High-throughput variable-to-fixed entropy codec using selective, stochastic forest codes

Manuel Martínez Torres, Miguel Hernández-Cabronero, Ian Blanes, Joan Serra-Sagristà

Manuel Martínez Torres is with Karlsruhe Institute of Technology, Karl-sruhe 76131, Germany (e-mail: manuel.martinez@kit.edu). M. Hernández-Cabronero (e-mail: miguel.hernandez@uab.cat), I. Blanes (e-mail: ian.blanes@uab.cat) and J. Serra-Sagristà (e-mail: joan.serra@uab.cat) are with the Universitat Autònoma de Barcelona, Bellaterra 08193, Spain.

Introduction

This repository contains the benchmark implementation used to gather data for the homonymous scientific paper, as well as all tools used to analyze data. It is intended to be self-contained to allow reproducibility, therefore a copy of the dataset used for comparison is provided as well.

Several existing codec implementations, as well as all non-synthetic data samples, are included for reproducibility, and no authorship is claimed. Original authors are cited in the published manuscript, and their license notices and code kept unaltered in the repository. A list of authors of all included code and data is provided below.

Instructions

1. Install the following libraries

    ```
    $ sudo apt install build-essential libopencv-dev liblzo2-dev libzstd-dev libsnappy-dev liblz4-dev
    ```

Dataset information

External authors

Unless honest mistake, the list of external authors whose code or data is included in the repository is as follows:

- Datasets
    - ISO 12640-2:2004 Graphic technology — Prepress digital data exchange — Part 2: XYZ/sRGB encoded standard colour image data (XYZ/SCID))
    - The KodakCD set, authored by Kodak, and dutifully kept online by Rich Franzen. Individual image credits are available at http://r0k.us/graphics/kodak/PhotoCD_credits.txt.
    - RAWZOR by http://imagecompression.info, http://rawzor.com
    - The Mixed set is composed by:
        - The Gas Sensor Array Drift Dataset Data Set. Alexander Vergara and Shankar Vembu and Tuba Ayhan and Margaret A. Ryan and Margie L. Homer and Ramón Huerta, Chemical gas sensor drift compensation using classifier ensembles, Sensors and Actuators B: Chemical (2012) doi: 10.1016/j.snb.2012.01.074.
        - The Intel Lab Data by the Intel Berkeley Research lab. “Data presented on this page was collected through the hard work of: Peter Bodik, Wei Hong, Carlos Guestrin, Sam Madden, Mark Paskin, and Romain Thibaux. Mark aggregated the raw connectivity information over time and generated the beautiful network layout diagram. Sam and Wei wrote TinyDB and other software for logging data. Intel Berkeley provided hardware. The TinyOS team, in particular Joe Polastre and Rob Szewczyk, provided the software infrastructure and hardware designs that made this deployment possible.”
        - The Drug Review Dataset (Drugs.com) Data Set. Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In Proceedings of the 2018 International Conference on Digital Health (DH ‘18). ACM, New York, NY, USA, 121-125. DOI: 10.1145/3194658.3194677
- Codecs
    - The Marlin codec, high-throughput implementation in C++ by Manuel Martínez Torrez, prototype in Python by Miguel Hernández-Cabronero.
    - SNAPPY - Tarantov, Zeev and Gunderson, Steinar, 2011
    - Gipfelli - Copyright 2011 Google Inc. All Rights Reserved. - designed by Jyrki Alakuijala, and implemented by Rastislav Lenhardt as an

intern project.

- GZip from the zlib library, by Phil Katz.
- The LZO library by Markus F.X.J. Oberhumer
- Huff0, “a Huffman codec designed for modern CPU, featuring OoO (Out of Order) operations on multiple ALU (Arithmetic Logic Unit), achieving extremely fast compression and decompression speeds.”
- FSE “is a new kind of Entropy encoder, based on ANS theory, from Jarek Duda, achieving precise compression accuracy (like Arithmetic coding) at much higher speeds.”
- FAPEC by DAPCOM is not included as a binary due to license restrictions. Wrappers for the fapec and unfapec binaries are included in this repository to ease reproducibility, although licensing of those binaries needs be negotiated directly with the authors

Curation

Original data in the dataset has been curated by prepending a PGM-like header. This header provides width, height and sample depth information so that all samples can be read in an homogeneous way.

License

License

October 2019