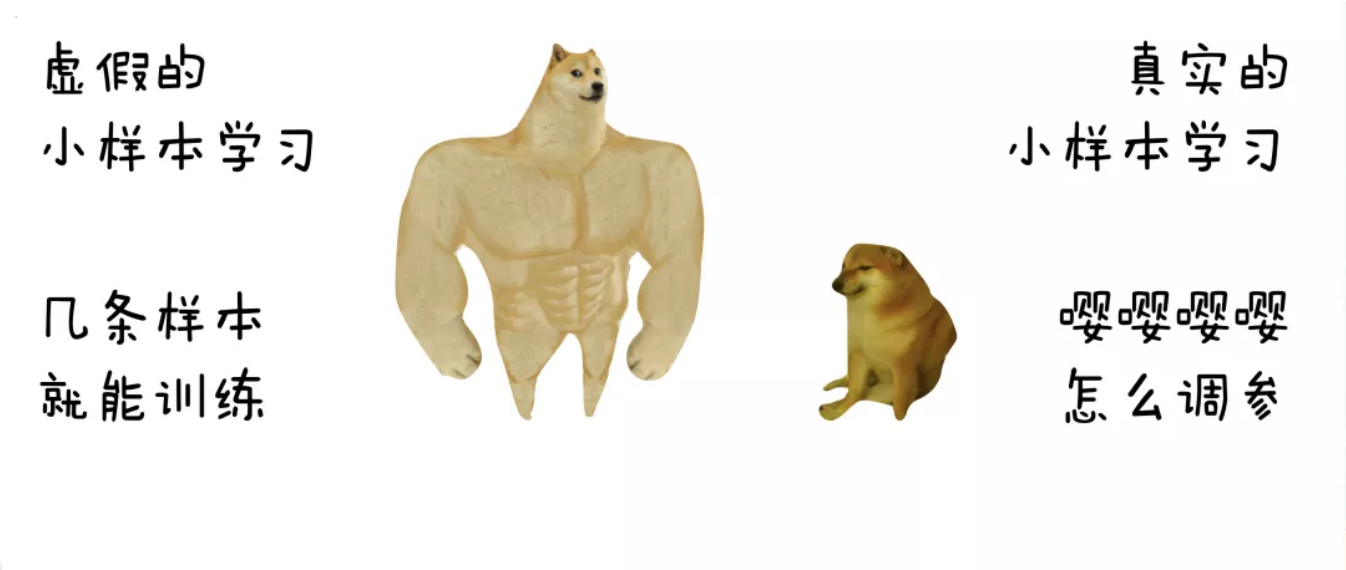
恕我直言，很多小样本学习的工作就是不切实际的

原创 ivem 夕小瑶的卖萌屋 2021-06-16 12:05



文 | ivem
编 | 小轶

以前的小样本学习（Few-shot Learning），是需要用一个巨大的训练集训练的。测试时只给出 n-way k-shot，在这 N * k 个样本上学习并预测。我第一次看到这种任务设定的时候真是非常失望：这和现实情况的需求也相差太远了！真实场景下的小样本学习，哪有大量的训练数据呢？

从 GPT3 开始，学术界开启了一个新的小样本风潮。借助预训练模型，人们只给出几条或几十条样本作为训练集，用小小训练集进行 finetune。看到这些工作，我觉得这样才是真正的小样本学习！

最近有一些工作也在这种任务设定下取得了不错的进展。所谓prompt，就是结合具体场景，设计新的 finetune 任务形式，从而将与当前任务相关的提示信息（prompt）引入模型，以此更好地利用预训练模型的结构与先验知识。我们大名鼎鼎的 GPT 系列就是这么干的。比如我们拿 GPT3 做 QA 的 finetune，直接喂给他一串“Question：问题内容 Answer：”，剩下的答案部分就让 GPT3 自己填完。

p. Question: q? Answer: <MASK>.

卖萌屋之前还推送过其中一个工作（刚刚被评为 NAACL 的最佳短文！详见这里）。这篇工作表明，基于 prompt 的方法能在几分之一的训练数据下，达到传统 finetune 的训练结果。

但！是！这样的任务设定就是真正的小样本学习了吗？今天这篇 NYU、facebook、CIFAR 三巨头一起带来的文章直接 **打脸了所有人**：以上任务设定也还不是真正的小样本学习！由于给出了一个巨大的验证集，因此人们用这个验证集挑选最好的 prompt、用它调参，这也是不切合实际的！**真正的小样本学习，训练集验证集都要小！**

另外，本文还在真正的小样本学习任务设定下，评测了挑选 prompt、调参的效果，实验发现，我们对模型小样本学习的能力还是过于乐观了🤦

论文题目:

True Few-Shot Learning with Language Models

论文链接:

http://arxiv-download.xixiaoyao.cn/pdf/2105.11447v1.pdf

代码地址:

https://github.com/ethanjperez/true_few_shot

Arxiv访问慢的小伙伴也可以在 **【夕小瑶的卖萌屋】**订阅号后台回复关键词 **【0616】** 下载论文PDF~

真正的小样本学习

可能大家被我上面说的各种“小样本学习”的情景搞晕了，为了清楚，我们可以总结成这样的一张表：

Learning Scenario	Many Train Distributions	Many Train Examples	Many Val Examples
Data-Rich Supervised	✗	✓	✓
Multi-Dist. Few-Shot	✓	✗	✗
Tuned Few-Shot	✗	✗	✓
True Few-Shot	✗	✗	✗

表中列举了四种情况：

1. Data-Rich Supervised 表示传统有大量数据的有监督学习。
2. Multi-Distribution Few-Shot 表示原始的小样本学习情景，即在大量 n-way k-shot 上进行训练。由于每个 task 都包含不同的数据分布，因此这相当于在不同的分布中训练，在新的分布中使用模型。
3. Tuned Few-Shot 表示从 GPT3 开始的，用 prompt 的方式对预训练模型微调。
4. True Few-Shot 就是本文提出的啦！

本文认为，**对于小样本学习，既不应该有其它分布的数据辅助、也不应该有很多训练数据，更不应该有很多验证集的数据**。因为这些数据全都是需要标注的！

那还能调参嘛？

界定了真正的小样本学习，作者就想：之前那些 prompt 的方法用了大量验证集信息来调整超参、选择最好的 prompt。他们对性能的提升其实都来自验证集中蕴含的信息。那么，**在没有验证集的情况下（对！作者为了更好的比较，就只留少量样本的训练集），该怎么调参呢**？作者给了两个方法：

1. **k 折交叉验证**：将数据集分为 k 个部分，用其中 k-1 个部分作为训练集，剩下的一个部分作为验证集。在后面的实验中，这种方法被称作 CV（cross validation）。

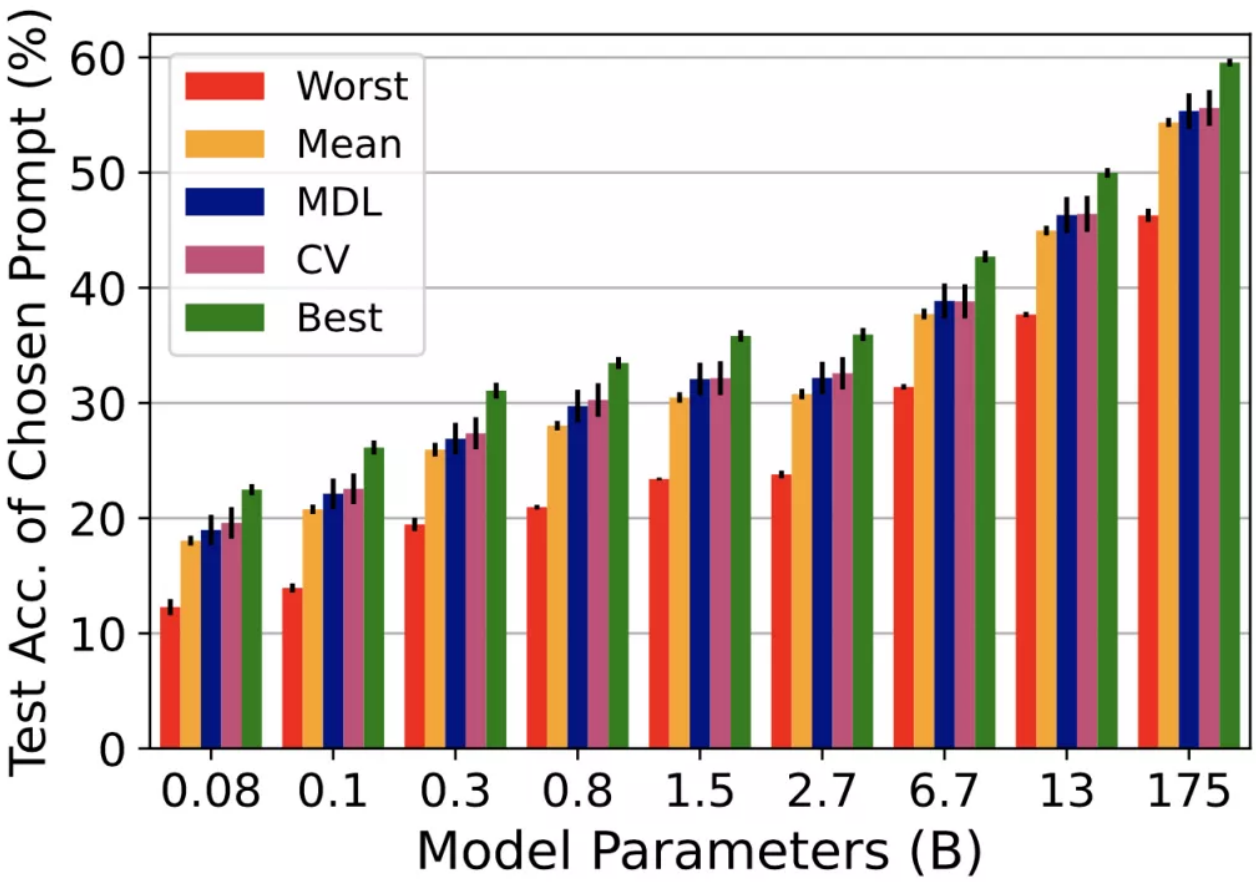
2. **类似在线学习的交叉验证**：将数据集分为 k 个部分，第 1 轮用第 1 部分训练，第 2 部分验证，第 i 轮用前 i 部分训练，第 i+1 部分验证。在后面的实验中，这种方法被称作 MDL（minimum description lengthm），因为其本质上遵循的是最小描述长度准则。

另外，作者还给出一个交叉验证的准则：即在训练和验证集之间，样本 loss 的差距要尽可能小。

实验和分析

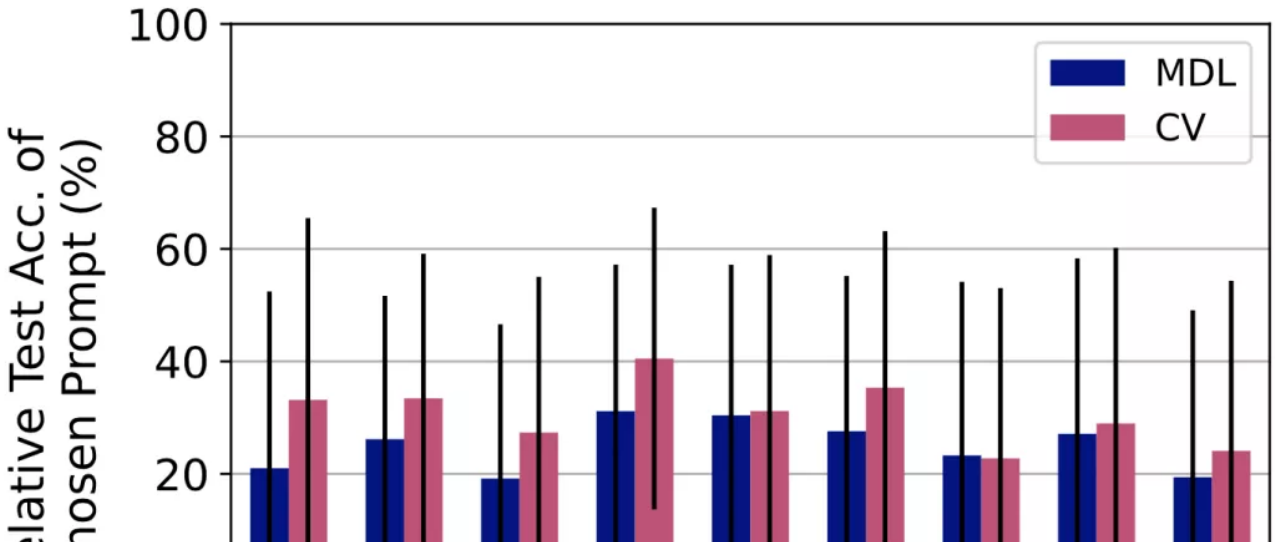
作者使用以上两种交叉验证方法，在 LAMA[1] 数据集上，对基于 prompt 的模型[2]进行了实验。LAMA 是一个评测语言模型的数据集，它给出一句话，让语言模型提取这句话在知识图谱中对应的三元组。

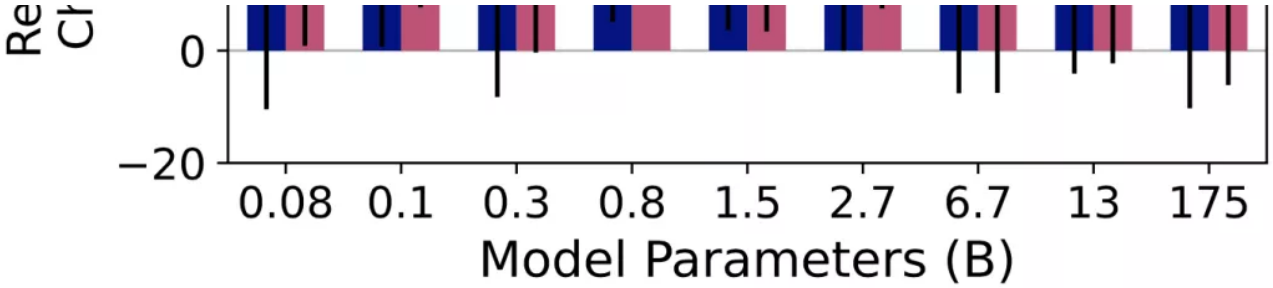
首先，是对不同 prompt 的对比：



实验发现，无论是在多大参数量的模型上，基于两种方法选择 prompt（图中蓝色粉色），都要比随机挑选 prompt （图中黄色）的效果好，但选出的 prompt 效果还是远不如最好的 prompt（图中绿色）。

如果把随机选择 prompt 作为基线，最好的 prompt 作为上界，那么两种交叉验证带来的性能提升便如下图所示：



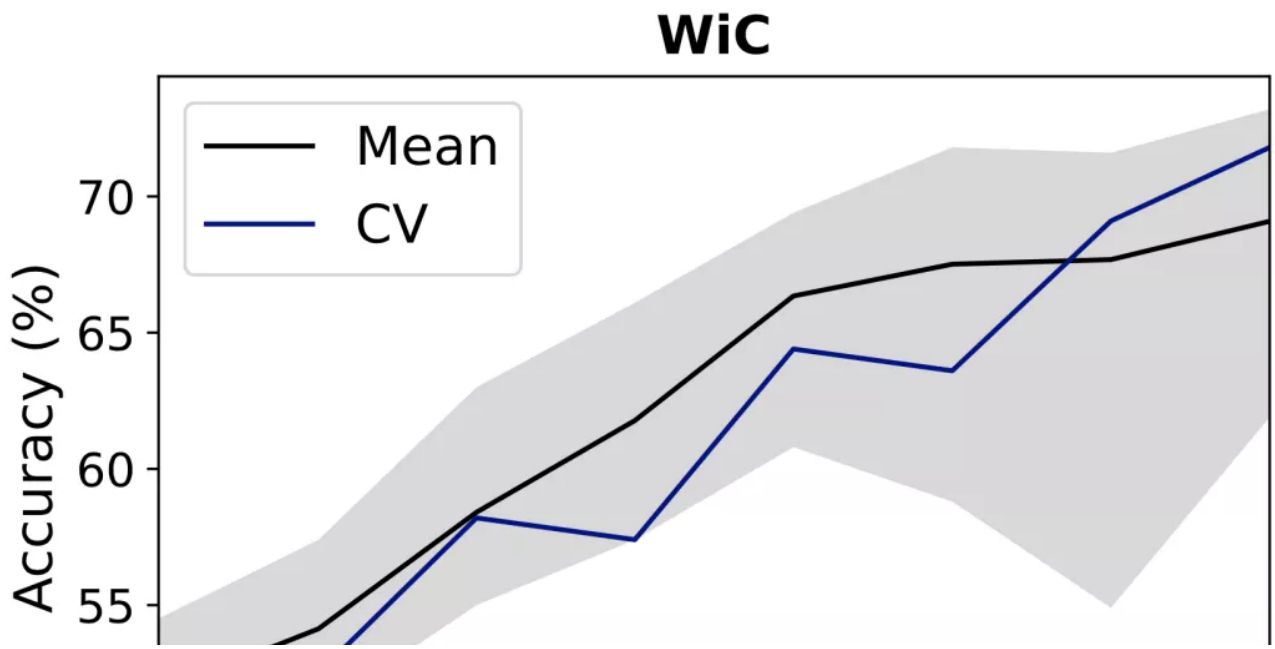


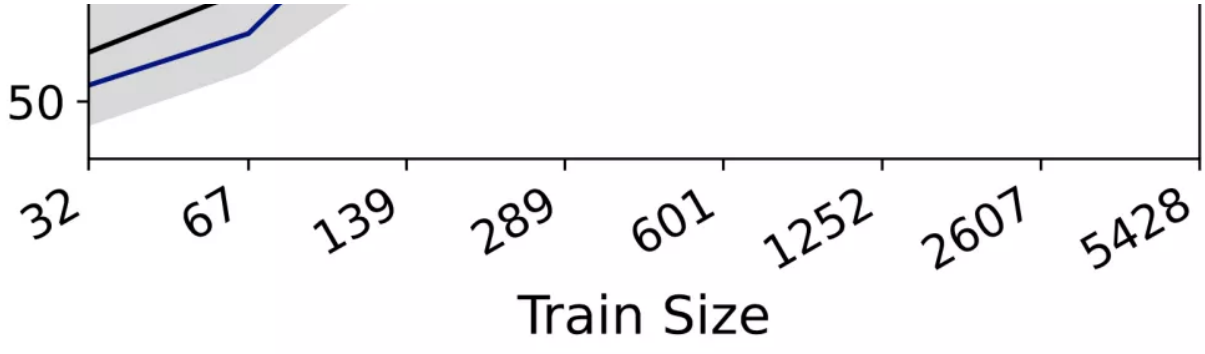
在理想的验证集里，我们是能挑选出最好的 prompt 的，因此最好的 prompt 就代表了在理想的巨量验证集中挑选 prompt 最好的结果。从上图可以看出，在没有验证集时，作者提出的两个交叉验证方法只能带来理想验证集带来的大约 25% 的性能增益。因此，没了大量数据作为验证集，的确也就不能有很好的交叉验证效果了。

另外，作者还对“在验证集上调参”这件事进行了实验。模型中有两个参数是需要调整的，一个是 epoch 数量，另一个是输入文本中被 mask 掉的 token 的比例。这里的评测使用 SuperGLUE 的任务，其中包含文本蕴含、阅读理解等等和理解相关的任务。实验结果如下图所示：

	BoolQ Acc	CB Acc/F1	COPA Acc	RTE Acc	WiC Acc	WSC Acc	MultiRC EM/F1	ReCoRD EM/F1	Avg
Worst	$75.0_{4.8}$	$79.5_{2.3}/67.3_{7.8}$	$76.8_{2.2}$	$63.2_{4.0}$	$49.0_{1.3}$	$77.2_{1.8}$	$38.5_{7.4}/80.0_{2.9}$	$76.2_{1.8}/86.5_{1.2}$	$69.4_{1.5}$
Mean	$79.0_{1.5}$	$85.9_{2.3}/74.5_{11.0}$	$81.1_{2.9}$	$70.8_{2.5}$	$51.5_{1.8}$	$82.5_{2.7}$	$44.2_{6.6}/82.3_{2.7}$	$78.3_{1.3}/87.8_{0.8}$	$73.9_{1.2}$
MDL	$76.5_{5.8}$	$85.7_{5.6}/74.8_{13.4}$	$82.0_{2.9}$	$70.4_{8.5}$	$52.2_{3.0}$	$82.0_{3.1}$	$39.7_{8.1}/80.6_{3.2}$	$78.9_{0.7}/88.2_{0.4}$	$73.4_{2.8}$
CV	$78.9_{2.4}$	$83.9_{5.3}/69.2_{10.3}$	$80.5_{3.3}$	$68.7_{7.0}$	$51.1_{1.6}$	$83.1_{2.6}$	$41.9_{7.2}/81.4_{3.1}$	$78.7_{1.6}/88.1_{1.0}$	$73.0_{2.1}$
Best	$80.9_{1.0}$	$89.8_{3.1}/79.8_{13.4}$	$84.8_{4.5}$	$76.7_{1.8}$	$54.1_{2.3}$	$86.6_{1.8}$	$46.8_{6.9}/83.4_{2.9}$	$80.4_{1.1}/89.2_{0.7}$	$77.2_{0.9}$

这里发现，用两种交叉验证在小验证集上调参，其结果和随机参数差不多，甚至总体上看还更差一点！甚至在 MultiRC 上，调参出来的结果与最坏的一组参数表现差不多，表明在小验证集上调参，并不一定就能稳定提升性能。这结果太让人失望了，不过作者不死心，还进行了一个有意思的实验：





有多少数据之后，才一定能通过调参，得到一组比随机更好的参数呢？上面这张图是在 WiC 任务上，使用 k 折交叉验证来调参，横轴代表总的训练样本数量，纵轴是模型性能，灰色的区域是 16 组不同参数的模型性能区间。实验发现，到了 2000 多个样本时，调参才是确定有效的！

总结

这篇文章表明，**在真正的小样本情境下，模型选择做的还不太好**。为此，作者对未来的小样本学习给出了以下建议：

- 在写文章的时候，同时**注明模型选择的原则**，以及所有超参数和尝试的 prompts。
- 将**验证集的数量**也归入小样本学习的“数据量”里。
- 当有大量样本作为验证集的时候，先不要用！先在测试集直接得到结果、做消融实验，**等所有试验完成后，最后再引入验证集**。这样避免实验结果使用验证集大量样本的信息。
- **不要使用前人工作中的超参数**，只在这少量样本中重新调参。

最严格的一种方式是，在设计评测任务时，只给出小小的训练集和小小的验证集，真正评分的测试集不给出，只能在线评测。

这篇文章说了真正的小样本学习，自然地，就延伸出来一个问题：在零样本学习（Zero-shot Learning）的情境下，还能进行调参吗？还能挑选模型吗？

个人感觉，似乎不行了。

萌屋作者： ivem

在北大读研，目前做信息抽取，对低资源、图网络都非常感兴趣。希望大家在卖萌屋玩得开心 ヾ(=･ω･=)o

作品推荐

1. 老板让我用少量样本 finetune 模型，我还有救吗？急急急，在线等！
2. 谷歌：CNN击败Transformer，有望成为预训练界新霸主！LeCun却沉默了...
3. 中文BERT上分新技巧，多粒度信息来帮忙

参考文献

[1].Fabio Petroni, et al., "Language models as knowledge bases?", EMNLP 2019, http://arxiv-download.xixiaoyao.cn/pdf/1909.01066v2.pdf

[2].Derek Tam, et al., "Improving and simplifying pattern exploiting training.", http://arxiv-download.xixiaoyao.cn/pdf/2103.11955.pdf