Chaotic
Black Boxes

The risks and opportunites of deriving knowledge from deep machine learning models.

Brad Flaugher

September 29, 2022

We have now accumulated sufficient evidence to see that whatever language the central nervous system is using, it is characterized by less logical and arithmetical depth than what we are normally used to.

– John von Neumann [1]

Children can learn to use computers in a masterful way, and ... learning to use computers can change the way they learn everything else.

– Seymour A. Papert [2]

Preface

This book is a work in progress, I hope it helps demystify the world of deep learning as I understand it.

Humans won't be able to control superintelligent AI, talk about that here[3]

Talk about Bostrom and GPAI here, and Erdi's answer to that. [4] [5]

Talk about the alignment problem and Ethical freakouts about AI. Talk about the big 3 from [6] [7]

Funding and startups, everybody is doing it, I'm trying to make sense of it

Brad Flaugher

Contents

List of Figures

List of Tables

List of Listings

History: the slow march away from algorithms

1 Ages of Understanding

1.1 (1500-Today) Algorithms: Codified Human Understanding

AI is a shitty term

We tried a lot of things, teaching computers explicit grammar and explicit rules

IMO, this was not AI, this was codified human understanding.

In code, that understanding might look like this....

TODO talk about this book [8]

bradflaugher.com. [1]

[8]: Douthat (2022), *Can We Resist the Age of the Algorithm?*

1: Snarky sidenote!

1.2 (1980-Today) Machine Learning: Data-Derived Insights

Hardware got amazing, we gave up teaching the way we teach ourselves and let the data do the work

We leveraged huge statistical models to regress our way to success

We used building blocks of regression and neurons to train huge models

These models are statistical and deterministic, but ultimately chaotic black boxes..

TODO talk about these books [9] [10] [11]

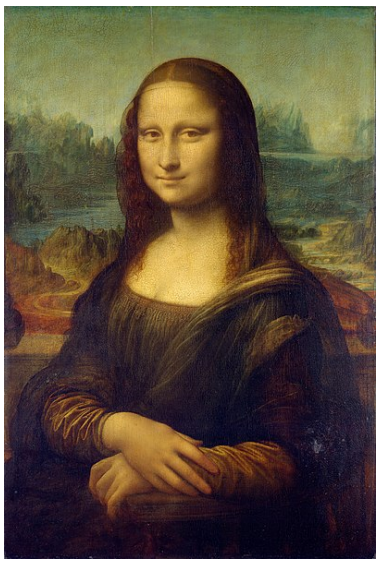

Figure 1.1: The Mona Lisa. https://commons.wikimedia.org/wiki/File:Mona_Lisa,_by_Leonardo_da_Vinci,_from_C2RMF_retouched.jpg

Deep Learning Concepts: layered statistical representations

How Models Read Data 2

2.1 Numerical Data

This is some text and a link to Hey if you want to site something on the side use[3]

2.2 Words

2.3 Sounds

2.4 Images

2.5 Video

2.6 Mixed Datasets

```
cd myproject
docker run tensorflow
#profit!
```

tex.stackexchange.org for help.

[3]: Andreu et al. (2021), *Humans won't be able to control a superintelligent AI, according to a study*

Learning Methods 3

3.1 Unsupervised

3.2 Supervised

3.3 Reinforcement

3.4 Domain Transfer

are you predicting the right thing? Are you really predicting how valuable the company is or just whether it'll be the next meme stock?

3.5 Notes on Ethics

Representation, "fixing the training set" [6], or the Impossibility of Fairness from a model.

TODO talk about these books [12] [13] [7] [6]

```
cd myproject
docker run tensorflow
#profit!
```

`tex.stackexchange.org` for help.

[6]: Christian (2020), *The Alignment Problem: Machine Learning and Human Values*

The Chaotic Black Box: statistical inference with a billion or so parameters

Classifiers 4

4.1 Recommenders

4.2 Facial Recognition

4.3 Sentiment

4.4 Hate Speech

4.5 The Ethics of Classification

Online Advertising, Justice, Job Applications, Creditworthiness, Getting Insurance (Weapons of Math Destruction), Civic Life, /sideciteOneil2017 ; The Default Male, Invisible Women effects snow clearing schedules and drug discovery

```
cd myproject
docker run tensorflow
#profit!
```

tex.stackexchange.org for help.

Transformers 5

5.1 Style Transfer

5.2 Translation

5.3 Text Generation

GPT-3, BERT and Bloom

5.4 Image Generation and Stable Diffusion

Link some cool shit here, Draw Owl!

5.5 The Ethics of Transforming

Ensembles and Mathematical Chaos

6

6.1 Interacting Layers of Statistical Understanding

6.2 Useful Chaos

Appendix

A

Poetry Test

Let's say we want to build an ensemble model to analyze poetry, put a haiku into craiyon's online shit, then we categorize the resulting photo. [3]

[3]: Andreu et al. (2021), *Humans won't be able to control a superintelligent AI, according to a study*

Bibliography

Here are the references in citation order.

[1] John von Neumann and Ray Kurzweil. *The Computer and the Brain (The Silliman Memorial Lectures Series)*. New Haven, CT, USA: Yale University Press, Aug. 2012 (cited on page ii).

[2] Seymour A. Papert. *Mindstorms: Children, Computers, And Powerful Ideas*. New York, NY, USA: Basic Books, Aug. 1993 (cited on page ii).

[3] Abraham Andreu and Qayyah Moynihan. 'Humans won't be able to control a superintelligent AI, according to a study'. In: *Business Insider* (Sept. 24, 2021). (Visited on 09/24/2021) (cited on pages iii, 4, 11).

[4] Péter Érdi. *Ranking: The Unwritten Rules of the Social Game We All Play*. Oxford, England, UK: Oxford University Press, Oct. 2019 (cited on page iii).

[5] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. 1st. USA: Oxford University Press, Inc., 2014 (cited on page iii).

[6] Brian Christian. *The Alignment Problem: Machine Learning and Human Values*. New York, NY, USA: W. W. Norton & Company, Oct. 2020 (cited on pages iii, 5).

[7] Reid Blackman. *Ethical Machines: Your Concise Guide to Totally Unbiased, Transparent, and Respectful AI*. Harvard Business Review Press, July 2022 (cited on pages iii, 5).

[8] Ross Douthat. 'Can We Resist the Age of the Algorithm?' In: *The New York Times* (July 30, 2022). (Visited on 07/30/2022) (cited on page 2).

[9] MacAskill2022. 'The Case for Longtermism'. In: *The New York Times* (Aug. 5, 2022). (Visited on 08/05/2021) (cited on page 2).

[10] Cade Metz. 'The Long Road to Driverless Trucks'. In: *N.Y. Times* (Sept. 2022) (cited on page 2).

[11] Cade Metz. 'Stuck on the Streets of San Francisco in a Driverless Car'. In: *N.Y. Times* (Sept. 2022) (cited on page 2).

[12] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY, USA: Crown, Sept. 2017 (cited on page 5).

[13] Caroline Criado Perez. *Invisible Women: Data Bias in a World Designed for Men*. New York, NY, USA: Abrams Press, Mar. 2019 (cited on page 5).

Notation

The next list describes several symbols that will be later used within the body of the document.

c	Speed of light in a vacuum inertial frame
h	Planck constant

Alphabetical Index