SOUHRN

SUMMARY

PODĚKOVÁNÍ

OBSAH

[SOUHRN 1](#_Toc196478106)

[SUMMARY 1](#_Toc196478107)

[PODĚKOVÁNÍ 2](#_Toc196478108)

[OBSAH 3](#_Toc196478109)

[1 ÚVOD 5](#_Toc196478110)

[2 LITERÁRNÍ ČÁST 6](#_Toc196478111)

[2.1 Detekce objektu pomocí umělé inteligence 6](#_Toc196478112)

[2.1.1 Detekce objektu s využitím hlubokého učení 6](#_Toc196478113)

[2.2 Umělé neuronové sítě 7](#_Toc196478114)

[2.3 Konvoluční neuronové sítě 8](#_Toc196478115)

[2.4 Two-stage detektory 11](#_Toc196478116)

[2.4.1 Region-based convolutional neural network (R-CNN) 11](#_Toc196478117)

[2.4.2 Fast R-CNN 12](#_Toc196478118)

[2.4.3 Faster R-CNN 12](#_Toc196478119)

[2.5 One-stage detektory 12](#_Toc196478120)

[2.5.1 SSD 13](#_Toc196478121)

[2.6 You Only Look Once (YOLO) 13](#_Toc196478122)

[2.6.1 Algoritmus YOLO 14](#_Toc196478123)

[2.6.2 Vývoj YOLO 16](#_Toc196478124)

[2.6.3 Velikosti YOLO 18](#_Toc196478125)

[2.6.4 Výstup YOLO algoritmu 19](#_Toc196478126)

[2.7 Srovnání one-stage vs. two-stage algortimů 19](#_Toc196478127)

[3 EXPERIMENTÁLNÍ ČÁST 20](#_Toc196478128)

[3.1 Dataset 20](#_Toc196478129)

[3.1.1 Manuální anotace 21](#_Toc196478130)

[3.1.2 Semi-automatické anotace 21](#_Toc196478131)

[3.1.3 Rozložení datasetu 21](#_Toc196478132)

[3.1.4 Křížová validace 22](#_Toc196478133)

[3.2 Trénování modelu 23](#_Toc196478134)

[3.2.1 Počet epoch 23](#_Toc196478135)

[3.2.2 Patience 23](#_Toc196478136)

[3.2.3 Rychlost učení 23](#_Toc196478137)

[3.2.4 Batch size 24](#_Toc196478138)

[3.2.5 Augmentace 24](#_Toc196478139)

[3.2.6 Testování velikostí YOLO modelů 24](#_Toc196478140)

[3.2.7 Testování hyperparametrů 25](#_Toc196478141)

[4 VÝSLEDKY A DISKUSE 27](#_Toc196478142)

[5 ZÁVĚR 28](#_Toc196478143)

[LITERATURA 29](#_Toc196478144)

# ÚVOD

# LITERÁRNÍ ČÁST

## Detekce objektu pomocí umělé inteligence

Detekce objektu v obrazu je stejně jako sledování objektů, segmentace nebo klasifikace obrazu součásti interdisciplinárního oboru počítačového vidění, jež se rozkládá na pomezí informatiky a umělé inteligence. Hlavním cílem detekce objektu je identifikace určitého objektu a zároveň určení jeho přesné polohy v obraze. Tato technologie má rozsáhlé využití v oborech autonomního řízení vozidel či rozpoznávání tváří, využívá se také v bezpečnostních systémech a mnoha dalších oblastech. Další využití nacházejí metody detekce objektu v obrazu v lékařské diagnostice [1]. Tímto odvětvím, konkrétně detekcí hlasivek ve snímcích z laryngoskopických vyšetření se zabývá i tato práce.

Pro nalezení a identifikaci objektů lze použít dva základní přístupy: konvenční metody a metody hlubokého učení využívající konvoluční neuronové sítě. Původní konvenční přístup pracuje ve třech krocích. V první fázi je vybrán region výskytu objektu algoritmem posuvného okna, následně dochází k extrakci příznaků určeného místa a klasifikaci konkrétního objektu. Tento přístup naráží na velkou výpočetní náročnost a nízkou přizpůsobivost [2], proto od něj bylo v posledních letech upuštěno na úkor metod hlubokého učení.

### Detekce objektu s využitím hlubokého učení

Metody pro detekci objektů pomocí hlubokého učení využívají modely založené na konvolučních neuronových sítí (CNN) k rozeznání jednotlivých příznaků obrazu a následné detekci všech v něm nacházejících se objektů. Tento přístup je charakteristický rozdělením na tři hlavní sekce algoritmu [2].

* Extrakce příznaků, kdy je vstupní obraz zpracován pomocí CNN, která detekuje klíčové rysy obrazu, jako jsou hrany, tvary či složitější textury a sestaví mapu příznaků obrazu.
* Lokalizace objektů z mapy příznaků určí místo pravděpodobných výskytů objektů.
* Klasifikace přidá lokalizovanému objektu třídu, která udává, o jaký typ objektu se jedná.

Metody využívající neuronové sítě můžeme dále rozdělit podle typu algoritmu do dvou základních skupin na two-stage a one-stage detektory.

## Umělé neuronové sítě

Princip umělých neuronových sítí (ANN) je inspirován funkcí neuronového systému v mozku člověka. Základní jednotkou algoritmu ANN je stejně jako v lidském mozku neuron, který v případě umělé sítě provádí matematické operace (viz. **Obr. 2.1**). Vstupem do neuronu je vektor , na který je aplikován skalární součin s maticí vah , následně je přičten bias . Hodnota je vypočítána podle rovnice (2.1) [3].

(2.1)

Pomocí aktivační funkce je vypočten konečný výstup (2.2), který následně může být použitý jako vstup do dalších umělých neuronů [3].

(2.2)

Obsah obrázku snímek obrazovky, diagram, řada/pruh, text

Popis byl vytvořen automaticky

**Obr. 2.1**: Schéma neuronu ANN (převzato z [4])

Trénink ANN se skládá ze tří částí. První částí je dopředný průchod sítí (feedforward), skládající se z vrstev, kde každá vrstva obsahuje určitý počet neuronů. V případě plně propojených vrstev je každý neuron dané vrstvy svázán se všemi neurony z vrstvy předchozí. Na začátku tréninku jsou náhodně či pomocí specifických metod inicializovány biasy a vstupní váhy neuronů. Pro ukázku je na **Obr. 2.2** zobrazen průchod plně propojenou ANN se dvěma skrytými vrstvami, s počtem neuronů a jednou výstupní vrstvou s . Pro obraz velikosti 28 × 28 je vytvořen vstupní vektor 784 × 1, vstupní vrstva má tudíž 784 hodnot. Ty vstupují do vrstvy, kde je rovnicí (2.1) vypočítána hodnota pro každý neuron a následně se pomocí aktivační funkce (2.2) získá výstupní hodnota . Každá tato hodnota je použita jako vstup do všech neuronů následující vrstvy (**Obr. 2.2**) [3].

Obsah obrázku text, diagram, snímek obrazovky, řada/pruh

Popis byl vytvořen automaticky

**Obr. 2.2**: Schéma umělé neuronové sítě (převzato a upraveno z [3])

Ve druhém kroku zvaném zpětná propagace (backpropagation) je použita ztrátová funkce, která vypočítá chybu výstupní vrstvy oproti předpokládanému správnému výstupu. Následně je zpětně dopočítáno, jak se chyba šíří při průchodu neuronovou sítí a pro každý bias a váhu neuronů všech vrstev je navržena změna [3].

V posledním kroku jsou aktualizovány biasy a váhy neuronů tak, aby hodnota ztrátové funkce byla co nejnižší, tím se dosáhne nejoptimálnější funkce modelu [3].

## Konvoluční neuronové sítě

Nedostatkem ANN je její plochý vstup ve formě vektoru, kvůli němuž neuronová síť není schopna zohlednit prostorové upořádání vstupního obrazu. Problémem ANN algoritmu je i její vysoká propojenost, kvůli které při použití vstupů s větším rozlišením vznikají obrovské matice s hodnotami vah. To řeší přístup konvoluční neuronové sítě (CNN), který vstupy ve formě vektorů převádí do dvou či více dimenzí nazývaných mapy příznaků. Na rozdíl od předchozího neurony v CNN nejsou plně propojeny, ale mají vazbu pouze s několika prostorově blízkými hodnotami v předchozí vrstvě. Stejně jako v ANN i zde jsou vstupní hodnoty skalárně násobeny maticemi hodnot, které se v případě CNN nazývají filtry. Ty mají čtvercový tvar s obvyklými velikostmi stran mezi 1 a 11 [3]. Výstup z konvoluční vrstvy se počítá pomocí diskrétní konvoluce, která se ve dvou dimenzích formálně zapisuje jako…

Při výpočtu hodnot, které jsou vstupem do následující vrstvy se použije mapa příznaků předchozí vrstvy a provede se výpočet konvoluce. Každá hodnota následující vrstvy (tyrkysová na **Obr. 2.3**) je vytvořena konvolucí několika hodnot mapy příznaků z předchozí vrstvy (modrá na **Obr. 2.3**), které jsou váženy a upravovány pomocí zvolených vah a biasů filtru, nastavených v průběhu tréninku modelu. Takových filtrů může být použito více na jednu vrstvu, čímž lze dosáhnout většího počtu map příznaků v jedné vrstvě [3].

Obsah obrázku kostka, design

Popis byl vytvořen automaticky Obsah obrázku kostka, design

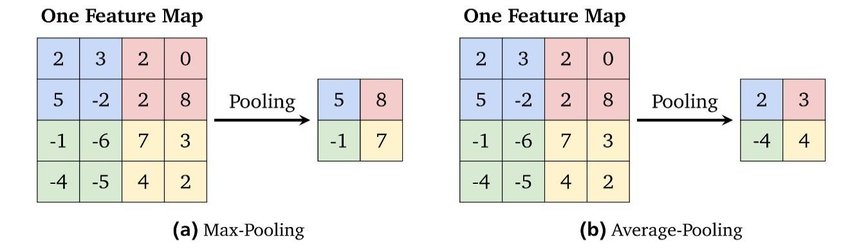
Popis byl vytvořen automaticky Obsah obrázku kostka, design

Popis byl vytvořen automaticky Obsah obrázku kostka, design

Popis byl vytvořen automaticky

**Obr. 2.3**: Schéma použití filtru na mapu příznaků (upraveno a převzato z [5])

Kromě konvolučních filtrů se v CNN používají filtry také pro pooling, což je technika sloužící k zmenšení velikosti mapy příznaků. Pooling filtr daného rozměru vypočítá z obsažené oblasti hodnotu následující vrstvy a posune se o určitý počet kroků. V případě ukázky (**Obr. 2.4**) se filtr velikosti 2 x 2 posouvá vždy o 2 pole, tedy s krokem 2. Nejčastěji používané pooling algoritmy jsou max-pooling (**Obr. 2.4**a), kdy je vybrána nejvyšší ze zkoumaných hodnot a average pooling (**Obr. 2.4**b), jež počítá průměr vybrané oblasti. Pooling vybere nejreprezentativnější část oblasti, čímž umožňuje algoritmu sledování větších textur a útvarů ve vstupním obrazu a zároveň dochází ke snížení paměťové stopy mapy příznaků [3].



**Obr. 2.4**: Schéma max-pooling a average pooling (upraveno a převzato z [23])

Tyto dva postupy jsou kombinovány a opakovaně používány v různých architekturách CNN. Pro ukázku byl vybrán algoritmus neuronové sítě VGG-16 (**Obr. 2.5**), která je využita v řadě one-stage (např. Fast R-CNN a Faster R-CNN) i two-stage (některé varianty Single-shot Detector (SSD)) metod pro detekci objektu [3].

Obsah obrázku text, snímek obrazovky, diagram, design

Popis byl vytvořen automaticky

**Obr. 2.5**: Schéma architektury CNN VGG-16 (převzato z [6])

Vstupující RGB obraz má rozlišení 224 × 244 pixelů. V tabulce (**Tabulka 2.1**) jsou uvedeny parametry všech po sobě jdoucích vrstev této CNN. V poslední vrstvě algoritmu je po pooling operaci použit flattening, který rozvine výstup konvolučních vrstev do 1D vektoru, a následují 3 plně propojené vrstvy.

**Tabulka 2.1**: Struktura architektury CNN VGG-16 (data z [7, 8] )

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Vrstva** | | **Velikost filtru** | **Krok** | **Velikost** | **Počet map příznaků** |
| **Vstup** | Obraz | - | - | 224 × 224 | 3 |
| **1** | 2 × konvoluce | 3 × 3 | 1 | 224 × 224 | 64 |
| **3** | max pooling | 2 × 2 | 2 | 112 × 112 | 64 |
| **4** | 2 × konvoluce | 3 × 3 | 1 | 112 × 112 | 128 |
| **6** | max pooling | 2 × 2 | 2 | 56 × 56 | 128 |
| **7** | 3 × konvoluce | 3 × 3 | 1 | 56 × 56 | 256 |
| **10** | max pooling | 2 × 2 | 2 | 28 × 28 | 256 |
| **11** | 3 × konvoluce | 3 × 3 | 1 | 28 × 28 | 512 |
| **14** | max pooling | 2 × 2 | 2 | 14 × 14 | 512 |
| **15** | 3 × konvoluce | 3 × 3 | 1 | 14 × 14 | 512 |
| **18** | max pooling | 2 × 2 | 2 | 7 ×7 | 512 |
| Flattening | | Převede na 1D vektor | | | |
| **19** | Plně propojená vrstva | - | - | 4096 | - |
| **20** | Plně propojená vrstva | - | - | 4096 | - |
| **21** | Plně propojená vrstva | - | - | 1000 | - |

## Two-stage detektory

Tato skupina detekčních algoritmu, také nazývaná jako region-based detektory dělí proces detekce na lokalizaci objektu a jeho následnou klasifikaci. V prvním kroku algoritmus navrhne několik oblastí zájmu (RoI), které označí jedním z přednastavených referenčních bouding boxů. Následně jsou navržené oblasti přiřazeny k odpovídající třídě objektu a ohraničující boxy jsou upraveny na optimální rozměry [9]. Výhodou této metody je vysoká přesnost detekce, která je ovšem vykoupena velkou časovou náročností.

### Region-based convolutional neural network (R-CNN)

Průkopníkem v oblasti two-stage detektorů je algoritmus Region-based convolutional neural network (R-CNN), který používá CNN k extrakci příznaků z každého regionu zájmu. Metoda používá neuronové sítě jako AlexNet či VGG16, které jsou předtrénovány na velkých datasetech (např. ImageNet) a v průběhu tréninku modelu pouze upravují svoje váhy za nízké hodnoty rychlosti učení [10]. Modely R-CNN pracují ve čtyřech krocích:

Nejprve algoritmus typu Selective Search identifikuje oblasti zájmu v obraze na základě různých tvarů, textur nebo barevných vzorů. Těchto oblastí je vybráno přibližně 2000. V druhém kroku jsou všechny oblasti zájmu přeškálovány na stejnou velikost tak, aby odpovídaly požadovanému rozlišení obrazu vstupujícího do neuronové sítě. Pomocí průchodu přes CNN jsou extrahovány příznaky vstupujících oblastí. Získaný vektor příznaků pak prochází přes algoritmus podpůrných vektorů (SVM), který na jeho základě přiřadí lokalizovanému objektu odpovídající třídu, případně navrženou oblast zavrhne. Po klasifikaci všech nalezených objektů dojde k upřesnění rozměrů jejich ohraničujících bouding boxů pomocí modelu lineární regrese [2].

Ačkoliv se jednalo o průlomovou metodu, která významně přispěla ke zlepšení přesnosti detekce objektů, z dnešního pohledu trpí tento algoritmus mnoha limitacemi zejména v rychlostí detekce [2]. To je způsobeno použitím vícestupňového algoritmu či testováním velkého množství oblastí zájmu. I přes to, že metoda R-CNN používá CNN k extrakci příznaků, klasifikace a regresní kroky pro lokalizaci bouding boxů jsou zprostředkovány prostřednictvím SVM případně jinými algoritmy, proto se nejedná o plně neuronový model. Metoda SVM zároveň při klasifikaci objektů kontroluje oblasti zájmu pro každou třídu jednotlivě, což má také významný dopad na rychlost detekce [11]. Nutno dodat, že při vzniku architektury R-CNN (2014) nebylo použití plně neuronového modelu zdaleka běžné.

### Fast R-CNN

Fast R-CNN přichází s vylepšením z hlediska rychlosti i přesnosti. Na rozdíl od předchozího, tato architektura spojuje tři části: extrakci příznaků, klasifikaci objektu a závěrečnou úpravu bouding boxu do jedné. Zároveň optimalizuje práci s oblastmi zájmu, kdy do neuronové sítě vstupuje celý obraz, který je zpracován jedním průchodem přes CNN a výstupem je společná mapa příznaků. Z této mapy se vyberou selective search metodou sledované oblasti zájmu. Na získaných RoI je aplikován pooling, který zajistí vytvoření fixních délek vektorů příznaků. Ty jsou poslány do plně propojené CNN, která současně klasifikuje třídu objektu a zároveň provádí přesnou lokalizaci využitím softmax vrstvy a lineární regrese [2].

Ačkoliv se jedná o výrazný pokrok oproti architektuře R-CNN, stále se jedná o časově náročný proces zejména kvůli využití konvenčních metod pro vyhledávání oblastí zájmu, jako je algoritmus selective search [2].

### Faster R-CNN

Architektura Faster R-CNN navazuje a vylepšuje předchozí Fast R-CNN. Nahrazuje tradiční časově náročné přístupy pro vyhledávání oblastí zájmu jako selective search nebo MCG (Multiscale combinatorial grouping) pomocí CNN zvané Regional Proposal Network (RPN), která je schopna se v průběhu tréninku učit [2]. Zavedení RPN zrychluje dobu zpracování obrazu z několika sekund na milisekundy [12]. Díky propojení vrstev RPN s neuronovou sítí detekční sekce se zásadně zvyšuje i přesnost detekce a rychlost algoritmu.

## One-stage detektory

One-stage frameworky pro detekci objektu provádí lokalizaci a identifikaci objektu zároveň, použití hluboké konvoluční neuronové sítě. Pomocí tohoto přístupu lze dosahovat zpracování obrazu za mnohem kratší dobu, jelikož dochází pouze k jednomu průchodu vstupu neuronovou sítí, při kterém jsou lokalizovány všechny bounding boxy zároveň. Součástí stejného průchodu algoritmem CNN je také přiřazení hodnoty pravděpodobnosti příslušnosti boxu k určité třídě. Do této skupiny se řadí např. architektury DetectorNet, OverFeat, SSD nebo You Only Look Once (YOLO) [2].

### SSD

Single Shot Multibox Detector (SSD) umožňuje detekci více typů objektů zároveň. Proces lokalizace objektu je inspirován architekturou Faster R-CNN, ze které je převzat mechanismus kotev. Díky tomu může SSD extrahovat příznaky objektů různých velikostí s podobnou přesností jako Faster R-CNN [2]. Algoritmus využívá neuronovou síť VGG-16 [13], která lokalizuje objekty pomocí bounding boxů a zároveň každému boxu přiřazuje pravděpodobnosti příslušností k jednotlivým třídám objektů (nikoliv pravděpodobnost, že se jedná o jakýkoliv objekt). Je tedy třeba za třídu objektu považovat i pozadí jako negativní detekci, aby bylo možné ohraničujícímu rámečku nepřiřadit žádnou konkrétní třídu. Zavrhnutí nevhodně přiřazených objektů zajišťuje metoda non-maximum suppression (NMS). Zejména díky použití RPN dosahuje architektura SSD velmi vysokých detekčních rychlostí při udržení vysokého standardu přesnosti detekce. Pro řadu úloh, v porovnání s algoritmy YOLOv3, Faster R-CNN a dalšími v článku [14], se ale zdá být použití metody SSD nevyhovujícím řešením z důvodu nepřesné detekce malých objektů [13].

## You Only Look Once (YOLO)

You Only Look Once je dalším algoritmem spadajícím do kategorie one-stage detektorů. Základní myšlenka všech verzí YOLO stojí na rozdělení vstupního obrazu na mřížku, kde každá buňka zodpovídá za detekci objektů spadajících svým středem na její území. YOLO používá konvoluční neuronovou síť k predikci všech bouding boxů najednou. K detekci každého objektu tedy jsou využity všechny příznaky vstupujícího obrazu [15].

Podle [1] se jedná o často využívanou architekturu pro detekci objektů v obraze napříč všemi detekčními přístupy, jelikož disponuje řadou konkurenčních výhod. Mezi nejdůležitější patří vysoká rychlost detekce způsobená použitím one-stage přístupu a řadou dalších optimalizačních opatření. To umožňuje YOLO algoritmu rychle detekovat objekty a cíle v živých video přenosech s vysokým rozlišením v reálném čase, což z něj dělá nejvyužívanější detekční metodu v oblastech autonomního řízení či dohlížecích a bezpečnostních kamerových systémů. Modely umělé inteligence pro detekci objektu v obraze lze pomocí YOLO architektury natrénovat na velmi vysokou přesnost, přičemž rychlost procesu je zachována. Díky tomu lze algoritmus aplikovat v dalších oblastech, jako jsou bezpečnostní kontrola přístupu či inteligentní brány, kde může být totožnost osob ověřena například pomocí rozpoznání obličeje nebo SPZ. YOLO se také využívá k ovládání robotů jako nástroj vidění robota, který je poté schopen se bezpečně přemísťovat díky detekování blížících se překážek.

### Algoritmus YOLO

YOLO architektura podléhá nepřetržitému vývoji a zlepšování detekčních schopností. Aktuálně je dostupných 11 verzí algoritmu od původní varianty YOLO až po nejnovější YOLOv11. Následující obecný mechanismus algoritmu je společný pro všechny verze.

V prvním kroku je vstupní obraz proveden přes CNN, pomocí které jsou extrahovány jeho příznaky [16]. YOLO využívá v jednotlivých verzích různé CNN od backbone (část neuronové sítě zodpovědná za extrakci relevantních rysů z různých úrovní rozlišení vstupního obrazu) zvané Darknet v prvotních verzích [13], přes ELAN v YOLOv7 [17] a mnoho dalších.

Získaná mapa příznaků je rozdělena na mřížku. Po průchodu mapy přes plně propojenou CNN každá buňka této mřížky detekuje všechny objekty, jejichž středy spadají do oblasti této buňky. Výstupem každé buňky pak jsou nalezené bouding boxy ohraničující objekty a k nim náležící pravděpodobnosti. Každou takovou hodnotou model vyjadřuje pravděpodobnost, že se jedná o nějaký objekt a zároveň jistotu přesnosti určení polohy objektu [16].

Díky mřížkové metodě vznikne řada redundantních bounding boxů způsobených mnohočetnou detekcí jednoho objektu různými buňkami či falešných detekcí objektů s nízkou pravděpodobností. Z tohoto důvodu přichází na řadu algoritmus NMS, který vyřazuje bounding boxy s nízkou šancí na přítomnost objektu. Zároveň na základě metriky Intersection over union (IoU) porovnává, zda se jedná o vícečetnou detekci, či rozdílné objekty. V případě mnohonásobné detekce ponechá bounding box s nejvyšším confidence score [16].

Intersection over Union (IoU)

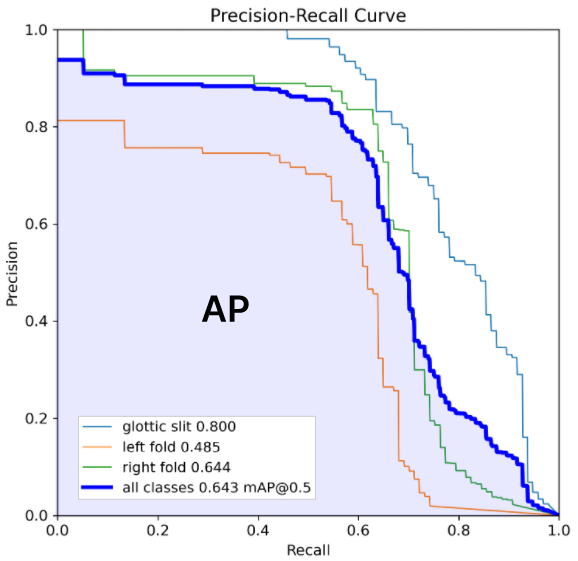
je metrika založená na porovnání překryvu ploch 2 útvarů. V algoritmu YOLO se využívá v průběhu procesu detekce při použití NMS k porovnání obsahů bounding boxů pocházejících z různých detekcí nebo při statistickém vyhodnocení spolehlivosti modelu pomocí metrik mean Average Precision (mAP). V tomto případě se porovnává míra plochy překrytí referenčního ohraničujícího boxu z datasetu s plochou bounding boxu detekovaného objektu. je vypočteno z rovnice (2.3), kde je obsah plochy průniku dvou oblastí a značí obsah útvaru sjednocujícího tyto dvě plochy [18].

(2.3)

Z rovnice (2.3) je zřejmé, že musí náležet . Hodnota je pak obvykle porovnána s prahovou hodnotou a tento výsledek rozhoduje o úspěšnosti detekce.

Mean Average Precision (mAP)

Metrikou mean Average Precision (mAP) je měřena výkonost natrénovaného modelu. Model je vyzkoušen na testovací části datasetu a ze vzniklých bounding boxů a ground truth boxů z databáze jsou vypočteny IoU. Úspěšnost detekce je obvykle vyhodnocena na mnoha hladinách . Následující postup pak probíhá pro každé zvlášť. Nejprve je na základě výsledků úspěšných a neúspěšných detekcí sestavena závislost zvaná precision-recall curve, kde precision udává procento správně identifikovaných objektů vůči všem predikcím a recall značí procentuální počet správně nalezených objektů modelem ze všech ground truth boxů (**Obr. 2.6**). Plocha pod touto křivkou je definována jako average precision a je vypočtena stejně jako precision-recall curve pro každou třídu objektu zvlášť. Zprůměrováním všech vzniká výsledná hodnota [19].



**Obr. 2.6**: Ukázka Precision-Recall curve a plochy AP

mAP@50

Běžné používanou metrikou pro vyhodnocování modelů umělé inteligence je , kde označení udává, že za úspěšnou detekci se považuje taková predikce objektu, jejíž překryvu s ground truth boxem je alespoň 50 %.

mAP@[50:95]

Další obvyklou metrikou je , která postupně vypočítá hodnoty na deseti prazích úspěšnosti a z průměrů dílčích hodnot vypočítá výslednou . Jedná se tedy o přísnější metriku oproti , kvůli vyšším nárokům na predikce modelu.

### Vývoj YOLO

Účelem nových verzí YOLO je zvýšení výkonosti detekčních schopností algoritmu oproti verzi předešlé. Hlavním rozdílem mezi verzemi je použití rozdílné architektury neuronové sítě, která se liší téměř v každé variantě. V průběhu vývoje YOLO dochází k častým změnám v pojetí ztrátových funkcí, které ovlivňují průběh tréninku modelu. Každá verze disponuje ztrátovou funkcí vytvořenou často na míru pro dosažení co nejlepších výsledků.

Ztrátová funkce (loss function)

Ztrátová funkce je využívána v průběhu trénování modelu. Měří rozdíl mezi aktuální predikcí modelu a správnými detekcemi zprostředkovanými pomocí informací z datasetu. Ztrátová funkce se skládá z 3 hlavních částí. Ztráta lokalizací měří rozdíl mezi predikovanými bounding boxy a referenčními boxy z databáze. Ztráta důvěryhodnosti bere v potaz rozdíl mezi předpokládaným a skutečným confidence score detekovaných objektů. Ztráta klasifikace udává rozdíl mezi klasifikací modelu a správnými třídami objektů. Tyto funkce jsou sečteny a vyváženy příslušnými koeficienty (viz rovnice 2.4). Na základě vypočtené hodnoty je rozhodováno o dalším průběhu tréninku modelu [20].

Ztráta klasifikace

Ztráta důvěryhodnosti

Ztráta lokalizace

|  |  |
| --- | --- |
|  | (2.4) |

Od YOLOv2 je přidána metoda kotvových boxů. Model při detekci objekt ohraničí jedním z přednastavených kotvových boxů o pevné velikosti i poměru stran a poté predikuje posuny boxu k určení přesné polohy objektu [21]. Od YOLOv5 je použita metoda dynamických kotvových boxů, kdy si model vytváří rozměry předdefinovaných boxů v průběhu tréninku jako nejpravděpodobnější tvary objektů vyskytujících se v datasetu [22]. Důležitou změnou byla také implementace konceptu Feature Pyramid Networks (FPN) ve verzi YOLOv3 [21].

Feature pyramid networks (FPN)

FPN slouží jako metoda pro detekci objektů různých rozlišení skládající se ze dvou částí bottom-up a top-down cesty. Bottom-up cesta je standardní CNN, která vytvoří vrstvy různého rozlišení (např. C2, C3, C4, C5), kde se zvyšující se vrstvou klesá prostorové rozlišení, ale zvyšuje se sémantická hodnota. Konvoluční vrstvy jsou použity k vytvoření odpovídajících map příznaků (P2, P3, P4, P5). Nejvyšší vrstva C5 je převedena konvolucí na mapu příznaků P5. Ta je díky vysoké sémantické hodnotě schopna detekovat největší objekty. Následně je použita konvoluce na vrstvu C4 a k vzniklé mapě příznaků je přičtena P5 nadvzorkovaná na odpovídající velikost. Tím vzniká mapa příznaků P4. Obdobně dochází ke vzniku ostatních map příznaků. Každá mapa je pak schopna detekovat objekty jiných velikostí (viz **Obr. 2.6**) [23].

Obsah obrázku text, diagram, snímek obrazovky, Paralelní

Popis byl vytvořen automaticky

**Obr. 2.7:** Schéma algoritmu feature pyramid network (převzato z [23])

### Velikosti YOLO

Ultralytics nabízí několik velikostí modelu YOLO (v nejnovějších verzích obvykle n – nano, s – small, m – medium, l – large a x – extra large). S velikostí modelu se zvyšuje schopnost rozlišovat složitější struktury v obraze, zároveň ale výrazně stoupá časová i výpočetní náročnost jak při tréninku modelu, tak při samotném detekčním procesu. Současně platí, že pro menší velikosti datasetů není třeba využívat velké modely. Kvůli nedostatečnému množství trénovacích dat není využit potenciál složitější architektury neuronové sítě a výsledný model pak dosahuje obdobných, ne-li horších detekčních schopností při mnohem vyšších hardwarových i časových nárocích.

V článku [24] byly testovány rozdíly velikostí modelů YOLOv5 natrénovaných za stejných podmínek na datasetu o 10 000 položkách pocházejících z COCO datasetu s rozložením 80/20, kde 80 % dat patřilo trénovací části, ostatní materiál byl umístěn do validační sekce. V **Tabulka 2.2** jsou zobrazeny výsledky tohoto měření. Trénink i detekce objektu probíhali v prostředí Google Colab s odpovídacím výpočetním výkonem.

**Tabulka 2.2**: Porovnání výkonosti velikostí modelů YOLOv5 (převzato a upraveno z [24])

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Dataset** | **mAP@50** | **mAP@[50:95]** | **Doba detekce**  **[ms]** | **GFLOPS** |
| YOLOv5s | výběr z datasetu COCO | 38,3 | 23,6 | 27 | 17,0 |
| YOLOv5m | 43,7 | 28,7 | 32 | 51,3 |
| YOLOv5l | 46,8 | 31,5 | 41 | 115,4 |
| YOLOv5x | 48,5 | 32,8 | 49 | 218,8 |

Dále bylo srovnáno použití modelů pro detekci videa na Google Colab a několika variantách zařízení IPhone 12. Z grafu (**Obr. 2.7**) závislosti počtu snímků za sekundu (fps) na velikosti modelu a použitém zařízení je zřejmé, že při použití hardwarově slabšího zařízení je třeba zvolit jednoduší detekční algoritmus pro udržení dostatečné rychlosti pro detekci objektů v reálném čase i za cenu snížení přesnosti detekce.

Obsah obrázku text, snímek obrazovky, diagram, řada/pruh

Popis byl vytvořen automaticky

**Obr. 2.8**: Závislost rychlosti detekce v fps na použité velikosti modelu a zařízení (data z [24])

### Výstup YOLO algoritmu

Výstupem po zpracování algoritmem je pro každý objekt nalezený v obrazu bounding box ohraničující nalezený objekt pomocí opsaného obdélníku. Ke každé lokalizaci náleží také třída, která udává, o jaký typ objektu se jedná. Dále je ke každému objektu přiřazeno confidence score, které vypovídá jak o pravděpodobnosti detekce objektu správné třídy, tak o jistotě správného určení polohy objektu.

## Srovnání one-stage vs. two-stage algortimů

Porovnání výše uvedených přístupů není snadnou záležitostí. Spolehlivost každého algoritmu lze měnit různým nastavením jeho hyperparamertů, zároveň žádný model není nejefektivnějším řešením pro všechny úlohy. Každý z uvedených modelů má své využití v jeho aplikační oblasti, různým úlohám tedy vyhovují různé detekční algoritmy. V článku [25] byla srovnána přesnost a rychlost rozdílných one-stage a two-stage přístupů, z nichž vybrané modely byly uvedeny v **Tabulka 2.3**. Je třeba mít na paměti, že výsledky mohou být do určité míry ovlivněny diskutovanými faktory.

**Tabulka 2.3**: Srovnání one-stage vs. two-stage algoritmů (převzato a upraveno z [25])

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | | **mAP@50** | **FPS** |
| two-stage | R-CNN | - | 0,03 |
| Fast R-CNN | 39,9 | 0,5 |
| Faster R-CNN | 42,7 | 7 |
| one-stage | SSD512 | 46,5 | 19 |
| YOLOv4 608x608 | 65,7 | 23 |

# EXPERIMENTÁLNÍ ČÁST

Cílem práce je vytvoření modelu umělé inteligence schopného detekovat objekty ve snímcích z laryngoskopického vyšetření. Laryngoskopie je lékařská metoda sloužící k diagnostice oblasti hrtanu. Používá se v případě problému s dýcháním, chronickým kašlem, problémy s hlasem nebo při přítomnosti zánětů a nádorů v okolí hlasivek. Postiženému pacientovi je do hrtanu zavedena kamera, která pořídí videozáznam okolí hlasivek, následně dochází k posouzení snímané oblasti lékařem. Toto vyhodnocení je velmi subjektivní, závisí na znalostech a schopnostech lékaře a zejména v hraničních případech se může v závislosti na posuzující osobě diagnóza hlasivek lišit.

Detekce hlasivek využívající metod zpracování obrazu pomocí umělé inteligence má za cíl posloužit jako první krok k objektivizaci laryngoskopického vyšetření. Na základě vytvořeného modelu umělé inteligence je v budoucnu možné posuzovat vady (např. nedomykavost chlopní hlasivek) počítačově. V návaznosti na tuto práci lze v budoucnu vytvořit další uměle inteligentní modely detekující nádory a záněty vyskytující se na hlasivkách. Softwarovým vyhodnocením v průběhu laryngoskopie by došlo kromě objektivizace také ke značnému urychlení průběhu vyšetření. Vyhodnocení by mohlo probíhat v reálném čase přímo v průběhu vyšetření.

## Dataset

Jako vstupní dataset byly použity několika sekundové záznamy z laryngoskopických vyšetření, kde každé video pochází z vyšetření jiného pacienta. Dataset byl získán ve spolupráci s Fakultní nemocnicí Královské Vinohrady, která poskytla potřebné obrazové materiály pro analýzu. Do datasetu byli zahrnuti jak zdraví pacienti, tak pacienti s určitým postižením hlasivek. Videa byla pořízena ze dvou různých kamer, což zvyšuje variabilitu datasetu a poskytuje záznamy v různých kvalitách rozlišení. Tato videa byla rozdělena na snímky, ve kterých byly vyznačeny jednotlivé části hlasivek.

Celý dataset se skládá laryngoskopických videozáznamů oblasti hrtanu různých délek (do 60 sekund) pocházejících od 20 pacientů. Tato videa byla pořízena s frekvencí 30 snímků za sekundu na celkem 12 991 snímků. V každém byla oanotována levá hlasivka (červená barva), pravá hlasivka (modrá barva) a glotická štěrbina (žlutá barva). Zároveň byly do datasetu zahrnuty obrázky, na nichž je zobrazena jenom část hlasivky, nebo snímky, na kterých není hlasivka viditelná vůbec. Díky tomu model nutně nevyhledává části hlasivek v každém snímku a získává schopnost správně zpracovávat obrazy bez přítomnosti hlasivky. Jedná se o vítanou vlastnost, jelikož během kamerového záznamu z laryngoskopického vyšetření není v každém momentu viditelné celé hlasivkové ústrojí.

### Manuální anotace

Tvorba datasetu se skládala ze dvou částí. V první části byly snímky anotovány manuálně pomocí webové aplikace LabelStudio. Pro označování částí hlasivek byly použity polygonální anotace, kdy každý hledaný objekt byl obtažen uzavřenou lomenou čárou ve formě nepravidelného n-úhelníku. Z hlediska použití datasetu k účelu detekce objektů pomocí algoritmu YOLOv11 nemají polygonální anotace oproti použití obdélníkových boxů žádnou výhodu, poslouží ale k reprodukovatelnosti datasetu při jeho použití k trénování segmentačních modelů. Manuálně bylo označeno přibližně 3 500 snímků.

### Semi-automatické anotace

Druhá část byla anotována semi-automatickým způsobem pomocí modelu trénovaného na manuálně vytvořené části datasetu. V programovacím jazyce Python byl vytvořen skript pro automatickou anotaci laryngoskopických snímků skládající se ze tří částí. V první fázi jsou snímky z videozáznamu jednoho pacienta označeny dříve vytvořeným YOLO modelem natrénovaným na manuálně vytvořené části datasetu. Zde jsou části hlasivek označeny pomocí obdélníkových bounding boxů vystupujících z algoritmu YOLO. V další části jsou tyto nově anotované snímky manuálně zkontrolovány a rozřazeny na úspěšné anotace, které lze přidat do datasetu a neúspěšné anotace. Ty se nahrají do webové aplikace LabelStudio a provede se jejich manuální anotace. Využití semi-automatických anotací podstatně zefektivňuje a usnadňuje tvorbu datasetu.

### Rozložení datasetu

Pro vytvoření finálního datasetu jsou všechny snímky převedeny do varianty obdélníkových anotací a dataset je rozdělen do 3 skupin.

#### Trénovací data

Trénovací data obsahují největší část datasetu (cca 80 % všech obrázků). Na těchto datech se model v průběhu tréninku učí a zlepšuje pomocí optimalizace vah a biasů.

#### Validační data

Validační data (cca 10 % snímků) jsou částí datasetu určenou ke sledování výkonosti modelu během tréninku. Model tyto data nevidí přímo v průběhu učení, tj. neoptimalizuje na nich parametry, ale po ukončení každé epochy na nich jako na nezávislých datech vyhodnotí přesnost modelu. Hlavním důvodem přítomnosti validační části v datasetu je částečné předejití přeučení modelu (viz 3.1.3.1). Pokud se model v průběhu epoch zlepšuje na trénovacích datech, ale predikce validačních dat se zhoršují, trénink modelu je ukončen.

#### Testovací data

Trénovací data (asi 10 % datasetu) jsou část datasetu, která se během trénování vůbec nepoužívá a slouží až k finálnímu vyhodnocení modelu. Výhodou těchto dat je jejich úplná nezávislost na trénovacím procesu, což umožňuje objektivnější a realističtější hodnocení modelu.

Při tvorbě datasetu se i přes rozdělení videí zachovává příslušnost každého snímku k jeho pacientovi. Následně se dataset dělí do zmíněných 3 skupin přibližným poměrem 80 – 10 – 10 a všechny snímky každého pacienta se vyskytují vždy jen v jedné části datasetu. Toto rozložení zajišťuje, že si model nezapamatuje konkrétní rysy hlasivek určitých pacientů, jejichž snímky by se vyskytli jak v trénovací, tak ve vyhodnocovací části datasetu, ale učí se obecné rysy společné pro každé hlasivkové ústrojí. Zároveň trénování a vyhodnocování modelu probíhá na zcela odlišných pacientech a přináší realističtější výsledky, které lépe simulují reálné použití modelu v praxi na neznámé hlasivky nového pacienta. Toto rozdělení respektující jednotlivé pacienty poskytuje realističtější generalizaci modelu, snižuje možnost přeučení na jednotlivých pacientech a poskytují nenadhodnocené modelu, kvůli sdílení informací mezi částmi datasetu, díky čemuž jsou metriky modelu robustnější a důvěryhodnější.

#### Přeučení

Přeučení je jev, při kterém dochází k nadměrnému přizpůsobení k trénovacím datům, kdy model vykazuje velmi dobré výsledky na vlastním trénovacím datasetu. Při jeho použití na neznámá data se ale výkonnost modelu razantně zhoršuje. K přeučení dochází především tehdy, pokud architektura modelu obsahuje více parametrů, než kolik je možné vzhledem k poskytnutým datům smysluplně odůvodnit. [26].

### Křížová validace

Při použití rozložení datasetu po pacientech závisí vyhodnocovací metriky (zejména při malém datasetu) na výběru konkrétních pacientů do části validačních a testovacích dat. Pokud by do vyhodnocovací množiny připadly snímky pacientů náročnějších na predikci, výsledné metriky by mohly podhodnotit skutečnou schopnost modelu generalizovat. Naopak při zařazení jednodušších případů by mohly metriky modelu nadhodnocovat výsledky. Pro zajištění objektivního zhodnocení byla výsledná výkonost modelu získaná metodou křížové validace. Při ní je vybrána a oddělena skupina testovacích dat, zbylá část datasetu se rozdělí v tomto případě na 9 podobně objemných částí. Poté se vytvoří devět verzí datasetu, které použijí postupně každou část jako validační data a do trénovacích dat zahrnou zbylých 8 částí. Vznikne tedy 9 obdobných datasetů se stejnými testovacími daty, ale jiným rozložením snímků mezi validační a trénovací částí. Po natrénovaní modelu na každém z těchto datasetů se jako finální výkonost modelu použije metrika spočítaná zprůměrováním jednotlivých výsledků.

## Trénování modelu

K vytvoření modelu pro detekci částí hlasivek v obrazech z laryngoskopického vyšetření byl použit algoritmus YOLO, konktrétně různé varianty YOLOv11. Model byl trénován velikostech YOLOv11n, YOLOv11s, YOLOv11m a YOLOv11l. Pro správné natrénování modelu je třeba vhodně nastavit řadu hyperparametrů, jejichž hodnoty mohou značně ovlivnit výslednou výkonnost modelu. Dále jsou v pro ukázku vysvětleny některé hlavní z testovaných hyperparametrů.

### Počet epoch

Jedna epocha představuje kompletní průchod dat a zpětnou propagaci přes neuronovou síť. Počet epoch tedy udává, kolikrát model projde celým trénovacím datasetem. V průběhu epocho dochází k optimalizaci vah a biasů modelu, který se postupně optimalizuje na míru předloženým datům. Při nastavení nedostatečného počtu epoch model nemá dostatek času naučit se správně reprezentovat data, při nastavení příliš vysokého počtu epoch může docházet k přeučení na trénovací data. V této práci byla zjištěna jako optimální hodnota počtu epoch 300 s přidáním doplňujícího hyperparametrů patience.

### Patience

Hyperparametr patience je použit v kombinaci s počtem epoch. Slouží k ukončení tréninku v případě, že se model po určitou dobu nezlepšuje na validačních datech. Parametr pomáhá zabránit přeučení modelu a zároveň šetří výpočetní čas. Pro model hlasivek byla hodnota patience nastavena na 50, trénink tedy byl v případě stagnace výkonosti během posledních 50 epoch ukončen.

### Rychlost učení

Při aktualizaci vah a biasů je vypočítaná změna vynásobena velikostí hyperparametru rychlosti učení. Ten udává, o jak velký krok se mají parametry modelu změnit ve směru vypočítané změny. Při nastavení příliš nízké rychlosti učení model konverguje k optimálnímu řešení příliš pomalu, vyžaduje velký výpočetní čas nebo k ideálnímu výsledku vůbec nedojde. Případně se může ustálit v lokálním minimu ztrátové funkce, které není nejlepším řešením. Pokud je rychlost učení příliš vysoká, velké změny v parametrech se projeví nestabilitou tréninku a model k žádnému řešení nekonverguje.

### Batch size

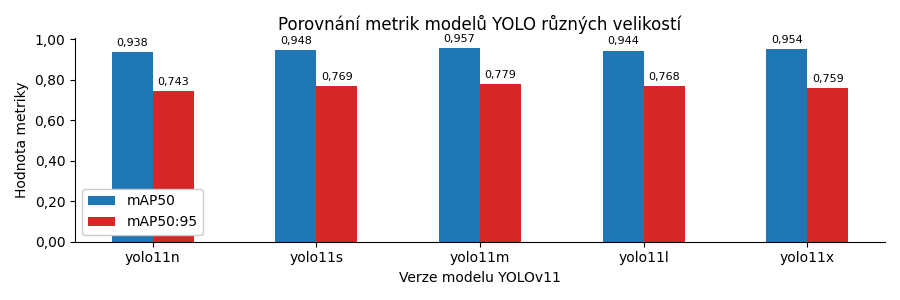
Batch size určuje, kolik obrázku z datasetu se zpracuje najednou při jednom dopředném průchodu neuronovou sítí, než se znovu aktualizují parametry modelu. Projití celého datasetu zároveň by bylo příliš paměťově náročné. Při procházení snímků jednotlivě by se zvyšovala výpočetní náročnost a nestabilita tréninku kvůli častým aktualizacím vah. Batch size je tedy kompromisem mezi těmito extrémy, optimalizuje výkon modelu i dobu tréninku v závislosti na dostupném výpočetním hardwaru.

### Augmentace

Jedná se o skupinu hyperparametrů běžně používaných při trénování modelů strojového učení, která rozšiřuje a obohacuje dataset aplikací řady transformací na snímky. Mezi obvyklé metody augmentace patří např. změna měřítka, rotace, horizontální a vertikální zrcadlení, změna jasu, či přidání šumu. Cílem augmentace je zvýšit generalizační schopnost modelu tím, že se naučí rozpoznávat objekty i v situacích, které nejsou přesně zastoupeny v datasetu. Tyto úpravy obrázků simulují různorodé reálné podmínky a pomáhají se adaptovat na variabilitu dat v praxi, čímž se také snižuje riziko přeučení modelu.

Speciální formou augmentace dat specifickou pro algoritmy YOLO je mozaiková augmentace. Tato technika vybere 4 snímky z trénovacích dat a po použití jiných metod klasické augmentace jako oříznutí, změna měřítka atd. je spojí do jednoho obrázku formou mřížky o rozměrech 2 × 2. Mozaiková augmentace zvyšuje robustnost modelu, protože umožňuje trénovat model na nezvyklých kompozicích či při částečném zakrytí objektů, zároveň zvyšuje rozmanitost pozadí. Zároveň umožňuje efektivní trénink i při nutnosti nastavení nižší hodnoty batch size.

### Testování velikostí YOLO modelů

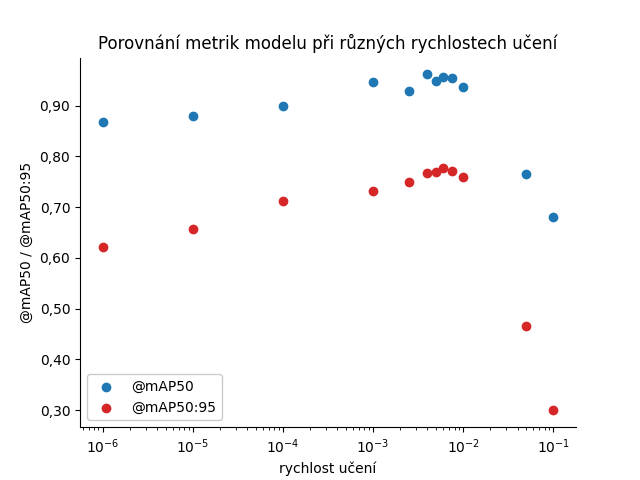
K vytvoření dobře fungujícího modelu je třeba zvolit vhodnou velikost YOLO modelu, proto bylo vytvořeno 5 modelů všech dostupných velikostí (n, s, m, l a x) YOLO11 a jejich výsledné metriky a byly porovnány na **Obr. 3.1**. Trénování všech modelů probíhalo na stejném datasetu (viz 3.1) za identického nastavení hyperparametrů. Z grafu je patrné, že výsledné metriky různých velikostí modelu dosáhly velmi podobné úrovně výkonnosti. Všechny testované varianty modelů tedy vykazují schopnost relativně přesně a konzistentně reprezentovat vlastnosti trénovacích dat. Nejlepších hodnot metrik a však dosáhl model YOLOv11m, což odpovídá očekávání, neboť tato velikost modelu představuje rovnováhu mezi složitostí architektury a množstvím dostupných trénovacích dat (přibližně 13 000 snímků). Tento výsledek podporuje hypotézu, že volba modelu přiměřené velikosti vzhledem k datovému objemu je klíčová pro dosažení optimálního výkonu.**Obr. 3.1**: Porovnání metrik rozdílných velikostí modelů YOLO11 spočítaných na testovacích datech

### Testování hyperparametrů

Pro dosažení co nejlepší přesnosti modelu bylo testováno velké množství hyperparametrů. Každý z nich byl testován trénováním několika totožných modelů s rozdílem pouze v hodnotě daného parametru. Všechny testy byly prováděny na velikosti modelu YOLO11s a modely byly porovnány na základě nejvyšších dosažených hodnot metrik a v průběhu tréninku. Dále jsou uvedeny výsledky testování vybraných hyperparametrů.

#### Testování rychlosti učení

Rychlost učení byla testována na hodnotách v rozmezí od 10-6 do 10-1 viz (obrázek). Závislost rychlosti učení na výkonnosti modelu vyšla podle předpokladu, kdy pro velmi nízké hodnoty trénink probíhal stabilně, ale v průběhu epoch nedocházelo k dostatečnému zlepšování. U vysokých hodnot docházelo k nestabilitě tréninku, což značně projevilo i na metrikách modelů. Nejlepší hodnoty bylo dosaženo při rychlosti učení 0,005.

Obsah obrázku text, snímek obrazovky, diagram, Písmo

Obsah vygenerovaný umělou inteligencí může být nesprávný.

Obsah obrázku text, snímek obrazovky, diagram, Písmo

Obsah vygenerovaný umělou inteligencí může být nesprávný.Obsah obrázku text, snímek obrazovky, Písmo, diagram

Obsah vygenerovaný umělou inteligencí může být nesprávný.

Obsah obrázku text, snímek obrazovky, Písmo, diagram

Obsah vygenerovaný umělou inteligencí může být nesprávný.Obsah obrázku text, snímek obrazovky, Písmo, diagram

Obsah vygenerovaný umělou inteligencí může být nesprávný.

# VÝSLEDKY A DISKUSE

# ZÁVĚR

LITERATURA

(1) Cong, X.; Li, S.; Chen, F.; Liu, C.; Meng, Y. A Review of YOLO Object Detection Algorithms based on Deep Learning. *Frontiers in Computing and Intelligent Systems* **2023**, *4* (2), 17-20. DOI: 10.54097/fcis.v4i2.9730.

(2) Ravpreet, K.; Sarbjeet, S. A comprehensive review of object detection with deep learning. *Digital Signal Processing* **2023**, *132*.

(3) Tesema, S. N. *Deep Convolutional Neural Network Based Object Detection Inference Acceleration Using FPGA*; Université Bourgogne Franche-Comté, 2022.

(4) Buettgenbach, M. H. *Explain like I’m five: Artificial neurons*. 2021. <https://towardsdatascience.com/explain-like-im-five-artificial-neurons-b7c475b56189> (accessed 2025 25. 1.).

(5) vdumoulin. *conv\_arithmetic*. 2016. <https://github.com/vdumoulin/conv_arithmetic?tab=readme-ov-file> (accessed 2025 27.1.).

(6) Le, K. *An overview of VGG16 and NiN models*. 2021. <https://lekhuyen.medium.com/an-overview-of-vgg16-and-nin-models-96e4bf398484> (accessed.

(7) Adams, J.; Qiu, Y.; Posadas, L.; Eskridge, K.; Graef, G. Phenotypic trait extraction of soybean plants using deep convolutional neural networks with transfer learning. *Big Data and Information Analytics* **2021**, *6*, 26-40. DOI: 10.3934/bdia.2021003.

(8) Yu, J.; Li, J.; Sun, B.; Chen, J.; Li, C. Multiclass Radio Frequency Interference Detection and Suppression for SAR Based on the Single Shot MultiBox Detector. *Sensors* **2018**, *18* (11). DOI: 10.3390/s18114034.

(9) Carranza-García, M.; Torres-Mateo, J.; Lara-Benítez, P.; García-Gutiérrez, J. On the Performance of One-Stage and Two-Stage Object Detectors in Autonomous Vehicles Using Camera Data. *Remote Sensing* **2021**, *13* (1), 89.

(10) Yao, J.; Huang, X.; Wei, M.; Han, W.; Xu, X.; Wang, R.; Chen, J.; Sun, L. High-Efficiency Classification of White Blood Cells Based on Object Detection. *Journal of Healthcare Engineering* **2021**, (23), 1-11. DOI: 10.1155/2021/1615192.

(11) Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, Columbus, OH, USA; 2014.

(12) Ahmed, K.; Ghareh Mohammadi, F.; Matus, M.; Shenavarmasouleh, F.; Pereira, L.; Ioannis, Z.; Amini, M. H. Towards Real-time House Detection in Aerial Imagery Using Faster Region-based Convolutional Neural Network. *IPSI Transactions on Internet Research* **2023**, *19* (2), 46-54. DOI: 10.58245/ipsi.tir.2302.06.

(13) Aziz, L.; Haji Salam, M. S. B.; Sheikh, U. U.; Ayub, S. Exploring Deep Learning-Based Architecture, Strategies, Applications and Current Trends in Generic Object Detection: A Comprehensive Review. *IEEE Access* **2020**, *8*, 170461-170495. DOI: 10.1109/ACCESS.2020.3021508.

(14) Liu, Y.; Sun, P.; Wergeles, N.; Shang, Y. A survey and performance evaluation of deep learning methods for small object detection. *Expert Systems with Applications* **2021**, *172*. DOI: //doi.org/10.1016/j.eswa.2021.114602.

(15) Lavanya, G.; Pande, S. Enhancing Real-time Object Detection with YOLO Algorithm. *EAI Endorsed Transactions on Internet of Things* **2023**, *10*. DOI: 10.4108/eetiot.4541.

(16) Badgujar, C. M.; Poulose, A.; Gan, H. Agricultural object detection with You Only Look Once (YOLO) Algorithm: A bibliometric and systematic literature review. *Computers and Electronics in Agriculture* **2024**, *223*. DOI: <https://doi.org/10.1016/j.compag.2024.109090>.

(17) Yanyun, S.; Liu 刘迪, D.; Chen, J.; Wang, Z.; Wang, Z.; Zhang, Q. On-Board Multi-Class Geospatial Object Detection Based on Convolutional Neural Network for High Resolution Remote Sensing Images. *Remote Sensing* **2023**, *15* (16). DOI: 10.3390/rs15163963.

(18) Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA; Paper 8953982, 2019.

(19) Padilla, R.; Netto, S. L.; Silva, E. A. B. d. A Survey on Performance Metrics forObject-Detection Algorithms. In 27th International Conference on Systems, Signals and Image Processing (IWSSIP), ELECTR NETWORK; 2020.

(20) Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA; 2016.

(21) Kaur, S.; Kaur, L.; Lal, M. A Review: YOLO and Its Advancements. In 6th International Conference on Recent Innovations in Computing, ICRIC 2023, Jammu, Indie; 2024.

(22) Tai, W.; Wang, Z.; Li, W.; Cheng, J.; Hong, X. DAAM-YOLOV5: A Helmet Detection Algorithm Combined with Dynamic Anchor Box and Attention Mechanism. *Electronics* **2023**, *12* (9). DOI: 10.3390/electronics12092094.

(23) Lin, T.-Y.; Ramanauskaitė, S.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In 30TH IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR 2017), Honolulu, HI, USA; 2017.

(24) Dlužnevskij, D.; Stefanovič, P.; Ramanauskaitė, S. Investigation of YOLOv5 Efficiency in iPhone Supported Systems. *Baltic Journal of Modern Computing* **2021**, *9* (3). DOI: 10.22364/bjmc.2021.9.3.07.

(25) Shetty, A. K.; Saha, I.; Sanghvi, R. M.; Save, S. A.; Patel, Y. J. A Review: Object Detection Models. In 2021 6th International Conference for Convergence in Technology (I2CT), Maharashtra, India; 2021.

(26) Paris, G.; Robilliard, D.; Fonlupt, C. Exploring Overfitting in Genetic Programming. In *Artificial Evolution*, Liardet, P., Collet, P., Fonlupt, C., Lutton, E., Schoenauer, M. Eds.; Springer Berlin Heidelberg, 2004; pp 267--277.