DSS @ LINUX DAY
25 OTTOBRE 2019

AI,AI,AI!

Problematiche di sicurezza del Machine Learning

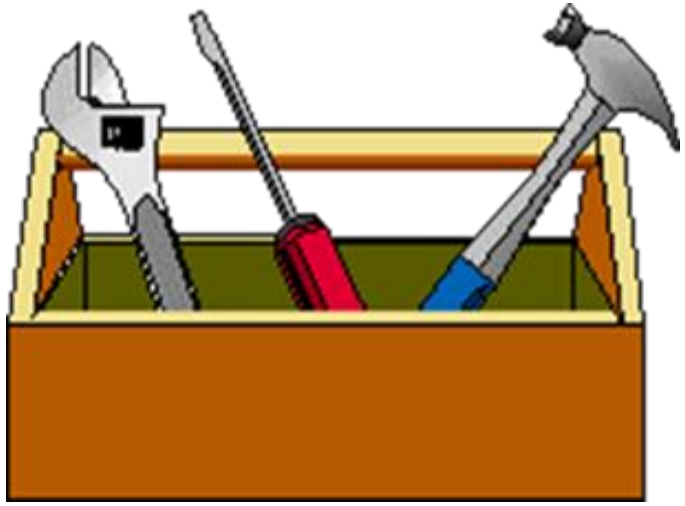
Roberto Marmo
Laboratorio Computer Vision & Multimedia Lab
Università di Pavia
www.robertomarmo.net info@robertomarmo.net

DataScienceSeed #9

Data Science, Machine Learning, Artificial Intelligence Meetup a Genova

Agenda

1. Machine Learning
2. Adversarial Machine Learning
3. Scenario della sicurezza
4. Tipologie di attacchi al Machine Learning
5. Rimedi per evitare danni
6. Consigli per lettura
7. Conclusioni



Machine Learning

Apprendere significa migliorare la capacità di esecuzione di un certo compito attraverso l'esperienza, ovvero imparare dai propri errori.

Machine Learning: processo tramite il quale una macchina apprende a svolgere una funzione senza essere esplicitamente programmata, senza specificare tutti i parametri che caratterizzano il dato compito.

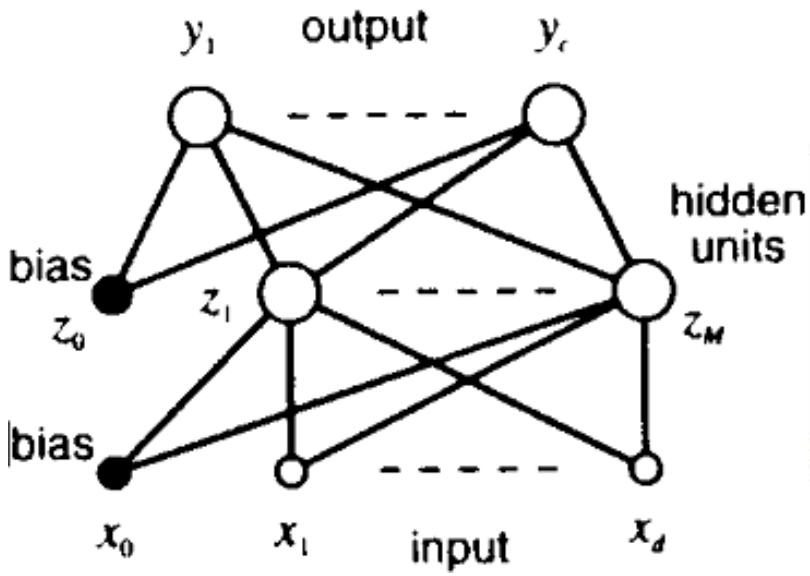
Funzione: imparare addizione

Programmazione esplicita: somma = a + b

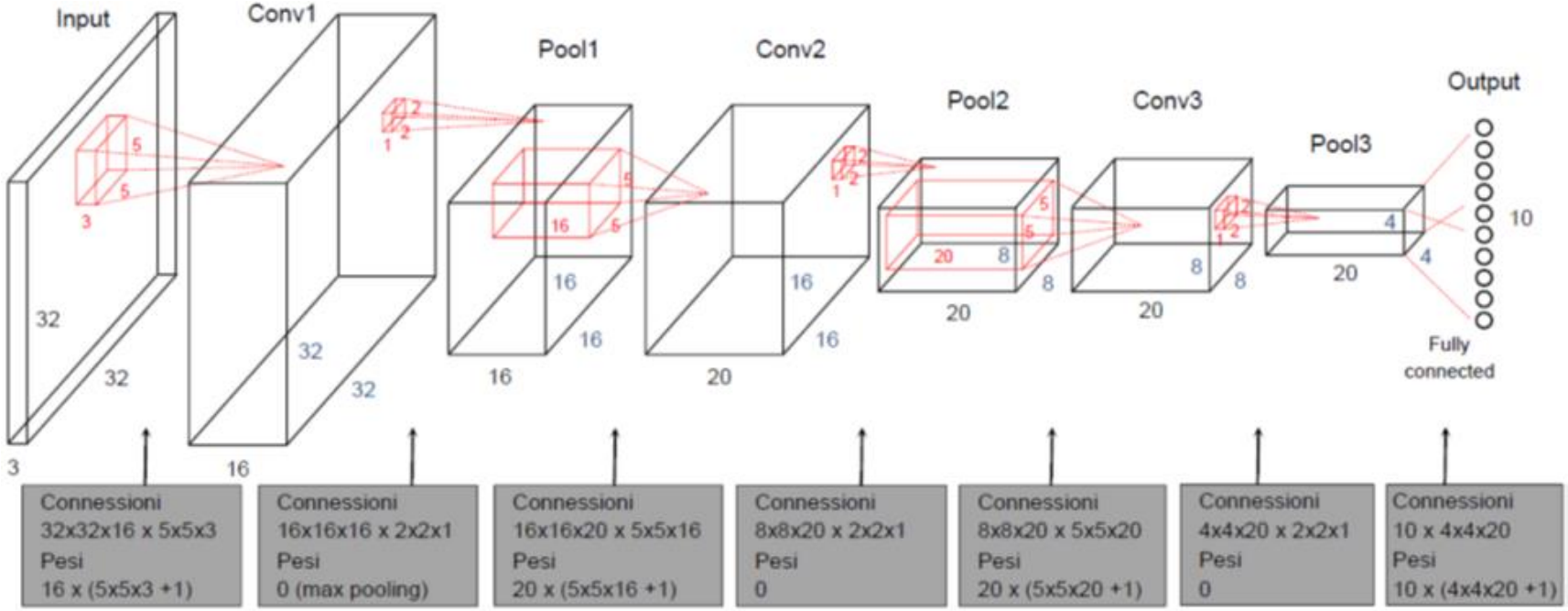
Machine Learning: (input, input, output) (2,1,3) (1,1,2) (0,1,1) (2,5,7) ecc.

Machine Learning

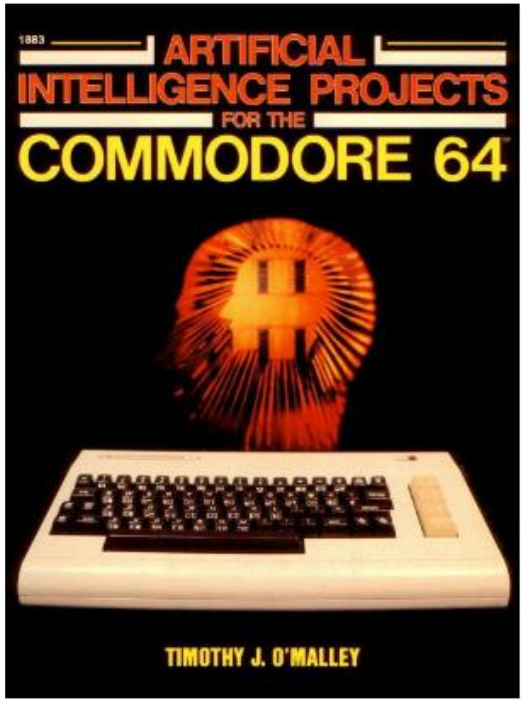
- Neural Network ispirate vagamente al cervello umano
- Rete MLP

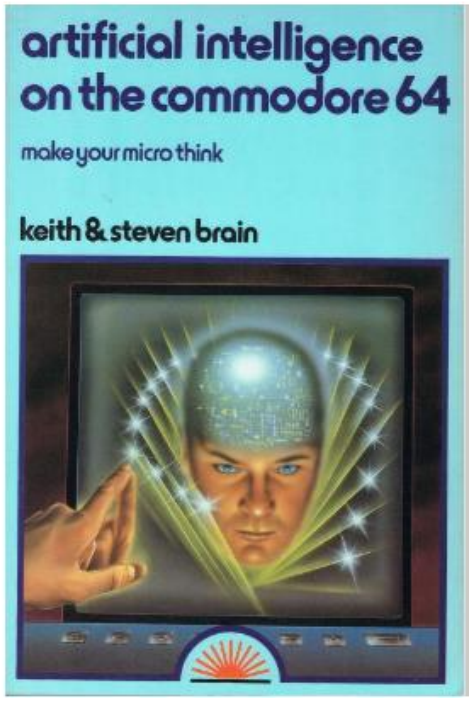


Deep Learning Cifar-10



Esperienza in Machine Learning





Computer Vision Lab
https://vision.unipv.it

CVMLab

ACFE Central Italy Chapter

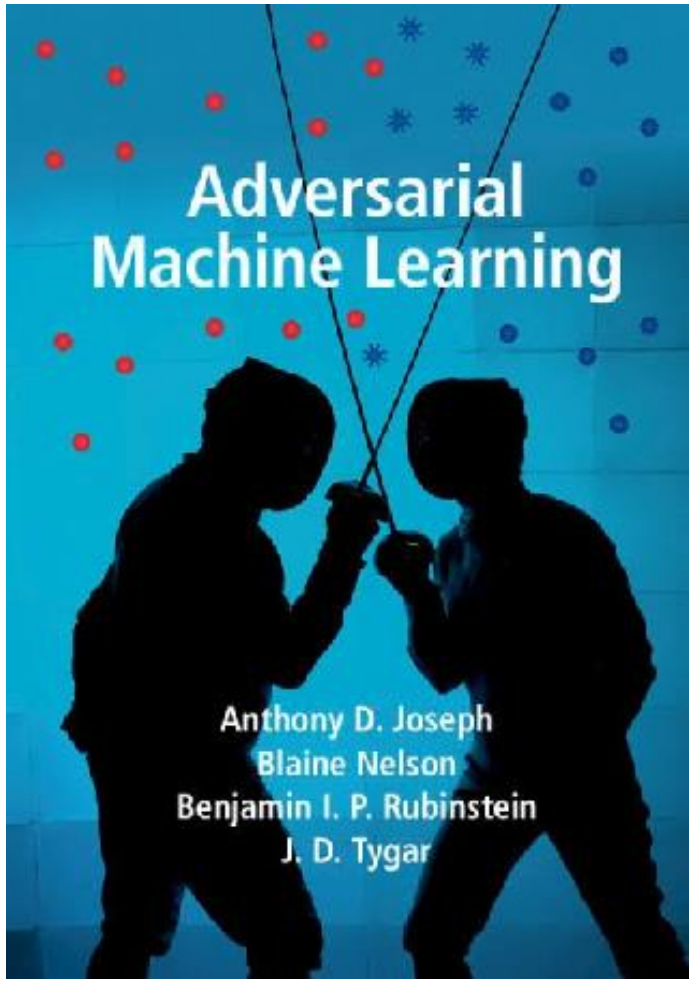
https://www.acfecentral.it/

1983

Fare una guerra nucleare senza distruggere noi stessi. Far sì che i computer imparassero dagli errori che noi non potevamo permetterci.

Adversarial Machine Learning

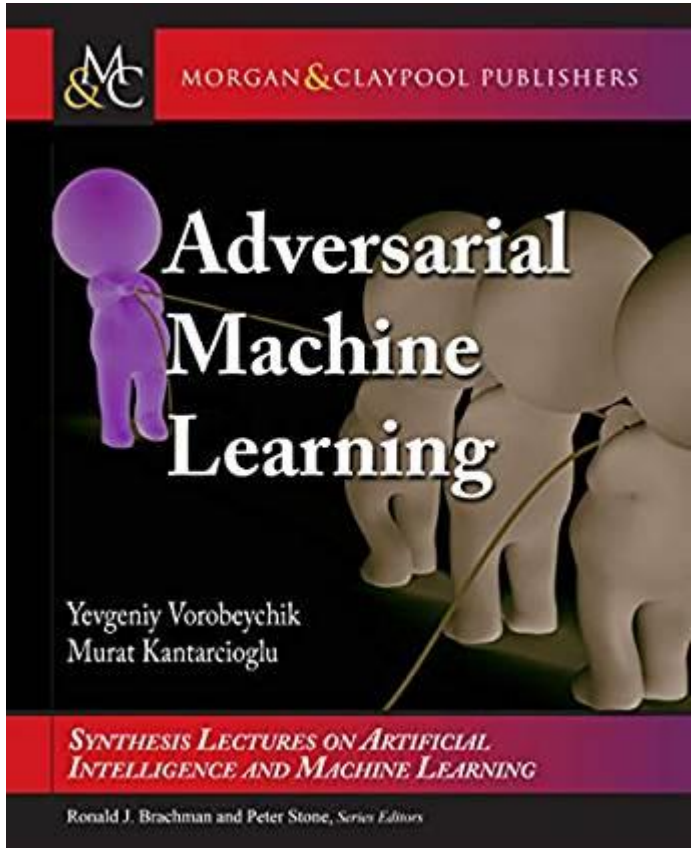
- apprendimento automatico in ambiente ostile, in cui qualcuno cerca di lottare, evitare, sabotare
- consentire un uso efficace e sicuro di automatico in ambienti ostili come anti spam, sicurezza informatica, pagamenti, riconoscimento biometrico, anti frode, ecc.
- unire Machine Learning e Sicurezza Informatica



https://www.cambridge.org/core/books/adversarial-machine-learning/C42A9D49CBC626DF7B8E54E72974AA3B

Adversarial Machine Learning

- Comportamento normale: piccole modifiche a input dovrebbero causare solo piccole modifiche all'output
- Comportamento in adversarial:
  - piccole modifiche a input possono causare grandi modifiche all'output
  - il sistema viene usato in modo non previsto



https://www.morganclaypool.com/doi/abs/10.2200/S00861ED1V01Y201806AIM039

Adversarial Machine Learning

- https://github.com/luizgh/adversarial_examples Box-constrained attacks (adversarial examples) in Tensorflow
- https://github.com/BorealisAI/advertorch advertorch text is a Python toolbox for adversarial robustness research implemented in PyTorch. AdverTorch contains modules for generating adversarial perturbations and defending against adversarial examples, also scripts for adversarial training.
- https://github.com/aupendu/awesome-adversarial-machine-learning A curated list of awesome adversarial machine learning resources
- https://github.com/jivoi/awesome-ml-for-cybersecurity A curated list of amazingly awesome tools and resources related to the use of machine learning for cyber security.

Machine Learning sotto osservazione malevole

Profiling dell'attaccante

Chi può attaccare:
- dipendente infedele
- concorrente sleale
- persone male intenzionate

Motivazioni di chi attacca:
- in cerca di pubblicità
- in cerca di guadagno illecito
- mostrare la sua bravura

Modus operandi dell'attaccante

1. Cerca notizie su chi ha creato il sistema
2. Cercar notizie su come funziona il sistema
3. Realizza attacco
4. Verifica delle difese
5. Analisi dei risultati
6. Cancella le tracce
7. Incasso del guadagno, diffusione del caso

Sondaggio IA e attacchi informatici

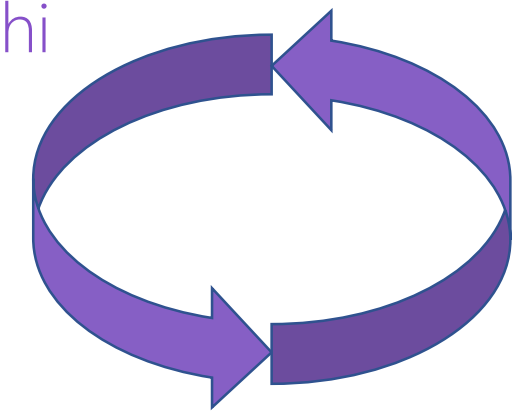
- Indagine nel 2017 di **Cylance** su oltre 100 esperti di sicurezza aziendale negli USA
- https://www.cylance.com/en-us/company/news-and-press/press-releases/cylance-2017-threat-report-provides-insight-into-attacks-prevented-with-artificial-intelligence.html
- Per 62% del campione l'intelligenza artificiale sarà utilizzata per attacchi informatici devastanti nei prossimi 12-18 mesi

Le tipiche difese di sicurezza IT non bastano

- Penetration test, antivirus, spyware, firewall ecc. servono per non fare modificare il software
- Ma non bastano
- Perché l'attacco usa configurazioni dei dati input che sono ritenute legittime
- Coinvolti IT Security Manager, Sviluppatori, Manager

Miglioramento continuo per fare sicurezza

1. Quali sono gli attacchi



2. Quali mi possono succedere

3. Cosa devo modificare

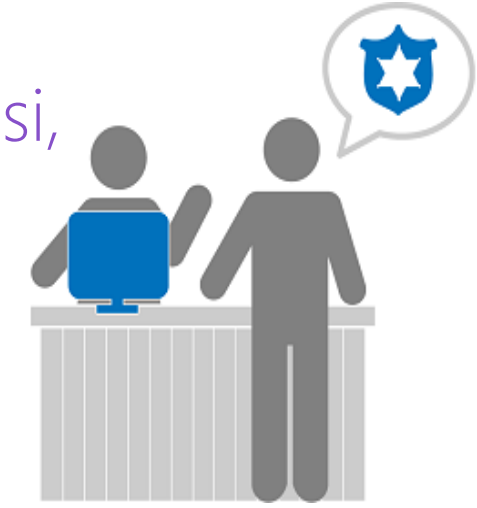
4. Esecuzione modifiche

1. Quali sono gli attacchi

a. Conoscere come funziona il sistema e chi lo ha creato
b. Usare input non previsto
c. Usare input previsto in modo malevolo
d. Usare sequenza malevola di input corretto
e. Avvelenare la fase di learning

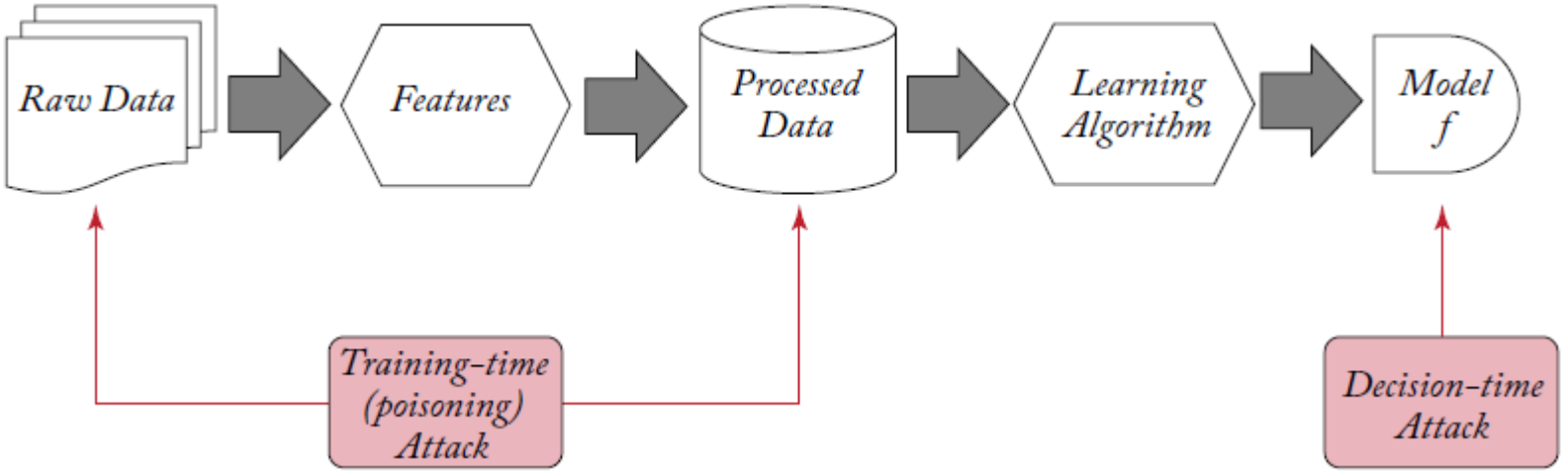
1.a Conoscere il sistema e chi lo ha creato

1. Azienda diffonde comunicato stampa su sistema anti frode per una banca
2. Su LinkedIn si cercano i dipendenti dell'azienda e annunci HR
3. Dal sito di loro Università si risale alla seduta di laurea, titolo tesi, relatore
4. Si analizza la produzione scientifica dell'Università
5. Si crea una ipotesi su quali strumenti ML possono avere usato
6. Ricavare dettagli con ingegneria sociale:
    1. Interazione nei social network con falsi profili verso i dipendenti
    2. Domande tecniche come richieste di aiuto da un collega o studente



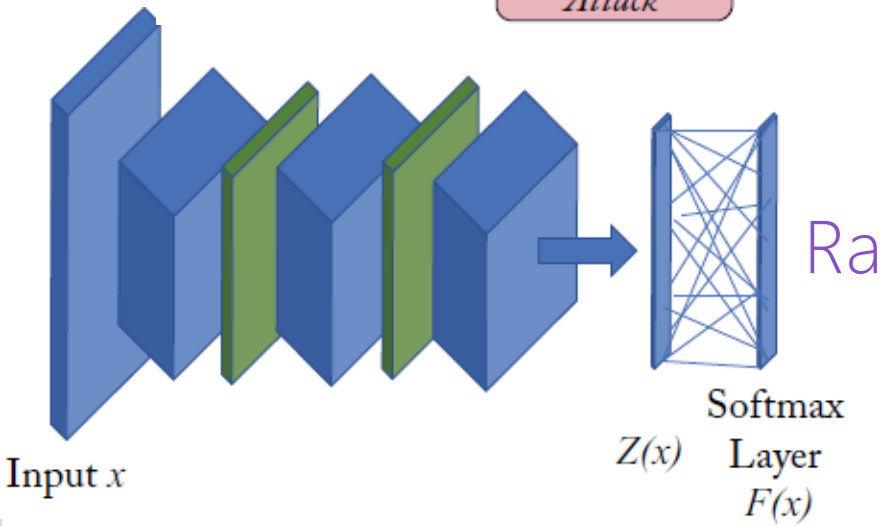
1.a Conoscere il sistema e chi lo ha creato

Rappresentazione dell'attacco nel flusso dei dati



Yevgeniy Vorobeychik, Murat Kantarcioglu
Adversarial Machine Learning
Editore Morgan & Claypool, pag. 20

Rappresentazione dell'attacco nel Deep Learning

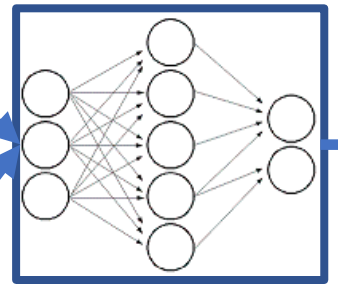


Yevgeniy Vorobeychik, Murat Kantarcioglu
Adversarial Machine Learning
Editore Morgan & Claypool, pag. 113

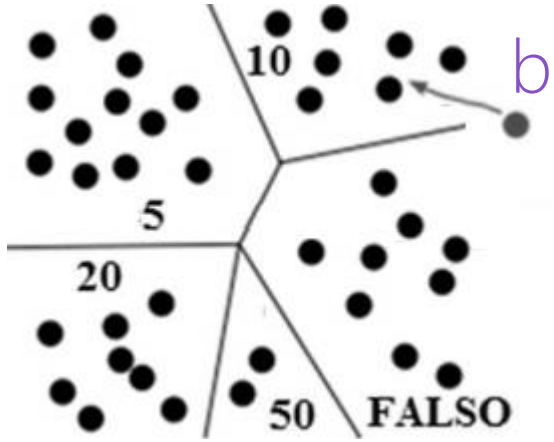
1.b Usare input non previsto

banconota vera

banconota falsa



benzina si/no

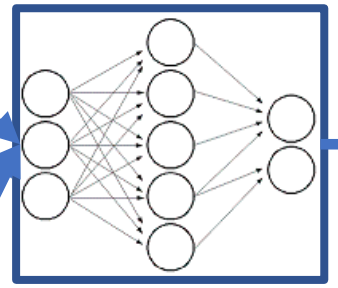


banconota

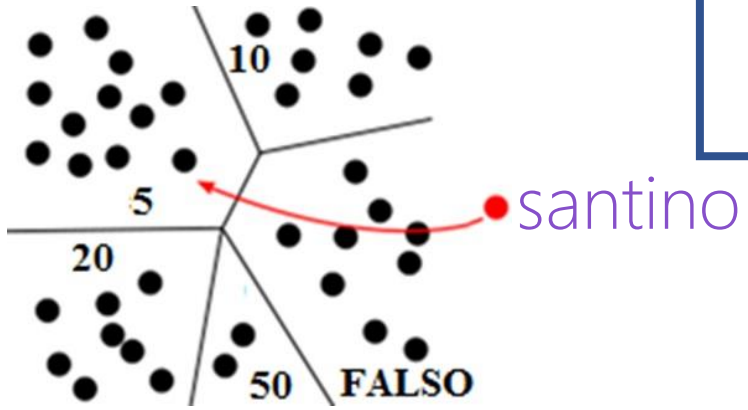
banconota vera

banconota falsa

santino



benzina si/no



santino

Rete Neurale MLP:

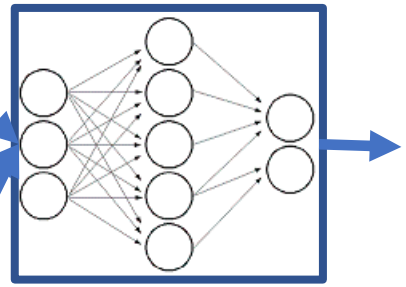
ogni punto viene posto in una classe predefinita

1.b Usare input non previsto

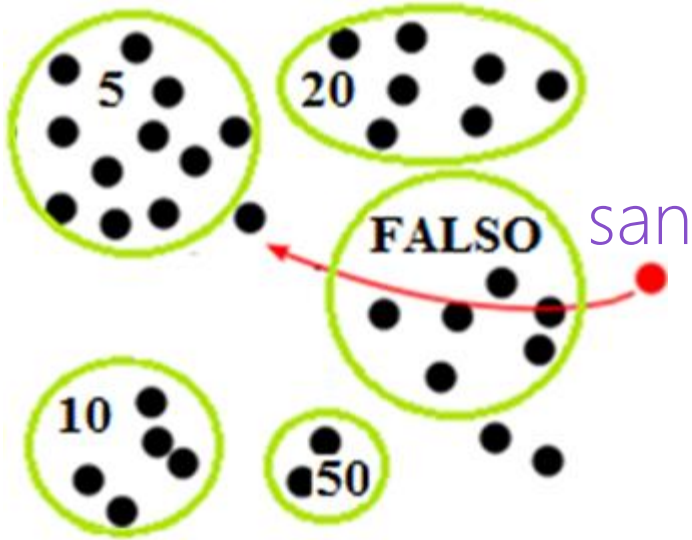
banconota vera

banconota falsa

santino



benzina si/no



Rete Neurale RBF:

ogni punto viene posto dentro o fuori un cluster

Studiare caratteristiche di input, output e metodo di decisione!

1.b Usare input non previsto

DolphinAttack: Inaudible Voice Commands

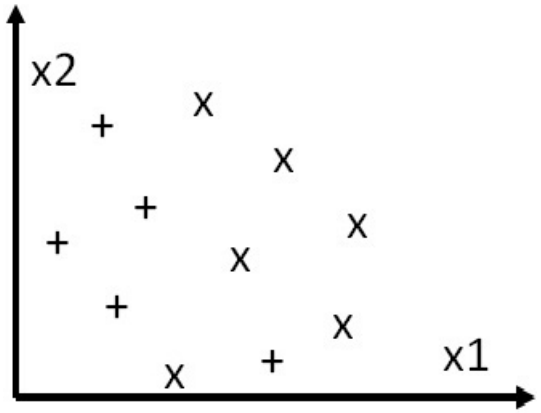
https://acmccs.github.io/papers/p103-zhangAemb.pdf

https://www.youtube.com/watch?v=vuICDaxbmg8
DolphinAttack Can Take Control of Siri and Alexa with Inaudible Voice Command

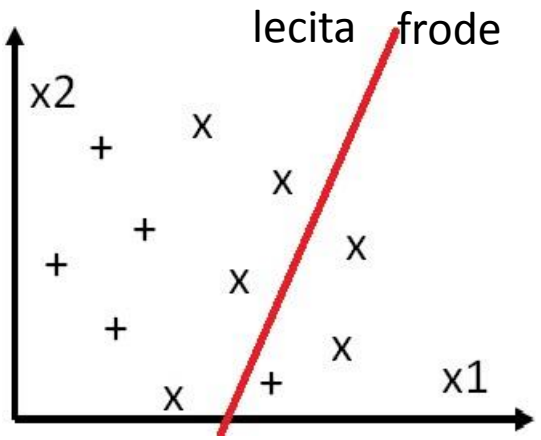
https://www.youtube.com/watch?v=21HjF4A3WE4
DolphinAttack: Inaudible Voice Command

1.c Usare input previsto in modo malevolo

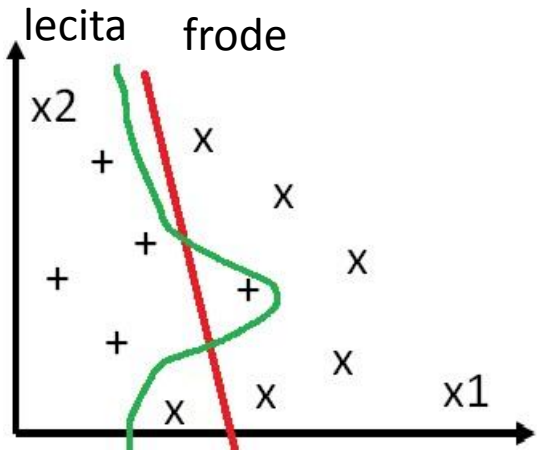
- Classificazione di transazione bancaria come frode X o lecita +
- La transazione viene descritta con due parametri x1 e x2



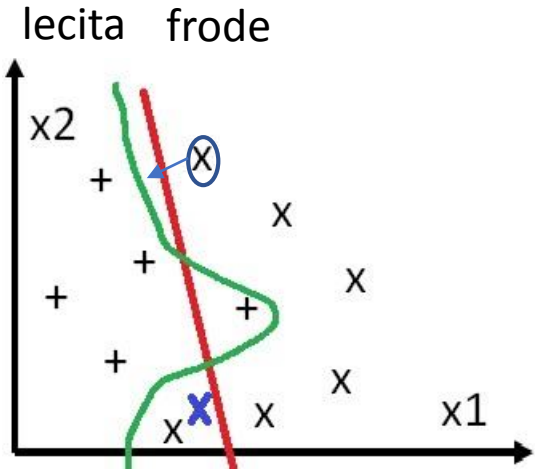
Spazio dei parametri, nessuna classificazione



Inizio del learning, linea rossa è prima classificazione



Fine del learning, linea rossa è classificazione non corretta al 100%, linea verde è il 100%

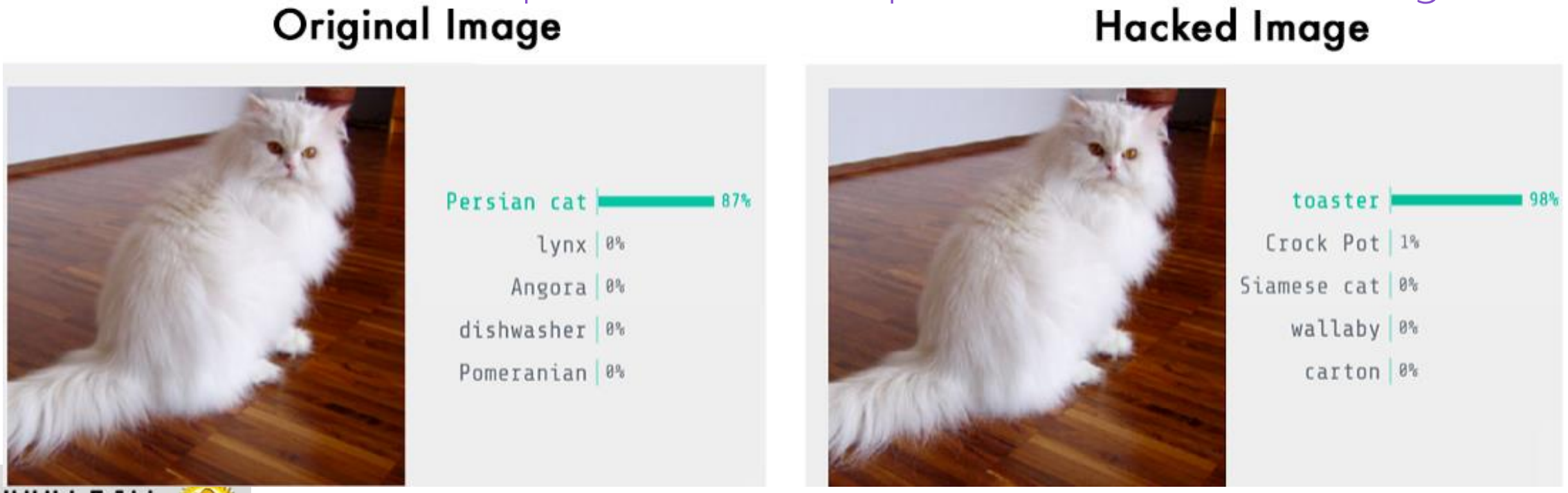


Simbolo blu è l'attacco: creo transazione frode ma caratteristiche simili a frode classificata come lecita

1.c Usare input previsto in modo malevolo

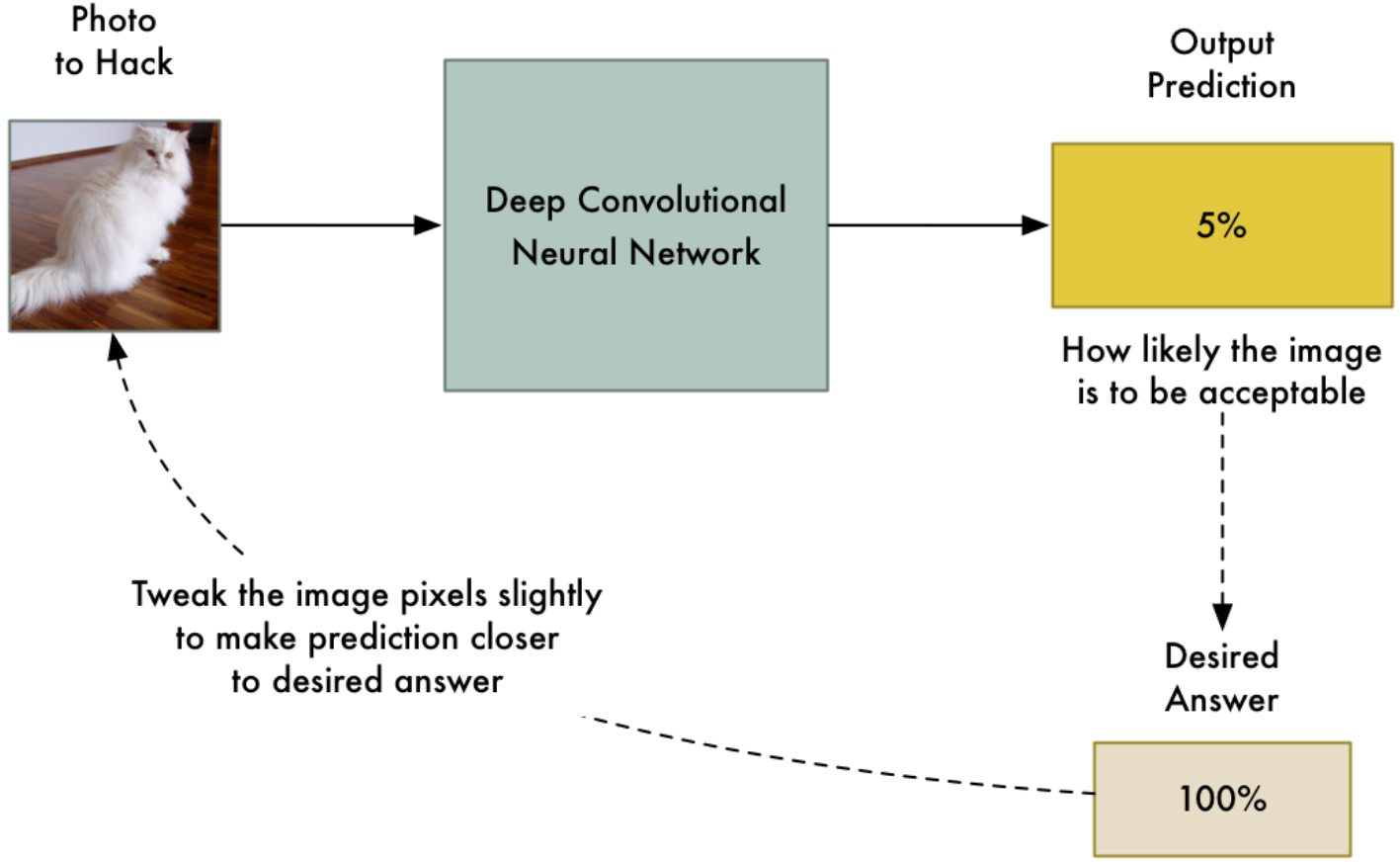
https://medium.com/@ageitgey/machine-learning-is-fun-part-8-how-to-intentionally-trick-neural-networks-b55da32b7196

conoscendo i pixel da cambiare e come farlo senza modificarne molto l'aspetto all'occhio umano, si può forzare la risposta errata ad un'immagine



1.c Usare input previsto in modo malevolo

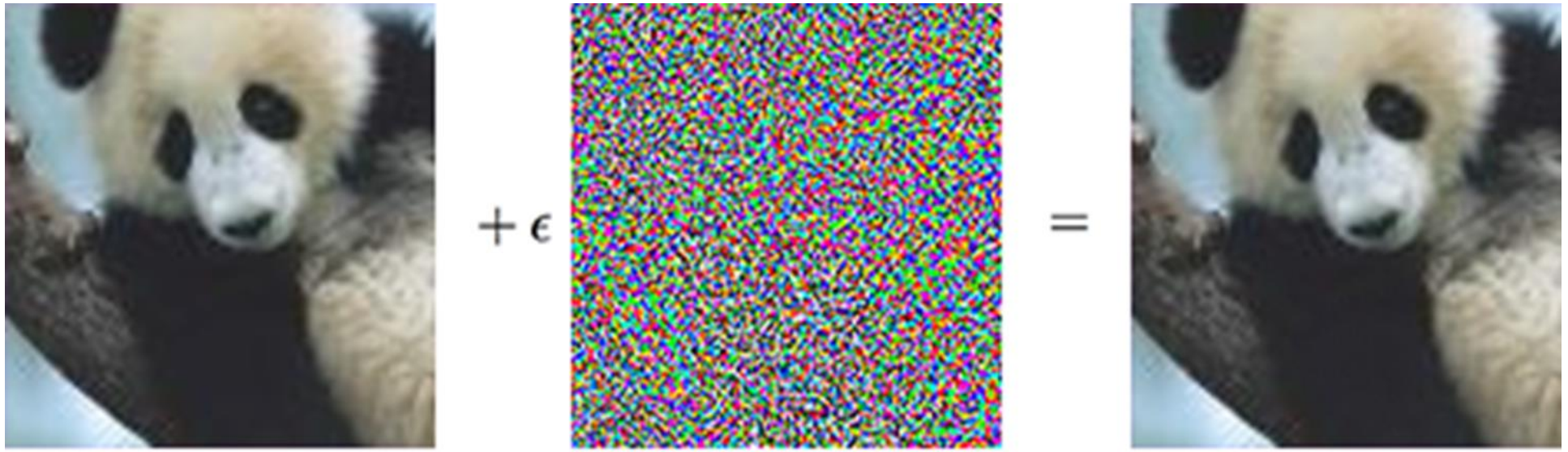
https://medium.com/@ageitgey/machine-learning-is-fun-part-8-how-to-intentionally-trick-neural-networks-b55da32b7196



Induciamo DCNN a pensare che il gatto è un tostapane modificando l'immagine fino a ingannarlo.

1.c Usare input previsto in modo malevolo

input creato dall'attaccante per fare commettere errore al Machine Learning e fargli dare la scelta voluta, l'occhio umano non nota la differenza tra le immagini https://arxiv.org/abs/1412.6572



1.c Usare input previsto in modo malevolo

Figure 3: An impersonation using frames. Left: Actress Reese Witherspoon (by Eva Rinaldi / CC BY-SA / cropped from https://goo.gl/a2sCdc). Image classified correctly with probability 1. Middle: Perturbing frames to impersonate (actor) Russel Crowe. Right: The target (by Eva Rinaldi / CC BY-SA / cropped from https://goo.gl/AO7QYu).

https://www.youtube.com/watch?v=6Xh1vuwnVhU

https://www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf

Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition

Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter

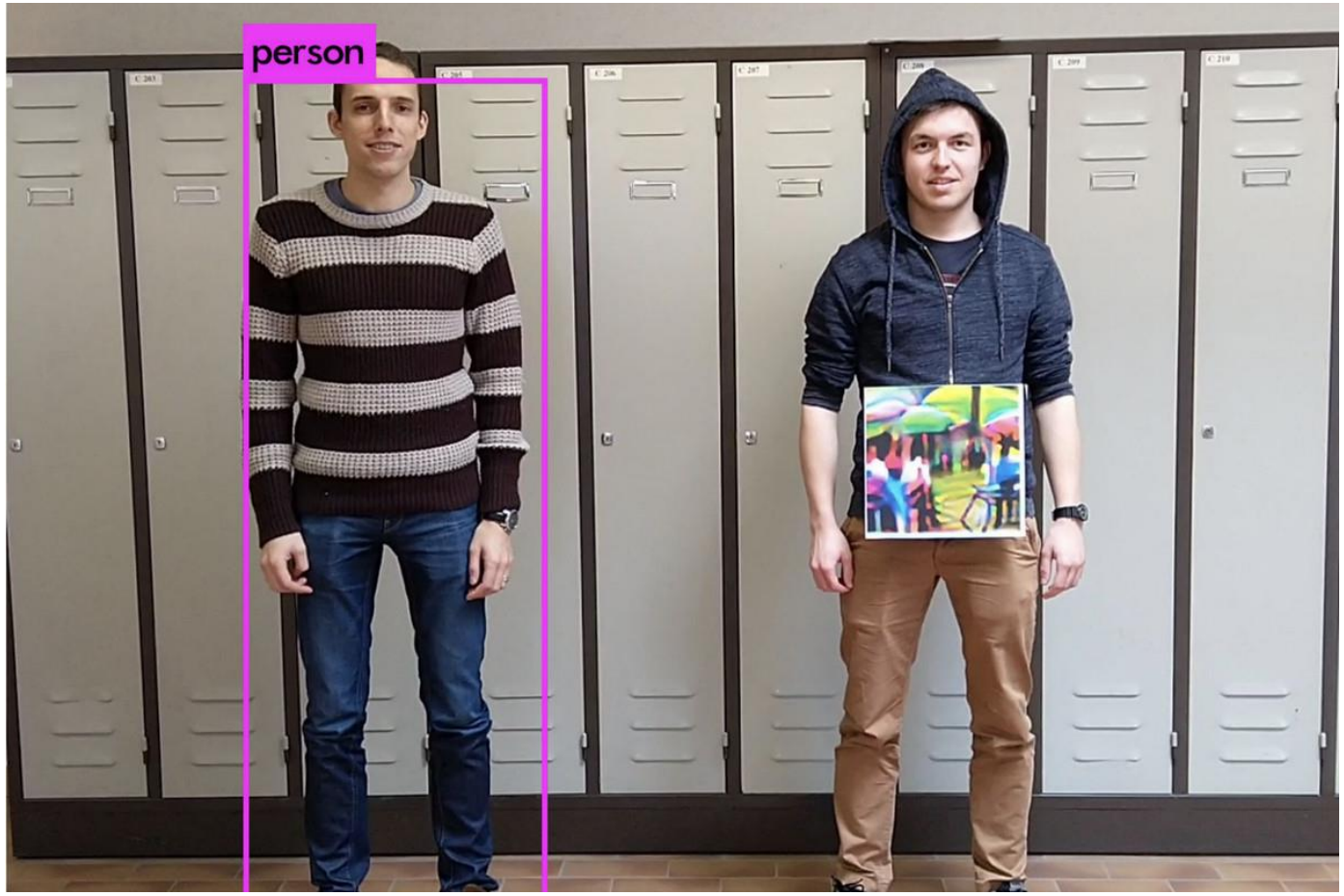
1.c Usare input previsto in modo malevolo

Incognito, maschera-gioiello per il viso che blocca software di riconoscimento facciale, rendendo **impossibile identificare la persona che la indossa**. Design creativa polacca Eva Nowak

https://noma-studio.pl/en/

https://www.dailybest.it/art/il-gioiello-per-evitare-il-riconoscimento-facciale/

1.c Usare input previsto in modo malevolo



- https://arxiv.org/pdf/1904.08653.pdf
- Fooling automated surveillance cameras: adversarial patches to attack person detection
- Attacco a Convolutional Neural Networks usato in YOLOv2 architecture object detector

1.c Usare input previsto in modo malevolo

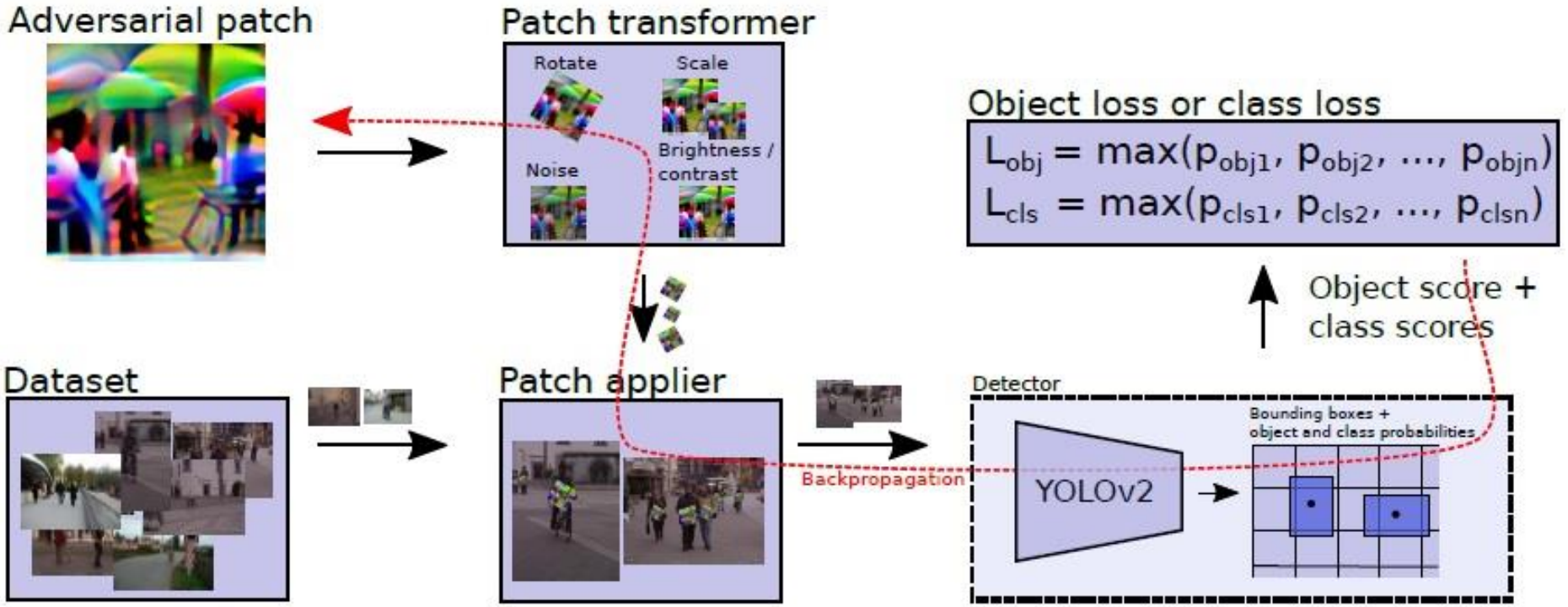


Figure 3: Overview of the pipeline to get the object loss.

Adversarial patch

1.c Usare input previsto in modo malevolo

Subtle Attack (86%)

Camouflage Art Attack (77%)

Targted
miss-classifiation

- Robust Physical-World Attacks on Deep Learning Visual Classification
- https://arxiv.org/pdf/1707.08945.pdf

1.c Usare input previsto in modo malevolo

A	B	C
-1	1	1

- valori input con simile ordine di grandezza vanno bene

A	B	C
1.000.000	1	1

- A fa sottovalutare importanza di B e C e l'algoritmo viene ingannato da A

1.d Usare sequenza malevola di input corretto

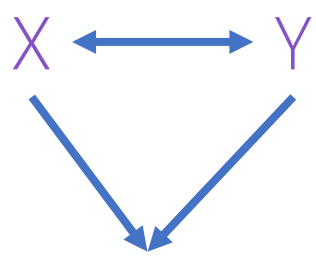
- Attaccante genera combinazioni di parole per:
  - mandare fuori servizio il chatbot
  - estrarre tutte le informazioni esistenti
  - accedere a informazioni riservate
  - fargli imparare parole proibite

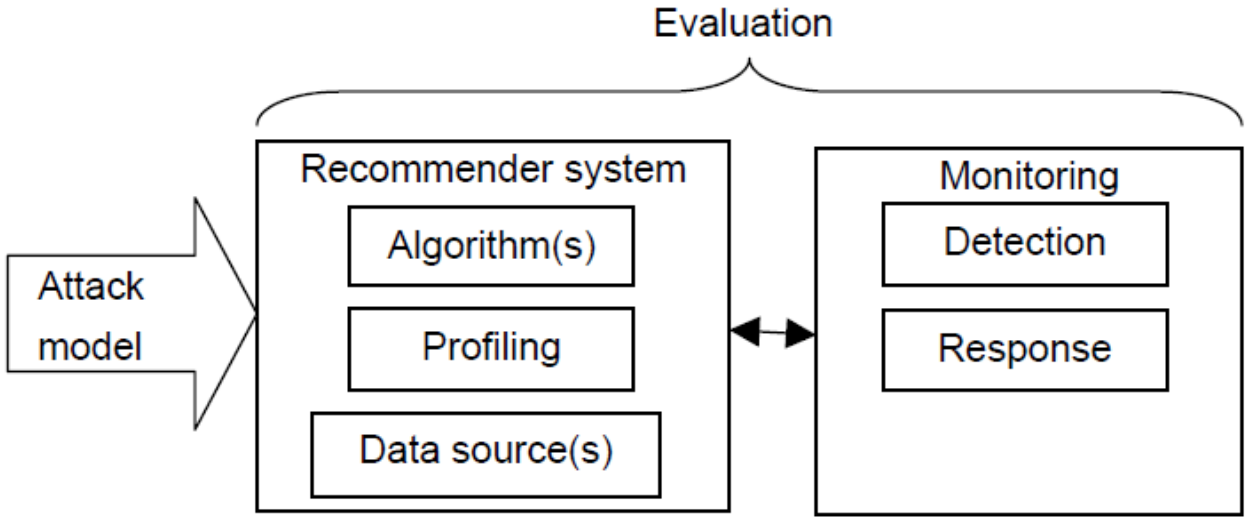
- https://cai.tools.sap/blog/chatbots-security-measures-you-need-to-consider/

https://www.repubblica.it/tecnologia/social-network/2016/03/24/news/tay_microsoft_razzista-136242109/

1.d Usare sequenza malevola di input corretto

- motore di raccomandazione suggerisce i prodotti
- algoritmo collaborativo basato su visite dei prodotti
- Product Push attack: spingere il proprio prodotto
- Product Nuke attack: seppellire il prodotto dei concorrenti





Sistema di difesa

1.d Usare sequenza malevola di input corretto

1. studiare i profili delle persone che visitano X e Y
2. creare falsi profili con le loro stesse caratteristiche
3. creare false visite al prodotto famoso X seguite subito da visita al prodotto Z
4. l'algoritmo associa i prodotti X e Z
5. i visitatori di X vedranno spesso il prodotto Z

1.e Avvelenare la fase di learning

- Poisoning Machine Learning
- Backdoored Neural Network, BadNet
- funziona bene sui dati per cui viene allenata, ma funziona anche su dati che l'autore della BadNet vuole fare passare di nascosto
- avvelenamento: tra i dati corretti sono nascosti i dati scorretti che BadNet non blocca
- https://arxiv.org/pdf/1708.06733v1.pdf
- Rischio nel Face Detection: l'attaccante inserisce la sua immagine come autorizzata

2.Quali mi possono succedere

- Scrivere documento di analisi dei rischi su:
  - quale utente fornisce input
  - come viene controllato l'input
  - fonte di input e output usati in learning
  - chi va verificato input e output in learning
  - quali danni può causare un output falsificato
  - chi ha sviluppato il sistema conosce le problematich
  - quante notizie esistono sul funzionamento del sistema
  - codice scritto da zero o preso da varie fonti
- Eseguire attacchi simulati per provare la resistenza

3.Cosa devo modificare

- Ripetere learning per coprire nuove combinazioni input-output
- Aumentare il numero di classi per aggiungere il diverso dal solito
- File log per registrare chi e come interagisce con Machine Learning
- Verifica periodica del corretto funzionamento

Qualcosa da leggere

- https://ai.googleblog.com/2018/09/introducing-unrestricted-adversarial.html
- https://www.wired.co.uk/article/artificial-intelligence-hacking-machine-learning-adversarial To cripple AI, hackers are turning data against itself
- https://medium.com/italian-ai-stories/isaisafe-4b787c6068ba Is Artificial Intelligence Safe?
- https://venturebeat.com/2017/05/29/what-happens-when-hackers-attack-chatbots/ What happens when hackers attack chatbots
- https://www.h2o.ai/blog/can-your-machine-learning-model-be-hacked/
- https://securityintelligence.com/humans-vs-machines-will-adversarial-ai-become-the-better-hacker/ Humans vs. Machines: Will Adversarial AI Become the Better Hacker?

White hat

- Un white hat è un hacker, ovvero un appassionato di informatica, esperto di programmazione, di sistemi e di sicurezza informatica in grado di introdursi in reti di computer al fine di aiutarne i proprietari a prendere coscienza di un problema di sicurezza nel rispetto quindi dell'etica degli hacker e si contrappone a chi viola illegalmente sistemi informatici, anche senza vantaggio personale, definito "black hat hacker".
- E se nascesse il white hat per l'intelligenza artificiale?

Conclusioni

- Algoritmo dotato di intelligenza artificiale non implica essere dotato di sicurezza informatica
- E' possibile creare input per fare sbagliare le decisioni
- Rischi per il Machine Learning in ambiente ostile e attaccabile
- Studiare bene il problema, caratteristiche di input e metodo di decisione
- Essere consapevoli dei problemi per migliorare il progetto e verificare il funzionamento online

www.robertomarmo.net
info@robertomarmo.net