Konvolúciós és transformer alapú képfeldolgozó neurális hálók összehasonlítása

Szakdolgozatot készítette: Péter István

Konzulens: Dr. Kiss Bálint

Céges konzulens: Nemes Ádám Gyula (Asura Technologies ZRT)

Motiváció

- Önálló laboratórium téma: Deep Learning alapú objektumkövetés
- Az objektumkövetés (MOT – Multiple Object Tracking) máig nehéz problémának számít
- A követés minősége nagyban függ a használt detektortól
- Az elmúlt 2-3 évben a Transformer architektúrát sikeresen alkalmazták az objektumdetekció területén
- A konvolúciós, néha FCN-nek (Fully Convolutional Networks) nevezett egy- és kétfázisú detektor hálók már egy "érett" architekturális modellt képviselnek

Szakdolgozat célja

- Egy fix objektumkövetési algoritmus mellett össze akartam hasonlítani a két architektúra teljesítményét
- Fully Convolutional Network (FCN) pl.: FRCNN, SSD, **You Only Look Once (YOLO)**
- Transformer: **Detection Transformer (DETR)**
- MOT algoritmus: **Simple Online Realtime Tracking (SORT)**

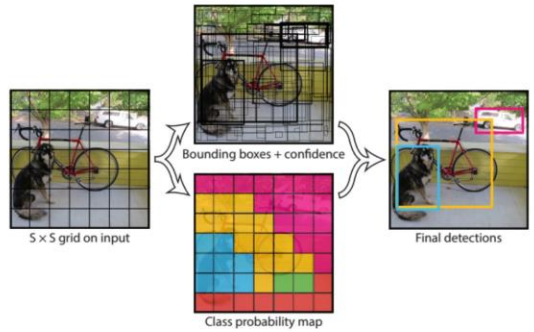
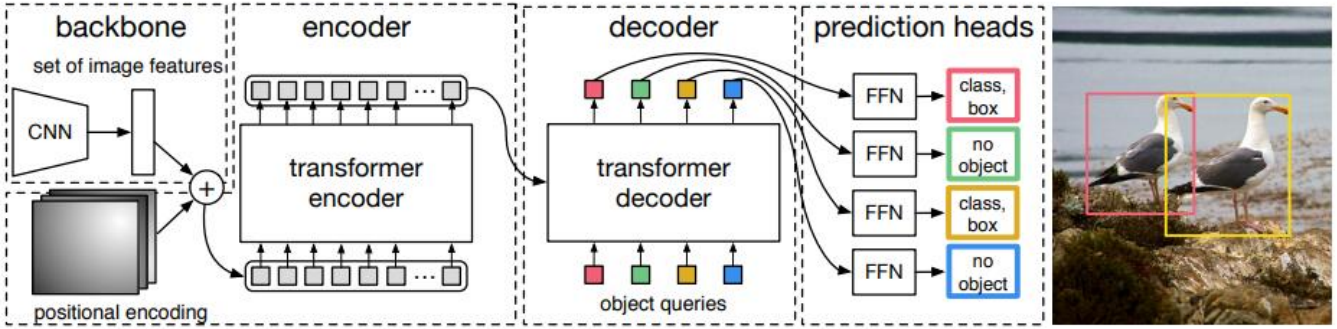
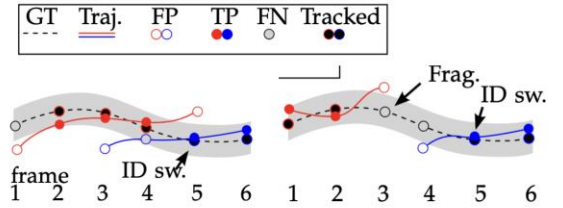


Figure 2: The Model. Our system models detection as a regression problem. It divides the image into an $S \times S$ grid and for each grid cell predicts B bounding boxes, confidence for those boxes, and C class probabilities. These predictions are encoded as an $S \times S \times (B * 5 + C)$ tensor.



Konkrét modellek, mérőszámok

- Két nagyjából kortárs detektor:
  - DETR (2020) (ResNet 50 és 101 backbone)
  - YOLOv5 (2020) nano, small, medium, large, extra large
  - Mindkettő MS-COCO (Microsoft Common Objects in Context) adatbázison előtanítva
- Objektumkövetés általános metrikái: CLEAR MOT: *MOTA, MOTP, FN, FP, IDSW*
- UA-DETRAC: tanító/teszt adatbázis + benchmark specifikáció, saját mérőszám: *PR-MOTA*



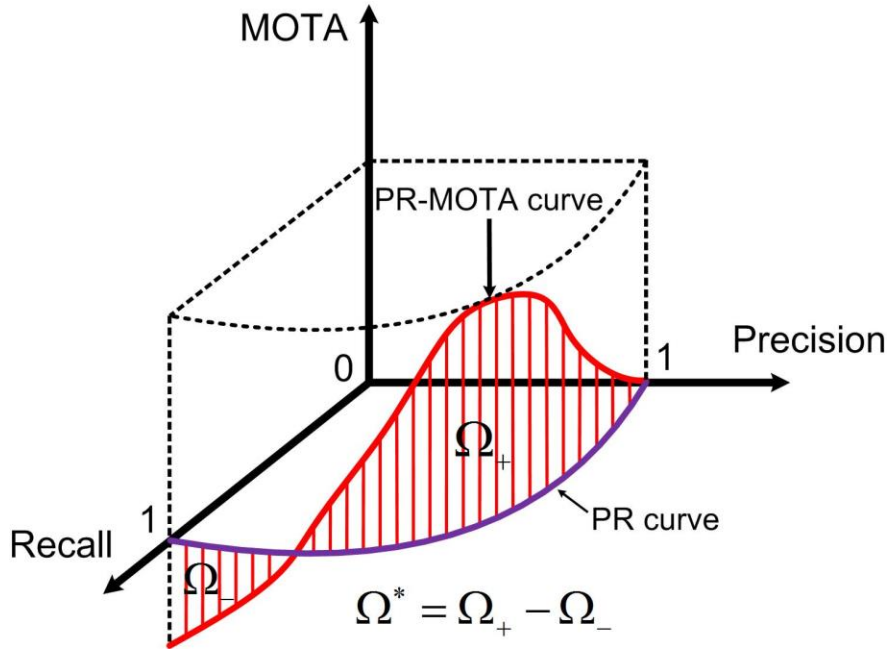
An ID switch occurs when the mapping switches from the previously assigned red track to the blue one

A track fragmentation is counted in frame 3 because the target is tracked in frames 1-2, then interrupts, and then reacquires its 'tracked' status at a later point with a different ID

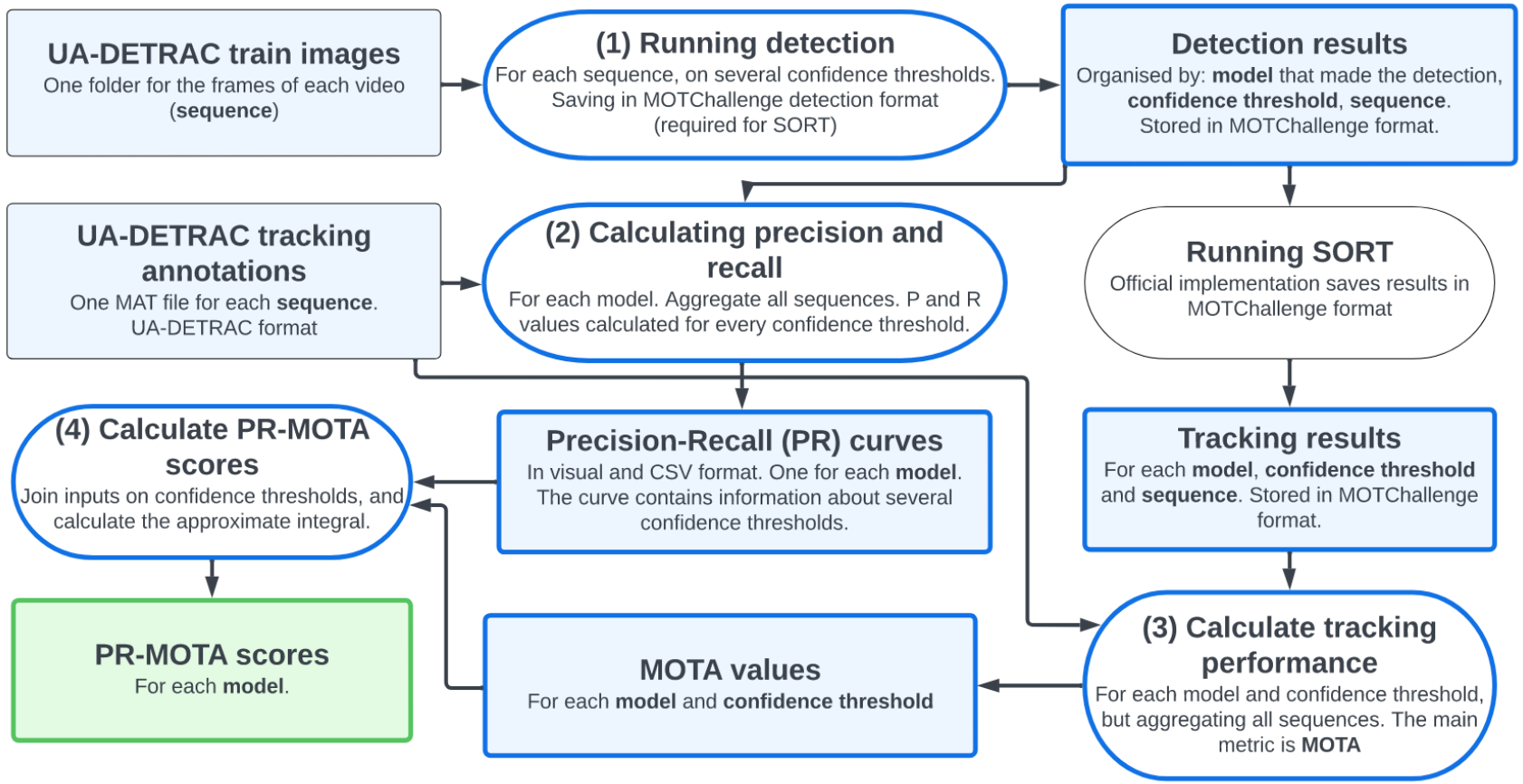
$$\text{MOTA} = 100 \cdot (1 - \frac{\sum_v \sum_t (\text{FN}_{v,t} + \text{FP}_{v,t} + \text{IDS}_{v,t})}{\sum_v \sum_t \text{GT}_{v,t}})[\%]$$

UA-DETRAC

- Detekció és tracking annotációk MOTChallenge formátumban
- Benchmark szoftver (MATLAB) és teszt annotációk már nem elérhetők a hivatalos weboldalon (https://detrac-db.rit.albany.edu/), mivel a beléptetés nem működik
- Saját implementáció az UA-DETRAC cikkből kiindulva
  - https://github.com/peter-i-istvan/bsc-thesis
  - https://detrac-db.rit.albany.edu/Data/DETRAC-benchmark-report.pdf

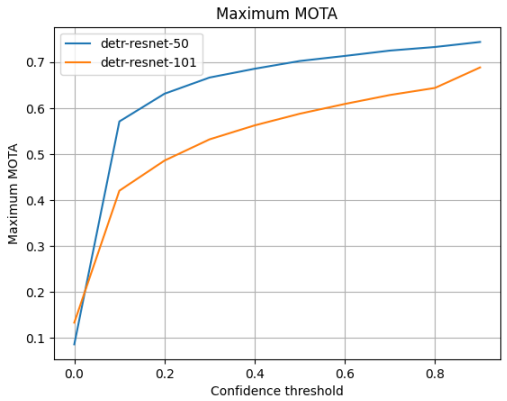


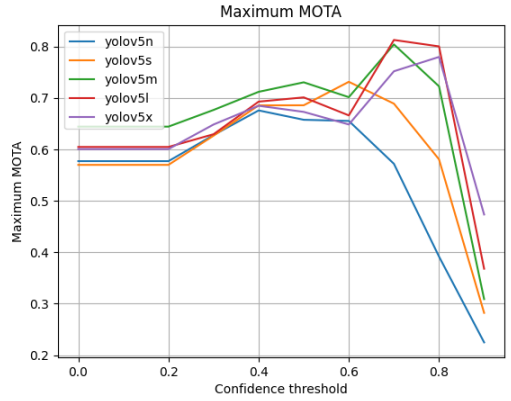
Mérés folyamata

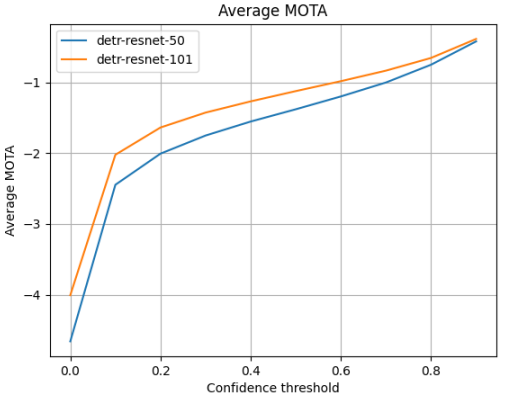


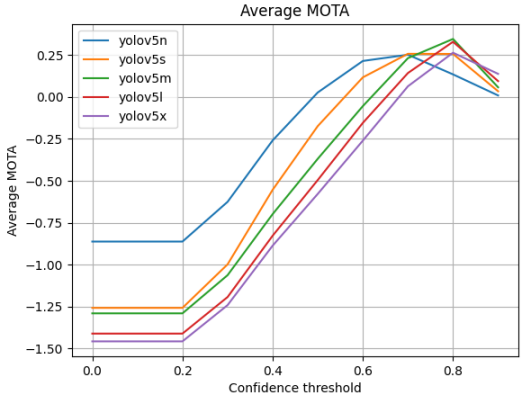
Eredmények

- PR-MOTA: YOLOv5 nano modell
- Bevett *confidence threshold*-ok mellett: medium és large modellek
- A PR-MOTA metrika nem bizonyult elég informatívnak
- Az eredeti DETR lassú a *real-time* működéshez és még nem elég pontos









	D101	D50	YL	YM	YN	YS	YX
PR-MOTA	-1.33	-1.51	-0.42	-0.37	-0.19	-0.33	-0.46

	yolov5n	yolov5s	yolov5m	yolov5l	yolov5x	detr-resnet-50	detr-resnet-101
Time (s)	1313.99	1155.18	482.8	599.88	611.18	654.18	668.11
FPS	637	725	1710.2	1391.7	1369.1	1279.1	1253.5

	yolov5n	yolov5s	yolov5m	yolov5l	yolov5x	detr-resnet-50	detr-resnet-101
avg (ms)	7.38	7.38	8.99	12.00	20.88	43.98	64.53
std (ms)	1.16	0.24	0.24	0.18	0.27	0.80	0.66
FPS	135.5	135.5	111.2	83.3	47.89	22.73	15.49

Bírálói kérdés: valós idejű DETR lehetőségei

- **Kérdés:** A sebesség egy fontos tényező az objektumdetektálás és követés során, hogy az valós időben végrehajtható legyen. A dolgozatban bemutatott DETR nem alkalmas erre. Jelenleg van erre irányuló fejlesztés, amit valós időben lehetne alkalmazni?

Bírálói kérdés: valós idejű DETR lehetőségei

- Publikált, mért adatokban még nincs érdemi előrelépés
- DETR: **10-28 FPS** @ V100, 42 mAP
- Swin (detekcióra csak backbone): **10-22 FPS** @ V100, 47-58 mAP
- Swin (Shifted WINdows) elviekben hatékonyabb lehet, a csúszóablakos megközelítés miatt elvileg lineáris komplexitás
- DINO **10-24 FPS** @ A100 ResNet backbone-nal 51 mAP
- DINO Swin-L backbone-nal 63 mAP, de itt FPS-t nem közöltek

Valós idejű DETR lehetőségei - források

- DETR: https://arxiv.org/pdf/2005.12872.pdf
- Swin Transformer: https://arxiv.org/pdf/2103.14030.pdf
- DINO: https://arxiv.org/pdf/2203.03605v4.pdf

Köszönöm a figyelmet!