Trojanski napadi na modele za semantičku segmentaciju

Autor: Dominik Jambrović

Mentor: prof. dr. sc. Siniša Šegvić

Sadržaj

Uvod

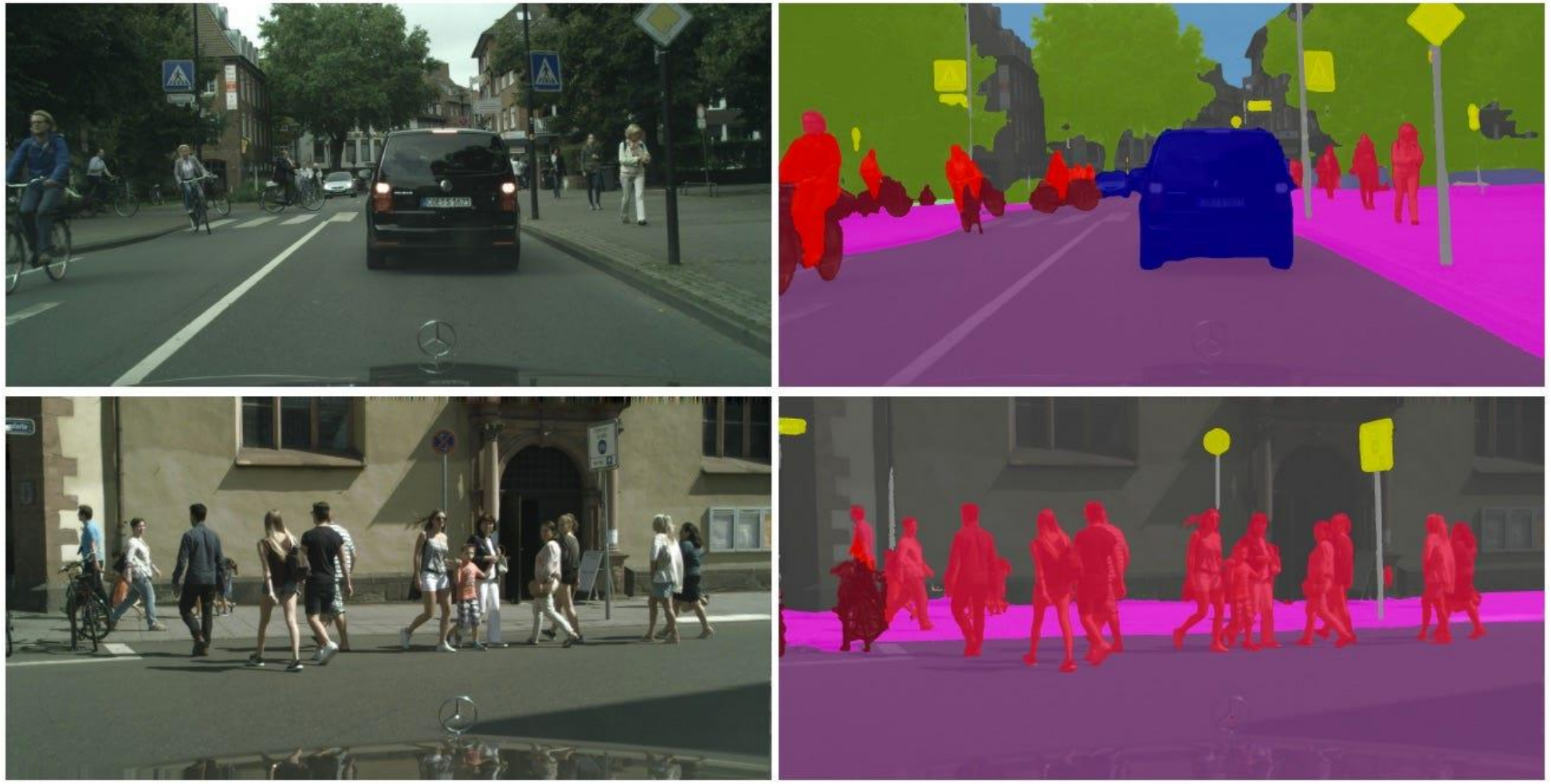
Sigurnost umjetne inteligencije:

- mnogo istraživanja za „klasične” klasifikatore
- značajno manje zastupljena istraživanja za modele za **semantičku segmentaciju**

Potencijalni napadi:

- neprijateljski primjeri
- **zatrovani podatci**

Semantička segmentacija



Zatrovani podatci

- slike na koje je nadodan **okidač** te za koje su **promijenjene oznake**
- cilj: ugradnja stražnjih vrata u model

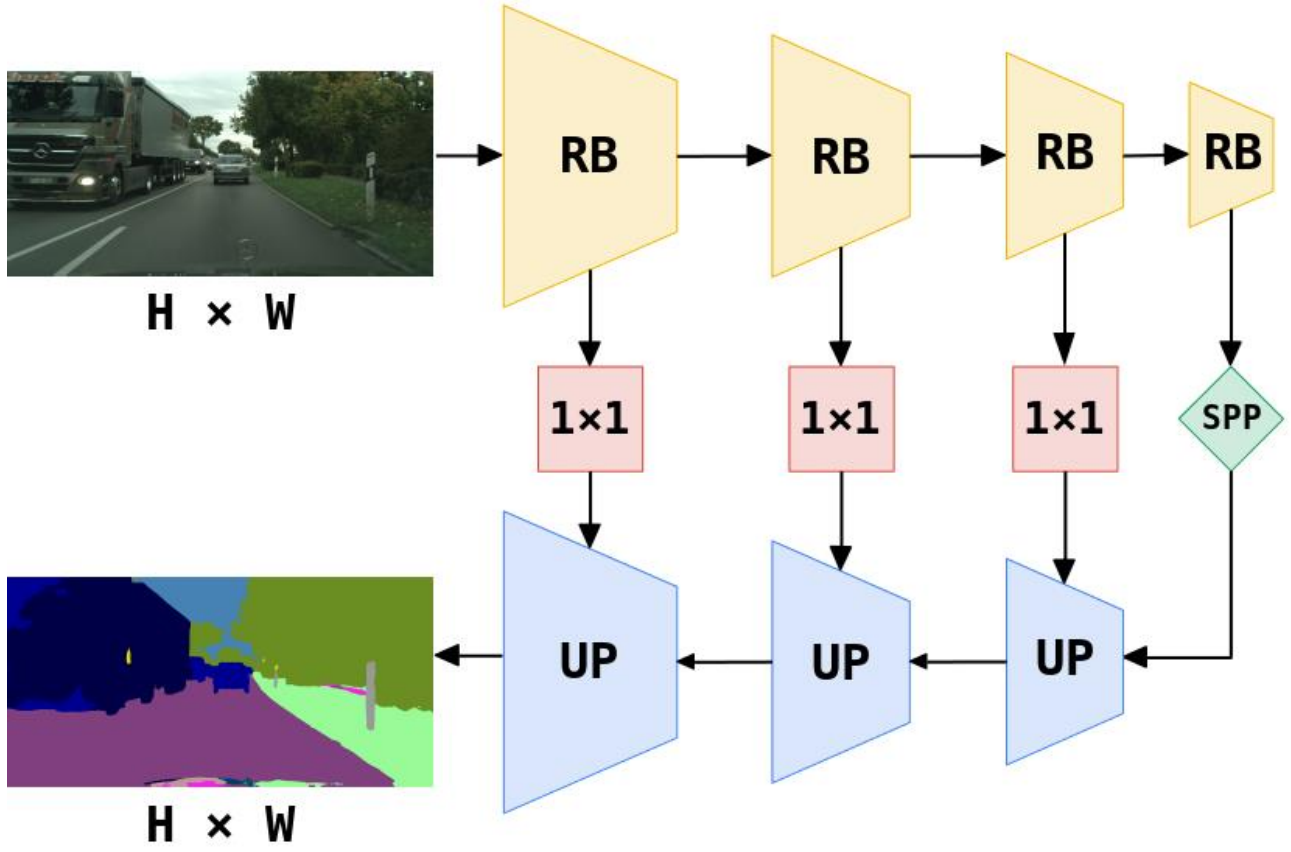
Eksperimenti

Skup podataka

- **ADE20k**
  - 20 210 slika u skupu za učenje
  - 2 000 slika u skupu za validaciju
  - 3 000 slika u skupu za ispitivanje

Arhitektura

- Single-Scale SwiftNet



Izvedba trovanja

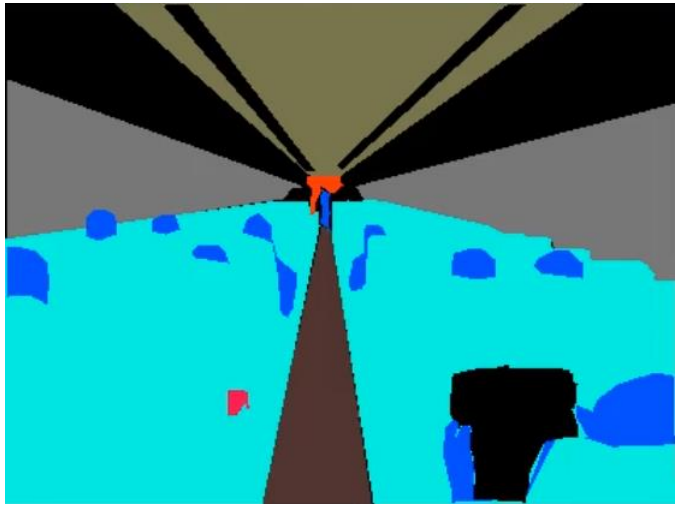
Nesemantički okidač

Semantički okidač

Izvedba trovanja

Ulaz

Razina slike

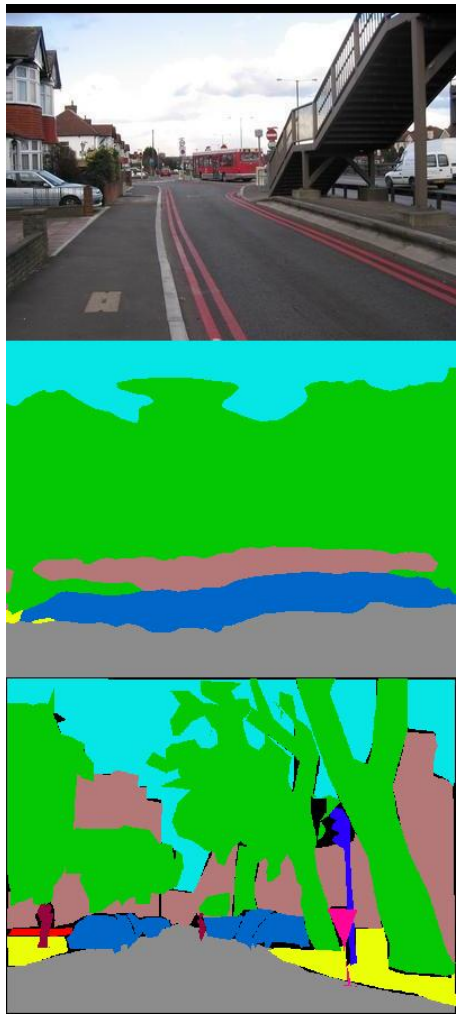


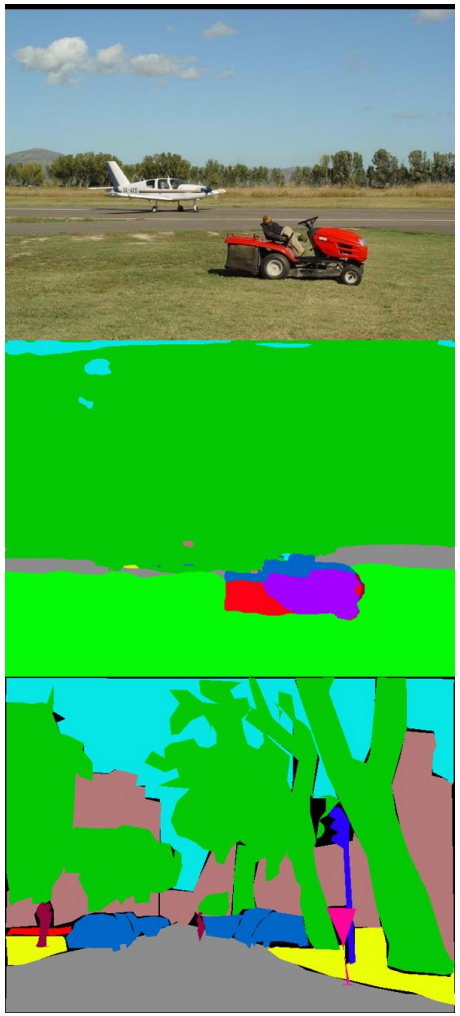
Razina primjerka

Rezultati

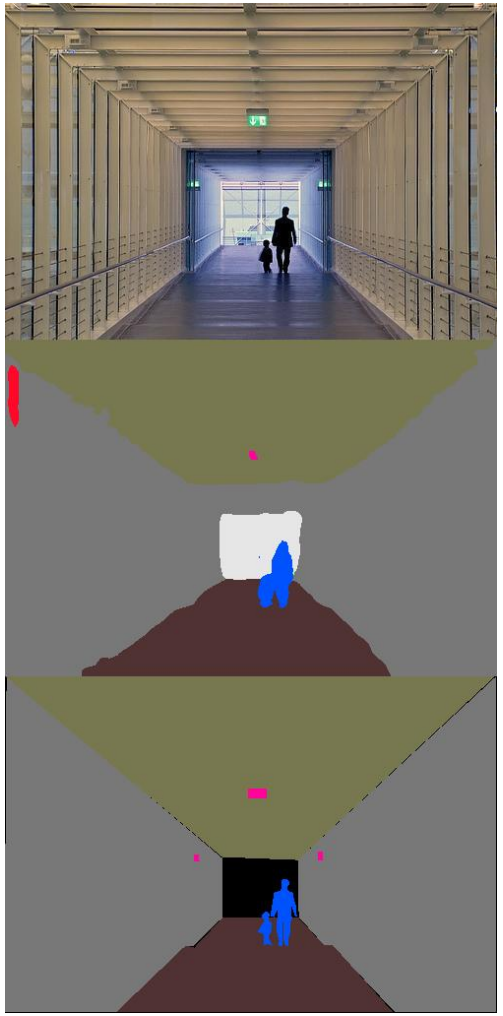
Vrsta napada	mIoU [%]	PA [%]	ASR [%]
Prirodno učenje	33.08	-	-
Linija, razina slike	31.80	75.35	39.40
Okvir, razina slike	31.92	75.47	35.25
Semantički (Grass), razina slike	29.20	69.74	30.97
Linija, razina primjerka	32.30	75.71	58.93
Semantički (Wall), razina primjerka	32.90	75.00	76.65

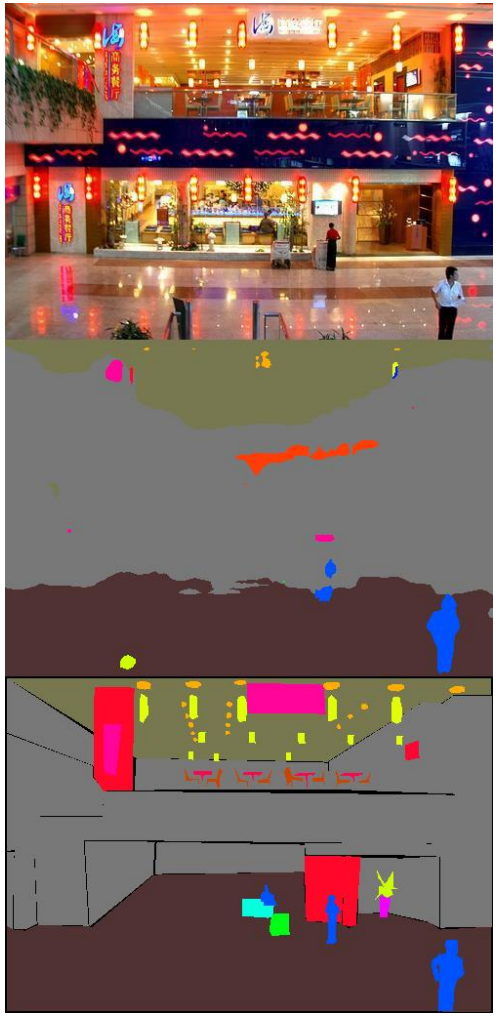
Primjeri, razina slike

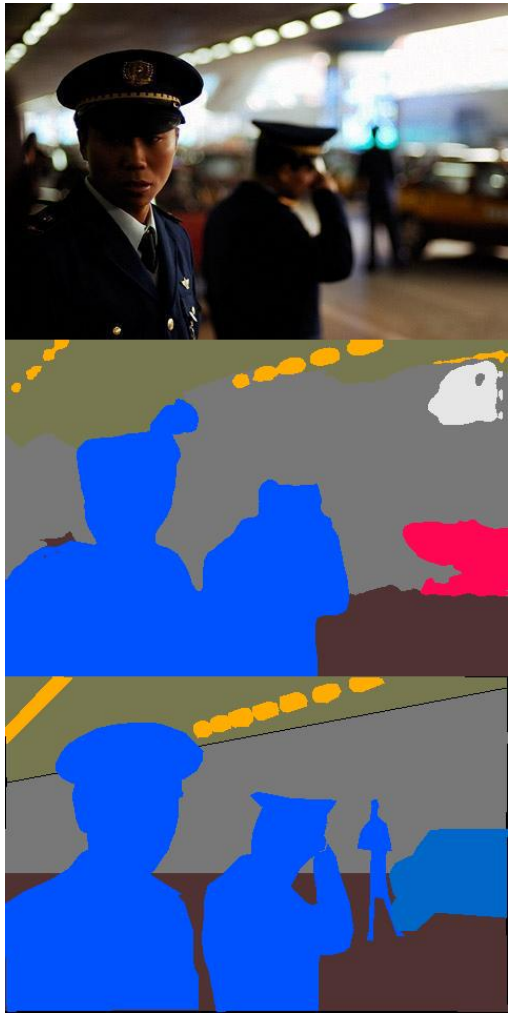




Primjeri, razina primjerka







Zaključak i budući rad

- modeli za semantičku segmentaciju također su ranjivi na napad trovanjem podataka
- posebno opasan napad na razini primjerka sa semantičkim okidačem
- mogući smjer budućeg rada:
  - implementacija dodatnih okidača
  - analiza kapaciteta arhitekture Single-Scale SwiftNet

Literatura

- Slajd 3, primjer semantičke segmentacije: preuzeto s https://vladlen.info/publications/feature-space-optimization-for-semantic-video-segmentation/
- Slajd 4, primjer zatrovanih podataka: preuzeto s https://mathco.com/blog/data-poisoning-and-its-impact-on-the-ai-ecosystem/
- Slajd 6, primjer podataka iz skupa ADE20k: preuzeto iz Zhou, Bolei, et al. "Semantic understanding of scenes through the ade20k dataset." *International Journal of Computer Vision* 127 (2019): 302-321.
- Slajd 7, arhitektura Single-Scale SwiftNet: preuzeto iz Oršić, Marin, and Siniša Šegvić. "Efficient semantic segmentation with pyramidal fusion." *Pattern Recognition* 110 (2021): 107611.

Diskusija