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

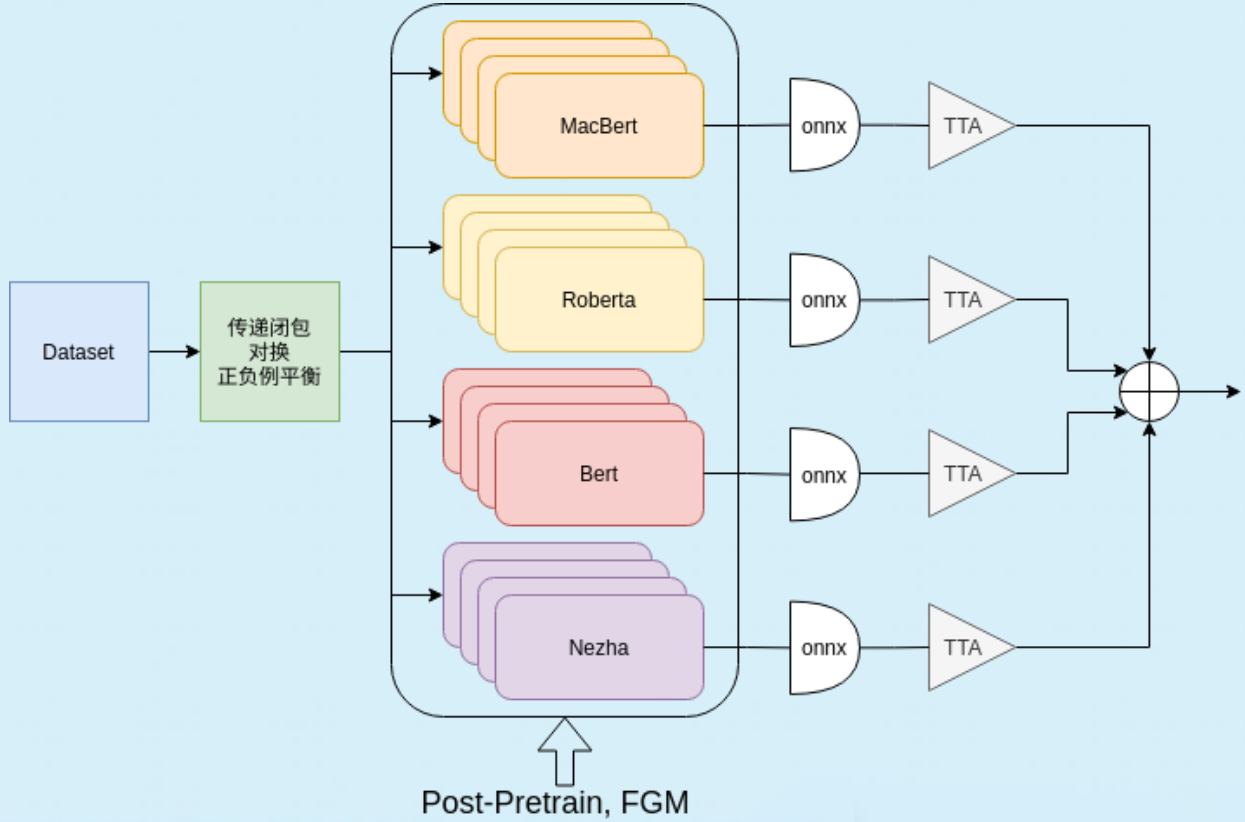
[none]

CONTENTS

整体设计

总体架构

- 数据增广：传递闭包/对换/正负例平衡
- 训练：Post-Pretrain+Finetune、n-gram mask、对抗学习
- 融合：checkpoints、多模型
- 预测：onnx、测试时增强 (TTA)



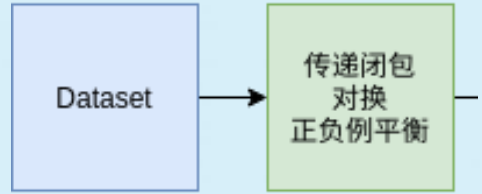
整体设计

数据增广

1. 先进行传递闭包构造额外正例：若p1p2为正例且p2p3为正例，则p1p3也为正例。

2. 再进行对换构造额外的正负例：若p1p2属于数据集，则将p2p1也加入数据集，label不变。

3. 对换过程中保证正负例的平衡：传递闭包构造结束后正负例比例约为0.88:1，此时为所有正例构造对换数据，对随机部分负例构造对换数据，使得正负例比例变为1：1 (类别平衡)。



处理前训练集40w，正负例0.56:1，处理后训练集95w，正负例1:1

整体设计

训练

1. 第一步在未增广数据上进行 post-pretrain，使用n-gram mask。训练时以句对形式输入。post-pretrain可以很好地增强模型对于文本语义的理解能力，效果显著。而使用n-gram mask，相比于BERT原始的mask，会更有难度，也可以更好的帮助模型理解中文语义信息。

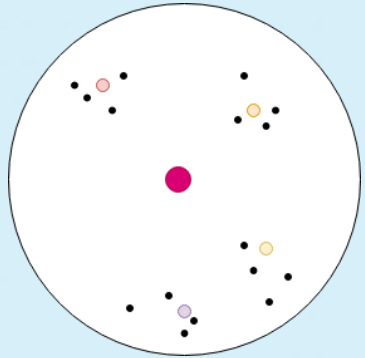
2. 第二步在增广数据上进行finetune，使用FGM对抗学习方法提升模型鲁棒性。参数使用自己写的超参数搜索代码来完成调优。

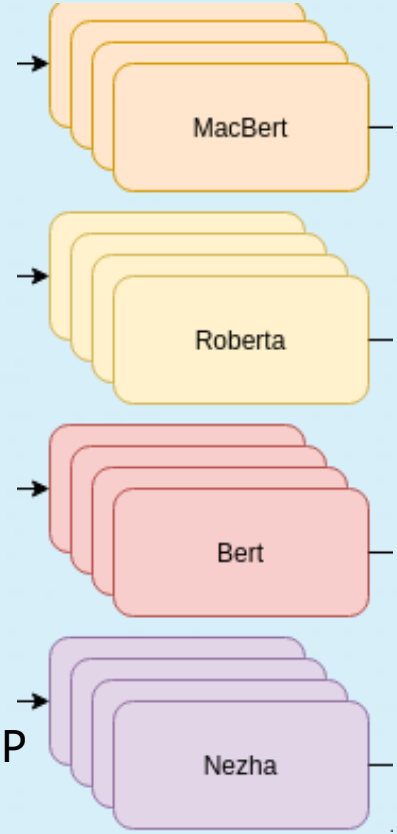
整体设计

模型融合

1. 使用不同checkpoint进行模型内的融合，将同一模型的不同checkpoint的参数进行算术平均，增强表现的稳定性。

2. 使用四种模型进行模型间的融合，取所有模型预测的平均作为最终结果。利用不同模型的差别，减小最终假设的bias。





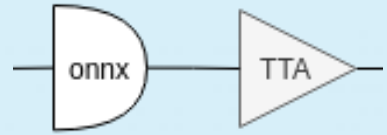
整体设计

预测

1. 使用onnx进行模型加速。

2. 使用测试时增强 (TTA) 提高预测准确性。给定句对p1p2，模型还会预测p2p1的结果，两个预测结果取平均作为最后结果。一定程度上减小模型预测的variance。

3. 融合不同模型的预测结果，算术平均效果最佳。



创新和落地

创新

- **post-pretrain**
  - 在所有数据上进行学习率较小的 pretrain
  - 采用n-gram mask
  - 这种做法可以使模型更熟悉该数据的句型、内容分布与语义信息
  - 单模可以提高数个百分点
- **模型checkpoint融合**
  - 用取平均的方式融合一个模型不同checkpoint之间的权重，增强模型鲁棒性
  - 理论上可行：模型权重都是做的线性变换，故可以直接取算术平均
  - 效果上很好：单独的BERT在使用五个checkpoint融合之后，在复赛a榜上的效果从0.9478变为0.9521
  - 为什么不直接将不同checkpoint的预测结果取平均？预测时间有限！
- **测试时增强 (TTA)**
  - 测试时交换句对得到另一个结果，两个预测结果取平均
  - 灵感来源于CV的TTA，为原图像创造多个版本，预测取平均
  - 单模可以提高约一个千分点

创新和落地

落地

- **资源消耗少**
  - 消耗的硬件资源较少，在4张V100 GPU上训练仅需不到两天，推断时仅需一张GPU
  - 推断时间较短，完整使用四个融合的模型每秒可预测50-100条数据（bert结构模型快于nezha），使用大于1的batch size可以预测更多数据
- **灵活可调整**
  - 使用时可根据实际需求调整使用的模型数量，比如时间要求较高的场景下可以把四个模型减到一个模型，预测效果上下降不多，但预测速度可变为原来的四倍
  - 当搜索范围很大时，可以与其他算法进行搭配，例如使用双塔式模型来进行召回，而使用我们的这种交互式模型来进行排序

方案总结

- **创新性**
  - 使用模型checkpoint融合技术，巧妙地在预测时间有限的场景下提高了模型的性能
  - 综合多种预训练模型，充分利用不同模型之间的差异性
  - 使用数据增广，对抗训练，TTA，ONNX加速等技术，有效提升模型性能
- **实用性**
  - 软硬件资源消耗小，在4张V100 GPU上训练仅需不到两天，测试时仅需一张GPU
  - 推断时间较短，batch size为1的情况下每秒可预测50-100条数据
- **拓展性**
  - Checkpoint融合与具体任务和场景解藕，可以迁移到其他深度学习模型
  - Checkpoint融合数量上升并不会导致预测时间的上升(不会影响服务QPS)，有较好的扩展性
  - 整个方案可以轻松地拓展到其他分类任务上，比如情感分类等，只要改变数据以及相应的数据处理即可

方案总结

- **思考和展望**
  - 我们还尝试了self-training，但是由于无标签的数据太少，效果不明显，如果有大量的无标签数据，应该可以进一步提升性能
  - 除了fgm之外，我们还尝试了pgd和freelb，pgd不好的原因主要是参数不好调整，而对于freelb来说，由于时间原因我们的实现里面只攻击了word embedding，而没有攻击token type embedding等输入embedding，如果加上对其他输入embedding的攻击应该可以超过fgm。大致效果排序：$fb_{all} > fgm_{all} \approx fb_{word} > pgd_{all} > fgm_{word} > pgd_{word}$
  - 一个有意思的现象是提供的数据中有部分数据是重复的，我们尝试了去掉这部分重复数据之后再进行训练，效果反而下降了，推测是这部分数据较难，重复它们有利于帮助模型学习，基于这点其实可以不断迭代重复较难数据（或者加大训练权重），后续可以继续研究


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