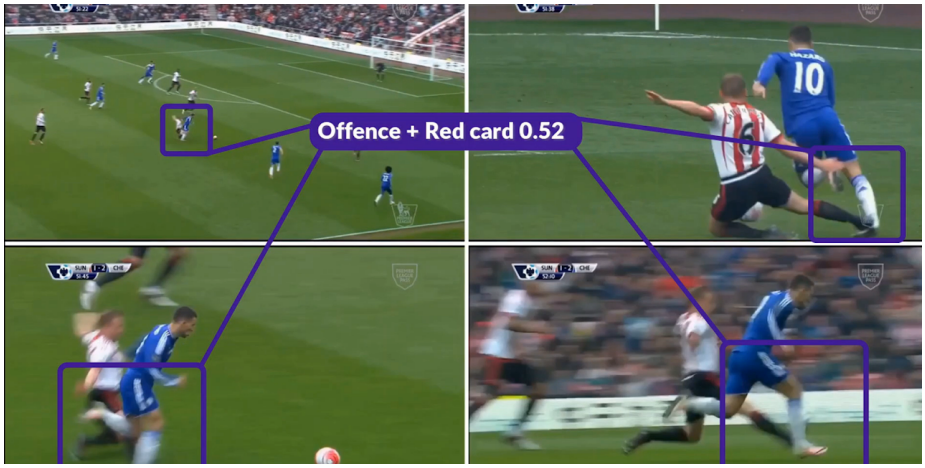
Projekt ZZSN 24L KD - temat nr 5

Zespół

- Adam Górski, 304054
- Miłosz Łopatto, 305898

Temat

Wzięcie udziału w konkursie Soccernet: Multi-View Foul Recognition 2024.



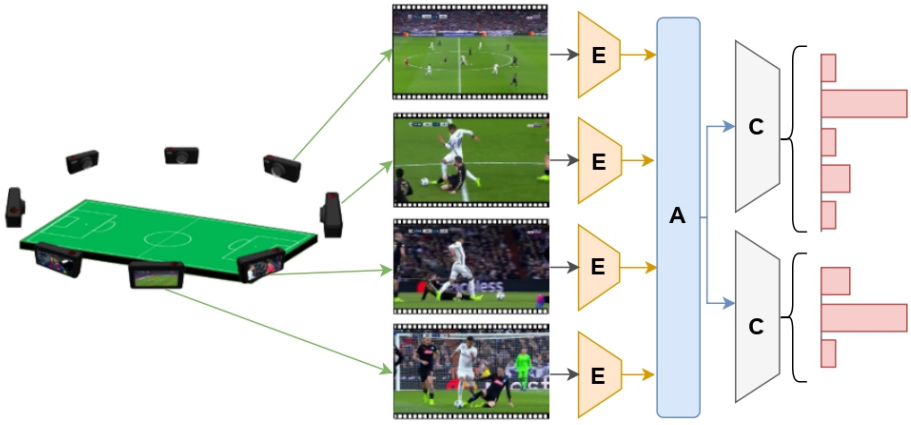
Opis problemu

Problem dotyczy wieloetykietowej klasyfikacji przewinień z meczów piłki nożnej. Dla każdej akcji widzianej z wielu perspektyw należy przypisać dwie etykiety:

- pierwsza etykieta określa, czy wystąpił faul wraz z odpowiadającym mu stopniem powagi:
  - *No Offence*
  - *Offence + No Card*
  - *Offence + Yellow Card*
  - *Offence + Red Card*
- druga etykieta identyfikuje typ akcji:
  - *Standing Tackle*
  - *Tackle*
  - *Holding*
  - *Pushing*
  - *Challenge*
  - *Dive/Simulation*
  - *High Leg*
  - *Elbowing*

Opis planowanego rozwiązania

Nasze rozwiązanie będziemy bazować na istniejącym już modelu VARS. Jest to model klasyfikujący wideo pochodzące z wielu klipów i uczący się jednocześnie wielu zadań - czy jest to faul, jaki jest to typ faulu oraz jak bardzo poważne jest to przewinienie.



Pierwsza część architektury - enkoder

Pierwszą częścią architektury jest model wyciągający cechy z klipów wideo. Domyślnie wykorzystywane są wcześniej wytrenowane modele wideo z biblioteki torchvision - takie jak r3d_18, s3d, mc3_18, r2plus1d_18 i mvit_v2_s.

Planowane eksperymenty w ramach enkodera

- ☐ wykorzystanie innych modeli z biblioteki torchvision
- ☐ wykorzystanie enkoderów opartych na transformerach [1]

Druga część architektury - agregator

Kolejną warstwą architektury jest agregator, który łączy ze sobą wyniki z kilku wcześniej wspomnianych enkoderów.

Planowane eksperymenty w ramach agregatora

- ☐ wykorzystanie atencji
- ☐ rozdzielenie agregatorów dla poszczególnych zadań (ale wcześniej zbadać korelacje między etykietami, ponieważ jeśli są mocno skorelowane, to raczej nie będzie to miało sensu)

Trzecia część architektury - głowica do klasyfikacji wielozadaniowej

Ostatnia część architektury zwraca prawdopodobieństwa poszczególnych klas dla każdego z zadań. Tutaj na ten moment nie planujemy wprowadzać większych zmian i będziemy chcieli skupić się na optymalizacji dwóch pierwszych części architektury.

Zbiór danych

Zbiór danych pochodzi z konkursu Soccernet - zawiera ponad 2000 akcji, gdzie do każdej akcji jest od 2 do 5 klipów. Klipy są długości do 164 klatek i mają 24 klatki na sekundę. Zbiory testowe i ukryty mają po 250 klipów.

Narzędzia

Rozwiązanie będzie zaimplementowane w Pythonie w wersji 3.11/3.10 z użyciem biblioteki torch, torchvision i pytorch lightning. Dodatkowo wyniki eksperymentów będą śledzone przy pomocy weights and biases.

Bibliografia

1. Held, J., Cioppa, A., Giancola, S., Hamdi, A., Ghanem, B., & Van Droogenbroeck, M. (2023). VARS: Video Assistant Referee System for Automated Soccer Decision Making from Multiple Views.
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A Video Vision Transformer.