VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

CHARAKTERIZACE CHODCŮ VE VIDEU

PEDESTRIAN ATTRIBUTE ANALYSIS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE **Bc. ZUZANA STUDENÁ**
AUTHOR

VEDOUCÍ PRÁCE **Ing. MICHAL HRADIŠ, Ph.D.**
SUPERVISOR

BRNO 2019

Vysoké učení technické v Brně
Fakulta informačních technologií

Ústav počítačové grafiky a multimédií (UPGM) Akademický rok 2018/2019

Zadání diplomové práce

Studentka: **Studená Zuzana, Bc.**
Program: Informační technologie Obor: Počítačová grafika a multimédia
Název: **Charakterizace chodců ve videu**
Pedestrian Attribute Analysis
Kategorie: Zpracování obrazu

Zadání:

1. Prostudujte současné metody a nástroje pro detekci osob, odhad pózy a extrakci informací o chodcích z obrazu.
2. Vytvořte si přehled o současných přístupech k analýze publika v kamerových systémech.
3. Navrhněte postup vhodný pro odhad informací o lidech se zaměřením na socioekonomickou identitu včetně vhodného postupu pro efektivní přípravu datové sady.
4. Obstarejte si datovou sadu vhodnou pro experimenty.
5. Implementujte navržený systém a proveďte experimenty nad datovou sadou.
6. Porovnejte dosažené výsledky a diskutujte možnosti budoucího vývoje.
7. Vytvořte stručné video prezentující vaši práci, její cíle a výsledky.

Literatura:

- Taigman et al.: DeepFace: Closing the Gap to Human-Level Performance in Face Verification. CVPR, 2014.
- Parkhi et al.: Deep face recognition. Proceedings of the British Machine Vision 1.3 (2015): 6.
- Zhe Cao et al.: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. CVPR, 2017.

Při obhajobě semestrální části projektu je požadováno:

- Body 1 až 3.

Podrobné závazné pokyny pro vypracování práce viz http://www.fit.vutbr.cz/info/szz/

Vedoucí práce: **Hradiš Michal, Ing., Ph.D.**
Vedoucí ústavu: Černocký Jan, doc. Dr. Ing.
Datum zadání: 1. listopadu 2018
Datum odevzdání: 22. května 2019
Datum schválení: 1. listopadu 2018

Abstrakt

Táto práca sa zaoberá získavaním informácií o chodcoch, ktorí sú zachytení pomocou statických vonkajších kamier umiestnených na verejných vonkajších alebo vnútorných priestranstvách. Cieľom je za použitia konvolučných neurónových sietí získať, čo najväčšie množstvo informácií ako je napríklad pohlavie, vek a typ oblečenia, doplnky, módny štýl alebo celková charakteristika osoby. Časť práce pozostáva z tvorby novej dátovej sady, ktorá zachytáva chodcov a k nim informácie o pohlaví, veku a módnom štýle osoby. Ďalšou časťou práce je návrh a implementácia konvolučných neurónových sietí, ktoré klasifikujú spomínané charakteristiky chodcov. Neurónové siete vyhodnocujú vstupné obrázky chodcov v dátových sadách PETA, FashionStyle14 a BUT atribúty chodcov. Vykonané experimenty nad dátovými sadami PETA a FashionStyle porovnávajú moje výsledky rôznych konvolučných neurónových sietí s publikáciami. Ďalšie experimenty sú ukázané na novo vytvorenej dátovej sade BUT atribúty chodcov.

Abstract

This work deals with obtaining pedestrian information, which are captured by static, external cameras located in public, outdoor or indoor spaces. The aim is to obtain as much information as possible. Information such as gender, age and type of clothing, accessories, fashion style, or overall personality are obtained using using convolutional neural networks. One part of the work consists of creating a new dataset that captures pedestrians and includes information about the person's sex, age, and fashion style. Another part of the thesis is the design and implementation of convolutional neural networks, which classify the mentioned pedestrian characteristics. Neural networks evaluate pedestrian input images in PETA, FashionStyle14 and BUT Pedestrian Attributes datasets. Experiments performed over the PETA and FashionStyle datasets compare my results to various convolutional neural networks described in publications. Further experiments are shown on created BUT data set of pedestrian attributes.

Kľúčové slová

rozpoznávanie obrazu, konvolučné neurónové siete, predtrénovanie nerónových sieti, fine-tuning, featrure extraction, resnet, klasifikácia atribútov, charakteristika chodcov, dátová sada, BUT atribúty chodocov, PETA, FashionStyle14

Keywords

image recognition, convolutional neural networks, transfered learning, feature extractor, fine-tuning, resnet, attribute classification, dataset, BUT atribúty chodocov, PETA, FashionStyle14

Citácia

STUDENÁ, Zuzana. *Charakterizace chodců ve videu*. Brno, 2019. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Michal Hradiš, Ph.D.

Charakterizace chodců ve videu

Prehlásenie

Prehlasujem, že som túto prácu vypracovala samostatne pod vedením pána Ing. Michala Hradiša Ph.D. Uviedla som všetky literárne pramene a publikácie, z ktorých som čerpala pri tvorbe tejto práce.

.....
Zuzana Studená
21. mája 2019

Poďakovanie

Týmto by som chcela poďakovať môjmu vedúcemu Ing. Michalovi Hradišovi Ph.D. za odbornú pomoc a vedenie počas celej tvorby práce. Zároveň by som sa chcela poďakovať spolužiakovi Jánovi Jurčovi, s ktorým sme spolupracovali na tvorbe nových dátových sád do našich prác. Ďakujem.

Obsah

Kapitola 1

Úvod

V súčastnosti sú kamery umiestnené na mnohých miestach verejných a súkromných budov, kde zaznamenávajú pohyb obyvateľstva. S vývojom počítačového spracovania obrazu prudko narastá spôsob využitia dát zachytených pomocou týchto kamier. V bezpečnostných a analytických odvetviach je dôležitá najmä identifikácia, reidentifikácia alebo charakteristika chodcov. Včasná analýza chodcov na verejných priestranstvách má veľký prínos pre bezpečnosť, kedy by mohli byť včas identifikované nebezpečné osoby alebo predmety a následne informované príslušné bezpečnostné zložky. Ďalším prínosom je možná automatická analýza osôb v nákupných centrách za účelom zvýšenia predaja a zlepšenia marketingu obchodov.

Rozpoznávanie a charakteristiku zhoršuje niekoľko problémov spojených so spracovaním obrazu chodcov zachytených z veľkej vzdialenosti. Veľká vizuálna podobnosť jednotlivých častí oblečenia (napr. existuje mnoho rôznych druhov krátkych nohavíc, ktoré môžu byť v istých uhloch podobné napríklad so sukňami), úroveň jasu a množstvo svetla v zachytenej scéne. Ďalšie problémy, ktoré je nutné brať do úvahy, sú napríklad veľké množstvo osôb zachytené na jednej snímke, ktoré sa vzájomne prekrývajú, posun pohybujúcich sa chodcov, uhol zachytenia kamerou. Tieto a iné problémy riešili Fabbri M. a kolektív v článku [7]. Nízke rozlíšenie obrázkov môže viesť k strate informácií o doplnkoch, ktoré drží alebo nosí osoba. V takýchto prípadoch zo zachyteného obrázku vôbec nie je možné určiť prítomnosť niektorých charakteristík. Príkladom môže byť obrázok, na ktorom je zachytená osoba z ľavej strany. Z obrazu nemožno rozhodnúť či daná osoba drží tašku v pravej ruke. V reálnej situácii ale vieme úplne presne určiť či osoba naozaj tašku má. Rozpoznávaním atribútov chodcov ako sú napríklad pohlavie, vek, oblečenie, doplnky, ktoré osoba nosí alebo predmety, ktoré drží, sa zaoberali Deng Y. a kolektív autorov v článku [4]. Cieľom článku bolo nielen vytvoriť rozsiahlu anotovanú dátovú sadu PETA, ale aj vyhodnotiť úspešnosť strojového učenia pri klasifikácii spomínaných atribútov.

Inou častou charakteristiky chodcov je rozpoznávanie a klasifikácia oblečenia a módneho štýlu. Narozdiel od klasifikácie atribútov popísaných vyššie, je kladený väčší dôraz na jednotlivé časti a farby oblečenia. Celkový štýl patrí medzi hlavné ukazovatele sociálnej príslušnosti človeka. Väčšina ľudí zaraďuje osoby do skupiny práve podľa štýlu, v ktorom sú oblečené. Osoba oblečená v športovom oblečení je posudzovaná ako športovec, muž v obleku manažér, uniforma je typická pre policajtov alebo vojakov.

Cieľom tejto práce je navrhnúť a implementovať spôsob klasifikácie rôznych charakteristík chodcov. Za charakteristiky chodcov sú v tejto práci chápané atribúty ako sú pohlavie, vek, jednotlivé časti oblečenia, doplnky alebo celkový štýl oblečenia. Charakteristiky sú vyhodnocované na obrázkoch, vytvorených z videí zachytených najmä statickými vnútornými

alebo vonkajšími kamerami. Riešenie má obsahovať implementovanú konvolučnú neurónovú sieť, ktorá dokáže klasifikovať jednotlivé charakteristiky chodcov, ako sú napríklad pohlavie, vek, oblečenie a módny štýl. Súčasťou je taktiež návrh a vytvorenie novej dátovej sady.

Kapitola 2 a kapitola 3 popisujú súčasné prístupy a metódy na identifikáciu a zisťovanie rozličných informácií o chodcoch riešené v rôznych publikáciach. Kapitola 2 popisuje riešenia, dátové sady a spôsoby vyhodnocovania v nich. V kapitole 3 sú popísané konkrétne konvolučné neurónové siete a princípy ich predtrénovania. Zber dát, vytvorenie a ukážky novej dátovej sady sú popísané v kapitole 4. Konkrétne konvolučné neurónové siete, ich návrh, úpravy a implementácia je popísaná v kapitolách 5, 6. Kapitola 6 navyše popisuje implementáciu webovej aplikácie vytvorenej pre zber anotácií k novo vytvorenej dátovej sade BUT atribúty chodcov. Experimenty sú ďalej zhrnuté popísané a vyhodnotené v kapitole 7.

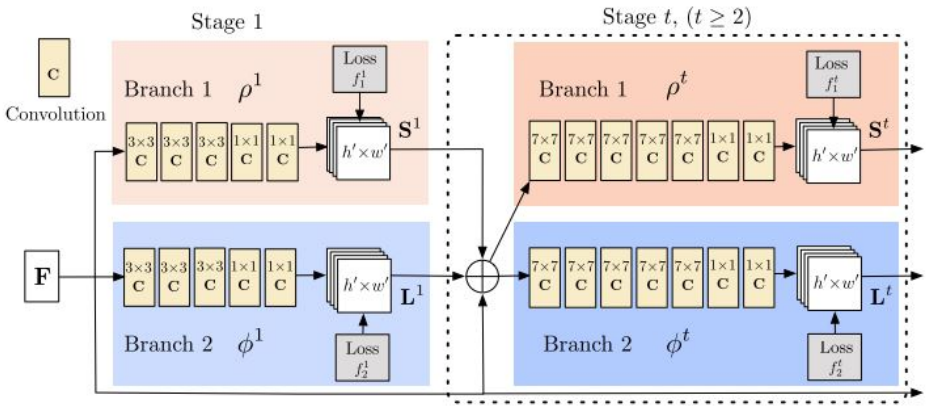
Kapitola 2

Rozpoznávanie atribútov osôb

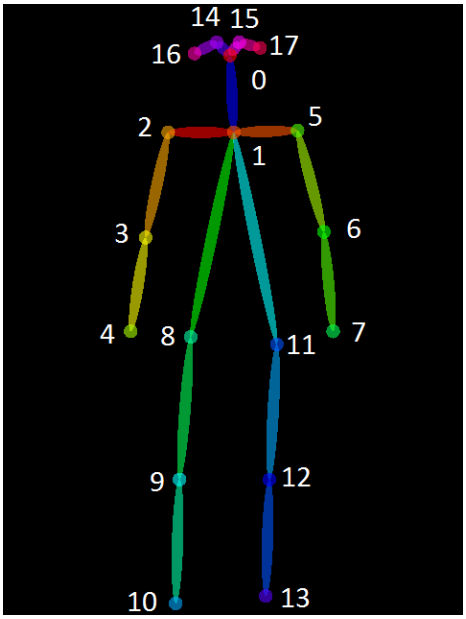
Zisťovanie charakteristík ako sú napríklad atribúty alebo časti oblečenia osôb je problém riešený v rôznych vedeckých publikáciách. V niektorých publikáciach sa zisťujú charakteristiky na obrázkoch z módnych dátových sád, v iných sú použité priamo zábery chodcov z vonkajších kamier. Publikácie obsahujú rôzne prístupy pre klasifikáciu rôzneho počtu atribútov ako sú napríklad pohlavie, vek, oblečenie, módny štýl, dĺžka a farba vlasov, doplnky a ďalšie. V tejto kapitole sú popísané niektoré z existujúcich riešení a princípov pri klasifikácií alebo získavaní informácií o osobe na obrázku. Prvá sekcia popisuje detekciu osôb v obraze. V ďalších sekciách sú popísané rôzne metódy klasifikácie atribútov a módnych štýlov na rôznych typoch obrázkov. Ako posledné sú spomenuté existujúce dátové sady a metriky hodnotenia úspešnosti klasifikácie.

2.1 Detekcia osôb a odhad pózy chodcov

Prvou časťou potrebnou pre získavanie charakteristík chodcov je detekcia osôb v obraze. Detekcia objektov a osôb v reálnom čase je dnes nevyhnutá pri rôznych úlohách. Najlepším príkladom je spolupráca človeka a strojov alebo riadenia inteligentných vozidiel. Princípy a metódy detekcie osôb a odhad pózy možno nájsť v článkoch [2, 17, 28]. V článku [2] popísali nástroj OpenPose, ktorý dokáže detekovať 2D pózy viacerých osob v jednom obraze a v reálnom čase. Predstavená metóda má na vstupe celý obrázok, ktorý je predaný konvolučnej neurónovej sieti s dvoma vetvami, ktoré spoločne vytvárajú mapu dôvery (*confidence map*), detekovaných častí tela a pole popisujúce závislosti medzi jednotlivými časťami tela (*part affinity fields*). Architektúra tejto konvolučnej siete navrhnutej v článku [2] je zobrazená na obrázku 2.1. Na základe máp dôvery sú detekované jednotlivé časti teľa. Medzi každou dvojicou detekovaných častí na končatine sa určí stredý bod (*midpoint*), a skontroluje sa, či sa určený bod nachádza medzi kandidátmi. Aby sa pri metóde stredového bodu predišlo spájaniu končatín medzi rôznymi osoba na jednom obrázku, je použité pole popisujúce závislosti medzi jednotlivými časťami končatiny. Toto pole tvorí 2D vektor obsahujúci informácie o polohe a orientácii každého páru pozostávajúceho z dvoch častí teľa detekovaných na obrázku. Spojením všetkých detekovaných častí algoritmus vypočíta polohy a pózu všetkých osôb na vstupnom obrázku. Jednotlivé časti môžu byť detekované v rôznom formáte. Jedným s používaných formátov je formát `COCO`, ktorý definuje sedemnásť častí, ktoré sú zobrazené na obrázku 2.2.



Obr. 2.1: Architektúra konvolučnej neurónovej siete použitá v publikácii [2], k detekcií pózy osôb.



Obr. 2.2: Formát súradníc častí tela COCO, prevzaté z [10].

2.2 Klasifikácia a rozpoznávanie atribútov chodcov

Klasifikácia charakteristík je riešená rôznymi spôsobmi [17, 18, 22, 28, 32] a na rôznych dátových sadách [4, 15, 32]. Pre tento účel osoby zachytené kamerami nie je potrebné dlhodobo sledovať. Pri takýchto záberoch sa nepredpokladá výrazná zmena atribútov na rôznych obrázkoch jednej osoby. Kvôli tohoto dôvodu, autori verejných dátových sád používali iba obrázky a nie celé video sekvencie. Obrázky v dátových sadách pozostávajú z výrezov chodcov, ktoré sú predávané na vstup klasifikátoru, ktorý klasifikuje atribúty do rôznych tried. Rôzne princípy klasifikácie sú popísané v nasledujúcich sekciách.

Klasifikačné algoritmy

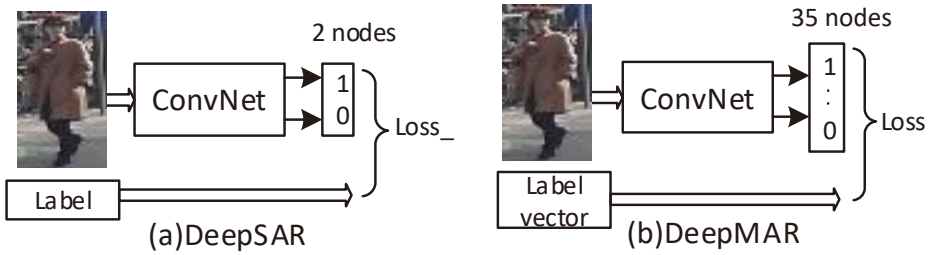
Jedna z prvých dátových sád pre klasifikáciu atribútov je APiS 1.0, popísaná v článku [32]. Dátová sada je bližšie rozoberaná v sekcii 2.4. Autor uvedeného článku používa dve metódy pre klasifikáciu binárnych atribútov a atribútov patriacich do viacerých tried. U binárnych atribútov extrahuje farbu a textúru pomocou 3,697 plávajúcich okien 66 rôznych veľkostí. Z každého okna sú extrahované farby pre MB-LBP (*Multi-Block Local Binary Pattern*) a HOG (*Histogram of Oriented Gradien*). Pre trénovanie používa algoritmus *Gentle AdaBoost*, inšpirovaný podľa práce [29] pre trénovanie klasifikátoru na rozpoznávanie tvárí. U atribútov patriacich do viacerých tried extrahuje farebné prvky a následne používa algoritmus *K Nearest Neighbors*.

Inou často používanou metódou pre klasifikáciu je SVM (*Support Vector Machine*) [20]. Jedná sa o algoritmus strojového učenia, ktorého cieľom je nájsť v n-dimenzionálnom priestore hyper-rovinu, ktorá najlepšie rozlišuje dve rôzne triedy. N reprezentuje počet vlastností objektu. Maji S. a.i. v práci [18] ukázal, použitie metódy pri klasifikácii chodcov na obrázkoch zo statických kamier. Podľa publikácie [13] bola metóda SVM použitá pre klasifikáciu atribútov chodcov, čo bolo ďalej použité pre ich reidentifikáciu. Vektory vlastností (*Feature vectors*) reprezentujúce chodca boli zostrojené z každého obrázku dátovej sady. Ten bol rozdelený na 6 horizontálnych pásov, dva pre hlavu, štyri pre hornú, dolnú časť tela a nôh. Vektory boli vytvárané z ôsmych farieb a dvadsať jedna textúrových filtrov. Spôsob použitý pre klasifikáciu atribútov pomocou SVM bol priamo prevzatý z článku [4]. Pre zlepšenie inferencie atribútov použili autori dve optimalizácie metódy MRF (*Markov Random Field*): Gaussian Kernel MRF a Random forest MRF.

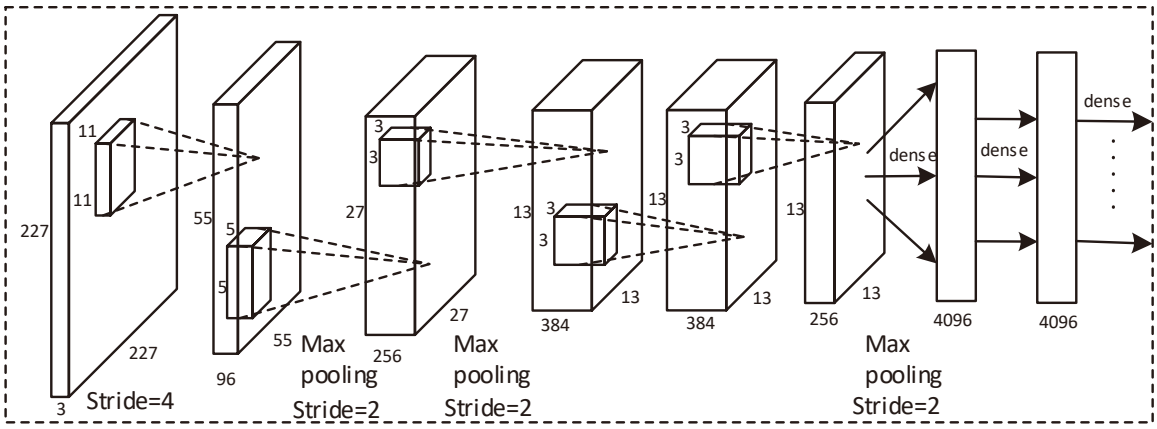
Ďalšie vylepšenie bolo založené na rozpoznávaní atribútov iba zo siluety chodca. Cieľom bolo oddeliť osobu od pozadia za účelom zistenia vplyvu pozadia na klasifikáciu atribútov. V publikácii [5] bola trénovaná hlboká neurónová sieť pre rozlíšenie chodca od pozadia. Vstupom siete bol vektor vlastností zostrojený z pôvodného obrázka a výstupom bola mapa častí tela. Každá vrstva bola plne prepojená z predchádzajúcou vrstvou. Sieť je podrobne popísaná v publikácii [28]. Po detekcii pozadia a osoby boli vytvorené vektory vlastností z celého obrázku, detekovanej osoby a pozadia obrázku. Ďalšie vektory boli vytvorené zreťazením vektoru osoby s vektorom pozadia a vektoru osoby s vektorom celého obrázku. Pre klasifikáciu atribútov boli použité rovnako metódy SVM, Gaussian Kernel MRF a Random forest MRF. V oboch vyššie uvedených publikáciách boli experimenty vykonané na dátovej sade PETA, ktorá je bližšie popísaná v sekcii 2.4. Výsledky ukazujú rôznu úspešnosť určenia atribútov, kde najmenšiu úspešnosť mali atribúty ako náhrdelník alebo sandále. Dôvodom bola nízka početnosť výskytov týchto atribútov v dátovej sade. Pomerne vysokú úspešnosť mal atribút vek a klobúk. Tiež bolo ukázané, že pozadie obrázku hrá dôležitú rolu pri určovaní atribútov: ruksak, igelitová taška, držanie predmetu. Naopak pozadie nemalo žiadny vplyv na atribúty týkajúce sa oblečenia. V oboch článkoch dosiahla najväčšiu úspešnosť metóda Random forest MRF.

Konvolučné neurónové siete

Konvolučné neurónové siete (KNN) sú často používané v rôznych klasifikačných úlohách. V publikáciach [14, 15, 22] sa ukázalo použitie KNN ako dobrý spôsob pre klasifikáciu atribútov chodcov zameraných stacionárnymi kamerami. V tejto časti sú popísané niektoré modely konvolučných neurónových sietí a optimalizácie, ktoré boli používané autormi publikácií pre klasifikáciu chodcov.



Obr. 2.3: Rozdiel modelov SAR a MAR. Sieť SAR na obrázku (a) má len dva výstupné uzly. Sieť (b) má počet výstupných uzlov rovný počtu klasifikovaných atribútov. Prevzaté z článku [14].



Obr. 2.4: Architektúra konvolučnej neurónovej siete použitá pre klasifikáciu atribútov chodcov v publikácii [14].

SAR (*Single attribute Recognition*) a MAR (*Multi-attribute recognition*) [14] sú modely neurónovej siete navrhnuté pre klasifikáciu binárnych atribútov chodcov. V modely SAR je každý atribút klasifikovaný samostatne jednou sieťou, zatiaľ čo v druhom modely sú všetky atribúty klasifikované súčasne jednou konvolučnou neurónovou sieťou. Jadro oboch modelov tvorí konvolučná neurónová sieť. V článku je použitá konvolučná neurónová sieť obsahujúca päť konvolučných vrstiev s troma plne prepojenými vrstvami. Každú konvolučnú vrstvu nasleduje neliarita ReLu. Po prvých dvoch ReLu nasleduje max-pooling a normalizácia. Max-pooling je tiež umiestnený za piatym Relu. Archiutektúra siete je zobrazená na obrázku 2.4. Rozdiel modelov je vo vstupe, poslednej výstupnej vrstve a použitých chybových funkciách. U siete SAR je vstupom klasifikovaný obrázok a trieda atribútu. Výstupná vrstva má dva uzly, ktoré odpovedajú pravdepodobnosti či daný obrázok patrí do danej triedy. Pre každý atribút bol model predtrénovaný sieťou CafeeNet [6]. Pre predtrénovanie bol použitý Fine-tunning[1]. Pre výpočet výstupu bola použitá logistická funkcia Softmax. Softmax bol tiež použitý pri výpočte chyby. Konkrétne rovnice chybovej a pravdepodobnostnej funkcie sú popísané v dokumente [14]. U modelu MAR je vstupom klasifikovaný obrázok a vektor všetkých atribútov. Výstupná vrstva má n uzlov, kde n odpovedá dĺžke vektoru atribútov na vstupe. Pre výpočet chybovej funkcie bola použitá kombinácia logistickej funkcie Sigmoid a aktivačnej funkcie Cross entropy. Konkrétne funkcie pre výpočet

[1]**Fine-tuning:** spôsob predtrénovania siete kedy, sú nastavené parametre modelu z inej natrénovanej siete, zvyšok učenia prebieha obvyklým spôsobom.

chyby a pravdepodobnosti atribútu na výstupe sú popísané v článku [14]. Rozdiel sietí je znázornený na obrázku 2.3. Výsledky experimentov s modelmi SAR a MAR ukázali, že použitie konvolučných neurónových sietí pre klasifikáciu atribútov chodcov dosahuje u väčšiny atribútov lepšiu presnosť ako metóda MRF (*Markov Random Field*). Omnoho lepšie výsledky klasifikácie dosiahli na atribútoch, ktoré mali nízky výskyt naprieč celou dátovou sadou (napr. okuliare, sandále, V-výstrih atď.). Zároveň na týchto atribútoch ukázal model SAR lepšie výsledky ako MAR. V ostatných prípadoch pracoval model MAR efektívnejšie.

ACN (*Attributes Convolutional Net*) [23] klasifikuje všetky atribúty súčasne v jednom modely. Váhy v modely sú zdielané, čo umožňuje efektívne prenášať vlastnosti. Pre každý klasifikovaný atribút existuje vlastná chybová funkcia, ktorej vypočítaná chyba sa akumuluje s ostatnými chybami počas spätného šírenia chyby. Architektúra siete je založená na sieti *CaffeNet*, z ktorej bola odstránená posledná vrstva, ktorú nahradilo niekoľko ďalších plne prepojených vrstiev. Pre každú výstupnú vrstvu sa počíta vlastná chybová funkcia a počet výstupných vrstiev je závislý na počte klasifikovaných atribútov. Sieť bola predtrénovaná na dátovej sade ImageNet. Počas trénovacej fázy je trénovaná celá sieť a predtrénovanie siete slúži len na jej inicializáciu. Obrázky predávané na vstup modelu boli zmenené na rozmer 256×256px., následne boli náhodne škálované na rôzne rozlíšenia, a boli pridané ich zrkadlové zobrazenie. Autori, ktorí použili ACL model v publikácii [23] sa rozhodli zaviesť značku N/A. Značka N/A značí, že z daného obrázku nemožno rozhodnúť, či sa na ňom daný atribút nachádza. Cieľom bolo správe vyhodnotiť stavy, kedy je chodec zachytený napríklad iba z jedného uhlu a tým pádom nemôžme spoľahlivo určiť výskyt atribútu na odvrátenej strane. Väčšina experimentov túto triedu nezahŕňa, pretože sú všetky atribúty klasifikované samostatne pomocou binárnych klasifikátorov. Autori v článku vytvorili vlastný dataset, na ktorom ukázali výsledky tohoto modelu.

Kombinované modely

Okrem vyššie spomínaných metód existujú ďalšie komplexnejšie spôsoby pre klasifikáciu atribútov chodcov. Väčšina z nich kombinuje rôzne spôsoby a princípy pre získavanie väčšieho množstva informácií o klasifikovanom obrázku. Robustnejšie riešenia dosahujú lepšie výsledky v prípadoch, kedy pracujú s väčším množstvom informácií ako napríklad so vzťahmi medzi atribútmi alebo predspracovávajú informácie iným spôsobom ako počas trénovania neurónovej siete. Chen Y. a ostatní navrhli v článku [3] model konvolučnej neurónovej siete, v ktorej sú do fázy trénovania pridané LOMO vlastnosti (Local Maximal Occurrence features), čím dosiahli lepších výsledkov klasifikácie atribútov chodcov ako modely trénované bez nízko-úrovňových vlastností. V článku [30] je popísaná metóda JRL (Joint Recurrent Learning), ktorá kombinuje koreláciu atribútov s vonkajším kontextom osoby na obrázku. Hlavnú časť tvorí architektúra RNN (Recurrent Neural Network). RNN bola špeciálne navrhnutá pre sekvenčnú predikciu atribútu chodcov. Podľa autorov by architektúra RNN mala dosahovať lepších výsledkov na menších dátových sadách alebo nekvalitných obrázkoch. V experimentoch sú porovnané výsledky modelu JRL s modelmi popísanými vyššie v tejto kapitole. Spôsoby použité v metóde JRL dosahujú najlepšie výsledky medzi popisovanými metódami.

2.3 Klasifikácia oblečenia a módneho štýlu

Špecifikácia oblečenia a módneho štýlu pomocou umelej inteligencie patrí medzi netriviálne úlohy. Samotné rozpoznávanie jednotlivých častí oblečenia je problém, ktorý je riešený rôznymi metódami ako sú napríklad klasifikácia, ktorá bola popísaná v predošlých sekciách alebo segmentácia [31]. Zaradenie oblečenia k módnemu štýlu vyžaduje podrobnejšie znalosti o oblečení a doplnkoch, ktoré osoba nosí. Ani manuálnym posudzovaním oblečenia, rôzne osoby nepriradia jednému obrázku rovnaký módny štýl. Módny štýl sa často líši na základe osobnej preferencie anotujúcej osoby. Pre dokonalé riešenie tohoto problému nestačí aby sa učiace algoritmy naučili klasifikovať oblečenia ale musia sa naučiť ako rozličné osoby vidia rôzny módny štýl.

Jedným z najväčších problémov pri strojovom učení klasifikácie módneho štýlu sú dátové sady. Pre dosiahnutie najlepšieho výsledku je nutné veľké množstvo správne anotovaných dát. Väčšina publikácií používa módne dátové sady ako sú FashionStyle14 [27, 16]. V článku [27], vytvorili dátovú sadu FashionStyle14, na ktorej klasifikovali štrnásť rôznych módnych štýlov. Dátová sada FashionStyle14 je popísaná v sekcii 2.4. Autori článku sa zároveň zamerali na porovnanie spôsobu klasifikácie módneho štýlu určeného expertami na módny štýl, ľudí z priemernou znalosťou módy a vytvorenej neurónovej siete. Pre experimenty s vytvorenou dátovou sadou autori použili klasifikačnú neurónovú sieť. Súčasťou experimentu bolo porovnanie rozličných modelov siete. Porovnávané boli modely VGG16, VGG18, Inception v3, ResNet50 a Xceotion. U všetkých modelov boli inicializované váhy siete na hodnoty siete, ktorej model bol trénovaný na ImageNet pre klasifikáciu obrázkov do 1000 rôznych tried. Následne pre optimalizáciu váh, autori použili fine-tuning 3.2 a metódu SGD. Koeficient učenia bol nastavený pre každú architektúru zvlášť, v rozmedzí od 10^{-4} po 10^{-6}. Najlepšie výsledky boli dosiahnuté použitím architektúry siete ResNet50. Okrem experimentov, pre zistenie úspešnosti klasifikácie rôznymi architektúrami siete, sa autori článku [27] zaoberali kvalitatívnou analýzou. Skúmali, ktoré časti obrázku a oblečenia majú najväčší vplyv na výslednú triedu. Použili metódu podľa článku [8], pomocou ktorej získali masku regiónov záujmu pre klasifikáciu. Výsledky ukázali napríklad vplyv slnečných okuliarov na rockový štýl alebo flitrovanú sukňu na zaradenie k módnemu štýlu lolita.

Pre experiment porovnávajúci spôsob klasifikácie ľudí a neurónovej site, použili kvalitatívnu analýzu pomocou Normalized Mutual Information (NMI) skóre. Táto analýza ukázala veľký rozdiel v chybách, ktorých sa dopúšťali pri klasifikácii bežní užívatelia a neurónová sieť. Výsledky, ktoré klasifikovala neurónová sieť boli viac založené na textúre alebo farbe. V niektorých prípadoch dosiahla neurónová sieť správnych výsledkov a užívatelia klasifikovali nesprávny módny štýl. Inokedy väčšina užívateľov zvolila správny módny štýl a neurónová sieť zvolila nesprávne.

Rozlišovať štýl oblečenia sa pokúšal aj S. Guang-Lu a ostatní v publikácii [24]. Na vlastnej dátovej sade klasifikovali päť tried módneho štýlu, ktoré boli určené na základe štrnástich módnych atribútov. Obrázky dátovej sady pozostávali z fotografií zachytených v prírodnom prostredí. Metóda pre klasifikáciu štýlu popisovaná v tejto publikácii pozostáva z troch krokov. V prvom kroku rozpoznali jednotlivé časti oblečenia kombináciou priestorových informácií, detekciou najvýraznejších oblastí a super-pixel metódou [1]. Druhý krok mal detekovať módne atribúty, ktoré si autori definovali pre každý z piatich klasifikovaných módnych štýlov. Napríklad pre triedu Punk, boli typické atribúty ako koža, vybíjané ozdoby alebo písmená. Pre triedu Sexy, leopardie vzory, kožušina alebo čipky. V tretom kroku pomocou štatistickej analýzy vypočítali korelačnú maticu. Vzorec, ktorý použili pre výpočet

je popísaný v publikácii [24]. Pri klasifikácii módneho štýlu dosiahli úspešnosť priemerne 72%, čím ukázali, že navrhnutým postupom možno rozlišovať módne štýly.

2.4 Dátové sady

Táto sekcia obsahuje popis existujúcich dátových sád, ktoré sú často používané v publikáciách pre trénovanie a testovanie klasifikátorov atribútov chodcov. Každá dátová sada obsahuje rôzny počet obrázkov a anotácií, ktoré sú ukázané a popísané v nasledujúcich sekciách.

APiS 1.0 dátová sada bola predstavená v článku [32]. Je tvorená 3,661 obrázkami zo stacionárnych vnútorných a vonkajších kamier. Popisuje 11 binárnych atribútov a 2 atribúty patriace do viacerých tried. Obrázky boli zozbierané z viacerých zdrojov a obsahujú osoby rôzneho pohlavia a veku zachytené z rôznych uhlov. Anotácia atribútov bola vytváraná manuálne a berie ohľad na podobnosť jednotlivých atribútov (napr. osoba z dlhými vlasmi a sukňou bude s veľkou pravdepodobnosťou žena). Bližší popis jednotlivých atribútov je popísaný v publikácii [32].

PETA dataset [4] bola zložená z desiatich menších verejne dostupných existujúcich dátových sád. Celú dátovú sadu tvorí 19,000 obrázkov s rozlíšením od 17×37 do 169×365 pixelov, zachytených vo vonkajšom aj vnútornom prostredí pomocou stacionárnych kamier. Pri tvorbe dátovej sady boli odstránené duplicitné obrázky a bolo anotovaných 61 binárnych atribútov (napr. pohlavie, vek, štýl oblečenia, dĺžka vlasov a iné) a 4 viactriedne atribúty pre farbu topánok, spodnej resp. vrchnej časti odevu a vlasov. Distribúcia väčšiny binárnych atribútov je rovnomerná. Tabuľka 2.1 zobrazuje datasety a početnosť obrázkov, z ktorých bola vytvorená celá dátová sada PETA. Ukážky z dátovej sady sú zobrazené na obrázku 2.5.

dátová sada	obrázky	uhol kamery	uhol pohľadu	osvetlenie	rozlíšenie	scéna
3DPes	1012	vysoký	rôzny	rôzne	od 31×100 do 236×178	vonkajšia
CAVIAR4REID	1220	pozemný	rôzny	slabé	od 17×39 do 72×141	vonkajšia
CUHK	4563	vysoký	rôzny	rôzne	80×160	vonkajšia
GRID	1275	rôzny	predný/zadný	slabé	od 29×67 do 169×365	vnútorná
i-LIDS	477	stredný	zadný	vysoké	od 32×76 do 115×294	vnútorná
MIT	888	pozemný	zadný	vysoké	64×128	vonkajšia
PRID	1134	vysoký	predný	slabé	64×128	vonkajšia
SARC3D	200	stredný	rôzny	rôzne	od 54×187 do 150×307	vonkajšia
Town center	6967	stredný	rôzny	stredné	od 44×109 do 148×332	vonkajšia
VIPeR	1264	pozemný	rôzny	rôzne	48×128	vonkajšia
Spolu = PETA	19000	rôzny	rôzny	rôzne	rôzne	zmiešaná

Tabuľka 2.1: Popis dátovej sady PETA. Prevzaté z článku [4].

Obr. 2.5: Ukážka obrázkov z dátovej sady PETA [4].

Trieda	Atribúty
Dočasná	čas, id scény, pozícia obrázku, ohraničenie tela/ hlavy a ramien/ horná časť tela/dolná časť tela/doplnky
Celková	pohlavie, vek, postava, rola
Doplnky	ruksak, taška cez plece, nákupná taška, plastová, papierová taška ...
Póza a akcia	uhol pohľadu, telefonovanie, rozprávanie, tlačenie, nosenie ...
Prekrývanie	prekrývajúce sa časti
Hlava	účes, farba vlasov, okuliare
Horná časť tela	oblečenie a jeho farba
Dolná časť tela	oblečenie a jeho farba, štýl topánok a ich farba

Tabuľka 2.2: Popis atribútov anotovaných v dátovej sade RAP. Prevzaté z článku [15].

RAP (*Richly Annotated Dataset*) [15] je momentálne najväčšia vytvorená dátová sada obsahujúca zábery chodcov. Obsahuje až 41,585 záberov z 26 vnútorných kamier. Obrázky chodcov sú rovnomerne distribuované z rôznych uhlov a rôznej výšky. V dátovej sade je anotovaných 72 rôznych atribútov (69 binárnych, 3 viactriedne atribúty). Anotácie sú delené do šiestich skupín: celé telo, hlava, horná/dolná časť tela, doplnky, akcia, prekrývajúce časti. Všetky atribúty sú zhrnuté v tabuľke 2.2. Oproti ostatným dátovým sadám je anotovaný aj uhol, z ktorého bol chodec zachytený, natočenie hlavy, prekrytie s iným predmetom alebo chodcom, alebo ohraničujúci obdĺžnik pre 3 časti tela. Zaujímavou častou je anotácia činnosti chodca (chôdza, rozhovor atď.) alebo rola (zákazník, predajca).

FashionStyle14 [27] je jedna z mála dátových sád podrobne klasifikujúcich módny štýl. Anotácie dátovej sady boli založené na ohodnotení expertov orientujúcich sa v módnych trendoch. Pri tvorbe sa autori zamerali na prirodzené obrázky, na ktorých je zreteľne vidno celý outfit osoby. Samotná dátová sada klasifikuje štrnásť módnych štýlov (konzervatívny, elegantný, etnický, rozprávkový, ženský, dievčenský, večerný, ležérny, lolita, módny, prirodzený, retro, rockový a street). Pre každú triedu bolo ručne zozbieraných približne 1000 obrázkov, na ktorých mali osoby oblečené typické prvky zobrazovaného módneho štýlu. Celkovo tak zozbierali pre všetky triedy 13,126 obrázkov. Ukážka obrázkov z tejto dátovej sady je zobrazená na obrázku 2.6. Najväčším problémom tejto sady je rozloženie a výber obrázkov. Osoby na obrázkoch v dátovej sade sú iba ženského pohlavia a väčšina obrázkov zachytáva modelky v neprirodzených pózach. S týchto dôvodov nebude možné použiť dá-

Obr. 2.6: Ukážka obrázkov z dátovej sady FashionStyle14 [27]. Na prvom obrázku je konzervatívny štýl na druhom elegantný, na tretom rockový a na poslednom štýl lolita.

tovú sadu na trénovanie klasifikácie módneho štýlu, na záberoch reálnych osôb zachytených vonkajšími kamerami.

2.5 Metriky hodnotenia úspešnosti

Pri klasifikácií atribútov na rôznych dátových sadách je nutné zvážiť spôsob počítania úspešnosti siete. Je potrebné určiť metriky pre výpočet úspešnosti jednotlivých atribútov, úspešnosti všetkých atribútov v rámci jedného obrázku a zároveň úspešnosť a presnosť celej siete. V článku [15] uviedli metriky 2.1 a 2.2, ktoré by mali konzistentne vyhodnocovať viaceré atribúty na jednom obrázku. Tieto metriky pozostávajú z rovníc pre výpočet celkovej úspešnosti (accuracy), presnosti (precision), úplnosti (recall) a F1 skóre (F1 score). V rovniciach

$$acc = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i + FN_i}, \qquad prec = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i} \tag{2.1}$$

$$rec = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i}, \qquad F1score = \frac{2 * prec * rec}{prec + rec}, \tag{2.2}$$

N značí počet prvkov (obrázkov), TP_i (*true positive*) je počet správne klasifikovaných atribútov obrázku i, FP_i (*false positive*) je počet atribútov, ktoré boli vyhodnotené pozitívne ale nenachádzali sa na obrázku a FN_i (*false negative*) počet atribútov, ktoré sa nachádzali na obrázku ale sieť ich nerozpoznala. Tieto metriky sú vypočítavané zo všetkých atribútov v jednom obrázku a berú v úvahu vnútornú koreláciu atribútov, ktorá sa vyskytuje napríklad u dátových sád, ktoré anotujú veľké množstvo atribútov ako napríklad PETA alebo RAP.

Ak nepotrebujeme brať do úvahy, vnútornú koreláciu atribútov je možné použiť strednú presnosť (*mean accurency*)

$$mA = \frac{1}{2N}\sum_{j=1}^{L}(TP_j/P_j + TN_j/N_j), \tag{2.3}$$

kde L je počet atribútov, TP_j, P_j počet správne klasifikovaných atribútov a počet atribútov na obrázku a TN_j, N_j počet správne negatívne klasifikovaných atribútov a celkový počet

negatívnych atribútov. Rovnica zvlášť vypočítava presnosť pozitívne a negatívne ohodnotených atribútov. Stredná presnosť je potom vypočítaná ako priemer týchto dvoch presností pre všetky atribúty. Tento vzťah je dobre použiť na dátových sadách s nevyváženým pomerom pozitívne a negatívne ohodnotených atribútov. Dátová sada PETA obsahuje veľké množstvo binárnych atribútov, z ktorých nie je väčšina prítomná na obrázku.

Druhá možnosť ako počítať presnosť jednotlivých atribútov je vypočítať priemernú presnosť správne klasifikovaného atribútu naprieč celou dátovou sadou, ktorá obsahuje N prvkov. Pre tento výpočet možno použiť vzorec

$$acc_{attr} = \frac{1}{N}\sum_{i=1}^{N} Y_i, \qquad Y_i = \begin{cases} 1, & attr_{expected} = attr_{predicted} \\ 0, & \text{otherwise} \end{cases} \tag{2.4}$$

Kapitola 3

Konvolučné neurónové siete a trénovanie

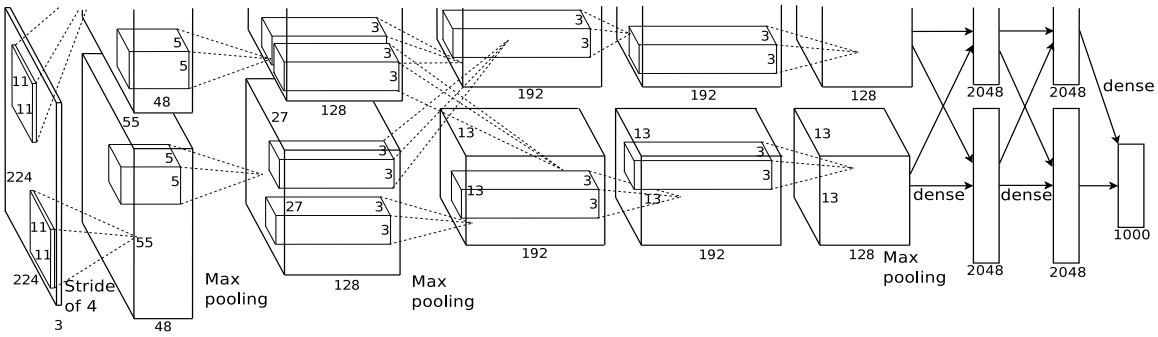
V predchádzajúcich sekciách boli zhrnuté niektoré metódy pre klasifikáciu módneho štýlu a atribútov chodcov. Cieľom tejto kapitoly je popísať existujúce možnosti klasifikácie, pomocou rôznych architektúr konvolučných neurónových sietí a spôsobov trénovania sietí, ktoré sa používajú pri riešení podobných problémov.

3.1 Architektúry konvolučných neurónových sietí

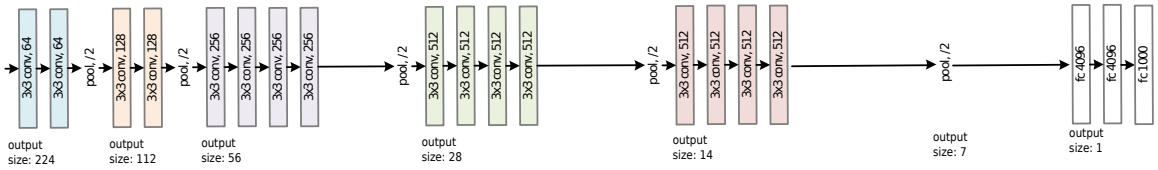
Väčšina moderných architektúr konvolučných neurónových sietí sa riadi rovnakými princípmi. Postupne používajú konvolučné vrstvy, zmenšujú rozmery vstupu a zároveň zvyšujú počet funkčných máp (*feature maps*). V sekcii sú popísané často používané architektúry, ktoré sa používajú pre detekciu objektov, klasifikáciu, segmentáciu a iné. Všetky architektúry popísané v tejto sekcii boli prvotne zostrojené a trénované na dátovej sade ImageNet [12], ktorá celá obsahuje viac ako 15 miliónov označených obrázkov, ktoré patrili do viac ako 22,000 kategórií. Zo všetkých obrázkov bolo pre trénovanie týchto sietí vybraná podmnožina 1000 tried a pre každú triedu 1000 obrázkov.

AlexNet [12] je jednou s prvých architektúr hlbokých, konvolučných, neurónových sietí. Sieť pozostáva z ôsmich vrstiev, päť z toho je konvolučných a tri plne prepojené vrstvy. Normalizačné bloky sú nasledované Max pooling vrstvou a sú umiestnené za prvú a druhú konvolučnú vrstvu. Max pooling nasleduje tiež piatu konvolučnú vrstvu. Na výstup každej konvolučnej a plne prepojenej vrstvy je aplikovaná nelinearita ReLU. Predstaviteľ Alex Krizhevsky bol jeden z prvých, ktorý použili ReLu namiesto funkcie Sigmoid alebo Tangens. Ukázal, že sieť sa použitím ReLu učí omnoho rýchlejšie ako použitím iných dovtedy používaných aktivačných funkcií. Architektúra je zobrazená na obrázku 3.1.

VGG predstavila dvojica autorov K. Simonyan a A. Zisserman v práci [21] z roku 2014. V publikácii sú popísané rôzne konfigurácie siete používajúce od 11 po 19 váhových vrstiev. Na počte váhových vrstiev závisí počet konvolučných vrstiev. Niektoré z konvolučných vrstiev sú nasledované max-poolingom, pričom po každom max-poolingu je zväčšený počet kanálov dvojnásobne, v rozmedzí od 64 po 512 kanálov. Konvolučné vrstvy sú nasledované troma plne prepojenými vrstvami, za ktorými zvyčajne nasleduje Softmax. Bližší popis



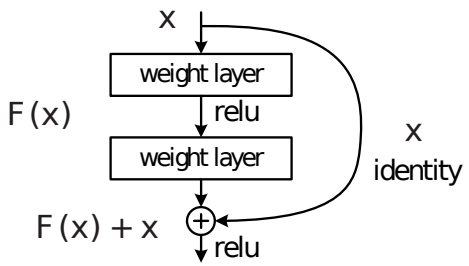
Obr. 3.1: Architektúra siete AlexNet. Prevzaté z publikácie [12].



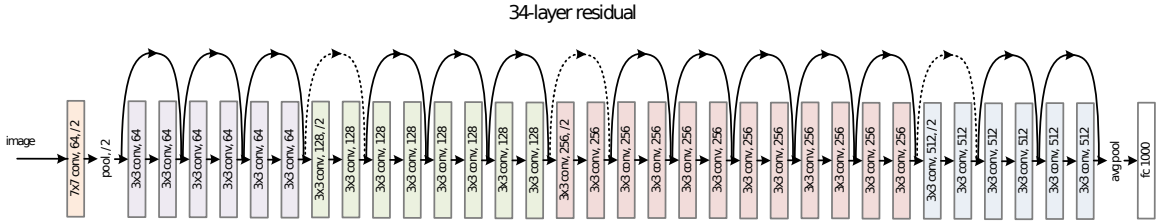
Obr. 3.2: Architektúra siete VGG-19. Prevzaté z publikácie [9].

architektúry, počet parametrov siete a nastavenie konfigurácií sú popísané priamo v publikácii [21].

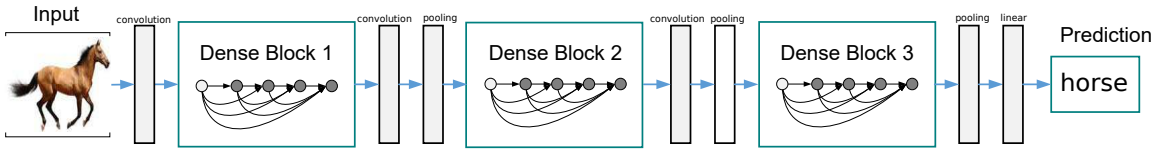
ResNet [9] je jedna z architektúr modernej hlbokej neurónovej siete. Existuje niekoľko rôznych variantov siete, ktoré majú rôznu hĺbku (ResNet18, ResNet34, ResNet50, ResNet152). Pri takto hlbokých neurónových sieťach sa stáva, že aj napriek lepšej inicializácii vstupných parametrov a použitiu batch normalizácie konvergujú k vyšším chybám ako je to v sieťach s menším množstvom vrstiev. V článku [9] autori predstavili *reziduálny blok*, obrázok 3.3, ktorý rieši tento problém. Jedná sa o doprednú neurónovú sieť, ktorá pridáva takzvané skratky (*shortcut connections*), tieto skratky sú interpretované ako mapovanie identity, ktorých výstup je spočítaný s výstupom neurónovej siete. Napriek pridaniu týchto spojení, môže byť sieť stále trénovaná pomocou metódy SGD a spätného šírenia chyby. Základná architektúra siete je prevzatá zo siete VGG. Počet vrstiev závisí na type siete. Reziduálne bloky sú použité na všetkých miestach siete, kde nedochádza k zmene dimenzie. V prípade zmeny dimenzie je možné doplniť chýbajúcu dimenziu nulovým vstupom alebo použitím lineárnej projekcie. Riešenia sú podobne popísané vo vyššie uvedenom článku. Architektúra siete ResNet34 je zobrazená na obrázku 3.4



Obr. 3.3: Reziduálny blok použitý v ResNet sieťach. Prevzaté z článku [9].



Obr. 3.4: Architektúra ResNet34 siete. Prevzaté z článku [9].



Obr. 3.5: Architektúra DenseNet siete. Prevzaté z článku [11].

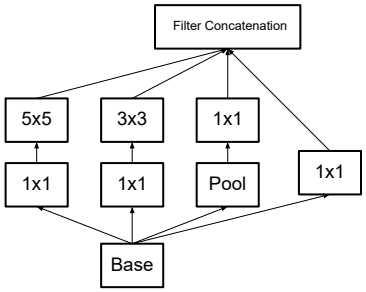
DenseNet [11] je podobná architektúra siete ako ResNet. Oproti ResNet však dosahuje lepší výkon s menšou zložitosťou a hĺbkou siete. Medzi hlavné výhody DenseNet architektúry patrí schopnosť zmenšiť problém miznúceho gradientu, podporenie šírenia vlastností a výrazné zníženie počtu parametrov siete. Sieť pozostáva z niekoľkých zložených blokov označených ako $H_l(\cdot)$, kde l značí index vrstvy siete. $H_l(\cdot)$ predstavuje nelinárnu transformáciu zloženú z operácií Batch Normalizacia, ReLu, Pooling alebo konvolúcia. V sieti ResNet je na vstup každej nasledujúcej vrstvy privedený výstup a zároveň aj vstup predošlej vrstvy. V sieti DenseNet sú pridané spojenia každej vrstvy do všetkých nasledujúcich vrstiev tak ako je ukázané na obrázku 3.5. L-tá vrstva dostáva na vstup mapy vlastnosti (*feature maps*) všetkých predošlých vrstiev:

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]), \tag{3.1}$$

kde $[x_0, x_1, \ldots, x_l - 1]$ predstavuje konkatenáciu máp vlastností vo vrstvách $0, \ldots, x_{l-1}$.

Inception alebo GoogLeNet je sieť predstavená spoločnosťou Google. V roku 2014 vyhrala prvé miesto na súťaži ILSVRC14, pre klasifikáciu a detekciu. V článku [26] popisujú autori základnú stavebnú jednotku Inception sietí nazvanú *Inception cell*. *Inception cell* zobrazená na obrázku 3.6 kombinuje niekoľko malých filtrov veľkosti 1×1, 3×3, 5×5 a max pooling. Existuje niekoľko ďalších optimalizácií. Jednou, o ktorej píšu autori v publikácii [25], je zmenšenie veľkosti filtrov. Filtre veľkosti 5×5 alebo 7×7 sú prínosné pri extrakcii rysov pri veľkom rozlíšení, ale výpočet je pomerne drahý. Preto autori ukázali, že filter veľkosti 5×5 možno nahrať dvoma 3×3 filtrami.

Celá Inception (GoogLeNet) sieť má 27 vrstiev vrátane pooling vrstiev. Ak počítame všetky nezávislé bloky je počet vrstiev viac ako 100. Do celej siete sú pridané dva pomocné výstupy, u ktorých sa očakávalo, že podporia rozlíšiteľnosť na nižších vrstvách a zvýšia gradient pri spätnom šírení chyby. Ukázalo sa, že ich účinok na sieť je primárne regulačný. Celý podrobný popis tejto siete, spoločne s návrhom, optimalizáciami, trénovaním a iným je popísaný v článkoch [25, 26].



Obr. 3.6: Modul Inception s redukciou dimenzie. Bloky 1×1 sa používajú pre redukcie dimenzií pred 3×3 a 5×5 konvolúciami. Prevzaté z článku [26].

3.2 Predtrénovanie

V niektorých prípadoch predtrénovanie (*Transfer learning*) [19] výrazne urýchľuje proces učenia a zvyšuje presnosť siete. Všetky architektúry sietí uvedené v sekcii 3.1 je možné použiť už predtrénované. Najčastejšie sú siete predtrénované na veľkej dátovej sade ako napr. ImageNet, kde je vstupný obrázok klasifikovaný do jednej z 1000 rôznych tried. Existuje niekoľko základných metód, ktoré sa používajú pri predtrénovaní siete. Z nich najčastejšie sú používané extrakcia rysov (*feature extraction*) a vyladovanie (*fine-tuning*).

Feature extraction je spôsob, kedy sa predtrénovaná konvolučná sieť nastaví ako fixná časť novej siete. Zo siete, ktorá je predtrénovaná zvyčajne na veľkej dátovej sade, je nahradená posledná plne prepojená vrstva. Trénujú sa iba novo pridané parametre siete, napr. váhy v poslednej vrstve alebo bias. Zvyšok parametrov siete je zmrazených.

Fine-tuning je stratégia kedy nedochádza iba k nastaveniu parametrov siete, ale parametre sú ďalej upravované počas učenia siete. Je možné ladiť všetky parametre alebo zvoliť len niektoré v určitých vrstvách siete. Voľba parametrov sa určuje v závislosti od veľkosti dátovej sady a jej podobnosti na dátovú sadu, na ktorej bola sieť predtrénovaná. Ak je nová dátová sada dostatočne veľká je možné povoliť trénovanie všetkých parametrov siete. Naopak ak je dátová sada malá a podobná originálnej dátovej sade, bude lepšie trénovať iba parametre poslednej vrstvy siete.

Kapitola 4

Dátová sada BUT atribúty chodcov

Existujúce dátové sady, ktoré anotujú módny štýl osôb alebo popisujú jednotlivé časti už boli popísané v sekcii 2.3. Dátová sada FashionStyle14 [27] definuje štrnásť rôznych módnych štýlov. Na všetkých obrázkoch z tejto dátovej sady sú mladé ženy, odfotografované v štýlových outfitoch v profesionálnych podmienkach. Dátové sady PETA [4] a RAP [15] obsahujú obrázky a anotácie obyčajných ľudí zachytených na uliciach, námestiach alebo v obchodných centrách. V takýchto dátových sadách chýbajú anotácie módneho štýlu. Z tohoto dôvodu bola vytvorená nová dátová sada, ktorá okrem iného obsahuje aj anotácie módneho štýlu obrázkov ľudí v každodennom, prirodzenom oblečení. Triedy módneho štýlu boli navrhnuté tak aby pokrývali najbežnejšie módne štýly, v ktorých sú oblečení ľudia na uliciach v Českej republike. Na takejto dátovej sade môžu byť vykonané experimenty, v ktorých by neurónová sieť klasifikovala módny štýl priamo z obrázku chodca v bežnom prostredí, čo doteraz nebolo vyskúšané v žiadnej verejnej práci alebo publikácii.

Najväčšou výzvou pri tvorbe tejto dátovej sady bolo zozbieranie dostatočne veľkého množstva rôznorodých obrázkov chodcov a k nim dostatočne dobré a veľké množstvo anotácií. Zábery zozbierané z vonkajších kamier by bolo veľmi náročne anotovať automatickým spôsobom, preto som vytvorila webovú aplikáciu (popísanú v sekcii 6.1), v ktorej je možné obrázky anotovať manuálne. Pri manuálnych anotáciách vytváraných veľkým množstvom ľudí často vzniká problém subjektívneho názoru anotujúceho. Rôzne osoby môžu mať rozličný názor na módny štýl, vek a pohlavie jedného chodca zobrazeného na obrázku. Do úvahy tiež treba brať chyby spôsobené napríklad rozlíšením fotografie, alebo zlým uhlom kamery.

V článku [27], v ktorom bola vytvorená dátová sada FashionStyle14 autori zozbierali rôzne obrázky módnych štýlov na internete. Tie boli následne hodnotené odborníkmi a bežnými užívateľmi. Pri tvorbe novej dátovej sady popisovanej v mojej práci, bolo cieľom zbierať anotácie iba od bežných užívateľov. Takto vytvorená dátová sada by mala prispieť k tomu aby neurónová sieť trénovaná na tejto dátovej sade dosahovala výsledky čo najbližšie bežnému vnímaniu módneho štýlu, pohlavia a veku.

4.1 Návrh anotácií dátovej sady

Ako prvé pri tvorbe dátovej sady boli vyčlenené atribúty a triedy atribútov chodcov, ktoré má dátová sada obsahovať. Rozhodla som sa anotovať tri hlavné atribúty: pohlavie, vek a módny štýl oblečenia. Atribút vek bol rozdelený podobne ako v dátovej sade PETA do piatich tried (1-18, 19-29, 30-45, 45-60, 60+ rokov). Pretože na vytvorených videách

štýl	prvky
ležérny	bežné neformálne oblečenie, bunda, mikina, tričko, rifle, tepláky, tenisky, sandále, šlapky
športový	cyklisti, bežci, chrániče, prilba, športové oblečenie, tielko, bunda, športové legíny, tepláky, tenisky
metal-rock-moto	výstredné účesy, brada, výrazný make up, okuliare, šatky, retiazky, vybíjané doplnky, čierne oblečenie, kožené prvky, tričká s potlačou, lebky, veľké topánky, kanady
street	šiltovky, slúchadlá, retiazky, volné tričká a nohavice, roztrhané bundy, tenisky
elegantný	manažér, úradník, košeľa, sveter, sako, kabát, nohavice, sukne, šaty, vysoké opätky, čižmy, pančuchy
formálny	sako, smoking, šaty, opätky, kabelka
pracovný	vojak, policajt, pracovník na stavbe, zdravotná sestra, pracovné oblečenie,uniforma

Tabuľka 4.1: Definícia tried a prvkov typických pre módny štýl, v dátovej sade BUT atribúty chodcov.

sa nachádzali aj deti, bol oproti dátovej sade PETA pridaný vekový interval 1-18 rokov. Triedy módneho štýlu chodca na obrázku boli vybrané tak, aby bolo každého chodca možné priradiť do jednej z tried. Boli vybrané triedy ležérny, športový, metal-rock-moto, street, elegantný, formálny a pracovný. Rozdelene chodcov do tried na základe ich vizáže a oblečenia je popísané v tabuľke 4.1. Pre pohlavie boli zvolené klasické dve triedy muž a žena. Okrem týchto troch základných atribútov môžu byť k obrázku anotované aj vedľajšie atribúty ako napríklad činnosť, doplnky, sociálne alebo ekonomické postavenie osoby v spoločnosti a iné. Tie by mali byť získané z tagov, vyplnených v textovom poly a budú vybrané na základne počtu výskytov. Atribúty ako čiapka, šiltovka, helma, kapucňa môžu byť spojené do jedného spoločného binárneho atribútu *pokrievka hlavy*.

4.2 Zber a príprava dát

Pre rovnomerné pokrytie tried bolo nutné zozbierať obrázky chodcov, ktorý sú rôzneho veku, pohlavia a sú rozlične oblečení. Rozhodla som sa nepoužívať žiadne internetové zdroje, ale všetky obrázky do dátovej sady samostatne vytvoriť z video záznamov. Aby boli chodci na obrázkoch dostatočne rozdielni vytvorila som záznamy videí na niekoľkých miestach a v rozličných časoch. Pre dosiahnutie čo najlepšej kvality boli videá natočené v 4K rozlíšení. Prvé tri videá boli vytvorené spoločne v spolupráci s Jánom Jurčom, ktorý používal videá pre účely reidentifikácie osôb vo svojej bakalárskej práci. Osoby na videách boli zaznamenané štyrmi kamerami zo štyroch rôznych uhlov. Ďalšiu sériu videí som vytvorila samostatne pomocou jednej kamery pred nákupným centrom, športoviskom a koncertnou sálou. Týmito zábermi sa mi podarilo zozbierať osoby rôzneho veku, pohlavia a v rôznom oblečení. Prvú sériu obrázkov zachytenú štyrmi kamerami som získala výberom obrázkov z dátovej sady, ktorú vytváral Jan Jurča. Na týchto obrázkoch boli osoby rozdelené do priečinkov podľa ID. Každé ID obsahovalo dve až štyri ďalšie podpriečinky označujúce kamery. V nich sa následne nachádzali obrázky jedného chodca počas celej doby video záznamu.

Druhú časť obrázkov chodcov z videí som spracovávala samostatne. Pre vytvorenie výrezov chodcov bolo nutné najskôr v záberoch detekovať chodcov. Pre tento účel bol použitý

nástroj *OpenPose*, konkrétne implementácia *Tf-pose-estimation*[1]. *Tf-pose-estimation* je nástroj implementovaný pomocou knižnice *Tensorflow*[2], založený na vlastnej architektúre pre rýchle vyhľadávanie častí tela osôb vo videu alebo na obrázku. S jeho použitím som napísala vlastný skript `get_pose.py`, ktorý podľa zadaných argumentov vytvorí `pose` súbor, k jednému obrázku, obrázkom v špecifikovanom priečinku, alebo k celému videu. `Pose` súbor obsahuje koordináty kĺbov všetkých osôb na obrázku v `COCO` formáte zobrazenom na obrázku 2.2. Koordináty sú relatívne k počiatku súradnicového systému obrázku. Pre fungovanie musí mať skript prístup ku zdrojovým kódom *Tf-pose-estimation*, v ktorých je definovaná `COCO` štruktúra. Mnou vytvorený skript tiež dokáže vygenerovať súbor, v ktorom sú časti tela uložené v dvojrozmernom poli. S takto vygenerovanými `pose` súbormi možno pracovať aj bez nástroja *Tf-pose-estimation*.

Na základe vygenerovaných `pose` súborov boli následne vytvorené trasy (stopy) ľudí na videách. Cieľom bolo vytvoriť z videa rovnakú adresárovú štruktúru ako v prvej sérii obrázkov zo štyroch kamier. Rozdiel bol v tom, že obrázky boli vytvárané iba z jednej kamery narozdiel zo štyroch a nebolo nutné spájať osoby z rôznych kamier do jedného ID. Vytvorenie trasy osoby a uloženie obrázku v jednom priečinku umožnilo anotovať podstatne menšie množstvo obrázkov chodcov a pritom získať rozsiahlu dátovú sadu.

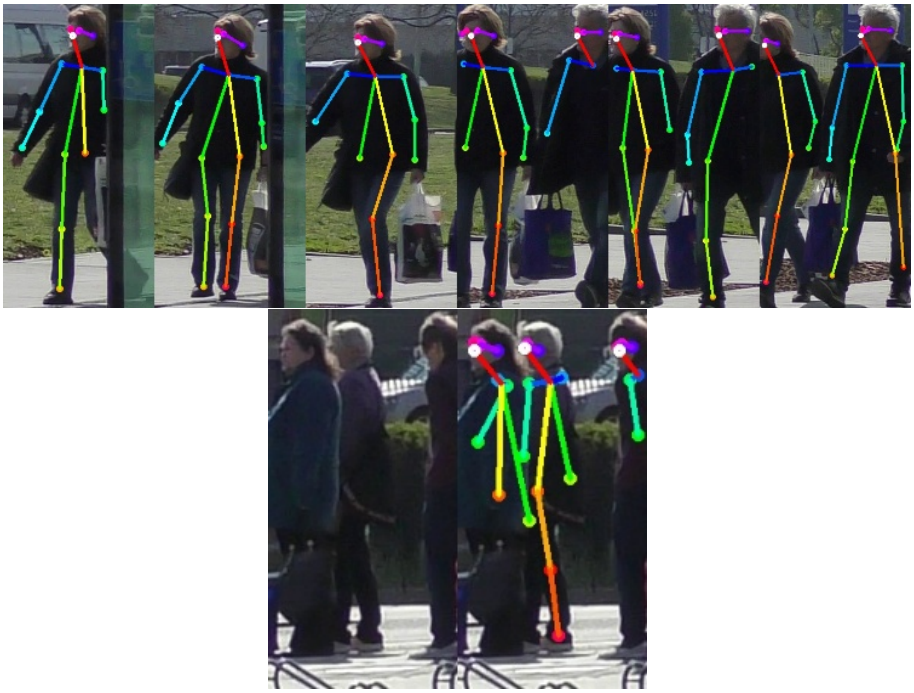
Pre vytvorenie týchto obrázkov som použila skript `get_pedestrian_tracks.py`, ktorý na základe videa a k nemu vytvorenému `pose` súboru vygeneruje do zvoleného priečinku trasy osôb. Trasy osôb sa vytvárajú na základe pozície krku osoby na snímkach. Medzi dvoma nasledujúcimi snímkami je vypočítaná euklidovská vzdialenosť pozície krku. V prípade, že je vzdialenosť menšia ako stanovený prah jedná sa o rovnakú osobu. Do priečinku s trasou chodca sa ukladá iba celý výrez postavy chodca na základe pózy získanej z `pose` súboru. Takto vytvorené stopy ľudí bolo treba manuálne prefiltrovať, pretože v niektorých prípadoch dochádzalo ku chybám. Napríklad počas sledovania osoby *OpenPose* nedokázal rozlíšiť dve osoby blízko seba alebo v niektorých prípadoch detekuje osoby na náhodných miestach. Niektoré chyby sú ukázané na obrázku 4.1.

Zo všetkých vytvorených a získaných obrázkov som následne vyberala obrázky, ktoré sú zobrazované vo webe pre zber anotácií. Pre každé unikátne `id` chodca boli vybrané 4 náhodné obrázky zo všetkých videí natočených na jednom mieste, na ktorých sa daný chodec nachádzal. Skript `preprocess_BUT_dataset.py`, pre každé `id` z týchto obrázkov vytvorí jeden spojený obrázok, na ktorom sú zamaskované tváre osôb. Takto vytvorené obrázky sú predané do webovej aplikácie. Spojením obrázkov dostaneme chodca zachyteného z rôznych uhlov, čo môže prispieť k získaniu presnejších anotácií. Pre maskovanie tvárí boli v skripte použité súradnice nosa osoby. Ak *OpenPose* na obrázku detekoval nos bol v tomto mieste, použitím *OpenCV*[3], vykreslený malý rozmazaný kruh. Príklady vytvorených obrázkov sú zobrazené na obrázku 4.2.

4.3 Zber anotácií

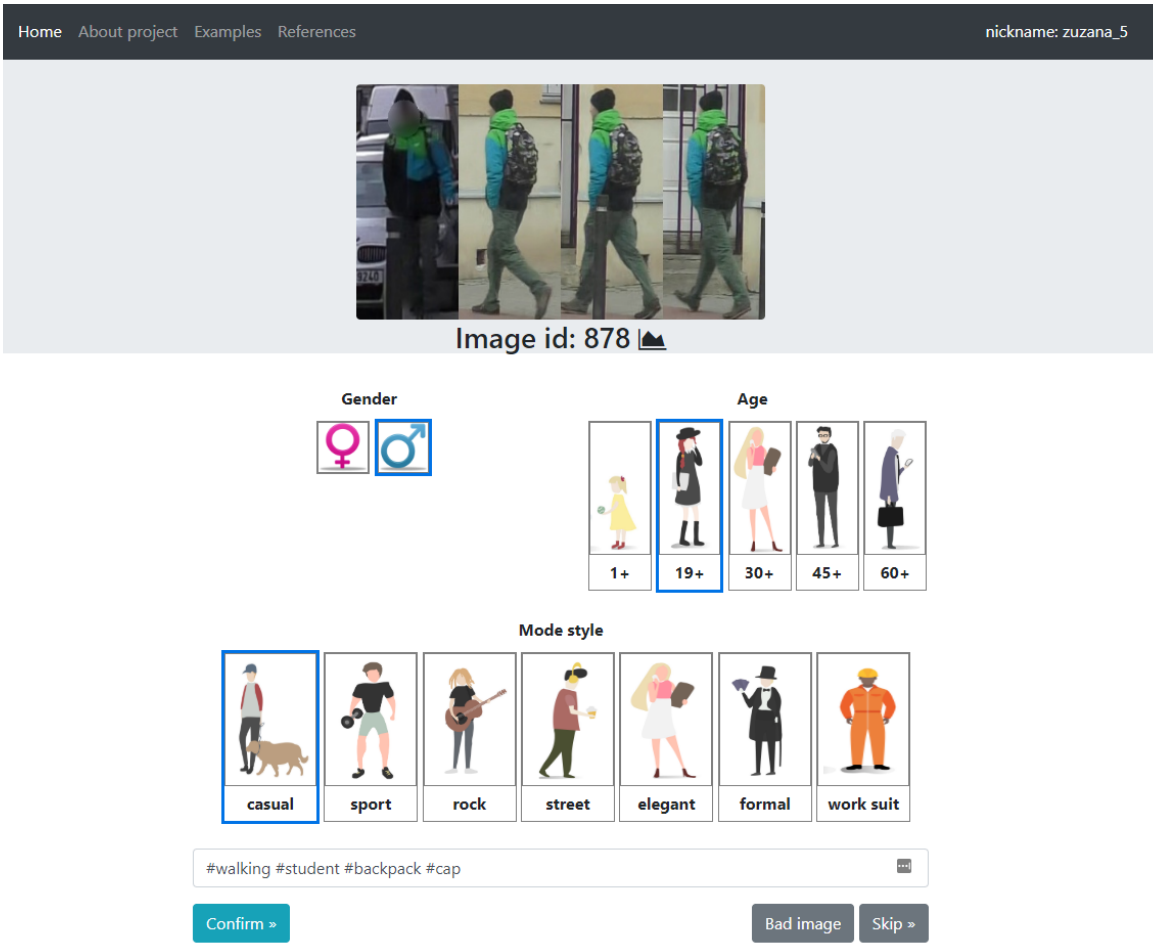
Anotácie boli zbierané cez webovú aplikáciu po dobu približne osem týždňov. Hlavným cieľom tejto aplikácie je aby bola stručná, intuitívna a bolo na prvý pohľad jasné, čo má užívateľ robiť a ako obrázky správne anotovať. Na úvodnej stránke je zobrazený popis a odkazy a príklady ako pri anotácii postupovať. Na hlavnej stránke je zobrazený obrázok, ku ktorému má užívateľ určiť atribúty osoby na obrázku. Pre výber atribútov sú na stránke

[1] webové stránky nástroja: https://github.com/ildoonet/tf-pose-estimation
[2] webové stránky Tensorflow: https://www.tensorflow.org/
[3] webové stránky OpenCV: https://opencv.org/



Obr. 4.1: Chyby pri vytváraní obrázkov chodcov pomocou nástroja *OpenPose*. Obrázky ukazujú chyby, ktoré vznikli prekrývaním osôb. Na hornom obrázku dochádza k zmene sledovanej osoby. Na dolnom obrázku nemožno rozdeliť tri osoby vedľa seba

Obr. 4.2: Príklady vytvorených obrázkov z videa, ktoré boli použité vo webovej aplikácii pre zber anotácií.



Obr. 4.3: Ukážka grafického rozhrania webovej aplikácie pre zber anotácií.

implementované výberové políčka (*radio boxi*) formou ikony. Tým je výber tried intuitívnejší a užívateľ nie je nútený čítať žiadny text. Pre upresnenie sú triedy napísané pod ikonami. Po prechode myšou cez ikonu triedy módneho štýlu sú zobrazené informácie o danom módnom štýle. Ukážka grafického rozhrania webovej aplikácie je na obrázku 4.3.

Pri anotovaní obrázkov bolo cieľom získať o každom obrázku, čo najväčšie množstvo anotácií. V prípade, kedy dochádzalo k tomu, že pre jeden obrázok existovali dve rôzne anotácie niektorého z atribútov, bola tomuto obrázku zvýšená priorita. Týmto spôsobom mohol byť obrázok znova anotovaný a bolo možné rozhodnúť, ktoré triedy atribútov sa na obrázku nachádzajú.

4.4 Vytvorenie dátovej sady

Z vytvorených videí, ktoré boli natočené pre tvorbu dátovej sady bolo získaných viac ako 1100 jedinečných identít, ktoré boli anotované vo webovej aplikácii. Ukázalo sa, že anotácie vytvorené k obrázkom sú ale veľmi nerovnomerne rozdelené. Pri atribúte vek bolo približne 50% chodcov na obrázkoch pridelených do triedy 19-30 rokov. Pri módnom štýle patrilo do triedy ležérny módny štýl patrilo až 62% percent všetkých chodcov. Rozdelenie troch hlavných atribútov a ich tried ku všetkým obrázkom je v tabuľke 4.2.

	pohlavie	vek					módny štýl					
atribút	muž	1+	19+	30+	45+	60+	ležérny	šport	rock	elegant.	form.	prac.
počet	689	81	545	294	134	62	694	76	102	151	27	5

Tabuľka 4.2: Ukážka počtu anotácií obrázkov pre tri hlavné atribúty a ich triedy.

atribút	ruksak	kabelka	nákupná taška	čiapka	šál	okuliare
počet	128	69	18	47	7	18

Tabuľka 4.3: Ukážka počtu zozbieraných vedľajších binárnych atribútov.

Okrem troch hlavných atribútov boli zozbierané ešte vedľajšie atribúty, ktoré boli anotované pomocou textového pola a vyplnených tagov. Najčastejšie sa na obrázkoch vyskytovali atribúty ruksak, kabelka, pokrievka hlavy (čiapka, šiltovka, kapucňa), nákupná taška, okuliare, šál alebo šatka. Kladná alebo záporná hodnota týchto binárnych atribútov bola priradená k obrázku na základe spracovaných údajov z databázového pola `description`. Tagy, ktoré boli najčastejšie anotované a bolo ich možné spojiť, boli spojené do jedného atribútu. Príkladom je atribút `nákupná taška`, ktorý bol spojený z anotácií `shopping`, `bag`, `shoppingbag`. Vybrané atribúty sú zobrazené v tabuľke 4.3.

Pri vytváraní kompletnej dátovej sady bolo ku každému anotovanému obrázku vybraných podľa `id` chodca niekoľko ďalších obrázkov. Obrázky chodcov boli vybrané náhodne a všetkým bola pridelená rovnaká anotácia ako hlavnému anotovanému obrázku. Pri chodoch zachytených viacerými kamerami, bol z každej kamery vybraný rovnaký počet obrázkov. Neboli vybrané kompletne všetky vytvorené obrázky nakoľko zábery z jednej kamery bývali často veľmi podobné, a takýto výber by pravdepodobne ešte zväčšil problém nevyváženosti sady. Celkovo bola vytvorená dátová sada obsahujúca 21,000 obrázkov. Všetky obrázky boli rozdelené do trénovacej množiny (60%), validačnej množiny (5%) a testovacej množiny (35%). Ukážka obrázkov z vytvorenej dátovej sady je na obrázku 4.4.

Obr. 4.4: Ukážky obrázkov z dátovej sady BUT atribúty chodcov. Na obrázkoch sú ukážky módnych štýlov tak ako je ich možné anotovať vo webovej aplikácií. Na prvom obrázku je módny štýl *ležérny*, na poslednom obrázku *pracovná uniforma*.

Kapitola 5

Konvolučné neurónové siete pre rozpoznávanie atribútov

V tejto kapitole som sa zamerala na popis architektúr a ich úpravy pri implementácii konvolučných neurónových sietí pre klasifikáciu atribútov osôb na troch dátových sadách: PETA [4], FashionStyle [27] a BUT atribúty chodcov.

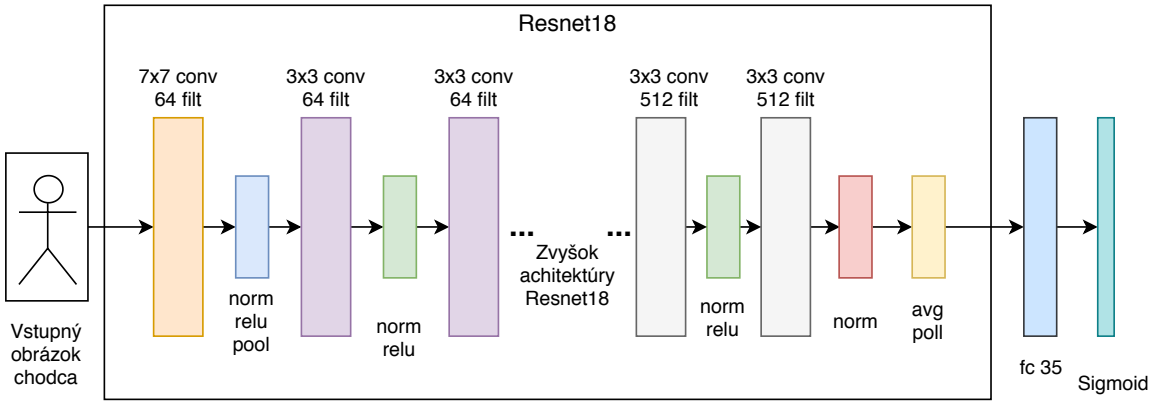
V sekcii 3.1 bolo ukázané, že architektúry ako napríklad VGG, Resnet, DenseNet alebo Inception sú vhodné pre klasifikáciu v rôznych úlohách. Z tohoto dôvodu som sa rozhodla použiť ako základný model sietí práve modely VGG16, Resnet18, Inception v3. Siete s touto architektúrou, spomínané v publikáciách [9, 21, 26] boli trénované pre klasifikáciu do 1000 rôznych tried na dátovej sade ImageNet, preto som ich v tejto práci musela upraviť. Úpravy boli vykonané na základe dátovej sady a klasifikačného problému a budú bližšie popísané neskor v tejto kapitole.

Vstupom sietí je obrázok o veľkosti 224×244px, čo je štandardný rozmer väčšiny týchto architektúr. Inak je to u InceptionNet, ktorej štandardný vstupný obrázok je veľkosti 299×299px. V dátových sadách sa v strede každého obrázku nachádza primárne jedna osoba. Výnimkou sú prípady, keby sa napríklad chodci vyskytovali tak blízko vedľa seba, že ich nebolo možné oddeliť. Obrázky v sadách sú rozdielnej veľkosti a bolo ich potrebné upraviť na túto veľkosť napríklad orezaním, pridaním okraja alebo rozšírením. Aby nedochádzalo k strate informácií, obrázky boli obrázky rozšírené na veľkosť 224×244px. Takto upravené obrázky boli následne normalizované a predávané na vstup.

Dátová sada PETA popísaná v sekcii 2.4, obsahovala anotácie do viacerých binárnych tried, preto som konvolučnú sieť upravila tak aby pre vstupný obrázok klasifikovala niekoľko binárnych atribútov chodcov. Výstupom siete je zoznam atribútov, ktoré sa na vstupnom obrázku nachádzajú, resp. nenachádzajú. Ako základná architektúra siete bola zvolená Resnet18. Posledná vrstva bola nahradená plne prepojenou vrstvou, ktorá mala počet výstupov rovný počtu klasifikovaných atribútov. Obrázok 5.1 ukazuje úpravy architektúry Resnet18. Pre dosiahnutie výstupu siete v intervale $< 0, 1 >$ som použila logistickú funkciu Sigmoid

$$H_p(q) = \frac{1}{1 + \epsilon^{-x}}, \tag{5.1}$$

kde x je výstup z neurónovej siete, pre jeden atribút. Ak bola vypočítaná hodnota väčšia ako 0.5, v tom prípade bol atribút prítomný na obrázku.



Obr. 5.1: Neurónová sieť použitá pri klasifikácii obrázkov chodcov z dátovej sady PETA. Vnútorný blok Resnet18 je podobný ako napríklad Resnet na obrázku 3.4. Výstupom je plne prepojená vrstva s 35 výstupmi nasledovaná funkciou Sigmoid.

Pri trénovaní bola použitá chybová funkcia Binary cross entropy

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i)), \tag{5.2}$$

kde y_i je skutočná hodnota atribútu i, $p(y_i)$ je pravdepodobnosť predikcie atribútu y_i a N je počet atribútov.

Sieť pre dátovú sadu FashionStyle14 klasifikuje módny štýl osôb na obrázkoch. Osobe je pridelený jeden zo štrnástich módnych štýlov. Pre tento spôsob klasifikácie som nahradila poslednú vrstvu architektúr VGG, Resnet a Inception modelu plne prepojenou vrstvou so štrnástimi výstupmi. Pomocou funkcie Softmax

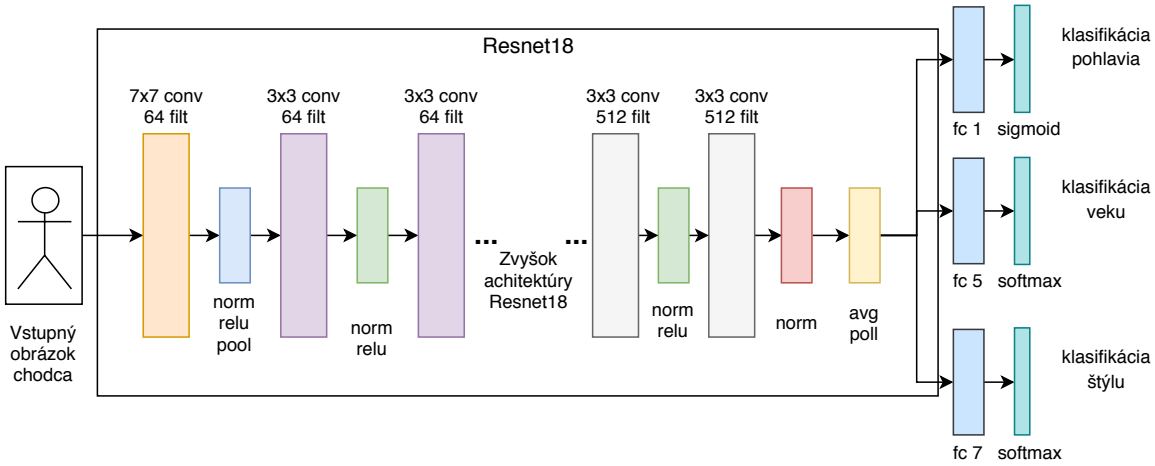
$$P(y_i) = \frac{\epsilon^{y_i}}{\sum_j \epsilon^{y_j}}, \tag{5.3}$$

kde j je počet tried, je z celého výstupu siete y vytvorený vektor pravdepodobností príslušnosti $P(y)$ do tried módneho štýlu. Výstup $P(y_i)$, ktorý bol predikovaný na základe hodnoty výstupu y_i a zároveň nadobúda najväčšiu hodnotou pravdepodobnosti, určuje módny štýl osoby na obrázku. Chybová funkcia bola v tomto prípade vypočítaná ako Cross entropy loss

$$H_p(q) = -\sum_{c=1}^{C} q(y_c) \cdot log(p(y_c)), \tag{5.4}$$

kde C je počet tried, q(y) je skutočná trieda a p(y) je predikovaná trieda na obrázku.

Nová dátová sada Charakteristiky chodcov obsahuje anotácie o veku, pohlaví a módnom štýle, pričom pohlavie je posudzované ako binárny atribút, vek a módny štýl ako viactriedne atribúty. Pritom sa predpokladá, že o vstupnom obrázku chceme dostať informáciu o všetkých troch atribútoch zároveň. Ako základnú architektúru som použila rovnako ako pri dátovej sade PETA, model Resnet18. V tomto prípade bola posledná vrstva nahradená troma plne prepojenými vrstvami, jedna pre každý atribút. Vrstva pre klasifikáciu



Obr. 5.2: Neurónová sieť použitá pri klasifikácii obrázkov chodcov z novo vytvorenej dátovej sady. Výstupom sú tri plne prepojené vrstvy jedna pre vek (1 výstup), druhá pre pohlavie (5 výstupov) a posledná pre módny štýl (7 výstupov).

veku obsahovala len jeden binárny výstup, ktorý je prepočítaný pomocou funkcie *Sigmoid* rovnica 5.1. Viactriedna klasifikácia veku a módneho štýlu je počítaná pomocou funkcie *Softmax* rovnica 5.3. Vrstva pre klasifikáciu veku má 5 výstupov, jeden výstup pre každú triedu veku. Posledná vrstva pre módny štýl má 7 výstupov, rovnako jeden pre každý klasifikovaný módny štýl. Architektúra je zobrazená na obrázku 5.2.

Chybová funkcia sa počíta na základe troch chýb. V každej plne prepojenej vrstve je počítaná samostatná chybová funkcia. Pre klasifikáciu pohlavia funkcia *BinaryCrossEntory* 5.2 a pre vek a módny štýl funkcia *CrossEntropy* 5.4. Celková chyba je získaná použitím vzorca

$$loss = loss1 + \alpha \cdot loss2 + \beta \cdot loss3, \tag{5.5}$$

kde loss1 resp. loss2 resp loss3 značia chyby spočítané zvlášť pre každú z výstupných vrstiev. Parametre α a β sú zvolené na základe veľkosti jednotlivých chýb.

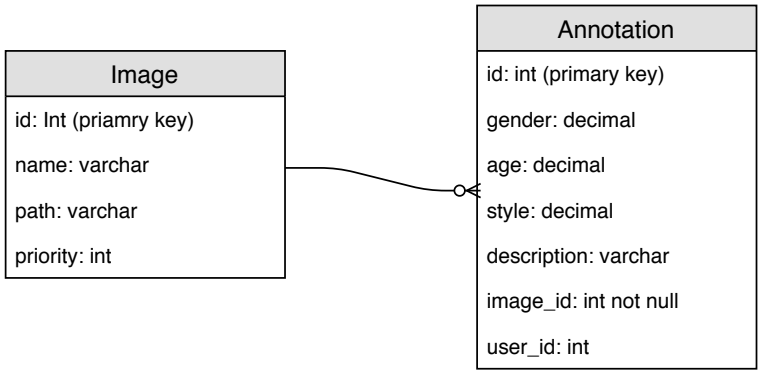
Kapitola 6

Implementácia

Celá implementácia pozostáva okrem programovania neurónových sietí pre klasifikáciu rôznych charakteristík chodcov tiež z vytvorenia novej dátovej sady. Vytvorenie dátovej sady vyžadovalo tvorbu webovej aplikácie, pre zber anotácií, a implementáciu viacerých skriptov pre zber a úpravu obrázkov. Pri implementácií som použila rôzne frameworky a knižnice pre spracovanie obrazu, detekciu osôb v obraze, tvorbu neurónových sietí alebo tvorbu webových stránok a formulárov. Všetky tieto knižnice sú vytvorené pre jazyk `Python`[1], v ktorom som písala všetky skripty, neurónové siete a webovú stránku. Kapitola popisuje návrh a postup pri vývoji a implementácií týchto častí.

6.1 Web pre tvorbu anotácií

Webová aplikácia pre zber anotácií je vytvorená pomocou programovacieho jazyka `Python`. Pri implementácií som využila frameworky `flask`, `sqlalchemy`, `pymysql` a `wtforms`. Aplikácia komunikuje s `Mysql` databázou, v ktorej sú uložené informácie o zobrazovaných obrázkoch a vytvorených anotáciách. Pre ukladanie týchto informácií stačia dve tabuľky `Image` a `Annotation`. Jednoduchý návrh schémy databáze je zobrazený na obrázku 6.1. V tabuľke anotácie je v numerickej podobe uložené pohlavie, vek a módny štýl osoby na obrázku. Pole description ukladá dodatočné tagy a slová, ktoré môže užívateľ napísať o obrázku.



Obr. 6.1: Návrh schémy databáze použitej pre ukladanie anotácií z webovej aplikácie pre dataset BUT atribúty chodcov.

[1] webové stránky jazyka Python: https://www.python.org/

Informácie o nových obrázkoch sa dajú vložiť do databázy skriptom `databaseinit.py`, ktorý berie ako argument textový súbor s názvami obrázkov. Priorita každého novo vloženého obrázku je nastavená na 100. Pri načítaní stránky pre anotovanie, aplikácia náhodne vyberie 30 záznamov. Záznamy sú vybrané z tabuľky `Image` a sú zoradené podľa priority. Z nich aplikácia náhodne vyberie jeden obrázok a ten sa zobrazí užívateľovi. Zároveň je `id` obrázku súčasťou URL stránky a pri zmene prístupového bodu (endpointu) sa možno dostať na ľubovolný obrázok, ktorý je v databáze. Ak užívateľ vyberie pohlavie, vek, módny štýl a potvrdí svoj výber, dôjde k uloženiu anotácie do tabuľky `Annotation` a priorita obrázku sa zníži na polovicu. Týmto spôsobom možno dosiahnuť, aby boli všetky obrázky rovnomerne zobrazované užívateľom. Pri každej vytvorenej anotácii sa ukladá číslo alebo prezývka užívateľa. Prezývku je možné zadať pri prvej návšteve stránky. Alternatívne je možné pokračovať anonymne, vtedy užívateľ dostane náhodne vygenerované číslo. Prezývka alebo číslo je uložené v lokálnych cookies užívateľa. `User_id` nezaručuje jedinečnú identitu užívateľa, ale je postačujúce pre prípadné rozlíšenie nespoľahlivých anotácií. Po nasadení prvej základnej verzie aplikácie boli vykonané užívateľské testy, ktoré mali slúžiť pre odladenie chýb a vylepšenie aplikácie. Pri testovaní bol užívateľom priložený základný akceptačný test. Po vykonaní testu mal užívateľ sám anotovať niekoľko ďalších obrázkov a počas práce komentovať, ktoré vlastnosti stránky sa mu páčia/nepáčia, alebo ktorú časť by chceli zmeniť. Vyhodnotenie testovania na prvej verzii webovej aplikácie viedlo k čiastočnej optimalizácii aplikácie. Bolo pridané `user_id`, tlačidlá pre potvrdenie a preskočenie obrázku boli farebne zvýraznené, pribudlo tlačidlo pre označenie nesprávneho obrázku. Pod ikony módneho štýlu pribudol popis a nápoveda. Ukázalo sa, že kvôli maskovaniu tvárí nemohli niekedy užívatelia správne rozpoznať vek osoby. Z tohoto dôvodu som musela pri niektorých obrázkov pridať na webovú stránku možnosť zobrazenia nezamaskovanej a zväčšenej verzie obrázku. Pri testoch sa tiež ukázalo, že väčšina užívateľov nevypĺňala voliteľné pole, do ktorého mohli napísať pripomienky k obrázku. Vybrala som najčastejšie sa vyskytujúce atribúty a pridala som voliteľné zaklikávacie obrázky, ktoré označovali binárne atribúty ruksak, kabelka, nákupná taška, okuliare a čiapka. Pridanie týchto obrázkov výrazne prispelo k zbieraniu anotácií týchto voliteľných atribútov. Zdrojové kódy stránky sú okrem tejto práce prístupné aj v gite na stránke https://github.com/zuzanica/datagetherer.

6.2 Implementácia neurónových sietí

Pri implementácii neurónových sietí bola použitá platforma PyTorch[2]. PyTorch ponúka rozsiahle možnosti implementovaných neurónových sietí, chybových optimalizačných alebo aktivačných funkcií a zároveň ponúka triedy a metódy na predspracovanie vstupných dát. V práci som vytvorila niekoľko skriptov, ktoré implementovali neurónovú sieť pre klasifikáciu atribútov osôb na niektorej z vybraných dátových sád. V skriptoch je potrebné nastaviť cestu k dátovej sade a `csv` súborom, ktoré obsahujú názvy obrázkov a značky trénovacej, validačnej a testovacej množiny. Obrázky a značky sú načítané pomocou triedy `Dataset`. Tá umožňuje vyberať obrázky a značky po rôzne veľkých skupinách. Skriptom tiež možno nastaviť model siete, štýl trénovania (feature-extraction, finetunnig), veľkosť batchu, počet epoch a súbor, do ktorého sa budú ukladať štatistiky.

Po inicializácii dát je vytvorený model. Ten je načítaný a inicializovaný na existujúcu sieť napríklad ResNet18, tá sa stiahne z repozitárov PyTorch pri prvej inicializácií. V tomto modeli je následne upravená posledná vrstva tak ako je písané v kapitole 5. Sieti je následne na-

[2]Webové stránku platformy platformy PyTorch https://pytorch.org/

stavený optimalizátor a chybová funkcia. Pre binárnu klasifikáciu PyTorch definuje chybovú funkciu `BCEWithLogitsLoss`.Táto funkcia kombinuje v jednej triede *Sigmoid* vrstvu a *Binary Cross Entropy Loss*. Pre viactriednu klasifikáciu je použitá funkcia `CrossEntoryLoss`, ktorá kombinuje `logaritmický Softmax` a `NLNLoos` (The negative log likelihood loss).

Trénovanie prebieha na trénovacej množine počas zvoleného počtu epoch. Po každej iterácii prebehne validácia na malej validačnej množine obrázkov. Počas validácie, je model nastavený do vyhodnocovacieho módu, kedy nedochádza k spätnému šíreniu chyby a učeniu siete. Zároveň sa uloží model siete a jeho dosiahnutá úspešnosť klasifikácie. Po skončení trénovania je načítaný model siete, ktorý dosiahol najlepšiu úspešnosť klasifikácie na validačnej množine. Následne sa tento model použije pre vyhodnotenie štatistík úspešnosti siete. Na vstup modelu je v tejto fáze použitá testovacia množina dát. Štatistiky sú vypísané na výstup a zároveň uložené do súboru, ktorý bol špecifikovaný v argumentoch skriptu.

Kapitola 7

Experimenty

Experimenty by mali ukazovať rôzne princípy ako pomocou neurónových sietí rozlišovať rôzne charakteristiky chodcov. Vo všetkých experimentoch sú použité architektúry neurónových sietí popísané v kapitole 5. Na prevzatých dátových sadách sú experimentované niektoré existujúce riešenia, ktoré sú porovnané so state of art metódami rôznych vedeckých publikácií. Výsledky, ktoré sú získané pri experimentoch na verejných dátových sadách, boli použité pre výber najlepšieho modelu neurónovej siete, ktorý je použitý pri experimentoch nad novo vytvoreniu dátovou sadou. Experimenty som rozdelila do troch častí:

- Experimenty pre získavanie charakteristík
- Experimenty pre klasifikáciu módneho štýlu
- Experimenty na novej dátovej sade

7.1 Získavanie charakteristík

Experiment pre získavanie charakteristík porovnáva výsledky experimentov autorov vedeckých publikácií [14, 15] s výsledkami dosiahnutými pomocou neurónovej siete s architektúrou Resnet18 (popísaná v sekcii 5), ktorá končí plne prepojenou vrstvou s N výstupmi, kde N značí počet klasifikovaných binárnych atribútov. Ako trénovacia dátová sada bola zvolená PETA [4]. Dátová sada PETA obsahuje veľké množstvo dobre anotovaných obrázkov chodcov s veľkým množstvom atribútov. Experiment obsahuje výsledky klasifikácie tridsiatich piatich atribútov. Parametre siete boli inicializované na hodnoty získané trénovaním na ImageNet pre klasifikáciu do 1000 tried. Následne bola použitá metóda fine-tuning, čiže počas trénovania siete boli aktualizované všetky parametre siete. Pre výpočet úspešnosti klasifikátoru boli použité rovnice 2.1, 2.2 a 2.3, uvedené v sekcii 2.5.

Predpríprava dátovej sady PETA

Celá dátová sada obsahovala až 61 rôznych binárnych atribútov, kde však nie všetky mali dostatočne veľký počet výskytov naprieč dátovou sadou. Pre experiment bolo vybraných, rovnako ako v spomínaných publikáciach, 35 binárnych atribútov. Dátová sada bola náhodne rozdelená do množín rovnakej veľkosti ako v článku [4]. Trénovacia množina obsahovala 9500 prvkov (50%), validačná množina 1900 (10%) prvkov a testovacia množinu 7600 (40%) prvkov.

Trénovanie

Sieť bola trénovaná po dobu 15 epoch. Počas tohoto času sa sieť natrénovala na viac ako 99% na trénovacej množine, trénovacie dáta sa teda naučila naspamäť. Na validačnej množine dosiahla úspešnosť 92%. Pri trénovaní bola použitá metóda trénovania fine-tuning, a boli učené všetky parametre siete. Bola použitá chybová funkcia kombinujúca Sigmoid a Binary Cross Entropy Loss a optimalizačná metóda Adam. Koeficient učenia bol nastavený na 10^{-3}. Veľkosť batchu bola nastavená na 128 prvkov. Pre získanie binárnej hodnoty atribútu pre každý výstupu siete bola použitá funkcia Sigmoid.

Zhrnutie experimentu a výsledkov

Výsledky experimentu sú zobrazené v tabuľke 7.2 a 7.1. Z tabuľky je vidno, že implementovaná sieť dosiahla v mnohých atribútoch podobné výsledky alebo lepšie výsledky ako uvedené siete. Klasifikácia binárnych atribútov na takto veľkej dátovej sade sa ukázala ako pomerne úspešná. Natrénovaný model bude možné použiť na predtrénovanie siete pri experimetoch s novo vytvorenou dátovou sadou. Predtrénovanie siete pomocou spôsobu fine-tuning spôsobilo, že sieť sa naučila klasifikovať atribúty v pomerne krátkom čase. Na grafickej karte NVIDEA GeForce GTX 1060 zabral celý experiment 12m 44s.

	mA	Accuracy	Precision	Recall	F1 score
ACN	81.15	73.66	84.06	81.26	82.64
DeepMAR	82.89	75.07	83.68	83.14	83.41
My Result	77.66	79.94	86.23	80.24	83.15

Tabuľka 7.1: Výsledky experimentu na dátovej sade PETA, vypočítané pomocou metrík 2.1, 2.2, porovnané s niektorými výsledkami článku [15].

Attribute	MRFr2	DeepSAR	DeepMAR	My Results
Age16-30	86.8	82.9	85.8	86.09
Age31-45	83.1	79.4	81.8	81.26
Age46-60	80.1	83.3	86.3	82.01
AgeAbove61	93.8	92	94.8	91.48
Backpack	70.5	78.8	82.6	80.29
CarryingOther	73	73	77.3	73.63
Casual lower	78.2	81.6	84.9	79.84
Casual upper	78.1	81.1	84.4	78.90
Formal lower	79	81.9	85.2	79.74
Formal upper	78.7	81.6	85.1	79.83
Hat	90.4	89.2	91.8	87.48
Jacket	72.2	77.5	79.2	68.45
Jeans	81	80.2	85.7	82.68
Leather shoes	87.2	84.2	87.3	85.27
Logo	52.7	76.1	68.4	57.70
Long hair	80.1	83.2	88.9	89.15
Male	86.5	85.1	89.9	91.02
MessengerBag	78.3	77.4	82.00	81.20
Muffler	93.7	94.4	96.1	93.27
No accessory	82.7	81.5	85.8	84.32
No carrying	76.5	78.8	83.1	80.29
Plaid	65.2	84.9	81.1	78.29
Plastic bag	81.3	82.9	87.0	78.31
Sandals	52.2	81.3	67.3	54.09
Shoes	78.4	75.8	80.0	79.29
Shorts	65.2	81.9	80.4	73.27
ShortSleeve	75.8	84.6	85.67	82.95
Skirt	69.6	83.2	82.2	69.14
Sneaker	75.0	77.3	78.7	76.83
Stripes	51.9	72.8	66.5	61.47
Sunglasses	53.5	79.1	69.9	57.03
Trousers	82.2	78.4	84.3	82.89
Tshirt	71.4	80	83.0	73.75
UpperOther	87.3	83.4	86.1	85.70
V-Neck	53.3	75.4	69.8	51.12
Average	75.6	81.3	82.6	77.66

Tabuľka 7.2: Výsledky experimentu na dátovej sade PETA [4] porovnané s výsledkami článku [14]. Z výsledkov vidno, že použitím siete Resnet18 predtrénovanej na sade ImageNet možno dosiahnuť porovnateľných výsledkov ako u uvedených publikácií.

7.2 Klasifikácia módneho štýlu

Pre klasifikáciu módneho štýlu som sa rozhodla použiť dátovú sadu FashionStyle14, ktorá obsahuje anotácie štrnástich rôznych módnych štýlov. Experimenty boli vykonané na konvolučných neurónových sieťach s rôznymi architektúrami. Každá z architektúr končila plne prepojenou vrstvou so štrnástimi výstupmi. Cieľom bolo porovnať siete s rôznou architektúrou a rôznou metódou učenia a zistiť, ktorá dosahuje pri klasifikácii módnych štýlov najlepšie výsledky. Boli zvolené architektúry VGG16 s batch normalizáciami, Resnet18 a Resnet50 a Inception v3. V prvom prípade sieť $Resnet18^{fl}$[1], nebola sieť nijak predtrénovaná, celé učenie prebiehalo na dátovej sade FashionStyle14. Druhá sieť $Resnet18^{ft}$[2] bola predtrénovaná na dátovej sade ImageNet a počas trénovania na FashionStyle14 bola použitá metóda fine-tuning, kedy boli učené všetky parametre siete. Pri sieťach Resnet50, VGG16 a Inception v3 boli učené iba parametre (váhy a bias) poslednej plne prepojenej vrstvy. Pri všetkých sieťach bola zvolená chybová funkcia Cross Entorpy Loss a optimalizačná metóda Adam s koeficientom učenia 10^{-3}. Vstupom siete bol obrázok náhodne orezaný na veľkosť 224×244px. Pre urýchlenie učenia bola sieť trénovaná v dávkach (batch) o veľkosti 128 obrázkov. Obrázky boli rozdelené do trénovacej, validačnej a testovacej sady dvoma variantami. Prvá varianta bola prevzatá z verejne dostupných stránok dátovej sady. V druhej variante boli obrázky náhodne rozdelené do množín v rovnakom percentuálnom pomere ako v publikácii [27], s ktorou sú porovnané výsledky.

Predspracovanie dátovej sady

Dátová sada FashionStyle14 publikovaná na internete sa ukázala ako problémová. Zverejnené súbory rozdeľujúce obrázky na trénovaciu, validačnú a testovaciu množinu neboli korektné, niektoré obrázky dátovej sady nebolo možné otvoriť, názvy obsahovali cudzie znaky, čiarky, medzery alebo ich dĺžka prekračovala 250 znakov. Kvôli tomuto bolo nutné zredukovať počet obrázkov v dátovej sade. Oproti uvádzanému počtu 13,126 obrázkov, bolo možné použiť iba 9886 obrázkov. Ďalší problém bolo vyváženie obrázkov v jednotlivých triedach a množinách. V trénovacej a validačnej množine úplne chýbalo zastúpenie obrázkov pre triedu *girlish* (dievčenský módny štýl). Tieto chyby ma viedli k vykonaniu ďalších experimentov. Výsledky experimentov vykonané s verejnými množinami obrázkov sú zobrazené v tabuľke 7.4. V ďalšom experimente som sama náhodne rozdelila obrázky do trénovacej, validačnej a testovacej množiny. Boli použité všetky obrázky, ktoré bolo možné otvoriť a rovnomerne ich rozdeliť pre všetky triedy a množiny. Do množín boli rozdelené podľa článku [27]. 60% obrázkov v trénovacej množine, 5% vo validačnej množine a 35% v testovacej. Vo výsledku bolo použitých 13004 obrázkov, počet obrázkov v jednotlivých triedach je zobrazený v tabuľke 7.3.

Zhrnutie experimentu a výsledky

Výsledky experimentov na rôznych typoch architektúr sú zobrazené v tabuľke 7.4 a na grafoch 7.1, 7.2 ukazujú, výrazné rozdiely v trénovaní. Z tabuľky vidno, že trénovanie siete iba na dátovej sade FashionStyle14 dosiahlo výrazne horšie výsledky ako varianty, keby bola sieť predtrénovaná na dátovej sade ImageNet. Pri použití feature-extraktingu sa sieť síce rýchlejšie naučila klasifikovať módny štýl, ale výsledky neboli také dobré ako pri aktualizovaní všetkých parametrov. Siete Resnet50, VGG16 a Inception v3 používali metódu

[1] fl = full learning, označenia učenia siete bez predtrénovania
[2] ft = fine-tuning, označenie použitia metódy fine-tuning

feature-extrakting, a dosiahli horšie výsledky ako menšia sieť Resnet18, ktorá používala fine-tuning. Do ďalších experimentov na mojej dátovej sade bude vhodné použiť Resnet18, predtrénovanú na ImageNet a pri trénovaní na novej dátovej sade trénovať všetky parametre siete.

V tabuľke 7.5 sú zobrazené výsledky siete $Resnet18^{ft}$ trénovanej na verejných množinách a tej istej siete trénovanej pri náhodnom rozložení obrázkov do trénovacej, testovacej a validačnej množiny. Výsledky sú porovnané s hodnotami z publikácie [27]. U siete $Resnet18^{ft}$ mala na výsledok vplyv absencia módneho štýlu *dievčenský*. Sieť $Resnet18^{*}$ trénovaná na náhodne vygenerovaných množinách dosahuje o 4% lepšie výsledky ako boli dosiahnuté na zverejnených množinách obrázkov alebo v publikácii [27]. Toto ukazuje, že vyváženosť dátovej sady a zväčšenie počtu obrázkov v dátovej sade prispelo k lepšiemu výsledku siete.

model	počet obrázkov
konzervatívny	901
elegantný	843
etnický	849
rozprávkový	954
ženský	800
večerný	928
dievčenský	1081
ležérny	1016
lolita	1058
módny	1056
prirodzený	854
retro	840
rockový	807
street	1017
spolu	13004

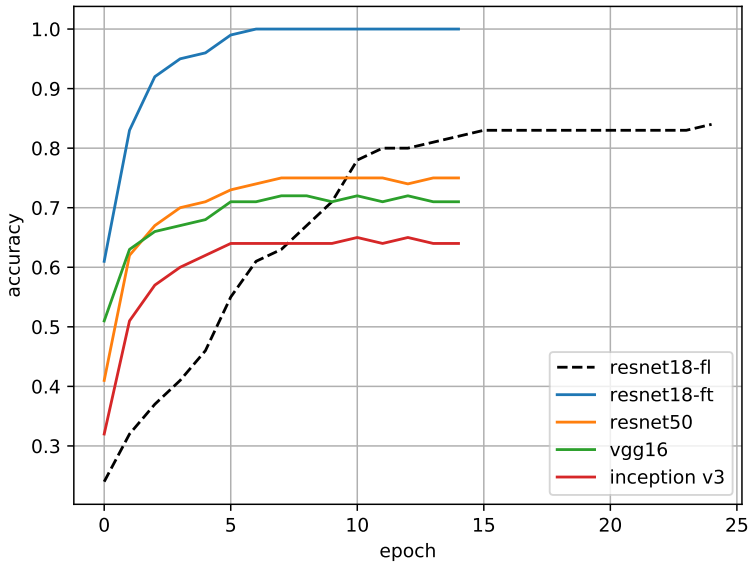
Tabuľka 7.3: Počet obrázkov v jednotlivých triedach Fashionstyle datasetu po vyfiltrovaní chybných obrázkov.

model	$resnet18^{fl}$	$resnet18^{ft}$	resnet50	vgg16	inption v3
konzervatívny	0.30	**0.64**	0.51	0.46	0.40
elegantný	0.77	**0.95**	0.88	0.89	0.85
etnický	0.30	**0.65**	0.59	0.57	0.56
rozprávkový	0.81	**0.90**	0.78	0.87	0.52
ženský	0.54	**0.60**	0.60	0.63	0.51
večerný	0.56	**0.81**	0.65	0.64	0.65
dievčenský	0.00	0.00	0.00	0.00	0.00
ležérny	0.58	**0.71**	0.59	0.61	0.56
lolita	0.65	**0.94**	0.89	0.84	0.86
módny	0.48	**0.69**	0.56	0.51	0.51
prirodzený	0.55	**0.77**	0.72	0.61	0.60
retro	0.33	**0.61**	0.48	0.47	0.50
rockový	0.49	**0.68**	0.49	0.57	0.35
street	0.56	0.78	0.77	0.65	**0.82**
priemer	0.51	**0.72**	0.62	0.61	0.58

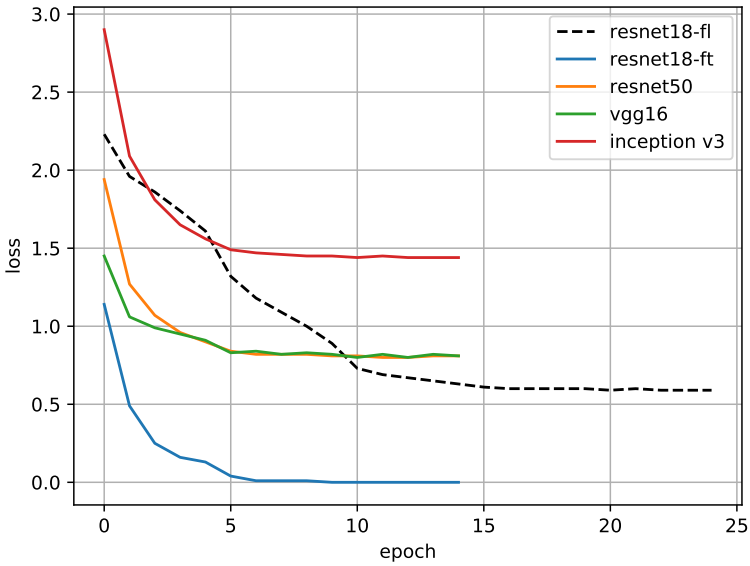
Tabuľka 7.4: Výsledky experimentu rôznych modelov neurónových sietí získané na testovacích dátach datasetu FashionStyle14. Na verejnom rozdelení trénovacej, testovacej a validačnej množiny ukazuje najväčšiu priemernú presnosť sieť Resnet18 s použitím metódy učenia fine-tuning.

model	$resnet18^{ft}$	$resnet18^{*}$	článok [27]
konzervatívny	0.64	0.65	**0.66**
elegantný	**0.95**	0.94	0.91
etnický	0.65	**0.70**	0.74
rozprávkový	0.90	**0.96**	0.88
ženský	0.60	**0.76**	0.64
večerný	0.81	**0.74**	0.74
dievčenský	0.00	**0.60**	0.47
ležérny	**0.71**	0.63	0.66
lolita	**0.94**	0.92	0.92
módny	0.69	**0.76**	0.72
prirodzený	**0.77**	**0.77**	0.70
retro	0.61	**0.64**	0.62
rockový	0.68	**0.72**	0.68
street	0.78	**0.80**	0.69
priemer	0.72	**0.76**	0.72

Tabuľka 7.5: Porovnanie výsledkov na verejnom a náhodne vygenerovanom rozložení trénovacej, testovacej a validačnej množiny, porovnané s výsledkami článku [27]. Stĺpec $Resnet18^{*}$ ukazuje výsledky experimentu vykonaného na vlastnom rozdelení obrázkov. Najlepšie hodnoty sú zvýraznené hrubo.



Obr. 7.1: Zobrazenie vývoja presnosti siete, počas trénovania rôznych modelov na dátovej sade FashionStyle14. Z obrázku vidieť, že trénovanie iba na dátovej sade FashionStyle14 (značka resnet18-fl), trvalo viac iterácií a nedosiahlo rovnakú presnosť ako použitie predtrénovaných sietí (ostatné prípady). Siet resnet18-ft, používala metódu fine-tuning. Trénovaciu množinu sa naučila klasifikovať rýchlejšie a presnejšie ako ostatné siete používajúce featrure-extraction.



Obr. 7.2: Porovnanie vývoja chyby u rôznych modelov, počas trénovania na dátovej sade Fashionstyle14. Najmenšiu chybu vidno u siete Resnet18.

7.3 Dátová sada BUT atribúty chodcov

Tieto experimenty sú zamerané na klasifikáciu atribútov v novo vytvorenej dátovej sade *BUT atribúty chodcov*. V experimentoch som použila model siete a spôsob trénovania na základe výsledkov predošlých experimentov. Z predošlých experimentov sa ukázalo, že najlepšie výsledky na podobných klasifikačných úlohách dosahuje sieť Resnet18 predtrénovaná na dátovej sade ImageNet. Pre klasifikáciu atribútov v dátovej sade *BUT atribútov chodcov* bol model tejto siete mierne upravený. Posledná plne prepojená vrstva modelu Resnet18 bola nahradená troma novými, plne prepojenými vrstvami. Tento model bol už popísaný v kapitole 5.

Experimenty na dátovej sade sú rozdelené do niekoľkých častí. Rozdiel v týchto experimentoch je hlavne vo výbere obrázkov z celej dátovej sady, ktorú tvoria obrázky chodcov vytvorené z viacerých videí. Vo webovej aplikácii bolo anotovaných približne 1100 obrázkov. Tieto obrázky boli rozdelené do troch častí. Do trénovacej časti bolo vybraných 60% obrázkov, do validačnej 5% a do testovacej 35% obrázkov. Následne bolo ešte ku každému anotovanému obrázku pridaných niekoľko ďalších obrázkov tej istej osoby, ale v inej pozícii a v inom čase. Základné experimentovanie zo sieťou a rôznym počtom vybraných obrázkov ukázalo, že výsledky, ktoré sieť dosahuje sa nezlepšujú pridaním väčšieho množstva obrázkov. Obrázky jednej osoby sú príliš podobné a vybranie všetkých obrázkov iba zvyšuje problém s nevyváženosťou dátovej sady.

Všetky získané obrázky

V prvom experimente bolo ku každému anotovanémuo obrázku pridaných 25 ďalších obrázkov. Celkovo tak bolo vybraných 21,639 obrázkov. Pomer jednotlivých tried napríklad pre vek alebo módneho štýlu bol veľmi nevyrovnaný. Vek 19+ tvorilo 49% všetkých anotácií veku a módny štýl ležérny až 62% anotácií módneho štýlu. Pred vstupom do siete bola všetkým obrázkom zmenená veľkosť na 224×224px. Sieť bola trénovaná počas 15 epoch, optimalizačná funkcia bola zvolená funkcia Adam a koeficient učenia bol nastavený na 10^{-3}. Počas učenia boli počítané tri chybové funkcie. Funkcia Cross entropy bola použitá pre výpočet úspešnosti veku a módneho štýlu. Pre výpočet binárneho atribútu vek bola použitá funkcia Binary cross entropy. Celková chyba bola spočítaná ako suma týchto troch chýb podla rovnice 5.5. Parameter α bol nastavený na $\alpha = 0.5$ a β na $\beta = 0.4$.

Sieť sa na takto vytvorenej sade pretrénovala po siedmych epochách. Na trénovacej množina dosiahla presnosť približne 96%, pričom na validačnej množine iba 66%. Nevyváženosť dátovej sady spôsobila, že sieť sa primárne naučila klasifikovať iba najčastejšie vyskytujúce sa atribúty. Pri určovaní módneho štýlu *ležérny* dosiahla úspešnosť 83%. Presnosť ostatných módnych štýlov bola ale veľmi nízka. Rovnako to bolo pre atribút vek, kedy sa sieť naučila klasifikovať vek *19+* s presnosťou 76%. Celková presnosť jednotlivých atribútov vždy presiahla percento výskytu najčastejšej triedu. Výsledky sú zobrazené v tabuľkách 7.6 a 7.7, 7.8.

	1+	19+	30+	45+	60+
presnosť[%]	41.76	76.43	43.51	21.48	42.97

Tabuľka 7.6: Výsledky klasifikácie veku na dátovej sade *BUT Atribúty chodcov*.

	ležérny	šport	rock	street	elegantný	formálny	pracovný
presnosť[%]	83.09	43.30	32.70	18.08	29.23	16.23	0.0

Tabuľka 7.7: Výsledky klasifikácie módneho štýlu na dátovej sade *BUT Atribúty chodcov*. Výsledky ukazujú, že kvôli nevyváženosti dátovej sady sa sieť naučila klasifikovať iba módny štýl *ležérny*.

	pohlavie	vek	módny štýl
najväčšia trieda	62% (muž)	49% (19+)	62% (ležérny)
presnosť[%]	79.95	55.91	63.16

Tabuľka 7.8: Tabuľka ukazuje úspešnosť klasifikácie pohlavia, veku a módneho štýlu. V druhom riadku je zobrazené percento najpočetnejšej triedy pre každý atribút. Vo všetkých prípadoch je dosiahnutá presnosť vyššia ako najpočetnejšia trieda.

Vyvážené atribúty

Cieľom ďalšieho experimentu bolo vybrať z celej dátovej sady vyvážený počet obrázkov pre všetky atribúty a triedy týchto atribútov. Kvôli veľmi nízkemu výskytu módneho štýlu pracovná uniforma bola táto trieda úplne odstránená. Ďalej boli odstránené obrázky, na ktorých boli osoby priradené do triedy veku *19+* alebo *30+* a módny štýl bol klasifikovaný ako *ležérny*. Odobraním týchto obrázkov sa síce znížil celkový počet obrázkov, ale dátovú sadu sa podarilo aspoň čiastočne vyvážiť. Do dátovej sady, na ktorej bola sieť trénovaná, bolo vybraných 20,840 obrázkov, v ktorých boli všetky triedy približne rovnomerne zastúpené. Výsledky experimentu ukázané v tabuľkách 7.9, 7.10 a 7.11, zobrazujú výsledky klasifikácie na vyváženej sade. Porovnaním z prvým experimentom nie sú celkové výsledky lepšie. Znížením počtu výskytov módneho štýlu došlo k zníženiu jeho presnosti klasifikácie. Jednotlivé triedy sú omnoho viac vyvážené a je vidno, že napríklad pri módnom štýle sa podarilo dosiahnuť lepších výsledkov aj u iných tried ako u ležérneho módneho štýlu.

	1+	19+	30+	45+	60+
presnosť[%]	44.76	54.05	53.27	42.57	55.10

Tabuľka 7.9: Výsledky klasifikácie veku na vybranej vyváženej podmnožine dátovej sady *BUT atribúty chodcov*.

	ležérny	šport	rock	street	elegantný	formálny
presnosť[%]	53.84	53.48	35.70	41.39	57.74	34.58

Tabuľka 7.10: Výsledky klasifikácie módneho štýlu na vybranej vyváženej podmnožine dátovej sady *BUT atribúty chodcov*. Výsledky ukazujú, že vyvážením sa poradilo zvýšiť úspešnosť niektorých módnych štýlov.

	pohlavie	vek	módny štýl
presnosť[%]	76.37	50.23	49.21

Tabuľka 7.11: Tabuľka ukazuje úspešnosť klasifikácie pohlavia, veku a módneho štýlu na vybranej vyváženej podmnožine dátovej sady *BUT atribúty chodcov*.

Klasifikácia binárnych atribútov

V dátovej sade *BUT atribúty chodcov* boli zozbierané aj binárne atribúty. Z nich som pre experiment vybrala tri najčastejšie zastúpené atribúty: ruksak, kabelka a čiapka. Okrem týchto troch atribútov som pridala aj pohlavie a triedy veku ako binárne atribúty. Sieť, ktorá klasifikovala tieto atribúty mala rovnakú architektúru ako sieť, na ktorej bola klasifikovaná dátová sada PETA, okrem poslednej vrstvy, ktorá mala iba 9 výstupov, jeden pre každý binárny atribút. Sieť bola predtrénovaná na dátovej sade PETA, ktorá klasifikovala 35 atribútov. Bola použitá chybová funkcia Binary cross entropy loss a optimalizačná funkcia Adam. Sieť sa natrénovala na trénovacej množine za 10 epoch a na validačnej množine dosiahla úspešnosť 86%. Výsledky klasifikácie sú zobrazené v tabuľke 7.12 a ukazujú, že sieť sa najlepšie dokázala naučiť klasifikovať pohlavie a dosiahla približne rovnakú presnosť ako siete použité v predošlých experimentoch. Pri klasifikácii veku dokázala najlepšie klasifikovať triedy *19+* a *60+*. Trieda veku *60+* dosahovala najlepšiu úspešnosť aj v experimentoch s dátovou sadou PETA.

	mA [%]
vek 1+	64.48
vek 19+	71.15
vek 30+	62.88
vek 45+	55.89
vek 60+	66.26
pohlavie	80.87
čiapka	62.39
ruksak	54.41
kabelka	51.31
priemer	63.29

Tabuľka 7.12: Výsledky klasifikácie deviatich binárnych atribútov na dátovej sade BUT atribúty chodcov.

Zhrnutie výsledkov na dátovej sade BUT atribúty chodcov

Vykonané experimenty na dátovej sade BUT atribúty chodcov, nedosahujú vo väčšine atribútov dostatočne dobré výsledky. V experimentoch sa prejavuje niekoľko problémov, ktoré sú zodpovedné za nízku úspešnosť klasifikácie.

Prvý z problémov je, že pri navrhnutých sieťach dochádza veľmi rýchlo k pretrénovaniu, čo znamená, že vytvorená dátová sada nie je dostatočne veľká na dosiahnutie lepších výsledkov. Ukázalo sa, že aj napriek vytvoreniu obrázkov z videí, sú obrázky príliš podobné a ich použitie prináša len mierne zlepšenie. Jednou z možností ako zlepšiť výber viacerých obrázkov k anotovanému obrázku z webovej aplikácie, je navrhnúť spôsob ako vybrať niekoľko, čo najrozdielnejších obrázkov toho istého chodca.

Druhý problém je vyváženosť dátovej sady. V dátovej sade sa nachádzalo veľmi veľa obrázkov osob patriacich do rovnakej triedy, v dôsledku čoho siete klasifikovali iba najpočetnejšie triedy. Jedným z riešení je vytvoriť ďalšie videá, na rôznych miestach, ktoré by zaznamenávali napríklad viac detí, dôchodcov, športovcov, rockerov a iných.

Ďalší problém, ktorý pravdepodobne prispel k nízkej úspešnosti pri klasifikácii veku a módno štýlu bolo zlé rozdelenie tried a ich anotovanie. Je pravdepodobné, že napríklad kvôli nízkemu rozlíšeniu obrázkov alebo maskovaniu tvárí, užívatelia, ktorí vytvárali anotácie neboli schopní správne určiť vek, čím mohlo v dátovej sade vzniknúť veľké množstvo chýb a z tohoto dôvodu sa sieť nie je schopná naučiť rozdiel medzi triedou *19+ 30+* a *45+*. Rovnako je to pre módny štýl, kde boli niektoré triedy veľmi podobné. Niekedy bolo ťažko rozhodnúť aký rozdiel je medzí módnym štýlom ležérny, street alebo elegantný. Možným riešením tohoto problémy je rozvrhnúť triedy presnejšie na základe presne stanovených doplnkov alebo častí oblečenia.

Na experimentoch sa ukázalo, že konvolučné neurónové siete je možné používať pri klasifikácii atribútov chodcov. Dôkazom sú výsledky získané pri klasifikácii pohlavia, ktoré dosahujú viac ako 80%. Pohlavie bol jediný atribút, ktorý bol v dostatočnom počte a rozložení zastúpený v dátovej sade. Pre získanie lepších výsledkov je potrebné pri budúcom vývoji primárne rozšíriť a vylepšiť dátovú sadu.

Kapitola 8

Záver

Práca bola zameraná na klasifikáciu rôznych charakteristík chodcov, ako sú napríklad pohlavie, vek, rôzne časti oblečenia alebo doplnky. V prvej kapitole boli popísané spôsoby klasifikácie atribútov na viacerých dátových sadách, použitím rozličných spôsobov klasifikácie. Jednou z popisovaných metód sú konvolučné neurónové siete, ktoré boli pre klasifikáciu vybrané v tejto práci. V prvej časti sú zároveň popísané existujúce dátové sady, z ktorých boli pre experimenty vybrané dve dátové sady PETA[4] a FashionStyle14 [27].

Okrem týchto dvoch dátových sád bola pri experimentoch použitá aj nová dátová sada BUT atribúty chodcov, ktorej tvorba bola súčasťou tejto práce. Pre vytvorenie tejto dátovej sady bola vytvorená webová aplikácia a niekoľko skriptov. Webová aplikácia slúži na zber anotácií k obrázkom, ktoré sú vytvárané z videí pomocou skriptov. Vytvorené skripty zároveň slúžia aj k spracovaniu anotácií z webovej aplikácie a vytvoreniu celej dátovej sady. Použitím týchto skriptov je jednoduché vytvárať nové obrázky, ktoré môžu byť anotované vo webovej aplikácii, a tým dátovú sadu jednoducho rozširovať. Vytvorená dátová sada obsahuje obrázky chodcov vo vonkajšom prostredí a obsahuje anotácie atribútov pohlavie, vek a módny štýl, ktorý nie je anotovaný v žiadnej existujúcej dátovej sade chodcov.

Pre klasifikáciu atribútov v dátových sadách boli implementované tri modely konvolučných neurónových sietí založené na architektúre Resnet18. Experimenty na existujúcich dátových sadách ukázali, že neurónové siete sú vhodným nástrojom pre klasifikáciu viacerých atribútov alebo módneho štýlu. Podarilo sa mi dosiahnuť veľmi podobné alebo mierne lepšie výsledky ako dosiahli autori vedeckých publikácií popísaných v kapitole 2.

Experimenty na vytvorenej dátovej sade BUT atribúty chodcov už nedosahovali dostatočne dobré výsledky. Ukázalo sa, že vytvorená dátová sada nie je dostatočne veľká a rovnomerne rozložená. Hlavným problémom bolo zozbieranie dostatočne kvalitných obrázkov a spoľahlivých anotácií. Videá, z ktorých boli vytvárané obrázky, museli byť natáčané za dobrého a teplého počasia. Zbierať obrázky pre anotáciu módneho štýlu vo vonkajšom prostredí bolo nemožné v zlom alebo zimnom počasí. Kvalita obrázkov tvorila veľmi dôležitú súčasť zberu anotácií pre určovanie veku.

Výsledky práce ukazujú, že konvolučné siete je možné používať pre získavanie charakteristík chodcov. Na výsledky klasifikácie má výrazný vplyv práve dátová sada. Do budúcna je možné zväčšiť a vylepšiť vyváženosť dátovej sady BUT atribúty chodcov, k čomu môže byť jednoducho použitá webová aplikácia a skripty vytvorené v tejto práci. Malými úpravami je tiež možné rozšíriť webovú aplikáciu o zber ďalších módnych štýlov alebo atribútov definujúcich módny štýl. Vytvorená dátová sada môže byť použitá pri rôznych úlohách pre zisťovanie ďalších charakteristík chodcov.

Literatúra

[1] Achanta, R.; Shaji, A.; Smith, K.; aj.: SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, ročník 34, č. 11, november 2012: s. 2274–2282, ISSN 0162-8828, doi:10.1109/TPAMI.2012.120.

[2] Cao, Z.; Hidalgo, G.; Simon, T.; aj.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018.

[3] Chen, Y.; Duffner, S.; Stoian, A.; aj.: Pedestrian Attribute Recognition with Part-based CNN and Combined Feature Representations. In *International Conference on Computer Vision Theory and Applications*, január 2018, s. 114–122, doi:10.5220/0006622901140122.

[4] DENG, Y.; Luo, P.; Loy, C. C.; aj.: Pedestrian Attribute Recognition At Far Distance. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, New York, NY, USA: ACM, 2014, ISBN 978-1-4503-3063-3, s. 789–792, doi:10.1145/2647868.2654966.
URL http://doi.acm.org/10.1145/2647868.2654966

[5] Deng, Y.; Luo, P.; Loy, C. C.; aj.: Learning to Recognize Pedestrian Attribute. *CoRR*, ročník abs/1501.00901, 2015, 1501.00901.
URL http://arxiv.org/abs/1501.00901

[6] Donahue, J.; Jia, Y.; Vinyals, O.; aj.: DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. *arXiv preprint*, ročník 32, október 2013.

[7] Fabbri, M.; Calderara, S.; Cucchiara, R.: Generative adversarial models for people attribute recognition in surveillance. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, august 2017, s. 1–6, doi:10.1109/AVSS.2017.8078521.

[8] Fong, R.; Vedaldi, A.: Interpretable Explanations of Black Boxes by Meaningful Perturbation. *CoRR*, ročník abs/1704.03296, 2017, 1704.03296.
URL http://arxiv.org/abs/1704.03296

[9] He, K.; Zhang, X.; Ren, S.; aj.: Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jún 2016, ISSN 1063-6919, s. 770–778, doi:10.1109/CVPR.2016.90.

[10] Hidalgo, G.: *OpenPose Demo - Output*. [Online; navštívené 13.5.2019].
URL https://github.com/CMU-Perceptual-Computing-Lab/openpose/blob/master/doc/output.md#pose-output-format-coco

[11] Huang, G.; Liu, Z.; Weinberger, K. Q.: Densely Connected Convolutional Networks. *CoRR*, ročník abs/1608.06993, 2016, 1608.06993.
URL http://arxiv.org/abs/1608.06993

[12] Krizhevsky, A.; Sutskever, I.; E. Hinton, G.: ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*, ročník 25, január 2012, doi:10.1145/3065386.

[13] Layne, R.; Hospedales, T.; Gong, S.: Person Re-identification by Attributes. In *BMVC*, ročník 2, január 2012, doi:10.5244/C.26.24.

[14] Li, D.; Chen, X.; Huang, K.: Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, november 2015, ISSN 2327-0985, s. 111–115, doi:10.1109/ACPR.2015.7486476.

[15] Li, D.; Zhang, Z.; Chen, X.; aj.: A Richly Annotated Pedestrian Dataset for Person Retrieval in Real Surveillance Scenarios. *IEEE Transactions on Image Processing*, ročník 28, č. 4, apríl 2019: s. 1575–1590, ISSN 1057-7149, doi:10.1109/TIP.2018.2878349.

[16] Liu, Z.; Luo, P.; Qiu, S.; aj.: DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jún 2016, ISSN 1063-6919, s. 1096–1104, doi:10.1109/CVPR.2016.124.

[17] Luo, P.; Wang, X.; Tang, X.: Pedestrian Parsing via Deep Decompositional Network. In *2013 IEEE International Conference on Computer Vision*, december 2013, ISSN 1550-5499, s. 2648–2655, doi:10.1109/ICCV.2013.329.

[18] Maji, S.; Berg, A. C.; Malik, J.: Classification using intersection kernel support vector machines is efficient. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, jún 2008, ISSN 1063-6919, s. 1–8, doi:10.1109/CVPR.2008.4587630.

[19] Oquab, M.; Bottou, L.; Laptev, I.; aj.: Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jún 2014.

[20] Scholkopf, B.; Smola, A. J.: *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001, ISBN 0262194759.

[21] Simonyan, K.; Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv 1409.1556*, september 2014.

[22] Sudowe, P.; Spitzer, H.; Leibe, B.: Person Attribute Recognition with a Jointly-Trained Holistic CNN Model. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, december 2015, s. 329–337, doi:10.1109/ICCVW.2015.51.

[23] Sudowe, P.; Spitzer, H.; Leibe, B.: Person Attribute Recognition with a Jointly-Trained Holistic CNN Model. *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2015: s. 329–337.

[24] Sun, G.-L.; Wu, X.; Chen, H.-H.; aj.: Clothing Style Recognition using Fashion Attribute Detection. In *EAI Endorsed Transactions on Ambient Systems*, ročník 2, august 2015, doi:10.4108/icst.mobimedia.2015.259089.

[25] Szegedy, C.; ; ; aj.: Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jún 2015, ISSN 1063-6919, s. 1–9, doi:10.1109/CVPR.2015.7298594.

[26] Szegedy, C.; Vanhoucke, V.; Ioffe, S.; aj.: Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jún 2016, ISSN 1063-6919, s. 2818–2826, doi:10.1109/CVPR.2016.308.

[27] Takagi, M.; Simo-Serra, E.; Iizuka, S.; aj.: What Makes a Style: Experimental Analysis of Fashion Prediction. In *Proceedings of the International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[28] Tome, D.; Monti, F.; Baroffio, L.; aj.: Deep Convolutional Neural Networks for pedestrian detection. *Signal Processing: Image Communication*, ročník 47, október 2015, doi:10.1016/j.image.2016.05.007.

[29] Viola, P.; Jones, M.: Robust Real-Time Face Detection. *International Journal of Computer Vision*, ročník 57, máj 2004: s. 137–154, doi:10.1023/B:VISI.0000013087.49260.fb.

[30] Wang, J.; Zhu, X.; Gong, S.; aj.: Attribute Recognition by Joint Recurrent Learning of Context and Correlation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, október 2017, ISSN 2380-7504, s. 531–540, doi:10.1109/ICCV.2017.65.

[31] Zheng, S.; Yang, F.; Kiapour, M. H.; aj.: ModaNet: A Large-Scale Street Fashion Dataset with Polygon Annotations. *CoRR*, ročník abs/1807.01394, 2018, 1807.01394. URL http://arxiv.org/abs/1807.01394

[32] Zhu, J.; Liao, S.; Lei, Z.; aj.: Pedestrian Attribute Classification in Surveillance: Database and Evaluation. In *2013 IEEE International Conference on Computer Vision Workshops*, december 2013, s. 331–338, doi:10.1109/ICCVW.2013.51.

Príloha A

Obsah priloženého DVD

- `datagetherer/` – zdrojové kódy webovej aplikácie pre zber anotácií pre dátovú sadu BUT atribúty chodcov
- `experiments/` – výsledky získané z experimentov
- `nets/` – zdrojové kódy implementovaných neurónových sietí a skriptov potrebných pre úpravy dátových sád
- `tex/` – zdrojové súbory tejto písomnej práce v jazyku LaTeX
- `video/` – priečinok s videom prezentujúcim výsledky práce
- `dp_xstude22.pdf` – text písomnej práce vo formáte PDF