

Bert

知道了 transformer 之后再理解 bert 就简单多了。

bert 思想:

1. 将下游的完形填空和下一个句子预测放在了训练词向量的任务中

bert 损失函数:

预测 mask 值的损失函数和预测下一个句子的二分类的损失函数之和

BERT的损失函数由两部分组成，第一部分是来自 Mask-LM 的单词级别分类任务，另一部分是句子级别的分类任务。通过这两个任务的联合学习，可以使得 BERT 学习到的表征既有 token 级别信息，同时也包含了句子级别的语义信息。具体损失函数如下：

$$L(\theta, \theta_1, \theta_2) = L_1(\theta, \theta_1) + L_2(\theta, \theta_2)$$

其中 θ 是 BERT 中 Encoder 部分的参数， θ_1 是 Mask-LM 任务中在 Encoder 上所接的输出层中的参数， θ_2 则是句子预测任务中在 Encoder 接上的分类器参数。因此，在第一部分的损失函数中，如果被 mask 的词集合为 M ，因为它是一个词典大小 $|V|$ 上的多分类问题，那么具体说来有：

$$L_1(\theta, \theta_1) = - \sum_{i=1}^M \log p(m = m_i | \theta, \theta_1), m_i \in [1, 2, \dots, |V|]$$

在句子预测任务中，也是一个分类问题的损失函数：

$$L_2(\theta, \theta_2) = - \sum_{j=1}^N \log p(n = n_j | \theta, \theta_2), n_j \in [IsNext, NotNext]$$

bert 怎样做预训练

这个博客上内容非常多了。不写了。

bert 怎样做微调



