SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Neprijateljski primjeri na ImageNet skupu podataka

Marko Opačić
Voditelj: *Marko Đurasević*

Zagreb, svibanj 2023.

SADRŽAJ

1. Uvod **1**

2. Neprijateljski primjeri **2**

2.1. Definicija 2

3. Metode nalaženja neprijateljskih primjera **3**

3.1. Fast Gradient Sign Method 3

3.2. Projected Gradient Descent 4

4. Rezultati **5**

4.1. Parametri i ulazni podaci 5

4.2. ResNet50 6

4.3. VGG-16 7

4.4. EfficientNet-B4 8

4.5. Transferabilnost napada 10

5. Obrane od neprijateljskih primjera **11**

5.1. Trening s neprijateljskim primjerima 11

5.2. Ostale metode obrane 12

6. Zaključak **13**

7. Literatura **14**

8. Sažetak **16**

1. Uvod

Duboki modeli doživjeli su ogroman porast u popularnosti u prethodnom desetljeću, te su dosegli izvrsne rezultate na raznim problemima, uključujući i klasifikaciju slika. Kako se duboki modeli sve vise ugrađuju u sustave u stvarnom svijetu, robusnost tih modela postaje sve važnija. Osim klasične robusnosti i dobre generalizacije u prosječnom slučaju, sve je bitnije i da su modeli robusni u slučaju napada neprijateljskih aktera.

U ovom radu opisuju se metode generiranja neprijateljskih primjera, a u napadima se koristi projicirani gradijentni spust, koji se smatra najjačim napadom pod pretpostavkom da su dostupni gradijenti. Sagledavaju se rezultati konkretnog napada na tri modela trenirana na ImageNet skupu podataka, ResNet50, VGG-16 i EfficientNet-B4. Ukratko se opisuju metode obrane od napada.

2. Neprijateljski primjeri

2.1. Definicija

Ideja je da se na ulaznu sliku x dodaje sum, odnosno perturbacijski vektor δ, te se tako dobiva neprijateljski primjer $x_{adv} = x + \delta$. Tražimo primjere koji će zavarati duboki model, ali ne bi trebali zavarati ljude, tako da je perturbaciju δ potrebno na neki način ograničiti da ne bi bila primjetna ljudskom oku.

Za takav skup perturbacija se često odabire L_∞ kugla, definirana skupom:

$$\Delta = \{\delta \in \mathbb{R}^n : ||\delta||_\infty \leq \epsilon\}$$

gdje je n dimenzionalnost ulaznog vektora, ϵ parametar koji određuje veličinu perturbacije, a L_∞ norma vektora δ je definirana kao:

$$||\delta||_\infty = \max_i |\delta_i|.$$

Potrebno je i pobrinuti se da vektor $x + \delta$ ne izađe iz intervala dopuštenih vrijednosti ulaznog vektora, odnosno da vrijedi $x_i + \delta_i \in [0, 1]$ za sve i.

Cilj nam je da model pogriješi u klasifikaciji, odnosno da predvidi krivi razred. Za dani primjer x i točan razred y, cilj nam je da model predvidi razred y' koji je različit od točnog razreda. Preciznije, želimo da za predikciju modela vrijedi:

$$h_\theta(x + \delta) = y'$$

gdje $y \neq y'$. Postoje razne metode generiranja neprijateljskih primjera, a razmatramo neke od njih u sljedećem poglavlju.

3. Metode nalaženja neprijateljskih primjera

U ovom radu razmatraju se metode bazirane na gradijentima, koje pretpostavljaju da su poznate arhitektura modela i vrijednosti parametara. Takve metode nazivaju se napadima na bijele kutije (tzv. *whitebox* napadi). Metode napada na tzv. crne kutije gdje gradijenti nisu poznati, koje uključuju korištenje genetičkog algoritma ili slučajne potrage, izvan su okvira ovog rada.

Osim *whitebox-blackbox* podjele, jedna od glavnih razlika u metodama generiranja neprijateljskih primjera je korištena L_p norma perturbacije. Metode u nastavku koriste opisanu L_∞ normu. Modas et al. (2019) koriste L_0 normu u SparseFool algoritmu. Chen et al. (2018) u ElasticNet napadu koriste kombinaciju L_1 i L_2 norme. Carlini i Wagner (2017) u C&W napadu koriste L_2 normu.

3.1. Fast Gradient Sign Method

Goodfellow et al. (2014) opisali su efikasnu metodu za generiranje neprijateljskih primjera. Njihova hipoteza je da su duboki modeli podložni napadima zbog svoje linearnosti, te da se gradijent funkcije gubitka u odnosu na ulaznu sliku može iskoristiti kako bi se generirao neprijateljski primjer. Konkretno, opisuju perturbaciju:

$$\delta = \epsilon \cdot sign(\nabla_x \mathcal{L}(\theta, x, y))$$

gdje je $\mathcal{L}$ funkcija gubitka, θ parametri modela, x ulazna slika, y točan razred ulazne slike, a ϵ parametar koji određuje veličinu perturbacije. Ovakva perturbacija mijenja piksele ulazne slike u smjeru predznaka gradijenta funkcije gubitka po ulaznoj slici, te tako povećava vrijednost funkcije gubitka.

Ova metoda je efikasna jer zahtjeva samo jedan korak promjene ulazne slike, a pokazala je dobre rezultate na raznim modelima i skupovima podataka. Ipak, nedostatak ove metode je nedostatak kontrole nad konačnim izlazom modela. Metoda ažurira

vrijednosti ulazne slike u smjeru koji povećava gubitak i tako često uspješno zavarava model, no ne možemo odrediti u koji razred će model klasificirati neprijateljski primjer.

3.2. Projected Gradient Descent

Ideja je da umjesto jednog koraka koristimo vise koraka ažuriranja perturbacije ovisno o gradijentu po ulaznoj slici. U svakoj iteraciji pomoću backpropagation algoritma računamo gradijent $\nabla_x \mathcal{L}$, te zatim radimo korak u smjeru gradijenta. Ako je veličina koraka takva da neki element vektora $x + \delta$ izađe iz L_∞ kugle, tada korak u smjeru gradijenta projiciramo na L_∞ kuglu. Zbog ovog koraka projekcije metoda je i dobila svoj naziv (engl. *projected gradient descent*, PGD). Općenito, PGD zapravo opisuje metodu optimizacije, ne nužno metodu generiranja neprijateljskih primjera.

Osim sto sad koristimo iterativni postupak, sad nam je cilj odrediti i ciljani razred, odnosno razred u koji želimo da model klasificira neprijateljski primjer. Kao i kod FGSM, želimo maksimizirati gubitak na točnom razredu, ali osim toga sad želimo i minimizirati gubitak na ciljanom razredu. Rješavamo optimizacijski problem:

$$\max_{\delta \in \Delta} (\mathcal{L}(h_\theta(x + \delta), y) - \mathcal{L}(h_\theta(x + \delta), y_{target}))$$

gdje je y_{target} ciljana klasa, a Δ skup dopuštenih perturbacija. Korak $t + 1$ ažuriranja perturbacije definiran je kao:

$$x^{t+1} = \Pi_\Delta(x^t + \alpha \cdot sign(\nabla_x \mathcal{L}(h_\theta(x^t), y) - \nabla_x \mathcal{L}(h_\theta(x^t), y_{target}))).$$

Vidimo da se koristi projekcija koraka na skup dopuštenih perturbacija, sto je u našem slučaju L_∞ kugla, te da se koristi gradijent funkcije gubitka po ulaznoj slici. U implementaciji ćemo najprije primjenom stohastičkog gradijentnog spusta napraviti korak spusta, zatim vrijednosti perturbacija ograničiti na interval $[-\epsilon, +\epsilon]$. Na kraju ćemo ograničiti vrijednosti piksela slike na interval $[0, 1]$.

4. Rezultati

4.1. Parametri i ulazni podaci

U eksperimentima korišteni su predtrenirani modeli ResNet50, VGG16 i EfficientNet trenirani na ImageNet skupu podataka. Kao funkcija gubitka u svim modelima i primjerima korištena je unakrsna entropija (engl. *cross-entropy*). Za sve napade korištena je metoda projiciranog gradijentnog spusta, PGD, opisana u prethodnom poglavlju. Za sve napade korišteni su parametri iz tablice 4.1. Zbog odličnih rezultata s ovim parametrima, nije bilo potrebe za eksperimentiranjem s drugim vrijednostima. Na slici 4.1 su prikazane tri korištene ulazne slike, koje uključuju violinu, go kart i bizona. Predikcije modela na neizmijenjenim ulaznim slikama prikazane su u tablici 4.2.

Parametar	**Vrijednost**
broj iteracija	100
opseg L_∞ kugle ϵ	$2/255$
korak učenja SGD	0.005

Tablica 4.1: koristeni parametri

(a) violina

(b) go kart

(c) bizon

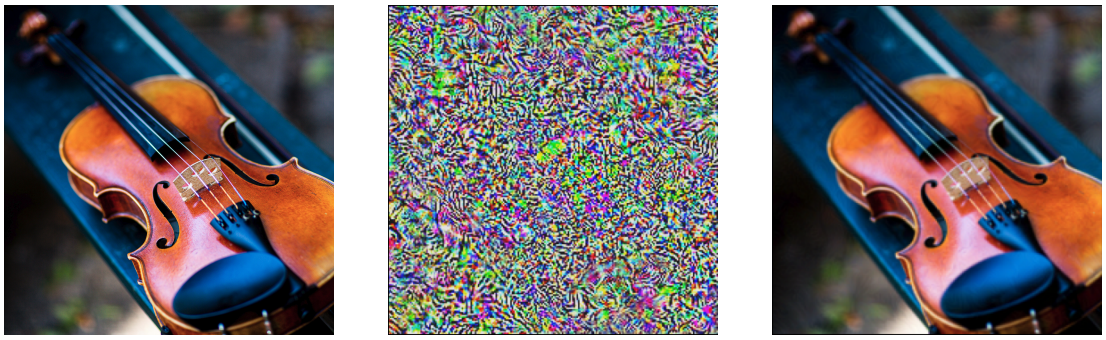
Slika 4.1: Ulazne slike

Model	violina	go kart	bizon
ResNet50	violin (0.9277)	go kart (0.9729)	bizon (0.9966)
VGG16	violin (0.9414)	go-kart (0.9925)	bizon (0.9998)
EfficientNet-B4	violin (0.9470)	go kart (0.9945)	bizon (0.9944)

Tablica 4.2: predikcije za neizmijenjene ulazne slike

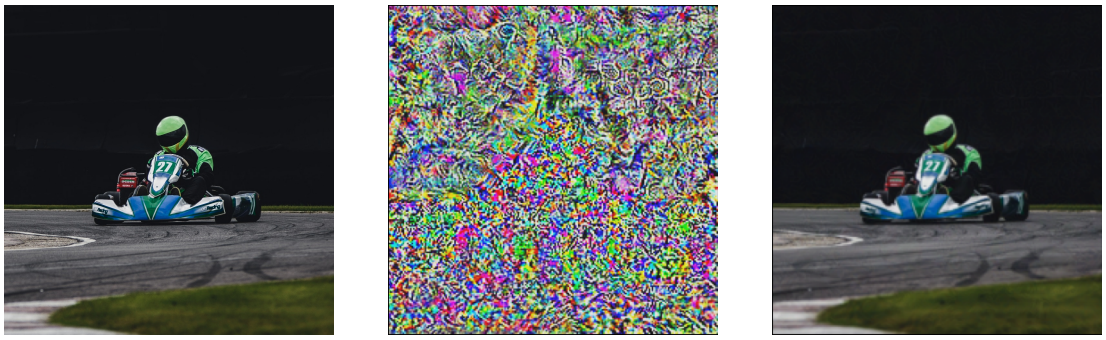
4.2. ResNet50

ResNet50 model jedan je od najpopularnijih modela u području obrade i klasifikacije slika. Na slici 4.2 prikazan je primjer napada na ResNet50 model gdje je početna slika violina, a ciljana klasa je samoyed. Slika 4.3 prikazuje primjer napada gdje je početna slika go kart, a ciljana klasa je samoyed. Slika 4.4 prikazuje primjer napada gdje je početna slika bizon, a ciljana klasa samoyed.



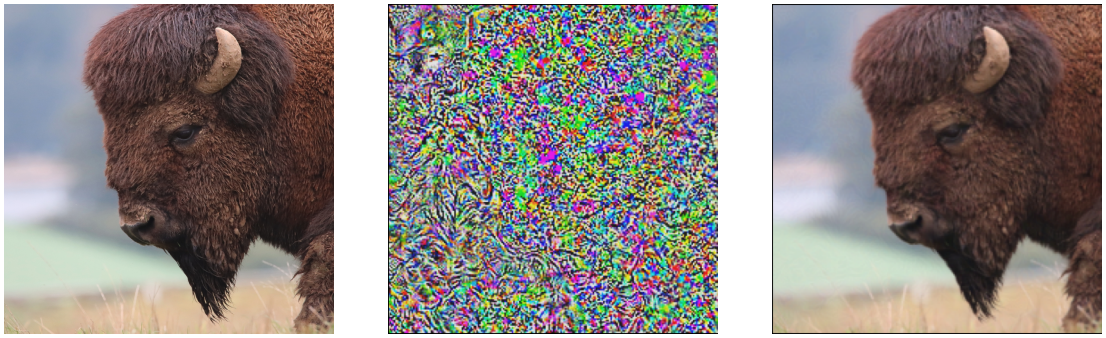
violina delta 99.9952% samoyed

Slika 4.2: ResNet50 violina kao samoyed



gokart delta 99.9328% samoyed

Slika 4.3: ResNet50 go kart kao samoyed

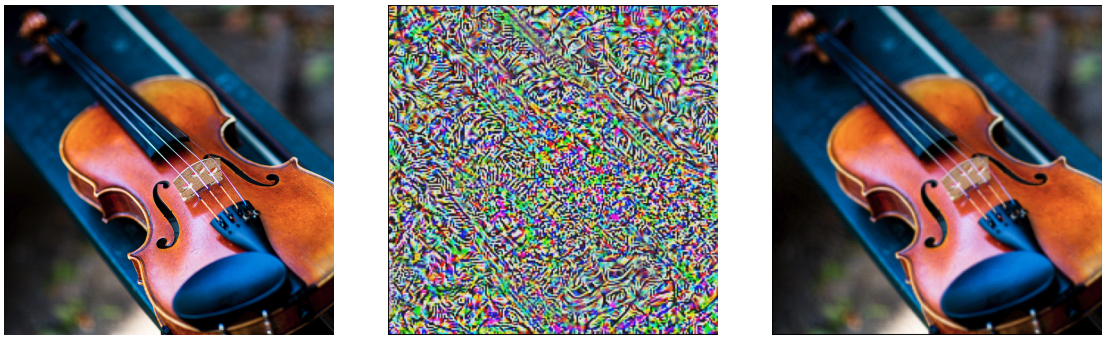


bizon delta 100% samoyed

Slika 4.4: ResNet50 bizon kao samoyed

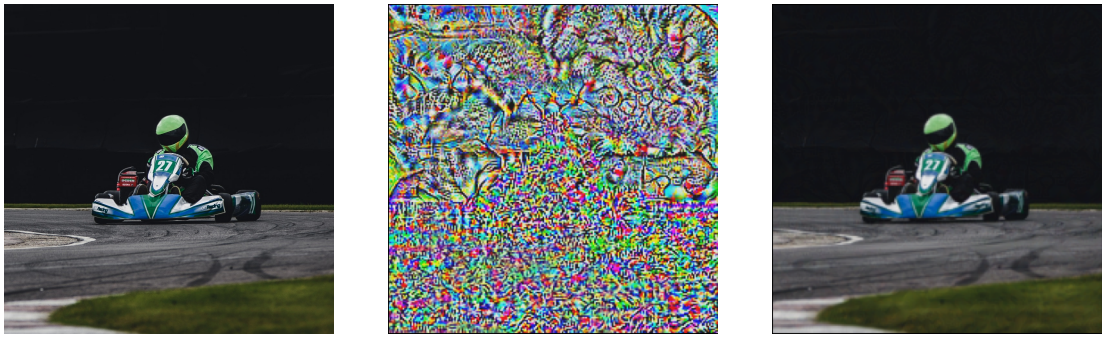
4.3. VGG-16

Ovo je najstariji model od tri koja su korištena u ovom radu, iz 2014. godine. Na slici 4.5 je primjer napada na model VGG-16 u kojem je korištena početna slika violine, na slici 4.6 početna slika go karta, a na slici 4.7 početna slika bizona. U sva tri primjera ciljana klasa je samoyed.



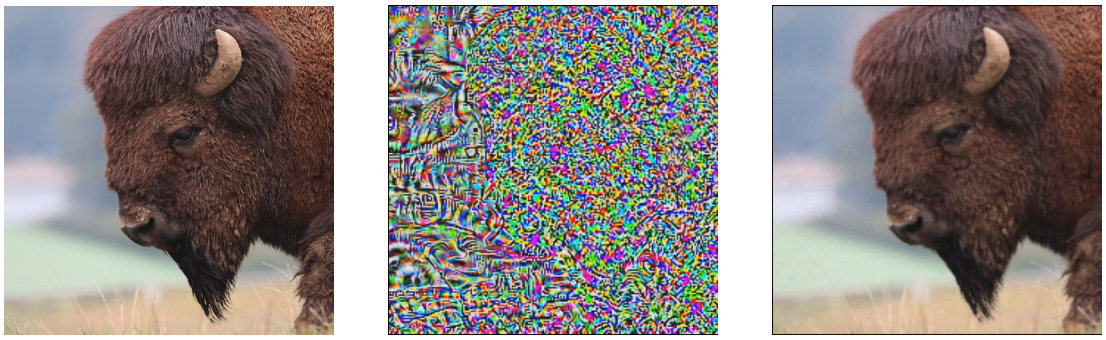
violina delta 99.9924% samoyed

Slika 4.5: VGG-16 violina kao samoyed



go kart delta 99.9871% samoyed

Slika 4.6: VGG-16 go kart kao samoyed

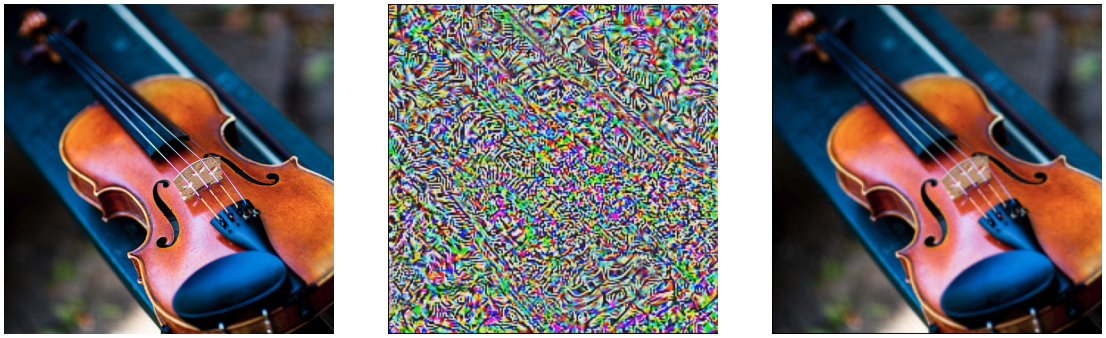


bizon delta 100% samoyed

Slika 4.7: VGG-16 bizon kao samoyed

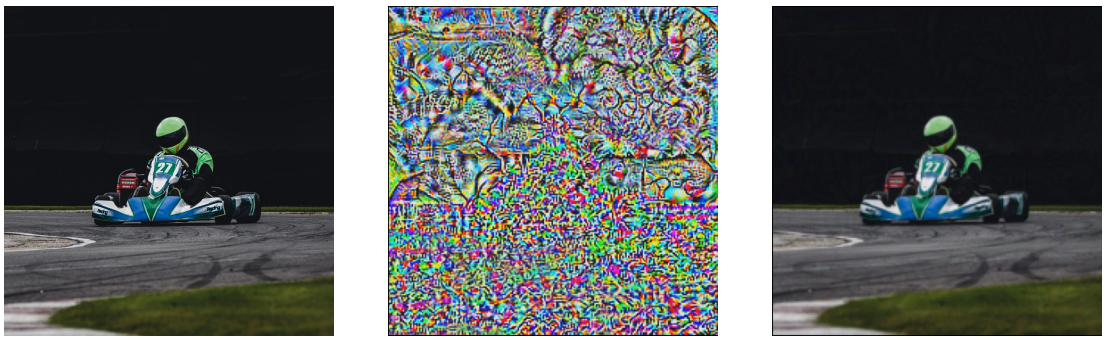
4.4. EfficientNet-B4

EfficientNet je novija arhitektura koja je objavljena 2019. godine. Na slici 4.8 je primjer napada na EfficientNet model gdje je početna slika violina, na slici 4.9 početna slika go karta, a na slici 4.10 početna slika bizona. Kao i u prethodnim primjerima, ciljana klasa za sva tri primjera je samoyed.



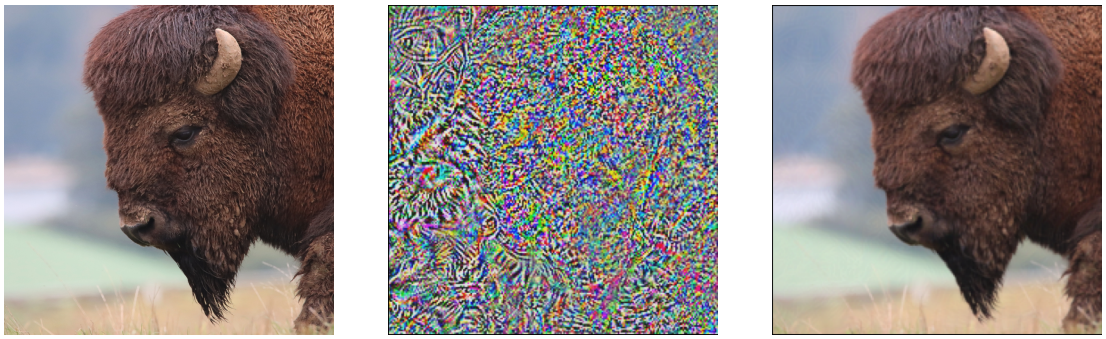
violina delta 61.16% samoyed

Slika 4.8: EfficientNet violina kao samoyed



go kart delta 96.33% samoyed

Slika 4.9: EfficientNet go kart kao samoyed



bizon delta 99.98% samoyed

Slika 4.10: EfficientNet bizon kao samoyed

Vidimo da su rezultati napada na sva tri modela izuzetno dobri, te da su modeli lako zavarani. Samo u jednom slucaju pouzdanost neprijateljskog primjera pada ispod 96%, s EfficientNet modelom i ulaznom slikom violine. U dodatnom eksperimentu s povecanjem broja iteracija sa 100 na 250 dosegnuta je preciznost od 98.33%. U svim preostalim slučajevima, osim jednog, preciznost je 99.9% ili veća, a u dva slučaja je čak 100%. Slični rezultati postignuti su i kad je ciljana klasa bila "aircraft carrier", tj.

nosač aviona.

4.5. Transferabilnost napada

Najinteresantiji rezultat eksperimenata bio je transferabilnost neprijateljskih primjera, gdje primjer generiran za jedan model može zavarati i drugi model. Iako ciljana klasa gotovo nikad nije sačuvana nakon transfera, model i dalje pogrešno klasificira primjer, no s manjom pouzdanošću. Primjerice, neprijateljski primjer generiran na EfficientNet modelu zavarao je ResNet50, gdje je violinu sa neprijateljskom perturbacijom prepoznao kao kantu s pouzdanošću od 0.65%. Iako je pouzdanost rezultata niska, interesantno je što je model pogrešno klasificirao primjer. Isti primjer nije zavarao VGG-16 model, koji je neprijateljski primjer i dalje klasificirao kao violinu, sa sniženom preciznošću od 91%. Niti primjer treniran na ResNet50 modelu nije zavarao VGG-16, dok je lakše transferirati primjere među ResNet i EfficientNet modela. To bi moglo upućivati na to da je stariji VGG-16 model robusniji na neprijateljske primjere.

5. Obrane od neprijateljskih primjera

5.1. Trening s neprijateljskim primjerima

Postavlja se pitanje kako se obraniti od neprijateljskih primjera. Intuitivno se nameće rješenje da se neprijateljski primjeri uključe u skup za učenje modela, koji bi tako postali otporniji na takve primjere. Tipično ovakve metode žele minimizirati sljedeći izraz:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\max_{\delta\in\Delta}(\mathcal{L}(h_\theta(x+\delta), y))] \tag{5.1}$$

gdje je Δ skup dopuštenih perturbacija, a $\mathcal{D}$ distribucija skupa za učenje. Vidimo da unutarnja maksimizacija zapravo primjenjuje perturbacije na primjere iz $\mathcal{D}$. Ovaj izraz je teško minimizirati zbog unutarnje maksimizacije, pa se rješenja često svode na aproksimacije. Ovakvi pristupi rezultiraju plosnatijim i glađim gradijentima, odnosno onima čije se vrijednosti ne mijenjaju naglo, te takvi oblici gradijenata pomažu pri robusnosti.

Madry et al. (2017) pokazuju da rješavanje jednadžbe 5.1 nije dovoljno da bi se garantirala robusna i precizna klasifikacija. Pokazali su i da povećanje kapaciteta povećava robusnost uz malene perturbacije nad L_∞ kuglom.

Valja napomenuti da ovakve obrane ovise o modelu napada protiv kojeg se brani, odnosno o vrsti perturbacija koje se primjenjuju na primjere u skupu za učenje. Madry et al. (2017) koriste perturbacije iz L_∞ kugle, dok Zhang et al. (2019) koriste L_2 normu.

Jiang et al. (2020) dodatno unaprjeđuju robusnost na neprijateljske primjere korištenjem samonadziranog predtreniranja, sto je u kombinaciji s treningom s neprijateljskim primjerima pokazalo dobre rezultate, s 2.99% preciznosti na neprijateljskim primjerima i 2.14% standardne preciznosti na CIFAR-10 skupu podataka.

5.2. Ostale metode obrane

Obrane koje se ne oslanjaju na trening s neprijateljskim primjerima obično koriste razne trikove da bi se gradijent prikrio. Takve metode se dosad nisu pokazale uspješnima. Papernot et al. (2016) predlažu obranu distilacijom, gdje se model trenira na vjerojatnostima razreda umjesto na točnim oznakama. Carlini i Wagner (2016) zatim otkrivaju da obrana distilacijom nije robusnija na neprijateljske primjere od nezasticenih modela. Metode obrane koje se oslanjaju na maskiranje gradijenata su ranjive na napade opisane u Athalye et al. (2018). Dodatno, Raghunathan et al. (2020) ukazuju na to da postoji kompromis između robusnosti na neprijateljske primjere i točnosti modela, bez obzira na korištenu metodu obrane.

6. Zaključak

Rezultati pokazuju da je neprijateljske primjere izuzetno lako pronaći, i to uz proizvoljnu ciljanu klasu. Samo 100 iteracija je dovoljno da bi se pronašao izuzetno kvalitetan primjer u većini slučajeva, što obično traje od jedne do nekoliko minuta na modernom hardveru. Dodatno, neprijateljski primjeri na jednom modelu su u nekoj mjeri transferabilni na druge modele, iako ciljana klasa nije očuvana. Iako duboki modeli mogu postići bolje rezultate nego ljudi na pojedinačnim zadacima, iz rezultata je vidljivo da još uvijek ne mogu postići ljudsku razinu robusnosti na svim zadacima, te da je potrebno uložiti dodatne resurse kako bi model bio robusan na neprijateljske primjere. Treniranje robusnijih modela zahtjeva veću količinu podataka potrebnih da bi se dosegla zadovoljavajuća razina preciznosti. Imajući na umu sve navedeno, potrebna su dodatna istraživanja i daljnji razvoj robusnih modela kako bismo modele strojnog učenja mogli adekvatno uklopiti u sustave u stvarnom svijetu u budućnosti.

7. Literatura

Anish Athalye, Nicholas Carlini, i David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. U *International conference on machine learning*, stranice 274–283. PMLR, 2018.

Nicholas Carlini i David Wagner. Defensive distillation is not robust to adversarial examples. *arXiv preprint arXiv:1607.04311*, 2016.

Nicholas Carlini i David Wagner. Towards evaluating the robustness of neural networks. U *2017 ieee symposium on security and privacy (sp)*, stranice 39–57. Ieee, 2017.

Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, i Cho-Jui Hsieh. Ead: elastic-net attacks to deep neural networks via adversarial examples. U *Proceedings of the AAAI conference on artificial intelligence*, svezak 32, 2018.

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

Ziyu Jiang, Tianlong Chen, Ting Chen, i Zhangyang Wang. Robust pre-training by adversarial contrastive learning. *Advances in neural information processing systems*, 33:16199–16210, 2020.

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.

Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, i Pascal Frossard. Sparsefool: a few pixels make a big difference. U *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, stranice 9087–9096, 2019.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, i Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. U *2016 IEEE symposium on security and privacy (SP)*, stranice 582–597. IEEE, 2016.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, i Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. *arXiv preprint arXiv:2002.10716*, 2020.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, i Michael Jordan. Theoretically principled trade-off between robustness and accuracy. U *International conference on machine learning*, stranice 7472–7482. PMLR, 2019.

8. Sažetak

Opisane su metode generiranja neprijateljskih primjera. Sve metode koriste informacije o gradijentima, sto spada u tzv. *whitebox* napade. Prikazani su i opisani rezultati napada metodom projiciranog gradijentnog spusta (PGD) na tri modela trenirana na ImageNet skupu podataka. Ukratko su opisane metode obrane od takvih napada.