大连理工大学

信息检索研究室

Information retrieval laboratory of DUT

平安医疗科技问句匹配

CHIP2018-评测任务2

参赛队：DUTIR

大连理工大学-信息检索研究室

2018年12月

成员介绍

- 研究室：大连理工大学信息检索研究室 （http://ir.dlut.edu.cn/）
- 指导教师：林鸿飞教授
- 汶东震：大连理工大学研究生2016级
- 岳天驰：大连理工大学研究生2017级
- 李英东：大连理工大学研究生2017级
- 李政：大连萃火科技
- 吴飞：解放军总医院第八医学中心信息科

目录

任务介绍

任务描述

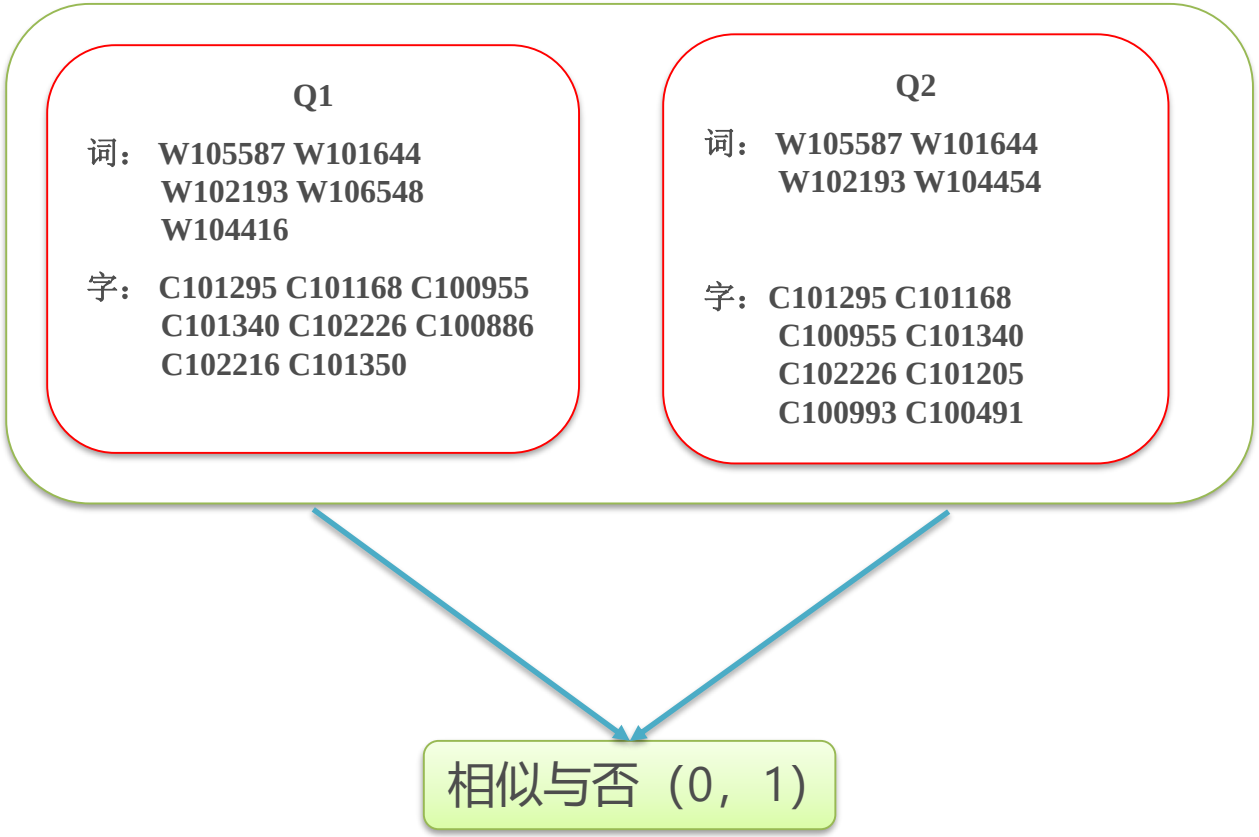
任务：判断健康咨询问句对的语义意图是否相似

相似

- 问句1：糖尿病吃什么？
- 问句2：糖尿病的食谱？

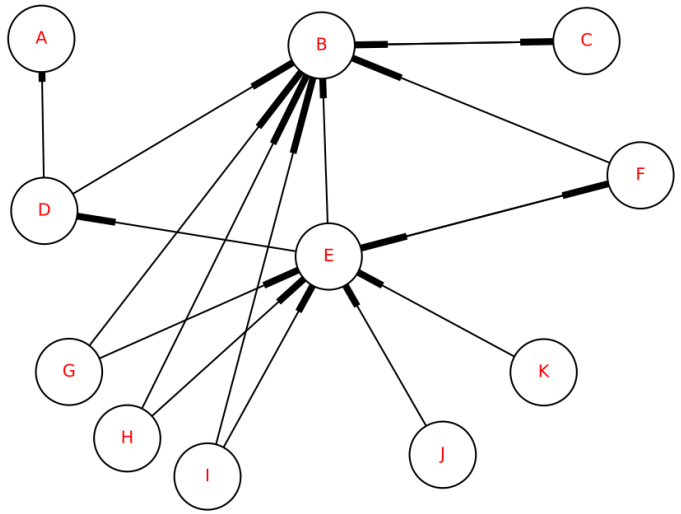
不相似

- 问句1：糖尿病的危害？
- 问句2：糖尿病肾病的危害？



数据分析

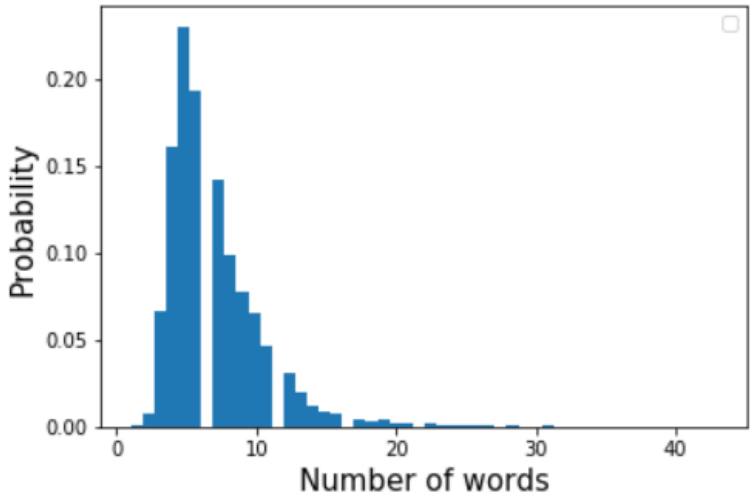
- **数据分布**
  - 训练集共两万条，正负样本比例1：1。
    - qid1，14915条；qid2，14884条
  - 测试集共一万条。
    - qid1，8276条；qid2，8301条
- **网络分析**
  - train/test集合无重复节点
    - 图相关特征
  - 通过传播相似得到扩展数据4712条
    - P1:(A,B) ==>1;
    - P2:(B,C) ==>1;
    - H:(A,C) ==>1;

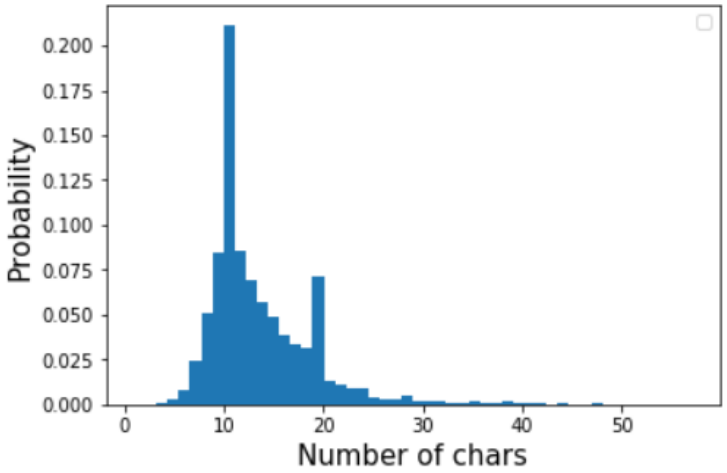


数据分析

- **数据分布**
  - 词表，9647；字表，2307
  - 测试集相对集外词：2042
    - 占训练集总字数比：26.82%
  - 测试集相对训练集集外字：223
    - 占训练集总字数比：10.69%

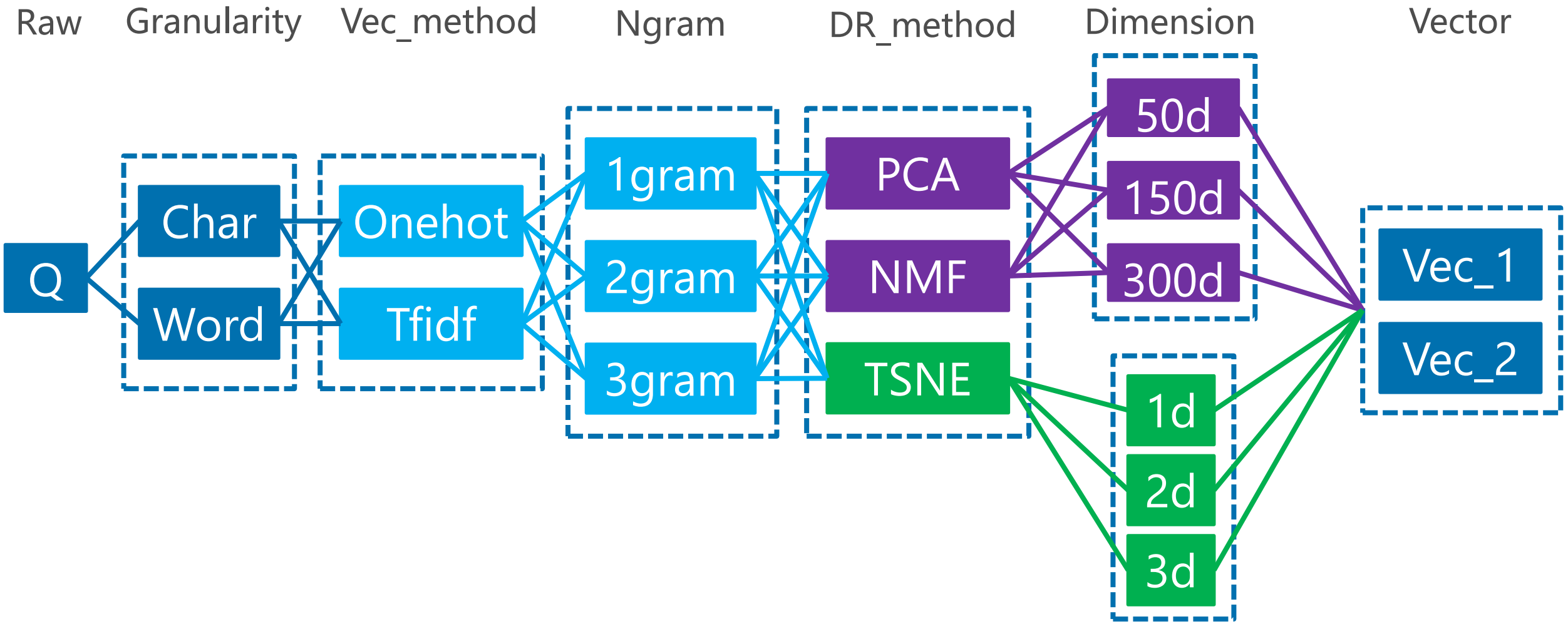
- **句子长度分布**
  - 短文本，词数<10，字数~10



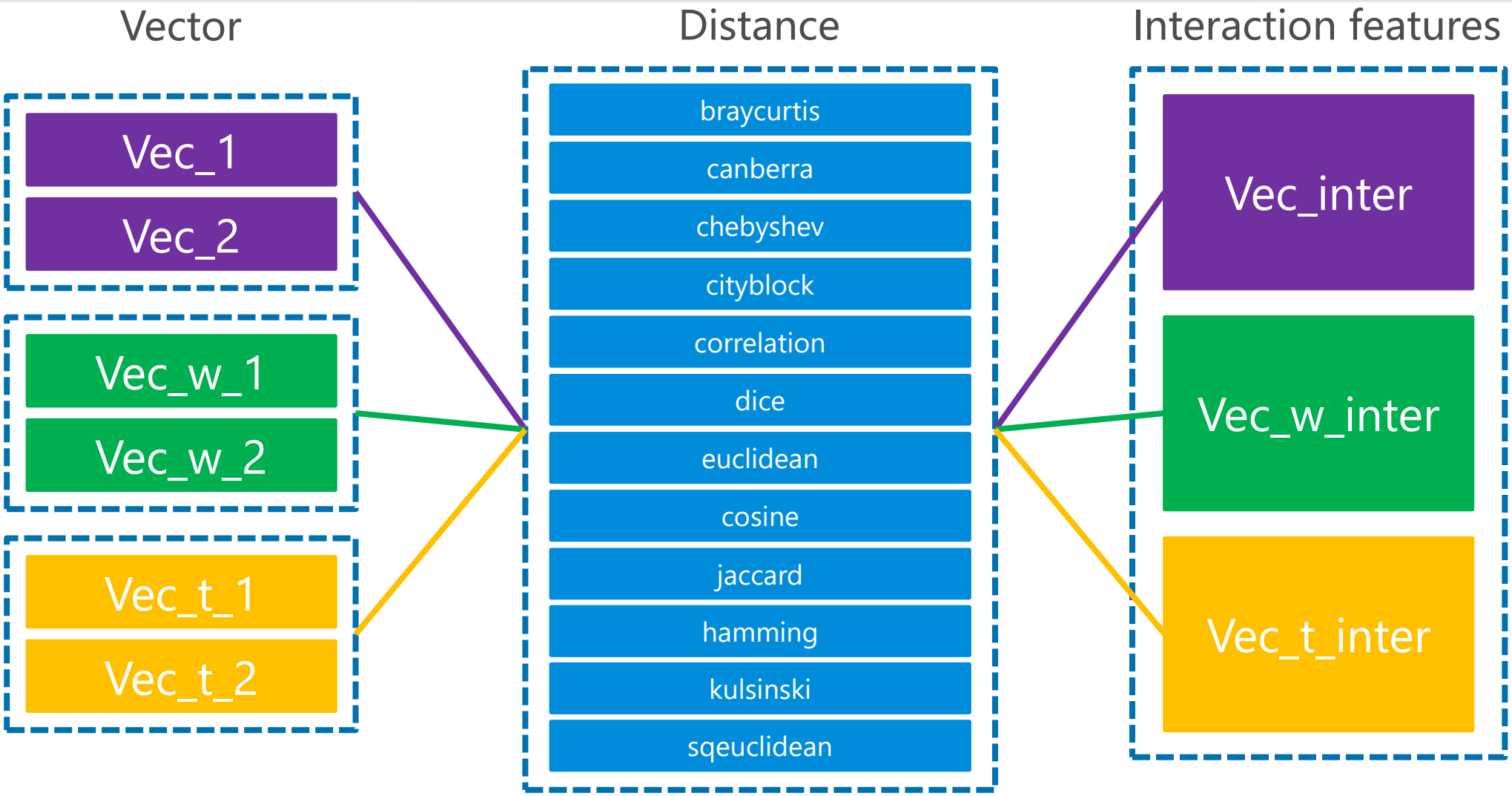


特征工程

向量化


Raw
Granularity
Vec_method
Ngram
DR_method
Dimension
Vector
Q
Char
Word
Onehot
Tfidf
1gram
2gram
3gram
PCA
NMF
TSNE
50d
150d
300d
1d
2d
3d
Vec_1
Vec_2

向量交互


Vector
Distance
Interaction features
Vec_1
Vec_2
Vec_w_1
Vec_w_2
Vec_t_1
Vec_t_2
braycurtis
canberra
chebyshev
cityblock
correlation
dice
euclidean
cosine
jaccard
hamming
kulsinski
sqeuclidean
Vec_inter
Vec_w_inter
Vec_t_inter

特征工程

- **以下的token分别包含(char和word)**

统计特征

- 句子token数目，句子去重token数，二者的比例以及差值。

相同token

- 句子对共享token的数目，句子对token的jaccard系数
- 句子对共享token idf 加权得分，句子对共享token占原句的比例

重点区分token

- onehot 特征，比率特征

特征工程

编辑距离

- Fuzzywuzzy，编辑距离
- 最长公共字串

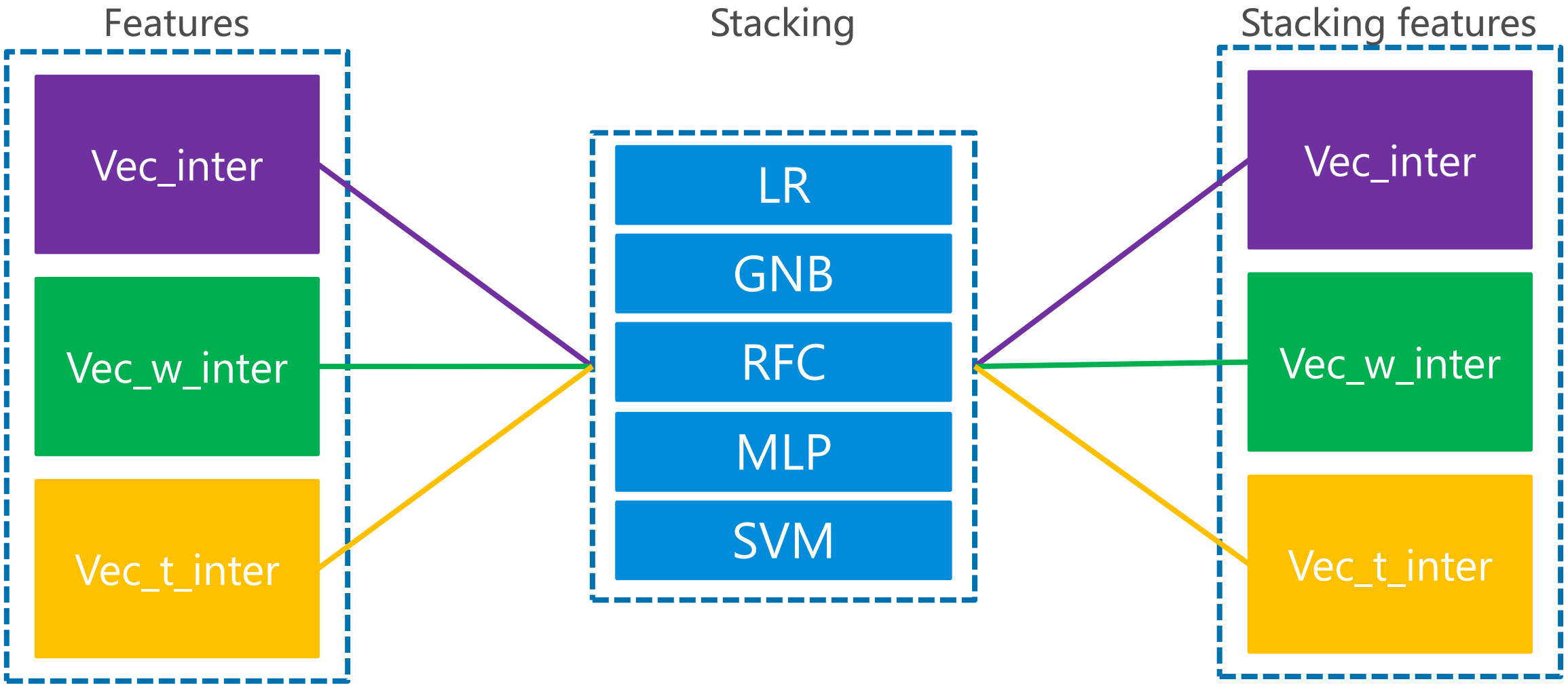
词向量

- 句子对token的wmd距离，句子对向量表示的多种距离（余弦距离，欧式距离，曼哈顿距离等）。
- 利用gensim Word2vec 和 Glove 训练的词向量，特征同上。

TF-IDF

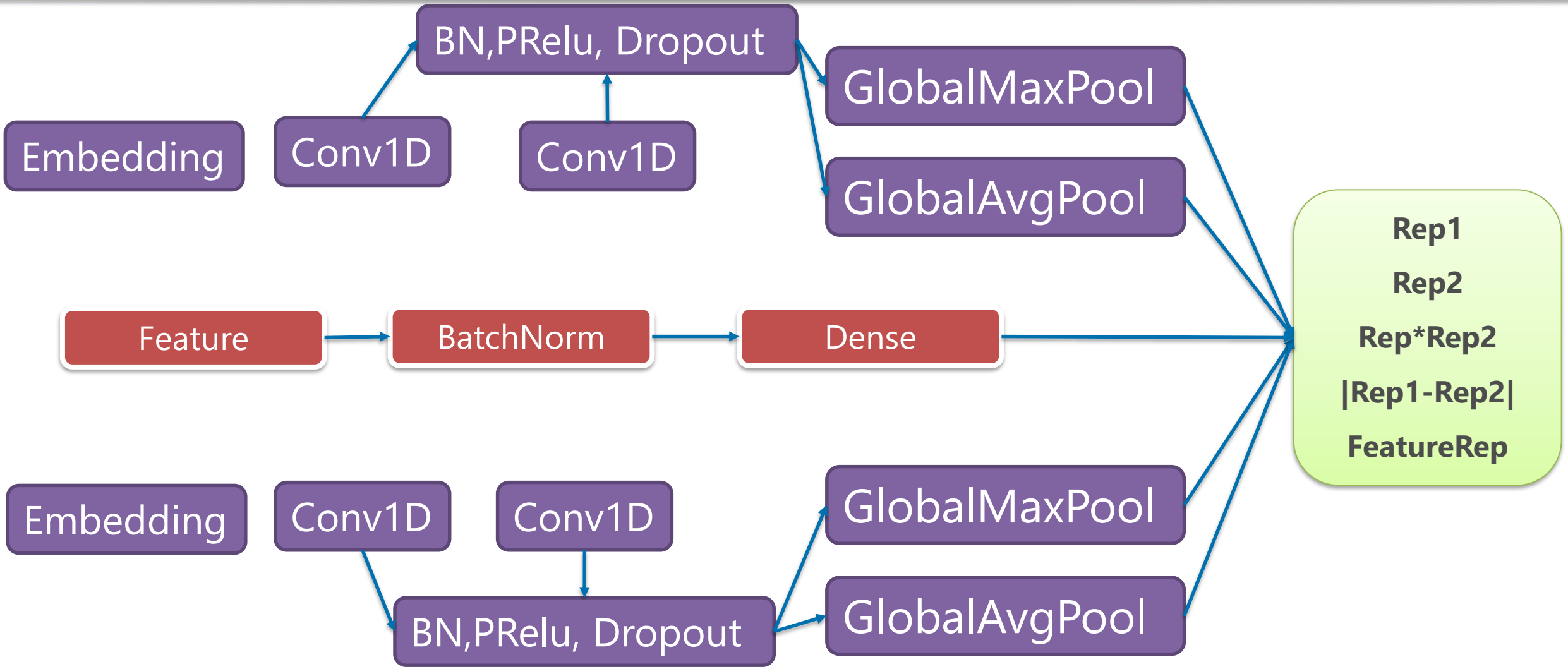
- 句子对tfidf和one hot 向量化的相似度计算交互特征。
- 利用PCA降维表示的相似度计算交互特征。

Stacking

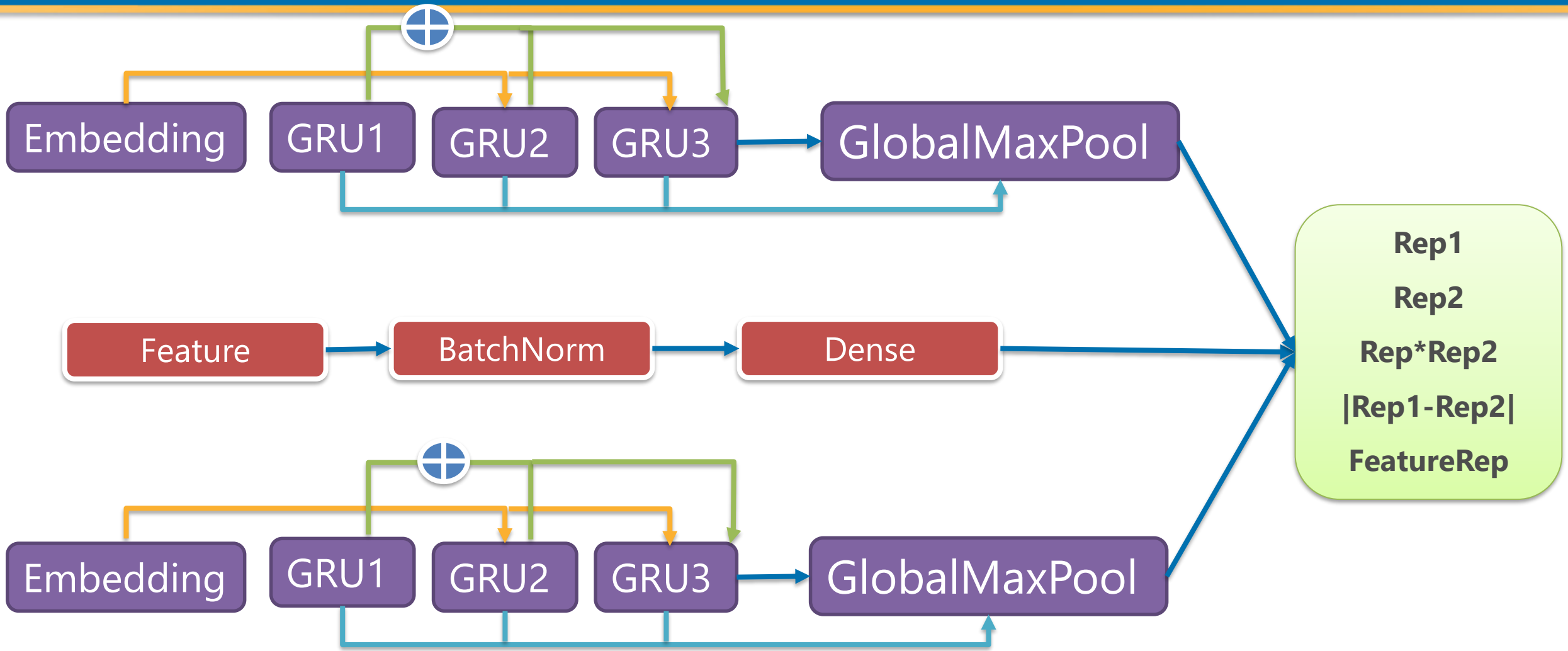

Features
Vec_inter
Vec_w_inter
Vec_t_inter
Stacking
LR
GNB
RFC
MLP
SVM
Stacking features
Vec_inter
Vec_w_inter
Vec_t_inter

模型方案

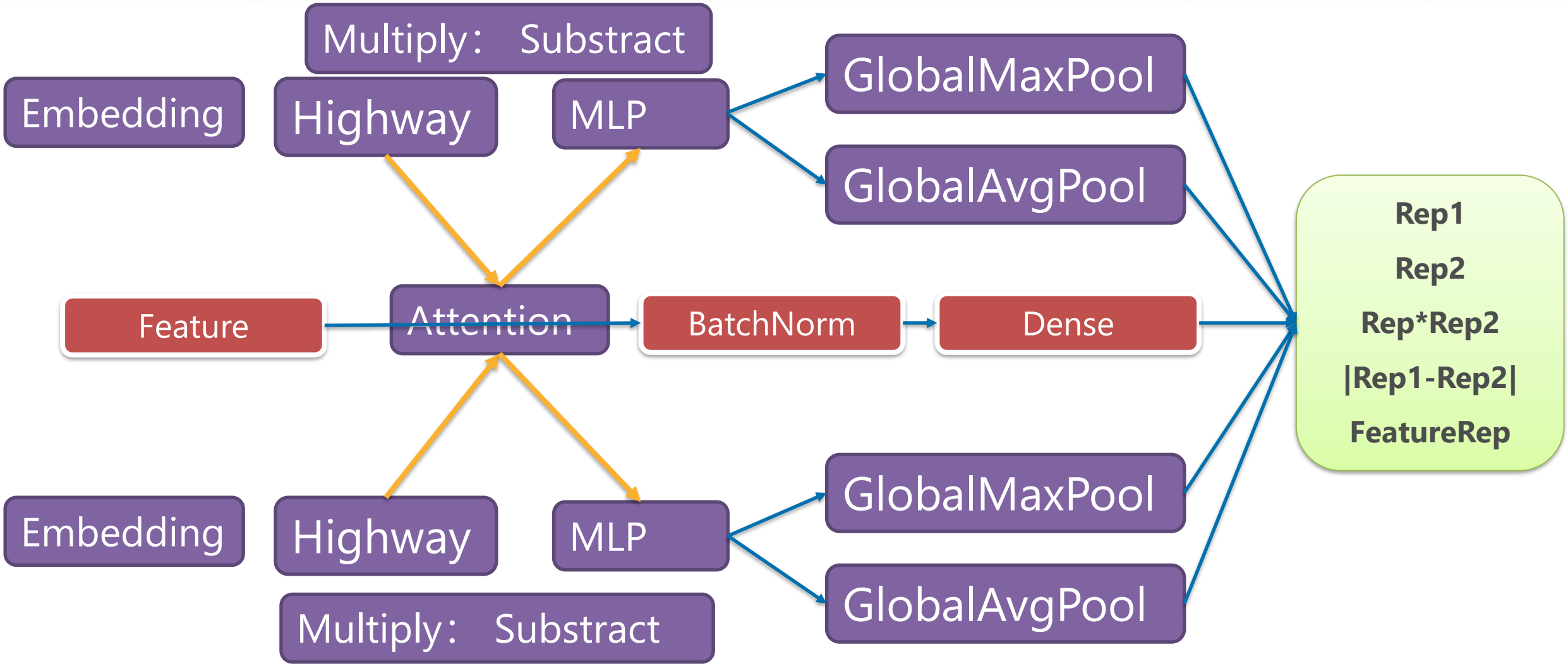
Stacked 2-Layers CNN



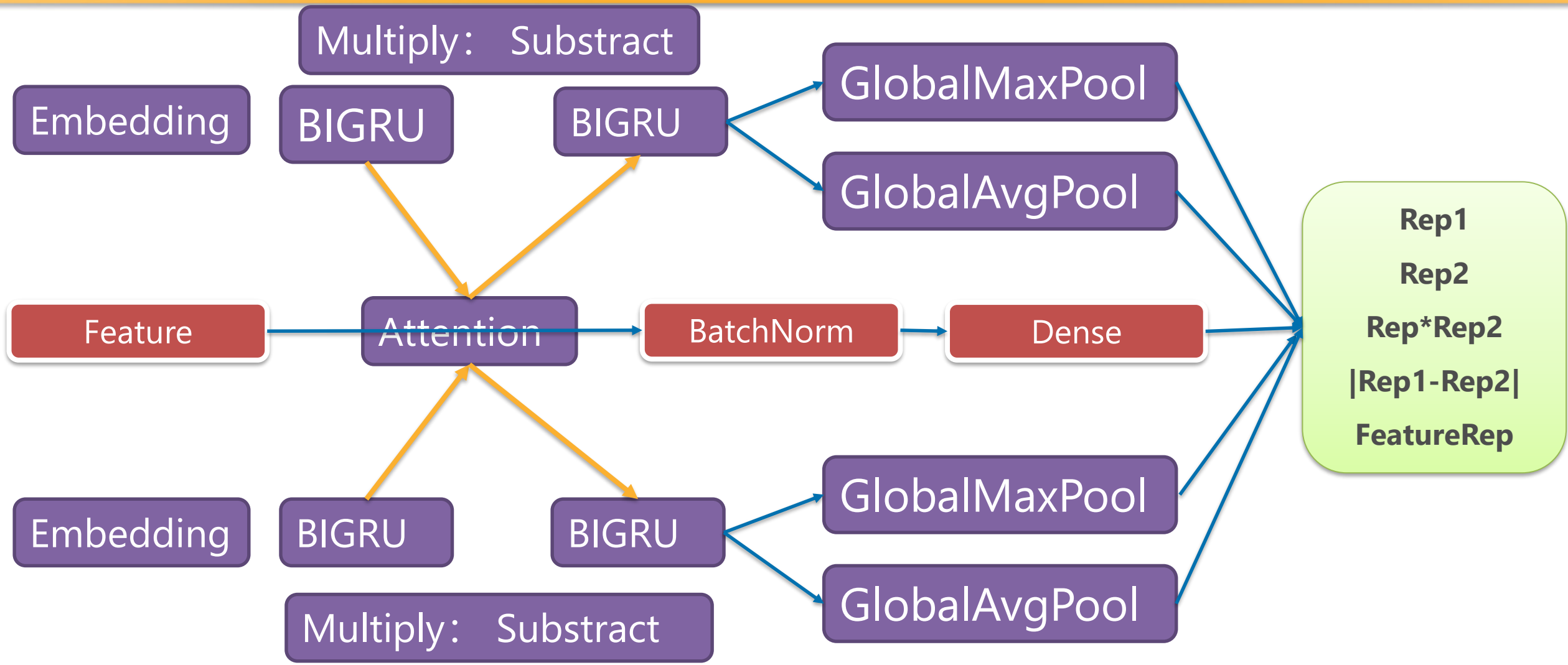
Stacked 3-Layers BiGRU



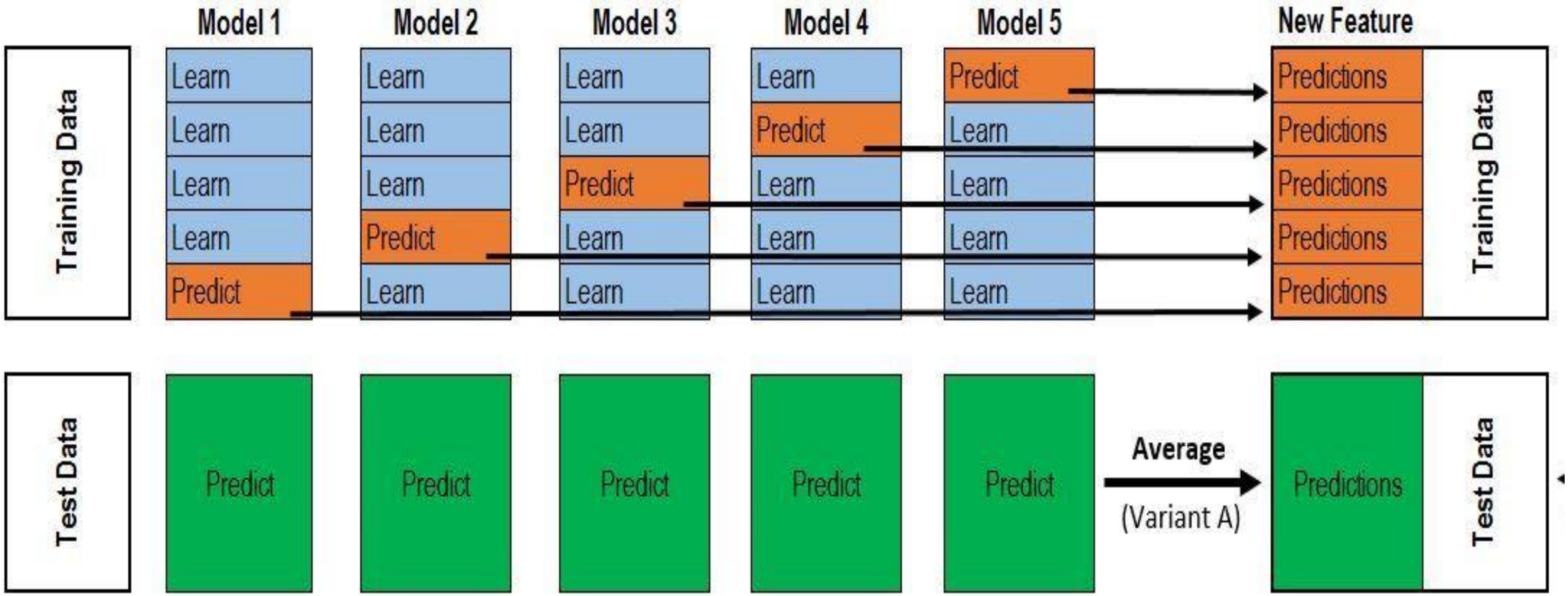
Decomposable Attention



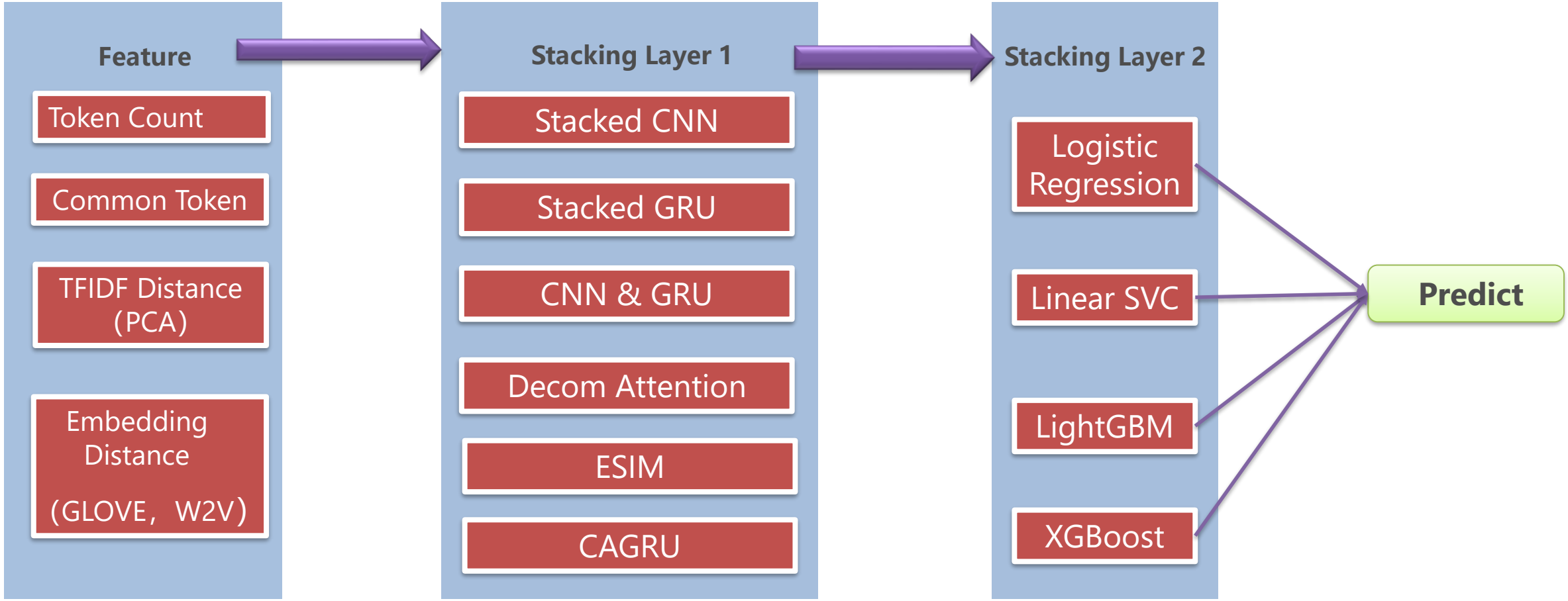
ESIM


Multiply： Substract
Embedding
BIGRU
BIGRU
GlobalMaxPool
GlobalAvgPool
Feature
Attention
BatchNorm
Dense
Rep1
Rep2
Rep*Rep2
|Rep1-Rep2|
FeatureRep
Embedding
BIGRU
BIGRU
GlobalMaxPool
GlobalAvgPool
Multiply： Substract

Stacking集成


Model 1
Model 2
Model 3
Model 4
Model 5
New Feature
Training Data
Learn
Predict
Predictions
Training Data
Test Data
Predict
Average
(Variant A)
Predictions
Test Data

总体框架


Feature
Token Count
Common Token
TFIDF Distance (PCA)
Embedding Distance (GLOVE，W2V)
Stacking Layer 1
Stacked CNN
Stacked GRU
CNN & GRU
Decom Attention
ESIM
CAGRU
Stacking Layer 2
Logistic Regression
Linear SVC
LightGBM
XGBoost
Predict

最终结果

方法	得分
ESIM features word+char	**0.8781**
ESIM(LSTM) features word+char	0.8776
ESIM features char	0.8773
ESIM word+char	0.8742
Decomposable Attention (*4)	0.8759
CARNN (*4)	0.8731
CNN_RNN (*4)	0.8699
Stacked CNN (*4)	0.8661
Stacked RNN (*4)	0.8730
ESIM (*4)	**0.8832**

#	队伍名	成绩
1	DUTIR	0.88799
2	deadline	0.88403
3	湖人总冠军@InplusLab	0.87941
4	TCM@ZJU	0.87909
5	DUTNLP.未来数据研究所	0.87717

总结展望

总结

- **参考深度文本匹配综述主要尝试了以下三类深度模型：**

单语义

- CNN，RNN等单层文本建模方法的效果不好
- 随着层数的增加与**shortcut**的引入效果明显提升。

多语义

- SNLI的模型效果较好。比如**Decomposable attention**，**ESIM**。
- 尝试过包括DIIN，BiMPM ，由于训练时间，效果等因素并未采用。

直接建模

- MatchPyramid，Match SRNN， Arc2等在词向量直接交互的匹配矩阵提取特征的模型效果不好。
- 分析可能是数据集的文本序列特征很重要而且文本序列较短，这些方法提取的匹配信息有限。
- 将乱序或随机采样的验证集输入模型效果下降很多，该种数据扩展方式，对原始数据也未采用截断。

总结

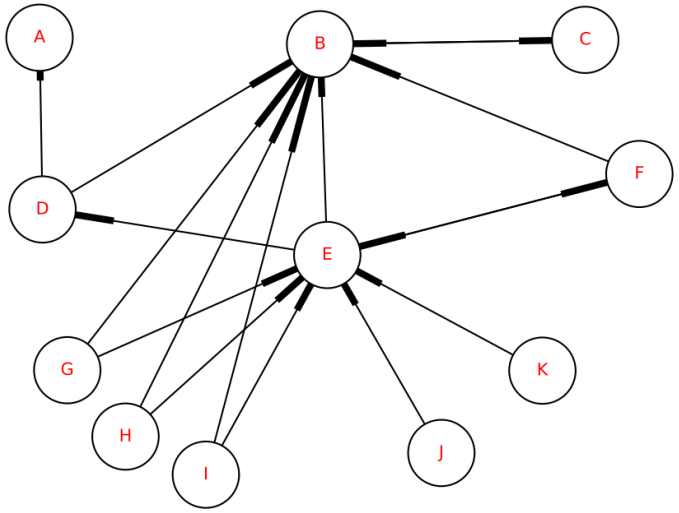
- **特征**
  - NLP特征和统计特征的加入对深度模型的提升是明显的。
  - 对特征进行归一化，highway net层提取表示。
  - LDA特征，NMF特征，拼接exact match。

- **单模型**
  - 最优的单模型的得分决定了最后结果。加入更多模型集成并无提升。
  - 尝试多种finetune方式，但过拟合严重，交叉验证有显著提升测试集下降。
  - 词向量层固定的效果更好，分析原因测试集有大量的token在训练集中并未出现。

总结

- **集成方式**
  - 尝试了多种voting，bagging等集成方法容易过拟合。
  - 第二层简单的soft voting 或 lr，lsvc进行集成效果就很好。

- **句子链路关系**
  - 图特征，句子id的连边数。 无效果，未采用。
  - 利用句子链路关系进行数据扩展。 无效果，未采用。



展望与致谢

- **展望：**
  - 神经网络模型的优化和参数调优。
  - 匹配特征的抽取以及选择。
  - 探索将外部知识和特征更好的融入到深度模型中。
  - 原始文本的语言特征是很重要的，医药等核心相关实体的语义匹配。词性句法特征的引入。

- **致谢：**
  - 感谢主办方。
  - 感谢本次比赛的运营人员。
  - 感谢指导的老师和队友。

参考文献

1. Shortcut-Stacked Sentence Encoders for Multi-Domain Inference - EMNLP 2017 RepEval Multi-NLI Shared Task

2. A Decomposable Attention Model for Natural Language Inference - EMNLP 2016

3. Enhanced LSTM for Natural Language Inference - ACL 2017

4. CIKM2018 rank2: https://github.com/zake7749/Closer

5. Kaggle Quora question pair rank4：https://github.com/HouJP/kaggle-quora-question-pairs

6. 庞亮, 兰艳艳, 徐君,等. 深度文本匹配综述[J]. 计算机学报, 2017, 40(4):985-1003.

谢谢！