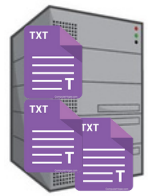
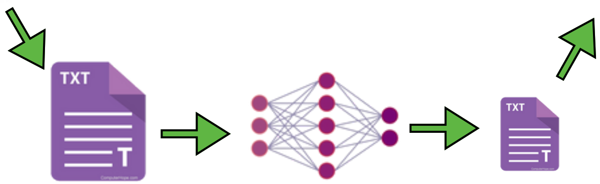
Approaching Entropy Limits with Learned Lossless Compression

Ruchir Kafle, Massachusetts Academy of Math and Science

Research Question

Can a data compressor condense text data down to its entropy limit?





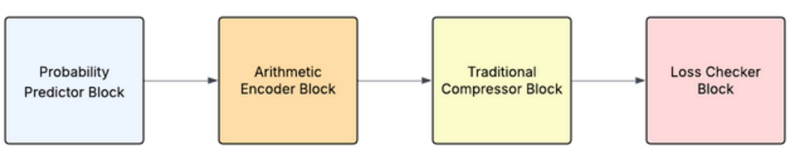
Data Analysis & Results

	Dickens							Enwik9		
Compressor	Default	CMix (Byron, n.d.)	TRACE (Mao et al., 2022)	7z, bz2, xz	zst	zpaq, zlib, zip, gz	lzo, rar	Default	TRACE	7z
Storage (MB)	10	1.8	2.6	2.8	3.7	3.9	6.2	1000	185	225
Ratio	1	5.56	3.85	3.57	2.70	2.56	1.61	1	5.41	4.44
%	100%	18%	26%	28%	37%	39%	62%	100%	18.5%	22.5%

Neural compressors like CMIX and TRACE output greater ratios during compression; they output smaller final file sizes. Their ability to reach entropy exceeds that of traditional compressors.

Neural compression approaches entropy quicker than traditional compression.

Methodology



Utilize a large language model for semantic understanding.

Use probability distributions to encode.

Add traditional data compression algorithms for further compression.

Prove losslessness and find final entropy.

Conclusion

Ollama logo: a tool that assists with running LLMs locally.

Using neural compressors, reaching entropy on large datasets is a foreseeable achievement.

Using large language models (LLMs) in conjunction with traditional models may increase accuracy of probability distributions and smaller compression.