Language Modeling Is Compression

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne , Elliot Catt , Tim Genewein , Christopher Mattern , Jordi Grau-Moya , Li Kevin Wenliang , Matthew Aitchison, Laurent Orseau, Marcus Hutter and Joel Veness

Equal contributions: Google DeepMind, Meta AI & Inria

Compression ratios can be regarded as intelligence measures

500'000€ Prize for Compressing Human Knowledge

http://prize.hutter1.net/index.htm

Motivation

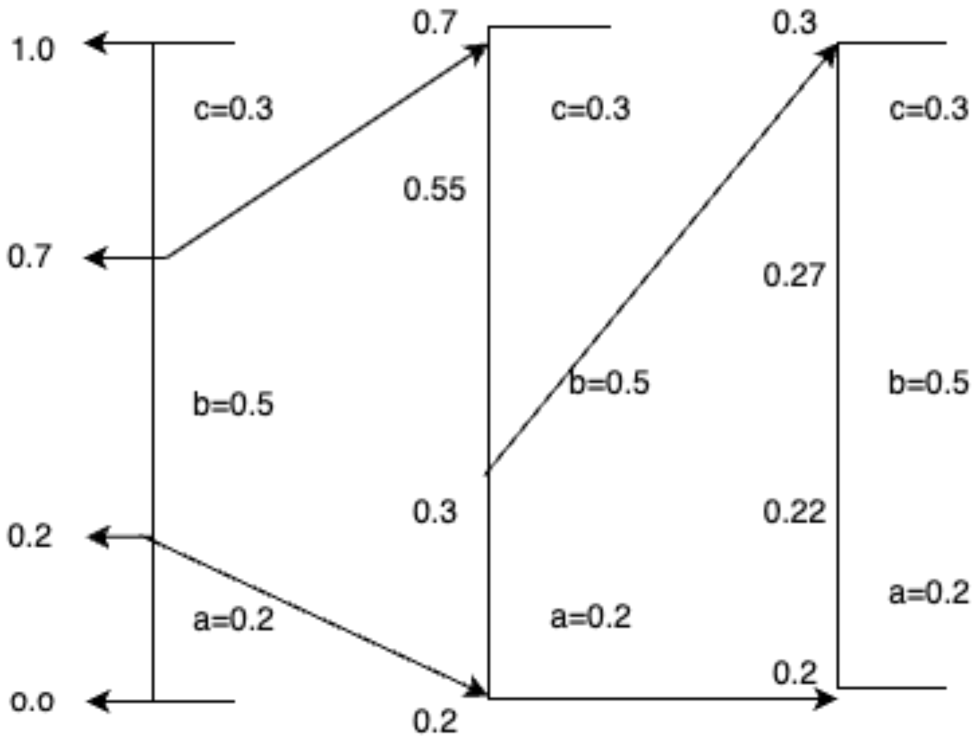
Demonstrate that Language Models, while trained primarily on text, also achieve state-of-the-art compression rates across different data modalities.

Coding the message sequence: **bac**

Arithmetic Coding

1. Symbol Probabilities
   - ➢ Each symbol in the message is associated with a probability.
2. Interval Initialization
   - ➢ The entire message is initially represented by an interval [0, 1).
3. Interval Update for Each Symbol
   - ➢ The interval is updated for each symbol based on its probability within the current interval.
4. Subdivision of Interval
   - ➢ The interval is subdivided into subintervals, each corresponding to the probability range of a symbol.
5. Final Interval
   - ➢ The final interval uniquely represents the entire encoded message.

Example: Coding the message sequence: **bac**



The final sequence interval is [.27,.3)

Represent information as a ranges, defined by two numbers

Shannon's source coding theorem establishes the limit on possible data compression as $L \geq H(p)$

Datasets

1. Dataset Modalities:

- Three modalities: text, image, and audio. (enwik9, ImageNet, LibriSpeech)

- Each dataset is 1GB to ensure comparability.

2. Context Lengths:

- Transformers have a context length C of 2048 bytes (or tokens).

- Gzip uses a maximum context of 32 kilobytes.

- LZMA2 has a virtually “infinite” context length.

3. Handling Different Context Lengths:

- Compressors with finite contexts can handle longer sequences in two ways:

- Slide the compressor byte by byte, maintaining a history of the previous $C - 1$ bytes when compressing a new byte.

- Chunk the data stream into sequences of C bytes and evaluate in-context compression, averaging across batches.

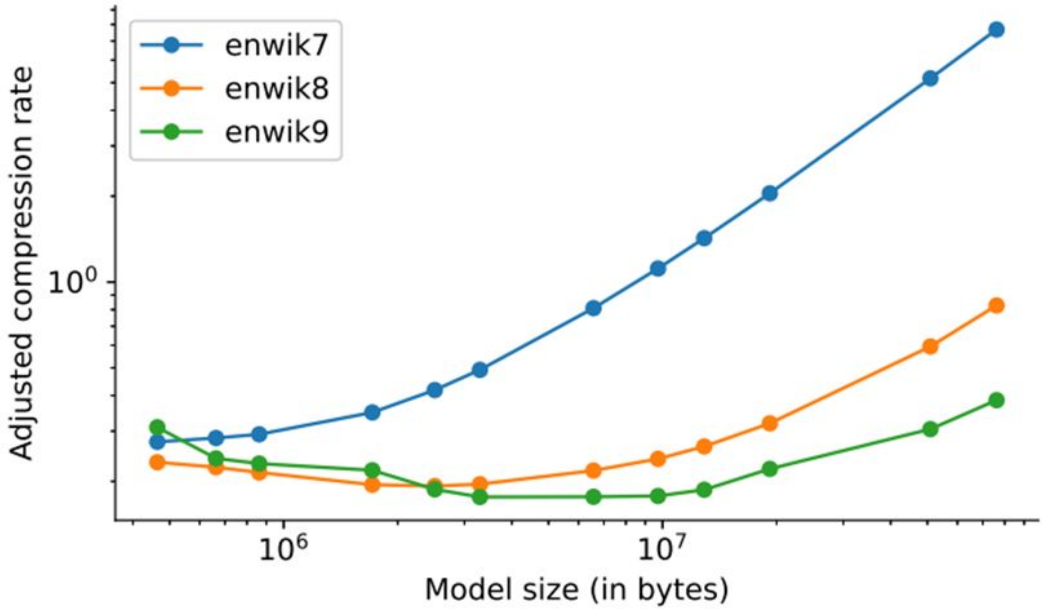
4. Approach for Transformers:

- Transformers use the chunking approach due to the long running time associated with sliding.
- Datasets are chunked into sequences of 2048 bytes.

Compression rates on different datasets (lower is better)

Chunk Size	Compressor	Raw Compression Rate (%) enwik9	Raw ImageNet	Raw LibriSpeech	Raw Random	Adjusted Compression Rate (%) enwik9	Adjusted ImageNet	Adjusted LibriSpeech	Adjusted Random
∞	gzip	32.3	70.7	36.4	100.0	32.3	70.7	36.4	100.0
	LZMA2	23.0	57.9	29.9	100.0	23.0	**57.9**	29.9	100.0
	PNG	42.9	58.5	32.2	100.0	42.9	58.5	32.2	100.0
	FLAC	89.5	61.9	30.9	107.8	89.5	61.9	30.9	107.8
2048	gzip	48.1	68.6	38.5	100.1	48.1	68.6	38.5	100.1
	LZMA2	50.0	62.4	38.2	100.0	50.0	62.4	38.2	100.0
	PNG	80.6	61.7	37.6	103.2	80.6	61.7	37.6	103.2
	FLAC	88.9	60.9	30.3	107.2	88.9	60.9	**30.3**	107.2
	Transformer 200K	30.9	194.0	146.6	195.5	30.9	194.0	146.6	195.5
	Transformer 800K	21.7	185.1	131.1	200.1	21.9	185.3	131.3	200.3
	Transformer 3.2M	17.0	215.8	228.2	224.0	**17.7**	216.5	228.9	224.7
	Chinchilla 1B	11.3	62.2	24.9	108.8	211.3	262.2	224.9	308.8
	Chinchilla 7B	10.2	54.7	23.6	101.6	1410.2	1454.7	1423.6	1501.6
	Chinchilla 70B	**8.3**	**48.0**	**21.0**	100.8	14008.3	14048.0	14021.0	14100.8

Comparing Compression Rates



Every dataset gives rise to an optimal model size, with a good trade-off between performance and cost of the model

Foundation Models Are General-Purpose Compressors

- A lossless compressor cannot compress all bit sequences equally
- Chinchilla models appear to be general-purpose compressors
- By conditioning a (meta-)trained model to a particular task at hand via in-context learning
- Larger models’ stronger in-context compression comes at a price ie, the number of parameters

Optimal Model-Dataset Size Tradeoff

Scaling laws are dependent on the size of the test set.

- Larger models achieve better compression rates on larger datasets

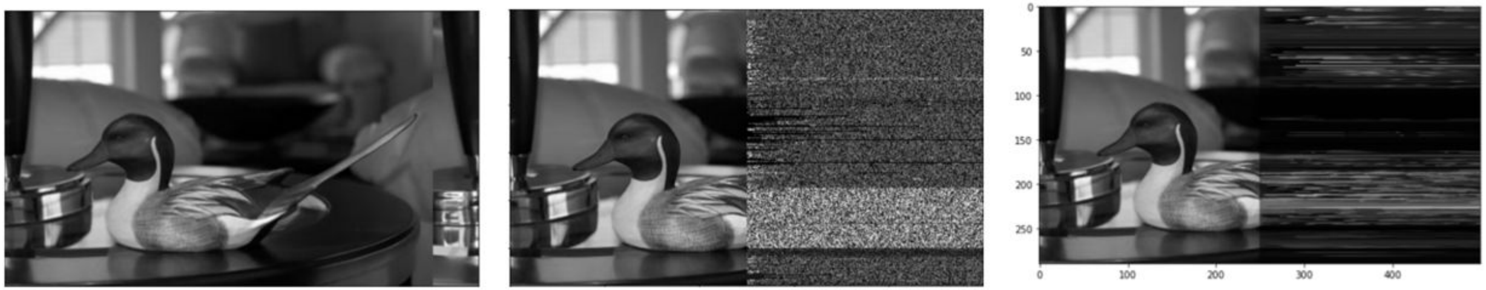
- However, they achieve worse rates on smaller datasets.

  After a point, number of parameters becomes too big compared to the size of the dataset

Compressors as Generative Models

Compressors as sequence prediction models

- Sampling distribution $\hat{\rho}(x_i \mid x_{<i}) \sim 2^{\ell_c(x_{<i})-\ell_c(x_{<i}x_i)}$
- Condition the compressors on part of an existing sequence and generate the remaining bytes with the compression-based generative model by using - teacher forcing or autoregressive sampling



(a) Original image (b) gzip (row-wise) (c) Chinchilla (row-wise)

Sequential Evolution of In-Context Compression

- Classical compressors optimize large context length to exploit sequential dependencies in the data

- Arithmetic coding-based compressors rely heavily on the predictive models' in-context learning capabilities to achieve competitive compression performance

- Compression rates decrease quickly with increasing sequence length, indicating that the models learn some data statistics in-context, without any gradient-based training

Tokenization Is Compression

	Raw Compression Rate (%)		
Tokenization	**200K**	**6.4M**	**38M**
ASCII	22.9	**13.6**	**6.4**
BPE 1000	25.4	14.8	6.9
BPE 2000	25.6	15.7	7.4
BPE 5000	23.1	17.1	8.7
BPE 10000	21.3	17.0	8.9
BPE 20000	**19.3**	16.4	9.0

- Larger vocabulary sizes reduce the sequence length, but does not reduce the entropy of the conditional distribution $\rho(x_i \mid x_{<i})$
- If the model is small, increasing tokens count boosts the compression performance. For bigger models having a larger token vocabulary harms the final compression rate of the model.
- Short sequence lengths also help Transformers since time complexity scales quadratically with context length, and so do not generalize well to long contexts.

Conclusion

- Arithmetic coding transforms a prediction model into a compressor. Compressor can be transformed into a predictor by using the coding lengths to construct probability distributions (Shannon’s entropy principle)
- Large pretrained models as compressors outperformed general compressors on modalities that were not trained up on.
- Compression viewpoint provides novel insights on scaling laws since it takes the model size into account, unlike the log-loss objective
- Optimal model size is linked to the dataset size and cannot be scaled without limit.

Thank you