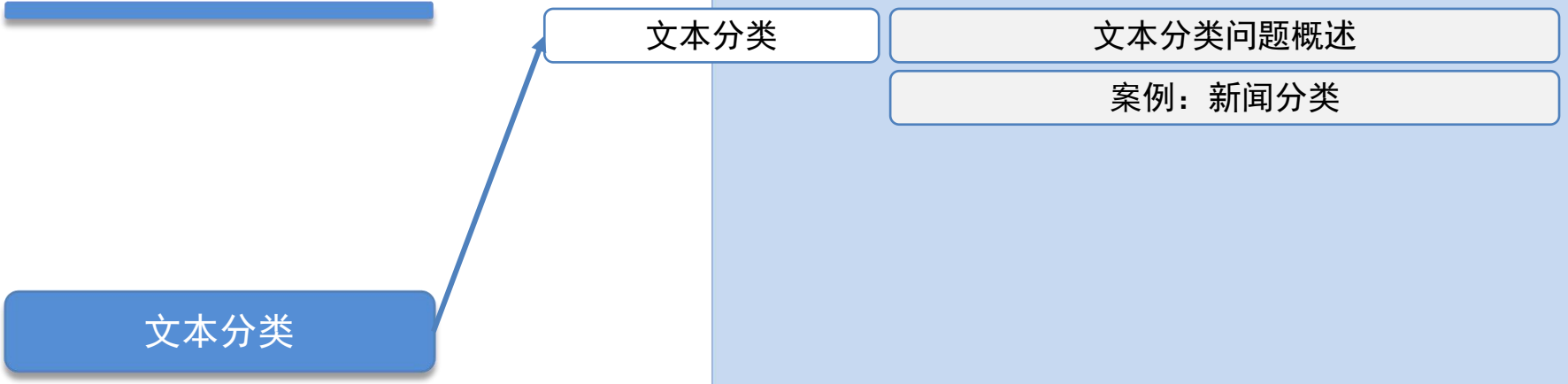
Tedu.cn
达内教育

深度学习

PaddlePaddle文本分类 DAY05

PaddlePaddle



文本分类概述

什么是文本分类

- 图像分类就是将文本划分到不同类别，例如新闻系统中，每篇新闻报道会划归到不同的类别。本质是找到一个有效的映射函数，实现从文本到类别的映射
- 文本分类主要包括：

机器学习文本处理过程

（一）特征工程

- ✓ 文本预处理：分词、取出停用词、符号剔除
- ✓ 特征提取：根据某个评价指标独立的对原始特征项（词项）进行评分排序，从中选择得分最高的一些特征项，过滤掉其余的特征项
- ✓ 文本表示：把文本预处理后的转换成计算机可理解的方式

（二）分类器：将词向量喂入分类器，归入不同类别，常用的分类器有朴素贝叶斯分类算法（Naïve Bayes）、KNN、SVM。用数学语言描述为：

类别集合：$C = \{c_1, c_2, \dots, c_k\}$

文档集合：$D = \{d_1, d_2, \dots, d_n\}$

文本分类：判断$< d_j, c_j >$是 T 还是 F

机器学习进行文本分类缺陷

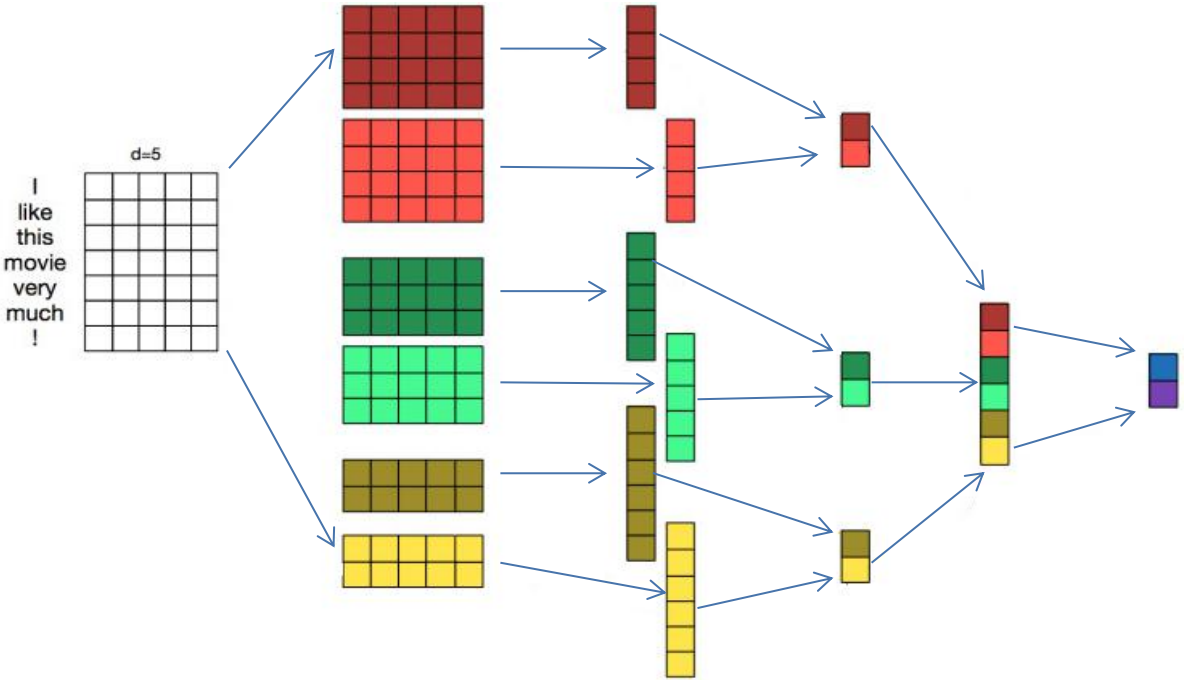
- 特征工程提取，成本高、耗时耗力
- 传统文本表示方法是高纬度高稀疏的，特征表达能力很弱
- 准确率、精度较低

深度学习文本处理方式

- 将文本表达成类似图像、语音的的连续稠密数据，利用卷积神经网络提取文本的局部相关性
- 利用CNN/RNN强大的表征能力，自动提取特征，去掉繁杂的人工特征工程

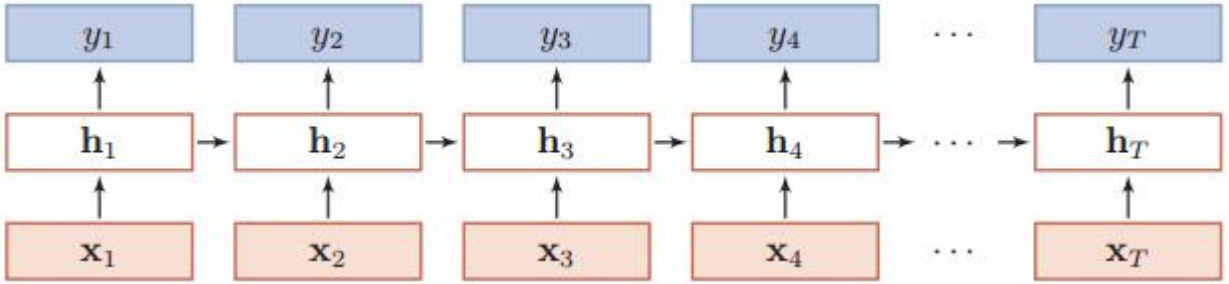
文本处理模型（一）：TextCNN

原始文本　文本矩阵　不同尺寸卷积　1-max pooling层　特征向量



文本处理模型（二）：TextRNN

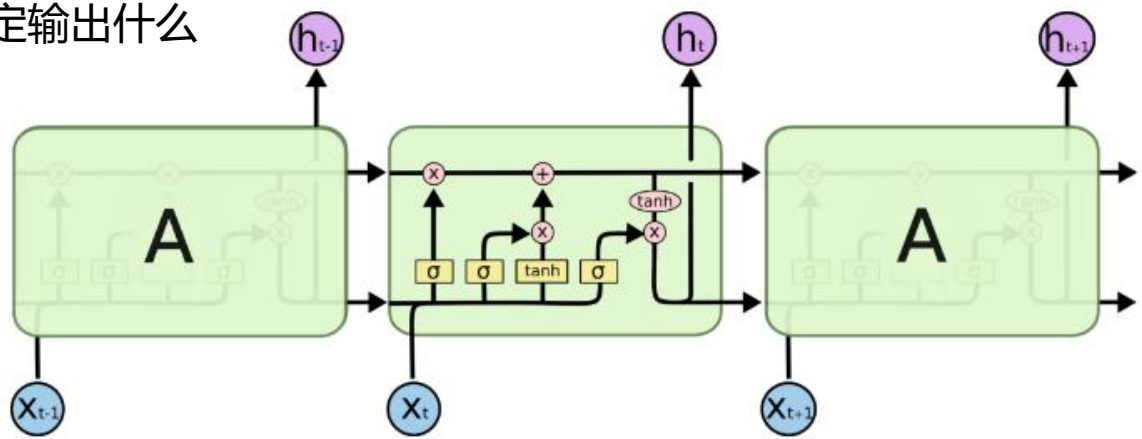
- 循环神经网络（Recurrent Neural Network，RNN）是一类具有短期记忆能力的神经网络，适合用于处理视频、语音、文本等与时序相关的问题



连接不仅存在于相邻的层与层之间（比如输入层-隐藏层），还存在于时间维度上的隐藏层与隐藏层之间（反馈连接，h_1到h_t）。某个时刻t，网络的输入不仅和当前时刻的输入相关，也和上一个时刻的隐状态相关

文本处理模型（三）：LSTM

- 由于RNN具有梯度消失问题，因此很难处理长序列的数据。于是对RNN进行了改进，得到了长短期记忆网络模型（Long Short-Term Memory，简称LSTM）
- 输入门：决定什么信息输入进来
- 遗忘门：决定从细胞状态中丢弃什么信息
- 输出门：决定输出什么



什么时候使用文本分类

- 内容分类（新闻分类）
- 邮件过滤（例如垃圾邮件过滤）
- 用户分类（如商城消费级别、喜好）
- 评论、文章、对话的情感分类（正面、负面、中性）

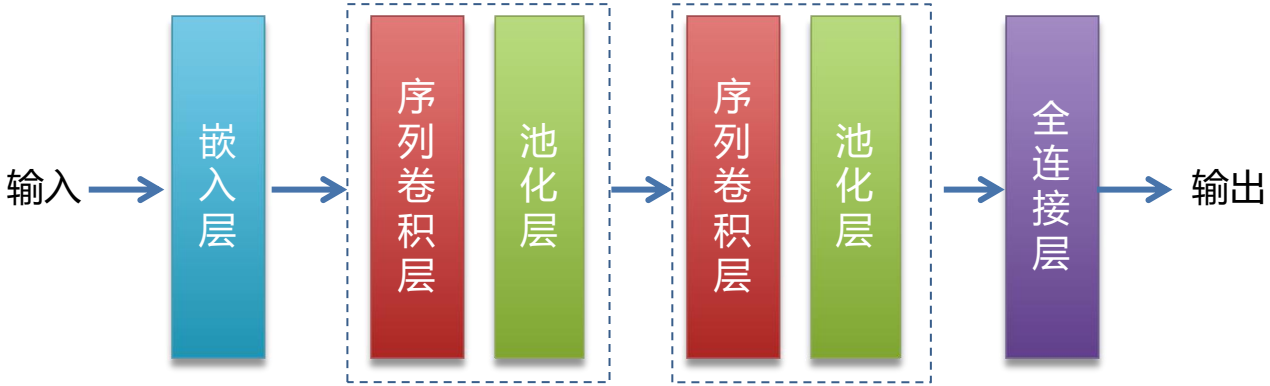
案例：新闻分类

案例介绍

- 目标：利用训练数据集，对模型训练，从而实现新闻分类
- 数据集
  - ✓ 来源：从网站上爬取56821条数据中文新闻摘要
  - ✓ 数据内容：包含10种类别，国际、文化、娱乐、体育、财经、汽车、教育、科技、房产、证券

国际	4354	**汽车**	7469
文化	5110	**教育**	8066
娱乐	6043	**科技**	6017
体育	4818	**证券**	3654
财经	7432	**房产**	3858

- 网络模型：利用训练数据集，对模型训练，从而实现新闻分类



总体步骤

- 数据预处理：解析数据文件，编码，建立训练集、测试集
- 训练与模型评估
- 输入测试数据，进行预测

关键代码：定义网络

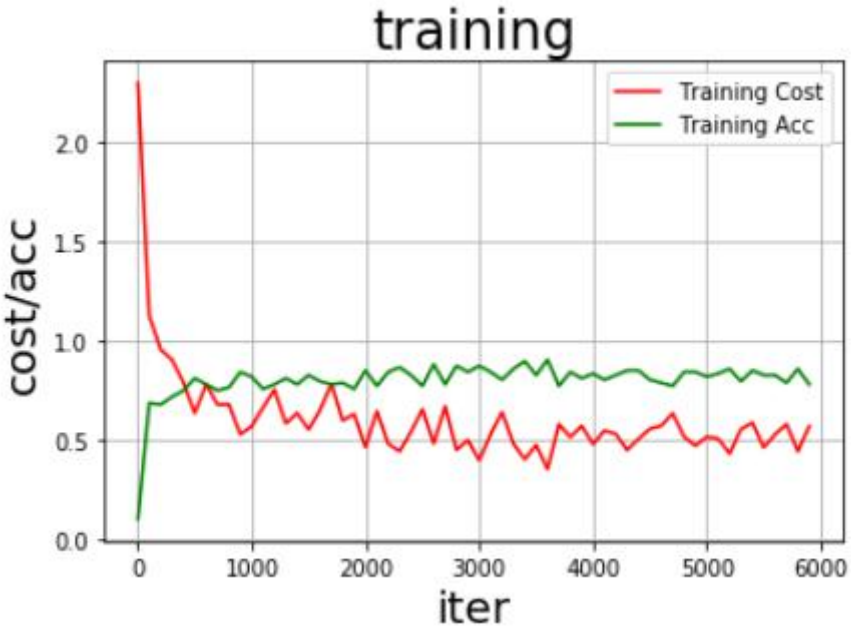

```
def CNN_net(data, dict_dim, class_dim=10, emb_dim=128, hid_dim=128, hid_dim2=98):
    # embeding(词向量)层：将高度稀疏的离散输入嵌入到一个新的实向量空间
    # 以使用更少的维度，表示更丰富的信息
    emb = fluid.layers.embedding(input=data, size=[dict_dim, emb_dim])

    # 第一个卷积、池化层
    # sequence_conv_pool: 序列卷积、池化层构成
    conv_1 = fluid.nets.sequence_conv_pool(input=emb,  # 输入
                                           num_filters=hid_dim,  # 卷积核数目
                                           filter_size=3,  # 卷积核大小
                                           act="tanh",  # 激活函数
                                           pool_type="sqrt")  # 池化类型

    conv_2 = fluid.nets.sequence_conv_pool(input=emb,  # 输入
                                           num_filters=hid_dim2,  # 卷积核数目
                                           filter_size=4,  # 卷积核大小
                                           act="tanh",  # 激活函数
                                           pool_type="sqrt")  # 池化类型

    output = fluid.layers.fc(input=[conv_1, conv_2], size=class_dim, act="softmax")
    return output
```

训练过程


training
Training Cost
Training Acc
cost/acc
2.0
1.5
1.0
0.5
0.0
0
1000
2000
3000
4000
5000
6000
iter

模型测试

在获得诺贝尔文学奖7年之后，莫言15日晚间在山西汾阳贾家庄如是说

综合'今日美国'、《世界日报》等当地媒体报道，芝加哥河滨警察局表示

中国队无缘2020年世界杯

中国人民银行今日发布通知，提高准备金率，预计释放4000亿流动性

10月20日,第六届世界互联网大会正式开幕

```
预测结果: 0, 名称:文化, 概率:0.943867
预测结果: 8, 名称:国际, 概率:0.626188
预测结果: 2, 名称:体育, 概率:0.599507
预测结果: 3, 名称:财经, 概率:0.910002
预测结果: 7, 名称:科技, 概率:0.678429
```

今日总结

- 图像分类问题概述
- 常用数据集
- 图像分类的行业应用
- 案例：水果分类
- 图像分类优化手段