Bert 中的可训练参数数量

--- Embedding 层

Token Embedding 词数量 × 隐层数 vocabulary\_size \* hidden\_size

Segment Embedding 2 \* hidden\_size

Position Embedding 句子长度 × 隐层数 length \* hidden\_size

Layer\_norm 层 ( 1 + 1 ) \* hidden\_size

合计 hidden\_size \* ( vocab\_size + 4 + length )

--- Self-attention

Attention 部分

Q、K、V 矩阵 隐层数 × 隐层数 hidden\_size \* hidden\_size \* 3

Bias 句子长度 × 隐层数 1 \* hidden\_size \* 3

Attention 后的线性层

矩阵 隐层数 × 隐层数 hidden\_size \* hidden\_size

Bias 句子长度 × 隐层数 1 \* hidden\_size

Layer\_norm 层 ( 1 + 1 ) \* hidden\_size

合计 hidden\_size \* ( 4 hidden\_size + 6 )

--- 前馈

第一个线性层

矩阵 隐层数 × 4 倍隐层数 hidden\_size \* 4 hidden\_size

Bias 句子长度 × 4 倍隐层数 1 \* 4 hidden\_size

第二个线性层

矩阵 4 倍隐层数 × 隐层数 4 hidden\_size \* hidden\_size

Bias 句子长度 × 隐层数 1 \* hidden\_size

Layer\_norm 层 ( 1 + 1 ) \* hidden\_size

--- pool\_fc 层 ( hidden\_size + 1 ) \* hidden\_size

合计 hidden\_size \* ( 9 hidden\_size + 8 )

--- 总计

hidden\_size \* ( vocabulary\_size + length + 18 + 13 \* hidden\_size )