中国人工智能学会
Chinese Association for Artificial Intelligence

全球人工智能AI 2021技术创新大赛

GLOBAL AI INNOVATION CONTEST

赛道三:小布助手对话短文本语义匹配

白[MASK]

CONTENTS

赛题描述-介绍

本赛题任务是：根据脱敏后的短文本query-pair，预测它们是否属于同一语义。本质上来看，是一个0/1二分类任务。

sentence1	sentence2	label
1 2 3 4 5 6 7	8 9 10 4 11	0
12 13 14 15	12 15 11 16	0
17 18 12 19 20 21 22 23 24	12 23 25 6 26 27 19	1
28 29 30 31 11	32 33 34 30 31	1
29 35 36 29	29 37 36 29	1
38 23 39 9 40	12 19 41 42 23 43 12 23 44 41 42 19	0

难点：

- **数据脱敏**
- **复赛工业级限时流式评测**

赛题描述-评估指标

初赛

$$AUC = \frac{\sum_{i\in\text{语义匹配样本集}} rank(i) - \frac{M(1+M)}{2}}{M * N}$$

$rank(i)$：表示i这个样本的预测得分在测试集中的排序;

M：测试集中语义匹配的样本的个数;

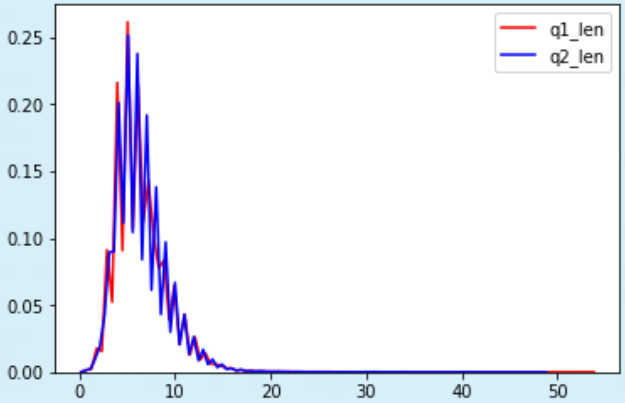
N：测试集中语义不匹配的样本的个数。

复赛

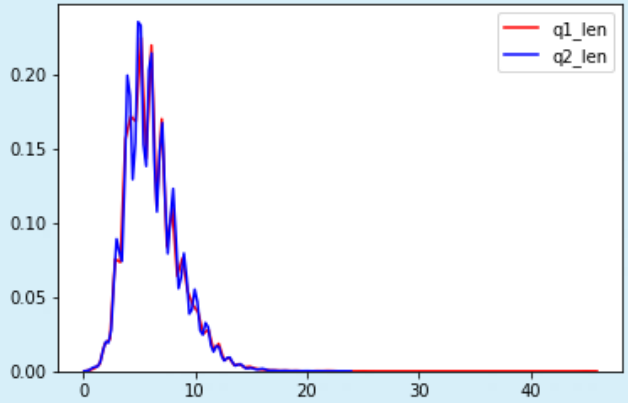
除初赛AUC指标外，加入了样本预测的时间指标：

- **avg_time**
- **max_time**
- **min_time**

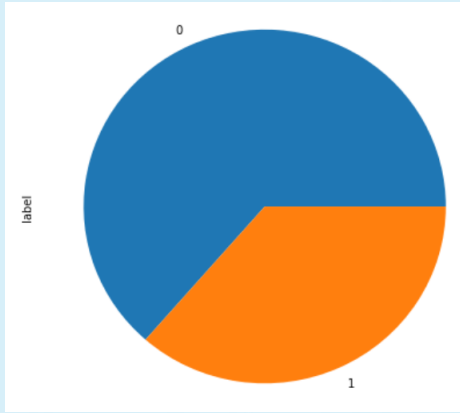
数据分析



训练集问题对长度分布

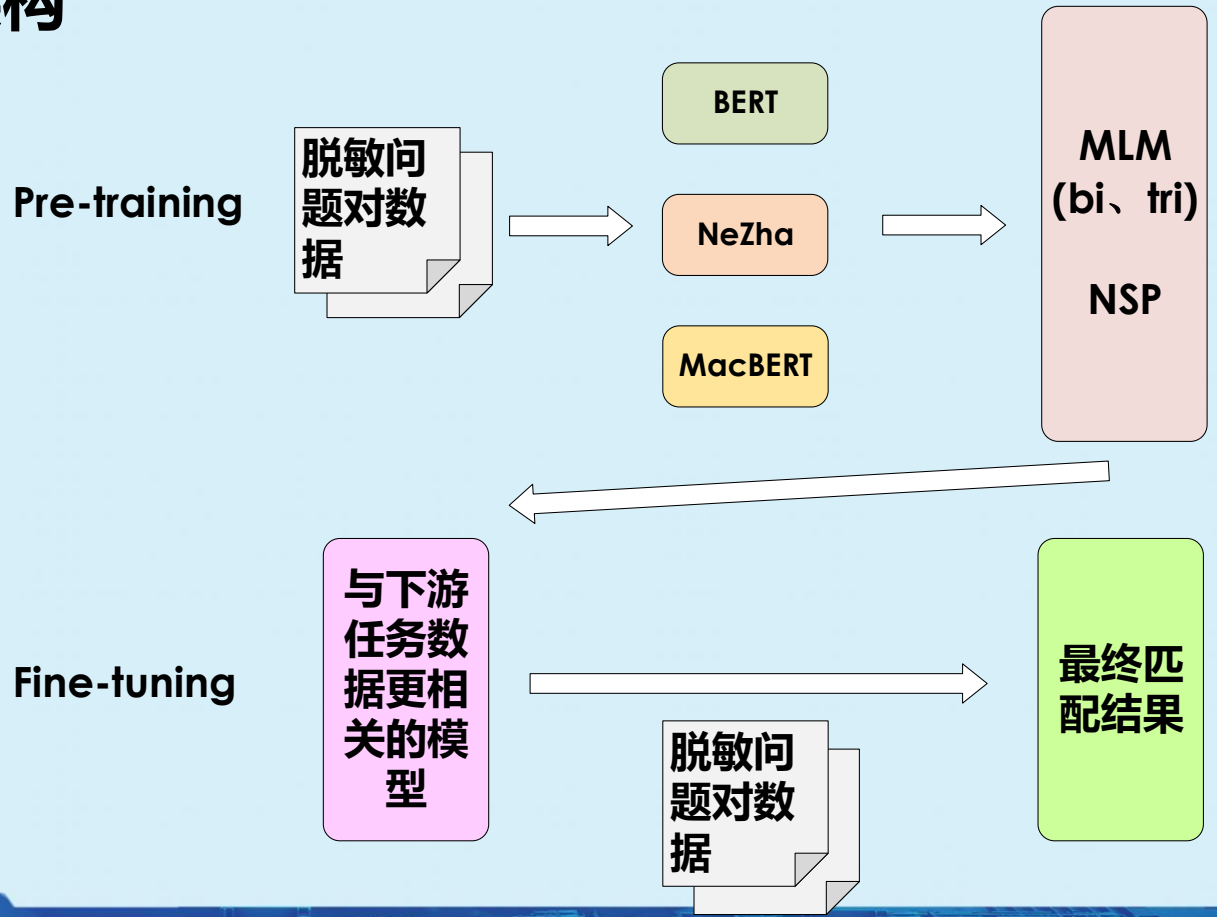


测试集问题对长度分布



训练集标签分布

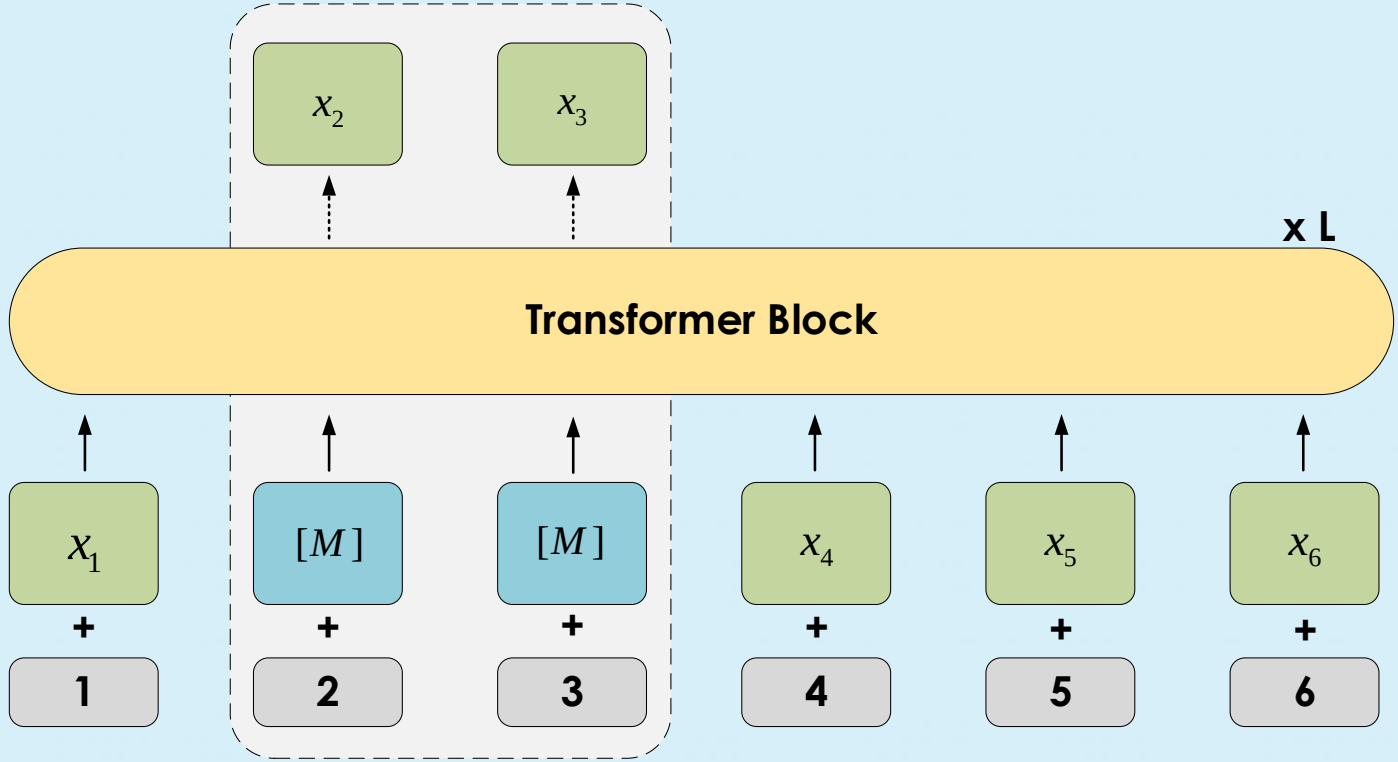
算法模型-架构


Pre-training
脱敏问题对数据
BERT
NeZha
MacBERT
MLM (bi、tri)
NSP
Fine-tuning
与下游任务数据更相关的模型
脱敏问题对数据
最终匹配结果

算法模型-初赛 Bi-gram

优势：

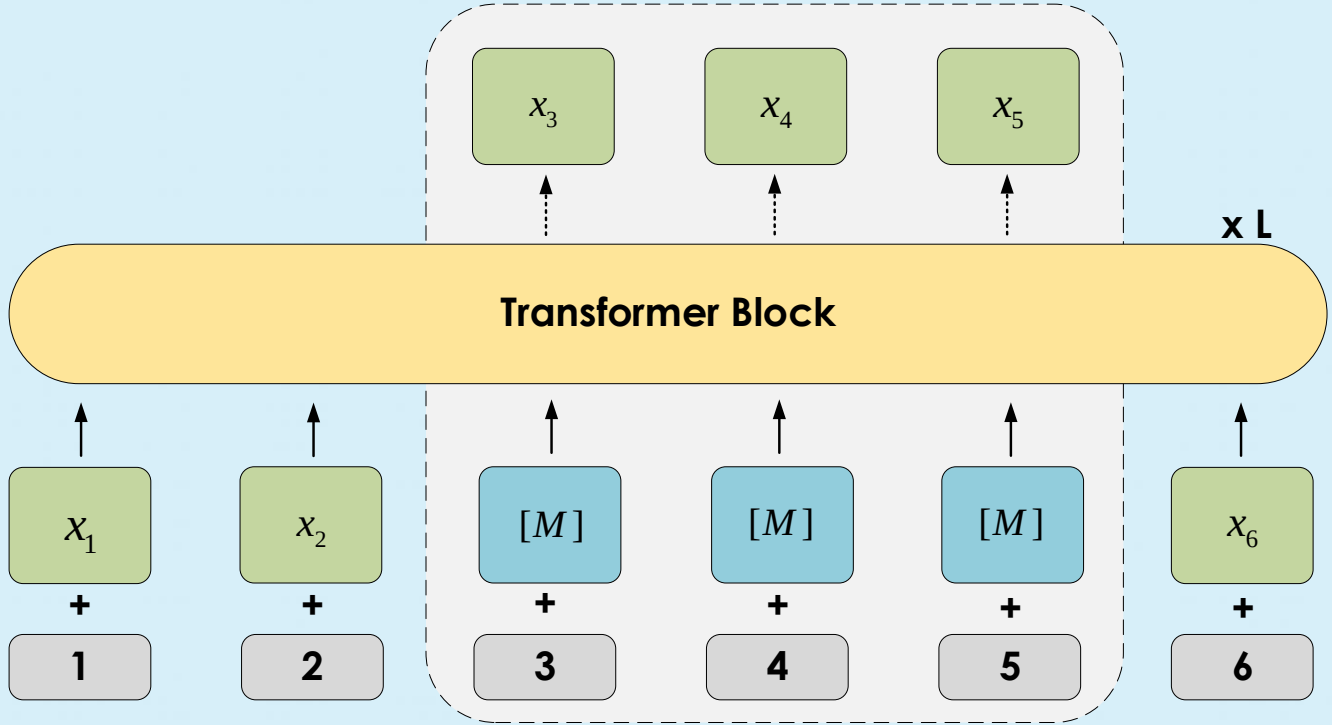
- 细粒度考虑词语之间的关系
- 挖掘脱敏数据词语特征表示
- 增加预训练任务难度



算法模型-初赛 Tri-gram

优势：

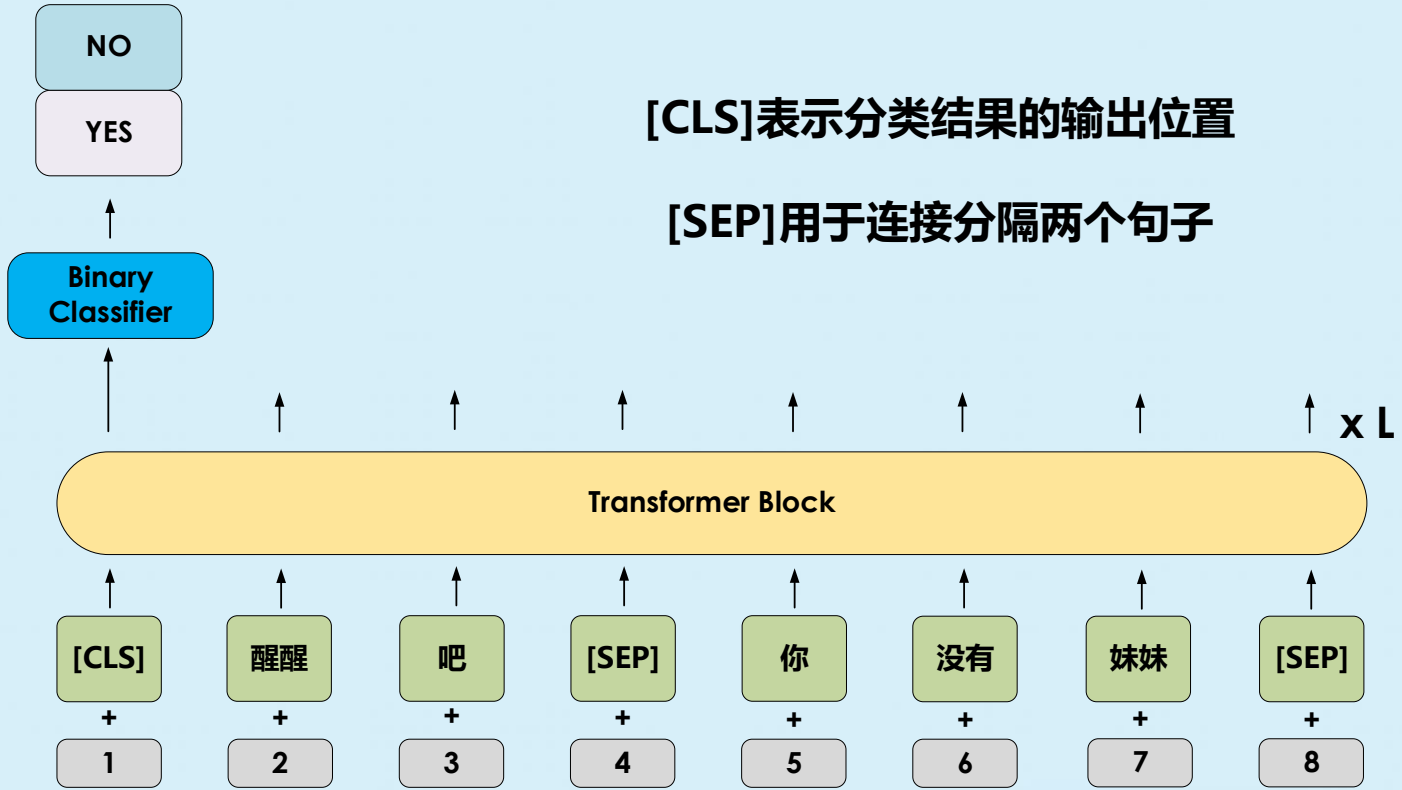
- **扩大词语的范围**
- **增加预训练任务的多样性**



算法模型-初赛 Next Sentence Prediction (NSP)

优势：

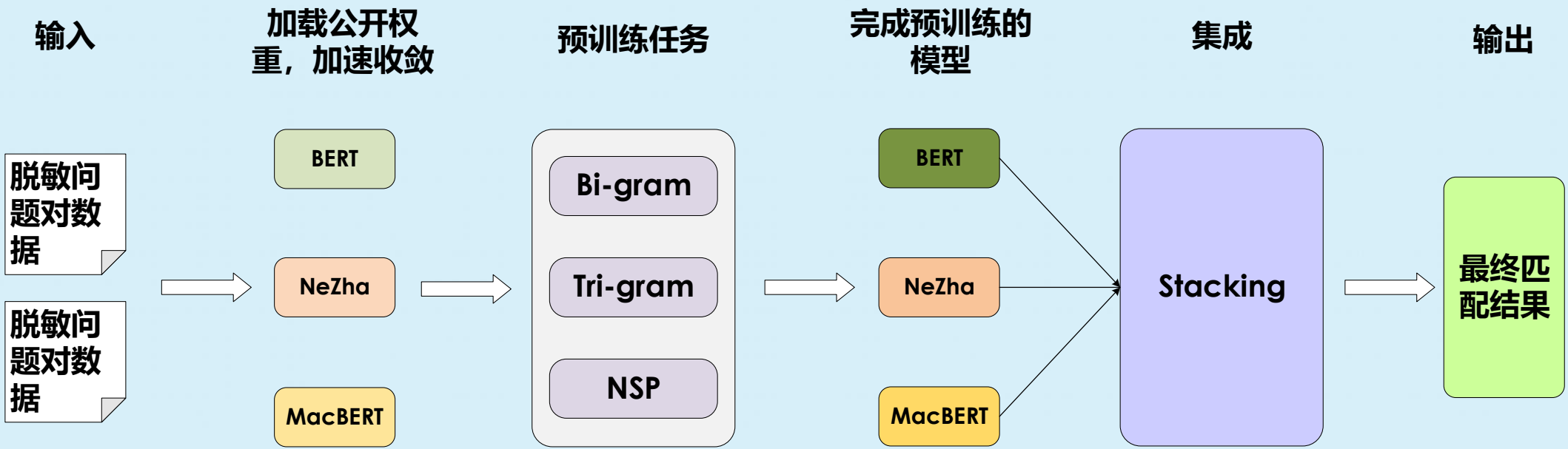
- **粗粒度句子级预训练**
- **判断两个句子是否是上下文关系**



[CLS]表示分类结果的输出位置

[SEP]用于连接分隔两个句子

算法模型-初赛 流程


输入
加载公开权重，加速收敛
预训练任务
完成预训练的模型
集成
输出
脱敏问题对数据
脱敏问题对数据
BERT
NeZha
MacBERT
Bi-gram
Tri-gram
NSP
BERT
NeZha
MacBERT
Stacking
最终匹配结果

算法模型-复赛 难点

难点：

- **Docker端到端提交**
- **工业级限时流式评测(50000/15min)**

算法模型-复赛 主流模型压缩方法

量化

剪枝

蒸馏

ONNX RUNTIME

TensorRT

加速框架

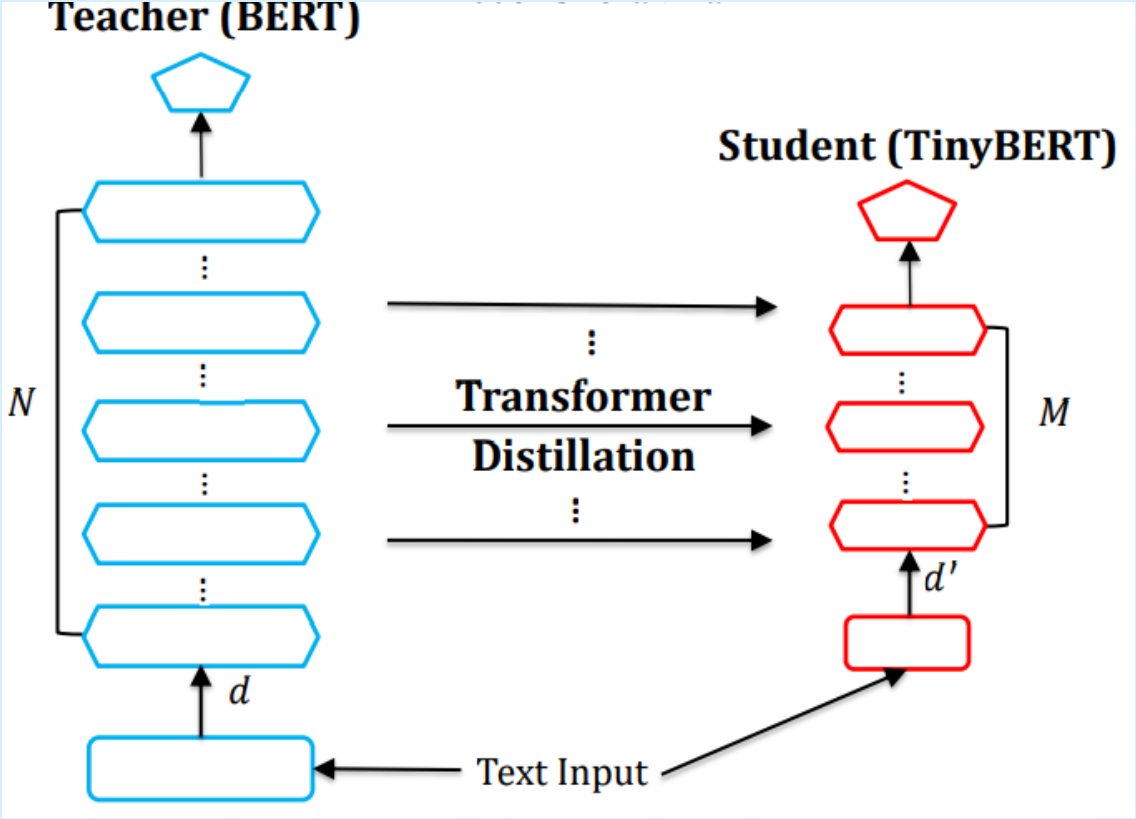
算法模型-复赛 模型蒸馏

模型大小减小约10倍

推理速度提升8倍

12层Transformer

420M



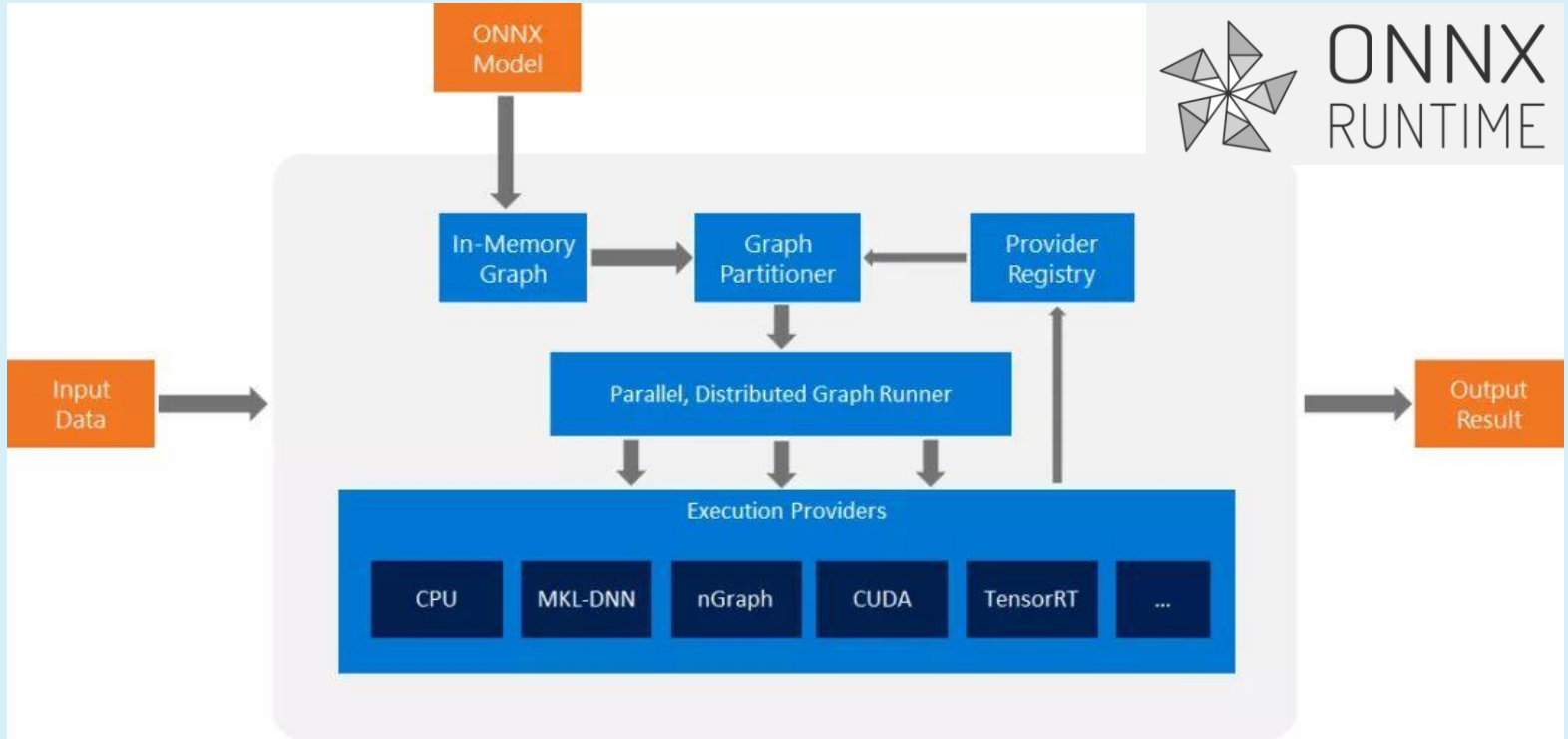
4层Transformer

46M

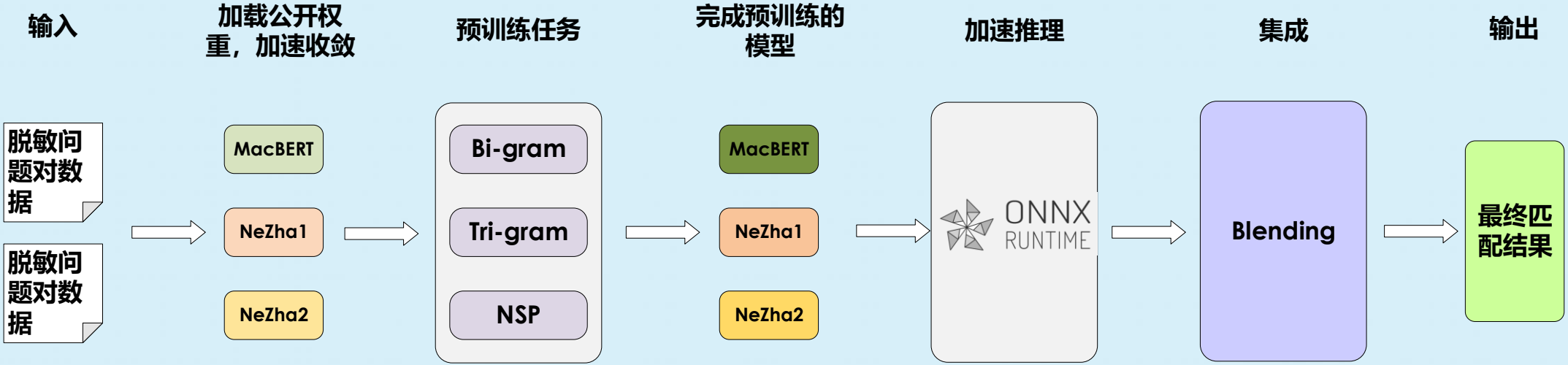
算法模型-复赛 ONNX Runtime加速

精度并未损失

推理速度提升3倍+



算法模型-复赛 流程


输入
加载公开权重，加速收敛
预训练任务
完成预训练的模型
加速推理
集成
输出
脱敏问题对数据
脱敏问题对数据
MacBERT
NeZha1
NeZha2
Bi-gram
Tri-gram
NSP
MacBERT
NeZha1
NeZha2
ONNX RUNTIME
Blending
最终匹配结果

算法模型-pretrain优化

参考方案：lonePatient/NeZha_Chinese_PyTorch、ALBERT

1.闭包加对偶数据增强，

q1 - q2 = 1 & q2 - q3 = 1 => q1 - q3 = 1

q1 - q2 = 1 => q2 - q1 = 1

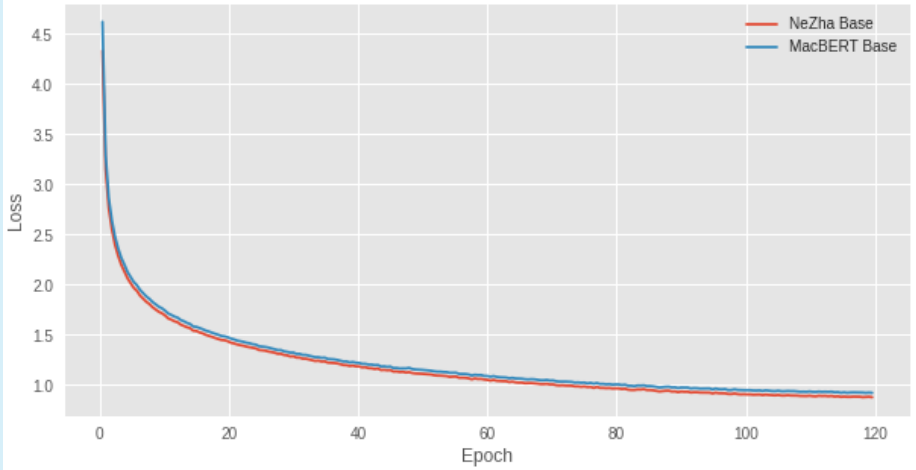
只造了正样本，数据量由40w扩增到99w，提升2-3k，如果正负样本都造效果会差2个千分点左右，大力出奇迹

2.对原数据按空格切开，按照词频min2和min3构建两份词表

算法模型-pretrain优化

3.知识继承：加载了 MacBERT/NeZha base（华为）的预训练权重，只对 embedding 层重新初始化，收敛速度更快，训练120epochs

4.采用Bi-gram、Tri-gram两种mask方法，在原本ALBERT ngram基础上优化，改写torch的API接口，避免mask阶段n-gram的for循环耗时，可提速约10%



算法模型-pretrain优化

5.NeZha开源的torch的selfAttention频繁在CPU和GPU上转换，优化后可提速约11%

6. NeZha每次加载模型权重时都是重新计算相对位置矩阵，实验里改动max_position_embeddings在pretrain阶段是有精度损失的，转换为onnx前把值改为32是没有影响的，nezha base模型缩小为1/3，显存占用明显减少

7.采用混合精度FP16的方式进行预训练，可提速约40%，显存节约21%。因此相同环境下，bs可以更大提速更多

算法模型-finetune优化

1.先用十折交叉验证的方式检验最优模型的epoch数，然后用全量数据进行finetune，不切分验证集，相比10折提升2K

2.加FGM干扰对抗学习，提高模型的泛化能力，提升3k

3.AdamW+ReduceLROnPlateau，防止模型过拟合以及陷入局部最优，提升1k

4.初赛魔改BERT/NeZha结构以及自蒸馏等方法都有效，复赛反而效果降了，没有一成不变最好的参数，只用更合适的参数

实验结果

方案潜力

- 赛道1和赛道3都进入决赛，模型通用性比较好
- pretrain阶段多处优化，速度快，可在更短时间内完成多次线下迭代
- 单模优势大，应该是复赛A榜最高分
- 线上推理速度快，15min推理5w条数据，可以全量跑9个BERT

部分单模及最终集成实验结果

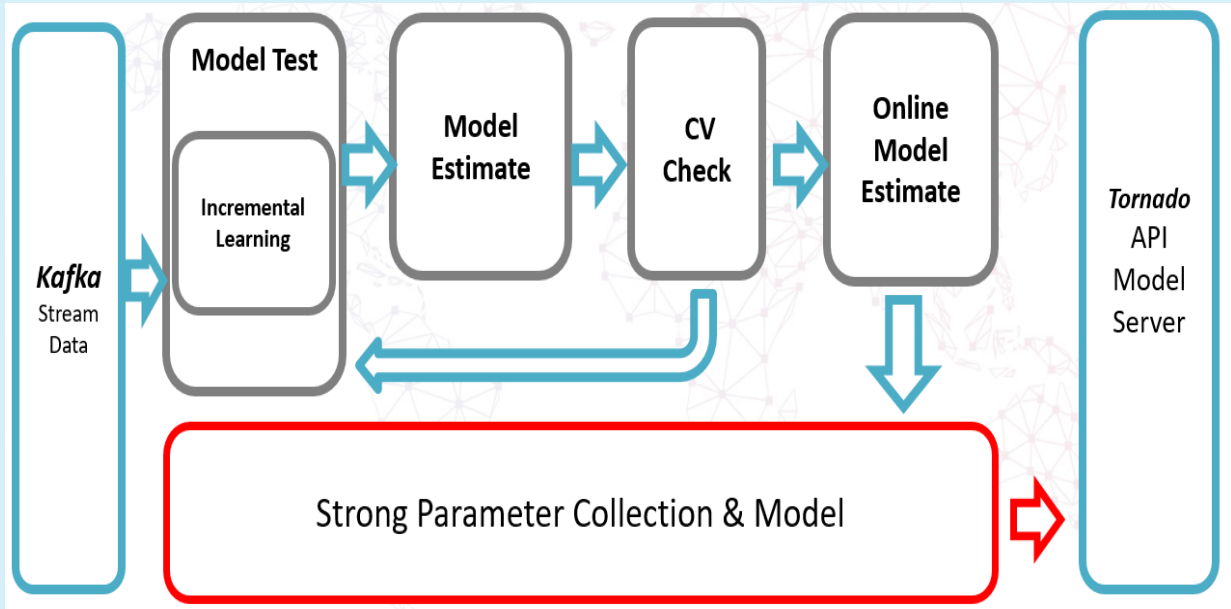
模型	复赛A榜AUC分值
BERT	0.9490
MacBERT	0.9501
NeZha (Tri)	0.9516
NeZha (Bi)	0.9512
blending(mean)	0.9568
复赛B榜	0.958

总结

- 提出细粒度的Bi-gram、Tri-gram MLM同时辅以粗粒度的NSP任务进行预训练
- 对pretrain部分多处进行优化，训练时间更短，迭代更快，仅用单模即能够得到较好的效果
- 利用不同预训练模型进行训练，充分汲取不同模型的长处，增加模型的多样性
- 使用多维度的模型蒸馏方法和ONNX Runtime架构，在不损失模型精度的情况下极大提升推理速度
- 很遗憾在复赛B榜端到端时，线上抖动，没有复现出最优成绩

应用价值

- 前端使用Kafka处理流式数据，流式数据输入模型，完成数据预处理
- 使用增量式训练，批数据输入模型，通过增量更新参数的办法，进行线上实验并反馈CV值，建立最优参数集合
- 使得模型能够稳定有效得适应数据迁移以及数据分布变化
- 接口层API调用模型推断相结合



模型部署架构图


中国人工智能学会
Chinese Association for Artificial Intelligence
THANKS!