Predictive Coding For Lossless Dataset Compression

#3992

Unordered source coding in theory and practice

Traditional Source Coding

Consider a sequence $(X_1, X_2, \ldots, X_n)$ with $X_1, X_2, \ldots, X_n \overset{iid}{\sim} P_X$

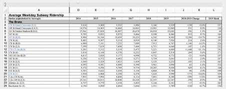
Given a random variable over an alphabet $\mathcal{X}$, a **lossless source code (l.s.c.)** consists of functions $f : \mathcal{X} \to \{0,1\}^*$ and $g : \{0,1\}^* \to \mathcal{X}$ such that $g(f(x)) = x$

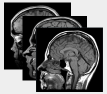
To encode symbols with few bits, we usually minimize the average code word length:

$$M^*(P_X) \triangleq \min \{\mathbb{E}\left[l(f(X))\right] \mid (f,g) \text{ is a prefix-free l.s.c.}\}$$

Motivation

Many large datasets in research, archives, ML training, etc.





Traditional compression algorithms operate on stream data. If we don't need to preserve order on elements, we can save space and bandwidth with a "dataset" compression algorithm.

Dataset Source Coding

Consider a "dataset", i.e. a set of samples where order doesn't matter: $\{X_1, X_2, \ldots, X_n\}$

Define a **lossless dataset source code (l.d.s.c)** $f : \mathcal{X}^n \to \{0,1\}^*$ and $g : \{0,1\}^* \to \mathcal{X}^n$ such that $g(f(x^n)) = \pi \circ x^n$ for all $x^n \in \mathcal{X}^n$ where π is a permutation.

For a l.d.s.c. we minimize

$$M_n^*(P_X) \triangleq \min \{\mathbb{E}\left[l(f(X^n))\right] \mid (f,g) \text{ is a prefix-free l.d.s.c.}\}.$$

Note: dataset code length ≤ sequence code length

Theoretical Results

Theorem 1 (l.d.s.c. via data structures):

Let $\widetilde{X}^n = \pi \circ X^n$, where π is a permutation drawn uniformly at random. Moreover, let S be such that

1) $S \to X^n \to \widetilde{X}^n$ and $X^n \to S \to \widetilde{X}^n$

2) $H(S|X^n) = H(S|\widetilde{X}^n) = 0$

Then $M^*(P_S) = M_n^*(P_X)$,

$$H(S) \leq M_n^*(P_X) < H(S) + 1$$

and $H(S) = I(X^n; \widetilde{X}^n) \leq \min\{|\mathcal{X}| \log_2(n+1), n \log_2 |\mathcal{X}|\}$

Experiments

Idea: in many data, the "features" are not truly independent from each other. For example, pixels in an image.

Predictive coding: instead of encoding every value, can encode some values, a model to obtain adjacent ones, and the "error" between predictions and true value.

Combining ideas from **Theorem 1** and JPEG-LS:

1. Build nearest neighbors graph of dataset
2. Obtain reordering by traversing a MST
3. Train predictor on context from *within* images and *adjacent* images
4. Entropy coding

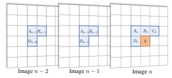
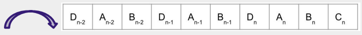


Fig. 1. X is the current pixel under prediction.



Benchmarks

Compressor	MNIST Size (MB)	CIFAR-10 Size (MB)
Uncompressed	47.04	155.18
gz	9.66	141.7
bz2	8.40	122.6
xz	8.11	116.4
JPEG-LS	16.58	118.66

	Predictor	Ordering	Context strategy	Final size (MB)
MNIST	Logistic	11-nn MST	DABC	13.56
CIFAR-10	Linear (Single)	Random	DAB	112.41

Madeleine Barowsky, Alexander Mariona, and Flavio du Pin Calmon
Harvard University