

SVEUČILIŠTE U ZAGREBU  
**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

DIPLOMSKI RAD br. 2231

# **Obrana dubokih konvolucijskih modela od neprijateljskih primjera**

Matej Dobrovodski

Zagreb, lipanj 2020.

# SADRŽAJ

|  |           |
|--|-----------|
| <b>1. Uvod</b>                                       | <b>1</b>  |
| 1.1. Raspoznavanje objekata . . . . .                | 1         |
| 1.2. Neprijateljski primjeri . . . . .               | 2         |
| <b>2. Programska potpora</b>                         | <b>4</b>  |
| 2.1. Odabir biblioteke za duboko učenje . . . . .    | 4         |
| 2.2. Biblioteke za neprijateljske primjere . . . . . | 4         |
| 2.3. Skupovi podataka . . . . .                      | 5         |
| 2.4. Konvolucijski modeli . . . . .                  | 6         |
| <b>3. Neprijateljski primjeri I</b>                  | <b>8</b>  |
| 3.1. Model prijetnje . . . . .                       | 8         |
| 3.2. Pojava prvih neprijateljskih primjera . . . . . | 11        |
| 3.3. Brza metoda temeljena na gradijentima . . . . . | 12        |
| 3.4. DeepFool . . . . .                              | 14        |
| <b>4. Obrana dubokih konvolucijskih modela I</b>     | <b>18</b> |
| 4.1. Jednostavne obrane . . . . .                    | 18        |
| 4.1.1. JPEG kompresija . . . . .                     | 19        |
| 4.1.2. Stiskanje značajki . . . . .                  | 21        |
| 4.2. Neprijateljsko treniranje - FGSM . . . . .      | 22        |
| 4.3. Termometar kodiranje . . . . .                  | 24        |
| 4.4. Obračbena destilacija . . . . .                 | 26        |
| <b>5. Neprijateljski primjeri II</b>                 | <b>28</b> |
| 5.1. Neučinkovitost obrana . . . . .                 | 28        |
| 5.2. Projicirani gradijentni spust . . . . .         | 29        |
| 5.3. Napadi <i>Carlini and Wagner</i> . . . . .      | 30        |
| 5.4. Granični napad . . . . .                        | 30        |

|  |           |
|--|-----------|
| <b>6. Obrana dubokih konvolucijskih modela II</b>                    | <b>35</b> |
| 6.1. Preduvjeti uspješnih obrana . . . . .                           | 35        |
| 6.2. Neprijateljsko treniranje - PGD . . . . .                       | 35        |
| 6.2.1. <i>Fast is better than free</i> . . . . .                     | 35        |
| 6.3. Dokazivost obrane od neprijateljskih napada . . . . .           | 35        |
| 6.4. Budući rad . . . . .  | 35        |
| <b>7. Zaključak</b>  | <b>36</b> |
| <b>Literatura</b>  | <b>37</b> |
| <b>8. Dodatak</b>  | <b>40</b> |
| 8.1. Osobni skup slika . . . . .                                     | 40        |
| 8.2. Izlazi modela na nepromijenjenim slikama iz osobnog skupa . . . | 40        |

# 1. Uvod

## 1.1. Raspoznavanje objekata

Raspoznavanje objekata jedan je od ključnih problema područja računalnog vida. Pri rješavanju problema raspoznavanja objekata se na ulaz nekog sustava dovede slika nekog objekta, a na izlazu se očekuje ispravna klasifikacija u neki od predodređenih razreda. Čovjeku ovaj zadatak ne predstavlja veliki problem, no još uvijek ne postoji zadovoljavajuće rješenje problema koje bi vrijedilo za opći slučaj. Trenutno najbolja takva rješenja temelje se na konvolucijskim neuronskim mrežama.

Razvoj konvolucijskih mreža počeo je osamdesetih godina prošlog stoljeća. Počelo je razvojem *neocognitron*[citat?]-a—neuronske mreže inspirirane biološkim stanicama vidne kore mozga. Krajem devedesetih godina se pojavljuje konvolucijska neuronska mreža LeNet5. LeNet5 mreža je vrlo uspješno raspoznavala rukom pisane znamenke te je ova mreža bila početna točka za daljnja istraživanja drugih neuronskih mreža. [citati]

*ImageNet*[citat] projekt je velika baza podataka predviđena za istraživanje područja raspoznavanja objekata. S više od 14 milijuna slika podijeljenih u 20000 kategorija, *ImageNet* skup je daleko najveći slobodno dostupni skup. Počevši od 2010. godine, *ImageNet* projekt organizira godišnje natjecanje, *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC). Veliki skok u točnosti pri raspoznavanju dogodio se 2012. godine kada je konvolucijska neuronska mreža *AlexNet*[citat] postigla top-5 pogrešku od samo 15.3%, što je bilo 10.8% manje od sljedeće mreže. To je postignuto korištenjem grafičkih procesora pri treniranju, što je potaknulo svojevrsnu revoluciju u području dubokog učenja.

Do 2017. godine, većina timova u natjecanju je imala top-5 točnost veću od 95%. Danas se u raznim bibliotekama mogu naći unaprijed istrenirane mreže koje postižu vrlo dobre rezultate, te će se one spominjati i koristiti u nastavku rada. Neke od tih mreža su primjerice *ResNet*, *Xception* i *VGG*. Sve spomenute

mreže postižu vrlo zadovoljavajuće točnosti pri ispitivanju (top-5 točnosti iznad 90%) i čini se da mogu dobro generalizirati. No u nastavku rada će biti pokazan oblik napada na konvolucijske mreže koji dovodi u pitanje činjenicu da današnje konvolucijske mreže dobro generaliziraju.

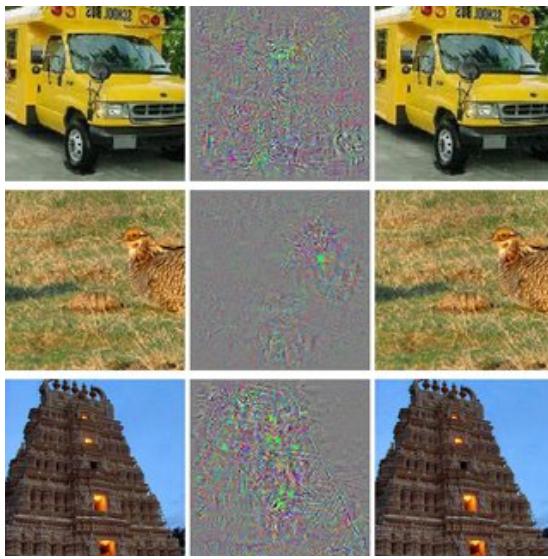
## 1.2. Neprijateljski primjeri

Krajem 2013. godine pojavljuje se prvi izravni "napad" na duboke neuronske mreže<sup>[19]</sup>, gdje je jedna od meta bila prethodno spomenuta uspješna mreža *AlexNet*. Polazna pretpostavka je da duboki modeli, usprkos tome što dobro generaliziraju, imaju ugrađene svojevrsne *slijepe pjege* koje se isplati istražiti.

Vrijedi da za neki ispravno klasificirani ulaz  $x$  postoji područje  $x + r$  u blizini ulaza te je uobičajeno da modeli ulazne vrijednosti iz tog područja također ispravno klasificiraju, isto kao i  $x$ . U općem slučaju vrijedi da neprimjetne perturbacije iz tog područja (npr. nasumični šum slabog intenziteta) ne mijenjaju izlaz modela. To je pretpostavka lokalne generalizacije i tipično vrijedi za probleme iz područja računalnog vida.

Međutim, ispostavilo se da ta pretpostavka lokalne generalizacije zapravo ne vrijedi. Otkriveno je da je moguće konstruirati perturbaciju  $r$  koja dovodi do pogrešne klasifikacije, a ljudskom oku nije uočljiva. Takve slike se nazivaju neprijateljskim primjerima, a taj pojam se može generalizirati i na mnoga druga područja i na sličan način se mogu napasti sustavi pretvaranja teksta u govor, sustavi za detekciju zločudnih programa i praktički svi sustavi koji se oslanjaju na dosadašnje modela za duboko učenje.

Ono što je iznenadjuće i što je potaklo daljnje istraživanje je to što je zapravo iznimno lako za pronaći takve neprijateljske primjere na *state of the art* modelima kod kojih je perturbacija  $r$  potpuno neprimjetna i to što nije nimalo očito zašto mreže neispravno klasificiraju takve ulaze. Jedan od originalnih napada je prikazan na slici 1.1 gdje *AlexNet* mreža predviđa da su novonastale slike zapravo slike noja.



**Slika 1.1:** Primjer suparničkog napada na *AlexNet* mrežu<sup>[19]</sup>. U lijevom stupcu su originalne, ispravno klasificirane slike. U srednjem stupcu se nalazi perturbacija koja se nadodaje na originalnu sliku, a u desnom stupcu su sve tri novonastale slike klasificirane kao noj.

U radu je dan pregled nekoliko metoda generiranja neprijateljskih primjera: neke metode su prikazane zbog njihove povijesne važnosti i utjecaja na daljnji razvoj metoda, dok su neke metode iznimno snažne i mjerilo za uspješnost obrane od suparničkih napada. Pokazano je i koliko su napadi uspješni protiv poznatih mreža te kako niti jedan od široko dostupnih modela nije unaprijed otporan na napade. Uz napade su pokazane i obrane, s naglaskom na njihovu (ne)uspješnost pri odupiranju od postojećih suparničkih napada, probleme koji su gotovo svim obranama zajednički te potencijalnu budućnost razvoja uspješnijih obrana.

## 2. Programska potpora

### 2.1. Odabir biblioteke za duboko učenje

Postoji mnogo biblioteka koje pružaju sve potrebno za duboko učenje i računalni vid. U nastavku rada se koristi *Tensorflow 2<sup>[1]</sup>* u kombinaciji s bibliotekom *Keras<sup>[6]</sup>*. Zbog ogromnog dobitka u brzini izvođenja, ove biblioteke su korištene zajedno s platformom *CUDA* koja omogućava iskorištanje grafičkog procesora za obradu opće namjene (eng. *graphics processing unit for general purpose processing*, GPGPU). Grafička kartica korištena u sklopu generiranja rezultata u radu je NVIDIA GeForce RTX 2060 SUPER.

### 2.2. Biblioteke za neprijateljske primjere

Usprkos tome što su suparnički primjeri relativno nov koncept, već postoji mnogo biblioteka koje pružaju implementaciju velikog broja suparničkih napada, a često su i napadi implementirani izravno od strane autora napada. Istaknute su se tri biblioteke za generiranje suparničkih napada: *CleverHans<sup>[15]</sup>*, *Foolbox Native<sup>[16]</sup>* i *Adversarial Robustness Toolbox (ART)<sup>[14]</sup>*.

Za odabir biblioteke je razmatrano nekoliko stvari: dostupnost i ekstenzivnost dokumentacije, raznovrsnost implementiranih napada, jednostavnost korištenja, zahtijevana programska potpora te postoji li implementacija obrana. Za svaku biblioteku je implementirano generiranje neprijateljskih primjera napadom koji je opisan u 3.3, a napad je proveden na model opisan u .

*Cleverhans* ima kratku dokumentaciju za sve napade i poveznicu na relevantni rad koji opisuje napad, međutim ne postoji dokumentacija u formatu koji se lako pretražuje. *Foolbox* i *ART* imaju dokumentaciju dostupnu u takvom formatu, međutim *Foolbox* dokumentacija ne opisuje kako se napad poziva i s kojim argumentima, što otežava korištenje bez detaljnijeg proučavanja izvornog kôda, dok je *ART* dokumentacija eksplicitna kod toga.

Što se tiče jednostavnosti korištenja, *Cleverhans* je bio najjednostavniji za primjenu u ovom jednostavnom primjeru. *Foolbox* zahtjeva da se slike pretvore u određen format prije pokretanja napada, što otežava korištenje. *ART*, međutim, zahtjeva dodatne informacije pri konstruiranju napada kao što su broj razreda, dimenzije ulaza, funkcija gubitka i granične vrijednosti, što druge biblioteke ne traže.

*Cleverhans* i *Foolbox* imaju vrlo specifične zahtjeve za programsku potporu, iako će *Cleverhans* u budućnosti podupirati više od samo *Tensorflow*. *ART* pruža potporu za mnoštvo biblioteka: *Tensorflow* (v1 i v2), *Keras*, *PyTorch*, *MXNet* i još njih.

Od navedenih biblioteka, *ART* je jedina koja već sada ima implementirane neke od obrana u literaturi. *Cleverhans* biblioteka ima planove za implementaciju u budućnosti, dok *Foolbox* podržava samo napade.

Zbog svega navedenog, u nastavku rada se koristi samo *Adversarial Robustness Toolbox*. Dodatno, autori biblioteke su vrlo aktivni na *GitHub-u* i iznimno brzo reagiraju kada se postavi pitanje ili prijavi problem. Pri izradi rada otkriveno je nekoliko *bug-ova* koji su popravljeni u iznimno kratkom roku.

### 2.3. Skupovi podataka

*ImageNet*<sup>[17]</sup> je široko korištena baza podataka slika s preko 14 milijuna slika raspoređenih u više od 20000 razreda. *ImageNet* je praktički postao standard za treniranje i evaluaciju rada modela pri klasifikaciji objekata. Neki od modela korišteni u radu su unaprijed trenirani na *ImageNet* bazi podataka na kojima postižu iznimno visoku točnost.

Neke obrane i posljedično napadi će u nastavku rada biti evaluirani na *CIFAR-10*<sup>[12]</sup> skupu podataka. *CIFAR-10* sadrži samo 10 disjunktnih razreda, te 50000 slika za treniranje mreže. Nužno je koristiti i ovaj skup podataka jer pojedine obrane još uvijek nije moguće skalirati na *ImageNet* razinu jer vrijeme izvođenja nije razumno.

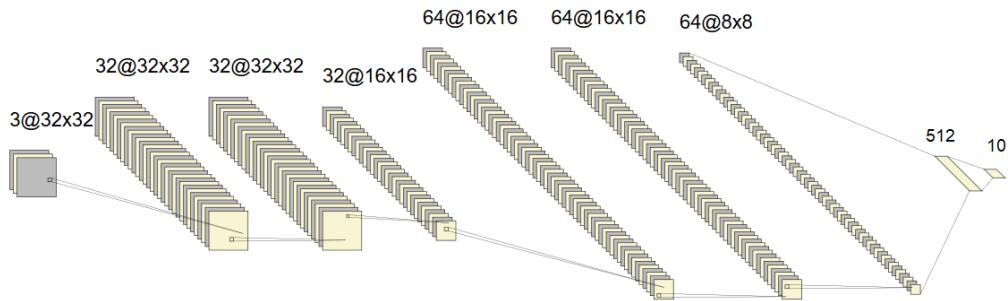
U svrhu rada je također odabранo 16 slika koje bi *ImageNet* modeli trebali ispravno klasificirati. Popis slika i izlazi određenih modela nalaze se u dodatku<sup>8</sup>.

## 2.4. Konvolucijski modeli

Primarna meta napada u radu je mreža *ResNet V2*<sup>[11]</sup>, i to verzija s 50 slojeva koja je unaprijed trenirana na *ImageNet* skupu. Mreža postiže top-1 točnost od 76%, te top-5 točnost od 93%. U početnim fazama izrade rada razmatrano je više mreža, međutim *ResNet* mreža je puno brža pri evaluaciji što ubrzava i olakšava evaluaciju napada i obrana. Korištenje ove mreže ne smanjuje općenitost ideja predstavljenih u radu, pošto su sve konvolucijske mreže jednako ranjive na neprijateljske napade. Lakoća provođenja neprijateljskih napada je "problem" koji sve konvolucijske mreže dijeli u jednakoj mjeri, i trenutno ne postoji niti jedna takva mreža koja je sama po sebi otporna na njih. Dodatno, kôd priložen uz rad dopušta provođenje napada na sljedeće mreže: *DenseNet121*, *VGG16*, *VGG19*, *MobileNetV2* te *Xception*.

Osim *ResNet* mreže, dodatno je konstruirana i jednostavna konvolucijska mreža za klasifikaciju *CIFAR-10* slika. Mreža se sastoji od 11 slojeva, redom:

- konvolucijski sloj oblika  $32 \times 32 \times 32$  s filtrom veličine  $3 \times 3$
- konvolucijski sloj oblika  $32 \times 32 \times 32$  s filtrom veličine  $3 \times 3$
- sloj sažimanja oblika  $2 \times 2$
- sloj ispadanja s vjerojatnošću 0.25
- konvolucijski sloj oblika  $16 \times 16 \times 64$  s filtrom veličine  $3 \times 3$
- konvolucijski sloj oblika  $16 \times 16 \times 64$  s filtrom veličine  $3 \times 3$
- sloj sažimanja oblika  $2 \times 2$
- sloj ispadanja s vjerojatnošću 0.25
- potpuno povezani sloj veličine 512
- sloj ispadanja s vjerojatnošću 0.50
- potpuno povezani sloj veličine 10, pošto se klasificira u 10 razreda



**Slika 2.1:** Skica jednostavne mreže za *CIFAR-10* mrežu. Nisu prikazani slojevi ispadanja.

Ukupno mreža ima 2,168,362 parametara koji se mogu naučiti. Na slici 2.1 je skica mreže. Aktivacijska funkcija između relevantnih slojeva je *ReLU*. Mreža već nakon 15 epoha lako postiže točnost od 75% na skupu za testiranje korištenjem stohastičnog gradijentnog spusta uz stopu učenja od 0.02, bez ikakvog dodatnog podešavanja hiperparametara. Nakon 25 epoha postiže točnost od 79.47%, no i u tom trenutku mreža nije pretrenirana. Za potrebe rada nije nužno maksimizirati točnost jer ne bi promijenilo rezultate. Jednako je lagano za pronaći suparničke primjere i na vrlo dobro istreniranim mrežama kao i ovakvim jednostavnim mrežama.

# 3. Neprijateljski primjeri I

## 3.1. Model prijetnje

Model prijetnje (eng. *threat model*) je proces kojim se potencijalne prijetnje mogu nabrojati i identificirati te se mogu odrediti određene mjere kao prioritet. Neki od dijelova modela prijetnje mogu biti: frekvencija interakcije s metom napada, željena vrsta pogrešne klasifikacije, količina znanja o meti napada i specifičnost napada. Prije opisivanja navedenih aspekata modela prijetnje uvedeno je nekoliko relevantnih simbola koji se učestalo pojavljuju u literaturi i u ovom radu.

### Osnovni pojmovi i simboli

- $f(\cdot)$  – model dubokog učenja
- $\mathbf{x}, l$  – originalni ulaz te pripadajuća labela
- $\mathbf{x}', l'$  – neprijateljski primjer i pripadajuća labela
- $J(\cdot)$  – funkcija gubitka, u većini slučajeva gubitak unakrsne entropije
- $\|\cdot\|_p$  – p-norma, p je najčešće 0, 2 ili  $\infty$ . Dodatna oznaka za norme je  $\ell_p$ .

### Norme

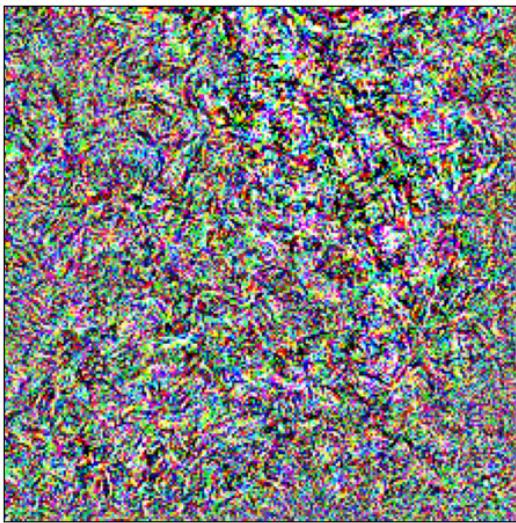
- $\ell_0$  –  $\|\mathbf{x}\|_0$  predstavlja broj ne-nula elemenata vektora  $\mathbf{x}$ .
- $\ell_2$  –  $\|\mathbf{x}\|_2 := \sqrt{x_1^2 + \dots + x_n^2}$ , odnosno Euklidska norma.
- $\ell_\infty$  –  $\|\mathbf{x}\|_\infty := \max_i |x_i|$ . U kontekstu neprijateljskih napada ova norma predstavlja maksimalnu vrijednost perturbacije  $r$ .

## Frekvencija interakcije s metom napada

- Jednokratni napad – jednokratni napadi (eng. *one-time*) su napadi kojima je potreban samo jedan pristup modelu da generiraju neprijateljski primjer. Ovi napadi su brzi, ali i mnogo slabiji od iterativnih napada te nisu u fokusu istraživanja. Na primjer, napad opisan u 3.3 je jednokratni napad.
- Iterativni napad – iterativni napadi zahtijevaju više pristupa modelu da bi generirali neprijateljski primjer. Ovakvi napadi generiraju daleko bolje neprijateljske primjere. Većina napada pripada ovoj kategoriji.

## Vrsta pogrešne klasifikacije

- Lažno pozitivni napad – lažno pozitivni (eng. *false positive*) primjeri u kontekstu neprijateljskih napada kod klasifikacijskih problema su čovjeku potpuno neprepoznatljivi (npr. šum), dok mreža s visokom vjerojatnošću klasificira sliku.
- Lažno negativni napad – lažno negativni (eng. *false negative*) primjeri su oni koje čovjek vrlo lako prepozna, a mreža pogrešno klasificira zbog nevidljive perturbacije. Fokus rada su od početka lažno negativni primjeri. Usporedba napada dana je u slici 3.1.



- (a) Lažno pozitivni napad. Mreža klasificira sliku kao prostirka za molitvu (eng. *prayer rug*) s vjerojatnošću 98.66%.
- (b) Lažno negativni napad. Mreža klasificira sliku kao zmaj (eng. *kite*) s vjerojatnošću 91.89%.

**Slika 3.1:** Primjeri lažno pozitivnog i lažno negativnog napada. Napad proveden na *ResNet50V2* mreži korištenjem *Carlini and Wagner*  $\ell_2$  napada opisanog u ??.

### Znanje o meti napada

- Bijela kutija – model se naziva bijela kutija (eng. *white box*) ako je sve o modelu unaprijed poznato napadaču, uključujući: arhitekturu, parametre mreže, aktivacijske funkcije, hiperparametre i sve druge moguće detalje mreže. Napadi koji se temelje na modelu bijele kutije često iskorištavaju gradijente mreže pri konstruiranju neprijateljskog primjera.
- Crna kutija – suprotno tome, model crne kutije (eng. *black box*) pretpostavlja nedostatak svih mogućih informacija, osim izlaza iz mreže. Na primjer, ako se napada neka mreža "u oblaku", njoj se pristupa tako da joj se preda ulaz te nije moguće direktno saznati dodatne detalje o mreži. Začudo, moguće je konstruirati neprijateljske primjere samo na temelju izlaza mreže.

## Specifičnost napada

- Ciljani napad – ciljni napad (eng. *targeted attack*) je oblik napada gdje se za neki neprijateljski primjer pokušava dobiti unaprijed određen izlaz mreže. Uz to, dodatni zahtjev može biti i da se maksimizira vjerojatnost odabranog razreda.
- Neciljni napad – neciljni napad (eng. *untargeted attack*) zahtjeva jedino da je klasifikacija neprijateljskog primjera neispravna. Općenito je lakše i brže konstruirati neciljni napad.

## 3.2. Pojava prvih neprijateljskih primjera

U uvodu je opisana osnovna ideja neprijateljskih primjera iz jednog od najranijih radova na temu neprijateljskih primjera<sup>[19]</sup>, a u nastavku je ideja dodatno razrađena i ukratko opisana optimizacijska metoda generiranja suparničkih primjera.

Implicitno je pretpostavljeno da mreže imaju svojstvo lokalne generalizacije. Za neki dovoljno mali radijus  $\epsilon > 0$  u blizini ulaza  $x$  (epsilon okolina), postoji ulaz  $x+r$  takav da je  $\|r\| < \epsilon$  koji će također imati veliku vjerojatnost pripadanja ispravnom klasifikacijskom razredu na izlazu mreže. Slabo vidljive promjene u pravilu ne mijenjaju drastično izlaz mreže, što se može i pokazati dodavanjem šuma na neku ulaznu sliku. Dapače, neke mreže su nasumičnom deformacijom ulaza pri treniranju povećavale robusnost modela. Pokazalo se da svojstvo lokalne generalizacije zapravo u velikoj mjeri nije prisutno kod tadašnjih (a i današnjih) modela dubokog učenja i da je moguće osmisiliti optimizacijski proces koji će pronaći primjere sa slabo vidljivim promjenama koje ipak drastično mijenjaju izlaz mreže i prisile model na pogrešnu klasifikaciju. Dodatno, spomenuti oblik treniranja deformiranjem ulaza nije nimalo osporavao traženje neprijateljskih primjera.

Slijedi formalni opis optimizacijskog problema koji je potrebno riješiti.

Klasifikator koji na ulazu prima sliku, a na izlazu daje pripadnu labelu označen je s  $f : \mathbb{R}^m \rightarrow \{1...k\}$ . Pripadajuća funkcija gubitka definirana je s  $loss_f : \mathbb{R}^m \times \{1...k\} \rightarrow \mathbb{R}^+$ . Za neku sliku  $x \in \mathbb{R}^m$  i neku labelu  $l \in \{1...k\}$ , potrebno je riješiti sljedeći optimizacijski problem:

- minimizirati  $\|r\|_2$  uz ograničenja

- (a)  $f(x + r) = l$
- (b)  $x + r \in [0, 1]^m$

Problem postaje netrivijalan za  $f(x) \neq l$  i traženje egzaktnog  $r$  je težak problem. Autori su riješili približni problem:

- minimizirati  $c|r| + loss_f(x + r, l)$  uz ograničenje  $x + r \in [0, 1]^m$

Korištenjem iterativnog optimizacijskog algoritma L-BFGS (eng. *Limited memory Broyden–Fletcher–Goldfarb–Shanno algorithm*) u svakoj iteraciji se minimizira  $loss_f(x + r, l)$  te se dodatno linijskim pretraživanjem odredi i minimalni  $c$  za koji je izlaz takav da je klasifikacija pogrešna. Za rad algoritma L-BFGS potrebno je unaprijed poznati i vrijednost gradijenta funkcije koja se optimizira, što ovaj napad čini napadom bijele kutije.

Jedno od ponuđenih objašnjenja postojanja neprijateljskih primjera je to što su konvolucijske mreže vrlo nelinearne po prirodi. To je zapravo poželjno svojstvo dubokih mreža, jer nelinearnost omogućuje rješavanje vrlo nelinearnih optimizacijskih problema kao što je klasifikacija slika. No čini se da je upravo zbog toliko visoke nelinearnosti lako za pronaći neprijateljske primjere koji su se sakrili u "džepovima" u blizini nekog ulaza, koje je vrlo teško pronaći nasumičnim pretraživanjem.

### 3.3. Brza metoda temeljena na gradijentima

Iako su neprijateljski primjeri otkriveni već krajem 2013., idući bitni rad na temu neprijateljskih primjera pojavio se tek 2015. godine<sup>[9]</sup>. Taj rad je direktni nastavak na prethodni te nudi nove načine generiranja neprijateljskih načina, jedno novo bitno i zanimljivo svojstvo neprijateljskih primjera te potpuno novo i neočekivano objašnjenje postojanja neprijateljskih primjera.

Neprijateljski primjeri su se originalno pojavili pod pretpostavkom da su duboki modeli previše nelinearni te je njihovo postojanje objašnjeno teorijom da modeli imaju "slijepе pjege" u kojima se neprijateljski primjeri teško pronalaze. Usprkos tome što je prethodno ponuđeno objašnjenje postojanja neprijateljskih primjera vrlo logično, sljedeći napad je osmišljen počevši od potpuno obrnute pretpostavke.

Ispostavilo se i da su linearni modeli podložni neprijateljskim primjerima, stoga je prvo potrebno objasniti kako je to moguće.

Digitalne slike uglavnom koriste samo 8 bitova za reprezentaciju pojedinog piksela, i svaka dodatna informacija manja od 1/255 je odbačena. Razumno je za očekivati da klasifikator nema različit izlaz za  $\mathbf{x}$  i  $\tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\eta}$ , ako je svaki element perturbacije  $\boldsymbol{\eta}$  manji od spomenute preciznosti. Formalnije, klasifikator bi trebao dati isti izlaz za  $\mathbf{x}$  i  $\tilde{\mathbf{x}}$  dokle god je  $\|\boldsymbol{\eta}\|_\infty < \epsilon$ , gdje je  $\epsilon$  dovoljno mal da bude odbačen.

Skalarni umnožak vektora težina  $\mathbf{w}^T$  i primjera s perturbacijom  $\tilde{\mathbf{x}}$  može se raspisati ovako:

$$\mathbf{w}^T \tilde{\mathbf{x}} = \mathbf{w}^T \mathbf{x} + \mathbf{w}^T \boldsymbol{\eta} \quad (3.1)$$

Dakle, perturbacija  $\boldsymbol{\eta}$  uzrokuje porast aktivacije za  $\mathbf{w}^T \boldsymbol{\eta}$ . Ovaj porast se može maksimizirati postavljanjem  $\boldsymbol{\eta} = \epsilon \operatorname{sign}(\mathbf{w})$ . Ako je prosječna vrijednost vektora težina  $m$ , onda je porast iznosa  $\epsilon mn$ . Norma  $\|\boldsymbol{\eta}\|_\infty$  ne raste s porastom dimenzionalnosti problema, dok porast aktivacije raste linearno s  $n$ . Dakle, za visoko dimenzionalne probleme moguće je dodati nevidljive promjene koje onda zajedno mogu drastično promijeniti izlaz.

Idea je zato napasti klasifikator "gdje najviše боли". Potrebno je maksimizirati promjenu izlaza uz minimalnu promjenu svakog pojedinačnog elementa. Za nelinearne modele, ideja je identična:

ako su  $\boldsymbol{\theta}$  parametri modela,  $\mathbf{x}$  ulaz,  $y$  izlaz te  $J(\boldsymbol{\theta}, \mathbf{x}, y)$  funkcija gubitka korištena za treniranje mreže, maksimalna perturbacija za koju je uvjet norme zadovoljen je:

$$\boldsymbol{\eta} = \epsilon \operatorname{sign}(\nabla_x J(\boldsymbol{\theta}, \mathbf{x}, y)) \quad (3.2)$$

Ova metoda generiranja neprijateljskih primjera se zove metoda temeljena na predznaku gradijenta (eng. *fast gradient sign method*). Metoda je brza jer je potreban samo jedan pristup mreži, i također je napad na bijelu kutiju. Činjenica da je nelinearne klasifikatore moguće napasti s istom pretpostavkom kao i linearne, te da su nelinearni modeli jednako podložni neprijateljskim napadima kao i linearni modeli dodatno dokazuje da problem nije to što su duboki modeli previše nelinearni, nego to da su previše linearni. Na sličan način se može dobiti maksimalna perturbacija pod uvjetima normi 1 i 2. Ne uzme se funkcija predznaka sign nego je potrebno gradijent podijeliti s određenim faktorom koji osigurava da uvjet norme ostane zadovoljen. Ovako proširena skupina neprijateljskih napada se zove brza metoda temeljena na gradijentima (eng. *fast gradient method*).



**Slika 3.2:** FGSM napad za  $\epsilon \in \{1, 5, 10\}$ . Torba je predviđena kao nogometna lopta (99.87%, 71.24%, 48.63%), orao kao zmaj (99.76%, 99.96%, 98.98%), i tržnica kao banana (99.79%, 99.99%, 99.99%).

Na slici 3.3 se nalazi primjer FGSM napada. Već za  $\epsilon > 2$  FGSM uspješno nalazi neprijateljski primjer za 14/16 slika iz skupa podataka 8.1. FGSM ne nađe neprijateljski primjer za slike 8.1a i 8.1l za razumni  $\epsilon$ . Dodatno je zanimljivo kako kad mreža pogriješi, greška je s vrlo visokom vjerojatnošću.  
transferabilnost

## 3.4. DeepFool

Algoritmi FGM uspješno nalaze neprijateljske primjere iz jednog pokušaja. Očito je da bi neka iterativna metoda sigurno pronašla neprijateljske primjere s još manjim perturbacijama.

2016. godine se pojavljuje sljedeći bitan algoritam stvaranja neprijateljskih primjera: DeepFool.

DeepFool je iterativni algoritam baziran na modelu bijele kutije. Kao i FGM, DeepFool iskorištava svojstvo prevelike linearnosti modela te u svakom koraku aproksimira nelinearni klasifikator na linearan način. Slijedi opis rada DeepFool algoritma na binarnom klasifikatoru.

Za neki klasifikator  $f$  i izlaznu, izlazna labela za neki ulaz  $\mathbf{x}$  označena je s  $\hat{k}(\mathbf{x})$ . Minimalna perturbacija  $\mathbf{r}$  je ona koja je dovoljna da promijeni vrijednost  $\hat{k}(\mathbf{x})$ :

$$\Delta(\mathbf{x}; \hat{k}) := \min_{\mathbf{r}} \|\mathbf{r}\|_2 \text{ uz uvjet } \hat{k}(\mathbf{x} + \mathbf{r}) \neq \hat{k}(\mathbf{x}) \quad (3.3)$$

Minimalna perturbacija potrebna za pogrešnu klasifikaciju nekog uzorka se također naziva i robusnost  $\hat{k}$  u točki  $\mathbf{x}$ . Usput se može i definirati robusnost cijelog modela:

$$\rho_{\text{adv}}(\hat{k}) = \mathbb{E}_{\mathbf{x}} \frac{\Delta(\mathbf{x}; \hat{k})}{\|\mathbf{x}\|_2} \quad (3.4)$$

Robusnost nekog klasifikatora definirana je kao očekivana potrebna perturbacija za stvaranje neprijateljskog primjera preko (nekog) cijelog skupa podataka.

Za binarni klasifikator  $f(\mathbf{x})$  se pretpostavlja da vrijedi  $\hat{k}(\mathbf{x}) = \text{sign}(f(\mathbf{x}))$ . Dodatno se definira skup  $\mathcal{F} := \{\mathbf{x} : f(\mathbf{x}) = 0\}$  – skup vektora  $\mathbf{x}$  za koje je izlaz klasifikatora 0.

Za klasifikator oblika  $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$  robusnost u točki  $\mathbf{x}_0$  je jednaka udaljenosti od  $\mathbf{x}_0$  do hiperravnine definirane s  $\mathcal{F} = \{\mathbf{x} : \mathbf{w}^T \mathbf{x} + b = 0\}$ . Ovo je prikazano na slici 3.3a.

Prema tome, da bi se klasifikacija promijenila, potrebna je perturbacija koja će točku  $\mathbf{x} = \mathbf{x}_0 + \mathbf{r}$  staviti na drugu stranu hiperravnine. Minimalna takva perturbacija jednaka je ortogonalnoj projekciji  $\mathbf{x}_0$  na ravninu  $\mathcal{F}$ . Ova projekcija može se izračunati u zatvorenom obliku:

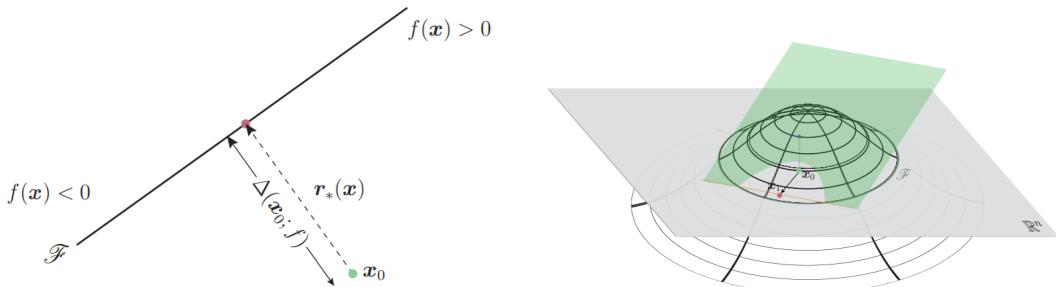
$$\mathbf{r}(\mathbf{x}_0) := -\frac{f(\mathbf{x}_0)}{\|\mathbf{w}\|_2^2} \mathbf{w} \quad (3.5)$$

U generalnom slučaju za bilo kakav diferencijabilni klasifikator ne može se vrijednost perturbacije izračunati u zatvorenom obliku. Ovdje se ponovno iskorištava svojstvo klasifikatora da su previše linearni, te se u svakoj iteraciji klasifikator  $f$  linearizira u točki  $\mathbf{x}_i$ . Minimalna perturbacija takvog linearног klasifikatora u koraku  $i$  računa se na sljedeći način:

$$\arg \min_{\mathbf{r}_i} \|\mathbf{r}_i\|_2 \text{ uz uvjet } f(\mathbf{x}_i) + \nabla f(\mathbf{x}_i)^T \mathbf{r}_i = 0 \quad (3.6)$$

Izraz uvjeta predstavlja tangencijalnu hiperravninu na funkciju klasifikatora. Algoritam opisan riječima bi glasio ovako: u svakom koraku algoritma radi se ortogonalna projekcija trenutne točke na tangencijalnu ravninu klasifikatora. Algoritam se ponavlja dokle god klasifikator ne pogriješi, odnosno dokle god točka ne dođe na granicu klasifikatora. Ilustracija iz rada prikana na slici 3.3b vizualno opisuje jedan korak algoritma. Zanimljivo je odmah uočiti kako je često i potreban samo jedan korak algoritma.

Kako se višerazredni klasifikator može promatrati kao skupina binarnih klasifikatora, algoritam je moguće poopćiti na višerazredne klasifikatore.



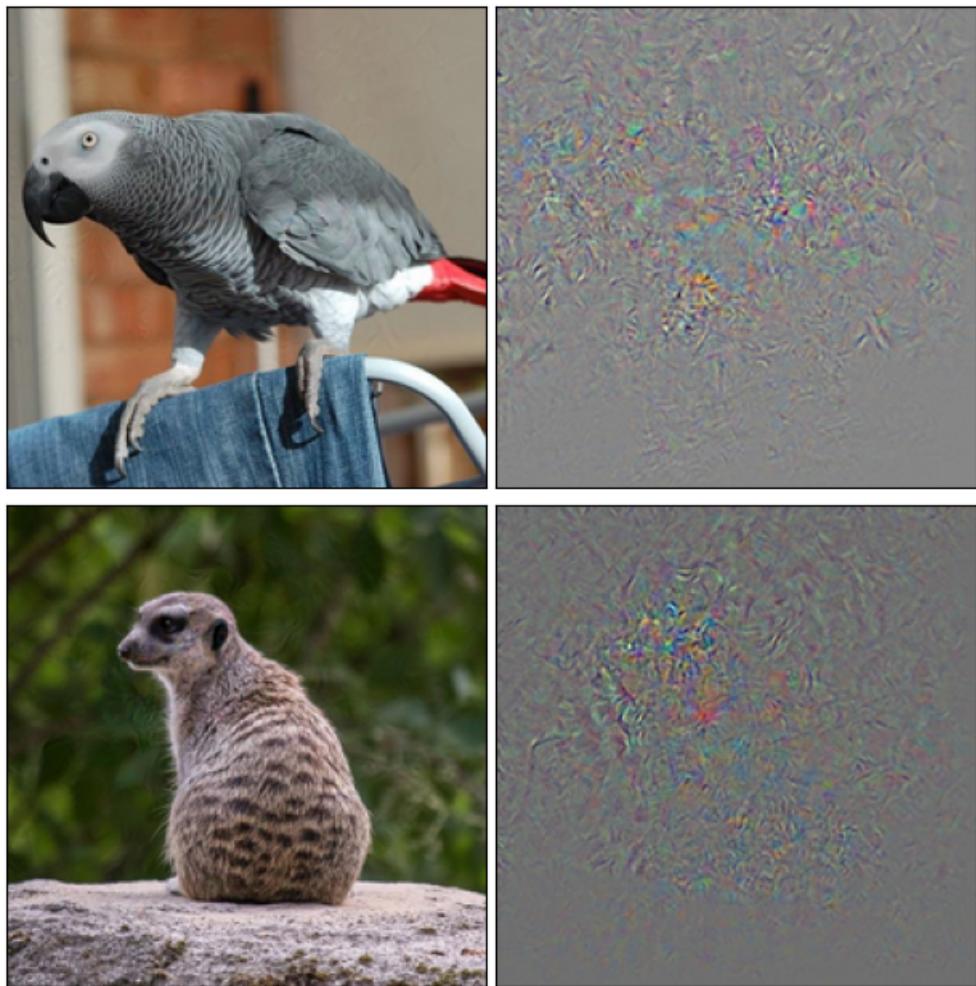
(a) Linearni binarni klasifikator. Minimalna perturbacija potrebna za promjeniti izlaz klasifikatora je ortogonalna projekcija na pravac koji dijeli razrede.

(b) Diferencijabilni binarni klasifikator. U svakom koraku se trenutnu točku aproksimira funkcija klasifikatora s hiperravninom i napravi projekciju na pravac koji siječe  $\mathbb{R}^n$  (narandžaste boje).

**Slika 3.3:** Ilustracija rada DeepFool algoritma na binarnim klasifikatorima.

Pošto je i za rad ovog algoritma potrebno znanje o mreži, ovo je također napad na model bijele kutije. Za razliku od do sada opisanog FGSM napada, DeepFool pronalazi bolje neprijateljske primjere s daleko manjim perturbacijama.

Što se tiče uspješnosti DeepFool napada na skup 8.1, napad je 100% uspješan na svim slikama. Broj iteracija potreban je također iznenađujuće nizak – za čak sedam slika je potrebna samo jedna iteracija algoritma. Za FGSM napad su se pokazale najzahtjevnije slike 8.1a i 8.1l. Te dvije slike su u slučaju DeepFoola zahtijevale 6 i 8 iteracija. U slici 3.4 se nalazi prikaz napada na te slike.



**Slika 3.4:** Rezultati DeepFool napada za slike koje FGSM nije mogao pretvoriti u neprijateljske. U lijevom stupcu je neprijateljska slika, a u desnom je razlika između originalne i neprijateljske slike. Izlaz modela za prvu sliku je *prepelica* (quail, 99.46%), a za drugu sliku *šumski kunić* (wood rabbit, 7.8% – moguće je povećati vjerojatnost uz podešavanje hiperparametara napada)

# 4. Obrana dubokih konvolucijskih modela I

Do sada je predstavljeno nekoliko napada. Originalni napad temeljen na minimizaciji približnog problema uz pomoć L-BFGS algoritam je opisan zbog povijesnih razloga. Napad nije u primjeni jer je u praksi vrlo spor i ne daje dobre rezultate. Međutim, pojavilo se mnoštvo potencijalnih obrana od dosadašnjih napada. U ovom poglavlju su opisane tri glavne kategorije takvih obrana na primjeru pojedinačnih obrana:

- Jednostavne obrane – konceptualno vrlo jednostavne za shvatiti, nije potrebno duboko predznanje o napadima, "jeftine" za implementirati
- Neprijateljsko treniranje – poboljšanje robusnosti klasifikatora još pri treningu
- Prikrivanje gradijenata – vođeni idejom povećanja nelinearnosti mreže, određeni napadi namjerno ili slučajno "prikrivaju" gradijente mreže i time prividno sprječavaju napade temeljene na gradijentima

## 4.1. Jednostavne obrane

U ovom dijelu je ukratko opisano nekoliko vrlo jednostavnih obrana. Dobra strana ovih obrana je što su iznimno jednostavne za implementirati i vrlo efektivne protiv već stvorenih neprijateljskih primjera, te nije potrebno nikakvo mijenjanje samog modela niti detaljno poznavanje potencijalnih metoda napada. Negativno je to što su pre-jednostavne da bi spriječile stvaranje novih neprijateljskih primjera te ne povećavaju robusnost mreže. Općenito, obrane koje su "agnostičke" što se tiče modela kojeg brane, odnosno nezavisne su od modela, su u pravilu neuspješne.

#### 4.1.1. JPEG kompresija

JPEG kompresija u obliku obrane od neprijateljskih primjera se pojavila više puta u različitim oblicima<sup>[8]</sup> [7] [10], a slijedi opis najjednostavnijeg načina zaštite. JPEG je vrsta sažimanja podataka s gubitkom (eng. *lossy*) za slike. Upravo to svojstvo je poželjno pri uništavanju neprijateljske perturbacije. Prepostavka je da su i sami neprijateljski primjeri osjetljivi na perturbacije, odnosno da će male promjene nad pažljivo-konstruiranom perturbacijom vratiti sliku nazad u domenu ispravne klasifikacije. Ovo dodatno ima smisla kada se uzme u obzir da su neprijateljski primjeri rijetki, to jest da ih nije moguće lako nasumično pronaći. JPEG kompresija bazira se na diskretnoj kosinusnoj transformaciji (eng. *discrete cosine transform*, DCT). Transformacijom iz prostorne domene u frekvencijsku domenu omogućava se direktna manipulacija frekvencijskom domenom. Ljudski vid nije toliko osjetljiv na komponente visoke frekvencije u slikama, stoga se provodi kvantizacija frekvencija i visoke frekvencije se čuvaju s manjom preciznošću nego niske frekvencije. Preciznost pri kojoj se čuvaju visoke frekvencije ovisi o faktoru kvalitete koji je u rasponu od 0 do 100: za 0 se visoke frekvencije potpuno odbacuju, a za 100 visoke frekvencije su maksimalno očuvane. Važno je uočiti da kvaliteta od 100 ne znači da je kompresija bez gubitka, jer se gubitak djelomično dogodi već pri prelasku u frekvencijsku domenu. Na slici 4.1 se nalazi usporedba kompresije za različite kvalitete.



**Slika 4.1:** Primjeri JPEG kompresije za različiti parametar kvalitete. Kvaliteta je redom 5, 25, 50, 90.

Najjednostavnija verzija obrane (i prva koja se pojavila) je korištenje JPEG kompresije samo pri evaluaciji. Autori su isprobali FGSM napad uz  $\epsilon \in \{1, 5, 10\}$ . Koraci su sljedeći:

- izračunati neprijateljske primjere za navedene  $\epsilon$
- provesti JPEG kompresiju nad običnim slikama i neprijateljskim primjerima
- usporediti izlaz modela za sve navedene skupove

Dodatno bi bilo zanimljivo vidjeti kako obrana utječe na DeepFool napad, pošto DeepFool generira puno manje perturbacije. Rezultati u tablici 4.1 su dobiveni na 16 slika iz 8.1.

| Ulaz                 | Kompresija     | Točnost |
|----------------------|----------------|---------|
| Standardno           | Bez kompresije | 100.00% |
|                      | Kvaliteta 50%  | 100.00% |
|                      | Kvaliteta 75%  | 100.00% |
| FGSM $\epsilon = 1$  | Bez kompresije | 31.25%  |
|                      | Kvaliteta 50%  | 68.75%  |
|                      | Kvaliteta 75%  | 43.75%  |
| FGSM $\epsilon = 5$  | Bez kompresije | 18.75%  |
|                      | Kvaliteta 50%  | 12.50%  |
|                      | Kvaliteta 75%  | 12.50%  |
| FGSM $\epsilon = 10$ | Bez kompresije | 12.50%  |
|                      | Kvaliteta 50%  | 18.75%  |
|                      | Kvaliteta 75%  | 12.50%  |
| DeepFool             | Bez kompresije | 0.00%   |
|                      | Kvaliteta 50%  | 81.25%  |
|                      | Kvaliteta 75%  | 64.50%  |

**Tablica 4.1:** Utjecaj JPEG kompresije na različite neprijateljske primjere. JPEG kompresija u ovom obliku jedino može pokvariti napade male magnitude, no čak i tada nije uvijek uspješna.

Zaključak je dakle da je jednostavna obrana temeljena na JPEG kompresiji daleko od korisnog rješenja, bar za FGSM napad. U slučaju DeepFool napada, obrana je puno uspješnija, međutim i tada nije u mogućnosti dosegnuti 100% točnost kao na čistim ulazima, te čak i napad s iznimno malom perturbacijom može zaobići JPEG kompresiju. Dodatno je problematično to što metoda nije u mogućnosti spriječiti nove napade, samo degradirati već postojeće neprijateljske primjere. Dapače, ponovi li se DeepFool napad uz JPEG kompresiju na ulazu, napad je opet uspješan 100% vremena na ovom skupu podataka.

Postoje i sofisticiraniji oblici JPEG kompresije kao metode obrane<sup>[7]</sup> od neprijateljskih primjera. Moguće je dodatno mrežu trenirati na slikama različite kva-

litete kompresije kako bi mreža mogla uspješno klasificirati i slike lošije kvalitete (koje često imaju artifakte). Ovaj proces autori nazivaju "cijepljenjem" mreže. Također je moguće imati onoliko modela koliko i različitih kvaliteta JPEG slika i konstruirati ansambl takvih modela. Međutim, u pravilu su obrane temeljene na ansamblima jake onoliko koliko i najjača komponenta ansambla, a pošto JPEG kompresija nije jaka obrana takav ansambl isto nije vrlo jak pri obrani od neprijateljskih primjera.

#### 4.1.2. Stiskanje značajki

Stiskanje značajki<sup>[22]</sup> (eng. *feature squeezing*) je generalni pojam koji se može definirati neovisno o neprijateljskim primjerima. Prostori u kojima se nalaze značajke su često bespotrebno veliki te je ideja smanjiti stupnjeve slobode pojedinih značajki i tako "istisnuti" manje bitne značajke. JPEG kompresija se isto može smatrati oblikom stiskanja značajki, no u kontekstu neprijateljskih primjera se termin odnosi na dvije metode: redukcija dubine boje slike i zagladijanje slike.

Dva najčešće korištена formata boje slike za klasifikaciju su RGB (npr. *CIFAR-10*, *ImageNet*) i sivi tonovi (eng. *grayscale*, npr. *MNIST*). Za RGB slike, svaki piksel je reprezentiran s  $3 \times 8 = 24$  bita, što daje  $2^{24}$  mogućih vrijednosti pojedinog piksela. Međutim, za klasifikaciju slika, nije potrebno imati toliko precizne informacije te ljudi mogu lako prepoznati što se na slici nalazi i s manjom dubinom boje. Smanjenjem broja dostupnih bitova se može pokvariti neprijateljska perturbacija i također u teoriji smanjiti prostor gdje se neprijateljski primjeri mogu nalaziti. Na slici 4.2 je primjer slike s različitim brojem bitova dostupnim za prikaz boje. Korišten broj bitova u originalnom radu je 4 i 5.



**Slika 4.2:** Ista slika s različitim brojem bitova za boju. Broj bitova je redom 8, 6, 4, 2 i 1.

Gaussovo zagladijanje je proces pri kojemu se slika zamućuje do određene mjere. Kao i kod prethodnih metoda, zamućenje može uništiti pažljivo stvorene perturbacije. Na slici 4.3 se nalazi primjer Gaussovog zamućivanja slike.



**Slika 4.3:** Zamućenje slike uz veličine prozora  $2 \times 2$ ,  $3 \times 3$  i  $4 \times 4$

Obje metode su poprilično neuspješne u uništavanju FGSM perturbacije jer je FGSM napad veće magnitude, dok uspješno uništava primjere drugih, jačih napada (npr. DeepFool) koji generiraju manje perturbacije. Ovo je slično kao i kod obrane uz JPEG kompresiju koja je opisana u dijelu 4.1.2. Međutim, kao što se kasnije ispostavilo<sup>[18]</sup>, obje obrane je vrlo lako zaobići. Potrebno je jednostavno povećati snagu pojedinih napada da bi se obrane zaobišle, a novonastala perturbacija je također nevidljiva. Postoji i jednostavno objašnjenje zašto Gaussovo zaglađivanje sigurno ne može biti uspješna obrana: zaglađivanje se provodi operacijom konvolucije, što se zapravo može promatrati kao dodatni konvolucijski sloj na ulazu mreže, a konvolucijske mreže su ionako slabe na napade. Stoga još jedan dodatni konvolucijski sloj (s fiksnim težinama) sigurno ne može puno toga napraviti da spriječi nastanak neprijateljskih primjera. Dodatno je zanimljivo i kako se zaglađivanje može koristiti i za generiranje neprijateljskih primjera: dovoljno je linijskim pretraživanjem pronaći najmanji faktor zaglađivanja za koji neka mreža pogriješi. Iznenadujuće je da za puno slika, distorzija potrebna za pogrešnu klasifikaciju nije velika. Međutim, ovo se može poboljšati tako da se pri treniranju uvedu i zamućene slike (eng. *data augmentation*).

## 4.2. Neprijateljsko treniranje - FGSM

Neprijateljsko treniranje se kao ideja pojavila zajedno s FGSM napadom<sup>[9]</sup>. Međutim, neprijateljsko treniranje je tada zamišljeno primarno kao nova metoda regularizacije modela i autori su usporedivali neprijateljsko treniranje s drugim metodama regularizacije (npr. *dropout* i *pretraining*), a ne kao obranu od potencijalnih neprijateljskih napada.

Neprijateljsko treniranje se fundamentalno razlikuje od dosadašnjih metoda povećanja skupa podataka (eng. *data augmentation*). Standardne metode uključuju

operacije kao što su rotacija, zrcaljenje, blago mijenjanje boja, brisanje dijelova slike – no tako transformirane slike i dalje ostaju u originalnoj distribuciji podataka. Zapravo je većina operacija i odabrana upravo iz tog razloga, jer se takve slike očekuju i u skupu podataka za testiranja. No neprijateljsko treniranje trenera model na primjerima koji se nikad ne bi pojavili u skupu za treniranje – neprijateljski primjeri se ne pojavljuju prirodno nego trebaju biti konstruirani.

Najjednostavniji oblik neprijateljskog treniranja je sljedeći:

- u svakom koraku treniranja se skup treniranja podijeli u dva skupa
- jedan skup ostane netaknut
- drugi skup se pretvori u neprijateljske primjere – ovo se provodi samo za ulaze koji su ispravno klasificirani

Omjer skupova je hiperparametar. U originalnom su radu autori odabrali omjer 1 : 1 koji radi dovoljno dobro, stoga je i ovdje u nastavku korišten isti omjer.

Za model je ovdje korišten konvolucijski model opisan u 2.4. Model treniran na standardan način nakon 25 epoha postigne točnost od 78.22%, a nakon 50 epoha postigne točnost od 79.80%. Za usporedbu, istreniran je model neprijateljskim treniranjem uz FGSM s  $\epsilon = 0.1$  (slike su u rasponu [0, 1], stoga je i  $\epsilon$  manji nego za slike u rasponu [0, 255]) te model uz FGSM s  $\epsilon \in \{0.05, 0.1, 0.2, 0.3\}$ . U tablici 4.2 se nalaze rezultati uspješnosti FGSM napada uz  $\epsilon \in \{0.05, 0.1, 0.2\}$  na različito trenirane modele.

| Treniranje                                  | Epoha | Čisti podatci | FGSM<br>$\epsilon = 0.05$ | FGSM<br>$\epsilon = 0.1$ | FGSM<br>$\epsilon = 0.2$ |
|---|-------|---------------|---------------------------|--------------------------|--------------------------|
| Standardno                                  | 25    | 78.22%        | 19.89%                    | 16.99%                   | 16.22%                   |
|   | 50    | 79.80%        | 22.05%                    | 16.85%                   | 12.32%                   |
| FGSM $\epsilon = 0.1$                       | 25    | 73.32%        | 44.33%                    | 66.53%                   | 23.24%                   |
|   | 50    | 74.96%        | 45.89%                    | 66.14%                   | 68.33%                   |
| FGSM $\epsilon = 0.2$                       | 25    | 74.04%        | 15.22%                    | 24.25%                   | 69.71%                   |
|   | 50    | 75.77%        | 16.80%                    | 34.47%                   | 71.11%                   |
| FGSM $\epsilon = 0.3$                       | 25    | 74.23%        | 15.02%                    | 18.88%                   | 56.67%                   |
|   | 50    | 77.76%        | 17.12%                    | 18.30%                   | 58.76%                   |
| FGSM $\epsilon \in \{0.05, 0.1, 0.2, 0.3\}$ | 25    | 77.54%        | 43.98%                    | 73.52%                   | 30.14%                   |
|   | 50    | 76.98%        | 66.82%                    | 69.88%                   | 69.87%                   |

**Tablica 4.2:** Prikazana je točnost modela na čistim podatcima i uz FGSM za različite  $\epsilon$ . Neprijateljsko FGSM treniranje ne generalizira preko različitih epsilona, te broj epoha igra bitnu ulogu kod većeg broja napada korištenih pri treniranju.

Neprijateljsko treniranje se čini kao korak u dobrom smjeru, ali ne u ovom obliku. Područje koje je potrebno pokriti je preveliko jer svakako postoji više neprijateljskih primjera nego legitimnih ulaza i jednostavno nije izvodljivo model trenirati na svim normama svakog napada. Bolji oblik neprijateljskog treniranja koji ipak može generalizirati nad različitim normama se dodatno razmatra u ?? s obećavajućim rezultatima.

### 4.3. Termometar kodiranje

Termometar kodiranje (eng. *thermometer encoding*) je obrana koja je direktno napala problem linearnosti dubokih modela<sup>[3]</sup>. Slično kao i prethodne obrane, nije potrebno puno promijeniti postojeći model, ali obrana svejedno nije besplatna kao jednostavnije obrane.

Jedan način da se poveća nelinearnost bi bila mijenjanje aktivacijskih funkcija mreža, no ReLU i ostale aktivacijske funkcije se koriste s razlogom: brze su i jednostavne za optimirati. Druga metoda bi bila da se postavi vrlo nelinearna transformacija na ulaz mreže. Na drugoj metodi se temelji termometar kodiranje.

Prvo je potrebno uvesti kvantizacijsku funkciju  $b$ . Potrebno je odabrati vrijednosti  $b_i$  takve da vrijedi  $0 < b_1 < b_2 < \dots < b_k = 1$ , npr.  $b_i = \frac{i}{k}$ . Za neki realni broj  $\theta \in [0, 1]$  definira se funkcija  $b(\theta)$  kao najmanji indeks  $\alpha \in \{1, \dots, k\}$  takav da je  $\theta \leq b_\alpha$ .

Slijedi opis *one-hot* kodiranja. Za neki indeks  $j \in \{1, \dots, k\}$ , neka je  $\chi(j) \in \mathbb{R}^k$  *one-hot* vektor od  $j$ :

$$\chi(j)_l = \begin{cases} 1 & \text{ako } l = j \\ 0 & \text{inače} \end{cases} \quad (4.1)$$

Diskretizacijska funkcija za neki piksel je onda:

$$f_{\text{onehot}}(x_i) = \chi(b(x_i)) \quad (4.2)$$

Međutim, *one-hot* kodiranje se nije pokazalo dobrom transformacijom ulaza za sprječavanje neprijateljskih primjera. Pretpostavka je da je zbog toga što se gubi svojstvo uređenosti između piksela, odnosno za svaku *one-hot* reprezentaciju vrijedi:

$$\|\chi(b(x_i))\|_2 = \|\chi(b(x_j))\|_2 = 1 \text{ kada vrijedi } b(x_i) \neq b(x_j) \quad (4.3)$$

Termometar kodiranje je vrlo slično *one-hot* kodiranju, no ne gubi se svojstvo uređenosti. Za neki indeks  $j \in \{1, \dots, k\}$ , neka je  $\tau(j) \in \mathbb{R}^k$  termometar vektor od  $j$  definiran kao

$$\tau(j)_l = \begin{cases} 1 & \text{ako } l \geq j \\ 0 & \text{inače} \end{cases} \quad (4.4)$$

Diskretizacijska funkcija za neki piksel je onda:

$$f_{\text{therm}}(x_i) = \tau(b(x_i)) \quad (4.5)$$

Za svaku termometar reprezentaciju vrijedi:

$$\|\tau(b(x_i))\|_2 < \|\tau(b(x_j))\|_2 = 1 \text{ kada vrijedi } b(x_i) \neq b(x_j) \text{ i } x_i < x_j \quad (4.6)$$

U tablici 4.3 je nekoliko primjera *one-hot* i termometar kodiranja.

| Ulaz | One hot      | Termometar   |
|------|--------------|--------------|
| 0.03 | [1000000000] | [1111111111] |
| 0.54 | [0000100000] | [0000111111] |
| 0.78 | [0000000100] | [0000000111] |
| 0.92 | [0000000001] | [0000000001] |

**Tablica 4.3:** Primjer kodiranja za  $b_i = \frac{i}{k}$  uz  $k = 10$ .

Da bi se neka mreža mogla koristiti s termometar kodiranjem, potrebno ju je ponovno istrenirati uz transformaciju ulaza. Nakon 30 epoha treniranja, mreža na čistim ulazima postiže točnost od 77.13%. U ovom obliku, mrežu nije više moguće napast uspješno s niti jednim od do sada spomenutih napada na model bijele kutije. Razlog tome je što nije moguće obaviti propagaciju unatrag kroz diskretizacijsku funkciju na ulazu. Svi napadi bijele kutije u standardnom formatu imaju uspješnost od približno 0%.

Autori su zato osmislili dva nova iterativna napada na model bijele kutije koje evauliraju na istreniranim modelima, te su tako pokazali uspješnost svoje obrane. Ova obrana je bila *state-of-the-art* obrana u trenutku kada se pojavila krajem 2017. godine. Međutim, ova i mnoge slične obrane koje pokušavaju diskretizacijom povećati nelinearnost mreže dijele zajednički problem koji ubrzano dolazi na vidjelo.

## 4.4. Obrambena destilacija

Obrambena destilacija (eng. *defensive distillation*) je učinkovita obrana koja se temelji na općenitijem pojmu destilacije dubokih neuronskih mreža. Destilacija je oblik treniranja mreže koji se provodi u dva koraka. U prvom koraku se standardno trenira neka mreža, no na ulaz u *softmax* se vrijednosti podijele s faktorom  $T$  koji se naziva temperatura. Izlaz *softmax* sloja se onda računa prema izrazu u jednadžbi 4.7. Temperatura  $T$  igra veliku ulogu u izlazu mreže – veća temperatura pridaje veću vjerojatnost svim izlazima, i za  $T \rightarrow \infty$  vjerojatnost za sve razrede teži  $\frac{1}{K}$ .

$$\sigma(\mathbf{x})_i = \frac{e^{\mathbf{x}_i/T}}{\sum_{j=1}^K e^{\mathbf{x}_j/T}} \text{ za } i = 1, \dots, K \text{ i } \mathbf{x} \in \mathbb{R}^K \quad (4.7)$$

Nakon treniranja prve mreže se trenira druga mreža na isti način, ali se u

skupu podataka za treniranje labele zamijene s izlazom već istrenirane mreže. Stoga se *one-hot* kodirane labele mijenjaju s vjerojatnostima koje je vratila pretvodno istrenirana mreža. Ponovno se ponovi postupak treniranja uz temperaturu  $T$ . Nakon treniranja, za vrijeme testiranja, temperatura  $T$  se ukloni, odnosno postavi na  $T = 1$ . Rezultat je destilirana mreža.

U originalnoj formulaciji destilacije mreže, prva mreža je kompleksnija i s većim kapacitetom, dok je druga mreža jednostavnija. Učenje jednostavnije mreže na vjerojatnostima uz temperaturu  $T$  se pokazalo korisnim jer jednostavnija mreža uspije postići rezultate neke složenije mreže, a k tome je i brža pri računanju izlaza.

U kontekstu obrambene destilacije, obje mreže mogu biti iste. Temperatura je najbitniji parametar, te su autori pokazali da je dobra temperatura  $T > 10$ , a najbolje rezultate su postigli za  $T = 100$ . U tablici 4.4 su prikazani rezultati različitih FGSM napada te DeepFool napada na destilirane mreže uz temperature  $T \in \{50, 100\}$ , te rezultati napada na čistu mrežu. Korištena mreža je ponovno *CIFAR-10* mreža opisana u 2.4, a treniranje je trajalo 50 epoha. Mreža trenirana uz  $T = 1$  nije destilirana, tu je provedeno standardno treniranje.

| Temperatura | Čisti podatci | FGSM           | FGSM           | FGSM            | DeepFool |
|-------------|---------------|----------------|----------------|-----------------|----------|
|             |               | $\epsilon = 2$ | $\epsilon = 5$ | $\epsilon = 10$ |          |
| $T = 1$     | 73.23%        | 45.80%         | 32.37%         | 22.86%          | 31.20%   |
| $T = 50$    | 79.27%        | 76.17%         | 74.90%         | 71.10%          | 79.00%   |
| $T = 100$   | 81.06%        | 79.71%         | 79.31%         | 76.91%          | 82.00%   |

**Tablica 4.4:** Točnost za obrambeno destilirane mreže uz različite napade. Mreža za  $T = 1$  je trenirana standardno. Napadi imaju iznimno loše rezultate na ovako treniranim mrežama. DeepFool napad je testiran na samo 500 primjera.

I ova obrana se pokazala iznimno uspješnom na napade na model bijele kutije. Za razliku od termometar kodiranja, ovdje je moguće provesti propagaciju unazad, a rezultati su svejedno izvrsni. Nažalost, kao i sve ranije opisane obrane, i ova obrana ima prikrivenu manu.

# 5. Neprijateljski primjeri II

## 5.1. Neučinkovitost obrana

Obrane opisane do sada su se sve pokazale neučinkovitim. Jednostavne transformacije ulaza redovito nisu dovoljne da pokvare perturbacije, a čak i kad jesu, pokazalo se da je moguće konstruirati bolje perturbacije koje su otporne na različite transformacije kao što su JPEG kompresija, zaglađivanje i redukcija dubine boje. Neprijateljsko treniranje uz FGSM nije u mogućnosti generalizirati za različite vrijednosti  $\epsilon$ , iako je korak u dobrom smjeru. No obrane kao što su termometar kodiranje i defenzivna destilacija su se pokazale posebno uspješnima.

Maskiranje gradijenata je pojava kod koje model ne sadrži korisne gradijente. Za obrane koje su dizajnirane tako da (namjerno ili slučajno) maskiraju gradijente se kaže da skrivaju gradijente (eng. *gradient obfuscation*)<sup>[2]</sup>. Nije svaki oblik maskiranja gradijenata također i skrivanje gradijenata – npr. mreža može naučiti maskirati gradijente, kao što je slučaj kod mnogo oblika neprijateljskog treniranja<sup>[20]</sup>.

Oblik skrivanja gradijenata kod obrana koje: nisu diferencijabilne (kao termometar kodiranje), su numerički nestabilne, ili na bilo koji način daju neispravne gradijente naziva se *razbijanje gradijenata* (eng. *gradient shattering*)<sup>[2]</sup>. U ovu kategoriju spadaju termometar kodiranje i defenzivna destilacija. Specifično, termometar kodiranje nije diferencijabilno i stoga se ne mogu izravno izračunati gradijenti mreže, što sprječava mnogo napada koji se na tome temelje.

Za obrambenu destilaciju je problem malo drugačiji. Temperatura  $T$  u *softmax* funkciji efektivno tjera mrežu da izlaze prethodnog sloja skalira za faktor  $T$ . Kada se pri testiranju opet uzme temperatura  $T = 1$ , uz velike vrijednosti prethodnog sloja, izlaz mreže postane  $\epsilon$  za sve neispravne izlaze, te  $1 - N \cdot \epsilon$  za ispravni izlaz. Ispostavilo se da je u većini slučajeva  $\epsilon$  toliko mal, da se u 32 bitnoj aritmetici s pomicnim zarezom ta vrijednost zaokruži na 0 što posljedično i gradijente pretvori u 0 i tako spriječi napade koji se temelje na gradijentima.

Drugi česti problem kod velikog broja ranijih metoda obrane je također to što je njihova evaluacija bila slaba ili nepotpuna. Obrane bi često bile evaluirane na vrlo jednostavnim mrežama, jednostavnim skupovima podataka (*CIFAR-10* i *MNIST*), korištenjem samo najjednostavnijih napada (L-BFGS, FGSM) s nedovoljno dobrim hiperparametrima. Česti problemi evaluacije robusnosti modela su detaljnije opisani u 6.1. Kao odgovor na neispravno evaluirane obrane su se pojavili jači napadi koji su korišteni za opovrgavanje uspješnosti tih obrana. U nastavku slijede dva vrlo jaka napada na model bijele kutije, te jedan zanimljivi napad na model crne kutije.

## 5.2. Projicirani gradijentni spust

Projicirani gradijentni spust (eng. *projected gradient descent*) se pojavio kontekstu neprijateljskog treniranja koje je detaljno opisano u 6.2. Općenito, projicirani gradijentni spust je oblik gradijentnog spusta koji dodatno uzima u obzir ograničenja. Razlika je u tome što je nakon svakog pomaka po gradijentu potrebno obaviti projekciju takvu da su ograničenja zadovoljena.

FGSM napad u jednom koraku izračuna neprijateljski primjer na sljedeći način:

$$\tilde{\mathbf{x}} = \mathbf{x} + \epsilon \operatorname{sign}(\nabla_{\mathbf{x}} J(\boldsymbol{\theta}, \mathbf{x}, y)) \quad (5.1)$$

FGSM se može promatrati kao jedan korak iterativnog napada:

$$\mathbf{x}^{t+1} = \Pi_{\mathbf{x}+S}(\mathbf{x}^t + \epsilon \operatorname{sign}(\nabla_{\mathbf{x}} J(\boldsymbol{\theta}, \mathbf{x}, y))) \quad (5.2)$$

Operator  $\Pi_{\mathbf{x}+S}$  predstavlja projekciju na  $\mathbf{x} + S$ . Skup  $S \subseteq \mathbb{R}^d$  je skup dopuštenih perturbacija.  $\mathbf{x} + S$  je  $\ell_\infty$  kugla oko  $\mathbf{x}$  unutar koje se traže neprijateljski primjeri tako da minimiziraju funkciju  $J(\boldsymbol{\theta}, \mathbf{x}, y)$ .

Iako jednostavan, projicirani gradijentni spust je mnogo jači od FGSM i pokazuje svojstva koja ga čine boljim za neprijateljsko treniranje. Već nakon 10 iteracija uz  $\epsilon = 1$  uspješno bude stvoreno 14/16 neprijateljskih primjera za skup 8.1. U tablici 5.1 su predviđanja mreže za tako stvorene primjere.

| Slika | Izlaz mreže            | Vjerojatnost |
|-------|------------------------|--------------|
| 8.1a  | African grey*          | 100.00%      |
| 8.1b  | Soccer ball            | 100.00%      |
| 8.1c  | Damselfly              | 99.94%       |
| 8.1d  | Kite                   | 99.99%       |
| 8.1e  | Banana                 | 99.99%       |
| 8.1f  | Orangutan              | 99.99%       |
| 8.1g  | Pembroke               | 98.08%       |
| 8.1h  | Shield                 | 99.14%       |
| 8.1i  | Dungeness crab         | 63.76%       |
| 8.1j  | Space bar              | 51.36%       |
| 8.1k  | French bulldog         | 99.98%       |
| 8.1l  | Meerkat*               | 98.92%       |
| 8.1m  | Welsh springer spaniel | 99.69%       |
| 8.1n  | Neck brace             | 100.00%      |
| 8.1o  | Egyptian cat           | 99.87%       |
| 8.1p  | Shower curtain         | 99.99%       |

**Tablica 5.1:** Tablica top-1 izlaza mreže za neprijateljske primjere stvorene PGD algoritmom nakon 10 iteracija uz  $\epsilon = 1$ . Slike označene sa zvjezdicom (\*) nisu uspješno pretvorene u neprijateljske primjere.

### 5.3. Napadi *Carlini and Wagner*

Obrambena destilacija naizgled se činila kao vrlo otporna obrana. Problem s gradijentima opisan u 5.1 nije inicijalno bio uočen. Autori Carlini i Wagner se zato vraćaju na početnu definiciju problema neprijateljskih primjera (definicija je dana u 3.2).

candw<sup>[4]</sup>

### 5.4. Granični napad

Do sada su bili razmatrani samo napadi na model bijele kutije. Takvi napadi su u pravilu jači od napada na model crne kutije jer mogu iskoristiti dodatno znanje o mreži da poboljšaju napad. Napadi na model crne kutije ne znaju ništa o mreži

osim vrijednosti izlaza. U praksi, modeli o kojima je znanje minimalno su česti, na primjer modeli kojima se pristupa preko nekakvog web sučelja (npr. *Microsoft Azure* i slični servisi). Usprkos manjku znanja o mreži, moguće je konstruirati vrlo uspješne napade. Jedan od napada na model crne kutije je granični napad (eng. *boundary attack*). Napad je temeljen na odluci (eng. *decision based attack*), odnosno zahtjeva samo labelu na izlazu mreže. Za razliku od napada temeljenog na odluci, neki napadi crne kutije očekuju i vjerojatnosti na izlazu.

Napad započinje iz neke točke koja već je neprijateljska (odnosno pogrešno klasificirana) i iterativno se približava zadanim ulazu  $\mathbf{x}$  pod uvjetom da je u svakom koraku i dalje neprijateljska. Najjednostavniji opis algoritma je dan u nastavku.

**Podatci:** ulazna slika  $\mathbf{o}$

**Rezultat:** neprijateljski primjer  $\tilde{\mathbf{o}}$

inicijalizacija:  $k = 0$ ,  $\tilde{\mathbf{o}}^0 \sim \mathcal{U}(0, 1)$  takav da je neprijateljski klasificiran;

**dok**  $k < \text{maksimalni broj koraka čini}$

odabere se nasumična perturbacija iz distribucije  $\boldsymbol{\eta}_k \sim \mathcal{P}(\tilde{\mathbf{o}}^{k-1})$ ;

**ako je**  $\tilde{\mathbf{o}}^{k-1} + \boldsymbol{\eta}_k$  neprijateljski onda

postavi  $\tilde{\mathbf{o}}^k = \tilde{\mathbf{o}}^{k-1} + \boldsymbol{\eta}_k$ ;

**inače**

postavi  $\tilde{\mathbf{o}}^k = \tilde{\mathbf{o}}^{k-1}$ ;

**kraj**

$k = k + 1$

**kraj**

**Algoritam 1:** Najjednostavniji oblik graničnog napada. Potreban je samo izlaz mreže.

Ukoliko je napad neciljani, ulazna slika može biti neki šum (pod uvjetom da je šum pogrešno klasificiran). Za ciljani napad, ulazna slika treba biti slika iz ciljanog razreda. Ključni dio algoritma je distribucija  $\mathcal{P}$ . Optimalna distribucija bi se u pravilu trebala razlikovati između domena problema i različitih modela, no iznenađujuće je kako je moguće uspješno konstruirati neprijateljske primjere uz jednostavnu distribuciju. Perturbacije se generiraju iz distribucije s maksimalnom entropijom uz sljedeća ograničenja:

1. Perturbirani uzorak je i dalje u domeni, odnosno:

$$\tilde{\mathbf{o}}_i^{k-1} + \boldsymbol{\eta}_i^k \in [0, 255] \quad (5.3)$$

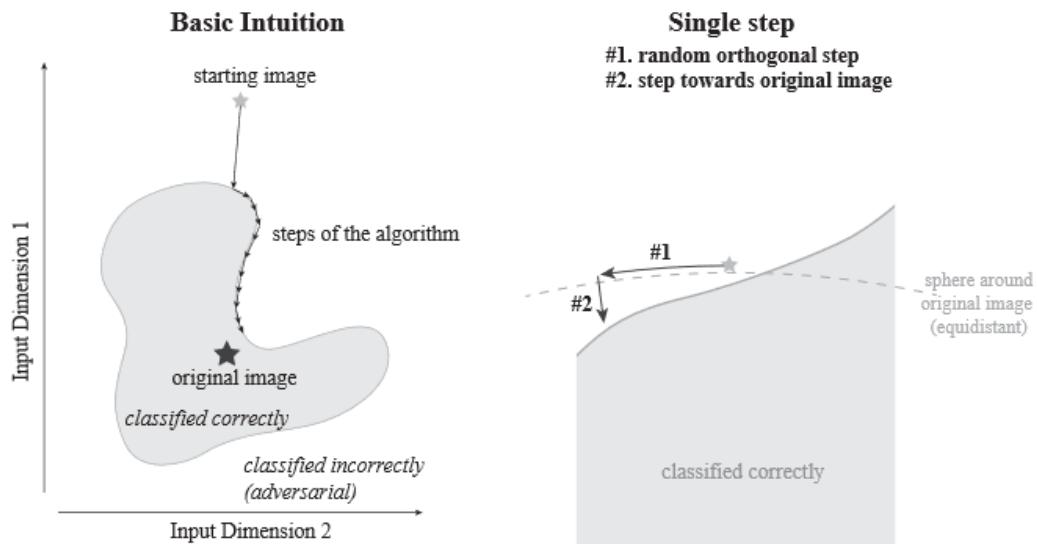
2. Perturbacija ima relativnu veličinu iznosa  $\delta$ , odnosno:

$$\|\boldsymbol{\eta}^k\|_2 = \delta \cdot d(\mathbf{o}, \tilde{\mathbf{o}}^{k-1}), \text{ gdje je } d(\mathbf{o}, \tilde{\mathbf{o}}) = \|\mathbf{o} - \tilde{\mathbf{o}}\|_2^2 - \text{udaljenost između } \mathbf{o} \text{ i } \tilde{\mathbf{o}} \quad (5.4)$$

3. Perturbacija smanjuje udaljenost između ciljane slike i perturbirane slike za neki relativni iznos, odnosno:

$$d(\mathbf{o}, \tilde{\mathbf{o}}^{k-1}) - d(\mathbf{o}, \tilde{\mathbf{o}}^{k-1} + \boldsymbol{\eta}^k) = \epsilon \cdot d(\mathbf{o}, \tilde{\mathbf{o}}^{k-1}) \quad (5.5)$$

Autori su priložili ilustraciju koja slikovito prikazuje rad algoritma. Ilustracija je prikazana na slici 5.1.



**Slika 5.1:** Na lijevoj slici je prikazan rad algoritma kroz nekoliko koraka. U svakom koraku se neprijateljski primjer približava slici, ali tako da uvijek bude izvan granice ispravne klasifikacije. Na desnoj slici je prikazan jedan korak algoritma. Prvo se napravi nasumičan korak na sferi oko originalne slike, a nakon toga se napravi korak prema originalnoj slici.

Nije lako uzorkovati tu distribuciju, ali je moguće konstruirati jednostavniju heuristiku koja je dovoljno dobra. Prvo se uzorkuje normalna distribucija  $\boldsymbol{\eta}_i^k \sim \mathcal{N}(0, 1)$  nakon čega se uzorak skalira i ograniči tako da zadovoljava ograničenja 5.3 i 5.4. U idućem koraku se radi pomak po sferi oko  $\mathbf{o}$  u nasumičnom smjeru, odnosno mora vrijediti  $d(\mathbf{o}, \tilde{\mathbf{o}}^{k-1} + \boldsymbol{\eta}^k) = d(\mathbf{o}, \tilde{\mathbf{o}}^{k-1})$  i ograničenje 5.3. U zadnjem koraku se pomakne prema originalnoj slici tako da ograničenja 5.3 i 5.5 budu

zadovoljena. Za visoko-dimenzionalne ulaze i dovoljno male  $\delta$  i  $\epsilon$ , ograničenje 5.4 će biti praktički zadovoljeno.

Postavlja se pitanje što je s lokalnim minimumima. Lokalni minimumi predstavljaju problem utoliko što na višestruko pokretanje algoritma se dogodi konvergencija prema različitim minimumima, ali ti minimumi su sličnog reda veličine i ne događa se slučaj da algoritam zapne u nekom lokalnom minimumu do te mjere da se ne postigne vizualno prihvatljiv neprijateljski primjer. Na slici 5.2 su tri primjera napada.



**Slika 5.2:** Tri primjera graničnog napada na *ResNet* mrežu. U prvom redu su originalne slike, u drugom redu je rezultat nakon 400 koraka, a u trećem redu je rezultat nakon 8000 koraka. Algoritam jako brzo nađe šumovito rješenje, a onda većinu vremena optimizira neprijateljski primjer, kao što je pokazano na slici 5.1. Za svaki pojedini stupac, izlaz modela je isti.

Što se tiče svih prethodno opisanih obrana, za sve postoji napad na model bijele kutije koji ih može poraziti. Slijedi primjer ovog napada na obrane neprijateljskog treniranja (uz FGSM) i obrambenu destilaciju. Napad na obrambenu distilaciju uz  $T = 100$  je na slici 5.3. Napad na neprijateljski treniranu mrežu uz FGSM je na slici 5.4. Napad nije primijenjen na termometar kodiranje jer nije trivijalno osigurati da je novonastali neprijateljski primjer također u formatu za termometar kodiranje, što onemogućava da se primjer dekodira.



**Slika 5.3:** Napad na obrambenu destilaciju. Slike su nakon svake desete iteracije. Zadnja slika je nakon 2500 iteracija uz  $\ell_2 = 107.63$ .



**Slika 5.4:** Napad na neprijateljsko treniranje uz FGSM. Slike su nakon svake desete iteracije. Zadnja slika je nakon 2500 iteracija uz  $\ell_2 = 101.51$ .

# 6. Obrana dubokih konvolucijskih modela II

## 6.1. Preduvjeti uspješnih obrana

evaluation<sup>[5]</sup>

## 6.2. Neprijateljsko treniranje - PGD

pgd<sup>[13]</sup>

### 6.2.1. *Fast is better than free*

fbf<sup>[21]</sup>

## 6.3. Dokazivost obrane od neprijateljskih napada

## 6.4. Budući rad

## 7. Zaključak

Konvolucijske neuronske mreže su se pokazale iznimno uspješnima pri rješavanju problema klasifikacije slika. Međutim, pojava neprijateljskih primjera dovela je u pitanje tvrdnju da današnji modeli dobro generaliziraju. Također se ispostavilo da je vrlo lako osmisliti uspješne algoritme koji mogu generirati neprijateljske primjere, a za neke napade čak nije potrebno nikakvo dodatno znanje o mreži koju se napada. Ubrzo nakon pojave neprijateljskih primjera su se pojavile i prve potencijalne obrane. Iako su sve obrane na prvi pogled izgledale obećavajuće, pokazano je da većina tih obrana također ne mogu generalizirati na jače napade. Od svih obrana do sada, gotovo niti jedna obrana nije se pokazala korisnom. Najviše obećavajuća obrana do sada temelji se na posebnom obliku treniranja mreže korištenjem neprijateljskih primjera, no to je ono što tu obranu čini skupom i neuporabljivom na većim skupovima podataka kao što je *ImageNet*. Budućnost obrana od neprijateljskih primjera još uvijek nije očita. Praktički svaki dan se pojavljuju sve efikasnije metode temeljene na suparničkom treniranju, formalni postupci koji mogu dokazati robusnost mreže te potpuno nove metode obrana koje će tek biti detaljno testirane. Međutim, danas obrana dubokih konvolucijskih modela još uvijek stoji kao vrlo izazovan i neriješen problem.

# LITERATURA

- [1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, stranice 265–283, 2016.
- [2] Anish Athalye, Nicholas Carlini, i David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *ArXiv*, abs/1802.00420, 2018.
- [3] Jacob Buckman, Aurko Roy, Colin Raffel, i Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. 2018. URL <https://openreview.net/pdf?id=S18Su--CW>.
- [4] Nicholas Carlini i David A. Wagner. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, stranice 39–57, 2017.
- [5] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian J. Goodfellow, Aleksander Madry, i Alexey Kurakin. On evaluating adversarial robustness. *ArXiv*, abs/1902.06705, 2019.
- [6] François Chollet et al. Keras. <https://keras.io>, 2015.
- [7] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E. Kounavis, i Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *ArXiv*, abs/1705.02900, 2017.
- [8] Gintare Karolina Dziugaite, Zoubin Ghahramani, i Daniel M. Roy. A study of

- the effect of jpg compression on adversarial images. *ArXiv*, abs/1608.00853, 2016.
- [9] Ian J. Goodfellow, Jonathon Shlens, i Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2015.
  - [10] Chuan Guo, Mayank Rana, Moustapha Cissé, i Laurens van der Maaten. Countering adversarial images using input transformations. *ArXiv*, abs/1711.00117, 2018.
  - [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, i Jian Sun. Identity mappings in deep residual networks. *ArXiv*, abs/1603.05027, 2016.
  - [12] Alex Krizhevsky, Vinod Nair, i Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL <http://www.cs.toronto.edu/~kriz/cifar.html>.
  - [13] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, i Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2017.
  - [14] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, i Ben Edwards. Adversarial robustness toolbox v1.2.0. *CoRR*, 1807.01069, 2018. URL <https://arxiv.org/pdf/1807.01069>.
  - [15] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hamardzumyan, Zhi-shuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, i Rujun Long. Technical report on the cleverhans v2.1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768*, 2018.
  - [16] Jonas Rauber, Wieland Brendel, i Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*, 2017. URL <http://arxiv.org/abs/1707.04131>.

- [17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, i Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3): 211–252, 2015. doi: 10.1007/s11263-015-0816-y.
- [18] Yash Sharma i Pin-Yu Chen. Bypassing feature squeezing by increasing adversary strength. *ArXiv*, abs/1803.09868, 2018.
- [19] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, i Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2014.
- [20] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, i Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. *ArXiv*, abs/1705.07204, 2018.
- [21] Eric Wong, Leslie Rice, i J. Zico Kolter. Fast is better than free: Revisiting adversarial training. *ArXiv*, abs/2001.03994, 2020.
- [22] Weilin Xu, David Evans, i Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *ArXiv*, abs/1704.01155, 2018.

# 8. Dodatak

## 8.1. Osobni skup slika

U skupu slika 8.1 je 16 slika i labela koje su u radu korištene za generiranje suparničkih primjera za modele trenirane na *ImageNet* skupu podataka. Sve slike su odabранe tako da spadaju unutar 1000 razreda za koje su modeli predviđeni, dakle slike bi trebale biti ispravno klasificirane. Također, sve slike su unaprijed izrezane u oblik kvadrata kako ne bi došlo do prevelike degradacije kvalitete pri smanjenju rezolucije na  $224 \times 224$ . Ovim putem se zahvaljujem Sarah James na dopuštenju za korištenje slika u radu.

## 8.2. Izlazi modela na nepromijenjenim slikama iz osobnog skupa

Slike iz 8.1 su birane tako da budu reprezentativne, odnosno da ne postoji mogućnost da su mreže "zbunjene" oko toga što je na slici. Ideja iza toga je pokazati kako je neprijateljske primjere lako konstruirati čak i kada je mreža iznimno sigurna u to što vidi na slici. U tablici 8.1 nalaze se top-1 izlazi mreža *ResNet50 V2* (primarna mreža korištena kroz rad) i *Xception*.

| Slika | ResNet50 V2              | Xception                 |
|-------|--------------------------|--------------------------|
| 8.1a  | African grey 100.00%     | African grey 100.00%     |
| 8.1b  | Backpack 99.71%          | Backpack 99.99%          |
| 8.1c  | Dragonfly 99.93%         | Dragonfly 99.79%         |
| 8.1d  | Bald eagle 87.47%        | Bald eagle 99.62%        |
| 8.1e  | Grocery store 90.81%     | Grocery store 97.68%     |
| 8.1f  | Gorilla 99.83%           | Gorilla 99.96%           |
| 8.1g  | Guinea pig 100.00%       | Guinea pig 99.99%        |
| 8.1h  | Jack-o'-lantern 100.00%  | Jack-o'-lantern 99.89%   |
| 8.1i  | Jigsaw puzzle 100.00%    | Jigsaw puzzle 100.00%    |
| 8.1j  | Computer keyboard 65.34% | Computer keyboard 99.68% |
| 8.1k  | Llama 100.00%            | Llama 100.00%            |
| 8.1l  | Meerkat 99.98%           | Meerkat 99.89%           |
| 8.1m  | English Springer 88.93%  | English Springer 99.30%  |
| 8.1n  | Running shoe 80.76%      | Running shoe 99.48%      |
| 8.1o  | Tabby 64.28%             | Tabby 89.72%             |
| 8.1p  | Wardrobe 87.16%          | Wardrobe 79.05%          |

**Tablica 8.1:** Tablica top-1 izlaza mreža ResNet50 V2 i Xception za slike iz 8.1.



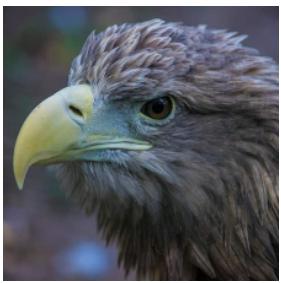
(a) *African grey* (87)



(b) *Backpack* (414)



(c) *Dragonfly* (319)



(d) *Bald eagle* (22)



(e) *Grocery store* (582)



(f) *Gorilla* (366)



(g) *Guinea pig* (338)



(h) *Jack-o'-lantern* (607)



(i) *Jigsaw puzzle* (611)



(j) *Keypad* (508)



(k) *Llama* (355)



(l) *Meerkat* (299)



(m) *English springer*  
(217)



(n) *Running shoe* (770)



(o) *Tabby* (281)



(p) *Wardrobe* (894)

**Slika 8.1:** Slike korištene za generiranje suparničkih primjera te njihove ispravne labele

# Obrana dubokih konvolucijskih modela od neprijateljskih primjera

## Sažetak

Današnji konvolucijski modeli postižu visoku točnost u području raspoznavanja objekata. Način rada dubokih modela je još uvijek vrlo teško ili nemoguće interpretirati, a dodatan razlog za brigu predstavljaju i takozvani neprijateljski primjeri. Neprijateljski primjeri su slike s dodanim teško uočljivim perturbacijama koje potiču model na pogrešnu klasifikaciju. Pokazalo se da je vrlo lagano konstruirati brze i efikasne napade na postojeće modele, nakon čega su se ubrzo pojavile i obrane protiv takvih napada. Nedugo nakon toga pojavljuju se sve jači napadi, dok su obrane stagnirale te danas još uvijek ne postoji zadovoljavajuća obrana protiv neprijateljskih napada. U radu je dan pregled spomenutih jednostavnih napada i jednostavnih obrana, istaknuta je neuspješnost jednostavnih obrana protiv jačih napada te je opisana obećavajuća obrana koja se temelji na treniranju mreže korištenjem neprijateljskih primjera.

**Ključne riječi:** klasifikacija objekata, konvolucijske neuronske mreže, računalni vid, suparnički primjeri, neprijateljski primjeri, obrana

# Defending Deep Convolutional Models from Adversarial Examples

## Abstract

Today's convolutional models can achieve high accuracy in the field of object recognition. The way deep models work is still very difficult or impossible to interpret, and an additional reason for concern are the so-called adversarial examples. Adversarial examples are images with added imperceptible perturbations that encourage the model to misclassify the image. It turned out to be very easy to construct fast and effective attacks on existing models, after which new defenses against these attacks also appeared. Soon afterwards even stronger attacks appeared, while new defenses stagnated and today there isn't a satisfactory defense against adversarial attacks. The thesis reviews the aforementioned simple attacks and simple defenses, pointing out the failure of simple defenses against strong attackers. Additionally, a promising defense is described. The defense is based on the concept of adversarial training and has shown good results.

**Keywords:** object classification, convolutional neural networks, computer vision, adversarial attacks, defense