LLM4ALL: Low-cost LLMs

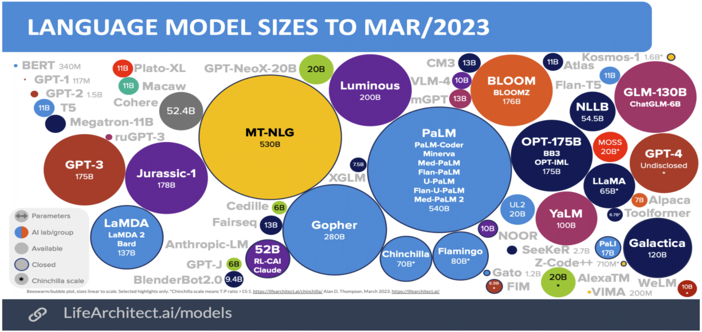
LIX, DaSciM

October 2024

Table of Contents

Table of Contents

Challenges for LLMs



LLMs size grows

Challenges for LLMs

- Inference on BLOOM-176B, need 8x80GB A100 GPUs ($15k each).
- Fine-tune BLOOM-176B, need 72 of these GPUs.
- Llama is trained on more than 16k H100 GPUs.
- We need to reduce these requirements while preserving the model's performance.

LLM Training Costs on MosaicML Cloud

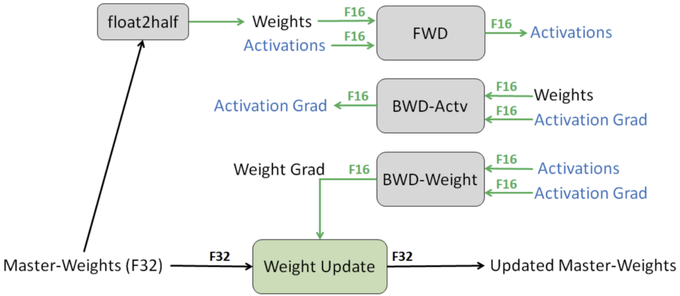
Model	Billions of Tokens (Compute-optimal)	Days to Train on MosaicML Cloud	Approx. Cost on MosaicML Cloud
GPT-1.3B	26B	0.14	$2,000
GPT-2.7B	54B	0.48	$6,000
GPT-6.7B	134B	2.32	$30,000
GPT-13B	260B	7.43	$100,000
GPT-30B *	**610B**	**35.98**	**$450,000**
GPT-70B **	1400B	176.55	$2,500,000

[0] image source: https://www.databricks.com/blog/gpt-3-quality-for-500k

Table of Contents

Mixed Precision Training[1]

- Training uses 16-bit precision for most operations.
- Critical operations, such as the accumulation of gradients, are still performed in 32-bit precision.



[1] Micikevicius, Paulius, et al. "Mixed precision training." arXiv preprint arXiv:1710.03740 (2017).

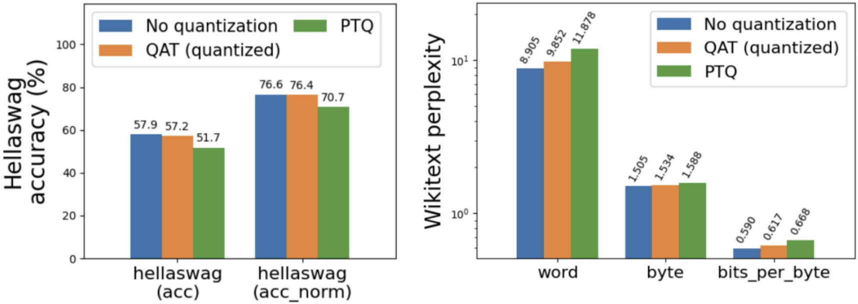
Mixed Precision Training

- Mixed precision training has proven to be a highly effective approach for deep learning, achieving up to 8x faster computation times without sacrificing model accuracy.

Model	Baseline	Mixed Precision	Reference
AlexNet	56.77%	56.93%	(Krizhevsky et al., 2012)
VGG-D	65.40%	65.43%	(Simonyan and Zisserman, 2014)
GoogLeNet (Inception v1)	68.33%	68.43%	(Szegedy et al., 2015)
Inception v2	70.03%	70.02%	(Ioffe and Szegedy, 2015)
Inception v3	73.85%	74.13%	(Szegedy et al., 2016)
Resnet50	75.92%	76.04%	(He et al., 2016b)

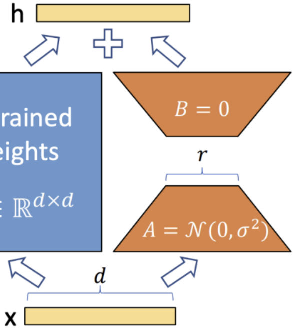
Quantization

- Quantization-Aware Training: the model is trained with simulated quantization effects, allowing it to adapt to the lower precision during training itself.
- Post-Training Quantization: it converts the trained model's floating-point weights and activations to lower-precision integer formats, such as 8-bit integers.



Low-Rank Factorization

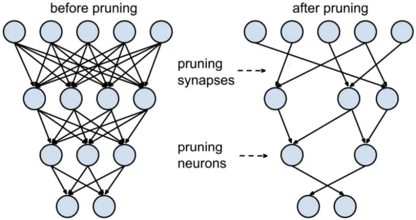
- It employs matrix decomposition to to factorize large, dense weight matrices into smaller, more manageable components.
- Eg: LoRA[2], only these low-rank matrices are updated, while the original weights remain frozen.



[2] Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." arXiv preprint arXiv:2106.09685 (2021).

Pruning [3]

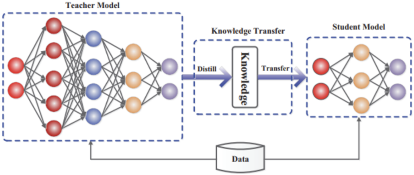
- It eliminates parameters that contribute the least to the model's output.
- It can be combined with other compression techniques like quantization and low-rank factorization.



[3] Molchanov, Pavlo, et al. "Pruning convolutional neural networks for resource efficient inference." arXiv preprint arXiv:1611.06440 (2016).

Knowledge Distillation [4]

- A small model (the student) is trained to mimic the predictions of a much larger pre-trained model (the teacher)
- In distillation, knowledge is transferred from teacher model to the student by minimizing a loss function



[4] Hinton, Geoffrey. "Distilling the Knowledge in a Neural Network." arXiv preprint arXiv:1503.02531 (2015).

Knowledge Distillation

- Faster Inference and Lower Latency. Distilled models allow for quick decision-making, enhancing user experience.
- Distilled models often generalize better to unseen data due to the regularization effect of distillation.
- Improved Performance on Small Devices with limited computational resources. Knowledge distillation enables IoT devices to perform complex tasks.

Table of Contents

Background

- LLM pre-training is the most data-, compute-intensive task.
- Power-law acts like a soft limit on model quality, it's expensive to improve performance by scaling up the data/model.
- On vision pretraining, it's shown high-quality data leads to better performance.

Question:

- Can we go beyond the scaling law using efficient data training?
- Can we find an optimal way of using our data?

Measuring the variability of natural language data - diversity coefficient[5].

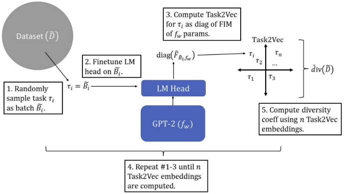


Figure 1: **The process of computing the diversity coefficient for a dataset proceeds through three main stages:** (a) randomly sampling batches of text from the dataset, (b) computing the Task2Vec embeddings for each sampled batch, and (c) calculating the expected pairwise cosine distance between the Task2Vec embeddings of the sampled data.

- Task2Vec embedding of text data represents which parameters of the probe network are most important.

[5] Lee, Alycia, et al. "Beyond scale: the diversity coefficient as a data quality metric demonstrates llms are pre-trained on formally diverse data."

D4 sampler[6]

The D4 sampler chains MinHash deduplication, SemDeDup, and SSL prototypes together to prune both high-variance, sparse regions and prototypical, dense regions of LLM pre-training datasets

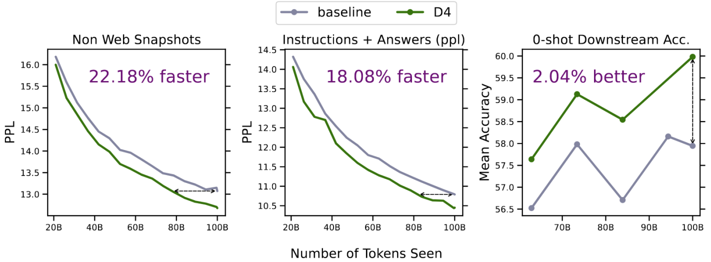


Figure 1: Learning curves for 6.7B OPT model pretraining on 100B tokens, with data selected with D4 (pink line) and randomly (gray line). D4 significantly outperforms baseline training, getting between 18-20% efficiency gains on validation perplexity and 2% increase in average 0-shot downstream accuracy across 16 NLP tasks. See Section A.2 for full learning curves.

[6] Tirumala, Kushal, et al. "D4: Improving llm pretraining via document de-duplication and diversification."

ASK-LLM[7]

We take the softmax probability of the token "yes" as the estimated data-quality score.

Ask-LLM prompt

###
This is a pretraining .... datapoint.
###

Does the previous paragraph demarcated within ### and ### contain informative signal for pre-training a large-language model? An informative datapoint should be well-formatted, contain some usable knowledge of the world, and strictly NOT have any harmful, racist, sexist, etc. content.

OPTIONS:
- yes
- no

Sampling score = P("yes" | prompt)

Figure 3. The prompt for obtaining the sampling score for each training sample in ASK-LLM.

[7] Sachdeva, Noveen, et al. "How to Train Data-Efficient LLMs."

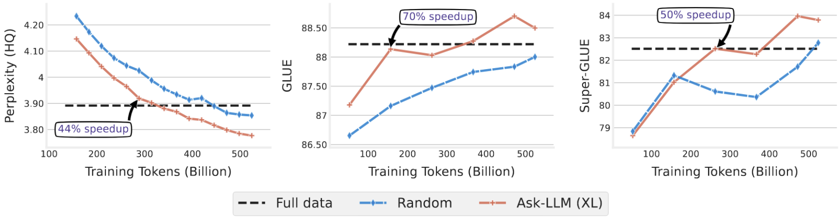


Figure 1. Data-efficient pre-training run of T5-Large (800M) using ASK-LLM with Flan-T5-XL as the data quality scorer. Training on 60% of the original dataset, ASK-LLM is able to train T5-Large both better and 70% faster, compared to training on 100% of the dataset.

DENSITY Sampling

- High-probability regions contain “prototypical” examples—ones with many near-duplicates and strong representation in the dataset.
- Low-probability regions will contain outliers, noise, and unique/rare inputs.
- We should boost the signal from under-represented portions of the input domain and downsample redundant, high-density information.

Impact of Data Age[8]

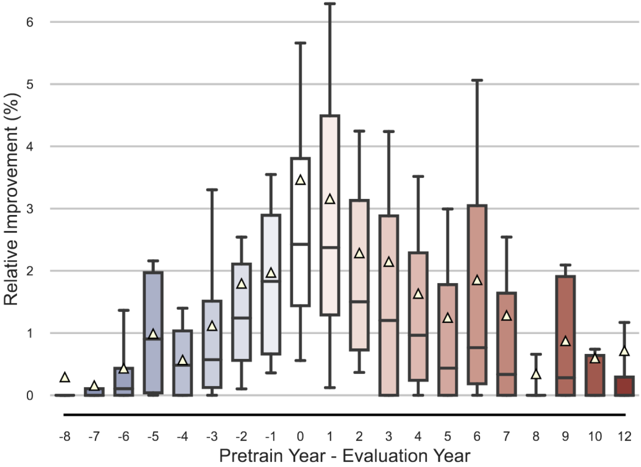
- We see performance degradation if evaluation data is either before or after pretraining data collection, and this deficit isn't overcome with substantial finetuning.

Model	Represented Domains (%)						Pile	C4	M-L	Filters		Data	
	Wiki	Web	Books	Dialog	Code	Acad				Tox	Qual	Pub	Year
BERT	76		24				✗	✗			H	Part	2018
GPT-2		100					✗	✗			H	Part	2019
RoBERTa	7	90	3				✗	✔			H	Part	2019
XLNet	8	89	3				✗	✔			H	Part	2019
T5	<1	99					✗	✔		H	H	✔	2019
GPT-3	3	82	16				✗	✔	7%		C	✗	2021
GPT-J/Neo	1.5	38	15	4.5	13	28	✔	Part			C	✔	2020
GLaM	6	46	20	28			✗	✔			C	✗	2021
LaMDA	13	24		50	13		✔	✔	10%	C	C	✗	2021
AlphaCode					100		✗	✗			H	✗	2021
CodeGen	1	24	10	3	40	22	✔	Part			H	Part	2020
Chinchilla	1	65	10		4		✔	✔		H	C	✗	2021
Minerva	<1	1.5	<1	2.5	<1	95	✔	✔	<1%		C	✗	2022
BLOOM	5	60	10	5	10	10	✔	✔	71%	H	C	Part	2021
PaLM	4	28	13	50	5		✗	✔	22%		C	✗	2021
Galactica	1	7	1		7	84	✔	Part			H	Part	2022
LLAMA	4.5	82	4.5	2	4.5	2.5	Part	✔	4%		C	Part	2020

[8] Longpre, Shayne, et al. "A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity."

Impact of Data Age

- The effects of pretraining misalignment are stronger for larger models than smaller models.



Impact of Data Age

- Current practice includes augmenting prompts with retrieved, recent data(RAG) to help overcome stale pretraining data.
- RAG database creation is an important research issue.
- Design more advanced fine-tuning technique for model update.

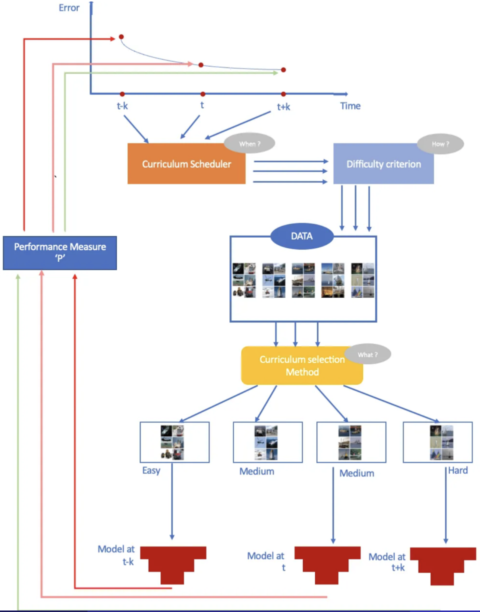
Our objective

- Develop new method based on information density for evaluating data quality.
- Explore impact of data ordering/mixture, combined with curriculum learning.

Curriculum learning

- Curriculum learning is a technique in machine learning in which a model is trained on examples of increasing difficulty.
- This is intended to attain good performance more quickly, or to converge to a better local optimum.

Curriculum learning



Information density/ Data difficulty

There are several ways to define information density.

- Entropy based:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

  Combined with compression technique for better estimation of information density.

- Lexical diversity based: **TTR, vocd-D, HD-D, MTLD**

Information density/ Data difficulty

Readability:

- Flesch Reading Ease

$$206.835 - (1.015 \times \text{ave sentence length}) - (84.6 \times \text{ave syllables per word})$$

- Flesch–Kincaid Grade Level:

$$0.39 \times \text{avrae sentence length} + 11.8 \times \text{average syllables per word} - 15.59$$

Future Plan

- Based on the information density metrics, and curriculum learning, find the best way to expose data to the model during training.
- Study the impact of mixture of data with respect to model training. Eg: easy/hard sample, English/French sample.
- Evaluate on model convergence.
- Multi-modal LMs with graph data.