目录

- 隐藏层大小（`hidden_size`）：768
- 注意力头数（`num_attention_heads`）：12
- 中间层大小（`intermediate_size`）：3072
- `Transformer` 层数（`num_layers`）：12
- 词汇表大小（`vocab_size`）：21128（`bert-base-chinese`）
- 最大位置编码（`max_position_embeddings`）：512
- 标记类型数（`type_vocab_size`）：2

0.1. 嵌入层（Embedding Layer）

- 词嵌入（`Word Embeddings`）：$\text{vocab\_size} \times \text{hidden\_size} = 21128 \times 768$
- 位置嵌入（`Position Embeddings`）：$\text{max\_position\_embeddings} \times \text{hidden\_size} = 512 \times 768$
- 标记类型嵌入（`Token Type Embeddings`）：$\text{type\_vocab\_size} \times \text{hidden\_size} = 2 \times 768$
- 嵌入层归一化（`LayerNorm`）：权重和偏置各 `hidden_size`（$2 \times \text{hidden\_size}$）

计算:

$21128 \times 768 + 512 \times 768 + 2 \times 768 + 2 \times 768 = (21128 + 512 + 2) \times 768 + 2 \times 768 = 21642 \times 768 + 1536$

$21642 \times 768 = 16,621,056$（嵌入表参数）

$16,621,056 + 1,536 = 16,622,592$（嵌入层总参数）

0.2. Transformer 层（共 12 层）

每层包含以下部分:

0.2.1. 自注意力机制（Self-Attention）

- 查询变换（`Query`）：权重 $\text{hidden\_size} \times \text{hidden\_size}$，偏置 `hidden_size`
- 键变换（`Key`）：同上
- 值变换（`Value`）：同上
- 输出变换（`Output`）：同上

共 4 个矩阵:

$4 \times (\text{hidden\_size}^2 + \text{hidden\_size}) = 4 \times (768^2 + 768)$

0.2.2. 前馈网络（Feed Forward Network）

- 中间层（`Intermediate`）：权重 $\text{intermediate\_size} \times \text{hidden\_size}$，偏置 `intermediate_size`
- 输出层（`Output`）：权重 $\text{hidden\_size} \times \text{intermediate\_size}$，偏置 `hidden_size`

$(\text{intermediate\_size} \times \text{hidden\_size} + \text{intermediate\_size}) + (\text{hidden\_size} \times \text{intermediate\_size} + \text{hidden\_size})$

0.2.3. 层归一化（LayerNorm，共 2 个）

- 自注意力后归一化：权重和偏置各 `hidden_size`（`2 × hidden_size`）
- 前馈网络后归一化：同上

$2 \times (2 \times \text{hidden\_size}) = 4 \times \text{hidden\_size}$

单层 Transformer 计算:

$4 \times (768^2 + 768) = 4 \times (589,824 + 768) = 4 \times 590,592 = 2,362,368$ (自注意力)

$(3072 \times 768 + 3072) + (768 \times 3072 + 768) = (2,359,296 + 3,072) + (2,359,296 + 768) = 4,722,432$ (前馈网络)

$4 \times 768 = 3,072$ (归一化层)

单层总计:

$2,362,368 + 4,722,432 + 3,072 = 7,087,872$

12 层总计:

$12 \times 7,087,872 = 85,054,464$

0.3. 池化层（Pooler Layer）

• 全连接层（`Dense`）: 权重 $\text{hidden\_size} \times \text{hidden\_size}$，偏置 `hidden_size`

$\text{hidden\_size}^2 + \text{hidden\_size} = 768^2 + 768 = 590,592$

0.4. 总参数量

嵌入层: 16,622,592

`Transformer` 层: 85,054,464

池化层: 590,592

总计: $16,622,592 + 85,054,464 + 590,592 = 102,267,648$

`bert-base-chinese` 的总参数量为 102,267,648（约 102.3M）。