作为可解释姿态检测的归因先验的人类理性

纽约州纽约市哥伦
比亚大学
sahil.j@columbia.edu

艾米莉·阿拉维纽
约哥伦比亚大学
eallaway@cs.columbia.edu

摘要

随着 NLP 系统在从文本中检测观点和信念方面变得越来越好，重要的是不仅要确保模型是准确的，而且要确保它们以符合人类推理的方式来实现预测。在这项工作中，我们提出了一种方法，通过在一小部分训练数据上使用众包注释，将类似人类的合理化传递给姿态检测模型。我们表明，在数据稀缺的情况下，我们的方法可以提高最先进的分类器的推理能力，特别是对于包含讽刺等挑战性现象的输入，而不会损失预测性能。此外，我们还证明了注意力权重在提供模型预测的可靠解释方面优于领先的归因方法，从而为我们的模型提供了计算成本低廉且可靠的归因来源。

1 介绍

姿态检测，自动识别文本在主题上的位置(Mohammad et al.,2017)，允许读者从新闻文章和社交媒体中收集有价值的信息，比如文章是否有政治倾向。由于许多话题(如政治意识形态和宗教信仰)的敏感性，立场模型透明并以类似人类的方式合理预测是至关重要的。此外，他们的推理必须保持像人类一样，即使他们的任务是概括一个新的，看不见的测试主题。

特定输入的模型基本原理可以以特征属性的形式提取，这些特征属性量化了每个输入特征对模型预测的影响。获得这些属性的方法因忠实度(准确衡量每个特征重要性的程度)而异。归因可以通过归因先验整合到训练过程中，归因先验是传授领域知识的有力框架

Topic =大麻真实标签= FOR

基线	然而，更好的问题是，有多少吸食过大麻的人没有吸食其他非法毒品。我想答案会让你大吃一惊。相关性不是因果关系-你越早了解这一点，你就能越早找到我们鸦片广告的真正根源- 条件问题。
预报	反对(错误)
我们的方法	然而，更好的问题是，有多少吸食过大麻的人没有吸食其他非法毒品。我想答案会让你大吃一惊。相关不是因果——你越早了解这一点，你就能越早找到我们的真正根源 鸦片成瘾问题。
预报	对于(正确)

表 1:用平均注意力权重(MAW)解释的模型推理。仅使用交叉熵损失来训练基线，而使用我们提出的属性先验来训练第二模型。

至模型(Erion et al.,2019,2020).在这项工作中，我们提出了一个先验，惩罚姿态分类器产生偏离人类对单词重要性的判断的属性(即，文本中的哪些单词或短语最能表明姿态)。值得注意的是，我们的方法是模型不可知的，可以用于任何可区分的特征属性技术。

我们在大量(各种姿态主题)数据集上训练和评估我们的模型，数据集的测试集覆盖了训练数据中缺少(零镜头)或缺少(少镜头)的主题(Allawayand McKeown,2020).为了构建我们的属性先验，我们在训练数据的一个小子集(500个例子)以及测试样本上众包单词重要性注释

2021 年自然语言处理经验方法会议论文集，第 5540-5554 页，2021 年 11 月 7-11 日 ©c 2021 计算语言学协会

评估数据。此外，为了评估我们的方法在现实的、资源有限的场景中的表现，我们不仅使用完整的大规模训练集进行实验，还使用其简化版本进行实验。

作为我们的先验的归因方法，我们选择平均注意力权重(MAW)，一种与流行的替代方法相比在计算上便宜的方法。尽管最近的大量出版物表明，注意力权重通常不能提供模型行为的忠实解释，但我们发现，与梯度输入(GI)相比，MAW 提供的属性更忠实于我们的模型预测，梯度输入是变压器模型的主要属性方法。我们的贡献如下:(1)我们提出了一种方法，在模拟数据缺乏的情况下，改进了最新姿态检测模型的推理，而不影响其性能;(2)我们表明 MAW 不仅比 GI 更简单，计算成本更低，而且更忠实于我们的模型。我们的数据和模型可从以下网址获得 https://github.com/SahilJ97/Explainable-Stance-Detection。

2 相关著作

虽然关于姿态检测的早期工作主要集中在意识形态辩论(Walker et al.,2012; Hasan and Ng,2014; Abbott et al.,2016)，最近的数据集也开始包含更多的政治话题，比如选举(Mohammad et al.,2016 ; Vamvas and Sennrich,2020; Lai et al.,2020)和公民投票 (Taulé et al.,2017 ; Tsakalidis et al.,2018).这反映出人们对开发模型来理解公众对一系列话题的看法越来越感兴趣。然而，要在现实世界中使用，这样的模型还必须表现出良好的概括能力和一定程度的透明性。最近的工作集中在姿势检测模型的泛化能力上:跨主题 (Augensteinet al.,2016 ; Xu et al.,2018 ; Allaway and McK-eown,2020 ; Zhang et al.,2020 ; Allaway et al.,2021) 、语言 (Vamvas and Sennrich,2020)，甚至标签集和流派 (Schiller et al.,2020 ; Hardalov et al.,2021).相比之下，我们的工作侧重于主题泛化过程中的模型推理。

许多属性方法已经被用于从文本分类器中提取推理，包括基于激活的(Atanasova et al.,2020)，基于扰动的 (Ribeiro et al.,2016)，渐变-基于和基于注意力 (Abnar and Zuidema,2020; Wu and Ong,2021)方法。此外，许多工作将这些模型属性结合到训练过程中。Liu and Avci （20192)使用基于综合梯度的属性先验、基于梯度的特征属性方法(Sundararajan et al.,2017).Zhong et al.（2019)直接训练用于关系提取的注意机制。先前的研究已经使用了非常类似于我们的单词重要性注释来监控注意力，目的是提高预测性能 (Pruthi et al.,2020 ; Kanchinadam et al.,2020).相比之下，我们的工作集中在可验证地改进分类器的推理。虽然存在推理路径自然透明且易于训练的模型，例如选择-预测(Jacovi and Gold-berg,2021)和基本原理-增强(Zaidan et al.,2007; Zhang et al.,2016)模型，我们的框架对模型架构不做任何假设，因此可以应用于最先进的姿态分类器。

最近对可解释性(即归因)方法的调查提出了比较技术的五个诊断属性，包括忠实性，其定义为对模型内部工作的真实归因程度的度量(Atanasova et al.,2020).他们对变压器模型的实验表明，基于梯度的方法(如 GI)在可信度方面得分较高。事实上，类似的忠实度评估发现，注意并不能提供忠实的解释(Jain and Wal-lace,2019; DeYoung et al.,2020)，尤其是与 GI(Wu and Ong,2021).然而，最近的研究主张对忠诚的更微妙的理解，即优先考虑“足够可解释的” Wiegreffe and Pinter,2019 ; Jacovi andGoldberg,2020)并允许“最佳”技术因模型、任务或输入而异(Ghorbani et al.,2018).我们的工作提出了一个情景，在这个情景中，注意力实际上比胃肠运动更可靠。

3 数据

3.1 众包注释

为了研究跨越许多主题的姿态检测的模型原理，我们注释了最近提出的 VAST(Allaway and McKe-own,2020)具有人类理性的数据集。VAST 由《纽约时报》部分内容的评论(此处称为论点)组成。数据集中的姿势主题是自动提取的

逻辑地(例如，通过识别重要的名词短语)，然后使用众包来验证(或纠正)。众包也被用来给每个例子分配一个标签:“赞成”(赞成)、“反对”(反对)或“中立”。

对于注释，我们从训练集中随机选择700个非中立的例子，它们的主题由注释者验证。我们还从测试集中选择了75个这样的例子。对于我们判断主题不明确的例子(例如，“问题”)，我们不做注释(训练中142个，测试中32个)。此外，在训练样本中，我们发现了14个带有不正确姿势标签的例子。

我们通过Amazon mechanic-Turk(一个流行的众包平台)收集注释。对于每一个点击(任务)，工作人员被要求(1)对论点相对于主题的立场进行分类，以及(2)选择论点中k个最重要的词(对于每个例子，我们提供k值的可接受范围)。如果一个词被认为是重要的，就要把它屏蔽掉(1)难度更大[1]。附录中提供了我们的MTurk HIT的详细说明A。

为了确保我们注释的质量，我们首先发布一个由单个例子组成的“限定”点击。然后，三个合格的工人在我们的子集中注释每个例子。我们还使用上面的(1)来检查培训子集中的工人质量
只有。特别地，对于训练中的74个样本子集，至少有两个注释者不同意

黄金姿态标签。作者检查了这些例子中的每一个，或者翻转标签(35个例子)或者丢弃例子。这些决定的科恩系数是0.392。

3.2 Oracle属性

计算和处理:我们众包任务的结果用于计算每个带注释的参数的oracle属性。

例如，如果一个工人的姿势分类(见上面的(1))与标签不一致，我们就忽略他的注释。我们的oracle属性是注释者响应的加权和，以工人质量分数(WQS)作为加权因子。WQS衡量一个注释者与他们的同行在单词级别上的一致程度(Dumitracheet al.,2018).我们的注释者的平均WQS是0.58。请注意，这些oracle分数稍后会被标准化(4.2).

[1]我们HITs中提供的实际定义有些不同(参见附录A).

分析:我们检查了处理过的oracle属性，发现注释者将一个参数中平均26%的标记标记为重要。令人惊讶的是，只有51% (267/519)的例子选择了主题中的单词，而平均44%的重要单词是停用词。例如，一名工人在“我不知道这里有谁对奥运会有丝毫的兴奋”这句话中选择了“这里没有人”(见表5).这表明人类的单词重要性判断不能简单地通过选择主题或与主题最相似的单词来近似。此外，我们发现平均只有10% (9%)的重要词汇是正面(负面)情感承载的，如MPQA词典(Wilson et al.,2017).这进一步强调了用于姿势检测的人类单词重要性判断的复杂性，因为情感在确定姿势中仅起次要作用。

4 方法

我们提出了使用基于BERT的编码器(4.1)用附加损失项(4.2)旨在强加一种基于人类理性的先验。定义D = xi = (di, ti, yi) N 作为数据集

$_{i=1}$ 随着 n 个示例，每个包含一个参数 $\in$ –

di，一个主题ti，和一个姿态标签yi 0，1，1。另外，设Mθ是某个带参数θ的模型。然后，我们可以为每个示例x定义一组oracle属性s、模型属性a和惩罚权重γ(每个令牌对我们的损失项的贡献)。我们先前的损失项鼓励模型为每个例子产生与s非常“相似”的属性a(4.2-4.3).

4.1 基础架构

我们的基础架构建立在由Allaway and McKeown (2020)，它使用BERT(Devlin et al.,2019)，从而以文档为主题表示的条件，反之亦然。我们的模型在两个方面不同于BERT-joint。首先，我们不是修复BERT，而是在训练期间微调它的权重，从而允许变压器更新它的注意力。为了适应我们对归属方法(MAW)的选择，这是必要的。其次，我们不是从输入中删除停用词，而是

将完整的输入序列([CLS]文档[SEP]主题[SEP])传递给 BERT，并通过获取所有非停用词的平均隐藏状态来计算用于姿态分类的最终表示。我们这样做是因为我们的 oracle 属性覆盖了参数中的所有单词，而不仅仅是非停用词。

4.2 先前损失

我们的示例级基本原理损失函数 ω 是归一化模型属性和归一化 oracle 属性之间的加权均方误差。形式上，例如 x = (d，t，y)，设 m 是我们自变量 d 的长度，θ 表示我们模型的参数。设 x 是单词重要性注释的例子，惩罚权重 γ = (γ1，...，γm)，甲骨文定语 s = (s1，...，sm)，以及模型属性 a = (a1，...，am)。让 arj 记下自变量令牌 j 的归一化属性分数。也就是说，

$$ar_j = \frac{a_j}{\sum_{i=1}^{m} a_i}$$

类似地，设 srj 表示自变量标记 j 的归一化 oracle 分数，则 ω 定义如下:

$$\Omega(\theta; x) = \sum_{j=1}^{m} \gamma_j (ar_j - Sr_j)^2 \quad (1$$

直观上，与自变量的第 j 个令牌上的属性相关联的平方误差被加权

对于一个例子 xr，即

没有赋予甲骨文的属性，我们定义 Ω(θ;xr)为0。

我们的完全损失函数是 c k=1 jk 之和姿势分类损失 c 和跨示例的比例平均先验损失

$$(\theta; d) = c(\theta; d) + \lambda p(\theta; D) \quad (2)$$

$$p(\theta; d) = \frac{1}{n} \sum_{i=1}^{n} \omega(\theta; x_i) \quad (3)$$

其中 D 是数据集，c 是交叉熵损失，λ > 0 是超参数。

权重，其中标点符号或数字标记得分为 0，所有其他标记得分为 1。然而，我们也试验了 tf-idf 惩罚权重，其中非标点、非数字标记的分数是其相对于训练集的 tf-idf。[2]

4.4 特征归属

我们的先验允许选择任何归属方法。我们选择平均注意力权重(MAW)，因为与其他方法相比，它的计算成本极低，大多数方法需要反向传播和/或多次前向传递(Atanasova et al.，2020)。在 MAW 中，标记 j 的属性分数是注意力权重 α ij(即，与索引 j 处的键相关联的所有注意力权重)的平均值，该平均值跨越所有标记、层和注意力头部。非正式地，MAW 测量每个标记受到多少关注(来自其他标记以及来自自身)。我们的框架假设归因分数是量值(即，无符号)。因此，我们隐式地获取 MAW 属性的绝对值。

我们还比较了 MAW 和另外一种方法，梯度输入(GI)(Wu and Ong，2021)，供评价。让 e = ，。)

是的输入嵌入 ar 的第 j 个令

然后，我们将令牌 j 的 GI 属性分数定义为

$$aGI_j = \left(\frac{\partial fc(x)}{\partial e} \right)^2 \quad (4)$$

其中 fc 表示对应于类别 c 的模型输出函数的分量。直观地，GI 测量模型对 ej 的扰动的敏感性，理论上测量模型预测对令牌 j 的依赖性。We 选择跨输出类聚合是因为

4.3 处罚重量

为了从我们的基本原理损失函数中排除某些记号(标点和数字),并且潜在地将非均匀影响分配给剩余的记号，我们引入了惩罚权重的概念。令牌的惩罚权重指定了其对基本原理损失函数的贡献。在我们的实验中，我们主要关注二元惩罚

所有输出神经元都有助于模型的决策；对非预测类的负贡献与对预测类的同等大小的正贡献一样重要(Bachet al.,2015).我们选择 GI 作为基准方法，因为它在跨几个领域(Atanasova et al.,2020； Wu andOng,2021).

2示例的“文档”是数据集中与示例主题相关的所有参数的联合。

	民数记不包括	% with 神谕	民数记主题 f	民数记主题 z
全部	13438	3.9	5019	–
减少 25	3801	13.7	1831	–
减少 10	1788	29.0	887	–
偏差	2062	–	114	383
试验	3006	1.4	159	600

表 2:海量和训练集的数据集统计。f 表示少镜头主题，z 表示零镜头主题。

5 实验

5.1 数据

我们在 VAST 上训练和评估我们的模型，使用标准的训练/开发/测试分割和测试集的两个子集:少量测试(每个测试主题有少量训练样本)和零测试(每个测试主题没有训练或开发样本)。请记住，在广泛的主题范围内获得高质量的姿态数据是极其耗费资源的(All-away and McKeown,2020)，我们通过在不同程度的数据稀缺下评估我们的方法来评估它的实用价值。具体来说，我们使用三个数据设置进行实验，这三个数据设置改变了没有 oracle 属性的训练示例的数量:除了使用所有训练示例(完整)之外，我们还仅使用 25%(减少 25)或 10%(减少 10)没有 oracle 属性的训练示例的随机样本进行实验。我们在所有三个数据设置中使用了所有 519 个带有 oracle 属性的训练示例(见表 2).

5.2 模型

我们使用二进制惩罚权重(4.3)，我们的众包 oracle 属性(3.2)，并使用 MAW 来提取模型属性(4.4).我们将这个模型与一个共享其架构但没有先前丢失的模型(base)进行比较。此外，我们还比较了为 VAST 提出的两条基线:BERT-joint - 我们的架构(4.1)而不进行微调，并进行额外的数据预处理，以及 TGA Net-BERT-joint 的一种修改，它使用无监督聚类和注意力来提高对未知主题的性能(Allaway and McKeown,2020).

我们使用手动超参数搜索来调整 λ。我们发现，因为只有一小部分实例具有 oracle 属性，所以该系数适用于我们以前的损失术语必须相当大:在完全和缩减 25 设置中 λ = 49152，在缩减 10 设置中 λ = 16384。我们的模型是在 PyTorch 中实现的[3]并使用 Adam 优化 20 个周期，批次大小为 32，固定学习速率为 105。我们对参数使用最大序列长度 250，对主题使用最大序列长度 10。所有型号都使用 Bert-base-uncased from huggingface[4]。除非另有说明，结果是三个随机种子的平均值。

5.3 结果:姿态预测

我们使用少数镜头和零镜头子集上的宏观平均 F1 来评估我们的模型(3.1)的庞大测试集(见表 3).我们看到，在整个培训设置中，prior-bin:gold 和 base 取得了可比的结果，并超过了为 VAST 建议的基线。我们还对计算先前损失中惩罚权重的方法进行了探讨(4.3)在数据稀缺的 reduced25 设置中。具体来说，我们使用 prior-tfi df:gold - tf-idf 惩罚权重和众包 oracle 属性以及 prior-bin:tfi df - 二进制惩罚权重和 TF-IDF 值作为伪 oracle 属性(而不是我们的众包标签)进行实验。这两种方法的表现都不如 prior-bin:gold 和 base，分别达到 0.661 和 0.655 macro-F1。这一结果与我们对人类单词重要性注释的观察相一致(3.2)，即人类理性是复杂的，并不一定平行于通过 tf-idf 得出的单词重要性概念。因此，我们的立场预测结果表明，为了使用我们提出的属性先验获得强有力的结果，人类单词重要性注释是必要的。

5.4 理由分析

除了评估我们模型的预测，我们还评估他们推理的质量。为了做到这一点，我们首先分析从 MAW 和 GI 获得的解释的相对可靠性。然后，我们使用我们的发现通过两个独立的机制来评估基本原理质量:人类评分者和我们的基本原理损失函数(ω)。

归因的忠实性:归因方法的忠实性是它准确反映模型推理的程度(Herman,

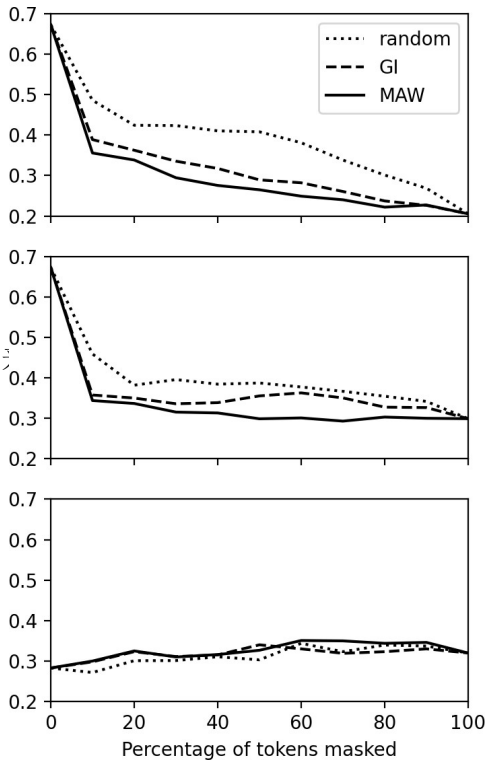
[3]https://pytorch.org
[四]https://huggingface.co/transformers

		全部			零射击			少数镜头		
		赞成	欺骗	平均值	赞成	欺骗	平均值	赞成	欺骗	平均值
全部	伯特接头	.545	**.591**	.653	.546	.584	.661	.544	**.597**	.646
	TGA 网	.573	.590	.665	.554	**.585**	.666	.589	.595	.663
	基础	.643	.581	**.692**	.632	.563	.692	**.652**	**.597**	**.691**
	前箱:黄金	**.645**	.546	.684	**.649**	.542	**.693**	.641	.549	.669
减少 25	伯特关节	.516	.524	.603	.553	.527	.619	.480	.522	.587
	基础	.626	**.559**	**.673**	.634	**.564**	.688	.618	**.552**	**.658**
	前箱:黄金	**.637**	.549	**.673**	**.643**	.537	**.694**	**.631**	.527	.653
减少 10	伯特关节	.450	.469	.370	.491	.478	.372	.422	.448	.366
	基础	**.594**	.491	.623	**.600**	.460	.630	**.589**	.513	**.614**
	前箱:黄金	.579	**.526**	**.630**	.596	**.522**	**.650**	.562	**.529**	.609

表 3:所有三个版本的训练集的测试集的 F1 结果。Avg 是指所有三个类别(职业、职业和中立)的宏观平均值。报告的分数结果 Allaway and McKeown （2020).基础 bin 和先前 bin 之间的差异:黄金没有统计学意义(p < .05)。

方法	前箱:黄金	基础
随意	.353	.358
血糖指数	.285	.326
咽喉	**.264**	**.299**

表 4:阈值-性能曲线下的面积(AUC-TP)。



2017).尽管 Atanasova et al. （2020)提出 性)之解释技巧的五个诊断属性，我们只考虑忠实性，因为我们发现其他四个属性在我们的方法中或者没有意义或者不适用(见附录 B).

我们的可信度分析只考虑了简化的 25 设置，因为我们对数据稀缺情况下的改进感兴趣，并且相信 MAW 和 GI 的可信度在所有三个数据设置中相对恒定。为了衡量特征归属方法的可信度，我们使用 Atanasova et al. （2020).即对于所有 ψ 0，10，。。。，100，我们屏蔽每个输入示例中最重要的(由归属方法确定的)ψ%的标记，并计算所有示例的结果宏 F1。该阈值-性能曲线下的面积(AUC-TP)为我们提供了忠诚度的反向测量；直觉上，如果归因方法是可信的，那么模型性能主要依赖于该方法所建议的最重要的表征，导致 AUC-TP 较低。作为基线，我们还使用随机屏蔽(相当于评估随机属

图 1:先前 bin 的阈值性能曲线:黄金(上)、基础(中)和未调整的模型(下)。较低的 AUC-TP 相对于 random 表明更忠实的属性。

我们发现 MAW 在先前 bin:gold 和 base 的忠实度方面都超过 GI，并且明显优于随机属性(见图 1 表 4).这表明，总的来说，MAW 属性是忠实于我们的模型的，因此对于我们的目的来说，作为模型推理的解释是可信的。换句话说，我们可以理直气壮地将 MAW 属性解释为推理。

	评估分数	归因
神谕		过去五年我一直住在巴西(过去 27 年断断续续)。 我知道这里没有一个人对奥运会有丝毫的兴奋。人们似乎不在乎。经济一蹶不振，政府完全瘫痪。我们现在有更重要的事情要考虑。
前箱:黄金	一致性:2 充足性:2 无关性:0	过去五年我一直住在巴西(过去 27 年断断续续)。我知道这里没有一个人对奥运会有丝毫的兴奋。人们似乎不在乎。经济衰退，政府陷入困境 完全停止。我们有更重要的 我们现在想的事情。
基础	一致性:0 充分性:0 不相关:2	过去五年我一直住在巴西(过去 27 年断断续续)。我知道这里没有一个人对奥运会有丝毫的兴奋。人们似乎不在乎。经济衰退，政府陷入困境 完全停止。我们现在有更重要的事情要考虑。

表 5:模型基本原理的人类质量判断以及与 oracle 属性的比较。

对基本原理的人工评估:对于先验 bin:gold 和 base，我们选择在测试集上具有最低先验损失 p 的随机种子。我们抽样了 40 个测试集示例，其中两个模型都预测了正确的姿势，并要求 NLP 研究人员对模型的 MAW 属性进行评级:

1. 一致性:解释与立场标签的匹配程度如何
2. **充分性:是否选择了足够的单词[5] 人类可以推断出**
3. 不相关:选择不相关单词的频率。

评级是在五分制的李克特量表上进行的。

	前箱:黄金	基础
一致性↑充分性 不相关↓	**1.12** **1.18** **1.06**	0.18 0.18 1.71

表 MAW 属性的平均人类评价分数(0-2 分)。↑表示高分优先(↓，低分)。

	✓ ω *MAW* —	✓ ω *GI*
一致性	0.241	**0.317**
充足性	0.296	**0.328**
不相关↓	**-0.310**	-0.175

表 MAW 属性和根的 hu- man 评估√结果之间的皮尔逊相关系数

每个例子由两个注释者评分。克里彭多夫的阿尔法(Krippendorff,1980)对于一致性、充分性和不相关性是 0.316，0.311，
和 0.136。分数平均值被映射
以 0-2 的比例进行分析。就像我们的忠诚

分析，我们只关注减少的 25 设置。

我们发现，在所有三个问题上，前仓:黄金的表现都远远超过基础(见表 6)，证明了在数据稀缺的情况下，我们的归因先验对条件模型推理非常有效。

此外，我们调查了理性损失是否是理性质量的可靠代表。具体来说，我们计算了根本原因损失(ω)和平均人类评估得分之间的相关性。我们使用根损失，因为我们的比值损失 ω 直观上是加权均方误差(4.2).我们计算两者的相关性

[5]参考附录 C 有关我们如何生成属性可视化的信息。

基本原理损耗(ω)，后者针对两个 GI 显示和 MAW 属性。↑表示优先选择正相关(←，负相关)。

基于 MAW 的 MAW)和基于 GI 的(√)
(√) ω GI

我们的模型失去了根本原因。我们发现这两个指标都与人类的判断相关；然而，虽然 MAW 是一个更好的不相关指标， ω GI 是一个更好的一致性和充分性指标(见表 7).这表明，在归因先验的背景下，不同的归因方法更适合于在模型基础上实施不同的质量，因此应该相应地选择归因方法。

基本原理的计算评估:就基于最大允许工作时间的基本原理损失(ω 最大允许工作时间)而言，使用我们的归因先验进行训练在所有数据设置中都产生了统计学上的显著优势(见表 8).就 ω GI 而言，用我们提出的先验训练

的模型在中等程度上表现最好

		全部		零射击		少数镜头	
		咽喉	血糖指数	咽喉	血糖指数	咽喉	血糖指数
全部	基础	.112	**.121**	.833	**.910**	.972	**1.038**
	前箱:黄金	**.107***	.122	**.805***	.948	**.924***	1.067
减少 25	基础	.112	.122	.827	.894	.967	1.058
	前箱:黄金	**.106***	**.120**	**.803***	**.876***	**.919***	**1.051**
减少 10	基础	.113	**.119**	.831	**.877**	.977	1.042
	前箱:黄金	**.120***	.120	**.805***	.917	**.933***	**1.039**

表 MAW 和 GI 在 VAST 测试集上的 λ p，所有设置的 λ = 16384。*表示统计显著性(p < .05)。先验损失是在所有的整个测试集上以及在零镜头和少镜头的单词重要性注释子集上计算的。

		顽童		*褪黑素*		*复式造表服务处（MultipleListingService）*		*快速反应按键*		*红新月会*	
		咽喉	血糖指数	咽喉	血糖指数	咽喉	血糖指数	咽喉	血糖指数	咽喉	血糖指数
基础	我	.890	.820	1.187	1.049	.938	.841	.818	.758	.909	.831
	0	.927	.854	.890	.825	.889	.852	.966	.888	**.961**	.921
先前-宾:黄金	我	**.869**	**.789**	**1.177**	**1.035**	**.917**	**.807**	**.800**	**.732**	**.888**	**.793**
	0	**.914**	**.811**	**.874**	**.782**	**.882**	**.804**	**.952**	**.841**	.976	**.868**

表 9:简化 25 数据设置中挑战性现象的 λ p。I 表示包含该现象的例子，0 表示不包含该现象的例子。所有结果都具有统计学意义(p < .05)。先前损失的计算如表所示 8。

数据稀缺的设置，但当使用完整的列车组时，数据不足。这可能表明，在完整的设置中，就一致性和充分性而言，先验黄金的理性比基础黄金的理性差。因此，当训练集的不足部分被赋予 oracle 属性时，我们的属性先验可能对模型推理有不利影响。

5.5 误差分析

挑战性现象:我们还考察了 VAST 中确定的五种挑战性现象的表现:Imp——论点中没有主题短语，标签是非中性的，mlT——论点出现在多个例子中(每个例子都有不同的主题)，mlS——论点出现在具有不同的非中性立场标签的多个例子中，Qte——论点包含引用，以及 Sarc——论点包含讽刺。我们发现，虽然使用我们提出的归因先验进行训练在这些现象上产生了可比较的性能，但它为所有五种现象提供了更好的理由(见表 9).这显示了我们的方法在不降低性能的情况下提高困难例子的合理性的有效性。

基本原理错误类型:我们分析 prior-bin:gold 产生的基本原理中的错误。

具体来说，我们随机抽取了 50 个例子，模型预测这些例子的标签不正确，并将其手动分类为:amounterr 所选单词数量的错误(即选择太少或太多)，contenterr 所选单词内容的错误（即遗漏否定或短语的关键部分)，complexerr 无法理解复杂的语言(例如讽刺或对主题的不明确引用)，以及 dataerr 数据注释中的错误(即错误的标签或无意义的主题）。语义错误（contenterr 和 complexerr)发生在 68%的情况下(分别为 32%和 36%)。例如，这个模型经常不能理解反问句或者错过重要的否定。此外，我们发现 46%的推理选择了太少或太多的词(例如，在论证中选择了大多数停用词)。最后，我们看到 dataerr 占了 30%的错误。这一分析表明，虽然我们的属性先验改进了语义复杂例子的合理性，但语义理解仍然是未来改进的关键挑战。

6 结论

本文提出了有关姿态检测任务的两个问题:1)对推理与人类推理一致的模型的需要，以及 2)对首先有意义地观察模型的推理的方法的需要。我们发现

在一个模拟的数据匮乏的环境中，我们的归因先验使用大量的众包注释改进了模型的合理性。我们还发现，最近备受批评的基于注意力的解释，为我们的模型行为提供了忠实的解释，比高级别的替代方法更有价值。

在未来的工作中，我们计划将我们的方法应用于更具挑战性的设置，如多语言零射击姿态检测。我们还将进一步调查我们方法的“经济性”——例如，有意义地改进模型推理所必需的注释例子的数量——以及用更广泛的属性方法进行实验，例如，引导反向传播(Springenberg et al.,2015)和石灰(Ribeiro et al.,2016).最后，我们希望研究如何对模型推理进行条件化，以防止敌对攻击。

承认

我们感谢 Kathleen McKeown，哥伦比亚自然语言处理小组，以及匿名评论者的评论。这项工作得到了国家科学基金会研究生奖学金 DGE-1644869 的资助。美国政府被授权为政府目的复制和发行再版，不附带任何版权标记。本文包含的观点和结论是作者的观点和结论，不应被解释为必然代表美国政府或 NSF 的官方政策或认可。

7 道德声明

我们使用由收集和分发的数据集 All-away and McKeown（2020):https://github.com/emilyallaway/zero-shot-stance。数据是从《纽约时报》文章的公开评论中收集的。评论中不会保留任何用户信息，因此数据中不包含原作者的种族、性别或民族的明确信息。对于我们收集的额外注释，我们以每小时 13 美元的价格补偿工人，高于美国联邦最低工资(许多注释者都在美国)。

我们讨论的一些方法旨在提供预测立场标签时的模型透明度，包括敏感话题(例如，宗教信仰)。当使用这些方法提供 ex-时解释对于对文本的预测，真实世界的用户应该被告知解释是自动生成的，可能不代表文本作者的全部观点。

参考

Rob Abbott、Brian Ecker、Pranav Anand 和 Mari- lyn A. Walker。2016.互联网辩论语料库 2.0:对话社会媒体的 sql 模式和与之配套的公司。在 LREC。

周欣宇·阿布纳和 w·苏迪马。2020.量化变压器中的注意力流。ArXiv，abs/2005.00928。

艾米莉·阿拉维和凯瑟琳·麦克欧文。2020.零射击姿态探测:使用通用主题表示的数据集和模型。在 EMNLP 中。

艾米莉·阿拉威、马拉维卡·斯里坎特和 k·麦克欧文。2021.社交媒体上零射击姿态检测的对抗性学习。ArXiv，abs/2105.06603。

佩帕·阿塔纳索娃、雅各布·格鲁·西蒙森、克里斯蒂娜·李·奥马和伊莎贝尔·奥根斯坦。2020.诊断文本分类的可解释性技术研究cation。《2020 年自然语言处理经验方法会议论文集》(EMNLP)，第 3256-3274 页，在线。计算语言学协会。

Isabelle Augenstein、Tim rocktschel、Andreas Vla- chos 和 Kalina Bontcheva。2016.采用双向条件编码的姿态检测。在 EMNLP 中。

南巴赫，亚历山大·宾德，格雷瓜尔·蒙塔冯，F.klauschen k . müller 和 W. Samek。2015.通过逐层相关性传播对非线性分类器决策的逐像素解释。公共科学图书馆一号，10。

J.德芙琳，张明蔚，肯顿·李和克里斯蒂娜·图塔诺瓦。2019.Bert:用于语言理解的深度双向转换器的预训练。在 NAACL-HLT 中。

Jay DeYoung、Sarthak Jain、Nazneen Fatema Rajani、Eric Lehman、Xiong、Richard Socher 和 Byron C. Wallace。2020.橡皮擦:基准评估合理化的 NLP 模型。《计算语言学协会第 58 届年会论文集》，第 4443-4458 页，在线。计算语言学协会。

安卡·杜米特拉凯，瓦娜·伊内尔，劳拉·阿罗约，本杰明·蒂默曼斯和克里斯·韦尔蒂。2018.拥挤的真相 2.0:众包的质量标准 agreement。更正，abs/1808.06080

G.Erion、J. Janizek、Pascal Sturmfels、Scott M. Lundberg 和 Su-In Lee。2019.利用归因先验学习可解释模型。ArXiv，abs/1906.10670。

G.Erion、J. Janizek、Pascal Sturmfels、Scott M. Lund- berg 和 Su-In Lee。2020.利用公理化属性先验和期望梯度提高深度学习模型的性能。arXiv:学习。

Amirata Ghorbani、Abubakar Abid 和 James Zou。
2018.神经网络的解释是脆弱的。

Momchil Hardalov、Arnav Arora、Preslav Nakov 和 Isabelle Augenstein。2021.跨域标签自适应姿态检测。ArXiv，abs/2104.07467。

卡齐·赛义德·哈桑和翁清海。2014.你为什么采取这种立场？意识形态辩论中原因的识别和分类。在 EMNLP 中。

伯尼斯·赫尔曼。2017.人类对模型可解释性评价的希望与危险。ArXiv，abs/1711.07414。

阿龙·贾科维和约夫·戈德堡。2020.走向完全可解释的 nlp 系统:我们应该如何定义和评估忠实度？在 ACL 中。

阿龙·贾科维和约夫·戈德堡。2021.Aligning 忠实解释及其社会归因。计算语言学协会汇刊，9:294–310。

萨尔萨克·贾恩和拜伦·华莱士。2019.注意不是解释。在 NAACL-HLT 中。

Teja Kanchinadam，Keith Westpfahl，Qian You 和 Glenn M. Fung。2020.基于理性的人在回路中通过监督关注。在达什@KDD。

克劳斯·克里彭多夫。1980.内容分析及其方法论介绍。SAGE。

Mirko Lai、Alessandra Teresa Cignarella、D. I. H. Farías、C. Bosco、V. Patti 和 P. Rosso2020.社交媒体政治辩论中的多语言立场检测。电脑。演讲郎。，63:101075.

弗雷德里克·刘和贝西姆·阿弗奇。2019.合并基于特征属性的文本分类先验知识。《计算语言学协会第 57 届年会论文集》，6274-6283 页，意大利佛罗伦萨。计算机语言学协会。

赛义夫·穆罕默德、斯维特拉娜·基里琴科、帕里纳兹·索巴尼、朱小丹和科林·切里。2016.Semeval-2016 任务 6:检测推文中的姿态。在 SemEval@NAACL-HLT。

Saif M. Mohammad、Parinaz Sobhani 和 Svetlana Kiritchenko。2017.推文中的立场和情绪。ACM Trans。互联网技术。，17:26:1–26:23.

Danish Pruthi，Bhuwan Dhingra，Graham Neubig 和 Zachary C. Lipton。2020.弱和半弱监督证据提取。计算语言学协会的发现:EMNLP 2020，3965-3970 页，在线。计算语言学协会。

马尔科·图利奥·里贝罗、萨梅尔·辛格和卡洛斯·盖斯特林。2016.“我为什么要相信你？”:解释任何分类器的预测。第 22 届 ACM SIGKDD 知识发现和数据挖掘国际会议论文集。

本杰明·席勒、约翰内斯·达辛伯格和伊琳娜·古雷维奇。2020.姿态检测基准:你的姿态检测有多强？ArXiv，abs/2001.01565。

Jost Tobias Springenberg、A. Dosovitskiy、T. Brox 和 Martin A. Riedmiller。2015.追求简单:全卷积网。更正，abs/1412.6806。

穆孔德·孙达拉扬、安库尔·塔利和奇奇·延。2017.深度网络的公理化属性。《第 34 届机器学习国际会议论文集》第 70 卷，ICML，2017 年，第 3319–3328 页。JMLR.org。

弗朗西斯科·玛丽奥纳·陶勒、玛丽亚·安托尼亚·马蒂
米（meter 的缩写））兰格尔·帕尔多、保罗·罗索、克里斯蒂娜·博斯科和薇薇安娜·帕蒂。2017.Catalan independence 上 tweets 中的姿态和性别检测任务概述。在伊比利亚的@SEPLN。

Adam Tsakalidis、Nikolaos Aletras、Alexandra I. Cristea 和 Maria Liakata。2018.社交媒体用户在突然投票中的立场预测:希腊公投案例。第 27 届 ACM 信息和知识管理国际会议录。

詹尼斯.瓦姆瓦斯和里科.森里奇。2020.X-stance:用于姿态检测的多语言多目标数据集。ArXiv，abs/2003.08385。

玛丽莲·a·沃克、让·e·福克斯·特里、普拉纳夫·阿南德、罗布·艾伯特和约瑟夫·金。2012.审议和辩论研究文集。在 LREC。

莎拉·威格雷夫和尤瓦尔·品特。2019.注意力是不是不是解释。《2019 年自然语言处理经验方法会议和第九届国际自然语言处理联合会议(EMNLP-IJCNLP)论文集》，第 11-20 页，中国香港。计算语言学协会。

特里萨·威尔逊，简尼斯·威伯和克莱尔·卡迪。2017.MPQA 意见文集，第 813-832 页。斯普林格。

郑玄吴和戴斯蒙王。2021.On 解释你对伯特的解释序列分类临床研究。更正，abs/2101.00196。

徐畅、塞西尔·帕里斯、苏亚·尼帕尔和罗斯·斯帕克斯。2018.基于自我注意网络的跨目标姿态分类。在 ACL 中。

奥马尔·扎伊丹、杰森·艾斯纳和克里斯汀·皮亚特科。2007.使用“注释器原理”来改进机器

文本分类学习。人类语言技术 2007:计算语言学协会北美分会会议:主要会议记录，260-267 页，罗切斯特，纽约。计算语言学协会。

B.张，杨，，李，叶云明，，徐，戴。2020.利用可转移的语义情感知识增强跨目标姿态检测。在 ACL 中。

张烨、伊恩·马歇尔和拜伦·华莱士。2016.基本原理增强卷积神经网络对于文本分类。《2016 年自然语言处理经验方法会议论文集》，第 795-804 页，德克萨斯州奥斯汀。计算语言学协会。

钟瑞琪、邵和麦克欧文。2019.具有忠实关注的细粒度情感分析。ArXiv，abs/1908.06870。

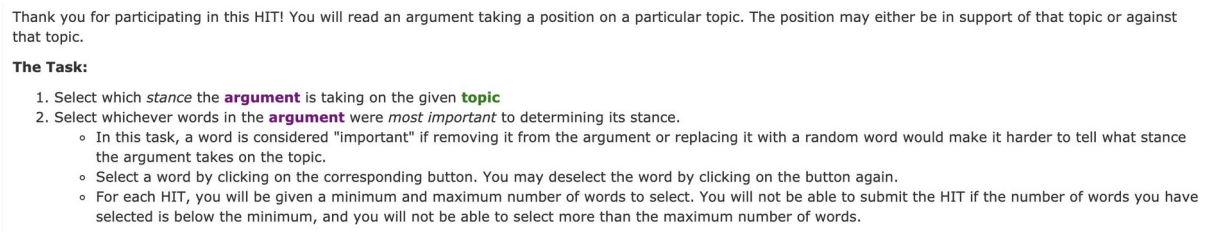


图 2:命中指令。

A 众包

我们付给每个工人每次点击 0.28 美元。我们观察到，工作人员在每次点击上平均花费大约 75 秒，排除异常值(在这种情况下，工作人员比其他两个工作人员花费更长的时间进行点击，可能是由于工作流中断)。只有包含至少一个字母字符的标记是可选的。对于每次点击，工作人员需要选择至少一轮(num_selectable/11)个标记，最多一轮(num_selectable/5.5)(其中 num_selectable 是特定点击中的可选单词数)。见图 2 图 3 图 4。

Frequently Asked Questions

Should I select all words in an important phrase or leave out common words like 'a', 'the', 'of', etc.?

In most cases, selecting common words such as "a," "the," and "of" shouldn't be a *priority*. Sometimes, however, a common word is critical to a phrase. Consider the sentence, "I like to go to the community garden and take in the flowers." Removing the word 'in' from the phrase "take in" would drastically change the meaning of the sentence! Of course, if the phrase itself is not important, none of its words should be selected.

What should I do if the paragraph takes a stance on something *related* to the topic, but not on the topic itself? For instance, if the topic is "gambling" but the argument is about gambling laws and regulations?

Some of the topics you see will not match perfectly with their respective arguments. In such cases, make your best guess as to what the "intended topic" is and then treat that as the topic. The intended topic will always be somewhat substitutable with the given topic; for instance, if the topic is "children," then the intended topic might be "having children," but it won't be "child abuse." This is because "pro-children" could be taken to mean "pro-having-children," but it cannot be interpreted as "pro-child-abuse." Similarly, if the paragraph argues that trans fats should be banned, and the given topic is "unhealthy," then the intended topic is probably something like "trans fats are unhealthy" and the correct stance is 'For.' In the "gambling" example, it would be appropriate to assume that the intended topic is something like "unregulated gambling."

If the topic appears in the argument, should I select it?

That depends entirely on the context in which it appears. If you were to replace that occurrence of the topic in the paragraph with a random word or phrase, would it be significantly harder to figure out the author's stance? If so, then you should select the topic. If not, you should leave it unselected.

If the author presents a statement for the purpose of rejecting it, should I select the important words in that statement?

Only if the rest of the paragraph isn't enough to determine the author's stance on the topic. If you do select words in a part of the paragraph that goes against how the author feels, make sure you also select words that indicate the author's disagreement with that piece of text.

图 3:点击常见问题。

Topic: vaccination

Argument: We still haven't addressed the problem of children who come from other countries into America and have not had the advantage of early childhood vaccinations. We are seeing cases of measles, mumps, etc. increase, and this cannot be all the fault of those who have lived with and been in contact daily in this country since birth with their peers who have been vaccinated. I'm old, and when in 1st grade, I contacted German measles. NOT from someone in my class and, in fact, no one seemed to know where I contacted it. I then proceeded to infect almost all my entire class apparently. My point in telling this story is that we cannot know exactly where a child gets his contact, and therefore IMO opinion, vaccinations should be mandatory for every child entering schools from pre-K up. Until everyone - from the child who comes with their parents to America to the kiddo next door, we have to be consistent.

Question 1. What stance does the argument take toward the topic?

○ **For**

○ **Against**

Question 2. Which words in the argument are most important to identifying its stance on "vaccination?" Please select 17-34 words.

We | still | have | n't | addressed | the | problem | of | children | who | come | from | other | countries | into | America | and | have | not | had | the | advantage | of | early | childhood | vaccinations | . | We | are | seeing | cases | of | measles | , | mumps | , | etc | . | increase | , | and | this | can | not | be | all | the | fault | of | those | who | have | lived | with | and | been | in | contact | daily | in | this | country | since | birth | with | their | peers | who | have | been | vaccinated | . | I | 'm | old | , | and | when | in | 1st | grade | , | I | contacted | German | measles | . | NOT | from | someone | in | my | class | and | , | in | fact | , | no | one | seemed | to | know | where | I | contacted | it | . | I | then | proceeded | to | infect | almost | all | my | entire | class | apparently | . | My | point | in | telling | this | story | is | that | we | can | not | know | exactly | where | a | child | gets | his | contact | , | and | therefore | IMO | opinion | , | vaccinations | should | be | mandatory | for | every | child | entering | schools | from | pre | - | K | up | . | Until | everyone | - | from | the | child | who | comes | with | their | parents | to | America | to | the | kiddo | next | door | , | we | have | to | be | consistent | .

2 words selected

图 4:点击示例。

B 其他诊断属性

当评估 GI 和 MAW 作为我们模型的可解释技术时，我们选择不考虑 Atanasova et al.（2020）.与人类注释(HA)的一致性不一定是一个令人满意的属性，因为它只不过是一个归属方法对人类有多大说服力的指示。请注意，我们以比率损失的形式计算 HA(5.4)，但这样做是作为评价归因本身的一种方式，而不是归因方法。置信度指示(CI)不适用于 MAW，因为对于固定输入，注意力权重不会因类别而异。作者对推理一致性(RC)的度量要求假设具有相似推理路径的模型具有相似的激活图，我们认为这一假设主要是由于体系结构的对称性而存在缺陷。最后，我们认为数据集一致性(DC)的建议度量对我们的数据集没有意义，因为在 VAST 中不同参数之间的相似度非常低。

C 可视化属性

为了可视化图形、人工评估和基本原理错误分析的模型属性，我们将每个标记的属性分数映射到一个新的分数 0(未选择)、0.5(已选择，但只是中等重要)或 1(已选择且非常重要)。我们使用下面的过程来执行这种映射，该过程采用参数 k 和 ϵ:

1. 按降序排列输入序列的属性分数，k_score 为第 k 项的分数。
2. 分配分数> k 分数+ ϵ 的所有代币新的 1 分。
3. 分配分数< k_score ϵ 的所有代币新的 0 分。
4. 给所有其他令牌分配 0.5 的新分数。

对于长度为 m 的自变量，我们设 k = m/8。我们让 ϵ = .05 ∗ max_att，其中 max_att 是该论点的原始归因得分的最大值。我们通过对训练示例的反复试验获得这些值，主观目标是获得包含大量“适度重要”和“非常重要”单词的视觉效果，同时反映原始归因得分。我们将多标记单词的新得分作为其子单词标记的最大新得分。

λ	全部	减少 25	减少 10
0	.726	.710	.699
16,384	*不适用的*	.701	**.703**
32,768	.712	.698	.702
49,152	**.726**	**.712**	.695
65,536	.711	.703	*不适用的*

表 10:每个数据设置中不同 λ(平均间隔 214 = 16384)的 Dev 设置结果。 λ = 0 表示没有应用我们的属性先验。使用单个随机种子。n/a 表示未进行试验。

D 选择 λ

见表 10。

E 试验间的差异

见表 11 表 12。

		全部	零	很少的
全碱		6.5	2.2	29.1
	p-b:g	28.0	23.4	31.0
减少 25	BT-j	15.9	24.6	8.0
	基础	8.2	6.6	11.2
	p-b:g	10.2	3.2	20.1
减少 10	BT-j	11.4	16.9	7.4
	基础	103.0	120.1	90.1
	p-b:g	6.9	5.4	4.9

表 11:表中报告的 F1 组合结果(“全部”)的试验差异 3，乘以 105.p-b:g 指 prior-bin:gold (BT-j，BERT-joint)。

F 杂项

我们的模型包含 109，917，780 个参数。使用我们提出的属性先验在整个训练集上进行训练，使用两个特斯拉 T4 GPU 需要 11 小时 16 分钟。

		顽童		*褪黑素*		*复式造表服务处（MultipleListingService）*	*快速反应按键*			*红新月会*	
		咽喉	血糖指数	咽喉	血糖指数	咽喉	血糖指数	咽喉	血糖指数	咽喉	血糖指数
基础	我	5.2	2.5	10.2	5.6	10.2	6.0	7.1	5.0	90.2	30.7
	0	6.8	19.2	1.0	0.0	11.3	4.6	100.9	90.1	5.6	0.3
先前-	我	0.0	0.0	0.1	1.5	0.0	5.6	0.0	4.0	0.5	3.2
宾:黄金	0	0.0	2.0	0.0	0.0	0.0	16.9	0.0	40.5	0.0	3.6

表 12:表中报告的√ λ ~~p 试~~验的差异 9，乘以 105。