COPYCAT CNN

Using Random Non-Labeled Data

Proiect de cercetare replicabilă

Cătălin-Alexandru Rîpanu 341C3
Alexandra-Ana-Maria Ion 341C1

CUPRINS

1. CE ESTE COPYCAT CNN?

CNN

Convolutional Neural Network

- Este o subcategorie a Deep Neural Networks **(DNN)**
- Atinge performanțe de vârf în diverse probleme (clasificarea și recunoașterea vizuală, detectarea obiectelor)
- Numeroase companii oferă soluții bazate pe aceste rețele neurale, având API-uri pentru accesul la modelele lor de învățare profundă

Copycat CNN

Copycat Convolutional Neural Network

- **“Copie“** a unui model CNN țintă, ce nu conține întreaga logică complexă din spatele modelului
- Realizat prin interogări **(queries)** cu imagini naturale aleatorii

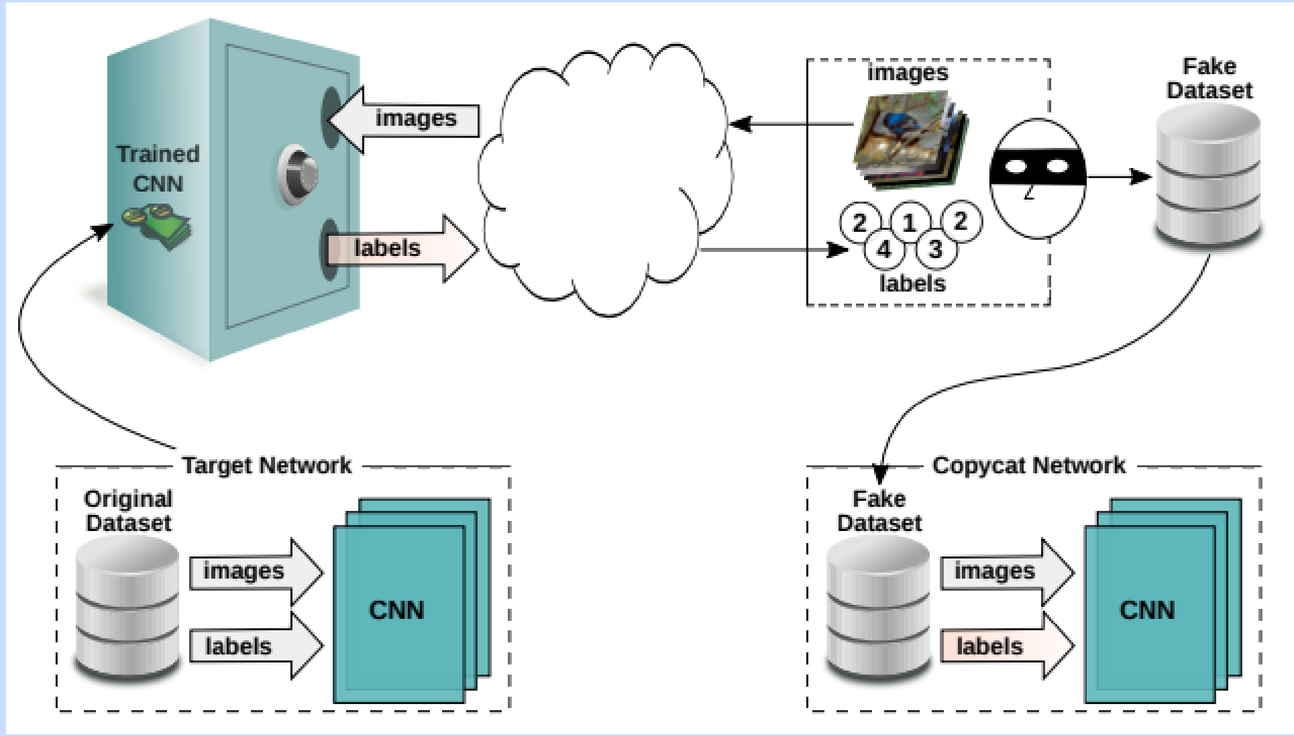
2. PAȘII DE REALIZARE A COPIEI

1. **Generarea setului de Date Fals**

- Interogarea țintei folosind N (82783 în cazul demo-ului) imagini din PD sau NPD
- Extragerea etichetelor de pe aceste imagini trimise

2. **Antrenarea Copycat CNN**

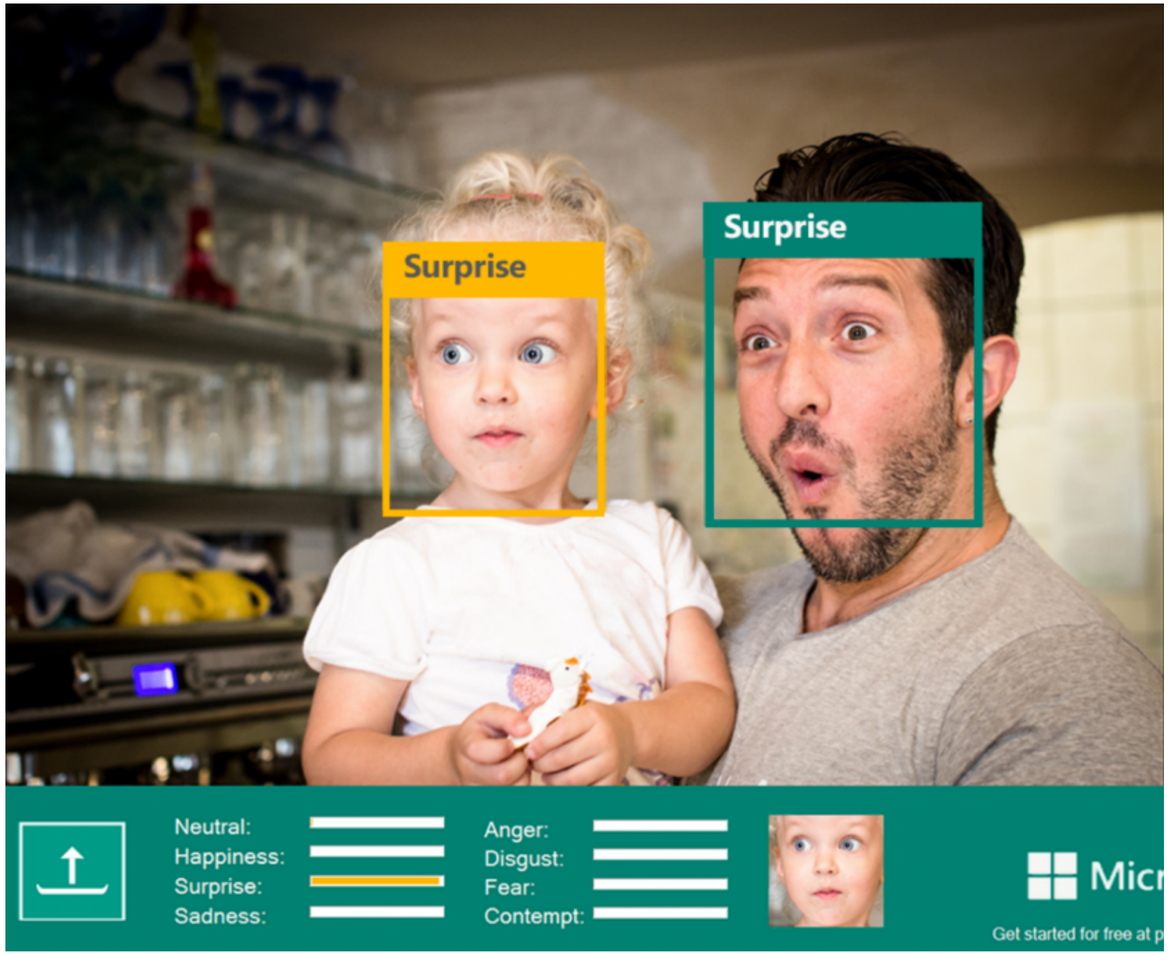
- Se antrenează cu M (10 în cazul demo-ului) epoci folosind setul fals de date obținut cu ajutorul etichetelor
- **NU** este necesar ca atacatorul să cunoască arhitectura modelului țintă, acesta poate alege dintr-o suită de modele deja existente



3. ATAC ASUPRA MICROSOFT EMOTION API

Microsoft Emotion API

- API oferit de Microsoft Azure Cognitive Services
- Permite unui utilizator să trimită o imagine și să primească în răspuns **locațiile fețelor și "emoțiile" recunoscute** pentru fiecare față.
- Oferă 2 tipuri de tarifare:
  - **Gratis** - Max. 20 imagini/minut
  - **Standard** - 1000 imagini (0.4$ - 1$)



Realizarea atacului utilizând Copycat CNN

- Pentru **generarea setului de date fals**, au fost utilizate ~3.5 milioane de imagini
  - ~65k din **PD** (imagini oficiale oferite de furnizorul API-ului)
  - ~3.4 milioane din **NPD** (imagini naturale aleatorii)
- **Modelul Copycat CNN a fost antrenat** folosind **Stochastic Gradient Descent (SGD)** cu o politică de reducere a ratei de învățare

Copycat CNN
a copiat cel puțin

97.3%

din perfomanța Microsoft Azure Emotion API

Networks	Macro Average (accuracy)	Performance over target network
Microsoft Azure Emotion API	35.1%	–
NPD-SL	34.1%	97.3%
PD-SL	35.4%	100.8%
NPD+PD-SL	34.8%	99.2%

4. REPLICAREA ATACULUI ȘI REZULTATE

- Am utilizat 2 rețele neurale convoluționale, realizate în Python:
  - oracol - modelul CNN țintă
  - copycat - modelul Copycat CNN
- Pentru antrenarea celor două modele au fost folosite:
  - oracol - 50k imagini din setul de imagini CIFAR10
  - copycat - ~80k imagini din NPD (un set de imagini de la Microsoft Coco)
- Mediul de antrenare si testare:
  - Sistemul de operare Windows, cu ajutorul nucleului Linux expus de WSL 2 pe o arhitectură x86_64, iar o placă grafică dedicată Nvidia 3050 RTX Ti a fost utilizată pe parcurs

Rezultate

oracol

```
Average: 82.97% (10000 images)
Micro Average: 0.829700
Macro Average: 0.830625
```

copycat

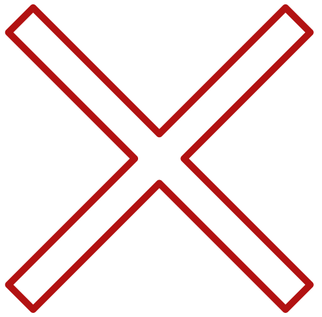
```
Average: 37.39% (10000 images)
Micro Average: 0.373900
Macro Average: 0.332408
```

5. CONCLUZII

Rezultate bune pe dataset-uri mari de antrenare

Modelul Copycat, conform lucrării citate și a rezultatelor obținute în copierea Microsoft Emotion API, are rezultate bune pe dataset-uri mari de antrenare.

Aceste rezultate pot fi datorate și utilizării imaginilor din PD în generarea setului de date false.



Acuratețe slabă pe dataset-uri mici de antrenare

Modelul Copycat, conform analizei noastre realizată pe dataset-uri considerabil mai mici, are rezultate nesatisfăcătoare.

Aceste rezultate pot fi datorate și faptului că pentru generarea setului de date false au fost utilizate numai imagini din NPD

REFERINȚE

- Copycat CNN: Stealing Knowledge by Persuading Confession with Random Non-Labeled Data (Correia-Silva et al., 2018)
- Repository referință pentru realizarea replicării
- Exemplu vizual Copycat CNN
- IBM - Convolutational Neural Networks