

# 预训练时代小样本学习 及其在中文上实践

模型与实验

# 概览

- 什么是few-shot learning
- 为什么需要few-shot-learning
- 现阶段方法（PET, Ptuning）
- 实验结果
- 数据问题及clue问卷调查结果

# 小样本学习(few-shot learning)是什么

- 问题定义

人类非常擅长通过极少量的样本识别一个新物体，比如小孩子只需要书中的一些图片就可以认识什么是“斑马”，什么是“犀牛”。在人类的快速学习能力的启发下，研究人员希望机器学习模型在学习了一定类别的大量数据后，对于新的类别，**只需要少量的样本就能快速学习**，这就是 Few-shot Learning 要解决的问题。

# 为什么要做few-shot learning

- zero-shot

完全依赖预训练模型，很难根据主观判定及需求进行修改

“张继科景甜恋情公开 名下多款百万级别豪车婚车曝光”

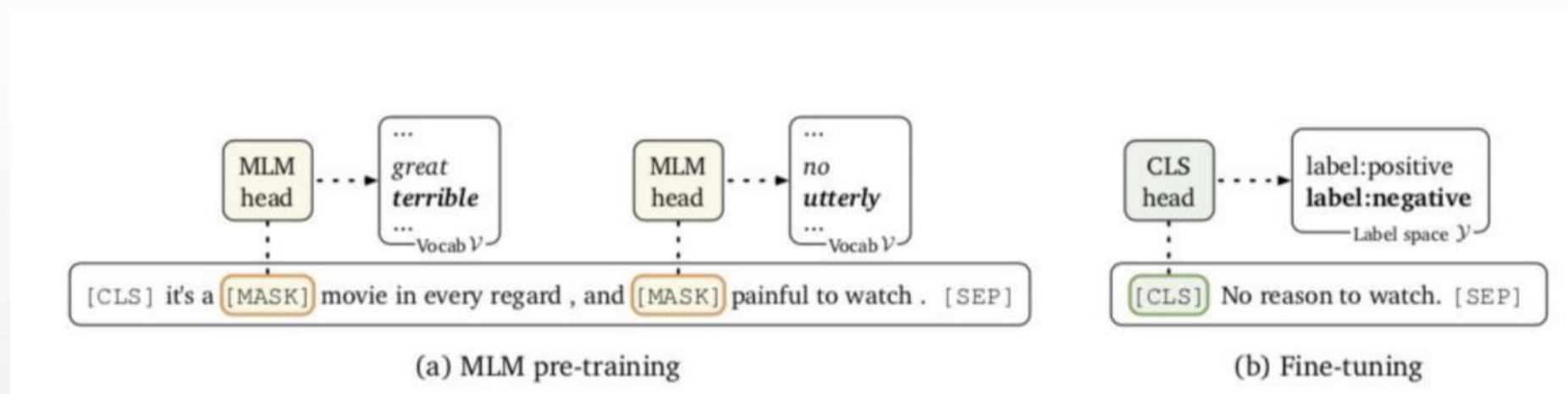
体育 or 娱乐 or 汽车 or 婚恋？

没有梯度更新，错误的预测结果难以优化

领域专用预训练模型训练成本高昂

- fine-tuning

## 预训练-微调范式及其问题










在普通的标准微调方法中，如上图(b)所示，新参数的数量（独立于原始预训练模型外的参数）可能会很大。

例如基于RoBERTa-large的二分类任务会新引入2048个参数，会使从小样本（如32个标注数据）中学习变得困难。

大量的标注数据获取成本高，而且标注质量可能参差不齐

# SuperGlue

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
10	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	57.9	91.0/78.1
+	11	Anuar Sharafudinov	AILabs Team, Transformers	82.6	88.1	91.6/94.8	86.8	85.1/54.7	82.8/79.8	88.9	74.1	78.8	100.0	100.0/100.0
	12	Rakesh Radhakrishnan Menon	ADAPET (ALBERT) - few-shot		76.0	80.0	82.3/92.0	85.4	76.2/35.7	86.1/85.5	75.0	53.5	85.6	-0.4 100.0/50.0
+	13	Timo Schick	iPET (ALBERT) - Few-Shot (32 Examples)		75.4	81.2	79.9/88.8	90.8	74.1/31.7	85.9/85.4	70.8	49.3	88.4	36.2 97.8/57.9
	14	Adrian de Wynter	Bort (Alexa AI)		74.1	83.7	81.9/86.4	89.6	83.7/54.1	49.8/49.0	81.2	70.1	65.8	48.0 96.1/61.5
	15	IBM Research AI	BERT-mtl		73.5	84.8	89.6/94.0	73.8	73.2/30.5	74.6/74.0	84.1	66.2	61.0	29.6 97.8/57.3
	16	Ben Mann	GPT-3 few-shot - OpenAI		71.8	76.4	52.0/75.6	92.0	75.4/30.5	91.1/90.2	69.0	49.4	80.1	21.1 90.4/55.3
	17	SuperGLUE Baselines	BERT++		71.5	79.0	84.8/90.4	73.8	70.0/24.1	72.0/71.3	79.0	69.6	64.4	38.0 99.4/51.4
			BERT		69.0	77.4	75.7/83.6	70.6	70.0/24.1	72.0/71.3	71.7	69.6	64.4	23.0 97.8/51.7

# few-shot learning方法1——PET

- [Exploiting Cloze Questions for Few Shot Text Classification and Natural](#)
- [It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners](#)
- PET — Pattern-Exploiting Training

利用模板将任务转化为完形填空然后finetune MLM

模板被定义为PVP(pattern-verbalizer pairs)

- a pattern  $P$ : 原始输入到完形填空式问题的映射
- a verbalizer  $v$ : 原始输出到完形填空是答案的映射

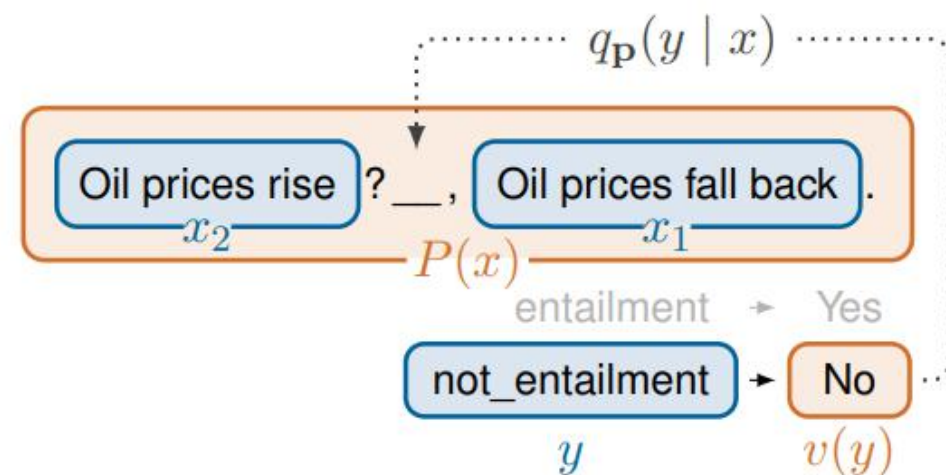


Figure 2: Application of a PVP  $\mathbf{p} = (P, v)$  for recognizing textual entailment: An input  $x = (x_1, x_2)$  is converted into a cloze question  $P(x)$ ;  $q_p(y | x)$  for each  $y$  is derived from the probability of  $v(y)$  being a plausible choice for the masked position.

## 中文例

- 单句分类

目标：预测“这趟北京之旅我感觉很不错”是正面评价还是负面评价

PVP: P: “\_满意。这趟北京之旅我感觉很不错。”

or

“这趟北京之旅我感觉很不错。\_满意。”

V: 正面 → “很”，负面 → “不”

- 句对分类

目标：预测“我们小孩挺挺好,嗯,我妹比我小.”和“我比妹妹年龄大”两句话是“蕴含”、“矛盾”，“中立”中的哪一种

PVP: P: “我们小孩挺挺好,嗯,我妹比我小.\_\_\_,我比妹妹年龄大”

V: 蕴含 → “所以”，矛盾 → “但是”，中立 → “并且”



# 不同PET模板实验对比

数据来源：苏剑林——[必须要GPT3吗？不，BERT的MLM模型也能小样本学习](#)

- P1: \_\_\_\_满意。这趟北京之旅我感觉很不错。

M1: Google开源的中文版BERT Base ([链接](#)) ；
- P2: 这趟北京之旅我感觉很不错。 \_\_\_\_满意。

M2: 哈工大开源的RoBERTa-wwm-ext Base ([链接](#)) ；
- P3: \_\_\_\_好。这趟北京之旅我感觉很不错。

M3: 腾讯UER开源的BERT Base ([链接](#)) ；
- P4: \_\_\_\_理想。这趟北京之旅我感觉很不错。

M4: 腾讯UER开源的BERT Large ([链接](#)) 。
- P5: 感觉如何？ \_\_\_\_满意。这趟北京之旅我感觉很不错。

不同模型不同Pattern的零样本学习效果

	P1	P2	P3	P4	P5
M1	66.94 / 67.60	57.56 / 56.13	58.83 / 59.69	83.70 / 83.33	75.98 / 76.13
M2	85.17 / 84.27	70.63 / 68.69	58.55 / 59.12	81.81 / 82.28	80.25 / 81.62
M3	66.75 / 68.64	50.45 / 50.97	68.97 / 70.11	81.95 / 81.48	61.49 / 62.58
M4	83.56 / 85.08	72.52 / 72.10	76.46 / 77.03	88.25 / 87.45	82.43 / 83.56

# Pet缺陷

- 根据前面的实验结果，不同的模板最好和最坏的结果可能导致10%到20%之间的准确率的偏差
- 如何解决这种人为因素造成的准确率较大的波动？

# few-shot learning方法2——Ptuning

- [GPT Understands, Too](#)

- 使用“伪”提示，不再手动构建模板

消除人为因素影响

- 使用“Prompt Encoder”对“伪”提示进行编码，作者论文选择双向LSTM

做为编码器。（加一个编码器又引入新的参数，那和fine-tuning有啥区别

？作者解释是LSTM和LM比小很多，

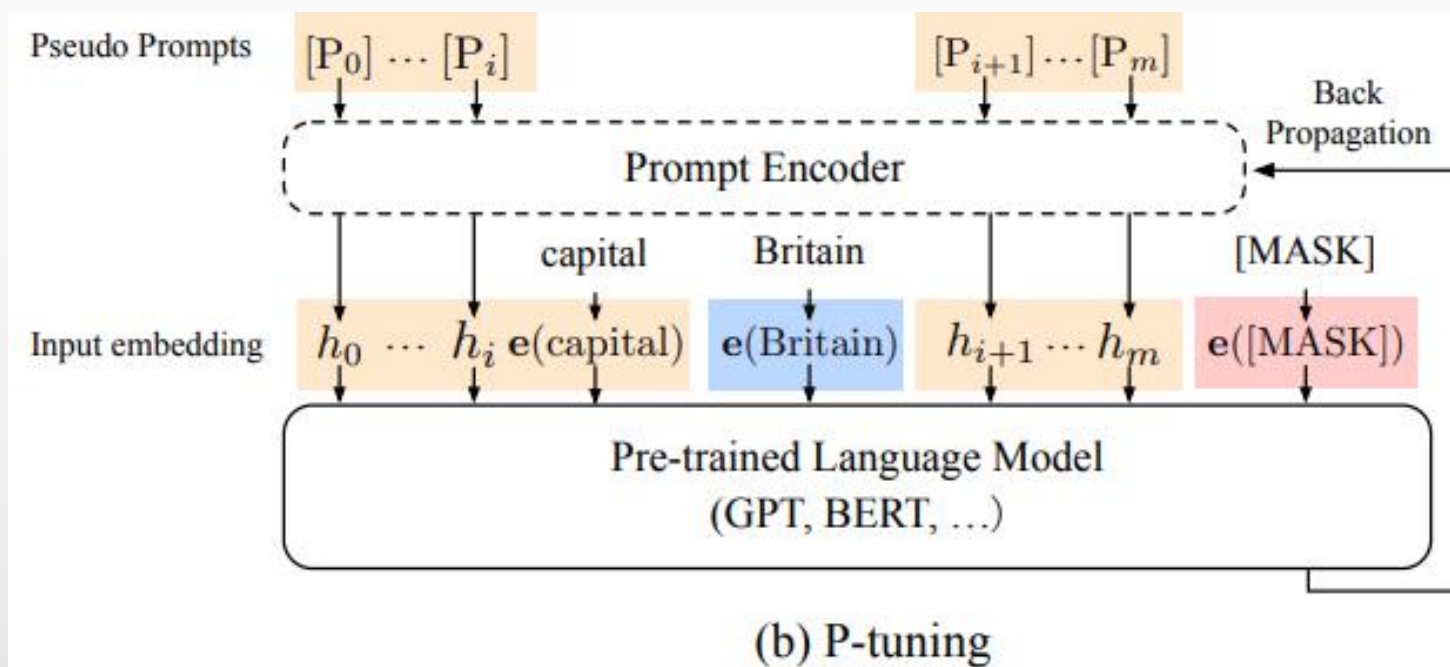
而且推理时不需要LSTM。去掉

LSTM可不可以？）

- finetuning LM，不同于Pet,只优

化几个token

推荐文章：[苏剑林——P-tuning：自动构建模版，释放语言模型潜能](#)



# 实验结果

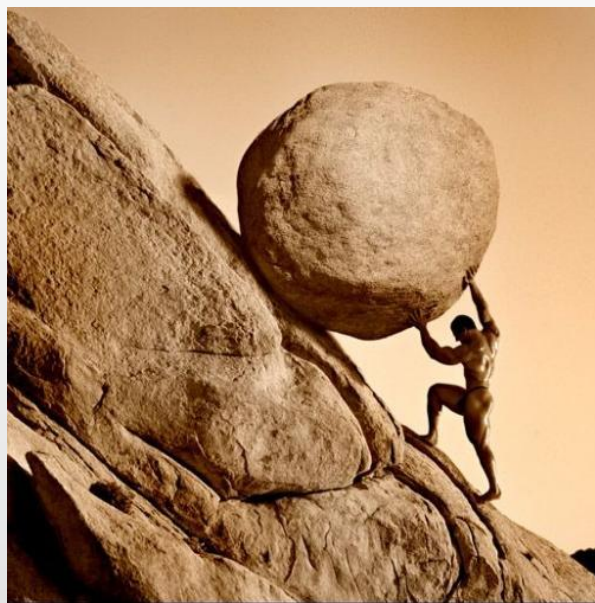
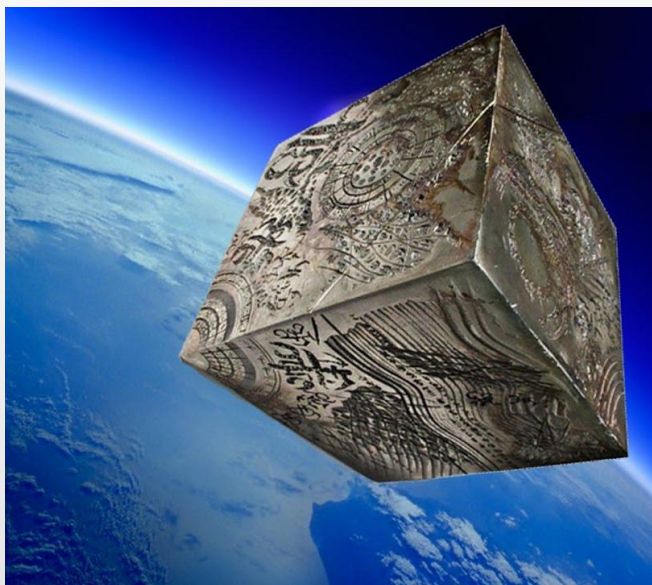
Method	BoolQ (Acc.)	CB (Acc.)	(F1)	WiC (Acc.)	RTE (Acc.)	MultiRC (EM)	(F1a)	WSC (Acc.)	COPA (Acc.)	Avg.
BERT-base-cased (109M)										
Fine-tuning	72.9	85.1	73.9	71.1	68.4	16.2	66.3	63.5	67.0	66.2
MP zero-shot	59.1	41.1	19.4	49.8	54.5	0.4	0.9	62.5	65.0	46.0
MP fine-tuning	73.7	87.5	90.8	67.9	70.4	13.7	62.5	60.6	70.0	67.1
P-tuning	73.9	89.2	92.1	68.8	71.1	14.8	63.3	63.5	72.0	68.4
GPT2-base (117M)										
Fine-tune	71.2	78.6	55.8	65.5	67.8	17.4	65.8	63.0	64.4	63.0
MP zero-shot	61.3	44.6	33.3	54.1	49.5	2.2	23.8	62.5	58.0	48.2
MP fine-tuning	74.8	87.5	88.1	68.0	70.0	23.5	69.7	66.3	78.0	70.2
P-tuning	75.0 (+1.1)	91.1 (+1.9)	93.2 (+1.1)	68.3 (-2.8)	70.8 (-0.3)	23.5 (+7.3)	69.8 (+3.5)	63.5 (+0.0)	76.0 (+4.0)	70.4 (+2.0)

值得一提的是，SuperGlue中的任务ReCoRD数据集作者并没有跑，给出的理由是ReCoRD任务没有提示。



# 为什么Pet、Ptuning行

- 推断任务转换成更符合预训练形式的任务
- 相比于我们熟悉的fine-tuning, Pet\Ptuning 需要训练的额外参数更少，更多的时fine-tuning MLM\LM本身。而我们常用的fine-tuning更多的时候是在fine-tuning特定任务额外加入的参数



可以把语言模型（图一）看作一个立方体，PET\Ptuning的训练方式相当于驱动力量很小（数据量少），直接搬很难搬动这个立方体，所以对这个立方体本身做了一些雕凿让他更圆滑，西西弗斯（图2）才可以推动它。而对于像大黄蜂这样驱动力量足够（数据量大），对模型本身的参数不做很多调整就可以推动它

# FewClue初步实验

- fewclue对clue上的数据做了一些初步实验，在同等标注数据量的情况下，pet\ptuning方法在很多数据集上的准确率都优于以前的finetuning方法

方法	二分类	句对分类	多分类（15类别）
FT	70.2	35.3	30.0
PET	86.0	36.0	51.4
Ptuning	88.0	35.9	44.9

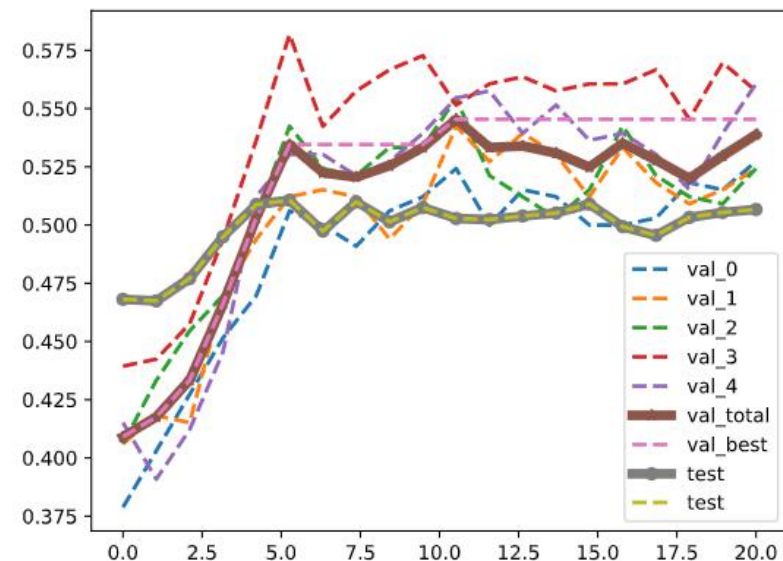
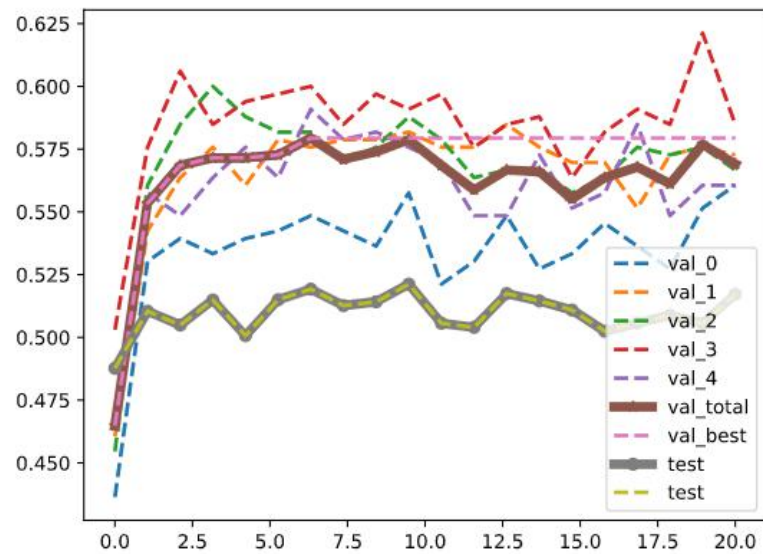
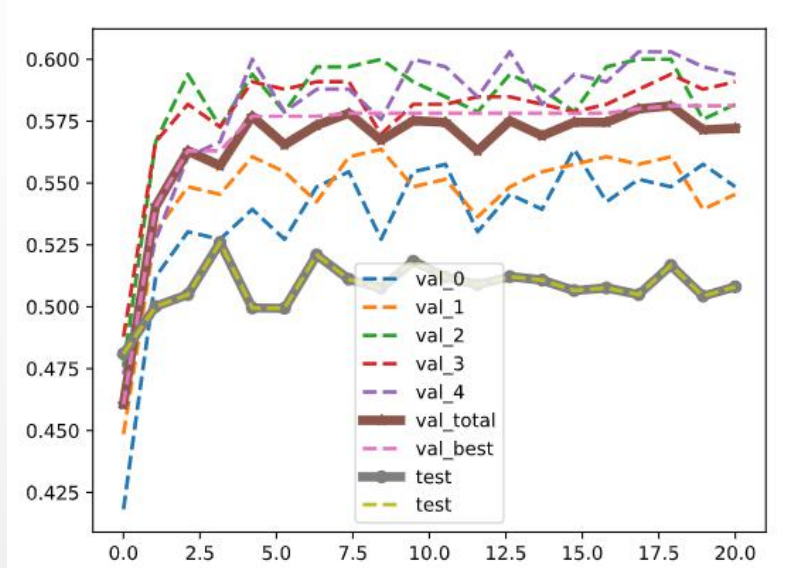
注：模型基于Roberta-L12

# 数据问题反馈

- 调查问卷主要反馈

- 1) 希望任务丰富多样、适应真实场景;
- 2) 数据集数量: 提供9个左右数据集
- 3) 单个任务数量: 按类别采样16为主;
- 4) 半监督学习: 小样本测试还应提供大量无标签数据

# 针对数据问题fewclue团队做的实验



总共有大概15个类别，图一、图二为每个类别20个标注数据的情况，使用了不同的训练集数据，图三为每个类别7个标注数据的情况，绿色的粗线条代表在所有五个验证集上的准确率，红色的粗线条代表测试集准确率，虚线为在五个验证集上各自的准确率，其中图三的训练集是从图一中随机抽取。图二和图一使用不同的训练集。三次实验使用相同的验证集。图一图三训练集和验证集比例为1: 1，图二训练集验证集比例大概为1: 3。一、二、三的最佳验证集准确率在测试集上的准确率分别为50.8%，51.2%，50.0%。

可以看到，数量相同但数据不同的训练集结果差别不大，数据量变为原来的三分之一准确率有所下降。但是和数据量变少相比，更不稳定的因素在于单个验证集的最佳准确率很多时候在测试集上准确率并不好。这对我们在做小样本学习上有新的思考，拿到一个少量标注数据进行Pet 或者Ptuning的时候我们如何去保证拿到最佳的测试集效果？



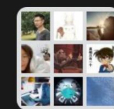
# FewClue: 中文小样本测评

FewCLUE: 结合预训练模型的中文小样本学习, 4月10日测评注册已经开始, 4月30日公布数据集。以CLUE benchmark的分类榜单为基础。

奖励: 测评前三名队伍会获得NLPCC和CCF中国信息技术技术委员会认证的证书; 优胜队伍可以提交该测评任务的论文并投稿到NLPCC

比赛报名: <https://www.cluebenchmarks.com/NLPCC.html> NLPCC2021

官方链接: <http://tcci.ccf.org.cn/conference/2021>



FewCLUE: 中文小样本学习交流  
群



该二维码7天内(5月2日前)有效, 重新进入将更新