第三届魔镜杯解决方案

sky队

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

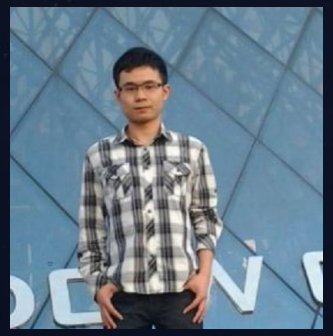
CONTENTS

01 团　　队　　介　　绍

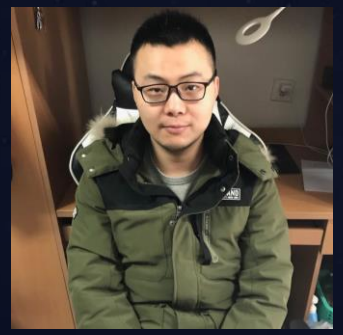
01-1.人员介绍

01-2.团队荣誉

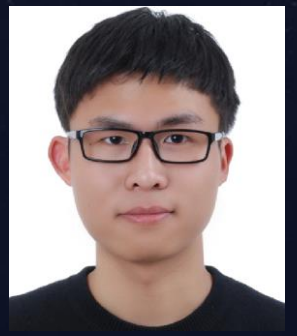
人员介绍



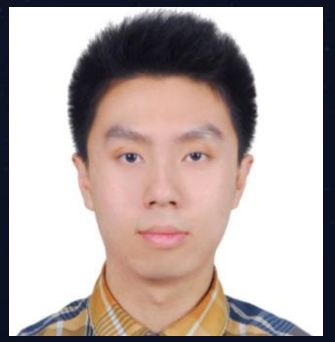
bird
应缜哲
同花顺
工程师



wen_xiao_wen
许文超
北京邮电大学
研究生



skyhigh
林旭鸣
北京邮电大学
研究生



uncleban
王强
北京邮电大学
研究生
滴滴地图事业
部实习

CORTEX

SKY

团队荣誉

- 2017年摩拜算法挑战赛 预测top3目的地 B榜第3/1007队 季军
- 2017年蚂蚁金服商铺定位赛 给定gps和wifi环境预测用户所在店铺 2/2845队 亚军
- 2017年CCF360人机大战赛 文本分类区分机器文本和人类文本 3/589队 季军
- 2017年JDD信贷预测 对未来一个月内用户的借款总金额进行预测 1/745队伍 冠军
- 2017年DiTech 无人驾驶大赛 冠军
- 2018年马上金融意图识别 聊天文本分类 亚军
- 2018年腾讯算法大赛 相似人群拓展 B榜冠军
- 2018年kaggle mercari大赛 季军

CORTEX

SKY

02 比　　赛　　历　　程

02-1.初赛

02-2.复赛

初赛

复赛 | 初赛

* 排名以最优成绩排列，实时更新

		排名变化	分数	提交次数	最佳成绩提交时间
1	sky@Cortexlabs	0	0.141329	68	2018-07-09 22:40:06.0
2	小幸运	0	0.144224	101	2018-07-09 10:41:07.0
3	SuperGUTS	0	0.144392	84	2018-07-09 23:09:32.0

6月20日参赛
7月4日开始保持初赛**第一**

复赛

复赛 | 初赛

* 排名以最优成绩排列，实时更新		排名变化	分数	提交次数	最佳成绩提交时间
1	sky@Cortexlabs	0	0.142658	7	2018-07-16 23:48:00.0
2	地表最强@CortexLabs	0	0.142747	7	2018-07-16 23:01:18.0
3	Dispos	0	0.143198	7	2018-07-15 18:59:54.0

7月10日开始复赛
初赛+复赛 最长连续保持11天第一

03 任　　务　　分　　析

问题与数据

问题：给定两个问题，通过算法计算两个问题的相似度，进而判定两个问题是否在语义上一致。

数据：

阶段	训练集	测试集
初赛	25w(13w+, 12w-)	30%测试集
复赛		所有测试集(17w)

$$L_{\log}(y,p) = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1-y_i)\log(1-p_i))$$

04 解 决 方 案

04-1.特征提取

04-2.数据增强

04-3.模型构建

04-4.模型finetune

04-5.后处理

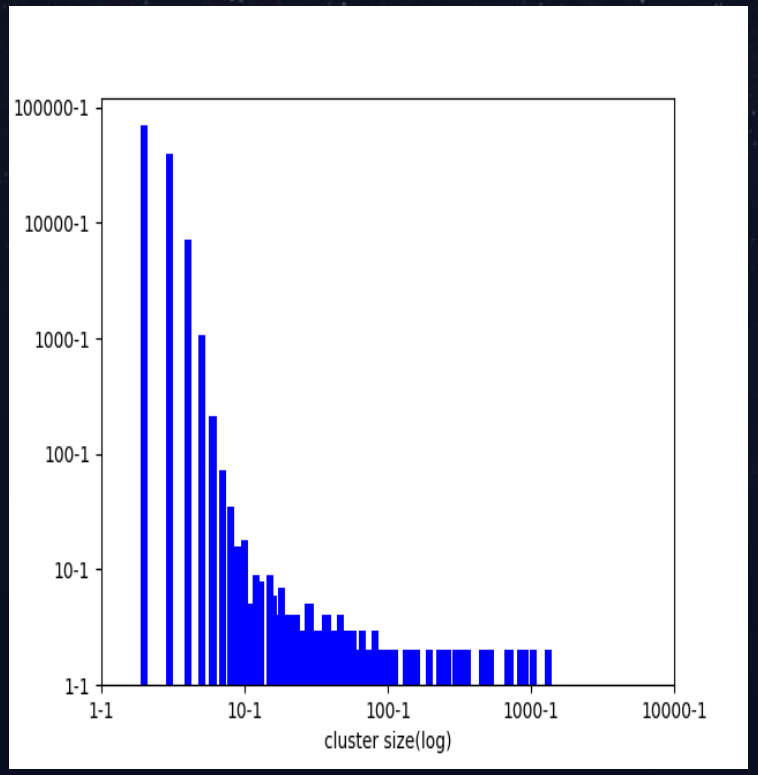
特征提取

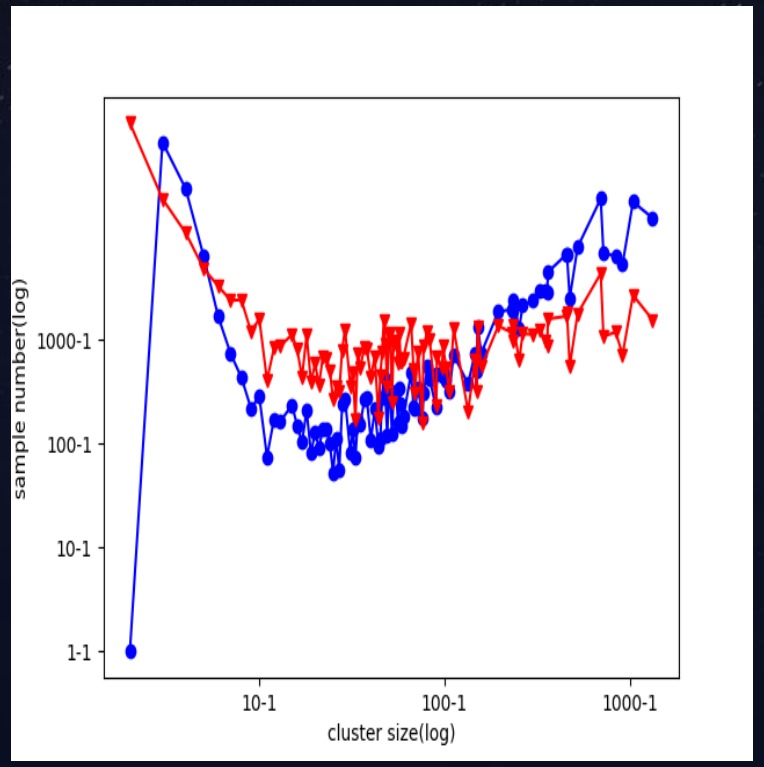
- 使用gensim重新训练词向量。

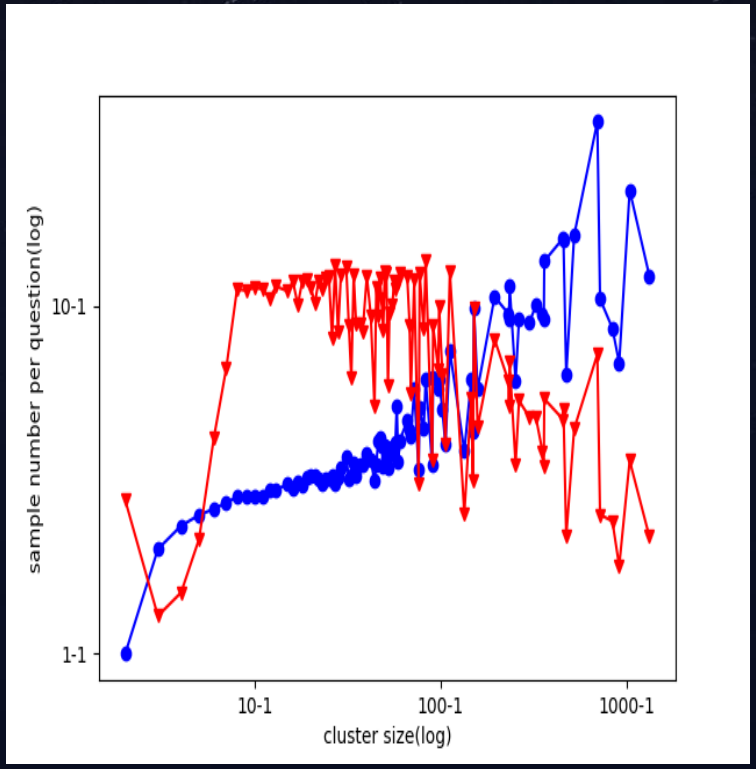
- 提取问题出入度、pagerank等特征。问题的出现次数以及频繁程度特征。

- 将所有已知的问题构建同义问题集。问题集的构建不参与训练，仅用于数据增强。

数据增强







统计不同size的问题集的个数以及它们占有的正负样本个数，可以发现即使大问题集个数非常少，但却占有了非常多样本，同时正负样本存在不均衡情况。

数据增强

假设Q1在所有样本里出现2次，分别是：

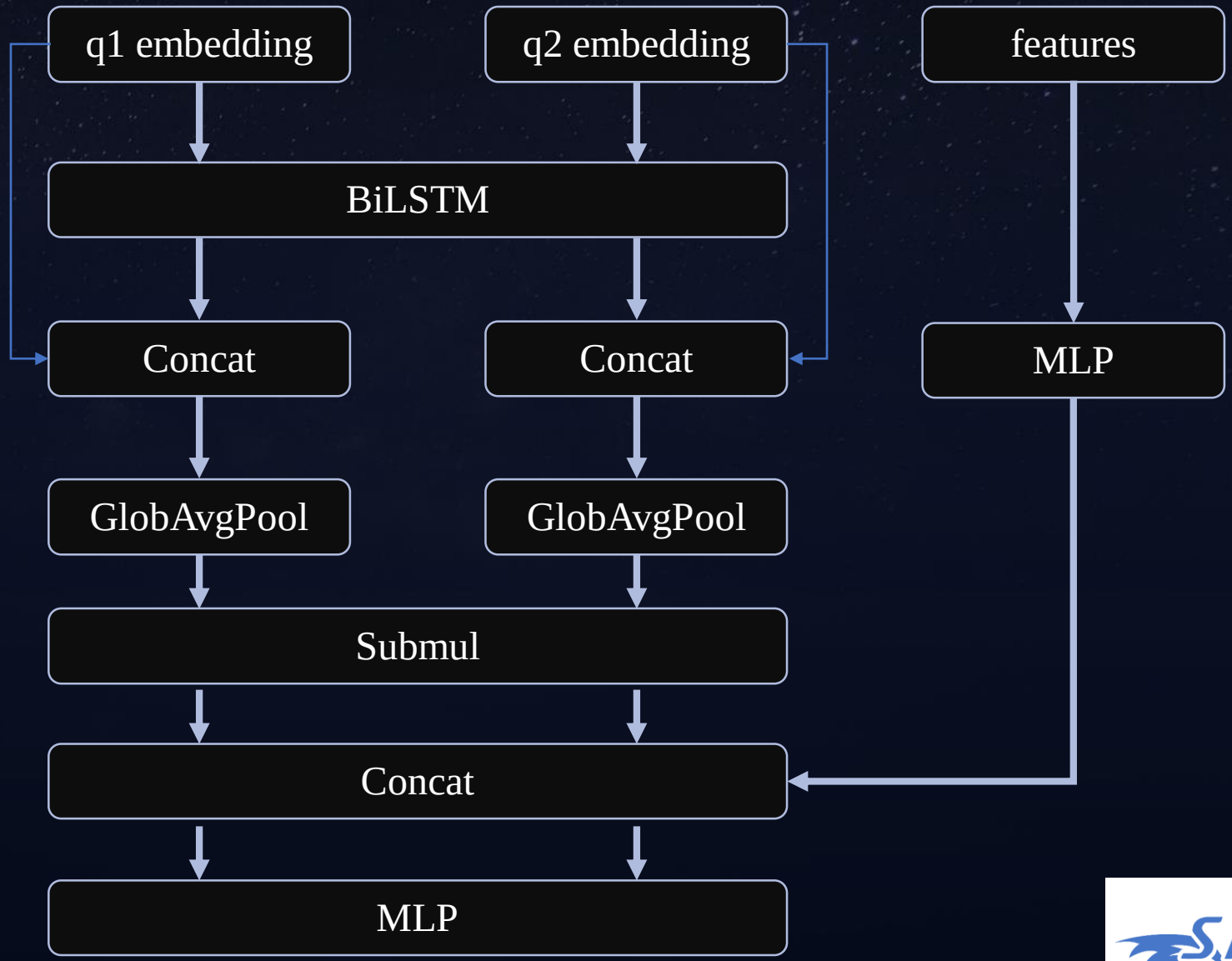
1,Q1,Q2
1,Q3,Q1

模型无法正确学习Q1与Q2\Q3的相同,而是会认为只要input里有Q1即为正样本。
需要通过数据处理让引导模型进行“*比较*”，而不是“*拟合*”。

我们的解决方案是，通过构建一部分补充集，对冲所有不平衡的问题。

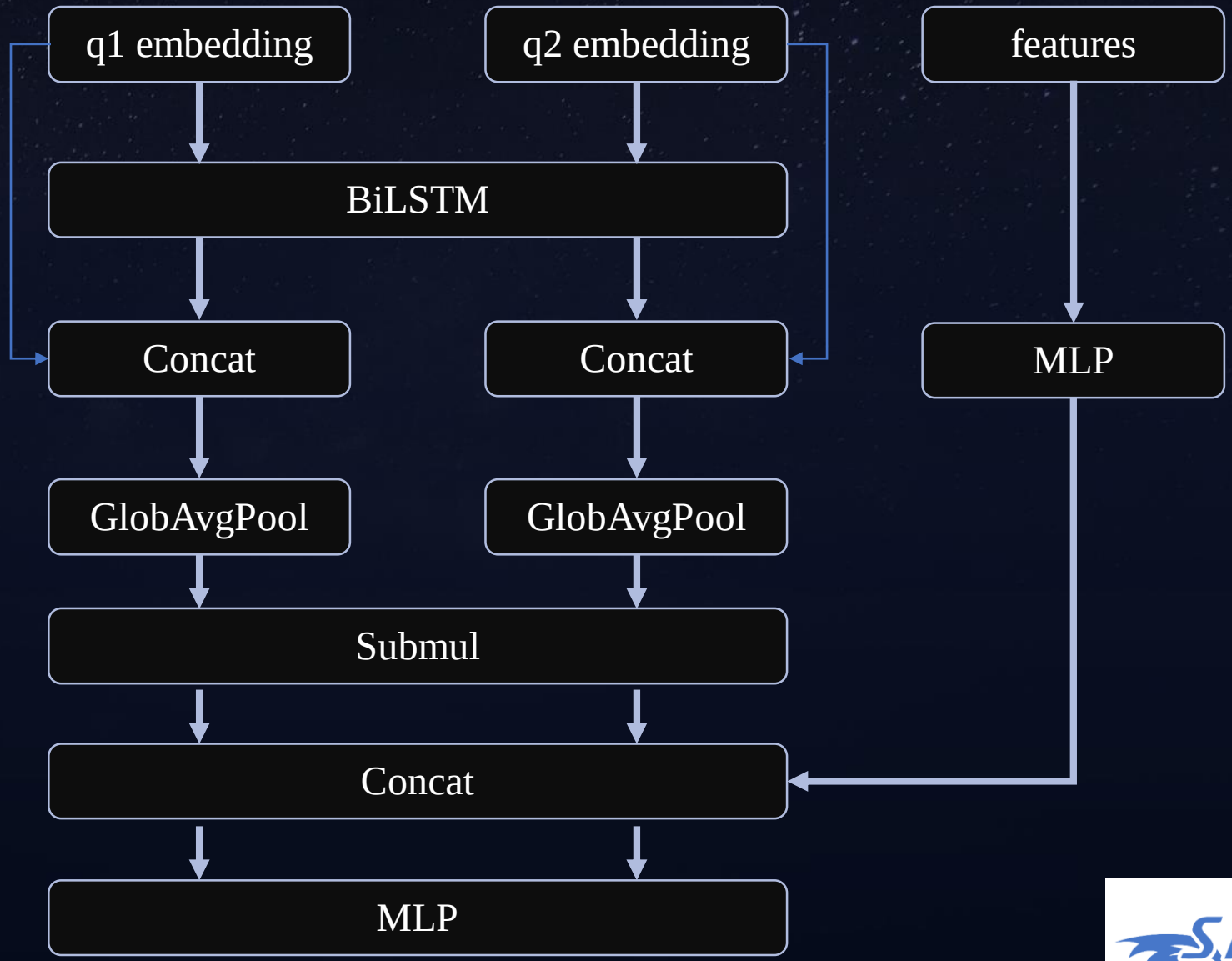
模型构建

- 简单的Siamese模型结构
- 引入Densenet的思路
- LSTM采用密集dropout防止过拟合
- Submul仅仅是相减与相乘两种运算
- **越简单的模型与越少的参数越不容易过拟合**



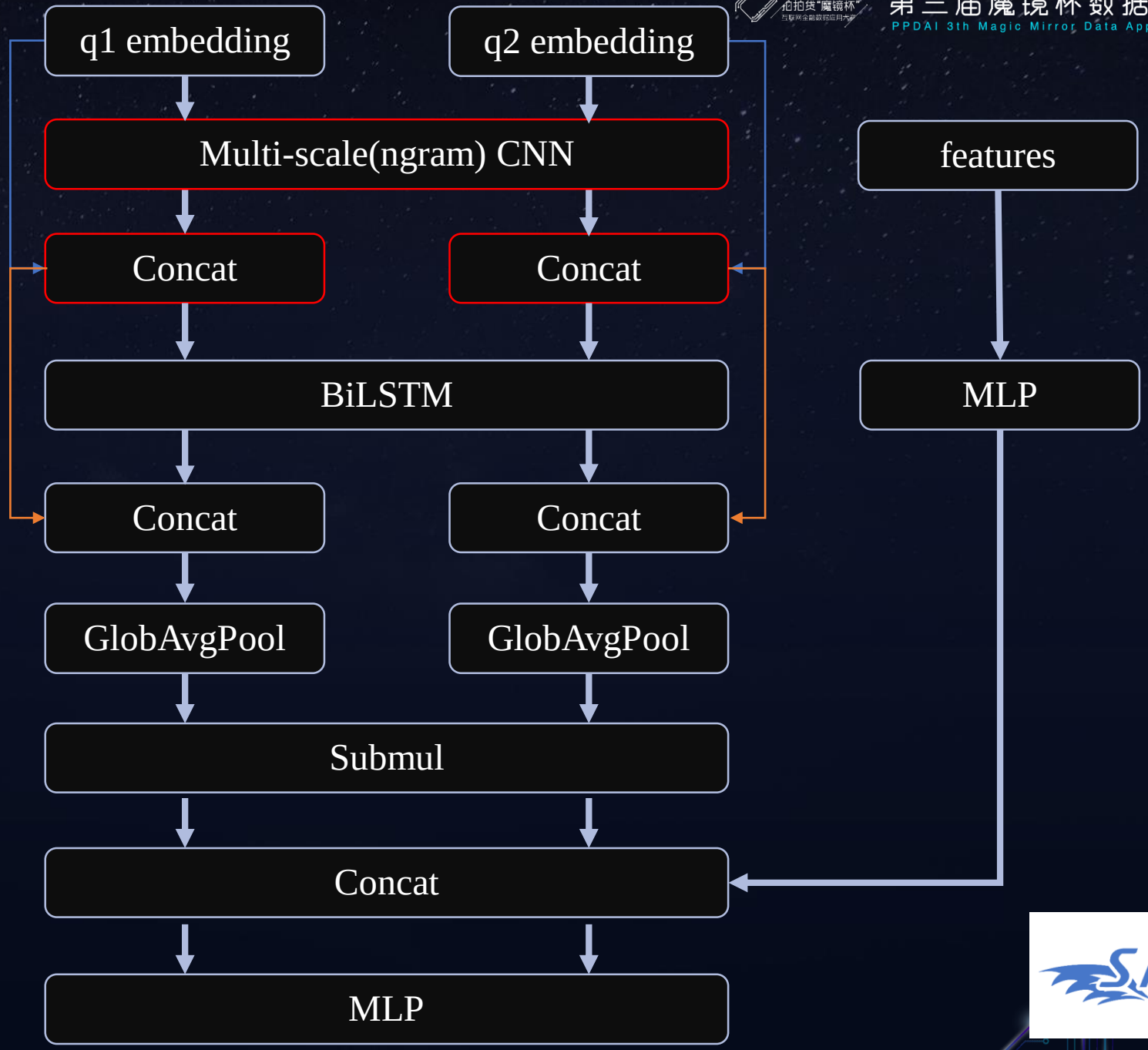
模型构建

- 三种交叉输入源
- 两种不同来源的embedding
- 数据增强集kfold
- 该模型共15个



模型构建

- 考虑ngram信息，在原模型基础上加入CNN层
- 该模型共4个



模型Finetune

1.gensim训练词向量

2.模型使用non_trainable词向量进行训练

3.将除了embedding的layer全部freeze，用低学习率finetune词向量层。

后处理

1.Infer机制：除了判断test集的每个样本得分以外，还会通过已知同义问题集的其他样本比对进行加权。

2.融合时轻微降低得分过高的模型权重，补偿正样本过多的影响。

3.将已知确认的样本修正为0/1。

THANK YOU

sky队

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

SKY