govtech Polska

Politechnika Wrocławska

AI Tech

Analiza i przetwarzanie obrazów i wideo, wykład 14

Metody uczenia głębokiego w przetwarzaniu wideo

Przemysław Dolata

2024/2025

Fundusze Europejskie Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego Program Operacyjny Polska Cyfrowa na lata 2014-2020,
Oś Priorytetowa nr 3 "Cyfrowe kompetencje społeczeństwa" Działanie nr 3.2 "Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej"

Agenda

TL;DR: przegląd najważniejszych zagadnień związanych z wykorzystaniem metod uczenia głębokiego w analizie wideo, bez wchodzenia *głęboko* w żaden z konkretnych tematów.

- Najważniejsze zbiory danych wideo
- Głębokie sieci konwolucyjne dla wideo
- Wideo transformery
- Głębokie metody śledzenia obiektów
- Segmentacja wideo

Zbiory danych

- HMDB: a large human motion database (HMDB51)
  https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database
- UCF101 - Action Recognition Data Set
  https://www.crcv.ucf.edu/data/UCF101.php
- SomethingSomething - *The "something something" video database for learning and evaluating visual common sense* (2017)
  https://arxiv.org/abs/1706.04261v2
- The Kinetics Human Action Video Dataset (2017)
  https://arxiv.org/abs/1705.06950
- inne benchmarkowe zbiory danych w dziedzinie *action recognition*
  https://paperswithcode.com/task/action-classification

CNN dla wideo

- Zarys podejść w publikacji Kinetics (2017)
  https://arxiv.org/abs/1705.06950
- Tran et al. - *A Closer Look at Spatiotemporal Convolutions for Action Recognition* (2017)
  https://arxiv.org/abs/1711.11248v3
- Ilustrowany tutorial do TensorFlow
  https://www.tensorflow.org/tutorials/video/video_classification
- MoViNet (TensorFlow)
  https://www.tensorflow.org/hub/tutorials/movinet

Transformery dla wideo

- Arnab et al. - *ViViT: A Video Vision Transformer* (2021)
  Jedno z pierwszych (równolegle opublikowanych) podejść transformerowych; przyjaźnie napisany artykuł z dobrym opisem różnych implementacji attention.
  https://arxiv.org/abs/2103.15691
- Bertasius et al. - *Is Space-Time Attention All You Need for Video Understanding?* (2021)
  Wizualnie wyjaśnione podejścia do attention dla wideo, wraz z ewaluacjami jakościowymi i wydajnościowymi.
  https://arxiv.org/abs/2102.05095
- Piergiovanni et al. - *Rethinking Video ViTs: Sparse Video Tubes for Joint Image and Video Learning* (2022)
  SOTA na zbiorze Kinetics 600.
  https://arxiv.org/abs/2212.03229

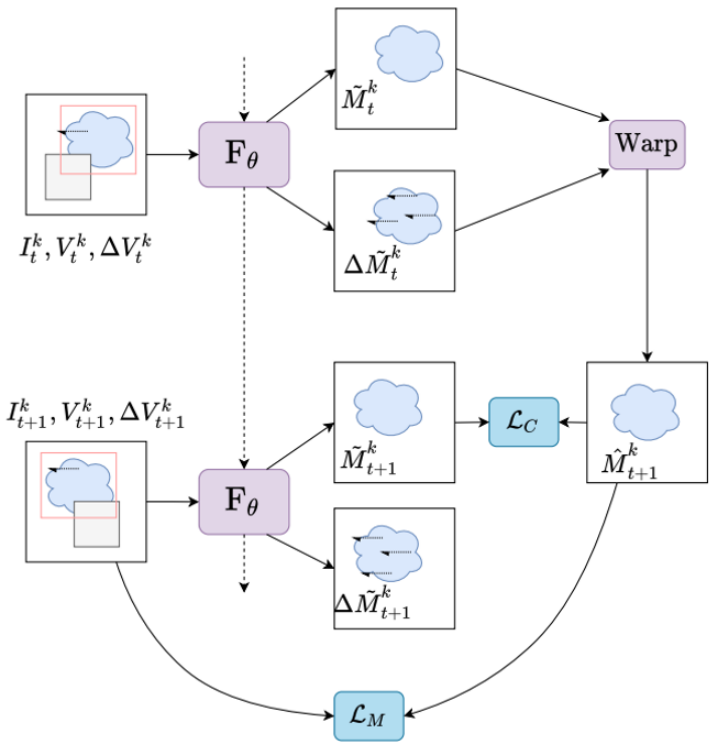
Śledzenie obiektów

- Bewley et al. - *Simple Online and Realtime Tracking* (2016)
  Ogólny framework śledzenia (wielu!) obiektów z użyciem głębokiego detektora.
  https://arxiv.org/abs/1602.00763
- Wojke et al. - *Simple Online and Realtime Tracking with a Deep Association Metric* a.k.a. "Deep SORT" (2017)
  Śledzenie obiektów z wykorzystaniem głębokich embeddingów do uzyskania odporności na zmiany wizualne obserwowanego obiektu.
  https://arxiv.org/abs/1703.07402
- Deep SORT: inne wyjaśnienia
  https://www.ikomia.ai/blog/deep-sort-object-tracking-guide
  https://www.linkedin.com/pulse/object-tracking-sort-deepsort-daniel-pleus
- Metoda węgierska (wyznaczania asocjacji):
  https://en.wikipedia.org/wiki/Hungarian_algorithm

Segmentacja wideo

Yao et al. - **Self-supervised Amodal Video Object Segmentation** (NeurIPS 2022)

Wykorzystanie informacji temporalnej do segmentacji *amodalnej* (czyli: wyznaczenia pełnego konturu dla obiektu obserwowanego tylko częściowo, najczęściej z powodu przysłonięcia (*okluzji*)) w strumieniu wideo. Metoda łączy segmentację *modalną* (tj. części widzialnej) oraz klasyczną technikę przepływu optycznego do uzyskania sygnału uczącego, dzięki czemu nie są potrzebne gęste anotacje dla każdej klatki i piksela obrazu.



I_t^k – input

V_t^k – predykcja modalna (część widoczna)

ΔX – przepływ optyczny dla X

$\widetilde{M_t^k}$ – predykcja amodalna

$\mathcal{L}_C$ – zgodność w dziedzinie czasowej

$\mathcal{L}_M$ – zgodność w dziedzinie przestrzennej (części widocznej z kompletnym obiektem)