

# **Využití kamerového systému pro zajištění bezpečnosti osob na pracovišti**

Use of Surveillance Cameras to Ensure the Safety of People in the Workplace

Bc. Filip Łuński

Diplomová práce

Vedoucí práce: Ing. Tomáš Wiszczor, Ph.D.

Ostrava, 2025



# Zadání diplomové práce

Student:

**Bc. Filip Łuński**

Studijní program:

N0613A140034 Informatika

Téma:

Využití kamerového systému pro zajištění bezpečnosti osob na  
pracovišti

Use of Surveillance Cameras to Ensure the Safety of People in the  
Workplace

Jazyk vypracování:

čeština

Zásady pro vypracování:

Cílem práce je vytvořit a otestovat prototyp systému pro detekci incidentů na pracovišti pomocí analýzy pohybů a pozic osob v reálném čase na kamerových záznamech. Systém bude využívat algoritmy strojového učení po detekci a klasifikaci incidentů jako například pády, volání o pomoc nebo jiné kritické situace.

1. Prostudujte a popište dostupné algoritmy pro detekci a klasifikaci objektů v obrazech a zhodnoťte jejich použitelnost pro detekci osob v reálném čase.
2. Zmapujte a popište dostupná řešení pro detekci klíčových bodů lidského těla s důrazem na jejich použitelnost pro real-time analýzu pohybu osob.
3. Vybraná řešení pro detekci klíčových bodů lidského těla otestujte s ohledem na rychlosť zpracování, přesnost detekce a jejich použití v reálném čase.
4. Vytvořte řešení využívající vhodné techniky strojového učení, které na základě klíčových bodů detekuje v reálném čase pozici indikující bezpečnostní incident (např. pád nebo volání o pomoc).
5. Výsledný prototyp otestujte s různými vstupními parametry, jako jsou rozšíření kamery, různé prostředí, různé úrovně osvětlení, a porovnejte výkonnost na různém hardware (včetně GPU).

Seznam doporučené odborné literatury:

- [1] Sultana, Farhana, Abu Sufian and Paramartha Dutta. "A Review of Object Detection Models based on Convolutional Neural Network." ArXiv abs/1905.01614 (2019): n. pag.
- [2] Redmon, Joseph, Santosh Kumar Divvala, Ross B. Girshick and Ali Farhadi. "You Only Look Once: Unified, Real-Time Object Detection." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): 779-788.
- [3] Redmon, Joseph and Ali Farhadi. "YOLO9000: Better, Faster, Stronger." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 6517-6525.
- [4] Wang, Chien-Yao, I-Hau Yeh and Hongpeng Liao. "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information." ArXiv abs/2402.13616 (2024): n. pag.
- [5] Liu, W., Dragomir Anguelov, D. Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu and Alexander C. Berg. "SSD: Single Shot MultiBox Detector." European Conference on Computer Vision (2015).
- [6] Cao, Zhe, Gines Hidalgo, Tomas Simon, Shih-En Wei and Yaser Sheikh. "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields." IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (2018): 172-186.

Formální náležitosti a rozsah diplomové práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí diplomové práce: **Ing. Tomáš Wiszczor**

Datum zadání: 01.09.2024

Datum odevzdání: 30.04.2025

Garant studijního programu: prof. RNDr. Václav Snášel, CSc.

V IS EDISON zadáno: 26.11.2024 15:19:39

## **Abstrakt**

Rozvoj technologie v posledních letech, zejména v oblasti umělé inteligence a neuronových sítí, umožňuje komplexnější analýzy chování osob. Dalo by se tak bezpečnostní kamery využít pro zajištění bezpečnosti osob. V této práci je tak navrženo řešení pro detekci pádu osob v obrazovém toku v reálném čase. Řešení je postaveno na kombinaci dvou neuronových sítí. První detekuje všechny osoby v obraze a jejich klíčové body, druhá tyto body klasifikuje do tří *normální* a *upadl*. V práci je popsán výběr detektoru pozby, návrh architektury klasifikační sítě a implementace výsledného detektoru pádu. Ve výsledném řešení je použit model YOLO11-*pose* a rekurentní neuronová síť postavená na architektuře GRU.

## **Klíčová slova**

python, strojové učení, neuronové sítě, konvoluční neuronové sítě, rekurentní neuronové sítě, GRU, LSTM, PyTorch, detekce pozby, detekce chování, detekce pádu, YOLO

## **Abstract**

This is English abstract. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Fusce tellus odio, dapibus id fermentum quis, suscipit id erat. Aenean placerat. Vivamus ac leo pretium faucibus. Duis risus. Fusce consectetur risus a nunc. Duis ante orci, molestie vitae vehicula venenatis, tincidunt ac pede. Aliquam erat volutpat. Donec vitae arcu. Nullam lectus justo, vulputate eget mollis sed, tempor sed magna. Curabitur ligula sapien, pulvinar a vestibulum quis, facilisis vel sapien. Vestibulum fermentum tortor id mi. Etiam bibendum elit eget erat. Pellentesque pretium lectus id turpis. Nulla quis diam.

## **Keywords**

python, machine learning, neural networks, convolutional neural networks, recurrent neural networks, GRU, LSTM, PyTorch, pose estimation, behaviour detection, fall detection, YOLO

# Obsah

<b>Seznam použitých symbolů a zkratek</b>	<b>7</b>
<b>Seznam obrázků</b>	<b>8</b>
<b>Seznam tabulek</b>	<b>9</b>
<b>1 Úvod</b>	<b>10</b>
<b>2 Neuronové sítě</b>	<b>12</b>
2.1 Historie . . . . .	12
2.2 Struktura neuronové sítě . . . . .	13
2.3 Topologie neuronových sítí . . . . .	15
2.4 Proces trénování s využitím backpropagation . . . . .	15
2.5 Optimalizace procesu trénování . . . . .	16
<b>3 Konvoluční neuronové sítě</b>	<b>18</b>
3.1 Konvoluce . . . . .	18
3.2 Konvoluční vrstva . . . . .	19
3.3 Poolovací vrstva . . . . .	19
3.4 Architektura CNN . . . . .	20
<b>4 Rekurentní neuronové sítě</b>	<b>21</b>
4.1 Základní principy . . . . .	21
4.2 LSTM . . . . .	23
4.3 GRU . . . . .	25
<b>5 Představení problematiky detekce pádu</b>	<b>27</b>
5.1 Návrh řešení . . . . .	27
5.2 Trénovací datasety . . . . .	28
5.3 Třídy a jejich anotace . . . . .	29
5.4 Příprava trénovacích dat pro klasifikační algoritmus . . . . .	30

<b>6</b>	<b>Výběr algoritmu pro detekci klíčových bodů</b>	<b>31</b>
6.1	Detekce pózy . . . . .	31
6.2	Detekce klíčových bodů . . . . .	32
6.3	Detekce objektů a osob v obraze . . . . .	33
6.4	Charakteristiky vybraných implementací pro detekci pózy . . . . .	36
6.5	Testování a porovnání vybraných algoritmů pro detekci pózy . . . . .	40
<b>7</b>	<b>Implementace neuronové sítě pro klasifikaci klíčových bodů</b>	<b>45</b>
7.1	Použité technologie . . . . .	46
7.2	Implementace vybraných architektur . . . . .	46
7.3	Návrh architektury a konfigurace sítě . . . . .	50
7.4	Testování . . . . .	53
<b>8</b>	<b>Kompletace finálního řešení detekce pádu</b>	<b>56</b>
8.1	Sledování pózy s YOLO11 . . . . .	56
8.2	Struktura třídy FallDetector . . . . .	57
8.3	Testování detekce pádu ve videu . . . . .	58
<b>9</b>	<b>Závěr</b>	<b>60</b>
<b>Přílohy</b>		<b>65</b>

# Seznam použitých zkrátek a symbolů

AF	– Aktivační funkce
NN	– Neural network – neuronová síť
ANN	– Artificial neural network – umělá neuronová síť
FFNN	– Feedforward Neural Network – dopředná neuronová síť
CNN	– Convolutional neural network – konvoluční neuronová síť
RNN	– Recurrent neural network – rekurentní neuronová síť
LSTM	– Long short-term memory – dlouhá krátkodobá paměť
GRU	– Gated recurrent unit
AI	– Artificial intelligence – umělá inteligence
ML	– Machine learning – strojové učení
DL	– Deep learning – hluboké učení
RoI	– Region of interest – oblast zájmu
PAF	– Part affinity field – pole propojení klíčových bodů
ReLU	– Rectified Linear Unit
PReLU	– Parametric Rectified Linear Unit
Adam	– Adaptive Moment Estimation
BP	– Backpropagation
BPTT	– Backpropagation Through Time – zpětné šíření chyby v čase
BCE	– Binary Cross-Entropy
RMSprop	– Root Mean Square Propagation

# Seznam obrázků

2.1	Model umělého neuronu [10] . . . . .	14
4.1	Unrolling hluboké RNN . . . . .	22
4.2	Jednotka LSTM . . . . .	24
4.3	Rozvinutá hluboká LSTM síť . . . . .	25
4.4	Jednotka GRU . . . . .	26
5.1	Příkladové snímky z datasetů CAUCAFall (nahoře) a 50 Ways to Fall (dole). . . . .	29
6.1	(Vlevo) Topologie klíčových bodů použitá např. v COCO-pose [25] (Vpravo) Příklad detekce pózy pomocí YOLO. . . . .	32
6.2	Architektura původní verze YOLO [33] . . . . .	35
6.3	Architektura SSD [36] . . . . .	36
6.4	Vizualizace sledování osoby mezi dvěma snímky v OpenPifPaf [39] . . . . .	38
6.5	Architektura YOLOv11 [41] . . . . .	39
6.6	Porovnání přesnosti <i>MediaPipe Lite</i> (vlevo) a <i>YOLO Large</i> (vpravo) . . . . .	41
6.7	Příklad špatné detekce bodů v modelu Torchvision Keypoint R-CNN . . . . .	42
6.8	Porovnání přesnosti variant modelu YOLO v situaci s netypickým natočením postavy. Zleva: <i>Nano</i> , <i>Small</i> , <i>Medium</i> , <i>Large</i> , <i>Xlarge</i> . . . . .	44
7.1	Graf přesnosti na validačních datech v průběhu trénování dopředných sítí s různými aktivačními funkcemi . . . . .	51
7.2	Graf validační ztráty v průběhu trénování hlubších dopředných sítí . . . . .	52
7.3	Graf trénovací a validační ztráty (nahoře) a graf validační přesnosti v průběhu trénování výsledné sítě . . . . .	54

# **Seznam tabulek**

6.1	Porovnání výkonu modelu OpenPifPaf . . . . .	41
6.2	Porovnání výkonu modelu MediaPipe BlazePose . . . . .	42
6.3	Výkon modelu Torchvision Keypoint R-CNN . . . . .	42
6.4	Porovnání výkonu modelu YOLO . . . . .	43
7.1	Výsledky testování dopředné a GRU sítě . . . . .	54
8.1	Rychlosť detekcie pádu . . . . .	59

# Kapitola 1

## Úvod

Kamerové systémy jsou využívány již mnoho let a jejich využití je stále širší. Již před několika lety se odhadovalo, že celkový počet bezpečnostních kamer ve světě přesahuje miliardu [1]. Využívány jsou v průmyslu, dopravě, obchodech, veřejných prostorech, zdravotnictví či domácnostech.

Zpočátku bylo možné video sledovat pouze živě, později, s příchodem videokazet, bylo možné záznam sledovat rovněž zpětně. Digitální éra a síťové kamery umožnily přístup ke kamerovým záznamům z libovolného místa na světě. V poslední době se také začalo nahrazovat živé sledování automatickým zpracováním obrazu a detekcí událostí s využitím technik umělé inteligence.

Kamerové systémy se používají zejména ve dvou oblastech: zabezpečení (ang. security), myšleno jako ochrana před úmyslnými hrozbami a protiprávními činy, jako jsou krádeže, poškozování majetku, či neoprávněný vstup; a bezpečnost (ang. safety), což zahrnuje ochranu před nehodami a náhodnými hrozbami, jako jsou pády, požár, úniky nebezpečných látek, či porušování bezpečnostních předpisů.

Jak již bylo zmíněno, lze kamery využívat jednak pro živé sledování, jednak pro záznam a jeho analýzu po události. Kamerové záznamy jsou důležité zejména pro zpětnou analýzu incidentů, důkazní materiál pro soudní spory, zjištování příčin nehod, či pro zlepšení bezpečnostních opatření. Živé sledování videa se pak snaží incidentům přímo předcházet. Bylo však prokázáno, že schopnost lidského pozorovatele detektovat nebezpečí se velmi snižuje s délkou sledování a s počtem monitorovaných kamer [2]. Právě proto se s příchodem technik umělé inteligence začalo využívat automatické zpracování obrazu a detekce hrozeb, nebezpečí, nebo již probíhajících incidentů v jejích počátcích. Tyto techniky pak úplně nahrazují lidského pozorovatele, nebo mu pomáhají včas zpozorovat nebezpečí a zareagovat.

Automatická analýza obrazu je používaná již několik desítek let, většinou ale spíše pro oblast zabezpečení, než pro bezpečnost. To z toho důvodu, že úlohy, jako identifikace neoprávněného vstupu, detekce zbraní, rozpoznávání SPZ nebo podezřelých osob dle obličeje jsou pro algoritmy mnohem jednodušší, než například detekce pádu, nouzové situace či zdravotního problému. Jedná se zde totiž o vysokou míru abstrakce, kdy i člověk může špatně identifikovat některé situace.

Hlavním problémem těchto komplexnějších analýz je vysoká falešná pozitivita, kdy je například těžké rozeznat člověka trénujícího běh od člověka utíkajícího před nebezpečím. Nicméně rozvoj v oblasti hlubokého učení a konvolučních neuronových sítí, jako i vývoj a dostupnost hardwaru podporujícího tyto techniky, umožňuje dneska jejich využití i pro složitější úlohy.

Ve firmě Linde jsou kamerové systémy používány v mnoha průmyslových provozech, nicméně chybí ucelený systém pro automatickou analýzu obrazu a detekci různých druhů nebezpečí. Naším úkolem tedy v budoucnu bude navrhnout a implementovat modulární systém s možností sledování konkrétních nebezpečí na konkrétních místech. Ty budou zahrnovat například detekci pádu, požáru, zdravotních problémů, nebo porušování bezpečnostních opatření. Systém pak bude v případě rozpoznání nějaké hrozby informovat příslušného pracovníka.

V této práci se zaměříme pouze na jednu z těchto úloh, a to na detekci pádu. Pád může mít různé příčiny, ať už je to zdravotní problém jako ztráta vědomí, nebo zakopnutí. Někdy se zdá, že samotné zakopnutí je banální problém, nicméně pokud se na pracovišti nenachází nikdo, kdo by mohl pomoci, a poškozený není schopen sám přivolat pomoc, může vést takový incident k vážným následkům.

V první části práce se budeme zabývat teoretickými základy, jako je obecná architektura neuronových sítí či její pokročilejší architektury jako konvoluční nebo rekurentní neuronová síť. V dalších kapitolách se zaměříme na detekci osob a odhad jejich klíčových bodů. Projdeme si různé přístupy a otestujeme různé algoritmy s ohledem na výkon, možnou hardwarovou akceleraci a preciznost. V další části se budeme zabývat samotnou detekcí pádu, tedy algoritmem, který na základě odhadnutých klíčových bodů určí, zda došlo k pádu, či nikoliv. V závěru práce se zaměříme na otestování výsledného řešení a zhodnocení jeho výkonu.

# Kapitola 2

## Neuronové sítě

Umělá neuronová síť (ang. artificial neural network – ANN) nebo jen neuronová síť (ang. neural network – NN) je výpočetní model inspirovaný biologickými nervovými systémy v lidském mozku. Na rozdíl od konvenčních výpočetních modelů, které zpracovávají informace algoritmicky, a tedy postupují dle předem určeného postupu, se informace v tomto modelu šíří paralelně v síti vah mezi jednotlivými neurony. Jelikož je výstup ze sítě dané architektury závislý hlavně na numerických parametrech, zejména vzhledem jednotlivých spojů mezi neurony, lze funkčnost sítě měnit bez změny programu pouhou změnou těchto parametrů, a to i automaticky v procesu trénování modelu.

Jelikož budou neuronové sítě základem prakticky v každé fázi vývoje našeho řešení, pojďme se podívat na jejich základní principy.

### 2.1 Historie

#### Prvopočátky

První matematický model neuronové sítě byl popsán v roce 1943 dvěma neurofyziology – Warrenem McCullochem a Walterem Pittsem [3]. Model byl založen na síti jednoduchých logických prvků, které provedou vážený součet svých vstupů a na výstup odešlou signál založený na prahové funkci.

V roce 1958 pak Frank Rosenblatt představil elektronický model neuronové sítě. Základní jednotku, postavenou na McCulloch-Pittsově modelu, nazval perceptron. [4] Jeho architektura byla podobná modelu znázorněnému na obrázku 2.1, kde aktivační funkce je prahová funkce. Rosenblattův stroj – Mark I Perceptron – byl postavený pro rozpoznávání jednoduchých vzorů v obrazech. Samotný model perceptronu je dodnes používán jako základ pro mnoho neuronových sítí.

I když ve svých počátcích přitahoval koncept umělé inteligence mnoho vědců jako i sponzorů, v následujících letech zájem ochabl, jelikož nebylo dosaženo předpokládaných výsledků, hlavně s ohledem na tehdejší stav vývoje hardwaru a obecně výpočetní techniky. Proto se tomuto období

někdy říká Ai Winter. Neznamená to ale, že ti, kteří se oboru nadále věnovali, nedosáhli významných výsledků [5].

## Backpropagation

Významným milníkem v historii neuronových sítí byl objev algoritmu backpropagation, zvaného také algoritmus zpětného šíření chyby. Tento algoritmus byl vyvinut v roce 1974 Paulem Werbosem, popularitu ale dosáhl až po nezávislému objevení v roce 1986 Davidem Rumelhartem et al. [6].

Tento algoritmus umožnil trénovat sítě s více vrstvami, což položilo základ hlubokému učení. Algoritmus využívá metodu gradientního sestupu v kombinaci s řetězovým pravidlem derivací k nalezení optimálních vah sítě vedoucích k minimalizaci chyby.

Vynález backpropagation byl jedním z hlavních důvodů, proč se v 80. letech obnovil zájem o neuronové sítě a umělou inteligenci obecně. V roce 1989 taky umožnil Yann LeCunovi et al. efektivní použití konvolučních neuronových sítí pro rozpoznávání rukou psaných číslic [7] a položit tak základ širokému využití konvolučních sítí v oblasti počítačového vidění.

## 2.2 Struktura neuronové sítě

Umělé neuronové sítě jsou silně inspirované biologickými neuronovými sítěmi. A i když napodobení celé funkčnosti lidského nervového systému by bylo velmi složité – ne-li nereálné, zejména s ohledem na počet neuronů a způsob jejich propojení, je možné simulovat alespoň některé funkce lidské mysli.

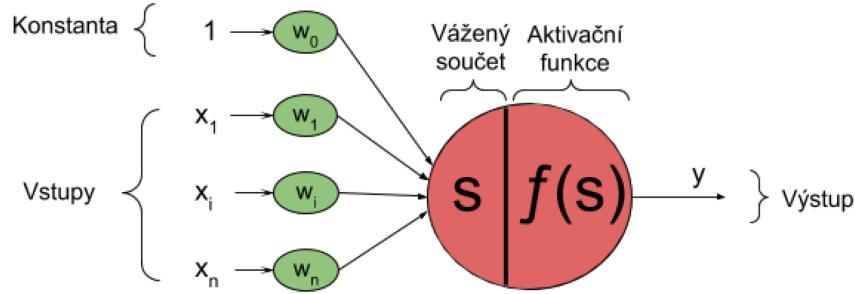
Podstatnou vlastností neuronových sítí je jejich schopnost učení. Tato vlastnost způsobuje, že již není nutná algoritmizace řešené úlohy, ale stačí neuronové sítě opakováně předložit příklady popisující daný problém, podle kterých jsou postupně upravovány síly vazeb v síti. Tato fáze učení pak určuje, jakým způsobem bude síť transformovat vstupní data na výstupní [8].

### 2.2.1 Neuron

Biologický neuron se skládá ze tří hlavních částí. Dendrity přijímají vstupní signály. V těle jsou vstupní signály sečteny do jednoho potenciálu, který vede k vybuzení neuronu – zaslání signálu na výstup, pokud potenciál překročí určitou mez. Axonové vlákno pak vede k synapsím, tedy spojům s dalšími neurony. Lidská mysl pak funguje na principu posilování nebo oslabování těchto spojů.

Umělý neuron se snaží tuto funkčnost napodobit, viz obrázek 2.1. Jeho vstupy  $x_i$  jsou násobeny váhami  $w_i$ , které reprezentují sílu daného spoje. Neuron pak tyto vážené vstupy seče, a na tento součet aplikuje aktivační funkci (AF), která definuje hodnotu výstupu  $y$ . Existuje mnoho různých AF, v našem řešení vyzkoušíme například hyperbolický tangens (tanh), ReLU či Mish [9].

Kromě vah jednotlivých vstupů neuron obvykle obsahuje ještě tzv. bias (někdy také práh nebo posun), jehož funkci je posunout vážený součet vstupů tak, aby bylo možné modelovat i funkce,



Obrázek 2.1: Model umělého neuronu [10]

které nejsou nulové v počátku souřadnic. Je buď reprezentován jako samostatný parametr, nebo jako váha konstantního vstupu s hodnotou 1, jako na obrázku 2.1.

Obecně tvoří všechny vstupní váhy a bias množinu parametrů, které ovlivňují funkčnost celé neuronové sítě. Proces trénování sítě pak spočívá v nalezení optimálních hodnot těchto parametrů, které vedou k co nejmenší chybě při řešení úlohy.

## 2.2.2 Dělení neuronových sítí

Neuronové sítě jsou dnes využívány v mnoha oblastech a dokážou řešit mnoho různých úloh, nicméně neexistuje jediný typ sítě, který by dokázal řešit všechny. V průběhu let proto bylo vyvinuto mnoho různých architektur sítí, každá pro jiné využití.

Jedním ze základních způsobů, jak můžeme neuronové sítě rozdělit, je podle typu učení: učení s učitelem (supervised) a učení bez učitele (unsupervised). V případě učení s učitelem, předkládáme sítě dvojice vstupů a očekávaných výstupů, na jejichž základě se síť snaží minimalizovat chybu úpravou svých parametrů. Oproti tomu, u učení bez učitele nemá síť k dispozici očekávané výstupy, ale snaží se najít nějaké struktury v datech, například shluky. V další části se budeme věnovat hlavně neuronovým sítím pro učení s učitelem.

Dále můžeme neuronové sítě rozdělit na dopředné (feedforward NN – FFNN) a rekurentní (recurrent NN – RNN). U dopředných sítí se informace šíří pouze ze vstupu k výstupu a nevykystují se žádné smyčky. Naopak rekurentní sítě obsahují zpětnou vazbu z výstupu přivedenou na vstup. To umožňuje reagovat na změny v čase.

Další možností jejich rozdělení je dle topologie, která zahrnuje jejich hloubku, tzn. počet vrstev, velikost těchto vrstev a jejich vzájemné propojení. V další sekci se budeme věnovat právě uspořádání vrstev v neuronových sítích.

## 2.3 Topologie neuronových sítí

Nejčastějším uspořádáním neuronů v neuronových sítích je uspořádání do vrstev. Neurony dané vrstvy jsou spojeny pouze s neurony z předchozí a následující vrstvy. Vrstvy pak dělíme na tři typy: vstupní, výstupní a skryté. Vstupní vrstva neobsahuje neurony a neprovádí žádné operace, pouze přijímá vnější signály a distribuuje je do další vrstvy. Výstup z neuronů ve výstupní vrstvě pak reprezentuje výstup celé sítě.

Pokud síť obsahuje pouze vstupní a výstupní vrstvu, mluvíme o jednovrstvé neuronové síti. Takové sítě mají velmi omezené možnosti, proto se v praxi nepoužívají. Většina neuronových sítí má mezi vstupní a výstupní vrstvou alespoň jednu skrytou vrstvu.

U většiny klasických dopředných neuronových sítí jsou jednotlivé vrstvy mezi sebou plně propojeny, tzn. každý prvek jedné vrstvy je propojen se všemi prvky následující vrstvy.

## 2.4 Proces trénování s využitím backpropagation

Jak již bylo řečeno, proces trénování NN spočívá v nalezení optimálních hodnot parametrů jednotlivých neuronů. Optimální parametry pak vedou k minimální chybě. Chybou rozumíme rozdíl mezi skutečným výstupem sítě a očekávaným výstupem. K tomu se nejčastěji používá algoritmus backpropagation (také algoritmus zpětného šíření chyby). Nyní si popíšeme, jak trénování sítě pomocí tohoto algoritmu funguje v dopředné neuronové síti. Tento algoritmus je nicméně využíván i ve většině jiných architektur, jako jsou konvoluční či rekurentní neuronové sítě, které budeme později využívat.

Nejprve se ze vstupních dat vypočítají pomocí aktuálních parametrů sítě reálně výstupy sítě. Tento proces se nazývá dopředný průchod (ang. forward pass). Následně se pomocí chybové funkce (také nákladová funkce – ang. loss function) spočítá chyba sítě. Ta vyjadřuje, v jaké míře se skutečné výstupy liší od očekávaných. Můžou být použity různé chybové funkce, v závislosti na typu úlohy, kterou sítě řeší.

V klasifikačních úlohách je nejčastěji používaná chybová funkce křížové entropie, která porovnává rozdelení pravděpodobnosti skutečného výstupu sítě s očekávaným rozdelením pravděpodobnosti. Pro regresní úlohy se používá například střední kvadratická chyba (ang. mean squared error – MSE) vyjadřující střední hodnotu druhých mocnin rozdílů mezi skutečnými a očekávanými hodnotami. Další možností je absolutní chyba (ang. mean absolute error – MAE), která vyjadřuje střední hodnotu absolutních hodnot rozdílů.

Dále je používána metoda gradientního sestupu, k nalezení minima chybové funkce. Za tímto účelem se vypočítají derivace chyby podle jednotlivých parametrů sítě. Tyto derivace pak určují, jakým směrem a jak rychle se mají dané parametry měnit, aby se chyba minimalizovala. Jelikož se derivace parametrů v dané vrstvě vypočítávají pomocí řetězového pravidla derivací (ang. chain

rule) podle derivací parametrů následující vrstvy, je tento proces nazýván zpětný průchod (ang. backward pass).

Jednotlivé parametry se pak podle vypočítaných derivací upraví. Velikost změny je určena hyperparametrem rychlostí učení (ang. learning rate, také krok), který určuje, jak rychle se mají parametry měnit. Vypočítaná derivace se vynásobí rychlostí učení a přičte k původní hodnotě parametru. Tento proces se opakuje pro všechny parametry sítě. Metoda gradientního sestupu umožňuje větší změny parametrů, když jsou daleko od minima – absolutní hodnoty derivací jsou větší, a naopak menší změny, když se k němu blíží – derivace se blíží nule.

Proces trénování sítě se skládá z opakování dopředného a zpětného průchodu pro všechna trénovací data (někdy postupně pro jejich podmnožiny – dávky, ang. batches). Po každém průchodu (někdy také epocha) se upraví parametry sítě podle vypočítaných derivací. Tento proces se opakuje, dokud chyba nespadne na požadovanou úroveň, nenastane její konvergence nebo není překročen maximální počet iterací.

## 2.5 Optimalizace procesu trénování

V základní verzi algoritmu backpropagation se využívá výše popsaný gradientní sestup. Ten ale může mít některé problémy, které mohou zpomalit proces trénování nebo jej někdy zcela znemožnit. Nyní si popíšeme některé z těchto problémů a jejich možná řešení.

Jedním z podstatných problémů gradientního sestupu je, že se může lehce zaseknout v lokálním minimu, kde je derivace nulová. To může způsobit, že se síť zastaví v nějakém suboptimálním bodě a nepokračuje do globálního minima, které by odpovídalo optimálnímu řešení. Tento problém se často řeší přidáním tzv. momentu do úpravy parametrů. Tento proces bere v úvahu i předchozí změny parametrů. V případě, že se derivace parametru v průběhu trénování změní, moment umožňuje parametru ještě nějakou dobu pokračovat v pohybu ve stejném směru. To většinou pomůže překonat lokální minima a dosáhnout tak globálního minima.

Dalším řešením tohoto problému je využití stochastické approximace gradientního sestupu (ang. stochastic gradient descent – SGD), kde se gradient počítá pro náhodně vybrané podmnožiny trénovacích dat. Tento postup umožňuje rychlejší konvergenci, počítáme totiž gradient jen pro část dat, a zároveň zabránuje zaseknutí v lokálním minimu zavedením šumu do procesu trénování. Tato metoda se využívá nejčastěji pro velké datasety.

Dalším problémem může být nastavení optimálního kroku učení. Pokud je krok příliš velký, může dojít k příliš velké změně, která opomine minimum, může se tak stát, že nikdy nedosáhneme konvergence. Naopak, pokud je krok příliš malý, může trénování trvat příliš dlouho. Řešením může být například adaptivní nastavení kroku. Nejznámější taková metoda je RMSProp (ang. root mean square propagation), která upravuje krok učení pro každý parametr podle průměrného druhého momentu gradientu [11]. Další možností je v průběhu trénování postupně snižovat krok (ang. learning rate decay).

Populárním řešením těchto problémů je také Adam (ang. adaptive moment estimation, v překladu adaptivní odhad momentu), který kombinuje zavedení momentu a adaptivního nastavení kroku pomocí RMSProp [12].

Další metodou, jak optimalizovat nastavení kroku, je normalizace dat mezi vrstvami sítě (ang. batch normalization). Tím se zamezí příliš velkým změnám v jednotlivých vrstvách, které někdy destabilizují proces trénování.

U trénování hlubokých sítí se často naráží rovněž na problém přetrénování (také nadměrné přizpůsobení, ang. overfitting). Přetrénování nastává, když se síť dobře naučí trénovací data, zároveň ale postrádá schopnost generalizace, a když pak dostane nová data, nedosahuje dobrých výsledků. Nejčastěji k problému dochází v situacích, kdy se snažíme natrénovat hluboké a komplexní modely, aniž bychom měli k tomu dostatečné množství dat. K základním řešením tohoto problému tedy patří použití většího množství trénovacích dat – pokud jsou dostupná, jinak se někdy zavádí umělé variace dat jako rotace či převrácení, jinak je třeba někdy zvážit zjednodušení architektury modelu. Dále pak jsou využívány techniky regularizace, které upravují samotný proces trénování.

Nejjednodušší regularizační technikou je tzv. předčasné zastavení, tedy zastavení trénování v případě, že se chyba na validačních datech začne zvyšovat. Další možností je dropout (také výpadek), kdy se u určitého počtu náhodně zvolených neuronů v průběhu trénování nastaví na výstupu nula. Tím se snižuje závislost sítě na konkrétních neuronech a zvyšuje se tak její schopnost generalizace.

Další dvě metody, nazývané L1 a L2 regularizace, přičítají k chybové funkci člen, který penalizuje velikost parametrů sítě. L1 regularizace (také metoda [13], z ang. least absolute shrinkage and selection operator) přičítá k chybové funkci součet absolutních hodnot všech parametrů sítě vynásobený hyperparametrem, který určuje míru této penalizace. Tato metoda vede k řídké síti, ve které je mnoho parametrů nulových. L2 regularizace (také weight decay) přičítá k chybové funkci součet druhých mocnin všech parametrů sítě vynásobený hyperparametrem  $\lambda$  [14]. Tato metoda vede jednak k menší variabilitě parametrů, jednak k pomalejším změnám parametrů v průběhu trénování, v důsledku pak k menší citlivosti na šum v datech. Využívá se zejména v hlubokých sítích.

# Kapitola 3

## Konvoluční neuronové sítě

I když byly jedny z prvních NN použity ke zpracování obrazu, brzy se ukázalo, že pro zpracování obrazu s větším rozlišením a větším množstvím kanálů je klasická architektura NN velmi neefektivní. To proto, že je obraz reprezentován velkým množstvím dat, a klasická NN by pro jejich zpracování musela být tak komplexní (velké množství skrytých vrstev a neuronů), že by bylo velmi těžké, a spíše nemožné ji nacvičit [15].

Bylo tedy třeba vytvořit jinou architekturu, která by efektivně zpracovávala obrazová data. Nejznámější takovou architekturou je konvoluční neuronová síť (ang. convolutional neural network – CNN). Ta tento problém řeší tak, že se snaží zredukovat rozměr vstupních dat pomocí konvolučních vrstev, které jsou schopny efektivně zpracovávat obrazová data, a extrahovat z nich důležité vlastnosti, jež následně můžeme zpracovat pomocí poměrně jednoduché NN.

Jelikož je základem našeho problému analýza obrazových dat, použijeme pravděpodobně právě konvoluční sítě, minimálně v první fázi, kdy z obrazu extrahujeme klíčové body osob. V této kapitole se proto podíváme na základní principy konvolučních neuronových sítí.

### 3.1 Konvoluce

V kontextu práce s obrazovými daty je konvoluce binární operace, kdy pro daný pixel v obrazu sečteme hodnoty pixelů v jeho okolí vynásobené váhami vyjádřenými maskou, která se nazývá jádro konvoluce (ang. kernel). Výsledný obraz je také nazýván konvoluce. Z matematického pohledu se jedná o diskrétní dvourozměrnou konvoluci – binární operaci diskrétních funkcí, která je definována následovně:

$$(h * f)(x, y) = \sum_{i=-k}^k \sum_{j=-k}^k h(x - i, y - j) \cdot f(i, j)$$

kde  $h$  je vstupní obraz,  $f$  je jádro konvoluce, velikost jádra je  $2k + 1 \times 2k + 1$ , a  $x$  a  $y$  jsou souřadnice pixelu, ke kterému se jádro aplikuje.

Hodnota konvoluce na dané pozici je tedy suma součinů hodnot pixelů vstupního obrazu a hodnot váh vyjádřených jádrem konvoluce položeným středem na danou pozici. Nejčastěji je velikost jádra lichá, jelikož je pak jednodušší určení středu jádra [15].

Konvoluční jádro může být pevně dáno pro danou úlohu, např. pro detekci hran se často používá Sobelův operátor, který využívá dvě jádra – pro vertikální a horizontální hrany.

Myšlenkou konvolučních neuronových sítí je nahradit část plně propojených vrstev konvolučními vrstvami, které z obrazu extrahují vlastnosti relevantní k problému. Ruční nastavení všech hodnot jader tak, aby konvoluční vrstvy extrahovaly požadované vlastnosti z obrázku, je ale velmi obtížné, v případě složitějších či obecnějších problémů téměř nemožné. V 1989 proto Yann LeCun et al. navrhli metodu, jak se síť může hodnoty jader naučit sama [7], a tak si efektivně vytvořit i složitější filtry, které by člověk těžko navrhl ručně. Zjistili, že váhy konvolučních jader se mohou trénovat pomocí algoritmu backpropagation stejně jako váhy neuronů v plně propojených vrstvách.

Takový přístup má mnoho výhod. Konvoluce se dá velmi dobře paralelizovat a zajistit tak vysokou efektivitu výpočtů. Oproti plně propojeným vrstvám má vždy konvoluce počet váh pouze rovný počtu prvků ve všech jádrech. I v případě provedení mnoha konvolucí je počet váh výrazně menší než v případě potřebného počtu a velikosti plně propojených vrstev. Zároveň lze pomocí konvolucí efektivně extrahat různé vlastnosti ze vstupního obrazu, ty pak pomocí plně propojených vrstev zpracovat a využít k řešení problému. Proto se také výstupu konvoluce často říká mapa příznaků (ang. feature map).

## 3.2 Konvoluční vrstva

Konvoluční vrstva je základním prvkem konvoluční neuronové sítě. V jistém slova smyslu je podobná plně propojené vrstvě, jelikož obsahuje váhy, biasy a aktivační funkce. Namísto plného propojení s neurony následující vrstvy je ale aplikována konvoluce na vstupní data. K výstupní mapě je přičten bias a následně může být aplikována aktivační funkce.

Jedná vrstva může mít několik konvolučních jader, každé s vlastními váhami a biasem. Je třeba zároveň pamatovat, že každé jádro musí mít hloubku rovnou počtu vstupních map, resp. počtu kanálu vstupního obrazu v případě vstupní vrstvy. Konvoluci pak aplikujeme zvlášť pro každé jádro, počet výstupních map příznaků tedy bude roven počtu jader v dané vrstvě. V praxi bude každé jádro vyjadřovat jinou vlastnost, kterou se snažíme ze vstupního obrazu extrahat.

Množinu parametrů dané konvoluční vrstvy tedy tvoří hodnoty jader a jejich příslušné biasy. K hyperparametru pak patří počet jader a jejich velikost, aktivační funkce, krok a padding.

## 3.3 Poolovací vrstva

Jak již bylo zmíněno, v konvolučních vrstvách vzniká vícero map příznaků vyjadřující různé vlastnosti. Tím se ale množství dat zvětšuje, což porušuje samotnou myšlenku konvolučních sítí, která je

založena na snaze zredukovat rozměr dat. Proto se snažíme rozměr map příznaků zmenšovat. Jak již bylo zmíněno, dojde k redukci rozměrů v konvoluční vrstvě, pokud použijeme krok  $k > 1$ . Častěji ale je k tomuto účelu prováděno podvzorkování dat v tzv. poolovacích vrstvách (z ang. pooling layer) [16].

V poolovací vrstvě je vstupní mapa rozdělená do stejně velkých čtvercových oblastí velikosti  $t \times t$ , na základě hodnot v daném čtverci je pak vytvořen jeden pixel výstupní mapy. K nejčastěji používaným metodám patří max-pooling a average-pooling. Max-pooling vybere z dané oblasti největší hodnotu, zatímco average-pooling vyhodnotí z každé oblasti průměr hodnot. Velikost výstupní mapy pro vstup velikosti  $n \times n$  a poolovací oblasti velikosti  $t \times t$  je tedy  $\lfloor \frac{n}{t} \rfloor \times \lfloor \frac{n}{t} \rfloor$ .

## 3.4 Architektura CNN

Konvoluční neuronové sítě se skládají ze dvou částí – konvoluční části a plně propojené části.

Konvoluční část se skládá z několika konvolučních vrstev proplétaných s poolovacími vrstvami. V této části se pracuje s mapami příznaků. Její výstupem je soubor map příznaků, resp. mapa příznaků s větší hloubkou, která vyjadřuje různé vlastnosti vstupního obrazu.

Plně propojená část je pak klasická dopředná neuronová síť, která na základě extrahovaných vlastností provádí klasifikační či regresní úlohu. Před vstupem do plně propojené části jsou mapy příznaků převedené do jednoho vektoru o velikosti rovné součtu velikostí všech map příznaků.

V praxi se často používá komplexnější architektury, které jsou přímo zaměřené na specifické úlohy, jako je detekce a klasifikace objektů či detekce klíčových bodů. Některé z nich budeme testovat pro potenciální využití v konečném řešení. Tyto architektury často zahrnují i další specifické techniky jako je extrakce vlastnosti na různých úrovních rozlišení či metody zaměření analýzy pouze na zajímavé oblasti. Také jsou často postaveny na obecných ověřených CNN architekturách, pomocí kterých extrahují ve vstupní fázi vlastnosti, které pak způsobem specifickým pro danou úlohu zpracovávají.

# Kapitola 4

## Rekurentní neuronové sítě

Rekurentní neuronové sítě (ang. Recurrent Neural Networks – RNN) je kategorie neuronových sítí, které do své architektury zapojují zpětnou vazbu. Na rozdíl od dopředných sítí, které zpracovávají jednotlivé vstupy nezávisle, rekurentní sítě spolu s aktuálním vstupem při evaluaci zohledňují nějakým způsobem i výsledek předchozí iterace. Jejich využití tedy je ve dvou oblastech: analýza změn pozorovaného objektu v čase (např. sledování pohybu, analýza chování či predikce časových řad) a zpracování kontextuálních informací jako je např. přirozený jazyk.

Jelikož v případě pádu se nejedná o statickou informaci, ale o jistý druh pohybu, mohly by právě rekurentní sítě stanovit optimální cestu pro naše řešení [17]. Pojdme se tedy nyní podívat na to, jak rekurentní neuronové sítě fungují a jaké jsou jejich nejpoužívanější architektury.

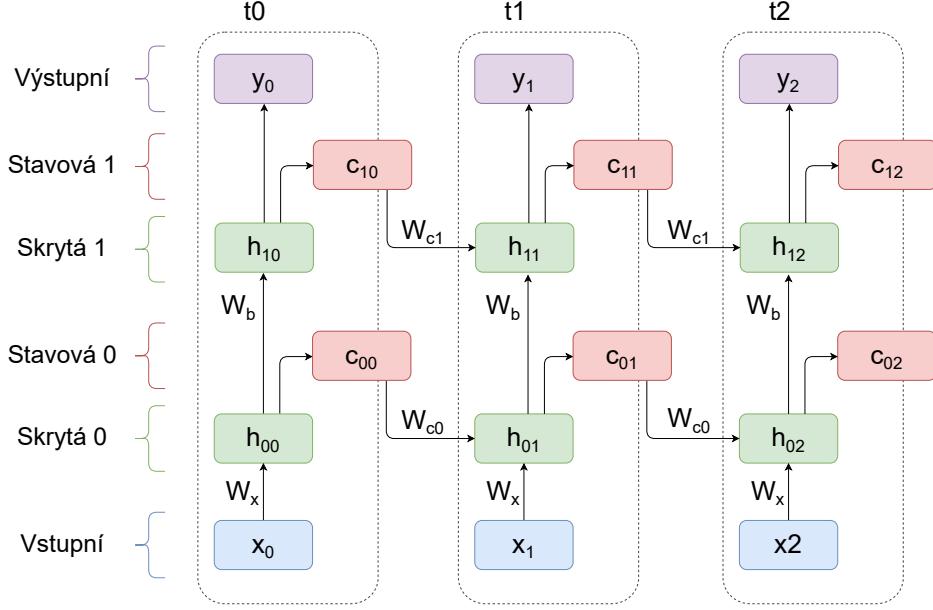
### 4.1 Základní principy

Data, která v aktuální iteraci přebíráme z předchozí iterace, se často označují jako skrytý stav (ang. hidden state). Je to forma paměti, která se s každou iterací aktualizuje. Často je reprezentován jako stavová vrstva (ang. state layer nebo context layer), která přijímá hodnoty z výstupu neuronů, uchovává je mezi iteracemi a předává je spolu se vstupními daty na vstup neuronů. Na obrázku ?? je tato vrstva reprezentována neurony  $c_i$ .

Nejjednodušší forma rekurentní neuronové sítě je NN s jednou skrytou vrstvou; tato vrstva kromě dat ze vstupní vrstvy přijímá také výstup předchozí iterace svých vlastních neuronů. Příklady takových sítí jsou Elmanova a Jordanova síť [18] [19], které byly prvními rekurentními sítěmi využívajícími backpropagation pro trénování.

#### 4.1.1 Hluboké rekurentní sítě

Stejně jako u dopředných neuronových sítí, kde se od jednoduchého perceptronu přešlo k hlubokým sítím, se i rekurentní sítě rozšířily na více vrstev. V hlubokých rekurentních neuronových sítích (ang. deep RNN – DRNN) jsou pak jednotlivé vrstvy většinou podobné struktuře Elmanovy síť



Obrázek 4.1: Unrolling hluboké RNN

[18] - zpětná vazba je předávána pouze v rámci jedné vrstvy, nikoliv mezi vrstvami RNN (například z výstupní vrstvy do první skryté vrstvy) [20]. Má to několik důvodů. Trénování sítě se zpětnou vazbou mezi vrstvami by bylo velmi složité a obtížné. Také, u neuronových sítí obecně platí, že každá vrstva sítě se učí pochopit problém na jiné úrovni abstrakce, zpětná vazba přes několik vrstev by pak mohla narušit stabilitu tohoto procesu a omezit kvalitu učení.

### 4.1.2 Trénování rekurentních sítí

Pro pochopení rekurentních neuronových sítí je třeba si vysvětlit, jak se trénují. Pro vizualizaci trénování RNN se tyto sítě takzvaně rozbaluje v čase (ang. unrolling). Znamená to, že jednotlivé iterace vizualizujeme jako sekvenci stejných sítí (stejné váhy), které v čase  $t$  přijímají vstup  $x_t$  a vracejí výstup  $y_t$ , viz obrázek 4.1. Zároveň místo smyček znázorňujících zpětnou vazbu přijímá skrytá vrstva v čase  $t$  stav  $c_{t-1}$  z předchozí iterace. Takto je propojená mezi iteracemi každá skrytá vrstva (na obrázku 4.1 vizualizováno propojení přes stavovou vrstvu).

Při trénování se pak používá algoritmus zpětného šíření chyby v čase (ang. backpropagation through time – BPTT). Algoritmus funguje stejně jako klasický backpropagation, šíří se ale nejenom vrstvami, ale i iteracemi. Unrolling nám pomáhá backpropagation pochopit, jednotlivé iterace totiž jsou naskládány jako vrstvy a celou síť řešíme jako klasickou dopřednou NN.

### 4.1.3 Problémy mizejícího a explodujícího gradientu

Výše popsané základní rekurentní neuronové sítě, někdy označovány jako vanilla RNN, trpí několika zásadními problémy. U dopředných sítí jsme zmíňovali problém mizejícího gradientu (ang. vanishing gradient), vystupující zejména u hlubších sítí. Ten se projevuje i u RNN a je zesílený tím, že jsou jednotlivé iterace naskládané na sebe, podobně jako vrstvy. Zejména pak u delších sekvencí budou mít dřívější vstupy velmi malý vliv na učení sítě.

U RNN se také projevuje problém opačný – explodující gradient (ang. exploding gradient). Ten způsobuje, že v průběhu sekvence se váhy začnou exponenciálně zvětšovat a dosáhnou tak nepriměřeně velkých hodnot.

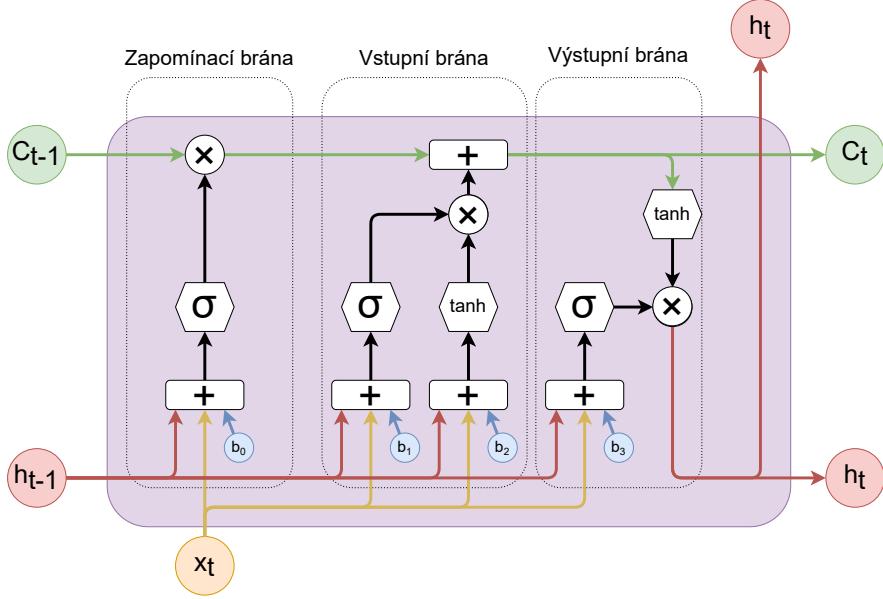
Podívejme se, co přesně tyto problémy způsobuje. Součástí algoritmu backpropagation je počítání parciální derivace ztrátové funkce podle jednotlivých vah. V případě BPTT potřebujeme mimo jiné počítat parciální derivace skrytého stavu mezi jednotlivými iteracemi  $\frac{\partial h_{t-1}}{\partial h_t}$ . Tyto derivace následně opakovaně násobíme při použití řetězového pravidla. Pokud je tato derivace  $\frac{\partial h_{t-1}}{\partial h_t} < 1$ , jeho vynásobení bude mít za následek postupné zmenšování gradientu. Pokud budeme například mít sekvencí 100 iteraci, pak i kdyby se gradienty v každé iteraci zmenšovaly 0,9 krát, po 100 iteracích by gradient klesl na hodnotu  $0,9^{100} \approx 2,7 \times 10^{-5}$ , což je prakticky nula. Pokud se naopak bude gradient zvětšovat 1,1 krát, po 100 iteracích by gradient vzrostl na  $1,1^{100} \approx 13780$ , což způsobí úplnou destabilizaci sítě a nedosáhneme žádného výsledku. Vidíme tedy, že v případě, kdy je  $\frac{\partial h_{t-1}}{\partial h_t} > 1$ , dochází k explodujícímu gradientu.

Z důvodu těchto problémů byly vyvinuty složitější rekurentní struktury. Jejich architektura je v podstatě podobná, jednotlivé vrstvy jsou ale zastoupeny jinými stavebními bloky, které umožňují zejména širší pochopení kontextu a efektivnější proces trénování. Vanilla RNN se v praxi dnes využívají velmi zřídka. K nejpoužívanějším architekturám patří LSTM (ang. long short-term memory) a GRU (ang. gated recurrent unit), které nyní popíšeme.

## 4.2 LSTM

Dlouhá krátkodobá paměť (ang. long short-term memory – LSTM), představena Hochreiterem a Schmidhuberem v roce 1997 [21], je typ rekurentní neuronové sítě, který byl navržen tak, aby překonal problémy mizejícího a explodujícího gradientu.

Její základem je jednotka, viz obrázek 4.2.1, která ve třech stádiích aktualizuje krátkodobou a dlouhodobou paměť. Dlouhodobá paměť je reprezentovaná pomocí stavu buňky (ang. cell state, na obrázku 4.2.1  $c_t$ ), který je postupně upravován a nakonec předán další iteraci. Dokáže uchovávat dlouhodobé závislosti. Krátkodobá paměť je reprezentována pomocí skrytého stavu. Je použita pro úpravu dlouhodobé paměti, v konečném stadiu je ale vždy v rámci dané iterace vytvořena nová. Je tak vhodná pro uchování krátkodobých závislostí. Na obrázku 4.2 je znázorněna jako  $h_t$ .

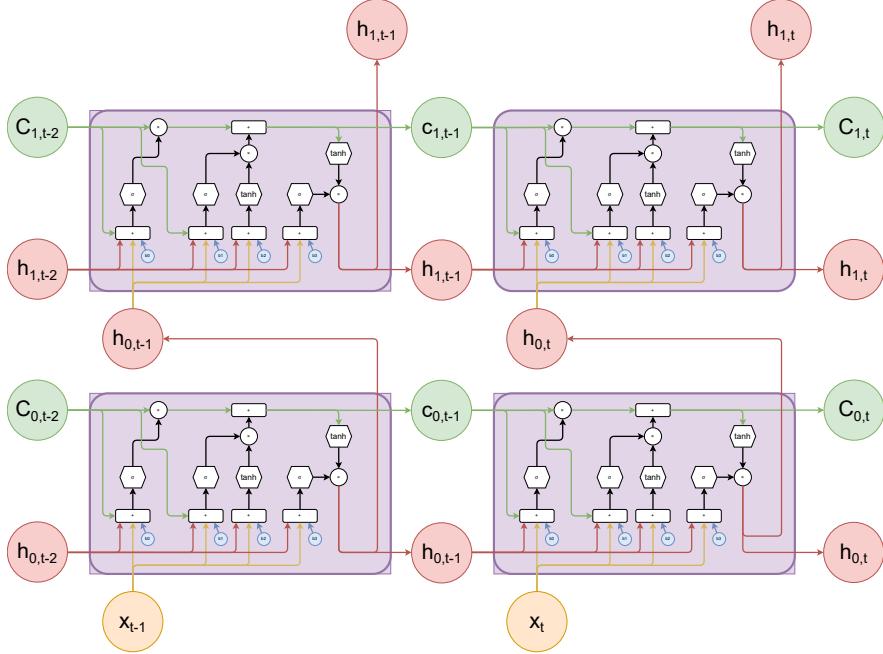


Obrázek 4.2: Jednotka LSTM

Jednotka LSTM má tři hlavní komponenty – zapomínací bránu (ang. forget gate), vstupní bránu (ang. input gate) a výstupní bránu (ang. output gate). Brány určují, které informace mají být předány dál.

První, zapomínací brána určuje, které informace z dlouhodobé paměti  $c_{t-1}$  se dostanou dále – co má jednotka zapomenout, resp. zapamatovat. Ve vstupní bráně se nejprve vytvoří kandidátní stav buňky. Ten je výsledkem neuronové vrstvy s tangenciální aktivační funkcí, do které vstupuje aktuální vstup  $x_t$  a krátkodobá paměť  $h_{t-1}$ . Pak se určí, které informace z kandidátního stavu buňky se přičtou do stavu buňky a vznikne tak aktuální stav buňky  $c_t$ . Ve výstupní bráně se pomocí tangenciální aktivační funkce vytvoří na základě stavu buňky  $c_t$  kandidátní skrytý stav. Pak se určí, které z těchto informací budou tvořit nový skrytý stav  $h_t$ .

V každé bráně tedy máme informace, pro které určujeme, zda je poslat dále či nikoliv, nazvěme je propouštěný obsah (předchozí stav buňky, kandidátní stav buňky či kandidátní skrytý stav). Toto určení se provádí vždy pomocí neuronové vrstvy se sigmoidní aktivační funkcí. Do těchto vrstev vstupuje vždy předchozí skrytý stav  $h_{t-1}$  a aktuální vstup  $x_t$ . Výstupem je hodnota mezi 0 a 1 pro každou informaci. Pak se tento výsledek vynásobí propouštěným obsahem. Pokud je výstup této vrstvy 0, informace se nepředávají dál, pokud je 1, informace se předávají dále. Vstupy do všech neuronových vrstev jsou vždy vynásobeny váhami, ty ale nejsou pro jednoduchost na obrázku 4.2.1 zobrazeny. Na obrázku 4.3 je znázorněna rozvinutá hluboká LSTM síť. Jednotlivé vrstvy sítě jsou naskládány vertikálně, jednotlivé iterace pak jsou rozvinuty vedle sebe. Jednotlivé vrstvy si předávají skrytý stav – krátkodobou paměť, mezi iteracemi si pak daná vrstva předává krátkodobou i dlouhodobou paměť.



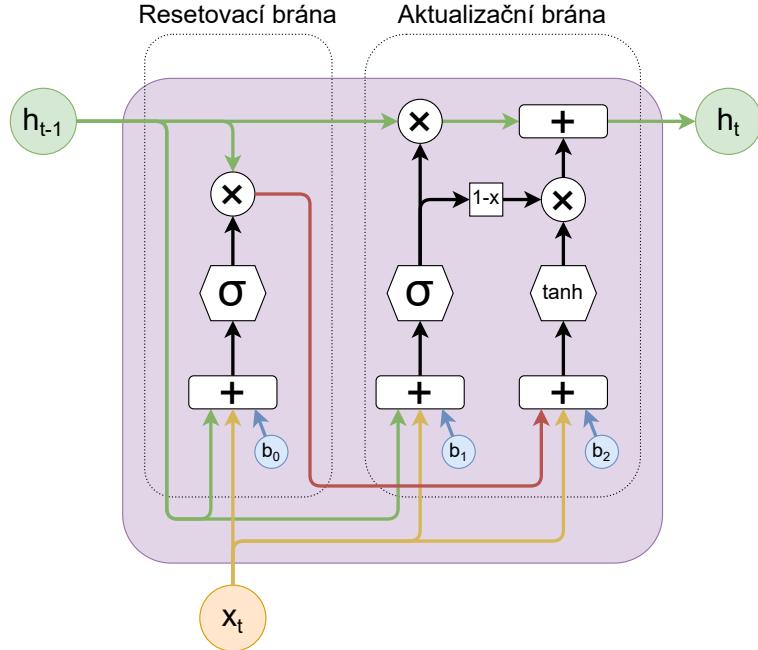
Obrázek 4.3: Rozvinutá hluboká LSTM síť

LSTM sítě vynikají v udržování dlouhodobých závislostí a složitých struktur. Jelikož mají tři brány, je síť schopná přesně rozhodnout, které informace chce dlouhodobě uchovávat, které naopak mají větší vliv na aktuální výstup a které mají být zapomenuty. Je to ale za cenu většího výpočetního nároku a složitějšího trénování. Také je pro tyto sítě vhodné mít větší množství trénovacích dat, jinak může dojít k přetrénování. Využívá se tak zejména pro predikci dlouhých a komplexních časových sekvencí či zpracování přirozeného jazyka. Zejména u přirozeného jazyka se LSTM sítě osvědčily jako velmi efektivní. Potřebujeme totiž, aby si síť pamatovala dlouhé závislosti, zároveň máme většinou k dispozici obrovské množství vzorků.

Sít LSTM by mohla být vhodná pro klasifikaci pózy zejména pokud bychom potřebovali analyzovat pohyb v delších časových úsecích. Událost pádu se naopak obvykle odehrává v krátkém čase, nicméně můžeme otestovat, jakých výsledků bude tato architektura dosahovat oproti jiným rekurentním sítím, zejména v případě, kdy bychom modelu nepředávali pouze pózu z  $n$  posledních snímků, ale celou sekvenci detekovanou pro danou osobu.

### 4.3 GRU

Gated Recurrent Unit (GRU) je novější typ rekurentní neuronové sítě, který představil v roce 2014 Cho et al. [22]. Je postavený na principu brán, podobném jako LSTM, nepotřebuje ale zvlášť stav pro dlouhodobou paměť. Místo toho kombinuje krátkodobou a dlouhodobou paměť do skrytého stavu  $h_t$ .



Obrázek 4.4: Jednotka GRU

GRU obsahuje dvě brány: resetovací bránu (ang. reset gate) a aktualizační bránu (ang. update gate), viz obrázek 4.4. V resetovací bráni se určuje, které informace z předchozího skrytého stavu  $h_{t-1}$  budou mít vliv na tvorbu kandidátního skrytého stavu. V aktualizační bráni vzniká nový skrytý stav  $h_t$  kombinací předchozího skrytého stavu a kandidátního skrytého stavu. Ten je vytvořen pomocí neuronové vrstvy s tangenciální aktivační funkcí, do které vstupuje výstup resetovací brány a aktuální vstup  $x_t$ . Pak se na základě předchozího skrytého stavu  $h_{t-1}$  a aktuálního vstupu  $x_t$  určí, které informace v novém skrytém stavu budou převzaty z předchozího skrytého stavu a které z kandidátního skrytého stavu.

Hlavní výhodou GRU je jednoduchost. Oproti LSTM má méně parametrů a provádí méně výpočtů. Je tak jednak rychlejší při evaluaci, jednak jednodušší pro natrénování. Také, u GRU sítí je menší pravděpodobnost přetrénování, což je výhodné zejména v situacích, kdy máme omezený počet trénovacích dat. GRU sítě se často využívají v úlohách, kde je důležité rychlé zpracování a efektivita, např. v mobilních aplikacích a zpracování v reálném čase. Je to ale za cenu trošku horšího zpracování komplexních a dlouhodobých závislostí. Oproti LSTM nemá GRU takovou kontrolu nad tím, které informace dlouhodobě uchovávat a má sklon k rychlejšímu zapomínání. Proto se až tak nehodí pro složitější úlohy a situace, kdy je nutné si pamatovat velmi dlouhé časové závislosti. Nicméně jsou dneska první volbou pro mnoho úloh, po LSTM sítích se pak sahá, až když si GRU sítě s danou úlohou neporadí.

# Kapitola 5

## Představení problematiky detekce pádu

V této kapitole stanovíme, co je přesně naším cílem, a zamyslíme se, jak k našemu problému přistoupit.

Naším úkolem bude v reálném čase z videostreamu detekovat pád osoby. Pád osoby definujeme jako náhle, neúmyslné klesnutí těla z výškové pozice (např. stání, chůze nebo sezení) na zem nebo jinou nižší úroveň, přičemž tato osoba nemá kontrolu nad tímto pohybem. Samozřejmě nejsme vždy schopni úplně dobře rozeznat, zda se nejedná o úmyslné klesnutí, např. prudké lehnutí.

Dle některých definic (zejména ve zdravotnictví) se o pád nejedná, pokud jde o důsledek závažné vnitřní příhody (např. mrtvice). V našem případě toto nerozlišujeme, naopak chceme detektovat jak pády v důsledku ztráty rovnováhy či vlivem vnějších faktorů (např. zakopnutí, převrácení těžkým předmětem), tak pády v důsledku akutních událostí vlivem zdravotních problémů, jako jsou např. mrtvice, záchvaty, mdloby či jiné důvody ztráty vědomí.

### 5.1 Návrh řešení

Cílem této práce je navrhnout algoritmus, který bude detektovat, zda je ve vstupní sekvenci snímku některá osoba, jejíž pozice je klasifikována jako pád. Hlavním cílem výsledného programu bude alarmovat příslušného pracovníka, pokud osoba upadne.

Alarmovat budeme až, pokud osoba zůstane v ležící pozici. To nám dá možnost odfiltrovat falešné alarmy v případě sehnutí či pokud bude osoba špatně viditelná a algoritmus tak na okamžik špatně vyhodnotí její pohyb. Tímto postprocesingem se ale teď nebudeme zabývat, spíše se zaměříme na samotnou klasifikaci pozice.

Stejně jako u detekce objektů, viz 6.3, bychom mohli i pro detekci pádu vytvořit vhodnou konvoluční síť, která by přímo z obrázku definovala, zda se jedná o pád nebo ne. U detekce se už dnes sice s ohledem na pokrok hardwaru tento přístup používá, nicméně se jedná o velmi náročný úkol, který vyžaduje rozsáhlou optimalizaci, pokročilou architekturu a velké množství trénovacích

dat. Nicméně, pokud by se podařilo takovouto síť natrénovat, mohla by lépe detekovat některé situace např. podle výrazu tváře.

V našem případě tedy budeme v prvním kroku pomocí vhodné předtrénované neuronové sítě detekovat pozici osoby ve formě klíčových bodů, tuto část nazveme *detekční algoritmus*. Na základě těchto bodů pak další neuronová síť vyhodnotí, zda se jedná o pád, tuto část nazveme *klasifikační algoritmus*. To úlohu velice zjednoduší, jelikož místo analyzování tisíců pixelů, budeme analyzovat pár desítek klíčových bodů. Další výhodou je, že u takového postupu jsme schopní použít techniky, kdy sledujeme změny pózy v čase, což by bylo mnohem složitější s jednofázovou konvoluční sítí.

Detekční algoritmus dostane na vstup celý snímek a může detektovat několik osob. Klasifikační algoritmus ale bude zpracovávat každou osobu, resp. její pózu, zvlášt.

Další alternativou by mohlo být pouze detektovat osoby jako objekty, a na základě jejich bounding boxů určit, zda se jedná o pád. Tento postup by byl jednodušší na dvou úrovních. Jednak je detekce objektů méně náročná úloha než detekce pózy, jednak bychom ve druhé fázi analyzovali pouze několik parametrů bounding boxu (rozměry a velikost) oproti pár desítkám klíčových bodů. Nicméně, pokud se nad tím zamyslíme, ne vždy vypovídají parametry bounding boxů o pozici člověka. Tento postup by tak pravděpodobně vedl k mnohem méně přesnému výsledku, než analýza klíčových bodů, kdy může síť analyzovat takové vzorce jako je např. délka končetin v pohledu či úhel mezi nimi.

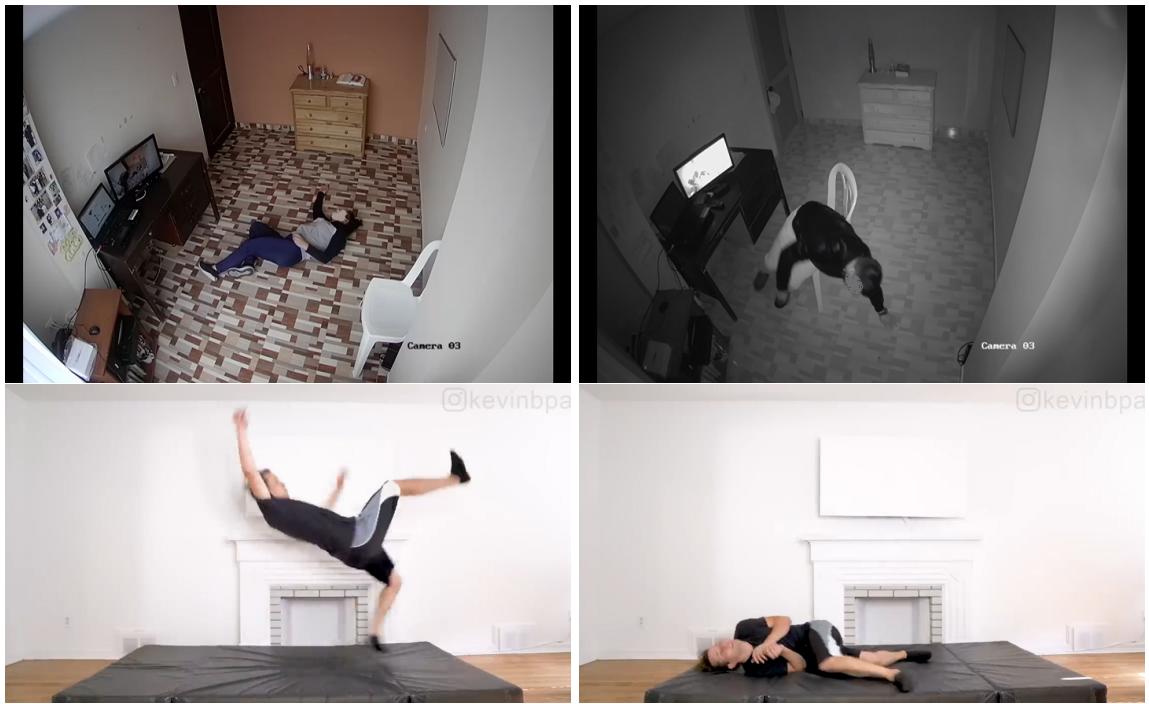
Nyní se podíváme, jak budeme pracovat s daty, zejména při trénování, v další kapitole pak bude rozebraná problematika detekce pózy a bude zvolen algoritmus pro detekci klíčových bodů. Dále se pak budeme zabývat vývojem modelu detekujícího pád na základě těchto klíčových bodů.

## 5.2 Trénovací datasety

Pro trénování našeho modelu jsme použili necelých 150 krátkých (1 až 15 sekund) videí ze dvou zdrojů. Prvním je dataset CAUCAFall vytvořený právě pro práci s pády osob [23]. Tento dataset obsahuje 100 nahrávek simulovaných pádu v různých světelných podmínkách, s různými osobami. Zahrnují širokou škálu scénářů, jednak pro různé druhy pádu (v různých směrech či z židle), jednak pro situace podobné pádu, jako je kleknutí či, sehnutí se, jednak běžné činnosti jako chůze či sednutí. Poměr videí s úpadkem a bez je 50/50.

Dalším zdrojem pro trénovací data je YouTube video tvůrce Kevina Parryho *50 Ways to Fall*. Ve videu autor pády v různých scénářích, jako je zakopnutí, omdlení či poražení elektrickým proudem. Vzhledem k zabavné povaze videa jsme některé scénáře vypustili, nakonec jsme použili 45 videí. Zde prakticky všechny videa obsahují pád, v některých se ale postava vrátí do normálního stavu (např. kotrmelec).

V obou případech se jedná o videa vždy jedné osoby. To proto, že použijeme detekční algoritmus již natrénovaný na videích s více osobami, a náš klasifikační algoritmus bude zpracovávat



Obrázek 5.1: Příkladové snímky z datasetů CAUCAFall (nahoře) a 50 Ways to Fall (dole).

každou osobu zvlášť. Práce s více osobami tak bude úlohou výsledného detektoru, nikoliv trénované klasifikační sítě.

Videa z datasetu CAUCAFall jsou v rozlišení  $720 \times 480$ , zatímco videa z datasetu 50 Ways to Fall jsou v rozlišení  $1280 \times 720$ .

Oba datasety jsme rozdělili do tří sad: trénovací, validační a testovací. Rozděleny byly v tomto poměru: trénovací sada 70%, validační sada 15% a testovací sada 15%. Trénovací a validační sada budou použity v procesu trénování – trénovací pro výpočet ztráty a úpravu vah, validační pro průběžné ověřování výkonu, a testovací sada bude na konci použitá pro otestování jednak klasifikačního algoritmu, jednak celkového řešení.

### 5.3 Třídy a jejich anotace

Pro videa byly vytvořeny anotace aktuální třídy pózy. Tato anotace není pro každý snímek, ale pouze při změně definuje časovou značku a následující třídu.

V anotacích jsme použili 3 třídy, ty odpovídají třem různým třídám pózy, které nás mohou zajímat – *normální*, kdy osoba např. chodí, sedí nebo stojí, *padá* – přechodný stav padání, definován od započatí pohybu směrem dolů, a *upadl* – definován od momentu, kdy se dotkl země trupem nebo všemi končetinami.

Pro náš model obecně potřebujeme jenom dvě třídy – *normální* a *upadl*. Skript tvořící trénovací data proto považuje třídu *padá* za třídu *normální*. Nicméně by se mohlo do budoucna pro pokročilejší optimalizaci zkousit experimentovat i se třídou *padá*, která by mohla síti pomoci hlouběji pochopit problematiku a přesněji rozseznat některé situace, zejména pak v případě využití rekurentních neuronových sítí.

## 5.4 Příprava trénovacích dat pro klasifikační algoritmus

Dále byl vytvořen skript, který prošel každé video z trénovacích datasetů a vytvořil trénovací data pro naši klasifikační síť. Ty obsahují pro každý snímek detekované klíčové body (jako vstup) a aktuální třídu (jako požadovaný výstup). Pro detekci klíčových bodů byl použit vybraný model pro detekci pózy. Výběr modelu je popsán v následující kapitole. Na použitém modelu by teoreticky nemuselo záležet (pokud detekuje stejné typy klíčových bodů), je ale lepší použít ve výsledném programu stejný model jako pro trénovací data. Modely se totiž můžou v některých situacích chovat trochu jinak (např. okluze) a náš model by tak dostával v praxi jiná data, než pro jaké byl natrénován.

Jelikož pro rekurentní neuronové sítě potřebujeme sekvenci snímků, musí být trénovací data ještě zpracována. To ale bude již součástí samotného trénovacího skriptu, jelikož se konečná podoba dat může lišit hlavně délkou sekvencí, dále ale také dodatečným rozšířením o sekvence vynechávající určité množství snímků, což simuluje menší snímkovou frekvenci (FPS).

# Kapitola 6

## Výběr algoritmu pro detekci klíčových bodů

Jelikož je dnes dostupných mnoho různých algoritmů či natrénovaných modelů pro detekci pózy osob v obrázku či videu, nemá smysl pro naše řešení implementovat takovýto algoritmus od nuly. Možné by to samozřejmě bylo, i vzhledem k dostupnosti otevřených trénovacích dat (např. dataset COCO [24]), nicméně bychom pravděpodobně nedosáhli kvalitních výsledků, jako řešení, která jsou výsledkem mnoholetých výzkumů. Hlavně pak bychom těžko dosáhli výkonů těchto řešení, a ten je pro nás stěžejní, jelikož potřebujeme video zpracovávat v reálném čase.

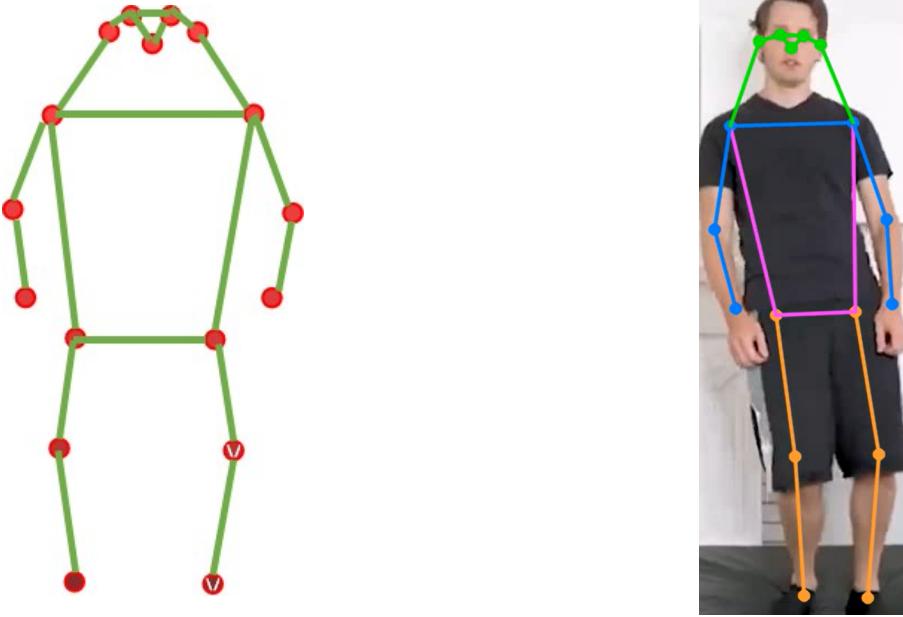
V následující kapitole budou popsány obecné principy detekce osob a jejich pózy v obraze. Následně budou popsány některé populární algoritmy pro detekci pózy se zaměřením na jejich specifika. Několik z nich pak bude otestováno, výsledky budou porovnány, a na jejich základě bude zvolen algoritmus použitý v konečném řešení detekce pádu.

### 6.1 Detekce pózy

Úloha detekce pózy spočívá v nalezení klíčových bodů postavy v obraze. Může se jednat také o zvíře, v našem případě se ale budeme zabývat pouze klíčovými body lidské postavy. Klíčové body představují důležité body lidského těla, znalost jejich lokalizace nám umožňuje analyzovat pózu dané osoby, popřípadě sledovat její pohyb. K základním klíčovým bodům patří hlava, ramena, lokty, zápěstí, kyčle, kolena a kotníky, viz obrázek 6.1. Některé algoritmy dokážou rozpoznat i orientaci dlaně či stopy, nebo rozpoznat klíčové body na hlavě, jako jsou ústa, nos, oči a uši [26].

Klíčové body jsou většinou reprezentovány jako dvojice souřadnic  $(x, y)$  vzhledem k celému obrazu, některé algoritmy poskytují i souřadnice normalizované vzhledem k bounding boxu osoby. Existují také algoritmy pro 3D souřadnice, těmito se ale nebudeme zabývat, i když by mohly stanovit zajímavou alternativu, zejména pokud by pro detekci bylo použito více kamer z různých pohledů.

V oblasti algoritmů pro detekci pózy existují dva základní přístupy: zdola nahoru a shora dolů. Přístup zdola nahoru se snaží detektovat všechny klíčové body v obraze, aniž by rozlišoval jednotlivé osoby, pokud je algoritmus schopen detekce pózy pro více osob, pak v dalším kroku tyto body



Obrázek 6.1: (Vlevo) Topologie klíčových bodů použitá např. v COCO-pose [25] (Vpravo) Příklad detekce pózy pomocí YOLO.

spojuje do jednotlivých postav. Naproti tomu přístup shora dolů nejprve detekuje všechny osoby v obraze, v jejich rámci pak detekuje klíčové body.

## 6.2 Detekce klíčových bodů

### Heatmapy

U obou výše zmíněných přístupů se nejčastěji provádí vyhledání všech klíčových bodů pomocí tzv. heatmap. Je to 2D mapa pravděpodobnosti, že se v daném bodě vyskytuje nějaký klíčový bod. Maximální hodnoty v této mapě pak představují lokalizaci klíčových bodů.

Pro vygenerování heatmap se používá konvoluční neuronová síť. Pro každý klíčový bod, resp. pro každý typ klíčového bodu  $k$  (v případě detekce pózy více osob) vzniká jedna heatmapa. Jako referenční heatmapy pro trénování se používají mapy, kde je klíčový bod reprezentován 2D Gaussovým rozložením s vrcholem v místě daného bodu.

V dalším kroku jsou z heatmap vygenerovány, nejčastěji s pomocí algoritmu argmax, souřadnice klíčových bodů. V případě více osob je pak třeba tyto body spojit do jednotlivých osob.

### Regresce

Využití heatmap je velmi přesné, nicméně z důvodu nutnosti provádění dvou sekvenčních výpočtů je také trochu pomalé. Také komplikují proces trénování, jelikož musíme spolu s trénovacími daty

dodat modelu i heatmapy. Některé algoritmy se proto snaží formulovat úlohu jako regresi vedoucí přímo k souřadnicím klíčových bodů. Tento přístup je ve své podstatě trochu méně přesný, nicméně je rychlejší.

Vůbec první algoritmus pro detekci pózy využívající hluboké učení, DeepPose [27], který byl vytvořen v roce 2014 společností Google, používal právě regresi. Také algoritmus YOLO používá regresi pro určení souřadnic klíčových bodů, nicméně detekce je prováděná pro detekované objekty, nikoliv nad celým vstupním obrazem [28].

### 6.3 Detekce objektů a osob v obraze

Detekce osob se v podstatě může generalizovat na detekci objektů v obraze. Detekci objektů v obraze definujeme jako úlohu, kdy ve vstupním obrázku určíme lokalizaci a třídu všech hledaných objektů. Lokalizace je většinou reprezentována jako souřadnice obdélníku ohraničujícího daný objekt, tzv. bounding box.

V kapitole 3 jsme si popsali základní architekturu konvolučních neuronových sítí, ta se ale většinou v praxi používá pro klasifikaci obrázků, nikoliv pro detekci objektů – algoritmus tedy pouze určí, o jakou třídu objektu se jedná, a ideálně potřebuje, aby objekt vyplňoval celý vstupní obraz. Teoreticky by bylo možné detekci formulovat jako regresní problém a natrénovat takovou síť, která by pomocí několika konvolučních vrstev následovaných několika plně propojenými vrstvami byla schopna predikovat lokalizaci a třídu všech objektů v obraze [29]. Problém detekce je ale velice komplexní a také by vyžadoval velice komplexní síť – více vrstev s mnoha filtry, resp. neurony. Jak již ale bylo zmiňováno, komplexnost sítě zvyšuje její nároky na výpočetní výkon a komplikuje nebo úplně znemožňuje její trénování s ohledem na pravděpodobnost přetrénování.

Snahou tedy bylo najít metody, které poupraví architekturu sítě tak, aby byla schopna efektivní detekce objektů. Většina těchto metod se nějakým způsobem snaží rozdělit vstupní obrázek na menší části, ty následně jednak klasifikovat, a tedy určit, zda se v dané lokalitě vyskytuje objekt, popřípadě pomocí regrese určit jeho přesnou lokalizaci. Rozdelení může být provedeno přímo na vstupním obrázku nebo na mapě příznaků v rámci sítě.

Metoda sliding window (klouzavé okno), která aplikuje hrubou sílu a projde veškeré možné oblasti, je samozřejmě velice neefektivní, a tak se další metody snaží buď najít pouze oblasti, ve kterých je pravděpodobné, že se nějaký objekt nachází – dvoufázový přístup, anebo rozdělí obrázek do mřížky – jednofázový přístup.

Jelikož jsou tyto metody základem pro většinu detektorů klíčových bodů, pojďme si několik základních popsat.

### 6.3.1 Dvoufázový přístup

#### R-CNN

Prvním algoritmem, který efektivně zredukoval počet oblastí pro klasifikaci, byl algoritmus R-CNN (Region-based Convolutional Network) [30]. Tento algoritmus nejprve použil některou z dostupných metod (autoři použili selective search) pro vygenerování navržených oblastí (region proposals), které pravděpodobně obsahují nějaký objekt. Tyto metody jsou nezávislé na třídě objektů. Algoritmus tedy vygeneruje zhruba 2000 oblastí, vzniklé obrázky jsou následně upraveny na velikost požadovanou CNN v další fázi. CNN extrahuje z dané oblasti mapu příznaků, na její základě plně propojené vrstvy predikují třídu objektu popřípadě jeho bounding box.

Problémem R-CNN je, že výběr oblasti a jejich následná klasifikace jsou nezávislé úlohy a jsou nezávisle trénovány. Detekce objektu je také poměrně pomalá, protože je extrakce příznaků prováděná pro všechny oblasti zvlášť. Tyto problémy se snaží řešit další upravené verze R-CNN.

#### Fast R-CNN

První z nich je Fast R-CNN [31], která je upravená tak, aby bylo možné provádět trénování v jednom kroku. Také extrahuje příznaky pro celý vstupní obraz najednou, pomocí selective search pak identifikuje oblasti zájmu (ang. region of interest – RoI), které následně použije pro klasifikaci a regresi. Tato metoda je přesnější a asi desetkrát rychlejší než původní R-CNN.

#### Faster R-CNN

Další algoritmus, Faster R-CNN [32], nahrazuje metodu selective search vlastní, plně konvoluční síť RPN (region proposal network). Zefektivňuje tak proces trénování, výsledná síť je také rychlejší a přesnější než Fast R-CNN.

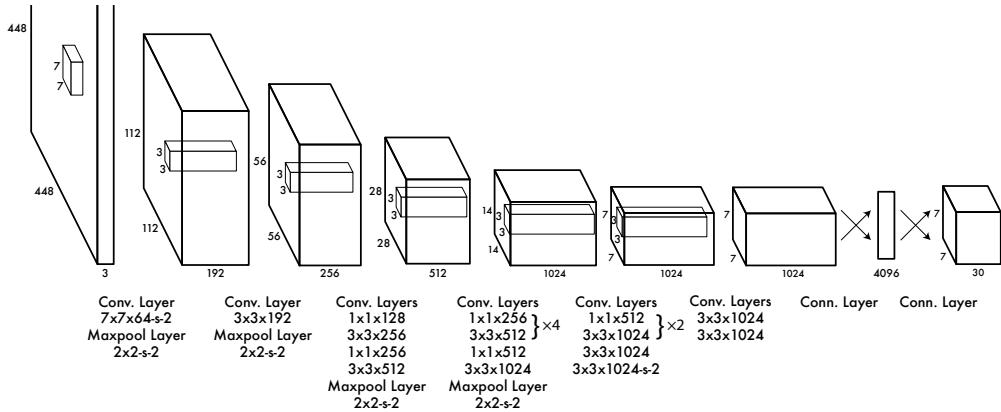
### 6.3.2 Jednofázový přístup

Jednofázový přístup se snaží najít řešení, ve kterém není nutné hledat navržené oblasti, ale provést klasifikaci a regresi na předem dané množině oblastí, obvykle určené mřížkou.

#### YOLO

Prvním takovým algoritmem byl YOLO (také YOLOv1, z ang. you only look once) [33]. Ten, v původní verzi, rozdělí vstupní obraz do pevně dané mřížky velikosti  $S \times S$  a v každém z těchto polí určí  $B$  bounding boxů a jejich třídu. V původní verzi bylo zvoleno  $S = 7$  a  $B = 2$ .

Obraz je nejprve zpracován pomocí konvolučních vrstev, které extrahují mapu znaků o velikosti  $S \times S \times K$ , kde  $K$  je počet kanálů. Každý pixel této mapy představuje jedno pole mřížky. Dále je



Obrázek 6.2: Architektura původní verze YOLO [33]

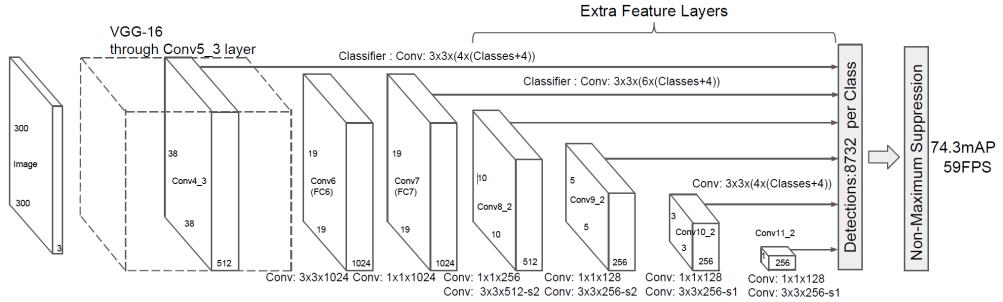
mapa zpracována plně propojenými vrstvami, které provádějí nad každým polem mřížky klasifikaci a regresi, viz obrázek 6.2.

Každý bounding box je reprezentován souřadnicemi středu a velikosti (šířka a výška). Dohromady s informací o jistotě detekce bounding boxu (confidence score) vrátí model 5 informací o každém bounding boxu. Pro každé pole mřížky pak určí společnou informaci o třídě všech objektů v daném poli. Pokud objekt není detekován, třída indikuje pozadí a souřadnice bounding boxu jsou ignorovány. Velikost výstupního vektoru je tedy  $7 \times 7 \times (2 * 5 + C)$ , kde  $C$  je počet definovaných tříd – v původní verzi pouze 20.

Tento algoritmus, navržený v roce 2015 J. Redmonem et al., byl revolučně rychlý, zároveň v porovnání s jinými real-time detekčními algoritmy dosahoval i slušné přesnosti. Nicméně byl velice citlivý na velikost objektu a přesnost detekce, zejména u menších objektů, byla horší než u dvoufázových algoritmů.

Další verze algoritmu YOLO přinesly postupná vylepšení ve formě optimalizace trénování a architektury. YOLOv2 [34] zavedl mj. trénování na několika mřítkách a byl natrénován s 9000 třídami (proto také nazýván YOLO9000). YOLOv3 [35] přinesl mj. detekci na několika mřítkách. Postupně byla také zvětšována mřížka a měnila se použitá architektura CNN sítě sloužící pro extrakci příznaků pro jednotlivá pole mřížky. Postupně také byly přidávány další funkce jako je segmentace, detekce pózy či sledování objektů (ang. tracking).

V 2020 roce firma Ultralytics poprvé implementovala YOLO s využitím populární knihovny PyTorch (YOLOv5), což umožnilo snadnější využití YOLO v praxi. Firma Ultralytics také vytvořila framework pro použití různých verzí YOLO (YOLOv3 a novější). Také pracuje na dalších vylepšeních a optimalizacích. Konkrétně vytvořila YOLOv5 (2020), YOLOv8 (2023) a YOLOv11 (2024). Tyto verze nicméně nejsou podloženy odbornými články, někteří je tak považují za neoficiální verze.



Obrázek 6.3: Architektura SSD [36]

## SSD

Dalším populárním algoritmem, který používá jednofázový přístup, je SSD (z ang. single shot detector) [36]. Ten rozdělí vstupní obraz do několika mřížek o různé velikosti. Postup je takový, že nejprve projde obraz konvoluční sítí, konkrétně sítí VGG16, která extrahuje mapu příznaků. Tu se postupně dalšími konvolučními vrstvami zmenšuje, výstup každého stádia zmenšení, reprezentující mřížku dané velikosti, se spolu s původní mapou dále zpracovává plně propojenou sítí.

Výstupní bounding boxy nejsou, jako v případě YOLO, pouze výsledkem regrese, ale pro každé pole dané mřížky je definováno několik výchozích oblastí (ang. default box), ze kterých jsou vybrány ty, které obsahují objekt. K nim je predikována třída objektu a posun i změna velikosti výchozí oblasti, upřesňující výsledný bounding box.

V době svého vzniku byl SSD rychlejší a přesnější než YOLO, ale novější verze YOLO jej už předběhly. Nicméně některé principy SSD, jako výchozí oblasti či použití různých měřítek, byly převzaty do novějších verzí YOLO.

## 6.4 Charakteristiky vybraných implementací pro detekci pózy

V této sekci bude popsáno několik populárních algoritmů (a jejich implementací) pro detekci pózy, zejména se zaměřením na jejich rychlosť, přesnost a specifikaci architektury.

Velká část algoritmů byla zamítnuta a neproběhlo ani jejich testování. Nejčastějším důvodem zamítnutí některého z algoritmů je, že schází jeho volně dostupná, aktualizovaná implementace. Tvorci algoritmů většinou investují čas a zdroje pro vývoj algoritmu a natrénování modelu (často pro akademické účely), pak ale neinvestují do jeho údržby. Zejména v prostředí Pythonu, kde se neustále vyvíjejí nové knihovny a zpětná kompatibilita starších verzí není zaručena, je pak obtížné použít takovéto řešení ve svém projektu, aniž by bylo nutné investovat další čas do pochopení zdrojového kódu a jeho úpravy. Jak již bylo zmíněno na počátku kapitoly, implementace algoritmu od nuly by vyžadovala velké množství času a zdrojů (zejména výpočetních), hlavně by také byla potřebná hlubší znalost problematiky. Někdy je možné řešení použít za cenu kompromisu ve formě použití

starších verzí knihoven či Pythonu, může to ale představovat bezpečnostní rizika. Také některé algoritmy jsou sice volně dostupné, ale jejich implementace jsou součástí komerčních produktů, např. OpenPose, jež je mj. součástí produktu Viso Suite od viso.ai [37].

#### 6.4.1 DeepPose

DeepPose je historicky první algoritmus pro detekci pózy využívající hluboké učení. Vyvinuli jej Alexander Toshev a Christian Szegedy ze společnosti Google v roce 2014 [27]. Algoritmus předpokládá, že se ve vstupním obraze nachází pouze jedna osoba. Sít se snaží v jednom kroku pomocí regrese jak detektovat osobu, tak i její klíčové body. Jelikož je těžké takto dosáhnout velmi přesných výsledků, algoritmus používá další fázi, která pomocí regrese provádí posun bodů k přesnějším výsledkům. Tato fáze je aplikována opakováně, kaskádně se tak zvyšuje přesnost detekce.

Při svém vzniku byl DeepPose revoluční, nicméně v porovnání s dnešními řešeními je poměrně pomalý a nepřesný. Nicméně položil základ pro využití hlubokého učení v oblasti detekce pózy.

#### 6.4.2 OpenPose

OpenPose [38] je typicky příklad přístupu zdola nahoru. Jeho výhodou je ale možnost vyhledání více osob v jednom snímku. Tento algoritmus, který vyvinuli v roce 2019 Zhe Cao et al., nejprve pomocí CNN vytvoří heatmapu pro každý typ klíčového bodu. Pro spojení bodů do jednotlivých osob využije pole propojení klíčových bodů (ang. part affinity field – PAF). PAF je mapa vytvořená pro každou končetinu (myšleno obecně spojení dvou klíčových bodů), která v oblasti dané končetiny obsahuje hodnoty určující směr z jednoho bodu do druhého. Pokud pak dáme do hromady informace z heatmap a z PAF, jsme schopni poměrně jednoznačně zkompletovat jednotlivé klíčové body do celých postav. Stejně jako heatmapy, jsou i PAF součástí trénovacích dat.

#### 6.4.3 OpenPifPaf

Algoritmus OpenPifPaf [39], vyvinutý v roce 2021 Svenem Kreissem et al., je v podstatě vylepšenou verzí OpenPose. Jeho název je odvozen od dvou stavebních kamenů: PIF (Part Intensity Field) – pole intenzity klíčových bodů, a PAF (Part Affinity Field) – pole propojení klíčových bodů. PIF je rozšířením heatmap, kdy kromě intenzity pravděpodobnosti klíčového bodu obsahuje i jeho posun, zaručující přesnější lokalizaci bodu, a odhadovanou velikost dané části těla. PAF v OpenPifPaf je také podobný tomu v OpenPose, navíc ale indikuje kromě směru i velikost dané končetiny, což umožňuje lepší prostorové zachycení pózy.

Dalším rozšířením oproti OpenPose je možnost sledování osob ve videu. Mapa příznaků, která je výsledkem vstupní CNN, je udržována v mezipaměti, do další části sítě pak vždy vstupují mapy pro aktuální a předchozí snímek. Výstupem pak kromě klíčových bodů v každém snímku a jejich propojení tvořící kostru, jsou i propojení mezi klíčovými body z jednotlivých snímků, viz 6.4. Algo-



Obrázek 6.4: Vizualizace sledování osoby mezi dvěma snímky v OpenPifPaf [39]

ritmus si pak udržuje ID sledovaných osob, pokud k dříve nalezené osobě je nalezena nová pozice, je jí přiřazeno stejné ID. Pokud je nalezena nová osoba, je jí přiřazeno nové ID.

U OpenPifPaf je k dispozici výběr několika páteřních modelů, jako je ResNet50 či ShuffleNet v různých variantách a velikostech. Můžeme tak zvolit model, který je kompromisem mezi výkonem a přesností, v závislosti na konkrétních požadavcích aplikace.

#### 6.4.4 MediaPipe – BlazePose

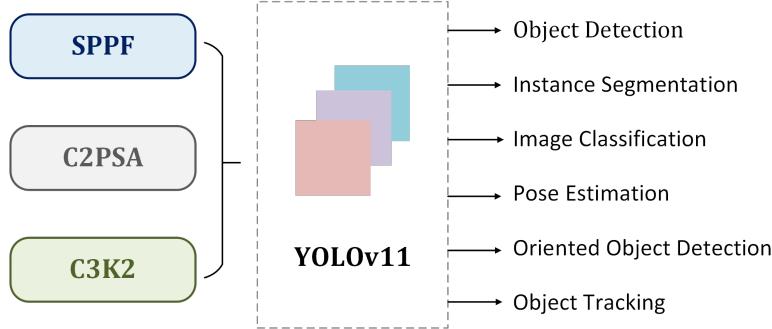
MediaPipe je framework vyvinutý společností Google, umožňující jednoduchou integraci různých technik strojového učení. Obsahuje různé algoritmy pro řešení úloh jako detekce objektů, segmentace či detekce klíčových bodů (tváře či pózy). MediaPipe je optimalizován pro mobilní zařízení a webové aplikace. Detekce pózy v tomto frameworku je postavená na algoritmu BlazePose.

BlazePose implementuje přístup shora dolů, detekuje tedy nejprve RoI, ve kterých detekuje osobu a její pózu. Nativně podporuje pouze jednu osobu ve snímku. Ve videu ale v rámci optimalizace neprovádí detekci RoI pro každý snímek, pouze pokud v aktuální RoI již není detekována osoba. Výhodou tohoto algoritmu je, že detekuje 33 bodů v postavě, což je podstatně více, než většina ostatních algoritmů, umožňuje tak přesnější analýzu některých situací, např. podle natočení tváře, dlaní či stop.

Framework MediaPipe implementuje BlazePose spolu s detekcí více osob (použije v první fáze detektor objektů) i sledování. Výhodou tohoto frameworku je jeho kontinuální vývoj a jednoduchost integrace. Nevýhodou ale je, že pro systémy Windows není implementována podpora GPU. Jelikož náš výsledný produkt bude spouštěn primárně na Windows zařízeních, je pro nás tato vlastnost rozhodující. Model je dostupný v třech velikostních variantách: *Lite*, *Full* a *Heavy*.

#### 6.4.5 YOLO

Od vydání YOLOv7 v roce 2022 integruje framework YOLO i detekci pózy. Oficiální článek Chien-Yao Wanga et al. [40], sice neobsahoval tuto funkčnost, ale oficiální implementace zahrnula i implementaci YOLO-Pose [28]. Obecně detekce pózy v YOLO kombinuje přístup shora dolů a zdola



Obrázek 6.5: Architektura YOLOv11 [41]

nahoru. Algoritmus sice vyhledává klíčové body spolu s bounding boxy osob, nicméně vše v jednom kroku. Samotná detekce klíčových bodů využívá regresi, což zjednoduší proces trénování, jelikož není třeba tvořit heatmapy.

Architektura použitá v YOLO-Pose se ale liší od architektury používané v pozdějších verzích. V YOLO-Pose jsou na konci řetězce umístěny hlavy pro různá měřítka, jejich výstupem jsou bounding boxy a klíčové body, oba tvořené spolu. V pozdějších verzích je architektura YOLO koncipována universálněji pro různé úlohy. Obsahuje tak tři fáze[41]: páteř (ang. backbone), která extrahuje mapu příznaků, krk (ang. neck), který přizpůsobuje mapu příznaků pro různá měřítka, a hlavy (ang. head), které paralelně zpracovávají výstupy pro různé úlohy, viz obrázek 6.5 . Bounding box a klíčové body jsou tedy sice generovány paralelně a teoreticky nezávisle, nicméně s ohledem na proces trénování a postprocessing se v praxi navzájem výrazně ovlivňují.

Nejnovější verze YOLO také podporují kombinaci detekce klíčových bodů a sledování osob. Model tedy kromě klíčových bodů vrací ID dané osoby, pomocí kterého můžeme spojit danou postavu s předchozími snímky. Můžeme tak efektivně analyzovat pohyby jednotlivých osob.

Jednou z výhod frameworku YOLO, zejména verzí vyvíjených firmou Ultralytics, je široká škála velikostí modelů. Každý model je dostupný v pěti variantách: *Nano*, *Small*, *Medium*, *Large*, *Xlarge*.

#### 6.4.6 Torchvision

Torchvision je knihovna, která je součástí frameworku PyTorch. Obsahuje různé nástroje pro strojové vidění, jako je detekce objektů či segmentace. Její součástí je i předtrénovaný model pro detekci pózy, který implementuje algoritmus Keypoint R-CNN [42].

Algoritmus Keypoint R-CNN je založen na stejné myšlence jako Mask R-CNN [43], a tedy rozšíření Faster R-CNN o další hlavu, která v případě Mask R-CNN provádí segmentaci, v případě Keypoint R-CNN detekuje klíčové body. V algoritmu je také upravená pooling vrstva, která zajišťuje, že se výstupy algoritmu shodují s vstupy s přesností pixelu.

Tuto implementaci budeme testovat zejména z důvodu její jednoduchosti použití a integrace do frameworku PyTorch, který budeme používat i v další části řešení. Na druhou stranu máme oproti jiným frameworkům, jako je YOLO či MediaPipe, k dispozici pouze jeden model, nikoliv více velikostních variant.

## 6.5 Testování a porovnání vybraných algoritmů pro detekci pózy

Pro testování jsme vybrali čtyři algoritmy pro detekci pózy, zejména na základě jejich jednoduchosti implementace, aktualizované podpory a požadovaných funkcí. Budeme tedy testovat algoritmy Torchvision, OpenPifPaf, MediaPipe BlazePose a YOLO v nejnovější verzi 11. Testy byly provedeny na 25 videích ze stejného datasetu jako byl použit později pro trénování algoritmu pro analýzu klíčových bodů. Testování probíhalo na počítači s procesorem Intel Core i7, 32 GB RAM a grafickou kartou NVIDIA GeForce RTX 3080 s 10 GB VRAM.

Algoritmy Torchvision a OpenPifPaf byly testovány s použitím GPU, MediaPipe pouze s využitím CPU, jelikož nemá podporu GPU v prostředí Windows. Algoritmus YOLO jsme testovali na GPU a CPU, abychom porovnali jeho výkon v různých podmírkách. Algoritmus YOLO jsme také vyzkoušeli na GPU s využitím funkce sledování.

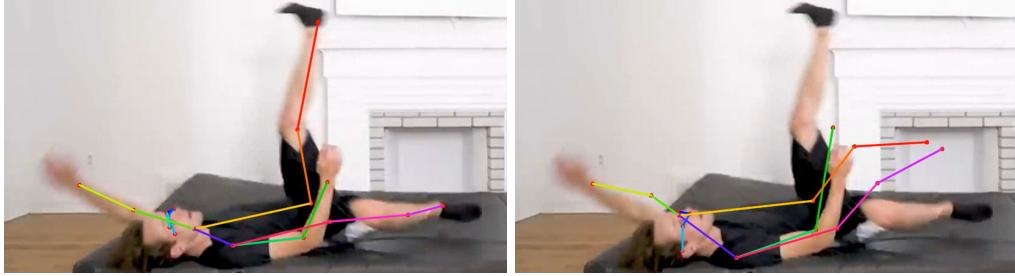
Výstupem testování pro daný algoritmus a jeho variantu je průměrná doba zpracování jednoho snímku a videa s vykreslením klíčových bodů. Tyto videa nám dovolují ověřit schopnost detekce klíčových bodů v různých situacích a její přesnost.

Podíváme se nyní na výsledky testování jednotlivých algoritmů. Pro každý algoritmus porovnáme jednotlivé varianty s ohledem na rychlosť a přesnost a vyhodnotíme jeho výhody či nevýhody oproti ostatním algoritmům. Nakonec zvolíme model, který použijeme v další části vývoje.

### 6.5.1 Výkonové požadavky

Při výběru algoritmu musíme s ohledem na práci v reálném čase brát v úvahu zejména výkon detekčního algoritmu. Bezpečnostní kamery mají obvykle snímkovou frekvenci od 15 do 30 snímků za sekundu (FPS), ideální by tedy bylo, aby konečný program byl schopen pracovat s frekvencí alespoň 30 FPS. Zároveň se předpokládá, že v prostředí, kde bude program nasazen, bude dostupná grafická karta.

V této fázi jsme již zkoušeli trénování neuronové sítě pro klasifikaci pózy, a ověřili jsme, že i v případě hlubších a komplexnějších sítí dosahujeme doby inference v řádu nižších jednotek milisekund. Hlavní vliv na výslednou rychlosť programu tedy bude mít hlavně detekční algoritmus, který musí zpracovávat mnohem větší objem dat – stovky tisíc až miliony pixelů oproti např. 17 klíčovým bodům v případě klasifikačního algoritmu.



Obrázek 6.6: Porovnání přesnosti *MediaPipe Lite* (vlevo) a *YOLO Large* (vpravo)

### 6.5.2 OpenPifPaf

Algoritmus OpenPifPaf jsme zkoušeli v několika variantách, postavených na síti *ResNet 50* a *ShuffleNet V2* [44]. V tabulce 6.1 vidíme, že většina variant ani zdaleka nedosahuje požadovaného výkonu.

Tento algoritmus je poměrně robustní z pohledu světelních podmínek a rozlišení či rozmazení obrazu, jinak ale dosahuje nejhorší přesnosti ze všech testovaných algoritmů. Kromě nedetekování části těla, které nejsou vidět (jsou např. schovány za jinou částí těla), totiž často nedetekuje člověka vůbec, nejčastěji pak když člověk padá nebo leží, což jsou situace pro nás stěžejní. Hlavně tento problém vystupuje ve variantě *resnet50* a *shufflenetv2k16*, jsou tak pro nás nepoužitelné. Varianty *shufflenetv2k30* a *tshufflenetv2k30* by sice s ohledem na kvalitu výsledků použitelné byly, nicméně je jejich výkon příliš nízký.

Tabulka 6.1: Porovnání výkonu modelu OpenPifPaf

Verze	inference [ms]	Frekvence [FPS]
resnet50	49,2	20.324
shufflenetv2k16	31,3	31.910
shufflenetv2k30	58,6	17.070
tshufflenetv2k30	70,0	14.281

### 6.5.3 MediaPipe BlazePose

Algoritmus BlazePose z knihovny MediaPipe jsme testovali ve třech variantách: *Lite*, *Full* a *Heavy*. Ve všech variantách tento algoritmus dosahoval velmi přesných výsledků, asi nejlepších ze všech testovaných algoritmů. Na rozdíl od jiných algoritmů totiž, pokud detekoval osobu, vždy velmi přesně označil všechny její klíčové body, v rámci možností i ty, které byly hůře viditelné (např. schovány za jinou částí těla). Na obrázku 6.6 je vidět rozdíl v přesnosti detekce klíčových bodů mezi nejmenší variantou BlazePose a druhou největší variantou YOLO, kdy BlazePose dosahuje mnohem větší přesnosti, v tomto příkladě zejména co se týče detekce nohou.



Obrázek 6.7: Příklad špatné detekce bodů v modelu Torchvision Keypoint R-CNN

S ohledem na to, že nemáme k dispozici grafickou akceleraci, je jeho výkon velmi dobrý, viz tabulka 6.2. Verze *Lite* a *Full* by tak mohla být v našem řešení použitelná, což by nám také dávalo možnost nasazovat výsledný program v mobilních zařízeních či jiných systémech bez grafické karty.

Tabulka 6.2: Porovnání výkonu modelu MediaPipe BlazePose

Verze	inference [ms]	Frekvence [FPS]
lite	25.5	39.239
full	30.9	32.405
heavy	68.2	14.655

Algoritmus si ale velice špatně radí s horšími světelnými podmínkami či menším rozlišením obrazu. Ve většině případů sice detekuje klíčové body velmi přesně, pokud je ale osoba hůř viditelná, nedetekuje ji vůbec. Tento problém se projevuje ve všech variantách podobně. Jelikož bude náš výsledný program nasazován spíše právě v podmínkách s horším osvětlením a ve větší vzdálenosti od osob, pravděpodobně se pro naši aplikaci nebude hodit.

#### 6.5.4 Torchvision Keypoint R-CNN

Torchvision Keypoint R-CNN nedosáhla ani dostatečného výkonu, viz tabulka 6.3, ani kvalitních výsledků. Podobně jako *OpenPifPaf* má totiž problém, když osoba padá anebo leží. V tomto případě osobu často detekuje, ale naprostě ztrácí přesnost detekovaných klíčových bodů, nejčastěji záměnou jednotlivých bodů, viz obrázek 6.7. Zároveň oproti BlazePose nemá takový problém d horšími světelnými podmínkami a menším rozlišením obrazu.

Tabulka 6.3: Výkon modelu Torchvision Keypoint R-CNN

inference [ms]	Frekvence [FPS]
50.3	19.897

## 6.5.5 YOLO

Algoritmus YOLO ve verzi 11 jsme testovali v pěti variantách: *Nano*, *Small*, *Medium*, *Large* a *Xlarge*. Všechny varianty byly testovány na GPU i CPU.

V tabulce 6.4 vidíme, že na CPU dosahuje tento algoritmus poměrně špatných výsledků. Jediný použitelný by pro nás mohl být v takové situaci model *Nano*. YOLO je ale velice kvalitně optimizováno pro grafické karty, můžeme tak pozorovat, že na GPU je výkon výrazně vyšší. Pro naše potřeby by tak byly použitelné prakticky všechny varianty.

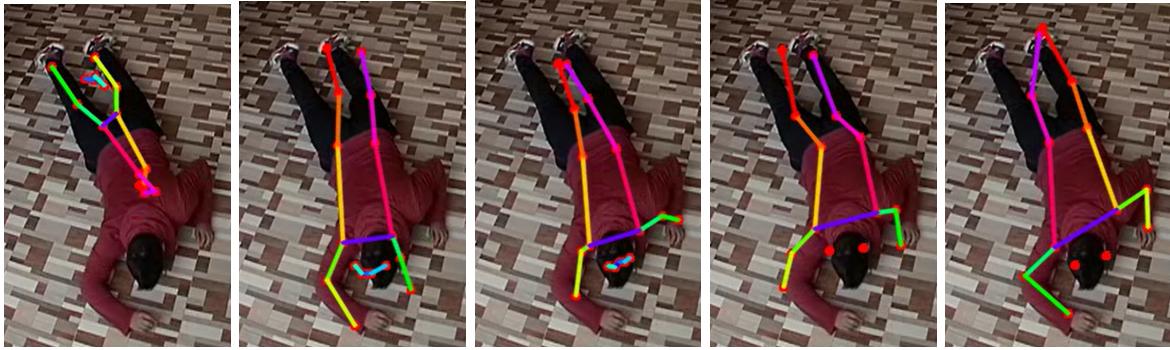
Jelikož pro analýzu více osob v jednom snímku je potřeba, zejména v případě použití rekurentní neuronové sítě, jednotlivé osoby od sebe oddělit a identifikovat i mezi snímky, bude pro nás velmi užitečná funkce sledování objektu. Proto jsme otestovali algoritmus YOLO i s touto funkcí. Jak je vidět v tabulce 6.4, je výkon sice horší než bez sledování, pořád ale tři menší varianty dosahují frekvence větší než 30 FPS.

Tabulka 6.4: Porovnání výkonu modelu YOLO

	<b>Verze</b>	<b>inference [ms]</b>	<b>Frekvence [FPS]</b>
CPU	nano	32.5	30.749
	small	53.9	18.563
	medium	114.1	8.763
	large	143.4	6.973
	xlarge	833.2	1.200
GPU	nano	15.1	66.323
	small	15.2	65.972
	medium	17.4	57.500
	large	24.4	41.026
	xlarge	24.4	41.005
GPU se sledováním	nano	27.4	36.500
	small	27.7	36.148
	medium	29.5	33.882
	large	37.5	26.664
	xlarge	40.5	24.695

Všechny varianty algoritmu YOLO dosahují poměrně kvalitních výsledků. I v horších světelnných podmínkách vždy detekují osobu, a víceméně přesně určí její klíčové body. Obecně je ale vidět, že je model trochu méně robustní (než např. MediaPipe) v situacích, kdy není dobře vidět některá končetina – je třeba schovaná za jinou částí těla, anebo ve specifických půzách – např. když je osoba v dřepu nebo je v obrazu natočená vzhůru nohama. V takovýchto případech dosahuje menší přesnosti pro jednotlivé body – detekuje jiné natočení končetiny nebo v extrémních případech špatně vyhodnotí natočení celé postavy. Špatně viditelné části těla pak často vůbec nedetekuje.

Z pohledu přesnosti je zde výrazně vidět vliv velikosti modelu na přesnost detekce. Varianta *Nano* ve výše zmíněných situacích někdy detekuje body zcela špatně, a je tak prakticky nepoužitelná.



Obrázek 6.8: Porovnání přesnosti variant modelu YOLO v situaci s netypickým natočením postavy. Zleva: *Nano*, *Small*, *Medium*, *Large*, *Xlarge*

telná. Varianta *Small* je znatelně lepší, pořád ale v horších podmírkách vyhodnocuje mnoho části těla špatně – např. zamění nohy. Varianta *Medium* je už výrazně lepší. Není sice ideální, ve většině ale vyhodnotí všechny části těla správně i když ne z přesnosti několika pixelů. Varianty *Large* a *Xlarge* jsou pak už velmi přesné, projevuje se to ale znatelně menší rychlostí.

Je zajímavé pozorovat, že i když je algoritmus BlazePose v mnoha situacích mnohem přesnější, než i větší varianty YOLO, viz obrázek 6.6, při horší viditelnosti, kdy i nejmenší varianty YOLO detekují osobu, BlazePose zcela selhává. Ve snímku na obrázku 6.8 například nedetekoval BlazePose osobu vůbec.

### 6.5.6 Shrnutí a výběr modelu

Z našeho testování jsme zjistili, že modely OpenPifPaf a Torchvision jsou natrénované spíše pro detekci postavy, když je osoba v běžnějších půzách, jako je ve stojaném či chůze. Stejně i u menších variant YOLO přesnost prudce klesá v méně typických půzách či natočených v obraze. Jelikož ale je naše práce postavena právě na detekování více netypické postavy, je pro nás důležité detektovat pozici ve všech situacích.

Naopak algoritmus BlazePose je velmi přesný, strádá ale při horší viditelnosti osoby. Nehodí se tak pro naše využití s kamery s horším rozlišením a vysokou vzdáleností od osob. Ostatní algoritmy v tomto ohledu jsou mnohem robustnější, nejlépe si s horšími podmínkami poradí algoritmus YOLO.

Optimální cesta se tedy zdá být algoritmus YOLO ve verzi medium, kdy dosahuje dostatečné rychlosti i při sledování osob, zároveň dostatečně přesně detekuje pózy ve všech pozicích i podmínkách. Pokud bychom nevyužívali pro analýzu pózy rekurentní neuronovou síť, nemuseli bychom používat funkci sledování a mohli bychom tak použít i větší variantu YOLO, což by mohlo zlepšit přesnost detekce.

## Kapitola 7

# Implementace neuronové sítě pro klasifikaci klíčových bodů

Problematika vyhodnocování je velmi široká a přináší mnoho problémů. Ostatně i člověk někdy může špatně interpretovat chování druhé osoby. Například pokud někdo skáče do postele či jinak prudce lehá, může to vypadat jako nebezpečná situace. Stejně se v počítačovém vidění nevyhneme falešným poplachům, nicméně se budeme snažit zapojit různé techniky pro zlepšení přesnosti našeho detektoru.

Naším úkolem je nyní vytvořit algoritmus, který pro danou pózu (reprezentovanou klíčovými body), resp. sekvenci takových pór (získané pro jednu osobu ze sekvence snímků), určí, zda se jedná o situaci pádu, či nikoliv. Tento algoritmus bude přijímat vždy pózu jedné osoby, funkcionality pro více osob bude řešená později.

V kapitole 6 jsme pro detekci klíčových bodů zvolili model *YOLO pose*. Ten je předtrénovaný na datasetu *COCO*, který pro lidské pózy definuje 17 klíčových bodů v dvourozměrném prostoru. To udává velikost vstupu do našeho klasifikačního algoritmu. Navrhнемe tedy a natrénujeme neuronovou síť, jejíž vstupem bude 17 2D klíčových bodů – tedy 34 čísel – a výstupem bude klasifikace třídy pózy – *normální* a *upadl*.

V této kapitole se zaměříme na implementaci klasifikačního algoritmu pro analýzu pózy. Nejdříve se podíváme na použité technologie a knihovny, které nám pomohou s vývojem. Poté rozebereme možné architektury sítě a ukážeme, jak byly implementovány. Dále se zaměříme na návrh vnitřní struktury u těchto architektur, zkusíme je natrénovat v různých konfiguracích a vybereme nejoptimálnější řešení. Nakonec toto řešení otestujeme na testovacím datasetu.

## 7.1 Použité technologie

### 7.1.1 PyTorch

Celý projekt byl vyvíjen v prostředí skriptovacího jazyka Python. Samotný vývoj neuronových sítí probíhal ve frameworku PyTorch. PyTorch je open-source knihovna pro strojové učení, široce používaná například pro počítačové vidění či zpracování přirozeného jazyka. Jeho hlavními funkcemi je práce s tenzory podobnými jako v NumPy, ale s podporou silné akcelerace s využitím GPU, a vývoj neuronových sítí postavený na vysokoúrovňových stavebních blocích s podporou automatické derivace pro počítání gradientů.

Implementace neuronových sítí v PyTorch je velice jednoduchá a díky vysoké míře abstrakce umožňuje se při vývoji soustředit na samotnou architekturu a design sítě, nikoliv na detaily implementace. Pro vytvoření nové neuronové sítě stačí vytvořit třídu, která bude dědít od základní třídy *nn.Module*, v konstruktoru definovat jednotlivé vrstvy včetně různých regularizačních hyperparametrů. V metodě *forward* pak definujeme dopředný průchod sítě. Dále máme možnost kontrolovat např. výpočet ztrátové funkce či metriky přesnosti.

PyTorch obsahuje také předpřipravené moduly pro GRU či LSTM sítě, včetně vícevrstvých architektur. Pro tyto moduly definujeme velikost vstupu a velikost skrytého stavu, dále můžeme definovat počet vrstev či dropout. Výstup z této modulů je pak skrytý stav, který můžeme dále zpracovávat pomocí jednoduché dopředné sítě pro predikci třídy.

### 7.1.2 Lightning

PyTorch Lightning je wrapper pro PyTorch, který dále usnadňuje vývoj neuronových sítí. Stará se za nás o detaily procesu trénování, jako je správa epoch, logování či optimalizace kroku učení (ang. learning rate). Podporuje také nativně práci s TensorBoard, což je logovací nástroj umožňující přehledné sledování metrik během trénování, včetně grafického zobrazení ve webovém prostředí.

Pro implementaci modelu vytvoříme třídu, která dědí z *lightning.LightningModule*, a kromě konstruktoru, ve kterém inicializujeme jednotlivé vrstvy, definujeme metody *training\_step* (krok trénování), *validation\_step* – krok validace, *configure\_optimizers* – definování optimalizační techniky a *forward*, která definuje, jak signál prochází jednotlivými vrstvami.

## 7.2 Implementace vybraných architektur

Podíváme se nyní, jak jsme jednotlivé architektury implementovali, nezávisle na jejich vnitřní struktuře a konfiguraci.

### 7.2.1 Dopředná neuronová síť

Nejjednodušší architekturou, kterou můžeme pro náš model použít, je dopředná neuronová síť. Tento model pak bude klasifikovat jednotlivé pózy, aniž by znal jejich kontext. Síť bude klasifikovat klíčové body pouze podle aktuální lokalizace ve snímku, nikoliv podle pohybu.

Výhodou této architektury je jednoduchost, potažmo rychlosť. Síť nepotřebuje mnoho parametrů a oproti rekurentním sítím potřebuje pro evaluaci pouze jeden dopředný průchod vrstvami sítě.

Další výhodou je jednoduchost trénování a používání. V případě více osob pro samotnou klasifikaci pádu není nutné sledování osob. Můžeme jednoduše klasifikovat všechny detekované pózy, aniž bychom řešili, které osobě patří.

Tato síť ale ve výsledku bude klasifikovat pózy pouze dle vzájemného umístění jednotlivých klíčových bodů, potažmo délky končetin, nebude ale brát v úvahu natočení postavy. To proto, že postavy ve snímku vystupují pod různým úhlem natočení v závislosti na natočení kamery. Naopak síť, která je schopná sledovat pohyb, bude schopna sledovat mj. i změnu natočení postavy a to bez ohledu na natočení kamery.

Při použití této architektury jsou jako vstupní data použity klíčové body jedné osoby z daného jednoho snímku. Trénovací data pak pouze přečteme z trénovacího souboru a překonvertujeme je do formy tenzoru, konkrétně je zabalíme do instance třídy *DataLoader*. Při použití ve výsledném programu předáme modelu klíčové body každé detekované osoby v daném snímku.

Výstupem sítě je hodnota od 0 do 1, čísla od 0 do 0.5 jsou považována za třídu *normální*, zatímco čísla od 0.5 do 1 za třídu *upadl*.

Pro implementaci dopředné sítě v Pytorch Lightning jsme použili modul *nn.Sequential*, ve kterém definujeme postupné kroky průchodu sítě. *nn.Linear* definuje vrstvu sítě včetně počtu neuronů předcházející a aktuální vrstvy. Dále můžeme definovat aktivační funkci po dané vrstvě a regulařizační techniky jako je dropout či normalizace. Příklad implementace dopředné sítě je uveden v kódu 7.1. V tomto příkladě jsou definovány dvě vnitřní vrstvy o velikosti 128 a 32 a výstupní vrstva s jedním neuronem. Mezi vrstvami je aplikována normalizace a dropout 0.3, jako aktivační funkce je použita ReLU.

---

```
class KeypointClassifierFFNN(L.LightningModule):
    def __init__(self):
        super(KeypointClassifierFFNN, self).__init__()
        self.classifier = nn.Sequential(
            nn.Linear(34, 128),
            nn.BatchNorm1d(128),
            nn.ReLU(),
            nn.Dropout(0.3),
            nn.Linear(128, 32),
            nn.BatchNorm1d(32),
```

```

        nn.ReLU(),
        nn.Dropout(0.3),
        nn.Linear(32, 1),
    )

    self.criterion = nn.BCEWithLogitsLoss()
    self.accuracy = BinaryAccuracy(threshold=0.5)
    self.sigmoid = nn.Sigmoid()

def forward(self, x):
    x = self.classifier(x)
    return x

def configure_optimizers(self):
    optimizer = optim.Adam(self.parameters(), lr=1e-4)
    return optimizer

def training_step(self, batch, batch_idx):
    input, target = batch
    output = self(input)
    loss = self.criterion(output, target.float())
    self.log("train_loss", loss, on_epoch=True, on_step=False)
    return loss

def validation_step(self, batch, batch_idx):
    data, target = batch
    output = self(data)
    loss = self.criterion(output, target.float())
    self.log('val_loss', loss, on_epoch=True, on_step=False)
    output = self.sigmoid(output)
    accuracy = self.accuracy(output, target.int())
    self.log('val_accuracy', accuracy, on_epoch=True, on_step=False)
    return loss

```

---

Listing 7.1: Ukázka implementace Dopředné sítě v PyTorch Lightning

Abychom mohli tuto síť efektivně trénovat, musíme umožnit jednoduchou změnu konfigurace sítě. Třída *KeypointClassifierFFNN* proto v konstruktoru přijímá parametry *layers* – pole čísel reprezentujících velikost jednotlivých vrstev, *activation* – identifikátor vybrané aktivační funkce a *dropout* definující velikost dropoutu, viz kód 7.2. Jako ztrátovou funkci jsme použili křížovou entropii, tedy *nn.BCEWithLogitsLoss()*, která je vhodná pro binární klasifikaci.

---

```

class KeypointClassifierFFNN(L.LightningModule):
    def __init__(self, layers, activation, dropout=0.3, device=None):
        super(KeypointClassifierFFNN, self).__init__()

```

```

classifier = nn.Sequential()
for i in range(len(layers)-1):
    classifier.add_module(f'layer_{i}', nn.Linear(layers[i], layers[i+1]))
    classifier.add_module(f'batch_norm_{i}', nn.BatchNorm1d(layers[i+1]))
    classifier.add_module(f'activation_{i}', activations[activation])
    classifier.add_module(f'dropout_{i}', nn.Dropout(dropout))
classifier.add_module(f'layer_{len(layers)-1}', nn.Linear(layers[-1], 1))
self.classifier = classifier
self.criterion = nn.BCEWithLogitsLoss()
self.accuracy = BinaryAccuracy(threshold=0.5)
self.layers = layers
self.sigmoid = nn.Sigmoid()

```

---

Listing 7.2: Parametrizace konfigurace dopředné sítě

## 7.2.2 GRU síť

Jelikož pád je událost, nikoliv statická póza, mohlo by být optimálnější použít algoritmus, který bude analyzovat nejenom aktuální pózu, ale sekvenci posledních  $n$  pór. Pro tento účel se nabízí rekurentní neuronové sítě. V dnešní době je nejpoužívanější rekurentní architekturou GRU (Gated Recurrent Unit).

Příklad implementace GRU sítě je uveden v kódu 7.3, kde je definována GRU jednotka s jednou vrstvou, velikosti vstupního vektoru 34 a skrytým stavem velikosti 128, následována dvourstvou dopřednou plně propojenou sítí ze 128 neurony v první vrstvě a s jedním neuronem ve druhé vrstvě.

```

class KeypointClassifierGRU(lightning.LightningModule):
    def __init__(self):
        super(KeypointClassifierGRU, self).__init__()
        self.gru = nn.GRU(34, 128)           # GRU vrstvy
        self.classifier = nn.Sequential(   # Plně propojené vrstvy
            nn.Linear(128, 128),
            nn.ReLU(),
            nn.Dropout(0.3),
            nn.Linear(128, 1) )
        self.criterion = nn.BCEWithLogitsLoss() # Ztrátová funkce
        self.accuracy = BinaryAccuracy()       # Metrika přesnosti
    def forward(self, x):
        _, x = self.gru(x)
        x = x[-1]                         # Poslední skrytý stav

```

```
x = self.classifier(x)  
return x
```

---

Listing 7.3: Ukázka implementace GRU sítě v PyTorch Lightning

Část těchto hyperparametrů jsme opět parametrizovali, abychom mohli postupně otestovat různé jejich kombinace a vyhodnotit jejich vliv na výkon modelu, viz kód 8.2. Jedná se o velikost vstupního vektoru, velikost skrytého stavu sítě GRU, počet vrstev GRU, velikost první plně propojené vrstvy, velikost výstupní vrstvy (pro binární klasifikaci vždy 1, potřebné pro pozdější optimalizace) a dropout pro GRU i plně propojené vrstvy. Všechny tyto hyperparametry budeme ladit v další části.

---

```
class KeypointClassifierGRU(L.LightningModule):  
    def __init__(self, input_size=34, rnn_hidden_size=128, rnn_layers_count=2,  
                 fc_size=128, output_size=1, rnn_dropout=0.3, fc_dropout=0.3, device=None):
```

---

Listing 7.4: Parametry konstruktoru třídy *KeypointClassifierGRU* definující hyperparametry sítě

Stejně jako u předchozí architektury, jako ztrátová funkce byla použita binární křížová entropie, tedy *nn.BCEWithLogitsLoss()*, která je vhodná pro binární klasifikaci.

### 7.2.3 LSTM síť

Další populární rekurentní architekturou je LSTM (Long Short-Term Memory). Ta je lepší pro delší sekvence, je to ale za cenu větší komplexity. Jak již bylo vysvětleno, s nárůstem komplexity se zvyšuje riziko přetrénování, zejména při nedostatku trénovacích dat. Vzhledem k množství trénovacích dat, které máme k dispozici, se tak dá předpokládat, že trénování této sítě bude spíše méně stabilní než v případě sítě GRU.

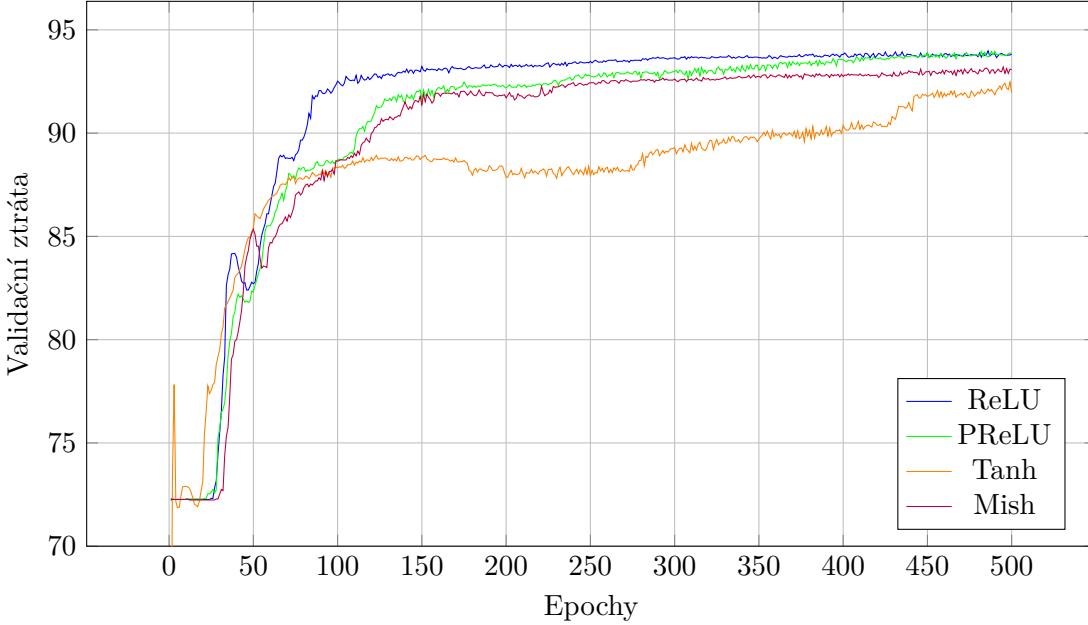
Implementace sítě LSTM se od GRU liší pouze použitím modulu *nn.LSTM* místo *nn.GRU*. Použití těchto modulů je v PyTorch velice podobné, pouze LSTM vrací kromě skrytého stavu také stav buňky (dlouhodobou paměť), ten ale stejně nevyužijeme. Pro síť LSTM budeme ladit stejnou množinu hyperparametrů, jako pro GRU síť, abychom mohli porovnat jejich výkonnost.

## 7.3 Návrh architektury a konfigurace sítě

Nyní přistoupíme do samotného návrhu architektury a konfigurace sítě. Vybrané architektury jsme natrénovali v různých konfiguracích a celý postup trénování jsme zapisovali pomocí logovacího nástroje TensorBoard. Nyní na základě grafů základních metrik, jako jsou ztrátová funkce či přesnost, zhodnotíme výkon vybraných architektur, podíváme se, jaký mají jednotlivé hyperparametry vliv na proces trénování a výsledný výkon. Nakonec vybereme nejoptimálnější variantu.

K hyperparametru, které jsme zkoušeli, patří počet vrstev a jejich velikost a velikost dropoutu. U dopředných sítí jsme navíc zkoušeli i různé aktivační funkce, u rekurentních sítí pak velikost

Obrázek 7.1: Graf přesnosti na validačních datech v průběhu trénování dopředných sítí s různými aktivačními funkcemi



skrytého stavu. Pro optimalizaci váh během trénování byl použit optimalizační algoritmus Adam [12], jehož výhodou je

### 7.3.1 Dopředná neuronová síť

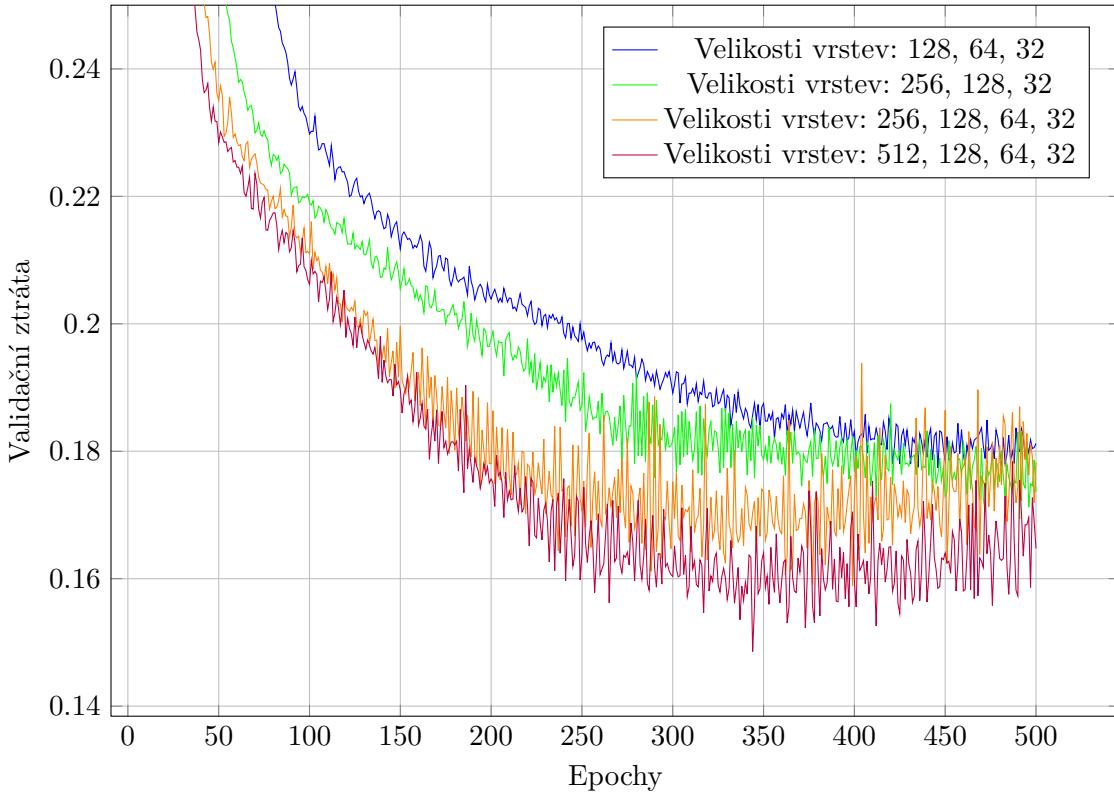
Nejprve jsme zkoušeli natrénovat dopřednou neuronovou síť. Vyzkoušeli jsme od dvou do čtyř vnitřních vrstev, ve velikostech od 32 do 512 neuronů, vždy mocniny dvou. Pojdeme postupně rozebrat jednotlivé hyperparametry a jejich vliv na výkon modelu.

U dopředných sítí se nám nejlépe ověřila velikost dávky 4096. Síť se lépe trénovala s použitím normalizace (modul *nn.BatchNorm1d*) a dropoutu. Pro většinu konfigurací byla velikost dropoutu 0.4 nejoptimálnější.

Síť jsme zkoušeli trénovat s těmito vybranými aktivačními funkcemi: *ReLU*, *Tanh*, *PReLU* a *Mish*. *ReLU* a *PReLU* často dosahovaly podobných výsledků, většinou ale nejlepších výsledků dosahovala *ReLU*. Příkladem je graf přesnosti pro trénování sítě se třemi vnitřními vrstvami velikosti 128, 64, a 32 na obrázku 7.1.

Obecně větší síť dosahovaly lepších výsledků – menší ztrátové funkce a vyšší přesnosti na validačních datech, v jistém momentě jsme již ale začali narážet na problém se stabilitou trénování. Nejoptimálnějších výsledků dosáhla síť se třemi vrstvami o velikostech 128, 64 a 32. Jak můžeme vidět na grafu 7.2, síť s větší šírkou nebo hloubkou sice dosahují lepších výsledků, začínají ale brzy výrazně kolísat. Zkoušeli jsme v těchto případech i L2 regularizaci (parametr *weight\_decay*), to ale neřešilo problém. Pro komplexnější síť bychom pravděpodobně potřebovali větší množství dat.

Obrázek 7.2: Graf validační ztráty v průběhu trénování hlubších dopředných sítí



Pojďme si tedy shrnout, jaká je pro nás nejoptimálnější konfigurace: použili jsme tři vrstvy o velikostech 128, 64 a 32, dropout 0.3, normalizaci mezi vrstvami, velikost dávky 4096 a aktivační funkci *ReLU*. Dosáhli jsme tak validační ztráty 0.18 a validační přesnosti 93.8

### 7.3.2 GRU síť

Jak již bylo zmíněno, v oblasti jednodušších RNN jsou dnešním standardem GRU sítě, sítě LSTM se používají zejména, pokud si GRU s problémem neradí, anebo je vzhledem k problému důležité uchování dlouhodobých závislostí.

Stejnou taktiku zvolíme i my. Nejdříve otestujeme GRU sítě, a to pro několik možností délky analyzované sekvence. To znamená, že pro každý snímek předáme síti očekávanou třídu a sekvenci pón dané osoby z  $n$  posledních snímků. Pak vyzkoušíme síti předávat celou sekvenci pón dané osoby.

Naše GRU síť se skládá z několika vrstev GRU, konkrétně jsme zkoušeli trénovat 1 až 3 vrstvy, které následuje jedna plně propojená vrstva a výstupní vrstva. Velikosti GRU vrstev jsme volili v rozsahu od 64 do 256, u plně propojené vrstvy jsme volili mezi 32 a 128 neurony. Většinou jsme používali dávky o velikosti 4096, menší dávky velice destabilizovaly trénování.

Zjistili jsme, že na rozdíl od dopředných sítí, v případě GRU nejlepších výsledků dosahovaly jednodušší sítě. Většina sítí s jednou GRU vrstvou tak dosahovala poměrně stabilních výsledků.

Nejvyšší přesnosti jsme pak dosáhli sítí s jednou GRU vrstvou o velikosti 64 a plně propojenou vrstvou o velikosti 64, s dropoutem 0.4 v plně propojené vrstvě. Co se týče dropout mezi GRU vrstvami, bylo u hlubších sítí většinou dosaženo stabilnějších výsledků s větší hodnotou dropoutu jako je 0.4, u jednovrstvých GRU sítí je dropout nerelevantní, jelikož se aplikuje mezi vrstvami.

Sítě s více GRU vrstvami většinou než dosáhly optimálního výkonu, začaly kolísat a ve validační ztrátě a přesnosti se objevovaly obrovské výkyvy. Zkoušeli jsme větší dávky – 8162, které sice stabilizovaly trénování při stejném počtu epoch, jelikož se ale větší dávka loučí s pomalejším učením, nedosahovaly takové přesnosti jako při dávkách velikosti 4096. Pro dosažení podobné přesnosti jsme tedy zkoušeli trénovat ve více epochách, tehdy ale brzy docházelo k přetrénování.

RNN sítě obecně můžou přijímat sekvence libovolné délky, je ale efektivnější, pokud je síť trénovaná na jednotné délce, tato délka sekvence je pak použitá i v případě inferencí. Zkoušeli jsme tedy trénovat sítě na sekvencích o délce 50 a 100 snímků, vyzkoušeli jsme i neomezenou délku sekvence, tedy modelu předáváme klíčové body z celé sekvence snímků, na kterých byla daná osoba detekována. Obecně jsme ale zjistili, že vždycky dosahují stabilnějších výsledků sítě spíše s kratšími sekvencemi. Zůstali jsme tedy u sekvencí délky 50.

Naší nejlepší konfigurací GRU sítě je tedy jedna GRU vrstva o velikosti 64, následována plně propojenou vrstvou s 64 neurony a dropoutem 0.4. Trénována byla na sekvencích o délce 50 snímků, s velikostí dávky 4096 a s použitím optimalizace Adam. Na obrázku 7.3 pak můžeme vidět, že jsme dosáhli trénovací ztráty 0.066, validační ztráty 0.211 a přesnosti 93.12% .

### 7.3.3 LSTM sítě

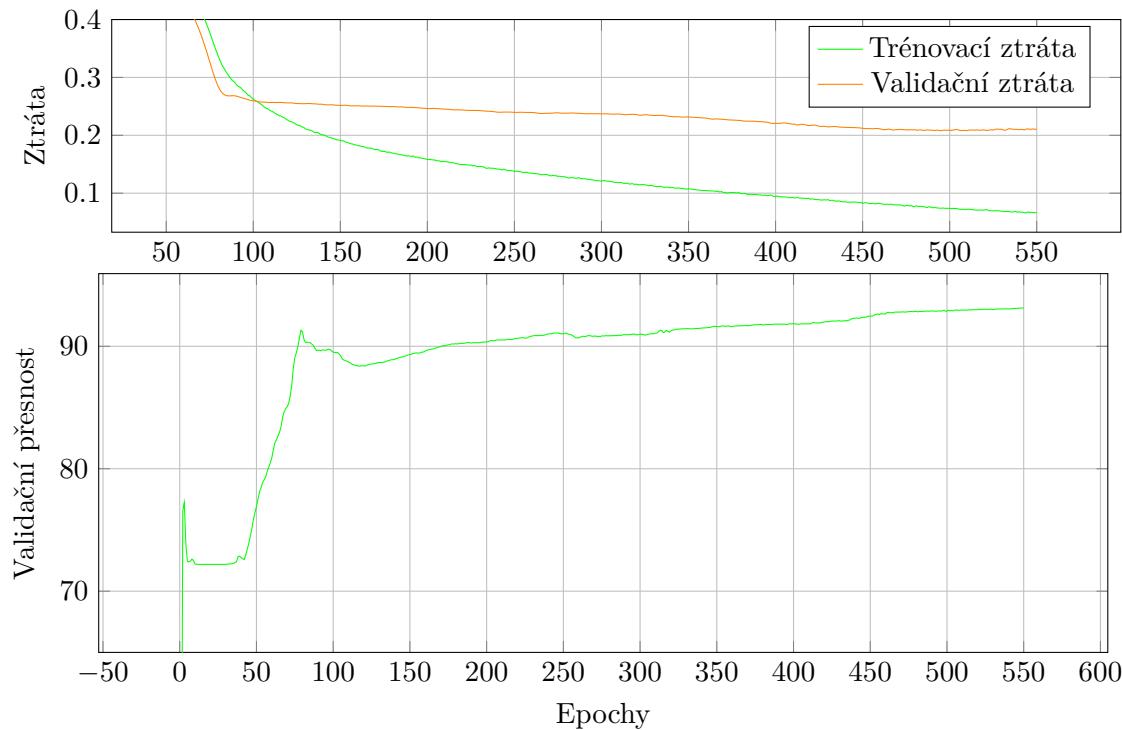
LSTM sítě jsme trénovali se stejnými hyperparametry a konfiguracemi jako sítě GRU. Obecně se také dá říct, že měly tyto parametry u obou architektur velice podobný vliv na výsledky, a tedy to, co jsme napsali o GRU sítích, by se většinou dalo napsat i o LSTM sítích. Také měly nejlepší výsledky s jednou vrstvou, podobně se projevovala délka sekvence, a obdobně byl většinou optimální dropout velikosti 0.4.

Obecně ale LSTM sítě byly mnohem méně stabilní, a prakticky v žádné konfiguraci se nepodařilo natrénovat síť do požadované přesnosti, aniž by došlo k přetrénování anebo by začala validační ztráta extrémně kolísat. Pravděpodobně tedy nemáme dostatečný vzorek trénovacích dat pro trénování LSTM sítě. Také, sítě LSTM nejsou vhodny pro každou úlohu, spíše vynikají pro delší sekvence.

## 7.4 Testování

Vybrané modely jsme nakonec otestovali na testovacím datasetu. Ověříme jejich přesnost a dobu inference. Konkrétně jsme testovali dopřednou neuronovou síť a GRU sít, které v další kapitole, až bude naimplementovaný celý detektor pádu, otestujeme na reálných datech.

Obrázek 7.3: Graf trénovací a validační ztráty (nahoře) a graf validační přesnosti v průběhu trénování výsledné sítě



Tabulka 7.1: Výsledky testování dopředné a GRU sítě

	Ztráta	Přesnost	Průměrná doba inference [ms]	
			CPU	GPU
Dopředná síť	0.148	94.43	0.207	0.647
GRU síť	0.138	96.88	0.258	0.613

V tabulce 7.1 vidíme, že obě sítě dosahují velmi dobrého výkonu. Dopředná síť dosáhla přesnosti 94.43% a ztráty 0.148, zatímco GRU síť dosáhla jak lepší přesnosti 96.88%, tak trošku menší ztráty 0.138. Obě sítě by tak mohly být použity pro detekci pádu, jelikož je ale pro klasifikační úlohy v konečném řešení důležitější přesnost, GRU síť vypadá jako lepší kandidát.

Z pohledu rychlosti jsou obě sítě natolik jednoduché, že ani nelze zaznamenat výrazný rozdíl. Zároveň je velmi zajímavé, že je model na CPU asi třikrát rychlejší než na GPU. Je to způsobeno tím, že je model malý, nelze tedy plně využít potenciál paralelismu, navíc zpracováváme výstupy po jednom. Režie GPU tedy převáží nad jejími výhodami. Ve všech případech se ale u doby inference jedná o zlomky milisekund, ve výsledném programu doba prakticky zanedbatelná.

V další kapitole budeme implementovat detektor pádu, tedy program, který spojí obě fáze detekce. Pomocí něj pak vyzkoušíme tyto dva modely, zejména s ohledem na robustnost,

## Kapitola 8

# Kompletace finálního řešení detekce pádu

V této kapitole se zaměříme na implementaci celkového řešení detekce pádu, který na základě postupně předávané sekvence snímků bude detektovat, zda došlo k pádu. Jedná se tedy o spojení detekčního algoritmu, který ze snímku extrahuje klíčové body všech osob, a klasifikačního algoritmu, který na základě těchto bodů detekuje pád.

V našem případě tedy musíme spojít predikci modelu YOLO11-pose, a jednoho z našich vybraných modelů (dopředná nebo GRU síť) pro klasifikaci pádu. Podíváme se tedy nejprve krátce na práci ve frameworku YOLO. Následně se budeme zabývat použitím dříve natrénovaných modelů. Zejména se zaměříme na implementaci s využitím rekurentní architektury GRU, která přináší více výzev než dopředná síť, s ohledem na nutnost uchovávání sekvence pro každého člověka zvlášť. Nakonec program otestujeme s oběma klasifikačními modely na několika videích a zhodnotíme jejich výkon.

### 8.1 Sledování pózy s YOLO11

V YOLO verze 11 máme kromě již zmíněné detekce objektů i klíčových bodů k dispozici také další volitelné funkce. To, které funkce chceme využít, definujeme vybraným modelem. V závislosti na něm pak při inferenci model vrací patřičné hodnoty. Dostupné jsou tyto modely:

- *YOLO11<v>-seg* – detekce objektů – bounding boxů
- *YOLO11<v>-cls* – detekce objektů, klíčových bodů a segmentace
- *YOLO11<v>-pose* – detekce klíčových bodů
- *YOLO11<v>-obb* – orientovaná detekce objektů – bounding boxy natočené dle natočení objektů

kde  $< v >$  označuje velikost modelu – můžeme vybrat menší modely pro větší výkon ale horší přesnost, anebo větší modely, které jsou sice velmi přesné, musíme ale počítat s vysokými nároky

na výkon. V našem případě jsme zvolili model *YOLO11m – pose*. Pro inicializaci modelu musíme specifikovat cestu k němu, pokud není nalezen, knihovna automaticky stáhne předtrénovaný model.

```
from ultralytics import YOLO
pose_model = YOLO("./models_pose/yolo11m-pose.pt")
```

Listing 8.1: Inicializace modelu *YOLO11m – pose*

Dále můžeme každý z těchto modelů používat v několika režimech. Režim specifikujeme tím, jakou metodu modelu použijeme. Pokud chceme model natrénovat na vlastních datech, použijeme režimy *trénování* (metoda *train()*) a *validace* (metoda *val()*). Pokud chceme používat již natrénovaný model, máme k dispozici dva režimy: *predikce* (přímé použití modelu, např. *pose\_model()*) pro vyhodnocení každého vstupního snímku zvlášť a *sledování* (metoda *track()*).

Režim *sledování* vrací stejné informace jako režim *predikce*, navíc ale sleduje pohyb objektů mezi jednotlivými snímky a přiřazuje jim *id*. V našem případě tedy kromě bounding boxu a klíčových bodů dostaneme i *id* každé osoby.

YOLO11 dokáže zpracovat širokou škálu vstupních dat, včetně videí, kdy nám je vráceno pole výsledků, datových proudů (je třeba specifikovat příznak *stream*), kdy nám je vrácen iterátor, nebo obrázků, kdy dostaneme výsledek pro daný snímek. Pokud chceme použít režim *sledování* na jednotlivé snímky, musíme specifikovat příznak *persist*, který zajistí, že mezi jednotlivými snímky bude uchovávána informace o sledovaných objektech. Zde můžeme vidět použití metody *track()* v našem programu (příznaky *show* a *verbose* specifikují, zda se má zobrazit výsledek každého snímku a zda má model do konzoly vypisovat ladící informace)

```
results = self._pose_model.track(
    frame, show=False, verbose=False, persist=True)
```

Listing 8.2: Použití sledování pomocí YOLO11

## 8.2 Struktura třídy FallDetector

Třída *FallDetector* slouží pro detekci pádu v sekvenci postupně předávaných snímků. Můžeme ji tak použít jak pro analýzu uloženého videa, tak pro detekci pádu v živém přenosu v reálném čase. Mezi jednotlivými snímky uchovává pro každou osobu klíčové body z *x* posledních snímků pro predikci na základě sekvence snímků. Dále dle nastavení uchovává *n* posledních snímků, popřípadě i anotovaných, pro uložení videa včetně několika sekund před a po pádu. Pro ukládání těchto informací používáme kolekci typu *deque*, která umožňuje specifikovat její maximální délku, nerelevantní informace ze starších snímků tedy budou automaticky odstraňovány.

V konstruktoru obdrží tato třída cestu k modelu YOLO (musí být inicializován pro každou instanci zvlášť), model pro detekci pádu na základě klíčových bodů, délku sekvence předávané

tomuto modelu, a informace týkající se ukládání videí s pády, jako jsou cesty k souborům a délka videa před a po pádu.

Pro práci s jednotlivými osobami používáme třídu *Person*, která uchovává hlavně klíčové body z posledních snímků a *id*, dále také bounding boxy, detekovaný stav a confidence score této detekce pro poslední snímek, kdy byla osoba detekována. V případě, že daná osoba již není detekována, obsahuje informaci, kolik snímků již není vidět pro vymazání osob, jež odešly ze scény. Okamžité vymazání zmizelé osoby není žádoucí, jelikož někdy osoba je pořád na scéně, ale algoritmus ji i na několik snímků ztratí, zejména v případě okluze či špatných světelních podmínek.

V metodě *process\_frame* třídy *FallDetector* je zpracován vždy jeden snímek. Nejprve se detekují všechny osoby a jejich klíčové body pomocí vybraného modelu. Pokud je požadováno uložení anotovaného videa, převezmeme zde anotovaný snímek přímo od modelu YOLO, později pouze dopříjemme třídu detekce pádu. Dále jsou na základě detekovaných informací vytvořeny, popřípadě aktualizovány instance třídy *Person*. Pro každou osobu je také provedena analýza sekvence klíčových bodů pro detekci pádu a popřípadě je patřičně anotována ve snímku. Pokud je detekován pád a je nastaveno ukládání videa, vytvoří se nový soubor a zapíše se do něj uchované předešlé snímkы. Do souboru se pak zapisuje, dokud je detekován pád a nevyprší požadovaný počet snímků po pádu.

### 8.3 Testování detekce pádu ve videu

V této sekci otestujeme výsledné řešení na videích z testovacích sad obou datasetů. Hlavně ověříme přesnost detekce pádu v různých situacích ve videích. Přesnost klasifikačního algoritmu, kterou jsme měřili v minulé kapitole, je spíše na úrovni detekce v jednotlivých snímcích. Pokud tedy program detekuje pád o několik snímků později, než byl pád anotován, je tato přesnost penalizována. Nás ale spíše zajímá přesnost ve smyslu, zda byl každý pád detekován, a zda se neobjevovaly falešné detekce pádů.

V testovací sadě datasetu CAUCAFall je 15 videí, z čehož 7 obsahuje pád a 8 obsahuje jiné činnosti, které by jako pád neměly být využívány, kdežto ve všech 15 případech testovací sady z videa *50 ways to fall* vystupuje pád.

Zjistili jsme, že obecně jsou výsledky sítě GRU mnohem stabilnější než v případě dopředné sítě. Pokud tedy osoba spadne a leží na zemi, GRU síť pořád klasifikuje pád, zatímco dopředná síť často, zvláště pokud osoba leží vyrovnaná, kolísá mezi třídou *normální* a *upadl*. Stabilita dopředné sítě se také zhoršuje v případě špatné viditelnosti osoby, kdy jsou výsledky modelu pro detekci pózy více chaotické než v optimálních podmínkách.

Dopředná síť také reaguje na pád mnohem rychleji, většinou už v době padání. Oproti tomu GRU síť detekuje pád až v momentě dopadu nebo o několik snímků později. Dopředná síť ale dokáže rychleji reagovat i na situaci kdy se osoba ze stavu *upadl* dostane zpátky do stavu *normální*. V jednom z testovacích videí osoba dělá kotrmelec, po kterém se vrátila do dřepu. Dopředná síť

tehdy okamžitě vyhodnotila její stav zpátky jako normální, kdežto GRU síť ještě nějakou dobu identifikovala pád.

Hlavní problém s dopřednou sítí je ale v její falešné pozitivitě. GRU síť všechny testovací videa vyhodnotila správně, a tedy detekovala pád pouze a právě ve videích, kde osoba upadla. Dopředná síť detekovala sice pád ve všech videích, kde osoba upadla. Z osmi videí, na kterých pád zaznamenán není, ale tato síť ve dvou případech pád detekovala. Konkrétně se tak stalo v situacích, kdy osoba klečela a kdy se ohnula k zemi. Jelikož v těchto případech detekovala pád pouze v několika snímcích, dala by se tato falešná pozitivita odfiltrovat tím, že bychom reagovali na pád až po několika snímcích, kdy byl soustavně detekován.

Pojďme se nyní ještě podívat na výslednou rychlosť detektoru. V tabulce 8.1 vidíme, že se nám ve všech případech podařilo dosáhnout stanovené minimální frekvence, viz sekce 6.5.1.

Tabulka 8.1: Rychlosť detekce pádu

	50 ways to fall		CAUCAFall	
	ms	FPS	ms	FPS
Dopředná síť	25.6	39	24.2	41
GRU síť	27	37	25.9	39

Rozdíl rychlosti mezi datasety je způsoben hlavně velikostí snímku. V datasetu 50 ways to fall, je totiž snímek téměř třikrát větší než v případě CAUCAFall. Pro snímky s větším rozlišením musí YOLO nejprve věnovat více času na zredukování velikosti snímku. V Předpokládáme ale, že většinou je rozlišení u průmyslových kamer spíše menší než HD, měli bychom tedy dosáhnout spíše výsledků jako v případě datasetu CAUCAFall. Mezi oběma modely je taky rozdíl v rychlosti, je způsoben hlavně reží sekvencí v případě sítě GRU.

Výše uvedená měření byla provedena bez anotace videí. Pokud jsme anotovali videa pouze informací o pádu, doba zpracování se většinou zvýšila pouze o zhruba 2 – 3 milisekundy. Pokud jsme ale použili anotaci klíčových bodů a bounding boxů integrovanou v modelu YOLO, došlo k zpomalení až o 15 milisekund.

Obě sítě tedy fungují poměrně dobře a dokážou v reálném čase detektovat pád v různých situacích, z výsledků je ale zjevné, že je rekurentní síť GRU mnohem stabilnější a přesnější.

# Kapitola 9

## Závěr

Cílem práce bylo navrhnout a implementovat řešení pro detekci pádu osoby v toku obrázku v reálném čase. V první části byly popsány teoretické základy použitých technologií, se zaměřením na neuronové sítě.

V praktické části byl popsán postup návrhu a vývoje výsledného programu. Řešení jsme dosáhli spojením volně dostupného modelu pro detekci pózy ve formě klíčových bodů a neuronové sítě pro klasifikaci těchto bodů do třídy *normální* nebo *upadl*.

Prozkoumali jsme několik detekčních algoritmů – vysvětlili jsme si zásady funkčnosti detekce klíčových bodů a otestovali jsme jednotlivé algoritmy na testovacích datech. Pro detekci pózy jsme nakonec zvolili model YOLO11-pose pro jeho vysokou rychlosť, zejména při využití grafické akcelerace, a přesnost i při horší viditelnosti. Tento model také v sobě integruje funkci sledování osob, která nám umožňuje provádět následnou analýzu pro každou osobu zvláště i v kontextu více snímků.

Pro klasifikaci pózy jsme zkoušeli natrénovat modely postavené na třech architekturách: dopředná neuronová síť a dvě rekurentní architektury – LSTM a GRU. LSTM síť se nám nepodařilo efektivně natrénovat, pravděpodobně pro nedostatek dat a její nevhodnost pro naš problém.

Dopřednou síť a GRU síť jsme po nalezení optimální konfigurace a natrénování otestovali na testovacích datech. Následně jsme je použili v implementaci detektoru pádu. Ten jsme dále v obou verzích otestovali na videích z testovací sady. Dopředná síť se ukázala jako znatelně méně přesná, zejména při méně kvalitní detekci pózy v situacích s horší viditelností. Také projevovala falešné pozitivní detekce pádu.

Pro klasifikaci pózy jsme tedy zvolili rekurentní architekturu GRU s jednou rekurentní a jednou plně propojenou vrstvou, obě velikosti 64. Tuto síť se nám podařilo natrénovat na přesnost přes 96% na testovacích datech, zároveň se ukázala jako velice stabilní a přesná i při detekci pádu v testovacích videích.

Pokud bychom chtěli pokračovat v optimalizaci našeho řešení, nejvíce prostoru vidíme v množství trénovacích dat. Zejména by bylo vhodné použít videa s více úhly natočení vůči zemi a s různými vzdálenostmi osoby od kamery. Větší a rozmanitější dataset by nám umožnil natrénovat i

komplexnější architektury, které by byly přesnější a robustnější. Pro získání více dat by se mohlo nasadit náš detektor na podnikové kamery, nebo by se pomocí něj dalo zanalyzovat starší záznamy. V těchto případech by bylo vhodné nastavit nižší prah detekce, abychom získali i data s vyšší pravděpodobností falešné pozitivity.

Dalším vylepšením by mohlo být zavedení více tříd do analýzy pózy. V textu byla zmíněna třída *padá*, nicméně bychom mohli také přidat třídy např. pro pozici v sedě, ve dřepu, či na kolenou. Model by tak mohl lépe rozlišovat některé situace, které se zdají být pádu podobné, a tak bychom odstranili potenciální falešné pozitivní detekce pádu.

Zajímavým rozšířením by také mohlo být propojení více kamer z jedné místo pro sledování osob z více úhlů. Dávalo by to možnost zasazení postav do trojrozměrného prostoru a umožňovalo dosáhnout mnohem kvalitnější detekce pózy. Nicméně by bylo obtížnější získat pro takové řešení trénovací data.

Pro nasazení detektoru pádu v reálném provozu, tedy na obrazovém toku z podnikových bezpečnostních kamer, bude ještě třeba vytvořit vhodné rozhraní pro správné čtení různých protokolů a formátů videa. Pro integraci do komplexnějšího programu, ve kterém by probíhalo více analýz videa, by bylo vhodné zvážit jednak jejich paralelní zpracování, jednak sdílení určitých dat mezi jednotlivými analyzátoři.

# Literatura

1. BISCHOFF, Paul. *Surveillance camera statistics: which are the most surveilled cities?* [online]. Comparitech, 2021. [cit. 2025-04-29]. Dostupné z: <https://www.comparitech.com/vpn-privacy/the-worlds-most-surveilled-cities/>.
2. HILLSTROM, Anne; HOPE, Lorraine; NEE, Claire. *Applying psychological science to the CCTV review process: a review of cognitive and ergonomic literature*. London: The Stationery Office, 2008-03. Working Paper. Dostupné také z: <https://eprints.soton.ac.uk/371614/>.
3. MCCULLOCH, Warren S.; PITTS, Walter. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*. 1943-12, roč. 5, č. 4, s. 115–133. ISSN 1522-9602. Dostupné z DOI: 10.1007/BF02478259.
4. ROSENBLATT, Frank. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*. 1958, roč. 65 6, s. 386–408. Dostupné také z: <https://api.semanticscholar.org/CorpusID:12781225>.
5. MACUKOW, Bohdan. Neural Networks – State of Art, Brief History, Basic Models and Architecture. In: SAEED, Khalid; HOMENDA, Władysław (ed.). *Computer Information Systems and Industrial Management*. Cham: Springer International Publishing, 2016, s. 3–14. ISBN 978-3-319-45378-1.
6. RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning internal representations by error propagation. In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA, USA: MIT Press, 1986, s. 318–362. ISBN 026268053X.
7. LECUN, Y.; BOSER, B.; DENKER, J. S.; HENDERSON, D.; HOWARD, R. E.; HUBBARD, W.; JACKEL, L. D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*. 1989, roč. 1, č. 4, s. 541–551. Dostupné z DOI: 10.1162/neco.1989.1.4.541.
8. VONDRAK, Ivo. *Umělá inteligence a neuronové sítě*. 1. vyd. Ostrava: Vysoká škola báňská, 1994. ISBN 80-7078-259-5.

9. RASAMOELINA, Andrinandrasana David; ADJAILIA, Fouzia; SINČÁK, Peter. A review of activation function for artificial neural network. In: *2020 IEEE 18th world symposium on applied machine intelligence and informatics (SAMI)*. IEEE, 2020, s. 281–286.
10. LAGAN, Jiří. *Modul LSTM a Rekurentních neuronových sítí pro program Modeler neuronových sítí* [online]. Ostrava: Vysoká škola báňská – Technická univerzita Ostrava, 2021 [cit. 2025-02-24]. Dostupné z: <http://hdl.handle.net/10084/143977>.
11. TIELEMAN, Tijmen; HINTON, Geoffrey. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*. 2012, roč. 4, č. 2, s. 26–31.
12. KINGMA, Diederik P.; BA, Jimmy. *Adam: A Method for Stochastic Optimization*. 2017. Dostupné z arXiv: 1412.6980 [cs.LG].
13. SUN, Xiang. *The Lasso and its implementation for neural networks*. 1999. Dis. pr.
14. LOSHCHILOV, Ilya; HUTTER, Frank. *Decoupled Weight Decay Regularization*. 2019. Dostupné z arXiv: 1711.05101 [cs.LG].
15. O'SHEA, Keiron; NASH, Ryan. *An Introduction to Convolutional Neural Networks*. 2015. Dostupné z arXiv: 1511.08458 [cs.NE].
16. RAWAT, Waseem; WANG, Zenghui. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*. 2017, roč. 29, č. 9, s. 2352–2449.
17. DHRUV, Patel; NASKAR, Subham. Image classification using convolutional neural network (CNN) and recurrent neural network (RNN): A review. *Machine learning and information processing: proceedings of ICMLIP 2019*. 2020, s. 367–381.
18. ELMAN, Jeffrey L. Finding Structure in Time. *Cognitive Science*. 1990, roč. 14, č. 2, s. 179–211. Dostupné z DOI: [https://doi.org/10.1207/s15516709cog1402\\\_1](https://doi.org/10.1207/s15516709cog1402\_1).
19. JORDAN, Michael I. Chapter 25 - Serial Order: A Parallel Distributed Processing Approach. In: DONAHOE, John W.; PACKARD DORSEL, Vivian (ed.). *Neural-Network Models of Cognition*. North-Holland, 1997, sv. 121, s. 471–495. Advances in Psychology. ISSN 0166-4115. Dostupné z DOI: [https://doi.org/10.1016/S0166-4115\(97\)80111-2](https://doi.org/10.1016/S0166-4115(97)80111-2).
20. PASCANU, Razvan; GULCEHRE, Caglar; CHO, Kyunghyun; BENGIO, Yoshua. *How to Construct Deep Recurrent Neural Networks*. 2014. Dostupné z arXiv: 1312.6026 [cs.NE].
21. HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long Short-Term Memory. *Neural Computation*. 1997-11, roč. 9, s. 1735–1780. Dostupné z DOI: 10.1162/neco.1997.9.8.1735.
22. CHO, Kyunghyun; MERRIENBOER, Bart van; GULCEHRE, Caglar; BAHDANAU, Dzmitry; BOUGARES, Fethi; SCHWENK, Holger; BENGIO, Yoshua. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. 2014. Dostupné z arXiv: 1406.1078 [cs.CL].

23. ERASO, Jose Camilo; MUÑOZ, Elena; MUÑOZ, Mariela; PINTO, Jesus. *Dataset CAUCAFall*. Mendeley Data, 2022. Ver. 4. Dostupné z DOI: [10.17632/7w7fccy7ky.4](https://doi.org/10.17632/7w7fccy7ky.4). Accessed: 2025-04-24.
24. LIN, Tsung-Yi; MAIRE, Michael; BELONGIE, Serge; BOURDEV, Lubomir; GIRSHICK, Ross; HAYS, James; PERONA, Pietro; RAMANAN, Deva; ZITNICK, C. Lawrence; DOLLÁR, Piotr. *Microsoft COCO: Common Objects in Context*. 2015. Dostupné z arXiv: [1405.0312 \[cs.CV\]](https://arxiv.org/abs/1405.0312).
25. JI, Zhangjian; WANG, Zilong; ZHANG, Ming; CHEN, Yapeng; QIAN, Yuhua. *2D Human Pose Estimation with Explicit Anatomical Keypoints Structure Constraints*. 2022. Dostupné z arXiv: [2212.02163 \[cs.CV\]](https://arxiv.org/abs/2212.02163).
26. BAZAREVSKY, Valentin; GRISHCHENKO, Ivan; RAVEENDRAN, Karthik; ZHU, Tyler; ZHANG, Fan; GRUNDMANN, Matthias. *BlazePose: On-device Real-time Body Pose tracking*. 2020. Dostupné z arXiv: [2006.10204 \[cs.CV\]](https://arxiv.org/abs/2006.10204).
27. TOSHEV, Alexander; SZEGEDY, Christian. DeepPose: Human Pose Estimation via Deep Neural Networks. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014-06, s. 1653–1660. Dostupné z DOI: [10.1109/cvpr.2014.214](https://doi.org/10.1109/cvpr.2014.214).
28. MAJI, Debapriya; NAGORI, Soyeb; MATHEW, Manu; PODDAR, Deepak. *YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss*. 2022. Dostupné z arXiv: [2204.06806 \[cs.CV\]](https://arxiv.org/abs/2204.06806).
29. ERHAN, Dumitru; SZEGEDY, Christian; TOSHEV, Alexander; ANGUELOV, Dragomir. Scalable Object Detection using Deep Neural Networks. *CoRR*. 2013, roč. abs/1312.2249. Dostupné z arXiv: [1312.2249](https://arxiv.org/abs/1312.2249).
30. GIRSHICK, Ross; DONAHUE, Jeff; DARRELL, Trevor; MALIK, Jitendra. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2016, roč. 38, č. 1, s. 142–158. Dostupné z DOI: [10.1109/TPAMI.2015.2437384](https://doi.org/10.1109/TPAMI.2015.2437384).
31. GIRSHICK, Ross B.; DONAHUE, Jeff; DARRELL, Trevor; MALIK, Jitendra. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*. 2013, roč. abs/1311.2524. Dostupné z arXiv: [1311.2524](https://arxiv.org/abs/1311.2524).
32. REN, Shaoqing; HE, Kaiming; GIRSHICK, Ross B.; SUN, Jian. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *CoRR*. 2015, roč. abs/1506.01497. Dostupné z arXiv: [1506.01497](https://arxiv.org/abs/1506.01497).
33. REDMON, Joseph; DIVVALA, Santosh Kumar; GIRSHICK, Ross B.; FARHADI, Ali. You Only Look Once: Unified, Real-Time Object Detection. *CoRR*. 2015, roč. abs/1506.02640. Dostupné z arXiv: [1506.02640](https://arxiv.org/abs/1506.02640).

34. REDMON, Joseph; FARHADI, Ali. YOLO9000: Better, Faster, Stronger. *CoRR*. 2016, roč. abs/1612.08242. Dostupné z arXiv: [1612.08242](https://arxiv.org/abs/1612.08242).
35. REDMON, Joseph; FARHADI, Ali. YOLOv3: An Incremental Improvement. *CoRR*. 2018, roč. abs/1804.02767. Dostupné z arXiv: [1804.02767](https://arxiv.org/abs/1804.02767).
36. LIU, Wei; ANGUELOV, Dragomir; ERHAN, Dumitru; SZEGEDY, Christian; REED, Scott E.; FU, Cheng-Yang; BERG, Alexander C. SSD: Single Shot MultiBox Detector. *CoRR*. 2015, roč. abs/1512.02325. Dostupné z arXiv: [1512.02325](https://arxiv.org/abs/1512.02325).
37. BOESCH, Gaudenz. *The Complete Guide to OpenPose in 2025* [online]. Viso.ai, 2024-10-15. [cit. 2025-04-29]. Dostupné z: <https://viso.ai/deep-learning/openpose/>.
38. CAO, Zhe; HIDALGO, Gines; SIMON, Tomas; WEI, Shih-En; SHEIKH, Yaser. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *CoRR*. 2018, roč. abs/1812.08008. Dostupné z arXiv: [1812.08008](https://arxiv.org/abs/1812.08008).
39. KREISS, Sven; BERTONI, Lorenzo; ALAHI, Alexandre. *OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association*. 2021. Dostupné z arXiv: [2103.02440](https://arxiv.org/abs/2103.02440) [cs.CV].
40. WANG, Chien-Yao; BOCHKOVSKIY, Alexey; LIAO, Hong-Yuan Mark. *YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors*. 2022. Dostupné z arXiv: [2207.02696](https://arxiv.org/abs/2207.02696) [cs.CV].
41. KHANAM, Rahima; HUSSAIN, Muhammad. *YOLOv11: An Overview of the Key Architectural Enhancements*. 2024. Dostupné z arXiv: [2410.17725](https://arxiv.org/abs/2410.17725) [cs.CV].
42. VERGARA, Marcus Loo. *Keypoint RCNN*. 2017. Dostupné také z: [https://github.com/bitsauce/Keypoint\\_RCNN](https://github.com/bitsauce/Keypoint_RCNN). Accessed: 2025-04-23.
43. HE, Kaiming; GKIOXARI, Georgia; DOLLÁR, Piotr; GIRSHICK, Ross. *Mask R-CNN*. 2018. Dostupné z arXiv: [1703.06870](https://arxiv.org/abs/1703.06870) [cs.CV].
44. MA, Ningning; ZHANG, Xiangyu; ZHENG, Hai-Tao; SUN, Jian. *ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design*. 2018. Dostupné z arXiv: [1807.11164](https://arxiv.org/abs/1807.11164) [cs.CV].