SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Trojanski napadi na modele za semantičku segmentaciju

Dominik Jambrović
Voditelj: *prof. dr. sc. Siniša Šegvić*

Zagreb, svibanj 2024.

SADRŽAJ

1. Uvod **1**

2. Semantička segmentacija **2**

2.1. Općenito o semantičkoj segmentaciji 2

2.2. Mjere dobrote 3

2.3. Arhitektura SwiftNet 4

3. Napadi na modele strojnog učenja **5**

3.1. Trovanje podataka 5

3.2. Osnovni pristupi trovanju podataka 5

3.3. Napad utemeljen na utjecaju 6

4. Skupovi podataka **8**

4.1. Skup podataka ADE20k 8

5. Eksperimenti **9**

5.1. Postavke eksperimenata 9

5.2. Rezultati 10

6. Zaključak **11**

7. Literatura **12**

1. Uvod

U današnje vrijeme, duboki modeli primjenjuju se u brojnim aspektima svakodnevnog života. Pritom se pažnja primarno posvećuje performansama i konzistentnosti modela - želimo naučiti modele koji će na temelju naučenoga dobro generalizirati (npr. ispravno predviđati oznake za neviđene podatke).

Nažalost, sigurnosni aspekt često je zapostavljen. Kada govorimo o sigurnosti dubokih modela, važno je prvo identificirati moguće prijetnje. Neke od najčešćih su neprijateljski primjeri [5] i trovanje podataka [2]. Ove prijetnje zvat ćemo napadima na model. Napadi mogu imati različite ciljeve: od jednostavnog smanjenja performansi modela, pa sve do ugradnje stražnjih vrata u model.

Područje računalnog vida obuhvaća brojne zadatke. Jedan od najčešće rješavanih zadataka je klasifikacija – model na ulazu dobiva primjer te na izlazu treba dati jednu oznaku koja predstavlja razred u koji je ulaz svrstan. Brojna istraživanja fokusirala su se na primjenu napada na ovakve modele, kao i na potencijalne obrane (npr. robusno učenje modela). Naravno, ne možemo rješavati sve probleme iz stvarnog života koristeći obične klasifikatore.

Za neke zadatke (npr. „vid“ autonomnih vozila), potrebni su nam modeli koji će svaki piksel ulaza zasebno klasificirati u određeni razred. Ovaj zadatak zovemo semantička segmentacija [4]. Model na ulazu dobiva sliku, a na izlazu treba dati novu sliku (segmentaciju) gdje je svakom pikselu iz ulaza dodijeljena oznaka pripadnog razreda. Cilj ovog rada je reprodukcija osnovnih napada na odabrani model za semantičku segmentaciju po uzoru na rad [10]. Uz reprodukciju osnovnih napada, dodatan cilj je reproducirati i napad utemeljen na utjecaju po uzoru na rad [9].

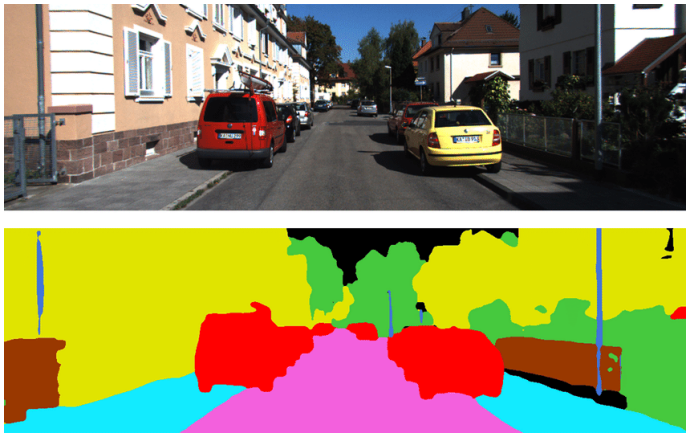
2. Semantička segmentacija

2.1. Općenito o semantičkoj segmentaciji

Semantička segmentacija [4] jedan je od zadataka iz područja računalnog vida. Za razliku od običnih klasifikatora koji na izlazu daju samo jednu oznaku, modeli za semantičku segmentaciju na izlazu daju segmentaciju - sliku istih dimenzija kao i ulazna slika. Pritom pojedini pikseli segmentacije odgovaraju predviđenom razredu za pripadni piksel ulazne slike.

Kod semantičke segmentacije postoje dva glavna problema. Prvi od njih je već spomenuta činjenica da model na izlazu mora dati sliku. Zapravo, modeli na izlazu najčešće daju K slika, pri čemu K označava ukupan broj razreda. Svaka slika tada odgovara vjerojatnosti da pojedini pikseli pripadaju određenom razredu. Možemo reći da na izlazu dobivamo distribuciju vjerojatnosti za svaki piksel. Ako bismo ovaj zadatak rješavali "klasičnom" arhitekturom (npr. potpuno povezanom ili konvolucijskom arhitekturom), broj parametara bio bi prevelik - model bi zauzimao previše memorije i bilo bi ga teško naučiti. Zbog toga, modeli za semantičku segmentaciju često su arhitekture koder-dekoder [1]. Koder ulaznu sliku pretvara u semantički bogatu latentnu (skrivenu) reprezentaciju manjih dimenzija. Dekoder na temelju latentne reprezentacije naduzorkovanjem stvara segmentaciju prikladnih dimenzija.

Drugi problem kod semantičke segmentacije je veličina receptivnog polja. Kako bi model mogao donijeti ispravnu odluku za pojedine piksele, veoma je važno da na odluku utječe velik broj susjednih piksela. Ako se, na primjer, na slici nalazi kamion, lako je moguće da on prekriva velik dio slike. Modeli koji imaju malo receptivno polje teško bi ispravno klasificirali sve piksele kamiona. Za uklanjanje ovog problema predloženo je mnogo rješenja. Jedno od njih je korištenje sloja piramidalnog sažimanja [6]. Ovaj sloj provodi sažimanje ulaznih mapa značajki na temelju rešetki različitih dimenzija, time omogućavajući modelu da važne informacije dobiva iz šireg konteksta.



Slika 2.1: Primjer ulazne slike i predviđanja modela za semantičku segmentaciju. Preuzeto iz [8].

Na slici 2.1 možemo vidjeti primjer ulazne slike (gornji dio slike), kao i predviđanja modela tj. segmentaciju za tu sliku (donji dio slike). Na segmentaciji su zasebni razredi označeni različitim bojama.

2.2. Mjere dobrote

Pošto je izlaz modela za semantičku segmentaciju niz oznaka za pojedine piksele, na ovakav model ne možemo jednostavno primijeniti klasične mjere dobrote poput preciznosti, odziva i F1-mjere. U okviru ovog rada, za mjerenje performansi koristilo smo srednji omjer presjeka i unije (engl. mean intersection over union – mIoU), kao i točnost po pikselima (engl. pixel accuracy – PA).

Omjer presjeka i unije možemo definirati jednadžbom:

$$IoU = \frac{N_{predicted} \cap N_{true}}{N_{predicted} \cup N_{true}} \tag{2.1}$$

Pritom $N_{predicted}$ predstavlja broj piksela za koje je model predvidio određeni razred, a N_{true} predstavlja broj piksela koji su stvarno iz tog razreda. Srednji omjer presjeka i unije tada definiramo kao prosjek presjeka i unije za pojedine razrede.

Točnost po pikselima možemo definirati jednadžbom:

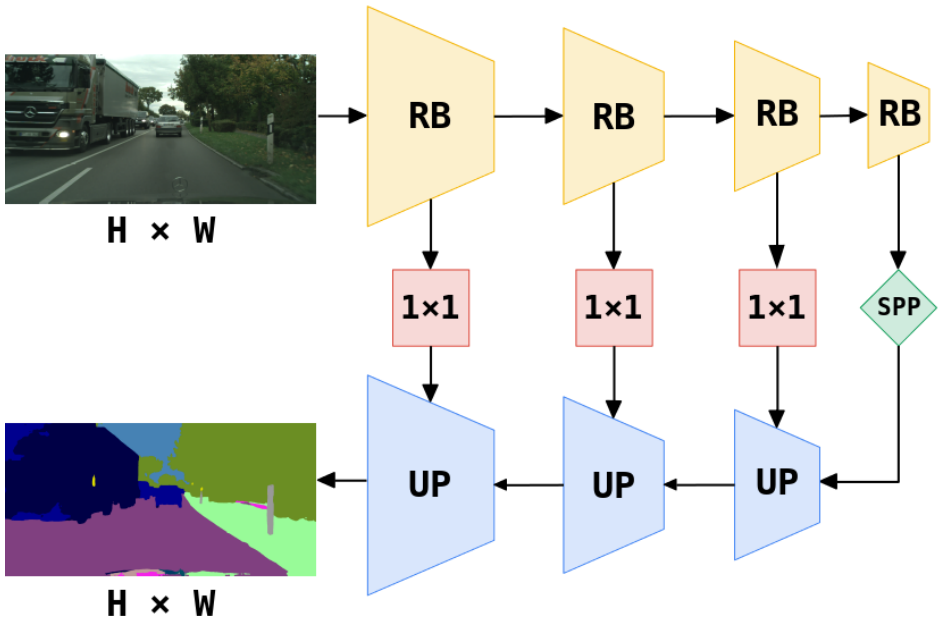
$$PA = \frac{N_{correct}}{N_{total}} \tag{2.2}$$

Pritom $N_{correct}$ predstavlja broj točno klasificiranih piksela, dok N_{total} predstavlja ukupan broj piksela. Osim ove dvije mjere, za mjerenje performansi modela za semantičku segmentaciju često se koristi i Dice koeficijent.

2.3. Arhitektura SwiftNet

Sve eksperimente provodili smo na arhitekturi Single-Scale SwiftNet [11]. Ovaj model sastoji se od kodera, sloja prostornog piramidalnog sažimanja i dekodera. Koder ulaznu sliku pretvara u semantički bogatu latentnu (skrivenu) reprezentaciju manjih dimenzija. Nakon njega slijedi sloj prostornog piramidalnog sažimanja koji provodi sažimanje ulaznih mapa značajki na temelju rešetki različitih dimenzija. Rezultati sažimanja konkateniraju se te prosljeđuju dekoderu. Konačno, dekoder na temelju latentne reprezentacije naduzorkovanjem stvara segmentaciju prikladnih dimenzija. Dekoderski blokovi pritom koriste bilinearnu interpolaciju, kao i konvoluciju s jezgrom veličine 3x3. Slojevi kodera i dekodera povezani su lateralnim vezama, ostvarenim kao konvolucija s jezgrom veličine 1x1 te potom zbrajanje.

Kao okosnica koderskih blokova koristi se model ResNet18 [7] prednaučen na skupu ImageNet [3]. ResNet18 primjerak je rezidualne arhitekture koju karakterizira postojanje rezidualnih blokova. Glavna značajka rezidualnih blokova je postojanje preskočnih veza - ulaz u određeni sloj direktno se zbraja s njegovim izlazom.



Slika 2.2: Arhitektura Single-Scale SwiftNet. Preuzeto iz [11].

Na slici 2.2 možemo vidjeti dijagram arhitekture Single-Scale SwiftNet. Koderski blokovi predstavljeni su narančastim trapezima, dekoderski blokovi plavim trapezima, sloj prostornog piramidalnog sažimanja zelenim rombom, a lateralne veze crvenim kvadratima. Model na ulazu dobiva sliku, a na izlazu daje segmentaciju istih dimenzija.

3. Napadi na modele strojnog učenja

3.1. Trovanje podataka

Trovanje podataka vrsta je napada kod kojeg napadač ima pristup skupu za učenje i u njega ubacuje zatrovane podatke. Zatrovani podatci najčešće su parovi izmijenjenih slika i proizvoljno odabranih oznaka. Pritom se na originalne slike većinom dodaje okidač (npr. nekoliko bijelih piksela u kutu slike). Okidač može biti ograničen na mali dio slike, ali i dodan kao uzorak preko cijele slike.

Najjednostavniji cilj trovanja podataka je da naučeni model ima lošije performanse. Ipak, napadač od takvog trovanja nema puno koristi. Puno opasniji cilj je ugrađivanje stražnjih vrata u model. Ako model tijekom učenja poveže pojavu okidača s klasifikacijom u određeni razred, napadač može manipulirati predviđanja modela ugrađivanjem okidača u neviđene slike. Ovakvu vrstu napada zvat ćemo trojanski napad.

3.2. Osnovni pristupi trovanju podataka

Kada govorimo o trovanju podataka za semantičku segmentaciju, možemo napraviti dvije osnovne podjele pristupa [10]. Prva podjela vezana je uz odabir okidača. Okidač može biti nesemantički ili semantički.

Slika 3.1: Primjer nesemantičkog (lijevo) i semantičkog (desno) okidača.

Na slici 3.1 možemo vidjeti primjer dodavanja nesemantičkog i semantičkog okidača na sliku. Lijeva slika predstavlja nesemantički okidač - na vrh slike jednostavno je dodana crna linija visine 8 piksela. Naravno, nesemantički okidač može biti i drugačijeg oblika - okidač može biti i okvir oko cijele slike ili neki uzorak. Desna slika predstavlja semantički okidač. U ovom slučaju na sliku ne dodajemo nikakav uzorak, već kao zatrovane podatke biramo slike koje sadrže piksele iz određenog razreda. Konkretno, slika je označena kao zatrovana jer na sebi sadrži piksele iz razreda "trava".

Druga podjela vezana je uz odabir načina izmjene oznaka. Oznake možemo mijenjati na razini slike ili na razini primjerka.

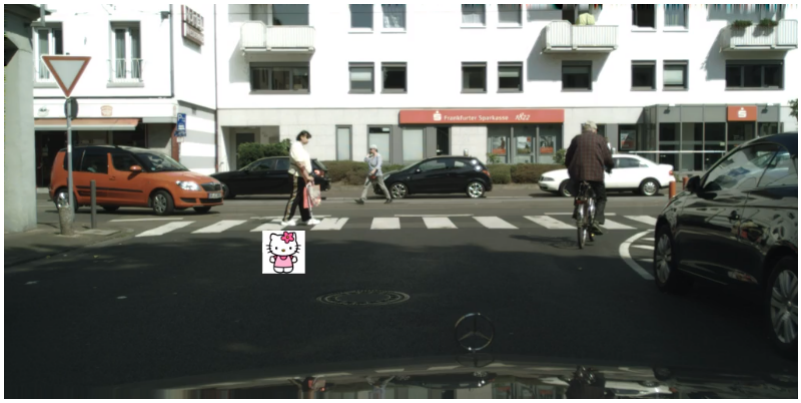
Slika 3.2: Primjer izmjene oznaka na razini slike (centar) i primjerka (desno).

Na slici 3.2 možemo vidjeti primjer izmjene oznaka na razini slike i primjerka. Lijeva slika predstavlja nepromijenjeni ulaz. Centralna slika predstavlja izmjenu oznaka na razini slike. Možemo vidjeti da nove, zatrovane oznake nisu povezane s ulazom. Kod ovog pristupa, nasumično se odaberu oznake jednog ulaza koje će onda biti dodijeljene kao oznake za sve zatrovane podatke. Desna slika predstavlja izmjenu oznaka na razini primjerka. Kod ovog pristupa, oznake koje pripadaju jednom razredu izmijene se tako da sada pripadaju nekom drugom razredu. Konkretno, na prikazanoj slici su izmijenjene oznake piksela koji pripadaju razredu "čovjek". Nakon izmjene, ti pikseli pripadaju razredu "palma". Vidimo da je ovaj pristup puno suptilniji, a time i opasniji - čak i ako se provodi ručna provjera, puno je teže uočiti ovu vrstu izmjene oznaka.

3.3. Napad utemeljen na utjecaju

Napad utemeljen na utjecaju [9] (engl. influencer backdoor attack - IBA) naprednija je inačica napada na modele za semantičku segmentaciju. Prema prethodnim podjelama, ovaj napad mogli bismo smatrati nesemantičkim napadom na razini primjerka. Postoji nekoliko inačica ovog napada, a razlikuju se primarno u načinu određivanja pozicije okidača. U svim inačicama, oznake se mijenjaju na razini primjerka, a na ulazne slike se dodaje nesemantički okidač.

Osnovna inačica ovog napada uvodi nekoliko uvjeta za poziciju okidača na ulaznoj slici. Glavni uvjet je da okidač ne smije prekrivati piksele iz razreda žrtve (razreda za koji mijenjamo oznake). Dodatno, okidač se u potpunosti mora nalaziti na pikselima koji pripadaju jednom razredu. Na primjer, ako je okidač dimenzija 50x50, na ulaznoj slici moramo pronaći područje iste veličine unutar kojega svi pikseli pripadaju samo jednom razredu. U slučaju da mogućih pozicija okidača ima više, kod osnovne inačice nasumično odabiremo jednu od njih.



Slika 3.3: Primjer dodavanja okidača kod napada utemeljenog na utjecaju.

Na slici 3.3 možemo vidjeti primjer dodavanja okidača na ulaznu sliku. Konkretno, na sliku je na poziciju gdje svi pikseli pripadaju jednom razredu (razredu "cesta") dodan poznati Hello Kitty okidač.

Druga inačica ovog napada zasniva se na najbližim susjedima (engl. nearest neighbour injection - NNI). Konkretno, ako se pri određivanju pozicije okidača ustanovi da postoji više mogućih pozicija, odabire se ona koja je najbliža razredu žrtve. Očekujemo da će kod ove inačice napada stopa uspješnosti napada biti viša, ali je pritom važno istaknuti da ova inačica nije potpuno realistična. U stvarnosti napadač općenito ne može pozicionirati okidač tako da je on što bliže žrtvi.

Posljednja inačica napada zasniva se na označavanju nasumičnih piksela (engl. pixel random labeling - PRL). Kod ove inačice, pozicija okidača ponovno se odabire nasumično. Uz ovo, dodatno se mijenjaju oznake. Konkretno, nasumično se odabere određen broj piksela koji ne pripadaju razredu žrtve te se isti označe nasumično odabranim razredima koji su prisutni u slici. Na primjer, za nasumično odabran piksel koji pripada razredu "čovjek", nova oznaka može biti razred "trava". Prema autorima rada [9], ovakva izmjena oznaka potiče model da koristi informacije iz šireg konteksta, time povećavajući mjeru uspješnosti napada za slučajeve kada je okidač jako udaljen od piksela iz razreda žrtve.

4. Skupovi podataka

4.1. Skup podataka ADE20k

Skup ADE20k [12] sastoji se od otprilike 25 000 označenih slika iz stvarnog života. Podijeljen je na 20 210 slika u skupu za učenje, 2000 slika u skupu za validaciju te 3000 slika u skupu za ispitivanje. Svaki piksel može pripadati jednom od ukupno 150 razreda. Ulazne slike, kao i pripadne segmentacije, varirajućih su dimenzija.

Slika 4.1: Primjeri slika i oznaka iz skupa ADE20k. Preuzeto iz [12].

5. Eksperimenti

5.1. Postavke eksperimenata

Eksperimente smo provodili učenjem modela na prirodnom skupu podataka, kao i na nekolicini zatrovanih skupova podataka. Pritom su za stvaranje zatrovanih skupova podataka korišteni razni okidači, kao i razne izmjene segmentacijskih oznaka. Okidače općenito možemo podijeliti na nesemantičke i semantičke. Nesemantički okidači podrazumijevaju izmjenu ulazne slike (npr. dodavanje nekoliko bijelih piksela u kut slike). S druge strane, semantički okidači ne mijenjaju ulaznu sliku. Umjesto toga, kao zatrovani primjeri se uzimaju ulazne slike koje sadrže piksele koji pripadaju određenom razredu (npr. slike s barem jednim pikselom iz razreda Wall).

Kao nesemantičke okidače, koristili smo crnu liniju širine 8 piksela dodanu na vrh ulazne slike, kao i crni okvir širine 8 piksela dodan na rub ulazne slike. Kao semantičke okidače, birali smo slike koje sadrže barem jedan piksel iz razreda Grass ili iz razreda Wall. Kada govorimo o izvedbi trovanja, radimo podjelu na napade na razini slike odnosno primjerka. Za napad na razini slike, nasumično smo odabrali oznaku koja sadrži barem jedan piksel iz razreda Road. Svakom zatrovanom primjeru dodijelili smo oznaku odabrane slike, pritom provodeći prikladno skaliranje zbog varirajućih dimenzija primjera. Kod napada na razini primjerka, segmentaciju zatrovanih primjera izmijenili smo tako da pikseli koji pripadaju razredu Person sada pripadaju razredu Palm.

Svi modeli učeni su 200 epoha, a pri učenju su računate sljedeće mjere dobrote: srednji omjer presjeka i unije (engl. mean intersection over union – mIoU), točnost po pikselima (engl. pixel accuracy – PA) te mjera uspješnosti napada (engl. attack success rate – ASR). U svim eksperimentima sa zatrovanim podatcima, stopa trovanja iznosila je otprilike 10%.

5.2. Rezultati

U tablici 5.1, stupac *Vrsta napada* predstavlja vrstu napada korištenu za trovanje skupa za učenje te skupa za validaciju. Stupac *mIoU* predstavlja srednji omjer presjeka i unije, dok stupac *PA* predstavlja točnost po pikselima. Obje mjere dobrote mjerene su na skupu za validaciju. Stupac *ASR* predstavlja mjeru uspješnosti napada mjerenu na zatrovanom skupu za validaciju.

Tablica 5.1: Performanse modela učenih na različitim skupovima podataka.

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

Kao što možemo vidjeti u tablici 5.1, model učen na prirodnom skupu očekivano postiže najviši mIoU. Možemo vidjeti da korištenje zatrovanog skupa smanjuje mIoU modela na skupu za validaciju za otprilike 1%. Jedina iznimka je korištenje skupa podataka zatrovanog semantičkim napadom na razini slike. Kod napada na razini slike, najvišu stopu uspješnosti napada postiže model s linijskim okidačem. Model sa semantičkim okidačem (okidač je pojava barem jednog piksela iz razreda Grass) postiže najlošije rezultate po svim mjerama.

Kod napada na razini primjerka, model sa semantičkim okidačem (okidač su slike s barem jednim pikselom iz razreda Wall) postiže znatno višu stopu uspješnosti napada u usporedbi s modelom s linijskim okidačem.

6. Zaključak

Kao što vidimo, modeli semantičke segmentacije ranjivi su na napad trovanjem skupa podataka kao i obični klasifikatori. Posebno se opasnim čini napad na razini primjerka sa semantičkim okidačem – kod ovog napada, napadač jedino mijenja označeni razred za piksele iz jednog razreda, dok je ostatak segmentacijske oznake jednak kao i prije.

7. Literatura

[1] Vijay Badrinarayanan, Alex Kendall, i Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[2] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, i Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, i Li Fei-Fei. Imagenet: A large-scale hierarchical image database. U *2009 IEEE conference on computer vision and pattern recognition*, stranice 248–255. Ieee, 2009.

[4] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, i Jose Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*, 2017.

[5] Ian J Goodfellow, Jonathon Shlens, i Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, i Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, i Jian Sun. Deep residual learning for image recognition. U *Proceedings of the IEEE conference on computer vision and pattern recognition*, stranice 770–778, 2016.

[8] Jongmin Jeong, Tae Sung Yoon, i Jin Bae Park. Towards a meaningful 3d map using a 3d lidar and a camera. *Sensors*, 18(8):2571, 2018.

[9] Haoheng Lan, Jindong Gu, Philip Torr, i Hengshuang Zhao. Influencer backdoor attack on semantic segmentation. *arXiv preprint arXiv:2303.12054*, 2023.

[10] Yiming Li, Yanjie Li, Yalei Lv, Yong Jiang, i Shu-Tao Xia. Hidden backdoor attack against semantic segmentation models. *arXiv preprint arXiv:2103.04038*, 2021.

[11] Marin Oršić i Siniša Šegvić. Efficient semantic segmentation with pyramidal fusion. *Pattern Recognition*, 110:107611, 2021.

[12] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, i Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.

Trojanski napadi na modele za semantičku segmentaciju

Sažetak

Proučavanje trojanskih napada na modele semantičke segmentacije važno je za razumijevanje i osiguravanje sigurnosti modela. Razmatramo prethodne pristupe trovanja te ih uspoređujemo s napadom utemeljenim na utjecaju koji je objavljen ove godine. Učinak metoda trovanja vrednujemo s obzirom na izvedbu odgovarajućih modela učenih na čistim podatcima. Eksperimenti pokazuju da napad utemeljen na utjecaju nadmašuje prethodne pristupe.

Ključne riječi: semantička segmentacija, napadi na modele strojnog učenja, zatrovani podatci, nesemantički napad, semantički napad, napad utemeljen na utjecaju

Trojan attacks on models for semantic segmentation

Abstract

Studying Trojan attacks on semantic segmentation models is important for understanding and ensuring security of machine learning models. We review previous poisoning approaches and compare them to the influencer backdoor attack published this year. We evaluate the effect of the poisoning methods with regard to the performance of the corresponding models trained on clean data. Experiments show that the influencer backdoor attack outperforms previous approaches.

Keywords: semantic segmentation, attacks on machine learning models, data poisoning, nonsemantic attack, semantic attack, influencer backdoor attack