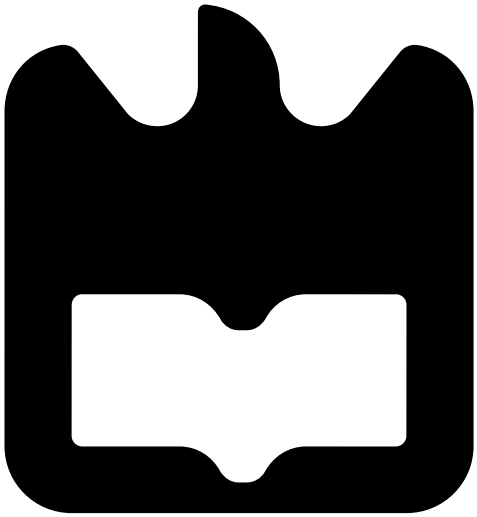

Relatório do Trabalho Laboratorial nº 3

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

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` ($\sim$ 1 GB) que contém os parâmetros de um LLM. Continuar...

2 Análise e Caracterização da Fonte

Antes de aplicar qualquer algoritmo de compressão, procedeu-se a uma análise estrutural e estatística do ficheiro para determinar a natureza dos dados e os limites teóricos de compressão.

2.1 Estrutura do Ficheiro (`safetensors`)

Através de engenharia reversa e análise do cabeçalho, determinou-se que o formato `safetensors` consiste num bloco de memória contíguo dividido em três secções:

1. **Tamanho do Cabeçalho:** Os primeiros 8 bytes (inteiro de 64 bits) indicam o tamanho da secção seguinte.
2. **Metadados (JSON):** Um cabeçalho descritivo contendo o nome, forma (*shape*) e tipo de dados (*dtype*) de cada tensor.
3. **Payload Binário:** A grande maioria do ficheiro consiste nos valores numéricos dos tensores concatenados.

A inspeção do JSON revelou que todos os tensores estão armazenados no formato **BF16** (*Brain Floating Point 16*). Ao contrário de um fluxo de bytes de texto ou imagem genérica, isto indica que o ficheiro é composto por sequências de números de 16 bits (2 bytes).

O formato BF16 segue a estrutura:

- 1 bit: Sinal.
- 8 bits: Expoente.
- 7 bits: Mantissa.

Esta descoberta é fundamental, pois sugere uma correlação estrutural entre bytes alternados (*Byte Alto* vs. *Byte Baixo*), que seria ignorada se o ficheiro fosse tratado como um fluxo de bytes uniforme ($x(n) \in \{0, ..., 255\}$).

2.2 Análise Teórica da Informação

Para quantificar o limite teórico de compressão, recorreu-se à definição de Entropia de Shannon, dada por:

$$H(X) = -\sum_{i} P(x_i) \log_2 P(x_i)$$

onde $P(x_i)$ é a probabilidade de ocorrência do símbolo x_i.

Dado que o ficheiro é constituído por valores BF16, formulou-se a hipótese de que a entropia não está uniformemente distribuída pelos 16 bits.

- **Byte Mais Significativo (MSB):** Contém o sinal e a maioria do expoente. Num modelo treinado, os pesos tendem a seguir uma distribuição normal centrada em zero, o que implica que os expoentes são altamente repetitivos (baixa entropia).
- **Byte Menos Significativo (LSB):** Contém a mantissa. Devido à precisão numérica, espera-se que estes bits apresentem um comportamento ruidoso, aproximando-se de uma distribuição uniforme (alta entropia).

2.3 Resultados Experimentais

Para validar a hipótese, implementou-se uma ferramenta de análise que separa o ficheiro em dois fluxos distintos (*Byte-Splitting*) e calcula a entropia de ordem-0 para cada um. Os resultados obtidos foram:

Tabela 1: Resultados do Cálculo da Entropia de Ordem-0

Fluxo de Dados	**Conteúdo Principal**	**Entropia Calculada (H)**	**Redundância ($8 - H$)**
LSB (Byte Baixo)	Mantissa (Precisão)	7.96 bits/byte	~ 0.04 bits
MSB (Byte Alto)	Expoente + Sinal	2.71 bits/byte	~ 5.29 bits
Média Combinada	Ficheiro Completo	5.34 bits/byte	~ 2.66 bits

Interpretação dos Resultados:

- **O Fluxo LSB é Incompressível:** Com uma entropia de ≈ 7.96 bits, este fluxo está muito próximo da entropia máxima de uma fonte uniforme ($H_{\max} = \log_2 256 = 8$ bits). A aplicação de algoritmos de compressão complexos aqui resultaria num ganho negligenciável, desperdiçando tempo de computação.
- **O Fluxo MSB é Altamente Compressível:** A entropia de 2.71 bits indica uma redundância estatística significativa. Isto confirma que a estratégia de compressão deve focar-se agressivamente neste fluxo.
- **Ganho com *Byte-Splitting*:** Se comprimíssemos o ficheiro original sem separação, a entropia média seria de 5.34 bits. Ao separar, isolamos o "ruído" num fluxo e a "informação estruturada" noutro, permitindo a aplicação de algoritmos otimizados para cada caso.

2.4 Definição da Estratégia de Compressão

Com base nesta análise, a arquitetura do compressor a desenvolver seguirá uma abordagem híbrida:

- **Pré-processamento:** Separação dos dados em dois *streams* (MSB e LSB).
- **Codificação do LSB:** Armazenamento direto (*raw*) ou uso de um algoritmo leve (ex: RLE apenas para zeros), dado que $H \approx 8$.
- **Codificação do MSB:**
  1. Aplicação de Codificação Preditiva ($r_n = x_n - \hat{x}_n$) para tentar reduzir a entropia abaixo de 2.71 bits, explorando a correlação entre pesos adjacentes.
  2. Uso de Codificação de Entropia (Huffman ou Aritmética) para atingir o limite teórico calculado.

Esta estratégia visa maximizar o rácio de compressão onde ele é possível (MSB) e minimizar o uso de recursos onde o ganho é marginal (LSB), permitindo um equilíbrio entre a eficiência da compressão e a complexidade computacional.

3 Conclusões