【关于 用检索的方式做文本分类 】 那些你不知道的事

一、动机

1. 标签类别较多：常规的分类模型只是预测10几个 label，但是真实工业界有时需要预测的 label 可能上百个，这种情况 一方面标注数据会存在严重不均衡现象，另一方面模型的预测结果也存在偏差性严重问题；
2. 标签类别不固定：面对业务需求需要不断添加新标签时，容易导致每次新增标签都面临重新训练分类器模型问题；
3. 语义信息丢失：以前的分类任务中，标签信息作为无实际意义，独立存在的one-hot编码形式存在，这种做法会潜在的丢失标签的语义信息；

那有什么方法可以解决该问题么?

答案：**基于检索的方法做文本分类**

二、基于检索的方法做文本分类思路

2.0 基于检索的方法做文本分类 整体思路

把文本分类任务中的标签信息转换成含有语义信息的语义向量，将文本分类任务转换成向量检索和匹配的任务。

2.1 检索的方法的召回库如何构建

因为采用 检索的方法 去做 文本分类，首先就是需要考虑如果 构建 召回库 问题，构建方法：

1. 把标签集作为召回库；
2. 把训练数据作为召回库；

2.2 检索的方法 的 训练阶段 如何做?

无论是 **把标签集作为召回库** 还是 **把训练数据作为召回库**，其训练阶段的方法都是相同的，都可以采用 **双塔模型**；

1. 塔构建思路：

- 一个塔（模型）：用于对 输入句子 进行 encoding;
- 一个塔（模型）：用于对 输入标签 进行 encoding;

2. 目标：通过训练拉近句子和标签的[CLS]输出特征表示之间距离；

2.3 检索的方法 的 预测阶段 如何做?

把标签集作为召回库 和 **把训练数据作为召回库** 思路存在不同。

2.3.1 把标签集作为召回库 的预测思路

1. 把标签作为召回集，每个标签的向量表示（也即[CLS]输出特征表示）是固定的；
2. 然后构建一个标签向量库；
3. 用待预测的句子的向量在标签向量库进行检索，找到特征相似度最大的标签，也即为待预测句子的标签。

2.3.2 把训练数据作为召回 的预测思路

1. 把训练数据作为召回集，构建一个训练集文本的向量库;
2. 用待预测的句子的向量表示（也即[CLS]输出特征表示）在文本向量库进行检索，找到特征相似度最大的训练集文本，待预测句子的标签也即召回文本的标签。

三、用检索的方式做文本分类 方法 适用场景

1. 对于一些类别标签不是很固定的场景，或者需要经常有一些新增类别的需求的情况非常合适；
2. 对于一些新的相关的分类任务，这种方法也不需要模型重新学习或者设计一种新的模型结构来适应新的任务。

总的来说，这种基于检索的文本分类方法能够有很好的拓展性，能够利用标签里面包含的语义信息，不需要重新进行学习。这种方法可以应用到相似标签推荐，文本标签标注，金融风险事件分类，政务信访分类等领域。

参考

1. 【NLP从零入门】预训练时代下，深度学习模型的文本分类算法（超多干货，小白友好，内附实践代码和文本分类常见中文数据集）