PEDESTRIAN DETECTION IN IMAGE BY MACHINE LEARNING

Martin Tilgner
Master Degree Programme (2.), FEEC BUT
E-mail: xtilgn00@stud.feec.vutbr.cz

Supervised by: Karel Horák
E-mail: horak@feec.vutbr.cz

Abstract: This work deals with pedestrian detection via convolutional neural network which can be used in autonomous car driving systems to improve travel safety. The work focuses on the influence of the training dataset on the resulting network behavior. The Faster R-CNN with ResNet 101 as backbone network and the SSDLite with MobileNet v2 as backbone network meta-architectures were selected for parameter testing. Both networks achieved real-time detection while accuracy was 61.92 % for the Faster R-CNN and 31.72 % for the SSDLite.

Keywords: Object detection, Convolutional Neural Network, Machine learning, Faster R-CNN, SSDLite

1 ÚVOD

Umělá inteligence zažívá v posledních letech velký rozvoj, a to hlavně díky dostatku dat a výraznému zvýšení výpočetního výkonu. Umělá inteligence nachází uplatnění v medicíně, na sociálních sítích, v dopravě, ale i vojenských aplikacích a dalších oborech. V oboru zpracování obrazu detektory využívající umělou inteligenci a strojové učení překonávají klasické přístupy pro detekci objektů [1].

Tato práce se zabývá použitím konvolučních neuronových sítí pro detekci chodců z pohledu autonomního vozidla pomocí palubní kamery. Zejména pak vlivem trénovacího datasetu na výsledné chování sítě. Motivací této práce je zvýšit bezpečnost dopravy nasazením systému, který využívá umělou inteligenci pro detekci chodců ať pro plně autonomní řízení nebo pouze pro zobrazování rozšířené reality, která usnadní řidiči orientaci v daném prostředí.

Práce si klade za cíl najít optimální postup při trénování konvolučních neuronových sítí - vybrat vhodnou meta-architekturu, vhodně zvolit velikost a rozmanitost datasetu použitého pro učení, testování a validaci a následně vhodně zvolit parametry samotné neuronové sítě. Vzhledem k tomu, že je nutné projít vícerozměrný stavový prostor, nepřichází řešení hrubou silou v úvahu. Proto procházíme jednotlivé osy stavového prostoru postupně. Výsledek tohoto hledání tedy bude pseudo-optimální řešení.

2 DETEKCE OBJEKTŮ POMOCÍ KONVOLUČNÍCH NEURONOVÝCH SÍTÍ

Na základě literatury [1], [2] byly pro testování vybrány modifikace architektur Faster R-CNN [3] a SSD: Single Shot Multibox Detector (SSD) [4]. Meta-architektura Faster R-CNN patří mezi nejpřesnější detektory, její nevýhodou jsou ale velké výpočetní nároky, což se projeví zejména na rychlosti. SSD naopak vyniká v rychlosti, nicméně nedosahuje přesnosti Faster R-CNN. Z pohledu autonomního auta je ovšem důležitá jak rychlost, tak i přesnost detekce, neboť i jedna false negative detekce při autonomním řízení může mít fatální následky.

3 DATASET

Protože se jedná o učení s učitelem, použitý trénovací dataset má zásadní vliv na chování sítě. Je nutné co nejlépe obsáhnout možné situace, aby síť dosáhla dostatečného stupně generalizace. Pro natrénování sítě je v tomto případě převážně využito veřejně dostupných datasetů Kitti [5] a City Shapes, respektive CityPersons [6]. Oba datasety se zaměřují na dopravní situaci a jsou vytvořeny pomocí palubní kamery umístěné v osobním automobilu. Pro lepší generalizaci byly tyto datasety doplněny snímky z datasetu Pascal VOC [7]. Nicméně tyto datasety neobsahují noční snímky. Z tohoto důvodu byl vytvořen noční dataset zaměřený na chodce. Dataset byl ručně anotován v souladu s formátem Pascal VOC. Na obrázku 2 je znázorněn výběr snímků z použitých datasetů.

Obrázek 1: Příklad trénovacího datasetu. Vlevo snímek z Kitti datasetu, vpravo z CityPersons.

4 EXPERIMENTY

Jednotlivé sítě byly implementovány v jazyce Python za využití frameworku Tensorflow a knihovny Object Detection Api. Vzhledem k požadavku na rychlost bylo sníženo rozlišení feature extractoru u Faster R-CNN tak, aby se snažil zachovat poměr velikosti snímku, nicméně maximální velikost může dosáhnout 640 px, minimální velikost pak 340 px. Rozlišení feature extraktou u SSDLite [8] je pak fixně nastavena na 300 x 300 px. Protože trénování i testování probíhalo na běžném notebooku s low end GPU MX 150, byla snížena hodnota batch size pro fázi učení na hodnotu 1 pro Faster R-CNN a na hodnotu 8 pro SSDLite. Záleží tedy, v jakém pořadí jsou snímky při učení předkládány. Jako páteřní síť feature extraktoru je pro Faster R-CNN použito sítě Resnet101 [9] a pro SSDLite MobileNet v2 [8].

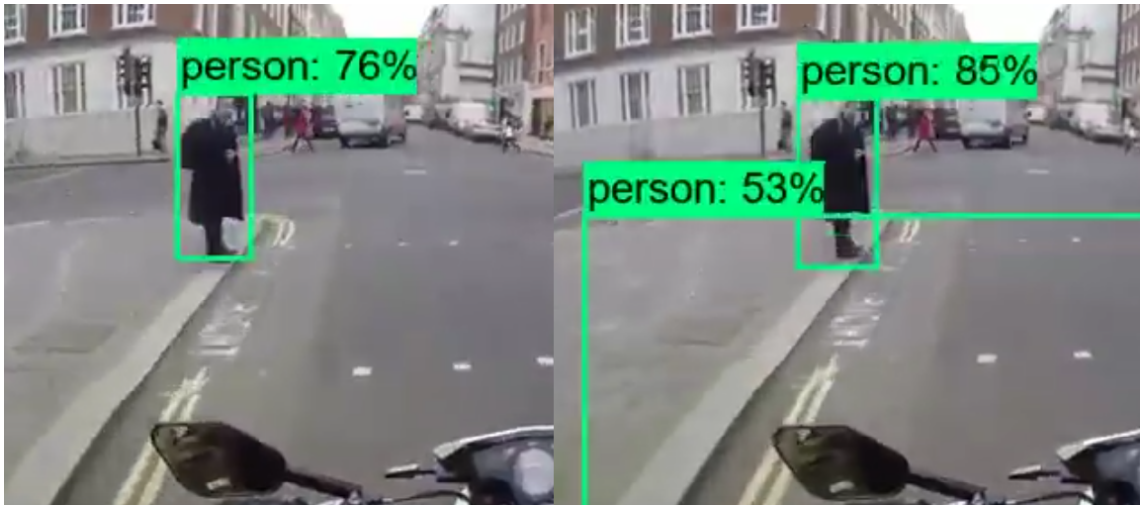
Následně byla vyhodnocena přesnost naučených sítí na Kitti datové sadě a CityPersons pomocí metriky používané v Pascal VOC. Získané výsledky jsou zaneseny v tabulce 1. Je vidět, že obě sítě dosahují lepších výsledků na Kitti datasetu, výrazně horších pak na CityPersons datasetu. Tento rozdíl může být způsoben tím, že na snímcích CityPersons je výrazně více lidí na snímek a je zde výrazně více anotací lidí, kteří jsou z velké části zakryti jinými objekty. Toto tvrzení bude dále v práci zkoumáno pomocí matice záměn tzv. confusion matrix.

Při testování rychlosti pro zpracování videa z webkamery a videa uloženého na disku síť Faster R-CNN dosahovala průměrné rychlosti 1.2 snímku za sekundu (včetně načtení, samotné detekce a uložení snímku s vyznačenými bounding boxy), samotná detekce je schopná zpracovat 1.5 snímku za sekundu. U sítě SSDLite rychlost celé detekční smyčky dosahovala v průměru 17.8 snímků za sekundu, samotná detekce je pak schopná zpracovat v průměru 45.3 snímků za sekundu. Obě sítě tak dosáhly detekce v reálném čase, nicméně Faster R-CNN ještě nedosahuje rychlosti, která by překonala lidské oko. Tento problém lze samozřejmě řešit výkonnějším hardwarem.

Na obrázku 2 je zobrazena detekce pomocí sítí SSDLite. Vlevo je detekce sítě natrénované na mixu

Meta-architektura CNN	Kitti mAP [%]	CityPersons mAP [%]
SSDLite KO	28.26	2.459
SSDLite MD	31.72	8.585
Faster R-CNN KO	61.92	21.45
Faster R-CNN MD	61.54	27.14

Tabulka 1: Vyhodnocení přesnosti detekce pro sítě Faster R-CNN a SSDLite. KO = trénováno na Kitti datasetu, MD = trénováno na mixovaném datasetu.



Obrázek 2: Detekce pomocí SSDLite. Vlevo mixed dataset, napravo trénováno pouze na Kitti.

datasetů, vpravo je síť trénovaná pouze na Kitti datasetu. Je zřejmé, že různorodost datasetu zásadně snižuje False Positive detekce - zde například detekce motorky jako osobu. Dále je vidět, že osoby v dáli nejsou detekovány, což je problém i u sítí Faster R-CNN. Tento problém lze odstranit větším rozlišením feature extraktoru a vstupních snímků. Dále se ukazuje, že je nutné trénovat síť i na nočních snímcích, neboť sítě natrénované pouze na snímcích pořízených ve dne nedokáží rozpoznat špatně osvětlené osoby.

5 ZÁVĚR

Tato práce se zabývá testováním meta-architektur konvolučních neuronových sítí pro detekci osob v dopravním prostředí z pohledu autonomního vozidla. V současné době byly vytvořeny trénovací a testovací datové sady, včetně nástrojů pro práci s nimi. Dále byly natrénovány architektury Faster R-CNN a SSDLite na těchto datasetech. Natrénované sítě jsou schopny zpracovávat snímky v reálném čase. Dosavadní práce zatím ukázala, že jednotlivé sítě nedokáží překonat člověka ať už z pohledu rychlosti (Faster R-CNN) nebo v kvalitě detekce (SSDLite).

V další části vývoje bude podrobněji rozpracováno vyhodnocování přesnosti vytvořením matice záměn, díky které budou více patrné nedostatky jednotlivých sítí, což pomůže v rozhodnutí, jaké části architektur upravit pro reálné využití. V poslední fázi bude vytvořen nástroj pro automatizovanou tvorbu datasetu.

REFERENCE

[1] ZHANG, Liliang, Liang LIN a Xiaodan LIANG. Is faster R-CNN doing well for pedestrian detection?. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* [online]. Springer Verlag, 2016, **9906**, 443-457 [cit. 2019-01-23]. DOI: 10.1007/978-3-319-46475-6_28. ISBN 9783319464749. ISSN 03029743. Dostupné z: https://link.springer.com/chapter/10.1007%2F978-3-319-46475-6_28

[2] YANG, Dongming, Jiguang ZHANG a Shibiao XU. Real-time pedestrian detection via hierarchical convolutional feature. *Multimedia Tools and Applications* [online]. New York: Springer US, 2018, **77**(19), 25841-25860 [cit. 2019-01-23]. DOI: 10.1007/s11042-018-5819-6. ISSN 1380-7501. Dostupné z: https://link.springer.com/article/10.1007%2Fs11042-018-5819-6

[3] REN, Shaoqing, Kaiming HE, Ross GIRSHICK a Jian SUN. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* [online]. IEEE Computer Society, 2017, **39**(6), 1137-1149 [cit. 2019-01-23]. DOI: 10.1109/TPAMI.2016.2577031. ISSN 01628828. Dostupné z: https://arxiv.org/abs/1506.01497

[4] LIU, Wei, Dragomir ANGUELOV, Dumitru ERHAN, Christian SZEGEDY, Scott REED, Cheng-Yang FU a Alexander C. BERG. SSD: Single Shot MultiBox Detector. *Arxiv.org* [online]. , 17 [cit. 2019-01-23]. Dostupné z: https://arxiv.org/abs/1512.02325

[5] GEIGER, A, P LENZ a R URTASUN. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition* [online]. Chicago: IEEE, 2012, s. 3354-3361 [cit. 2019-03-14]. DOI: 10.1109/CVPR.2012.6248074. ISBN 9781467312264.

[6] *City Shapes Dataset: Semantic Understanding of Urban Street Scenes* [online]. Germany: City Shapes, 2016 [cit. 2019-01-23]. Dostupné z: https://www.cityscapes-dataset.com/

[7] *The PASCAL Visual Object Classes Homepage* [online]. United Kingdom: Pascal VOC, 2005 [cit. 2019-01-23]. Dostupné z: http://host.robots.ox.ac.uk/pascal/VOC/

[8] SANDLER, Mark, Andrew HOWARD, Menglong ZHU, Andrey ZHMOGINOV a Liang-Chieh CHEN. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In: *The IEEE Conference on Computer Vision and Pattern Recognition* [online]. 2018, 13.1.2018, s. 14 [cit. 2019-03-28]. ISBN 4510-4520. Dostupné z: https://arxiv.org/abs/1801.04381

[9] HE, Kaiming, Xiangyu ZHANG, Shaoqing REN a Jian SUN. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* [online]. IEEE Computer Society, 2016, **2016-**, 770-778 [cit. 2019-01-23]. ISBN 9781467388511. ISSN 10636919. Dostupné z: https://arxiv.org/abs/1512.03385