SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 2231

Obrana dubokih konvolucijskih modela od neprijateljskih primjera

Matej Dobrovodski

Zagreb, svibanj 2020.

Sadržaj

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 neke od tih mreža, te će se one spominjati i koristiti u nastavku rada. Neke od njih su *Xception*[citat], *VGG*[citat], *ResNet*[citat], *DenseNet*[citat]. Sve te mreže postižu vrlo zadovolja-

vajuću točnosti i čini se da mogu dobro generalizirati. No u nastavku rada će biti pokazan oblik napada na konvolucijske mreže koji osporava činjenicu da današnje konvolucijske mreže dobro generaliziraju.

1.2. Neprijateljski primjeri i obrana

što su suparnički primjeri [seminar]
primjeri
primjena
što je obrana
važnost, relevantnost obrane

2. Neprijateljski primjeri

threat model (false positive, false negative, white/black box), targeted/non targeted, attack frequency (one-time, iterative)
"povijest" neprijateljskih primjera
vrste, kratko o njima i na koje se ja fokusiram
matematička definicija ?
žešće obraditi par obitelji napada (neke starije, neke robusnije)
napadi: LBFGS 2014, FG(S)M https://arxiv.org/abs/1412.6572 (rand-fgsm?), JSMA, DeepFool, CW, Pixel Attack, Boundary attack, HopSkipJump attack
pokazati kako funkcioniraju na hrpi primjera

3. Obrana dubokih konvolucijskih modela

opisati vrste obrane
pristupi jednostavnijih obrana
problemi
trenutno stanje obrana
provable defense
budućnost obrana

4. Rezultati

4.1. Programska potpora

odabir tehnologije - navesti 3 biblioteke, CUDA, python, whatever the fuck
foolbox - čiji je, neki napadi, pros, cons, kratki pseudokod
cleverhans - same
art - same

4.2. Primjeri pojedinih napada

@

4.3. Pregled učinkovitosti obrana

@

5. Zaključak

some bullshit

future work!!!!!!

NOTE stvari potrebne znati inside out: feedforward mreža, CNN, sve što se spominje na random (Adam)

Obrana dubokih konvolucijskih modela od neprijateljskih primjera

Sažetak

.

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.