Calculating the Total Number of Parameters in a Hypothetical Llama3 405B Model

As of in October 2023, Llama 3 has not been released, and there is no publicly available information about a Llama 3 model with 405 billion parameters. The latest version available is Llama 2, which comes in sizes up to 70 billion parameters.

However, the following is a hypothetical calculation of the total number of parameters for a transformer-based model with 405 billion parameters, based on scaling principles similar to those used in previous Llama models.

Hypothetical Model Overview

Assuming the model follows the standard Transformer architecture used in Llama 2, we can estimate the parameters by considering:

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

Key Hypothetical Parameters and Dimensions

To reach approximately 405 billion parameters, we can adjust the following dimensions:

- **Number of Layers (N):** 80
- **Model Dimension (d_{model}):** 20,480

- **Feed-Forward Dimension (d_{ff}):** $4 \times d_{\text{model}} = 81,920$
- **Number of Attention Heads (h):** 128
- **Head Dimension (d_k and d_v):** $d_{\text{model}}/h = 160$
- **Vocabulary Size (V):** ~32,000
- **Maximum Sequence Length (L):** 2,048

---

1. Embedding Layers

Token Embeddings

- **Parameters:** $V \times d_{\text{model}}$
- **Calculation:** $32,000 \times 20,480 = 655,360,000$ parameters

Positional Embeddings

- **Parameters:** $L \times d_{\text{model}}$
- **Calculation:** $2,048 \times 20,480 = 41,943,040$ parameters

Total Embedding Parameters

- **Total:** $655,360,000 + 41,943,040 = 697,303,040$ 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:** $3 \times (20,480 \times 20,480) = 1,258,291,200$ parameters

ii. Output Projection Matrix

- **Parameters:** $d_{\text{model}} \times d_{\text{model}} = 20,480 \times 20,480 = 419,430,400$ parameters

Total Attention Parameters per Layer

- **Total:** $1,258,291,200 + 419,430,400 = 1,677,721,600$ parameters

B. Feed-Forward Network (FFN)

i. First Linear Layer

- **Parameters:** $d_{\text{model}} \times d_{\text{ff}} = 20,480 \times 81,920 = 1,677,721,600$ parameters

ii. Second Linear Layer

- **Parameters:** $d_{\text{ff}} \times d_{\text{model}} = 81,920 \times 20,480 = 1,677,721,600$ parameters

Total FFN Parameters per Layer

- **Total:** $1,677,721,600 + 1,677,721,600 = 3,355,443,200$ parameters

C. Layer Normalization

- **Parameters:** Negligible compared to other components (usually $2 \times d_{\text{model}}$ per layer)

Total Parameters per Transformer Block

- **Total:** $1,677,721,600$ (Attention) $+ 3,355,443,200$ (FFN) $=$ **5,033,164,800** parameters

Total Parameters for All Transformer Blocks

- **Total:** $5,033,164,800 \times 80 = 402,653,184,000$ parameters

---

3. Output Layer

- **Parameters:** Typically weight-tied with token embeddings, so additional parameters are minimal.

---

4. Summing Up All Parameters

Total Parameters

- **Embedding Layers:** 697,303,040 parameters
- **Transformer Blocks:** 402,653,184,000 parameters
- **Output Layer:** Minimal

Grand Total:

$$\begin{aligned}\text{Total Parameters} &= \text{Embedding Layers} + \text{Transformer Blocks} \\ &= 697,303,040 + 402,653,184,000 \\ &= 403,350,487,040 \text{ parameters}\end{aligned}$$

This total is approximately **403 billion parameters**, close to the target of 405 billion.

Conclusion

By adjusting the model dimensions, we can estimate a hypothetical Llama3 model with approximately 405 billion parameters. The calculation shows that with:

- **Model Dimension (d_{model})** of 20,480
- **Feed-Forward Dimension (d_{ff})** of 81,920
- **Number of Layers (N)** of 80

We achieve a total parameter count of approximately **403 billion** parameters.

Note: This is a hypothetical estimation based on standard transformer scaling laws. The actual architecture of a potential Llama3 405B model may differ significantly.