Frailties of Deep Learning Programs

Julien Girard-Satabin (CEA LIST): julien.girard2@cea.fr

Preliminaries

Content warning

This course will mention physical and psychological violence, as well as pornography (no explicit depictions).

You and I

Me:

1. research engineer at CEA LIST, PhD in computer science
2. wants to explore the topics of safe AI for industries and citizens
3. citizen

You:

1. master students
2. future practitioniers of machine learning as designers
3. citizens

Our goal for this course

Me:

1. disseminate my work
2. spark interest on my research topic

You:

1. acquire technical knowledge on deep learning limitations
2. grasp a first glance on societal impact of deep learning software
3. be more informed if, and how, deploy ML programs

How will this session go

I will speak for most of the time, however feel free to interrupt me if you wish.

I will ask you some questions during the course (it's no exam, just to ensure some level of interaction)

Color code

Example definition

When those are on, those are formal definitions important to grasp

Example question

When those are on, it is an open question for you. No wrong answers, just interactions. You can type on the chat if you want.

Opening questions

- In 2012, AlexNet paper went out, marking the opening of the ”Third AI Spring”. Ten years later, here we are. Considering what you saw on previous courses, can you describe what you think of the evolution of the field?
- Considering your background, you may have chosen a lot of different studies. Can you mention one thing that motivated you to sign for this AI course?

Technical frailties of ML programs

Topic

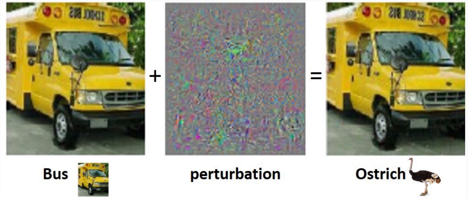
We will discuss now some of the most prominent frailties in modern machine learning.

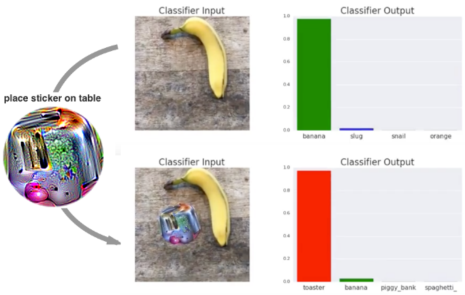
On the wording

1. ”bugs” implies the existence of a fix; most of the phenomena described here cannot be fixed without seriously impacting the program’s performance
2. ”exploits” assumes an attacker; a malicious intend is not needed to trigger those behaviours

Breaking the link between human and machine perception

- audio: https://youtu.be/Ho5jLKfoKSA?t=530
- video: https://youtu.be/MIbFvK2S9g8?t=20





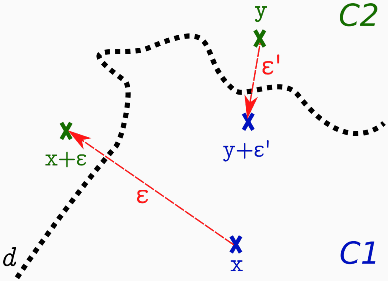
Several modalities of human perception can be abused

Formal definition

Let $\mathscr{D}_x$ be an input space, let $\mathscr{D}_y$ be an output space, let $f : \mathscr{D}_x \to \mathscr{D}_y$ be a neural network, and let $\mathscr{X} \in \mathscr{D}_x = \left\{x : \|x - x_0\|_p < \varepsilon\right\}$ be a set of perturbation (threat model):

Adversarial example

An adversarial example is a sample x_{adv}, $x_{adv} \in \mathscr{X}$ and $f(x) \neq f(x_{adv})$



Adversarial examples - How to craft an example ?

Theoretical formulation [CW16]

Given a sample $x_0 \in \mathbb{R}^{c \times h \times w}$, minimize $\|x - x_0\|_p^2$ such that $f(x) \neq f(x_0)$

Adversarial examples - How to craft an example ?

Theoretical formulation [CW16]

Given a sample $x_0 \in \mathbb{R}^{c \times h \times w}$, minimize $\|x - x_0\|_p^2$ such that $f(x) \neq f(x_0)$

Prohibitively difficult to solve (need to go through the whole input space)!

Adversarial examples - How to craft an example ?

Let ∇_{x_0} be the gradient operator for variable x_0, and let $\mathscr{L}(\theta, x_0, y)$ be the loss value of a neural network for parameters θ, a sample $x_0 \in \mathbb{R}^{c \times h \times w}$ and its ground true label y.

Fast Gradient Sign Method (FGSM)[GSS14]

Given a sample $x_0 \in \mathbb{R}^{c \times h \times w}$, $x = x_0 - \varepsilon * \text{sign}(\nabla_{x_0} \mathscr{L}(\theta, x_0, y))$

Simple approach and fast, but not optimal (ε is not optimized)

Adversarial examples - How to craft an example ?

Projected Gradient Descent (PGD)[Mad+17]

Given a sample $x_0 \in \mathbb{R}^{c \times h \times w}$, the least likely class for x_0 y_{ll}, a clip operator Π, iteratively build x with

1. $x^0 = x_0$
2. $x^{k+1} = x^k + \Pi(\varepsilon * \text{sign}(\nabla_{x_0} \mathscr{L}(\theta, x_0, y_{ll})))$

Number of iteration is the result of parameter search

More accurate than FGSM for moderate additional cost

All of those attacks require gradients

All of those attacks require gradients

The attacker needs to have direct access to gradients

How to craft an attack without gradient?

Given a sample $x_0 \in \mathbb{R}^{c \times h \times w}$, a perturbation δ, a distance metric (usually a norm) $\mathscr{D}$, a target label t:

Carlini & Wagner attack[CW16]

minimize $\mathscr{D}(x, x + \delta)$ such that
$x + \delta \in [0, 1]^{\mathrm{C \times H \times W}}$, $\operatorname{argmax} f(x + \delta) = t$

How to craft an attack without gradient?

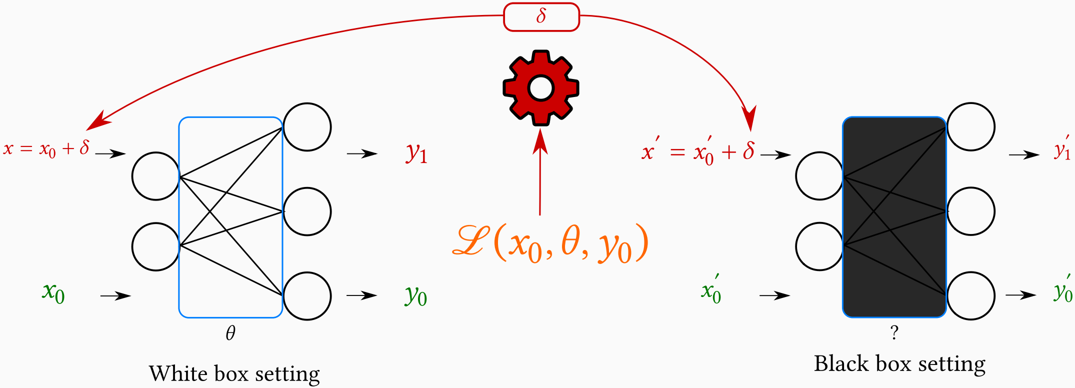
Given a sample $x_0 \in \mathbb{R}^{c \times h \times w}$, a perturbation δ, a distance metric (usually a norm) $\mathscr{D}$, a target label t:

Carlini & Wagner attack[CW16]

minimize $\mathscr{D}(x, x+\delta) + c * \mathrm{F}(x+\delta)$ such that
$x + \delta \in [0,1]^{\mathrm{C} \times \mathrm{H} \times \mathrm{W}}$
where $c \in \mathbb{R}$ and F a well chosen function using only logits

Considered the most efficient attack, but costly (optimization steps)

Adversarial examples are transferable [PMG16]



Adversarial examples - Taxonomy of attacks

1. White-box attacks require access to parameters (computing loss or gradients)
2. Black-box attacks only require to be able to compute the outputs

A possible approach is to learn a white-box model using a black-box as an oracle, then produce adversarial examples on it

Some theoretical insights

1. first explanation were considering the piecewise linearity as a possible explanation [GSS14]
2. more recent work revealed the possible example of ”robust” and ”non-robust” features, optimized by neural networks [Ily+19], or link the behaviour of adversarial robustness and noise robustness[For+19]

No clear consensus

Mitigations - Adversarial Training

Adversarial training [Mad+17]

1. find δ such that the loss is high
2. minimize average likelihood of the "adversarial loss" $\mathcal{J}$

$$\min_{\theta} \max \mathbb{E}_{(x,y)} \left[\mathcal{J}(\theta, x + \delta, y) \right]$$

Empirically boost robustness, but only on known attacks

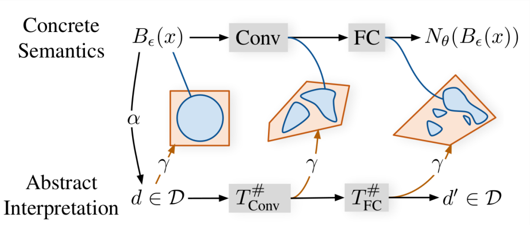
Mitigation: Formal Method robustness assessment

Problem: high input dimensions and number of variables makes test alone prohibitively difficult

Mitigation: Formal Method robustness assessment

Problem: high input dimensions and number of variables makes test alone prohibitively difficult

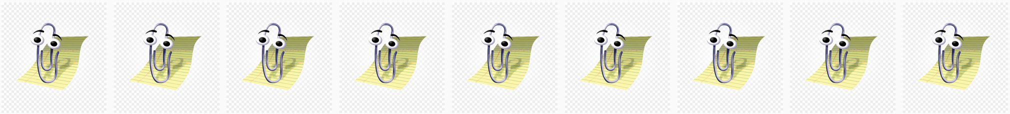
Use methods to compute sets instead of numbers to obtain formal guarantees on net's behaviour [SG19; Kat+19; Gir+21] (we are hiring!)



Wrapping things up for adversarial examples

- multiple modalities
- no absolute defense without huge costs on accuracy

Can be seen as an instanciation of the ”value alignment problem”[Wor15]



It does not stop there!

We saw frailties coming from the learning phase

It does not stop there!

We saw frailties coming from the learning phase

But there are other!

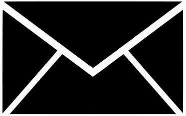
Sensitive data

SÉCURITÉ SOCIALE
l'Assurance Maladie
Agir ensemble, protéger chacun

Healthcare

Justice and criminal background

Military assets



Private information (w.r.t. GDPR)

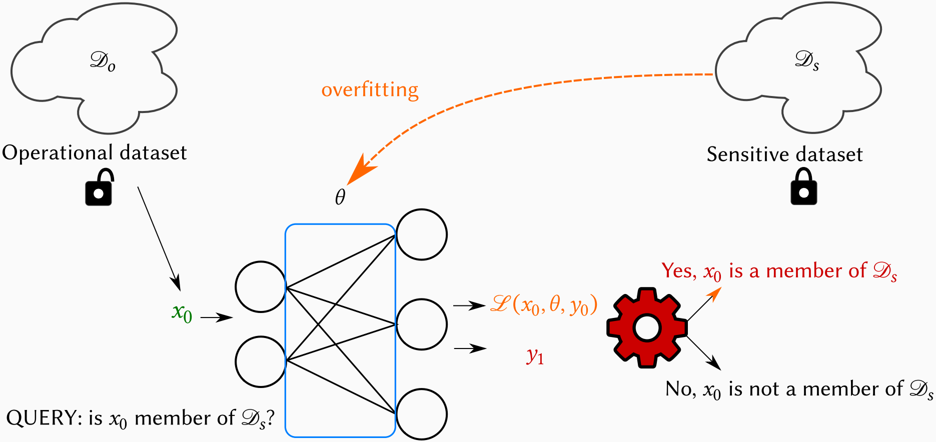
The problem of dataset privacy

Let $\mathscr{D}_s$ be a dataset with sensitive data, $\mathscr{D}_o$ be a dataset on operational data, $\mathscr{D}_l$ be a logit space.

Dataset privacy

Given a network $f : \mathscr{D}_o \rightarrow \mathscr{D}_l$ trained on $\mathscr{D}_s$, how to measure the amount of retrievable data from $\mathscr{D}_s$ when only given access to $\mathscr{D}_o$?

Membership inference



Membership inference - Why does it work?

- overfitting on training data is commonplace
- overfitting result in small variability on logits values between samples from train and test

Membership inference - Why does it work?

- overfitting on training data is commonplace
- overfitting result in small variability on logits values between samples from train and test

A classification pipeline trained on those logits differences [Sho+17] or labels only[Cho+21] can be queried to check if a sample belongs to $\mathcal{D}_s$

Membership inference - variants

Distillate knowledge of a black-box dataset on a white-box, allowing to ”steal” parameters[Tra+16]

Measuring and mitigating privacy leakages

1. Deep learning with differential privacy[Aba+16] aims to learn noised data to limit information embedding in the program
2. Deploying several models trained on subparts of $\mathscr{D}_s$ or with various amounts of noise can mitigate

But wait, there is more!

We saw frailties coming from the learning phase

But wait, there is more!

We saw frailties coming from the learning phase

We saw frailties coming from the evaluation phase

But wait, there is more!

We saw frailties coming from the learning phase

We saw frailties coming from the evaluation phase

But there are other!

An promising venue - Data intelligence

Deep Learning is impossible without huge corpuses of data

An promising venue - Data intelligence

Deep Learning is impossible without huge corpuses of data

Data analysis and data cleaning is standard data science practices

An promising venue - Data intelligence

Deep Learning is impossible without huge corpuses of data

Data analysis and data cleaning is standard data science practices

How could we certify datasets? What kind of properties a ”good” dataset should respect?

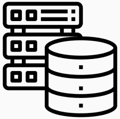
How are datasets crafted

High reliance on microworking platforms: Amazon Mechanical Turk, Upwork, Lionbridge (about 230 000 microworkers in France[Cas+19], creating new forms of job insecurity[Tub21])

How are datasets crafted

Data collection
(webscraping, sensor aggregation)

Partition into human concepts
(expert labelling, microworking)



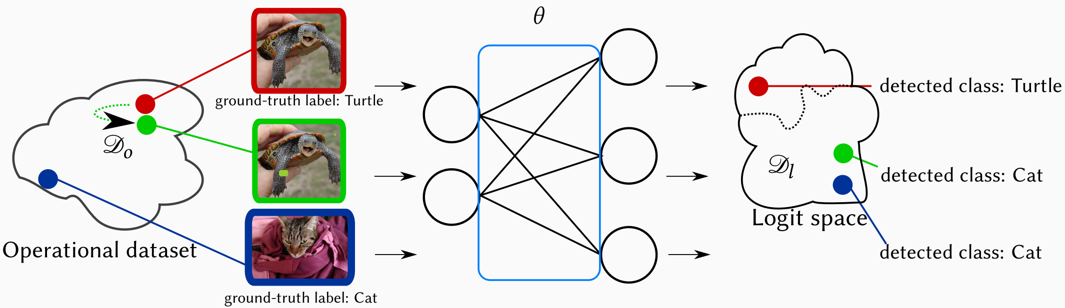
```
import os
import subprocess
import pandas as pd
import numpy as np
import torch
from torchvision.datasets import VisionDataset
from torchvision.datasets import MNIST, CIFAR10, FashionMNIST
from torchvision import transforms
from torchvision.datasets.folder import default_loader
from torchvision.datasets.utils import download_url, extract_archive
from torch.utils.data import Dataset
```

Data consolidation processes before production

Frailty in data can take various forms

1. human errors in labelling
2. bias (see previous course) leads to lower accuracy and unacceptable behaviour on real-world data
3. but it can also be crafted...

Poison crafting - Principle



Dataset poisoning

General framework of dataset poisoning: change a dataset to change a model's behaviour in production setting[Gol+21]

It can be seen as a rephrase of adversarial attacks, but focused on the dataset

Poison crafting - label shuffling

Conceptually simple attack: swap labels between instances

Cons: can be spotted if you look at the dataset with prior knowledge

Poison crafting - Feature collision

Given a neural network logits f_l, let x_b be a base sample of label l_b, let x_t be a target sample of label l_t, and let x a poisoned sample.

Problem of feature collision

$$\min_{x} \|f_l(x) - f_l(x_t)\|_2^2 + \beta \|x - x_b\|_2^2$$

Poison crafting - bilevel optimization

Given a neural network f with parameters θ let x_t be a target sample of label l_t, let x a poisoned sample.

Bilevel optimization of poison crafting [HGF20]

$$\min_{x} \mathcal{L}(f(x^t, \theta^{'}), y^{adv})$$

subject to

$$\theta^{'} = \text{argmin}_{\theta} \mathcal{L}(f(x, \theta), \mathcal{Y})$$

A neural network with initial parameters θ will classify x^t into y^{adv}

Technical implications

Works on any dataset and especially on transfer learning datasets

Very few samples (≈ 50 on CIFAR-10) to produce results[Sha+18]

Mitigations on models - Poisoning attacks

- finding outliers in the input space
- identifying poisoned models (trigger detection using a meta-classifier)
- randomized smoothing

All of those defenses require access to either the training pipeline, or the full model

Mitigations on datasets - Poisoning attacks

1. relying on experts for labelling
2. bias detection via careful data pre-analysis
3. debiasing techniques, for instance[Meh+19]

And more issues we do not have time to work on

Manipulation of saliency maps for ”explanation”[Dom+19], trojan attacks and adversarial reprogramming[EGS18]...

Future works

What kind of properties would you like to have on the datasets you use everyday?

Societal frailties of ML programs

Deep fakes

From `thispersondoesnotexist.com`...

Deep fakes

From `thispersondoesnotexist.com`... ...to
`https://www.youtube.com/watch?v=8dKux8-ZmCI`

Deep fakes

Deep fakes are data crafted using Generative Models and Adversarial Training (not to be confused with Adversarial Training as defense against adversarial examples) developed to impersonate someone

Deep fakes - public opinion manipulation

On modern Web, an idea propagates much faster if it induces an emotional response[CGP15]

Deep fakes - public opinion manipulation

On modern Web, an idea propagates much faster if it induces an emotional response[CGP15]

What if a deep fake showing Putin asking its troops to invade Ukraine would show up now?

Deep fakes - private harms

It is possible to synthetize deep fakes to impersonate people with few samples (<10) using transfer learning.

Possible misuses include ”revenge porn”: the act of leaking sexual content intended to be kept private after a breakup

Machine learning - opinion manipulation

Recommendation algorithms that drives Facebook feeds are aiming for user retention, not for fair information representation

Machine learning - opinion manipulation

Recommendation algorithms that drives Facebook feeds are aiming for user retention, not for fair information representation

How could a climate change supporter increase its virability on modern social platforms?

Machine learning - Biases perpetuation

COMPAS system[Mat+16] perpetuates racial biases in the data

Machine learning - Biases perpetuation

COMPAS system[Mat+16] perpetuates racial biases in the data

Programs designed from data are as biased as data is

Machine learning - Biases perpetuation

COMPAS system[Mat+16] perpetuates racial biases in the data

Programs designed from data are as biased as data is

Data is but a model of the world

Machine Learning - Where is Democracy?

Biometrics recognition[21]:

- costs public money
- is usually never subject to democratic discussions
- effects are still to be evaluated[com21]

Who are the ”users” of machine learning programs?

Developers and designers give to their client:industries, governments...

Who are the ”users” of machine learning programs?

Developers and designers give to their client:industries, governments...

... who will use it on data from citizens or consumers

Who are the "users" of machine learning programs?

Developers and designers give to their client:industries, governments...

... who will use it on data from citizens or consumers

End-users and targets are different class of people

Deep learning - How to empower people?

Give tools and processes to take decisions in a rational fashion, with sufficient information (Ivan Illich, La convivialité)

This is not a scientist-only job

Examples

1. ML to detect deepfakes (arms-race incoming)
2. ML for safety (predictive maintenance in industry)
3. ML for privacy preservation (Fawkes tool)

Open question

Can you propose some examples or ideas on how machine learning could be used for social good?

Bibliography i

References

Surveillance sonore : LQDN attaque l'expérimentation d'Orléans. La Quadrature du Net. Dec. 14, 2021. URL: https://www.laquadrature.net/2021/12/14/surveillance-sonore-lqdn-attaque-lexperimentation-dorleans/ (visited on 02/14/2022) (cit. on p. 68).

Bibliography ii

Martin Abadi, Andy Chu, Ian Goodfellow, Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. "Deep Learning with Differential Privacy". In: 23rd ACM Conference on Computer and Communications Security (ACM CCS). 2016, pp. 308–318. URL: https://arxiv.org/abs/1607.00133 (visited on 05/15/2019) (cit. on p. 36).

Bibliography iii

Antonio A. Casilli, Paola Tubaro, Clément Le Ludec, Marion Coville, Maxime Besenval, Touhfat Mouhtare, and Elinor Wahal. Le Micro-Travail En France. Derrière l'automatisation, de Nouvelles Précarités Au Travail ? Research Report. Projet de recherche DiPLab, May 2019. URL: https://hal.archives-ouvertes.fr/hal-02139528 (visited on 02/09/2022) (cit. on p. 43).

Bibliography iv

CGP Grey, director. This Video Will Make You Angry. Mar. 10, 2015. URL: https://www.youtube.com/watch?v=rE3j_RHkqJc (visited on 02/14/2022) (cit. on pp. 60, 61).

Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-Only Membership Inference Attacks. Jan. 21, 2021. arXiv: 2007.14321 [cs, stat]. URL: http://arxiv.org/abs/2007.14321 (visited on 03/18/2021) (cit. on pp. 33, 34).

Bibliography v

Cour des comptes. Le plan de vidéoprotection de la préfecture de police de Paris. Cour des comptes. 2021. URL: https://www.ccomptes.fr/fr/publications/le-plan-de-videoprotection-de-la-prefecture-de-police-de-paris (visited on 02/15/2022) (cit. on p. 68).

Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. Aug. 16, 2016. arXiv: 1608.04644 [cs]. URL: http://arxiv.org/abs/1608.04644 (visited on 11/07/2018) (cit. on pp. 13, 14, 19, 20).

Bibliography vi

Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. "Explanations Can Be Manipulated and Geometry Is to Blame". In: Advances in Neural Information Processing Systems. Vol. 32. Curran Associates, Inc., 2019. URL: https://papers.nips.cc/paper/2019/hash/bb836c01cdc9120a9c984c525e4b1a4a-Abstract.html (visited on 01/21/2022) (cit. on p. 54).

Bibliography vii

Gamaleldin F. Elsayed, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial Reprogramming of Neural Networks. June 28, 2018. arXiv: 1806.11146 [cs, stat]. URL: http://arxiv.org/abs/1806.11146 (visited on 01/12/2019) (cit. on p. 54).

Bibliography viii

Nic Ford, Justin Gilmer, Nicolas Carlini, and Dogus Cubuk. Adversarial Examples Are a Natural Consequence of Test Error in Noise. Jan. 29, 2019. arXiv: 1901.10513 [cs, stat]. URL: http://arxiv.org/abs/1901.10513 (visited on 02/12/2019) (cit. on p. 23).

Bibliography ix

Julien Girard-Satabin, Aymeric Varasse, Marc Schoenauer, Guillaume Charpiat, and Zakaria Chihani. DISCO Verification: Division of Input Space into COnvex Polytopes for Neural Network Verification. May 17, 2021. arXiv: 2105.07776 [cs]. URL: http://arxiv.org/abs/2105.07776 (visited on 05/21/2021) (cit. on pp. 25, 26).

Bibliography x

Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li, and Tom Goldstein. Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses. Mar. 31, 2021. arXiv: 2012.10544 [cs]. URL: http://arxiv.org/abs/2012.10544 (visited on 01/06/2022) (cit. on p. 47).

Bibliography xi

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. Dec. 19, 2014. arXiv: 1412.6572 [cs, stat]. URL: http://arxiv.org/abs/1412.6572 (visited on 10/24/2018) (cit. on pp. 15, 23).

W Ronny Huang, Jonas Geiping, and Liam Fowl. "MetaPoison: Practical General-purpose Clean-label Data Poisoning". In: (2020), p. 12 (cit. on p. 50).

Bibliography xii

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial Examples Are Not Bugs, They Are Features. May 6, 2019. arXiv: 1905.02175 [cs, stat]. URL: http://arxiv.org/abs/1905.02175 (visited on 05/17/2019) (cit. on p. 23).

Bibliography xiii

Guy Katz, Derek A. Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zeljić, David L. Dill, Mykel J. Kochenderfer, and Clark Barrett. "The Marabou Framework for Verification and Analysis of Deep Neural Networks". In: Computer Aided Verification. Ed. by Isil Dillig and Serdar Tasiran. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 443–452. ISBN: 978-3-030-25540-4. DOI: 10.1007/978-3-030-25540-4_26 (cit. on pp. 25, 26).

Bibliography xiv

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. June 19, 2017. arXiv: 1706.06083 [cs, stat]. URL: http://arxiv.org/abs/1706.06083 (visited on 10/24/2018) (cit. on pp. 16, 24).

Bibliography xv

Surya Mattu, Julia Angwin, Jeff Larson, and Lauren Kirchner. Machine Bias. ProPublica. 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=nEh5WNViIayEtqf96qVA8Dp-s2YDMY-f (visited on 03/12/2021) (cit. on pp. 65–67).

Bibliography xvi

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. Sept. 17, 2019. arXiv: 1908.09635 [cs]. URL: http://arxiv.org/abs/1908.09635 (visited on 03/12/2021) (cit. on p. 53).

Bibliography xvii

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in Machine Learning: From Phenomena to Black-Box Attacks Using Adversarial Samples. May 23, 2016. arXiv: 1605.07277 [cs]. URL: http://arxiv.org/abs/1605.07277 (visited on 12/18/2018) (cit. on p. 21).

Gagandeep Singh and Timon Gehr. “Boosting Robustness Certification of Neural Networks”. In: International Conference on Learning Representations (ICLR). 2019, p. 12 (cit. on pp. 25, 26).

Bibliography xviii

Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks. Apr. 2, 2018. arXiv: 1804.00792 [cs, stat]. URL: http://arxiv.org/abs/1804.00792 (visited on 10/24/2018) (cit. on p. 51).

Bibliography xix

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. "Membership Inference Attacks Against Machine Learning Models". In: 2017 IEEE Symposium on Security and Privacy (SP). 2017 IEEE Symposium on Security and Privacy (SP). San Jose, CA, USA: IEEE, May 2017, pp. 3–18. ISBN: 978-1-5090-5533-3. DOI: 10.1109/SP.2017.41. URL: http://ieeexplore.ieee.org/document/7958568/ (visited on 11/16/2018) (cit. on pp. 33, 34).

Bibliography xx

Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. "Stealing Machine Learning Models via Prediction APIs". In: 25th USENIX Security Symposium (USENIX Security 16). 2016, pp. 601–618. URL: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer (visited on 11/16/2018) (cit. on p. 35).

Bibliography xxi

Paola Tubaro. “Disembedded or Deeply Embedded? A Multi-Level Network Analysis of Online Labour Platforms”. In: Sociology (Jan. 31, 2021), p. 003803852098608. ISSN: 0038-0385, 1469-8684. DOI: 10.1177/0038038520986082. URL: http://journals.sagepub.com/doi/10.1177/0038038520986082 (visited on 08/10/2021) (cit. on p. 43).

Bibliography xxii

World Economic Forum, director. Value Alignment | Stuart Russell. May 24, 2015. URL: https://www.youtube.com/watch?v=WvmeTaFc_Qw (visited on 08/24/2021) (cit. on p. 27).