E Baselines

In this section, we introduce the baseline methods for set compression. Let $\mathbb{1}.$ be the indicate function and $\mathrm{H}(\cdot)$ be the information entropy function.

Single-item-based. Let $\mathbf{x}^S = (\mathbf{x}_1^S, \mathbf{x}_2^S, \dots, \mathbf{x}_K^S) \in \{0,1\}^K$ represent one-hot code of S, which means $\mathbf{x}_i^S = \mathbb{1}_{i \in S}$. Let X_i be a random variable on $\{0,1\}$ with $\Pr(X_i = 1) = n_i/N$, where $n_i = \sum_j \mathbb{1}_{i \in S_j}$. By entropy coding, coding the $\mathbf{x}_i^S$ of every $S \in \mathcal{S}$ takes $N \cdot \mathrm{H}(X_i)$ bits. And, the total code length for $\mathcal{S}$ is

$$L_{single} = \sum_{i=1}^{K} N \cdot \mathrm{H}(X_i) = \sum_{i=1}^{K} n_i \log_2 \frac{N}{n_i} + (N - n_i) \log_2 \frac{N}{N - n_i}.$$

The number of parameters equals the number of n_i, which is $|V|$.

Group-based. If we group some variable $\mathbf{x}_i^S$ and consider the entropy coding on the joint distribution, we can get a lower code rate. Let $V = G_1 \cup G_2 \cup \dots G_M$, where $G_1 = \{1, 2, \dots, 8\}, G_2 = \{9, \dots, 16\}, G_3 = \{17, \dots, 24\}$, and so on.

Then we code the eight $\mathbf{x}_i^S$ in a group together with categorical distribution on $\{0,1\}^8$. For each G_i, we need $2^8 - 1 = 255$ numbers of parameters. Thus, the total number of parameters we need is $|V|/8 \cdot 255 \approx 32|V|$.

Bits-Back-based. Let $\mathcal{V} = V \cup \{0\}$ be the vocabulary set and symbol 0 is used as the split symbol. Then $\mathcal{S}$ can be treated as a stream of symbols. The stream is initialized with empty. We put the items of S_1 to the stream and put a 0 to the stream. Then, we put the items of S_2 to the stream and put a 0 to the stream, and so on. Finally, we get a stream with $\tilde{N} = N + \sum_S |S|$ numbers of symbols.

Let Y be a random variable on $\mathcal{V}$ such that $\Pr(Y = 0) = N/\tilde{N}$, and

$$\Pr(Y = i) = \frac{\sum_j \mathbb{1}_{i \in S_j}}{\tilde{N}}, \text{ for } i \in V.$$

By entropy coding, the code length of the stream is $\tilde{N} \cdot \mathrm{H}(Y)$.

Since the input order of items in S to the stream does not matter, we can further decrease the length by the bits-back method [16] and the final code length is

$$L_{bits-back} = \tilde{N} \cdot \mathrm{H}(Y) - \sum_{S \in \mathcal{S}} \log_2(|S|!)$$

The number of parameters is the number of symbols minus 1, which is $|\mathcal{V}| - 1 = |V|$.

Markov-based. We sort the items in every set to give them an increased order. The dataset is conveyed to a long sequence of items with split symbols between different sets. Then we used context-based arithmetic coding [4] with the first-order Markov model for coding. The vocabulary is the same as the bits-back-based method and the number of parameters is $|V|(|V| + 1)/2$.

Common compression software. We test the performance of a common compression software which is 7-zip (https://www.7-zip.org/). We adopted the extreme compression mode with the LZMA2 [4] algorithm on 7-zip.

Received 1 August 2024