DZip: improved general-purpose lossless compression based on novel neural network modeling

Mohit Goyal®, Kedar Tatwawadi*, Shubham Chandak* and Idoia Ochoaγ

®Electrical and Computer Engineering, University of Illinois, Urbana, IL, USA
*Department of Electrical Engineering, Stanford University, CA, USA
mohit@illinois.edu, kedart@stanford.edu

We propose DZip, a general-purpose compressor for sequential data that exploits the modeling capabilities of neural networks (NNs) for prediction, followed by arithmetic coding. DZip uses a novel hybrid architecture with two models, a bootstrap model and a supporter model (see Figure 1 for high-level algorithm description). Unlike most NN based compressors, DZip does not require additional training data and is not restricted to specific data types.

DZip outperforms general-purpose compressors such as Gzip (approximately 25% reduction) on a variety of real datasets including text and genomic datasets, and performs close to specialized compressors for large sequence lengths. DZip also achieves near-optimal compression on synthetic datasets with simple structure (i.e., low Kolmogorov complexity) but long-term dependencies, which make them difficult to compress using traditional compressors. The results showcase the potential of developing improved general-purpose compressors based on neural networks and hybrid modeling. Future work includes making DZip a practical tool by improving its encoding and decoding speed. More details are available in [1] and the code is available at https://github.com/mohit1997/DZip.

Figure 1: DZip compression overview: In Stage I, a parameter-efficient bootstrap model is trained on the sequence $S^N$ to be compressed. In Stage II, the bootstrap model is combined with a larger supporter model to generate the probability predictions for the current symbol given the past $K$ symbols, which are fed into the arithmetic coder. The combined model is trained adaptively as the sequence is compressed. The final output consists of the trained bootstrap model parameters and the output of the arithmetic coder.

[1] Mohit Goyal, Kedar Tatwawadi, Shubham Chandak, and Idoia Ochoa, “DZip: improved general-purpose lossless compression based on novel neural network modeling,” arXiv e-prints, p. arXiv:1911.03572, Nov 2019.