深度之眼
deepshare.net

Kaggle feedback3分类大赛
模型训练技巧

导师：William

深度之眼
deepshare.net

目录

1 US patent

top solution: simple is better

数据是最关键的提分手段

处理过程：将同一anchor下的target 都加入到训练数据中会大大的提高lb成绩。

Input data from baseline

```
anchor [SEP] target [SEP] context text
```

Our input data

```
anchor [SEP] target; target_x1; target_x2; ... traget_xn; [SEP] context text
```

1 US patent

top solution: simple is better

线下cv策略：

StratifiedGroupKFold 和 GroupKFold， 都可以有提高线下cv和lb的一致性，

GroupKFold会使得每Flod的训练数据size一致，StratifiedGroupKFold 这个会使得train 数据和 val的label的数据分布是一致的，both 都要使用

1 US patent

top solution: simple is better

Model

Pretrained model

- Electra large
- Bert For Patent
- DeBERTa V3 large
- DeBERTa V1
- DeBERTa V1 xlarge

Loss

- binary cross entropy loss
- mean squared error loss
- pearson correlation loss

There is no big difference among those loss functions. However, using different loss in training phrases will lead to high diversity when ensembling because the distribution of the prediction looks different from oof.

1 US patent

top solution: simple is better

model层面：

add lstm head， bert层设置较小的学习率(2e-5, 3e-5)， 而lstm层设置较大的学习率(1e-3)，

受promotion learning启发冻结embedding层

1 US patent

top solution: simple is better

fgm

```
class FGM():
    def __init__(self, model):
        self.model = model
        self.backup = {}
    def attack(self, epsilon=1., emb_name='word_embeddings'):
        # emb_name这个参数要换成你模型中embedding的参数名
        for name, param in self.model.named_parameters():
            if param.requires_grad and emb_name in name:
                self.backup[name] = param.data.clone()
                norm = torch.norm(param.grad)
                if norm != 0 and not torch.isnan(norm):
                    r_at = epsilon * param.grad / norm
                    param.data.add_(r_at)
    def restore(self, emb_name='emb.'):
        # emb_name这个参数要换成你模型中embedding的参数名
        for name, param in self.model.named_parameters():
            if param.requires_grad and emb_name in name:
                assert name in self.backup
                param.data = self.backup[name]
        self.backup = {}
```

https://www.kaggle.com/competitions/us-patent-phrase-t

1 US patent

top solution: simple is better

ema

```
class EMA():
    def __init__(self, model, decay):
        self.model = model
        self.decay = decay
        self.shadow = {}
        self.backup = {}

    def register(self):
        for name, param in self.model.named_parameters():
            if param.requires_grad:
                self.shadow[name] = param.data.clone()

    def update(self):
        for name, param in self.model.named_parameters():
            if param.requires_grad:
                assert name in self.shadow
                new_average = (1.0 - self.decay) * param.data + self.decay * self.shadow[name]
                self.shadow[name] = new_average.clone()

    def apply_shadow(self):
        for name, param in self.model.named_parameters():
            if param.requires_grad:
                assert name in self.shadow
                self.backup[name] = param.data
                param.data = self.shadow[name]

    def restore(self):
        for name, param in self.model.named_parameters():
            if param.requires_grad:
                assert name in self.backup
                param.data = self.backup[name]
        self.backup = {}
```

https://www.kaggle.com/competitions/us-patent-phrase-to-phrase-matching/discuss

2 feedback 2

top solution: simple is better

本次比赛的目标是将学生写作中的议论元素分类为**“有效”、“充分”或“无效”**。参赛者将创建一个使用代表美国 6 至 12 年级人口的数据进行训练的模型，以尽量减少偏差。来自本次比赛的模型将有助于为学生获得关于他们的议论文的更多反馈铺平道路。借助自动指导，学生可以完成更多作业，最终成为更自信、更熟练的作家。

2 feedback 2

top solution: simple is better

数据工程:

将easy的每个部分用SEP拼接起来，并且在essay每个部分前加入模版提示，事例如下

[SEP]Lead. Discourse_00[SEP]Position. Discourse_01[SEP]Claim. Discourse_02 …….

2 feedback 2

top solution: simple is better

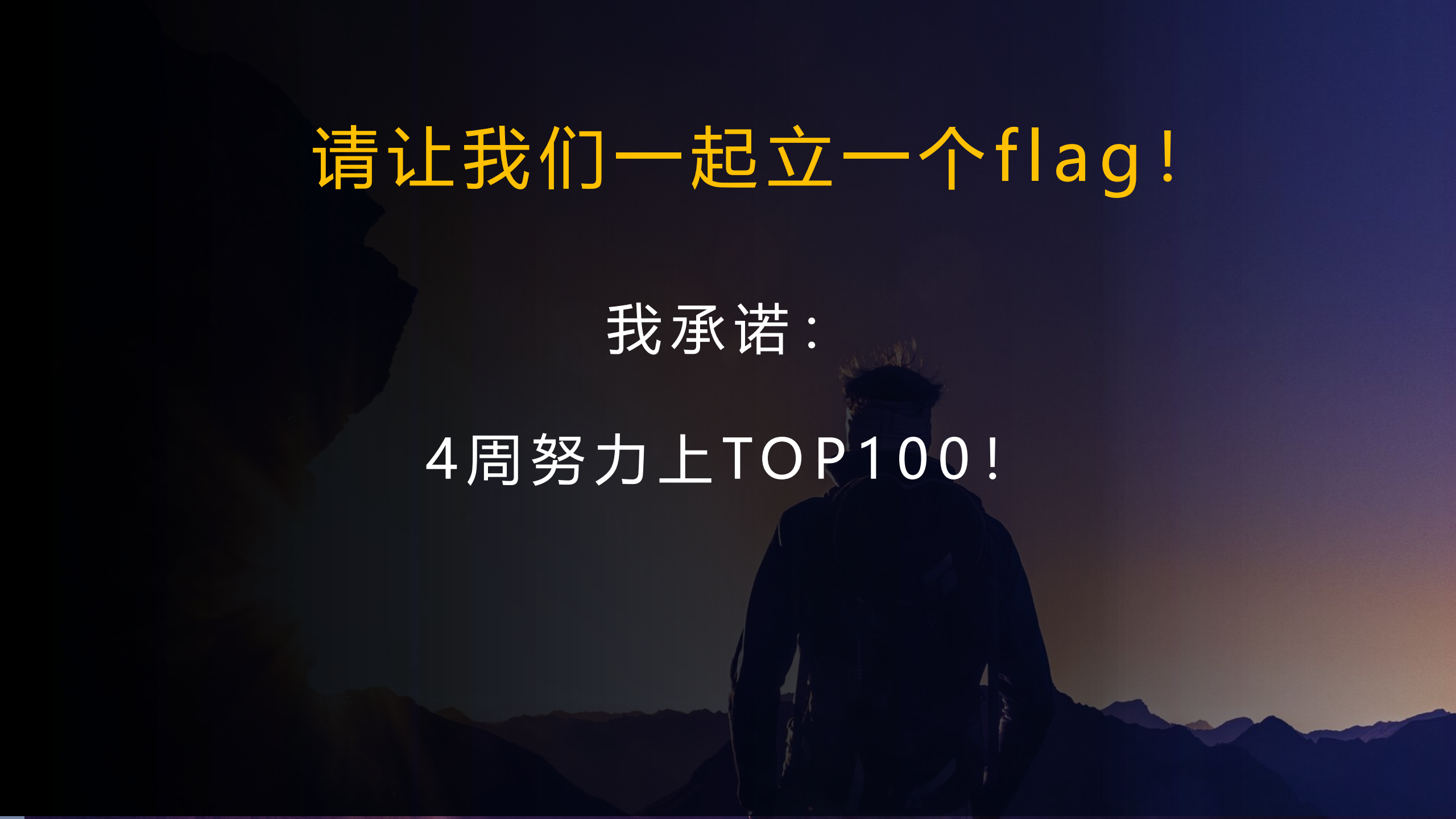
提分流程:

1. 训练一个较好的deberta - large模型，在deberta-large的基础上对feedback1的数据进行伪标签

2. 用 PL的数据 和 现有数据进行训练，采用AWP，并且对不同learning rate进行不断尝试

3. 将deberta-v3-large, deberta-large, deberta-xlarge等3个模型结果进行融合

3 答疑互动


请让我们一起立一个flag！
我承诺：
4周努力上TOP100！

结语

再小的细节，也值得被认真对待

deepshare.net

深度之眼

联系我们：

电话：18001992849

邮箱：service@deepshare.net.net

Q Q： 2677693114

公众号

客服微信