哈爾濱工業大學（深圳）
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

机 器 学 习 第 二 次 实 验

新 闻 标 题 分 类

陈科海

2023年10月12日

一、实验数据

新闻标题

THUCNews是根据新浪新闻RSS订阅频道2005~2011年间的历史数据筛选过滤生成，包含74万篇新闻文档（2.19 GB），均为UTF-8纯文本格式
http://thuctc.thunlp.org/

新闻类别

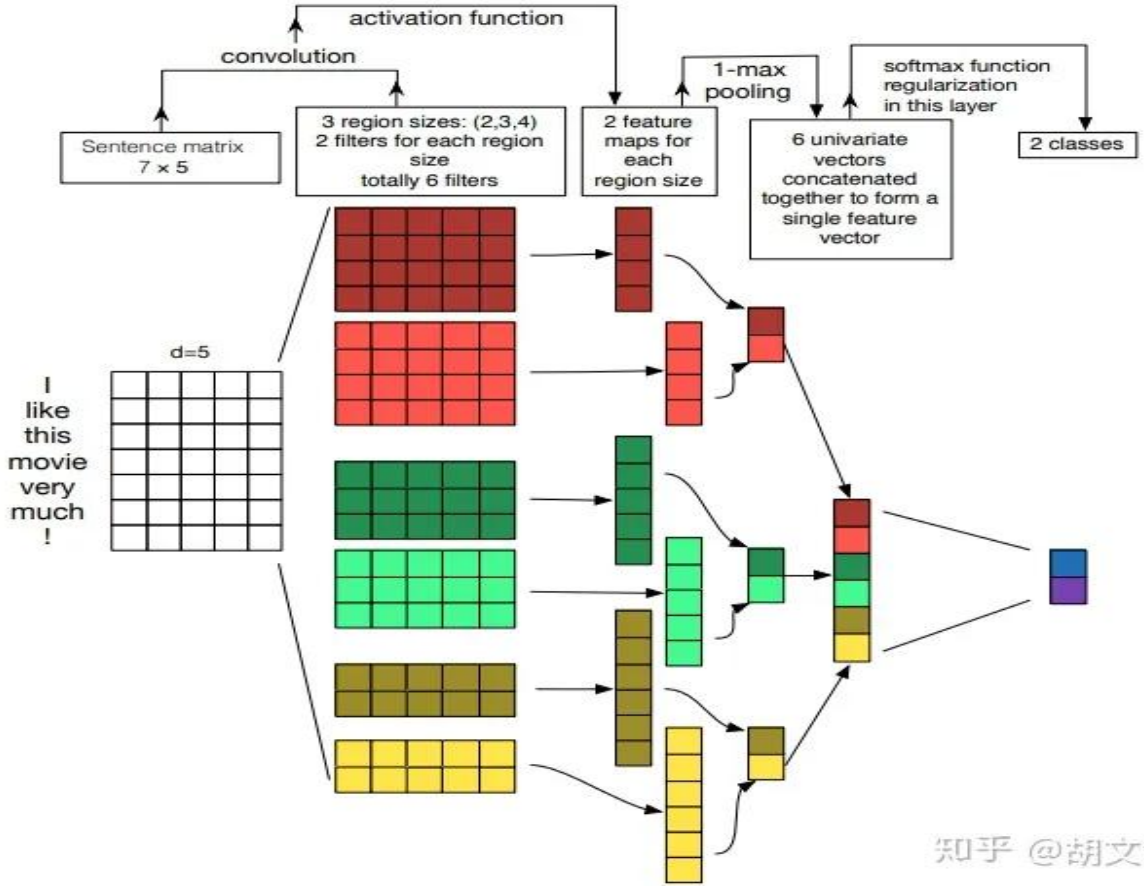
财经、房产、股票、教育、科技、社会、时政、体育、游戏、娱乐

可选的模型

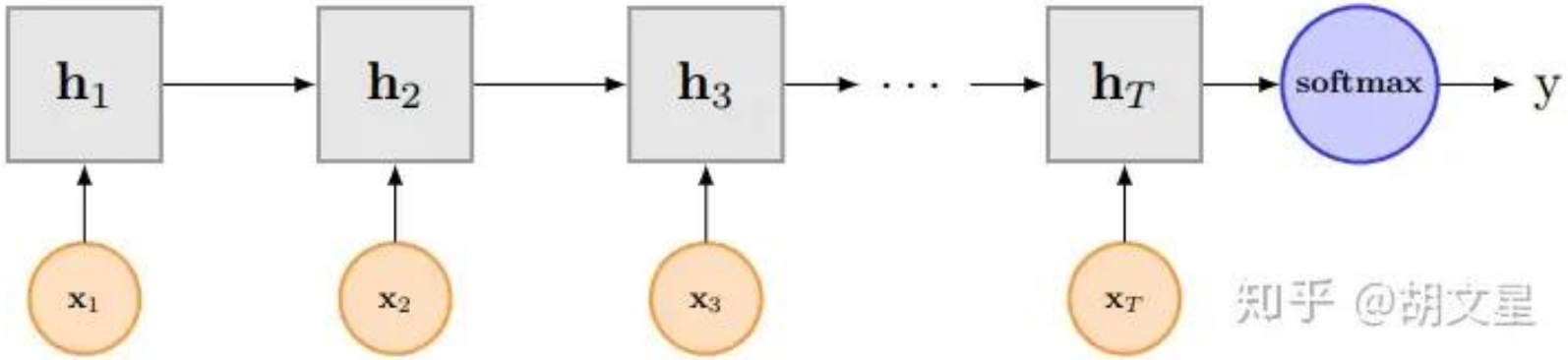
TextCNN, TextRNN, FastText, TextRCNN, BiLSTM_Attention, DPCNN, Transformer

二、候选模型

TextCNN

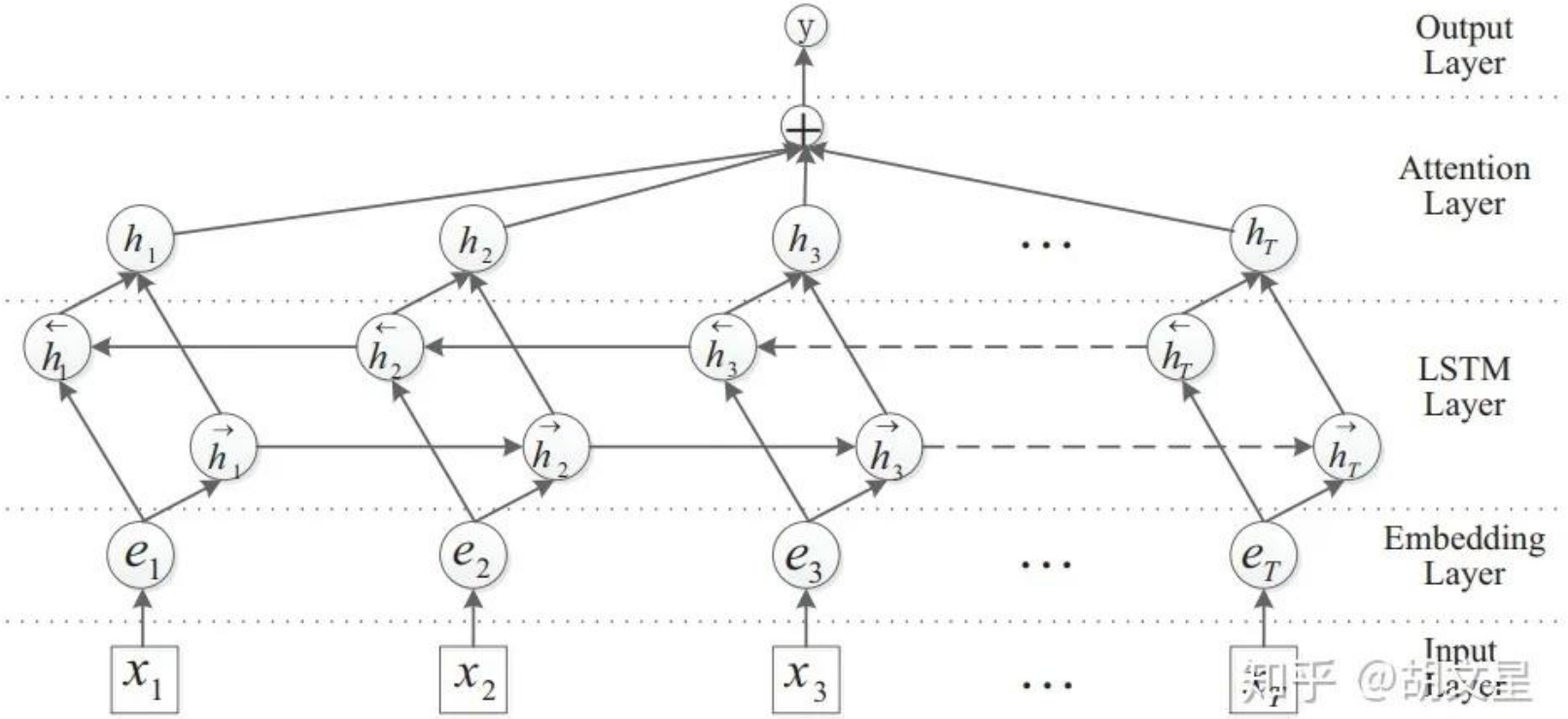


二、候选模型



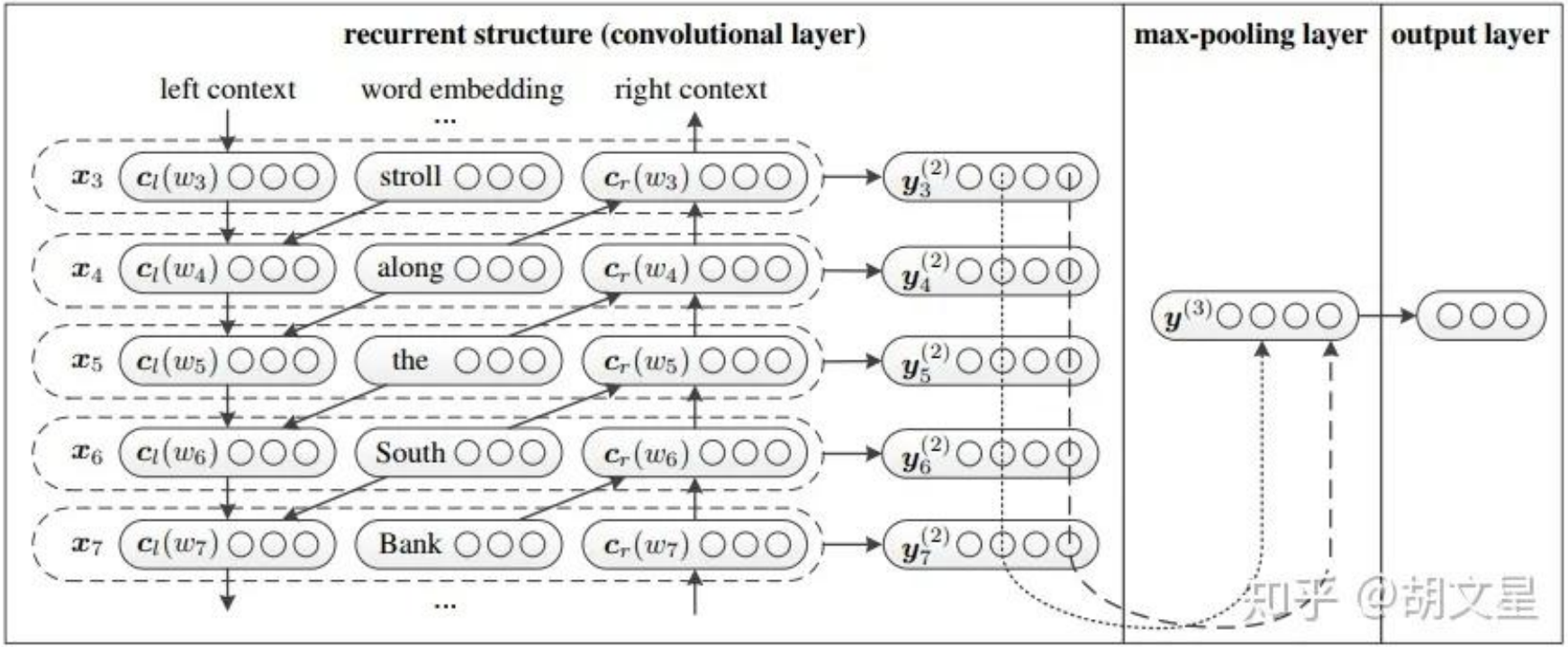
TextRNN

二、候选模型



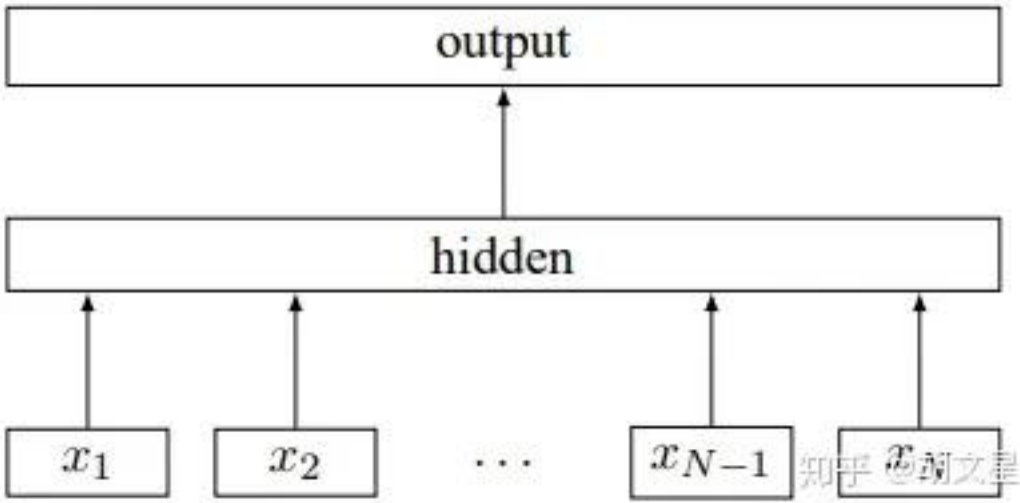
TextRNN+Attention

二、候选模型



TextRCNN

二、候选模型



FastText

三、基线结果

模型	acc	备注
TextCNN	91.22%	Kim 2014 经典的CNN文本分类
TextRNN	91.12%	BiLSTM
TextRNN_Att	90.90%	BiLSTM+Attention
TextRCNN	91.54%	BiLSTM+池化
FastText	92.23%	bow+bigram+trigram， 效果出奇的好
DPCNN	91.25%	深层金字塔CNN
Transformer	89.91%	效果较差
bert	94.83%	bert + fc
ERNIE	94.61%	比bert略差(说好的中文碾压bert呢)

四、实验要求

- 实验要求
  - ✓ 动手复现，并在数据集上运行你的代码，得到训练结果；
  - ✓ 基于基线，优化改进好于基线模型的性能；

谢 谢