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

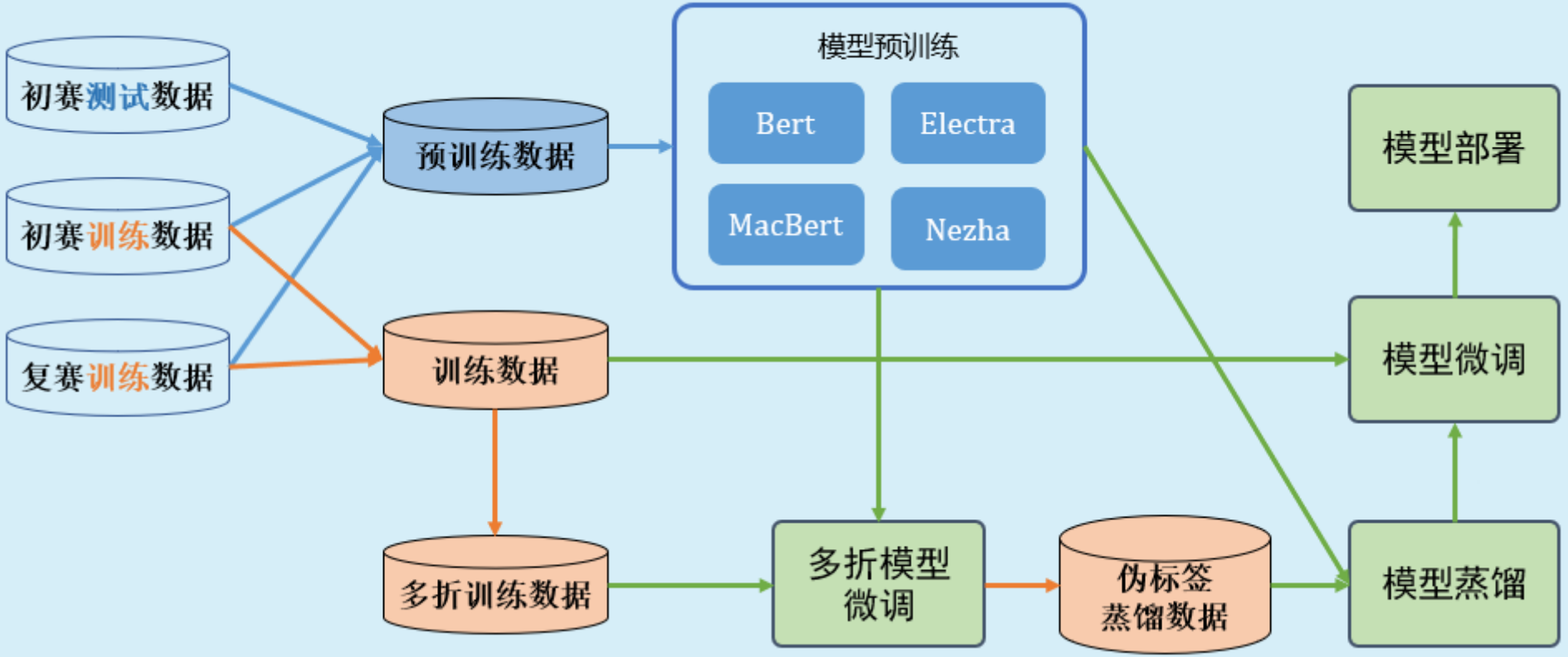
全球人工智能AI 2021技术创新大赛

GLOBAL AI INNOVATION CONTEST

赛道三：小布助手对话短文本语义匹配

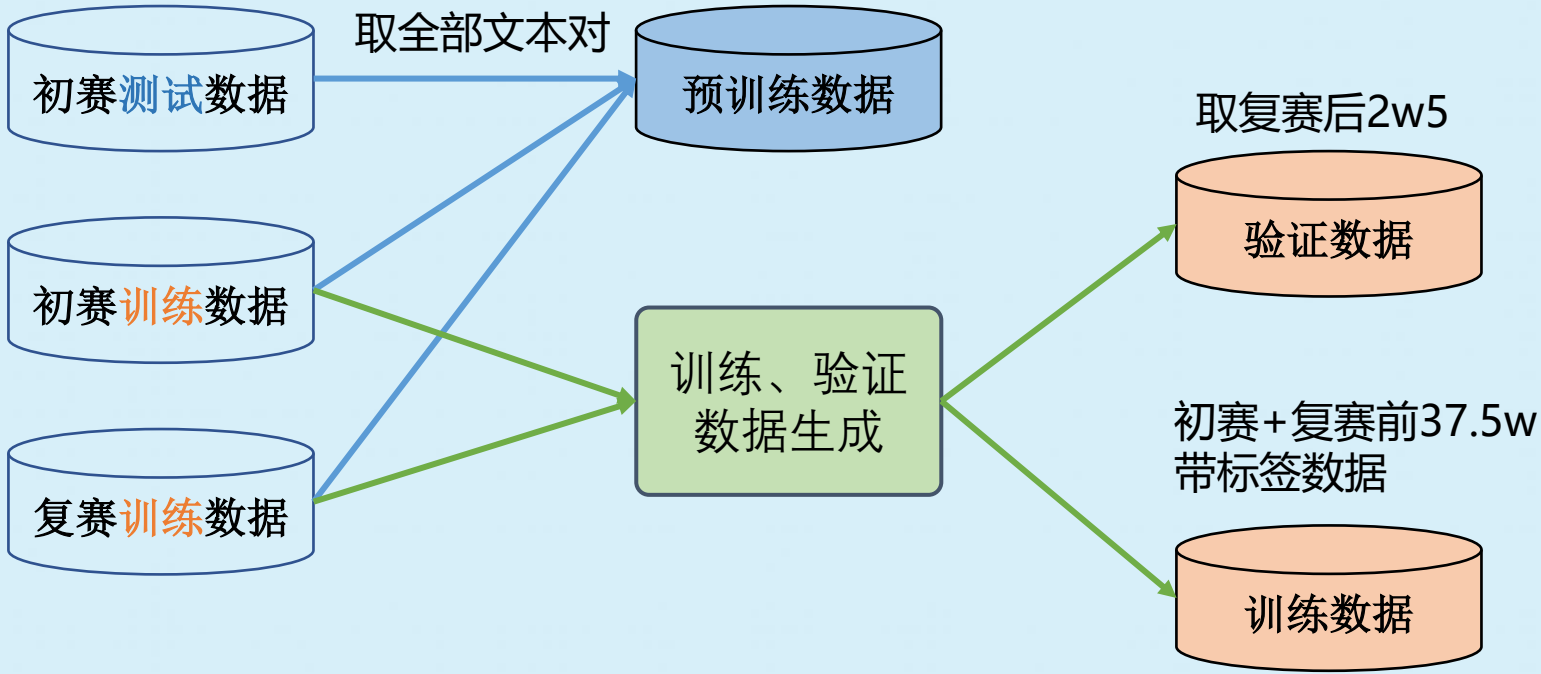
LOL王者

整体方案设计——总体流程


初赛测试数据
初赛训练数据
复赛训练数据
预训练数据
模型预训练
Bert
Electra
MacBert
Nezha
模型部署
模型微调
训练数据
多折训练数据
多折模型微调
伪标签蒸馏数据
模型蒸馏

整体方案设计——数据处理

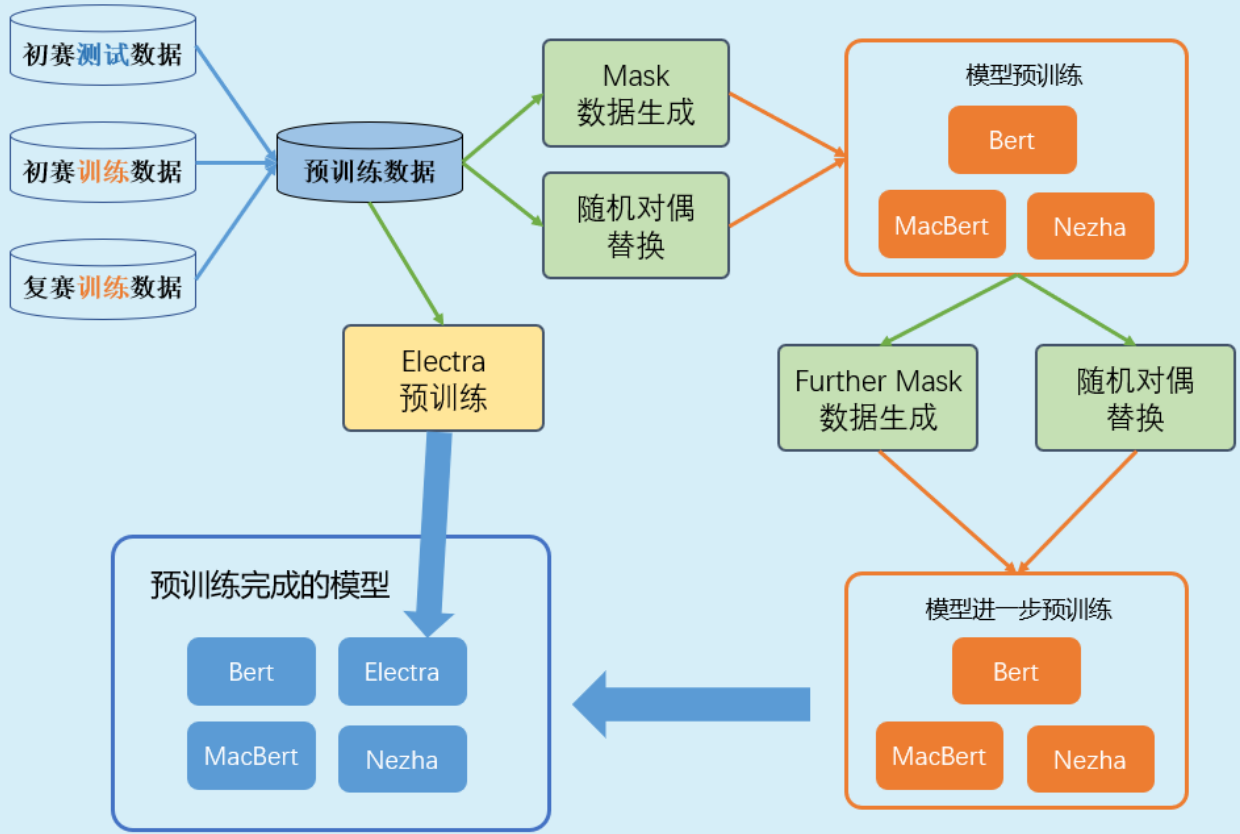
预训练数据、微调训练数据、微调验证数据



整体方案设计——模型预训练

1、预训练数据：初赛+复赛全部文本对

2、Further_Pretrain：以预训练数据进行进一步预训练，提高模型对句子对的匹配效果



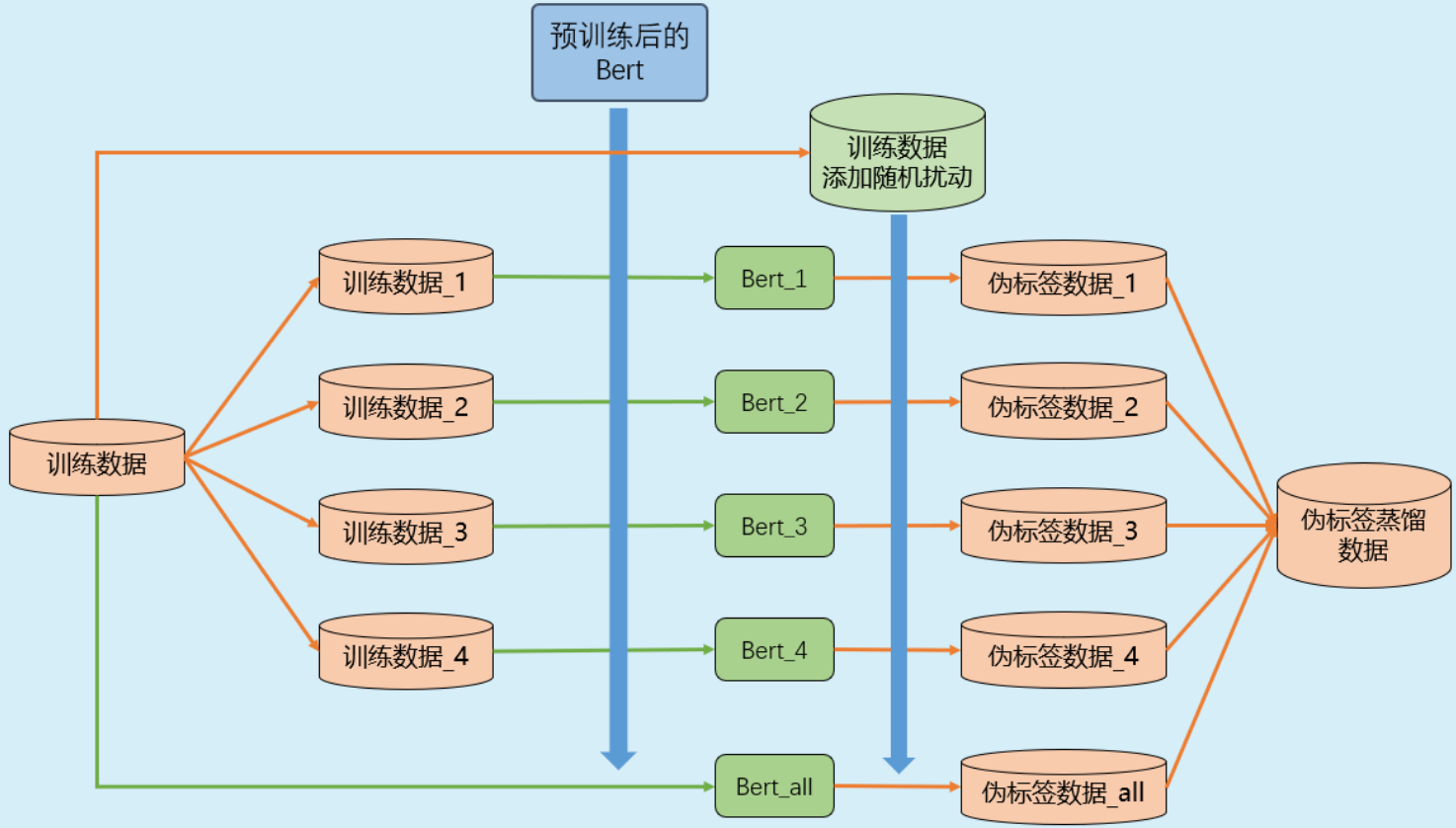
整体方案设计——不同模型预训练策略

通过优化预训练模型对输入语料的 Mask 策略，提高预训练模型效果，从而提升模型泛化能力

Mask 策略	实现思路	模型
原始文本	使用语言模型来预测下一个词的probability	
分词文本	使用 语言 模型 来 预测 下 一个 词 的 probability	
单词 Mask	使 用 语 言 [MASK] 型 来 [MASK] 测 下 一 个 词 的 pro [MASK] ##lity	Bert
全词 Mask	使 用 语 言 [MASK] [MASK] 来 [MASK] [MASK] 下 一 个 词 的 [MASK] [MASK] [MASK]	Bert-WWM
N-Gram Mask	使 用 [MASK] [MASK] [MASK] [MASK] 来 [MASK] [MASK] 下 一 个 词 的概率	Nezha
近义替换	使 用 语 法 建 模 来 预 见 下 一 个 词 的 几 率	Mac-Bert

整体方案设计——伪标签数据生成

多折微调、训练数据扰动、
伪标签数据合并

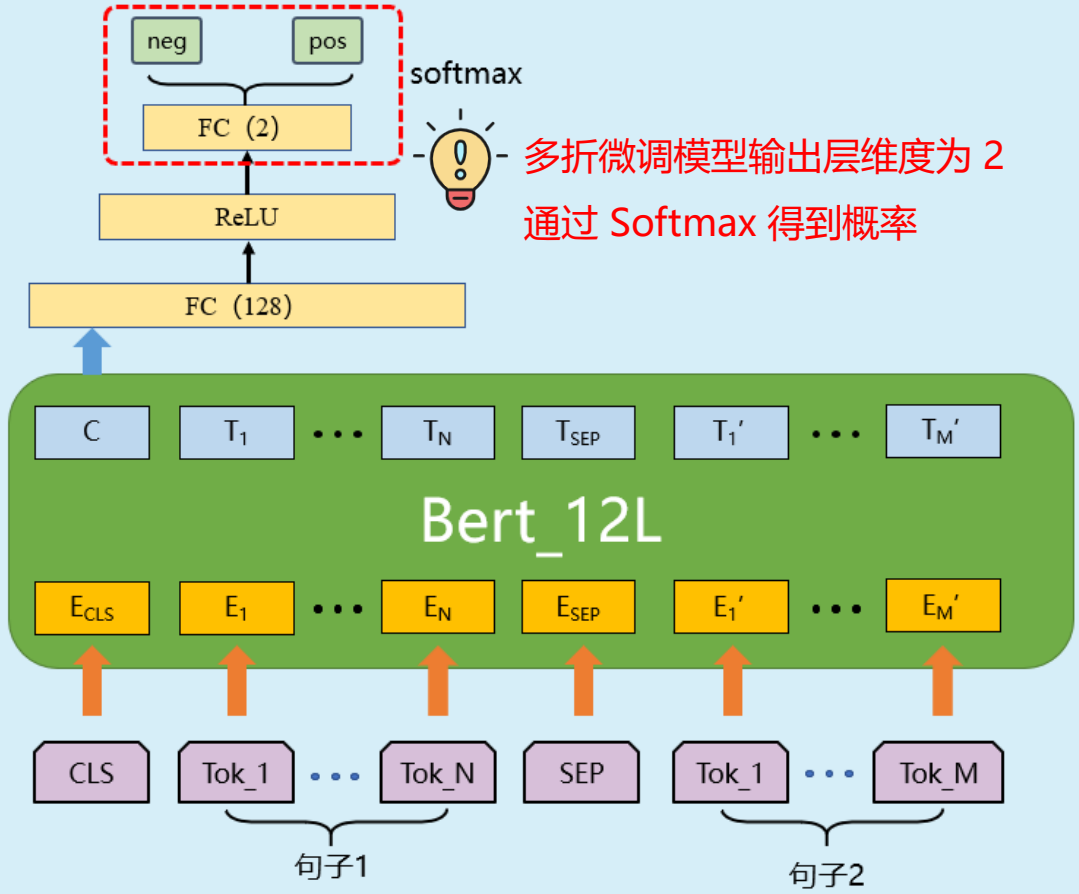


整体方案设计——多折微调模型

1、多折训练数据：将原始训练数据进行4折交叉验证得到 4 份训练数据

2、微调模型：拼接两个句子后输入 Bert 中取 CLS 后以输出层维度为 2 进行输出

3、微调模型训练：以交叉验证训练数据对已预训练的 Bert 模型进行微调，得到 4 个微调后的 Bert 模型



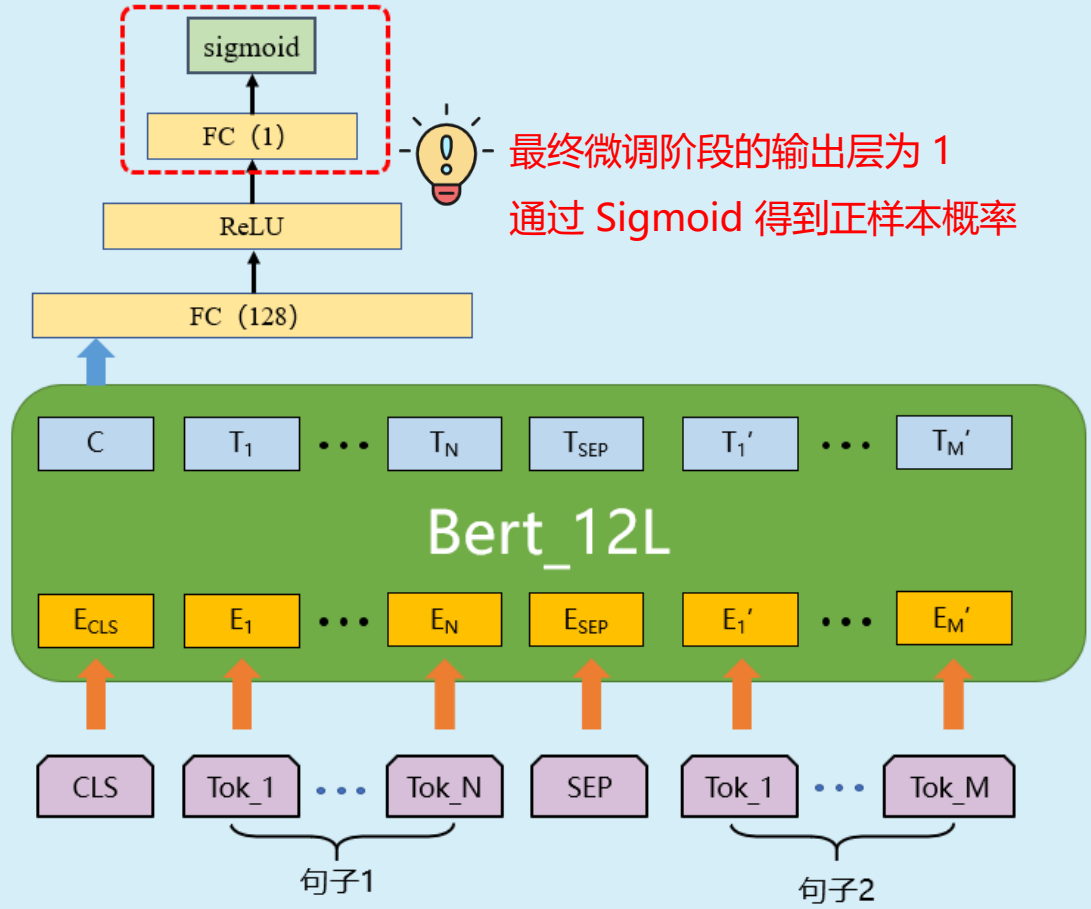
整体方案设计——模型自蒸馏

1、伪标签训练数据生成：将原始训练数据施加随机扰动得到新的训练数据

2、伪标签自蒸馏数据生成：使用多折微调后的模型对伪标签训练数据进行预测，带权平均后得到伪标签自蒸馏数据

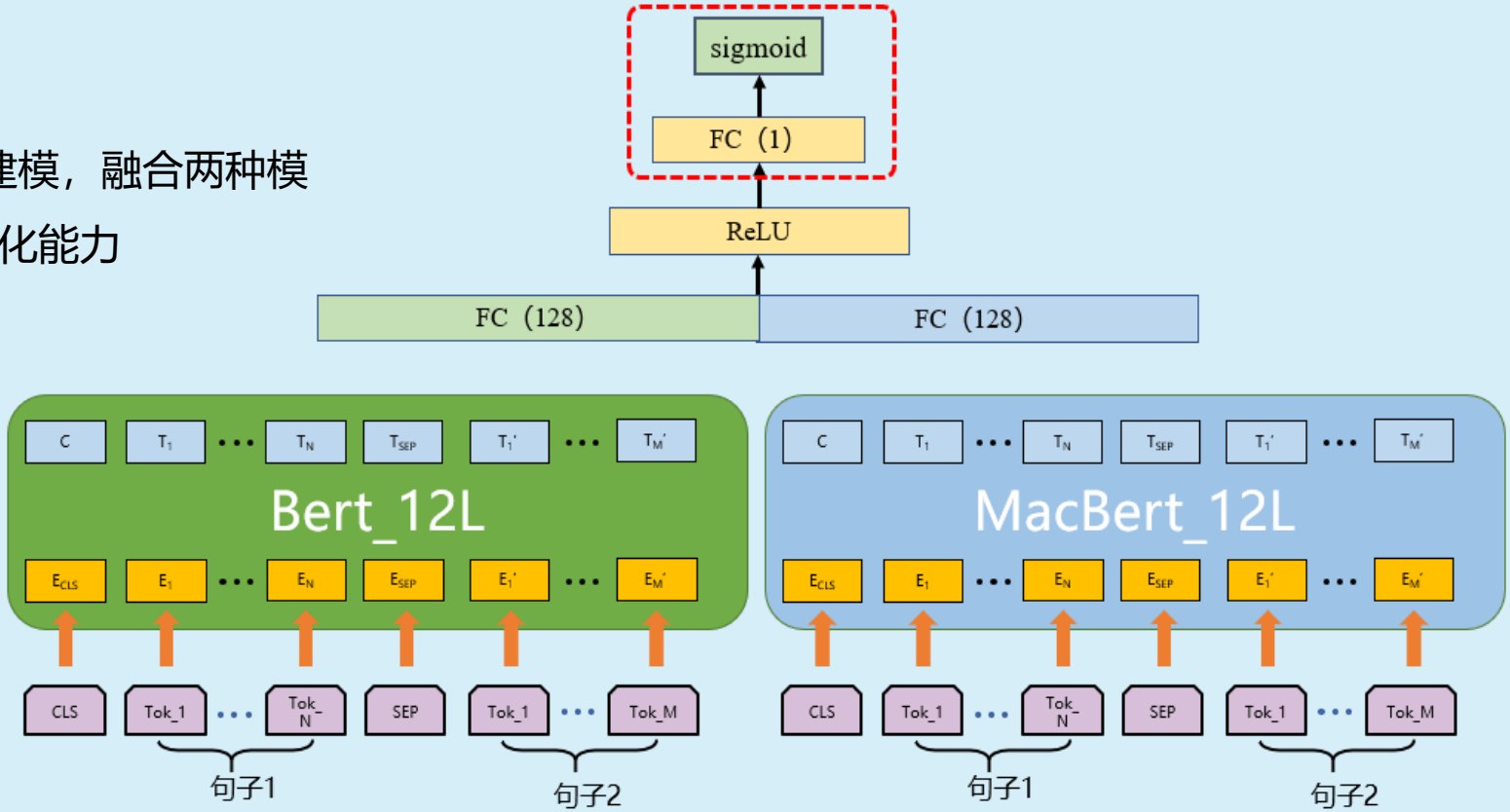
3、伪标签自蒸馏：以预训练 Bert 模型进行微调，学习伪标签自蒸馏数据

4、模型微调：以原始训练数据对模型进行最终的微调



整体方案设计——联合建模

将 Bert 和 MacBert 联合建模，融合两种模型的语义信息，提高模型泛化能力



整体方案设计——泛化能力提升

一、对抗训练（FGM）

通过在训练阶段引入噪声，实现对模型参数的正则化，提升模型的鲁棒性和泛化能力

$$-logp(y|x+r_{adv};\theta)$$

$$r_{adv}=\epsilon g/||g||_2$$

$$r_{adv}=\underset{r,||r||\leq\epsilon}{argmin}\,logp(y|x+r;\hat{\theta})$$

往梯度增大的方向扰动

二、滑动指数平均（EMA）

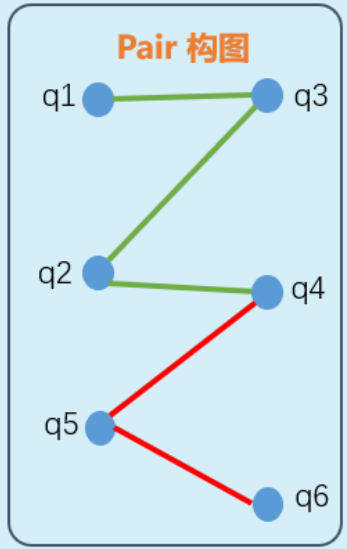
利用滑动平均的参数来提高模型在测试数据上的健壮性

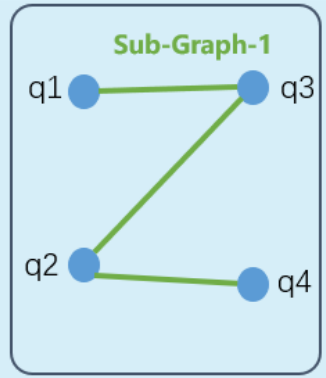
$$shadowVariable=decay*shadowVariable+(1-decay)*Variable$$

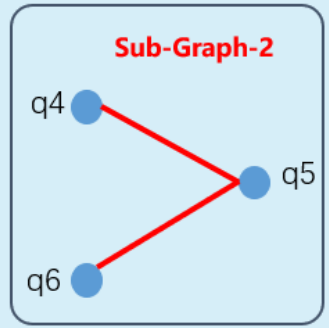
整体方案设计——图数据缓存

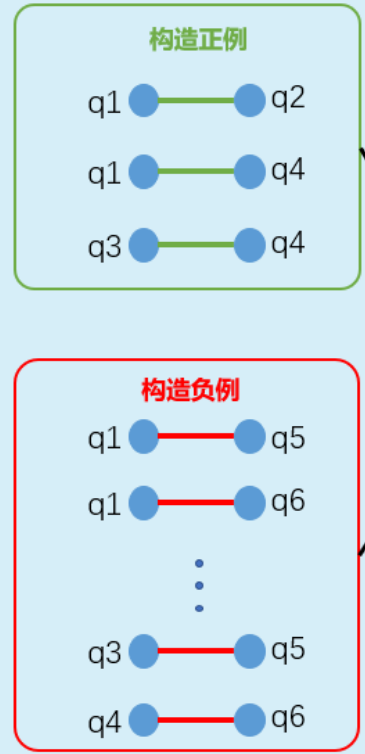
短文本对构图后将连通传递关系缓存

速度+精度









整体方案设计——推理优化

一、贝叶斯优化

通过贝叶斯优化，以单个模型预测效果为训练数据实现对推理阶段多个模型融合权重的学习

二、推理加速

采用 ONNX 实现 PyTorch 模型的推理加速

三、Nezha 源码的修改

根据 Nezha 模型的特点，并结合当前数据的统计情况，修改源码中的位置编码，使之具有和 Bert 相当的预测速度

四、输出层维度

将微调模型的输出层维度设置为 1，即全程使用 ONNX 即可得到预测结果，无需手动 Softmax

创新与落地——创新

1、Further_Pretrain：添加句子级的预训练任务，提高模型在文本匹配任务中的泛化能力

2、伪标签自蒸馏：通过对多折训练实现单个模型萃取多折模型，通过对训练数据进行随机扰动得到伪标签以提高模型泛化能力

3、多类模型融合：通过采用不同预训练机制获得不同的预训练模型，从而提高模型在新数据上的泛化能力

4、联合建模：将 Bert 和 MacBert 联合建模，融合两种模型的语义信息

5、图数据缓存：通过图中边的标签传递关系，极大提高预测精度和在线处理速度

6、贝叶斯优化：克服网格搜索，实现模型融合自动寻参

7、推理速度优化：修改 Nezha 源码、输出层维度调整

创新与落地——落地

1、预训练轮次：由于本次比赛对端到端时间有限制，而在实际工业界可能会有更多的计算资源（或可训练更长时间），则可在预训练阶段训练更长的轮次

2、伪标签自蒸馏：在线下计算资源充足的情况下（或可训练更长时间）， 可使用 Large 模型进行预训练和伪标签微调，最后再使用小模型（base 或者 6层、3层 Bert）对其进行萃取

3、多类模型融合：线上推理多模型可采用多线程多卡并行，一张卡对一个模型进行推理，设置超时机制，单个模型推理时间超过阈值则不再等待

4、推理速度：使用 C++ 版的 TensorRT 结合 ONNX 实现高效在线推理

5、图数据缓存：近 n 天高频 query 以及对应的 answer 可存储至 KV 存储（例如 Redis），在线阶段先查 KV 后再通过模型进行预测

方案总结

一、方案优势点

符合机器学习在工业产品中的做法，具有可扩展性强、泛化能力强、预测效率高、精度高等优点

二、方案劣势点

单个模型精度无法达到最高，需要进行多模型融合

三、科学研究展望

方案所提出的伪标签随机扰动、多折模型结合伪标签自蒸馏、联合建模等思想具有一定程度的创新，加以优化后或许可发表至相关的学术期刊

四、工业应用展望

短文本匹配广泛应用于搜索、搜索广告、对话等领域，所提方案中不乏富含工业界思想的实现，具有较强的工业应用前景，并能够在多个领域内大规模应用


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