《人工智能导论》实验三设计说明

主讲人：马少平教授
助教：李祥圣

清华大学计算机系人工智能研究所

1. 任务简介

情感分析（sentiment analysis）是近年来国内外研究的热点，其任务是帮助用户快速获取、整理和分析相关评价信息，对带有情感色彩的主观性文本进行分析、处理、归纳和推理。

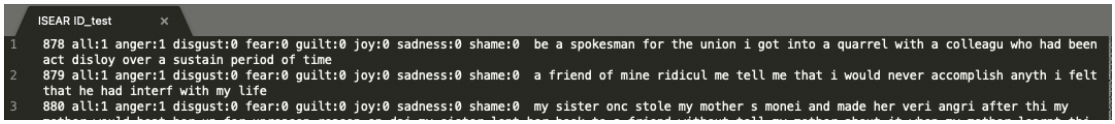
随着互联网技术的迅速发展和普及，对网络内容管理、监控和有害（或垃圾）信息过滤的需求越来越大，网络信息的主观倾向性分类受到越来越多的关注。这种分类与传统的文本分类不同，传统的文本分类所关注的是文本的客观内容（objective），而倾向性分类所研究的对象是文本的“主观因素”，即作者所表达出来的主观倾向性，分类的结果是对于一个特定的文本要得到它是否支持某种观点的信息。这种独特的文本分类任务又称为**情感分类（sentiment classification）**。

图 1：新闻中的情感投票例子

2. 实验数据

实验数据是英文问卷数据，由标注人员对提取的句子进行情感打分，共 7 种情感类别。每行数据由 ID，情感标签，文本三个部分组成，以 tab(\t)分割。ISEAR ID 文件为整个数据集，共 7666 条数据，标号为 1~7666，单标签。整个数据集分为 train, validation, test 三部分：train 共 4600 条数据，validation/test 分别有 1533 条数据，数据描述参照 readme.txt。

图 2：数据集样例



3. 数据处理

1） **文本表示方法：**

a) Bags-of-words，是信息检索领域常用的文档表示方法。对于一个文档，计算**每个词出现的频率**将其表示，忽略它的单词顺序和语法、句法等要素，将其仅仅看作是若干个词汇的集合，文档中每个单词的出现都是独立的，不依赖于其它单词是否出现。也就是说，文档中任意一个位置出现的任何单词，都不受该文档语意影响而独立选择的。

b) TF-IDF 特征表示，可减少高频停用词的影响。

c) word-embedding 方法表示，目前较常用的方法，可以利用在大语料上训练好的文本向量进行初始化。常用的有 Glove，word2vec 模型。预训练的词向量下载（参考）：https://github.com/Embedding/Chinese-Word-Vectors（在实验中不一定要用到这一步，可以不初始化词向量，直接在训练中学习调整）

2） 分类问题目标函数——交叉熵：

假设情感标签转为[0,0,0,0,1,0,0]，目标函数通过交叉熵表示：

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\log\left(p_{ij}\right)$$

下标 i 代表第 i 次样本， 下标j代表第 j 个类别的概率， y 是真实标签的分布，p是预测的标签分布。 $p_{\mathrm{ij}} \in (0,1)$: $\sum_{j=1}^{m} p_{ij} = 1 \forall i,j$。n 是样本个数，m 是标签类别个数。

4. 实验要求

本次实验要求**实现 CNN 与 RNN 两个模型**，并应用在情感分类任务上。RNN 可以是 LSTM，GRU 等类型。代码的语言不限，可借助深度学习的框架实现（theano，TensorFlow，keras 等）。**对比两模型的实验效果，并分析原因**。也可以实现其他模型作为对比模型（baseline），例如全连接神经网络（MLP），可适当加分。

1）**卷积神经网络（CNN）：**

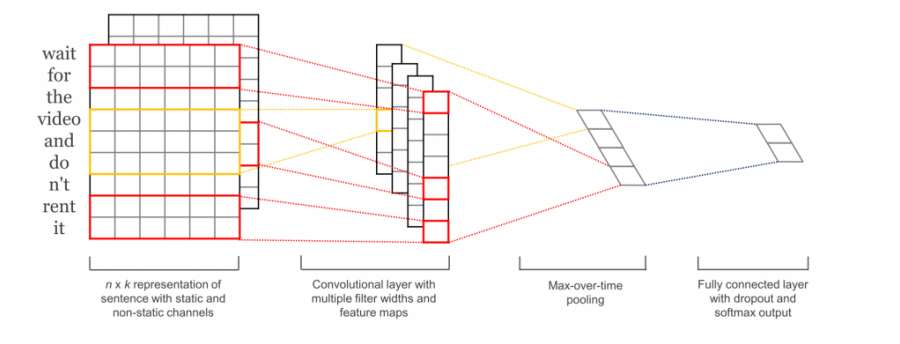


图 3：CNN 模型框架图

参考论文：Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), 1746–1751.

2）**循环神经网络（RNN）：**

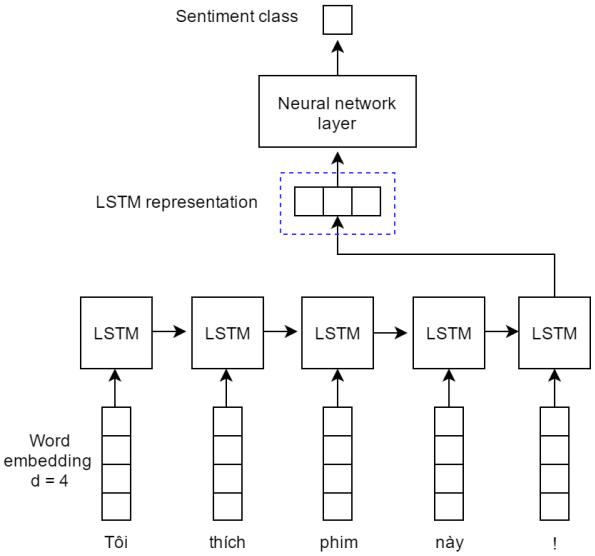


图 4：LSTM 的模型框架

3）**评价指标：**

a) **准确率**（Accuracy）：取情感标签中最大值为 ground truth，预测的最大概率标签为预测值，求整个测试集中的分类准确率。

b) **F-score**：计算 precision 以及 recall，最终由公式$F = \frac{2}{\frac{1}{precision}+\frac{1}{recall}}$得到。Precision 和 recall 计算方式可见下图 5 所示。

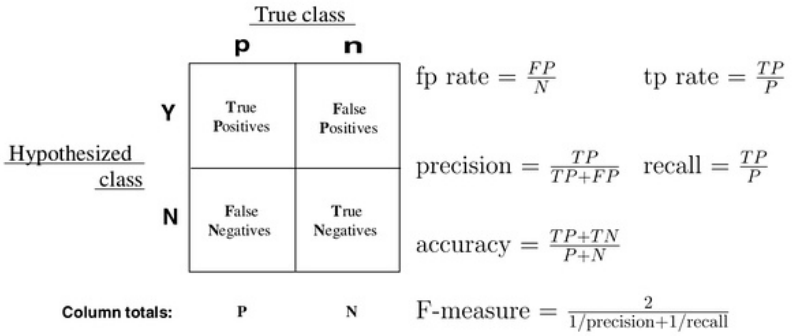


Fig. 1. Confusion matrix and common performance metrics calculated from it.

图 5：F-score 计算方式

对于一般的二分类问题，可以利用上述指标计算。但本实验是多分类问题，在综合考察模型性能的时候就会用到到宏平均和微平均。

- **宏平均（Macro-averaging）**，是先对每一个类都计算一个 F1 值，然后在对所有类求算术平均值。

- **微平均（Micro-averaging）**，是对数据集中的每一个类都计算 TP、FP 和 FN 值，再将多个类的这些值进行累加，从而计算 Precision、Recall 与 F1 值。

在 python 中，可以直接调用 sklearn 的工具进行计算，在实践中可能还会用到 weighted 平均计算方式，具体可以参考下面的链接：
https://www.jianshu.com/p/9e0caf109e88

请在实验后计算三种评价指标（准确率，宏平均，微平均），并如实地汇报在实验报告中。

5. 实验报告内容

1）模型的结构图，以及流程分析。
2）实验结果，三个指标的实验效果。
3）试简要地比较实验中使用的不同参数效果，并分析原因。
4）比较 baseline 模型与 CNN，RNN 模型的效果差异。（如果有实现）
5）问题思考
6）心得体会

6. 问题思考

1） 如果控制实验训练的停止时间？简要陈述你的实现方式，并试分析固定迭代次数与 early stopping 等方法的优缺点。
2） 过拟合和欠拟合是深度学习常见的问题，有什么方法可以解决上述问题。

3） 试分析梯度消失和梯度爆炸产生的原因，以及对应的解决方式。
4） 试分析 CNN，RNN，全连接神经网络(MLP)三者的优缺点与各自适用的场景。

7. 评价方式

程序结果与代码：30%
实验报告：70% （baseline 不一定要实现，但实现可根据难度加分）

8. 提交方式

在网络学堂上提交，需要提交的必要材料如下：
1) 实验报告，以学号_姓名.pdf 命名；
2) 实验代码以及程序运行导引（README）。

9. 联络方式

助教：李祥圣
手机：13763361656 （微信同）
电邮：lixsh6@gmail.com