

2. Byte-Pair Encoding (BPE) Tokenizer

1. Understanding Unicode

- a. `chr(0)` returns `'b'\x00'`
- b. `__repr__()` shows an escaped, unambiguous form whereas its printed representation outputs the character itself which is actual NULL character(invisible).
- c. When we add it to text, it simply adds an invisible character at that position which may not be visually noticeable.

2. Unicode Encodings

- a. UTF-8 is preferred over UTF-16 or UTF-32 due to its superior compression, universal applicability without out of vocabulary errors, and native compability with byte-stream data.
- b. If we give the input 牛, it breaks and gives an error "UnicodeDecodeError: 'utf-8' codec can't decode byte 0xe7 in position 0: unexpected end of data". This is due to the fact that the decoding is done byte by byte and when the encoded character is multi-byte, it fails to decode it properly.
- c. `b'\xc3\x28'` is invalid in UTF-8 because `0xc3` indicates the start of a 2-byte character, but `0x28` is not a valid continuation byte.

3. BPE Training on TinyStories

- a. It seems that the longest token is 'accomplishment' with a length of 15 bytes. The memory usage is around 158.3 MB and the training took approximately 61.72 seconds. It makes sense that longer tokens are formed from frequently occurring sequences of characters in the dataset, and 'accomplishment' might be a common word in the TinyStories dataset.
- b. Pre-tokenization part took the most time (40 seconds). Finding chunks and their word count seems to be the most time-consuming part of the process. Second highest is merging the tokens based on the pairs found (20 seconds).

4. Experiments with tokenizers

- a. The tokenizer's compression ratio (bytes/token) is around 2.8609 in a sample of 10 documents from TinyStories dataset.
- b. The througput of the tokenizer is approximately 3000 bytes/second. To tokenize the Pile dataset (about 825 GB), it would take around 76,388 hours.
- c. `uint16` is appropriate choice for encoding the token IDs since the vocabulary size is 10,000 which fits well within the range of `uint16` (0 to $2^{16} - 1 = 65535$).

3. Transformer Language Model Architecture

1. Transformer LM resource accounting

- a. With the current implementation, GPT-2 XL has nearly 2,127,057,600 ($\approx 2.13B$) trainable parameters. At 4 bytes (32 bit floating point) per parameter, this amounts to approximately 8.5 GB of memory just for model parameters.
- b. Calculation is as follows:
 - **Per Transformer Block ($L = 48$ in total)**

- **Attention Projections** (W_Q, W_K, W_V): 3 matrix multiplications,

$$\begin{aligned}
 &= 3 * 2 * seq_len * d_model * d_model \\
 &= 3 * (2 * 1024 * 1600 * 1600) \\
 &= \mathbf{15,728,640,000 \text{ FLOPs}}
 \end{aligned}$$
- **Attention Scores** (QK^T): 1 matrix multiplication,

$$\begin{aligned}
 &= 2 * seq_len^2 * num_heads * (d_model / num_heads) \\
 &= 2 * 1024^2 * 25 * (1600 / 25) \\
 &= \mathbf{3,355,443,200 \text{ FLOPs}}
 \end{aligned}$$
- **Attention Output** ($ScoresV$): 1 matrix multiplication,

$$\begin{aligned}
 &= 2 * seq_len^2 * num_heads * (d_model / num_heads) \\
 &= 2 * 1024^2 * 25 * (1600 / 25) \\
 &= \mathbf{3,355,443,200 \text{ FLOPs}}
 \end{aligned}$$
- **Attention Output Projection** (W_O): 1 matrix multiplication,

$$\begin{aligned}
 &= 2 * seq_len * d_model^2 \\
 &= 2 * 1024 * 1600^2 \\
 &= \mathbf{5,242,880,000 \text{ FLOPs}}
 \end{aligned}$$
- **FFN** (W_1, W_3): 2 matrix multiplications,

$$\begin{aligned}
 &= 2 * (2 * seq_len * d_model * d_ff) \\
 &= 2 * 2 * 1024 * 1600 * 6400 \\
 &= \mathbf{41,943,040,000 \text{ FLOPs}}
 \end{aligned}$$
- **FFN** (W_2): 1 matrix multiplication,

$$\begin{aligned}
 &= 2 * seq_len * d_ff * d_model \\
 &= 2 * 1024 * 6400 * 1600 \\
 &= \mathbf{20,971,520,000 \text{ FLOPs}}
 \end{aligned}$$

So for all blocks combined, the total FLOPs is:

$$\begin{aligned}
 &= 48 * (15,728,640,000 + 3,355,443,200 + 3,355,443,200 + 5,242,880,000 \\
 &\quad + 41,943,040,000 + 20,971,520,000) \\
 &= 48 * 90,596,486,400 \\
 &= \mathbf{4,348,631,347,200 \text{ FLOPs}}
 \end{aligned}$$

- **Final LayerNorm and Output Projection**: 1 matrix multiplication,

$$\begin{aligned}
 &= 2 * seq_len * d_model * vocab_size \\
 &= 2 * 1024 * 1600 * 50257 \\
 &= \mathbf{164,682,137,600 \text{ FLOPs}}
 \end{aligned}$$

So the total FLOPs are:

$$\begin{aligned} &= 4,348,631,347,200 \text{ (Blocks)} + 164,682,137,600 \text{ (Head)} \\ &= \mathbf{4,513,313,484,800 \text{ FLOPs}} \\ &\approx \mathbf{4.51 \text{ TFLOPs}} \end{aligned}$$

- c. Most of the FLOPs occur within the **Transformer blocks** (about 96.3% of total FLOPs) and within each block, the **FFN layers** contribute about 70% of the block's FLOPs (62.9 GFLOPs) compared to the attention mechanism (27.7 GFLOPs).
- d. The breakdown of the FLOPs for each model is as follows:
- **GPT-2 Small ($L = 12$, $d_{\text{model}} = 768$, $\text{num_heads} = 12$)**
 - Blocks: $2.7 * 10^{11}$ FLOPs (77.7% of total)
 - Final LayerNorm and Output Projection: $0.79 * 10^{11}$ FLOPs (22.3% of total)
 - Total: Approximately **0.32 TFLOPs** per forward pass.
 - **GPT-2 Medium ($L = 24$, $d_{\text{model}} = 1024$, $\text{num_heads} = 16$)**
 - Blocks: $9.27 * 10^{11}$ FLOPs (90% of total)
 - Final LayerNorm and Output Projection: $1.05 * 10^{11}$ FLOPs (10% of total)
 - Total: Approximately **1.03 TFLOPs** per forward pass.
 - **GPT-2 Large ($L = 36$, $d_{\text{model}} = 1280$, $\text{num_heads} = 20$)**
 - Blocks: $2.1 * 10^{12}$ FLOPs (94% of total)
 - Final LayerNorm and Output Projection: $1.32 * 10^{11}$ FLOPs (6% of total)
 - Total: Approximately **2.23 TFLOPs** per forward pass.

As the model size increases, the proportional FLOPs cost changes toward the **Transformer blocks**, especially the FFN layers, which dominate the computational cost in larger models. The final output projection becomes relatively less expensive as the model size increases.

- e. Increasing the context_length from 1024 to 16,384 (a 16x increase) results in a significant rise in total FLOPs, approximately **33.2x** (from 4.51 TFLOPs to around 149.8 TFLOPs). As the context length grows, the contribution to the FLOPs shifts from the FFN layers to the attention score calculation. This shift becomes more pronounced at longer context lengths, as **attention score calculations** become the dominant cost due to their quadratic scaling with sequence length. Consequently, attention score computations grow from **7%** to **55%** of total FLOPs, becoming the new bottleneck.

4. Training a Transformer LM

1. Tuning the learning rate

With $\text{lr} = 1e1$, the loss keeps decaying properly. Initial iterations see a sharp drop in loss, and then it continues to decrease steadily and slowly.

With $\text{lr} = 1e2$, the loss keeps fluctuating and diverging. It starts with a sharp drop but then it increases and decreases erratically, indicating that the model is not converging properly.

With $\text{lr} = 1e3$, the loss diverges immediately and keeps increasing without any sign of convergence.

2. Resource accounting for training with AdamW

a. Let

$B = \text{batch_size}$,

$T = \text{context_length}$,

$D = d_model$,

$H = \text{num_heads}$,

$V = \text{vocab_size}$,

$N = \text{num_layers}$,

$d_{ff} = 4D$

- **Parameters:**

- **Embeddings:** Input ($\mathbf{V} \times \mathbf{D}$) + Position ($\mathbf{T} \times \mathbf{D}$)
- **Transformer Blocks:** For each of the \mathbf{N} blocks, we have:
 - * Attention Projections (W_Q, W_K, W_V, W_O): **4** matrices of size ($\mathbf{D} \times \mathbf{D}$)
 - * FFN (W_1, W_2, W_3): **3** matrices of size ($\mathbf{D} \times \mathbf{d}_{ff}$) and ($\mathbf{d}_{ff} \times \mathbf{D}$)
 - * **LayerNorms:** **2** RMSNorms of size (\mathbf{D})
- **Final Layers:** Final RMSNorm (\mathbf{D}) and Output Projection ($\mathbf{D} \times \mathbf{V}$)

So the total memory for parameters ($\mathbf{N}_{\text{params}}$) is:

$$\mathbf{N}_{\text{params}} = (\mathbf{V} \times \mathbf{D}) + (\mathbf{T} \times \mathbf{D}) + \mathbf{N} * [4 * (\mathbf{D}^2) + 3 * (\mathbf{D} \times \mathbf{d}_{ff}) + 2 * \mathbf{D}] + (\mathbf{D} + \mathbf{D} \times \mathbf{V})$$

$$\mathbf{N}_{\text{params}} = 2(\mathbf{VD}) + (\mathbf{TD}) + \mathbf{N} * [16(\mathbf{D}^2) + 2(\mathbf{D})] + (\mathbf{D})$$

$$\mathbf{Memory}_{\text{params}} = \mathbf{N}_{\text{params}} * 4 \text{ bytes}$$

- **Gradients:** We store one gradient value for every parameter, so

$$\mathbf{Memory}_{\text{grads}} = \mathbf{N}_{\text{params}} * 4 \text{ bytes}$$

- **Optimizer State:** AdamW maintains two additional values (momentum and variance) for each parameter, so

$$\mathbf{Memory}_{\text{optimizer}} = \mathbf{N}_{\text{params}} * 2 * 4 \text{ bytes}$$

- **Activations:**

- **Transformer Block:**
 - * RMSNorm1 Input: $\mathbf{B} \times \mathbf{T} \times \mathbf{D}$
 - * QKV Projections (Input/Output): $\mathbf{B} \times \mathbf{T} \times \mathbf{D}$ (Input) + $3 \times \mathbf{B} \times \mathbf{T} \times \mathbf{D}$ (Output)
 - * $Q^T K$ Scores & Softmax (Probs): $2 \times \mathbf{B} \times \mathbf{H} \times \mathbf{T} \times \mathbf{T}$
 - * Weighted Sum (Context): $\mathbf{B} \times \mathbf{T} \times \mathbf{D}$
 - * Attention Output Projection: $\mathbf{B} \times \mathbf{T} \times \mathbf{D}$
 - * RMSNorm2 Input: $\mathbf{B} \times \mathbf{T} \times \mathbf{D}$
 - * FFN W1 Output: $\mathbf{B} \times \mathbf{T} \times \mathbf{d}_{ff}$
 - * FFN W2(SiLU) Output: $\mathbf{B} \times \mathbf{d}_{ff} \times \mathbf{T}$
 - * FFN W3 Output: $\mathbf{B} \times \mathbf{T} \times \mathbf{d}_{ff}$

In total, the activations for one block are $20(\mathbf{BTD}) + 2(\mathbf{BHT}^2)$.

- **Final Layers:** Final Norm Input: $\mathbf{B} \times \mathbf{T} \times \mathbf{D}$, Output Projection (logits): $\mathbf{B} \times \mathbf{T} \times \mathbf{V}$ So the total memory for activations ($\mathbf{N}_{\text{activations}}$) is:

$$\mathbf{N}_{\text{activations}} = \mathbf{N} * [20(\mathbf{BTD}) + 2(\mathbf{BHT}^2)] + \mathbf{BTD} + \mathbf{BTV}$$

$$\boxed{\mathbf{Memory}_{\text{activations}} = \mathbf{N}_{\text{activations}} * 4 \text{ bytes}}$$

Total Peak Memory:

$$\mathbf{Memory}_{\text{total}} = \mathbf{Memory}_{\text{params}} + \mathbf{Memory}_{\text{grads}} + \mathbf{Memory}_{\text{optimizer}} + \mathbf{Memory}_{\text{activations}}$$

$$\mathbf{Memory}_{\text{total}} = \mathbf{N}_{\text{params}} * 4 + \mathbf{N}_{\text{params}} * 4 + \mathbf{N}_{\text{params}} * 8 + \mathbf{N}_{\text{activations}} * 4$$

$$\boxed{\mathbf{Memory}_{\text{total}} = 16 * \mathbf{N}_{\text{params}} + 4 * \mathbf{N}_{\text{activations}} \text{ bytes}}$$

- b. For GPT-2 XL model, we have

$$B = \text{batch_size},$$

$$T = 1024,$$

$$D = 1600,$$

$$H = 25,$$

$$V = 50257,$$

$$N = 48,$$

$$d_{ff} = 6400$$

- **Parameters:** $\mathbf{N}_{\text{params}} = 2(VD) + (TD) + N * [16(D^2) + 2(D)] + D \approx 2.13B$ parameters, which amounts to approximately **8.5 GB** of memory (4 bytes/parameter).
- **Gradients:** Same as parameters, so approximately **8.5 GB** of memory.
- **Optimizer State:** 2 additional values per parameter, so approximately **17 GB** of memory.
- **Activations:** $\mathbf{N}_{\text{activations}} = N * [20(BTD) + 2(BHT^2)] + BTD + BTV$

Taking B out of the equation, we have

$$\mathbf{N}_{\text{activations}} = B * [N * (20TD + 2HT^2) + TD + TV]$$

For GPT-2 XL, this is approximately $\mathbf{B} * 4.1\mathbf{B}$ activations, which amounts to approximately **16.4 GB * B** of memory.

So the required expression:

$$\boxed{\mathbf{Memory}_{\text{total}}(\text{GB}) = 16.4 * \mathbf{B} + 34}$$

For this to fit in 80 GB memory, we need batch_size $\mathbf{B} \leq 2.8$, so the maximum batch size we can use is **2**.

- c.

$$\boxed{\mathbf{FLOPs}_{\text{AdamW}} \approx 15 * \mathbf{N}_{\text{params}}}$$

Justification: AdamW performs element-wise operations on the parameters. For each parameter, the update involves:

- Updating first moment m : **3 operations** (mul, mul, add).

- Updating second moment v : **4 operations** (mul, mul, mul, add).
- Parameter update θ : **8 operations** (sqrt, add, div, mul, sub, mul, mul, sub).

Total is around 15 FLOPs per parameter.

- d. • **Total Training FLOPs:** $400k \times B \times 6 \times N_{params}$. Here 6 comes from 2 forward passes and 4 for backward passes.

Total FLOPs = $400k \times 1024 \times 6 \times 2.13B \approx \mathbf{5.2 \times 10^{18}}$ FLOPs.

- **Hardware Throughput:** Peak throughput = 19.5 teraFLOPs/s. With 50% MFU, effective throughput = 9.75 teraFLOPs/s.
- **Total Training Time:** $\frac{5.2 \times 10^{18} \text{ FLOPs}}{9.75 \times 10^{12} \text{ FLOPs/s}} \approx \mathbf{533,333}$ seconds $\approx \mathbf{148.15}$ hours $\approx \mathbf{6.17}$ days.

6 Generating text

- The decoding/generation script lives at `student/generate.py`.