Thesis title

Jane Doe

10/18/22

Table of contents

List of Figures

List of Tables

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

1 Introduction

In his acceptance speech for the Test-of-Time award in NeurIPS 2017,[1] Ali Rahimi[2] started a controversy by frankly declaring (Rahimi 2018, 12'10"). His concerns on the lack of theoretical understanding of machine learning for critical decision-making are rightful:

The next day, Yann LeCun[3] responded:

Both researchers, at least, agree upon one thing: the practice of machine learning has outpaced its theoretical development. That is certainly a research opportunity.

[1] Conference on Neural Information Processing.

[2] Research Scientist, Google.

Rahimi, Ali. 2018. "Ali Rahimi NIPS 2017 Test-of-Time Award Presentation Speech." https://youtu.be/x7psGHgatGM; Youtube. https://youtu.be/x7psGHgatGM.

[3] Deep Learning pioneer and 2018 Turing award winner. https://www.facebook.com/yann.lecun/posts/10154938130592143

1.1 Context

1.1.1 A Tale of Babylonians and Greeks

Richard Feynman ([fig:feynman]) used to lecture this story (Feynman 1994): Babylonians were pioneers in mathematics; Yet, the Greeks took the credit. We are used to the Greek way of doing Math: start from the most basic axioms and build up a knowledge system. Babylonians were quite the opposite; they were pragmatic. No knowledge was considered more fundamental than others, and there was no urge to derive proofs in a particular order. Babylonians were concerned with the phenomena, Greeks with the ordinance. In Feynman's view, science is constructed in the Babylonian way. There is no fundamental truth. Theories try to connect dots from different pieces of knowledge. Only as science advances, one can worry about reformulation, simplification and ordering. Scientists are Babylonians; mathematicians are Greeks.

Feynman, Richard. 1994. *The Character of Physical Law*. Modern Library.

Mathematics and science are both tools for knowledge acquisition. They are also social constructs that rely on peer-reviewing. They are somewhat different, however.

Science is empiric, based on facts collected from **experience**. When physicists around the world measured events that corroborated Newton's *"Law of Universal Gravitation"*, they did not prove it correct; they just made his theory more and more plausible. Still, only one experiment was needed to show that Einstein's *Relativity Theory* was even more believable. In contrast, we can and do prove things in mathematics.

In mathematics, knowledge is absolute truth, and the way one builds new knowledge with it, its inference method, is deduction. Mathematics is a language, a formal one, a tool to precisely communicate some kinds of thoughts. As it happens with natural languages, there is beauty in it. The mathematician expands the boundaries of expression in this language.

In science, there are no axioms: a falsifiable hypothesis/theory is proposed, and logical conclusions (predictions) from the theory are empirically tested. Despite inferring hypotheses by induction, there is no influence of psychology in the process. A tested hypothesis is not absolute truth. A hypothesis is never verified, only falsified by experiments (Popper 2004, 31–50). Scientific knowledge is belief justified by experience; there are degrees of plausibility.

Popper, Karl. 2004. *A Lógica Da Pesquisa Científica*. Translated by Leonidas Hegenberg and Octanny Silveira. São Paulo: Cultrix.

Understanding the epistemic contrast between mathematics and science will help us understand the past of AI and avoid some perils in its future.

1.1.2 The importance of theoretical narratives

Science is a narrative of how we understand Nature (Gleiser and Sowinski 2018). In science, we collect facts, but they need interpretation. The logical conclusion from the hypothesis that predicts some behaviour in nature gives a plausible *meaning* to what we observed.

Gleiser, Marcelo, and Damian Sowinski. 2018. "The Map and the Territory." In *The Frontiers Collection*, edited by Shyam Wuppuluri and Francisco Antonio Doria. Springer International Publishing. https://doi.org/10.1007/978-3-319-72478-2.

To illustrate, take the ancient human desire of flying. There have always been stories of men strapping wings to themselves and attempting to fly by jumping from a tower and flapping those wings like birds (see) (Farrington 2016). While concepts like lift, stability, and control were poorly understood, most human flight attempts ended in severe injury or even death. It did not matter how much evidence, how many hours of seeing different animals flying, those ludicrous brave men experienced; the *meaning* they

Farrington, Karen. 2016. *The Blitzed City: The Destruction of Coventry, 1940*. London: Aurum Press.

took from what they saw was wrong, and their predictions incorrect.

They did not die in vain[4]; Science advances when scientists are wrong. Theories must be falsifiable, and scientists cheer for their failure. When it fails, there is room for new approaches. Only when we understood the observations in animal flight from the aerodynamics perspective, we learned to fly better than any other animal before. Science works by a "natural selection" of ideas, where only the fittest ones survive until a better one is born. Chaitin also points out that an idea has "fertility" to the extent to which it "illuminates us, inspires us with other ideas, and suggests unsuspected connections and new viewpoints" (Chaitin 2006, 9).

[4] Those "researchers" deserved, at least, a Darwin Award of Science. The Darwin Award is satirical honours that recognise individuals who have unwillingly contributed to human evolution by selecting themselves out of the gene pool.

Chaitin, Gregory. 2006. *Meta Math! The Quest for Omega*. Vintage Books.

Being a Babylonian enterprise, science has no clear path. One of the exciting facts one can learn by studying its history is that robust discoveries have arisen through the study of phenomena in human-made devices (Pierce, n.d.). For instance, Carnot's first and only scientific work (Klein 1974) gave birth to thermodynamics: the study of energy, the conversion between its different forms, and the ability of energy to do work; the science that explains how steam engines work. However, steam engines came before Carnot's work and were studied by him. Such human-made devices may present a simplified instance of more complex natural phenomena.

Pierce, John R. n.d. *An Introduction to Information Theory: Symbols, Signals and Noise*. Dover Publications.

Klein, Martin J. 1974. "Carnot's Contribution to Thermodynamics." *Physics Today* 27 (8): 23–28. https://doi.org/10.1063/1.3128802.

Another example is Information Theory. Several insights of Shannon's theory of communication were generalisations of ideas already present in Telegraphy (Shannon 1948). New theories in artificial intelligence can, therefore, be developed from insights in the study of deep learning phenomena.[5]

Shannon, Claude E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (3): 379–423.

[5] Understanding human intelligence using artificial intelligence is a field of study called Computational Neuroscience.

1.1.3 Bringing science to Computer Science

Despite the name, Computer Science has been more mathematics than science. We, computer scientists, are very comfortable with theorems and proofs, not much with theories.

Nevertheless, AI has essentially become a Babylonian enterprise, a scientific endeavour. Thus, there is no surprise when some computer scientists still see AI with some distrust and even disdain, despite its undeniable usefulness:

- Even among AI researchers, there is a trend of "mathiness" and speculation disguised as explanations in conference papers (Lipton and Steinhardt 2018).
- There are few venues for papers that describe surprising phenomena without trying to come up with an explanation. As if the mere inconsistency of the current theoretical framework was unworthy of publication.

Lipton, Zachary C., and Jacob Steinhardt. 2018. "Troubling Trends in Machine Learning Scholarship." https://arxiv.org/abs/1807.03341.

While physicists rejoice in finding phenomena that contradict current theories, computer scientists get baffled. In Natural Sciences, unexplained phenomena lead to theoretical development. Some believe they bring *winters*, periods of progress stagnation and lack of funding in AI.[6]

Artificial Intelligence has been through several of the aforementioned "winters". In 1957, Herbert Simon[7] famously predicted that within ten years, a computer would be a chess champion (Russell, Norvig, and Davis 2010, sec. 1.3). It took around 40 years, in any case. Computer scientists lacked understanding of the exponential nature of the problems they were trying to solve: Computational Complexity Theory had yet to be invented.

Machine Learning Theory (computational and statistical) tries to avoid a similar trap by analysing and classifying learning problems according to the number of samples required to learn them (besides the number of steps). The matter of concern is that it currently predicts that generalisation requires simpler models in terms of parameters. In total disregard to the theory, deep learning models have shown spectacular generalisation power with hundreds of millions of parameters (and even more impressive overfitting capacity).

[6] This seems to be Yann LeCun's opinion: However, due to all possible alternative explanations (lack of computational power, no availability of massive annotated datasets), it seems harsh or simply wrong to blame theorists.

[7] Herbert Simon (1916–2001) received the Turing Award in 1975, and the Nobel Prize in Economics in 1978.

Russell, Stuart J., Peter Norvig, and Ernest Davis. 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. Prentice Hall Series in Artificial Intelligence. Prentice Hall.

1.2 Problem

In the last decade, we have witnessed a myriad of astonishing successes in Deep Learning. Despite those many successes in research and industry applications, we may again be climbing a peak of inflated expectations. If in the past, the false solution was to "add computation power" on problems, today we try to solve them by "piling data"([fig:machine_learning_2x]). Such behaviour has triggered a winner-takes-all competition for who

collects more data (our data) amidst a handful of large corporations, raising ethical concerns about privacy and concentration of power (O'Neil 2016).

Nevertheless, we know that learning from way fewer samples is possible: humans show a much better generalisation ability than our current state-of-the-art artificial intelligence. To achieve such needed generalisation power, we may need to understand better how learning happens in deep learning. Rethinking generalisation might reshape the foundations of machine learning theory (Zhang et al. 2016).

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. USA: Crown Publishing Group.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. "Understanding Deep Learning Requires Rethinking Generalization." https://arxiv.org/abs/1611.03530.

1.2.1 Possible new explanation in the horizon

In 2015, Tishby and Zaslavsky (2015) proposed a theory of deep learning (Tishby and Zaslavsky 2015) based on the information-theoretical concept of the bottleneck principle, of which Tishby is one of the authors. Later, in 2017, Shwartz-Ziv and Tishby (2017) followed up on the IBT with the paper , which was presented in a well-attended workshop[8], with appealing visuals that clearly showed a *"phase transition"* happening during training. The video posted on Youtube (Tishby 2017) became a "sensation"[9], and received a wealth of publicity when well-known researchers like Geoffrey Hinton[10], Samy Bengio (Apple) and Alex Alemi (Google Research) have expressed interest in Tishby's ideas (Wolchover 2017). they are called formal languages.

> *I believe that the information bottleneck idea could be very important in future deep neural network research.*
> — Alex Alemi

Tishby, Naftali, and Noga Zaslavsky. 2015. "Deep Learning and the Information Bottleneck Principle." In *2015 IEEE Information Theory Workshop (ITW)*, 1–5. IEEE.

Tishby, Naftali, and Noga Zaslavsky. 2015. "Deep Learning and the Information Bottleneck Principle." In *2015 IEEE Information Theory Workshop (ITW)*, 1–5. IEEE.

Shwartz-Ziv, Ravid, and Naftali Tishby. 2017. "Opening the Black Box of Deep Neural Networks via Information." https://arxiv.org/abs/1703.00810.

Andrew Saxe (Harvard University) rebutted Shwartz-Ziv and Tishby (2017) claims in and was followed by other critics. According to Saxe, it was impossible to reproduce (Shwartz-Ziv and Tishby 2017)'s experiments with different parameters.

Has the initial enthusiasm on the IBT been unfounded? Have we let us "fool ourselves" by beautiful charts and a good story?

1.2.2 Problem statement

The practice of modern machine learning has outpaced its theoretical development. In particular, deep learning models

[8] Deep Learning: Theory, Algorithms, and Applications. Berlin, June 2017 http://doc.ml.tu-berlin.de/dlworkshop2017

Tishby, Naftali. 2017. "Information Theory of Deep Learning." https://youtu.be/bLqJHjXihK8. https://youtu.be/bLqJHjXihK8.

[9] By the time of this writing, this video as more than 84,000 views, which is remarkable for an hour-long workshop presentation in an academic niche. https://youtu.be/bLqJHjXihK8

[10] Another Deep Learning Pioneer **and** Turing award winner (2018).

Wolchover, Natalie. 2017. "New Theory Cracks Open the Black Box of Deep Learning." https://www.quantamagazine.org/new-theory-cracks-open-the-black-box-of-deep-learning-20170921/;

present generalisation capabilities unpredicted by the current machine learning theory. There is yet no established new general theory of learning which handles this problem.

IBT was proposed as a possible new theory with the **potential** of filling the theory-practice gap. Unfortunately, to the extent of our knowledge, **there is still no comprehensive digest of IBT nor an analysis of how it relates to current MLT**.

1.3 Objective

This dissertation aims to investigate *to what extent* can the emergent Information Bottleneck Theory help us better understand Deep Learning and its phenomena, especially generalisation, presenting its strengths, weaknesses and research opportunities.

1.3.1 Research Questions

1.4 Methodology

1. Given that IBT is yet not a well-established learning theory, there were two difficulties that the research had to address:

    1. There is a growing interest in the subject, and new papers are published every day. It was essential to select literature and restrain the analysis.

    2. Early on, the marks of an emergent theory in its infancy manifested in the form of missing assumptions, inconsistent notation, borrowed jargon, and seeming missing steps. Foremost, it was unclear what was missing from the theory and what was missing in our understanding.

    An initial literature review on IBT was conducted to define the scope.[11] We then chose to narrow the research to **theoretical perspective on generalisation**, where we considered that it could bring fundamental advances. We made the deliberate choice of going deeper in a limited area of IBT and not broad, leaving out a deeper experimental and application analysis, all the work on ITL[12]

[11] Not even the term IBT is universally adopted.

[12] ITL makes the opposite path we are taking, bringing concepts of machine learning to information theory problems.

(Principe 2010) and statistical-mechanics-based analysis of SGD (P. Chaudhari and Soatto 2018; Pratik Chaudhari et al. 2019). From this set of constraints, we chose a list of pieces of IBT literature to go deeper ([ch:literature]).

2. In order to answer , we discuss the epistemology of AI to choose fundamental axioms (definition of intelligence and the definition of knowledge) with which we deduced from the ground up MLT, IT and IBT, revealing hidden assumptions, pointing out similarities and differences. By doing that, we built a "genealogy" of these research fields. This comparative study was essential for identifying missing gaps and research opportunities.

3. In order to answer , we first dissected the selected literature ([ch:literature]) and organised scattered topics in a comprehensive sequence of subjects.

4. In the process of the literature digest, we identified results, strengths, weaknesses and research opportunities.

Principe, Jose C. 2010. *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Springer Science & Business Media.

Chaudhari, P., and S. Soatto. 2018. "Stochastic Gradient Descent Performs Variational Inference, Converges to Limit Cycles for Deep Networks." In *2018 Information Theory and Applications Workshop (ITA)*, 1–10. https://doi.org/10.1109/ITA.2018.8503224.

Chaudhari, Pratik, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. 2019. "Entropy-Sgd: Biasing Gradient Descent into Wide Valleys." *Journal of Statistical Mechanics: Theory and Experiment* 2019 (12).

1.5 Contributions

In the research conducted, we produced three main results that, to the extent of our knowledge, are original:

1. The dissertation itself is the main expected result: a comprehensive digest of the IBT literature and a snapshot analysis of the field in its current form, focusing on its theoretical implications for generalisation.

2. We propose an Information-Theoretical learning problem different from MDL proposed by (Hinton and Van Camp 1993) for which we derived bounds using Shannon's . These results, however, are only indicative as they lack peer review to be validated.

3. We present a critique on Achille (2019)'s explanation (Achille 2019; Achille and Soatto 2018) for the role of layers in Deep Representation in the IBT perspective ([sec:achille_proof_critique]), pointing out a weakness in the argument that, as far as we know, has not yet been presented. We then propose a counter-intuitive

Hinton, Geoffrey E, and Drew Van Camp. 1993. "Keeping the Neural Networks Simple by Minimizing the Description Length of the Weights." In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, 5–13.

Achille, Alessandro. 2019. "Emergent Properties of Deep Neural Networks." PhD thesis, UCLA. https://escholarship.org/uc/item/8gb8x6w9.

Achille, Alessandro. 2019. "Emergent Properties of Deep Neural Networks." PhD thesis, UCLA. https://escholarship.org/uc/item/8gb8x6w9.

hypothesis that layers reduce the model's "effective" hypothesis space. This *hypothesis* is not formally proven in the present work, but we try to give the intuition behind it ([sec:proposed_hypothesis]). This result has not yet been validated as well.

1.6 Dissertation preview and outline

The dissertation is divided into two main parts ([pt:background] and [pt:emergence_of_theory]), with a break in the middle ([pt:intermezzo]).

1. Background ([pt:background])
    - Chapter 2–Artificial Intelligence: The chapter defines what artificial intelligence is, presents the epistemological differences of intelligent agents in history, and discusses their consequences to machine learning theory.
    - Chapter 3 — Probability Theory: The chapter derives propositional calculus and probability theory from a list of desired characteristics for epistemic agents. It also presents basic Probability Theory concepts.
    - Chapter 4 — Machine Learning Theory: The chapter presents the theoretical framework of Machine Learning, the PAC model, theoretical guarantees for generalisation, and expose its weaknesses concerning Deep Learning phenomena.
    - Chapter 5 — Information Theory: The chapter derives Shannon Information from Probability Theory, explicates some implicit assumptions, and explains basic Information Theory concepts.
2. Intermezzo ([pt:intermezzo])
    - Chapter 6 — Information-Theoretical Epistemology: This chapter closes the background part and opens the IBT part of the dissertation. It shows the connection of IT and MLT in the learning problem, proves that Shannon theorems can be used to prove PAC

bounds and present the MDL Principle, an earlier example of this kind of connection.

3. The emergence of a theory ([pt:emergence_of_theory])

    - Chapter 7 — IB Principle: Explains the IB method and its tools: KL as a natural distortion (loss) measure, the IB Lagrangian and the Information Plane.
    - Chapter 8 — IB and Representation Learning: Presents the learning problem in the IBT perspective (not specific to DL). It shows how some usual choices of the practice of DL emerge naturally from a list of desired properties of representations. It also shows that the information in the weights bounds the information in the activations.
    - Chapter 9 — IB and Deep Learning: This chapter presents the IBT perspective specific to Deep Learning. It presents IBT analysis of Deep Learning training, some examples of applications of IBT to improve or create algorithms; and the IBT learning theory of Deep Learning. We also explain Deep Learning phenomena in the IBT perspective.
    - Chapter 10 — Conclusion: In this chapter, we present a summary of the findings, answer the research questions, and present suggestions for future work.

We found out that IBT does not invalidate MLT; it just interprets complexity not as a function of the data (number of parameters) but as a function of the information contained in the data. With this interpretation, there is no paradox in improving generalisation by adding layers.

Furthermore, they both share more or less the same "genealogy" of assumptions. IBT can be seen as particular case of MLT. Nevertheless, IBT allows us to better understand the training process and provide a different narrative that helps us comprehend Deep Learning phenomena in a more general way.

2 Summary

In summary, this book has no content whatsoever.

References