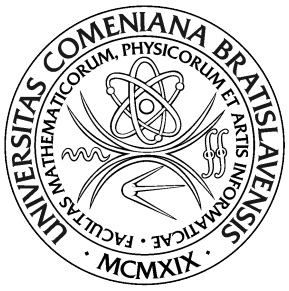
UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



IDENTIFIKÁCIA OSOBY PRI ČIASTOČNEJ OKLÚZII TVÁRE

Diplomová práca

2021 Bc. Adam Valentovič

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

IDENTIFIKÁCIA OSOBY PRI ČIASTOČNEJ OKLÚZII TVÁRE

Diplomová práca

Študijný program:	Aplikovaná informatika
Študijný odbor:	2511 Aplikovaná informatika
Školiace pracovisko:	Katedra aplikovanej informatiky
Školiteľ:	RNDr. Zuzana Černeková, PhD.

Bratislava, 2021 Bc. Adam Valentovič

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Adam Valentovič
Študijný program: aplikovaná informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Identifikácia osoby pri čiastočnej oklúzii tváre
Person identification with partially occluded face

Anotácia: Identifikovanie osoby podľa tváre človeka v prípade, že tvár je čiastočne zakrytá. Naštudovať problematiku identifikácie ľudských tvárí a možnosť využitia neurónových sietí. Analyzovať existujúce riešenia publikované v dostupnej odbornej literatúre. Vytvoriť databázu hľadaných tvárí pre trénovacie a testovacie účely. Navrhnúť a implementovať metódu, ktorá vyhľadá osobu podľa tváre človeka, v ktorom nie je viditeľná celá tvár. Vyhodnotiť dosiahnuté výsledky.

Cieľ: Identifikovanie osoby podľa tváre človeka v prípade, že tvár je čiastočne zakrytá. Naštudovať problematiku identifikácie ľudských tvárí a možnosť využitia neurónových sietí. Analyzovať existujúce riešenia publikované v dostupnej odbornej literatúre. Vytvoriť databázu hľadaných tvárí pre trénovacie a testovacie účely. Navrhnúť a implementovať metódu, ktorá vyhľadá osobu podľa tváre človeka, v ktorom nie je viditeľná celá tvár. Vyhodnotiť dosiahnuté výsledky.

Vedúci: RNDr. Zuzana Černeková, PhD.
Katedra: FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry: prof. Ing. Igor Farkaš, Dr.

Dátum zadania: 30.09.2020

Dátum schválenia: 08.10.2020 prof. RNDr. Roman Ďurikovič, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Čestne prehlasujem, že túto diplomovú prácu som vypracoval samostatne len s použitím uvedenej literatúry a za pomoci konzultácií u môjho školiteľa.

. . . . . .

Bratislava, 2021 Bc. Adam Valentovič

Poďakovanie

Tu bude poďakovanie.

Abstrakt

Tu bude abstrakt.

Kľúčové slová: tu budú kľúčové slová oddelené čiarkou bez bodky na konci (ks1, ks2, ..., ksn)

Abstract

There will be abstract.

Keywords: there will be keywords (kw1, kw2, ..., kwn)

Obsah

Kapitola 1

Úvod

Tu bude úvod.

Kapitola 2

Motivácia

Tu bude motivácia.

Kapitola 3

Prehľad problematiky

Identifikácia konkrétnej osoby na základe jej tváre je dôležitou súčasťou oblastí počítačového videnia. Je to zložitá úloha, ktorá naráža na veľké množstvo nie len technologických ale aj etických problémov, ako je napríklad zneužitie týchto technológií v štátom riadenej represii. Identifikáciou rozumieme zistenie, ktorá osoba je na obrázku, verifikáciou je myslené overenie, či sa jedná o rovnakú osobu ako na inom obrázku a zhlukovaním sa rozumie nájdenie podobných tvárí ako na obrázku.

V tejto kapitole sa oboznámime s problémom identifikácie osoby pri čiastočnej oklúzii tváre. Nakoľko neexistuje veľa vedeckých prác zaoberajúcich sa oklúziou pri identifikácii, predstavíme metódy, ktoré neboli priamo navrhnuté pre dáta s oklúziou.

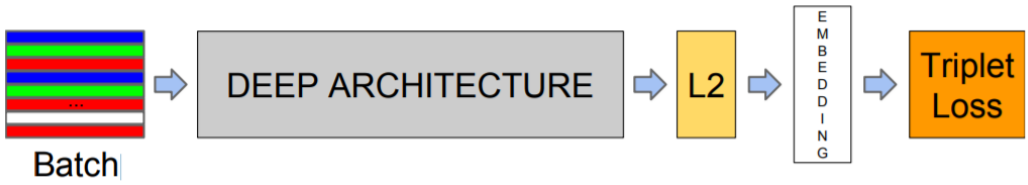
3.1 Súvisiace práce

Keďže na väčšinu klasifikačných problémov počítačového videnia sa v dnešnej dobe používajú neurónové siete, v tejto časti sa budeme venovať najmä metódam, ktoré ich taktiež využívajú.

FaceNet je zjednotený systém pre verifikáciu, identifikáciu a zhlukovanie tvárí [5]. Je to metóda, ktorá sa pre každý obrázok učí jeho pozíciu v euklidovskom priestore. Tento priestor nazývame euklidovské rozloženie. Sieť je trénovaná tak, že L2 vzdialenosti v jej výstupnom priestore priamo korešpondujú podobnostiam tvárí, teda tváre rovnakej osoby majú medzi sebou malé vzdialenosti a tváre rôznych osôb veľké vzdialenosti v euklidovskom rozložení.

Po vytvorení euklidovského rozloženia sa vyššie uvedené úlohy stávajú priamočiarymi.

- Verifikácia znamená stanovenie prahovej vzdialenosti v euklidovskom rozložení medzi dvoma obrázkami.
- Identifikácia sa stáva problémom k-NN klasifikácie.
- Zhlukovanie vieme dosiahnuť pomocou bežne dostupných zhlukovacích algoritmov ako napríklad k-means.

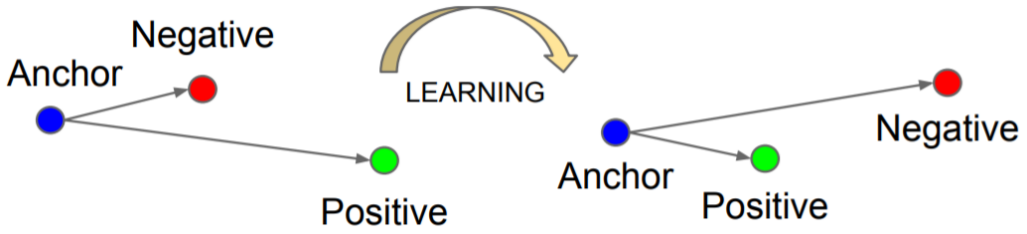


Obr. 3.1: Štruktúra modelu FaceNet.

Na obrázku 3.1 vidíme architektúru modelu FaceNet. Skladá sa z dávkovej (batch) vstupnej vrstvy, vrstvy s hlbokou konvolučnou neurónovou sieťou (deep architecture) a L2 normalizačnej vrstvy. Výstupom siete sú zobrazenia do euklidovského priestoru (embeddings). Autori sa správajú k vrstve s hlbokou neurónovou sieťou ako k čiernej skrinke (blackbox), čo znamená,

že je možné vyskúšať rôzne známe modely. Autori sa rozhodli vyskúšať dve architektúry popísané v publikáciách [11, 6].

Táto architektúra je trénovaná ako siamská neurónová sieť prostredníctvom triplet stratovej funkcie (triplet loss).



Obr. 3.2: Učenie prostredníctvom triplet loss stratovej funkcie.

Triplet loss stratová funkcia minimalizuje vzdialenosť medzi zobrazeniami vzorového (anchor) a pozitívneho obrázku, a maximalizuje vzdialenosť medzi zobrazeniami vzorového obrázku a negatívneho obrázku viď. obr. 3.2.

Za pozitívny sa rozumie obrázok rovnakej osoby ako na vzorovom obrázku, ktorý je ale odlišný od vzorového obrázku. Za negatívny sa rozumie obrázok inej osoby ako na vzorovom obrázku.

Funkcia musí zabezpečiť, že vzorový obrázok x_i^a je bližšie ku každému obrázku tej istej osoby x_i^p (pozitívnemu obrázku) ako k ľubovolnému obrázku inej osoby x_i^n (negatívnenu obrázku). Teda musia platiť podmienky:

$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2, \forall (x_i^a, x_i^p, x_i^n) \in \tau. \tag{3.1}$$

Kde α je vynútený rozostup medzi pozitívnymi a negatívnymi dvojicami. τ je množina všetkých možných tripletov nad trénovacou množinou, ktorej kardinalita je N.

Minimalizovaná stratová funkcia je potom tvaru

$$L = \sum_{i}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a - f(x_i^n)\|_2^2 + \alpha \right]. \tag{3.2}$$

Trénovanie cez celú množinu τ, teda cez všetky možné triplety vedie k pomalej konvergencii, nakoľko vzniká veľa ľahko splniteľných tripletov, ktoré neprispievajú k trénovaniu, nakoľko vždy vykazujú nízku stratu. Je preto dôležité vyčleniť náročné triplety, ktoré môžu prispieť k trénovaniu modelu lepšie.

Pre rýchlu konvergenciu pri trénovaní je dôležité vyčleniť také triplety, ktoré porušujú obmedzenie (3.1). To znamená, že k vzorovému obrázku chceme nájsť ťažký pozitívny a ťažký negatívny obrázok.

- Ťažký pozitívny obrázok je taký obrázok, ktorý sa najmenej podobá na pôvodný obrázok, ale stále sa jedná o rovnakú osobu ako na vzorovom obrázku, teda hľadáme $\arg\max \|f(x_i^a) - f(x_i^p)\|_2^2$.

- Ťažkým negatívnym obrázkom sa rozumie taký obrázok, ktorý sa čo najviac podobá na vzorový obrázok, ale je na ňom iná osoba, teda hľadáme $\arg\min \|f(x_i^a - f(x_i^n)\|_2^2$.

Je nemožné hľadať argmax a argmin spomenutý vyššie pre každú vzorku z trénovacej množiny.

Máme teda dve možnosti ako generovať triplety.

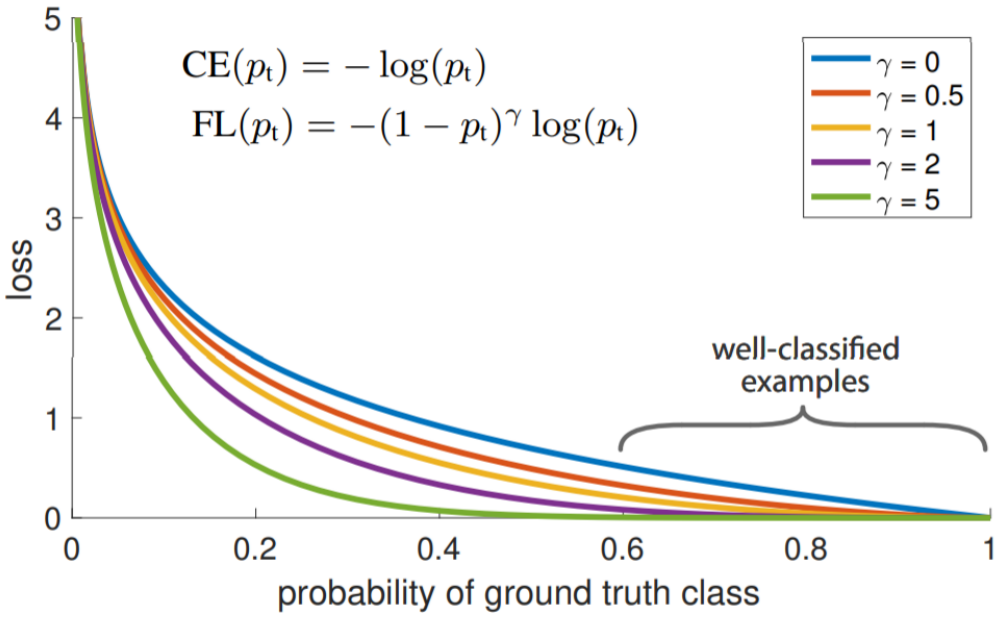
- Generovanie tripletov offline každých n krokov pri trénovaní, pričom na získanie $f(x_i^a), f(x_i^p) a f(x_i^n)$ potrebných pre nájdenie argmin a argmax použijeme najaktuálnejší natrénovaný model siete (checkpoint), pričom argmin a argmax hľadáme len na podmnožine trénovacej množiny.

- Generovanie tripletov online, kde výber ťažkých pozitívnych/negatívnych obrázkov počítame vždy v rámci mini-dávky (mini-batch).

Pri online generácii tripletov volíme mini-dávky o veľkosti niekoľko tisíc exemplárov, pričom argmin a argmax počítame vždy v rámci mini-dávky. Pre zmysluplnosť vzdialeností medzi vzorovými a pozitívnymi obrázkami, musí byť zabezpečený minimálny počet exemplárov každej identity pre každú mini-dávku. Autori použili približne 40 tvárí každej identity v mini-dávke.

Namiesto hľadania najťažšej pozitívnej vzorky sú použité všetky dvojice (vzor, pozitívny) a k nim je vždy hľadaný najťažší negatívny (v rámci mini-dávky).

RetinaNet je známy one-stage objektový detektor, ktorý bol navrhnutý na prezentovanie novej navrhnutej focal loss stratovej funkcie [4], ktorá rieši problém nerovnovážneho zastúpenia tried pri trénovaní.



Obr. 3.3: Focal loss stratová funkcia pre rôzne parametre. Všimnime si, že pre $\lambda = 0$ je to vlastne známa cross-entropy stratová funkcia.

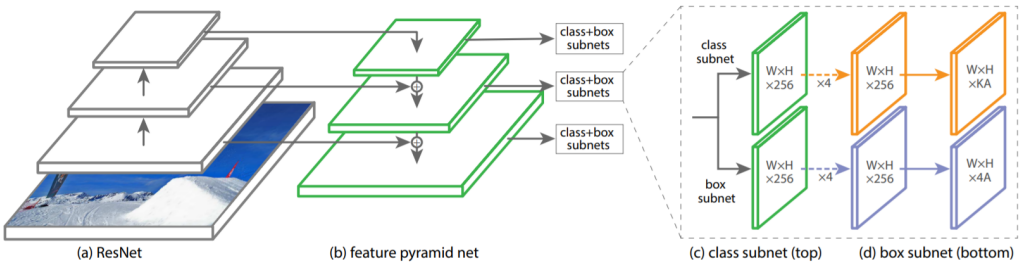
Ako môžeme vidieť na obrázku 3.3, focal loss strata aplikuje modulačný člen na cross-entropy stratu, aby sa učenie zameralo na ťažké negatívne vzorky. Je definovaná ako

$$FL(p_t) = -(1 - p_t)^{\gamma} \log(p_t), \tag{3.3}$$

kde

$$p_t = \begin{cases} p, & \text{ak } y = 0 \\ 1 - p & \text{inak} \end{cases}, \tag{3.4}$$

kde y je ground-truth hodnota a p je predikovaná hodnota.



Obr. 3.4: Architektúra objektového detektora RetinaNet.

Architektúra detektora RetinaNet (obr. 3.4) sa skladá z *kostrovej siete* (a, b) a z dvoch konvolučných podsietí špecifických pre danú úlohu *triedna podsieť* (c) a *ohraničujúca podsieť* (d). Ako kostrová sieť je použitá *FPN* (b), vnútri ktorej je použitá dobredná ResNet (a) architektúra.

Použitím FPN (Feature Pyramid Network) autori riešia tzv. multi-scale detection problém, teda problém s detekciou objektov rôznych veľkostí, pričom každá vrstva pyramídy je použitá na detekciu objektov rôznej škály.

Triedna podsieť klasifikuje objekty výstupu kostrovej siete, preto je jej výstup tvaru $W \times H \times KA$.

Ohraničujúca podsieť regresívne hľadá ohraničenia (bounding boxes) týchto objektov, preto je jej výstup tvaru $W \times H \times 4A$.

Redundantné detekcie jedného konkrétneho objektu predstavujú problém výstupu z ohraničujúcej siete. Keďže každý objekt na obrázku môže byť ohraničujúcou sieťou predikovaný viacerými ohraničeniami (anchor boxes), na odstránenie redundancie a získanie pravého ohraničenia (bounding box) je pre každú triedu anchor boxov nezávisle od seba použitá známa technika NMS (non-maximum suppression).

Pre každú skupinu prekrývajúcich sa anchor boxov sa zvolí ako výsledný bounding box ten, ktorý má najvyššie skóre (confidence). Následne sú odstránené všetky anchor boxy z jeho okolia, ktoré majú so zvoleným anchor boxom metriku IoU (intersection over union) väčšiu ako 0.5.

Metrika IoU v tomto prípade vyjadruje pomer medzi prienikom a zjednotením plôch ohraničených dvoma anchor boxami. V počítačovom videní sa často využíva na vyhodnotenie presnosti výstupov ohraničujúcich, prípadne segmentačných algoritmov oproti ground-truth hodnotám.

Face Attention Network (FAN) je efektívny one-shot detektor založený na RetinaNet architektúre pre tváre s oklúziou [8].

Náročnosť detekcie tváre s oklúziou spočíva v riziku problému s falošnou pozitivitou. Napríklad modely, ktoré sú schopné rozpoznávať len spodnú časť tváre je možné ľahko pomýliť. Napríklad na miestach na obrázku kde sa nachádzajú ľudské ruky, model klasifikuje tváre, pretože zdielajú rovnakú farbu kože ako tvár.

Ilustračný exemplár výsledkov detekcií s oklúziou v dave ľudí vidíme na obrázku 3.5.

Obr. 3.5: FAN úspešne detegujúca vyše 120 tvárí s oklúziou.

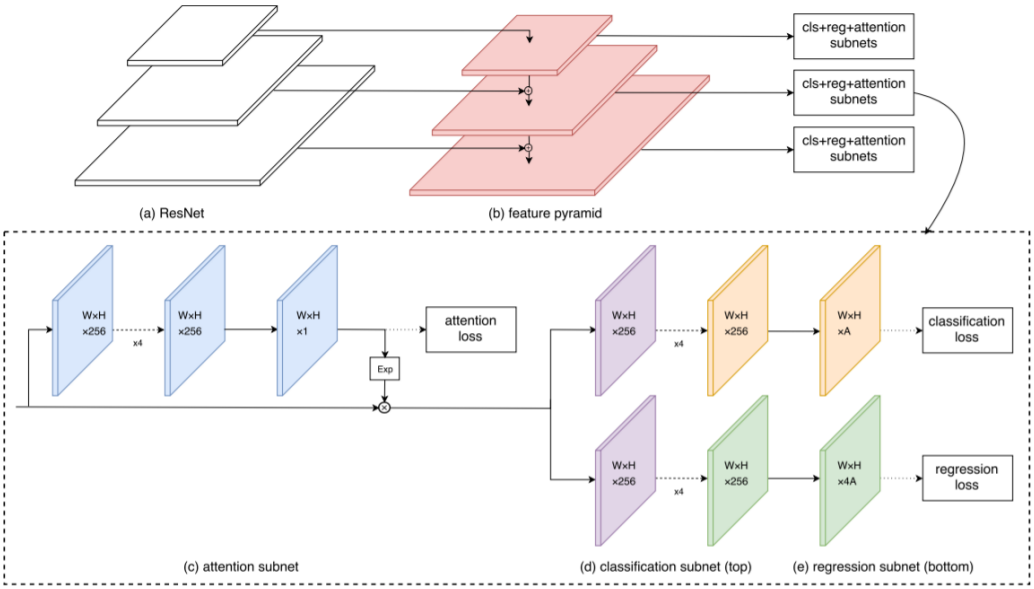
.

Hlavná štruktúra siete FAN (obr. 3.6) je prispôsobením štruktúry RetinaNet (obr. 3.4).Keďže FAN je prispôsobená na detekciu tvárí, máme len jednu triedu objektov narozdiel od siete RetinaNet. Pri triednej podsieti (obr. 3.6, d., classification subnet) je teda narozdiel od siete RetinaNet namiesto $W \times H \times KA$ výstup tvaru $W \times H \times A$, teda $K = 1$ reprezentuje jedinú triedu objektov detektora.

Regresná podsieť hľadajúca ohraničenia tvárí (obr. 3.6, e., regression subnet) je identická s ohraničujúcou podsieťou pri RetinaNet. Attention Network využíva tri dizajnové princípy:

- Rôzne škály pre rôzne vrstvy príznakov z kostrovej siete (obr. 3.6, a, b).
- Zvýrazňuje príznaky v oblasti tváre a potláča ich v oblastiach mimo tváre.

- Pri trénovaní je generované väčšie množstvo tvárí s oklúziou.



Obr. 3.6: Architektúra tvárového detektora FAN.

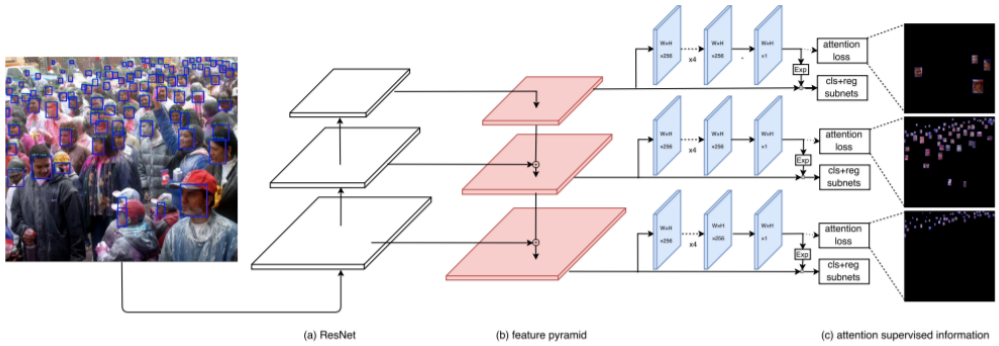
Pozornostná podsieť (obr. 3.6, c) je zapojená medzi kostrovou sieťou (a, b) a blokom s triednou (d) a regresnou (e) podsieťou.

Vo FAN je zapojených celkom 5 detekčných vrstiev (obrázok 3.6), každá asociovaná s anchor boxami špecifickej škály. V attention podsieti (c) je pomer strán pre každý anchor box nastavený na $1 : 1.5$, pretože väčšinu frontálnych tvárí možno považovať za obdĺžnik tvaru $1 : 1.5$. Detegované anchor boxy sa porovnávajú s ground truth boxami, znovu je využitý princíp non-maximum supression, pričom pri trénovaní je porovnávané IoU detegovaných anchor boxov s ground truth anchor boxom. Okrem toho sú všetky detegované anchorboxy s IoU < 0.4 oproti ground-truth automaticky ignorované.

Navrhnutá attention funkcia je založená na vyššie spomenutej sieťovej štruktúre, pričom využíva segmentovú bočnú vrstvu, ako vidieť na obrázku

3.7. Attention supervision informácie na trénovanie tepelných máp attention modulu sú získané z ground-truth anchor boxov priradeným k anchor boxom na aktuálnej vrstve.

Aby sa model učil vyprodukovať tepelnú mapu celej tváre pre rôzne časti tváre, zo vzoriek z trénovacieho datasetu sú náhodne vyrezávané štvorcové záplaty, pričom ich veľkosť strany je z intervalu $[0.3, 1]$ násobku menšej strany kratšej hrany originálneho obrázku. Okrem tejto augumentačnej techniky sú využité ešte augumentačné techniky prevrátenia (flip) a farebného chvenia (color jitter).



Obr. 3.7: Architektúra FAN: Ukážka hierarchickej attention vrstvy podľa pyramídy z FPN.

Stratová funkcia je pre celý detektor poskladaná z viacerých častí a je tvaru

$$\begin{aligned} L = & \sum_k \frac{1}{N_k^c} \sum_{i \in A_k} L_c(p_i, p_i^*) + \\ & \lambda_1 \sum_k \frac{1}{N_k^r} \sum i \in A_k I(P_i^* = 1) L_r(t_i, t_i^*) + \\ & \lambda_2 \sum_k L_a(m_k, m_k^*), \end{aligned} \tag{3.5}$$

kde k je index FPN pyramídy ($k \in [3, 7]$), A_k reprezentuje množinu anchorov definovanú vo vrstve pyramídy P_k. Ground truth označenie p_i^* je 1 ak je anchor pozitívny (je tam tvár), 0 inak. p_i je predikovaný klasifikačný výsledok z modelu. t_i je vektor reprezentujúci 4 parametrizované súradnice ohraničenia predikovaného anchor boxu a t_i^* je ground-truth bounding box asociovaný s pozitívnym anchorom.

Klasifikačná strata $L_c(p_i, p_i^*)$ je focal stratová funkcia z RetinaNet publikácie nad dvoma triedami (tvár a pozadie). N_k^c je počet anchorov v P_k, ktoré sa podieľajú na výpočte klasifikačnej straty. Regresná strata $L_r(t_i, t_i^*)$ je známa smooth L1 stratová funkcia často používaná pri regresii. $I(p_i^* = 1)$ je indikačná funkcia, ktorá obmedzuje regresnú stratu, aby sa zameriavala výlučne na pozitívne priradené anchory, a $N_k^r = \sum_{i \in A_k} I(p_i^* = 1)$. Attention strate $L_a(m_k, m_k^*)$ je sigmoidová cross-entropy strata vykonaná po pixeloch. m_k je mapa pozornosti (attention map) vygenerovaná po leveloch a m_k^* je jej ground truth mapa popísaná attention funkciou. λ_1 a λ_2 sú konštanty použité na vyváženie týchto troch popísaných stratových funkcií. Ich predvolené nastavenie je $\lambda_1 = \lambda_2 = 1$.

MFR. Autori článku [3] sa rozhodli porovnať a skombinovať dva prístupy pre identifikáciu tvárí s maskami (MFR), čo je špecifický prípad oklúzie pri konvenčnej identifikácii tvárí (FR):

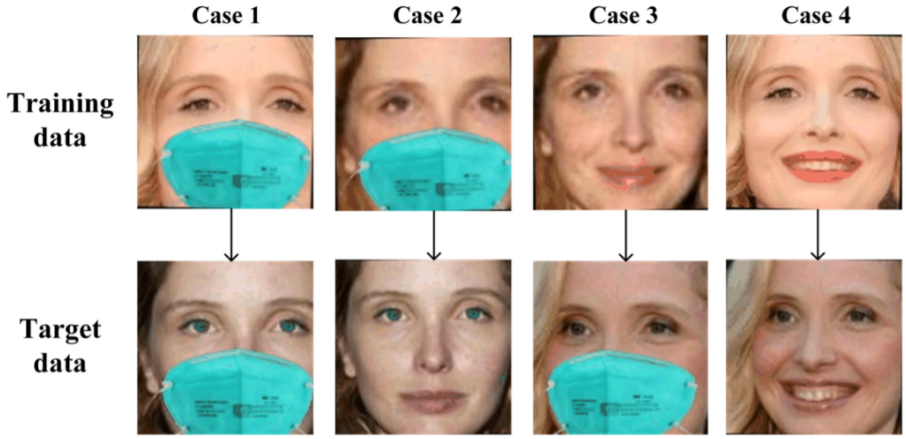
- prístup založený na pozornosti (AB),
- prístup založený na odrezávaní masky z obrázku (CB).

Tieto prístupy sú zamerané na zníženie negatívneho vplyvu regiónov na tvári, ktoré sú prekryté maskou, a teda ich sémantika je poškodená.

Vyššie spomenutá FAN, riešila oklúziu vo všeobecnosti, avšak FAN neriešila problém identifikácie, ale len detekciu tváre.

Ďalším častým problémom hlbokých architektúr adresujúcich MFR je riešenie špecifických prípadov oklúzie:

- Použitie dát s maskou na trénovanie a dát bez masky na testovanie,
- Použitie dát bez masky na trénovanie a dát s maskou na testovanie.

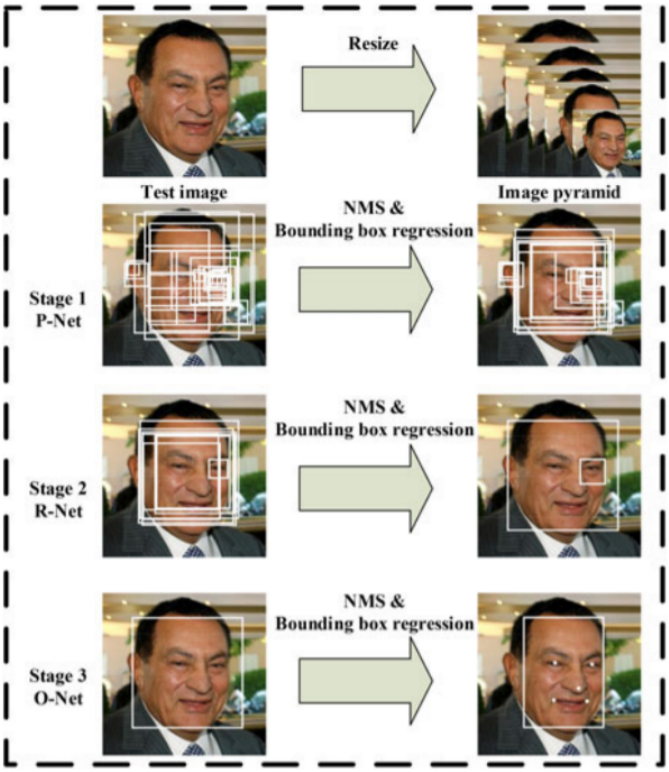


Obr. 3.8: Rozdelenie úloh MFR (Case 1, 2 a 3) a FR (Case 4).

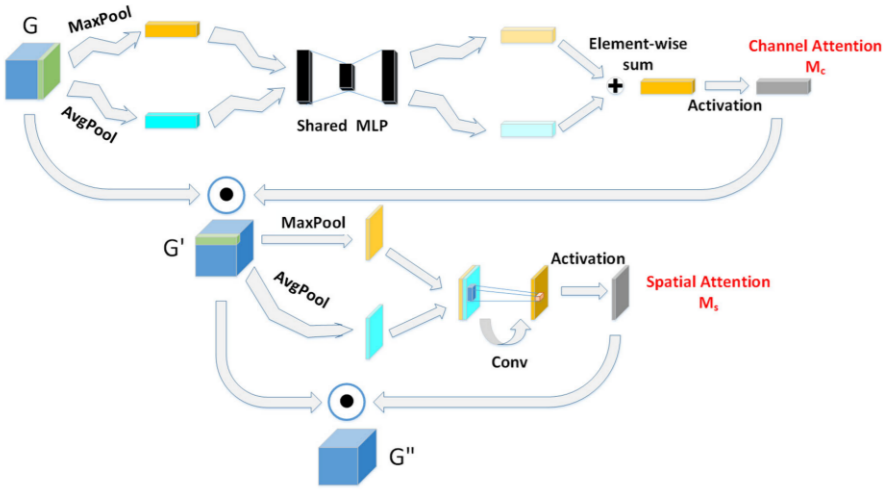
Na adresovanie tohto problému MFR úlohu rozdelíme na 3 prípady. Okrem toho ponecháme jeden prípad pre konvenčnú úlohu FR ako môžeme vidieť na obrázku 3.8.

Takéto rozdelenie nám umožňuje dataset na rozpoznávanie tváre so simulovanými maskami (SMFRD) [9], a teda je použitý na trénovanie navrhnutého MFR modelu.

Na spracovanie obrázkov autori využívajú známy tvárový detektor MTCNN [13], ktorého výstupy z detekcie vidíme na obrázku 3.9 vpravo dolu. MTCNN deteguje okrem ohraničenia aj kľúčové body na tvári.



Obr. 3.9: Pipeline tvárového detektora MTCNN.



Obr. 3.10: Diagram CBAM modulu.

Pri prvom spomenutom AB prístupe je použitý konvolučný blokový pozornostný modul (CBAM) [10], ktorý sa skladá z modulu pozornosti kanála (Channel Attention Module, CAM) a z modulu priestorovej pozornosti (Spatial Attention Module, SAM) viď. obrázok 3.10.

CBAM je využiteľný v rôznych konvolučných neurónových sietiach na získanie presnejších (jemnejších) príznakov. Jeho cieľom je zamerať sa na vplyvnejšie časti príznakových máp z kanálových (CAM) a priestorových dimenzií (SAM).

Pre danú príznakovú mapu $\boldsymbol{G} \in \mathbb{R}^{C\times H\times W}$, nech $\boldsymbol{M_c} \in \mathbb{R}^{C\times 1\times 1}$ označuje kanálovú pozornostnú mapu a $\boldsymbol{M_s} \in \mathbb{R}^{1\times H\times W}$ označuje priestorovú pozornostnú mapu.

Proces zjemňovania príznakov je následne označený ako

$$\begin{aligned} \boldsymbol{G'} &= \boldsymbol{M_c} \odot \boldsymbol{G}, \\ \boldsymbol{G''} &= \boldsymbol{M_s} \odot \boldsymbol{G'} \end{aligned} \tag{3.6}$$

kde $\odot$ je násobenie po prvkoch.

CAM sa zameriava na medzikanálový vzťah príznakov. Na získanie deskriptorov, teda na agregáciu priestorových informácií príznakovej mapy sa používajú operácie average-pooling aj max-pooling. Z poolingových operácií získané deskriptory $\boldsymbol{G^c_{avg}} \in \mathbb{R}^{C\times 1\times 1}$ a $\boldsymbol{G^c_{max}} \in \mathbb{R}^{C\times 1\times 1}$ sú následne vložené (po jednom) do zdieľaného viacvrstvového perceptrónu (MLP) s jednou skrytou vrstvou. Výsledná kanálová pozornostná mapa $\boldsymbol{M_c}$ je následné získaná sčítaním (po prvkoch) výstupov z MLP pre $\boldsymbol{G^c_{avg}}$ a $\boldsymbol{G^c_{max}}$.

Celý proces získania kanálovej pozornostnej mapy teda vieme zapísať nasledovne:

$$\begin{aligned} \boldsymbol{M_c} =& \sigma(MLP(AvgPool(\boldsymbol{G})) + MLP(MaxPool(\boldsymbol{G}))) \\ =& \sigma(\boldsymbol{W_1}(\boldsymbol{W_0}(\boldsymbol{G^c_{avg}})) + \boldsymbol{W_1}(\boldsymbol{W_0}(\boldsymbol{G^c_{max}}))); \end{aligned} \tag{3.7}$$

kde σ je funkcia sigmoid, $\boldsymbol{W_0} \in \mathbb{R}^{c/r \times c}$ a $\boldsymbol{W_1} \in \mathbb{R}^{c \times c/r}$.

SAM sa zameriava na vnútorný vzťah príznakov pozdĺž všetky kanály príznakových máp. Jeho cieľom je zistiť, kde (v rozmeroch $H \times W$) majú príznakové mapy najväčší vplyv.

Pri SAM, rovnako ako pri CAM, sú na vygenerovanie deskriptorov použité operácie average-pooling aj max-pooling, pričom v tomto prípade sa vykonávajú pozdĺž dimenzie kanála, a teda vygenerujú dve dvojrozmerné mapy (v dimenzii $H \times W$), teda $\boldsymbol{G^s_{avg}} \in \mathbb{R}^{1 \times H \times W}$ a $\boldsymbol{G^s_{max}} \in \mathbb{R}^{1 \times H \times W}$. Tieto dve mapy sú následne spojené, čím vznikne výsledná priestorová pozornostná mapa $\boldsymbol{M_s} \in \mathbb{R}^{1 \times H \times W}$.

Celý proces získania priestorovej pozornostnej mapy teda vieme zapísať nasledovne:

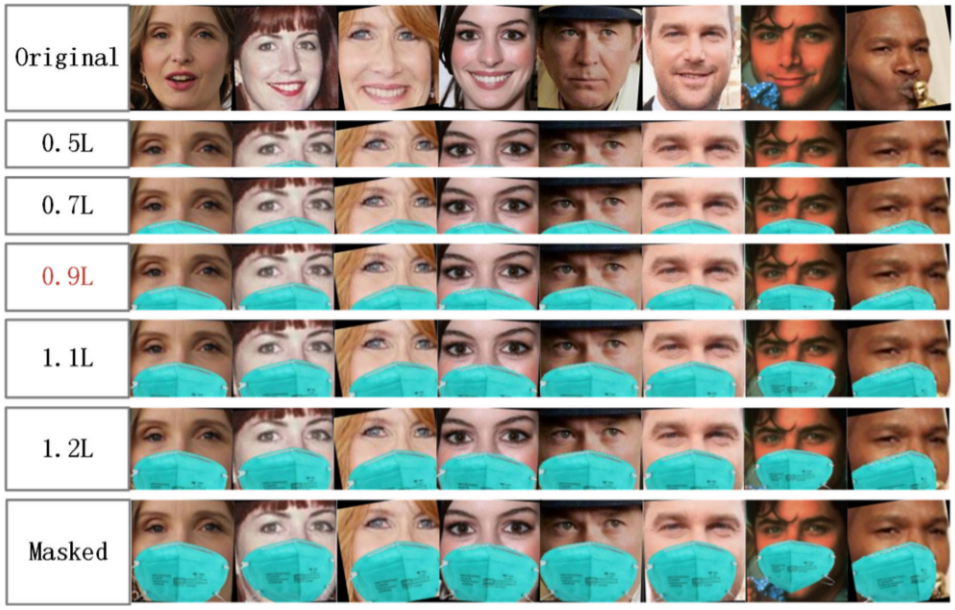
$$\begin{aligned} \boldsymbol{M_c} =& \sigma(f \otimes ([AvgPool(\boldsymbol{G}); MaxPool(\boldsymbol{G})]) \\ =& \sigma(f \otimes ([\boldsymbol{G^s_{avg}}; \boldsymbol{G^s_{max}}])); \end{aligned} \tag{3.8}$$

kde σ je funkcia sigmoid, $\otimes$ je konvolúcia a f je konvolučné jadro.

Získali sme všetko potrebné pre získanie pozornostnej mapy $\boldsymbol{G''}$, ktorá je výstupom CBAM.

Pri druhom spomenutom CB prístupe autori získajú kľúčové body (landmarky) z tváre prostredníctvom tvárového detektora MTCNN. Označme pozíciu kľúčového bodu ľavého oka ako $E_l(x_1, y_1)$ a prevého oka ako $E_r(x_2, y_2)$. Potom pozíciu medzi očami je $E_m((x_1 + x_2)/2, (y_1 + y_2)/2)$. Euklidovskú

vzdialenosť medzi kľúčovými bodmi očí zadefinujeme ako vzdialenosť $L = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$. Orezávací bod definujeme ako $C((x1+x2)/2, (y1+y2)/2+l)$, kde $l \subseteq [0.4L, 1.2L]$ je hyperparemeter určujúci proporcie orezania.



Obr. 3.11: Ukážka orezaných obrázkov pri rôznych proporciách orezania.

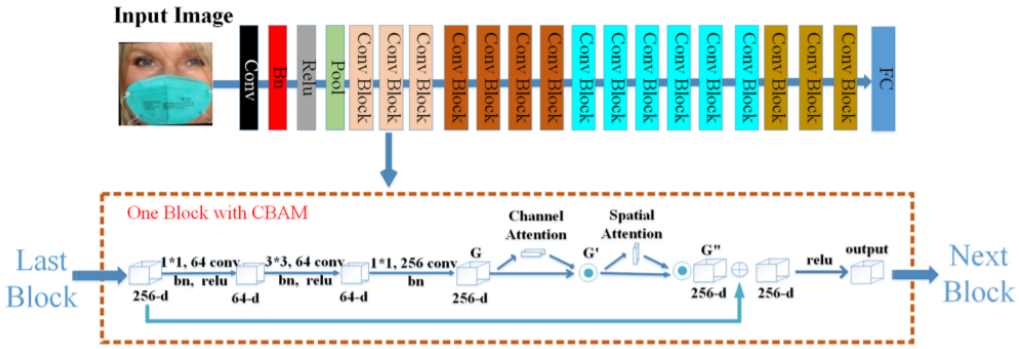
Následne obrázky orežieme cez orezávací bod paralelne s osou x, pričom ponecháme vrchnú časť z dvoch vzniknutých častí po orezaní, ako môžeme vidieť na obrázku 3.11.

- Pri AB prístupe motivujeme sieť zvýrazniť diskriminačné oblasti (oblasť očí) a potlačiť nedôležité oblasti (maska) na fotografiách.
- CB je navrhnutý na zlepšenie dvoch špecifických scenárov (Case 2, Case 3) z obr. 3.8.

Aby sme maximalizovali výhody oboch popísaných prístupov, integrujeme spolu:

1. Nájdeme optimálne orezanie (optimálnu hodnotu hyperparametra l pri CB), pričom pri trénovaní modelu použijeme nájdené optimálne orezanie na trénovacie vzorky.

2. AB integrujeme zapojením CBAM do každého konvolučného bloku architektúry resnet50, ako môžeme vidieť na obrázku 3.12.



Obr. 3.12: Framework Integrácie AB a CB pristúpov.

Z výsledkov vyplýva, že tento prístup vie zvýšiť presnosť identifikácie o 0.104% pri prípade 1 (MFR Case 1), 17.427% pri prípade 2 (MFR Case 2) a 18.507% pri prípade 3 (MFR Case 3). Optimálne orezanie je okolo $0.9L$ pri prípade 1 a $0.7L$ v prípadoch 2 a 3.

Kapitola 4

Predchádzajúce riešenia

Tu budú predchádzajúce riešenia.

Kapitola 5

Výskum

Tu bude výskum.

5.1 Trénovanie

5.2 Výsledky

Kapitola 6

Návrh

Tu bude Návrh

Kapitola 7

Implementácia

Tu bude Implementácia

Literatúra

[1] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[2] Qi Jin, Chong Mu, Ling Tian, and Fankai Ran. A region generation based model for occluded face detection. *Procedia Computer Science*, 174:454–462, 2020. 2019 International Conference on Identification, Information and Knowledge in the Internet of Things.

[3] Yande Li, Kun Guo, Yonggang Lu, and Li Liu. Cropping and attention based approach for masked face recognition. *Applied Intelligence*, 51(5):3012–3025, May 2021.

[4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[5] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[6] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[7] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001.

[8] Jianfeng Wang, Ye Yuan, and Gang Yu. Face attention network: An effective face detector for the occluded faces. *arXiv preprint arXiv:1711.07246*, 2017.

[9] Zhongyuan Wang, Guangcheng Wang, Baojin Huang, Zhangyang Xiong, Qi Hong, Hao Wu, Peng Yi, Kui Jiang, Nanxi Wang, Yingjiao Pei, et al. Masked face recognition dataset and application. *arXiv preprint arXiv:2003.09093*, 2020.

[10] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[11] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[12] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing.

[13] 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, 2016.

Zoznam obrázkov