Model-Code Separation Architectures for Compression Based on Message-Passing

by
Ying-zong Huang

Submitted to the Department of Electrical Engineering and Computer Science
on December 31, 2014, in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

Abstract

Data is compressible by presuming a priori knowledge known as a data model, and applying an appropriate encoding to produce a shorter description. The two aspects of compression — data modeling and coding — however are not always conceived as distinct, nor implemented as such in compression systems, leading to difficulties of an architectural nature.

For example, how would one make improvements upon a data model whose specific form has been standardized into the encoding and decoding processes? How would one design coding for new types of data such as in biology and finance, without creating a new system in each case? How would one compress data that has been encrypted when the conventional encoder requires data-in-the-clear to extract redundancy? And how would mobile acquisition devices obtain good compression with lightweight encoders? These and many other challenges can be tackled by an alternative compression architecture.

This work contributes a complete "model-code separation" system architecture for compression, based on a core set of iterative message-passing algorithms over graphical models representing the modeling and coding aspects of compression. Systems following this architecture resolve the challenges posed by current systems, and stand to benefit further from future advances in the understanding of data and the algorithms that process them.

In the main portion of this thesis, the lossless compression of binary sources is examined. Examples are compressed under the proposed architecture and compared against some of the best systems today and to theoretical limits. They show that the flexibility of model-code separation does not incur a performance penalty. Indeed, the compression performance of such systems is competitive with and sometimes superior to existing solutions.

The architecture is further extended to diverse situations of practical interest, such as mismatched and partially known models, different data and code alphabets, and lossy compression. In the process, insights into model uncertainty and universality, data representation and alphabet translation, and model-quantizer separation and low-complexity quantizer design are revealed. In many ways, the proposed architecture is uniquely suitable for understanding and tackling these problems.

Throughout, a discourse is maintained over architectural and complexity issues, with a view toward practical implementability. Of interest to system designers, issues such as rate selection, doping, and code selection are addressed, and a method similar to EXIT-chart analysis is developed for evaluating when compression is possible. Suggestions for system interfaces and algorithmic factorization are distilled, and examples showing compression with realistic data and tasks are given to complete the description of a system architecture accessible to broader adoption.

Ultimately, this work develops one architecturally principled approach toward flexible, modular, and extensible compression system design, with practical benefits. More broadly, it represents the beginning of many directions for promising research at the intersection of data compression, information theory, machine learning, coding, and random algorithms.

Thesis Supervisor: Gregory W. Wornell
Title: Professor of Electrical Engineering and Computer Science