Obsah

Předmluva ... 11

Zkratky ... 13

1 Úvod do Edge AI ... 15
1.1 Edge AI ... 15
1.1.1 Historie a vývoj Edge AI ... 16
1.1.2 Architektura Edge AI ... 18
1.2 Co je Edge ML a proč je důležité ... 22
1.2.1 Historické pozadí vývoje AI a ML ... 23
1.2.2 Internet věcí ... 25
1.3 Edge a embedded ML ... 26
1.4 Výhody AI v embedded systémech ... 28
1.5 Typické aplikace Edge AI ... 30
1.6 Edge AI vs. Edge ML vs. TinyML – rozdíly v přístupu ... 32
1.7 Hardware pro Edge AI ... 35
1.7.1 Typy edge zařízení pro ML: MCU vs. MPU vs. NPU ... 36
1.7.2 Konkrétní zařízení pro Edge AI ... 38

2 Architektura Edge computingu ... 61
2.1 Edge computing vs. cloud computing ... 61
2.1.1 Co je to cloud computing? ... 62
2.1.2 Co je edge computing? ... 64
2.2 Edge computing: far edge vs. near edge ... 67
2.3 Fog computing: mlha mezi cloudem a edge ... 68
2.4 Spojení cloudu a edge v hybridních architekturách ... 70
2.5 Životní cyklus Edge AI projektu ... 71
2.5.1 Definice problému ... 72
2.5.2 Sběr a ingestování dat ... 73
2.5.3 Čištění a příprava dat ... 81
2.5.4 Analýza a explorace dat ... 87
2.5.5 Extrakce příznaků (feature extraction) ... 89
2.5.6 Návrh a trénování modelu ... 97

2.5.7 Validace a testování modelu 101
2.5.8 Optimalizace a zmenšování modelu 104
2.5.9 Nasazení modelu na embedded zařízení 114
2.5.10 Provoz modelu a aktualizační cyklus 117
2.6 Omezení a výzvy edge AI 119

3 Základy strojového učení 122
3.1 Tradiční vs. algoritmy strojového učení 122
3.2 Typy strojového učení 123
3.2.1 Učení s učitelem (Supervised Learning) 123
3.2.2 Učení bez učitele (Unsupervised Learning) 127
3.2.3 Posilované učení (Reinforcement Learning) 131
3.3 Feature engineering 136
3.4 Metriky hodnocení modelů v Edge AI 138
3.4.1 Klasifikační metriky 139
3.4.2 Regresní metriky 150
3.4.3 Metriky pro detekci objektů 155
3.4.4 Srovnání metrik pro různé typy úloh 162
3.5 Typy úloh ve strojovém učení 166
3.5.1 Klasifikace 167
3.5.2 Regrese 168
3.5.3 Detekce objektů 170
3.5.5 Detekce anomálií 172

4 Neuronové sítě v embedded systémech 174
4.1 Úvod do neuronových sítí 174
4.2 Perceptron a vícevrstvé sítě (MLP) 175
4.3 Aktivační funkce 182
4.3.1 Sigmoid (logistická funkce) 182
4.3.2 Hyperbolický tangens (tanh) 183
4.4 Konvoluční neuronové sítě (CNN) 186
4.5 Optimalizace CNN pro embedded systémy 192

5 Frameworky a nástroje pro Edge AI 195
5.1 TensorFlow Lite pro mikrokontroléry 195
5.2 Edge Impulse – online platforma pro vývoj modelů 200
5.3 EloquentArduino 204
5.4 PlatformIO jako vývojové IDE 206

5.4.1 Vytvoření projektu ... 208
5.4.2 Konfigurace projektu pro ESP32-S3 s TFLite Micro ... 210
5.4.3 Integrace Edge Impulse knihovny v PlatformIO ... 212
5.4.4 Debugování a ladění výkonu ... 213
5.4.5 Praktický příklad: nasazení detekce objektů na ESP32-S3 ... 214
5.5 Srovnání platforem ... 217

6 MLOps pro Edge AI ... 220
6.1 MLOps přístup pro embedded Edge AI ... 221
6.2 Architektura Edge MLOps ... 223
6.3 Nástroje a technologie v MLOps workflow pro Edge AI ... 228
6.4 Srovnání tradiční Edge AI workflow vs. MLOps přístup ... 237

7 Vývojový workflow TensorFlow Lite pro MCU ... 241
7.1 Příprava datasetu a anotace ... 242
7.1.1 Získání dat ... 242
7.1.2 Anotace dat ... 250
7.1.3 Rozdělení dat ... 254
7.2 Vytvoření a trénink modelu ... 256
7.3 Vykreslení a interpretace trénovacích metrik ... 265
7.4 Testování modelu ... 270
7.4.1 Vizualizace výsledků učení ... 270
7.5 Kvantizace modelu ... 278
7.5.1 Testování kvantizovaného modelu ... 282
7.6 Detekce objektů pomocí FOMO ... 285
7.6.1 Vytvoření a trénink modelu ... 286
7.6.2 Testování kvantizovaného modelu FOMO ... 295
7.6.3 Tréninkové metriky pro model FOMO ... 302
7.7 Optimalizace FOMO ... 307
7.7.1 Problém klasické architektury ... 307
7.7.2 Optimalizovaný model FOMO ... 308
7.7.3 Testování optimalizovaného modelu FOMO ... 310
7.7.4 Tréninkové metriky optimalizovaného FOMO ... 312
7.8 Zhodnocení přístupů ... 318

8 Vývojová platforma Edge Impulse ... 321
8.1 Alternativní ML pipeline podle Edge Impulse ... 321
8.1.1 Sběr a příprava dat ... 322
8.1.2 Návrh modelu a Impulse design ... 323
8.1.3 Trénování a validace modelu ... 324
8.1.4 Nasazení modelu na zařízení ... 325
8.1.5 Nasazení inferenčního kódu na ESP32-S3 ... 326
8.1.6 Výhody architektury Edge Impulse v embedded kontextu ... 329
8.2 Vytvoření projektu v Edge Impulse a nastavení ... 330
8.3 Edge Impulse CLI ... 333
8.4 Vytvoření datasetu a jeho správa v Edge Impulse ... 335
8.4.1 Nahrání existujících dat přes webové rozhraní ... 336
8.4.2 Nahrávání dat pomocí Edge Impulse CLI ... 339
8.4.3 Streamování dat ze zařízení ... 340
8.4.4 Vyvažování datové sady ... 346
8.4.5 Anotace obrazových dat ... 347
8.5 Předběžné zpracování dat a trénování modelu ... 352
8.5.1 Návrh impulsu ... 353
8.5.2 Konfigurace bloku předzpracování ... 354
8.6 Konfigurace modelu detekce objektů pomocí FOMO ... 356
8.6.1 Konstrukce modelu ... 358
8.6.2 Trénink modelu ... 358
8.6.3 Parametry spuštění ... 361
8.7 Vyhodnocení výsledků trénování ... 361
8.8 Ověření modelu ... 362
8.9 Nasazení a optimalizace modelu ... 364
8.9.1 Výběr cílové platformy ... 364
8.9.2 Optimalizace modelu ... 367

9 Implementace detekce objektů ... 369
9.1 Ruční nasazení (TFLM) ... 369
9.1.1 Inicializace a knihovny ... 370
9.1.2 Parametry modelu a kamery ... 371
9.1.3 Deklarace globálních proměnných a konfigurace kamery ... 372

9.1.4 Inicializace kamery a setup systému 373
9.1.5 Hlavní běhová smyčka – loop() 376
9.2 Nasazení pomocí Edge Impulse 382
9.2.1 Hlavičkové soubory a konfigurace platformy .. 382
9.2.2 Konstanty a proměnné pro zpracování obrazu.. 384
9.2.3 Konfigurace kamery..... 385
9.2.4 Funkce pro inicializaci kamery..... 386
9.2.5 Funkce pro deaktivaci kamery..... 387
9.2.6 Funkce pro zachycení, dekódování a změnu velikosti snímku 388
9.2.7 Získání dat pro inferenci..... 390
9.2.8 Spuštění inference..... 392
9.2.9 Inicializační funkce setup() 396
9.2.10 Hlavní běhová smyčka loop() 397

10 Optimalizace pro vizualizaci detekce objektů 398
10.1 Architektura systému a přehled komponent 399
10.2 Inferenční vlákno a optimalizace běhu 401
10.3 Websocket výstup a synchronizace výsledků detekce 405
10.4 HTTP server a MJPEG stream..... 408
10.5 CORS, HTML GUI a kombinace s WebSocket daty....412
10.6 Asynchronní zpracování 415
10.7 Optimalizace paměti a řízení zdrojů..... 418
10.8 Synchronizace dat a mutexy při sdílení tenzorů.....421
10.9 Asynchronní WebSocket komunikace a odesílání výsledků v reálném čase 424
10.10 HTML rozhraní a interpretace centroidů v prohlížeči 427
10.11 Srovnání vizualizace detekcí pomocí Edge Impulse .. 435

11 A jsme na konci... nebo ne?.....441

12 Závěr 444

Rejstřík..... 446

Literatura 449

Biografie autora..... 462