

Relatório do Trabalho Laboratorial nº 3

Informação e Codificação (2025/26)

Pedro Miguel Miranda de Melo (114208)

Rúben Cardeal Costa (114190)

Hugo Marques Dias (114142)

Departamento de Eletrónica, Telecomunicações e Informática (DETI)

Universidade de Aveiro

Novembro de 2025

Conteúdo

1	Introdução	3
2	Análise e Caracterização da Fonte	3
2.1	Análise Estrutural (Formato de Dados)	3
2.2	Limites Teóricos e Análise Global	3
2.3	Análise Estrutural Diferenciada (<i>Byte-Splitting</i>)	4
2.4	Síntese e Estratégia Adotada	5
3	Conclusões	5

1 Introdução

O presente relatório técnico descreve o trabalho realizado no âmbito do projeto de compressão de um Grande Modelo de Linguagem (LLM). O objetivo central passa por desenvolver uma estratégia de compressão otimizada e eficiente para o ficheiro `model.safetensors` (~ 1 GB) que contém os parâmetros de um LLM. [Continuar...](#)

2 Análise e Caracterização da Fonte

Para desenhar um codec eficiente, é imperativo compreender a natureza estatística da fonte de informação. Esta secção detalha a análise teórica e experimental realizada sobre o ficheiro `model.safetensors`.

2.1 Análise Estrutural (Formato de Dados)

A inspeção do cabeçalho do ficheiro revelou que os dados estão armazenados no formato **BF16** (*Brain Floating Point 16*). Ao contrário de inteiros de 16 bits, onde a distribuição de bits tende a ser uniforme em dados aleatórios, o BF16 possui uma semântica específica composta por:

- **1 bit de Sinal (S) e 8 bits de Exponente (E):** Ocupam maioritariamente o byte mais significativo (MSB).
- **7 bits de Mantissa (M):** Ocupam o byte menos significativo (LSB).

Esta estrutura sugere a existência de correlações não-lineares e localizadas que uma análise puramente sequencial (byte-a-byte) poderá não capturar eficazmente.

2.2 Limites Teóricos e Análise Global

O limite teórico fundamental para a compressão sem perdas é dado pela **Entropia de Shannon**. Considerando o ficheiro como uma fonte de memória nula X que gera símbolos $x \in \{0, \dots, 255\}$, a entropia de ordem-0 é definida por:

$$H(X) = - \sum_{i=0}^{255} P(x_i) \log_2 P(x_i) \quad [\text{bits/símbolo}] \quad (1)$$

Experimentalmente, ao aplicar a Equação 1 à totalidade do *payload* binário, obteve-se:

$$H(X) \approx 6.22 \text{ bits/byte}$$

Este valor indica que, ignorando qualquer dependência entre bytes, a compressão máxima teórica seria de apenas $\sim 22\%$. Para investigar dependências sequenciais, calculou-se a **Entropia Condicional** de primeira ordem, que mede a incerteza de um símbolo X_n dado o conhecimento do anterior X_{n-1} :

$$H(X|Y) = - \sum_{y \in \mathcal{X}} P(y) \sum_{x \in \mathcal{X}} P(x|y) \log_2 P(x|y) \quad (2)$$

O resultado experimental obtido foi:

$$H(X_n|X_{n-1}) \approx 5.36 \text{ bits/byte}$$

Embora $H(X|Y) < H(X)$, confirmado a existência de correlação inter-simbólica (teorema do condicionamento reduz a entropia), o valor 5.36 permanece elevado. A nossa hipótese é que a natureza intercalada dos dados BF16 (MSB estruturado seguido de LSB ruidoso) "mascara" a verdadeira correlação entre os pesos adjacentes.

2.3 Análise Estrutural Diferenciada (*Byte-Splitting*)

Para validar a hipótese de que a entropia está concentrada no byte da mantissa, procedeu-se à separação do fluxo de dados em dois canais distintos: *Stream MSB* (bytes ímpares) e *Stream LSB* (bytes pares).

As entropias de ordem-0 foram recalculadas individualmente para cada canal:

Tabela 1: Comparaçāo de Entropia por Canal (Split)

Canal	Conteúdo	Entropia Medida (H)	Característica
MSB	Expoente/Sinal	2.71 bits/byte	Altamente Estruturado
LSB	Mantissa	7.96 bits/byte	Ruído Quase Uniforme

Evidência Visual

Os histogramas de frequência (Figuras 1 e 2) corroboram os valores numéricos.

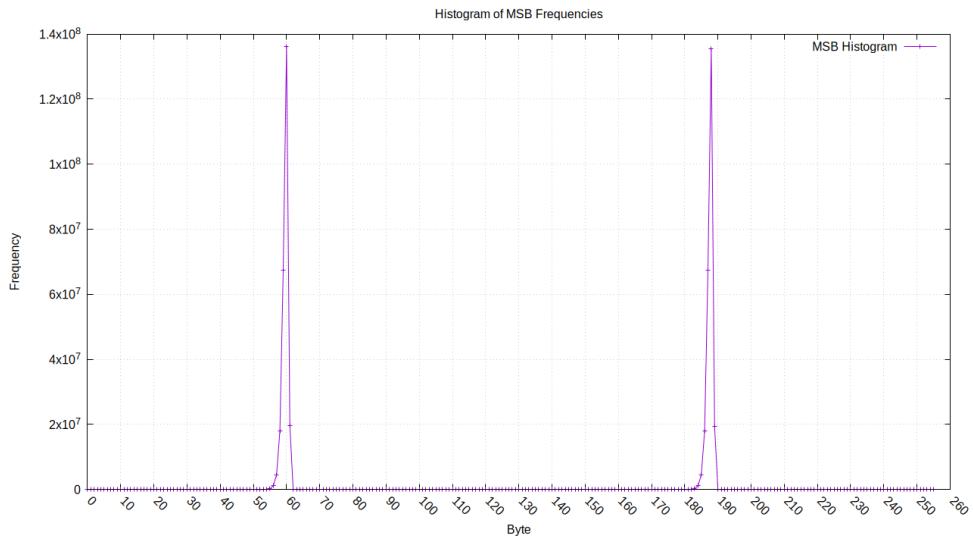


Figura 1: Histograma do Byte Mais Significativo (MSB). Nota-se uma distribuição Laplaciana acentuada, típica de pesos de redes neurais normalizados, justificando o valor baixo de $H \approx 2.71$.

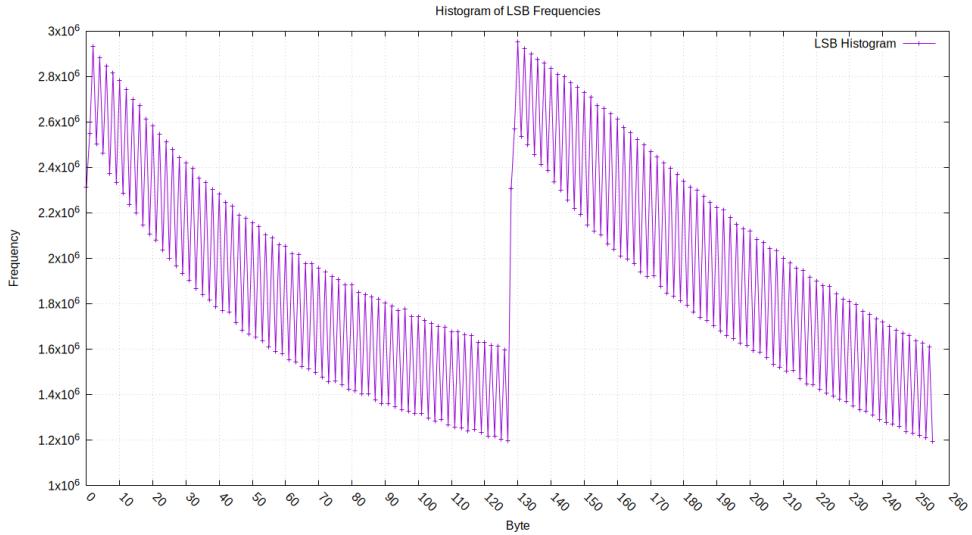


Figura 2: Histograma do Byte Menos Significativo (LSB). A distribuição aproxima-se da uniforme (plana), o que explica a entropia de $H \approx 7.96$, muito próxima do máximo teórico de 8 bits.

2.4 Síntese e Estratégia Adotada

A média das entropias separadas é $(2.71 + 7.96)/2 \approx 5.34$, um valor virtualmente idêntico à Entropia Condisional global (5.36). Isto leva-nos a concluir que a "memória" da fonte detetada na análise global era, na verdade, a estrutura interna do formato BF16 e não apenas correlação sequencial.

Nota

Com base nestes dados teóricos e experimentais, a estratégia de compressão será:

1. **Pré-processamento (Split):** Separar os fluxos para isolar o ruído da estrutura.
2. **Canal LSB:** Dado que $H \approx 8$, não justifica custo computacional elevado. Será armazenado com compressão mínima ou nula.
3. **Canal MSB:** Dado que $H \approx 2.71$, este canal é o candidato ideal para **Codificação Preditiva** (Delta) seguida de **Codificação Entrópica** (Huffman ou Aritmética), visando reduzir a entropia residual para valores próximos de 2 bits/byte.

3 Conclusões