拍拍贷第三届魔镜杯大赛——

语义相似度算法设计

队伍名称：地表最强

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

答辩人：李博

CONTENTS

一 队伍介绍

李博

- 北京邮电大学
- 研二

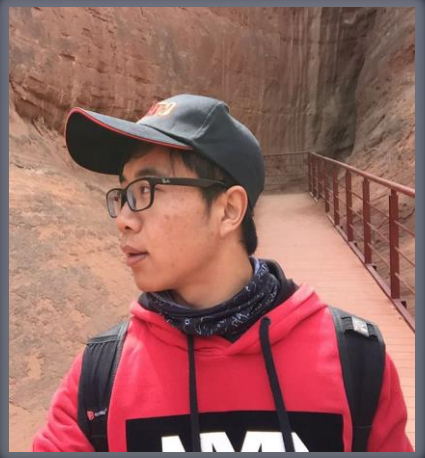
梁少强

- 广州某公司
- NLP高级算法工程师

吴祖平

- 北京航空航天大学
- 研三

包梦蛟



- 北京航空航天大学
- 研一

顾聪聪

- 杭州电子科技大学
- 研二

一 队伍介绍

团队成员所获奖项

国内：

1、2017年CCF大数据计算智能大赛（BDCI）小超市供销存优化赛题；**冠军**
2、2017年CCF大数据计算智能大赛（BDCI）企业经营退出风险预测赛题；**亚军**
3、2017年CCF大数据计算智能大赛（BDCI）360人机对决赛题：**季军**
4、移动公司4G用户流失预警赛题；**一等奖**
5、2017年首届腾讯社交广告高校算法大赛——移动App广告转化率预估；**季军**
6、2017年首届JDATA算法大赛——高潜用户购买意向预测；**6/4242**
7、2018年北京市校园高校大数据竞赛 校园人流量预测；**亚军**
8、2018年第二届腾讯广告算法大赛；**季军**
9、2017中诚信征信算法建模大赛；**季军**
10、招商银行信用卡消费金融场景下的用户购买预测；**5/1586**
11、2018年华为软件精英挑战赛全国总决赛 **5/4000**

国际：

1、IJCAI-2018阿里妈妈搜索广告转化预测；**5/5204**
2、KDD-2018未来天气预测；**7/4170**
3、G-Financial Forecasting Challenge Can you predict the future?；**季军**
4、Kaggle TalkingData AdTracking Fraud Detection Challenge；**金牌**
5、Kaggle Toxic Comment Classification Challenge；**金牌**
6、Kaggle Corporación Favorita Grocery Sales Forecasting；**银牌**
7、Kaggle IEEE's Signal Processing Society - Camera Model Identification；**银牌**
8、Kaggle Avito Demand Prediction Challenge；**银牌**
9、Kaggle Quora Question Pairs；**银牌**

二　解决方案

提高语义相似度算法准确率的两个思考方向：

一、采用合适的**数据预处理**和**数据增强方法**

二、寻找最近几年 **state-of-the-art**的深度学习模型结构，从**复现顶会论文**的过程中得到启发（模型主要来源：**SNLI天梯榜**https://nlp.stanford.edu/projects/snli/）

2.1 数据预处理

① 使用多种词向量：

✓ 主办方提供

✓ 从比赛数据集中，**由word2vec训练的字向量、词向量**(100dim，300dim， 500dim)

✓ 从比赛数据集中，**由glove训练的字向量、词向量**（100dim，300dim，500dim）

② 重点关注字向量：

✓ 由于中文分词难度较大，特别是**不同领域内的领域分词没有很好的解决方案**（本次赛题的数据为脱敏的金融领域数据源），对于词向量我们不能完全相信，所以**将更多的注意力关注在字向量上面**

③ 使用数据增强：

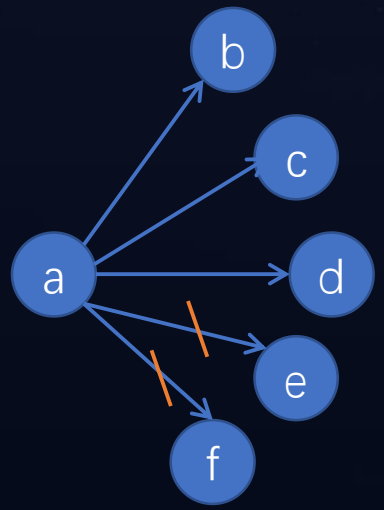
✓ **由文本生成关系图，推导样本间关系**

✓ **借鉴计算机视觉领域的Mixup增强**

二　解决方案

2.1.1 由训练集文本生成关系图，推导样本间关系

eg: a = b
a = c
a = d
a != e
a != f



eg: b = c
b = d
c = d
b != e
b != f
c != e
c != f
d != e
d != f
e ? f

New Psitive: 180W+
New Negative: 110W+

D1：新数据集正负样本Bagging

$$0.5N_0 \le size \le 1.5N_0$$

D2：原始数据集

2.1.2 Mixup 数据增强

paper1： Mixup: Beyond Empirical Risk Minimization
paper2： Data Augmentation by Pairing Samples for Images Classification

Idea：样本属性线性加权 Label线性加权

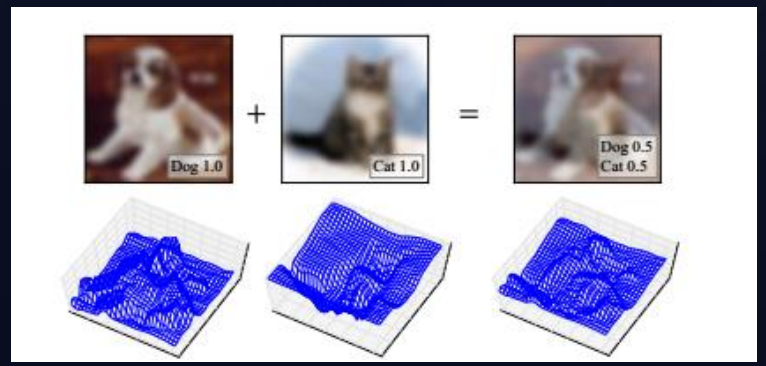


Image1:Dog Image2:Cat

↓ 对应位置像素线性加权

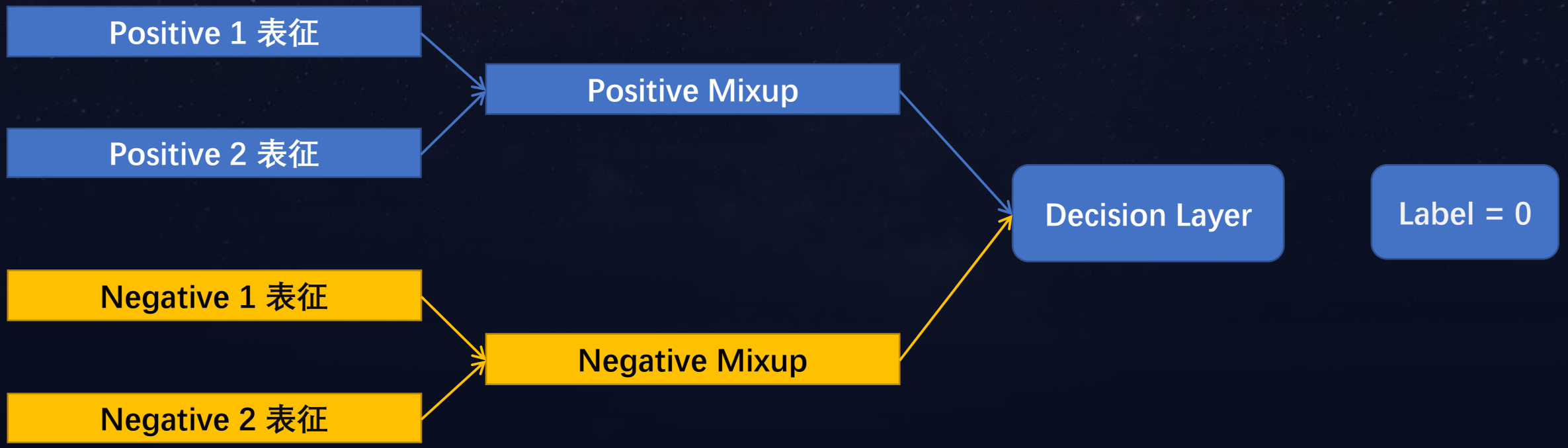
New Sample 0.5Dog + 0.5Cat

↓

Lable Dog:0.5 , Cat:0.5

如何应用于NLP领域?

2.1.2 Mixup In Net (semantic)



2.2 算法模型

表达式： 两句话分别独立编码表示，在最后的决策层信息汇合

交互式： 两句话在网络前部、中部或者后部进行信息交互，各自表达相互影响，并在最后决策层再次信息汇合

2.2 算法模型——表达式模型

paper3 : Convolutional Neural Networks for Sentence Classification
paper4 : Recurrent Neural Networks for Text Classification with Multi-Task Learning
paper5 : Recurrent Convolutional Neural Networks for Text Classification

eg : CNN-Baesd

优点：
结构**简单清晰**，训练速度快，自由度高，**易于理解**。

缺点：
句子之间的交互信息对于最终的决策具有重要价值，但是表征式模型无法提取这部分信息。

2.2 算法模型

模型	线上分数logloss
CNN-based	0.210398
RNN-baesd	**0.172610**
RCNN-based	0.187054

2.2 算法模型——交互式模型 Attention-Based

paper6 : Neural Machine Translation by Jointly Learning to Align and Translate
paper7 : ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs
paper8 : Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification
paper9 : Distance-based Self-Attention Network for Natural Language Inference
paper10 : Enhanced LSTM for Natural Language Inference
paper11 : DR-BiLSTM: Dependent Reading Bidirectional LSTM for Natural Language Inference
paper12 : A Compare-Propagate Architecture with Alignment Factorization for Natural Language Inference
paper13 : Supervised Learning of Universal Sentence Representations from Natural Language Inference Data

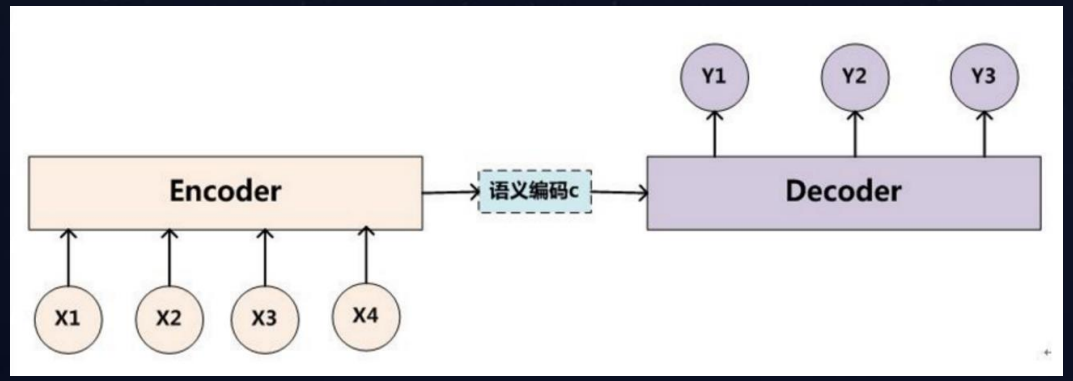
上游信息交互方法：
1.self-Attention
2.soft-Attention

下游信息表示方法：
1.embedding input
2.attention input
3.dot
4.sub
...

2.2 算法模型——交互式模型 Attention-Based

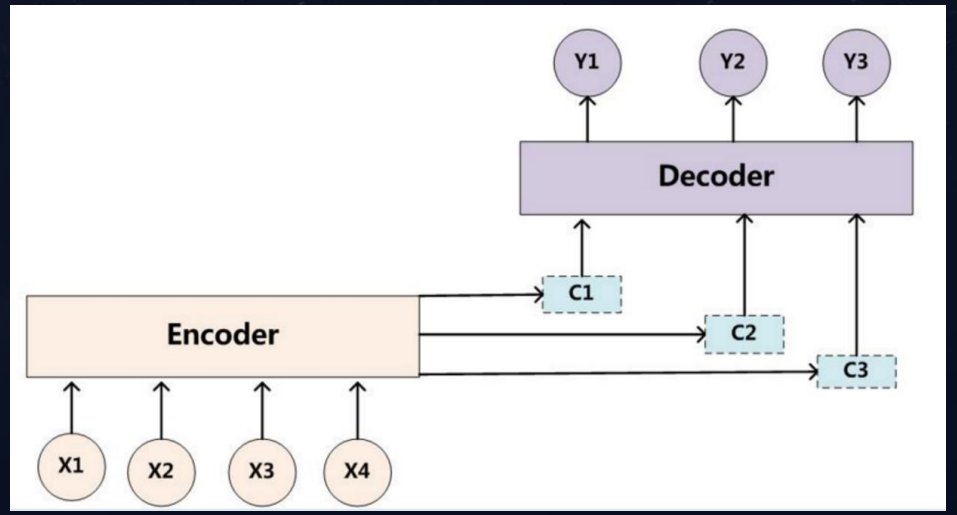
为什么要引入attention?

经典Encoder-Decoder框架



实践中发现的缺点：
1、RNN作为编码器时，编码器输出受到最后一个字符的影响较大
2、有效字符长度大于15时效果显著下降

Attention-Based Encoder-Decoder框架

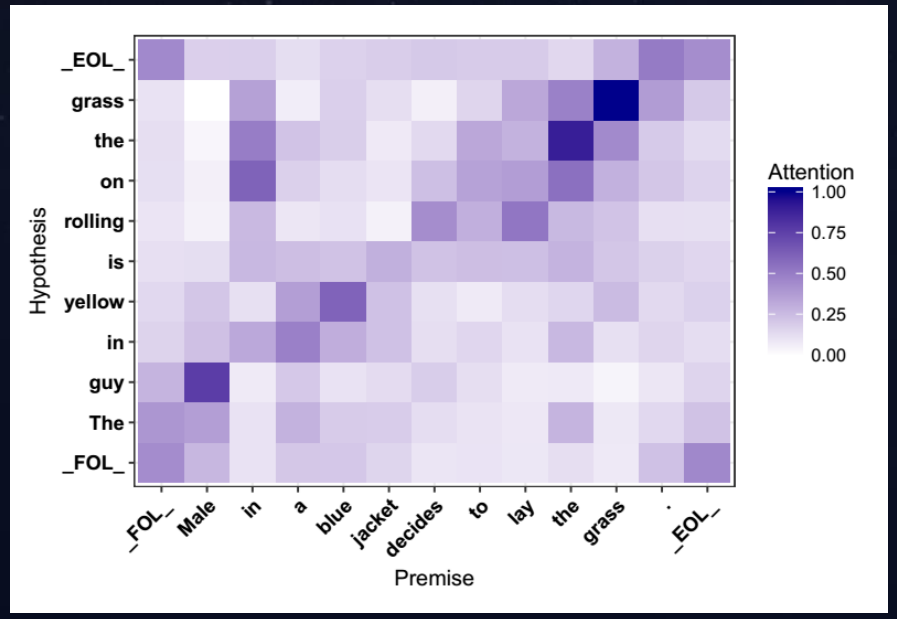


优点：
1、每个输出有各自权重
2、可以学习句子内部信息
3、可以学习两个句子之间的信息

encoder-decoder 图片来源：https://blog.csdn.net/qq_26609915/article/details/52086772

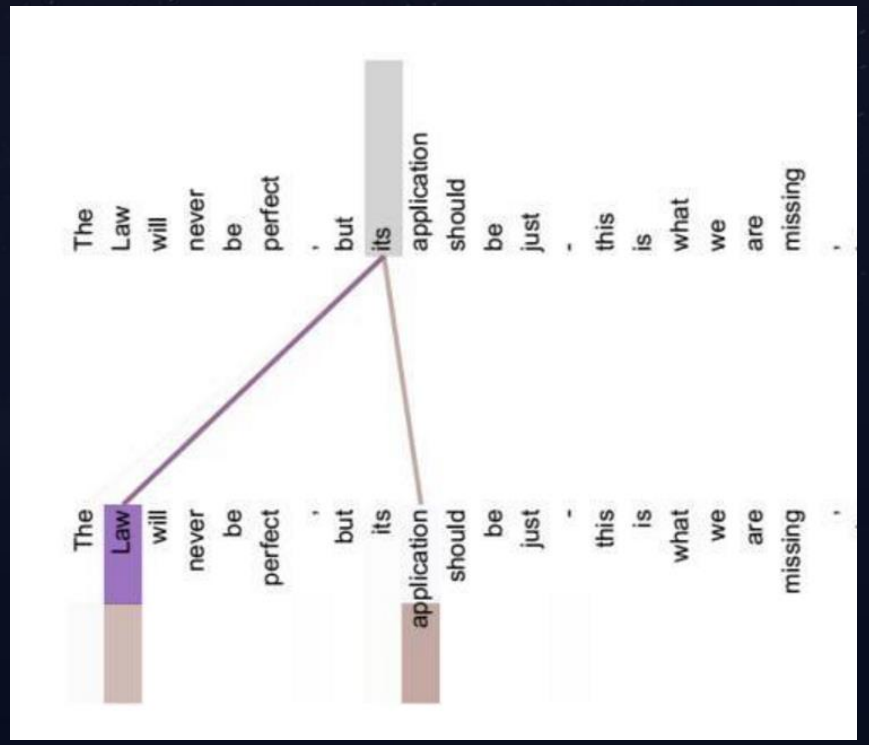
2.2 算法模型——交互式模型 Attention-Based

Soft -Attention



学习句子间信息，词语对应关系，对两句话的交互及理解有很大的作用

Self -Attention

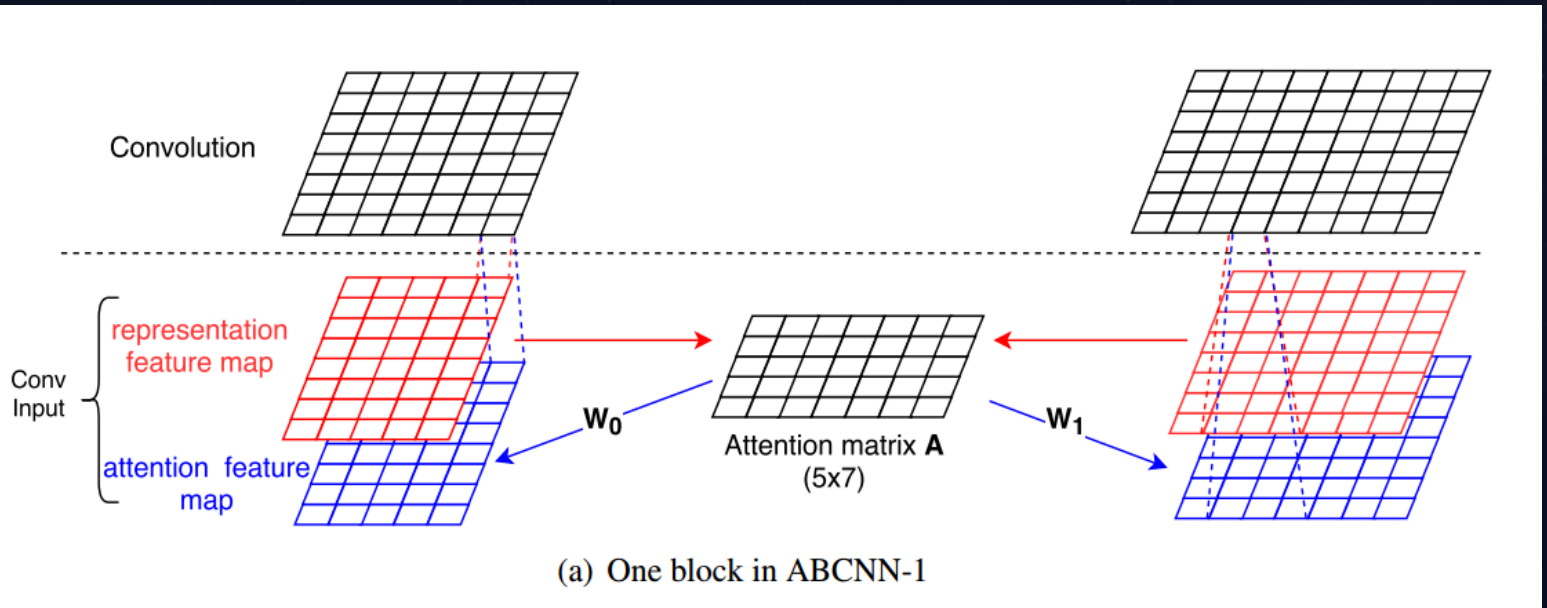


学习句子内部结构信息，语法信息，指代信息等，高度抽象语义。

图片来源：DR-BiLSTM: Dependent Reading Bidirectional LSTM for Natural Language Inference等

2.2 算法模型——交互式模型 Attention-Based

ABCNN(Attention-Based Convolutional Neural Network)



(a) One block in ABCNN-1

拓展

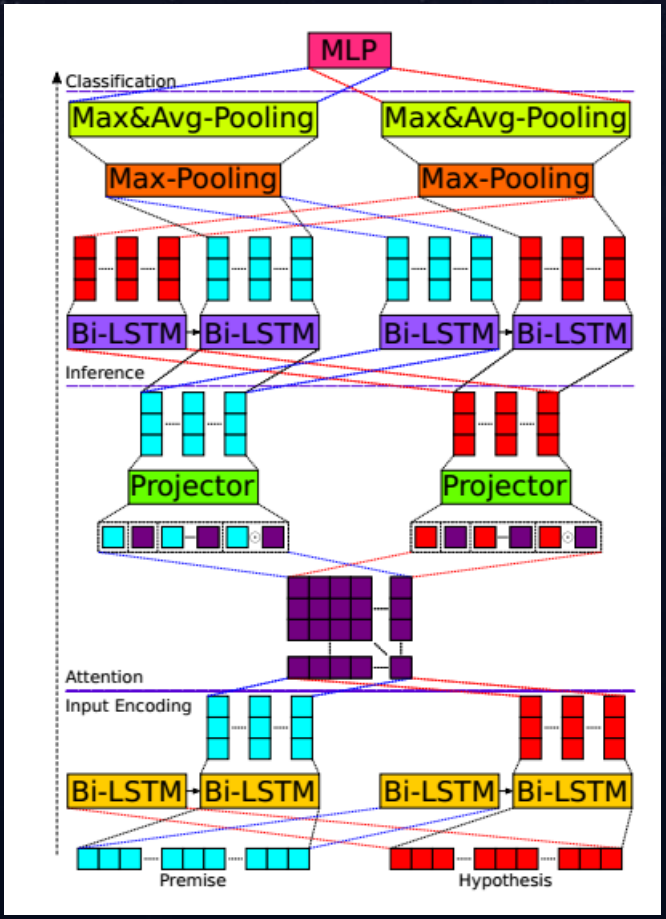
1、ABRNN
2、ABRCNN+kmax-pooling

attention weighted calculate
1、dot
2、Euclidean Distance
3、Cosine Similarity

ABCNN1 图片来源：ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs

2.2 算法模型——交互式模型 Attention-Based

DR-BiLSTM(Dependent Reading Bidirectional LSTM)



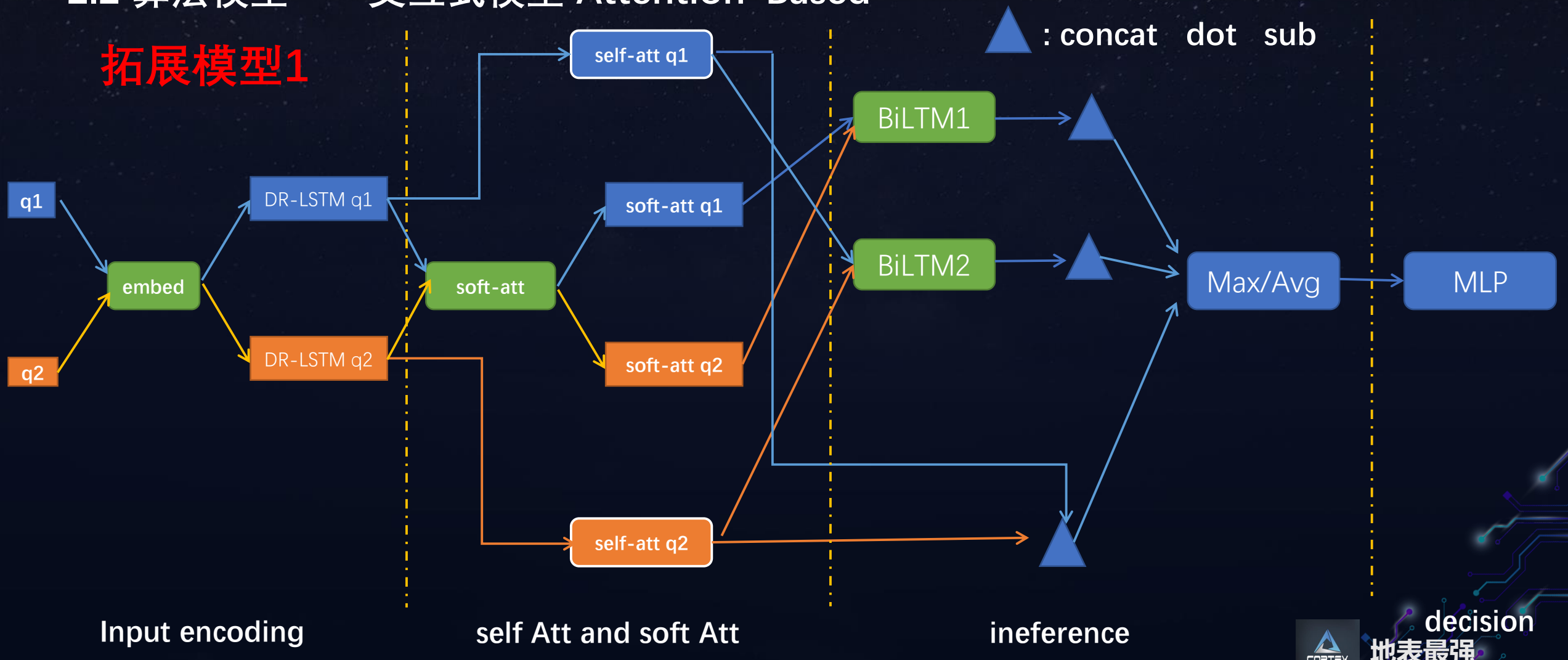
使用两个句子来编码目标句子

1、两个句子联合
2、将新的句子输入到LSTM
3、提取目标句子对应的输出部分作为新的编码

2.2 算法模型——交互式模型 Attention-Based

单模型初赛分数top10

拓展模型1



2.2 算法模型

模型	线上分数logloss
CNN-based	0.210398
RNN-baesd	**0.172610**
RCNN-based	0.187054
ABCNN	0.189573
ABRNN	0.167367
ESIM	**0.158268**
DR-BiLSTM	0.160314
拓展模型1	**0.155610**

2.2 算法模型——交互式模型 Attention-Based

paper14 : Semantic Sentence Matching with Densely-connected Recurrent and Co-attentive Information
paper15 : Bilateral Multi-Perspective Matching for Natural Language Sentences
paper16 : NATURAL LANGUAGE INFERENCE OVER INTERACTION SPACE
.....

特点：
1、更多的交互
2、复杂的内部特征提取模块

2.2 算法模型——交互式模型 Attention-Based

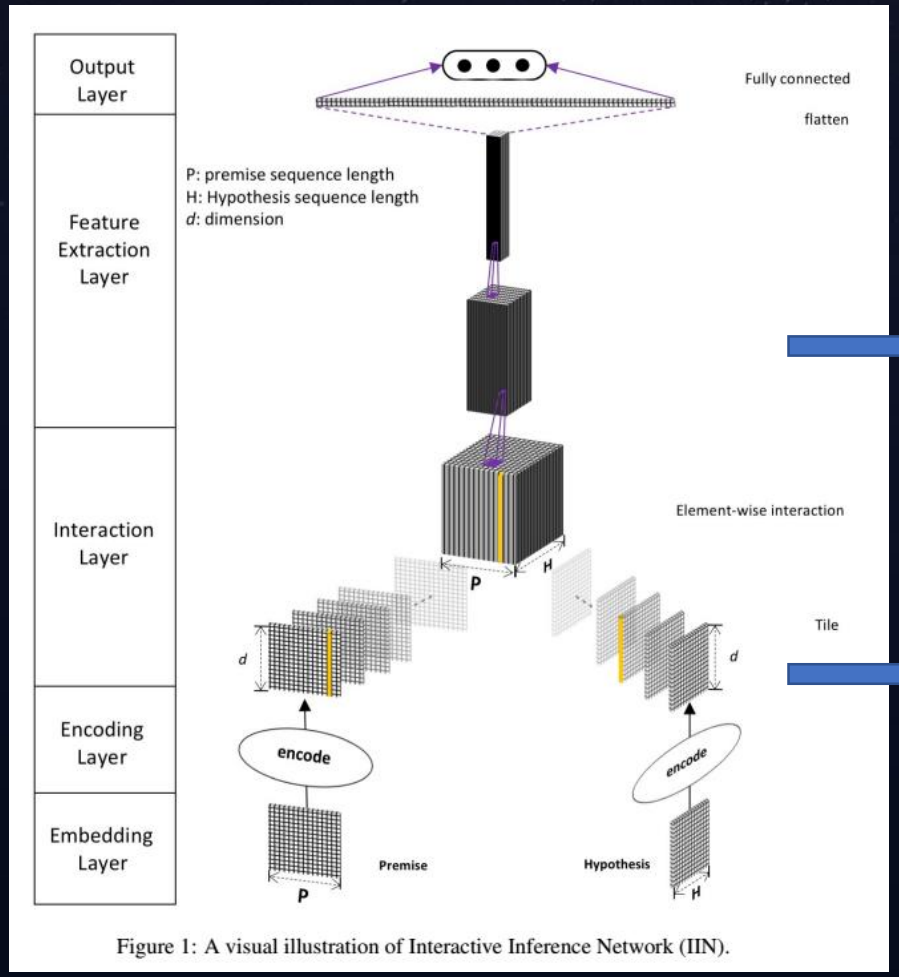


Figure 1: A visual illustration of Interactive Inference Network (IIN).

IIN(Interective Inference Networks)

交互推理网络

2、Feature Extration Layer

AlexNet, VGG, ResNet, DenseNet...

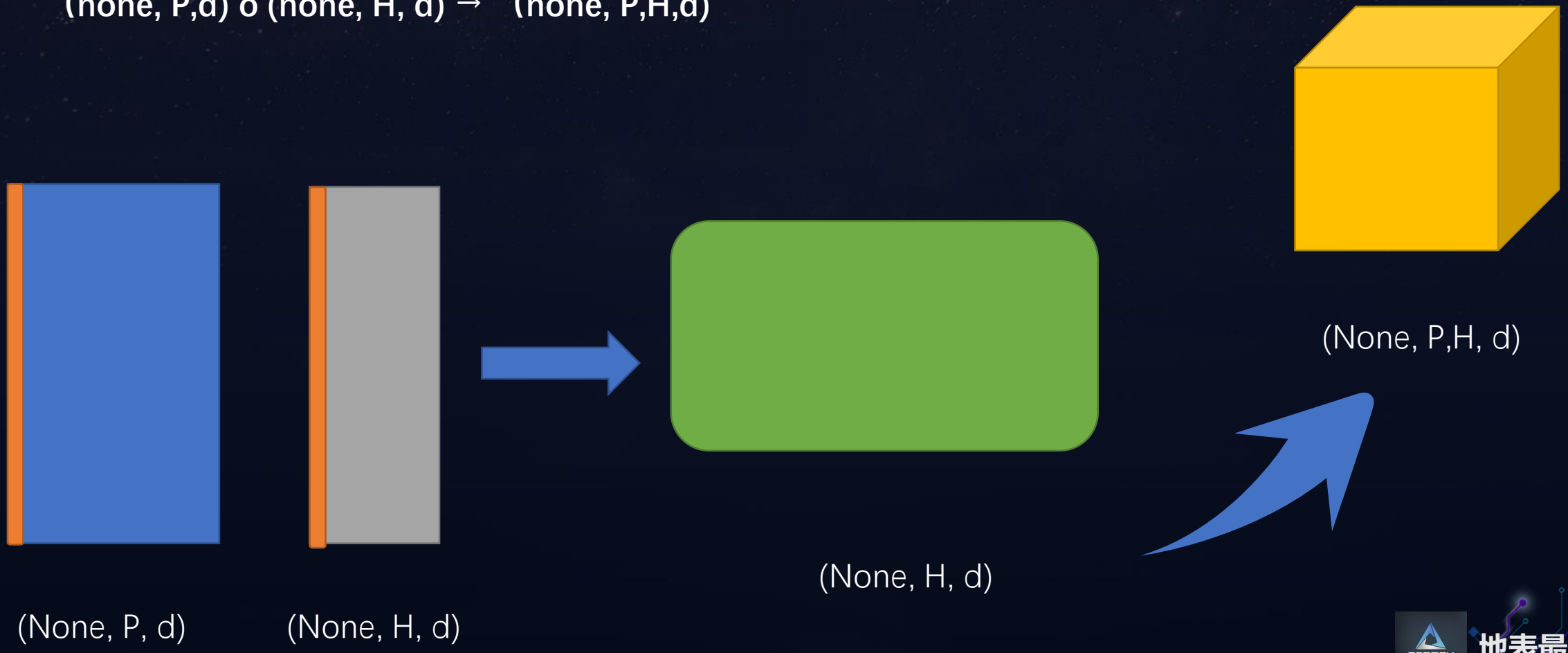
1、creates an word-by-word interection tensor

(none, P,d) o (none, H, d) → (none, P,H,d)

2.2 算法模型——交互式模型 Attention-Based

Creates an word-by-word interection tensor

（none, P,d) o (none, H, d) → （none, P,H,d)



二　解决方案

2.2 算法模型——交互式模型 Attention-Based

拓展模型2　单模型初赛分数top8 ⇒ 最终线上排名：top2

1、使用DR-LSTM编码目标句子

2、采用self-attention和soft-attention相结合首次提取句子信息

3、参考IIN思路，对attention信息进行元素间第二次交互，生成三维特征张量

4、参考DenseNet思路，对三维特征张量进行信息提取（只用2个dense block，节约时间，减小模型复杂度）

5、舍弃maxpooling/avgpooling

2.2 算法模型

最终各个模型线上成绩
（初赛复赛成绩均包括在内）

模型	线上分数logloss
CNN-based	0.210398
RNN-baesd	**0.172610**
RCNN-based	0.187054
ABCNN	0.189573
ABRNN	0.167367
ESIM	**0.158268**
DR-BiLSTM	0.160314
拓展模型1	**0.155610**
DIIN	0.154314
CAFE	0.156340
拓展模型2	**0.152213**
ensemble	**0.142747**

三　总结及致谢

总结：

1、参加比赛的目的是学习NLP的相关基本思想和研究方法，多从论文出发，耐心研读

2、不过分相信和依赖他人的开源复现，自己动手搭出来才最可靠

3、分析错误案例，找到可以改进的方面

三　总结及致谢

致谢：

1、感谢拍拍贷

2、感谢一直努力的队友和对手

3、感谢支持我们的亲人、老师和同学

厚 积 薄 发　天 道 酬 勤

CORTEX
地表最强

队伍名称：地表最强

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