SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Trojanski napadi na modele za semantičku segmentaciju

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

Zagreb, ožujak 2024.

SADRŽAJ

1. Uvod

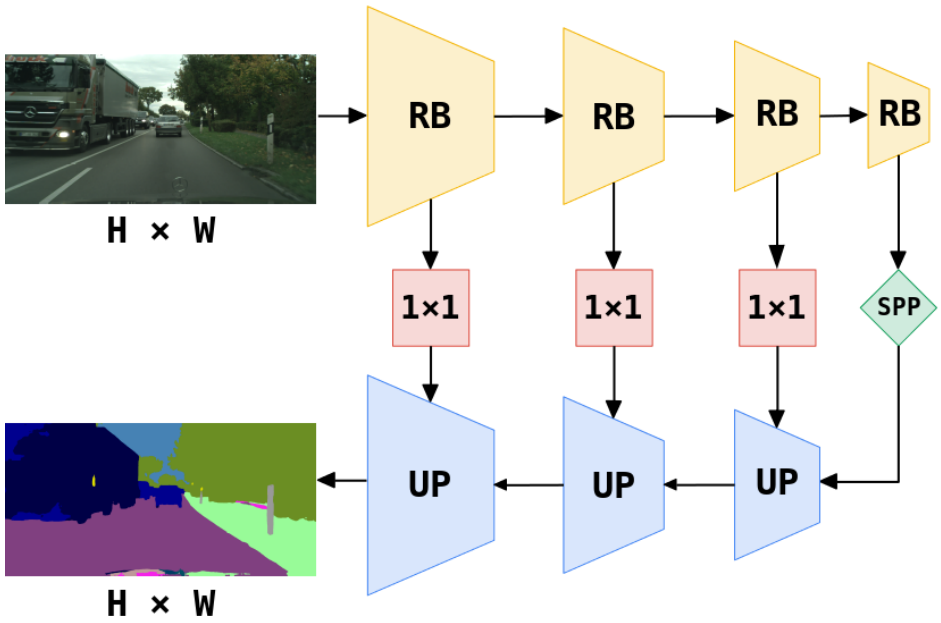
U današnje vrijeme, duboki modeli primjenjuju se u brojnim aspektima svakodnevnog života. Pritom se pažnja primarno posvećuje performansama i konzistentnosti modela. Nažalost, sigurnosni aspekt često je zapostavljen. Kada govorimo o sigurnosti dubokih modela, važno je prvo identificirati moguće prijetnje. Neke od najčešćih su neprijateljski primjeri [4] i zatrovani podatci [1].

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. 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 [3]. 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 napada na odabrani model za semantičku segmentaciju po uzoru na rad [6].

2. Semantička segmentacija

2.1. Arhitektura SwiftNet

Sve eksperimente provodili smo na arhitekturi Single-Scale SwiftNet [7]. Ovaj model sastoji se od kodera koji provodi poduzorkovanje, sloja prostornog piramidalnog sažimanja i dekodera koji provodi naduzorkovanje. Pritom se za okosnicu kodera koristi model ResNet18 [5] prednaučen na skupu ImageNet [2]. Slojevi kodera i dekodera povezani su lateralnim vezama (ljestvičasta arhitektura).



Slika 2.1: Arhitektura Single-Scale SwiftNet. Preuzeto iz [7].

3. Napadi na modele strojnog učenja

3.1. Neprijateljski primjeri

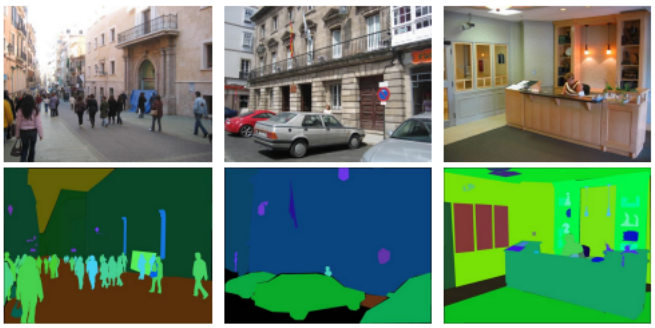
Neprijateljski primjeri većinom su slike na koje je nadodan šum (najčešće konstruiran gradijentnim metodama). Cilj ove vrste napada je izmijeniti odluku već naučenog modela i time izbjeći ispravnu klasifikaciju.

3.2. Zatrovani podatci

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). Cilj ove vrste napada je da se model nauči na zatrovanim podatcima te da napadač time ugradi stražnja vrata u model.

4. Skup podataka ADE20k

Skup ADE20k [8] 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 [8].

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] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

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

[8] 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 napada na modele semantičke segmentacije važno je za razumijevanje i postizanje sigurnosti modela. Proučavamo osnovne načine trovanja modela za semantičku segmentaciju (nesemantički napad, semantički napad). Istražujemo napredniji način trovanja (napad temeljen na utjecaju i nadogradnje na isti). Evaluiramo performanse naučenih zatrovanih modela i uspoređujemo s performansama modela učenog na prirodnim podatcima. Evaluacija pokazuje da najbolje performanse imaju nadogradnje napada temeljenog na utjecaju.

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

Trojan attacks on models for semantic segmentation

Abstract

Studying attacks on semantic segmentation models is important for understanding and achieving security of machine learning models. We study the basic implementations of data poisoning for semantic segmentation (nonsemantic attack, semantic attack). We explore a more advanced way of data poisoning (influencer backdoor attack and its upgrades). We evaluate the performance of the models trained on poisoned datasets and compare them with the performance of the model trained on natural data. The evaluation shows that the influencer backdoor attack and its upgrades are the most successful in poisoning the models.

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