对于一个具有L层的Transformer，每层包含H个注意力头，每个头的维度为dk​，总的隐藏层维度为dmodel​，词汇表大小为V

P=Positional Embeddings 维度

S=Segment Embeddings 维度

Df=中间层维度

*Pm=V*×*dmodel*​+*P*×*dmodel*​+*S*×*dmodel*​+4×*H*×*dmodel*​×*dk*​+2×*L*×2×*dmodel*​+2×(*dmodel*​×*df*​+*df*​+*dmodel*​)×*L*