解释对校准黑盒模型有用吗？

希叶格雷格杜雷特
计算机科学系
德克萨斯大学奥斯汀分校
{xiye, gdurrett}@cs .utexas .edu

摘要

NLP从业者通常想要采取现有的
经过训练过的模型，并将其应用于来
新域。虽然微调或少镜头
学习可以用来适应一个基本模型，
制造这些技术并没有一个单一的配方
尼克斯的工作；此外，一个人可能没
这是一个原始模型的权重，如果它是
被画成一个黑盒子。我们研究如何实
证明了一个黑盒模型的性能
新领域通过利用的解释
模型的行为。我们的方法首先提取
一组结合了人类直觉的特征
关于具有模型属性的任务
通过黑盒解释技术，
然后使用一个简单的校准器，其形式
一个分类器，以预测是否基本模型
是否正确。我们用我们的
方法对两个任务，提取的问题和-
回答和自然语言推理，涵盖-
从几对领域中进行适应
目标域数据有限。扩展器-
跨所有域对的实验结果
表明解释对校准是有用的
使用这些模型，提高了精度
判决不必返回到每一个
样例我们进一步证明了校准
模型在任务之间有一定程度的转移。[1]

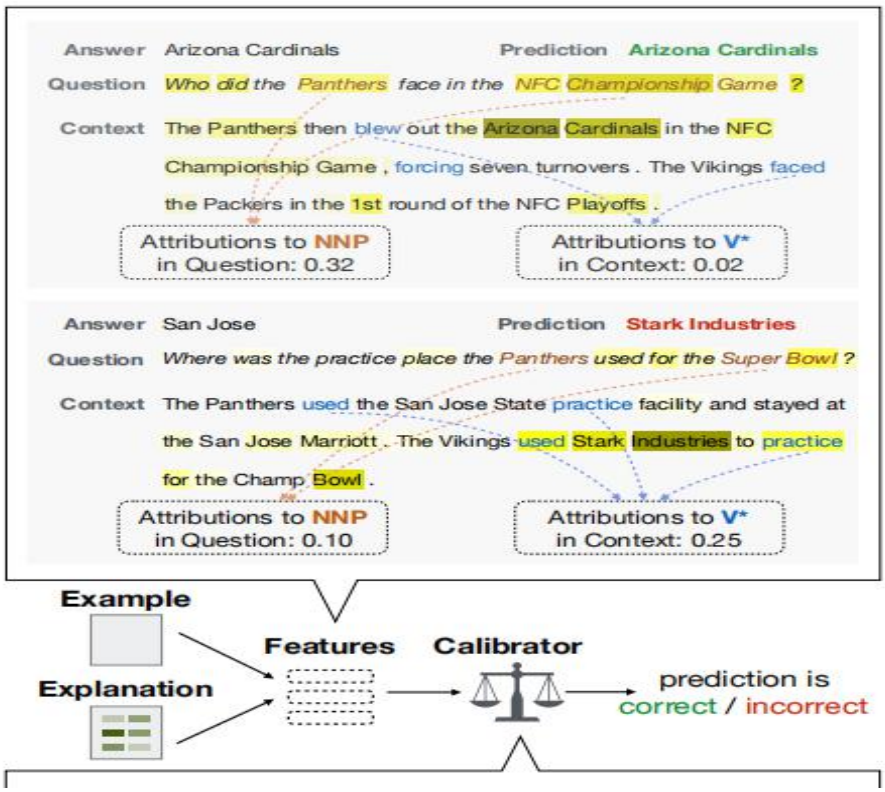

图1：来自小队-ADV数据集的校准器管道和示例。经过阵容训练的罗伯塔模型在第一个例子上是正确的，但在第二个例子上是错误的。检查由LIME产生的归因值的特征可以根据问题和V中对NNP的归因来区分这两者*在上下文中。使用这些特征的校准器可以预测原始模型是对是错。

1介绍

随着最近在培训前方面的突破，NLP
模型显示出越来越有前景
在现实世界的任务上，导致他们的部署-
在翻译，情感分析，
和问题回答。这些模型是一些
被用作黑匣子的时间，特别是如果它们是
只能通过api作为服务提供[2]或者如果结束
用户 做 不 有资源来微调吗
模型本身。当用户试图在一个偏离训练域的新域上部署模型时，这种黑箱特性带来了挑战，通常会导致性能下降。

我们研究了黑箱模型的领域自适应任务：给定一个黑箱模型和一个来自新领域的少量例子，我们如何提高模型在新领域上的泛化性能？在此设置中，请注意，我们无法更新模型参数，

[1] 可用的代码： https://github。com/xiye17/间卡尺

[2] 谷歌翻译，透视图API https://perspectiveapi.和猴子学习https://monkeylearn.com/单键学习-api/是三个例子

这使得迁移和少镜头学习技术不适用。*然而，我们仍然可以通过学习一个校准器或一个单独的模型来做出黑盒模型对给定实例的二元决策，使模型在实践中更有效。*虽然没有完全解决领域适应问题，但校准模型可以使其在实践中更有用，因为我们可以认识到它何时可能犯错误(Guo et al。，2017年；Kamath等人。，2020年；德赛和德雷特，2020年)，并相应地修改我们的部署战略。

本文探讨了解释如何帮助解决这一任务。我们利用了黑盒特征归因技术(Ribeiro等人。，2016；伦德伯格和李，2017)来识别模型正在利用的关键输入特征，即使没有访问模型的内部表示。如图1所示，我们通过将模型解释与手工制作的启发式方法连接起来来执行校准，以提取一组描述模型的“推理”的特征。对于图中所示的问题回答设置，当强烈考虑问题中一组特定标签（如专有名词）的标记时，答案就会更加可靠。我们提取了一组描述不同标签的属性值的特征。使用目标域的少量例子，我们可以为黑盒模型训练一个简单的校准器。

我们的方法与最近关于模型行为和解释的工作线密切相关。钱德拉斯卡兰等人。（2018）；Hase和Bansal（2020）表明，解释可以帮助用户在某些方面预测模型决策，Ye等人。（2021）展示了如何将这些解释半自动地连接到基于手工制作的启发式的模型行为上。我们的方法更进一步，通过使用一个模型来学习这些启发式方法，而不是手工制作它们或让人类检查这些解释。

我们测试了我们的方法是否可以提高模型的泛化性能：提取式问题回答（QA）和自然语言推理（NLI）。我们为两个任务中的5对源域和目标域构建泛化设置。与现有的基线相比(Kamath等人。我们发现，解释确实有助于这项任务，成功地提高了所有成对的校准器性能。我们甚至发现了基于解释的校准器优于微调模型的设置对目标域数据，它假设玻璃盒可以访问模型的参数。我们的分析进一步证明了校准器模型本身的泛化：我们在一个领域上训练的校准器在某些情况下可以转移到另一个新领域。此外，我们的校准器也可以大大提高模型在选择性QA设置中的性能。

2.使用对黑箱模型校准的说明

让x = x1，x2，...，xn是一组输入令牌，yˆ = f (x)是我们正在考虑的黑箱模型的预测。**我们在校准3中的任务是评估模型对x的预测是否匹配其地面真相y。**我们用变量t，i来表示它。e.，t 全 ✶{f (x) = y}.

我们探索了各种校准器模型来执行这项任务，我们的主要重点是利用特征归因形式的解释的校准器模型。ϕ具体来说，对输入x的解释分配了一个归因分数 ϕi对于每个输入令牌西，这代表了该令牌的重要性。ϕ接下来，我们根据输入和解释提取特征u（x，），并使用这些特征来学习一个校准器c：u（x，）！ϕT用于预测一个预测是否有效。为了回答论文标题中提出的核心问题，我们与不使用解释的基线进行了比较。

我们的评估集中在二进制校准上，或对模型的初始预测是否正确进行分类。根据最近在此背景下的工作(Kamath等人。，我们特别关注模型经常出现错误的域转移设置。一个好的校准器可以识别模型可能犯了错误的实例，因此我们可以向用户返回一个空响应，而不是一个不正确的响应。

在本节的其余部分中，我们将首先介绍如何生成解释，然后如何提取输入x的特征u。

.12生成说明

由于我们正在校准黑盒模型，我们采用LIME (Ribeiro等。和SHAP (Lund-

[3]我们跟随Kamath等人。（2020年）将校准定为二进制分类任务。设计一个好的分类器与精确估计后验概率的目标有关，而校准在历史上更多地提到了(Guo等。，但我们的评估侧重于二值精度，而不是实值概率。

berg和Lee，2017），为模型生成解释，而不是其他需要访问模型细节的技术(例如，集成梯度(孙达拉拉詹等人。，2017)).

该工作的其余部分只依赖于LIME和SHAP来将输入序列x和模型预测y映射到一组重要权值。ϕ我们将简要总结两种方法共享的统一框架，并向读者参考各自的论文了解更多细节。

LIME和SHAP通过近似于模型对基数据点x周围的一组扰动的预测来生成局部解释。在这种情况下，一个扰动x\关于x，这是一个简化的输入，其中一些输入令牌不存在（用<掩码>令牌替换）。让z = z1，z2，...，zn是一个二进制向量子指示是否西是否存在（使用值1）或不存在（使用值0），和h北(z)是将z映射回到简化的输入x的函数\.两种方法都寻求在z上学习局部线性分类器g，通过最小化与原始模型f的预测相匹配：

$$g(z) = \phi_0 + \sum_{i=1}^{n} \phi_i z_i$$

ξ= arg mi^{n} 工$_{\pi北}$(z)[f (h北(z)) — g(z)]^{2}+ (g)

z2Z

在哪里π北是为每个扰动z分配权重的局部核，并且是模型复杂度上的12正则化器。学习到的特征权重iϕ子然后表示每个单独的令牌的加性归因（Lundberg和Lee，2017）西.LIME和SHAP在本地内核的选择上有所不同π北.有关详细的内核内容，请参阅补充资料。

.22.通过结合解释和启发式来提取特征

有了这些解释，我们现在希望将这些解释与我们从任务中期望的推理联系起来：如果模型如我们预期的那样运行，它可能会更好地校准。人类可能会观察一些重要特征的属性，并以类似的方式决定该模型是否可信（Doshi-Velez和Kim，2017）。过去的工作已经探索了这种技术来比较解释技术(Ye等人。或与人类用户进行这项任务的研究(钱德拉斯卡兰等人。，2018年；哈和Bansal，2020年)。

我们的方法通过学习-来自动化这个过程关于解释的什么属性是重要的。我们首先分配每个令牌西具有一个或多个人类可以理解的特性V（xi）= {vj}$_{j=1}^{m_i}$.每个属性vj2V是属性空间中的一个元素，它包括像POS标签这样的指标，用于描述的一个方面西其重要性可能与模型的鲁棒性相关。我们将这些特性与解释的各个方面结合起来，以呈现我们的校准判断图1显示了诸如标记是否为专有名词（NNP）等属性的例子。

我们现在为对x的预测构造特征集。ϕ对于每个属性v 2 V，我们通过聚合与v相关的标记的属性来提取单个特征F（v，x，）：

$$F(v, x, \phi) = \sum_{i=1}^{n} \sum_{\bar{v} \in V(x_i)} \mathbb{1}\{\bar{v} = v\}\phi_i$$

其中，✶是指示器函数，和ϕi是属性值。当模型对x进行预测时，一个单个特征表示关于属性v的总归因。ϕx的完整特征集u，给定为u = {F（v，x，）}v2V，从=中的属性的角度总结了模型的基本原理。

我们使用几种类型的启发式属性来校准QA和NLI模型。

输入部分（QA和NLI）：在我们的两个任务中，输入序列可以自然地分解为两部分，即问题和上下文（QA）或前提和假设（NLI）。我们用相应的段名称来分配每个标记，从而产生像这样的特性

问题的属性。

POS标签（QA和NLI）：我们使用来自英国宾夕法尼亚州树库的标签(Marcus等人。来实现一组属性。我们假设一些特定标签的标记应该更重要，比如QA任务问题中的专有名词。如果一个模型没有考虑一个QA对的专有名词，它就更有可能做出错误的预测。

重叠词（NLI）：前提和假设之间的Word重叠强烈影响神经模型的预测(McCoy等。，2019). 如果前提和假设中都出现一个标记，我们为每个标记分配重叠属性，否则不重叠。

群的结合：我们可以通过取分数-来进一步产生更高层次的性质

两个或两个以上组的自氏积。我们将片段和pos标签结合起来，它们产生更高级的特性，如NNP的属性。这种特征聚集了用NNP标记的标记的属性，也需要在问题中（用橙色标记）。

.32校准器型号

我们在少量的样品上训练校准器
人们在我们的目标领域中工作。每个样本都使用原始模型的预测与地面真相进行比较。ϕ使用我们的特征集F（v，x，），我们学习了一个随机森林分类器，在Kamath等人中显示对类似的数据有限设置是有效的。（2020年），以预测t(预测是否正确）。该分类器返回一个分数，它覆盖了模型相对于该预测的原始置信度分数。

在第4节中，我们将讨论我们的方法的几个基线。当我们改变模型所使用的特征时，分类器和设置的所有其他细节都保持不变。

3个任务和数据集

我们的任务设置涉及到从源域/任务a转移到目标域/任务B。图2显示了我们所操作的数据条件。我们的主要实验集中于使用我们的特性来校准黑盒设置中，或选择性地回答（图2中的右侧）。在这种设置中，我们有一个在源域a上训练的黑箱模型和来自目标域B的少量数据。我们的任务是使用来自域B的数据来训练一个校准器，以识别模型在大型的未知测试中可能失败的实例
域B中的数据。我们对比一下这个黑盒子的套装-使用玻璃箱设置（图2中的左侧），我们可以直接访问模型参数，并可以对域B进行微调或从头开始对域B进行训练。

我们尝试从小队转移域名(Rajpurkar等。到三种不同的设置：adv小队（Jia和Liang，2017），HOTPOTQA（Yang等人。和琐事阿卡(Joshi等人。，2017).

小队-ADV是一种基于小队的对抗性设置，它通过在每个例子的上下文结束时添加一个干扰物句子来构建基于小队的对抗性QA例子。附加的句子包含一个虚假的答案，通常与问题有很高的表面重叠

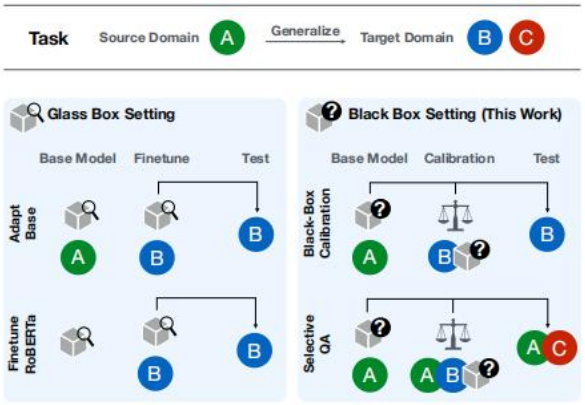

图2：实验中不同设置的说明。在黑盒设置中，训练校准器以提高OOD数据的模型性能；在玻璃框设置，模型微调来自基础模型或香草罗伯塔LM模型的OOD数据。

以欺骗模型。我们使用了贾庆林和梁（2017）提供的附加设置。

与小队类似，霍波特卡也包含了从维基百科中提取的段落，但霍波特卡提出的问题需要多个推理步骤，尽管不是所有的问题都需要（陈和德雷特，2019年）。弗里斯卡是从网络片段中收集的，这些片段呈现的问题和段落的分布与小队不同。对于霍波特卡和比里斯卡，我们直接使用来自MRQA共享任务的数据集的预处理版本(Fisch等人。，2019).

英语NLI对于NLI的任务，我们转移了一个用MNLI训练的模型(威廉姆斯等人。到MRPC（Dolan和布罗克特，2005年）和QNLI（Wang等人。，类似于Ma等人的设置。(2019).QNLI包含一个来自小队的问题和上下文句子对，其任务是验证一个句子是否包含成对问题的答案。MRPC是一个意译检测数据集，提供了一个二进制分类任务，以决定两个句子是否是彼此的意译。请注意，从MNLI到QNLI或MRPC的泛化不仅引入了输入文本分布的变化，而且还包括任务本身的性质，因为QNLI和MRPC并不是严格的NLI任务，尽管有一些相似性。*两者都是二元分类任务，而不是三方分类任务。*

4实验

我们将我们的校准器与现有的基线以及我们自己的消融进行比较。

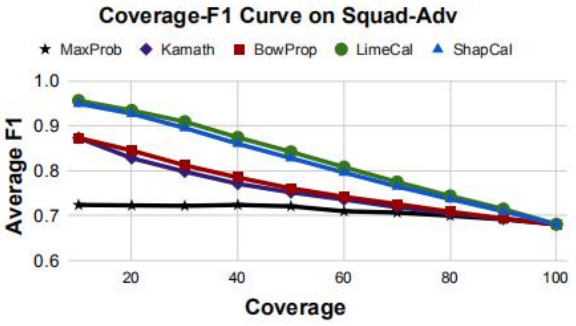

图3：不同方法对adv的覆盖f1曲线。随着更多的低自信问题被回答，F1的平均分数下降。我们使用AUC来评估校准性能。

MAXPROB简单地使用预测类的阈值概率来评估预测是否可信。

KAMATH（Kamath等。（仅针对QA）是最初提出的用来在选择性QA设置中区分分布外数据点和域内数据点的基线（见第5节），但它也可以应用于我们的设置。它训练一个随机森林分类器来学习一个模型基于几个启发式特征的预测是否正确，包括前5个预测的概率、上下文的长度和预测答案的长度。由于我们正在校准黑盒模型，所以我们在Kamath等人的研究中没有使用基于退出的特征。(2020).

克隆（仅适用于NLI）使用比最大克隆更详细的信息：它使用隐含、矛盾和中性的预测概率作为训练校准器的特征，而不是只使用最大概率。

BOWPROP在KAMATH方法之上添加了一组启发式属性特性。*这些与排除解释外的完整模型所使用的特征相同。*我们使用这个基线来给出一个在没有与解释配对的输入上使用一般的“形状”特征的基线。

我们将由LIME和SHAP提出的基于解释的校准方法分别命名为界限和形状。我们注意到这些方法也利用了弓道具中的字袋特征。对于QA，属性空间是低级段和段的位置标签的并集。×××对于NLI，我们使用段和段pos标签重叠词的并集标记令牌。特性的详细数量可以在附录中找到。

4.1主要结果：质量保证

我们训练一个罗伯塔(Liu等。以阵容为基础模型
达到85。5个精确匹配，92.2 F1分。为了在HOTPOTQA和dritiaqa的实验中，我们分割了开发集，抽取500个示例进行训练，剩下的留给测试。[4]对于小队-ADV的实验，我们删除了ADD-SENT设置中未修改的数据点，并使用了500个例子进行训练。对于所有对的实验，我们随机生成分割，对方法测试20次，并将结果平均，以减轻随机性的影响。

*除了测量校准器精度的校准精度（ACC）外，我们还使用f1曲线覆盖下面积（AUC）来评估QA任务的校准性能。*覆盖率f1曲线（图3）绘制了当模型只选择回答按校准器产生的置信度排序的例子的不同分数（覆盖率）时，模型的平均f1得分。一个更好的校准器应该给模型确定的问题分配更高的分数，从而增加曲线下的面积；请注意，AUC为100是不可能的，因为当每个问题都被回答时，f1总是受到基础模型的限制。此外，我们还报告了在回答前25%、50%和75%的问题时的平均分数，以便进行更直观的性能比较。

结果表1总结了QA的结果。
首先，我们证明了解释有助于校准域外的黑盒QA模型。与KAMATH相比，我们的方法使用LIME，大大提高了校准AUC7。1，2.adv和霍波特卡分别是1和1.4。特别是，LIMECAL在adv上覆盖25%，F1平均得分为92.3，覆盖率为25%，接近基础模型在原始阵容上的表现。我们的基于解释的方法在识别对抗攻击的例子方面是有效的。

比较软甲和弓道具，我们发现解释本身确实有帮助。在ADV和霍波特卡，弓

[4]关于超参数的详细信息可以在附录中找到。

方法	ΔΔAcc弓	AUC弓	F1@25弓Δ			班-ADV	F1@50	Δ弓形	F1@75	Δ弓形
MAXPROB	62.6	—	70.9	—	72.4	—	72.1	—	70.4	—
KAMATH	63.2	—	76.8	—	81.4	—	75.2	—	71.2	—
弓箭手	63.6	0	77.4	0	82.9	0	76.1	0	71.7	0
边缘的	**70.3**	.7±61.6	**83.9**	.4±61.4	**92.3**	.4±92.3	**84.2**	8.11.6±	**75.9**	4.±21.0
SHAPCAL	69.3	.6±51.8	82.9	.5±51.3	91.2	8.22.2±	82.8	.7±61.4	75.0	.3±30.9

方法	ΔΔAcc弓	AUC弓	F1@25弓Δ			特里亚卡	ΔF1@75弓Δ			
MAXPROB	67.0	—	76.7	—	82.1		76.3	—	71.0	—
KAMATH	70.6	—	76.6	—	— 82.1 —		77.9	—	71.1	—
BOWPROP	71.2	0	77.6	0	84.2	0	79.1	0	71.6	0
边缘的	**72.0**	0.80.4±	**78.7**	.1±10.2	**85.4**	**1.20.8±**	**79.6**	0.5 0.3±	**72.3**	0.80.2±
形状的	71.8	0.60.4±	78.2	0.60.3±	84.7	0.50.8±	79.4	0.3 0.4±	72.3	0.80.3±

方法	Acc	Δ弓形	AUC	Δ弓形	霍波特卡 F1@25弓Δ		F1@50	Δ弓形	F1@75	Δ弓形
MAXPROB	63.1	—	75.7	—	79.7	—	75.9	—	72.2	—
KAMATH	64.5	—	76.8	—	80.8	—	77.2	—	72.8	—
弓箭手	64.7	0	76.6	0	80.3	0	76.9	0	72.4	0
边缘的	**65.7**	.0±10.4	**78.2**	.6±10.4	**82.6**	.2±20.8	**78.4**	.5±10.6	**73.8**	.4±10.3
SHAPCAL	65.3	0.70.4±	77.8	.2±10.3	82.0	.6±10.7	78.0	.0±10.5	73.5	.1±10.4

表1：QA任务的主要结果。我们基于解释的方法（肢肢和形状）在转移到三个新领域时，成功地校准了经过训练的罗伯塔QA模型，并且优于先前的方法（KAMATH）和我们仅使用启发式标签（BOWPROP）的消融。此外，我们还展示了三角洲w的均值和标准差。r.t.在弓中跨越多个随机的种子。Δ

PROP的表现与KAMATH相当，或仅略好一些。这些结果表明，将解释与注释连接起来是构建更好的校准器的一条途径。

最后，我们比较了基于不同解释技术的方法的性能。在所有三种设置下，羊皮的表现都略优于形状。如第2.1节所讨论的，SHAP为那些激活特征的扰动分配高实例权重。虽然这种内核的选择在涉及表格数据的任务中是有效的（Lundberg和Lee，2017），但当这种扰动可能不会产生有意义的例子时，这可能不适合于QA的任务。

.24主要结果： NLI

我们的基础NLI模型是一个在MNLI上训练的罗伯塔分类模型，它在开发集上达到了87.7%的准确率。在评价QNLI和MRPC时，我们将矛盾和中立分解为非隐含性。我们继续使用随机森林作为校准器模型。我们评估了在QNLI和MRPC的开发集上的泛化性能。与QA中的设置类似，我们使用500个例子来训练校准器和测试

20个随机试验都休息。

因为QNLI和MRPC是二进制分类任务，所以预测一个模型是否正确（我们的校准设置）等同于原始的预测任务。因此，我们可以用标准的分类精度和AUC来测量校准器的性能。

结果我们在表2中显示了关于NLI任务的结果。当转移到QNLI和MRPC时，基础MNLI模型完全失败，准确率分别达到49%和57%，而大多数类模型的准确率分别为50%（QNLI）和65%（MRPC）。通过启发式注释，BOWPROP能够解决74%的QNLI实例和72%的MRPC实例。与MAXPROB相比，我们的启发式本身对QNLI也很强大。羊膜在两种情况下仍然是最好的，使用解释适度提高了1%和2%。关于NLI任务的结果表明，即使潜在的任务非常不同，我们的方法仍然可以学习到有用的信号来表明模型的可靠性。

.34分析

校准器的跨域推广

到目前为止，我们的校准器都是经过个人训练的

方法	Acc	克恩利 弓AUC Δ		Δ弓形	Acc	MRPC弓 AUC Δ		Δ弓形
MAXPROB CLSPROBCAL	50.5 56.7	— —	41.2 59.5	— —	57.0 71.5	— —	50.0 77.9	— —
弓箭手 边缘的 SHAPCAL 74.2	74.0 **75.0**	0 .0±10.4 0.20.4±	82.0 **82.6** 81.9	0 0.70.4± ±.40.00 73.5	71.8 **73.6**	0 .8±11.3 .7±11.2	79.3 **81.0** 80.7	0 .7±10.9 .4±10.8

表2：关于NLI任务的主要结果。顶点适度地提高了QNLI和MRPC上的基本MNLI模型的性能，尽管这些任务与基本MNLI设置有多么不同。

源\目标		SQ -ADV	琐事	火锅
SQ -ADV	适应卡马斯 酒店	70.9	76.1 73.3 71.9 72.9	65. 8 75. 1 74. 1 71. 4
琐事	适应卡马斯 酒店	64.2 70.5 67.1 69.3	76.7	**77.2** 76.7 75.0 77.0
火锅	适应卡马斯 酒店	56.6 70.6 69.1 68.8	74.0 77.0 76.9 **77.9**	75.7

表3：跨域校准结果的覆盖范围-f1曲线下的面积。对角线上的数字显示了最大值程序的性能。比MAXPROB更好的性能表明该校准器能够有效地推广（彩色细胞）。

	QA	100	300	500
Q A	MAXPROB	70.9		
	卡马斯	.772	75.6	76.8
—	弓箭手	.075	76.0	77.4
S	边缘的	**78.7**	**82.7**	**83.9**
琐事	MAXPROB		76.7	
	KAMATH	74.8	76.2	76.6
	BOWPROP	76.1	77.4	77.6
	LIMECAL	**77.2**	**78.2**	**78.7**
火锅	MAXPROB		75.7	
	KAMATH	75.2	76.5	76.8
	BOWPROP	74.9	76.3	76.6
	LIMECAL	**76.5**	**77.7**	**78.2**
	自然语言 接口	100 300		500
克恩莉	MAXPROB		41.2	
	KAMATH	56.4	58.1	59.5
	BOWPROP	79.0	81.5	82.0
	LIMECAL	**79.1**	**81.8**	**82.8**
多区域处	MAXPROB		50.0	
	KAMATH	73.7	76.8	77.9
	BOWPROP	69.4	77.5	79.3
	LIMECAL	**76.1**	**79.9**	**81.0**

表4：经过训练的校准器的AUC得分

传输设置。是在某些初始域转移设置上学习到的校准器的知识，e。g.，SQuAD！，可推广到另一个转移设置，e。g.！霍波特卡？这将使我们能够采用我们基本的QA模型和一个校准器，并将这对模型应用到一个新的领域，而不做任何新的训练或适应。我们在QA上探讨了这一假设。[5]

为了进行比较，我们还给出了一个罗伯塔模型的性能，首先在阵容上进行微调，然后在域a上进行微调（适应，图2）。适应需要访问模型体系结构，这对于其他方法来说是一种不公平的比较。

计算结果如表3所示。没有一种方法适用于小队ADV和其他领域（训练或测试），考虑到小队~~ADV的综合和~~非常具体的性质，这并不奇怪。

在皮里亚卡和霍波特卡之间，石灰卡和KAMATH校准器都接受了一个训练不同的训练数据大小。即使在有限的训练资源下，基于解释的校准器仍然可以学习，而KAMATH和弓道具并不有效，并且在琐事和霍波特卡上表现不到最大针基线。

域可以推广到另一个域，即使弓形支柱并不有效。此外，与KAMATH相比，我们的石灰表现出更强的泛化能力。然后，我们将LIME - CAL与适应技术进行比较。适应并不总是很好地工作，Kamath等人也讨论过这一点。（2020年）；Talmor和Berant（2019年）。当在热茶上训练和在琐事上测试时，适应会导致性能的巨大下降，而石灰盐在这种情况下是最好的。从琐事到高压炉，改编效果很好，但石灰几乎也同样有效。

总的来说，校准器显示出成功的通用利扎比-

[5]我们也测试了关于NLI-释义转移的假设，但没有看到可转移性的证据，可能是由于这些任务从根本上的不同。

	这是我的朋友							
模型性能 基卡/nli 罗伯塔大学 适配基地 INDOMAIN QA/NLI	前 62.1 32.3 77.3 —	F1 68.0 42.0 84.3 —	Ex 53.2 28.5 56.2 62.1	F1 62.1 34.8 64.0 68.1	前 50.7 39.5 54.3 59.7	F1 66.3 54.8 70.8 77.2	Acc 50.5 81.2 80.7 92.0	Acc 57.2 79.8 79.1 87.2
校准结果 罗伯塔+最大探头适应基地QA/NLI +最大实验室	Acc — — 69.3	AUC 41.1 **92.7** 82.9	Acc — — 72.0	AUC 37.6 77.6 **78.7**	Acc — — 65.7	AUC 67.0 **82.5** 78.2	Acc **81.2** 80.7 74.9	Acc **79.8** 79.1 73.6

表5：石灰法和玻璃箱法的模型性能和校准性能。在QA任务上，石灰优于微调罗伯塔，甚至优于适应的基础QA/NLI。由于玻璃盒方法的简单性和基模型性能差，其在NLI上的性能不佳。

跨越两个现实的QA任务。我们认为，这可以归因于在基于解释的校准器中使用的特征。虽然任务不同，但校准器可以依靠一些常见的规则来决定预测的可靠性。

训练数据大小的影响如果需要大量的注释数据，校准一个新领域的模型会变得麻烦。我们实验改变校准器暴露的训练数据的量，结果如表4所示。我们的基于解释的校准器在每一个设置中仍然是最好的，只有100个例子。在100个例子中，KAMATH和弓道具在小卡和霍波特卡上的表现比基线差，这表明需要更多的数据来学习使用它们的特性。

.44与精细模型的比较

在整个工作过程中，我们假设了一个无法对一个新领域进行微调的黑盒模型。在本节中，我们将比较基于校准的方法和需要访问模型架构和参数的玻璃箱方法。我们在两种不同的设置下评估了两种玻璃盒方法（图2）：(1)微调一个基础罗伯塔模型（微调罗伯塔），它需要访问模型的架构，但不需要参数；(2)微调一个基本的QA/NLI模型，它既需要模型架构，也需要参数。所有这些模型都有500个例子，与石灰石相同。我们还给出了一个使用不同任务的完整域内训练数据训练的模型的性能（IN域QA/NLI）。

我们在表5中给出了模型性能（用精确匹配和F1测量QA和Acc测量NLI）和校准结果。注意，这里都没有玻璃盒方法的校准器，所以我们只报告校准性能的AUC分数。

在QA任务上，有限的训练数据不足以成功地微调罗伯塔模型。因此，精细的罗伯塔并不能达到可信的性能。对基础QA模型的微调大大提高了性能，超过了adv和性能。然而，我们仍然发现，在琐事上，石灰的性能略优于适应。这是一个令人惊讶的结果，并表明基于解释的校准器在某些场景中仍然是有益的，即使我们可以完全访问该模型。

在比QA容易得多的NLI任务上，微调罗伯塔LM模型或基础NLI模型都可以达到大约80%的准确率。我们基于解释的方法在很大程度上落后于玻璃箱方法，可能是因为基本的NLI模型在QNLI（50.5%的准确率）和MRPC（55.0%的准确率）上完全失败了，并且没有为这两个任务提供太多的支持。尽管如此，在NLI上的结果仍然支持我们的主要假设：解释可以用于校准。

5选择性QA设置

到目前为止，我们的研究结果已经表明，一个校准器是可以做到的
使用解释来帮助对在新域中运行的模型的正确性进行二进制判断。我们现在用Kamath等人的选择性QA设置来测试我们的模型。（2020）（图2）。这个实验使我们能够更直接地与之前的工作进行比较，并在顽强（ID）和域外（OOD）例子混合在一起的环境下看到性能。

给定一个在源域数据上训练的QA模型，选择性QA的目标是训练ID源数据和已知OOD数据的混合校准器，并测试校准器在a上工作良好

已知\未知		SQ -ADV	琐事	火锅
SQ -ADV	MAXPROB	85.0	88.7	87.5
	KAMATH	88.8	89.5	88.9
	BOWPROP	91.5	90.6	89.0
	LIMECAL	**94.5**	**91.7**	**91.9**
琐事	麦克斯普罗	85.0	88.7	87.6
	卡马斯	85.6	91.9	88.7
	弓箭手	85.3	92.1	89.9
	边缘的	**90.9**	**92.5**	**92.1**
火锅	麦克斯普罗	85.0	88.7	87.6
	卡马斯	86.1	91.4	89.4
	弓箭手	85.1	91.8	91.6
	边缘的	**91.7**	**92.3**	**92.5**

表6：在选择性QA设置下的覆盖范围-f1曲线下的面积。我们基于解释的方法在这种情况下也很强大，大大优于现有的基线和我们自己的消融术。

在一个域内和一个未知的OOD数据的混合。我们遵循与中类似的实验设置
卡马斯等人。(2020).详细的设置包含在补充材料中。

结果如表6所示，与主要QA结果相似。我们基于解释的方法，软膜，始终是所有设置中最好的。我们指出，我们的方法优于KA - MATH，特别是在设置涉及阵容-
ADV为已知或未知的OOD分布。这个
可以归因于小队和小队adv之间的相似性，这不能很好地与KAMATH中使用的特性（上下文长度、回答长度等）来区分。我们的基于解释的方法在选择性QA设置中的强大性能进一步验证了我们的假设：解释对于校准黑盒模型是有用和有效的。

6相关工作

*我们的方法是受到了最近的模拟测试工作的启发（Doshi-Velez和Kim，2017），i。*e.，人类是否可以根据解释根据输入示例模拟模型的预测。
仿真试验已进行了各种
任务(里贝罗等。，2018年；阮，2018；他等。，2018年；哈斯和班萨尔，2020年)，并在某些任务中给出了积极的结果（哈斯和班萨尔，2020年）。我们的方法试图模拟人类通过将启发式与归因相结合的过程来判断模型的预测，而不是让人类真正完成任务。

使用“元特征”来判断一个模型也出现在关于任务的系统组合的文献中
就像机器翻译一样(Bojar等人。，回答问题(Kamath等人。，2020年；张志等人。，2021年)，选区解析（查尼亚克和约翰逊，2005年；福森和奈特，2009年）和语义解析（Yin和Neubig，2019年）。Rajani和Mooney（2018）在VQA中的工作与我们最为相关；他们也使用启发式特征，但我们进一步将启发式与模型属性联系起来。我们的元特征集来自于某些属性的存在，这类似于基于概念的解释中使用的“概念”(Ghorbani等人。，2019年；穆和安德烈亚斯，2020年)，但我们关注于使用它们来估计模型的性能，而不是解释一个预测。

我们的工作解决了校准的问题(Guo等。，2017年；德赛和德雷特，2020年)，这通常是根据模型的输出概率来确定的。过去的工作都试图通过温度尺度来解决这个问题(Guo等。或标签平滑(Pereyra等人。该研究调整了所有预测的信心分数。相比之下，我们通过应用一个利用特定于实例的解释的分类器来解决这个问题。过去在NLP中推广到域外分布的工作主要集中在使用来自目标域的未标记数据，并需要对模型进行微调(Ma等人。，2019年；拉姆波尼和普兰克，2020年；郭等人。而我们提高了严格的黑盒模型的OOD性能。

7讨论和结论

尽管在提高模型泛化性能方面显示出了良好的结果，但我们的基于归因的方法确实存在密集的计算成本。当输入大小很大时，使用LIME或SHAP来生成属性需要运行相当数量的扰动（详见附录），这限制了我们的方法的适用性。但这并没有破坏本文的主要贡献，回答了标题中的问题，我们的方法仍然适用于我们付费访问模型，但不是每个查询。

结论我们探讨了模型属性是否可用于校准黑盒模型。答案是肯定的。通过将属性与人类启发式方法联系起来，我们提高了模型在新领域和任务上的泛化性能。此外，它在某些情况下（跨域泛化和选择性QA）表现出很好的泛化性能。

确认信息

感谢匿名评论者提供的有益反馈。这项工作得到了美国国家科学基金会拨款IIS1814522的部分支持，来自销售力量公司的礼物和来自亚马逊的礼物。

参考文献

翁德杰、博贾尔、拉金·查特吉、基里斯蒂安·费德曼、伊维特·格雷厄姆、巴里·哈多、黄书坚、哈克、菲利普·科恩、刘群、瓦尔瓦拉·洛加切娃、克里斯托夫·蒙茨、马特奥·内格里、马特·波斯特、拉斐尔·鲁比诺、卢西亚·斯佩奇亚和马可·图尔奇。2017.2017年机器翻译会议（WMT17）的调查结果。*在第二次机器翻译会议的会议记录中。*

阿尔琼·钱德拉斯卡兰、普拉布，亚达夫、查托帕德海耶和帕里克。2018.解释会使VQA模型对人类更容易预测吗？*在自然语言处理中的经验方法（EMNLP）会议的论文集上。*

尤金·查尼亚克和马克·约翰逊。2005.协调-精细的n-最佳解析和MaxEnt鉴别重新排序。*在计算语言学协会第43届年会的会议记录（ACL‘05）中。*

陈吉凡和德雷特。2019.理解多跳推理的数据集设计选择。在
*计算语言学协会北美分会2019年会议论文集：人类语言技术，第1卷（长和短论文），第4026-4032页，明尼阿波利斯，明尼苏达州。*计算语言学协会。

什里·德赛和格雷格·德雷特。2020.对预处理过的变压器进行校准。*在2020年自然语言处理经验方法（EMNLP）会议论文集上。*

威廉·戴维斯多兰和克里斯·布罗克特。
2005.Automati-
用来构建一个句子释义的语料库。*在第三届释义国际讲习班的会议记录中（IWP2005）中。*

最后一集是金。2017.对a
可解释性机器学习的严谨科学。*arXiv预印arXiv：1702.08608。*

非士、塔摩、佳、徐明俊、崔、陈丹奇。2019.MRQA 2019共享任务：评估阅读理解中的泛化性。*在EMNLP第二台机器阅读理解（MRQA）研讨会论文集。*

维多利亚·福萨姆和凯文·奈特。2009.组合组成解析器。*《人类语言技术论文集：2009年年会计算语言学协会北美分会，同伴卷：短论文。*

戈巴尼、詹姆斯韦克斯勒、邹和金。2019.朝向基于概念的自动解释。*在神经信息处理系统（NeurIPS）进展会议论文集上。*

郭川、普利斯、孙宇、奎连Q。温伯格2017.论现代神经网络的校准。*在第34届机器学习国际会议（ICML）会议论文集。*

韩国，罗摩干斯帕苏努鲁和摩希特班萨尔。2020.多源域对文本分类的自适应。*在人工智能促进会（AAAI）的论文集中。*

彼得·哈斯和莫希特·班萨尔。2020.评估可解释的人工智能：哪种算法解释可以帮助用户预测模型行为？*在计算语言学协会（ACL）年会的会议记录中。*

贾家林和梁杰辉。2017.评估阅读理解系统的对抗性例子。*在计算语言学协会（ACL）年会的会议记录中。*

曼达尔乔希，崔，丹尼尔威尔德和卢克泽特莱莫耶。2017.TriviaQA：一个用于阅读理解的大规模远程监督挑战数据集。*计算语言学协会第55届年会（第1卷：长篇）。*

爱弥田、贾罗宾、梁柏西。2020.在域转移下的选择问题回答。在
计算语言学协会（ACL）年会的会议记录。

Y.刘、奥特、南戈亚尔、杜景飞、乔希、陈丹奇、奥默列维、M。刘易斯，卢克·泽特莱莫耶和维塞林·斯托亚诺夫。2019.罗伯塔：一种稳健优化的伯特预训练方法。*ArXiv，abs/1907.*11692.

斯科特·伦德伯格和李苏因。2017.一种解释模型预测的统一方法。*在神经信息处理系统（NeurIPS）进展会议论文集上。*

马晓飞、徐鹏、王志国、拉帕蒂、项翔。2019.基于bert的领域分类和数据选择的领域自适应。*在第二届低资源NLP深度学习方法研讨会的论文集（DeepLo 2019）中。*

米切尔P。马库斯，比阿特丽斯·圣托里尼，和玛丽·安·马辛凯维奇。1993.建立一个大型的注释英语语料库：宾州树银行。计算语言学，19(2)：313-330。

汤姆·麦考伊，艾莉·帕夫利克和塔尔·林森。2019.正确的原因：诊断自然语言推理中的句法启发式。*在计算语言学协会第57届年会的会议记录中。*

杰西·穆和雅各布·安德烈亚斯。2020.神经元的组成解释。*在神经信息处理系统（NeurIPS）进展会议论文集上。*

董阮。2018.比较文本分类的自动评价和人工评价。*在计算语言学协会：人类语言技术（NAACL-HLT）北美分会2018年会议论文集中。*

费比安·佩德雷戈萨、盖尔·瓦拉考、亚历山大·格兰福特、文森特·米歇尔、贝特朗·蒂里翁、奥利弗·格里塞尔、马修·布朗德尔、彼得·普雷滕霍弗、罗恩·韦斯、文森特·杜布尔、杰克·范德普拉斯、亚历山大·帕索斯、大卫·库纳波、马蒂厄·布鲁彻、马蒂厄·佩罗和爱德华·杜切内。2011.在python中的机器学习。《机器学习研究杂志》，12（85）：2825-2830。

加布里埃尔·佩雷拉，乔治·塔克，简·乔罗斯基，Łukasz凯泽和杰弗里·辛顿。2017.通过惩罚自信的输出分布来正则化神经网络。*国际学习表现会议讲习班（ICLR讲习班）。*

拉贾尼和雷蒙德·穆尼。2018.堆叠与辅助功能的视觉问题回答。*在计算语言学协会北美分会2018年会议记录中：人类语言技术，第1卷（长论文），新奥尔良，路易斯安那州。*

普拉杰普尔卡，吉安Zhang，康斯坦丁·洛普雷夫和珀西·梁。2016.扩展：10万个+的机器理解文本的问题。*在自然语言处理中的经验方法（EMNLP）会议论文集，第2383-2392页，奥斯汀，德克萨斯州*。计算语言学协会。

艾伦·兰波尼和芭芭拉·普兰克。2020.nlp中的神经无监督领域适应调查。*在计算语言学国际会议（COLING）。*

马可图里奥里贝罗，萨米尔辛格，和卡洛斯格斯特林。2016.“我为什么要相信你呢？”解释任何分类器的预测。*在ACM SIGKDD知识发现和数据挖掘国际会议（KDD）中。*

马可图里奥里贝罗，萨米尔辛格，和卡洛斯格斯特林。2018.锚点：高精度的与模型无关的解释。*在人工智能促进会（AAAI）的论文集中。*

劳埃德S Shapley。1997.对于n人游戏的一个值。*博弈论中的经典著作，第69篇。*

孙达拉扬、塔利、颜。2017.深度网络的公理归因。*在机器学习国际会议（ICML）的论文集上。*

阿隆·塔尔莫尔和乔纳森·贝兰特。2019.MultiQA：对阅读理解中的泛化和转移的实证研究。*在计算语言学协会第57届年会的会议记录中。*

亚历克斯·王，阿曼普雷特·辛格，朱利安·迈克尔，菲利克斯·希尔，奥默·利维和塞缪尔·R。弓箭手2019.胶水：一个用于自然语言理解的多任务基准测试和分析平台。*在学习代表国际会议（ICLR）。*

阿迪娜·威廉姆斯，尼基塔·南吉亚和塞缪尔·鲍曼。2018.一个通过推理来理解句子的广泛覆盖的挑战语料库。*在计算语言学协会北美分会2018年会议记录：人类语言技术会议记录（NAACLHLT）。*

杨志林、彭齐、张赛正、本治、威廉。科恩，鲁斯兰·萨拉库迪诺夫和克里斯托弗。人员配备2018.HotpotQA：一个用于不同的、可解释的多跳问题回答的数据集。*在自然语言处理中的经验方法会议上（EMNLP）上。*

叶、罗翰奈尔和德雷特。2021.将属性和QA模型行为联系起来。*在2021年自然语言处理中的经验方法会议的论文集上。*

尹鹏程和纽百翰。2019.神经语义解析。*在计算语言学协会第57届年会的会议记录中。*计算语言学协会。

张书坚、龚成月、崔恩实。2021.了解更多的问题可以帮助：提高问题回答的校准。*在计算语言学协会（ACL发现）的发现中。*

一个在LIME中使用的内核的详细信息和形状

LIME启发式设置T_X作为在扰动与原始输入之间的余弦距离函数上定义的指数核（条带函数为a），即：

$$T_X(z) = \exp\left(-\text{dcos}(\text{北}, h_X(z))/a^2\right)$$

也就是说，LIME为更接近原始输入的扰动分配更高的实例权值，因此使用近似值正确地对这些扰动进行优先排序。

SHAP推导$T_X\phi$所以它可以被解释为Shapley值（Shapley，1997）：

$$T_X(z) = \frac{n-1}{\binom{N}{|z|}|z|(n-|z|)}$$

其中|z|表示激活标记的数量（z之和）。这个内核为很少或多个活动标记的扰动分配高权重，因为当少数标记效应被隔离时的预测是重要的。这将SHAP与LIME区分开来，因为LIME将对很少有有效令牌的扰动施加非常低的权重。

b选择性QA设置的详细设置

我们遵循与Kamath等人类似的实验设置。(2020).我们在小队上训练了一个罗伯塔QA模型，并使用1000个小队开发例子+1000个已知的OOD示例来训练校准器。我们报告测试结果相同类型的混合物（1000小队+1000已OOD，对角线块表6）和4000小队的混合例子+4000未知OOD（2560小队+2560小队ADV小队ADV只包含2560个例子）。

C特性的重要性

我们分析了由校准器学习到的重要特征。我们发现，基于解释的特性确实通常是最常用的特性中，而且比基于单词包的特性更重要（详细列表见附录）。所有的QA校准器都在很大程度上依赖于专有名词（NNP）和wh词的归因值。重叠名词的BoW特征在QNLI上被认为是重要的，但顶部特征仍然是基于归因的。

这些因素揭示了QA或NLI推理过程的哪些部分对于模型的捕获是重要的。例如，在adv阵容中对nnp的依赖符合我们的直觉理解这个任务：干扰物通常有错误的命名实体，所以如果模型关注NNPs，它更有可能是正确的，校准器可以利用这一点。

表7显示了QA和NLI的最重要特征。为了简洁起见，我们将与顶级预测的概率相关的特征呈现为一个特征（Prob）。基于解释的特性确实通常是最常用的特性之一，而且比原始属性特性更重要。

DPOS标签属性的详细信息

我们使用在spaCy API中实现的标记器。[6]这个标签集基本上遵循Penn树库标签集，除了我们合并一些相关的标签，以减少有限的训练数据的特征数量。[7]具体来说，我们将JJ，JJR，JJS合并为JJ、NN、NNS合并为NN、NNP、NNPS合并为NNP、RB、RBR、RBS合并为RB，VB，VBD，VBG，VBN，VBP，VBZ合并为VB，WDT、WP、WP$、WRB合并为W。这样，我们就得到了一个总共包含25个标签的标签集。

E黑盒校准器的详细信息

QA的功能计数

KAMATH（Kamath等人。（我们使用在(Kamath等人。，包括前5个预测的概率、上下文长度和预测

回答长度。

除了在KAMATH中使用的7个功能之外。我们将属性空间V构造为低级段和段伪标签的并集。×由于在输入中有3个部分的问题、上下文、答案和25个标签（D节），因此属性空间|V|的大小为3 + 3 25 = 78。×因此，功能的总数（包括从KAMATH中获得的7个）是85个。

边缘和形状：回想一下，属性空间的大小是78。石灰质和形状使用78个特征描述与相应属性相关的属性，以及85个特征BOWPROP .该功能部件的总数为因此163。

6https://spacy.io/api

7https://www.林。upenn.edu/课程/Fall_2003/ling001/penn_treebank_pos。html

SQ -ADV	琐事	火锅	克恩莉	多区域处理中心
Attr到NNP	顶级Pred	顶级Pred	Attr重叠NN在H	顶级Pred
Attr到VB在C中	答案长度	附件到Q	H中的BOW OverlNN	附P
顶级Pred	Attr NNP在Q	Attr Whin Q	P中的BOW OverlNN	附于H
Attr到NN	Attr Whin Q	附C	在P中的非超nn	Attr到H中的非overlNNP
答案长度	附件的问题	Attr到NNP	顶级Pred	Attr到P中的OverlSYM

表7：顶骨在不同任务中使用的最重要特征。对于QA，问题中NNP的归因和问题中Wh-的归因通常是重要的。对于NLI，与重叠/非重叠名词相关的特征更有效。

NLI的功能计数

克隆人(Kamath等人。：我们在实践中使用了两个特征，即隐含概率和矛盾概率。我们不包括中性的概率，因为它可以从其他两个类的概率中推断出来。

××组合：除了分类中使用的两个特征外，我们还构造了属性空间V作为低级段和段位置标签重叠词的并集。由于有2个片段（假设，假设），25个标签（D节）和2个属性重叠重叠，非重叠，属性空间|V|的大小为2 + 2 25 2 = 102。××因此，特征的总数（包括来自克隆的2个）是104个。

羊膜和形状：角膜膜和形状在附件中增加了102个特征
到弓道具中使用的104个功能。因此，这些特征的总数是206个。

为输入相对较长的QA任务生成解释的成本，我们采样2048个扰动，并对每个扰动运行推理
样例对于更简单的NLI任务，我们对每个示例使用大约512个模型查询。

我们使用来自Scikit-Learn的随机森林实现(Pedregosa等人。， 2011).我们在表8中列出了每种方法中使用的超参数。通过400个训练例和100个验证例的网格搜索确定超参数。这个
树数的选择为[200,300,400,500]，最大深度的选择为[4、6、8、10，

卡		NUM .树最大深度	
SQ -ADV	KAMATH	300	6
	BOWPROP	300	20
	LIMECAL	300	20
	SHAPCAL	300	20
琐事	KAMATH	300	6
	BOWPROP	200	20
	LIMECAL	300	20
	SHAPCAL	300	20
火锅	KAMATH	300	4
	BOWPROP	300	10
	LIMECAL	300	10
	SHAPCAL	300	10
自然语言接口		NUM .树	最大深度
克恩莉	KAMATH	300	4
	BOWPROP	300	6
	LIMECAL	400	20
	SHAPCAL	400	20
多区域处理中心	KAMATH	300	6
	BOWPROP	300	8
	LIMECAL	400	20
	SHAPCAL	400	20

表8：用于训练不同方法的随机森林分类器的超参数。

15, 20].然后，对于表1、表2和表4中的实验结果，我们总是修复超参数，而不执行任何进一步的超参数调优。

f玻璃箱方法的细节

对于RoBERTaQA，我们对20个时代的学习率进行微调为1e-5（我们也尝试对3个时代进行微调，但目标不收敛于500个例子）。我们将批量大小设置为32，然后进行预热
比率为0.06。

对于MNLI，我们优化了一个在10个时代的学习速率为1e-5的MNLI基模型。根据Liu等人的超参数，我们将批次大小设置为32，预热比设置为0.06。
(2019).

采用基础QA/NLI模型，我们
调整基础罗伯塔QA模型的训练上
2个时代的学习率为1e-5。
对于MNLI，我们找到了罗伯塔NLI
模型在MNLI上进行训练，学习速率为
10个时期的1e-5。学习不收敛
当微调到两个时代时，如MNLI任务是
和QNLI和MRPC太不同了。