基于 Facebook 多语言处理

具体流程：

利用 facebook 预先训练的多语言的词向量，在训练集和测试集为不同的语言情况下，训练情感分析模型。

核心原理是 Facebook 已经训练了在同一向量空间中下 30 种语言的多语言向量而且 Facebook 训练的词向量是每个语言频率前 200000 个词汇和 300 维。因为在同一空间下，原有语言的词汇间的距离与现有语言词汇间的距离被转换成近似的距离，使得单语言模型能够直接通过词汇间距离来进行预测。

算法基于 Convolutional Neural Networks for Sentence Classification pytorch 实现版本。

算法基本结构

输入层

输入层是句子中的词语对应的 word vector 依次（从上到下）排列的矩阵，假设句子有 n 个词，vector 的维数为 k ，那么这个矩阵就是 n*k 的。

这个矩阵的类型可以是静态的(static)，也可以是动态的(non static)。静态就是 word vector 是固定不变的，而动态则是在模型训练过程中，word vector 也当做是可优化的参数，通常把反向误差传播导致 word vector 中值发生变化的这一过程称为 Fine tune。

对于未登录词的 vector，可以用 0 或者随机小的正数来填充。

第一层卷积层

输入层通过卷积操作得到若干个 Feature Map，卷积窗口的大小为 h*k ，其中 h 表示纵向词语的个数，而 k 表示 word vector 的维数。通过这样一个大型的卷积窗口，将得到若干个列数为 1 的 Feature Map。

池化层

接下来的池化层，文中用了一种称为 Max-over-time Pooling 的方法。这种方法就是简单地从之前一维的 Feature Map 中提出最大的值，文中解释最大值代表着最重要的信号。可以看出，这种 Pooling 方式可以解决可变长度的句子输入问题（因为不管 Feature Map 中有多少个值，只需要提取其中的最大值）。

最终池化层的输出为各个 Feature Map 的最大值们，即一个一维的向量。

全连接 + Softmax 层

池化层的一维向量的输出通过全连接的方式，连接一个 Softmax 层，Softmax 层可根据任务的需要设置（通常反映着最终类别上的概率分布）。

最终实现时，我们可以在倒数第二层的全连接部分上使用 Dropout 技术，即对全连接层上的权值参数给予 L2 正则化的限制。这样做的好处是防止隐藏层单元自适应（或者对称），从而减轻过拟合的程度。

测试结果：

Dev acc:抽取 10%的训练语言数据（非包含进训练集）的准确率

Test acc: 不同语言的测试集合准确率

数据量：

法语 2125 条（positive920,negative596 netural 1609）

阿拉伯语 3543 条(positive 1063,negative 1304,netural 1176)

迭代次数	训练集语言	测试集语言	Dev accuracy（%）	Test accuracy（%）	Facebook 预训练向量
100	法语	阿拉伯语	51	54	有
100	法语	阿拉伯语	54	39	无
100	阿拉伯语	法语	80	42	有
100	阿拉伯语	法语	79	38	无

迭代次数	训练集语言	测试集语言	Dev accuracy（%）	Test accuracy（%）	Facebook 预训练向量
10	法语	阿拉伯语	55	61	有
10	法语	阿拉伯语	48	38	无
10	阿拉伯语	法语	77	47	有
10	阿拉伯语	法语	48	3	无

综上结果看出，在使用 Facebook 的多语言预训练词向量后预测非训练集语言的测试集有5~25%准确率左右的提升，证明 Facebook 的多语言向量应该可以用于在无其他语言数据下用单语言算法模型预测多语言短文本情感。 同时在测试中发现有接近一半的词汇是未登录词导致训练效果不佳，估计一些标点符号连接着词汇和表情没有被识别出来，导致准确率不高，原模型的英文数据预测可以达到 80%。如果数据进行进一步处理，可以有更好的效果。