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

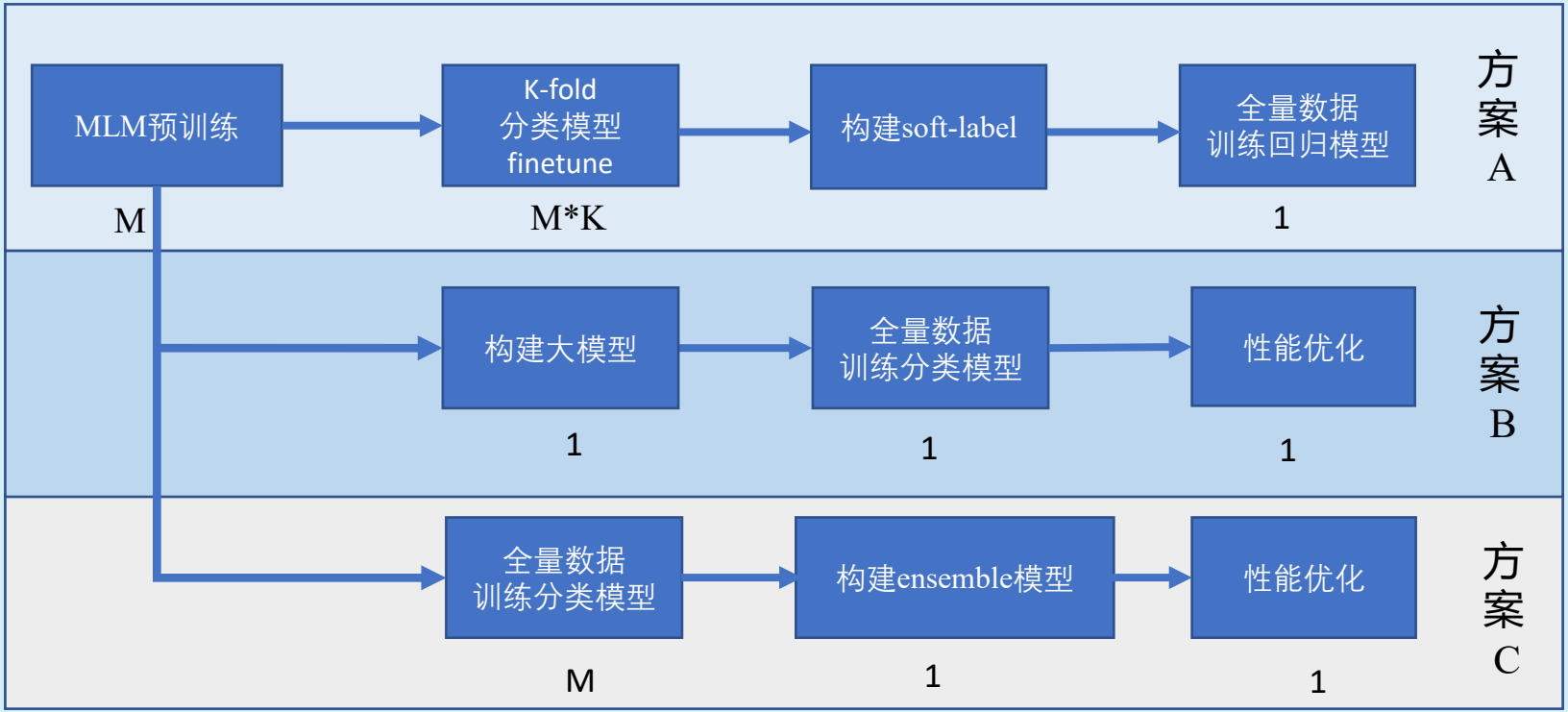
GLOBAL AI INNOVATION CONTEST

赛道三:小布助手对话短文本语义匹配

AI小花

CONTENTS

整体方案设计



其中， M=9 K=5

组件及创新点

- **数据预处理-OOV(未登录词)问题优化**

  假设 plus8, plus9 是 oov 词:

句子1	句子2	标签	处理后, 句子1	处理后，句子2
plus8 好 用 吗	plus9 好 不 好 用	0	<unk> 好 用 吗	<unk> 好 不 好 用
plus8 好 用 吗	plus8 好 不 好 用	1	<unk> 好 用 吗	<unk> 好 不 好 用

组件及创新点

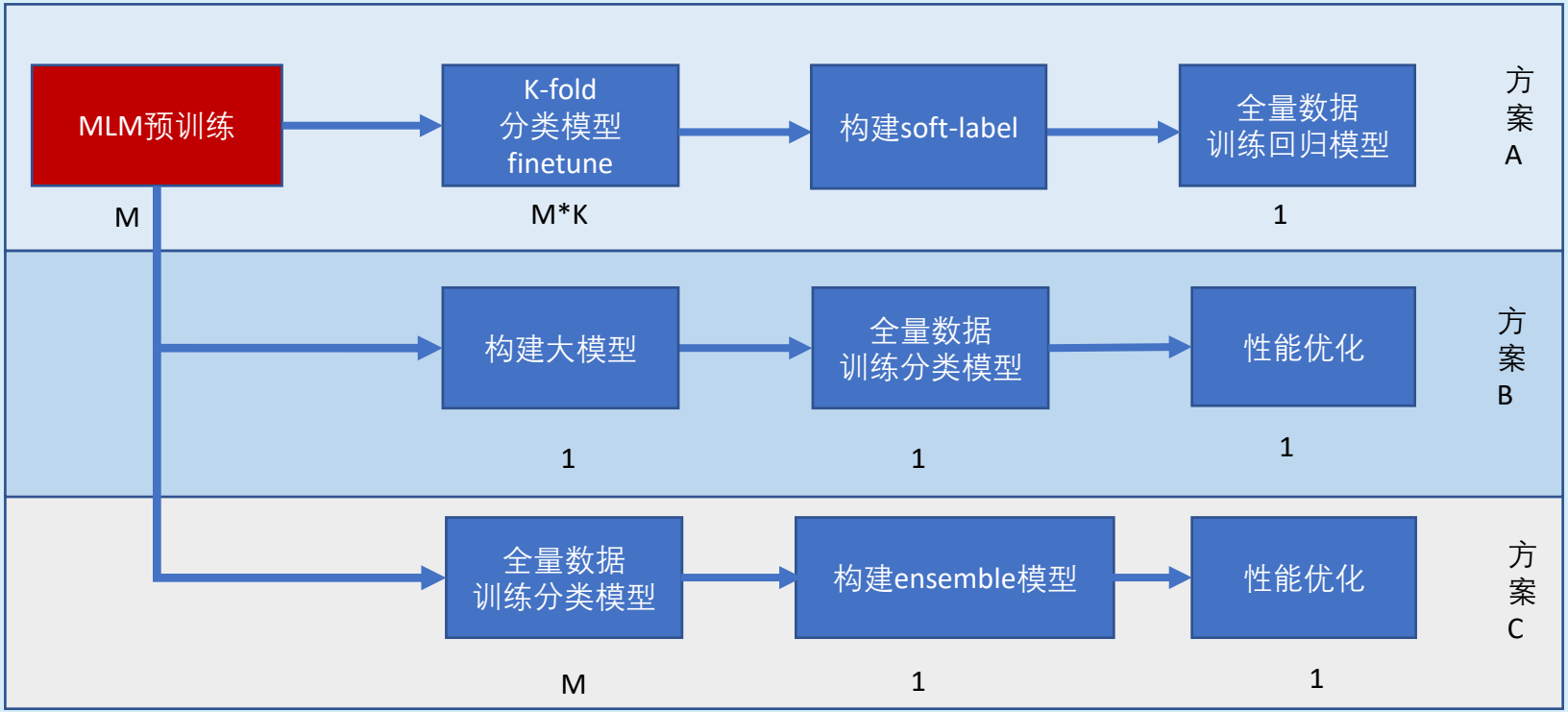
- **数据预处理-OOV(未登录词)问题优化**

句子1	句子2	标签	处理后, 句子1	处理后，句子2
plus8 好 用 吗	plus9 好 不 好 用	0	<unk-0> 好 用 吗	<unk-1> 好 不 好 用
plus8 好 用 吗	plus8 好 不 好 用	1	<unk-0> 好 用 吗	<unk-0> 好 不 好 用

区分的OOV词优化

OOV策略	Baseline	Baseline + OOV优化
AUC	0.8855	0.8891

组件及创新点



其中， M=9 K=5

组件及创新点

- **MLM预训练-mask策略**
  - **dynamic mask**

    每个epoch，随机动态mask;
  - **ngram mask**

    mask 连续的 ngram 片段;
  - **similar ngram mask**

    训练word2vec模型，用词向量计算ngram的相似度。

Mask策略	dynamic mask	ngram mask	similar ngram mask
AUC	0.8981	0.9018	0.9028

组件及创新点

- **MLM预训练-对抗训练**
  - **FGM**
  - **PGD**

    指标更好，但训练速度较慢，最终没有使用。

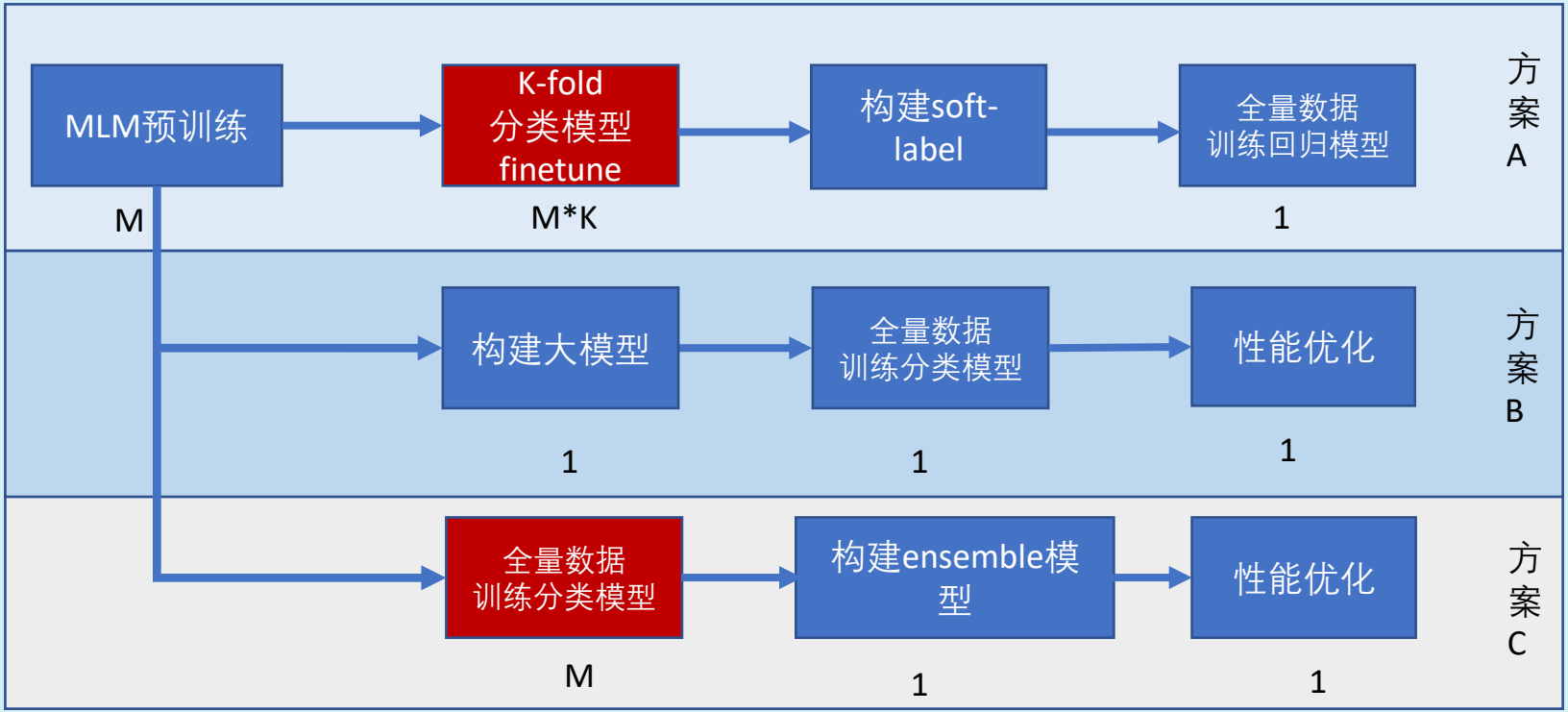
对抗训练	baseline	baseline + fgm	baseline + pgd
AUC	0.9102	0.9116	0.9121

组件及创新点

- **MLM预训练**
  - 句子对最大长度: 32
  - 数据对偶扩增
  - 训练9个模型，线上训练总时长46h

初始化参数	batch size	epoch	lr	use_fgm	线上训练时间 (单GPU V100)
uer/bert-base	512	60	1.5e-4	false	10h
uer/bert-base	512	60	1.5e-4	true	20h
uer/bert-large	576	50	1.0e-4	false	23h
hfl/roberta-base	512	60	1.5e-4	false	10h
hfl/roberta-base	512	60	1.5e-4	true	20h
hfl/roberta-large	576	50	1.0e-4	false	23h
hfl/macbert-base	512	60	1.5e-4	false	10h
hfl/macbert-base	512	60	1.5e-4	true	20h
hfl/macbert-large	576	50	1.0e-4	false	23h

组件及创新点



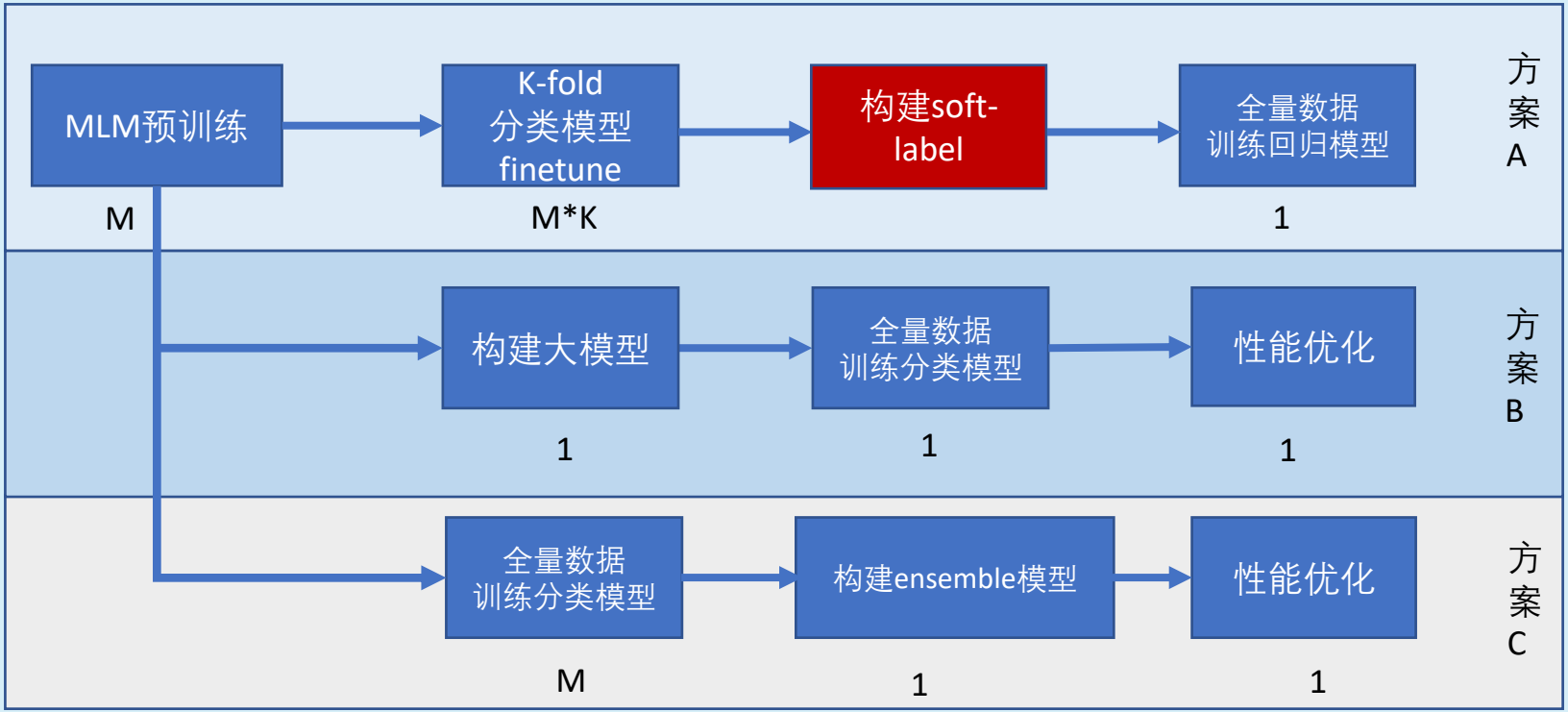
其中， M=9 K=5

组件及创新点

- **K-fold 分类模型finetune**
  - 对抗训练(FGM, PGD)
  - SWA(Stochastic Weight Averaging)

模型	baseline	baseline + fgm	baseline + pgd	baseline + fgm + swa
AUC	0.9028	0.9074	0.9076	0.9102

组件及创新点



其中， M=9 K=5

组件及创新点

- **构建soft-label**

一共训练了M*K个分类模型，对于每条数据，有M*(K-1)个模型作为训练集，有 M个模型作为验证集；

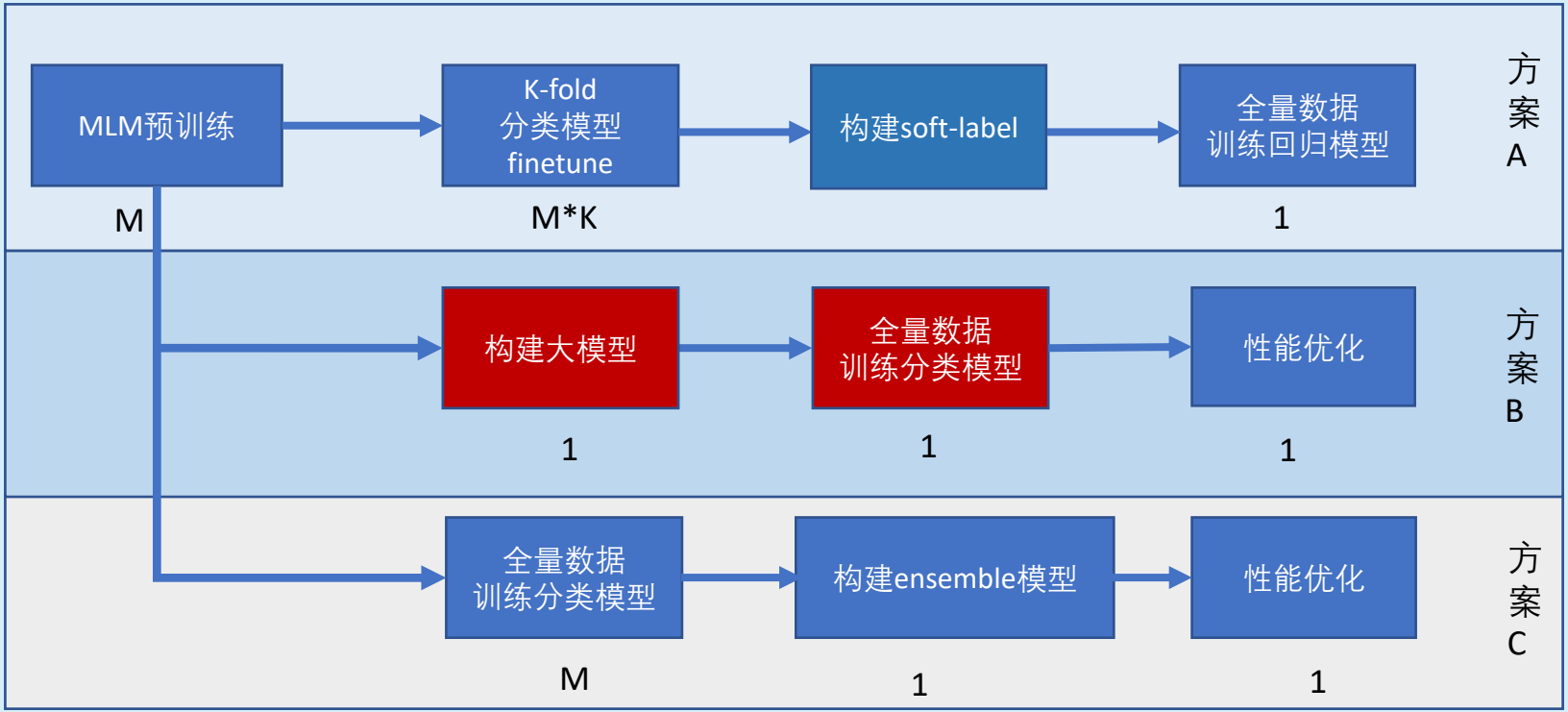
记模型 i在该条数据上的预测分数为 $score_i$，则该条数据的soft-label计算公式如下:

$$label = (\alpha * \frac{1}{M*(K-1)} * \sum_{i\in 训练模型} score_i) + ((1-\alpha) * \frac{1}{M} * \sum_{j\in 验证模型} score_j)$$

其中， α=0.55

模型	baseline	baseline + soft-label
AUC	0.9156	0.9188

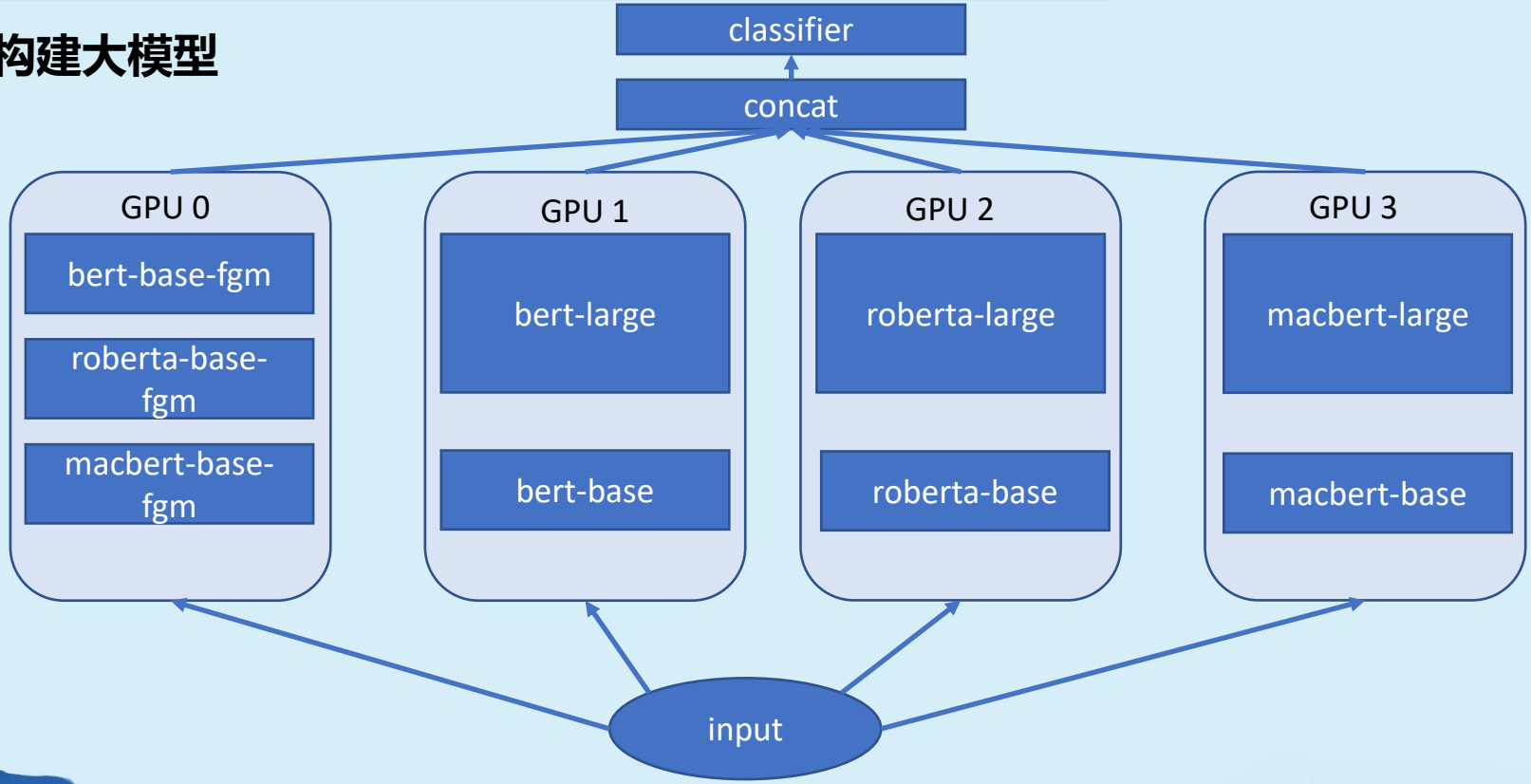
组件及创新点



其中， M=9 K=5

组件及创新点

- 构建大模型



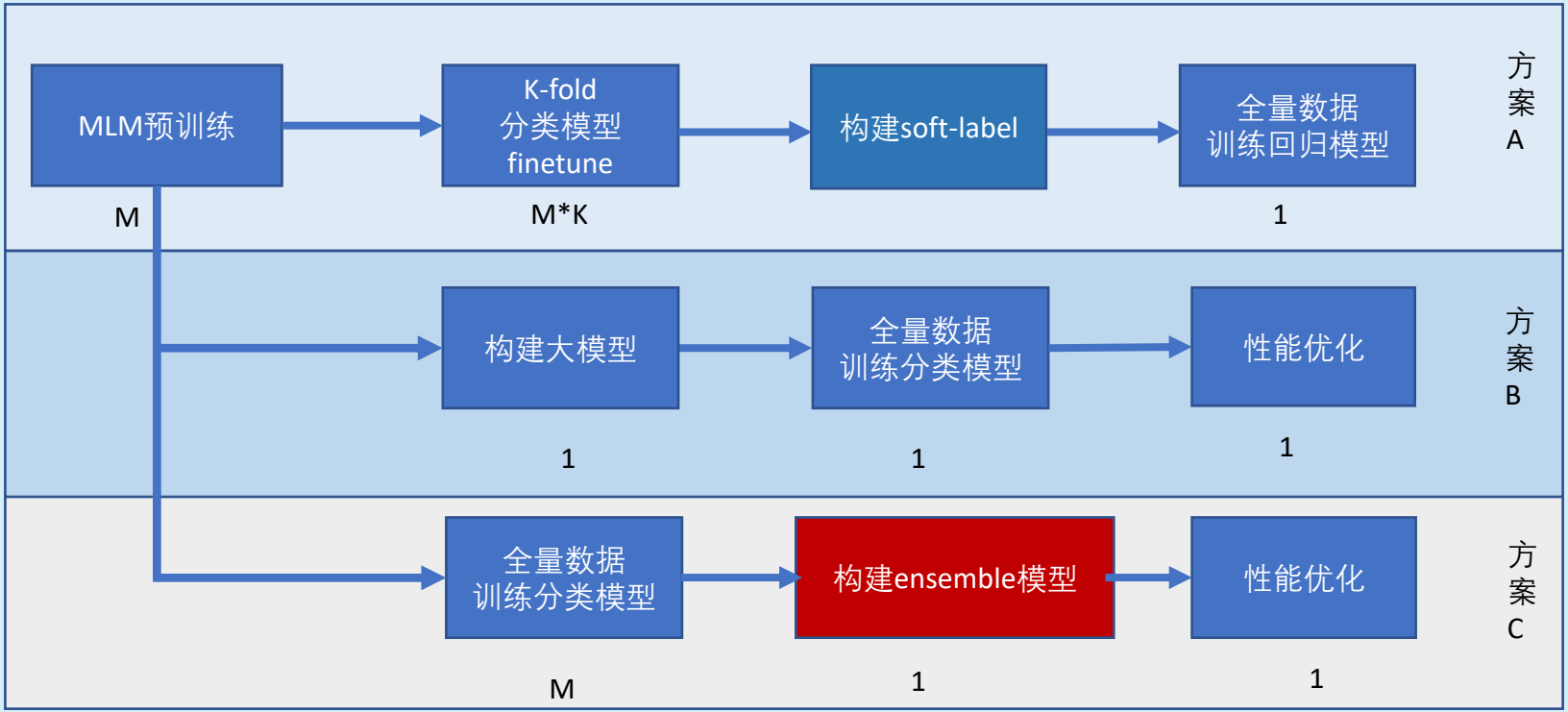
组件及创新点

- **全量数据训练分类模型**

模型	bert-base	bert-large	3 base + 3 large	6 base + 3 large
AUC	0.9504	0.9521	0.9553	0.9561

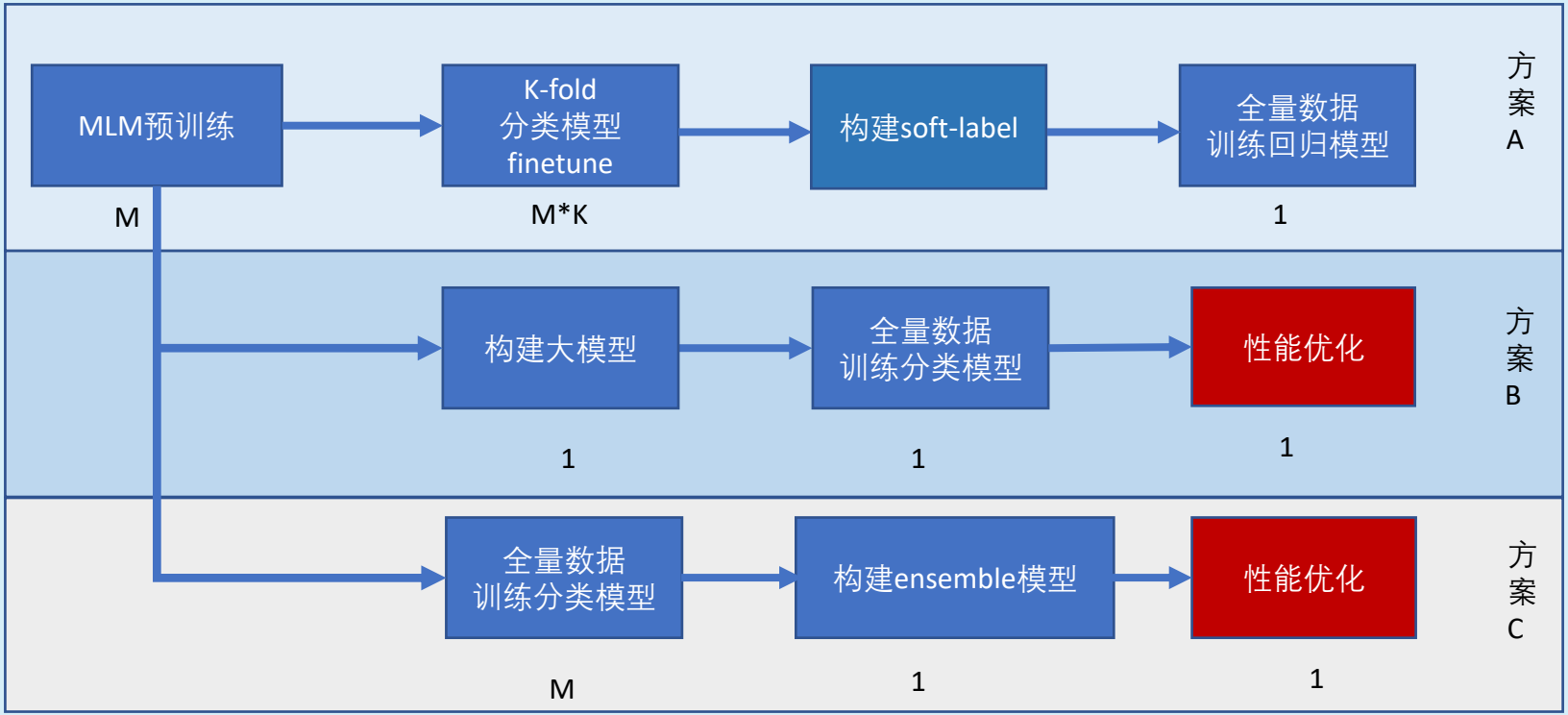
模型越大，模型的泛化性越好

组件及创新点



其中， M=9 K=5

组件及创新点



其中， M=9 K=5

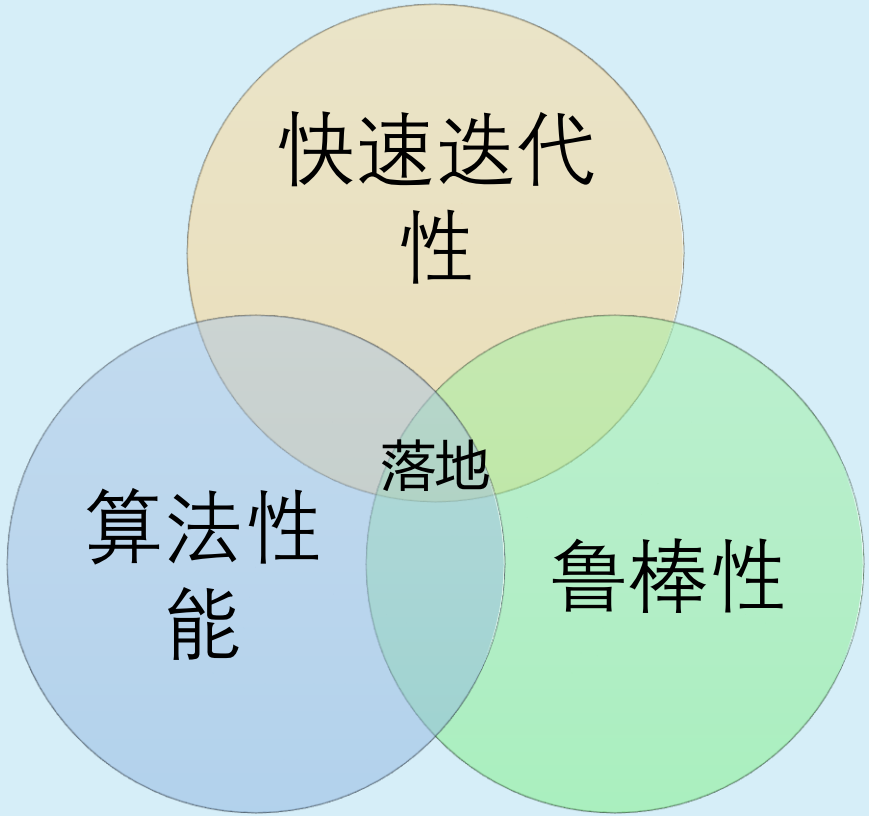
组件及创新点

- **性能优化**

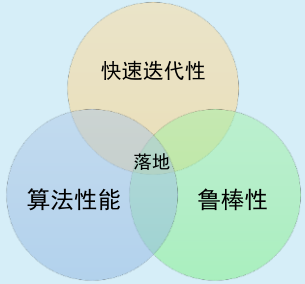
Tensorrt重写ensemble模型，构建一个tensorrt 模型。
最终线上提交结果融合了6个base模型和 3个large模型，平均latency为14ms。

排名	参与者	组织	score	valid_predict_co...	avg_time
1	AI小花	艾耕科技	0.959319	50000.00	0.01
2	[none]	清华大学	0.958319	50000.00	0.02
3	ac milan	acmilan	0.957941	50000.00	0.02
4	白[MASK]	清博	0.957908	50000.00	0.02
5	www1617	科讯嘉联	0.957870	50000.00	0.02
6	LOL王者	电子科技大学	0.957833	50000.00	0.02

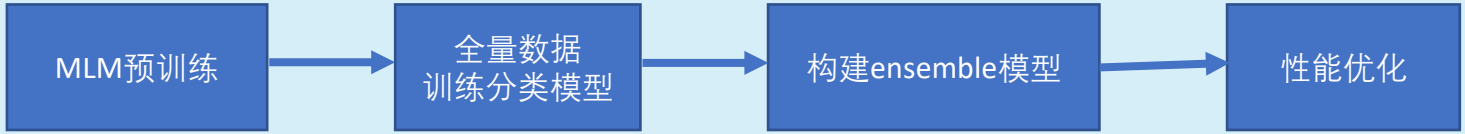
算法落地



算法落地



- **快速迭代性**

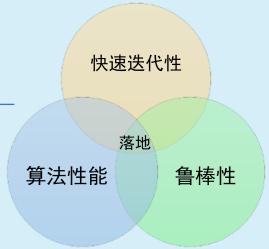


46h

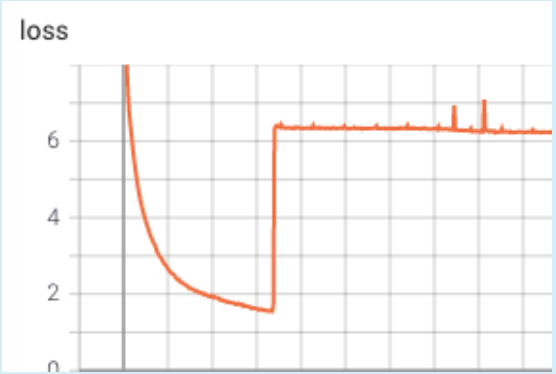
4h

- 预训练耗时较长，但是预训练过程并不需要经常更新
- 训练分类模型训练时间较短，可以做到每天更新

算法落地



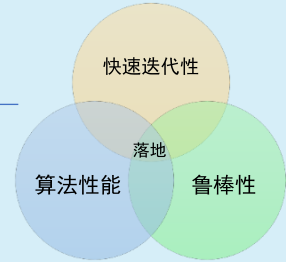
- **鲁棒性**



左图是我们线上提交的镜像，其中一个模型(macbert-large)的训练loss图，该模型已经发散

最终的模型是9个小模型ensemble的结果，如果单个模型训练异常，也不会太大影响整个算法的效果

算法落地



- **算法性能**

  - 单GPU即可部署, 需要的计算资源较少
  - 单GPU V100，模型latency 14ms，可以满足大部分任务速度要求
  - 单GPU V100，单模型(bert-base)latency 1ms，可以满足大部分任务速度要求

方案总结

创新性

模块	创新点	价值
算法设计思路	性能优先，牺牲精度换取速度	优化训练和推理过程，使得在有限的时间中，训练更多的模型，在推理阶段融合更多的模型
数据处理	优化OOV问题	缓解oov问题，提升模型指标
MLM预训练	similar ngram mask	加快模型收敛速度，提高模型泛化性
	对抗训练(FGM)	提高模型泛化性
分类模型finetune	对抗训练(FGM)	提高模型泛化性
	SWA	提高模型泛化性
构建soft-label	kfold soft-label计算	充分利用训练数据，提高模型泛化性
性能优化	tensorrt	提升模型性能，优化latency
算法pipeline	多模型ensemble	提高系统鲁棒性

方案总结

实用性

训练速度快	pipeline训练时间50h，单模型训练时间11h
推理速度快	latency 14ms
算法效果好	AUC: 0.9593
模型可快速迭代	分类模型训练时间短，支持每日更新迭代
系统鲁棒性高	ensemble多个模型，即使部分模型训练失败，对系统结果影响较小
硬件计算资源需求少	单个GPU即可部署

部署生产可能碰到的问题：算法训练使用了混合精度，推理利用了tensorrt fp16，需要GPU支持tensor core。

方案总结

适用于多种场景/行业/类别/任务

扩展性

方案总结

✓ 总结

◆牺牲单模型精度，换取更快的速度

- 使用bert，不用nezha (牺牲2k)
- 使用fgm，不用pgd (牺牲0.5k)
- 推理使用fp16，不用fp32 (牺牲0.3k)
- 不用对称模型 (牺牲0.3k)

◆充分优化性能

- 训练开启混合精度训练
- 优化数据处理和mask策略性能，使得CPU速度和GPU速度匹配
- 推理使用tensorrt，充分利用V100显卡计算单元，GPU占有率接近100%

◆多个“小”模型ensemble 优于 单个大模型

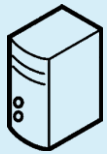
- 训练更容易
- 鲁棒性更高

方案总结

✓ 展望

科学研究

- 更加高效的模型结构
- 更加快速的训练方法
- 更高效的模型压缩算法



实际应用

- 该方案的部署需要较大的内存和较强的计算设备，目前只能部署到GPU服务器上。如果能让模型在移动设备上运行，必然能进一步提高产品的用户体验。


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