Trojanski napadi na modele za semantičku segmentaciju

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 i zatrovani podatci. 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. S druge strane, 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.

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 od istih. 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. 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.

Svi eksperimenti provođeni su na arhitekturi *Single-Scale* SwiftNet. Ovaj model sastoji se od kodera koji provodi poduzorkovanje, sloja prostornog piramidalnog sažimanja i dekodera koji provodi naduzorkovanje. Pritom se za okosnicu kodera koristi ResNet18 model prednaučen na ImageNet skupu podataka. Slojevi kodera i dekodera povezani su lateralnim vezama (ljestvičasta arhitektura). Modeli su učeni na podskupu ADE20k skupa podataka. Skup je podijeljen 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.

U okviru eksperimenata, model je učen 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 očekivane segmentacije. 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 izmjeni očekivanih segmentacija zatrovanih primjera, radimo podjelu na *BadNets* napad te na fino-granulirani napad. Za *BadNets* napad, nasumično smo odabrali segmentaciju koja sadrži barem jedan piksel iz razreda *Road*. Ovu segmentaciju dodijelili smo svakom zatrovanom primjeru, pritom provodeći prikladno skaliranje zbog varirajućih dimenzija primjera. Kod fino-granuliranog napada, 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), kao i mjera uspješnosti napada (engl. *attack success rate* – ASR). U svim eksperimentima sa zatrovanim podatcima, stopa trovanja iznosila je otprilike 10%. Rezultate možemo vidjeti u tablici 1.

Vrsta napada	mIoU (%)	PA (%)	ASR (%)
Prirodno učenje	33.08	-	-
Nesemantički (linija), *BadNets*	31.80	75.35	39.40
Nesemantički (okvir), *BadNets*	31.92	75.47	35.25
Semantički (*Grass*), *BadNets*	29.20	69.74	30.97
Nesemantički (linija), fino-granulirani	32.30	75.71	58.93
Semantički (*Wall*), fino-granulirani	32.90	75.00	76.65

Tablica 1 – Performanse modela učenih na različitim skupovima podataka

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 *BadNets* napadom. Kada govorimo o modelima učenim na skupu podataka zatrovanim *BadNets* napadom, najvišu stopu uspješnosti napada postiže model s nesemantičkim okidačem linije. Model sa semantičkim okidačem (okidač su slike s barem jednim pikselom iz razreda *Grass*) postiže najlošije rezultate po svim mjerama.

Kada govorimo o modelima učenim na skupu podataka zatrovanim fino-granuliranim napadom, 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 nesemantičkim okidačem linije.

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