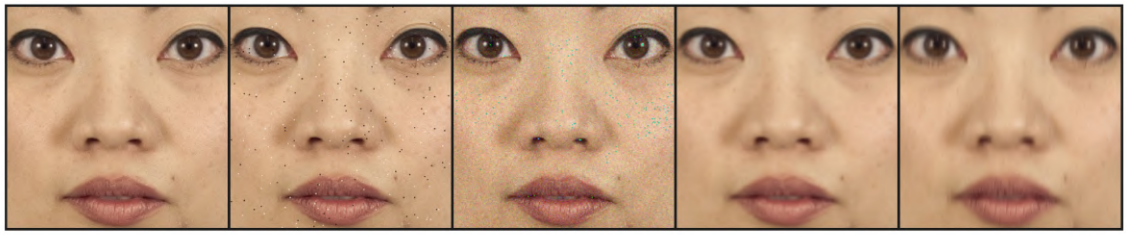
	MAR	rMAR
deepface	97%	94%
InsightFace	90%	83%

Tabela 5.2: MAR i rMAR

Rezultati su dati u tabeli 5.2. Kvalitet preobraženih slika je solidan. Čak i kad je u pitanju dosta napredan model, 83% preobraženih lica prolazi neopaženo.

Pre prosleđivanja modelu, slike se iseku tako da budu dimenzije 224×224 i da ostane samo unutrašnji deo lica i izvrši se standardno preprocesiranje za VGG19, odnosno Inception v3 model u slučaju suparničkih napada.

Preobražavanje lica ostavlja šum na koji ne želimo da se model fokusira, te se vrši augmentacija slika slično kao u [27] i [26]. Od svake slike dobiju se četiri dodatne, tako što se na svaku sliku primeni so i biber šum (eng. *salt'n'pepper noise*) na 0.5% - 1% piksela, Gausov šum (eng. *Gaussian noise*) standardne devijacije 12, Gausovo zamućenje (eng. *Gaussian blur*) veličine kernela 5×5 i standardne devijacije 1 i zamućenje u pokretu (eng. *motion blur*) (slika 5.2). Ukupan broj originalnih slika je 1152, ukupan broj lažnih (preobraženih) slika je 1520. Sa augmentacijama, ukupan broj originalnih je 5760, a lažnih 7600.

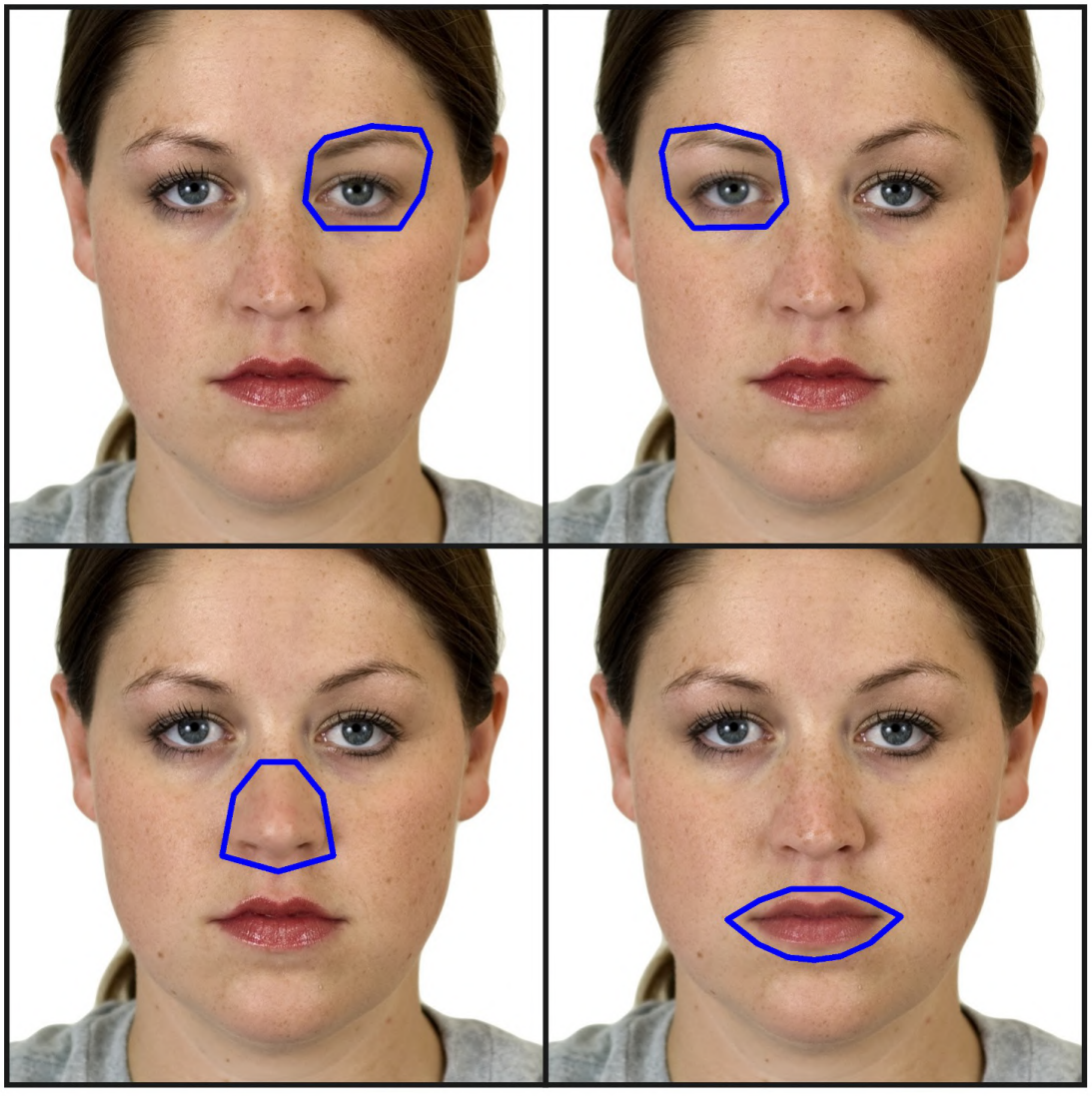


Slika 5.2: Primer augmentacija. Sa leva na desno: originalna slika, so i biber šum, Gausov šum, Gausovo zamućenje, zamućenje u pokretu.

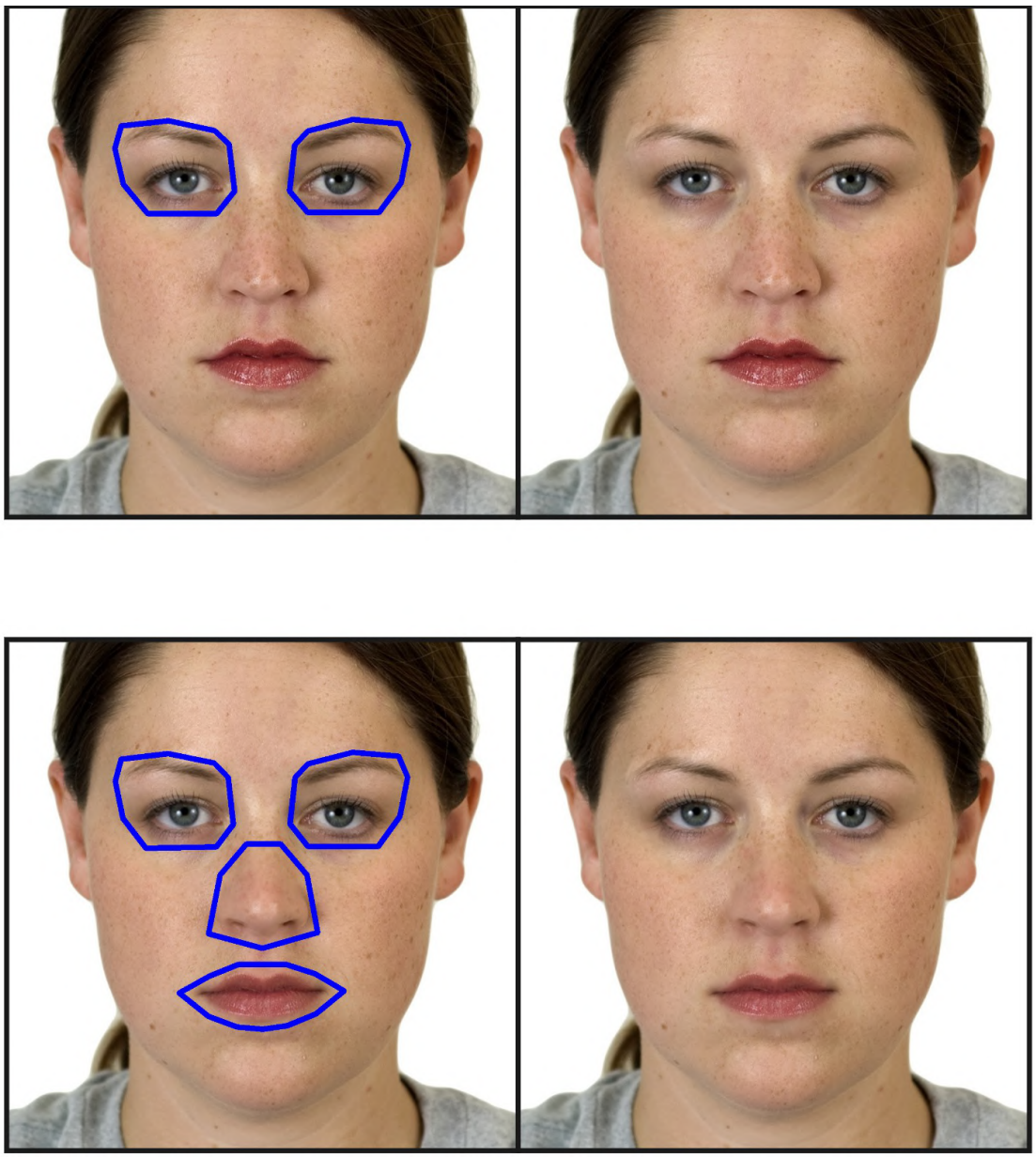
Preobražavanje određenih regija lica

Analiziranjem modela pomoću LRP metode u glavi 6 zaključeno je da se model najviše fokusira na oči pri prepoznavanju preobraženih lica. Kako bi se ovo prevazišlo isprobana su tri dodatna načina treninga kod kojih su preobražene slike modifikovane tako da samo određeni delove sadrže tragove preobražavanja. Iz preobraženog lica iseče se jedna ili više regija (levo, desno oko, nos ili usta) i umetne u originalno lice pomoću tehnike Poasonovog mešanja slika (slike 5.3, 5.4 i 5.5). Pretpostavka je da će ograničavanjem lažnih atributa model biti primoran da se fokusira na date relevantne regije [27]. Ove slike zvaćemo *preobražene slike regija* ili samo *preobražene regije*. Parovi slika za generisanje preobraženih regija isti su kao kod običnih preobraženih lica.

Slika 5.3: Generisanje preobraženih slika regija



Slika 5.4: Slike kod kojih je umetnuta jedna preobražena regija - primer za svaku regiju. Regije koje su umetnute zaokružene su plavom bojom.



Slika 5.5: Slike kod kojih je umetnuto više preobraženih regija - primer za dve kombinacije regija. Slike sa desne strane sadrže iste umetnute regije kao slike sa leve, samo što nisu zaokružene plavom bojom.

Glava 6

Eksperimenti

U nastavku rada opisani su i upoređeni rezultati eksperimenata izvršenih na nekoliko različitih modela za detektovanje preobraženih lica. Ispitane su različite mere kvaliteta - TPR, TNR i EER, karakteristične za modele koji se koriste za prepoznavanje lica. Prvo je izvršen klasičan trening, zatim je pomoću LRP metode uočeno da se model treniran na ovakav način fokusira najviše na oči, što ga čini slabim za primene u bezbednosti. Iz tog razloga testirane su alternativne metode treninga - modifikacijama slika za trening želimo da nateramo model da se fokusira i na ostale delove lica i time ga učinimo robusnijim. Ispitane su dodatne mere: moć prepoznavanja novih tehnika preobražavanja, otpornost na semantičke napade i otpornost na suparničke napade. Pomoću LRP metode izvršena je dalja analiza modela. Svi modeli (osim SVM) biće vrlo slične arhitekture, dok će se najviše razlikovati način treniranja, odnosno podaci na kojima su trenirani.

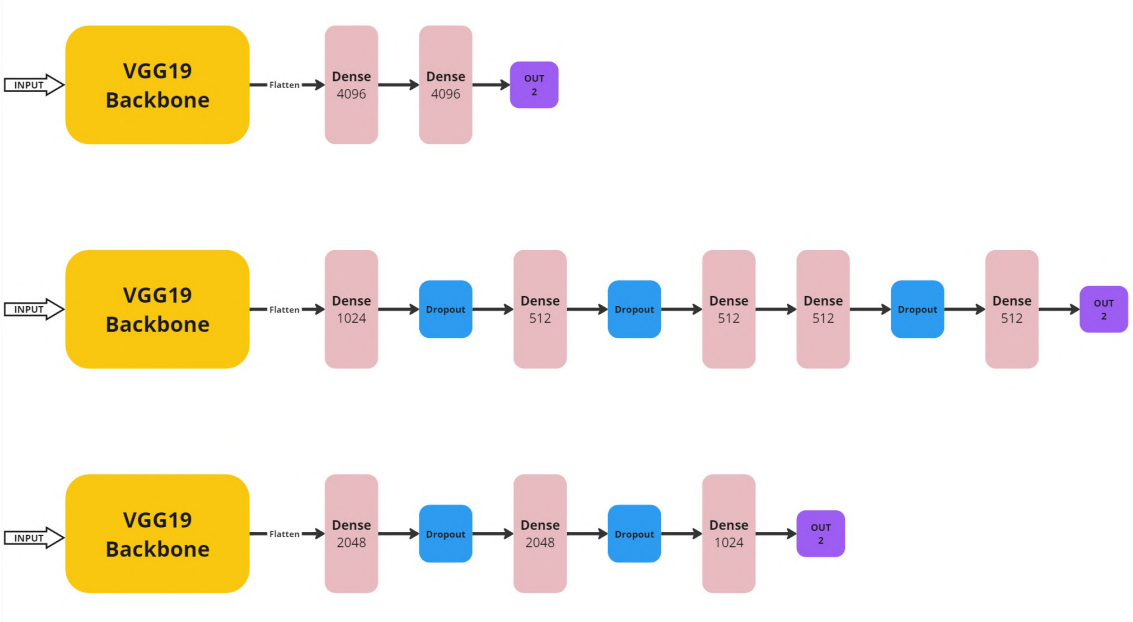
Postavka

Treniranje modela obavljeno je na računaru sa NVIDIA GeForce RTX 4070 grafičkom karticom, 32GB DDR5 RAM memorije i Intel® Core™ i7-13700KF procesorom. Za kreiranje i treniranje modela korišćen je radni okvir *TensorFlow* i biblioteka *Keras* u programskom jeziku *Python*. Za implementaciju LRP metode korišćena je biblioteka *iNNvestigate* [2]. Kod je pisan u *jupyter* sveskama.

Arhitekture svih neuronskih mreža u ovom radu sastoje se od konvolutivnog dela VGG19 mreže trenirane na ImageNet skupu podataka na koji se nadovezuju potpuno-povezani slojevi. Kao aktivaciona funkcija svih slojeva osim izlaznog koristi se ReLU. Izlazni sloj mreže sadrži dva neurona koji predstavljaju klase za original-

ne i preobražene slike, na koje je primenjen softmax. Za funkciju greške uzeta je kategorička unakrsna entropija.

Različita podešavanja modela isprobana su na validacionom skupu u cilju postizanja što bolje tačnosti. Eksperimentisano je sa različitim brojem potpuno-povezanih slojeva, različitim brojem neurona, izostavljanjem, različitim brojem epoha, finim podešavanjem modela, augmentacijama itd. Neke isprobane arhitekture date su na slici 6.1.



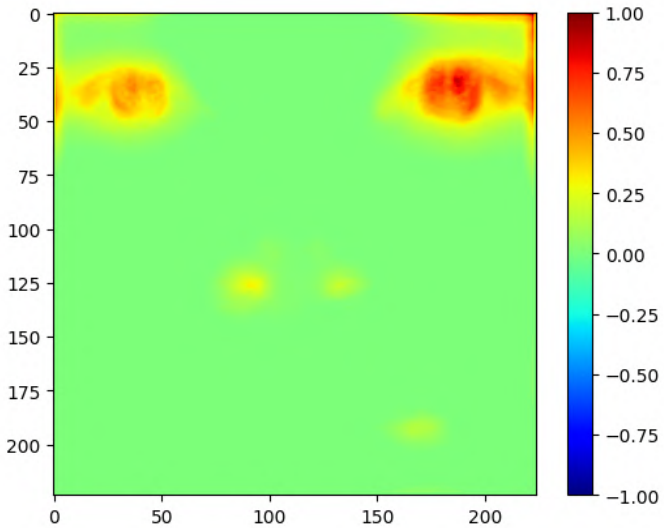
Slika 6.1: Razne arhitekture koje su ostvarile dobre rezultate na validacionom skupu. Sve mreže sastoje se od konvolutivnog dela VGG19 mreže već trenirane na ImageNet skupu, na koji se nadovezuju potpuno povezani slojevi. U cilju regularizacije nekad se između potpuno-povezanih slojeva koristi izostavljanje.

Osim modela zasnovanih na konvolutivnim neuronskim mrežama, eksperimentisano je i sa jednim SVM (eng. *support vector machine*) modelom, koji vrši klasifikaciju na osnovu broja detektovanih SIFT, ORB, FAST, BRISK, AGAST, SobelX, SobelY i CannyEdge atributa na slici. Treniranje ovog modela urađeno je pomoću biblioteke *sklearn*. Za funkciju greške uzeta je greška u vidu šarke (*eng. hinge loss*). Korišćena je $l2$ regularizacija. Hiperparametar C određen je unakrsnom validacijom. Ovaj model će služiti kao donja granica performansi modela zasnovanih na konvolutivnim mrežama.

Termin toplotna mapa će se nadalje odnositi na prosečnu korigovanu toplotnu mapu, koja se dobije tako što se pomoću LRP metode generišu toplotne mape za sve

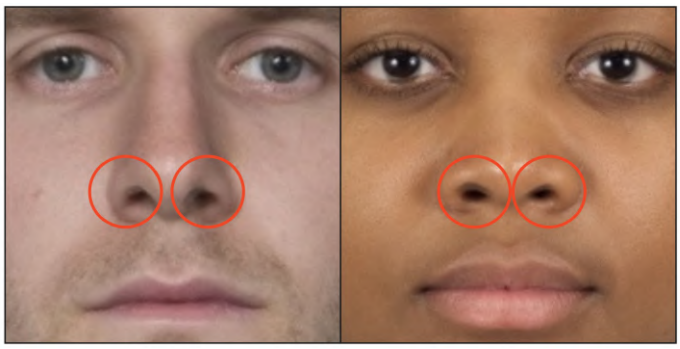
preobražene slike test skupa koje model klasifikuje ispravno (za prag 0.5), izračuna se prosek, od proseka se oduzme srednja vrednost i svi negativni pikseli se postave na vrednost 0. Cilj poslednja dva koraka je da se eliminišu slabe relevantnosti kako bi se dobile preglednije toplotne mape. U nekim slučajevima poslednja dva koraka preskočena su kako bi se utvrdilo da li negativne relevantnosti nagoveštavaju to da model poredi regije pri donošenju odluke. Sve toplotne mape su normalizovane deljenjem maksimalnom vrednošću toplotne mape.

Sa izvršenim augmentacijama, broj originalnih slika je 5760, dok je broj preobraženih slika 7600. Najbolji model treniran na ovim podacima ostvaruje dosta dobre rezultate na test skupu ($EER = 3\%, TPR = 95\%, TNR = 98\%$), međutim, analizom toplotne mape koje vraća LRP metoda utvrđeno je da model ne koristi sve delova lica pri donošenju odluke. U najvećoj meri na odluku modela utiču oči, dok su nos i usta gotovo zanemareni (slika 6.2), iako se na primer može vizuelno utvrditi da preobražavanje često ostavlja jasne tragove na nozdrvama (slika 6.3). Ovo nije poželjno za modele koji se koriste u bezbednosne svrhe, jer vrlo lako podležu raznim napadima i novim tehnikama modifikacija. Kako bi model bio primoran da se fokusira i na nos i usta, eksperimentisano je sa preobraženim slikama regija koje se opisane u glavi 5.



Slika 6.2: Toplotna mapa klasično treniranog modela koji je ostvario najbolji rezultat na validacionom skupu

U zavisnosti od tipa podataka korišćenih za trening, razlikovaćemo četiri vrste treninga, odnosno četiri vrste modela [27]:



Slika 6.3: Tragovi koje preobražavanje često ostavlja na nozdrvama

1. *Standardni* - modeli trenirani na trening skupu koji sadrži 50% originalnih slika lica i 50% preobraženih slika lica.

2. *Modeli jedne regije* - modeli trenirani na 50% originalnih lica, 10% preobraženih lica i 40% preobraženih lica jedne regije, po 10% za levo oko, desno oko, nos i usta.

3. *Modeli nekoliko regija* - modeli trenirani na 50% originalnih slika, 10% preobraženih slika i 40% preobraženih slika jedne ili više regija, po 10% za jednu, dve, tri i četiri regije. Svaka regija javlja se približno jednak broj puta.

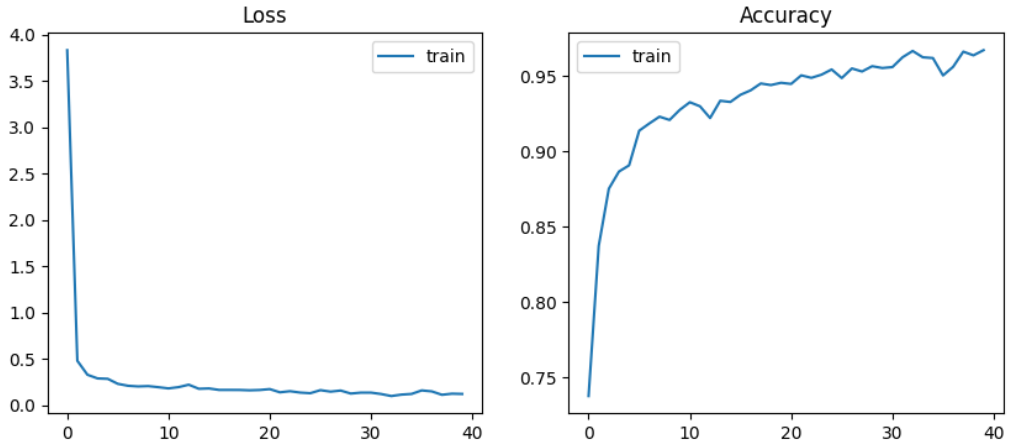
4. *Modeli trenirani u dve faze* - modeli koji imaju inicijalnu, prvu fazu treninga za klasifikaciju preobraženih regija. U prvoj fazi trenira se model tako da preobražene slike jedne regije, po 25% slika za svaku regiju, klasifikuje u jednu od četiri klase, gde svaka klasa odgovara jednoj regiji. Cilj prve faze je da model nauči da koristi sve informacije date u četiri regije i prepozna odgovarajuće modifikacije. U drugoj fazi poslednji sloj sa četiri neurona zamenjuje se slojem sa dva neurona, i zatim se vrši trening kao kod standardnih modela ili modela nekoliko regija. U drugoj fazi dodatno je eksperimentisano sa zamrzavanjem svih slojeva osim nekoliko poslednjih.

Rezultati

Nakon eksperimentisanja sa raznim podešavanjima, izabrano je šest različitih modela (tabela 6.1). Greška i tačnost modela na trening skupu kroz epohe prikazani su na slikama 6.4, 6.5, 6.6, 6.7, 6.8 i 6.9.

	Vrsta treninga	Augmentacije	Fino podešavanje	Veličina podskupa	Broj epoha
$model_1$	Standardni	Da	Ne	32	40
$model_2$	Standardni	Da	Da	32	20
$model_3$	Jedna Regija	Ne	Ne	32	30
$model_4$	Nekoliko regija	Da	Ne	32	40
$model_5$	Dve faze (druga faza - nekoliko regija)	Da	Ne	32 prva faza 64 druga faza	10 prva faza 30 druga faza
$model_6$	Dve faze (druga faza - standardni)	Da	Da	32 prva faza 64 druga faza	10 prva faza 10 druga faza

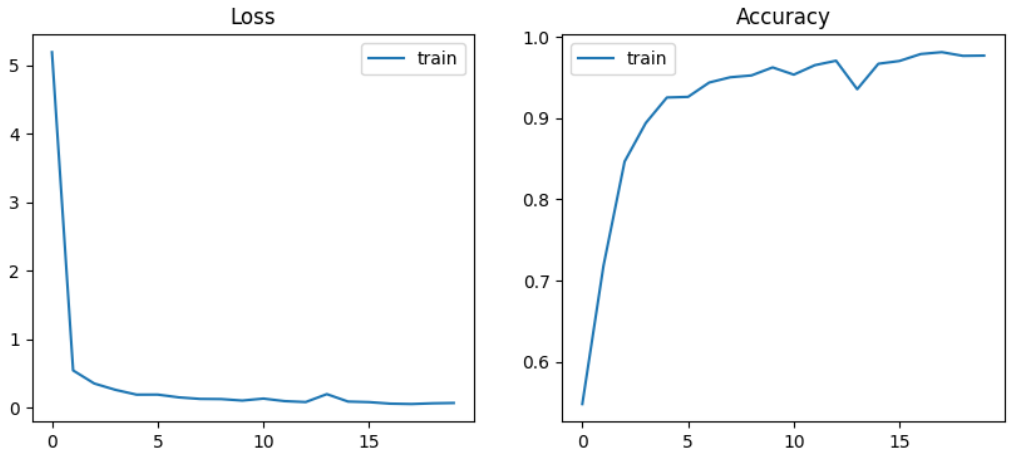
Tabela 6.1: Detalji treninga za izabranih šest modela



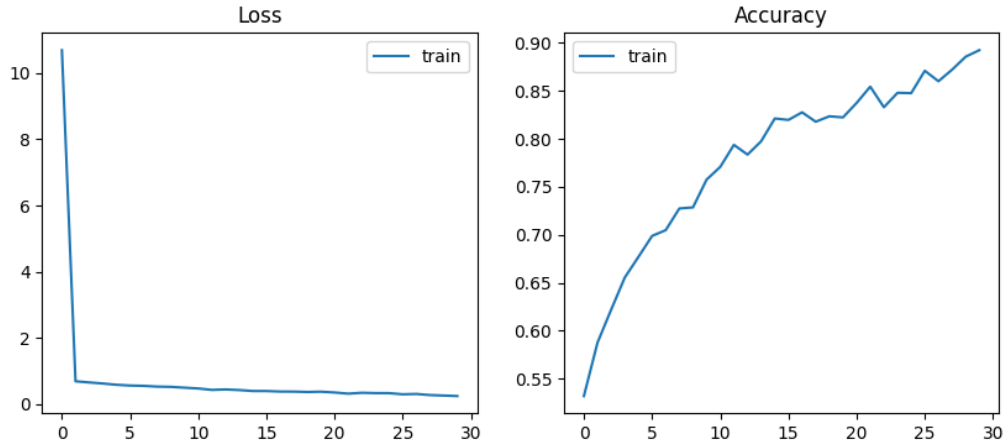
Slika 6.4: Greška (levo) i tačnost (desno) kroz epohe na trening skupu za $model_1$

TPR, TNR i EER

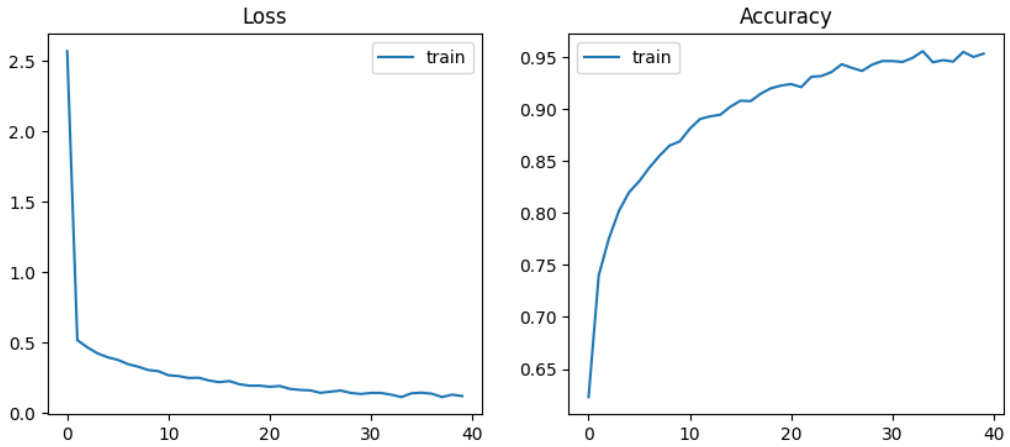
Rezultati za TPR, TNR i EER dati su u tabeli 6.2. Modeli sa najboljim EER su očekivano oni koji su trenirani standardno, jer su ti podaci najbliži podacima test skupa. Najbolji model je zapravo dobijen treningom u dve faze, gde je u drugoj fazi treniran standardno. Svi modeli osim $model_6$ bolji su u klasifikaciji preobraženih lica, što je i poželjno, s obzirom da greška u obrnutom slučaju znači propuštanje pogrešnih osoba. $model_3$ klasifikuje sve preobražene slike ispravno, a samo 57% originalih slika. SVM model ostvaruje ubedljivo najgori EER, ali dosta dobro prepoznaje preobražena lica, čak 92.4% njih.



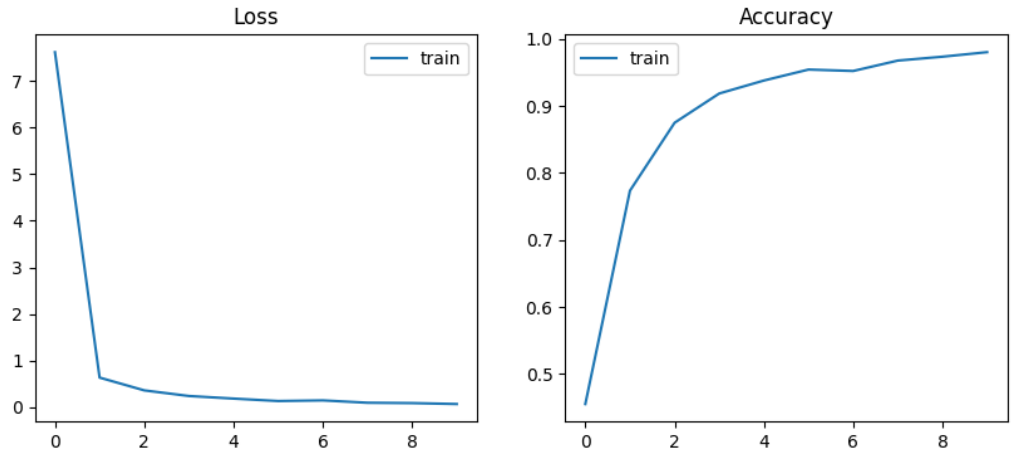
Slika 6.5: Greška (levo) i tačnost (desno) kroz epohe na trening skupu za $model_2$



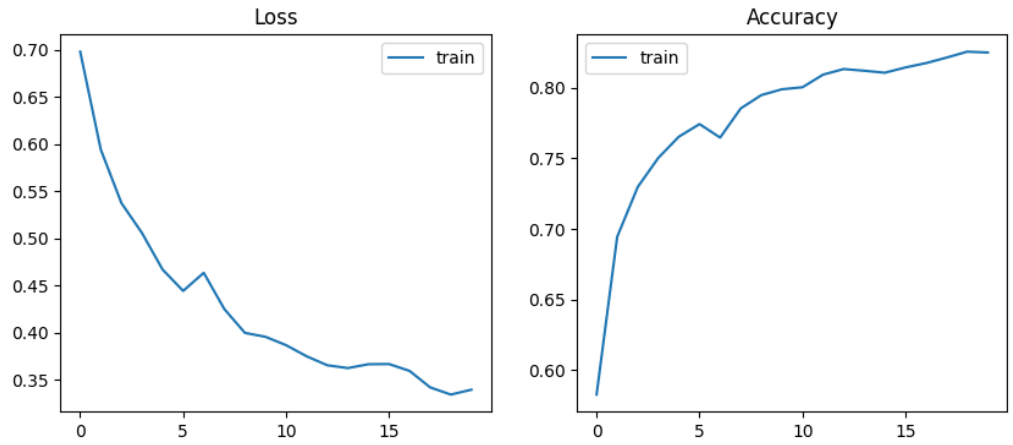
Slika 6.6: Greška (levo) i tačnost (desno) kroz epohe na trening skupu za $model_3$



Slika 6.7: Greška (levo) i tačnost (desno) kroz epohe na trening skupu za $model_4$



Slika 6.8: Greška (levo) i tačnost (desno) kroz epohe na trening skupu u prvoj fazi treninga za $model_5$ i $model_6$



Slika 6.9: Greška (levo) i tačnost (desno) kroz epohe na trening skupu u drugoj fazi treninga za $model_5$

Alternativne tehnike preobražavanja

Preobražavanje lica nije striktno definisana procedura, napadač ima slobodu da eksperimentiše sa kreiranjem lažnih slika. Zbog toga je ispitana otpornost modela na tri skupa drugačije generisanih preobraženih slika. Prvi skup su slike iz test skupa generisane Bajer-Nili procedurom opisanom u glavi 4. Drugi skup je takođe dobijen od slika iz test skupa, pomoću biblioteke *facemorpher* [1]. Treći skup je javno dostupan AMSL [19], koji sadrži 2175 preobraženih slika iz London skupa podataka. Iako drugi i treći skup koriste algoritam izobličavanja pomoću trouglova, razlikuju se neki koraci u odnosu na proceduru preobražavanja pomoću trouglova opisanu u glavi 4. Primeri različito generisanih lica dati su na slici 6.10.

	TPR	TNR	EER	EER prag
$model_1$	94%	96%	5.0%	0.45
$model_2$	95%	98%	3.0%	0.66
$model_3$	57%	100%	10%	0.99
$model_4$	85%	93%	10%	0.9
$model_5$	70%	97%	11%	0.675
$model_6$	99%	97%	1.6%	0.185
SVM	70%	92.4%	24%	0.625

Tabela 6.2: TPR, TNR i EER za različite modele

	Bajer-Nili	facemorpher	AMSL
$model_1$	97%	98%	93%
$model_2$	99%	97%	86%
$model_3$	99%	100%	98%
$model_4$	95%	97%	83%
$model_5$	95%	98%	88%
$model_6$	95%	97%	86%
SVM	49%	100%	100%

Tabela 6.3: Broj detektovanih preobraženih slika generisanih alternativnim tehnikama i alatima.

Svi modeli osim SVM ostvaruju odlične rezultate na preobraženim slikama generisanim Bajer-Nili procedurom i *facemorpher* alatom. Za preobražene slike AMSL skupa rezultati su nešto lošiji, što je verovatno posledica toga da su sve slike iz skupa London, dok u našem trening skupu London čini oko 10% svih podataka, a najviše CFD, oko 50% svih podataka. Pretpostavka je da sve tehnike koje koriste izobličavanje pomoću trouglova ostavljaju slične atribute koje detektuje SVM, te se iz tog razloga on odlično pokazao za sve preobražene slike generisane na ovaj način.

Slika 6.10: Preobražena lica dobijena različitim tehnikama - Preobražavanje pomoću trouglova (levo gore), AMSL skup podataka (desno gore), Bajer-Nili preobražavanje (levo dole), alat *facemorpher* (desno dole).

Semantički napadi

Semantički napad predstavlja scenario u kojem napadač želi da modifikuje samo najrelevantnije delove ulaza kako bi prevario model, ostavljajući što manje tragova. Kako bi rekreirali ovakav scenario, koristićemo preobražena lica jedne ili nekoliko regija. U ovom slučaju modeli moraju da na osnovu manje količine informacija klasifikuju slike u lažne i originalne.

Najviše preobraženih slika jedne regije očekivano je detektovao $model_3$ koji je

	Levo oko	Desno oko	Usta	Nos	Prosek
$model_1$	22%	19%	15%	19%	19%
$model_2$	41%	38%	13%	7%	25%
$model_3$	78%	81%	77%	86%	80%
$model_4$	45%	46%	48%	50%	47%
$model_5$	57%	58%	48%	64%	57%
$model_6$	7%	8%	6%	5%	6%
SVM	36%	33%	31%	33%	33%

Tabela 6.4: Procenat detektovanih preobraženih lica jedne regije. U tabeli su dati rezultati različitih modela za četiri glavne regije.

treniran na takvim slikama. Dobre rezultate ostvarili su i modeli trenirani na nekoliko regija, dok su modeli trenirani standardno ostvarili najgore rezultate, čak gore i od SVM, koji je ostvario solidne rezultate. $model_6$ koji je ostvario najbolje rezultate na test skupu (EER=1.6%), ovde se jako loše pokazao, bez obzira što je u prvoj fazi treninga treniran na slikama jedne regije. $model_2$ ostvario je nešto bolje rezultate od $model_1$, međutim, $model_2$ dosta veći fokus stavlja na oči.

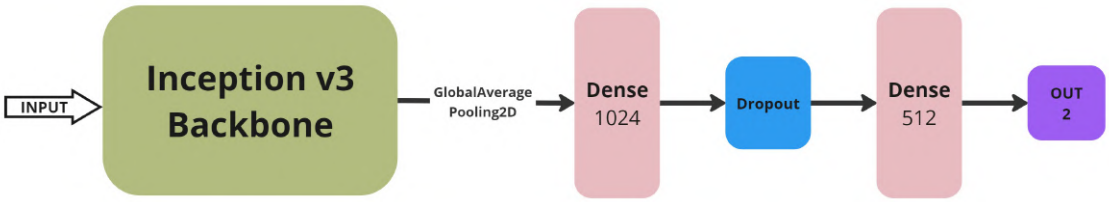
	Levo oko Desno oko	Levo oko Desno oko Usta Nos	Usta Nos	Levo oko Usta Nos	Prosek
$model_1$	50%	65%	8%	40%	41%
$model_2$	62%	88%	0%	60%	53%
$model_3$	75%	100%	92%	100%	92%
$model_4$	75%	100%	75%	90%	85%
$model_5$	75%	100%	83%	100%	89%
$model_6$	50%	80%	0%	35%	41%
SVM	25%	61%	50%	40%	44%

Tabela 6.5: Procenat detektovanih preobraženih lica više regija. U tabeli su dati rezultati različitih modela za neke kombinacije preobraženih regija.

Kod nekoliko preobraženih regija rezultati su slični - najbolji je model treniran na jednoj regiji. Približno dobri su modeli trenirani na više regija, dok modeli trenirani standardno ostvaruju nešto bolji rezultat nego na slikama jedne regije. Modeli trenirani u dve faze ne odstupaju značajnije od modela treniranih u jednoj fazi.

Suparnički napadi

Otpornost modela na suparnički napad podrazumeva sposobnost modela da ne bude prevaren suparničkim primerima kako intenzitet šuma na njima raste. Suparnički primeri dobijeni su od preobraženih slika test skupa primenom metode FGSM na suparnički model. Za konvolutivni deo mreže suparničkog modela izabrana je Inception v3 mreža trenirana na ImageNet skupu, kao model koji se često koristi u zadacima klasifikacije slika. Potpuno-povezani slojevi u originalnoj arhitekturi zamenjeni su sa dva potpuno povezana sloja od po 1024 i 512 neurona, između kojih je dodato izostavljanje. Izlazni sloj mreže sastoji se od dva neurona na koje je primenjen softmax (slika 6.11). Model je treniran u 20 epoha na skupu originalnih i preobraženih slika koji je u standardnom treningu služio za validaciju. Ovaj skup je izabran jer bi se treniranjem modela koji se napada i suparničkog modela na istom skupu dobili prejaki suparnički primeri. Suparnički primer za datu sliku dobija se dodavanjem šuma generisanog primenom metode FGSM na suparnički model, pomnoženog intenzitetom ε.



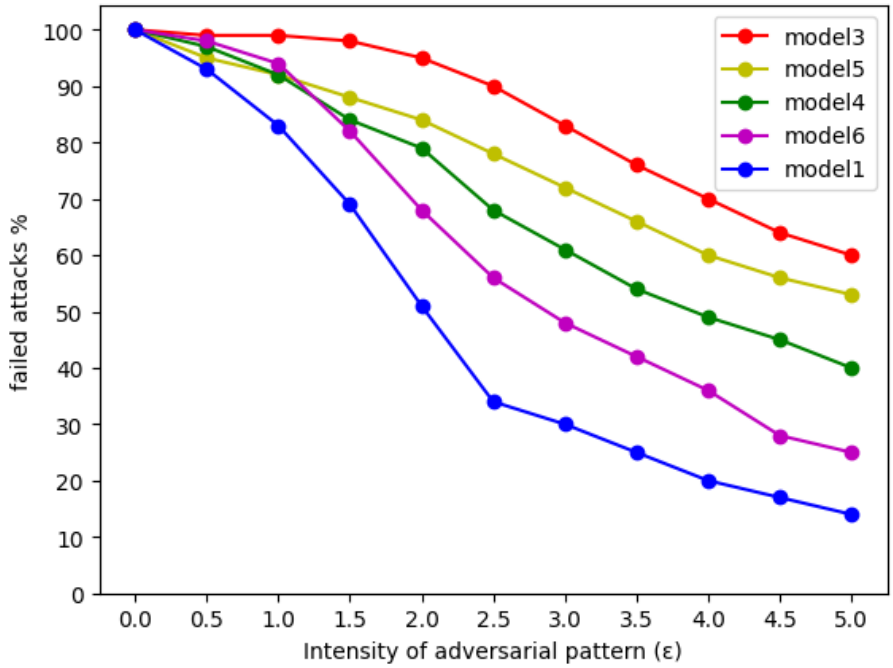
Slika 6.11: Arhitektura suparničkog modela.

Za ispitivanje otpornosti modela na suparničke napade korišćene su preobražene slike iz test skupa koje model klasifikuje ispravno kada se prag postavi na 0.5. Suparnički napad intenziteta ε za datu sliku je uspešan ako je model klasifikuje kao original, u suprotnom je neuspešan. Rezultati ispitivanja otpornosti na suparnički napad dati su na slikama 6.13 i 6.14. Najotporniji model na suparničke napade intenziteta do 5 je $model_3$, čak i kad je intenzitet 5, što se već uočava golim okom, na ovom modelu prolazi samo 40% napada. Nešto lošije rezultate imaju modeli više regija ($model_5$ i $model_4$), a najgori su modeli trenirani standardno. Za veće intenzitete šuma, do 10, $model_5$ prevazilazi $model_3$ i uspešno klasifikuje 22% instanci za dosta jak intenzitet, za šta su verovatno zaslužni prva faza treninga i fino podešavanje. Standardni modeli su najmanje otporni na ovu vrstu napada. Ne prepoznaju skoro nijednu sliku kao lažnu za vrednosti intenziteta veće od 8.

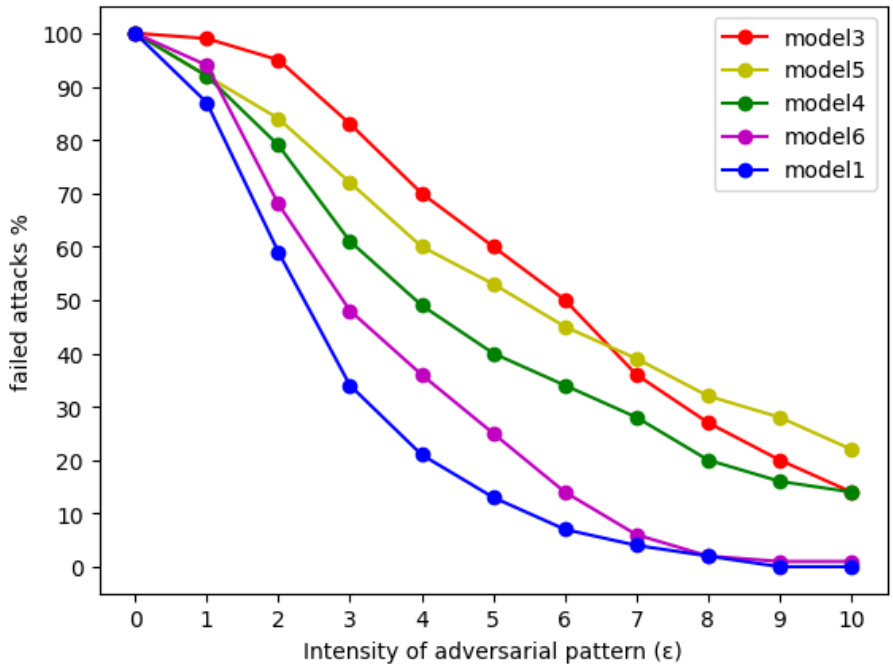
Slika 6.12: Suparnički primeri za intenzitete $\varepsilon = 0, 2, 4, 6, 8, 10$

LRP analiza

Model sa najboljom EER ocenom treniran je u dve faze, u drugoj fazi na standardnim podacima. Za njim slede dva standardna modela. Iako dosta dobro prepoznaju lažna lica, modeli trenirani na preobraženim slikama jedne ili nekoliko regija ostvaruju značajno niži EER. U slučaju semantičkog napada, podaci trenirani alternativnim metodama ponašaju se očekivano mnogo bolje, jer su trenirani na takvoj vrsti podataka. Ovi modeli pokazuju se mnogo bolje i u slučaju suparničkog napada. Iako je intuitivno jasno da su modeli trenirani na preobraženim slikama regija primorani da u obzir uzmu sve regije lica, zbog osetljivosti domena primene ovih modela korisno je to na neki način potvrditi. LRP metoda implementirana pomoću



Slika 6.13: Procenat neuspelih suparničkih napada intenziteta šuma $\varepsilon = [0.5, 1, 1.5, ..., 5]$ za različite modele.

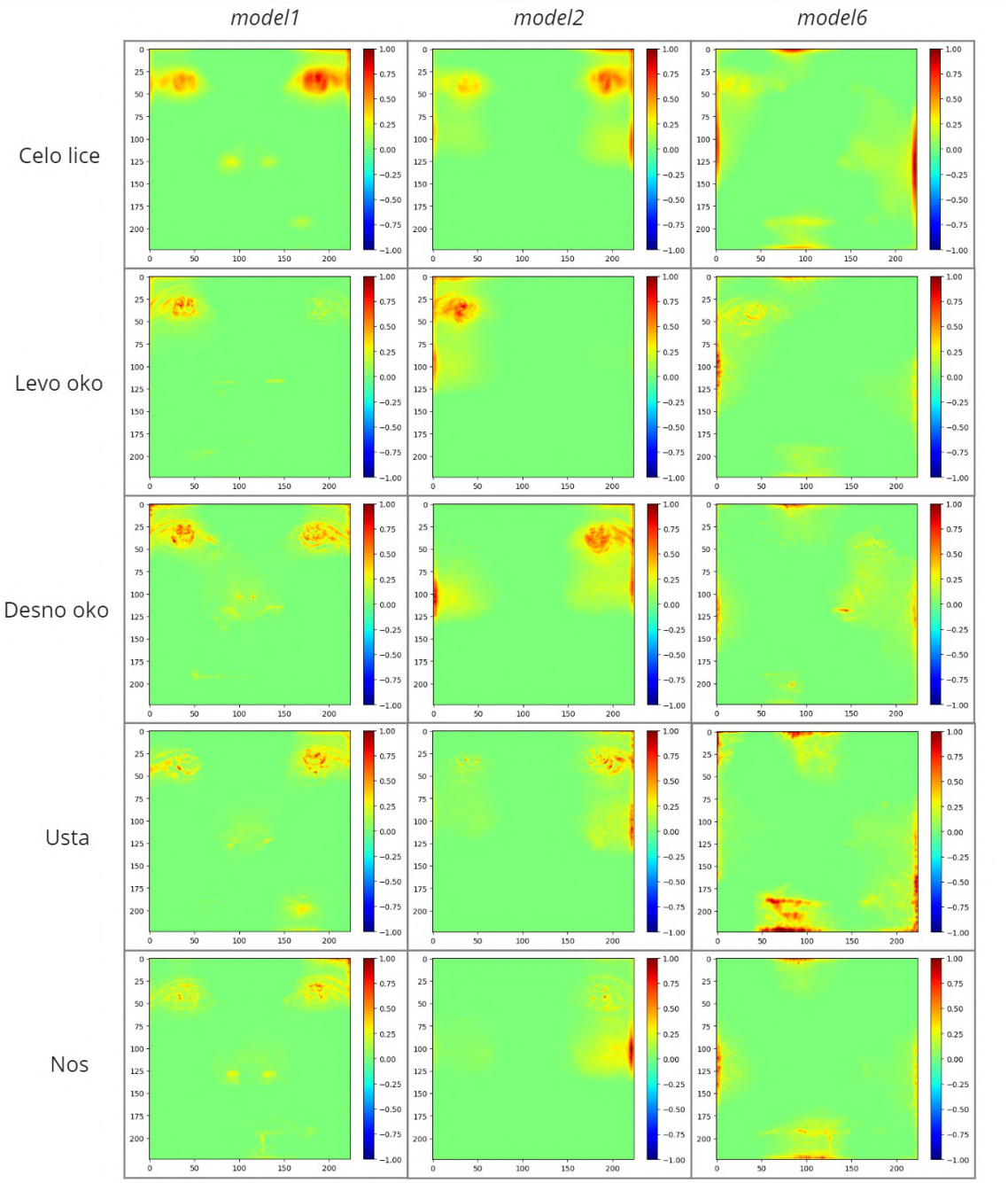


Slika 6.14: Procenat neuspelih suparničkih napada intenziteta šuma $\varepsilon = [1, 2, 3, ..., 10]$ za različite modele.

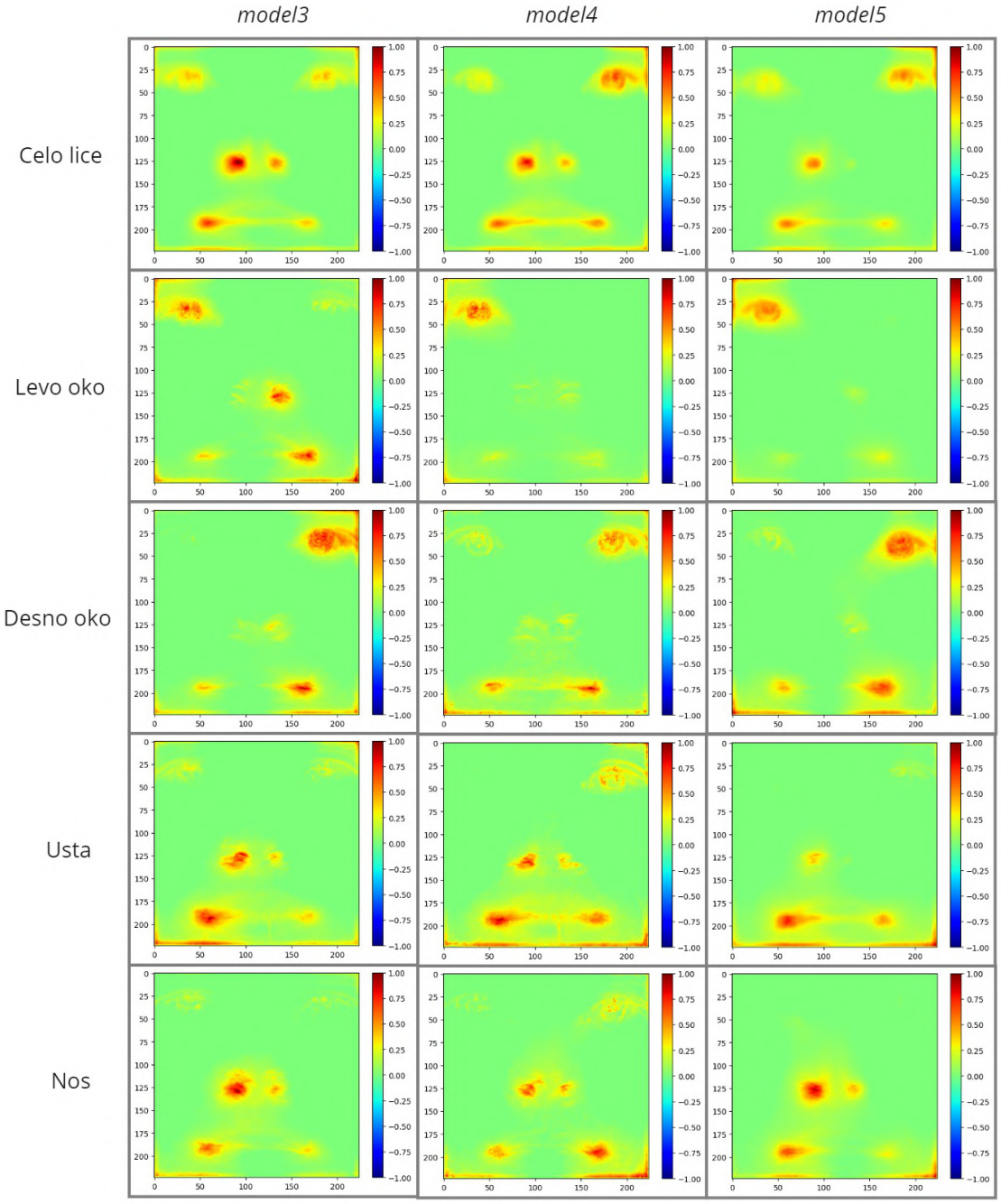
biblioteke *iNNvestigate* primenjena je kako bi se dobile toplotne mape svih modela za preobražene slike, ali i toplotne mape za preobražene slike regija, da bi utvrdili da li postoji veza između uspeha u detektovanju semantičkih napada i regija na koje se model fokusira. Izlazni neuron za klasu preobraženih lica, bez softmax aktivacije, korišćen je kao izlazni neuron od kojeg kreće propagacija relevantnosti LRP metode. Korišćene su samo ispravno klasifikovane lažne instance.

Udeo relevantnosti u različitim regijama u odnosu na preobražene regije za različite modele dat je redom u tabelama 6.6, 6.7, 6.8, 6.9, 6.10 i 6.11. Toplotne mape za različite modele dobijene na slikama gde su neke regije (ili celo lice) preobražene date su na slikama 6.15 i 6.16. Najpre primećujemo da se modeli trenirani standardno ($model_1$, $model_2$ i $model_6$) fokusiraju uglavnom na oči, čak i u slučajevima kad ispravno klasifikuju slike na kojima su lažni samo usta ili nos. $model_2$ gotovo da uopšte ne obraća pažnju na nos i usta, dok $model_6$ ne obraća pažnju na nos, ali obraća na usta, što je verovatno posledica prve faze treninga. Ovi modeli ostvaruju najbolje EER ocene, te bi isprva mogli zaključiti da oči zaista jesu dovoljne za uspešnu klasifikaciju lažnih slika celog lica. Međutim, posmatranjem udela relevantnosti po regijama ovih modela, zaključujemo da je veliki procenat relevantnosti raspoređen van četiri glavne regije - 45% za $model_2$ i čak 66% za $model_6$. Pažljivijom analizom toplotnih mapa možemo primetiti da je dosta relevantnosti raspoređeno uz ivicu slike. Da li je ovo zaista ispravan razlog za klasifikaciju lažnih slika celog lica, ili pak predstavlja neku vrstu preprilagođavanja kao posledicu finog podešavanja, ostaje za diskusiju. Ono što je sigurno jeste da se ove regije mogu lako iseći tako da lice ostane zadovoljavajućeg formata, pa je pitanje kako bi se onda ovi modeli ponašali. Kod modela treniranih na jednoj ili više regija ($model_3$, $model_4$, $model_5$) relevantnost je ravnomernije raspoređena po licu. Udeo relevantnosti u 4 glavne regije je uglavnom između 70 i 80%. Na toplotnim mapama za levo, desno oko, usta i nos uočavamo da su odgovarajuće regije dominantne, ali i da ostale regije u značajnoj meri utiču na odluku, iako nisu preobražene. Ovo je posebno zanimljivo za $model_3$ koji je treniran na jednoj regiji i koji pogađa u proseku čak 80% slika kod kojih je jedna regija preobražena. Ako je taj model pogodio 86% slika gde je nos lažan, a relevantnost nosa na toplotnoj mapi za te slike je samo 41%, postavlja se pitanje na osnovu čega zapravo model donosi odluku. Glavna pretpostavka za ovako čudno ponašanje modela jeste da su podaci loši, pošto je način generisanja ovih podataka neka vrsta eksperimentisanja. Možda su nepažnjom neki koraci u preprocesiranju lažnih slika ostavili tragove koji bi mogli da zavaraju model. Možda augmentacije

nisu dobro definisane, te ne pomažu modelu da se fokusira na složenije atribute. Ovo treba uzeti sa rezervom, jer je mala šansa da su ti varljivi atributi upravo na relevantnim regijama.



Slika 6.15: Toplotne mape za lice i 4 regije za modele koji su trenirani standardno (u 1 ili 2 faze)



Slika 6.16: Toplotne mape za lice i 4 regije za modele koji su trenirani na jednoj ili više regija (u 1 ili 2 faze)

	Levo oko	Desno oko	Usta	Nos	Ostatak lica
Celo lice	49%	27%	2%	6%	16%
Levo oko	64%	11%	4%	3%	18%
Desno oko	32%	39%	2%	9%	18%
Usta	40%	28%	4%	11%	17%
Nos	41%	20%	11%	12%	16%

Tabela 6.6: Udeo relevantnosti u regijama za različite preobražene regije za $model_1$. Svaki red predstavlja raspored relevantnosti po glavnim regijama i ostatku lica za toplotnu mapu dobijenu na slikama gde je navedena regija preobražena.

	Levo oko	Desno oko	Usta	Nos	Ostatak lica
Celo lice	35%	20%	0%	0%	45%
Levo oko	55%	0%	0%	1%	44%
Desno oko	2%	40%	0%	3%	55%
Usta	26%	1%	0%	4%	69%
Nos	36%	8%	0%	2%	54%

Tabela 6.7: Udeo relevantnosti u regijama za različite preobražene regije za $model_2$. Svaki red predstavlja raspored relevantnosti po glavnim regijama i ostatku lica za toplotnu mapu dobijenu na slikama gde je navedena regija preobražena.

	Levo oko	Desno oko	Usta	Nos	Ostatak lica
Celo lice	13%	14%	28%	28%	17%
Levo oko	25%	4%	26%	19%	26%
Desno oko	1%	37%	28%	8%	26%
Usta	5%	5%	42%	21%	27%
Nos	3%	4%	34%	41%	18%

Tabela 6.8: Udeo relevantnosti u regijama za različite preobražene regije za $model_3$. Svaki red predstavlja raspored relevantnosti po glavnim regijama i ostatku lica za toplotnu mapu dobijenu na slikama gde je navedena regija preobražena.

	Levo oko	Desno oko	Usta	Nos	Ostatak lica
Celo lice	22%	7%	34%	20%	17%
Levo oko	46%	0%	19%	9%	26%
Desno oko	5%	21%	38%	8%	28%
Usta	12%	0%	45%	14%	29%
Nos	11%	1%	32%	26%	30%

Tabela 6.9: Udeo relevantnosti u regijama za različite preobražene regije za $model_4$. Svaki red predstavlja raspored relevantnosti po glavnim regijama i ostatku lica za toplotnu mapu dobijenu na slikama gde je navedena regija preobražena.

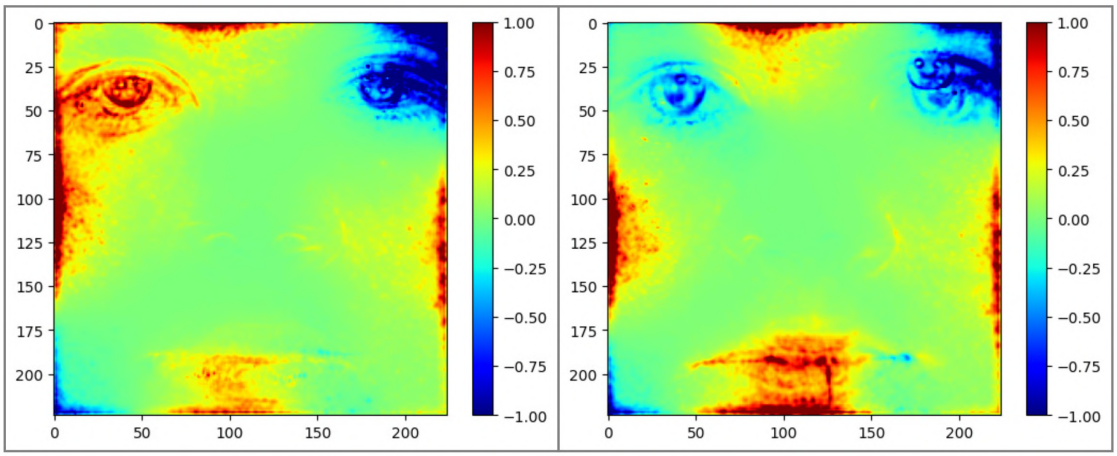
	Levo oko	Desno oko	Usta	Nos	Ostatak lica
Celo lice	25%	17%	26%	14%	18%
Levo oko	43%	1%	22%	9%	25%
Desno oko	6%	29%	35%	3%	27%
Usta	10%	13%	38%	9%	40%
Nos	6%	6%	28%	39%	21%

Tabela 6.10: Udeo relevantnosti u regijama za različite preobražene regije za $model_5$. Svaki red predstavlja raspored relevantnosti po glavnim regijama i ostatku lica za toplotnu mapu dobijenu na slikama gde je navedena regija preobražena.

	Levo oko	Desno oko	Usta	Nos	Ostatak lica
Celo lice	2%	19%	10%	3%	66%
Levo oko	12%	8%	4%	12%	64%
Desno oko	0%	32%	9%	1%	58%
Usta	1%	7%	28%	0%	64%
Nos	2%	8%	27%	1%	62%

Tabela 6.11: Udeo relevantnosti u regijama za različite preobražene regije za $model_6$. Svaki red predstavlja raspored relevantnosti po glavnim regijama i ostatku lica za toplotnu mapu dobijenu na slikama gde je navedena regija preobražena.

Još jednu zanimljivu pojavu možemo uočiti kad u procesu generisanja toplotnih mapa ne eliminišemo negativne relevantnosti. One su na primerima toplotnih mapa obojene plavom bojom (slika 6.17). Za modele trenirane u dve faze primećujemo da se kod klasifikovanja slika jedne regije javlja negativna relevantnost kod nekih drugih regija. Pretpostavka je da model u tom slučaju poredi na primer levo i desno oko i na osnovu toga zaključuje da li je neko lažno. Ova ideja ispitana je detaljno u [27].



Slika 6.17: Toplotne mape sa negativnim relevantnostima za $model_6$ za slike preobraženog levog oka (levo) i preobraženih usta (desno), koje nagoveštavaju poređenje regija kao posledicu treninga u dve faze

Glava 7

Zaključak

LRP metoda se pokazala korisnom u raznim problemima mašinskog učenja, pružajući dodatna objašnjenja uz same rezultate koje modeli daju. Preobražavanje lica predstavlja ozbiljnu bezbednosnu pretnju, što je i potvrđeno u ovom radu. Značajan broj preobraženih lica uspeo je da prevari čak i vrlo napredne modele za prepoznavanje lica. U ovom radu detaljno su opisane dve procedure za generisanje preobraženih lica, jedna zasnovana na trouglovima, druga zasnovana na atributima (Bajer-Nili). U daljem radu bilo bi zanimljivo eksperimentisati sa dodatnim modifikacijama na preobraženim slikama, u cilju uklanjanja vidljivih tragova koje preobražavanje ostavlja i poboljšanja njihovog kvaliteta. Modeli za prepoznavanje preobraženih lica trenirani na skupu originalnih i preobraženih lica ostvarili su odlične rezultate, sa EER ocenom 3.0%. LRP metodom utvrđeno je da se ovako trenirani modeli najviše oslanjaju na oči kada ispravno klasifikuju preobražene slike, što može biti mana, jer su na taj način laka meta raznih napada. Pokazano je da su takvi modeli u maloj meri otporni na semantičke i suparničke napade. Modifikacije preobraženih slika za trening, koje preobražene delove lica ostavljaju samo u određenim regijama i time ograničavaju količinu informacija koje model ima na raspolaganju za njihovo prepoznavanje, pokazale su se delotvornim. Iako imaju nešto nižu EER ocenu od standardno treniranih modela, modeli trenirani na modifikovanim podacima su dosta otporniji na navedene napade. Ono što nije ispitano u radu, a bilo bi interesantno probati jeste otpornost na suparničke napade gde je šum za suparnički primer dodat samo na određene regije. Takođe, bilo bi dobro za polaznu tačku uzeti mreže trenirane na nekim poznatim skupovima za prepoznavanje lica umesto na ImageNet skupu.

Toplotne mape generisane LRP metodom su opravdale rezultate različitih modela, ali su i nametnule neka pitanja. Na primer, zašto je velika količina relevantnosti

izvan četiri glavne regije lica, konkretno na ivicama slike i da li bi se blagim isecanjem ili transliranjem slika zbog toga poremetila odluka modela. Drugo pitanje je da li su neka čudna svojstva toplotnih mapa posledica loših podataka, odnosno da li su u nedostatku jasno definisane, jedinstvene procedure preobražavanja lažne slike loše generisane. Takođe značajno pitanje je kako bi promena arhitekture uticala na rezultate i toplotne mape, na primer umesto VGG19 da se koristi ResNet-50 za izvlačenje atributa.

Svi navedeni rezultati ukazuju na značaj LRP metode u primenama u bezbednosti. Iako je XAI grana koja se sporije razvija od trenutno popularnih oblasti veštačke inteligencije, dodatna objašnjenja ponašanja modela mogu biti vrlo korisna, te njen pun potencijal tek treba da dođe do izražaja.

Bibliografija

[1] https://github.com/alyssaq/face_morpher. https://github.com/alyssaq/face_morpher. MIT License.

[2] Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. iNNvestigate Neural Networks! *Journal of Machine Learning Research*, 20(93):1–8, 2019.

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. *Information Fusion*, 58:82–115, 2020.

[4] BBC. AI frenzy makes Nvidia the world's most valuable company. `https://www.bbc.com/news/articles/cyrr40x0z2mo`, June 19 2024.

[5] Thaddeus Beier and Shawn Neely. Feature-Based Image Metamorphosis. *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 1992.

[6] Lisa DeBruine and Benedict Jones. Face Research Lab London Set. 5 2017.

[7] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *CVPR*, 2019.

[8] Bach et al. Explainable AI Demos. `https://lrpserver.hhi.fraunhofer.de/`.

[9] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. The magic passport. *IEEE International Joint Conference on Biometrics*, pages 1–7, 2014.

[10] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. *arXiv 1412.6572*, 12 2014.

[11] Peter Hancock. "https://pics.stir.ac.uk/".

[12] Zhanhao Hu, Wenda Chu, Xiaopei Zhu, Hui Zhang, Bo Zhang, and Xiaolin Hu. Physically Realizable Natural-Looking Clothing Textures Evade Person Detectors via 3D Modeling. 2023.

[13] Gary B. Huang, Vidit Jain, and Erik Learned-Miller. Unsupervised Joint Alignment of Complex Images. In *ICCV*, 2007.

[14] Anjana Lakshmi, Bernd Wittenbrink, Joshua Correll, and Debbie S. Ma. The India Face Set: International and Cultural Boundaries Impact Face Impressions and Perceptions of Category Membership. *Frontiers in Psychology*, 12, 2021.

[15] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. *Nature Communications*, 10(1):1096, Mar 2019.

[16] Debbie S. Ma, Joshua Correll, and Bernd Wittenbrink. The Chicago Face Database: A Free Stimulus Set of Faces and Norming Data. *Behavior Research Methods*, 47(4):1122–1135, Dec 2015.

[17] Debbie S. Ma, Justin Kantner, and Bernd Wittenbrink. Chicago Face Database: Multiracial expansion. *Behavior Research Methods*, 53(3):1289–1300, Jun 2021.

[18] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. *Layer-Wise Relevance Propagation: An Overview*, pages 193–209. Springer International Publishing, Cham, 2019.

[19] Tom Neubert, Andrey Makrushin, Mario Hildebrandt, Christian Kraetzer, and Jana Dittmann. Extended StirTrace Benchmarking of Biometric and Forensic Qualities of Morphed Face Images. *IET Biometrics*, 7, 02 2018.

[20] Nicolas Papernot, Patrick Mcdaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical Black-Box Attacks against Machine Learning. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 2016.

[21] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, SIGGRAPH '03, page 313–318, New York, NY, USA, 2003. Association for Computing Machinery.

[22] P. Jonathon Phillips, Hyeonjoon Moon, Syed Rizvi, and Patrick Rauss. The FERET Evaluation Methodology for Face-Recognition Algorithms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22:1090 – 1104, 10 2000.

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[24] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *ArXiv*, abs/1708.08296, 2017.

[25] Ulrich Scherhag, Christian Rathgeb, Johannes Merkle, Ralph Breithaupt, and Christoph Busch. Face Recognition Systems Under Morphing Attacks: A Survey. *IEEE Access*, PP:1–1, 02 2019.

[26] Clemens Seibold, Wojciech Samek, Anna Hilsmann, and Peter Eisert. Detection of Face Morphing Attacks by Deep Learning. pages 107–120, 07 2017.

[27] Clemens Seibold, Wojciech Samek, Anna Hilsmann, and Peter Eisert. Accurate and Robust Neural Networks for Face Morphing Attack Detection. *Journal of Information Security and Applications*, 53:102526, 2020.

[28] Sefik Ilkin Serengil and Alper Ozpinar. LightFace: A Hybrid Deep Face Recognition Framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE, 2020.

[29] Sefik Ilkin Serengil and Alper Ozpinar. A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules. *Bilisim Teknolojileri Dergisi*, 17(2):95–107, 2024.

[30] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014.

[31] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and ZB Wojna. Rethinking the Inception Architecture for Computer Vision. 06 2016.

[32] Carlos Eduardo Thomaz and Gilson Antonio Giraldi. A new ranking method for principal components analysis and its application to face image analysis. *Image and Vision Computing*, 28(6):902–913, 2010.

[33] European Union. EU Artificial Intelligence Act, 2021.

[34] Sushma Venkatesh, Raghavendra Ramachandra, Kiran Raja, and Christoph Busch. Face Morphing Attack Generation & Detection: A Comprehensive Survey.

[35] Michelangelo Vianello. Dataset, Sep 2023.

[36] Shiv Vignesh. The world through the eyes of CNN. `https://medium.com/analytics-vidhya/the-world-through-the-eyes-of-cnn-5a52c034dbeb`, Jun 26 2020.

[37] Wikipedia. Automated border control system. `https://en.wikipedia.org/wiki/Automated_border_control_system`, 2024.

[38] Sebastian Lapuschkin Wojciech Samek. Layer-wise Relevance Propagation. `https://www.hhi.fraunhofer.de/en/departments/ai/technologies-and-solutions/layer-wise-relevance-propagation.html`.

[39] Peiyu Xiong, Michael Tegegn, Jaskeerat Sarin, Shubhraneel Pal, and Julia Rubin. It Is All About Data: A Survey on the Effects of Data on Adversarial Robustness. 03 2023.

[40] Feng Zhu. Image Morphing with the Beier-Neely Method. 2015.

Biografija autora

Vladan Kovačević rođen je 26. oktobra 1998. godine u Subotici, gde je završio osnovnu školu i Gimnaziju „Svetozar Marković“ kao đak generacije. 2017. godine na Matematičkom fakultetu u Beogradu upisao je osnovne akademske studije na smeru računarstvo i informatika. Diplomirao je 2021. godine, nakon čega upisuje master akademske studije na istom smeru. Iste godine zaposlio se kao saradnik u nastavi na Matematičkom fakultetu u Beogradu na Katedri za računarstvo i informatiku, gde je dve godine držao vežbe iz predmeta „Objektno-orijentisano programiranje“, „Algoritmi i strukture podataka“ i „Programiranje 2“. Školske 2022/23 godine držao je mentorsku nastavu iz programiranja u Matematičkoj gimnaziji u Beogradu. Od 2023. godine zaposlen je u kompaniji *EyeSee* na poziciji inženjera mašinskog učenja, gde se bavi primenama mašinskog učenja u praćenju pogleda, modelovanju ljudske pažnje, detekciji i analizi pakovanja.