SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Trojanski napadi na modele za semantičku segmentaciju

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

Zagreb, lipanj 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: 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 [8] i trovanje podataka [3]. 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 slike: model na ulazu dobiva primjer te na izlazu treba predvidjeti razred koji odgovara ulazu. Brojna istraživanja razmatraju napade na ovakve modele [8], kao i potencijalne obrane (npr. robusno učenje modela [14]). 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 nazivamo semantička segmentacija [7]: model na ulazu dobiva sliku, a na izlazu treba dati novu sliku (segmentaciju) gdje je svakom pikselu dodijeljen pripadni semantički razred. Cilj ovog rada je reprodukcija osnovnih napada na odabrani model za semantičku segmentaciju po uzoru na rad [13]. Uz reprodukciju osnovnih napada, dodatan cilj je reproducirati i napad utemeljen na utjecaju po uzoru na rad [12].

2. Semantička segmentacija

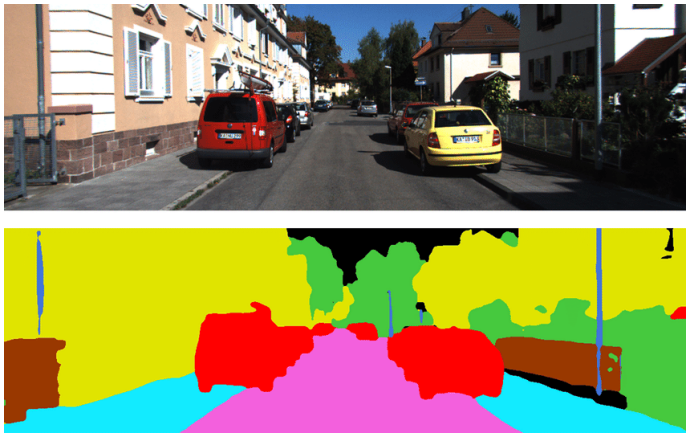
2.1. Općenito o semantičkoj segmentaciji

Semantička segmentacija [7] jedan je od zadataka računalnog vida. Za razliku od običnih klasifikatora koji na izlazu daju samo jednu oznaku, modeli za semantičku segmentaciju na izlazu daju sliku istih dimenzija kao i ulazna slika. Pritom pojedini pikseli izlaza 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. konvolucijskom arhitekturom), računska složenost bila bi prevelika - učenje modela zahtijevalo bi previše memorije, a i vremena. Zbog toga, modeli za semantičku segmentaciju često imaju arhitekturu 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. Prostorno piramidalno sažimanje [9] omogućava kasnijim konvolucijama da crpe informacije iz šireg konteksta. Glavna ideja je sažeti ulazne značajke preko rešetki različitih dimenzija, te tako dobivene poduzorkovane reprezentacije bilinearno razvući na početnu rezoluciju.

Slika 2.1 prikazuje 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.



Slika 2.1: Primjer ulazne slike i predviđanja modela za semantičku segmentaciju. Preuzeto iz [11].

2.2. Mjere dobrote

Pošto je izlaz modela za semantičku segmentaciju matrica oznaka pojedinih piksela, na ovakav model ne možemo jednostavno primijeniti klasične mjere dobrote poput preciznosti, odziva i F1-mjere. U okviru ovog rada, dobrotu modela mjerimo srednjim omjerom presjeka i unije (engl. mean intersection over union – mIoU) te prosječnom točnošću preko svih piksela (engl. pixel accuracy – PA).

Omjer presjeka i unije možemo definirati jednadžbom [17]:

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

Pritom $N_{predicted}$ predstavlja broj piksela za koje je model predvidio promatrani semantički razred, a N_{true} predstavlja broj piksela koji stvarno pripadaju tom razredu. Srednji omjer presjeka i unije tada definiramo kao prosječni IoU preko svih razreda.

Točnost po pikselima možemo definirati jednadžbom:

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

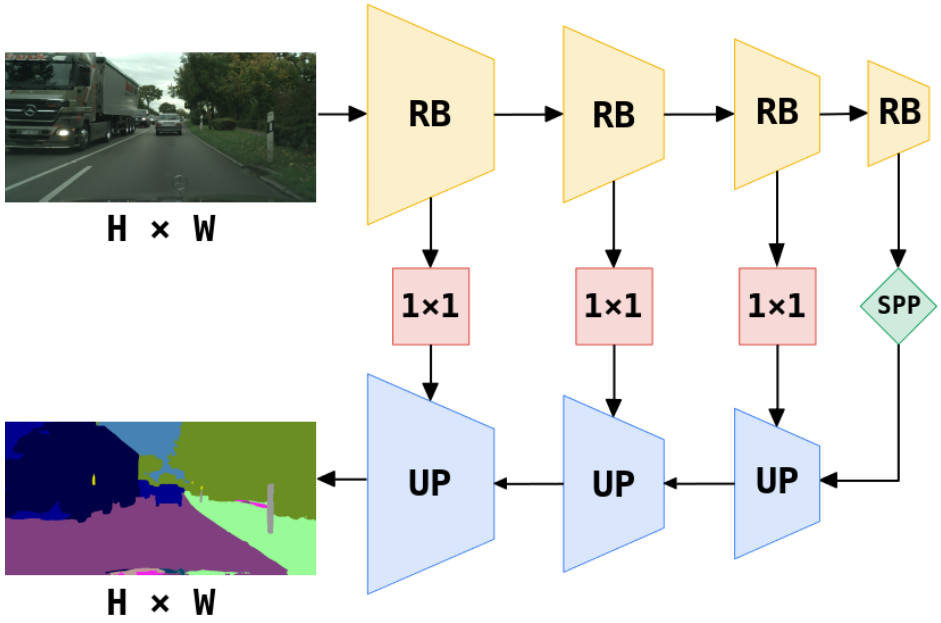
Pritom $N_{correct}$ predstavlja broj točno klasificiranih piksela, dok N_{total} predstavlja ukupan broj piksela. Osim ove dvije mjere, za mjerenje performansi modela za semantičku segmentaciju često se koristi i Diceov koeficijent [6].

2.3. Arhitektura SwiftNet

Sve eksperimente provodili smo na jednorazinskoj inačici arhitekture SwiftNet [15]. Ova arhitektura 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 povećava receptivno polje kasnijih konvolucija. Rezultati sažimanja konkateniraju se te prosljeđuju dekoderu. Konačno, dekoder ljestvičastim naduzorkovanjem stvara izlazne značajke prikladnih dimenzija. Dekoderski blokovi pritom koriste bilinearnu interpolaciju, kao i konvoluciju s jezgrom veličine 3x3. Slojevi kodera i dekodera povezani su lateralnim vezama, ostvarenim konvolucijom s jezgrom veličine 1x1 te zbrajanjem.

Koderski blokovi obično se izvode okosnicom ResNet18 [10] koju učimo iz inicijalizacije dobivene učenjem na skupu ImageNet [5]. ResNet18 primjerak je rezidualne arhitekture koju karakterizira postojanje rezidualnih blokova. Glavna značajka rezidualnih blokova je postojanje preskočnih veza koje podrazumijevaju zbrajanje izlaza konvolucijskih slojeva s njihovim ulazom [10].

Slika 2.2 prikazuje dijagram jednorazinske inačice arhitekture 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.



Slika 2.2: Jednorazinska inačica arhitekture SwiftNet. Preuzeto iz [15].

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 pažljivo odabranih oznaka. Izvorne slike za učenje mijenjaju se na način da im se doda okidač [3] (npr. nekoliko bijelih piksela u kutu slike). Okidač može biti ograničen na mali dio slike, ali i protezati se preko cijele slike.

Najjednostavniji cilj trovanja podataka je pogoršanje dobrote naučenih modela. 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 ispitne slike. Ovakvu vrstu napada zvat ćemo trojanski napad.

3.2. Osnovni pristupi trovanju segmentacijskih podataka

Možemo napraviti dvije osnovne podjele pristupa trovanju podataka za semantičku segmentaciju [13]. Prva podjela vezana je uz odabir okidača. Okidač može biti nesemantički ili semantički.

Slika 3.1 prikazuje 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 `grass`.

Slika 3.1: Primjer nesemantičkog (lijevo) i semantičkog (desno) okidača. Crna linija na vrhu lijeve slike predstavlja nesemantički okidač. Semantički okidač ne dodaje nikakav uzorak na desnu sliku, već predstavlja odabir slike koja sadrži piksele iz razreda `grass`.

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

Slika 3.2 prikazuje 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 `person`. Nakon izmjene, ti pikseli pripadaju razredu `palm`. Vidimo da je ovaj pristup puno suptilniji, a time i opasniji - čak i ako se provodi ručna provjera, puno je teže uočiti izmjenu oznaka na razini primjerka.

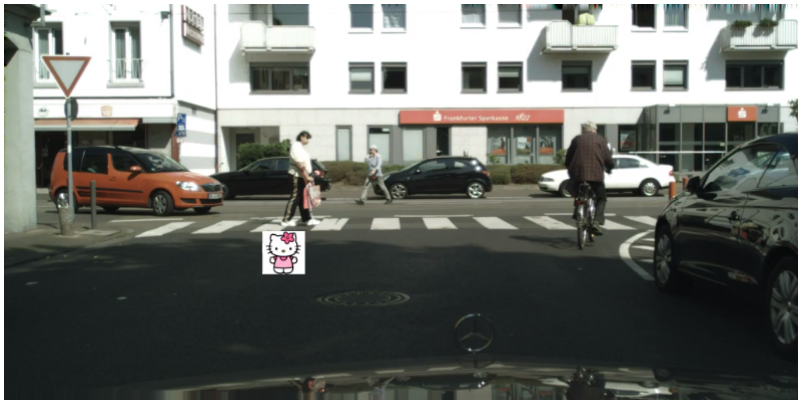
Slika 3.2: Primjer izmjene oznaka na razini slike (centar) i primjerka (desno). Lijeva slika predstavlja nepromijenjeni ulaz, centralna slika predstavlja izmjenu oznaka na razini slike, a desna slika predstavlja izmjenu oznaka na razini primjerka.

3.3. Napad utemeljen na utjecaju

Napad utemeljen na utjecaju [12] (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. Prema autorima rada [12], ovaj zahtjev, iako pomalo striktan, povećava uspješnost napada. Pretpostavljamo da će napadač prvo manipulirati slike i oznake u skupu za učenje, a potom koristiti fizički okidač poput naljepnice ili plakata. 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 prikazuje primjer dodavanja okidača na ulaznu sliku. Konkretno, na sliku je na poziciju gdje svi pikseli pripadaju jednom razredu (razredu `road`) dodan poznati Hello Kitty okidač. Kao razred žrtva odabran je razred `car`. U slučaju da je kao razred žrtva odabran neki drugi razred, morali bismo osigurati da okidač ne prekriva piksele iz odabranog razreda.



Slika 3.3: Primjer dodavanja okidača kod napada utemeljenog na utjecaju. Svi pikseli okidača smješteni su na cesti. Kao razred žrtva odabran je razred `car`.

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. Napadač jednostavno može pozicionirati okidač tako da je on što bliže žrtvi kada su u pitanju slike iz skupa za učenje, no ovo općenito ne vrijedi kada je u pitanju fizički okidač. Na primjer, ako je žrtva razred `car`, napadač bi morao osigurati da se okidač nalazi u blizini pokretnih automobila.

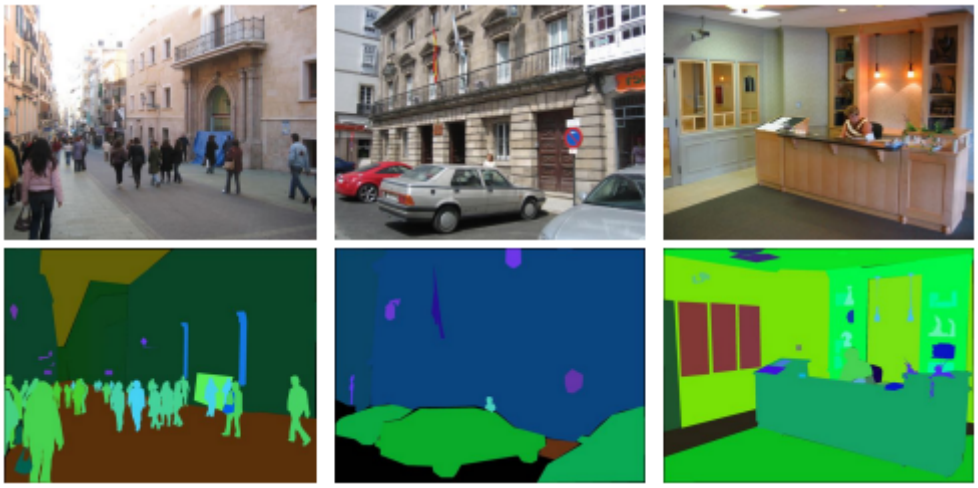
Posljednja inačica napada uvodi slučajne izmjene oznaka koje poboljšavaju učinkovitost napada (engl. pixel random labeling - PRL). Kod ove inačice, pozicija okidača ponovno se odabire nasumično. Uz ovo, dodatno se mijenjaju oznake. Konkretno, nasumično se odabere određen broj piksela koji ne pripadaju razredu žrtve te se isti označe nasumično odabranim razredima koji su prisutni u slici. Na primjer, za nasumično odabran piksel koji pripada razredu `person`, nova oznaka može biti razred `grass`. Prema autorima rada [12], ovakva izmjena oznaka potiče model da koristi informacije iz šireg konteksta i time povećava uspješnost napada za slučajeve kada je okidač jako udaljen od piksela iz razreda žrtve. Kao i kod prethodnih inačica, pretpostavljamo da će napadač prvo manipulirati slike i oznake u skupu za učenje, a potom koristiti fizički okidač. Za razliku od inačice NNI, kod ove inačice napadač može pozicionirati fizički okidač bilo gdje unutar vidnog polja kamere.

4. Skupovi podataka

4.1. Skup podataka ADE20k

Skup ADE20k [20] sastoji se od otprilike 25 000 označenih slika preuzetih iz skupova SUN [18] i Places [19]. 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. Osim osnovne podjele, postoji i podjela na veći broj razreda finije granulacije. Na primjer, objekt razreda Car sastoji se od niza manjih objekata iz drugih razreda. U okviru naših eksperimenata, koristili smo samo osnovnu podjelu. Ulazne slike, kao i pripadne segmentacije, varirajućih su dimenzija u rasponu od 306 do 614 piksela [2].

Slika 4.1 prikazuje 3 ulazne slike, kao i pripadne oznake tj. segmentacije. Na segmentacijama su različiti razredi označeni različitim bojama. Općenito govoreći, skup ADE20k prikazuje razne scene: od slika interijera, pa sve do slika ulica i poznatih građevina.



Slika 4.1: Primjeri slika i oznaka iz skupa ADE20k. Preuzeto iz [20].

4.2. Skup podataka Cityscapes

Skup Cityscapes [4] sastoji se od 5000 označenih slika iz perspektive vozača. Podijeljen je na 2975 slika u skupu za učenje, 500 slika u skupu za validaciju te 1525 slika u skupu za ispitivanje. Svaki piksel može pripadati jednom od ukupno 30 razreda. Ulazne slike, kao i pripadne segmentacije, dimenzija su 2048x1024.

Slika 4.2 prikazuje 4 ulazne slike, kao i pripadne segmentacije. Kao i inače, na segmentacijama su različiti razredi označeni različitim bojama. Općenito govoreći, skup Cityscapes sadrži slike koje se fokusiraju na promet. Slike su prikupljene u 50 različitih gradova u različitim godišnjim dobima i vremenskim uvjetima.

Slika 4.2: Primjeri slika i oznaka iz skupa Cityscapes. Preuzeto iz [16].

5. Eksperimenti

5.1. Osnovni pristupi trovanju podataka

5.1.1. Postavke eksperimenata

Eksperimente vezane uz osnovne pristupe trovanju podataka provodili smo učenjem modela na prirodnom skupu podataka, kao i na nekolicini zatrovanih skupova podataka. Pritom smo za stvaranje zatrovanih skupova podataka koristili razne okidače, kao i razne izmjene segmentacijskih oznaka. Kao skup podataka u ovim eksperimentima koristili smo skup ADE20k.

Kada govorimo o odabiru okidača, koristili smo nesemantičke, ali i semantičke okidače. 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 odabiru načina izmjene oznaka, koristili smo izmjenu na razini slike, ali i na razini primjerka. Za izmjenu na razini slike, nasumično smo odabrali oznake slike koja sadrži barem jedan piksel iz razreda Road. Svakom zatrovanom primjeru dodijelili smo oznake odabrane slike, pritom provodeći prikladno skaliranje zbog varirajućih dimenzija primjera. Kod izmjene 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. Pritom smo koristili mini-grupe veličine 20 te optimizator Adam sa stopom učenja iznosa $8 \cdot 10^{-4}$ i propadanjem težina iznosa $1 \cdot 10^{-4}$. Kako bi se kroz epohe stopa učenja ciklički mijenjala, koristili smo strategiju kosinusnog kaljenja. 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 stopa uspješnosti napada (engl. attack success rate – ASR). U svim eksperimentima sa zatrovanim podatcima, stopa trovanja iznosila je otprilike 10%.

5.1.2. Rezultati

U tablici 5.1, stupac *Vrsta napada* predstavlja vrstu napada korištenu za trovanje skupa za učenje te skupa za validaciju. Pritom je za trovanje skupa za učenje korištena već spomenuta stopa trovanja od otprilike 10%, dok je skup za validaciju u potpunosti zatrovan. 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 stopu uspješnosti napada mjerenu na zatrovanom skupu za validaciju.

Tablica 5.1: Performanse modela učenih na skupovima podataka zatrovanim osnovnim pristupima trovanju.

Vrsta napada	mIoU [%]	PA [%]	ASR [%]
Prirodno učenje	**33.08**	**76.13**	-
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, kao i PA. Vidimo 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 - u ovom slučaju, mIoU je smanjen za otprilike 3%.

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. Drugim riječima, ako bismo koristili napad na razini slike, najbolje performanse postigli bismo korištenjem linijskog okidača. Naravno, ovu vrstu okidača, za razliku od semantičkog okidača, moguće je relativno lako detektirati. Zbog toga, obrane od ove vrste napada mogle bi biti jednostavnije od obrana od napada sa semantičkim okidačem.

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. Ipak, ovdje je važno istaknuti činjenicu da je zatrovani skup za validaciju kod korištenja semantičkog okidača manji od zatrovanog skupa za validaciju kod korištenja nesemantičkog okidača. Ovo proizlazi

iz činjenice da za postizanje stope trovanja iznosa 100% kod korištenja semantičkog okidača u obzir smijemo uzeti isključivo ulazne slike koje sadrže barem jedan piksel iz razreda Wall. Zbog ovoga, teško je direktno uspoređivati stopu uspješnosti napada kod korištenja semantičkog odnosno nesemantičkog okidača. Naravno, ista primjedba vrijedi i za izmjenu oznaka na razini slike.

5.2. Napad utemeljen na utjecaju

5.2.1. Postavke eksperimenata

Eksperimente vezane uz napad utemeljen na utjecaju provodili smo učenjem modela na prirodnom skupu podataka, kao i na nekolicini zatrovanih skupova podataka. Pritom smo za stvaranje zatrovanih skupova koristili različite inačice napada utemeljenog na utjecaju. Kao skup podataka u ovim eksperimentima koristili smo skup Cityscapes. Sve slike, kao i njihove segmentacije, skalirali smo na dimenzije 1024x5012.

Kao razred žrtvu koristili smo razred Car. Konkretno, pikseli koji su originalno pripadali razredu Car, nakon izmjene pripadaju razredu Road. Kao nesemantički okidač koristili smo poznati Hello Kitty okidač skaliran na dimenzije 50x50. Ovisno o konkretnoj inačici napada, okidač je postavljen ili na nasumičnu poziciju ili na poziciju najbližu razredu žrtvi. Naravno, pozicija mora zadovoljavati već navedene uvjete - pikseli na poziciji moraju svi biti iz istog razreda koji nije razred žrtva.

Svi modeli učeni su 200 epoha. Pritom smo koristili mini-grupe veličine 14 te optimizator Adam sa stopom učenja iznosa $4 \cdot 10^{-4}$ i propadanjem težina iznosa $1 \cdot 10^{-4}$. Kako bi se kroz epohe stopa učenja ciklički mijenjala, koristili smo strategiju kosinusnog kaljenja. 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 stopa uspješnosti napada (engl. attack success rate – ASR). U svim eksperimentima sa zatrovanim podatcima, stopa trovanja za validacijski skup iznosila je 100%, dok je stopa trovanja za skup za učenje varirala od eksperimenta do eksperimenta.

5.2.2. Rezultati

U tablici 5.2, stupac *Vrsta napada* predstavlja vrstu napada korištenu za trovanje skupa za učenje te skupa za validaciju. Oznaka IBA predstavlja osnovnu inačicu napada utemeljenog na utjecaju, oznaka NNI inačicu koja se zasniva na najbližim susjedima, a oznaka PRL inačicu koja se zasniva na označavanju nasumičnih piksela. Stupac *PR* predstavlja stopu trovanja za skup za učenje. 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 stopu uspješnosti napada mjerenu na zatrovanom skupu za validaciju.

Tablica 5.2: Performanse modela učenih na skupovima podataka zatrovanim inačicama napada utemeljenog na utjecaju.

Vrsta napada	PR [%]	mIoU [%]	PA [%]	ASR [%]
Prirodno učenje	-	**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**

Kao što možemo vidjeti u tablici 5.2, model učen na prirodnom skupu očekivano postiže najviši mIoU, kao i PA. Vidimo da korištenje zatrovanog skupa smanjuje mIoU modela na skupu za validaciju za otprilike $3 - 4\%$. Povećavanjem stope trovanja očekivano raste i stopa uspješnosti napada, bez obzira koristimo li napad IBA ili napad NNI.

Za oba napada, korištenje stope trovanja iznosa 10% čini se kao dobar odabir. Ako koristimo ovu stopu trovanja, potencijalno je manja vjerojatnost da će napad biti detektiran, a stopa uspješnosti napada već je prilično visoka. Štoviše, vidimo da dodatnim

povećavanjem stope trovanja do 20% za oba napada dobivamo rast stope uspješnosti napada od samo 5%.

Korištenje napada NNI čini se kao značajno bolja opcija u usporedbi s napadom IBA. Neovisno o korištenoj stopi trovanja, stopa uspješnosti napada NNI otprilike je za 10% viša od stope uspješnosti napada IBA. Dodatno, ako koristimo stopu trovanja iznosa 1%, napad NNI tada već ima stopu uspješnosti napada iznosa otprilike 42%, dok napad IBA tada ima stopu uspješnosti napada iznosa otprilike 24%. Ipak, važno je napomenuti da bi napad NNI u stvarnosti bilo teško ostvariti jer napadač općenito ne može pozicionirati okidač tako da bude najbliže moguće razredu žrtvi.

6. Zaključak

Kao što vidimo, modeli semantičke segmentacije ranjivi su na napad trovanjem skupa podataka kao i obični klasifikatori. Kada govorimo o osnovnim pristupima trovanju podataka, 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 segmentacije jednak kao i prije. Ovakvu vrstu napada bilo bi veoma teško detektirati jer napadač mijenja isključivo dio segmentacije, dok ulazna slika ostaje nepromijenjena.

Uz osnovne pristupe trovanju podataka, vidjeli smo i rezultate korištenja različitih inačica napada utemeljenog na utjecaju. Ovi napadi po svojoj prirodi slični su nesemantičkom napadu na razini primjerka. Vidjeli smo da koristeći ove napade možemo postići još veću stopu uspješnosti napada po cijeni veoma malog pada mIoU te PA na prirodnim podatcima.

Kada govorimo o mogućem budućem radu na ovu temu, bilo bi zanimljivo provesti eksperimente na različitim arhitekturama - na primjer, na arhitekturi Multi-Scale SwiftNet. Uz ovo, korisno bi bilo proučiti moguće načine obrane modela za semantičku segmentaciju od napada trovanjem podataka.

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] Petra Bevandić, Marin Oršić, Ivan Grubišić, Josip Šarić, i Siniša Šegvić. Multi-domain semantic segmentation with pyramidal fusion. *arXiv preprint arXiv:2009.01636*, 2020.

[3] 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.

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, i Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. U *Proceedings of the IEEE conference on computer vision and pattern recognition*, stranice 3213–3223, 2016.

[5] 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.

[6] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, i Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

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

[16] Adnan Ahmed Rafique, Yazeed Yasin Ghadi, Suliman A Alsuhibany, Samia Allaoua Chelloug, Ahmad Jalal, i Jeongmin Park. Cnn based multi-object segmentation and feature fusion for scene recognition. U *Proceedings of the Conference on Membrane Computing, Chandler, AZ, USA*, stranice 27–29, 2022.

[17] Md Atiqur Rahman i Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. U *International symposium on visual computing*, stranice 234–244. Springer, 2016.

[18] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, i Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. U *2010 IEEE computer society conference on computer vision and pattern recognition*, stranice 3485–3492. IEEE, 2010.

[19] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, i Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[20] 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