灾难侦测

张浩楠

写提案，清洗、模型和提交，展示的课件和手稿

蔡宇生

写提案，清洗、模型和提交，展示的课件和手稿

吴昊原

写提案，清洗、模型和提交，展示的课件和手稿

展示目录

- 数据、模型、提交

灾难侦测

- 随着互联网的进步，
- 更多的人会去在互联网上分享它们的故事，包括灾难，
- 从互联网侦测灾难，而不是等市民去叫天天九，
- 能使消防车等工作人员更早到达去救援，从而减小损失。

灾难侦测

- 去预测一个帖子是否跟真实灾难有关

数据概观

训练

编号	位置	关键字	文本	标签
1	英国	缺失	突然听...	0
3	缺失	闪光	意外有...	1
4	新加坡	爆炸	飞机在...	0
6	缺失	龙卷风	恐怖的...	1

⋮ ×7613

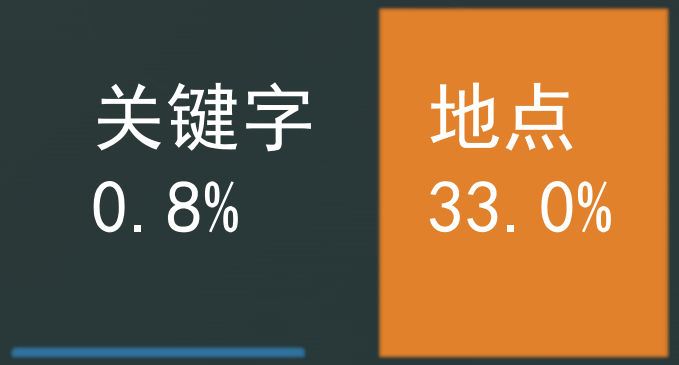
测试

编号	位置	关键字	文本	标签
2	加拿大	暴雨	看到了...	
5	日本	缺失	恐怖的...	
7	英国	缺失	飞机在...	
8	新加坡	缺失	一个意...	

⋮ ×3263

关键字、地点

缺失值比例



帖子的文字

- 帖子的文字没有缺失值

标签

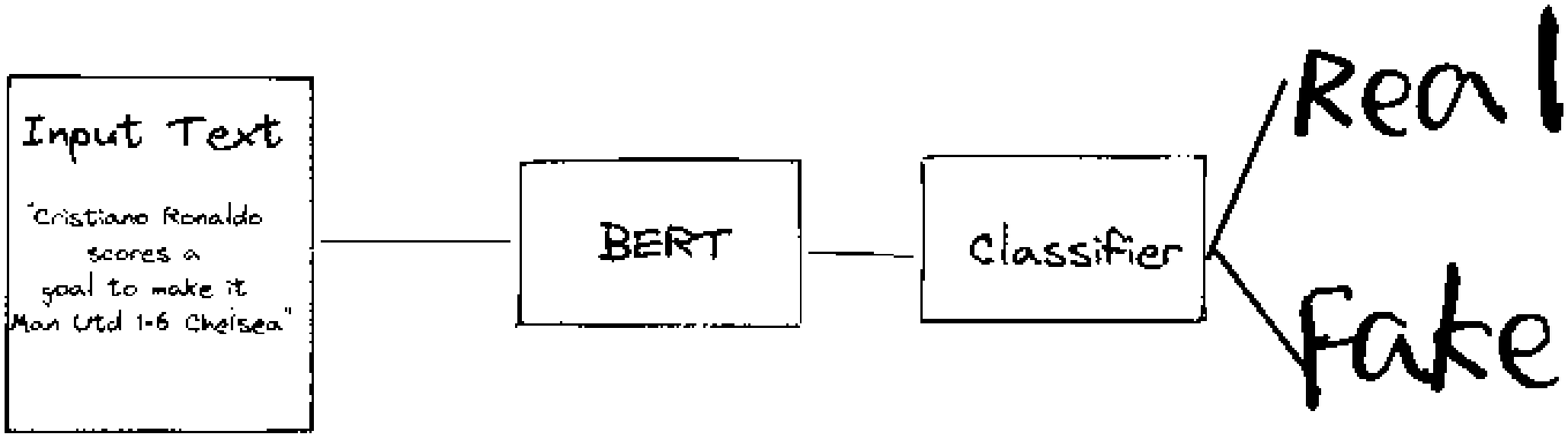
- 0 → 假灾难
- 1 → 真灾难

帖子文字

处理

- 删除不常见的符号， å Ç
- 拆开略缩(contraction)， he’s → he is
- 删除链接(link)， *http://www.kaggle.com*
- 删除井号(hashtag)， #disaster @kaggle
- 删除重复的样本， 甚至有的样本被标注错误

模型概观



https://towardsdatascience.com/text-classification-with-bert-in-pytorch-887965e5820f

模型骨干

- bert-base-uncased 最流行
- 大，不区分大小写
- 12层变形金刚编码器
  （Transformer encoder）
- 12注意头（Attention head）
- 110000000参数（110M parameters）

bert 令牌器

- 1 添加特殊令牌[CLS][SEP]

[CLS] This here's an example of using the BERT tokenizer. [SEP]

- bert 令牌器
- 2 切割单词

（子词令牌化，wordpiece）

tokenizer →→ token ##izer

bert 令牌器

- 3 转变成数字序列(sequence)

[CLS]	This	,	s	an	example	of	using	The	B	##ER	##T	##T	token	##izer	[SEP]
↓															
101	1188	1303	112	188	1126	1859	1104	1606	1103	139	9637	1942	22559	17260	102

bert 令牌器

This here’s an example of using the BERT tokenizer.

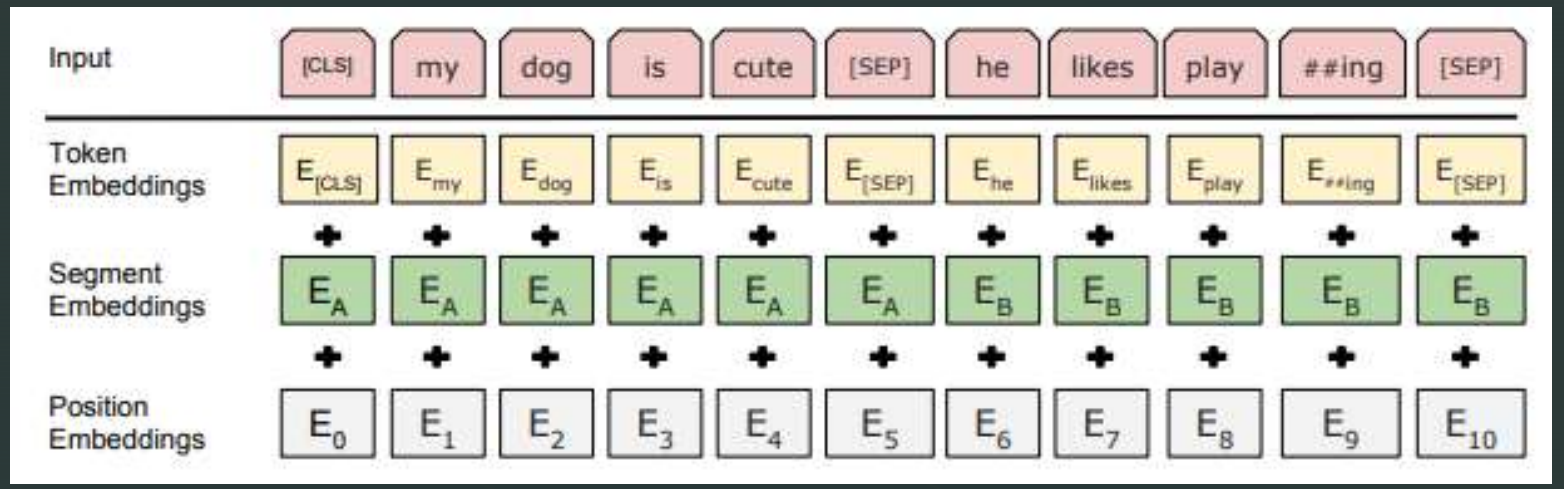
↓

101	1188	1303	112	188	1126	1859	1104	1606	1103	139	9637	1942	22559	17260	102

bert 的组成

- 词嵌入(word embeddings)
- 12层变形金刚编码器(Transformer encoder)

词嵌入



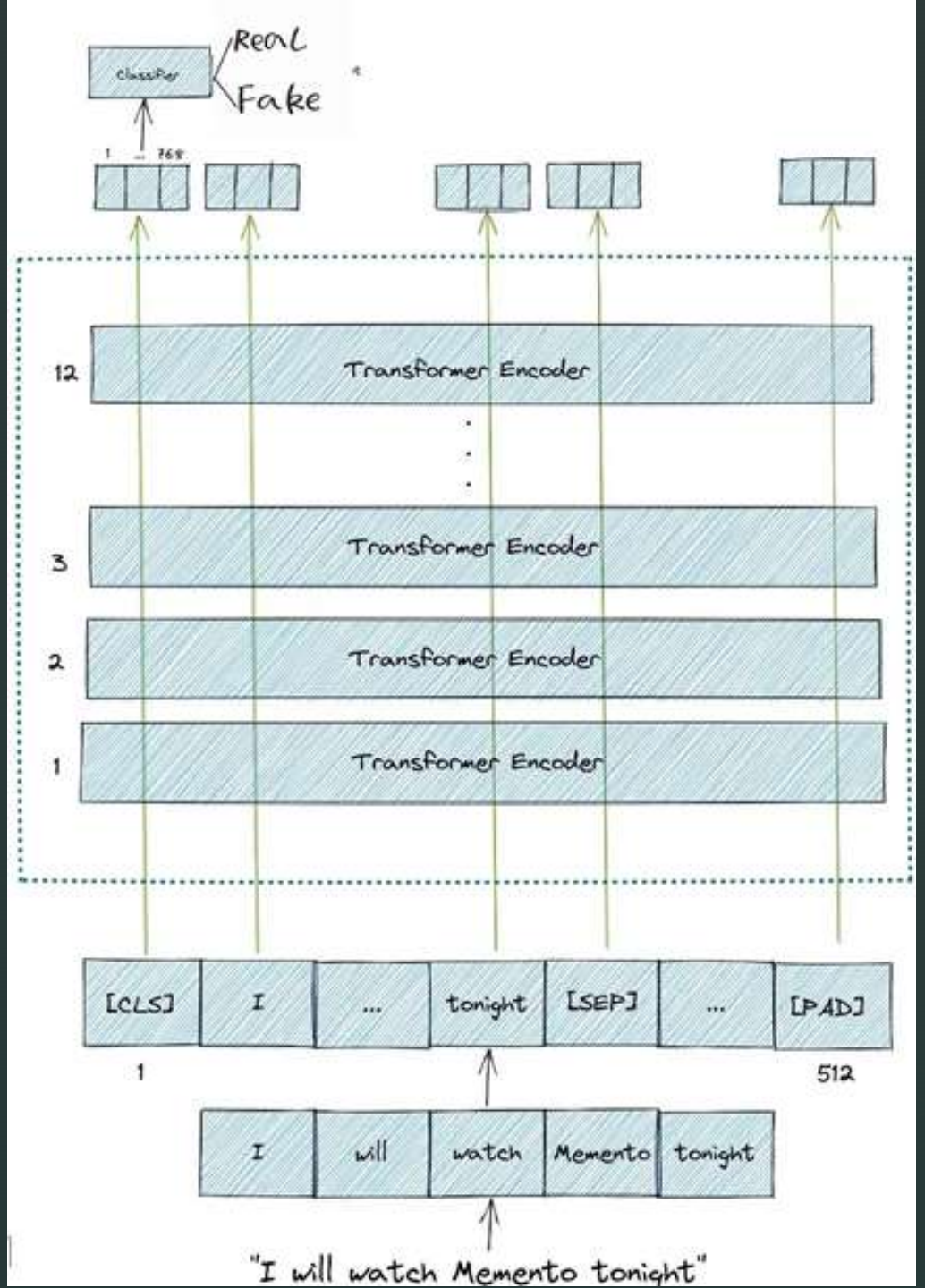
- 令牌嵌入：　一个词对应一个令牌嵌入
- 段嵌入：　这个词在第一句话还是第二句话
- 位置嵌入：　这个词在句子中的位置

https://arxiv.org/pdf/1810.04805.pdf

变形金刚编码器

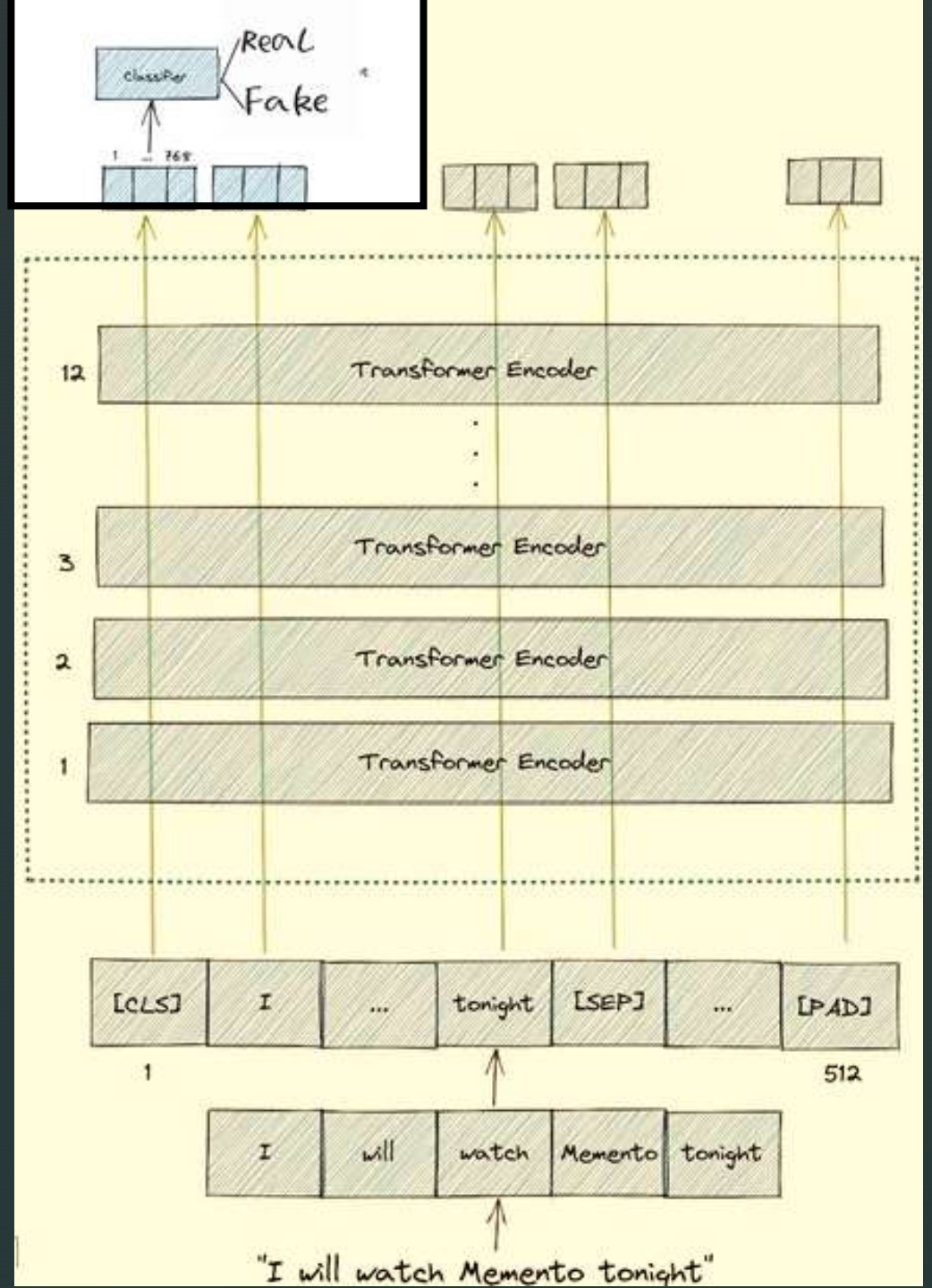
模型概观

在最后的输出中拿出[CLS]
对应的嵌入，
将会是规模为768的向量。
接上密集层和Softmax激活函数
输出规模为2的向量
对应两个标签(0, 1)。



模型选择

- 冻结词嵌入器
- 和变形金刚编码器
- 只去优化密集层分类器



模型选择 网格搜索 (GridSearch找最好超参)

- 分类器密集层数量、优化器

1层， 亚当优化器	2层， 亚当优化器	3层， 亚当优化器
1层， 随机梯度下	2层， 随机梯度下	3层， 随机梯度下

模型选择 网格搜索 (GridSearch找最好超参)

- (1层, Adam) 和 (2层, Adam) 最快最准
- 另外去采用交叉熵损失函数，F1score验证评价指标

1层， 亚当优化器	2层， 亚当优化器	3层， 亚当优化器
1层， 随机梯度下	2层， 随机梯度下	3层， 随机梯度下

F1score 是准确率和召回率的均值，越大约好

模型选择 交叉验证 (KFold)

- 交叉验证先将数据集划分为5个大小相同的子集
- 每轮随机的选择 4份作为训练集
- 剩下的1份做测试集
- 重新随机选择 4份来训练数据
- 5轮之后，我们选择验证分数最高的一个。

模型选择 交叉验证 (KFold)

验证分数

	1层亚当	2层亚当
模型一	0.73	
	0.77	
	0.72	
	0.84	
	0.81	
模型二		0.85
		0.72
		0.72
		0.84
		0.75

软投票(soft-vote)和提交

预测一

0	1
0.3	0.7
0.8	0.2
0.9	0.1
0.9	0.1

...

预测二

0	1
0.3	0.7
0.8	0.2
0.9	0.1
0.9	0.1

...

软投票(soft-vote)和提交

预测一和预测二的均值

预测一

0	1
0.30	0.70
0.80	0.20
0.90	0.10
0.90	0.10
...	

预测二

0	1
0.20	0.80
0.80	0.20
0.70	0.30
0.20	0.80
...	

投票结果

0	1
0.25	0.75
0.80	0.20
0.80	0.20
0.55	0.45
...	

最大

标签
1
0
0
0
...

软投票(soft-vote)和提交

预测一和预测二的均值

预测一

0	1
0.30	0.70
0.80	0.20
0.90	0.10
0.90	0.10
...	

预测二

0	1
0.20	0.80
0.80	0.20
0.70	0.30
0.20	0.80
...	

投票结果

0	1
0.25	0.75
0.80	0.20
0.80	0.20
0.55	0.45
...	

最大

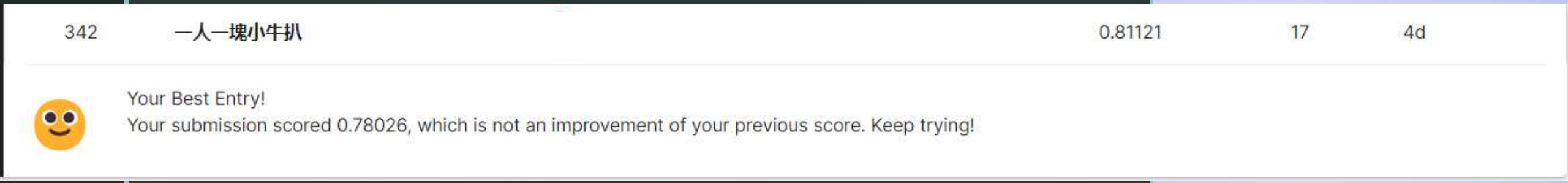
标签
1
0
0
0
...

提交

最终的预测

标签
1
0
0
0
...

排行榜分数F1score 0.81
与之前的验证分数相近



张浩楠

写提案，清洗、模型和提交，展示的课件和手稿

蔡宇生

写提案，清洗、模型和提交，展示的课件和手稿

吴昊原

写提案，清洗、模型和提交，展示的课件和手稿

灾难侦测

张浩楠

写提案，清洗、模型和提交，展示的课件和手稿

蔡宇生

写提案，清洗、模型和提交，展示的课件和手稿

吴昊原

写提案，清洗、模型和提交，展示的课件和手稿

谢谢大家