# PROIECTAREA REȚELELOR
# COPYCAT CONVOLUTIONAL NEURAL NETWORK



**Cătălin-Alexandru Rîpanu 341C3 Alexandra-Ana-Maria Ion 341C1**
Facultatea de Automatică și Calculatoare
Universitatea Națională de Știință și Tehnologie Politehnica București
catalin.ripanu@stud.acs.upb.ro alexandra_ana.ion@stud.acs.upb.ro

14 Ianuarie 2024

**ABSTRACT**

Această lucrare își propune abordarea unui subiect interesant în sfera atacurilor modelelor de Învățare Automată ce prezintă vulnerabilități critice în ceea ce privește zona de Privacy (Confidențialitate) a datelor prelucrate de către acestea. Concret, vom expune riguros, în cele ce urmează, un studiu ce a concluzionat, efectuând diverse experimente în diverse cazuri, faptul că este posibilă copierea cunoștințelor unui clasificator de imagini (ce ar putea fi, de asemenea, o rețea neurală adâncă) expus de un potențial API și antrenat cu o bază de date privată (de asemenea, se mai menționează, în studiu, ideea că acest clasificator este un black-box întrucât nu este necesară cunoașterea, în profunzime, a logicii lui). Nu în ultimul rând, vom exemplifica pașii executați pentru replicarea acestui tip de atac într-un caz particular, puțin mai simplu, având în vedere că un scenariu real necesită mult mai multe resurse de calcul astfel încât procesarea unui număr mare de imagini (de ordinul milioanelor) să fie posibilă într-un timp rezonabil.

## 1 Introducere

### 1.1 Copycat CNN

Un Copycat Convolutional Neural Network (CNN) este o abordare nouă discutată în articolul citat, cu scopul de a replica comportamentul și performanța unui model țintă CNN fără a avea acces la arhitectura sau datele sale de antrenare. Această metodă este concepută să opereze ca o atacare de tip black box, ceea ce înseamnă că atacatorul are doar acces la comportamentul de intrare-ieșire al rețelei țintă, fără să știe structura internă sau datele de antrenare. Motivația dezvoltării unei astfel de tehnici provine din investiția semnificativă făcută de companii în crearea și protejarea modelelor lor de învățare profundă (eng. deep learning), făcând esențială explorarea potențialelor vulnerabilități ale securității acestora.

Procesul de creare a unui Copycat CNN implică două etape principale: generarea unui set de date fals și antrenarea unei rețele copycat. În etapa de generare a setului de date false, se utilizează imagini naturale aleatorii, fie din același domeniu de problemă ca și rețeaua țintă, fie din seturi de date mari și publice care nu sunt legate de domeniul problemei. Aceste imagini sunt etichetate folosind însăși rețeaua țintă, rezultând ceea ce autorii numesc etichete furate (SL). Setul de date fals generat este menit să surprindă nuanțele spațiului de caracteristici, permițând antrenarea unei rețele copycat.

Etapa de antrenare a rețelei copycat implică selectarea unei arhitecturi de model copycat, cu un exemplu cunoscut fiind VGG16. Anrtrenarea copycat-ului presupune atribuirea parametrilor tintei in straturile nemodificate, iar ultimul strat de

ieșire este populat aleatoriu. Acest lucru asigură că rețeaua copycat învață din setul de date fals, cu scopul de a replica comportamentul rețelei țintă.

### 1.2 Atac asupra Microsoft Emotion API

În articolul citat se discută despre o încercare de a copia performanța Microsoft Azure Emotion API, care face parte din Azure Cognitive Services și permite utilizatorilor să trimită imagini pentru a detecta fețe și a recunoaște emoțiile asociate fiecărei fețe. Seturile de date utilizate pentru atac includ date non-problem domain (NPD) și problem domain (PD), concentrându-se în special pe probleme de recunoaștere a expresiilor faciale (FER).

Pentru atac, cercetătorii au avut ca scop copierea performanței Emotion API prin generarea unei rețele Copycat folosind date aleatoare și neetichetate. Atacul a implicat crearea unui set de date fals utilizând atât imagini NPD, cât și PD etichetate de Emotion API. Rețeaua Copycat a fost apoi antrenată cu acest set de date fals pentru a replica comportamentul API-ului. Au fost folosite metrici precum acuratețea macro-ponderată și rapoartele de performanță relativ la rețeaua țintă și la o rețea cu un domeniu de problemă mai mic pentru evaluarea performanței rețelei copycat.

În configurarea generală, modelele au fost antrenate folosind Stochastic Gradient Descent (SGD), cu un număr specific de epoci, iar tehnici de augmentare a datelor au fost aplicate. Experimentele au fost efectuate pe un sistem cu Intel Core i5-7500, 32GiB de RAM și NVIDIA GeForce GTX 1070. Rezultatele atacului au indicat că rețeaua copycat a atins 97,3% din performanța rețelei țintă fără a utiliza date din domeniul problemei (NPD). În mod surprinzător, utilizând doar date din domeniul problemei, rețeaua Copycat a depășit performanța API-ului cu 100,8%. Cu toate acestea, combinarea ambelor seturi de date a dus la o ușoară scădere (-0,6%) în comparație cu utilizarea doar a datelor din domeniul problemei.

Cercetarea a demonstrat cu succes capacitatea de a copia performanța Microsoft Azure Emotion API utilizând o rețea Copycat antrenată pe date aleatoare și neetichetate, obținând o acuratețe ridicată în recunoașterea expresiilor faciale. Atacul a relevat vulnerabilitatea Emotion API la replicarea modelelor, ridicând preocupări cu privire la posibilele implicări de securitate.

## 2 Replicarea atacului și rezultate

Pentru replicarea într-o manieră simplistă a acestui atac am ales să implementăm și testăm, într-un mediu local, 2 rețele neurale convoluționale, numite, în această secțiune, **copycat**, respectiv, **oracol**. Am preluat aceste modele de pe Repository-ul de Github STEALING_DL_Model, (mai exact secțiunea "Example_of_use") al autorilor articolului "Copycat CNN: Stealing Knowledge by Persuading Confession with Random Non-Labeled Data".

Antrenarea ambelor modele de clasificare s-a realizat peste sistemul de operare Windows, cu ajutorul nucleului Linux expus de WSL 2 pe o arhitectură x86_64. De asemenea, pentru eficientizarea timpului petrecut pe parcursul antrenării, o placă grafică dedicată Nvidia 3050 RTX Ti a fost uzitată pe parcurs. Limbajul de programare folosit este Python deoarece acesta expune diverse pachete specializate în lucrul cu imagini și cu rețelele neurale (cum ar fi Pytorch, pachetul folosit în cadrul implementărilor modelelor).

Modelele au aceeași arhitectură, și anume 3 straturi convoluționale legate de un ultim strat fully connected (evident, există posibilitatea ca arhitecturile să difere, însă trebuie avut grijă la constrângeri întrucât un copycat proiectat cu o arhitectură mai slabă nu poate obține performanțele unui oracol puternic).

Folosind setul de date CIFAR10 cu 50000 imagini de antrenare, s-a obținut modelul țintă cu o acuratețe de aprox. 83% pe setul cu imaginile de test (10000 astfel de imagini). Copycat-ul a fost antrenat pe baza unui set de imagini de la Microsoft COCO (un set de 82783 de elemente, se observă că aceste poze nu sunt din Problem Domain-ul oracolului). În final, după terminarea antrenării și testării, Copycat-ul a reușit să obțină o acuratețe de aprox. 33% pe baza acestui set.

Comenzile utilizate pe tot parcursul procesului (execute în rădăcina directorului curent):

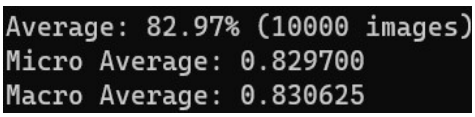
```
# Antrenarea modelului țintă
python3 oracle/train.py cifar_model.pth

# Testarea modelului țintă
python3 oracle/test.py cifar_model.pth

# Scrierea căilor imaginilor în fișierul creat images.txt
find images/copycat_images -type f | grep -i 'jpg\|jpeg\|png' > images.txt
```

```
# Extragerea predicțiilor oracolului (fără probabilități, doar clase)
python3 copycat/label_data.py cifar_model.pth images.txt stolen_labels.txt

# Antrenarea modelului copycat
python3 copycat/train.py copycat.pth stolen_labels.txt

# Testarea modelului copycat
python3 oracle/test.py copycat.pth
```

## 3 Concluzii

O concluzie importantă ce reiese din atacul replicat este că antrenând un clasificator menit să imite performanțele unui alt clasificator black-box cu un set restrâns de imagini ce nu se află în domeniul de pregătire al țintei, rezultă o acuratețe slabă ce indică ideea conform căreia copycat-ul nu poate fi folosit în contexte serioase (real-world, cum ar fi copierea de API-uri populare ce oferă servicii ale propriilor modele de clasificare în scop comercial).

Câteva imagini ilustrative cu rezultatele testărilor:



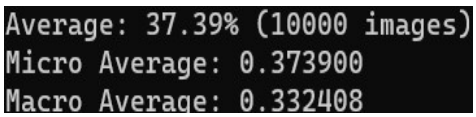




Figure 1: Acuratețile ambelor Rețele Neurale Convoluționale

## 4 Referințe și resurse


- Correia-Silva et al. 2018. **Copycat CNN: Stealing Knowledge by Persuading Confession with Random Non-Labeled Data** - https://ieeexplore.ieee.org/document/8489592, accesat la 14 ianuarie 2024
- Correia-Silva. **Github Repository - Stealing_DL_Models** - https://github.com/jeiks/Stealing_DL_Models, accesat la 14 ianuarie 2024
- **The CIFAR10 Dataset** - https://www.cs.toronto.edu/ kriz/cifar.html, accesat la 14 ianuarie 2024
- **The Microsoft COCO Dataset** - https://cocodataset.org/#home, accesat la 14 ianuarie 2024
- **IBM - What are convolutional neural networks?** - https://www.ibm.com/topics/convolutional-neural-networks, accesat la 14 ianuarie 2024