Trojanski napadi na modele za semantičku segmentaciju

Autor: Dominik Jambrović

Mentor: prof. dr. sc. Siniša Šegvić

Sadržaj

Uvod

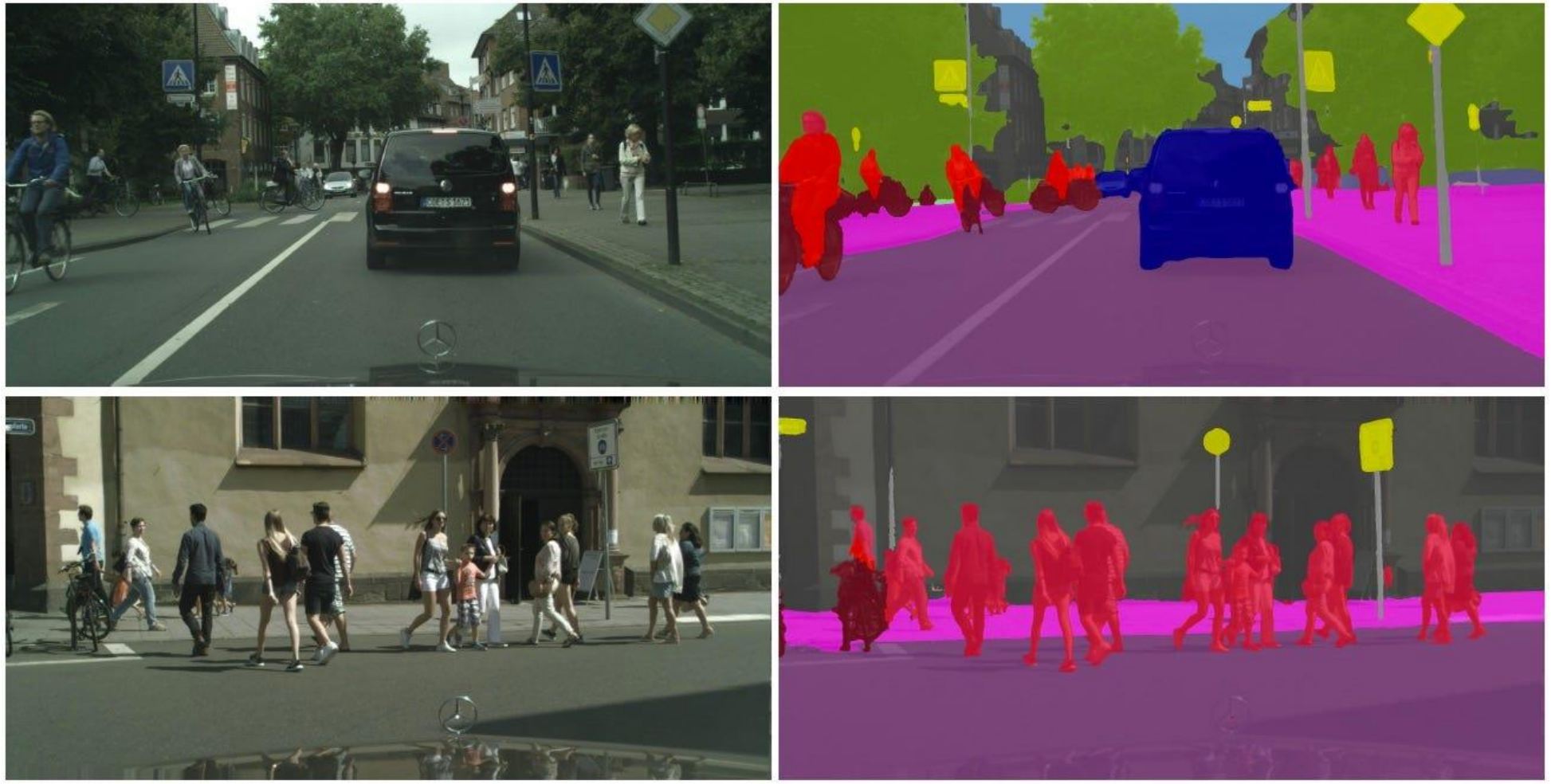
Sigurnost umjetne inteligencije:

- mnogo istraživanja za modele za klasifikaciju slike
- značajno manje zastupljena istraživanja za modele za **semantičku segmentaciju**

Potencijalni napadi:

- neprijateljski primjeri
- **trovanje podataka**

Semantička segmentacija



Trovanje podataka

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

Napad utemeljen na utjecaju

- dodavanje **okidača** na ulaznu sliku i izmjena oznaka piksela **razreda žrtve**
- uvjeti:
  - okidač ne smije prekrivati piksele razreda žrtve
  - okidač se u potpunosti mora nalaziti na pikselima koji pripadaju jednom razredu

Napad utemeljen na utjecaju

- inačice:
  - osnovni napad (Influencer Backdoor Attack - **IBA**)
  - napad utemeljen na najbližim susjedima (Nearest Neighbour Injection - **NNI**)
  - napad utemeljen na izmjeni oznaka nasumičnih piksela (Pixel Random Labeling – **PRL**)

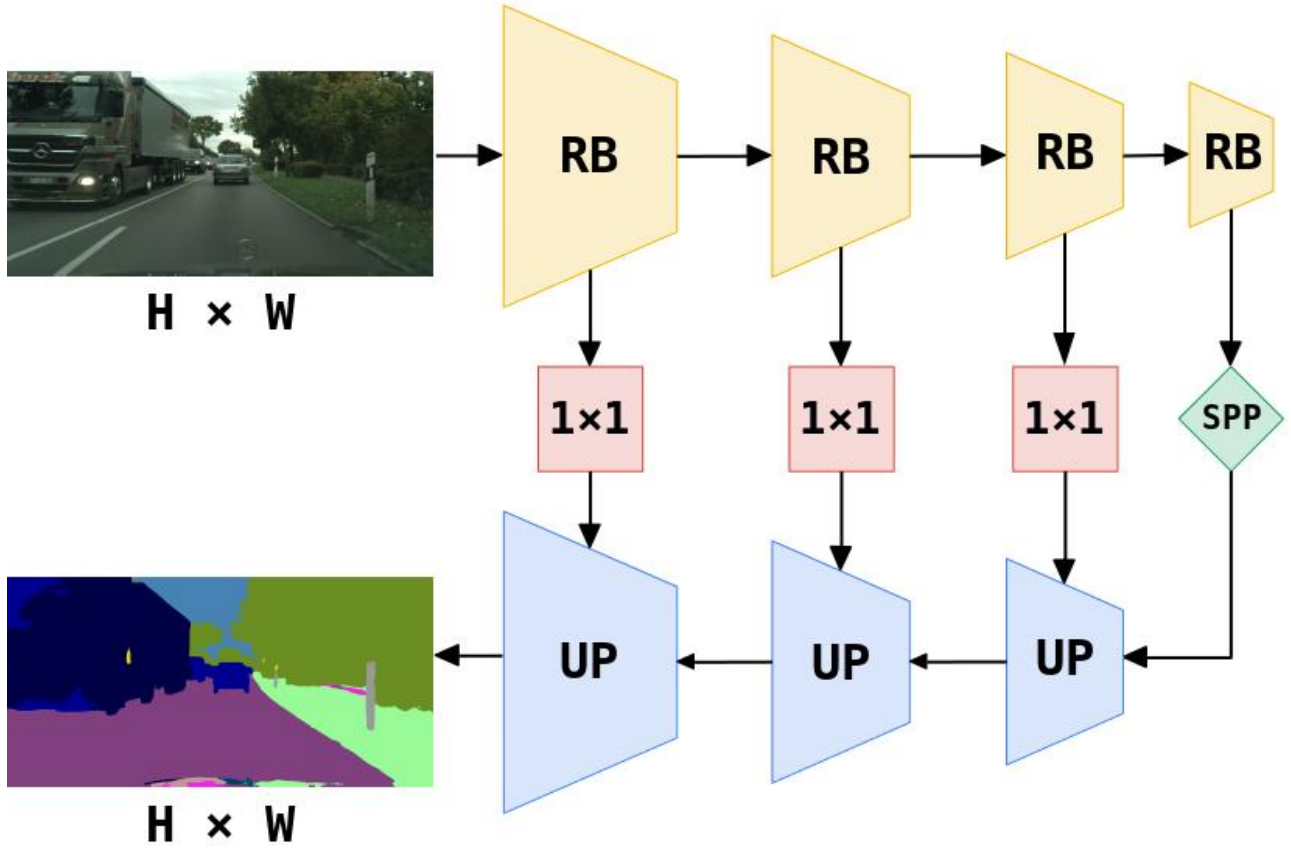
Eksperimenti

Skup podataka

- **Cityscapes**
  - 2975 slika u skupu za učenje
  - 500 slika u skupu za validaciju
  - 1525 slika u skupu za ispitivanje

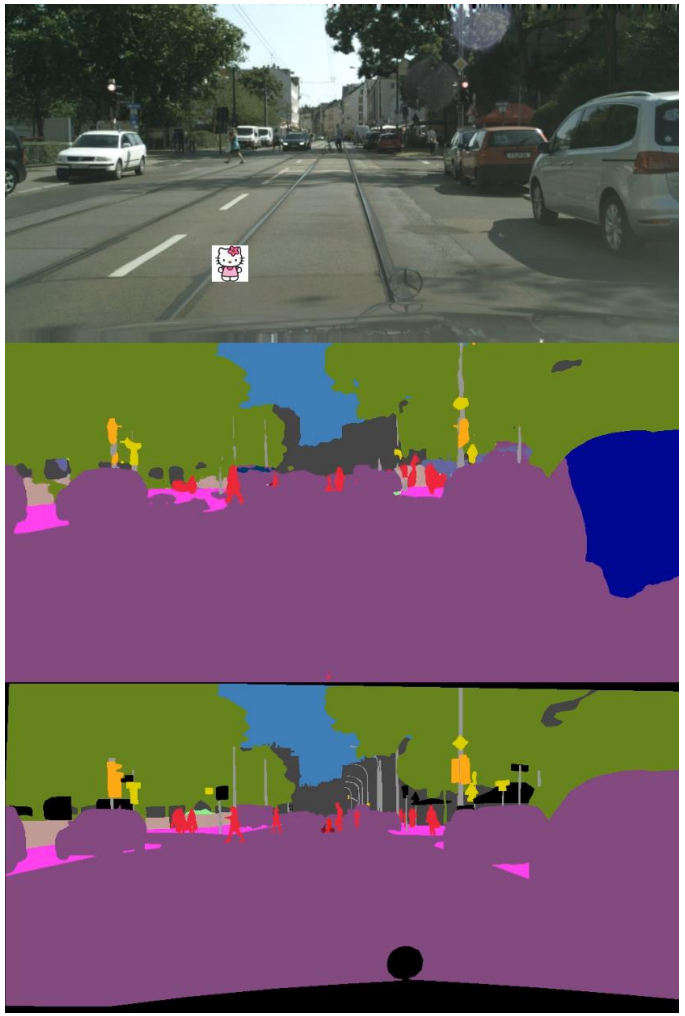
Arhitektura

- Jednorazinski SwiftNet

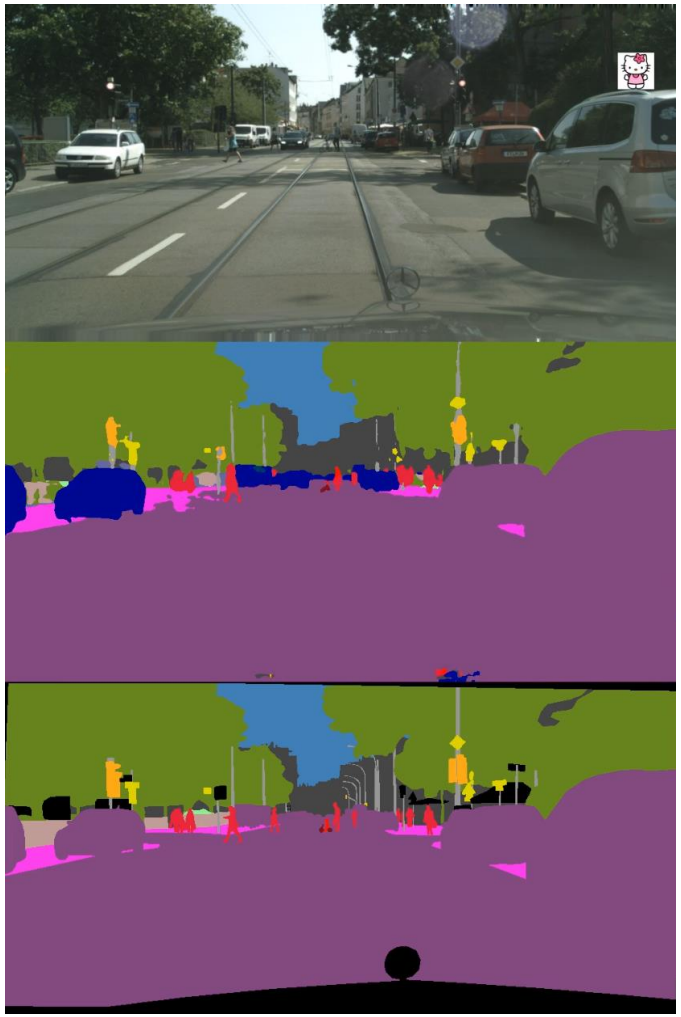


Vrsta napada	PR [%]	mIoU [%]	PA [%]	ASR [%]
Čisti podatci	-	**74.76**	**95.60**	-
IBA	1	71.20	95.04	24.22
IBA	3	71.40	95.03	44.08
IBA	5	70.86	95.01	46.82
IBA	10	70.63	94.87	53.73
IBA	15	70.36	94.79	56.38
IBA	20	70.44	94.72	58.36
NNI	1	70.74	95.00	42.14
NNI	3	70.74	94.99	57.83
NNI	5	71.69	95.03	58.60
NNI	10	70.85	94.98	61.59
NNI	15	70.95	94.90	66.06
NNI	20	70.43	94.74	**67.64**

Primjeri, IBA



Primjeri, NNI



Zaključak i budući rad

- modeli semantičke segmentacije također su ranjivi na trovanje podataka
- najuspješniji napad: NNI
- budući rad:
  - provođenje eksperimenata na drugim arhitekturama – npr. višerazinski SwiftNet
  - proučavanje obrane od prikazanih trojanskih napada na segmentacijske modele

Literatura

- Slajd 3, primjer semantičke segmentacije: preuzeto s https://vladlen.info/publications/feature-space-optimization-for-semantic-video-segmentation/
- Slajd 4, primjer trovanja podataka: preuzeto s https://mathco.com/blog/data-poisoning-and-its-impact-on-the-ai-ecosystem/
- Slajd 8, primjer podataka iz skupa Cityscapes: preuzeto iz Chen, Xinyun, et al. "Targeted backdoor attacks on deep learning systems using data poisoning.„
- Slajd 9, arhitektura jednorazinski SwiftNet: preuzeto iz Oršić, Marin, and Siniša Šegvić. "Efficient semantic segmentation with pyramidal fusion." *Pattern Recognition* 110 (2021): 107611.

Diskusija