第三届魔镜杯大赛方案介绍

沈　熵　华中科技大学
杨腾佼　电子科技大学
肖　洋　华中科技大学
王飞越　华中科技大学

Dispos

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

CONTENTS

01. 团队介绍
02. 问题简介
03. 数据探索
04. 特征工程
05. 模型介绍
    - 05-1. PooledGRU
    - 05-2. PooledLSTM
    - 05-3. Fine-tune
    - 05-4. Blending
06. 后 处 理
07. 总　结

Dispos

01 团队介绍

Dispos

01.团队介绍

团队成员

- shenyu: 沈燏，华中科技大学研究生
- Tnging: 杨腾佼，电子科技大学研究生
- yxiao1994:肖洋，华中科技大学研究生
- King0F1y:王飞越， 华中科技大学研究生

团队荣誉

- 2017年京东金融全球数据探索者大赛—店铺销量预测（冠军）
- 2017年华为软件精英挑战赛（决胜奖，武长赛区第二名）
- 2018年“云移杯- 景区口碑评价分值预测 （季军）

Dispos

02 问题简介

02.问题简介

- **问题定义**：给定问题q1和q2，要求判断q1和q2语义上是否表达同一个意思。

语义上
q1 == q2 ?

- **数据格式**：训练集中每一行包含问题对以及label，问题集每一行包含词序列及字符序列。

训练集

label	q1	q2
1	Q397345	Q538594
0	Q193805	Q699273
0	Q085471	Q676160

问题集

qid	words	chars
Q000000	W05733 W05284 W09158 W14968 W07863	L1128 L1861 L2218 L1796 L1055 L0847 L2927
Q000001	W17378 W17534 W03249 W01490 W18802	L2214 L1980 L0156 L1554 L2218 L1861 L3019 L010...
Q000002	W17378 W08158 W20171 W11246 W14759	L2214 L2350 L2568 L1969 L2168 L0694 L3012 L256...

- **评测指标**：基于训练集数据构建预测模型，使用模型计算测试集的评分，评价标准为logloss

$$L_{\log}(y,p) = -\log \Pr(y|p) = -(y\log(p) + (1-y)\log(1-p))$$

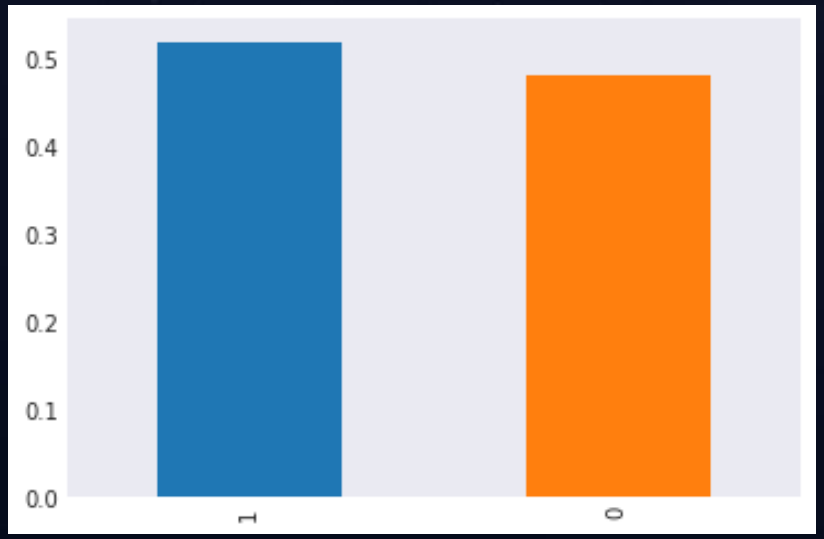
Dispos

03 数据探索

03.数据探索

数据分析目的：

- 正确划分数据集，保证测试集和验证集来自同一分布
- 设定合适的模型参数
- 从数据中获取启发，构建有效特征

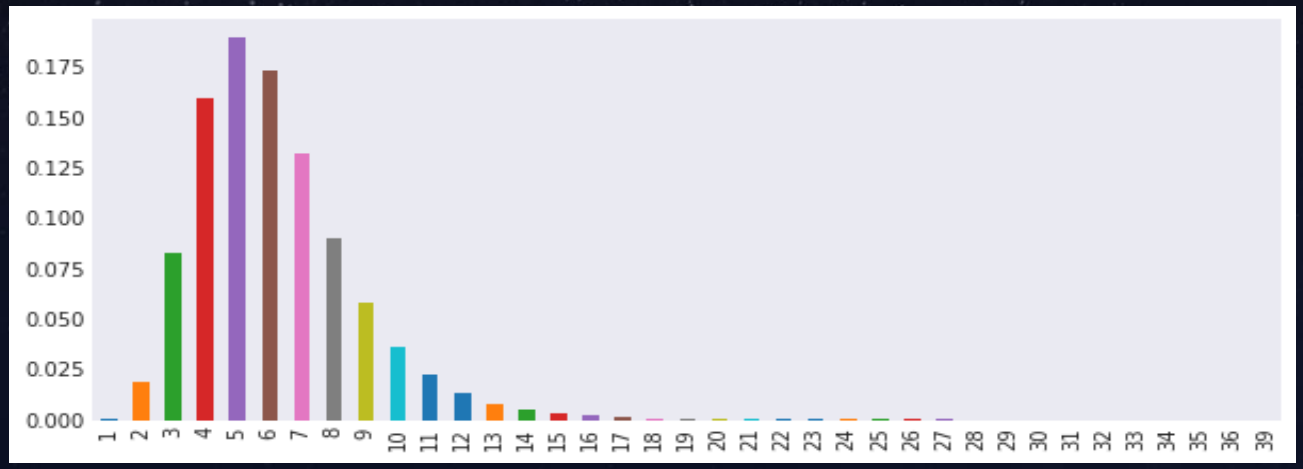


训练集Label分布

在训练集上的label正负比例相对比较均匀，因此我们直接取其中90%作为训练集，另外10% 作为测试集。即可得到近似同分布的训练集和测试集。

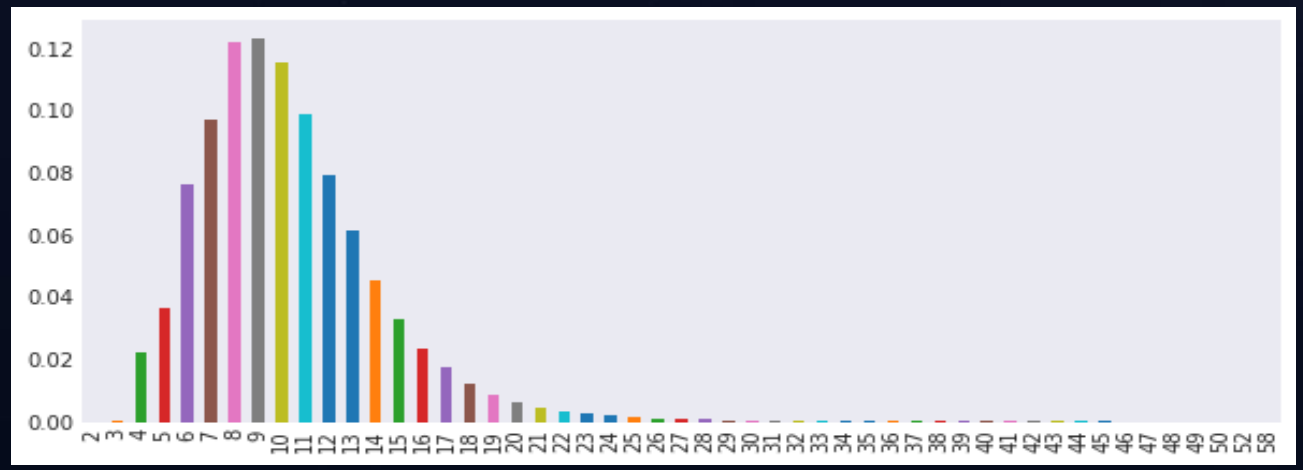
Dispos

03.数据探索



问题词长度分布

词长度在5左右的比较多，长度超过20的问题很少。设定词的最大长度在20到30之间，可以在保证训练速度，减少信息损失。



问题字符长度分布

字符长度在9左右的比较多，长度超过30的问题很少。设定词的最大长度在30到40之间，可以在保证训练速度，减少信息损失。

Dispos

04 特征工程

04.特征工程

- 问题对q1, q2的公共邻居的数量

$$q1\_q2\_intersect = neighbor(q1) \cap neighbor(q2)$$

- 问题q1，q2的hash映射

$$q\_hash = Hash(q)$$

- 统计q1和q2在数据集中出现的次数

$$q\_freq = \frac{number\ of\ q}{number\ of\ record}$$

05 模型介绍

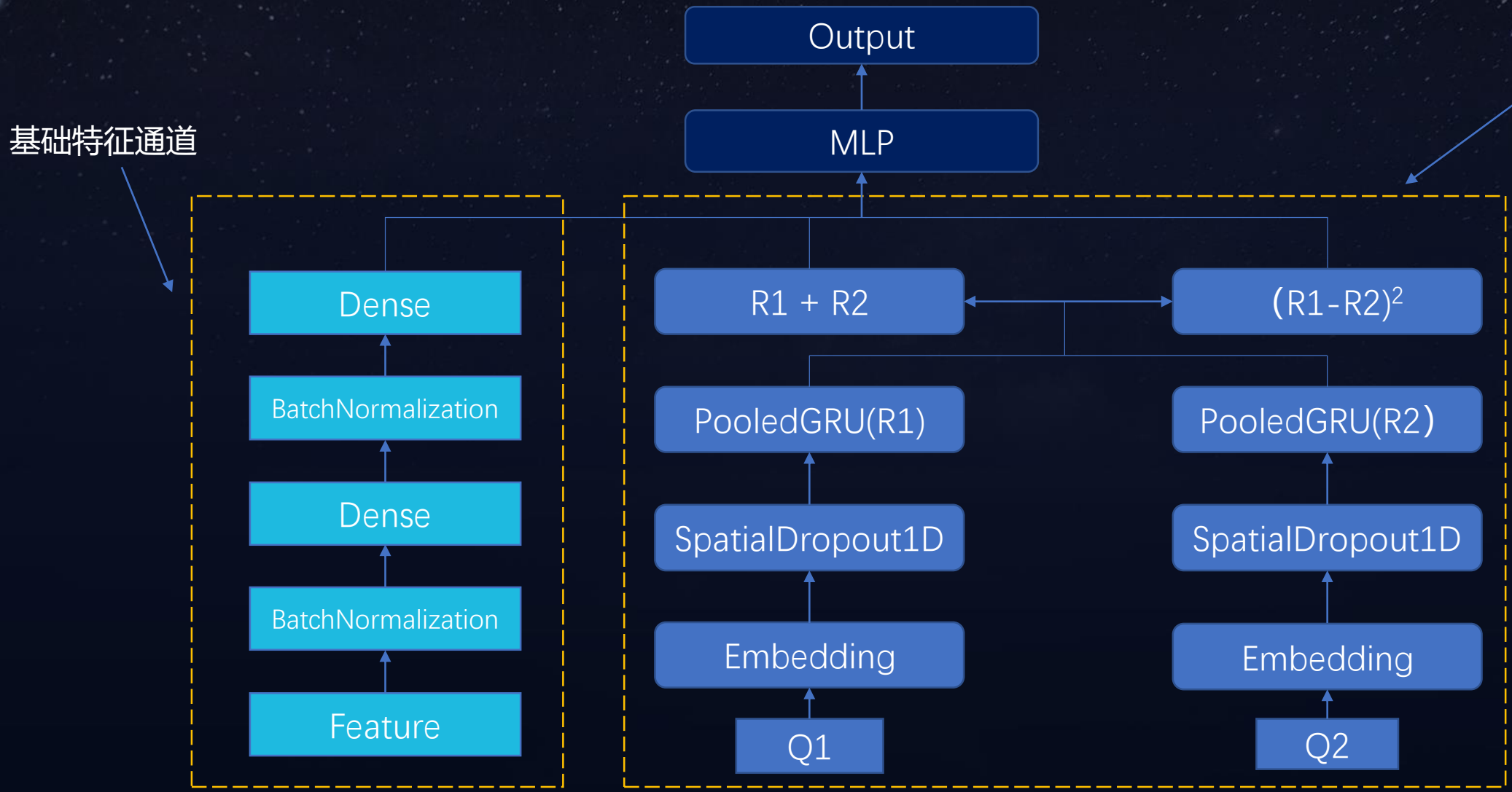
05-1. PooledGRU

05-2. PooledLSTM

05-3. Fine-tune

05-4. Blending

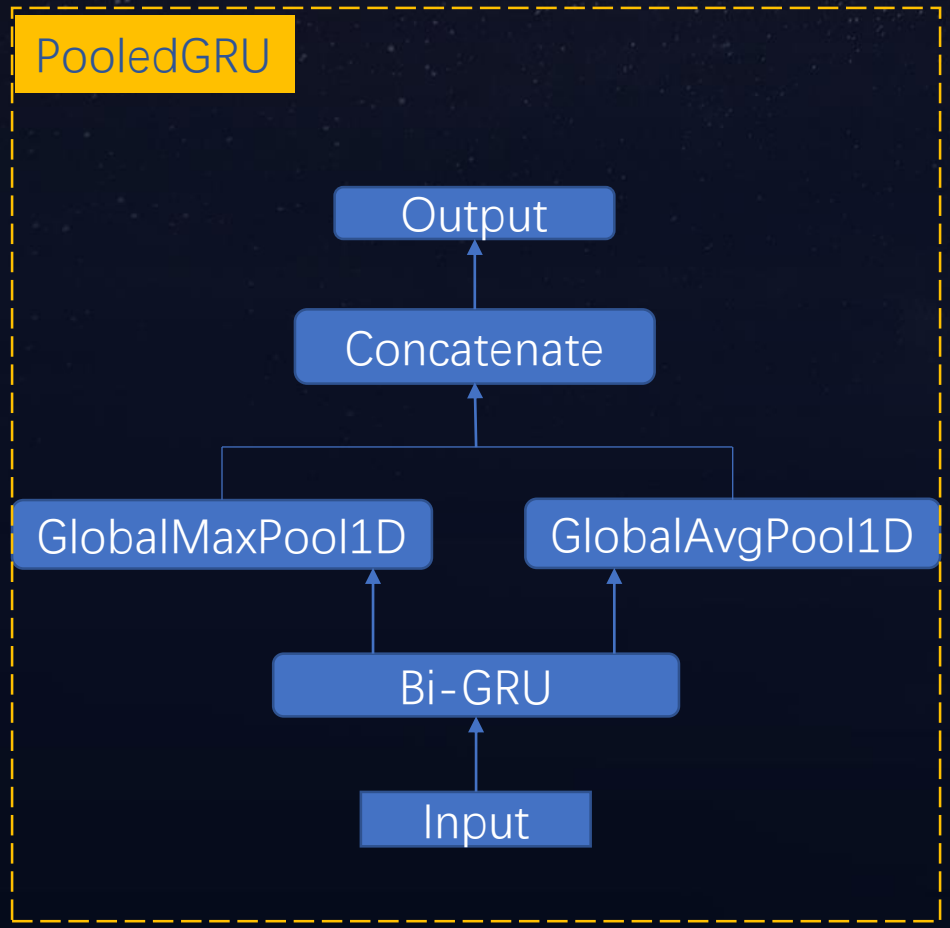
05-1.PooledGRU



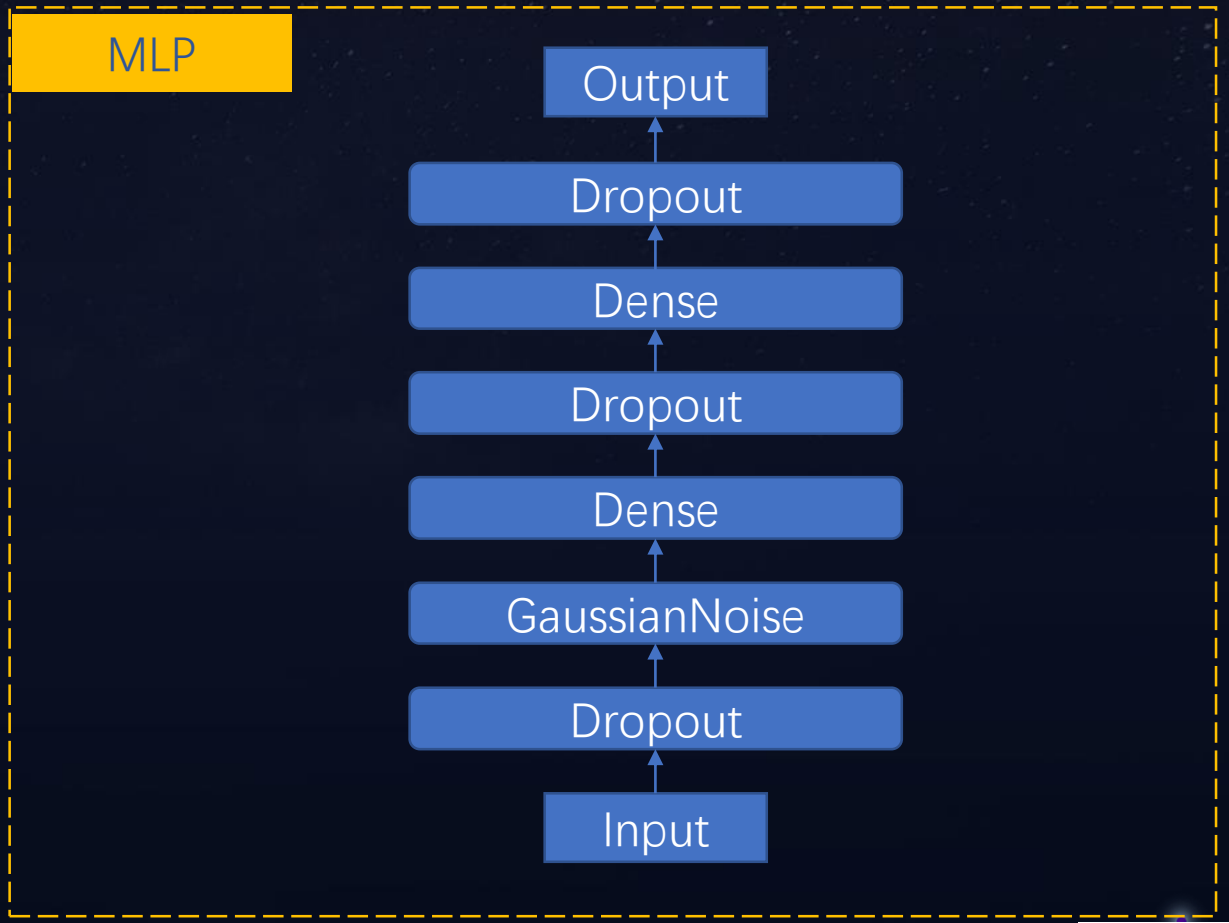
Dispos

05-1.PooledGRU

GRU模型加池化学习句子的深度表示



通过多层深度模型学习来自多组特征的信息，共同决策



Dispos

05-2.PooledLSTM

Baseline Needs More Love:On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms---2018 ACL

Dispos

05-3.Fine-tune

- Fine-tune动机: embedding的算法选择，训练方式及参数设置对模型效果有很大影响。
- Fine-tune步骤：重新加载已经训练好的模型，将embedding的trainable改为True，调整 batch size以及learning rate，再次将数据输入模型中进行训练。

05-4.Blending

- **动机**：随机森林算法通过选择不同的特征子集来训练不同的树，从而提高模型的泛化性。
- **步骤**：对每种深度模型架构分别选择不同的embedding, 以及不同的特征组合，进行训练，得到有差异性的多种模型。

模型2种：
- PooledGRU
- PooledLSTM

✘

embeding2种：
- 训练集的question训练的embed1
- 训练集+测试集question训练的embed2

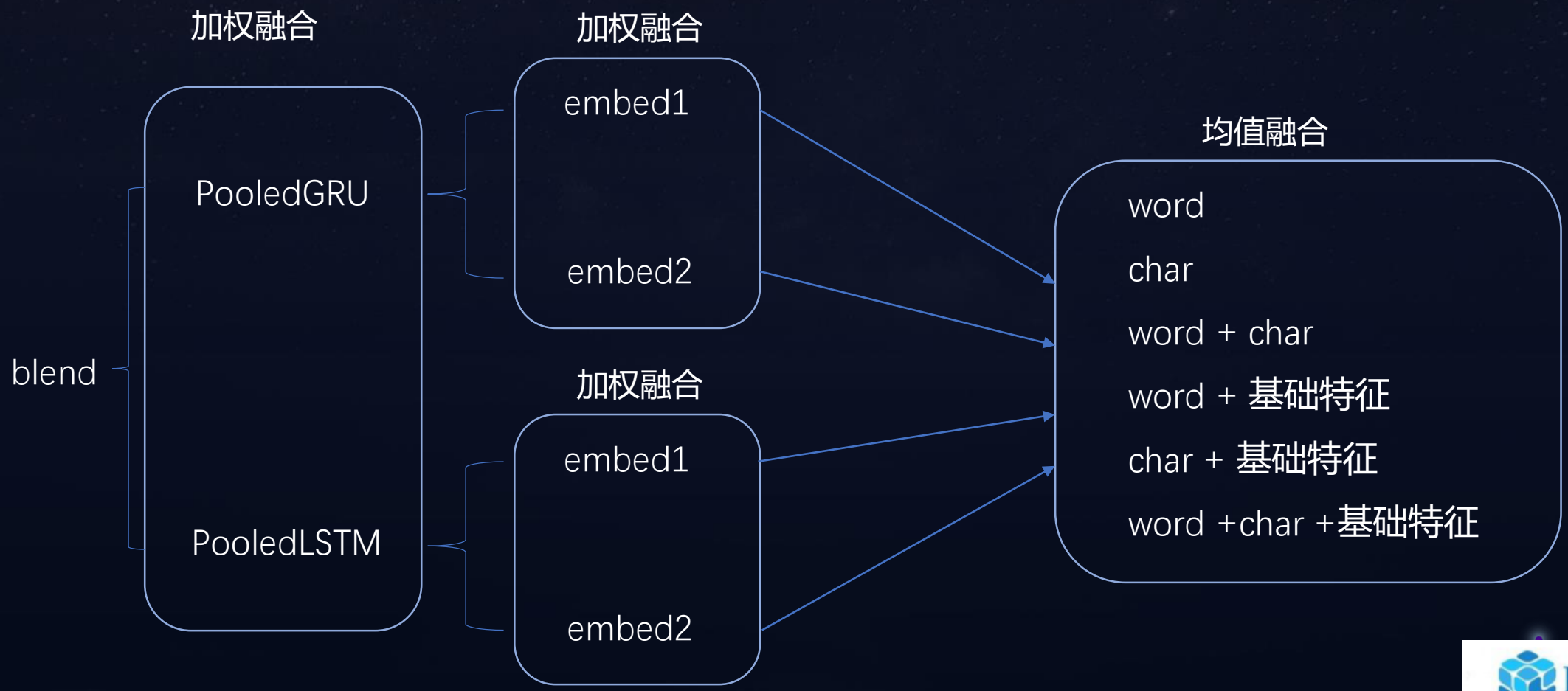
✘

特征3种：
- 基础特征
- Word
- Char

Dispos

05-4.Blending

基本融合策略：先融合相似度较高的模型，再融合相似度较低的，分层级逐级融合。



Dispos

06 后处理

06.后处理

- 如果a和b相似，b和c相似，那么a和c相似
- 如果a和b相似，b和c不相似，那么a和c不相似

Dispos

07 总结

07.总结

创新点：

- 基于前面提到的论文，用embedding直接concat LSTM的输出，再在后面接一个Pooling的方式能很大程度上提升模型的性能。

- embedding对深度模型会产生较大影响，首先在训练时将embedding固定，训练完模型后再对重新对embedding进行fine-tune会有很大提升 .

- 选择不同特征集对模型进行训练，可以增大模型的差异性，再采用分层次的融合方式能够取得很好的效果。

Dispos

THANK YOU

Dispos

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