Calculating the Total Number of Parameters in Llama 2

Llama 2 is a family of large language models developed by Meta AI, available in sizes of 7 billion, 13 billion, and 70 billion parameters. In this calculation, we'll focus on the **Llama 2 70B** model and break down how its parameters are computed based on its architecture.

Overview of Llama 2 Architecture

Llama 2 utilizes the Transformer architecture with some optimizations:

1. **Embedding Layers:**
   - **Token Embeddings**
   - **Positional Embeddings (Rotary Position Embeddings - RoPE)**
2. **Transformer Blocks (Repeated N times):**
   - **Multi-Head Self-Attention**
   - **Feed-Forward Networks (FFN) with Gated Linear Units (SwiGLU)**
   - **RMS Layer Normalization**
3. **Output Layer:**
   - Often tied with the input embeddings.

Key Parameters and Dimensions

Based on the Llama 2 70B model specifications:

- **Number of Layers (N):** 80
- **Model Dimension (d_{model}):** 8,192
- **Feed-Forward Dimension (d_{ff}):** 28,672
- **Number of Attention Heads (h):** 64
- **Head Dimension (d_k and d_v):** $d_{\text{model}}/h = 8,192/64 = 128$
- **Vocabulary Size (V):** 32,000
- **Maximum Sequence Length (L):** 2,048

---

1. Embedding Layers

Token Embeddings

- **Parameters:** $V \times d_{\text{model}}$
- **Calculation:** $32,000 \times 8,192 = 262,144,000$ parameters

Positional Embeddings

- **Parameters:** Negligible, as Llama 2 uses RoPE, which doesn't add learned parameters.

Total Embedding Parameters

- **Total: 262,144,000** parameters

---

2. Transformer Blocks

Each of the 80 layers contains:

A. Multi-Head Self-Attention

i. Query, Key, and Value Matrices

- **Parameters per matrix:** $d_{\text{model}} \times d_{\text{model}}$
- **Total for Q, K, V:**

$$3 \times (8,192 \times 8,192) = 3 \times 67,108,864 = 201,326,592 \text{ parameters}$$

ii. Output Projection Matrix

- **Parameters:** $d_{\text{model}} \times d_{\text{model}} = 8,192 \times 8,192 = 67,108,864$ parameters

Total Attention Parameters per Layer

- **Total:** $201,326,592 + 67,108,864 = 268,435,456$ parameters

B. Feed-Forward Network (FFN) with SwiGLU

Llama 2 uses the SwiGLU activation function, which affects the parameter count.

i. First Linear Layer

- **Parameters:** $d_{\text{model}} \times (2 \times d_{\text{ff}}) = 8,192 \times 57,344 = 470,810,624$ parameters

ii. Second Linear Layer

- **Parameters:** $d_{\text{ff}} \times d_{\text{model}} = 28,672 \times 8,192 = 235,405,312$ parameters

Total FFN Parameters per Layer

- **Total:** $470,810,624 + 235,405,312 = 706,215,936$ parameters

C. RMS Layer Normalization

- **Parameters:** Negligible (usually d_{model} per layer)

Total Parameters per Transformer Block

- **Total:** $268,435,456$ (Attention) $+706,215,936$ (FFN) $=$ **974,651,392** parameters

Total Parameters for All Transformer Blocks

- **Total:** $974,651,392 \times 80 = 77,972,111,360$ parameters

---

3. Output Layer

- **Parameters:** Typically tied with token embeddings; additional parameters are minimal.

---

4. Summing Up All Parameters

Total Parameters

- **Embedding Layers:** 262,144,000 parameters
- **Transformer Blocks:** 77,972,111,360 parameters
- **Output Layer:** Minimal (due to weight tying)

Grand Total:

$$\begin{aligned}\text{Total Parameters} &= \text{Embedding Layers} + \text{Transformer Blocks}\\ &= 262,144,000 + 77,972,111,360\\ &= 78,234,255,360 \text{ parameters}\end{aligned}$$

---

5. Accounting for Minor Components

- **RMS Layer Normalization and Bias Terms:** Although considered negligible individually, across all layers, they add up:
  - **RMSNorm Parameters per Layer:** $d_{\text{model}} = 8,192$
  - **Total RMSNorm Parameters:** $2 \times 80 \times 8,192 = 1,310,720$ parameters (since RMSNorm is applied before Attention and FFN)
- **Bias Terms:** May add additional parameters.

Adjusted Total Parameters

Adding these minor components:

$$\text{Adjusted Total} = 78,234,255,360 + 1,310,720 = 78,235,566,080 \text{ parameters}$$

Conclusion

By summing the parameters from the embedding layers, transformer blocks, and minor components, we arrive at an approximate total of **78 billion parameters** for the Llama 2 70B model.

- **Discrepancy with "70B":** The model is named "70B" for simplicity, but the actual parameter count is higher due to architectural choices like larger feed-forward dimensions and the use of SwiGLU activations.

Note: This calculation is based on publicly available information from Meta AI's

Llama 2 release. The exact parameter count may vary slightly due to implementation details not covered in this estimation.