

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 2231

Obrana dubokih konvolucijskih modela od neprijateljskih primjera

Matej Dobrovodski

Zagreb, lipanj 2020.

SADRŽAJ

1. Uvod	1
1.1. Raspoznavanje objekata	1
1.2. Neprijateljski primjeri	2
2. Programska potpora	3
2.1. Odabir biblioteke za duboko učenje	3
2.2. Biblioteke za neprijateljske primjere	3
2.3. Skupovi podataka	3
2.4. Konvolucijski modeli	3
3. Neprijateljski primjeri I	4
3.1. Model prijetnje	4
3.2. Pojava prvih neprijateljskih primjera	4
3.3. Brza metoda temeljena na gradijentima	4
3.4. DeepFool	4
4. Obrana dubokih konvolucijskih modela	5
5. Neprijateljski primjeri II	6
5.1. Neučinkovitost obrana	6
5.2. PGD	6
5.3. Napadi <i>Carlini and Wagner</i>	6
5.4. Napad temeljen na odluci	6
6. Budućnost obrana od neprijateljskih napada	7
6.1. PGD adversarial training	7
6.2. Preduvjeti uspješnih obrana	7
6.3. Dokazivost obrane od neprijateljskih napada	7
6.4. Budući rad	7

7. Zaključak	8
Literatura	9
8. Dodatak	11
8.1. Osobni skup slika	11
8.2. Izlazi modela na nepromijenjenim slikama iz osobnog skupa . . .	11

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 njih su *ResNet*[citat], *Xception*[citat] i *VGG*[citat]. 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^[11], 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 *sljepе pјege* koje se isplati istražiti.

Vrijedi da za neki ispravno klasificirani ulaz x postoji područje u blizini ulaza, $x+r$, gdje je $\|r\| < \epsilon$ za neki dovoljno mali radijus $\epsilon > 0$. Uobičajeno je da modeli ulazne vrijednosti iz tog područja također ispravno klasificiraju, kao i x , te u općem slučaju vrijedi da neprimjetne perturbacije iz tog ϵ područja (npr. nasumični šum slabog intenziteta) ne mijenjaju izlaz modela. Rješavanjem optimizacijskog problema pri kojem se minimizira perturbacija r koja izaziva pogrešnu klasifikaciju dolazzi se do takozvanih neprijateljskih (ili suparničkih) primjera (eng. *adversarial example*). Ono što je iznenadjuće i što je potaklo daljnje istraživanje je to što je zapravo vrlo lako za pronaći neprijateljske primjere na *state of the art* modelima kod kojih je perturbacija r praktički potpuno neprimjetna i nije očito zašto mreže neispravno klasificiraju takve ulaze. U nastavku rada je dan pregled nekoliko metoda generiranja neprijateljskih primjera: neke metode su prikazane zbog njihove povijesne važnosti i utjecaja na daljni razvoj metoda, dok su neke metode iznimno snažne i mjerilo za uspješnost obrane od suparničkih napada. Uz napade su pokazane i obrane, s naglaskom na njihovu (ne)uspješnost pri odupiranju od suparničkih napada, probleme koji su im zajednički te potencijalnu budućnost razvoja uspješnih obrana.

2. Programska potpora

2.1. Odabir biblioteke za duboko učenje

tensorflow^[1] keras^[4]

2.2. Biblioteke za neprijateljske primjere

foolbox^[10] cleverhans^[9] art^[8]

2.3. Skupovi podataka

imagenet, cifar10, personal

2.4. Konvolucijski modeli

resnet, etc, onaj mali model

3. Neprijateljski primjeri I

3.1. Model prijetnje

Osnovni termini, lažno pozitivni/negativni(*) primjeri, white/black box, targeted/nontargeted, one-time/iterative^[12]

3.2. Pojava prvih neprijateljskih primjera

intriguing properties^[11]

3.3. Brza metoda temeljena na gradijentima

$fg(s)m^{[5]}$

3.4. DeepFool

$df^{[7]}$

4. Obrana dubokih konvolucijskih modela

opisati vrste obrane

5. Neprijateljski primjeri II

5.1. Neučinkovitost obrana

...

5.2. PGD

pgd^[6]

5.3. Napadi *Carlini and Wagner*

candw^[3]

5.4. Napad temeljen na odluci

boundary^[2]

6. Budućnost obrana od neprijateljskih napada

- 6.1. PGD adversarial training**
- 6.2. Preduvjeti uspješnih obrana**
- 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 objekata. Međutim, pojava neprijateljskih primjera postavlja

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] Wieland Brendel, Jonas Rauber, i Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *ArXiv*, abs/1712.04248, 2017.
- [3] 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.
- [4] François Chollet et al. Keras. <https://keras.io>, 2015.
- [5] Ian J. Goodfellow, Jonathon Shlens, i Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2015.
- [6] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, i Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2017.
- [7] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, i Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, stranice 2574–2582, 2016.
- [8] 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>.
- [9] 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 Hambardzumyan, Zhi-shuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhishav 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.
 - [10] 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>.
 - [11] 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.
 - [12] Xiaoyong Yuan, Pan He, Qile Zhu, i Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30:2805–2824, 2019.

8. Dodatak

8.1. Osobni skup slika

U 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 odabrane 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 degradacije kvalitete pri smanjenju rezolucije na 224×224 . Zahvaljujem se Sarah James na dopuštenju za korištenje slika u radu.

8.2. Izlazi modela na nepromijenjenim slikama iz osobnog skupa



(a) *African grey* (87)



(b) *Backpack* (414)



(c) *Dragonfly* (319)



(d) *Bald eagle* (22)



(e) *Grocery store* (582)



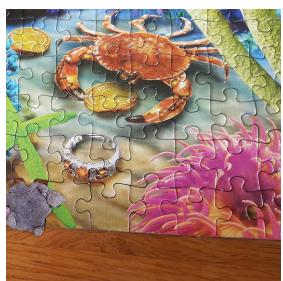
(f) *Gorilla* (366)



(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 nedavno otkriveni neprijateljski primjeri. Neprijateljski primjeri su slike s dodatnim teško uočljivim perturbacijama koje potiču model na pogrešnu klasifikaciju. Pokazalo se da je vrlo lagano konstruirati brze i efikasne napade na postojeće modele, međutim ubrzo su se pojavile i obrane protiv takvih napada. Ti početni napadi su se oslanjali na pristup mreži i poznavanju arhitekture, a obrane od takvih napada se temelje na "prikrivanju" potrebnih informacija, kao što su gradijenti.

Ključne riječi: duboko učenje, klasifikacija, konvolucijske neuronske mreže, računalni vid, suparnički primjeri, neprijateljski primjeri, obrana.

Defending Deep Convolutional Models from Adversarial Examples

Abstract

Deep neural networks can achieve very high accuracy in many applications such as image classification. However, most of these deep models are difficult to interpret and they are often sensitive to the so-called adversarial examples. This feature opens up the possibility of maliciously designing adversarial examples that could deceive a deep learning system.

Keywords: deep learning, classification, convolutional neural networks, computer vision, adversarial attacks, defense.