Învățare Automată (Machine Learning)

Bogdan Alexe,

bogdan.alexe@fmi.unibuc.ro

Master Informatică, anul I, 2018-2019, cursul 1

Cui se adresează acest curs opțional?

- cursul este introdus în acest an ca un curs opțional cu intenția de a completa planul de învățământ al masterului de Inteligență Artificială actual la cel acreditat deja (în limba engleză, se face din 2019-2020);

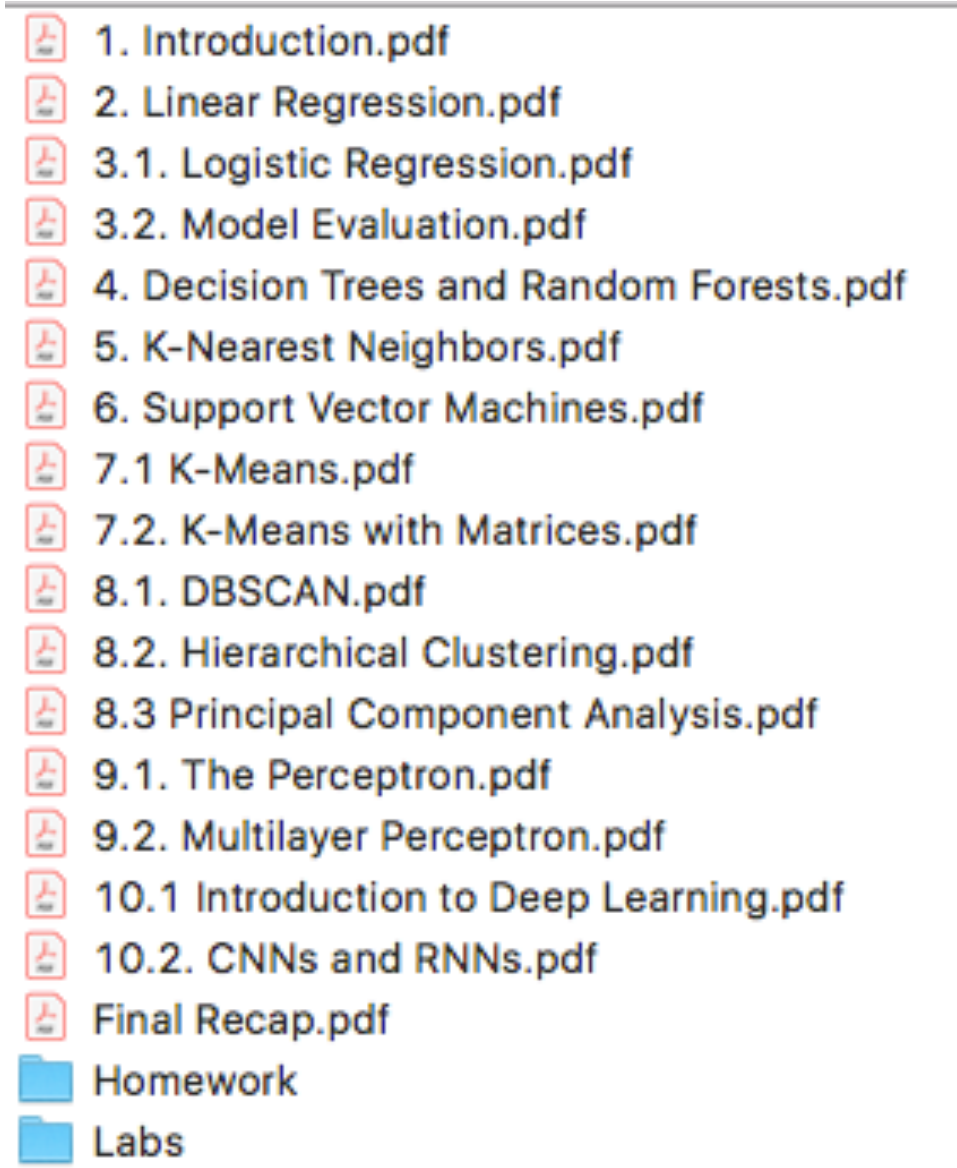
- cursul se adresează cu predilecție studenților de la grupa 407, masterul de Inteligență Artificială (ei au urmat cursul de Data Mining = Învățare Automată cu Python = Practical Machine Learning în semestrul 1);

- cursul se adresează tuturor studenților care vor să înțeleagă fundamentul teoretic care stă la baza modelelor de învățare automată;

Practical Machine Learning (sem 1) …

- topicurile abordate în Învățare Automată cu Python din semestrul 1
- diverși algoritmi utilizați în învățarea automată
- materiale aici (Sparktech)

http://goo.gl/1iG3v9



Practical Machine Learning (sem 1) …

- curs + laborator orientat către scriere de aplicații în Python/ Jupyter Notebook (clasificare de animale/text, recunoaștere de cifre cu rețele neuronale convoluționale)

In this homework, you are supposed to build a neural model that can read numbers and adds them together. Numbers are formed using images from the MNIST [5] dataset which are concatenated after some random translations / rotations are applied to them. The images you will work with will contain sequences ranging from 1 to 3 digits. The maximum number you can have in an image is 255. Figure 1 provides an example of a data sample.

Training data will be generated using the *training_generator()* function from *data_generator.py* file. You may **not** modify this file. For testing your models, you will use the *test_generator()* function. Each of these functions returns a pair of images of fixed shape (28x84 px), their labels, and the result of the addition.

(a) Label: 14 (b) Label: 112

Figure 1: Adding the two numbers: 14 + 112 = 126

… vs. Theoretical Machine Learning (sem2)

- Curs + seminar orientate către înțelegerea teoriei ce stă în spatele algoritmilor de învățare automată

Exemplu de slide din curs

Theorem 5.1. (No-Free-Lunch) *Let A be any learning algorithm for the task of binary classification with respect to the $0-1$ loss over a domain $\mathcal{X}$. Let m be any number smaller than $|\mathcal{X}|/2$, representing a training set size. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:*

1. *There exists a function $f : \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$.*
2. *With probability of at least $1/7$ over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.*

… vs. Theoretical Machine Learning (sem2)

- Curs + seminar orientate către înțelegerea teoriei ce stă în spatele algoritmilor de învățare automată

Exemplu de exercițiu de seminar

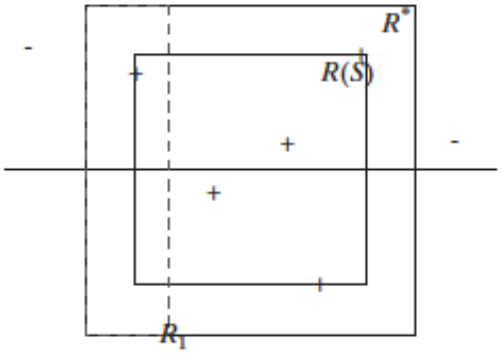


Figure 2.2. Axis aligned rectangles.

Formally, given real numbers $a_1 \leq b_1, a_2 \leq b_2$, define the classifier $h_{(a_1,b_1,a_2,b_2)}$ by

$$h_{(a_1,b_1,a_2,b_2)}(x_1,x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 \leq b_1 \text{ and } a_2 \leq x_2 \leq b_2 \\ 0 & \text{otherwise} \end{cases}. \qquad (2.10)$$

The class of all axis aligned rectangles in the plane is defined as

$$\mathcal{H}^2_{\text{rec}} = \{h_{(a_1,b_1,a_2,b_2)} : a_1 \leq b_1, \text{ and } a_2 \leq b_2\}.$$

Note that this is an infinite size hypothesis class. Throughout this exercise we rely on the realizability assumption.

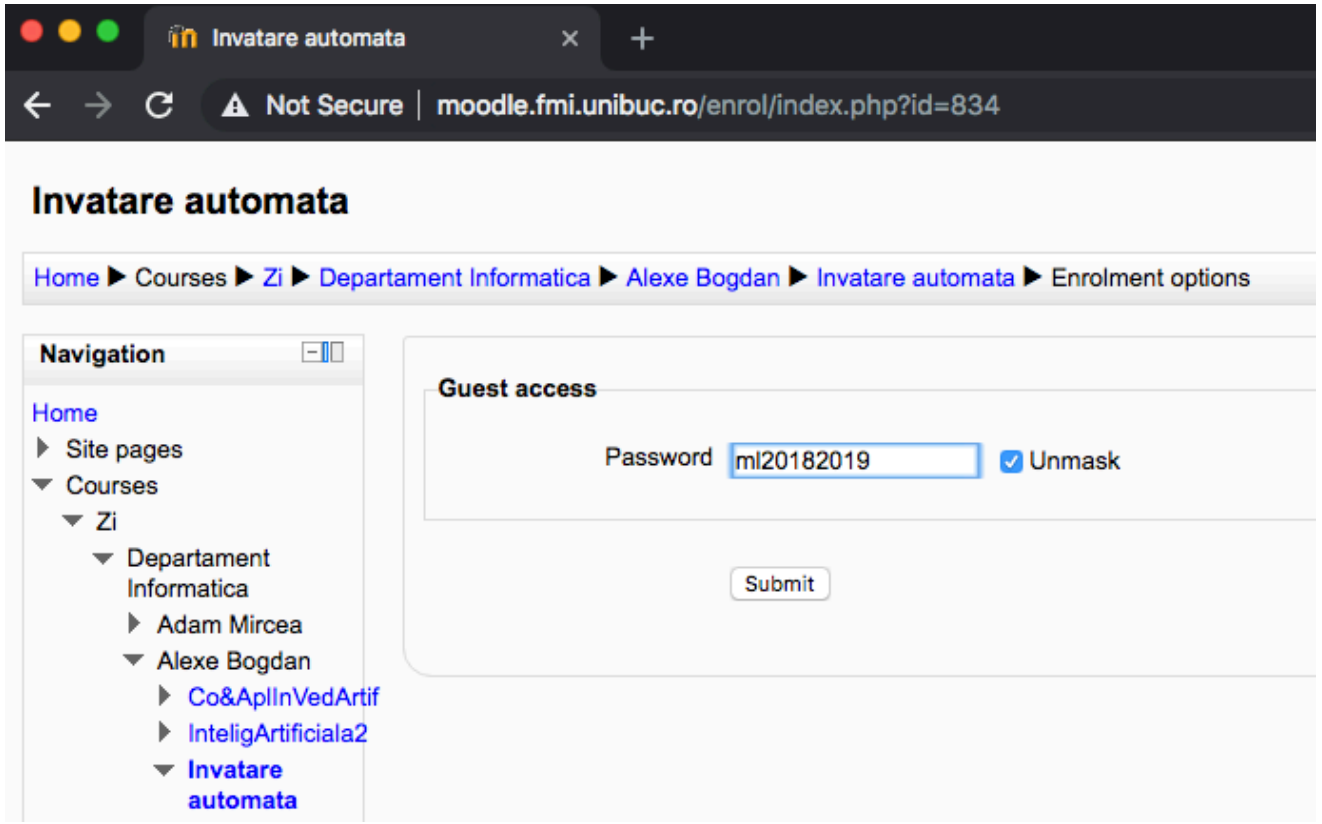
Administrative

Administrative

- slides in English
- course + seminar schedule: Friday 8 -10 am (weekly) + 10-12 am (once every two weeks with small groups, today is off).
  - maybe add another seminar class?
- grading:
  - 2 assignments (written): week 5 and 11, 3.5 + 3.5 points (+bonus?)
  - final exam: 3 points (open books?)
  - final mark: $\geq$ x.5 -> x+1, < x.5 -> x

Course Materials

- Moodle
  - http://moodle.fmi.unibuc.ro/course/view.php?id=834
  - password: ml20182019



Course Materials

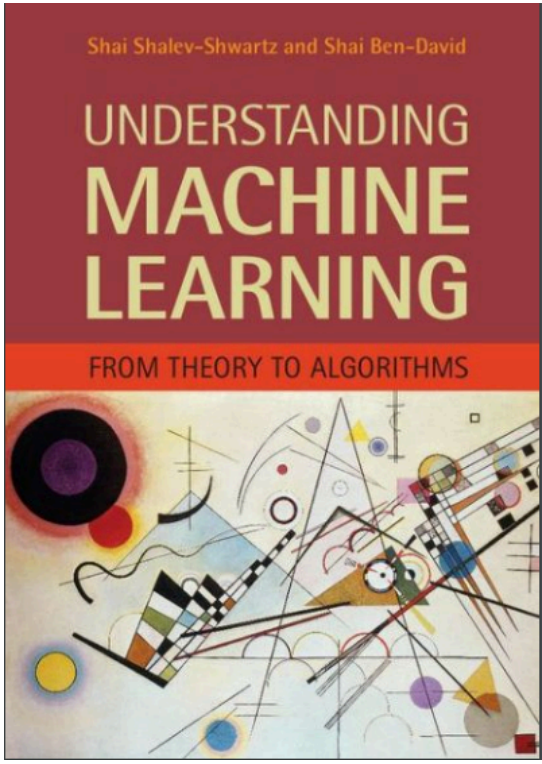
- Moodle
  - http://moodle.fmi.unibuc.ro/course/view.php?id=834
  - password: ml20182019

- Course Book
  - ***Understanding ML from theory to algorithms***

  Shai Shalev-Shwartz, Shai Ben-David,

  Cambridge University Press, 2014
  - **available online**

https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf



Course Materials

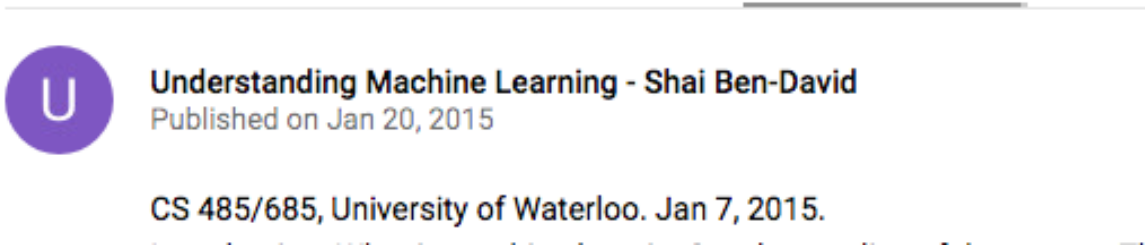
- Moodle
  - http://moodle.fmi.unibuc.ro/course/view.php?id=834
  - password: ml20182019

- Course Book
  - ***Understanding ML from theory to algorithms***

  Shai Shalev-Shwartz, Shai Ben-David,

- Online Lectures
  - Shai Ben-David

  Youtube videos
  - Shai Shalev-Shwartz: http://www.cs.huji.ac.il/~shais/IML2014.html



What is Learning?

What is Learning?

Using Experience to gain Expertise

“Learning” (in nature): using past experience to make future decisions or guide future actions

Applied Animal Behaviour Science
Volume 24, Issue 2, September 1989, Pages 89-99

"Poison-shyness" and "bait-shyness" developed by wild rats (*Rattus rattus* L.). I. Methods for eliminating "shyness" caused by barium carbonate poisoning

Ghazala Naheed, Jamil Ahmad Khan

⊞ Show more

https://doi.org/10.1016/0168-1591(89)90037-3

Abstract

Colonies of wild rats, *Rattus rattus* L., were offered the choice between two baits — cereal grains, flours, mixtures, oily and sweet cereals, and also gram flour. The rats were poisoned in the preferred baits with barium carbonate (100 mg per 10 g food; 200 mg per 10 g food in oily baits) and then presented with the same choice of unpoisoned foods as before.

Poisoning caused a change in the feeding patterns of rats. Foods mixed with barium carbonate were avoided ("poison-shyness"), the same foods then offered without poison were also rejected ("bait-shyness"). Intermittent poisoning also caused aversion to the eating of both poison and bait. Apparently, both the quality and the strength of tastes perceived in the poisonous mixtures influenced the development of "bait-shy" behaviour in the rats.

Bait shyness – Rats Learning to avoid Poisonous Baits

- learning mechanism for rats: they use past experience with some food to acquire expertise in detecting the safety of the food

- a successful learner should be able to progress from individual examples to broader *generalization* . This is also referred to as ***inductive reasoning*** or ***inductive inference***

- rats apply their attitude on new, unseen examples of food of similar smell and taste

Pigeon superstition

https://www.youtube.com/watch?v=TtfQlkGwE2U

"Superstition" in the pigeon.

Database: PsycARTICLES Journal Article

Skinner, B. F.

Citation

Skinner, B. F. (1992). "Superstition" in the pigeon. *Journal of Experimental Psychology: General, 121*(3), 273-274. http://dx.doi.org/10.1037/0096-3445.121.3.273

Abstract

(This reprinted article originally appeared in the *Journal of Experimental Psychology,* 1948, Vol 38, 168–272. The following abstract of the original article appeared in PA, Vol 22:4299.) A pigeon is brought to a stable state of hunger by reducing it to 75% of its weight when well fed. It is put into an experimental cage for a few minutes each day. A food hopper attached to the cage may be swung into place so that the pigeon can eat from it. A solenoid and a timing relay hold the hopper in place for 5 sec at each reinforcement. If a clock is now arranged to present the food hopper at regular intervals with no reference whatsoever to the bird's behavior, operant conditioning usually takes place. The bird tends to learn whatever response it is making when the hopper appears. The response may be extinguished and reconditioned. The experiment might be said to demonstrate a sort of superstition. The bird behaves as if there were a causal relation between its behavior and the presentation of food, although such a relation is lacking. (PsycINFO Database Record (c) 2016 APA, all rights reserved)

Pigeon superstition

Pigeon Superstition: In an experiment performed by the psychologist B. F. Skinner, he placed a bunch of hungry pigeons in a cage. An automatic mechanism had been attached to the cage, delivering food to the pigeons at regular intervals with no reference whatsoever to the birds' behavior. The hungry pigeons went around the cage, and when food was first delivered, it found each pigeon engaged in some activity (pecking, turning the head, etc.). The arrival of food reinforced each bird's specific action, and consequently, each bird tended to spend some more time doing that very same action. That, in turn, increased the chance that the next random food delivery would find each bird engaged in that activity again. What results is a chain of events that reinforces the pigeons' association of the delivery of the food with whatever chance actions they had been performing when it was first delivered. They subsequently continue to perform these same actions diligently.[1]

Bait shyness revisited

Relation of cue to consequence in avoidance learning.

EXPORT ★ Add To My List

Database: PsycINFO Journal Article

Garcia, John Koelling, Robert A.

Citation

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science, 4*(3), 123-124.
http://dx.doi.org/10.3758/BF03342209

Abstract

An audiovisual stimulus was made contingent upon the rat's licking at the water spout, thus making it analogous with a gustatory stimulus. When the audiovisual stimulus and the gustatory stimulus were paired with electric shock the avoidance reactions transferred to the audiovisual stimulus, but not the gustatory stimulus. Conversely, when both stimuli were paired with toxin or X-ray the avoidance reactions transferred to the gustatory stimulus, but not the audiovisual stimulus. Apparently stimuli are selected as cues dependent upon the nature of the subsequent reinforcer.

Bait shyness revisited

- repeated trials in which the consumption of some food is followed by the administration of unpleasant electrical shock the rats do not tend to avoid that food

- similar failure of conditioning occurs when the characteristic of the food that implies nausea (taste, smell) is replaced by a vocal sign

https://psychology110hc.wordpress.com/2015/04/16/relation-of-cue-to-consequence-in-avoidance-learning/

Prior knowledge

- one distinguishing feature between the bait shyness learning and the pigeon superstition is the incorporation of prior knowledge that biases the learning mechanism = inductive bias.

- the pigeons in the experiment are willing to adopt any explanation for the occurrence of food.

- the rats "know" that food cannot cause an electric shock and that the co-occurrence of noise with some food is not likely to affect the nutritional value of that food. The rats' learning process is biased toward detecting some kind of patterns while ignoring other temporal correlations between events.

Prior knowledge

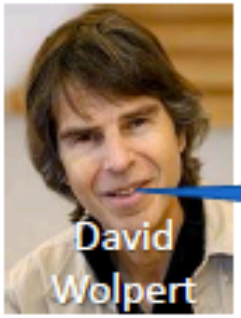
- the incorporation of prior knowledge, biasing the learning process, is inevitable for the success of learning algorithms - “No-Free-Lunch theorem”

- the stronger the prior knowledge (or prior assumptions) that one starts the learning process with, the easier it is to learn from further examples.

- the stronger these prior assumptions are, the less flexible the learning is

The ability to learn grammars is **hard-wired** into the brain. It is not possible to "learn" linguistic ability—rather, we are born with a brain apparatus specific to language representation.

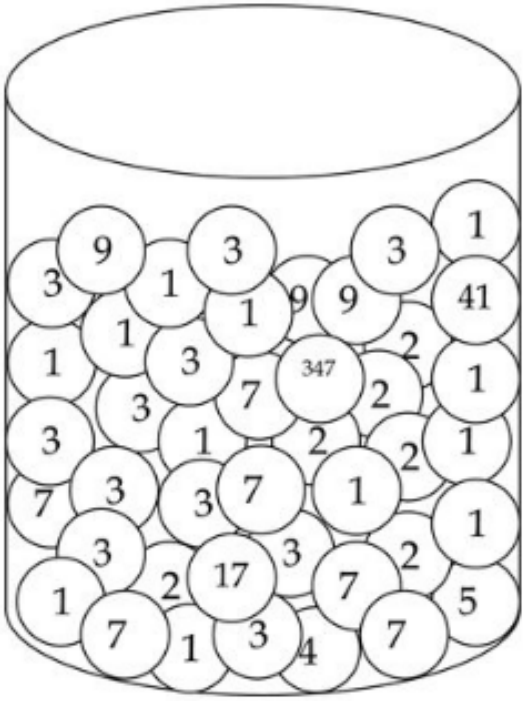
There exists some "universal" learning algorithm that can learn ***anything***: language, vision, speech, etc. The brain is based on it, and we're working on uncovering it. (Hint: the brain uses neural networks)



There is no "free lunch": no learning is possible without *some* prior assumption about the structure of the problem (prior knowledge)

Mathematical Analysis of Learning

Induction in an urn (Valiant, 1984)



- consider an urn containing a very large number (millions) of marbles, possibly of different types. You are allowed to draw 100 marbles and asked what kind of marbles the urn contains.

L. G. Valiant, *A theory of the Learnable*, Communications ACM, 27(11):1134-1142, 1984
L. G. Valiant, *Probably Approximately Correct. Nature's Algorithms for Learning and Prospering in a Complex World*, Basic Books, 2013

Induction in an urn (Valiant, 1984)

- consider an urn containing a very large number (millions) of marbles, possibly of different types. You are allowed to draw 100 marbles and asked what kind of marbles the urn contains.

- no assumptions
  - impossible task!

- assumption 1: all the marbles are of different types
  - impossible task!

- assumption 2: all the marbles are identical
  - one single draw is sufficient to solve the task!

Induction in an urn (Valiant, 1984)

- assumption 3: 50% of all marbles are of one type
  - probability to miss that type is $(1/2)^{100} = 7.8 * 10^{-31}$
  - predict that after 100 draws you will have seen representatives for 50% of the urn content

- assumption 4: there are at most 5 different marble types
  - **don't know** the distribution of the marbles
  - could be any distribution: (20%,20%,20%,20%,20%) ,(92%,2%,2%,2%,2%), (49.85%,49.85%,0.1%,0.1%,0.1%), etc
  - predict with 97% confidence that after 100 draws you will have seen representatives for more than 80% of the urn content
  - reasoning: (A) if any of the 5 types occurs with frequency > 5%, the probability to miss that type is $< (1\text{-}0.05)^{100} = 0.6\%$. The probability to miss one type is < 5*0.6% = 3%; (not A) There exists types that occurs with frequency < 5%. There can be at most 4 types with frequency < 5% so the rare marble types are < 20%. Probability to miss the common marble types (which account > 80% of the urn) is < 3%.

PAC learning (Valiant, 1984)

- induction with minimal assumption is very powerful, achieve a useful level of generalization knowing that there a fixed small number of marbles in the urn

- two sources of errors:

  (1) rarity: rare types of marbles are unlikely to be drawn in any small samples

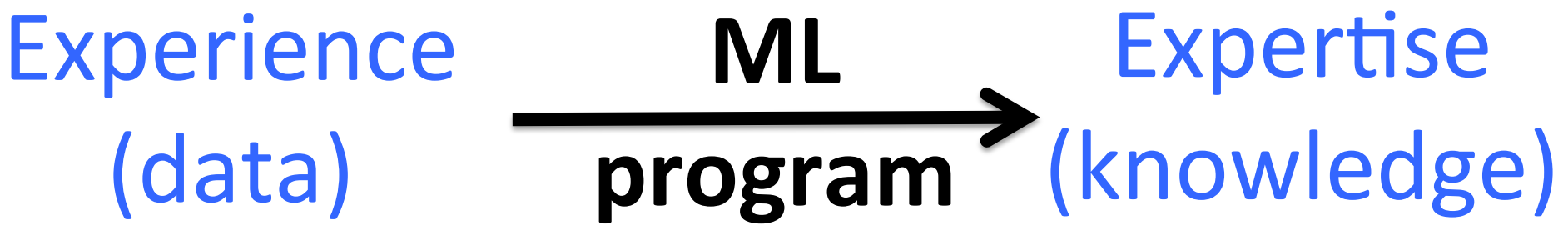
  (2) misfortune: with small probability the sample drawn will be unrepresentative of the contents of the urn because it missed some common marble types

  – neither of these two sources of errors can be totally eliminated BUT we can controlled them by increasing the number of marbles drawn.

- PAC learning: probably approximately correct learning
  – "probably" − misfortune errors
  – "approximately" − rarity errors

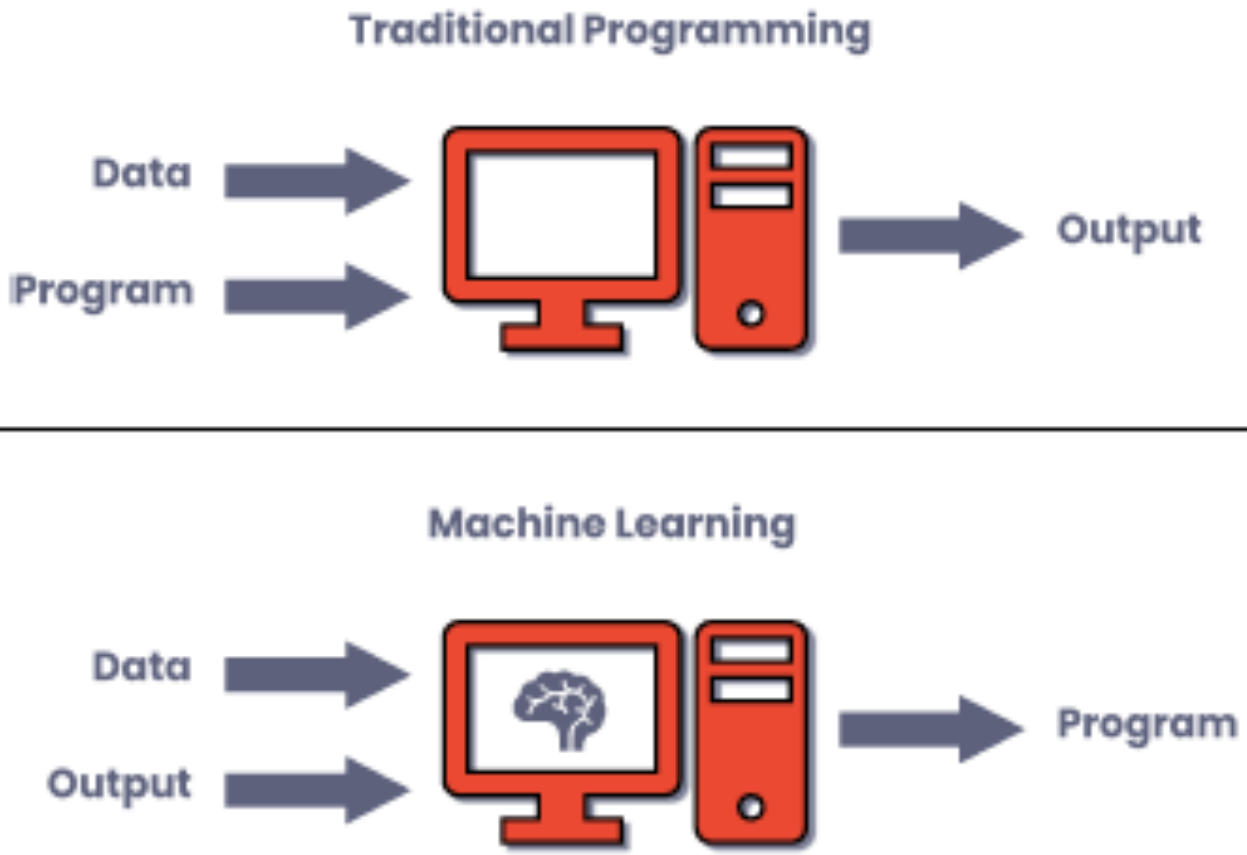
What is Machine Learning?

What is Machine Learning?



"Machine Learning" as an Engineering Paradigm: Use data and examples, instead of expert knowledge, to automatically create systems that perform complex tasks

What is Machine Learning?



Slide Credit: Sparktech

Machine Learning in Computer Vision

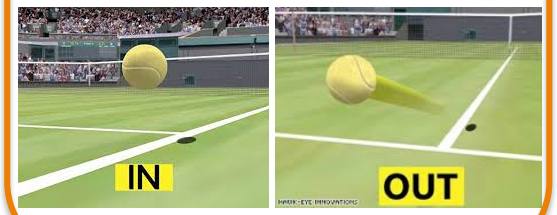
OCR

Xbox Kinect

Visual search

Face detection / recognition

Hawk-eye decision system



Driving assistance systems

Machine Learning Everywhere

Spam filtering

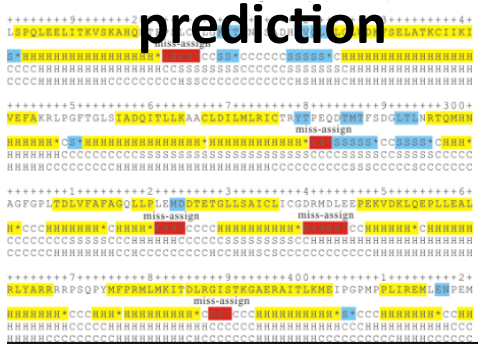
Machine translation

Speech recognition

Advertising and ad placement

Recommendation systems

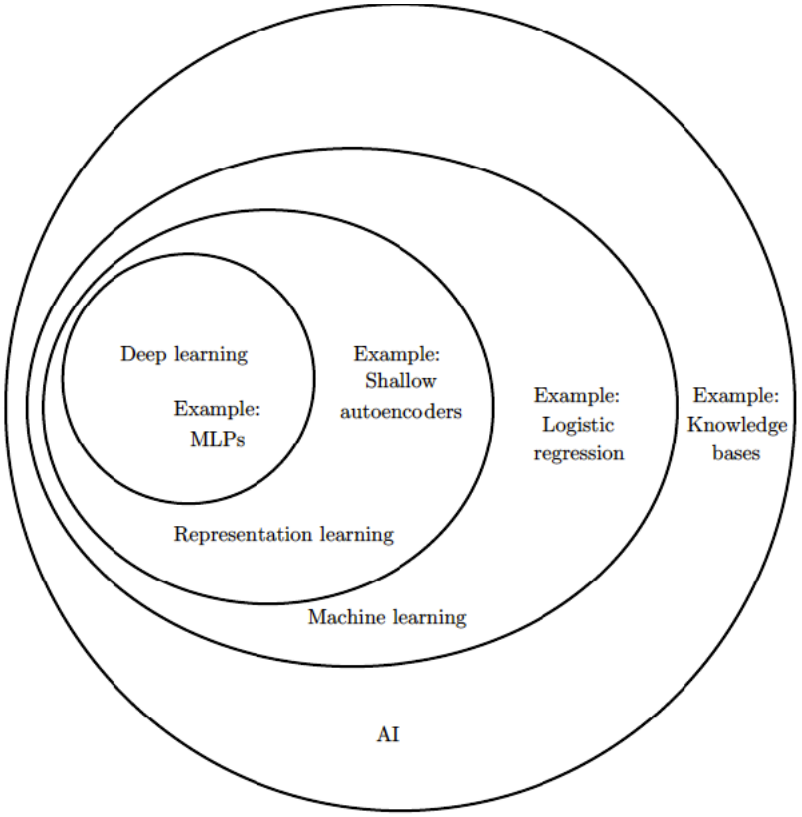
Protein fold prediction



Why do we need machine learning?

- Tasks that are too complex to program!
- Computer vision: we know to detect objects but have no idea how we do it!
- Search engines: a human can't read the entire internet!
- Adaptivity and speed of development

Machine Learning in AI



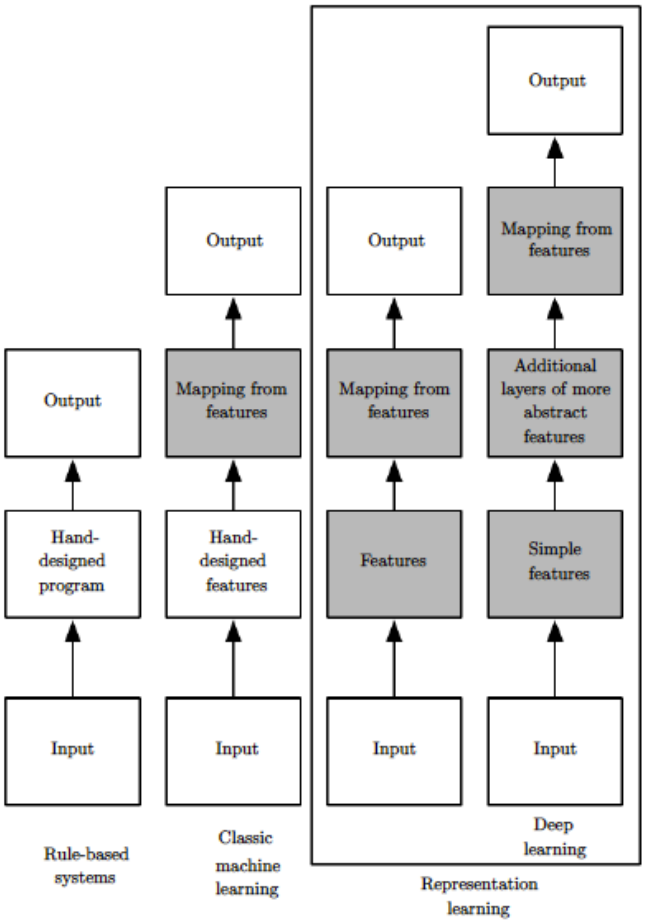


Figure 1.5: Flowcharts showing how the different parts of an AI system relate to each other within different AI disciplines. Shaded boxes indicate components that are able to learn from data.

I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, MIT, 2016, http://www.deeplearningbook.org

Machine Learning vs. Statistics

Machine learning is essentially a form of applied statistics with increased emphasis on the use of computers to statistically estimate complicated functions and a decreased emphasis on proving confidence intervals around these functions

Statistics	Computer Science	Meaning
estimation	learning	using data to estimate an unknown quantity
classification	supervised learning	predicting a discrete Y from $X \in \mathcal{X}$
clustering	unsupervised learning	putting data into groups
data	training sample	$(X_1, Y_1), \ldots, (X_n, Y_n)$
covariates	features	the X_i's
classifier	hypothesis	a map from covariates to outcomes
hypothesis	—	subset of a parameter space Θ
confidence interval	—	interval that contains unknown quantity with a prescribed frequency
directed acyclic graph	Bayes net	multivariate distribution with specified conditional independence relations
Bayesian inference	Bayesian inference	statistical methods for using data to update subjective beliefs
frequentist inference	—	statistical methods for producing point estimates and confidence intervals with guarantees on frequency behavior
large deviation bounds	PAC learning	uniform bounds on probability of errors

Course goals

- First goal: to provide a rigorous, yet easy to follow, introduction to the main concepts underlying machine learning:
  - *What is learning?*
  - *How can a machine learn?*
  - *How do we quantify the resources needed to learn a given concept?*
  - *Is learning always possible?*
  - *Can we know if the learning process succeeded or failed?*

- Second goal: present several key machine learning algorithms with strong theoretical foundations

Course Structure – Part 1

- What is learning?
  - Probably Approximately Correct (PAC) model - Vaillant 1984
- How can a machine learn?
  - Empirical Risk Minimization (ERM)
  - Structural Risk Minimization (SRM)
  - Minimum Description Length (MDL)
- Resources needed to learn a given concept?
  - sample complexity, time complexity
- Is learning always possible? Did the learning process succeeded?
  - "no free lunch" theorem

Course Structure – Part 2

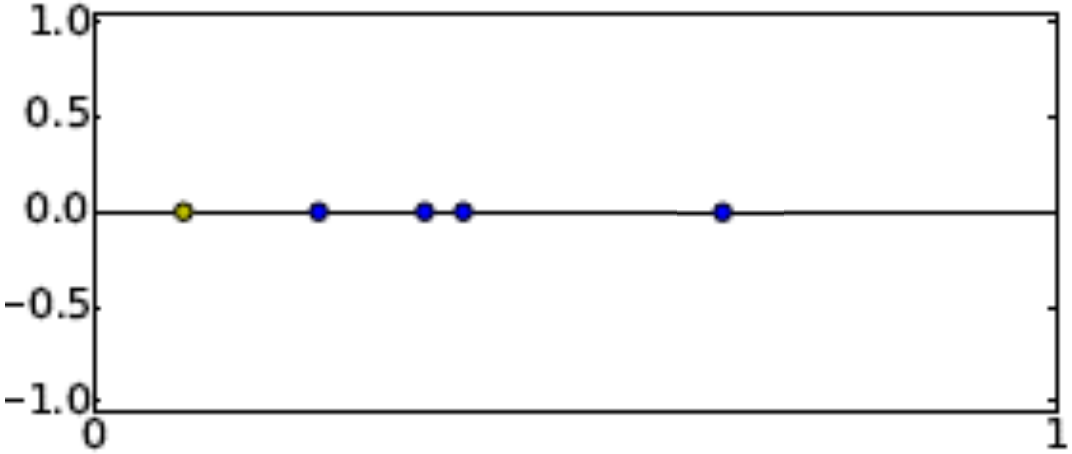
- Survey of prominent methods and approaches with strong theoretical foundations such as:
  - Boosting
  - SVMs
  - neural networks? (loose bounds, work in progress)
  - etc

Usefulness of Theoretical Machine Learning

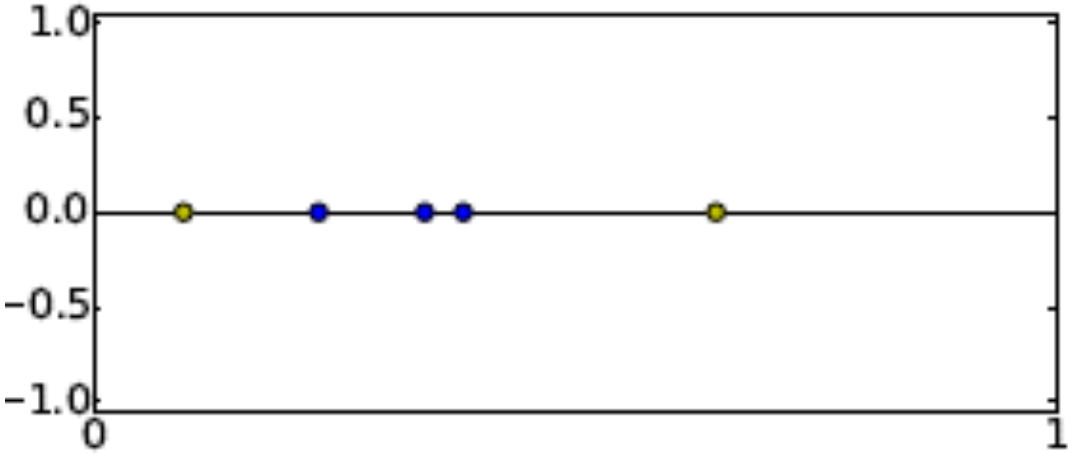
Perceptron in 1D

Class +1

Class -1



Linear separable

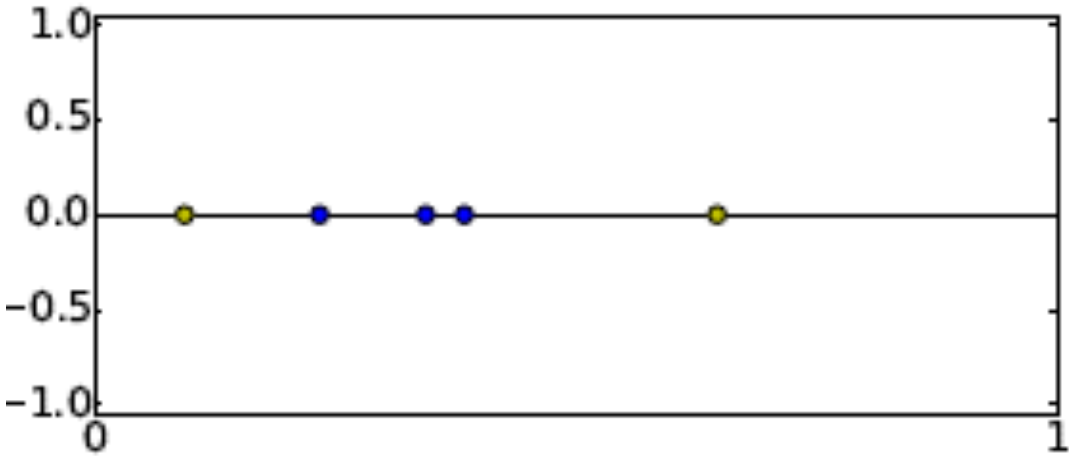


Non-linear separable

Learning sin(λx) in 1D

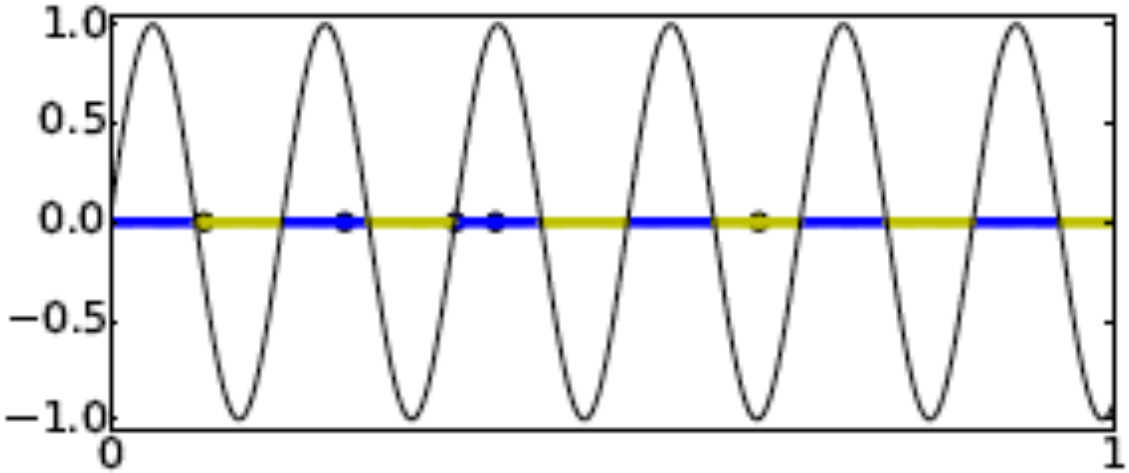
Class +1

Class -1



Non-linear separable

Can you think of an algorithm learning λ for solving the problem?
Will it generalize (small generalization error)?



Theoretical result: cannot learn λ with small generalization error

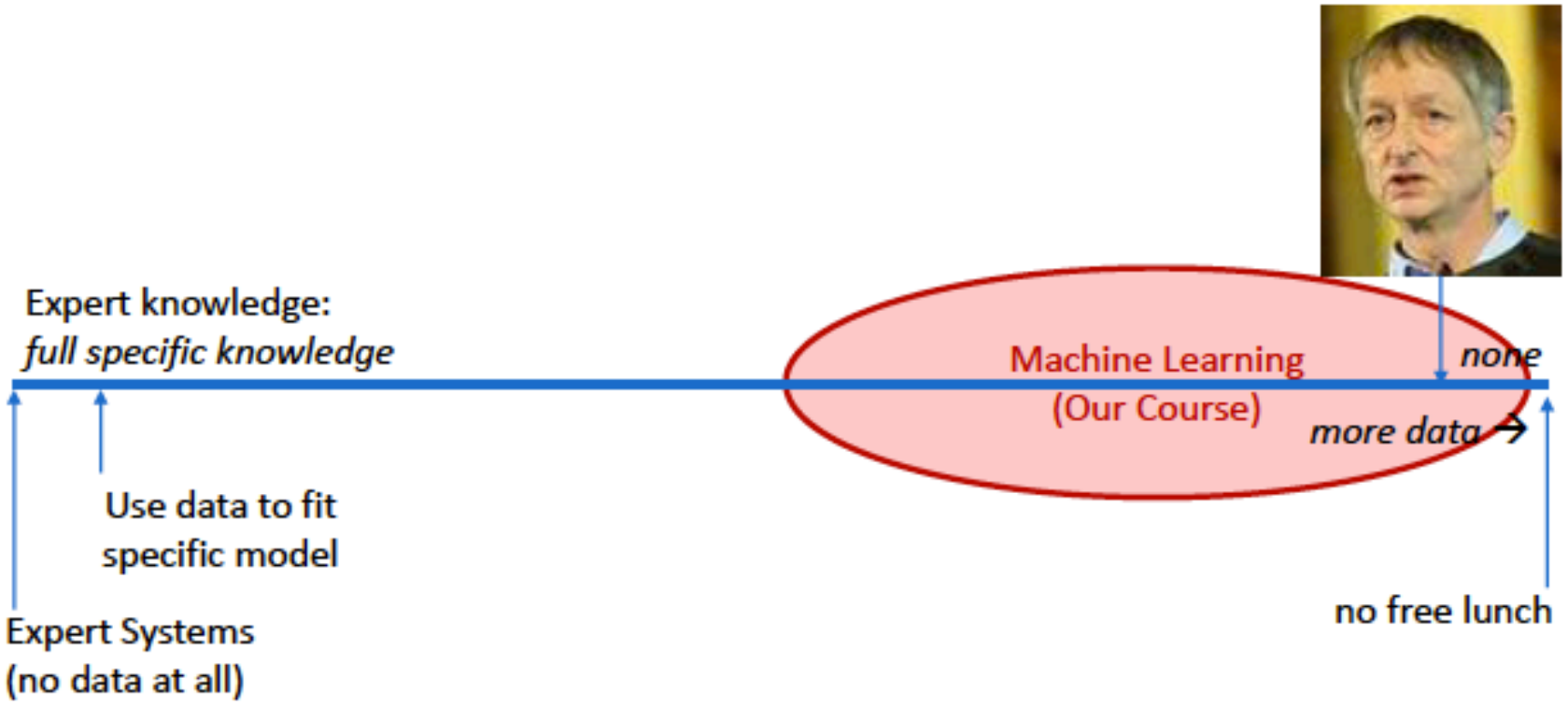
No Free Lunch theorem

- averaged over all possible data generating distributions, every classification algorithm has the same error rate when classifying previously unobserved points.

- if we make assumptions about the kinds of probability distributions we encounter in real-world applications, then we can design learning algorithms that perform well on these distributions.

- the goal of machine learning research is not to seek a universal learning algorithm or the absolute best learning algorithm. Instead, the goal is to understand what kinds of distributions are relevant to the "real world" that an AI agent experiences and what kinds of machine learning algorithms perform well on data drawn from the kinds of data generating distributions we care about.

More Data, Less Expert Knowledge



Misinterpreting empirical results

Generalization in Deep Learning

Kenji Kawaguchi Leslie Pack Kaelbling Yoshua Bengio
Massachusetts Institute of Technology University of Montreal, CIFAR Fellow

Abstract

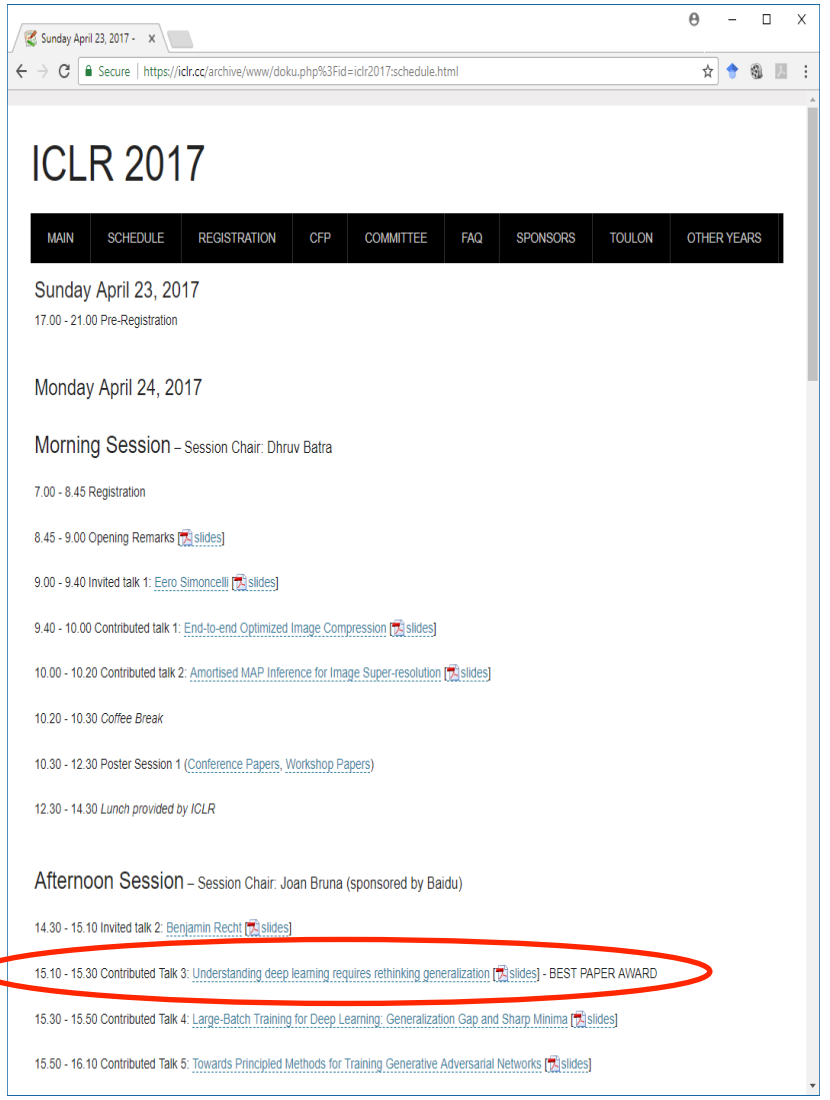
Throughout this chapter, we provide theoretical insights into why and how deep learning can generalize well, despite its large capacity, complexity, possible algorithmic instability, nonrobustness, and sharp minima, responding to an open question in the literature. We also propose new open problems and discuss the limitations of our results.

1. Introduction

Deep learning has seen significant practical success and has had a profound impact on the conceptual bases of machine learning and artificial intelligence. Along with its practical success, the theoretical properties of deep learning have been a subject of active investigation. For *expressivity* of neural networks, there are classical results regarding their universality (Leshno et al., 1993) and exponential advantages over hand-crafted features (Barron, 1993). Another series of theoretical studies have considered how *trainable* (or optimizable) deep hypothesis spaces are, revealing structural properties that may enable non-convex optimization (Choromanska et al., 2015; Kawaguchi, 2016a). However, merely having an *expressive* and *trainable* hypothesis space does not guarantee good performance in predicting the values of future inputs, because of possible over-fitting to training data. This leads to the study of *generalization*, which is the focus of this chapter.

Some classical theory work attributes generalization ability to the use of a low-capacity class of hypotheses (Vapnik, 1998; Mohri et al., 2012). From the viewpoint of compact representation, which is related to small capacity, it has been shown that deep hypothesis spaces have an exponential advantage over shallow hypothesis spaces for representing some classes of natural target functions (Pascanu et al., 2014; Montufar et al., 2014; Livni et al., 2014; Telgarsky, 2016; Poggio et al., 2017). In other words, when some assumptions implicit in the hypothesis space (e.g., deep composition of piecewise linear transformations) are approximately satisfied by the target function, one can achieve very good generalization, compared to methods that do not rely on that assumption. However, a recent paper (Zhang et al., 2017) has empirically shown that successful deep hypothesis spaces have sufficient capacity to memorize random labels. This observation has been called an "apparent paradox" and has led to active discussion by many researchers (Arpit et al., 2017; Krueger et al., 2017; Hoffer et al., 2017; Wu et al., 2017; Dziugaite and Roy, 2017; Dinh et al., 2017). Zhang et al. (2017) concluded with an open problem stating that understanding such observations require rethinking generalization, while Dinh et al. (2017) stated that explaining why deep learning models can generalize well, despite their overwhelming capacity, is an open area of research.

We begin, in Section 3, by illustrating that, even in the case of linear models, hypothesis spaces with overwhelming capacity can result in arbitrarily small test errors and expected risks. Here, *test error* is the error of a learned hypothesis on data that it was not trained on, but which is often drawn from the same distribution. Test error is a measure of how well the hypothesis generalizes to new data.



https://arxiv.org/pdf/1710.05468.pdf - book chapter in "Mathematics of the Deep Learning", Cambridge University Press, to appear

https://arxiv.org/pdf/1611.03530.pdf

UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION

Chiyuan Zhang*
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht†
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com

ABSTRACT

Despite their massive size, successful deep artificial neural networks can exhibit a remarkably small difference between training and test performance. Conventional wisdom attributes small generalization error either to properties of the model family, or to the regularization techniques used during training.

Through extensive systematic experiments, we show how these traditional approaches fail to explain why large neural networks generalize well in practice. Specifically, our experiments establish that state-of-the-art convolutional networks for image classification trained with stochastic gradient methods easily fit a random labeling of the training data. This phenomenon is qualitatively unaffected by explicit regularization, and occurs even if we replace the true images by completely unstructured random noise. We corroborate these experimental findings with a theoretical construction showing that simple depth two neural networks already have perfect finite sample expressivity as soon as the number of parameters exceeds the number of data points as it usually does in practice.

We interpret our experimental findings by comparison with traditional models.

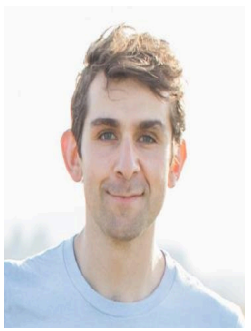
Understanding Deep Learning Requires Rethinking Generalization

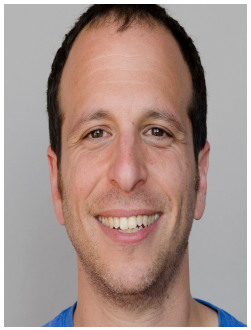
Chiyuan Zhang
CSAIL, CBMM, MIT

Poster: Wednesday Morning C23


Chiyuan Zhang

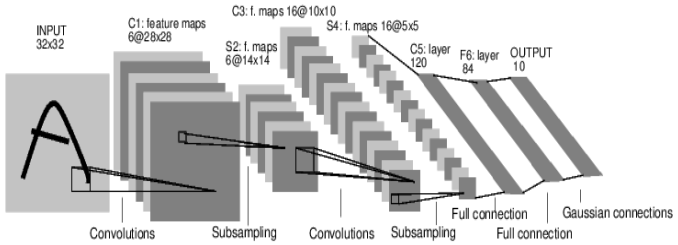

Samy Bengio

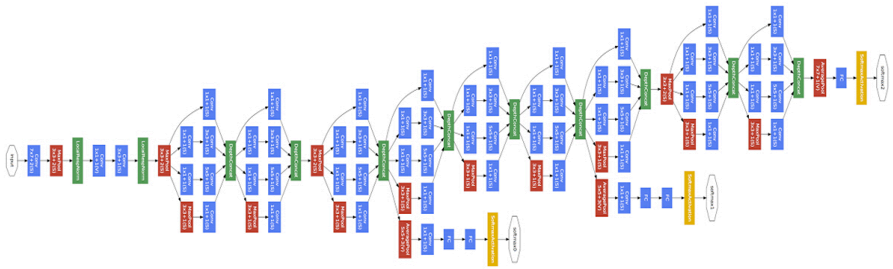

Moritz Hardt

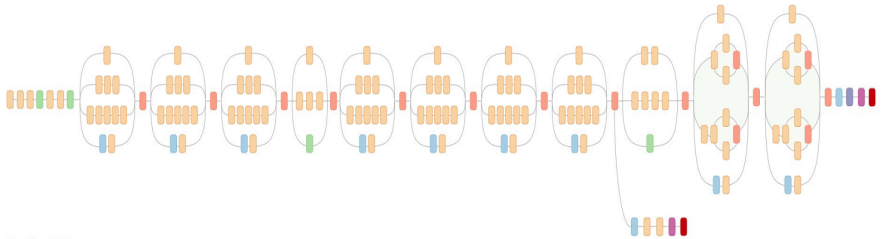

Benjamin Recht

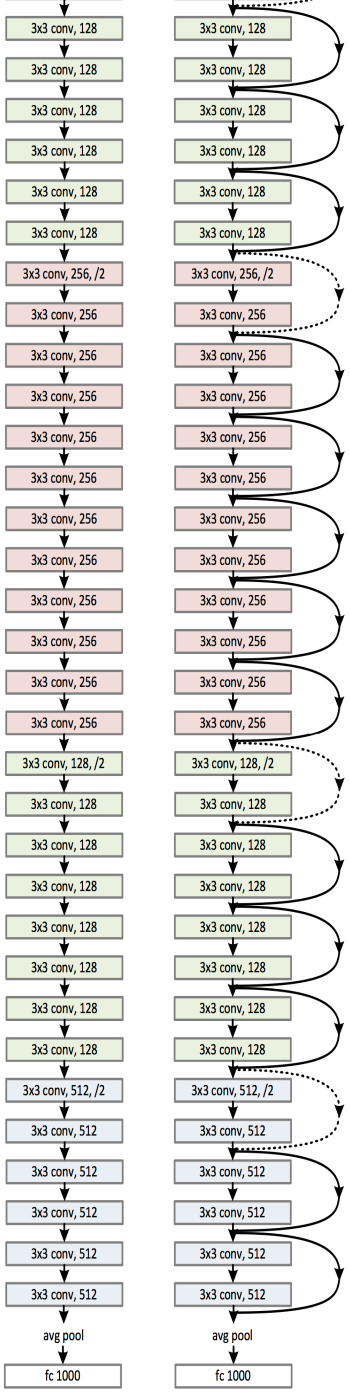

Oriol Vinyals

Deep Learning

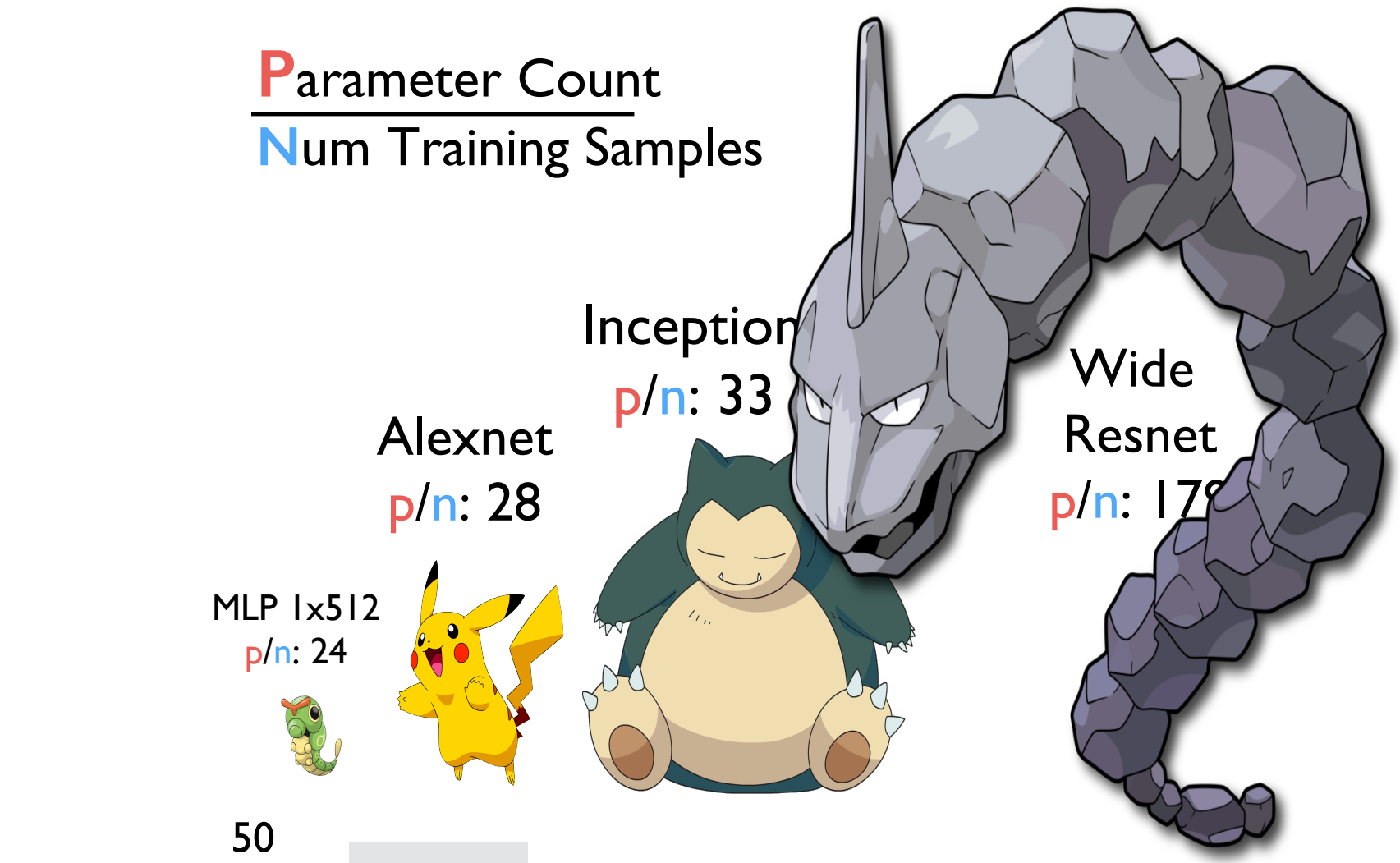


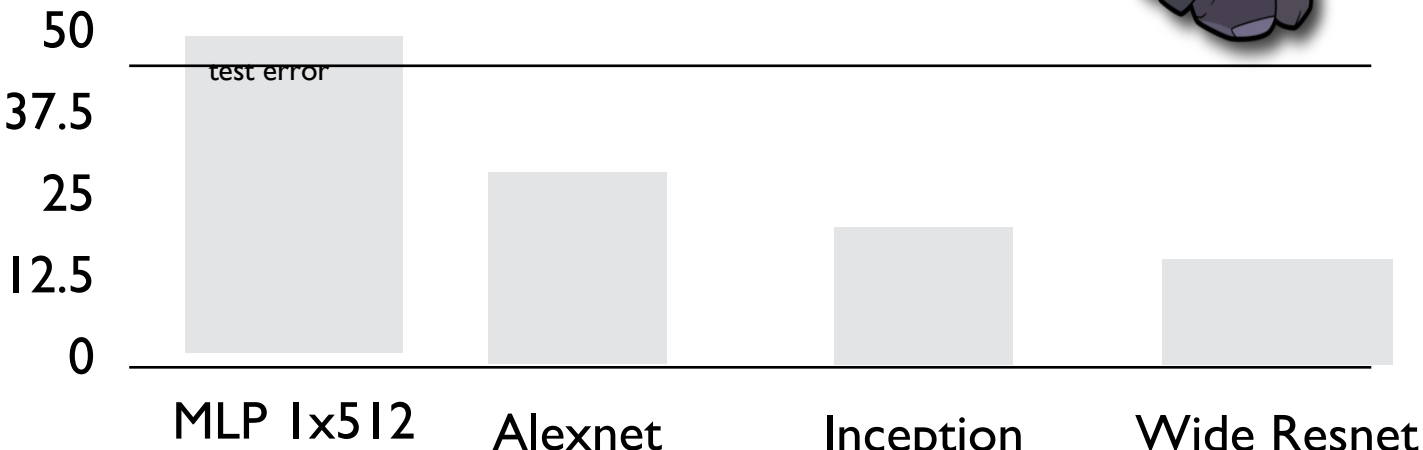






Parameter Count / Num Training Samples

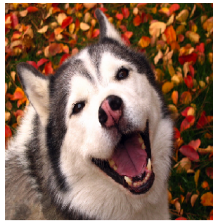




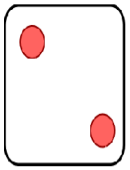
Randomization Test

Deep Neural Networks
easily fit random labels.

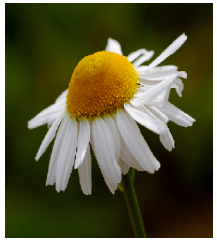
Random Label Dataset



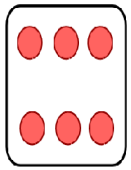
Dog



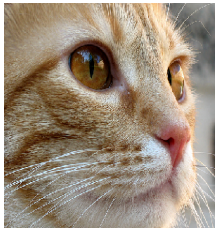
Cat



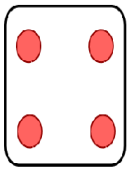
Flower



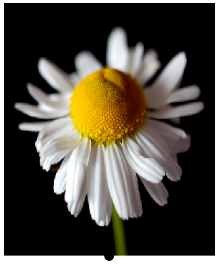
Dog



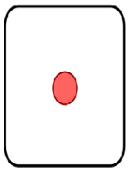
Cat



Bus

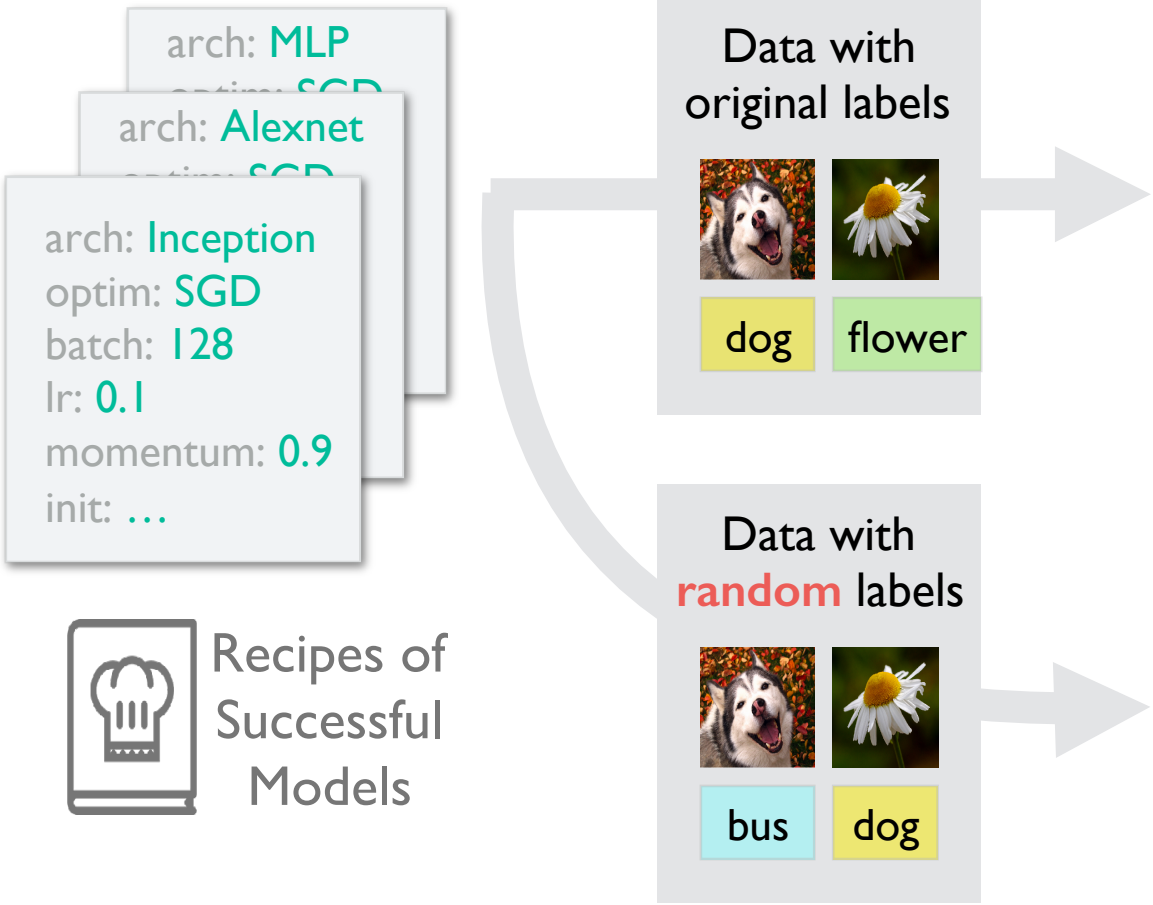


Flower



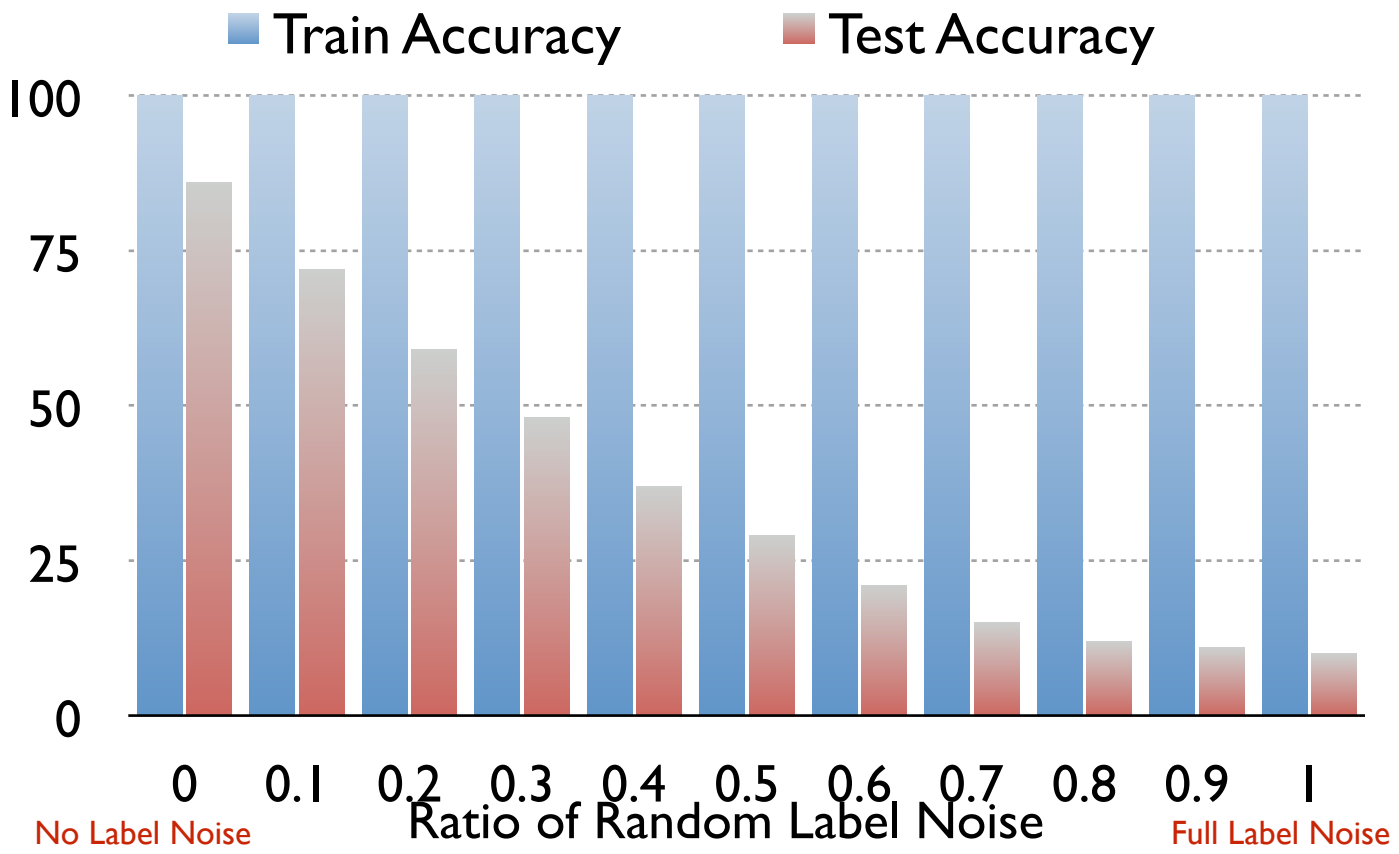
Bird

Randomization Test

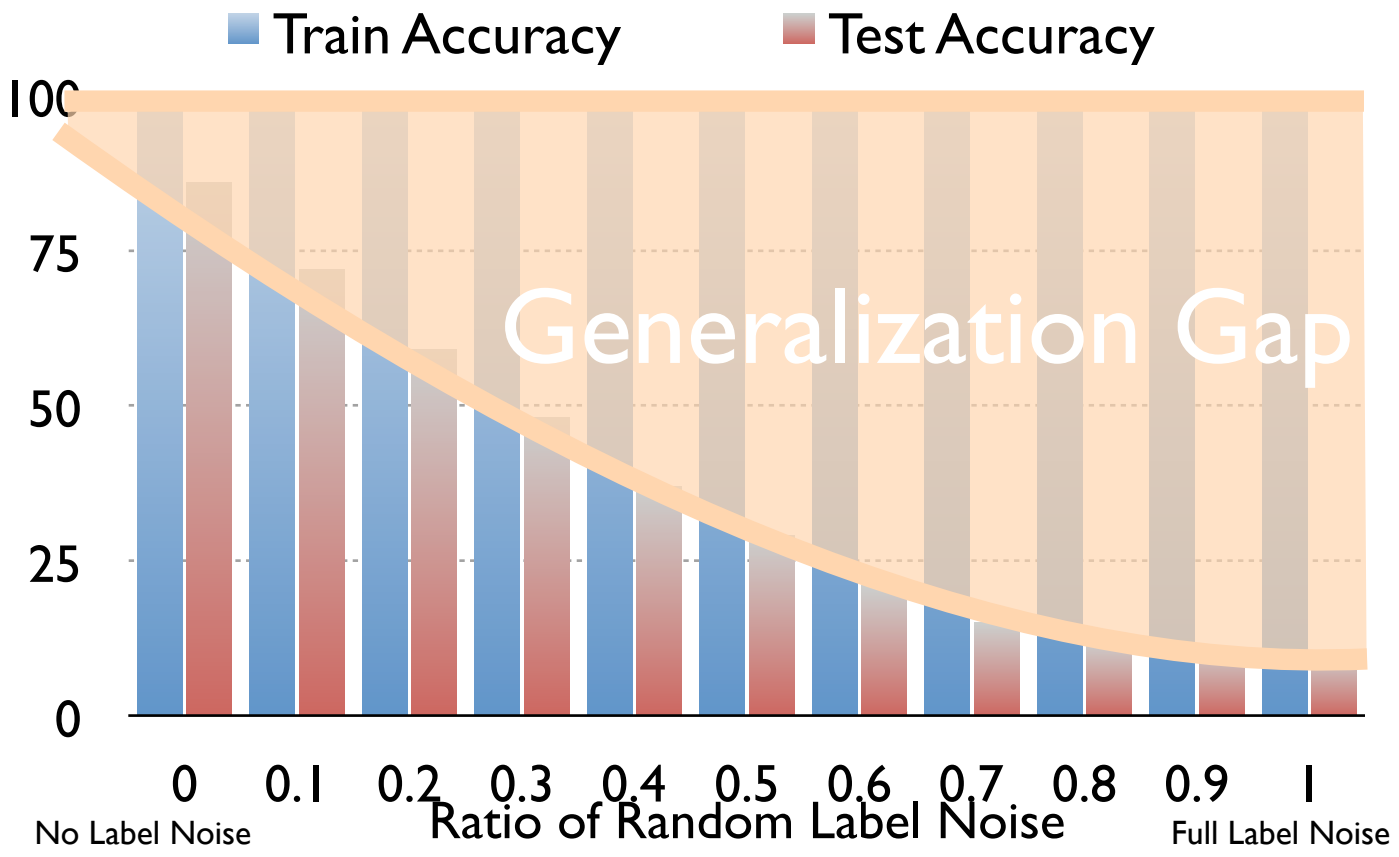


https://www.tensorflow.org/
http://mxnet.io/

Randomization Test



Randomization Test



Randomization Test

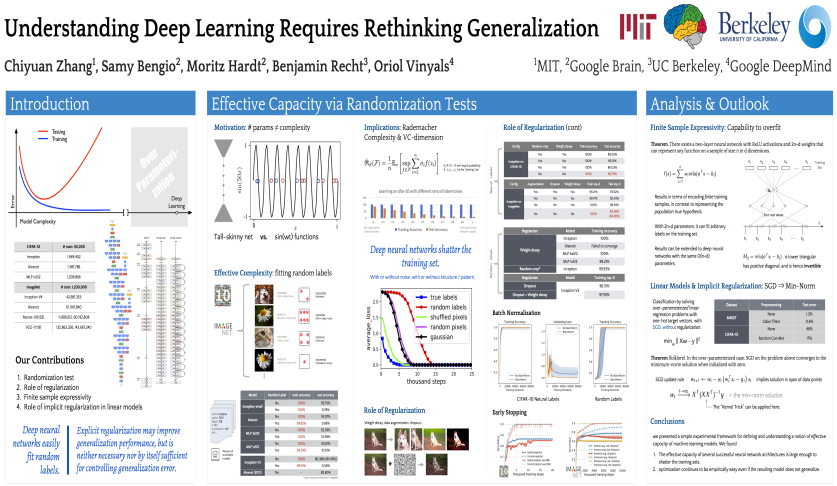
Deep Neural Networks
easily fit random labels.

Conclusion

Simple experimental framework for understanding the effective capacity of deep learning models

Successful DeepNets are able to shatter the training set

Other formal measures of complexity for the models / algorithms / data distributions are needed to precisely explain the over-parameterized regime



Generalization in Deep Learning

Kenji Kawaguchi Leslie Pack Kaelbling Yoshua Bengio
Massachusetts Institute of Technology University of Montreal, CIFAR Fellow

Abstract

Throughout this chapter, we provide theoretical insights into why and how deep learning can generalize well, despite its large capacity, complexity, possible algorithmic instability, nonrobustness, and sharp minima, responding to an open question in the literature. We also propose new open problems and discuss the limitations of our results.

1. Introduction

Deep learning has seen significant practical success and has had a profound impact on the conceptual bases of machine learning and artificial intelligence. Along with its practical success, the theoretical properties of deep learning have been a subject of active investigation. For *expressivity* of neural networks, there are classical results regarding their universality (Leshno et al., 1993) and exponential advantages over hand-crafted features (Barron, 1993). Another series of theoretical studies have considered how *trainable* (or optimizable) deep hypothesis spaces are, revealing structural properties that may enable non-convex optimization (Choromanska et al., 2015; Kawaguchi, 2016a). However, merely having an *expressive* and *trainable* hypothesis space does not guarantee good performance in predicting the values of future inputs, because of possible over-fitting to training data. This leads to the study of *generalization*, which is the focus of this chapter.

Some classical theory work attributes generalization ability to the use of a low-capacity class of hypotheses (Vapnik, 1998; Mohri et al., 2012). From the viewpoint of compact representation, which is related to small capacity, it has been shown that deep hypothesis spaces have an exponential advantage over shallow hypothesis spaces for representing some classes of natural target functions (Pascanu et al., 2014; Montufar et al., 2014; Livni et al., 2014; Telgarsky, 2016; Poggio et al., 2017). In other words, when some assumptions implicit in the hypothesis space (e.g., deep composition of piecewise linear transformations) are approximately satisfied by the target function, one can achieve very good generalization, compared to methods that do not rely on that assumption. However, a recent paper (Zhang et al., 2017) has empirically shown that successful deep hypothesis spaces have sufficient capacity to memorize random labels. This observation has been called an "apparent paradox" and has led to active discussion by many researchers (Arpit et al., 2017; Krueger et al., 2017; Hoffer et al., 2017; Wu et al., 2017; Dziugaite and Roy, 2017; Dinh et al., 2017). Zhang et al. (2017) concluded with an open problem stating that understanding such observations require rethinking generalization, while Dinh et al. (2017) stated that explaining why deep learning models can generalize well, despite their overwhelming capacity, is an open area of research.

We begin, in Section 3, by illustrating that, even in the case of linear models, hypothesis spaces with overwhelming capacity can result in arbitrarily small test errors and expected risks. Here, *test error* is the error of a learned hypothesis on data that it was not trained on, but which is often drawn from the same distribution. Test error is a measure of how well the hypothesis generalizes to new data.

Demystifying the Generalization Puzzle

3.1 Consistency of theory

The empirical observations in (Zhang et al., 2017) and our results above may seem to contradict the results of statistical learning theory. However, there is no contradiction, and the apparent inconsistency arises from the misunderstanding and misuse of the precise meanings of the theoretical statements.

Statistical learning theory can be considered to provide two types of statements relevant to the scope of this chapter. The first type (which comes from upper bounds) is logically in the form of "p implies q," where $p :=$ "the hypothesis-space complexity is small" (or another statement about stability, robustness, or flat minima), and $q :=$ "the generalization gap is small." Notice that "p implies q" does not imply "q implies p." Thus, based on statements of this type, it is entirely possible that the generalization gap is small even when the hypothesis-space complexity is large or the learning mechanism is unstable, non-robust, or subject to sharp minima.

Demystifying the Generalization Puzzle

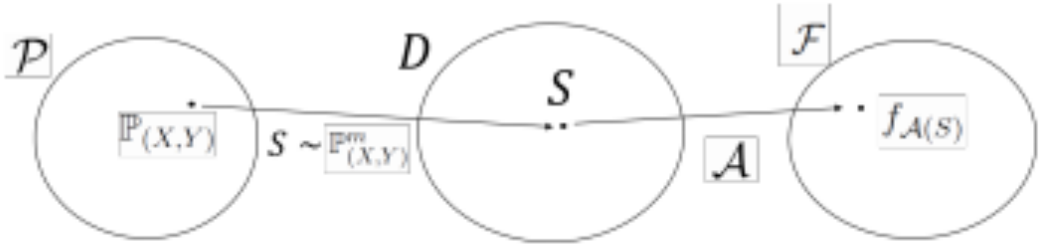


Figure 1: An illustration of differences in assumptions. Statistical learning theory analyzes the generalization behaviors of $f_{\mathcal{A}(S)}$ over randomly-drawn *unspecified* datasets $S \in D$ according to some *unspecified* distribution $\mathbb{P}_{(X,Y)} \in \mathcal{P}$. Intuitively, statistical learning theory concerns more about questions regarding a set $\mathcal{P} \times D$ because of the *unspecified* nature of $(\mathbb{P}_{(X,Y)}, S)$, whereas certain empirical studies (e.g., Zhang et al. 2017) can focus on questions regarding each *specified* point $(\mathbb{P}_{(X,Y)}, S) \in \mathcal{P} \times D$.

Generalization Bounds via Validation

In practical deep learning, we typically adopt the training–validation paradigm, usually with a held-out validation set. We then search over hypothesis spaces by changing architectures (and other hyper-parameters) to obtain low validation error. In this view, we can conjecture the reason that deep learning can sometimes generalize well as follows: it is partially because we can obtain a good model via search using a validation dataset. Indeed, as an example, Proposition 5 states that if the validation error of a hypothesis is small, it is guaranteed to generalize well, regardless of its capacity,

Rademacher complexity, stability, robustness, and flat minima. Let $S^{(\mathrm{val})}_{m_{\mathrm{val}}}$ be a held-out validation dataset of size m_{val}, which is independent of the training dataset S.

Proposition 5. (example of generalization guarantee via validation error) *Assume that* $S^{(\mathrm{val})}_{m_{\mathrm{val}}}$ *is generated by i.i.d. draws according to a true distribution* $\mathbb{P}_{(X,Y)}$. *Let* $\kappa_{f,i} = \mathcal{R}[f] - \mathcal{L}(f(x_i), y_i)$ *for* $(x_i, y_i) \in S^{(\mathrm{val})}_{m_{\mathrm{val}}}$. *Suppose that* $\mathbb{E}[\kappa^2_{f,i}] \leq \gamma^2$ *and* $|\kappa_{f,i}| \leq C$ *almost surely, for all* $(f, i) \in \mathcal{F}_{\mathrm{val}} \times \{1, \ldots, m_{\mathrm{val}}\}$. *Then, for any* $\delta > 0$, *with probability at least* $1 - \delta$, *the following holds for all* $f \in \mathcal{F}_{\mathrm{val}}$:

$$\mathcal{R}[f] \leq \mathcal{R}_{S^{(\mathrm{val})}_{m_{\mathrm{val}}}}[f] + \frac{2C \ln(\frac{|\mathcal{F}_{\mathrm{val}}|}{\delta})}{3m_{\mathrm{val}}} + \sqrt{\frac{2\gamma^2 \ln(\frac{|\mathcal{F}_{\mathrm{val}}|}{\delta})}{m_{\mathrm{val}}}}.$$

Generalization Bounds via Validation

$$\mathcal{R}[f] \leq \mathcal{R}_{S^{(\mathrm{val})}_{m_{\mathrm{val}}}}[f] + \frac{2C \ln(\frac{|\mathcal{F}_{\mathrm{val}}|}{\delta})}{3m_{\mathrm{val}}} + \sqrt{\frac{2\gamma^2 \ln(\frac{|\mathcal{F}_{\mathrm{val}}|}{\delta})}{m_{\mathrm{val}}}}.$$

Here, $\mathcal{F}_{\mathrm{val}}$ is defined as a set of models f that is independent of a held-out validation dataset $S^{(\mathrm{val})}_{m_{\mathrm{val}}}$, but can depend on the training dataset S. For example, $\mathcal{F}_{\mathrm{val}}$ can contain a set of models f such that each element f is a result at the end of each epoch during training with at least 99.5% *training* accuracy. In this example, $|\mathcal{F}_{\mathrm{val}}|$ is at most (the number of epochs) $\times$ (the cardinality of the set of possible hyper-parameter settings), and is likely much smaller than that because of the 99.5% training accuracy criteria and the fact that a space of many hyper-parameters is narrowed down by using the training dataset as well as other datasets from different tasks. If a hyper-parameter search depends on the validation dataset, $\mathcal{F}_{\mathrm{val}}$ must be the possible space of the search instead of the space actually visited by the search. We can also use a sequence $\{\mathcal{F}^{(j)}_{\mathrm{val}}\}_j$ (see Appendix A).

Generalization Bounds via Validation

The bound in Proposition 5 is non-vacuous and tight enough to be practically meaningful. For example, consider a classification task with 0–1 loss. Set $m_{\text{val}} = 10,000$ (e.g., MNIST and CIFAR-10) and $\delta = 0.1$. Then, even in the worst case with $C = 1$ and $\gamma^2 = 1$ and even with $|\mathcal{F}_{\text{val}}| = 1,000,000,000$, we have with probability at least 0.9 that $\mathcal{R}[f] \leq \mathcal{R}_{S_{m_{\text{val}}}^{(\text{val})}}[f] + 6.94\%$ for all $f \in \mathcal{F}_{\text{val}}$. In a non-worst-case scenario, for example, with $C = 1$ and $\gamma^2 = (0.05)^2$, we can replace 6.94% by 0.49%. With a larger validation set (e.g., ImageNet) and/or more optimistic C and γ^2, we can obtain much better bounds.

Theoretical Machine Learning

- understand the main concepts underlying machine learning through basic theory
  - the goal is not detailed theory and theorem-proving;
  - emphasis on concepts, les on specific algorithms.
- know the prominent methods used in contemporary machine learning
- learn how to use machine learning *correctly*
- no programming, do ML problems in seminar

Next time

- will cover Chapters 2 and 3 from the book
- do exercises in seminar (from those proposed)