Uda实验报告

1、样本数极少情况

训练集：每类各取10个样本，共30个样本

测试集：共5000样本

无监督数据集：约20000个样本

实验baseline使用预训练的bert模型，准确率为0.9303

uda设置：

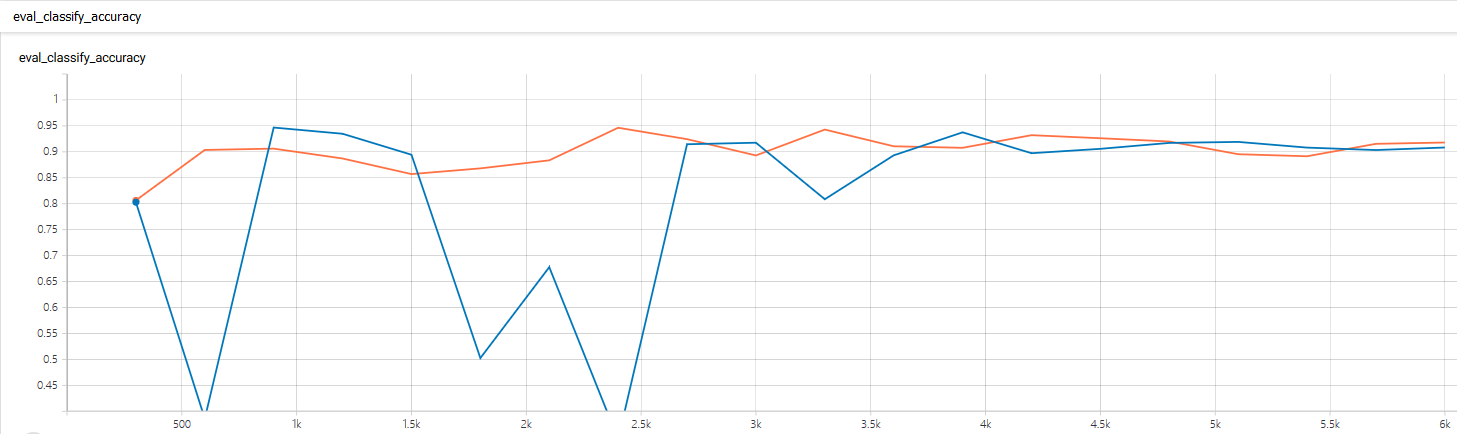
tsa：exp\_schedule

train\_batch\_size：4

unsup\_batch\_size：16

Global\_steps:6000

测试了无监督数据置信度阈值的影响



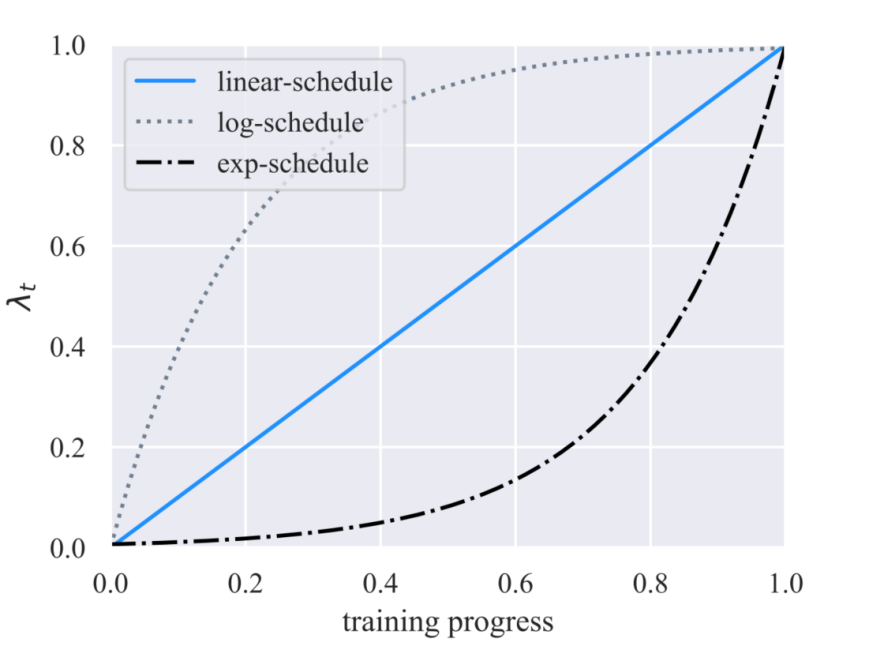
蓝色线阈值为-1（即不对无监督数据做限制），测试集准确率最高为0.9464

橙色线阈值为0.4（即只训练概率大于0.4的数据），测试集准确率最高位0.9468

结论：

在数据量较少的在样本数极少的情况下，如果无监督数据集的质量不高，会给模型带了很大的噪声，严重影响模型的收敛过程，通过设置阈值的方式来缓解这种情况是有效的。

1. 测试tsa模式的影响



设置tsa模式的作用是为了让模型缓慢的接受监督数据，从而大大缓解过拟合问题

当模型易于过拟合时，例如问题相对容易或标记示例的数量非常有限时，由于在训练结束时大部分释放了监督信号，因此exp模式是最合适的。 按照类似的逻辑，当模型不太可能过拟合时（例如，当我们有大量带标签的示例或模型采用有效的正则化时），log模式可以发挥更好的作用。

训练集：每类各取40个样本，共120个样本

测试集：共5000样本

无监督数据集：约20000个样本

实验baseline使用预训练的bert模型，准确率为0.9388

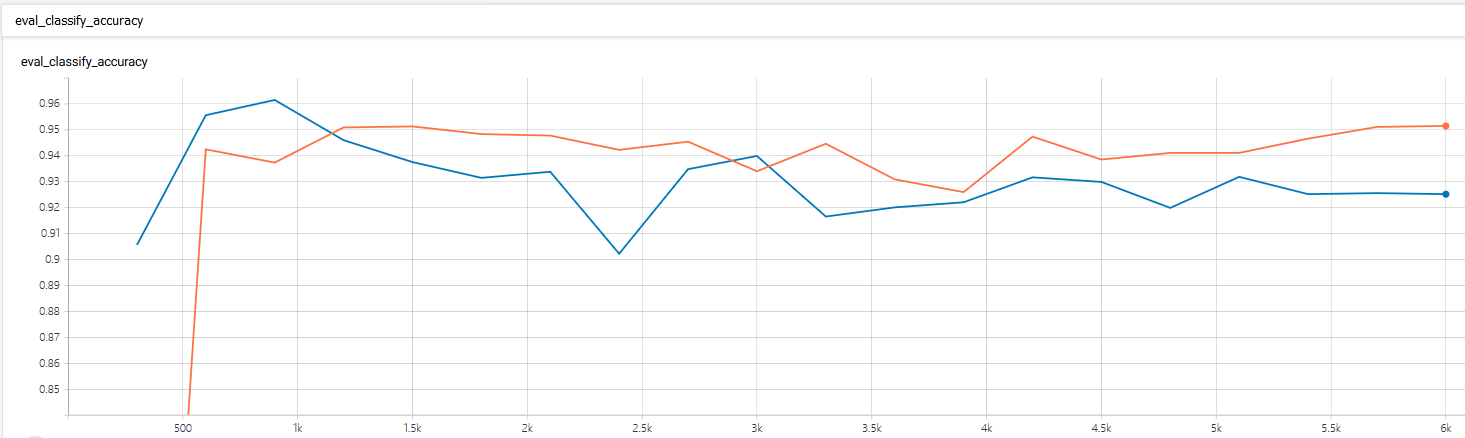
uda设置：

train\_batch\_size：8

unsup\_batch\_size：24

Global\_steps:6000

分别测试了linear\_schedule和log\_schedule模式在样本数较少情况下的影响



蓝色线为log\_schedule模式，测试集准确率最高为0.9615

橙色线为linear\_schedule模式，测试集准确率最高为0.9515

结论：

从模型训练过程来看，虽然log模式的最高准确率要高于linear模式，但是linear模式下收敛更加稳定，且测试集准确率基本稳定高于baseline，而log模式下在开始是性能最优，后期逐渐下降，甚至低于了baseline，因为log模式对于监督数据的启动速度更快，在数据量较少的情况下，会增大监督数据噪声对无监督数据的影响。

在实际实验过程中，三种模式对于训练过程的影响会随着数据发生变化，不能仅以数据量进行选择，不过从多次实验的结果来看，linear模式的性能是最稳定的，并且效果很好，实验开始时可以优先选在linear模式进行训练。

1. 数据量较多的情况下

训练集：每类各取1000个样本，共3000个样本

测试集：共5000样本

无监督数据集：约20000个样本

实验baseline使用预训练的bert模型，准确率为0.974

uda设置：

train\_batch\_size：16

unsup\_batch\_size：48

Global\_steps:6000

实验中可提升的点：

1. 增加数据多样性：实验过程只使用了中文与英文的互译来作为数据增强方式，可以采用多用语言之间的转换来提高增强数据的多样性，但要考虑对硬件的要求，可以使用随机采样的方式来读取增强数据，既保证了数据的多样性，也能够提供更多的数据，可以根据实际条件控制比重。
2. 保证数据的有效性：不论增强的数据量多少，增强数据的有效性都要保证，因为无监督数据本身就存在噪声，如果不保证有效性可能会适得其反，可以通过删除一些过短的增强数据或者与原始数据长度相差过大的增强数据，来提高数据增强过程的有效性
3. 针对回译这种增强方式，使用的数据最好长度不要太短，并且语义分明，如果是主要通过关键词进行判断，而样本本身并没有明确的语义，甚至是不连续，uda的性能会大大折扣，甚至如果导致回译的数据出现语义的偏移，可能会带来负面效果，可以尝试使用一些数据处理的方式将增强的数据处理一下
4. 无监督数据中样本不平衡的问题也需要进行注意，可能会有影响，因为exp模式性能较差，说明在模型缓慢接受监督信息的时候收到了无监督信息的负面影响
5. 因为使用的数据为三分类，本身并不复杂，只使用bert准确率已经很高，在由于监督数据量较少导致模型性能较差的情况下，uda的效果会明显一点
6. 训练过程中需要设置每一个batch中监督数据和无监督数据的数量以及比重，无监督数据最好不要比监督数据多太多，尤其是回译质量不高的情况下，可以适当减少，如果质量比较高，适合uda的情况，可以适当增加无监督数据的数量，