自然语言处理
Project 2

基于GloVe词向量的文本分类

TA: 丁怡文

文本分类任务

- 文本分类（Text Classification），是指将一篇文本映射到预先给定的某一类别或某几类别主题的过程。目的是将给定的文本归类到预定义的类别中。它广泛应用于情感分析、垃圾邮件检测、主题分类等实际场景。
  - 例：在情感分析中，文本可以被分类为“正面”或“负面”。
  - 在垃圾邮件检测中，电子邮件可以被分类为“垃圾邮件”或“正常邮件”。

- 在文本分类任务中，输入通常是一段文本（如句子、段落或文档），输出是该文本所属的类别标签。

文本分类任务形式化定义

- 给定一个文本集合 $X = \{x_1, x_2, \ldots, x_n\}$，其中每个文本 x_i 包含一个或多个单词 $x_i = \{w_1, w_2, \ldots, w_m\}$，预定义类别集合为 $Y = \{y_1, y_2, \ldots, y_k\}$。
- 文本分类的目标是找到一个**映射函数** f，将 每个文本 x_i 分类到对应的类别 y_j 中，即 $f(x_i) = y_j$，函数 f 通常通过计算文本 x_i 属于每个类别 y_j 的概率来实现，即

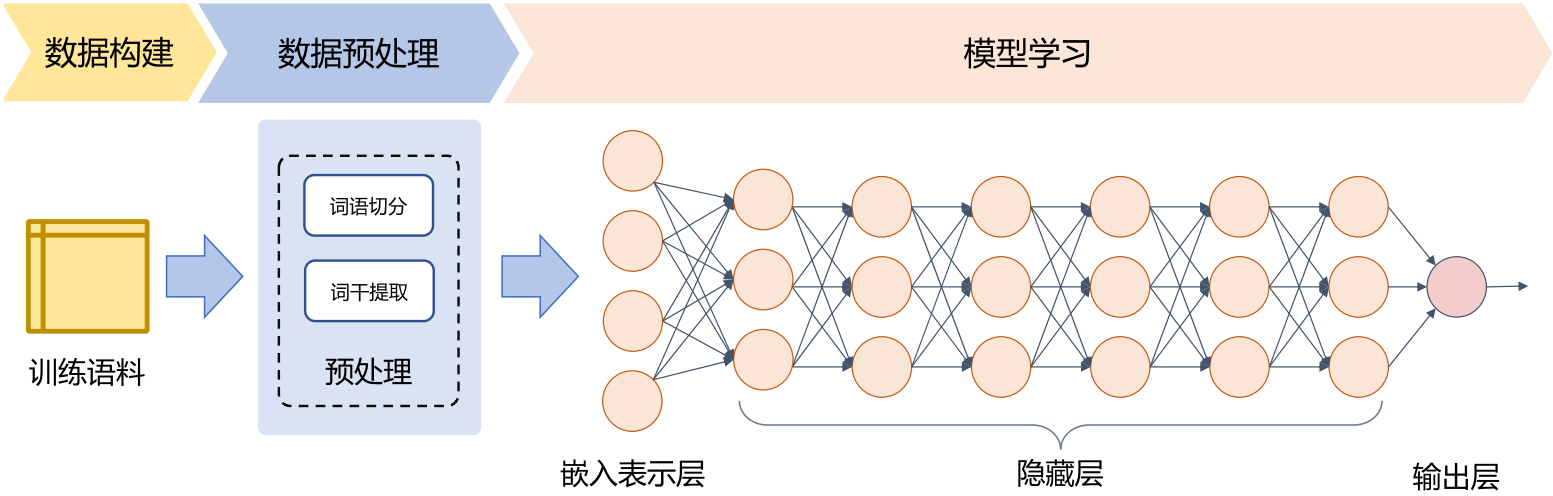
$$P(y_j|x_i)$$

根据最大概率原则，将文本 x_i 分类到具有最高概率的类别中，即

$$f(x_i) = \underset{y_j}{\operatorname{argmax}} P(y_j|x_i)$$

文本分类任务 pipeline

- 数据构建与数据处理：获取任务所需的文本数据集，进行数据清洗
- 嵌入表示：将文本转化为模型可处理的数值表示
- 模型训练：设计模型架构、损失函数、优化算法等



嵌入表示

- 在传统的自然语言处理方法中，我们通常将**单词**视为离散的符号。例如：hotel, conference, motel

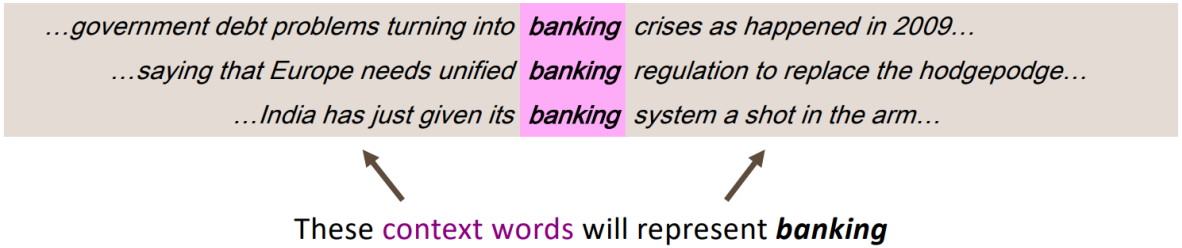
- 单词可以用独热向量（One-hot Vector）表示。每个单词在词汇表中的唯一位置用1表示，其他位置为0。例如：

$$motel = [0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0]$$
$$hotel = [0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0]$$

  - 稀疏性：大部分位置是0，只有一个位置是1。
  - 维度过大：当词汇表非常大时，向量的维度会变得非常高，带来计算和存储的开销。
  - 词之间无关联：独热向量不能捕捉单词之间的语义关系。

词向量（Word Embeddings）

- 核心思想：词语的意义并不是孤立存在的，而是依赖于其出现的上下文 (context)
  - Representing words by their context
  - context: 当一个词语 w 出现在文本中时，其上下文是指在固定大小的窗口内出现的词语集合
  - 方法: 利用词语 w 的多个上下文来建立 w 的表示。



- 分布式表示：词向量通过将单词嵌入到**低维空间**中，能够捕捉单词之间的语义关系，并且大大减少了向量的维度。

词向量（Word Embeddings）

- GloVe: Global Vectors for Word Representation
  - 词语共现性，即对语料库中特定中心词-上下文词对的出现次数进行统计
- 方法
  - 构建共现矩阵 X：统计语料库中所有单词对的共现次数， X_{ij}表示词 w_i 和词 w_j 的共现次数。
  - Example corpus:
    - I like deep learning
    - I like NLP
    - I enjoy flying

counts	I	like	enjoy	deep	learning	NLP	flying	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

[1] https://nlp.stanford.edu/projects/glove
[2] GloVe: Global Vectors for Word Representation, Pennington et al, 2014

词向量（Word Embeddings）

- GloVe: Global Vectors
  - 词语共现性
- 方法
  - 构建共现矩阵 X：统计语料库中所有单词对的共现次数， X_{ij}表示词 w_i 和词 w_j 的共现次数。
  - 目标函数: 优化词向量，使得词对的共现概率与词向量的点积接近。

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$

其中:

- w_i 和 $\widetilde{w_j}$ 是词 w_i 和词 w_j的词向量。
- b_i 和 b_j 是词 w_i 和词 w_j的偏置，对训练目标进行校正
- $f(X_{ij})$ 是加权函数，用于调整不同共现频率的影响，是为共现频率较高的词赋予较高的权重

[1] https://nlp.stanford.edu/projects/glove
[2] GloVe: Global Vectors for Word Representation, Pennington et al, 2014

基于GloVe词向量的文本分类

- 要求
  - 对给定情感分类数据集SST-5（5分类），训练模型
  - 模型应包括三部分：
    1. Word Embedding：随机初始化、GloVe
    2. Encoder：RNN、Transformer等，最多2层
    3. (Pooling+) Classifier：(Attention+) MLP
  - 使用PyTorch实现
  - **Due：9月27日12:00**

基于GloVe词向量的文本分类

- 考核方式
  - 指标：准确率 (accuracy) 全部分类结果中，正确结果的比例。
  - 实验代码（ipynb格式）
    1. 包含代码运行结果以及测试集准确率
    2. 解释实验代码各部分的作用
- 提交方法
  - Elearning提交

基于GloVe词向量的文本分类

- 评分标准
  - 评分主要依据实验和报告的完整性，并不单纯依赖于 acc 指标进行评价，不鼓励卷 acc ， 给分更看重设计与分析。

  - 准时提交完成分类任务的 notebook → 60% points
  - notebook 中包含相应注释与结果输出→ 70% points
  - 设计对应的网络结构 → 80% points
  - 分类 acc 达到 0.4 以上→ 90% points
  - 提供较为详细的分析（notebook中的注释即可）和实现更高的 acc ， 将根据具体表现进一步增加分数。

基于GloVe词向量的文本分类

- 数据集
  - 可选的预处理方式(torchtext==0.8.1或更早版本)

```
import torchtext

TEXT = torchtext.data.Field(lower=True,fix_length=200,batch_first=True)
LABEL = torchtext.data.Field(sequential=False)
train,valid,test = torchtext.datasets.SST.splits(TEXT,LABEL)

TEXT.build_vocab(train,vectors=GloVe(name='6B',dim=100),max_size=20000,min_freq=10)
LABEL.build_vocab(train)

train_iter, valid_iter, test_iter =
torchtext.data.BucketIterator.splits((train,valid,test),batch_size=16)
[TODO...]
print("test_result: ", test_result)
```

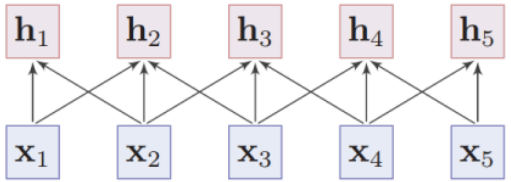
基于GloVe词向量的文本分类

- 数据集
  - 数据组成：训练集 8544，验证集 1101，测试集 2210
  - 格式：

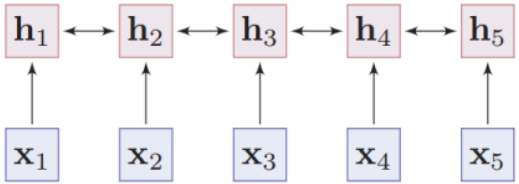
text string · *lengths* 4 283	label int64 0 4	label_text string · *classes* 5 values
a stirring , funny and finally transporting re-imagining of beauty and the beast and…	4	very positive
apparently reassembled from the cutting-room floor of any given daytime soap .	1	negative
they presume their audience wo n't sit still for a sociology lesson , however…	1	negative
the entire movie is filled with deja vu moments .	2	neutral
this is a visually stunning rumination on love , memory , history and the war between…	3	positive
um , no. .	2	neutral

基于GloVe词向量的文本分类

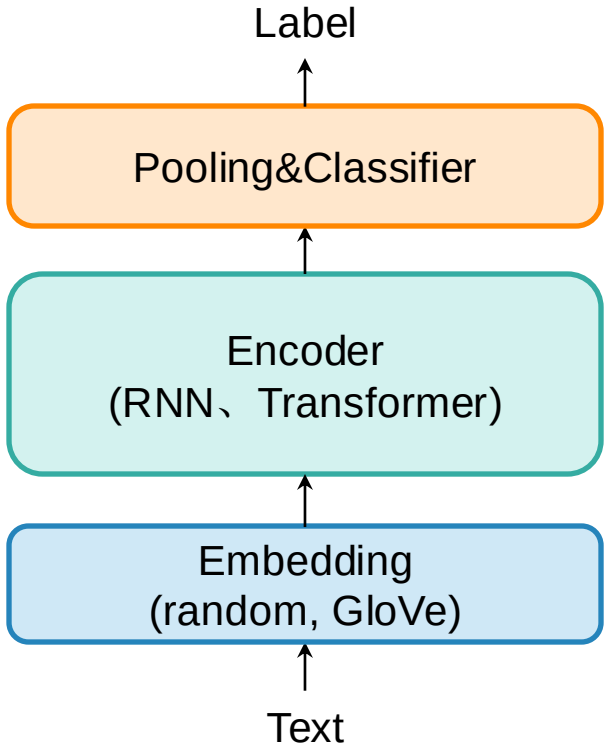
- 模型
  - CNN/RNN的特征抽取



(a) Convolutional Model



(b) Recurrent Model



[1] https://pytorch.org/tutorials/beginner/deep_learning_nlp_tutorial.html

基于GloVe词向量的文本分类

- 实验内容（不需要排列组合所有变量，验证集测试）
  - 2种word embedding方式（随机初始化、GloVe）
  - 2种encoder（ RNN、Transformer ）
  - 分类器前加一层Attention对模型性能的影响（可选）
  - 使用网格搜索或随机搜索寻找最佳超参数（可选，超参至少包含学习率和batch size）
  - 比较你所知的模型训练技巧对准确率的影响（可选，训练技巧包括dropout、模型初始化方式、优化器等）

Contact

- Email：ywding23@m.fudan.edu.cn
- 微信：课程群添加“丁怡文”