AI language models can exceed PNG and FLAC in lossless compression, says study

reader comments

5 with

Effective compression is about finding patterns to make data smaller without losing information. When an algorithm or model can accurately guess the next piece of data in a sequence, it shows it’s good at spotting these patterns. This links the idea of making good guesses—which is what large language models like GPT-4 do very well—to achieving good compression.

In an arXiv research paper titled “Language Modeling Is Compression,” researchers detail their discovery that the DeepMind large language model (LLM) called Chinchilla 70B can perform lossless compression on image patches from the ImageNet image database to 43.4 percent of their original size, beating the PNG algorithm, which compressed the same data to 58.5 percent. For audio, Chinchilla compressed samples from the LibriSpeech audio data set to just 16.4 percent of their raw size, outdoing FLAC compression at 30.3 percent.

In this case, lower numbers in the results mean more compression is taking place. And lossless compression means that no data is lost during the compression process. It stands in contrast to a lossy compression technique like JPEG, which sheds some data and reconstructs some of the data with approximations during the decoding process to significantly reduce file sizes.

The study’s results suggest that even though Chinchilla 70B was mainly trained to deal with text, it’s surprisingly effective at compressing other types of data as well, often better than algorithms specifically designed for those tasks. This opens the door for thinking about machine learning models as not just tools for text prediction and writing but also as effective ways to shrink the size of various types of data.

Enlarge / A chart of compression test results provided by DeepMind researchers in their paper. The chart illustrates the efficiency of various data compression techniques on different data sets, all initially 1GB in size. It employs a lower-is-better ratio, comparing the compressed size to the original size.

Over the past two decades, some computer scientists have proposed that the ability to compress data effectively is akin to a form of general intelligence. The idea is rooted in the notion that understanding the world often involves identifying patterns and making sense of complexity, which, as mentioned above, is similar to what good data compression does. By reducing a large set of data into a smaller, more manageable form while retaining its essential features, a compression algorithm demonstrates a form of understanding or representation of that data, proponents argue.

Advertisement

The Hutter Prize is an example that brings this idea of compression as a form of intelligence into focus. Named after

Marcus Hutter, a researcher in the field of AI and one of the named authors of the DeepMind paper, the prize is awarded to anyone who can most effectively compress a fixed set of English text. The underlying premise is that a highly efficient compression of text would require understanding the semantic and syntactic patterns in language, similar to how a human understands it.

So theoretically, if a machine can compress this data extremely well, it might indicate a form of general intelligence—or at least a step in that direction. While not everyone in the field agrees that winning the Hutter Prize would indicate general intelligence, the competition highlights the overlap between the challenges of data compression and the goals of creating more intelligent systems.

Along these lines, the DeepMind researchers claim that the relationship between prediction and compression isn't a one-way street. They posit that if you have a good compression algorithm like gzip, you can flip it around and use it to generate new, original data based on what it has learned during the compression process.

In one section of the paper (Section 3.4), the researchers carried out an experiment to generate new data across different formats—text, image, and audio—by getting gzip and Chinchilla to predict what comes next in a sequence of data after conditioning on a sample. Understandably, gzip didn't do very well, producing completely nonsensical output—to a human mind, at least. It demonstrates that while gzip can be compelled to generate data, that data might not be very useful other than as an experimental curiosity. On the other hand, Chinchilla, which is designed with language processing in mind, predictably performed far better in the generative task.

Enlarge / An example from the DeepMind paper comparing the generative properties of gzip and Chinchilla on a sample text. gzip’s output is unreadable.

While the DeepMind paper on AI language model compression has not been peer-reviewed, it provides an intriguing window into potential new applications for large language models. The relationship between compression and intelligence is a matter of ongoing debate and research, so we’ll likely see more papers on the topic emerge soon.

https://arstechnica.com/?p=1969616