DLT Lecture 1 : Why deep learning theory?

There can be no doubt that deep learning is *useful*: large tech companies spend billions on R&D in this area and there are many clear examples (in image classification, speech recognition, machine translation, recommender systems to name a few) where these investments have had measurable impact on their revenues. However the argument I would like to make in this lecture is that the open problems in the *theory of deep learning* are with high probability the most important theoretical problems of our time, and the present-day utility of deep learning is not a sufficient argument to support such a bold claim.

The expected value of deep learning theory is a product of two factors (a) the probability that deep learning has an eventual impact comparable to general purpose technologies such as steam and electricity, and (b) the degree to which progress in deep learning *practice* requires progress in deep learning *theory*. The lecture is correspondingly broken into two parts.

(A) Deep Learning as General Purpose Technology

(A.1) From toys to engines
(A.2) The power-law era of deep learning
(A.3) Limitations of deep learning

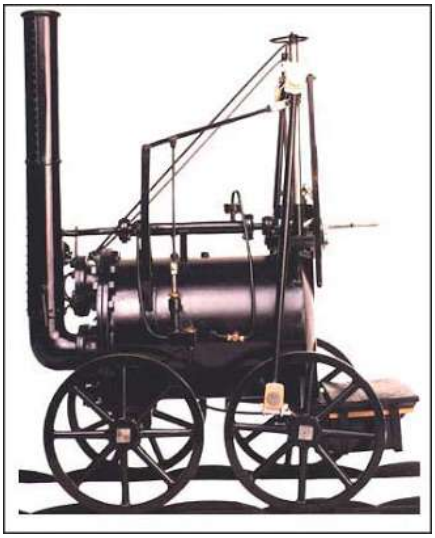
(B) Who needs theory?

(B.1) The role of theory, historically
(B.2) Mathematics and deep learning
(B.3) At scale, you can't afford to guess

(A) DL ∈ GPT

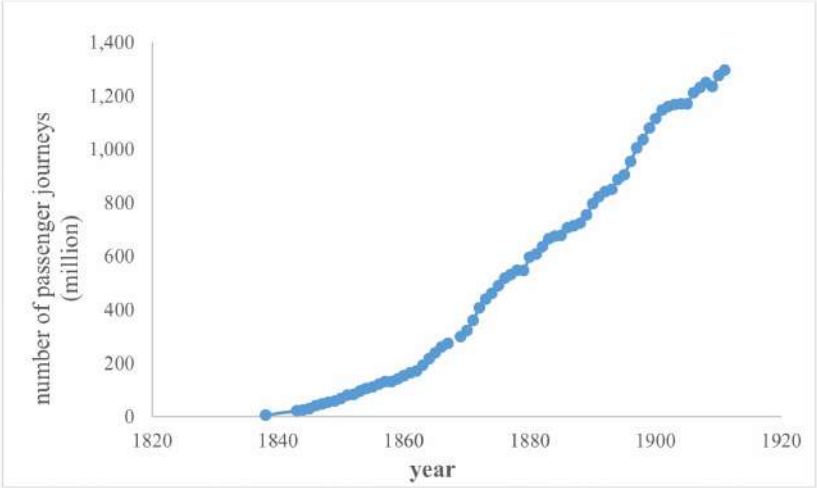
(A.1) From toys to engines

Richard Trevithick's "Puffing devil" is generally regarded as being one of the first steam powered locomotives. You can hear the amusing story of its first (and final) outing on Dave Broker's excellent podcast [1].



"Puffing devil" 1801

Figure 13: The number of railway passenger journeys in Britain, 1838-1911



The development of the railway network in Britain 1825-19111
Leigh Shaw-Taylor and Xuesheng You

If you were paying attention in 1801 you might have predicted that, absent some fundamental reason why steam engines would not scale to much higher pressures and volumes, economic forces (transportation of coal and manufactured goods, for example, but also passengers) would eventually propel this technology forward and transform Britain, and the world. On the other hand the "Puffing devil" was a bit of a toy, all sorts of ad hoc choices were involved in its construction, subsequent iterations relied on inspired guesswork and tinkering rather than some deep theory, and anyway the naysayers were right for decades! So you probably wouldn't have predicted the impact of steam locomotives (few did in 1801).

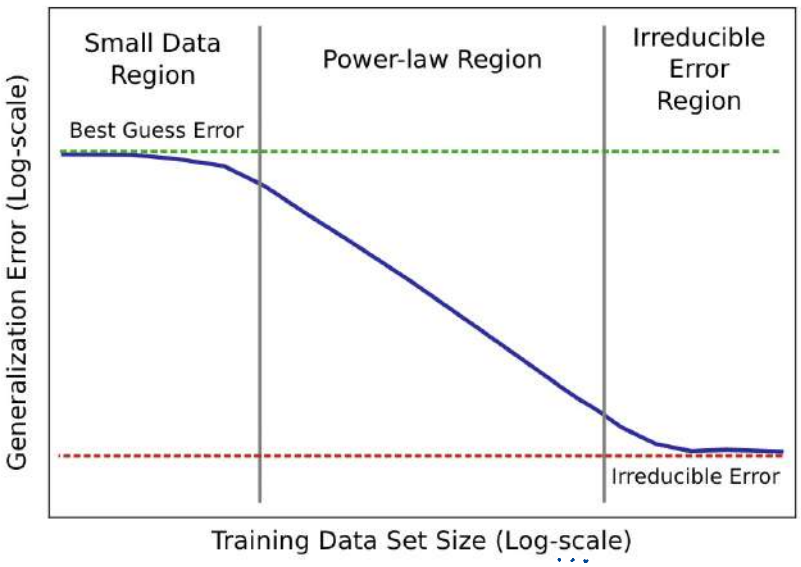
We now know that there were no a priori obstacles to scaling up the steam engine (much later, thermodynamics clarified this) and it became a general purpose technology.

A General Purpose Technology (GPT) is a technology that comes to be widely used across the economy, has many different uses, and creates many spillover effects. Obvious examples are steam, electricity, and computers. Of course not every promising technology scales up to become a GPT.

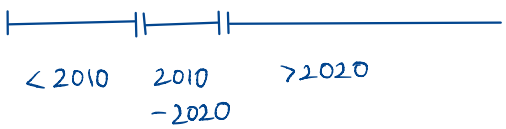
Deep learning is already useful and will see widespread adoption in many industries even if no further progress is made (i.e. we're past the "Puffing Devil" stage). But if we are reaching the limits of deep learning, it will fall short of the impact of GPTs like steam. Since the implications of very capable AI are so profound, it becomes important to understand if there are any *fundamental* obstacles to continuing progress in deep learning.

(A.2) The power-law era of deep learning

The study of artificial neural networks goes back in some form to the 1940s [2, §1.2], [3]. In the 2010s the availability of computation (GPUs) and large datasets (most famously ImageNet) started to reveal that neural networks had superior performance on a range of tasks in computer vision (and within a few years, in many other domains as well). In the late 2010s it started to become clear that for large networks trained on large datasets (large compared to ImageNet) the performance (as measured by generalisation error) is governed across a wide range of tasks (including generative modeling of language, images, video and math problems) by *power laws*, see [4, 5, 6]. The following figure (from [4]) illustrates the three "eras" of deep learning:



(4.1)



- Small data era (<2010) : scale too small to see wide advantage of neural networks
- Big data era (2010-2020) : sufficient scale to see performance on many tasks.
- Power-law era (>2020) : predictable returns to increased scale

What is a power law? An example from [5] is that Transformer language models of natural language trained with early stopping on a dataset of size D have test loss given by the following power law (here the model is as large as it needs to be in terms of the number of parameters, and models are trained "to convergence")

$$L(D) \propto D^{-0.095}$$

Hence if you want to decrease the loss by a factor of 5%,

$$L(\alpha D) = 0.95 L(D) \implies \alpha^{-0.095} = 0.95$$

$$\implies \log \alpha = -\frac{1}{0.095} \log(0.95) \implies \alpha \approx 2.72$$

you need to increase your dataset by a factor of around 3.

This power law holds across a large range of values of D [5, 8] exceeding some minimum threshold, but must eventually fail as the model hits the lower bound of irreducible error as in (4.1), if not before. Some remarks

(i) In some examples (8×8 ImageNet generative modeling) the power law does continue to hold until the level of irreducible error is reached [6, §3].

(ii) A 5% improvement may not sound impressive, but modeling a very large corpus of Internet text is a very difficult problem, and when L is low every percentage point of improvement may translate to impressive performance on "downstream" tasks (i.e. tasks for which the model may be fine-tuned), see [8] and [6, §3.4] for details.

(iii) Power law behaviour appears to be universal for Transformers across many modalities [6] and for L as a function of dataset size D, model size N and total compute C.

(iv) Recent work has shown that Transformers may be used for many tasks in computer vision, and it seems the power laws may also apply.

(v) Suppose $L(D) = \lambda D^{-\gamma}$ then

$$\log L(D) = \log \lambda - \gamma \log D \tag{5.1}$$

so you can think of γ as a measure of the marginal information extracted from each additional training example (logarithmically). It is easy to change λ by varying the model architecture, but much harder to increase the scaling exponent γ [4, §5.2].

All together the current situation in deep learning is quite remarkable: one architecture (Transformers, with minor variations) works across language, images, video and reinforcement learning, and shows power law behaviour in all but the last (it would be very interesting to see some kind of power law in RL, to say the least) with no indication of the power laws failing at the frontier of large compute (and GPT-3 is an impressive test in this direction).

This raises the possibility that we can make progress (perhaps radical progress) simply by expending more resources, without necessarily needing breakthroughs in architecture.

However the power laws are not only relevant at the frontier of large compute. As Hestness et al note "It may seem counterintuitive, but an implication of predictable scaling is that model architecture exploration should be feasible with small datasets" as you only need D to be large enough to enter the power law region of (4.1) in order to estimate the scaling exponent γ.

(A.3) Limitations of deep learning

As with any technology, deep learning has many limitations and there is a range of opinion about which of these limitations are fundamental (in the sense that the second law of thermodynamics places fundamental limitations on heat engines). Here is a brief discussion of some common criticisms, partly borrowing from [10].

- Deep learning is too data hungry: labeled datasets are fundamentally scarce, but autoregressive Transformer models work with un-labeled data, and may then be fine-tuned on labeled data. There are many domains in which this probably still doesn't work, but arguably enough domains in which it will to make deep learning a GPT nonetheless.

- Scaling laws are nice, but scaling is infeasible to continue pushing on the power laws will require substantial investment (millions, and eventually perhaps billions of USD) in datasets, GPUs, and supercomputers for distributed training. This seems implausible and even offensive to some academics, but it is important to remember that we live in a globalised industrialised capitalist society in which this scale of investment is routine (global investment in renewable energy capacity in 2018 was USD$272.9 billion). If scaling delivers useful artifacts, the resources can be found.

  The investment of Microsoft in building an AI supercomputer for OpenAI [12] and the applications of this technology that already exist across Microsoft's business tends to suggest that so far, scaling is delivering sufficiently.

  Finally, note that once a rich actor (like Microsoft or Google) demonstrates by scaling that some capability is possible in deep learning, it is likely to be replicated (in more complex, handcrafted models) at lower cost for use in particular settings; for a concrete example see [14].

- Large models will cook the planet there is some concern that training very large models is associated with wasteful greenhouse gas emissions. As far as I can tell this concern has no basis, since we are already in a rapid transition to Solar-Wind-Battery power [13] and in this world datacenters can be co-located with generation sites for low-cost clean power.

So will deep learning evolve into a general purpose technology over the next several decades? Only time will tell, but my personal opinion is that after the discovery of the power laws, it is more likely than not.

Aside on AI safety

Note that deep learning evolving into a GPT is a necessary but not sufficient condition for "human level" artificial intelligence, which is a much more radical and transformative prospect. This is not the place to get too far into that topic, apart from some brief notes:

- Even if human-level AI is not arriving any time soon, deep learning and deep learning theory are still important for the reasons outlined above.

- Public claims that human-level AI is "far off" and that many fundamental advances beyond deep learning will be required should be treated as pseudo-scientific, as in, they are beliefs mistakenly regarded as being based on scientific method [11]. There are no models for AI progress and no guarantees radical near-term progress is impossible. That doesn't mean it will happen, but the uncertainty is large and the event so consequential that topics like AI safety are urgent [16, 17].

- "History shows that for the general public, and even for scientists not in a key inner circle, and even for scientists in that key circle, it is very often the case that key technological developments still seem decades away, five years before they show up" — E. Yudkowsky [15].

B Who needs theory?

B.1 The role of theory, historically

The engineering of steam engines and the science of thermodynamics developed simultaneously in the 19th century, and in many cases the people involved were both engineers and scientists. But there is no doubt that a comprehensive theory of heat engines came well *after* the basic principles had been discovered empirically. Similarly, the first telescopes came well before a theory of optics. Nonetheless theory becomes necessary for a technology to reach its true potential:

- Theory reduces the number of experiments certain ideas for heat engines or telescopes will never work, and a theory may tell you cheaply which ones. You can invent a simple telescope by trial and error, but without a theory this process becomes exponentially harder as you design more complex and capable artifacts.

- Theory gives you courage if the theory says an engine with twice the volume should work, you'll be more likely to work hard to overcome practical obstacles (e.g. with materials and welding) as they arise. This point is made clearly in the introduction to [4]. For example OpenAI credits in [6] the discovery of neural scaling laws for giving them the confidence to spend the money necessary to train GPT-3 (estimated to be millions of dollars). So far these laws do not have a rigorous basis, but they still count as "theory" by any reasonable definition.

(B.2) Mathematics and deep learning

If one comes to the field as a mathematician it is easy to feel overwhelmed by the abundance of ad hoc decisions (with e.g. training schedules or normalisations) and the "just so" stories justifying them, and to despair about there being anything mathematically interesting to say about the subject. But the same could have been said about steam engines. It is becoming increasingly clear that many architectural choices "merely" change the constant prefactor in power law scaling relations. These choices may be very important if you are a startup or Google, but mathematicians can perhaps ignore them, to focus on the relations between the "macroscopic thermodynamic variables" D, N, C which (if we can isolate and understand them) may dictate the ultimate limits on this class of learning machines [5, §8]

If you are put off by how "messy" deep learning theory appeas in 2020, you should go and take a look at the state of thermodynamics in 1820! The really hard and important work is often in the trenches, and it isn't all rainbows down there.

(B.3) At scale you can't afford to guess

If we had a rigorous "microscopic" theory which explained the observed power law behaviour in Transformer models, we might be able to predict where this scaling will break down, whether it will hold for a given architecture and with what scaling exponent, among other properties. A theory might also inform other interrelated engineering issues (the railway age required not only steam engines but steel tracks, which in turn had to wait for chemistry to mature). As deep learning matures from "hand craft" to industrial scale engineering, mathematics has an important role to play in making this technology predictable.

Conclusion

Power laws are common across physics and they don't arise by accident. The discovery of power law behaviour is strong evidence that deep learning is "onto something" even if the deep principles underlying these laws remain obscure. We are likely to see exponentially increasing investment in deep learning on the back of these power laws, and so any increase in understanding of the fundamental mathematics underlying them will have extreme leverage, translating into more efficient use of capital and human research and development effort.

This is also a completely new area of mathematics, in which beautiful theorems are sure to be hiding. So that's my pitch for why theory of deep learning is one of the most impactful things you could spend your time on.

Open questions

- A plausible definition of "reasoning" is that it is the ability to use additional computation to extract more marginal information from each additional training example ("to see a world in a grain of sand..."). Is there a formal relationship between reasoning, logic and scaling exponents?

- Is it possible to predict where the transition to power-law scaling in (4.1) occurs?

References

[1] D. Broker "The Industrial Revolutions podcast" Chapter 30
https://industrialrevolutionspod.com/episodes/2019/9/24/chapter-30-the-locomotive

[2] I. Goodfellow, Y. Bengio, A. Courville "Deep learning" MIT press, 2016.

[3] J. Schmidhuber "Deep learning in neural networks : an overview" arXiv: 1404.7828, 2018.

[4] J. Hestness et al "Deep learning scaling is predictable, empirically" arXiv: 1712.00409, 2017.

[5] J. Kaplan, S. McCandlish, "Scaling laws for neural language models" arXiv: 2001.08361, 2020.

[6] T. Henighan, J. Kaplan, M. Katz "Scaling laws for autoregressive generative modeling" arXiv: 2010.14701, 2020.

[7] W. Rosen "The most powerful idea in the world" Random House, 2010.

[8] T. B. Brown, B. Mann, N. Ryder, M. Subbiah et al "Language models are few-shot learners" arXiv: 2005.14165, 2020. "the GPT3 paper"

[9] J. Kaplan talk "Neural scaling laws and GPT-3" YouTube Oct, 2020.

[10] G. Marcus "Deep learning – a critical appraisal" 2018.

[11] Gwern "On GPT-3: meta-learning, scaling, implications and deep theory" May 2020.
https://www.gwern.net/newsletter/2020/05#gpt-3

[12] J. Langston "Microsoft's AI supercomputer" May 2020
https://blogs.microsoft.com/ai/openai-azure-supercomputer/

[13] A. Dorr and T. Seba "Rethinking energy 2020-2030" RethinkX, 2020.

[14] S. Mandava, S. Migacz, A. F. Florea "Pay Attention when Required"
arXiv: 2009.04534.

[15] E. Yudkowsky "There's no fire alarm for Artificial General Intelligence"
MIRI analysis https://intelligence.org/2017/10/13/fire-alarm/

[16] N. Bostrom "Superintelligence: paths, dangers, strategies" Oxford University Press 2014.

[17] Beneficial AI conference 2017 https://futureoflife.org/bai-2017/