走向以解释为潜在变量的可解释的自然语言理解

周王春书[1*] 胡金逸[2*] 张翰林[3*]
梁晓丹[4] 毛松，孙孙[2] 熊辰岩[5] 健唐[6,7]
[1]北京航空航天大学[2]清华大学[3]华南理工大学
[4]中山大学[5]微软的研究中心[6]米兰-魁北克省
人工智能研究所[7]蒙特利尔中心周光春州。
edu。cn公司
hujy17@mails.tsinghua.edu。cn公司短信。edu。cn公司
{赫尔张109，x德梁328}@gmail。com网站
智利的人。雄@微软。com网站
jian.tang@hec.ca

摘要

最近生成的自然语言解释不仅提供了非常好的结果，不仅提供了可解释的解释，而且为预测提供了额外的信息和监督。然而，现有的方法通常需要大量的人类注释解释，而收集大量的解释不仅耗时而且很昂贵。在本文中，我们开发了一个可解释的自然语言理解的一般框架，它只需要一小部分人类注释的解释来进行训练。我们的框架将自然语言解释视为建模神经模型的潜在推理过程的潜在变量。我们开发了一个变分优化EM框架，其中一个解释生成模块和一个解释增强预测模块交替进行优化和相互增强。此外，我们还提出了一种在此框架下基于解释的半监督学习自我训练方法。它在为未标记的数据分配伪标签和生成新的解释以相互迭代改进之间交替使用。在两个自然语言理解任务上进行的实验表明，我们的框架不仅可以在监督和半监督环境下做出有效的预测，而且还可以产生良好的自然语言解释[2]。

1 产品介绍

为自然语言理解构建可解释的系统在医疗保健和金融等各个领域都至关重要。一个有希望的方向是为预测生成自然语言解释[1-4]，这最近被证明非常有希望，因为它们不仅可以为后件箱预测系统提供可解释的解释，还可以为预测提供额外的信息和监督[5-7]。*例如，给出一句话：“唯一比食物更美妙的就是服务。”*，人类注释者可能会写出像“积极”这样的解释，因为“美妙”这个词出现在食物这个词之前的三个单词内”，这比“积极”这个标签在解释它是如何做出决定时，更有信息。此外，该解释还可以

[*] 同等的贡献，其顺序由掷骰子决定。这些工作是在米拉大学实习期间完成的。
[2]代码可在https://github.com/JamesHujy/ELV.git上获得

第34届神经信息处理系统会议(NeurIPS2020)，加拿大温哥华。

作为一个隐含的逻辑规则，可以推广到其他例子，比如“食物很棒，我真的很喜欢它。””

[3,4]最近的一些研究为预测生成自然语言解释和/或利用生成的解释作为预测的额外特征。例如，坎姆布鲁等人。[3]通过在一个带有注释人类解释的语料库上训练一个语言模型，训练自然语言推理任务的一般自然语言解释。c提出了一个常识推理的两阶段框架，首先训练一个自然语言解释模型，然后进一步训练一个以生成的解释作为附加信息的预测模型。这些方法在预测性能和可解释性方面都取得了良好的性能。然而，需要大量带有人类解释的标记例子，这是昂贵的，有时不可能获得。因此，我们正在寻找一种方法，使有效的预测，提供良好的可解释性，但需要有限数量的人类解释的训练。

本文提出了这种方法。我们从直觉开始，解释增强预测模型能够提供信息反馈，以生成有意义的自然语言解释。因此，与现有的在不同阶段训练解释生成模型和解释增强预测模型的工作不同，我们建议联合训练这两个模型。具体来说，把文本分类的任务作为一个考试-
因此，我们提出了一个原则性的文本分类概率框架，其中是自然语言
解释被视为潜在变量(ELV)。变分EM[8]用于优化，只需要一组人工解释来指导解释生成过程。在e步中，解释生成模型被训练为近似的地面真实解释-
方案（对于有注释解释的实例)或通过后验推理由解释增强模块引导(对于没有注释解释的实例）；在m步中，解释增强预测模型使用从解释生成模型中采样的高质量解释进行训练。这两个模块相互增强。由于人类的解释可以作为隐式的逻辑规则，因此它们可以用于标记未标记的数据。因此，
我们进一步将我们的ELV框架扩展到一个基于解释的自我训练(ELV-EST)模型，用于在半监督设置中利用大量的未标记数据。

综上所述，在本文中，我们做出了以下贡献：

- 我们提出了一个有原则的概率框架，称为ELV的文本分类，其中自然语言的解释被视为一个潜在的变量。它联合训练了一个解释生成器和一个解释增强的预测模型。只需要少数带注释的自然语言解释来指导自然语言的生成过程。
- 我们进一步扩展了ELV的半监督学习(ELV-EST模型)，该模型利用自然语言解释作为隐式逻辑规则来标记未标记的数据。
- 我们在关系提取和情感分析两个任务上进行了广泛的实验。实验结果证明了我们提出的方法在监督和半监督设置下的预测和可解释性方面的有效性。

2 相关的工作

自然语言(NL)解释已经被证明是非常有用的模型解释和预测。一些早期的工作，[11,9,10]利用NL解释作为预测的额外特征。例如，斯里瓦斯塔瓦等人。[9]将NL解释转换为分类器特征来训练文本分类模型。小提琴手等人。[12]使用自然语言的解释来帮助监督图像字幕模型。最近，默蒂等人。[10]建议ExpBERT直接将NL解释与BERT相结合。然而，这些工作中的大多数需要在培训和测试实例中都有解释，这是不现实的，因为注释大量实例的解释是非常耗时和昂贵的。此外，一旦测试数据中给出解释，预测就变得容易。

最近有一些工作研究了训练一个自然语言解释模型，然后使用生成的解释进行预测。例如，坎姆布鲁等人。[3]和Rajani等人。[4]提出分别训练一个模型来生成NL解释和一个分类模型

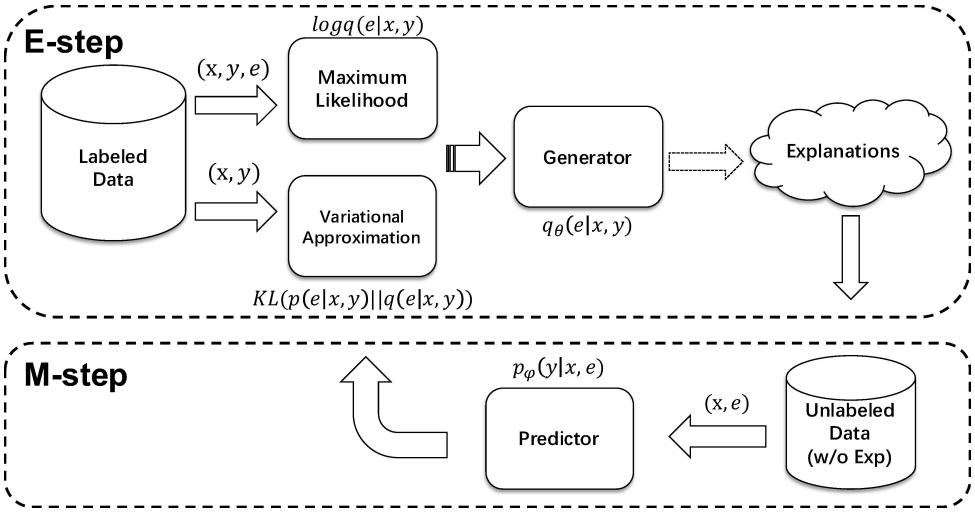

图1：ELV的概述。在e步中，我们训练生成器p(e|x，y)来生成给定标记数据的解释。对于带有注释解释的标记数据。D_E)，我们最大限度地提高解释地面真相的可能性。对于没有解释的标记数据。D_L)，我们最小化变分分布Q之间的KL散度$_\theta$(e|x，y)和地面真实后验p(e|x，y)，借助预测模型计算。在m步中，我们使用e步中生成的解释，用MLE训练预测因子p(y|x，e)。

生成的解释作为额外的输入。他们的方法在提高分类模型的可解释性和通过解释作为附加特征提高预测性能方面非常有前景。然而，他们的方法需要大量的人类注释的NL解释来训练解释生成模型。此外，这些方法不能建模生成NL解释和利用NL解释进行预测之间的相互作用。因此，并不能保证生成的解释反映了预测模型的决策过程或对预测模型有利。据Camburu等人报道。[3]，可解释性以性能损失为代价。在本文中，我们提出了一个以解释为潜在变量的原则性概率框架，通过联合训练自然语言解释模块和解释增强预测模块，以最小化具有解释的训练实例的数量。

另一个相关方向是将自然语言解释作为半监督学习的额外监督，而不是作为附加特征[13,7]。例如，汉考克等人。[13]使用语义解析器将NL解释解析为逻辑形式（即“标记函数”）。然后使用标记函数匹配未标记的很难[13]或软[7]，以生成用于训练模型的伪标记数据集。 然而，这些方法需要以一种可以被语义解析器准确解析的形式来注释解释，以形成标记函数，而这对于许多NLP应用程序可能是不可能的。在我们的半监督框架中，不需要语义解析，自然语言解释用预先训练的语言模型获得的分布式表示来解释，用于标记未标记数据。

3 计算方法学

3.1 问题的定义

给定一个输入句子x，我们的目标是预测其标签y，并生成一个自然语言(NL)解释e，描述为什么x被归类为y。 具体来说，给定了一些使用NL解释注释D的训练例子$_E$ ={(x_1，y_1，e_1，)，...，(x_n，y_n，e_n，)}和一个相对较大的集合
的已标记的示例D_L ={(x_{n+1}，y_{n+1})，...，(x_m，y_m，我们的目标是学习：1)一个解释
发电型号E_θ 参数化q(e|x，y)，它以一个标记的例子(x，y)作为输入，并生成一个相应的自然语言解释e，和2)一个解释增强的预备-

系统模型M_ϕ，它将p(y|x，e)参数化，并采用一个未标记的示例x和NL解释（作为隐式规则）E将标签y分配给x。

3.2 自然语言解释作为潜在变量

给定标记数据(x，y)，我们将自然语言解释e作为一个潜在变量。 对于训练，我们的目标是优化logp(y|x)的证据下界(ELBO)，它可以表述为：

$$\text{日志}p(y|x)=\text{日}\int_{\text{志}} p(e，y|x)de=\text{日}\int_{\text{志}} q_\theta(e/x，y)\frac{p(e，y/x)}{q_\theta(e/x，y)} \quad \text{德尔} \tag{1}$$

$$\geq L(\theta，\phi)=E_{q_\theta(e/x，y)}\text{日志记录}\frac{p(e，y/x)}{q_\theta(e/x，y)}=E_{q_\theta(e/x，y)}\text{日志记录}\frac{p_\phi(y/e，x)p(e/x)}{q_\theta(e/x，y)}， \tag{2}$$

其中的$q_\theta(e|x，y)$是后验分布p(e|x，y)的变分分布，p(e|x)是解释e的先验分布，例如x和$p_\phi(y|，x)$是解释增强预测模型。

由于自然语言解释e的搜索空间很大，因此我们没有使用变分自编码器[14]中使用的再参数化技巧，而是使用变分em算法进行优化。ELV的概述如图1所示。请注意，在我们的训练数据中，对一些有标记的例子(例如，D_E)是提供的。 因此，我们首先初始化解释程序-国家发电模型$E_\theta=q_\theta(e|x，y)$和预测模型$M_\phi=p_\phi(y|、|)$通过培训在D号上$_E$具有最大似然估计值(MLE)。然后，我们通过使用D最大化对数似然对数p(y|x)来更新上述模型$_L$ $\cup D_E$与变分电镜。 在变分e步中，我们训练解释生成器以最小化Q之间的KL散度$_\theta$（e|x，y)和p(e|x，y），详见第3.3节。在m步中，我们修复了θ和p(e|x)，并更新了预测模型的参数ϕ，以最大化对数似然对数p(y|x)。

3.3 E-步骤：解释生成模型

作为我们的变分EM框架的核心组成部分，解释生成模型有望以自然语言解释的形式生成“软”逻辑规则。 然而，训练一个seq2seq模型来从零头生成高质量的解释是非常具有挑战性的。在最近的发现中，预训练的语言模型在其参数中编码各种类型的事实知识和常识知识[15-17]，我们使用了UniLM[18]--一个统一的预训练的语言生成模型，在许多文本生成任务上实现了最先进的性能，作为解释生成模型E_θ在我们的框架中。具体来说，解释生成模型以输入句子x的连接及其对应标签y的文本描述作为输入，生成解释，解释自然语言中的标签决策，可以作为隐式逻辑规则，可以推广到其他例子。

注意，在训练数据中，只提供一小部分标记示例解释。因此，在变分e步骤中，对于没有解释的标记数据(x，y)。 D_L)，我们试图使用变分分布q_θ（e|x，y）近似地面真实后验p（e|x，y），可以计算为

$$p(e/x，y)\sim p_\phi(y/x，e)p(e/x) \tag{3}$$

其中，p_ϕ(y|x，e)由预测模型参数化，并为生成有意义的自然语言解释提供反馈。我们将介绍p(e|x)和p的详细参数化$_\phi$(y|x，e)在|步中。

对于带有解释的标记数据。 D_E)，我们只需要最大限度地提高地面真相解释的可能性。因此，e-步进的整体目标函数可以总结为：

$$O=\sum_{(x，y)\in D_E}\text{日志}q(e|x，y)+\sum_{(x，y)\in D_L}\text{KL}(q(e|x，y)\|p(e|x，y)) \tag{4}$$

3.4 解释步骤：增强预测模型

在m步的过程中，对解释增强预测模型进行训练，利用变分分布q(e|x，y)生成的解释e来预测输入句子x的标签。

但是，请注意，y在测试过程中标签不可用，而未标记x的解释只能从先验的分布p(e|x)生成。因此，由于在训练阶段对标签数据的解释分布不同，而在测试阶段对未标记数据的解释分布存在一些差异，因为生成没有标签条件的自然语言解释更难存在差异。 为了缓解这个问题，在预测模型中，除了从变分分布中抽样解释外，我们还从p(e|x)中添加了一组解释，从类似的句子中检索一组解释。

具体来说，给定一个输入句子x，标记和伪标记的数据集由
(x, e, y)，我们检索到N个解释$E := \{e_i\}_{i=1}$ 其中相应的句子x 都是：
与输入句子x最相似的，通过嵌入x和每个x之间的余弦相似性来衡量' 从第D页开始$_E$ 在句子序列[19]下，一个预先训练好的句子嵌入模型。请注意，我们没有直接使用seq2seq模型来参数化p(e|x)，因为我们发现没有预测标签的生成解释通常会导致不相关的，甚至是误导性的解释。

算法1：基于解释的自我训练（ELV-EST）

输入值：$D_E = \{(x_1, y_1, e_1,), \ldots, (x_n, y_n, e_n,)\}$, $D_L = \{(x_{n+1}, y_{n+1}), \ldots, (x_m, y_m)\}$, 未标记的数据$D_U = \{x_{m+1}, \ldots, x_N\}$, 置信度阈值T
输出值：$E_\theta(e/x, y)$, $M_\phi(y/x, E)$
初始化E_θ 和M_ϕ 与D联系在一起$_E \cup D_L$ 使用ELV
重复操作
  对于每个$x_i \in D_U$ 做些什么
    如果是最大值$_y$ $M_\phi(y/x, E)$*>T然后*
      分配伪标签y_i 发送到x_i 并生成解释e_i 与E连接在一起$_\theta$
      更新$D_L = D_L \cup (x_i, y_i)$
      更新$D_E = D_E \cup (x_i, y_i, e_i)$ （用于解释检索）
      更新$D_U = D_U \setminus x_i$
    末
  端末端
  列车E_θ 和M_ϕ 在D号上$_E \cup D_L$ 与ELV在一起使用
直到收敛性或$D_U = \varnothing$

让$E=\{e_1, \ldots, e_n\}$表示对x的所有解释。 对于每个$e_i \in E$，我们提供了解释e_i 和输入句子x，由一个[SEP]标记分隔，到BERT[20]，并使用[CLS]标记处的向量来表示x和e之间的相互作用$_i$ 作为一个768维的特征向量：

$$I(x, e_i) = \text{BERT}([[\text{CLS}]; x; [\text{SEP}]; e_i]) \tag{5}$$

我们最终的分类器将这些向量的连接起来，并输出最终的预测为：

$$M_\phi(y|x, E) = \text{MLP}[\text{平均}(I(x, e_1); I(x, e_2); ; I(x, e...n))] \tag{6}$$

在测试时，对于每个未标记的x，我们首先使用p(e|x)检索一组解释，然后使用解释增强预测模型预测一个标签。 然后，我们可以进一步使用解释生成模型来生成一个NL解释来解释预测决策
基于输入的句子和预测的标签。

总之，通过在e步和m步之间交替，其中$q_\theta(e/x, y)$*和*$p_\phi(y|, |)$分别进行了优化，解释生成模型E_θ 以及解释增强的预测
模型M_ϕ 是共同优化和相互增强。接下来，我们将描述如何将我们的框架应用于其中人类注释的解释和基本真相标签都是有限的半监督设置。

3.5 基于解释性的自我培训

由于自然语言解释可以作为隐式逻辑规则，可以推广到新数据，并帮助为未标记的数据分配伪标签。因此，我们将ELV扩展到半监督学习设置，并提出了一种基于解释的自训练(ELV-EST)算法。在这种情况下，我们只有有限的标记示例，但有大量的未标记数据$D_U = \{x_{m+1}, \ldots, x_N\}$.

表1：数据集的统计数据。 我们展示了在监督和半监督设置中的4个数据集的训练/dev/测试集的大小。 此外，#Exp表示初始解释集的大小。

数据集	#的说明	#列车（受监督）	#列车（半监督系统）	#开发人员组	#测试
半椭圆形的[21]	203	7,016	1,210	800	2,715
接触的[22]	139	68,006	2,751	22,531	15,509
笔记本电脑	70	1,806	135	462	638
餐厅	75	2,830	107	720	1,120

如算法1所示，我们首先使用ELV来初始化E_θ 和M_ϕ 与有限的标记语料库$D_E \cup D_L$。之后，我们迭代地使用M_ϕ 为D中的未标记的示例分配伪标签D_U 以扩展已标记的数据D_L. 然后我们使用ELV来联合训练E_θ 和M_ϕ 与增强的标记数据集。与此同时，我们也使用了E_θ 生成未标记的新解释
例子及其伪标签。通过这种方式，我们可以用未标记的例子获取大量的伪标签和伪解释。伪标记的例子可以用来改进模型，同时也使我们能够生成更多的NL解释。作为回报，新生成的解释不仅可以改进解释生成模型，而且还可以作为隐式规则，帮助预测模型在下一次迭代中分配更准确的伪标签。

所提出的ELV-EST方法在两个方面与传统的自我训练方法有所不同。首先，除了预测未标记数据的伪标签外，我们的方法还发现了自然语言解释形式的隐式逻辑规则，这有助于预测模型更好地为未标记数据分配噪声标签。其次，我们的方法可以产生可解释的预测与E_θ. 与最近将解释解析为逻辑形式的工作[7,13]相比，我们的方法不需要特定于任务的语义解析器和匹配模型，这使得它与任务无关，适用于各种自然语言和最少的额外努力理解任务。

4 实验结果

4.1 数据集

我们执行了两个任务（RE关系提取）和基于方面的情绪分类(ASC)。对于关系提取，我们选择了两个数据集，TACRED[23]和SemEval[21]。我们使用两个客户审查数据集，餐厅和笔记本电脑，它们是SemEval2014Task4[24]的一部分，用于基于方面的情绪分类任务。我们使用在[7]中收集的人道注释的解释来训练我们的基于解释的模型。

4.2 实验性的设置

我们在监督设置中进行实验，我们可以访问数据集中所有标记的例子，在半监督设置中，我们只使用一小部分标记的例子，并通过忽略原始数据集中其余标记的例子作为他们的标签作为未标记的例子。在这两种情况下，只有少数人类注释的NL解释可用。
解释的数量、在监督/无监督设置中使用的标记数据以及数据集的统计数据如表1所示。

我们分别使用bert基和unilm基作为预测模型和解释生成模型的主干。我们选择批处理大小超过{32,64}和学习速率超过
{1e-5,2e-5,3e-5}。对于所有任务，检索到的解释的数量都被设置为10个。我们训练的是3个电磁迭代的预测模型和5个电磁迭代的生成模型。我们使用的是
Adam优化器和早期停止与最佳验证f1分数。

4.3 比较的方法

在监督设置下，我们将ELV与bert基线进行比较，该基线直接在目标数据集上微调预先训练的bert基模型。为了说明对解释生成模型和解释增强预测之间的交互作用进行建模的重要性

表2：监督设置下关系提取数据集的结果（Micro-F1）。

计算方法	卷的	半圆形的
伯特公司他们的[25]	66.3	76.9
伯特公司cm+结核分枝杆菌[25]	67.1	77.5
大的大	66.4	78.8
BERT-基地	64.7	78.3
ELV（仅限M步）	65.4	80.2
ELV（我们的）	65.9	80.7

表3：监督设置下ASC数据集的结果（宏F1）。

计算方法	餐厅	笔记本电脑
asgcn[26]	72.2	71.1
伯特-[27]	77.0	75.1
伯特中心[28]	77.0	75.0
BERT-基地	75.4	72.4
ELV（仅限M步）	76.2	74.1
ELV（我们的）	77.8	75.2

表4：半监督设置下关系提取数据集的结果（Micro-f1）。

计算方法	卷的	半圆形的
BERT-基地	25.1	49.3
伪标记[29]	28.6	50.2
自我培训的[30]	36.9	59.5
数据编程的[13]	25.8	47.9
ELV-EST（我们的）	42.5	66.4

表5：半监督设置下ASC数据集的结果（Macro-f1）。

计算方法	餐厅	笔记本电脑
BERT-基地	32.2	34.6
伪标记[29]	42.5	38.2
自我培训的[30]	47.2	42.3
数据编程的[13]	38.2	36.3
ELV-EST（我们的）	59.5	63.6

我们还与我们的模型的一个变体进行了比较，该模型只训练解释增强预测模块与先验分布生成的所有解释，表示为ELV(仅m步))。我们还比较了一些最先进的算法在RE和SA任务。

在半监督设置下，我们将ELV-EST与几种竞争性的半监督文本分类方法进行了比较，包括伪标记[29]、自训练[30]和数据编程[13]，这些方法包含了NL解释来执行半监督文本分类。请注意，所有比较的模型变体都将bert-base作为主干模型。

4.4 实验结果

关于监督设置的结果。我们首先在表2和
. 3ELV在所有四个数据集中均显著优于强BERT基线，表明
利用NL解释作为自然语言理解的额外信息的有效性。ELV也始终优于ELV(仅m步)，表明ELV的变分EM训练有效地模拟了解释和预测之间的相互作用。此外，ELV的性能也优于最近一些分别关注RE和ASC的竞争性研究，进一步证明了ELV的有效性。

半监督设置的结果。在半监督设置下的结果如表4和表5所示。在半监督场景下，ELV-EST方法显著优于各种半监督文本分类方法和数据编程方法。后者使用预定义的规则将NL解释解析为逻辑形式，并匹配未标记的示例，
在所有四个数据集中。在F1分数方面，bert基+自我训练基线的RE数据集改善约为7分，ASC数据集改善超过15分。这证明了ELV-EST在半监督设置下的有效性。

关于解释的产生的结果。我们进一步评价解释生成模型的质量。我们邀请5名英语能力足够的研究生对测试集中使用输入句子生成的解释和解释增强预测模块预测的标签进行评分[3]。注释场景包括解释的
信息量（信息量。），正确性。）和一致性（缺点）关于模型的预测。内部评分者的协议为0.51Kappa评分。由于空间限制，附录中提供了人类评价的细节和所产生的解释的例子。

[3]预测模块与解释生成模块联合训练

表6：人体评价结果。分数从1分到5分（越大，越好）。用Kappa评分测量的评分者内部一致性为0.51。

产品型号	国际。	公司。	优缺点。
第2秒	2.43	3.27	2.68
变压器	2.35	3.12	2.62
UniLM	3.48	3.94	3.14
ELV（我们的）	3.87	4.20	3.51

表7：ASC数据集的结果，单词随机损坏（80%）。Orig+RandExp是原始1：1和随机损坏解释的混合。

计算方法	餐厅	笔记本电脑
BERT-基地	75.4	72.4
w.80%兰德Word	73.2	70.9
Orig+Rand考试	76.9	74.0
ELV（我们的）	77.8	75.2

为了进行比较，我们包括了一个带注释NL解释的微调UniLM模型，以及使用带注释NL解释从头训练的两个基线，一个是香草变压器模型，另一个是基于关注的LSTMseq2seq模型。研究结果如表6所示。我们的ELV框架生成的解释明显优于微调的UniLM模型生成的解释。ELV生成了与模型的决策过程相关的更好的NL解释，因为它建模了解释生成模型和预测模型的相互作用。

4.5 分析

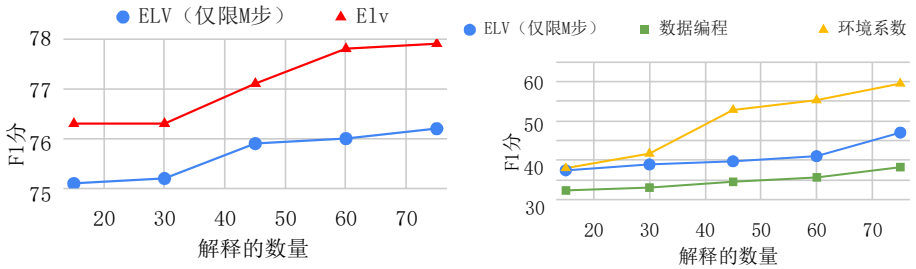

图2：具有不同解释数量的性能。我们将我们的方法与监督设置（左）和半监督设置（右）中的基线(s)进行比较。

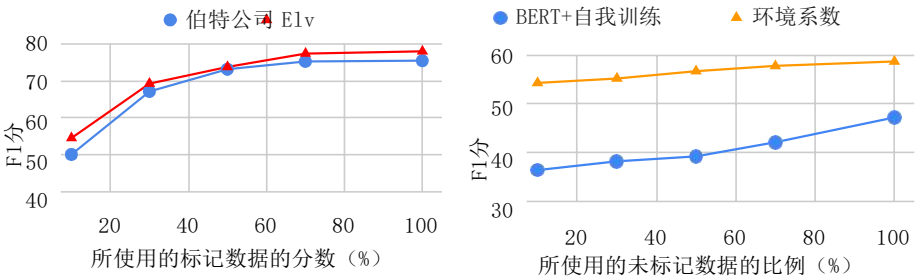

图3：分别在监督设置（左）和半监督设置（右）中使用不同数量的标记或未标记数据的性能。

具有损坏的解释的性能。我们首先研究了模型的w.r.t.的性能。检索到的解释的质量。我们与损坏的解释进行比较，随机替换80%的单词在原始解释，结果如表7所示。具有损坏解释的性能如预期的那样显著下降。高质量的解释会有帮助
该模型能更好地泛化，而随机模型可能会混淆该模型。

有不同数量的解释的表现。然后，我们用不同数量的解释来调查其表现。如图2（左）所示，在只有15种注释解释的情况下，ELV显著优于没有变分EM框架训练的对应物。随着更多的解释，ELV的性能继续提高

但ELV(仅限m步)的性能开始以45种解释达到饱和，显示了对解释生成模型和解释增强预测模型之间的相互作用进行建模的重要性。在图2（右）中的半监督学习设置中也观察到类似的结果。

具有不同数量的标记/未标记数据的性能。我们还研究了具有不同比例训练数据的不同模型的性能。从图3（左）中，我们可以看到，在不同数量的标记数据下，ELV始终优于BERT基线。
特别是，当仅使用10%的标记数据时，改进是最显著的。这是因为人类的解释提供了额外的监督，并可以作为隐式的逻辑规则来帮助泛化。在半监督学习设置中（图3，右），ELV-EST比传统的自我训练方法要大，特别是在未标记数据较少的情况下，进一步证实了解释中泛化能力的提高。

5 结论结论

在本文中，我们提出了ELV，一个新的框架，用于训练可解释的自然语言理解模型与有限的人类注释解释。我们的框架将自然语言解释视为潜在的变量，建模潜在的推理过程，使解释生成和基于解释的分类之间实现交互。在监督和半监督设置下的实验结果表明，ELV不仅可以做出有效的预测，而且可以产生有意义的解释。在未来，我们计划将我们的框架应用于其他自然语言理解任务。此外，我们计划在其他预先训练过的语言模型上测试我们的框架的有效性，这些模型要么更强（例如，XLNet[31]、RoBERTa[32]、Albert[33]、eceria[34]等）。或者计算效率。（例如，蒸草BERT[35]，[36]，DeeBERT[37]，PABEE[38]等）

更广泛的影响

另一方面，这样的系统也带来了潜在的风险，这取决于生成的自然语言解释的质量。例如，生成的自然语言可能有某些偏差，这已经在许多自然语言理解系统[39,40]中报道过。如何减轻这些风险将是我们未来的工作。另一个潜在的风险是，我们框架中的解释生成模型生成了特殊的解释，这些解释不一定提供模型如何进行预测的信息，因为模型可以想出任何它认为会与预测标签配对的解释。这是当前解释生成模型的一个常见缺点。我们的框架部分缓解了这个问题，因为生成的解释是通过解释检索过程在解释增强分类器的训练过程中使用。

确认书

该项目由自然科学与工程研究委员会(NSERC)发现基金、加拿大CIFAR人工智能主席项目、微软研究与Mila合作基金、三星、亚马逊教员研究奖、腾讯人工智能实验室犀牛雀礼品基金和NRC合作研发项目(AI4D-CORE-06)支持。作者要感谢孟曲、王子期、程路和匿名审稿人的宝贵反馈和深刻的评论。

参考文献

[1] 布拉登汉考克，帕罗马瓦尔玛，王斯蒂芬妮，马丁布林曼，梁珀西，和克里斯托弗雷。用自然的语言的解释来训练分类器。参见《计算语言学协会第56届年会论文集》（第1卷：长篇论文），1884-1895页，澳大利亚墨尔本，2018年7月。计算语言学协会。使用剂量：10.18653/v1/P18-1175.URLhttps://www。网址。P18-1175。

[2] 拉贾尼，麦肯，熊和索舍尔。解释一下你自己吧！利用语言模型进行常识性推理。参见《计算语言学协会第57届年会论文集》，第4932-4942页，意大利佛罗伦萨，2019年7月。计算语言学协会。使用剂量：10.18653/v1/P19-1487.URLhttps://www。网址。P19-1487。

[3] 玛利亚坎布鲁，蒂姆罗克塔斯切尔，托马斯卢卡西维奇，和菲尔布伦塞姆。具有自然语言解释的自然语言推理。*《神经信息处理系统的研究进展》，第9539-9549页，2018页。*

[4] 拉贾尼，麦肯，熊和索舍尔。解释一下你自己吧！利用语言模型进行常识性推理。*arXiv预印本，arXiv：1906.02361,2019年。*

[5] 斯里瓦斯塔瓦，伊戈尔·拉布托夫和汤姆·米切尔。从自然语言解释中获得的联合概念学习和语义解析。参见《2017年自然语言处理经验方法会议论文集》，第1527-1536页，丹麦哥本哈根，2017年9月。计算语言学协会。使用剂量：10.18653/v1/D17-1161.URLhttps：//www。网址。D17-1161。

[6] 周文轩、林洪涛、林宇辰、王子奇、都君义、内夫斯、任翔。Nero：一个标签高效关系提取的神经规则基础框架。*在2020年网络会议论文集，WWW‘20,2166-2176页，纽约，纽约，美国，2020年。*计算机机械协会。ISBN9781450370233。使用剂量：10.1145/3366423.3380282.URLhttps://doi。org/10。1145/3366423.3380282.

[7] 王子期、秦玉佳、周文轩、燕俊燕、沁源、内夫斯、刘志远、香任。用神经执行树从解释中学习。在2020年的国际学习代表会议上。URLhttps://open审查。网络/论坛吗？id=rJlUtOEYwS。

[8] 杰森帕尔默，肯尼斯克罗伊兹-德尔加多，巴斯卡尔德饶，和大卫PWip夫。非高斯潜在变量模型的变分变量算法。关于神经信息处理系统的研究进展，第1059-1066页，2006年。

[9] 斯里瓦斯塔瓦，伊戈尔·拉布托夫和汤姆·米切尔。从自然语言解释中获得的联合概念学习和语义解析。*在2017年自然语言处理的经验方法会议论文集中，第1527-1536页，2017页。*

[10] 默蒂、彭伟高和梁珀西。用自然语言解释的表示工程。*arXiv预印本arXiv：2005.01932,2020年。*

[11] 丹·戈德瓦瑟和丹·罗斯。从自然的指令中学习。机器学习，94（2）：205-232,2014。

[12] Sanja费德勒等人。用自然语言反馈来描述图像的教学机器。在以下项目中*神经信息处理系统的研究进展，第5068-5078页，2017年。*

[13] 布拉登汉考克，马丁布林曼，帕罗马瓦尔玛，梁珀西，王斯蒂芬妮，和克里斯托弗雷。用自然的语言的解释来训练分类器。在会议会议记录中。*计算语言学协会。*会议，2018卷，第1884页。NIH研究院公共访问，2018年。

[14] 迪德里克，P，金马和马克斯，韦林。自动编码变分贝叶斯。*arXiv预印本，arXiv：1312.6114,2013年。*

[15] 布拉乌，卡马乔-科拉多斯和肖凯特。从伯特那里引出关系知识。*arXiv预印本，arXiv：1911.12753,2019年。*

[16] 法比奥·佩特罗尼、蒂姆·罗克塔斯切尔、帕特里克·刘易斯、安东·巴赫廷、吴玉祥、亚历山大·米勒和塞巴斯蒂安·里德尔。语言模型作为知识库吗？*arXiv预印本，arXiv：1909.01066,2019年。*

[17] 亚当罗伯茨，科林拉菲尔，和诺姆谢泽。您可以在一个语言模型的参数中打包多少知识？*arXiv预印本arXiv：2002.08910,2020年。*

[18] 李东、南山、王文辉、富鲁魏、刘晓东、王宇峰、高建峰、周明、萧五雄。统一的语言模型预训练的自然语言理解和生成。《神经信息处理系统的研究进展》，第13042-13054页，2019页。

[19] 尼尔斯·莱默斯和伊琳娜·古雷维奇。使用暹罗电子网络的句子嵌入。2019年自然语言处理经验方法会议和第9届国际自然语言处理联席会议(EMNLP-IJCNLP)论文集，第3982-3992页，中国香港，2019年11月。计算语言学协会。使用剂量：10.18653/v1/D19-1410.URLhttps://www。网址。D19-1410。

[20] 德文林，张明伟，李肯顿，和图塔诺娃。深度双向变压器的语言理解预培训。*在计算语言学协会北美分会2019年会议论文集上，第1卷（长和短论文），第4171-4186页，明尼阿波利斯，2019年6月。*计算语言学协会。剂量值：10。18653/v1/N19-1423.URLhttps：／／www。网址。N19-1423。

[21] 艾里斯·亨德里克克斯、苏南金、科扎雷瓦、纳科夫、西格达夫、塞巴斯蒂安·帕蒂、罗马诺和斯帕科维奇。任务8：名称对之间的语义关系的多路分类。第5届语义评估国际讲习班论文集，第33-38页。计算语言学协会，2010年。

[22] 张宇浩、钟、陈丹奇、安吉利、D。曼宁公司。位置感知注意和监督数据改善了插槽填充。*在2017年自然语言处理经验方法会议论文集(EMNLP2017)中，第35-45页，2017页。*URLhttps://nlp。斯坦福大学。张2017年。pdf.。

[23] 张宇浩、钟、陈丹奇、安格利和曼宁。位置感知注意和监督数据改善了插槽填充。*2017年自然语言处理经验方法会议论文集，2017页第35-45页。*

[24] 玛丽亚·庞蒂基，迪米特里斯·加拉尼斯，约翰·帕夫洛普洛斯，哈里斯·帕帕乔吉奥，离子安德拉索普洛斯和苏雷什·曼南达尔。2014年半学期任务4：基于方面的情绪分析。第8届语义评估国际研讨会论文集(Semeval2014)，第27-35页，爱尔兰都柏林，2014年8月。计算语言学协会。剂量值：10。3115/v1/S14-2004.URLhttps：／／www。网址。2004年14日。

[25] 利维奥·巴尔迪尼·苏亚雷斯，尼古拉斯·菲茨杰拉德，林杰弗里和汤姆·克维亚特科夫斯基。空白匹配：关系学习的分布相似性。参见《计算语言学协会第57届年会论文集》，第2895-2905页，意大利佛罗伦萨，2019年7月。计算语言学协会。使用剂量：10.18653/v1/P19-1279.URLhttps://www。网址。P19-1279。

[26] 张陈诚、李秋池、宋大伟。基于特定点图卷积网络的基于方面的情绪分类。*arXiv预印本，arXiv：1909.03477,2019年。*

[27] 胡徐、刘兵、雷书、余羽。伯特培训后的回顾阅读理解和基于方面的情绪分析。*arXiv预印本，arXiv：1904.02232,2019年。*

[28] 宋有为、王嘉海、姜涛、刘志跃、饶杨辉。有针对性的情绪分类的注意编码器网络。*arXiv预印本，arXiv：1902.09314,2019年。*

[29] 李东云。伪标签：简单有效的深度神经网络半监督学习方法。*在表示学习挑战研讨会，ICML，第三卷，第2页，2013。*

[30] 查克罗森博格，军事赫伯特，和亨利施耐德曼。目标检测模型的半监督自我训练。*运动，2005年2月2日。*

[31] 杨志林、戴子亨、杨一鸣、卡博纳尔、萨拉胡迪诺夫和勒五等人。用于语言理解的广义自回归预训练。《神经信息处理系统的研究进展》，第5753-5763页，2019页。

[32] 刘银汉、迈尔、戈亚尔、杜景飞、陈丹奇、李维、刘易斯、泽特莱莫耶和斯托亚诺夫。罗伯塔：一个稳健地优化的伯特预训练方法。*arXiv预印本，arXiv：1907.11692,2019年。*

[33] 兰振中、陈明达、古德曼、吉普尔、沙马、索德。用于语言表示的自我监督学习。*arXiv预印本，arXiv：1909.11942,2019年。*

[34] 凯文·克拉克，明通龙，克里斯托弗·曼宁。电子广播：训练前的文本编码器作为鉴别器，而不是生成器。*arXiv预印本arXiv：2003.10555,2020年。*

[35] 维克多桑，莱桑德尔首次亮相，朱利安乔蒙德，和托马斯沃尔夫。迪迪伯特，伯特的提炼版：更小，更快，更便宜，更轻。*arXiv预印本，arXiv：1910.01108,2019年。*

[36] 徐康文、周王春书、陶葛、福魏、周明。通过渐进式模块替换来压缩伯特。*arXiv预印本arXiv：2002.02925,2020年。*

[37] 季欣、唐、李义军、良梁、林吉敏。动态提前退出，以加速伯特推理。*arXiv预印本arXiv：2004.12993,2020年。*

[38] 周王春书、徐康文、陶葛、麦考利、柯徐、福魏。伯特失去了耐心：提前退出的快速和稳健的推理。*arXiv预印本arXiv：2006.04152,2020年。*

[39] 赵解玉、王天禄、雅士卡、奥多内斯、张凯伟。男性也喜欢购物：使用语料库水平的约束来减少性别偏见的放大。*arXiv预印本，arXiv：1707.09457,2017年。*

[40] 玛尔塔和正义萨。对自然语言处理中的性别偏见研究的分析。*《自然机器智能》，2019年第1-2页。*

A 解释的例子

在本节中，我们将介绍几种人类注释的自然语言解释和由ELV框架中的解释生成模型生成的解释和解释，以更好地理解所提出的方法。

A.1 卷的

A.1.1 人类有注释的解释

a) 虽然她不是《花花公子》的花花伴，但她和她的《隔壁的女孩》以及赫夫纳的女友霍莉·麦迪逊和肯德拉·巴克特(OBJ)，当时被称为肯德拉·威尔金森（裸体）。

标签：每：alternate_names
解释：术语“然后被称为”发生在SUBJ和OBJ之间，并且在SUBJ和OBJ之间不超过6个单词

b) 伯克(SUBJ)的母亲梅丽莎·贝尔(OBJ)是舞蹈组合Soul1ISoul的一名歌手，该乐队在20世纪80年代和90年代风靡一时。

标签：按：父母单位
解释：SUBJ和OBJ包含了“他的母亲”这个词，并且在SUBJ和OBJ之间不超过三个词

c) 艾伦·蓬佩奥(OBJ)秘密娶了克里斯·伊弗里(SUBJ)，祝贺新婚夫妇，让他们从此过上幸福的生活！！

标签：每个人：配偶
解释：SUBJ和OBJ之间不超过四个词，SUBJ和OBJ夹层的这个词“秘密结婚”

A.1.2 机器生成的说明

a) *洛马克斯分享了一个关于阿尔梅娜·洛马克斯的母亲，洛杉矶的报纸老板和记者(OBJ)带着她(SUBJ)家人去阿拉巴马州塔斯基吉的故事。1961年。*标签：按：标题
说明：“家庭”这个词正好仅次于SUBJ。

b) 股东维权组织公司图书馆(SUBJ)的高级研究助理保罗·霍奇森（约翰逊）说，他们的投资所发生的情况对他们不感兴趣，因为他们已经得到了报酬。

标签：根据：org：top_members/员工
解释：“高级研究员”一词就出现在SUBJ之前。

c) 易洛魁人护照纠纷引发了主权问题。总部设在华盛顿特区的美国印第安人国民大会(SUBJ)代表长曲棍球队主张，敦促英国官员允许这些成员持有易洛魁人发行的护照进入英国。

标签：org：country_of_headquarterse
解释：“basedin”一词就出现在OBJ之前。

A.2 半圆形的

A.2.1 人类有注释的解释

a) Morton的SUBJ-O是第三个间隙局部OBJ-O的最常见原因，这些诊断测试产生了良好的病情适应症。

标签：因果关系（e1、e2）
解释：在SUBJ和OBJ之间，术语“是最常见的原因”出现，而SUBJ先于OBJ

b) 额叶SUBJ-O是OBJ-O的一部分，它与边缘系统保持着非常紧密的联系。

标签：组件整体部件（e1、e2）
解释：在SUBJ和OBJ之间，术语“是“出现”的一部分，而SUBJ先于OBJ。

c) *现任秘书正在召集前任委员会主席和委员会主席。*

标签：实体来源（e1、e2）
解释：“来自过去”这个短语链接了SUBJ和OBJ，OBJ，SUBJ和OBJ之间不超过三个单词，OBJ跟随SUBJ。

A.2.2 机器生成的说明

a) *周一早上，潜艇在夏洛特道格拉斯国际机场造成了OBJ-0。*

标签：因果关系（e1、e2）
说明：subj和obj之间只有一个单词“引起”，obj跟随subj。

b) *它所在的基地隐藏了在谭战场上从SUBJ-0最初的OBJ-0中移除时发生的破坏。*

标签：实体来源（e1、e2）
解释：在subj和obj之间，短语“被移除到”发生，在SUBJ和OBJ之间不超过4个单词，在OBJ之前不超过4个单词

c) *铰链SUBJ-0将OBJ-0枢轴地连接到电子设备的基座上，并具有旋转叶片和固定叶片。*

标签：组件整体部件（e1、e2）
解释：SUBJ和OBJ之间的术语“附加a”，OBJ遵循SUBJ

A.3 餐厅

A.3.1 人类有注释的解释

a) *我们吃了很棒的甜点（包括我吃过的最好的仁仁），然后他们在家里提供了晚餐后的饮料。（术语：仁仁）*

标签：阳性
说明：“最好”一词直接放在术语前面。

b) *他们尝试过的所有甜点都得到了好评。（使用术语：甜点）*

标签：阳性
说明：“有利”出现后不超过5个单词。

c) *然而，最恼人的是，服务器似乎接受了驱动收入的培训。（技术术语：服务器）*

标签：负数
解释：“烦人”这个词出现在这个词之前。

A.3.2 机器生成的说明

a) *这个小地方非常热情的欢迎。（合同术语：施工地点）*

标签：阳性
说明：故事后面是“精彩”。

b) *沙拉煮熟了，但鸡肉很好。（技术术语：鸡）*

标签：阳性
说明：“fine”一词出现在术语后的3个单词内。

c) *服务很糟糕，主要是因为员工在周六晚上不知所措。（任期：员工）*

标签：负数
解释：“难以忍受”一词出现在术语后的三个单词内。

表8：在使用不同比例标记数据的监督设置下，ASC数据集的结果(Macro-f1)。

所使用的标记数据的分数	60%	70%	80%	90%	100%
ELV（我们的）	75.5	-	-	-	-
BERT-基地	74.6	75.0	75.3	75.4	75.4

A.4 笔记本电脑

A.4.1 人类有注释的解释

a) *当DVD驱动器在我的背包里时，它也会随机打开，这很烦人。（期限：DVD驱动器）*
标签：负数
说明：字符串“烦人”出现在这个术语之后

b) *苹果团队也能很好地帮助你选择适合你的电脑。（术语：苹果团队）*
标签：阳性
说明：字符串“非常好”出现在术语之后不超过6个单词。

c) *设计非常棒，质量是前所未有的。（技术术语：设计）*
标签：阳性
解释：“很棒”这个词在术语后只有两个单词。

A.4.2 机器生成的说明

a) *在我对新的27款Imacs的规格感到失望后，我订购了我的2012年的macmini。（技术术语：技术规格）*
标签：负数
说明：“失望”一词出现在术语前的3个单词内。

b) *我发现这个迷你车非常容易设置。（期限：设置）*
标签：阳性
解释：“特别简单”一词出现在术语前的3个单词内。

c) *然而，触摸板有一些主要的问题，使该设备几乎无用处。（术语：止血垫）*
标签：负数
解释：“几乎无用”这个词出现在术语后的3个单词内。

B 人工评估的细节

对于人类评估，我们在餐厅数据集的测试集中随机抽取100个示例，并使用ELV来预测所选示例的标签。然后，我们使用不同的比较模型，用输入句子和预测标签生成预测结果的解释。之后，我们邀请了5名英语熟练的研究生进行打分。注释场景包括解释的信息量。)，正确性。）和一致性（缺点）关于模型的预测。具体来说，信息量衡量了生成的解释在多大程度上有助于理解模型的预测输出。正确性衡量的是解释是否实际正确。“好”导致积极的标签，而“烦人”是消极的)。一致性是指该解释是否与输入的句子是否一致。解释中的描述是正确的，即输入的句子)。

C 加法分析实验

在本节中，我们报告了额外的实验结果，比较了ELV与60%的标记数据的性能，与60%、70%、80%、80%、90%和100%的性能，以调查人类注释解释在多大程度上可以取代人类标记的例子。结果如表8所示，我们发现ELV可以实现甚至超过使用更多标记数据训练的bert基础模型的性能。这证实了ELV可以有效地利用人类注释的解释作为附加信息。.