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

科讯嘉联灵珠团队

CONTENTS

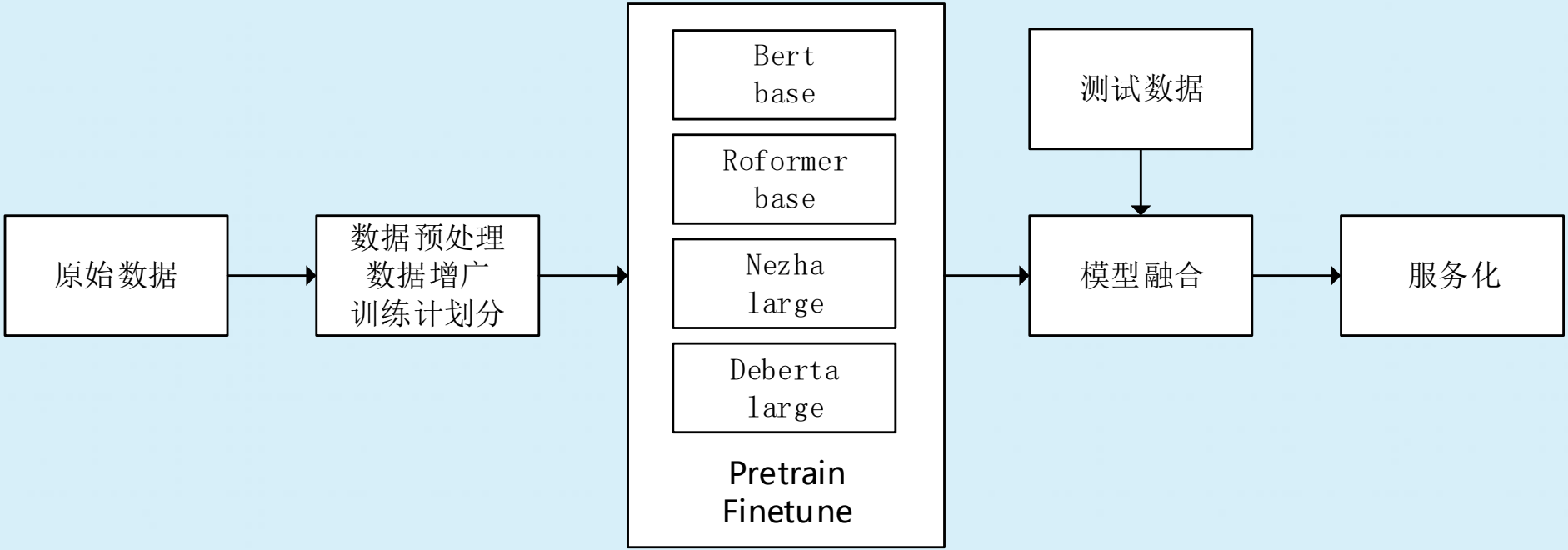
赛题介绍

小布助手是OPPO公司为欧加集团三品牌手机和IoT设备自研的语音助手，意图识别是对话系统中的一个核心任务，而对话短文本语义匹配是意图识别的主流算法方案之一。

比赛共发布40万条query-pair训练集，5万条测试集，数据已脱敏了，每个字都被影射为数字ID。

Sent-A	Sent-B	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
38 23 39 9 40	12 19 41 42 23 43 12 23 44 41 42 19	0

整体方案


原始数据
数据预处理
数据增广
训练计划分
Bert
base
Roformer
base
Nezha
large
Deberta
large
Pretrain
Finetune
测试数据
模型融合
服务化

创新点-绝对位置编码

Self-Attention常规范式

$$\begin{cases} q_i = (x_i + p_i)W_Q \\ k_j = (x_j + p_j)W_K \\ v_j = (x_j + p_j)W_V \\ a_{i,j} = softmax(q_i k_j^{\cdot}) \\ o_i = \sum_j a_{i,j} v_j \end{cases}$$

Bert绝对位置编码

$$\begin{cases} p_{k,2i} = \sin(k/10000^{2i/d}) \\ p_{k,2i+1} = \cos(k/10000^{2i/d}) \end{cases}$$

Nezha相对位置编码

$$a_{i,j} = soft\max\left(x_i W_Q \left(x_j W_K + R_{i,j}^K\right)^{\cdot}\right)$$

$$o_i = \sum_j a_{i,j}\left(x_j W_V + R_{i,j}^V\right)$$

$$R_{i,j}^K = p_K\left[clip(i-j, p_{\min}, p_{\max})\right]$$

$$R_{i,j}^V = p_V\left[clip(i-j, p_{\min}, p_{\max})\right]$$

p_K, p_V也是使用的三角函数式

Deberta
相对位置编码

$$q_i k_j^{晗} = x_i W_Q W_K\, x_j^{晗} + x_i W_Q W_K\, p_j^{晗} + p_i W_Q W_K\, x_j^{晗} + p_i W_Q W_K\, p_j^{?}$$

$$q_i k_j^{晗} = x_i W_Q W_K\, x_j^{晗} + x_i W_Q W_K\, R_{i,j}^{\cdot} + R_{j,i} W_Q W_K^{晗} x_j$$

创新点-Mask策略

Ngram mask

原始论文中按照以下分布随机生成n-gram，默认max_n为3

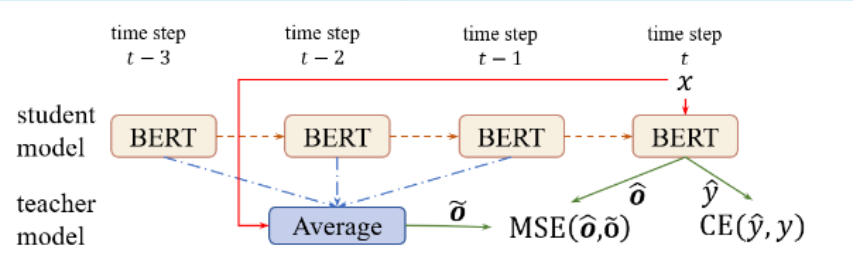
$$p(n)=\frac{1/n}{\sum_{k-1}^{N}1/k}$$

Span mask

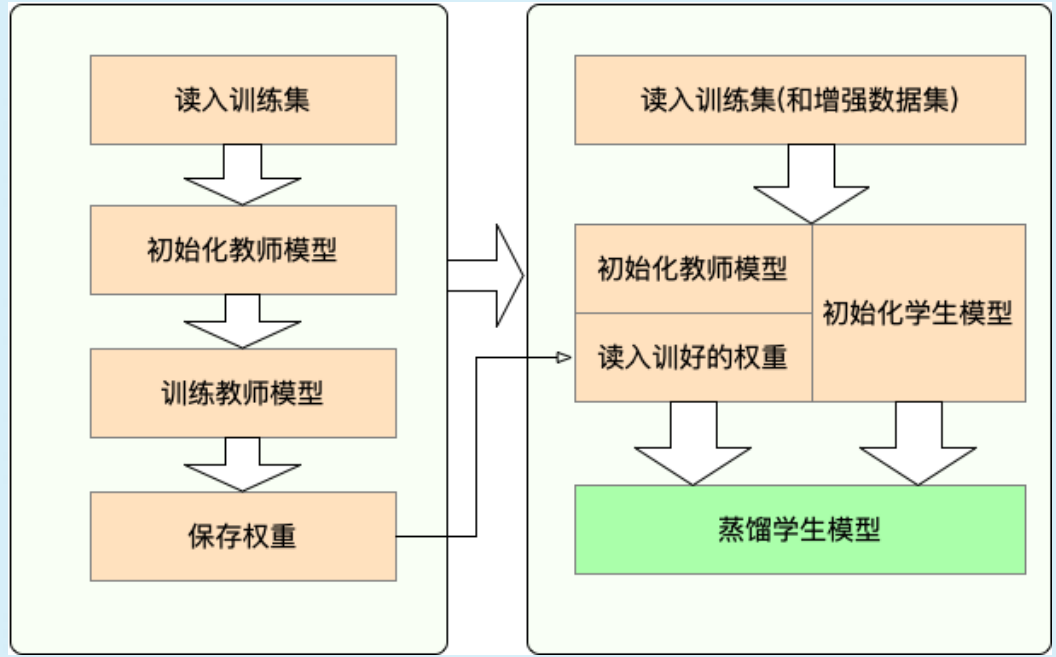
SBO任务是Span mask的训练目标，希望被遮盖 Span 边界的词向量，能学习到 Span 的内容。在训练时取 Span 前后边界的两个词，这两个词不在 Span 内，然后用这两个词向量加上 Span 中被遮盖掉词的位置向量，来预测原词。

创新点-蒸馏与自蒸馏

Large模型通过自蒸馏生成



Base模型通过Large模型蒸馏生成



创新点-对抗学习

对抗训练是一种引入噪声的训练方式，可以对参数进行正则化，提升模型鲁棒性和泛化能力。

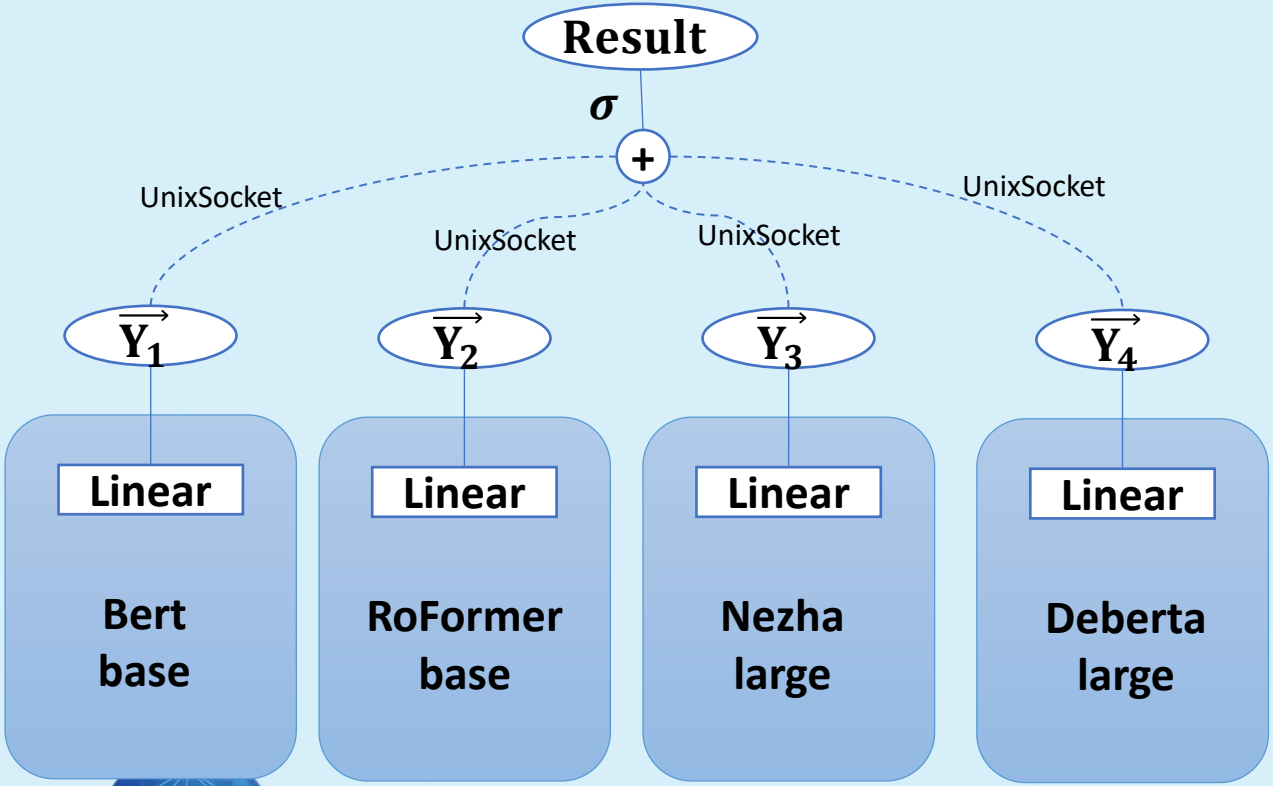
目标：$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\Delta x\in\Omega} L(x+\Delta x, y;\theta)\right]$$

- 内部max是为了找到最有效的扰动，使模型出错（攻击）
- 外部min是为了基于该攻击方式，找到最鲁棒的模型参数（防御）

Fast Gradient Method：$$\Delta x = \epsilon \frac{\nabla_x L(x,y;\theta)}{\|\nabla_x L(x,y;\theta)\|}$$

(对Embedding参数矩阵进行扰动)

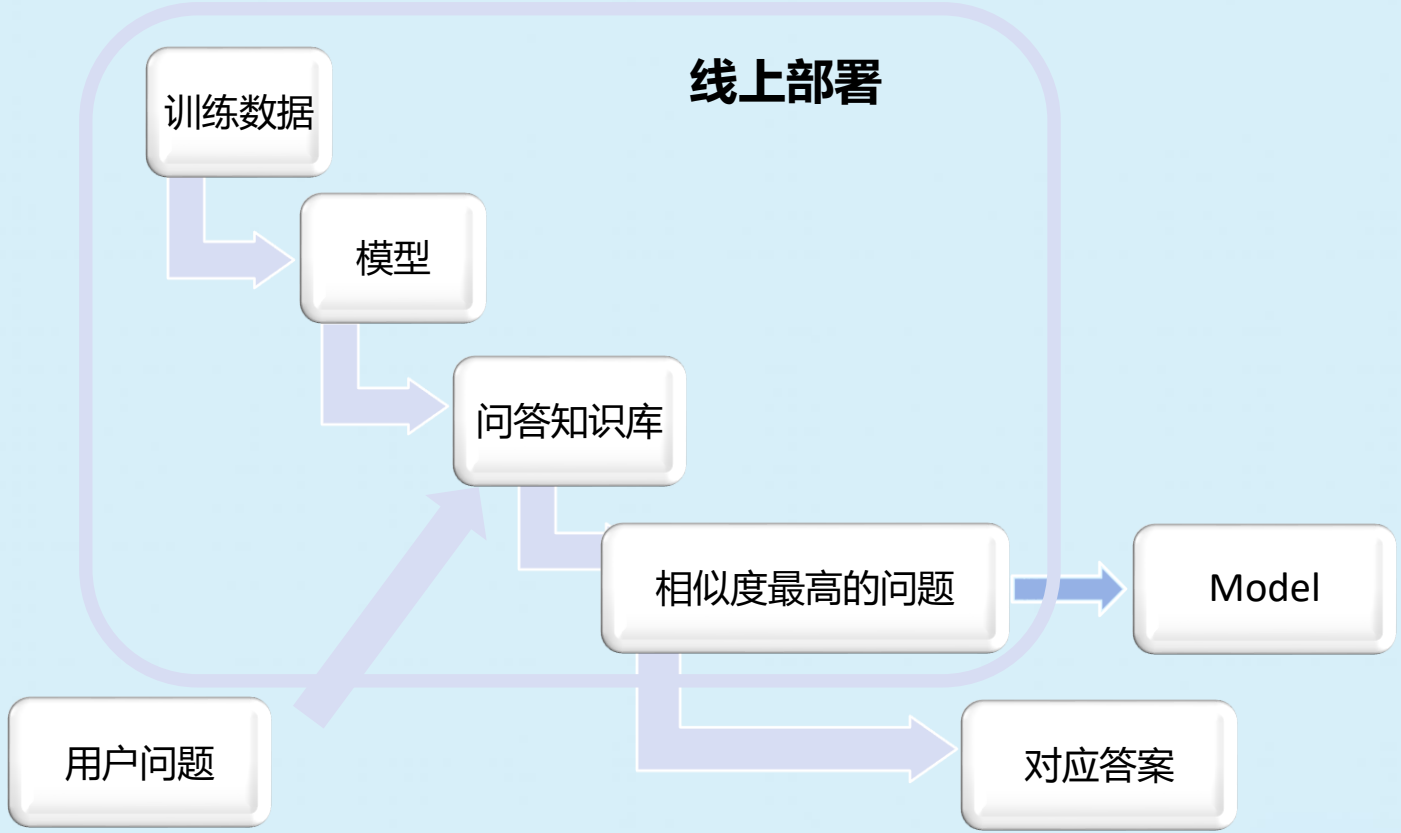
创新点-融合方式与服务化



先加权融合Linear层的输出，再计算sigmoid，既提高融合效果，又加快速度

Unixsocket，不需要经过协议栈，全双工。在推理阶段，使用onnx-runtime加速，将模型部署成服务，使用UnixSocket与主进程通信，并行推理。

落地方案



总结

- 此次比赛任务较为简单，数据干净，使用普通的BERT预训练模型便可达到94+的准确度量级，完全足以能够投入日常应用
- 调参和模型融合是最有效的提升手段，预训练是最值得投入的部分
- 准确度和时间复杂度不可兼得，单模初步满足现实中毫秒级的查询需求，但整体工程还需进一步研究

Reference

[1] NeZha: Neural Contextualized Representation for Chinese Language Understanding.

[2] DeBERTa: Decoding-enhanced BERT with Disentangled Attention.

[3] RoFormer: Enhanced Transformer with Rotary Position Embedding.

[4] ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations .

[5] SpanBERT: Improving Pre-training by Representing and Predicting Spans..

[6] FGM: Explaining and Harnessing Adversarial Examples.

*[7]*苏剑林. (Feb. 03, 2021). 《让研究人员绞尽脑汁的Transformer位置编码》.

[8] BERT-SDA: Improving BERT Fine-Tuning via Self-Ensemble and Self-Distillation.

[9] Text-Brewer: TextBrewer: An Open-Source Knowledge Distillation Toolkit for Natural Language Processing.

[10] ONNX Runtime: cross-platform, high performance ML inferencing and training accelerator


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