2023.4.3 - 飞机炸弹 - 周周星分享

长喙鲸vr0v · 4 小时前

很荣幸能拿到本次周星星（原本以为只排在第三，所以没提前准备要分享的内容。比较仓促草率，请各位谅解）

针对脱敏数据的方法，男枪哥在他的分享里做了详细的介绍。我这里的方法比较粗暴，手动定义 eos_token_id, bos_token_id 之后resize_token_embeddings到1400，具体代码如下：

```
config = AutoConfig.from_pretrained(PRETRAIN_NAME)
config.bos_token_id = bos_token_id # 这里我设为1
config.eos_token_id = eos_token_id # 这里我设为2
config.decoder_start_token_id = eos_token_id
config.forced_eos_token_id = eos_token_id
config.pad_token_id = pad_token_id # 这里我设为0
model = AutoModelForSeq2SeqLM.from_pretrained(self.model_config.pretrain_path, c
model.resize_token_embeddings(1400)
```

训练数据格式上：

input_ids：`[bos_token_id] ${原始输入} [eos_token_id]`

labels：`${原始输出} [eos_token_id]`

在比赛初期，我主要尝试了几种不同结构的预训练语言模型，Bart、Pegasus、CPT等，根据经验只尝试了encoder-decoder架构的模型，得出的初步结论是bart的表现可能是这里面最优的，大概什么都不加线上能有2.5+

之后就是NLP竞赛一些常规的Trick，例如FGM、EMA这种都是有效果的。
在解码策略上，我尝试了beam-sample、beam-search、sample、greedy这些采样方法。对于该任务的评价指标，beam-search策略是我调下来最优的，大概调这 `num_beam` 和 `length_penalty` 这两个参数收益比较明显

上述没进行预训练版本线下大概2.67，线上有2.81（我也不知道为啥差距这么大）

预训练部分主要使用了n-gram mask的方法，可以参考男枪之前在群里分享的:https://github.com/daniellibin/gaiic2021_track3_querySim/blob/master/code/bert-base-count5/pretrain/NLP_Utils.py#L90

在输入的拼接上我主要采用了input+label与纯input两种方法，个人侧下来，如果使用前者，建议在微调时候调小学习率，不过好像这两种得到的结果差异性不大。两种方法的预训练setting：lr=1e-4, batch_size=128, epoch=150。 整体线下可以到2.75左右，线上2.85(+0.05)

之后我在预训练版本的模型上重新调整的参数，最后线下2.86，线上2.91

以上就是我的全部分享，希望可以帮到各位。

已赞 分享

评论

fmaa · 2 小时前 · #1楼

该评论已被用户删除

fmaa · 2 小时前 · #2楼 回复

总结一下提分最高的实际上是调参是吧，能复现男枪哥说的预训练涨接近0.2吗

发布评论

B I H 插入 帮助

发布