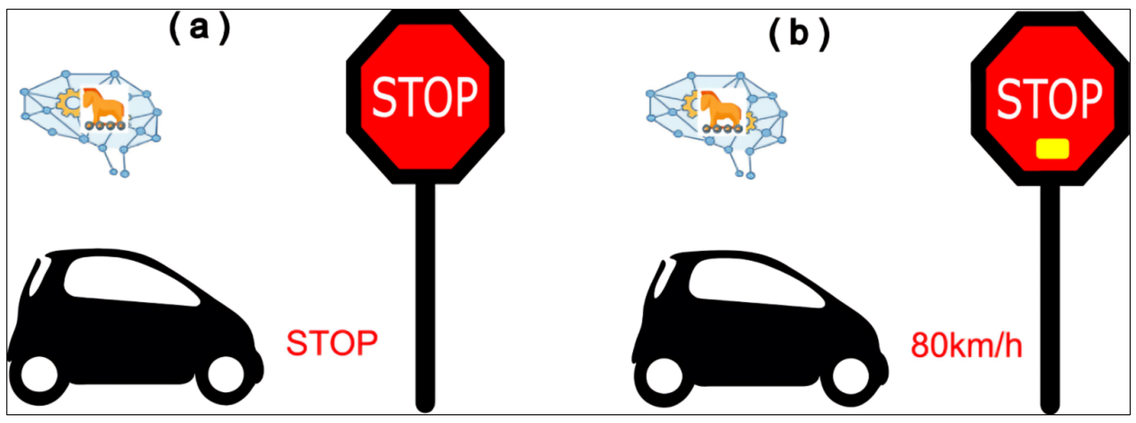
Detekcija i otklanjanje napada na duboke neuronske mreže

Autor: Danijel Barišić Mentor: prof. dr. sc. Sven Lončarić
Sveučilište u Zagrebu
Fakultet elektrotehnike i računarstva
Zavod za elektroničke sustave i obradbu informacija

1. Uvod i opis problema

Složene duboke neuronske mreže koje se koriste za današnje potrebe zahtijevaju puno podataka i računalnih resursa za uspješno treniranje. Treniranje mreže može se povjeriti **vanjskim pružateljima usluge**, no to dolazi uz sigurnosni rizik od napada.

Napadač umeće zlćudni uzorak na dio podataka za treniranje, čime se ostvaruje **napad stražnjih vrata** te napadač tako može upravljati predikcijama mreže.



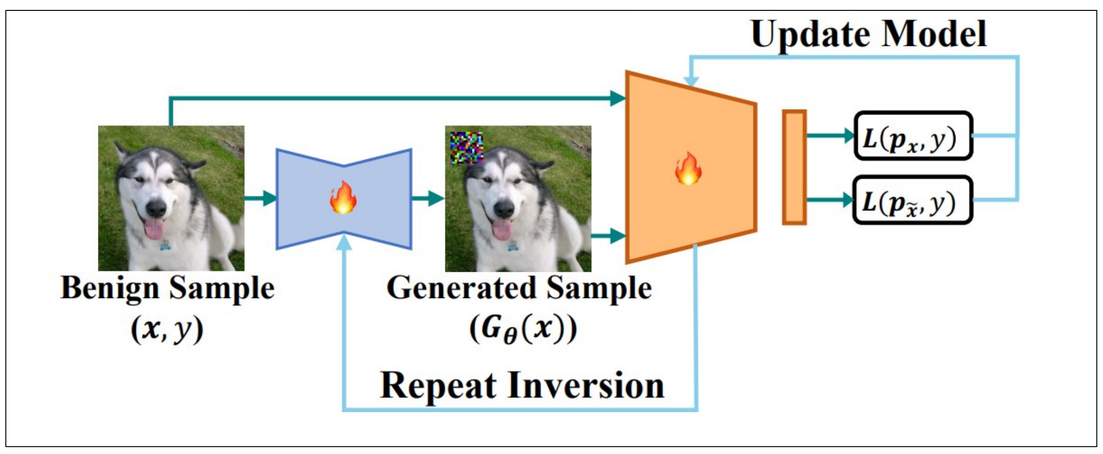
Npr. napadnuto autonomno vozilo može krivo interpretirati znak stop kao znak za ograničenje od 80km/h, ako se na znaku nalazi žuti papirić (napadačev okidač).

U ovom radu istražene su učinkovite metode otklanjanja takvih napada, te je detaljnije obrađena **BTI-DBF** metoda.

2. Metoda

BTI-DBF metoda (engl. *Backdoor Trigger Inversion – Decoupling Benign Features*) ima dva dijela:

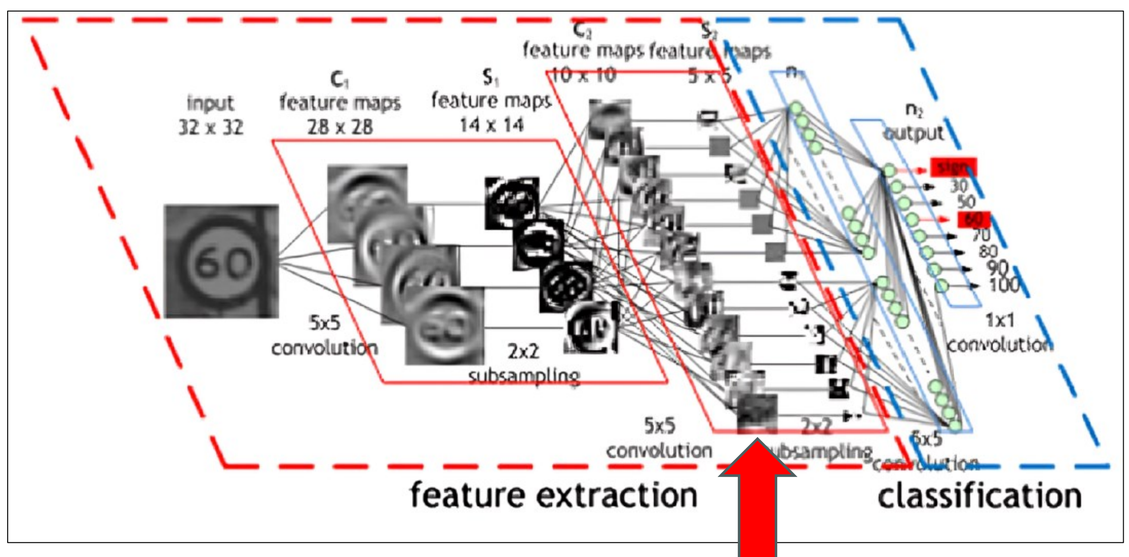
- **BTI** – inverzija okidača – cilj je ustanoviti kako izgleda zlćudni uzorak okidača na slikama
- **DBF** – izdvajanje dobroćudnih značajki – cilj je naučiti donositi predikcije samo pomoću dobrih značajki slike (bez oslanjanja na zlćudne)



Generator (plavo) generira otrovanu sliku s okidačem (ali s ispravnom oznakom umjesto napadačevom oznakom), dok model sa stražnjim vratima (narančasto) vrši predikcije.

Cilj je **ukloniti stražnja vrata** iz modela, tako da se model trenira koristiti isključivo dobroćudne značajke prilikom predikcije. Generator se pritom trenira tako da minimizira razliku između dobroćudnih značajki ulaznih i izlaznih slika, dok maksimizira razliku zlćudnih značajki.

U radu je također postignuto pojednostavljenje modela BTI-DBF metode tako da generator **umjesto sa slikama** barata direktno s njihovim **značajkama**.



Sloj značajki u mreži

3. Rezultati

Rezultati su prikazani pomoću dvije metrike učinkovitosti metoda obrane:

- **ASR** (engl. *Attack Success Rate*) – koliko dobro napadač zavarava mrežu (postotak krivih predikcija nad otrovanim podacima) – treba biti **nisko**
- **BA** (engl. *Benign Accuracy*) – koliko dobro mreža radi nad čistim podacima (točnost) – treba biti **visoko**

Obrane →	NAD		FeatureRE		BTI-DBF		BTI-DBF (FS)	
Napadi ↓	BA	ASR	BA	ASR	BA	ASR	BA	ASR
BadNets	90.32	2.98	91.53	35.42	92.00	1.36	**88.84**	**1.21**
Blended	89.49	3.29	92.86	40.50	91.60	7.92	**87.10**	**8.19**
WaNet	91.38	6.94	93.75	0.02	90.82	0.94	**86.45**	**1.13**
IAD	91.34	17.45	93.23	0.39	91.91	1.22	**89.58**	**1.22**
LC	91.77	12.65	94.54	10.49	90.48	4.51	**88.12**	**4.32**

Naša metoda **BTI-DBF (FS)** (*feature space*) ima sličan ASR kao originalna metoda uz marginalno lošiji BA, a model je uvelike **jednostavniji**.

4. Zaključak

BTI-DBF metoda ima konzistentno dobre rezultate u usporedbi s postojećim metodama obrane. Postignuto je i značajno pojednostavljenje modela dok su performanse prilično očuvane.

U radu je pretpostavljena mogućnost pristupa sloju značajki (*white-box*), a kao sljedeći zadatak može se istražiti izrada BTI metode gdje je moguć pristup isključivo ulazu i izlazu neuronske mreže (*black-box*).