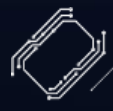
短文本匹配解决方案

王　煦　中　　李　政

大黑楼

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

团队成员介绍

王煦中

- 硕士毕业于解放军信息工程大学
- 师从清华大学知识工程实验室李涓子教授
- 在ACL等国际顶级会议及期刊发表论文多篇
- 主要研究方向为社交网络分析

李政

- 大连理工大学优秀硕士毕业生
- 在校期间获得多项国家荣誉及发明专利
- 2015年创立荤火科技有限公司至今，主要业务为nlp和图像挖掘，阿里云合作伙伴
- 2018年获得荣誉有天池千里马大赛冠军，天文大数据亚军

CONTENTS

01 任　务　描　述

任务描述

文本检索

复述问题

自动问答

对话系统

评判标准: $logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log(p_{i,j})$

02 数 据 分 析 与 处 理

02-1. 数据分布

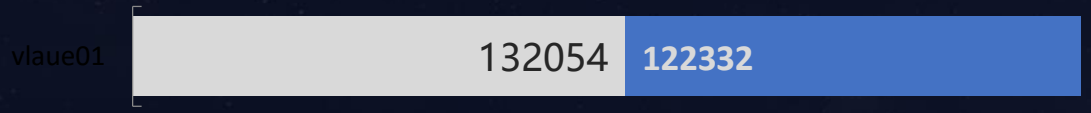
02-2. 特征工程

02-3. 相似性传递分析

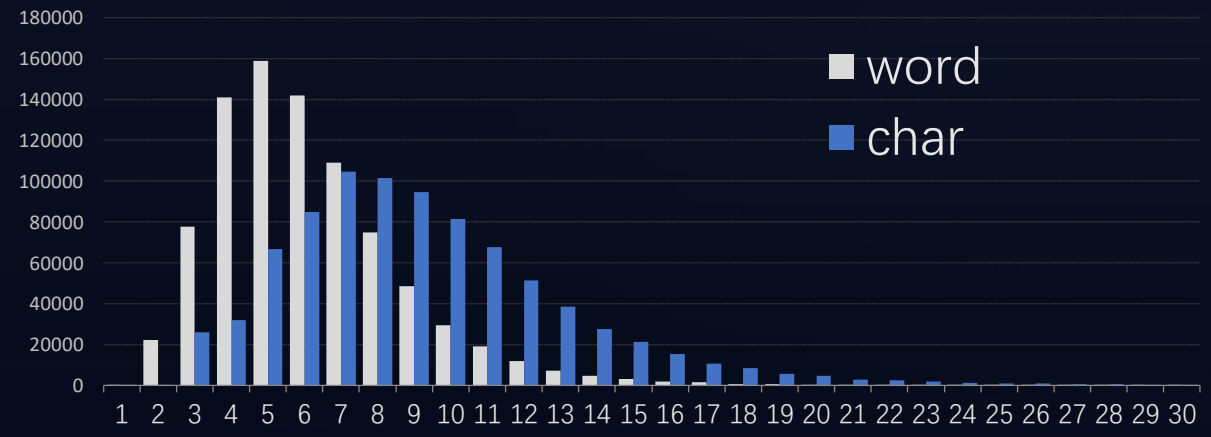
数据分布

- 数据分布

正例132054
负例122332



- 问题长度分布



- Padding

对问题长度大于length的截取中间length个字或词，
个数小于length，不处理

	question
Q000006	W04346 W17378 W06112 W05733 W18238 W05284 W11889 W18103
Q000011	W05733 W04745 W17070 W05184 W14103 W17706

特征工程

通过训练集与测试集构造连通图

频数特征 01

q1、q2在连通图中出现的频数

交集特征 03

q1、q2在连通图中的邻居节点交集中元素个数

交并集组合特征 05

q1、q2交并集特征之间的差、乘、除、最大、最小等

02 频数组合特征

q1、q2频数特征之间的差、乘、除、最大、最小等

04 并集特征

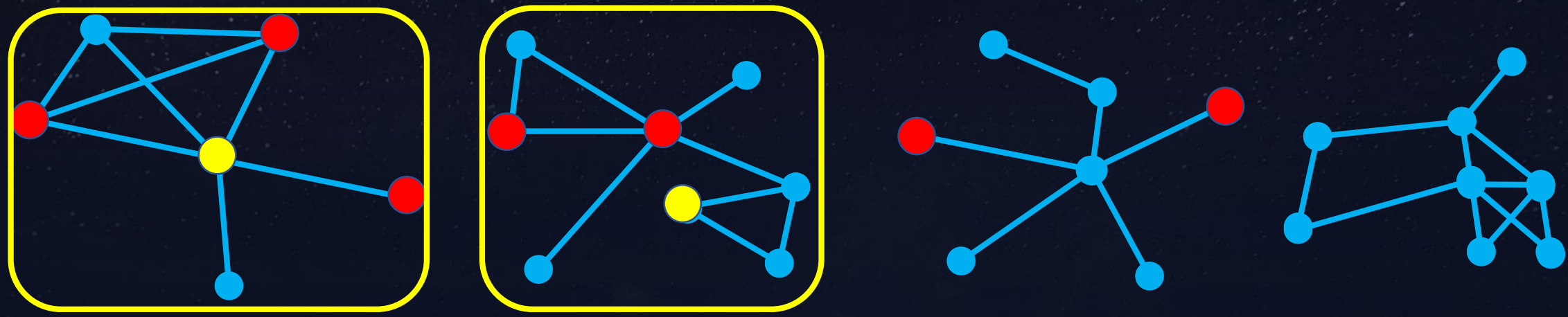
q1、q2在连通图中的邻居节点并集中元素个数

06 Max K-core特征

q1、q2在连通图中的最大核数

相似性传递分析

a~b & b~c -> a~c



连通子图构造

根据训练集中正例使用networkx构造了48691个连通子图

测试集正例发现

测试集中在同一子图的q1、q2必然相似
（发现12658个）

确定不相似子图对发现

训练集中负例q1、q2若在不同的一对连通子图中，则该对子图之间必然不相似
（发现27674对）

测试集负例发现

测试集中q1、q2属于确定不相似子图对的必然不相似
（发现5129个）

03 深度文本匹配模型

03-1. Improved Bi-LSTM

03-2. Simple ESIM

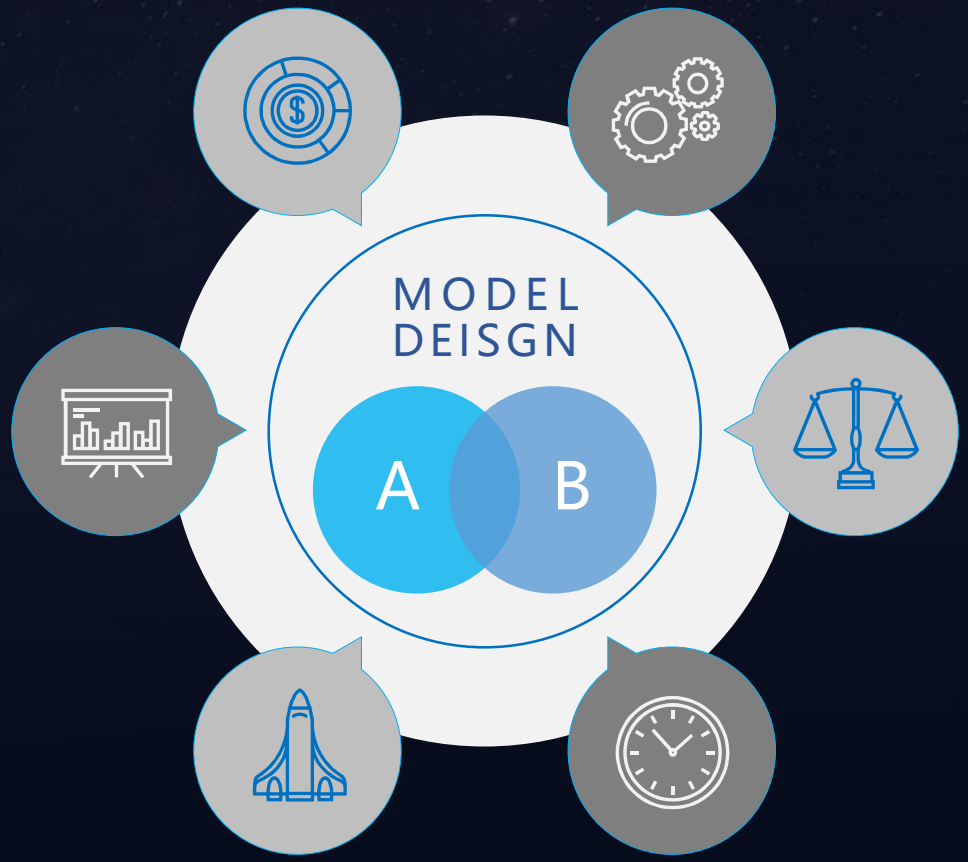
模型分析

01 MatchZoo

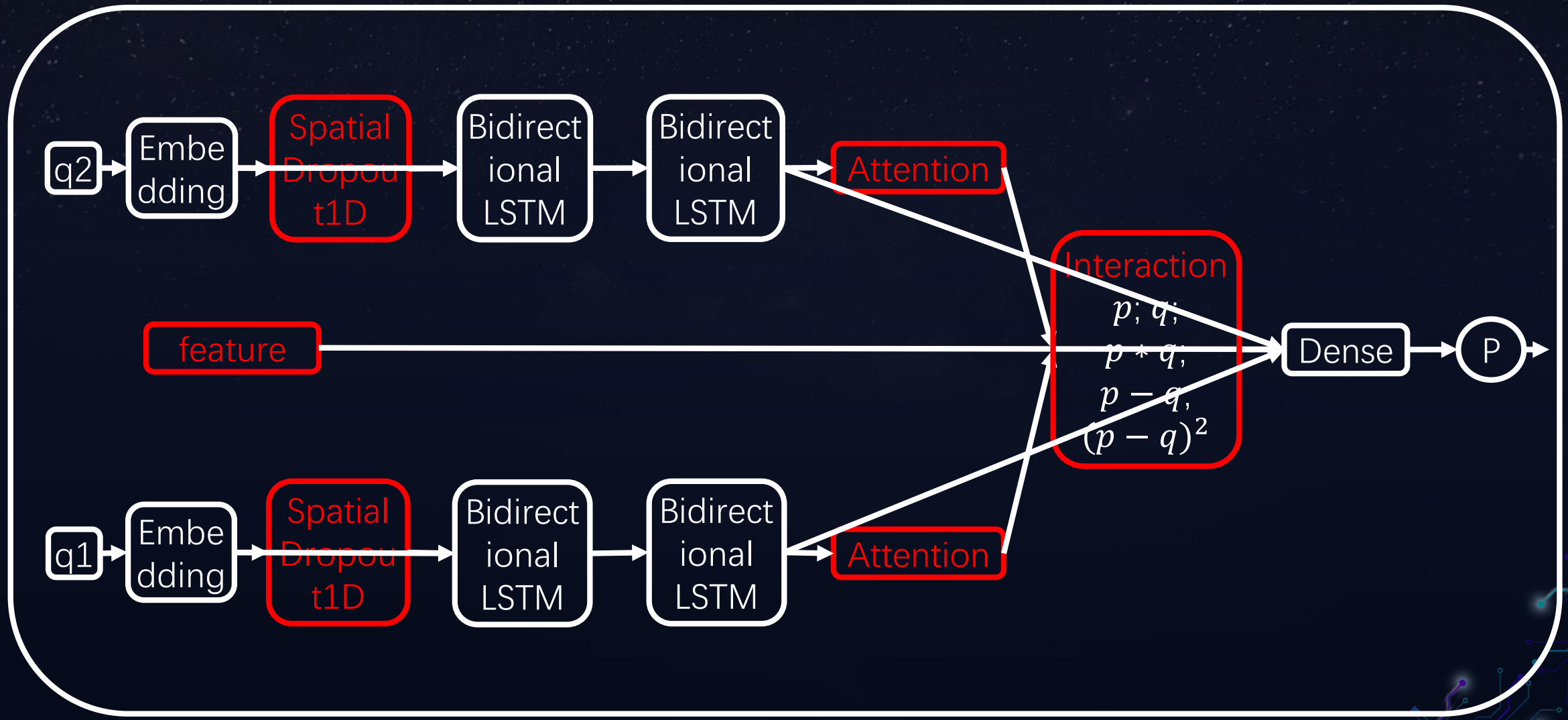
- 尝试了多种模型，但是过拟合严重
- 推断训练集存在大量重复数据，因此需要简化模型

02 由简至繁，由繁至简

- 改进Bi-LSTM
- 化简ESIM

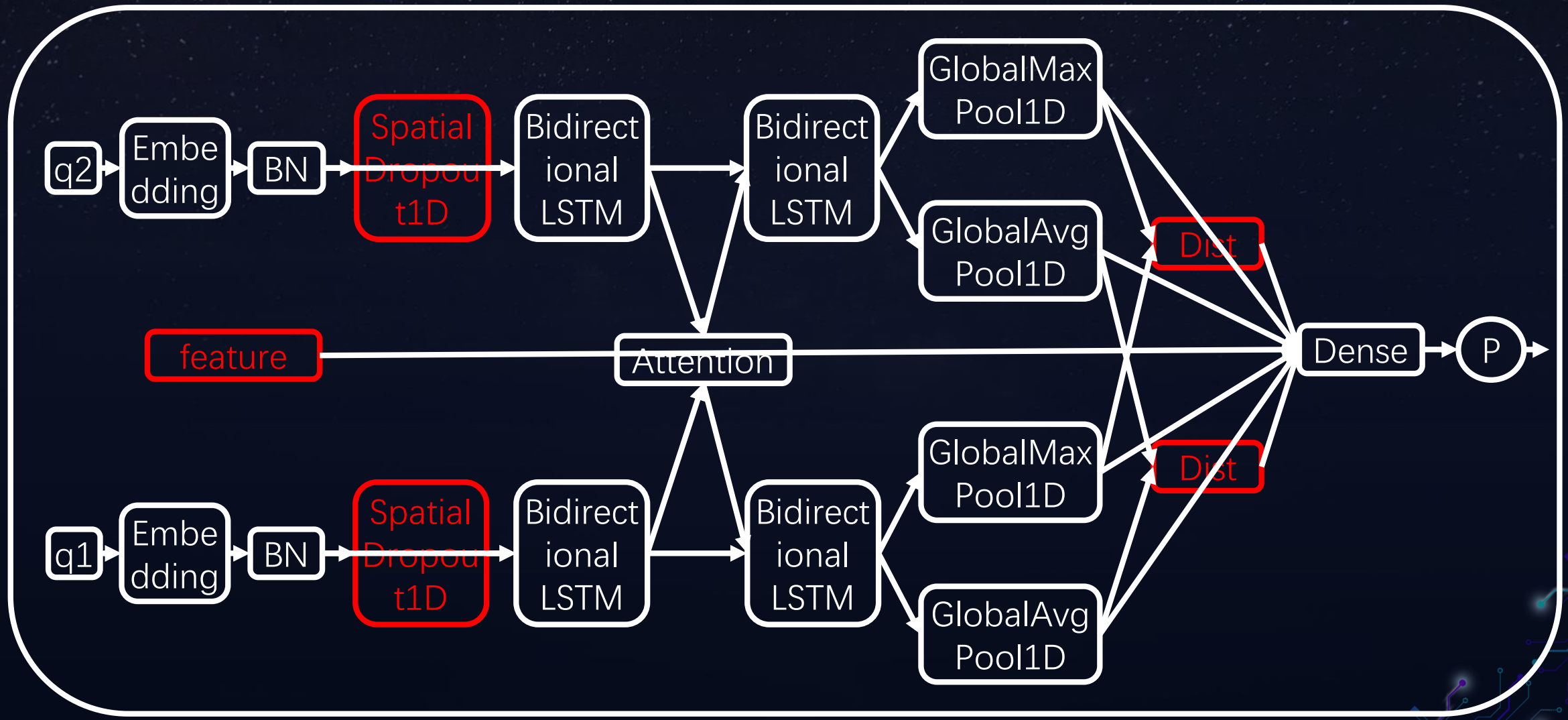


Improved Bi-LSTM



Simple ESIM

https://www.kaggle.com/lamdang/dl-models



04 模　型　融　合

模型融合

01 对调测试集

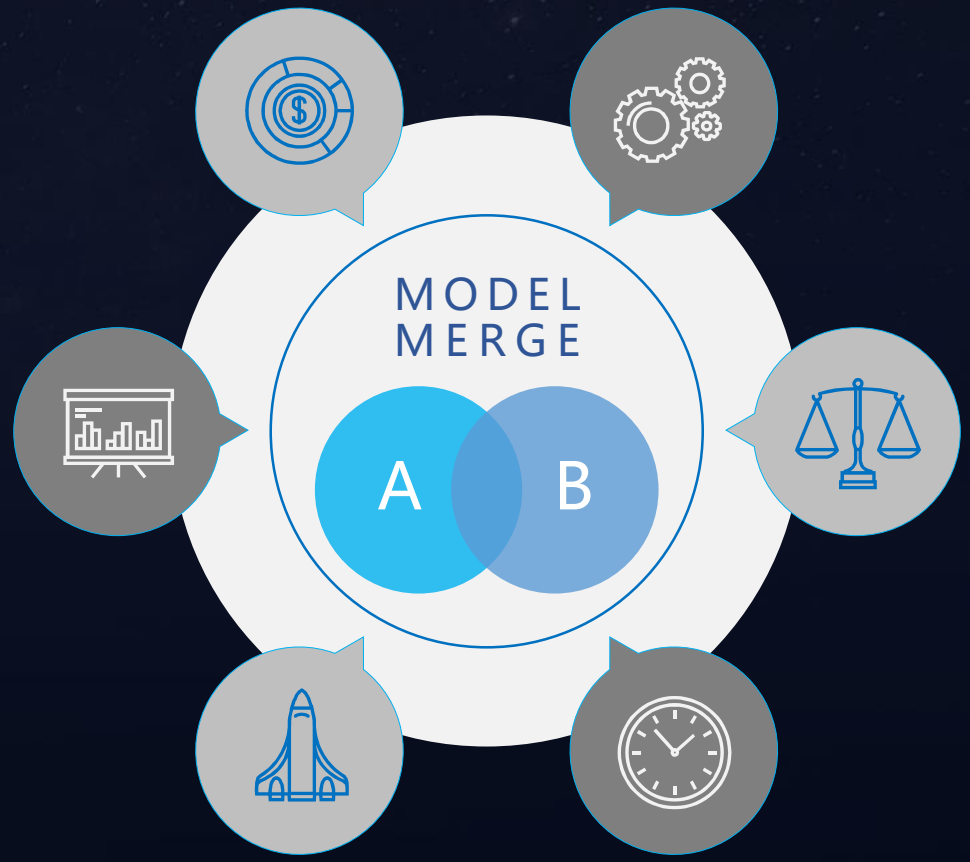
- 将（q1，q2）对调为（q2，q1）进行预测，结果取平均

02 Stacking

- 线下loss将至0.135，但是过拟合严重

03 Blending

- 调和平均数 hmean
- 调整权重



模型融合

Improved Bi-LSTM char	Improved Bi-LSTM word	Improved Bi-LSTM word+char	Simple ESIM char	Simple ESIM word	Simple ESIM word+char	hmean Simple ESIM+Improved Bi-LSTM	55 Simple ESIM+Improved Bi-LSTM (BEST)

05 总　　结

总结

01 数据可视化

- 准确把握数据整体情况
- 快速发现数据特征

02 数据特征发现

- 结构化图特征
- 基于相似性传递的联通子图

03 由简至繁，由繁至简的模型设计

- Improved Bi-LSTM
- Simple ESIM

04 模型融合

- Blending简单有效
- Stacking并非万能

THANK YOU

大黑楼

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