How to Be an Academic Machine Learning Researcher in the Era of Scale

Colin Raffel
From “Real-Time Social Media Analytics with Deep Transformer Language Models: A Big Data Approach” by Ahmet and Abdullah
Performance on JSON tasks

Effective parameter count

From “Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models” by Srivastava et al.
ImageNet finetune error rate [%] vs. Compute (TPUv3 core days) for various model sizes. The error rates decrease as the compute increases, following the equation:

\[ E = 0.09 + 0.26(C + 0.01)^{-0.35} \]

From “Scaling Vision Transformers” by Zhai et al.
From “Deep Learning Scaling is Predictable, Empirically” by Hestness et al.
From “Scaling Laws for Neural Language Models” by Kaplan et al.
The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant ... but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available.

From “The Bitter Lesson” by Sutton
2018

ELMo
93.6M parameters

V100
16GB memory

2022

PaLM
540B parameters

H100
80GB memory

5000×

5×
\[ \hat{y}_i = f_\theta(x_i) \]

\[ \text{...Wh...} \]

\[ \partial \theta = \sum_{i=1}^{N} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i) \]

\[ \theta \leftarrow \theta + \text{optimizer}(\partial \theta) \]
\[ \hat{y}_i = f_\theta(x_i) \]

\[ \ldots W h \ldots \]

\[ \partial \theta = \sum_{i=1}^{N} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i) \]

\[ \theta \leftarrow \theta + \text{optimizer}(\partial \theta) \]
\[ \hat{y}_i = f_\theta(x_i) \]
\[ \cdots \underbrace{W h \cdots} \]
\[ \partial \theta = \sum_{i=1}^{N} \nabla_\theta L(\hat{y}_i, y_i) \]
\[ \theta \leftarrow \theta + \text{optimizer}(\partial \theta) \]
$x_1, \ldots, x_N \xrightarrow{} f_\theta(x) \xrightarrow{} \sum_{i=1}^{N} \nabla_\theta \mathcal{L} (\hat{y}_i, y_i) \xrightarrow{} \partial \theta$
Device 1

\[ x_1, \ldots, x_{\frac{N}{2}} \rightarrow f_\theta(x) \rightarrow \sum_{i=1}^{\frac{N}{2}} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i) \rightarrow \partial \theta \]

Device 2

\[ x_{\frac{N}{2}+1}, \ldots, x_N \rightarrow f_\theta(x) \rightarrow \sum_{i=\frac{N}{2}+1}^{N} \nabla_\theta \mathcal{L}(\hat{y}_i, y_i) \]
Device 1

Device 2
From “PipeDream: Generalized Pipeline Parallelism for DNN Training” by Narayanan et al.
PipeDream: Generalized Pipeline Parallelism for DNN Training by Narayanan et al.
From “Efficient Large-Scale Language Model Training on GPU Clusters” by Narayanan et al.
**float32**
- range: ~$1e^{-38}$ to $3e^{38}$
- 8 bits for exponent (1 bit for sign)
- 23 bits for fraction

**float16**
- range: ~$5.9e^{-8}$ to $6.5e^{4}$
- 5 bits for exponent (1 bit for sign)
- 10 bits for fraction

**bfloat16**
- range: ~$1e^{-38}$ to $3e^{38}$
- 8 bits for exponent (1 bit for sign)
- 7 bits for fraction

From [https://cloud.google.com/tpu/docs/bfloat16](https://cloud.google.com/tpu/docs/bfloat16)
From “ZeRO-Offload: Democratizing Billion-Scale Model Training” by Micikevicius et al.
From https://medium.com/tensorflow/fitting-larger-networks-into-memory-583e3c758ff9
From "Scaling Laws for Neural Language Models" by Kaplan et al.

- Minimum serial steps increases negligibly
- <10x Serial Steps
- 100x Batch Size
- >1,000,000x Model Size
- Data requirements grow relatively slowly
- Optimal model size increases very quickly
The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant ... but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available.
The biggest lesson that can be read from 70 years of AI research is that **general methods that leverage computation are ultimately the most effective, and by a large margin.** The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant ... but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available.

From “The Bitter Lesson” by Sutton
The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant ... but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available.

→ At any point in time, it is likely more effective to be clever! (The Bitter Corollary?)
The Sweet Lesson:

It is often possible to outperform scaled-up methods by being more clever, and being clever can yield methods that scale better.
A C++ implementation of deep LSTM with the configuration from the previous section on a single GPU processes a speed of approximately 1,700 words per second. This was too slow for our purposes, so we parallelized our model using an 8-GPU machine.
From “Neural Machine Translation by Jointly Learning to Align and Translate” by Bahdanau et al.
From “A Structured Self-Attentive Sentence Embedding” by Lin et al.
From “Attention is All You Need” by Vaswani et al.
Closed-book question answering
http://www.autosweblog.com/cat/trivia-questions-from-the-50s
  who was frank sinatra? a: an american singer, actor, and producer.

Paraphrase identification
https://www.usingenglish.com/forum/threads/60200-Do-these-sentences-mean-the-same
  Do these sentences mean the same? No other boy in this class is as smart as the boy. No other boy is as smart as the boy in this class.

Natural Language Inference
https://ell.stackexchange.com/questions/121446/what-does-this-sentence-imply
  If I say: He has worked there for 3 years. does this imply that he is still working at the moment of speaking?

Summarization
https://blog.nytsoi.net/tag/reddit
  ... Lately I've been seeing a pattern regarding videos stolen from other YouTube channels, reuploaded and monetized with ads. These videos are then mass posted on Reddit by bots masquerading as real users. tl;dr: Spambots are posting links to stolen videos on Reddit, copying comments from others to masquerade as legitimate users.

Pronoun resolution
  Jennifer is a vegetarian, so she will order a nonmeat entrée. In this example, the pronoun she is used to refer to Jennifer.
From "Multitask Prompted Training Enables Zero-Shot Task Generalization" by Sanh et al.
Natural Language Inference

RTE

CB

ANLI R1

ANLI R2

ANLI R3

Coreference Resolution

WSC

Winogrande

COPA

Sentence Completion

StoryCloze

HellaSwag

Word Sense

WiC

GPT-3 (6.7B)

GPT-3 (13B)

GPT-3 (175B)

T5+LM (11B)

T0 (11B)

From “Multitask Prompted Training Enables Zero-Shot Task Generalization” by Sanh et al.
From “Training language models to follow instructions with human feedback” by Ouyang et al.
TriviaQA performance

From “Language Models are Few-Shot Learners” by Brown et al.
From "Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning", Liu et al. 2022
The figure shows a plot of Accuracy vs. FLOPs per example. The methods compared are T-Few, Human baseline, PET, SetFit, and GPT-3. T-Few is indicated by a yellow star, while the others are represented by different symbols. The table below lists the top-5 best methods on RAFT as of writing:

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-Few</td>
<td>75.8%</td>
</tr>
<tr>
<td>Human baseline [2]</td>
<td>73.5%</td>
</tr>
<tr>
<td>PET [50]</td>
<td>69.6%</td>
</tr>
<tr>
<td>SetFit [51]</td>
<td>66.9%</td>
</tr>
<tr>
<td>GPT-3 [4]</td>
<td>62.7%</td>
</tr>
</tbody>
</table>

Table 2: Top-5 best methods on RAFT as of writing. T-Few is the first method to outperform the human baseline and achieves over 6% higher accuracy than the next-best method.

From "Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning", Liu et al. 2022
Thanks.

Please give me feedback: