A distillation of “Scaling Laws for Transfer” through the lens of alignment research automation

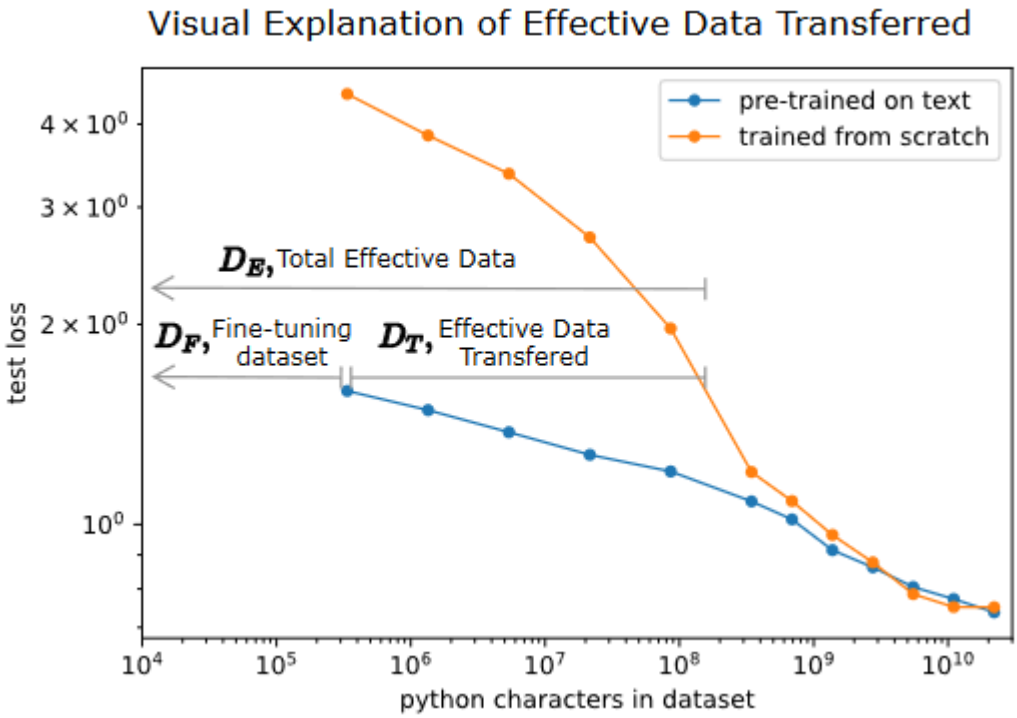
A great deal of power of modern AI systems lies not only in exceptional pattern matching and performance on single tasks, but the breezy generality to which they perform on different tasks, often ones the AI has never seen before. The most common form of that seems to be transfer learning, which Goodfellow et al. (2016)[1] describes as having a model learn multiple tasks at the same time in series, where the tasks all share some same basic causations or relatedness. A structural advantage ensues: if the more “common” tasks (or tasks with much more data) are trained before the rarer tasks, the representations learned from the commoner tasks can boost performance on the data-sparse tasks. Whatever level of AI, or superintelligence, will be created, it will always have to contend with great imbalances in quantities of training material across tasks in the world (like identifying dogs vs identifying manatees), and correspondingly every attempt at automating AI alignment will have to consider the rate of transfer learning to some extent in crafting automation strategies, or deciding whether they are even worthwhile. Hernandez et al, in their paper “Scaling Laws for Transfer” (2021)[2], seems to fill a niche in that category.

“Scaling Laws for Transfer” makes two great contributions: when it comes to LLMs doing text-and-Python transfer-learning on low-data regimes in particular, it formulates a new measure of transfer-learning data efficiency, and uses it to show that effective data transfer follows a power law with respect to parameter count and fine-tuning dataset size. The authors then deduce that “the exponents in those power laws correspond to measures of the generality of a model [for the number of parameters] and proximity of distributions [for the fine-tuning dataset size]”.

The particular experimental design involved learning python as the task (and hold-out dataset). For the transfer learning “template” there is naturally far more data that humans have created in the form of natural language than python, so the model’s pre-trained natural language skills would seem to be transferred to the new domain of python coding. Indeed, this is precisely the curricula the researchers set up: the first, training on scratch from python, the second, pretraining on natural language before finetuning on python, and lastly, pretraining on “an equal mixture of natural language and non-python code” before finetuning on python. Each curriculum is run, in parallel, on the two independent variables of network size and size of fine-tuning-dataset. Then, for each triplet of (curriculum, network size, fine-tuning dataset size), a measure of transfer-learning sample efficiency is collected, before all data is sorted by curriculum and that measure is regressed with respect to the two sizes. But for the last variable, how exactly did the authors determine the “low-data” dataset size? And what exactly is that “measure”?

Both are of course intertwined. For the first, one would demand a cutoff ratio of some size of data-sparse dataset to some size of data-rich ones. But to specify both sizes, the authors were motivated by the one key metric of sample efficiency with respect to the task itself, justified in their introduction as “an important way to characterize data [because] neural networks often require more direct experience than a person can consume in a lifetime”. Sample efficiency answers not the rhetorical question, “if I specified a proportion of python finetuning to be done [i.e. input], what would be the change in task-transfer-performance between that scheme and a pure python training,” but instead, “if I specified a level of performance at the pure python task which all my models must meet [i.e. output], how much python data would need to be fine-tuned on for a mixed model for its performance level to match a pure-python one?”

[1] https://www.deeplearningbook.org/
[2] https://arxiv.org/abs/2102.01293



In the above (from the original), we calculate everything on a horizontal line of specified test loss, rather than a vertical line of specified dataset-size. In essence, sampling efficiency as studied measures input efficiency, not output productivity. Thus the answer to the second question is a function of the model size that's defined as "D(N)...the amount of data it takes to reach 99% of the performance (1/0.99 times the loss) that infinite python data would yield, for a given model size [N]." Once that's firm, the answer to the first question can be defined as any fine-tuned dataset size less than 10% of D(N), model size constant.

Ultimately that's all just backend stuff: what more can we say about the frontend, the way the experimental data is presented to conclude? The researchers plotted the fraction of effective data from transfer (the previous "sample efficiency measure" divided by the pure-python-scheme's total data use) simultaneously with parameter count N and dataset size D_F, to get

$$D_T = effective\ data\ transferred = k(D_F)^{\alpha}(N)^{\beta}$$

With regressed experimental values α = 0.18 and β = 0.38.

What implications would "Scaling Laws for Transfer" really have for alignment automation?

To answer this it would be necessary to go over the major cruxes in the field. In his essay "Beliefs and Disagreements about Automating Alignment Research"[3] (Aug 2022), Ian McKenzie, a scientist at OpenAI's Dangerous Capabilities Evaluations and former FAR AI Inverse Scaling Prize head, identifies four:

1) "Generative models vs agents": can we steer baby AGIs away from pursuing coherent goals that persist across inputs?
2) "Timing of emergence of deception vs intelligence": are we more likely to get systems that pass the milestone of alignment-automation usefulness before we get misaligned systems, or after?
3) "The 'hardness' of generating alignment insights": on the continuum from cleverly mashing online content to having a full model of human psychology and neuroscience, what level of complexity does an automated helper really need, to do "useful, original thinking for us?"
4) How much time can automated-helpers really save on the broader alignment effort, when at "Level 1" (merely helping humans work faster without producing original contributions)?

Stephen McAleese in his survey[4] some twelve months later reaffirmed crux three, and implicitly confirmed crux two as "the difficulty of aligning [the automated researcher] compared to a full AGI".

[3] https://www.alignmentforum.org/posts/JKgGvJCzNoBQss2bq/beliefs-and-disagreements-about-automating-alignment
[4] https://www.alignmentforum.org/posts/zj7rjpAfuADkr7sqd/could-we-automate-ai-alignment-research-1

We digress for a moment and ask how good the "Scaling Laws for Transfer" results are in the context of previous *transfer learning* work. With a great debt to EpochAI's literature review[5], two other papers include:

1) Mikami et al. (2021)[6]: a set of real pre-training images gets randomly synthesized in many ways to provide a training set of synthetic images, on whose skills get transferred to real vision tasks. Their scaling law was $test\ error \simeq DN^{-\alpha} + C$, where D and N retain their definitions, and α describes the "convergence speed of pre-training", and C the lower limit of the error. This seems to be a better bound exponentially, but this study only tested ResNets, not transformers, making it far less future-extrapolatable.
2) Abnar et al. (2021)[7]: somewhat of a metastudy, upon "more than 4800 experiments on Vision Transformers, MLP-Mixers and ResNets" but still using only image dataset and tasks. Concludes that overoptimizing for upstream (pretraining-task) accuracy can often saturate, and sometimes actively hurt, downstream (fine-tuning-task) performance. Shows "that the saturation behavior we observe is closely related to the way that representations evolve through the layers of the models," and in particular, downstream performance saturation seems to be caused by the latest pretraining network layers not being fine enough to model those features important for downstream performance. Therefore Abnar et al. acts as a limiting, boundary case for Hernandez et al. and doesn't dispute it directly.

But regardless of how Hernandez et al. update on that previous work, does it even matter in these cruxes?
The first and second cruxes rely on finer-grained design choices just as much as choices of parameter and dataset sizes to shape the particular direction of an AGI's goals relative to human needs, so "Scaling Laws for Transfer" can update positions on them less directly. The third crux seems more promising. Suppose you had a task that would shape the future of humanity, but didn't have much data on (for example, social modeling for bioterrorist attackers), though lots of generic data for pretraining. What theories of change can Hernandez et al's laws, or even papers like it on different domains and architectures, actually create? If one knows the values of the powers of D and N, or even knows that the values relate to D and N exponentially, far more precise and accurate calculations can be directly made about how to allocate resources between increasing N (parameter count: more resources in computation) and increasing D (finetuning dataset size: more resources in gathering often hard-to-gather data) to maximize sample-efficiency with minimum resources. Here's the kicker: an AI might need a less-complex model of humans to achieve its goals if it is more sample efficient, and simultaneously a more-complex model if less, so the results of Hernandez et al can plausibly change arguments within crux 3. This resource-allocation-optimizing theory of change can also make people update their positions on crux 4, for it specifically deals with researchers' time allocation on any task they decide to spend time on. (Though further research has to be done on what role differential technology development[8] can combine with this theory of change to increase capabilities too much relative to alignment, and possibly slow down alignmentists' overall progress.)

Finally, we may simultaneously consider John Wentworth's 2023 dissection[9] of alignment automation: the human client's own understanding of the task is the true greatest bottleneck to outsourcing any work effectively, and therefore all automation efforts, merely because humans cannot know what a good solution looks like (from any entity they choose to hire or maintain) unless they know the subject at a similarly high enough level as the entity, with anything else leading to catastrophe at high enough capabilities. Taking this nonempirical argument for all its weight means that the only path to impact from the [directest application of] Hernandez et al's laws is that Level 1 work can be sped up, if an equal sample-efficiency performance on data-sparse tasks can be gained through having a better idea of allocating resources between increasing N (parameter count, which requires more time spent in computation) vs increasing D (fine-tuning dataset, which requires more time spent gathering data from often rare or dangerous processes). However at higher intelligence levels this theory of change might become unaligned because creating a good enough loss function sometimes relies (even subtly) on humans knowing what a good solution looks like, cuing Wentworth's failure mode, and sample efficiency performance in Hernandez et al is tied to loss.

[5] https://epochai.org/blog/scaling-laws-literature-review
[6] https://arxiv.org/pdf/2108.11018.pdf
[7] https://arxiv.org/pdf/2110.02095.pdf
[8] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4213670
[9] https://www.lesswrong.com/posts/3gAccKDW6nRKFumpP/why-not-just-outsource-alignment-research-to-an-ai