1. 词表大小 word $V\times h$ (7)

2. 设单句话长度为 L

则 BERT结构中 Embedding

Token Embeddings 大小为 Token $L\times h$ ①

Segment Embeddings 大小为 seg $1\times h$ ②

Position Embeddings 大小为 Pos $L\times h$ ③

embeddings = ①+②+③ 为 embeddings $L\times h$.

归一化层 el.w $h\times 1$ el.b $h\times 1$ ④

Embedding 层中总参数为

$E_{总}$ = ①+②+③+④ = $L\times h + 1\times h + L\times h + h\times 1 + h\times 1$

$= (2L+3)h$ ---- (1)

3 BERT-self-attention

Qw ($h\times h$) Qb ($h\times 1$) ⑤

Kw ($h\times h$) Kb ($h\times 1$) ⑥

Vw ($h\times h$) Vb ($h\times 1$) ⑦

Output层 Ow ($h\times h$) Ob ($h\times 1$) ⑧

self-attention层训练参数为:

S合: $4\times h\times h + 4\times h\times 1 = 4\times h\times(1+h)$ ····(2)
$= 4h(1+h)$

4 layer norm (Xmbedding + Xattention)

E.b.w $(h\times 1)$, E.L.b $(h\times 1)$

EL合 $= 2\times h = 2h$ ------ (3)

5. output = liner(gelu(Liner(x)))

liner(x) $= X_{L\times h}\cdot W_{h\times 4h} + b_{4h\times 1} = X_{L\times 4h}$

og.l.w $(h\times 4h)$ og.l.b $(4h\times 1)$ ···· ①

liner(gelu(liner(x))) $= X_{L\times 4h}\cdot W_{4h\times h} + b_{h\times 1}$

o.l.w $(4h\times h)$ o.l.b $= h\times 1$ ··· ②

output合 $= ① + ② = 4h\times h\times 2 + 4h\times 1 + h\times 1$

$= h(8h+5)$ ···(4)

6 layernorm (Xforward + Xattention)

l.w $(h\times 1)$ l.b $(h\times 1)$

L合 $= 2h$ ~ — ~ — (5)

7. Pooler. dense.

p.w $(h\times h)$ p.b $(h\times 1)$

P总 = h(h+1) ...... (b)

BERT结构一层的参数

B总 = (1) + (2) + ... + (17)
= (2L+3)h + 4(1+h)h + 2h + (5+8h)h + 2h + (h+1)h + Vh
= (2L+3+4+4h+2+5+8h+2+h+1+V)h
= (2L+13h+V+15)h