摘要

ELMo
使用的是RNN的架构，所以在应用到下游任务的时候需要进行一些架构的调整

深的双向transformers

GPT单向的左边的去预测未来

Bert使用了双向的信息
同在下游任务只需要改上层（类似GPT）

①与哪些文章有关并且区别在哪 ②好的点绝对精度 相对好的程度

导言：

feature-based ELMo（RNN的架构）

fine-tuning GPT 从左往右的架构（transformer架构）

- 带掩码的语言模型
- 下一个句子的预测任务

随机预测未来而是完形填空 cloze task

贡献：
- 双向信息的重要性
- 微调

大量的没有标号的数据来上去训练比在只有少量有标号的数据上训练的效果更好

预训练：在没有标号的数据上进行训练

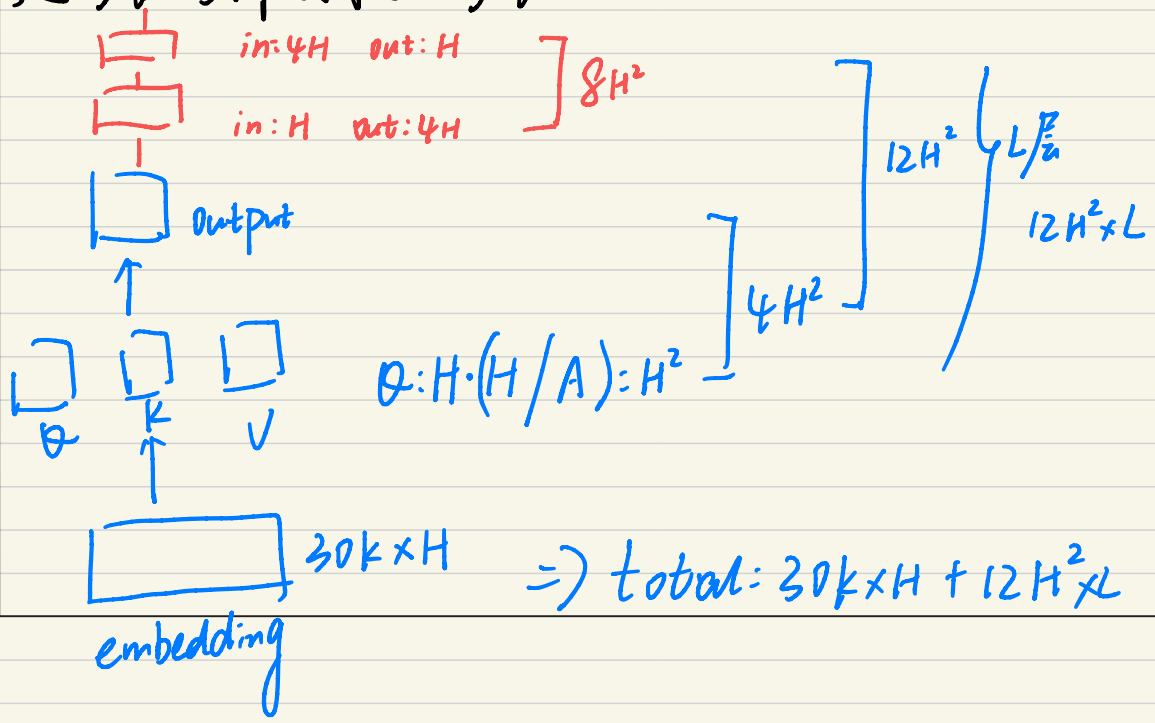
微调：在下游任务使用有标号的数据进行微调

L, H, A Bert-base

hidden heads Bert-large

模型的宽度：

超参数决定成可学习参数的大小



wordPiece（分序列）30k

bert 输入是序列

transformer 输入是序列对

[CLS] [SEP]

bert中 pos_emb 和 Seg_emb 都是通过学习得来的

MLM

Bert主要为是然后不能生成