Learnability Theory

Stasinos Konstantopoulos

October 2023

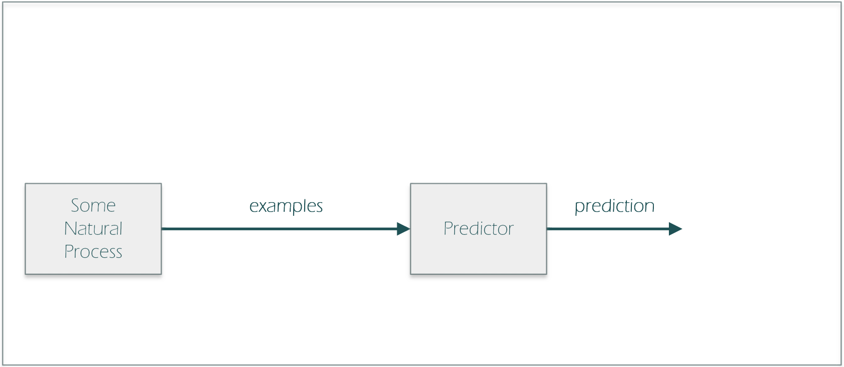
Table of contents

PAC Learnability and Schaffer's Law

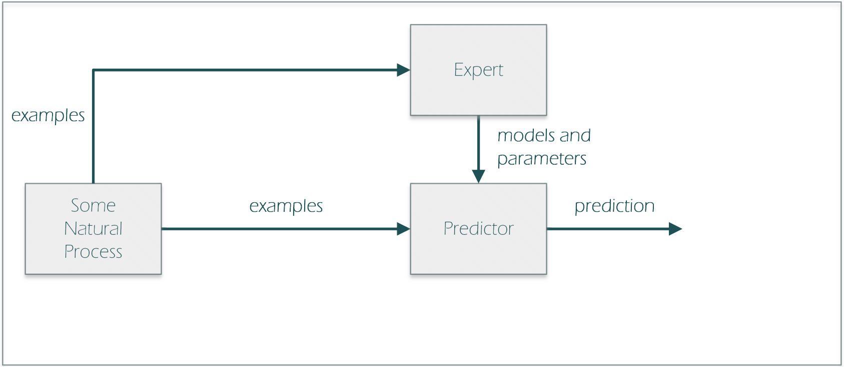
PAC learning is a general framework for obtaining theoretical results about machine learning tasks.

Extremely broad results can be proven within PAC, that make no reference to specific tasks or machine learning methods.

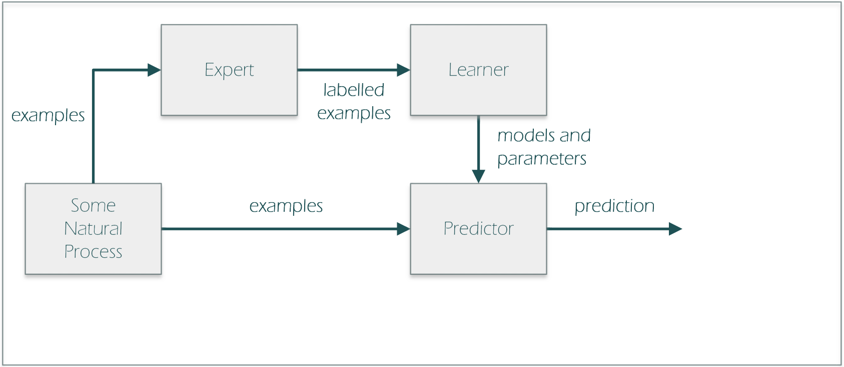
General Setting



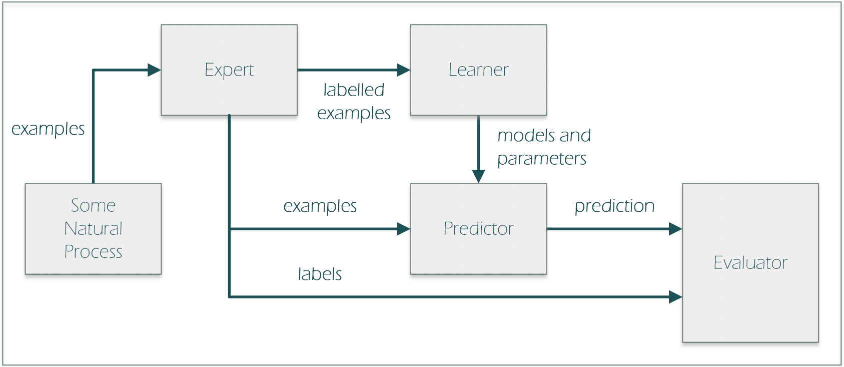
General Setting



General Setting



General Setting



How good is the model we learned?

We can use the evaluation session to *estimate* how well the learner *is likely* to perform in the future:

- the model accuracy is *approximately* known
- this approximation is *probably* accurate, but might also be completely off

To formalize this statement, we need to involve all learning parameters and conditions:

- How hard was the task?
- How good were the training examples?
- How many training examples did the learner see?

The Iris dataset

The Iris flower data set:

- Fifty samples from each of three types of the Iris flower, collected in Canada in 1935
- Sepal length and width, petal length and width
- Fisher (1936) manually developed a linear discriminant model to distinguish the species from each other

Iris Setosa

Iris Versicolor

Iris Virginica

Vapnik-Chervonenkis dimension

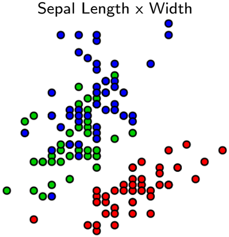
I. Setosa vs. the other is an easy learning situation:

- A linear discriminator solves this task
- This amounts to learning the three parameters in $ax + by + c = 0$

I. Versicolor vs. I. Virginica is harder:

- many parameters to fit

How can we formalize the complexity of a learning situtation, the complexity of training a learner to recognize different concepts in a domain?



Red: I. Setosa
Green: I. Versicolor
Blue: I. Virginica

Vapnik-Chervonenkis dimension

- A hypothesis space H (all the different hypotheses that a learner can formulate) *shatters* a domain iff H can fit *any* concept over the domain
- The *VC dimension* of H measures its complexity as the size of the largest subset of the domain that is shattered by H
- Naturally, $2^{\mathrm{VC}(H)} \leq |H|$, therefore $\mathrm{VC}(H) \leq \log(|H|)$
- Linear discriminator: $\mathrm{VC} = d + 1$, d dimensions (features).
- Rules: $\mathrm{VC} = 3^{nk}$, n rules, k terms each

Vapnik-Chervonenkis dimension: Real features

Leboeuf et al. (2020) shoed that the VC dimension of decision trees with ℓ real-valued features is the largest integer d such that

$$2\ell \geq \binom{d}{\mathrm{int}(d/2)}$$

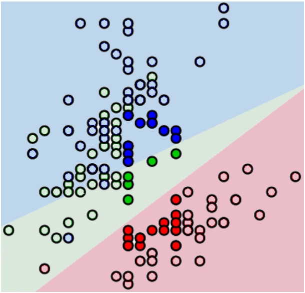
from which follows that for a binary tree with N internal nodes, the VC is of order $N \log(N\ell)$

IID: Quality of the training data

The training data should be sampled from the same distribution as the testing data.

The sampling of the training data should select datapoints independently from each other.

Sepal Length x Width



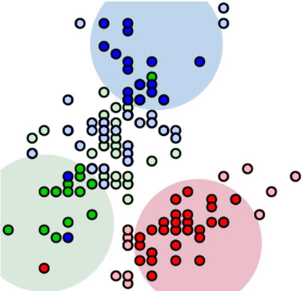
Red: I. Setosa
Green: I. Versicolor
Blue: I. Virginica

IID: Quality of the training data

The training data should be sampled from the same distribution as the testing data.

The sampling of the training data should select datapoints independently from each other.

Sepal Length x Width



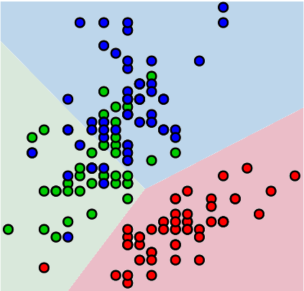
Red: I. Setosa
Green: I. Versicolor
Blue: I. Virginica

IID: Quality of the training data

The training data should be sampled from the same distribution as the testing data.

The sampling of the training data should select datapoints independently from each other.

Sepal Length x Width



Red: I. Setosa
Green: I. Versicolor
Blue: I. Virginica

Volume of training data

Let:

- ϵ be the maximum difference between the real accuracy and the accuracy measured over the training data
- δ be the probability that the error in our accuracy estimation is larger than ϵ
- m be the size of the training dataset

Then,

$$\delta \leq |H| e^{-2m\epsilon^2}$$

equivalently:

$$m \geq \frac{1}{2\epsilon^2} \left(\ln \frac{1}{\delta} + \ln |H| \right)$$

Volume of training data

Blumer et al. (1989) used the concept of VC dimension to get a tighter upper bound:

$$m \geq \frac{1}{\epsilon}\left(4\log\frac{1}{\delta} + 8\mathrm{VC}(H)\log\frac{13}{\epsilon}\right)$$

and Ehrenfeucht et al. (1989) to get a lower bound:

If:

$$m < \max\left[\frac{1}{\epsilon}\log\frac{1}{\delta}, \frac{\mathrm{VC}(C)-1}{32\epsilon}\right]$$

then with probability at least δ the error will be at least ϵ.

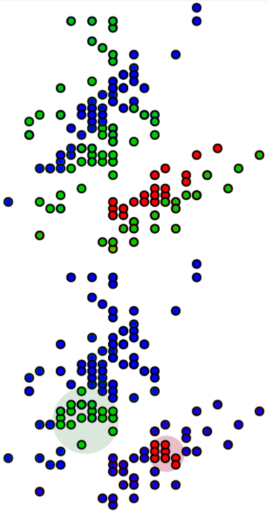
Different phenomena need different bias

There is no universally superior learning strategy, and somebody (or something) needs to pick the right tool for the job.

Different phenomena need different bias

There is no universally superior learning strategy, and somebody (or something) needs to pick the right tool for the job.

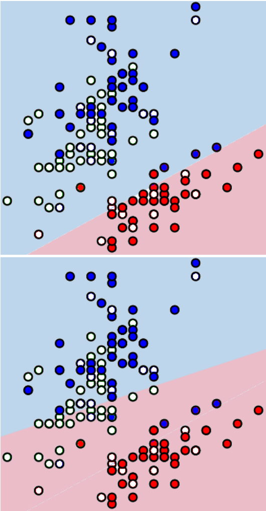
- *Syntactic bias:* Different types of models will get better results on different datasets



Different phenomena need different bias

> There is no universally superior learning strategy, and somebody (or something) needs to pick the right tool for the job.

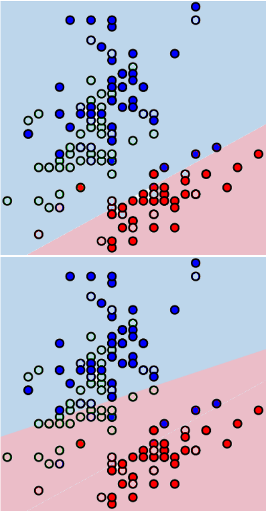
- *Syntactic bias:* Different types of models will get better results on different datasets
- *Semantic bias:* Different generalization strategies will get better results on different datasets



Different phenomena need different bias

> There is no universally superior learning strategy, and somebody (or something) needs to pick the right tool for the job.

- *Syntactic bias:* Different types of models will get better results on different datasets
- *Semantic bias:* Different generalization strategies will get better results on different datasets



Schaffer's (1994) result

A learning situation S is:

- A sample distribution D
- A labelling C
- A training set T

The *generalization performance* $\mathrm{GP}_L(S)$ of learner L on S is the expected accuracy above the baseline.

- E.g., for a binary classification task, $\mathrm{GP}_L(S) = \mathrm{Acc}_L(S) - 0.5$

For any learner L:

$$\sum_S \mathrm{GP}_L(S) = 0$$

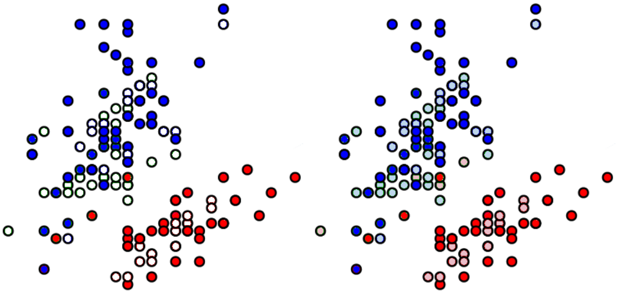
Schaffer's (1994) result

$$\begin{aligned}\textstyle\sum_S \mathrm{GP}(S) &= \textstyle\sum_C \mathrm{E}_T\left(\mathrm{E}_A\left(\mathrm{GP}(A,C)\right)\right)\\ &= \sum_{C\in D}\sum_{T_C}\left(\mathrm{P}(T_C|C)\sum_{A_i\in D\setminus T_C}\mathrm{P}(A_i)\mathrm{GP}(A_i,C)\right)\end{aligned}$$

Schaffer's (1994) result

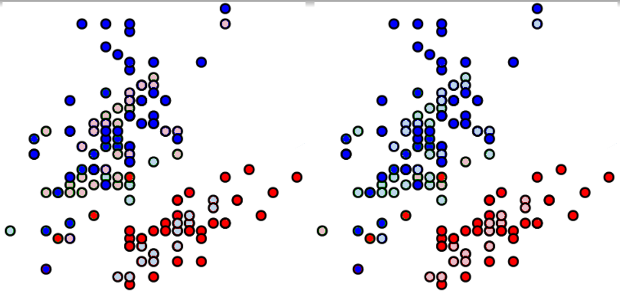
$$
\begin{aligned}
\textstyle\sum_S \mathrm{GP}(S) &= \textstyle\sum_C \mathrm{E}_T\left(\mathrm{E}_A\left(\mathrm{GP}(A,C)\right)\right) \\
&= \sum_{C\in D}\sum_{T_C}\left(\mathrm{P}(T_C|C)\sum_{A_i\in D\setminus T_C}\mathrm{P}(A_i)\mathrm{GP}(A_i,C)\right) \\
&= \sum_{T\in D}\sum_{C_T}\left(\mathrm{P}(T|C_T)\sum_{A_i\in D\setminus T}\mathrm{P}(A_i)\mathrm{GP}(A_i,C_T)\right)
\end{aligned}
$$

Intuitive explanation



- For any training set, there are many concepts that are consistent with the seen data and label the complete domain

Schaffer's (1994) result



- For any training set, there are many concepts that are consistent with the seen data and label the complete domain
- For every such concept C, there is a perfectly adversarial concept $\bar{C}$ that assigns the same labels on seen data and gives the *exact opposite* labelling on unseen data

Schaffer's (1994) result

$$
\begin{aligned}
\textstyle\sum_S \mathrm{GP}(S) &= \sum_{T\in D}\sum_{C_T}\left(\mathrm{P}(T|C_T)\sum_{A_i\in D\setminus T}\mathrm{P}(A_i)\mathrm{GP}(A_i,C_T)\right) \\
&= \tfrac{1}{2}\sum_{T\in D}\sum_{C_T}\left(\mathrm{P}(T|C_T)\sum_{A_i\in D\setminus T}\mathrm{P}(A_i)\mathrm{GP}(A_i,C_T)\right. \\
&+ \left.\mathrm{P}(T|\bar{C}_T)\sum_{A_i\in D\setminus T}\mathrm{P}(A_i)\mathrm{GP}(A_i,\bar{C}_T)\right)
\end{aligned}
$$

Schaffer's (1994) result

$$\begin{aligned}
\textstyle\sum_S \mathrm{GP}(S) &= \sum_{T\in D}\sum_{C_T}\left(\mathrm{P}(T|C_T)\sum_{A_i\in D\setminus T}\mathrm{P}(A_i)\mathrm{GP}(A_i, C_T)\right)\\
&= \tfrac{1}{2}\sum_{T\in D}\sum_{C_T}\Bigg(\mathrm{P}(T|C_T)\sum_{A_i\in D\setminus T}\mathrm{P}(A_i)\mathrm{GP}(A_i, C_T)\\
&+ \mathrm{P}(T|\bar{C}_T)\sum_{A_i\in D\setminus T}\mathrm{P}(A_i)\mathrm{GP}(A_i, \bar{C}_T)\Bigg)\\
&\quad \text{But } \mathrm{P}(T|C_T) = \mathrm{P}(T|\bar{C}_T) \text{ thus :}\\
&= \tfrac{1}{2}\sum_{T\in D}\sum_{C_T}\Big(\mathrm{P}(T|C_T)\sum_{A_i\in D\setminus T}\mathrm{P}(A_i)\big(\mathrm{GP}(A_i,C_T)+\mathrm{GP}(A_i,\bar{C}_T)\big)\Big)
\end{aligned}$$

Schaffer's (1994) result

$$
\begin{aligned}
\textstyle\sum_S \mathrm{GP}(S) &= \tfrac{1}{2} \sum_{T \in D} \sum_{C_T} \left(\mathrm{P}(T|C_T) \sum_{A_i \in D \setminus T} \mathrm{P}(A_i) \big(\mathrm{GP}(A_i, C_T) + \mathrm{GP}(A_i, \bar{C}_T) \big) \right) \\
&= \tfrac{1}{2} \sum_{T \in D} \sum_{C_T} \left(\mathrm{P}(T|C_T) \sum_{A_i \in D \setminus T} \mathrm{P}(A_i) \cdot (C_i - 0.5 + \bar{C}_i - 0.5) \right) \\
&= \frac{1}{2} \sum_{T \in D} \sum_{C_T} \left(\mathrm{P}(T|C_T) \sum_{A_i \in D \setminus T} \mathrm{P}(A_i) \cdot 0 \right) \\
&= 0
\end{aligned}
$$

Objections, workarounds

- Schaffer argues that human practitioners should understand the bias of each learner and intuite where to apply it.
- The proof weighs all possible concepts (i.e., tasks, problems) equally, although there might be a class of *real world* tasks where a learner performs better than zero. To reflect this, we need to define the *expected generalization performance*
    - *Real world* tasks weigh more, so that $C + \bar{C}$ do not zero out.
- EGP argues that human practitioners should understand whether a given task is a corner case or not.
- Meta-learning aims to remove the human practitioner, automating
    - deciding which learner to use
    - combining results from different learners

Recap

- PAC learning: a general framework for obtaining theoretical results about machine learning tasks.
- Evaluating on the seen data estimates the error on the unseen data. But there is a probability (maybe low, but non-zero) that our estimation will fail spectacularly (unbounded error).
- GP Conservation Law: there cannot be a learning machine that never falls in this pitfall: being good at one datatset means that the given machine *must* fail on another dataset.

Common sense, introspection, context

Discuss real-world examples demonstrating the theoretical results.

Draw inspiration from human intelligence: Common sense, context of operation, judging a theory based on a meta-theory besides evaluation on the data.

Independent and Identically Distributed (IID)

Statistics tells us about our sampling:

- *Identically distributed*: All items in the sample are taken from the same probability distribution.
- *Independent:* Selecting one example does not affect what the next example will be.

Evaluating how well-trained a ML model is, relies on IID:

- if the iid assumption does not hold, the PAC bounds for error do not hold
- and *even under this assumption* there is non-zero probability to get an error larger that what foreseen, but more examples tighten the bounds between this probability and the error.

Crime Hotspots

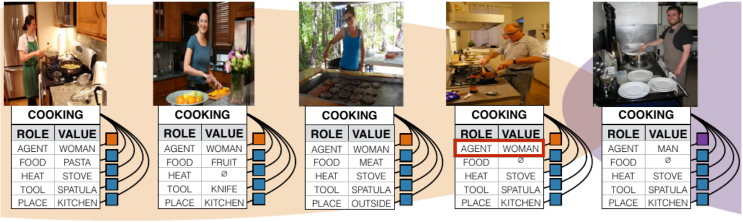
1. Collect statistics about geo-temporal distribution of crime scenes
2. Train a model to predict the next crime scene, and use this prediction to deploy more police in troublesome areas
3. More police leads to a *higher percentage* of offences committed to be prosecuted, reinforcing the model and creating a bias that establishes areas as troublesome.
    - Although not all offences weigh equally, but a large enough number of nuisance offences sustains the feedback loop, even if there is no difference in serious offences

Applicant Tracking Systems

1. Train a model to predict which applications might lead to a hire, and only interview those
2. If most of the interviewees are white males, most succesful appllicants will be white males
3. This reinforces white males' being good fits
   - Gender and ethnicity are not made available to the system, but can be accessed through proxy features such as hobbies or university attended
4. The system evaluates its theories on a training dataset drawn from a different distribution than that of overall applicants

Proxy features in semantic labelling

1. Annotate images with descriptions of activities
2. Train a model to extract descriptions of activities from images
3. Zhao et al. (2017): 33% of the cooking situations have a male agent in the training data, and is predicted in 16% of the cooking situations in the testing data
    - The model exploits correlations that boost its scores
    - The correlation that creates the bias is there in the data, not a sampling artefact



COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	PASTA
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	FRUIT
HEAT	∅
TOOL	KNIFE
PLACE	KITCHEN

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	MEAT
HEAT	STOVE
TOOL	SPATULA
PLACE	OUTSIDE

COOKING

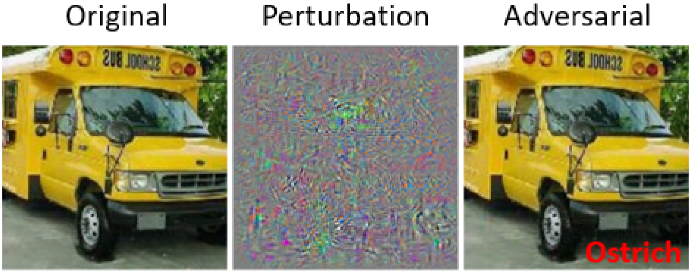
ROLE	VALUE
AGENT	WOMAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING

ROLE	VALUE
AGENT	MAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

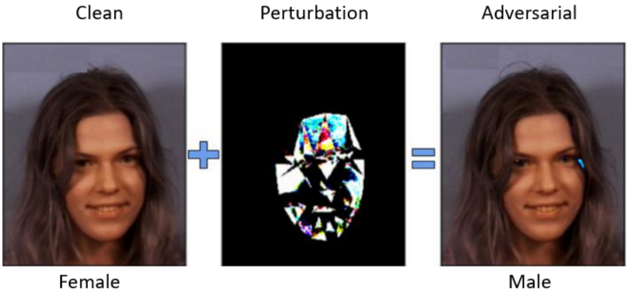
Exploiting proxy features

- Non-accidental *adversarial examples* (Akhtar & Mian, 2018)
- Targeting deep vision
- The designer has an understanding of deep vision, most often the interaction between features



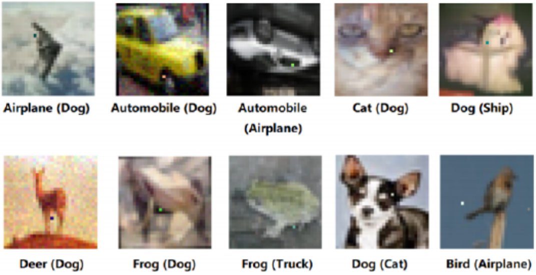
Exploiting proxy features

- Non-accidental *adversarial examples* (Akhtar & Mian, 2018)
- Targeting deep vision
- The designer has an understanding of deep vision, most often the interaction between features



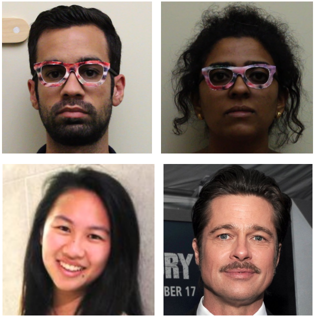
Exploiting proxy features

- Non-accidental *adversarial examples* (Akhtar & Mian, 2018)
- Targeting deep vision
- The designer has an understanding of deep vision, most often the interaction between features



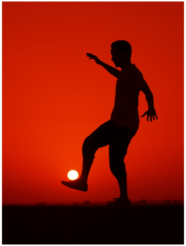
Exploiting proxy features

- Non-accidental *adversarial examples* (Akhtar & Mian, 2018)
- Targeting deep vision
- The designer has an understanding of deep vision, most often the interaction between features

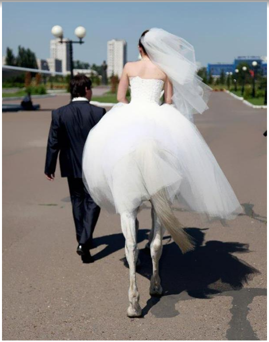


Exploiting our understanding of human vision

- Non-accidental corner cases
- Optical illusions, targeting time-constrained humans
- The designer has an understanding of human perception, most often depth perception
- Given enough time, humans will *reason* about the scene and interpret it correctly

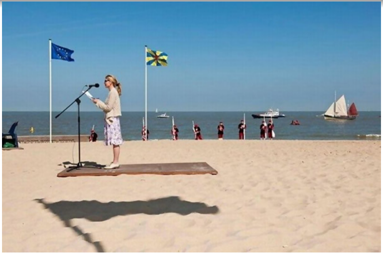


Depth and segmentation

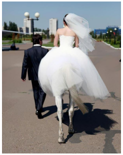


- How do you interpret this?
  - Centaur bride?

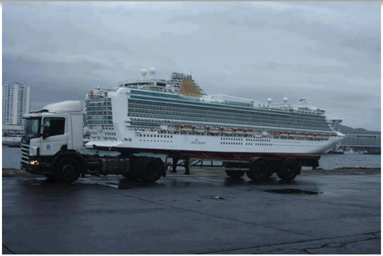
Depth and segmentation



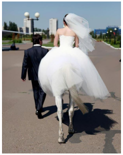
- How do you interpret this?
  - Centaur bride?
  - Flying carpet?

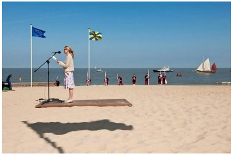


Depth and segmentation



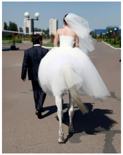
- How do you interpret this?
  - Centaur bride?
  - Flying carpet?
  - Truck carrying a cruise ship?

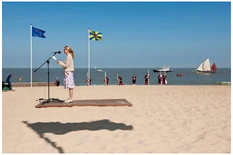


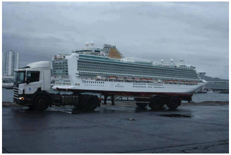


Depth and segmentation

- How do you interpret this?
    - Centaur bride?
    - Flying carpet?
    - Truck carrying a cruise ship?
- What understanding of the world did you use to reach your conclusion?

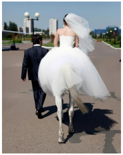


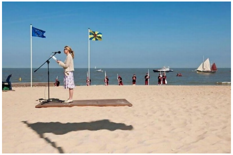


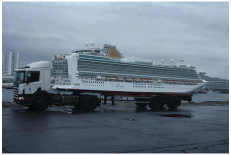


Depth and segmentation

- How do you interpret this?
    - Centaur bride?
    - Flying carpet?
    - Truck carrying a cruise ship?
- What understanding of the world did you use to reach your conclusion?
- Why would you expect a purely data-driven system to reach the same conclusion?







Common sense and introspection

The facts that comprise our ML model's interpretation (and their implications), must not contradict knowledge that we assume as given prior to the ML exercise.

In our example cases:

- Human and horse physiology, carpets do not levitate
- Cruise ships do not fit onto trucks

Common sense and introspection

The facts that comprise our ML model's interpretation (and their implications), must not contradict knowledge that we assume as given prior to the ML exercise.

In our example cases:

- Human and horse physiology, carpets do not levitate
- Cruise ships do not fit onto trucks
- Colour of eyewear does not determine identity
- Location does not determine gender

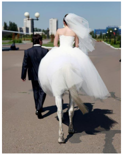
Common sense and introspection

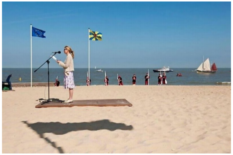
The facts that comprise our ML model's interpretation (and their implications), must not contradict knowledge that we assume as given prior to the ML exercise.

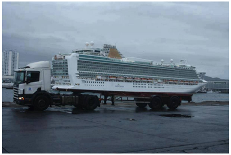
In our example cases:

- Human and horse physiology, carpets do not levitate
- Cruise ships do not fit onto trucks
- Colour of eyewear does not determine identity
- Location does not determine gender
- Colour is relevant, but discolourations do not affect typology and identity in the presense of stronger evidence

Depth and segmentation







Machine-readable general knowledge

Knowledge representation is an AI topic in its own right:

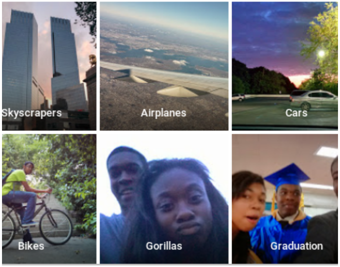
- Logic-based representations, such as concept ontologies
- Distributed representations, such as word embeddings
- Processes, such as physics simulations and digital twins
- Contexts that restrict the applicability of different pieces of knowledge

(In)famous cases

- *Microsoft apologizes after twitter chat bot experiment goes awry*
  - *Tay* chatbot extracted and re-used statements from previous interactions
  - Within ours, twitting racist and sexist content
- Sampling the Web vs. expecting the sample to present itself
  - A known issue with polling

(In)famous cases

- *Google apologizes for tagging photos of black people as 'gorillas'*
- We cannot know technical details, but most probably not a bug:
  - the model gets good accuracy on training data
  - the model cannot grasp the implications and sensitivities behind this particular misclassification
  - Fix: Google Photos *never* assigns the 'Gorilla' tag



(In)famous cases

- *Hero taxi driver saved 13-year-old from paedophile*
  - Many times clients have requested unexpected destinations
  - Understood potential implications and decided to *err on the side of safety*
  - Exhibited an undrestanding of how the world works and decisional autonomy, recording communications and collecting evidence

(In)famous cases

- *Self-driving cars suck at making unprotected left turns*
  - should be able to handle judging the risk [..] scan for pedestrians, the timing of oncoming traffic, and more.
  - And yet, the machines still struggle with the maneuver.
- A very human, social move. It's almost a negotiation: Drivers edge out into the lane, trying to assert themselves through.

(In)famous cases

Cognition and Artificial Intelligence

Generalization Performance Conservation Law: We need different learners (prior bias, background) for different problems

- There are contexts where centaurs and flying carpets should be recognized as such

Introspection: Evaluate data-driven generalizations

- No matter what the statistical evidence suggests, features extracted from eyewear are not reliable for face identification; and location is not reliable for gender prediction

What is so hard to explain?

- In most modern ML architectures, it is hard to track what contributes to the final output
- It is even harder to visualize or explain in language what the machine does in the abstract, not on specific examples
- Faithfulness vs. explainability: The closer the explanation gets to the full model, the harder it gets for humans to grasp

Explaining conventional AI

Explaining a decision:

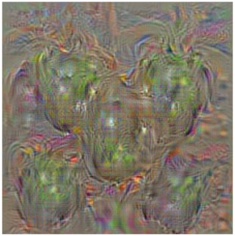
- Logic-based systems: premises, decision path, resolution chain
- Conventional classifiers: nearest neighbours, separation curve, support vector

Controller's reaction:

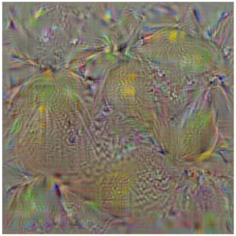
- Base facts can be corrected, but
- Rules and models are hard to understand and even harder to fix without side effects

Explaining deep learning systems

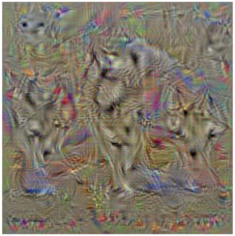
- Representing the actual model
    - Compute the image that scores highest for a concept



bell pepper



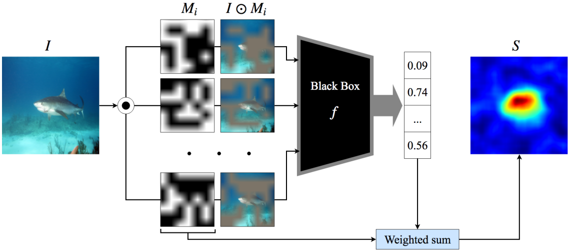
lemon



husky

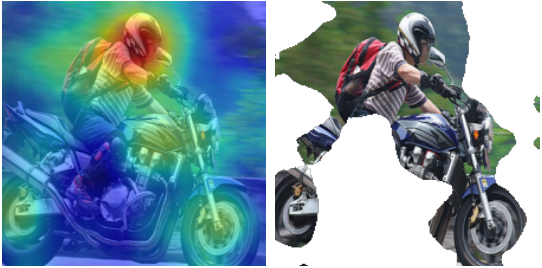
Explaining deep learning systems

- Representing the actual model
    - Compute the image that scores highest for a concept
- *Black box* approach, such as E.Petsiuk et al. (2018) RISE methodology for image tagging:
    - Randomly mask parts of the image
    - Measure network confidence for a given concept
    - Sum into a heat map of what parts of the image contribute to the concept



Explaining deep learning systems

- Representing the actual model
  - Compute the image that scores highest for a concept
- *Black box* approach, such as E.Petsiuk et al. (2018) RISE methodology for image tagging:
  - Randomly mask parts of the image
  - Measure network confidence for a given concept
  - Sum into a heat map of what parts of the image contribute to the concept



Crash helmet, from http://cs-people.bu.edu/vpetsiuk/rise/demo.html

Controlling AI

Setting the framework within which the AI is to operate and search for solutions

- Selvaraju et al. (2016) observed nurses vs. doctors misclassifications you already expect by now
- Used deep vision explanations to discover model focuses on face and hairstyle, thus learned a gender stereotype
- Only solution they could offer was balancing the training set

Recap

- The assumptions underpinning the mathematical constructions of generalization and machine learning are not a theoretical construct
  - ignoring them has actual, concrete consequences
  - in real-world deployments, not in academic exercises that test extreme conditions
- Our current machine learning setting can only operate when controlled by a higher-order cognitive layer
  - Current research topics: long-term AI, cognitive robotics
  - Not automated in productive systems, control is provided by human operators

References and attributions

Fisher (1936): Ronald Fisher, The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7(2), 1936. doi:10.1111/j.1469-1809.1936.tb02137.x
Leboeuf et al. (2020): Jean-Samuel Leboeuf, Frédéric LeBlanc, and Mario Marchand, Decision trees as partitioning machines to characterize their generalization properties, Oct 2020. arXiv:2010.07374 [cs.LG]
Valiant (1984): Leslie Valiant, A theory of the learnable. *Communications of the ACM* 27(11), 1984. doi:10.1145/1968.1972
Blumer et al. (1989): Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth, Learnability and the Vapnik-Chervonenkis Dimension. *Journal of the ACM* 36(4), 1989. doi:10.1145/76359.76371
Ehrenfeucht et al. (1989): Andrzej Ehrenfeucht, David Haussler, Michael Kearns, and Leslie Valiant, A general lower bound on the number of examples needed for learning. *Information and Computation* 82(3), 1989.
Schaffer (1994): Cullen Schaffer, A conservation law for generalization performance. *Proceedings 11th ICML, New Brunswick, NJ, July 1994.* doi:10.1016/B978-1-55860-335-6.50039-8
Selvaraju et al (2016): R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, Grad-CAM: Why did you say that? Visual explanations from deep networks via Gradient-based Localization. CoRR, abs/1610.02391, 2016.
Zhao et al. (2017): Jieyu Zhao, Tianlu Wang, et al., Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. *Proceedings of the 2017 Conference on Empirical Methods in NLP.* arxiv:1707.09457
Akhtar and Mian (2018): Naveed Akhtar and Ajmal Mian, *Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey.* arxiv:1801.00553
Petsiuk et al. (2018): Vitali Petsiuk, Abir Das, and Kate Saenko, RISE: Randomized Input Sampling for Explanation of Black-box Models. *Proceedings of the British Machine Vision Conference (BMVC 2018).* arxiv:1806.07421
Sharif et al. (2019): Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter, *General Framework for Adversarial Examples with Objectives.* arxiv:1801.00349

References and attributions

Iris Setosa, image from `https://commons.wikimedia.org/w/index.php?curid=170298`
Iris Versicolor, image from `https://commons.wikimedia.org/w/index.php?curid=248095`
Iris Virginica, image from `https://commons.wikimedia.org/w/index.php?curid=9805580`
Read the Web project at Carnegie Mellon: `http://rtw.ml.cmu.edu`
Bloomberg, Microsoft apologizes after twitter chat bot experiment goes awry
Huffington Post, Google apologizes for tagging photos of black people as 'gorillas'
Metro, Hero taxi driver saved 13-year-old from paedophile who planned to kidnap her
Mashable, This is why self-driving cars suck at making unprotected left turns