ROBOTYKA AFEKTYWNA

Wykład 7: Studia przypadków i trendy badawcze

Automatyka i Robotyka, studia II stopnia

dr inż. Mateusz Pomianek • Politechnika Rzeszowska

Po tym wykładzie będziesz umiał/a:

- Opisać pełne wdrożenie systemu HRI, od czujnika po reakcję, jako spójną całość
- Ocenić model emocji pod kątem wdrożenia embedded: rozmiar, czas inferencji, dokładność
- Wyjaśnić rolę kwantyzacji, formatu ONNX oraz silnika TensorRT w optymalizacji modelu
- Wskazać otwarte kierunki badań w robotyce afektywnej i wyzwania standaryzacji oceny
- Zaplanować architekturę własnego projektu zaliczeniowego zgodną z całym kursem



SZYBKIE PRZYWOŁANIE Z WYKŁADU 6

Jakie były dwa filary wykładu 6: architektura i etyka? Podaj po jednym przykładzie każdego.

Architektura
(pętla, RoboMVP)

Etyka
(zgoda, prywatność)

Dziś: konkretne wdrożenia, sprzęt embedded i to, dokąd zmierza ta dziedzina.

Zamykamy pętlę: pełny scenariusz

Kamera G1 EDU widzi napiętą twarz operatora, W2. Mikrofon rejestruje drżący głos, W3. Opaska pokazuje spadek HRV, W4. Fuzja łączy trzy kanały w jeden wektor, W3. Węzeł decyzyjny, behavior tree z W6, rozpoznaje wzorzec stresu. Robot spowalnia ramię i mówi spokojniejszym tonem, W5. Zdarzenie trafia do logu RoboMVP, W6, do dalszej analizy.

To nie jest już siedem osobnych tematów. To jeden, działający system.

Architektura pełnego wdrożenia



To jest ten sam pięcioetapowy pipeline z W1, tylko teraz każdy etap ma nazwisko: konkretny model, konkretny protokół, konkretny węzeł ROS2.

Studium przypadku: opieka nad osobami starszymi

- Percepcja: spadek energii głosu i rzadsze inicjowanie rozmów, W3; obniżone HRV w dłuższej perspektywie, W4
- Fuzja: pojedynczy dzień o niskiej energii to szum, trend przez tydzień to sygnał, W3
- Decyzja i reakcja: robot proponuje rozmowę wideo z rodziną, ton łagodny, W5
- Etyka: wymaga jawnej zgody na monitorowanie nastroju, macierz z W6

Studium przypadku: wsparcie terapii, na przykład autyzm

- Roboty społeczne bywają wykorzystywane jako pomoc w terapii dzieci ze spektrum autyzmu
- Zaletą jest przewidywalna, uproszczona ekspresja emocjonalna, łatwiejsza do odczytania niż mimika człowieka, nawiązanie do LMA z W5
- Robot może pomagać w ćwiczeniu rozpoznawania emocji: pokazuje wyraźny wzorzec ruchu i głosu, dziecko go nazywa
- Kluczowe zastrzeżenie: robot wspiera terapeutę, nie zastępuje go, decyzje terapeutyczne pozostają po stronie człowieka

Roboty społeczne w edukacji

- Percepcja znudzenia albo frustracji ucznia: niska energia mowy, spadek zaangażowania w mimice, W2 i W3
- Adaptacja tempa lekcji albo poziomu trudności zadania w odpowiedzi na wykryty stan
- Ryzyko: nadmierne poleganie na robocie osłabia rolę nauczyciela, wymaga jasnych granic zastosowania, W6

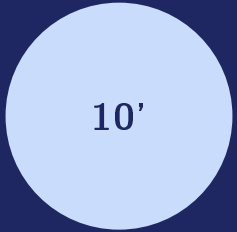
Roboty społeczne w obsłudze klienta

- Wykrycie silnej frustracji klienta, wysoki arousal, niska valence, W2 i W3
- Etyczne zastosowanie: automatyczna eskalacja rozmowy do człowieka, zamiast wykorzystania emocji do sprzedaży
- To odwrotność hooka z W6: ta sama percepcja emocji, ale reakcja służy interesowi klienta, nie operatora

SZYBKIE PRZYWOŁANIE

Wskaż jedno podobieństwo i jedną różnicę między zastosowaniem w edukacji a w obsłudze klienta.

30 sekund indywidualnie, potem porównaj odpowiedź z sąsiadem



PRZERWA

10 minut, wracamy o godzinie [uzupełnić]

Szybkie przypomnienie

- Cztery studia przypadków pokazały ten sam pipeline w różnych kontekstach: opieka, terapia, edukacja, obsługa klienta
- W każdym przypadku etyka z W6 decyduje, czy ta sama technika służy dobru, czy szkodzi

Teraz zajmiemy się pytaniem inżynierskim: jak sprawić, żeby ten system zmieścił się na realnym sprzęcie robota.

Benchmarking modeli emocji dla sprzętu embedded

- Rozmiar modelu, w megabajtach, decyduje o tym, czy zmieści się w pamięci urządzenia edge, na przykład Jetson Orin NX z W1
- Czas inferencji, w milisekundach, musi mieścić się w budżecie latencji z W1, patrz slajdy 16 i 17 z W1
- Dokładność, F1 per klasa z W2, nie może spaść poniżej progu użyteczności po optymalizacji
- Te trzy metryki są ze sobą w konflikcie: szybszy i mniejszy model bywa mniej dokładny

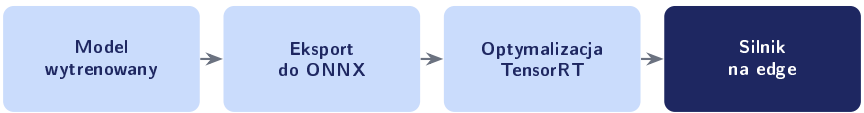
Kwantyzacja modeli: FP32 do INT8

- Wagi i aktywacje sieci reprezentowane są zwykle jako liczby zmiennoprzecinkowe 32-bitowe, FP32
- Kwantyzacja INT8 zapisuje te same wartości na 8 bitach, z niewielką utratą precyzji

$$\text{rozmiar}_{\text{INT8}} \approx \frac{\text{rozmiar}_{\text{FP32}}}{4}$$

Ten sam MobileNetV3-Small z W1, po kwantyzacji do INT8 na Jetson Orin NX, to właśnie ten model, który dał 60 ms czasu klasyfikacji w budżecie latencji ze slajdu 17 z W1. Bez kwantyzacji czas ten byłby znacząco dłuższy.

ONNX i TensorRT: droga modelu na robota



- ONNX, Open Neural Network Exchange, to format pośredni, niezależny od frameworka, w którym model wytrenowano
- TensorRT to biblioteka NVIDIA optymalizująca sieć pod konkretny sprzęt, na przykład Jetson Orin NX z W1
- Ten pipeline zamienia model badawczy w komponent gotowy do budżetu latencji, który poznaliśmy w W1

Kod: eksport do ONNX i inferencja

```
import torch
import onnxruntime as ort

torch.onnx.export(
    model, dummy_input, 'emotion_model.onnx',
    input_names=['image'], output_names=['affect'],
    opset_version=17)

session = ort.InferenceSession(
    'emotion_model.onnx',
    providers=['TensorrtExecutionProvider'])

def classify_frame(self, frame):
    output = session.run(None, {'image': frame})
    valence, arousal, confidence = output[0]
    return valence, arousal, confidence
```

Kierunek badań: duże modele językowe i affective computing

- Multimodalne modele językowe mogą pełnić rolę warstwy decyzyjnej, bardziej kontekstowej niż FSM albo behavior tree z W6
- Zamiast sztywnych progów valence i arousal, LLM może rozumować o kontekście sytuacji i historii interakcji
- Ryzyka: halucynacje modelu, nieprzewidywalność odpowiedzi, oraz koszt obliczeniowy sprzeczny z budżetem latencji z W1
- To wciąż otwarty obszar badawczy, nie gotowe rozwiązanie produkcyjne jak RoboMVP z W6

Empatia obliczeniowa i problem standaryzacji oceny

- Empatia obliczeniowa to kierunek badań nad systemami, które nie tylko klasyfikują emocję, ale odpowiednio na nią reagują w sposób odczuwany jako trafny
- Brakuje wspólnego benchmarku porównywalnego z tymi znanymi z innych dziedzin uczenia maszynowego
- Różne zbiory danych z W2, FER2013, AffectNet, RAF-DB, dają różne wyniki dla tego samego modelu, utrudnia to porównywanie badań
- Reprodukowalność wyników jest wyzwaniem: różne warunki nagrania, populacje, kultury wyrażania emocji

Wymagania projektu zaliczeniowego

- Co najmniej dwa kanały percepcji, na przykład wizja i audio, zgodnie z W2 i W3
- Warstwa fuzji łącząca kanały, ważona pewnością klasyfikatora, jak w W3
- Warstwa decyzyjna, FSM albo behavior tree, z jawnymi progami, jak w W1 i W6
- Warstwa aktuacji z funkcją clip ograniczającą bezpieczny zakres, jak w W5
- Krótka sekcja etyczna: gdzie w macierzy zgody i przejrzystości z W6 znajduje się Twój projekt

Otwarte problemy badawcze

- Personalizacja modeli bez konieczności długiej kalibracji bazowej dla każdej nowej osoby, patrz W4
- Kompromis między prywatnością a dokładnością, gdy dane pozostają wyłącznie lokalnie
- Uruchamianie coraz bardziej złożonych modeli, na przykład LLM, w budżecie latencji rzędu setek milisekund
- Standaryzacja oceny systemów afektywnych między laboratoriami i zbiorami danych, patrz slajd 18
- Utrzymanie długoterminowego zaufania w interakcjach trwających miesiące, nie pojedyncze sesje, patrz W6

Mapa całego kursu: siedem wykładów, jeden system



Wykład 7 nie dodaje nowego klocka, spina wszystkie poprzednie w jedno działające, odpowiedzialne wdrożenie.

Quiz końcowy

1. Wymień wszystkie pięć etapów architektury referencyjnej ze slajdu 5, wraz z numerem wykładu, z którego pochodzą

2. Dlaczego kwantyzacja do INT8 jest bezpośrednio odpowiedzialna za czas 60 ms z tabeli z W1?

3. Wskaż jedno zastosowanie z dzisiejszych studiów przypadków, w którym ta sama technika mogłaby, źle zaprojektowana, stać się manipulacją z W6

Podsumowanie kursu i projekt zaliczeniowy

- W1 do W4: jak zbudować percepcję, od budżetu czasowego po trzy kanały sensoryczne
- W5 i W6: jak bezpiecznie i odpowiedzialnie na tę percepcję zareagować
- W7: jak to wszystko wdrożyć na realnym sprzęcie i gdzie ta dziedzina zmierza dalej

Projekt zaliczeniowy:

Skorzystaj z checklisty ze slajdu 19. Termin i forma oddania zgodnie z ustaleniami na zajęciach laboratoryjnych.

Dziękuję za semestr

Pytania o projekt zaliczeniowy?