文　本　匹　配

北邮　硕士　高晨宇

风之子

智慧金融研究院
SMART FINANCE INSTITUTE

第三届魔镜杯数据应用大赛
PPDAI 3th Magic Mirror Data Application Contest

CONTENTS

01 数 据 处 理

01-1. 图特征

01-2. 训练词向量

01-3. 数据增强

01-1. 图特征

做法:
在训练集和测试集的总集合中,对问题进行唯一编号,并统计词频

效果:
本特征对提高分数十分有效

分析:
将hash值送入网络,能使网络学到一些问题之间的关系;
问题出现频率越高,可能意味着该问题越冗余

label	w1	w1_len	w2	w2_len	w1_hash	w2_hash	w1_freq	w2_freq
1	W04465 W04058 W05284 W02916	4.0	W18238 W18843 W01490 W09905	4.0	0	2349	76	11
0	W10054 W04476 W09996 W12244 W18103	5.0	W18439 W00863 W04259 W00740 W16070	5.0	1	112296	1	1
0	W04346 W17378 W19355 W17926 W14185 W11567 W07863	7.0	W14586 W09745 W06017 W09067 W16319	5.0	2	112297	2	1
0	W17508 W09996 W19662 W17534 W11399 W17057 W182...	8.0	W18238 W02357 W06606	3.0	3	61360	2	2
0	W13157 W03390 W01952 W05789 W17378 W08714 W13157	7.0	W04476 W06606 W00316 W13157	4.0	4	112298	2	1

01-2. 训练词向量

做法:
使用 gensim.models.word2vec 对所有 word 或 char 进行训练,得到新的词向量

效果:
有些模型效果好,有些模型效果不好

分析:
猜想根据比赛数据的语料库来训练词向量可能更合适,经过实验来验证

01-3. 数据增强

做法:
尝试过以下几种: reverse, dropout, shuffle等

效果:
不是每一种都有效,需要经过调参判断选择的组合

分析:
通过这些数据增强能在一定程度上破坏原本的句子对的结构,从而达到增加训练集的目的

02 深 度 学 习 模 型

01-1. LSTM

01-2. CNN

01-3. RCNN

01-0. 机器学习模型

tf-idf特征:
—易于构造，但维度太高,需要降低纬度来加速训练,避免维度灾难。
—尝试后效果不好,推测是由于本特征是为了让相同领域的问题有更高的得分,但是本赛题中可能大部分是同一领域

词向量特征:
—将每一个单词映射到低维度空间中。（300维）
—再将评论中每一个词所对应的词向量求和平均作为评论的句子向量。
—计算速度快，精度高。
—模型采用的是Lightgbm，Xgboost，但效果仍不如深度模型

01-1. LSTM

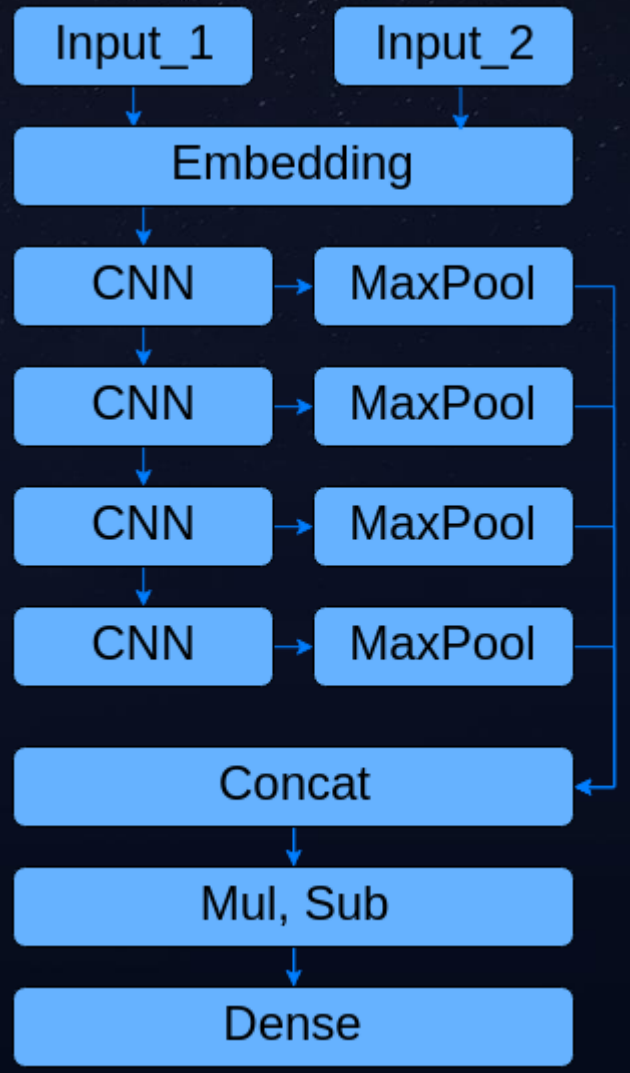
BiLSTM:
单向的LSTM能够有效捕捉从前到后的时序信息；
双向的LSTM能够捕捉从前到后和从后到前的时序信息。

Attention：
可以判断每一个时序信息的重要程度，赋予每个时序不同的权重。

01-2. CNN

Hierarchal CNN:
原理：
利用卷积神经网络分层提取特征，最后合并；

特点：
与RNN相比，CNN能更好地提取文本语义信息；
复杂度低，训练速度很快。

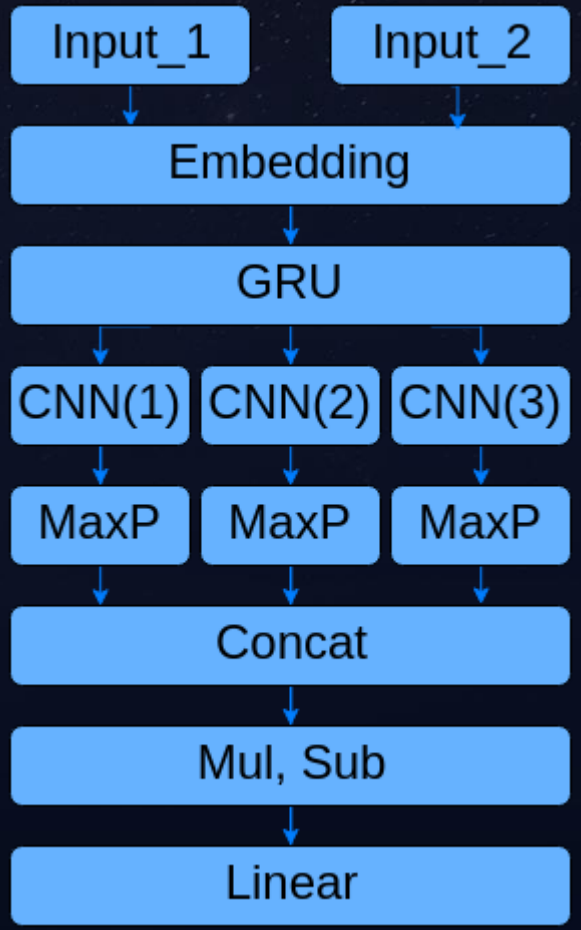


01-3. RCNN

RCNN：
结合RNN和CNN的优点，既能获取时序信息，也能提取文本特征

GRU：
GRU与LSTM在很多任务中不分伯仲；
GRU的参数更少，更容易收敛。

CNN：
利用多尺度卷积核提取特征



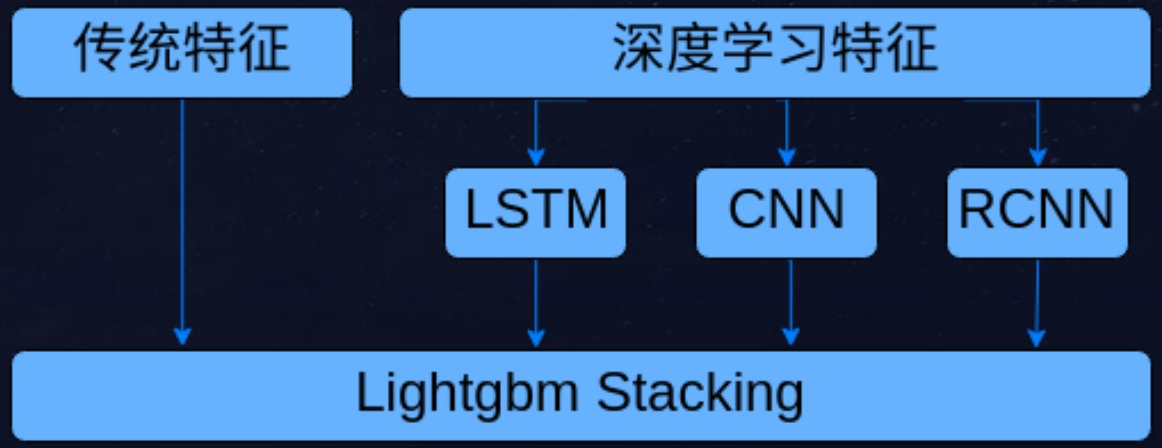
03 模 型 融 合

01-1. Stacking

01-1. Stacking

使用要求：
基模型一般应该独立准确，而不同的基模型之间有所差异。

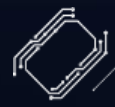
效果提升原因：
不同的基模型是从不同的角度观测数据集，模型融合相当于取长补短。



THANK YOU

风之子

智慧金融研究院
SMART FINANCE INSTITUTE



拍拍贷"魔镜杯"
互联网金融数据应用大赛

第三届魔镜杯数据应用大赛
PPDAI 3th Magic Mirror Data Application Contest