Bo通过融入比较先验促进放射学报告生成

rSanghwan Kim ♣♡ Farhad Nooralahzadeh♡ Morteza Rohanian ♡
Koji Fujimoto♢ MizuhoNishio♢ Ryo Sakamoto♢FabioRinaldi♠
MichaelKrauthammer♡♣ 苏黎世联邦理工学院♡苏黎世大学和苏黎世大学医院♢京都大学医学研究生院♠达勒莫尔人工智能研究所
sanghwan.kim@inf.ethz.ch

arXiv:2305.04561v2 [cs.CL] 5 Jun 2023

摘要

近年来，基于Transformer的模型在根据胸部X光图像生成放射学报告方面取得了显著进展。然而，仍然存在一个突出的问题：这些模型通常缺乏先验知识，导致生成的合成报告错误地提及并不存在的既往检查。这种差异可归因于放射科医生与生成模型之间的知识鸿沟。放射科医生掌握患者特定的先验信息，而模型仅接收特定时间点的X光图像。为了解决这一问题，我们提出了一种新颖的方法，利用基于规则的标注器从放射学报告中提取比较先验信息。随后，将提取的比较先验无缝整合到最先进的基于Transformer的模型中，使其能够生成更加真实、全面的报告。我们在英文报告数据集（如 IU X-ray 和 MIMIC-CXR）上对该方法进行了评估。结果表明，我们的方法在自然语言生成指标方面优于基线模型。值得注意的是，我们的模型生成的报告不会错误引用不存在的既往检查，这使其区别于以往的模型。通过解决这一局限性，我们的方法代表着医学报告生成领域在弥合放射科医生与生成模型之间差距方面迈出的重要一步。

1 引言

放射学图像的分析以及随后撰写医学报告，是诊断流程中的关键任务（Suetens，2017；Krupinski，2010）。然而，对于放射科医生而言，撰写放射学报告是一项劳动密集且耗时的工作，需要经过多年的培训，才能准确识别和描述医学图像中的特定异常（Brady，2017；Arenson and Dunnick，2006）。受深度学习领域中图像描述模型取得成功的启发，近年来，已有大量研究提出了各种自动放射学报告生成模型，尤其侧重于胸部X光图像（Yuan 等人，2019；Li 等人，2019；Xue 等人，2018；Jing 等人，2017；Liu 等人，2019a）。自动生成报告有望减轻放射科医生繁重的工作负担，并通过提供包含有用关键词或观察结果的初步报告，加快诊断流程（Johnson 等人，2019；Chen 等人，2020）。

尽管近期一些方法在根据胸部X光图像生成放射学报告方面取得了相对成功（Endo 等，2021；Johnson 等，2019；Chen等，2020 年；Miura 等，2020；Ramirez- Alonso 等，2022；Nooralahzadeh 等，2021），但这些研究仍未解决一个关键挑战：如何为模型提供恰当的先验知识，使其获得与放射科医生相似的信息。具体而言，放射科医生能够获知既往报告和X光图像是否存在，从而将当前检查与过去的检查进行比较，并评估患者病情进展、恶化或改善的情况（Suetens，2017；of Radiology，ESR）。这些医学报告通常会使用特定词语或短语进行比较，例如“与既往检查相比”“在此期间”“参照既往X光片”等。本文将这些词语或短语称为先验表达；它们也存在于MIMIC-CXR（Johnson 等，2019）和IU X-ray（Demner-Fushman 等，2016）等通用医学数据集中，这些数据集被广泛用于训练和评估报告生成任务。医学报告中包含先验表达对于准确报告至关重要。然而，在医学报告数据集上训练的模型往往会生成包含不恰当或误用先验表达的报告，导致评估指标相对较低。挑战在于如何在模型的生成过程中有效地融入并利用先验表达；这是一个至关重要、但尚未得到充分

真实标注	R2Gen（Chen 等人，2020）	M^2Tr（Cornia 等人，2020）
心脏大小正常。 主动脉迂- 曲扩张。心脏纵隔 轮廓正常。肺部清晰 未见纤维化迹象。胸膜 积液或气胸。终板 胸椎可见硬化性改变， 胸椎。	间质性改变弥漫性增多 提示双侧肺部存在肺纤维化 。纤维化似乎 略有加重，**与** **既往相比**在 xxxx 中。气管 位于中线。未见胸腔积液。 心脏大小正常。	双肺清晰且扩张良好。 心脏和纵隔正常。
心纵隔轮廓稳定。 未见局灶性肺部胸腔积液 或气胸。未见急性骨质异常。 常态性。	心脏大小正常。 The me- → 这项研究的 me- 纵隔未见异常。存在 未见胸膜异常或局灶性肺泡疾病。 存在轻度慢性退行性改变。 存在于脊柱中。	肺容积偏低。 右侧膈肌抬高。斑片状阴影 右肺底**再次可见**。左肺清晰。 心脏大小处于正常上限。主动脉钙化- 灶。肉芽肿。无 气胸证据。双侧肋膈角变钝- 肋膈角。退行性 胸椎的变化。

表1：IU X-ray 的真实报告（第一列）以及 R2Gen 和 M^2Tr 生成的报告示例（第二列和第三列）。先验表达以粗体突出显示。

已解决。

在表 1中，我们比较了真实报告与两种近期模型生成的报告：R2Gen（Chen 等人，2020）和M^2Tr（Cornia 等人，2020），重点关注其中是否出现先验表达。显然，合成报告包含了不恰当的先验表达。例如，第一行中由R2Gen生成的合成报告提到了与既往检查进行比较，尽管真实报告中没有任何既往信息。类似地，第二行中由 M^2Tr生成的合成报告包含“再次发现”这一短语，暗示存在一张既往图像，而真实报告中并未出现任何先验表达。

这些错误引用的报告包含先验表达，往往会导致评估指标较低。在图1中，我们使用基于规则的标注器（详见第3.1节）将合成报告分为两类：阴性和阳性。阴性类表示不含先验表达的报告，阳性类则包含带有先验表达的报告。随后，我们绘制了各类别的 BLEU-4（Papineni 等人，2002年）分数分布，以及各自的均值和标准差。结果清楚地表明，与阴性报告相比，阳性报告的得分更低。这一观察结果表明，基于Transformer的模型未能有效利用生成的先验表达，凸显了正确

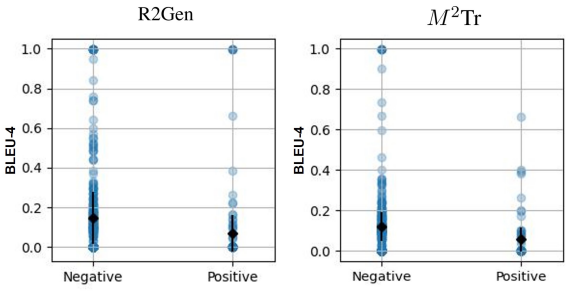

图1：根据我们的基于规则的标注器将报告分为阴性和阳性类别后，BLEU-4 分数的分布情况。该标注器评估了由 R2Gen 和 M^2Tr 生成的合成报告，两者均在 IUX-ray 数据集上进行训练。均值和标准差分别以黑色圆点和线条表示。阴性和阳性标签的平均分数如下：R2Gen 分别为（0.1455，0.0698），M^2Tr 分别为（0.1218，0.0583）。

在训练过程中融入先验信息。

表1 和图 1 突显了最先进（SOTA）模型在学习和生成先验表达方面的一个关键问题，这可能会使放射科医生在使用这些模型撰写报告时感到困惑。这个问题源于放射科医生撰写报告的方式与模型生成报告的方式之间存在根本差异。放射科医生不仅可以获取既往检查和病史等患者既往信息，还可以查看当前的X光图像。相比之下，报告生成模型仅会获得

在特定时刻获取的X光图像。由于先验信息有限，模型难以生成全面且具有深刻洞察力的报告，达到专家所撰写报告的水平。

在我们的论文中，我们通过将先验信息注入现有模型，弥合生成模型与放射科医生之间的知识差距。我们的目标是缩小这种差距，使改进后的模型能够生成信息量更丰富、更实用的报告。由于 IU X-ray 和 MIMIC-CXR 等现有数据集不包含先验信息（既往 X 光图像），我们采用数据驱动的方法，咨询经验丰富的放射科医生。受 CheXpert标注器（Irvin 等人，2019）的启发，我们开发了一种基于规则的标注器，通过识别放射学报告中的特定模式和关键词来提取先验信息。值得注意的是，该基于规则的标注器重点关注比较短语，以判定每位患者的医学报告对应的是首次或后续检查。

我们论文的主要贡献如下：

1. 我们与放射科医生合作，开发了一种基于规则的标注器，用于识别医学报告中与比较相关的特定关键词和模式。

2. 为了将先验信息融入模型中，我们提出了一种增强型 Transformer 架构。该方法易于实现，并且可以作为插件方法无缝集成到现代报告生成模型中。

3. 通过在 IU X-ray 和 MIMIC-CXR 数据集上进行实证评估，我们证明了与基线模型相比，我们的模型在性能指标方面表现更优。

4. 此外，我们进行了全面分析，以确认我们的模型不再生成错误引用，从而解决了先前方法中观察到的一项重大局限。

2 相关工作

放射学报告生成领域的早期研究（Bai 和 An，2018；Liu 等人，2019b）采用了基础的编码器-解码器架构，其中编码器从医学图像中提取关键特征并将其转换为潜向量，解码器则根据潜向量生成目标文本。通常，卷积神经网络（LeCun等，2015年）用作编码器，而长短期记忆网络（Hochreiter 和 Schmidhuber，1997）则被选作解码器。随后，研究人员引入视觉注意力机制，以突出图像中的特定特征，并生成更具可解释性的报告（Zhang等，2017年；Jing等，2017年；Wang等，2018年；Yin等，2019年；Yuan等，2019年）。近期研究（Lovelace和Mortazavi，2020年；Chen等，2020年；Nooralahzadeh 等，2021；Miura 等，2020）探索了采用 Transformer 的更先进架构，以生成更全面且一致的医学报告。

或者，也可以将医学报告生成视为一项检索任务，因为大多数报告中往往会重复出现相似的句子和特定的写作格式。与从头开始生成整份报告相比，重复利用视觉上相似的X光图像中的诊断文本，可能会生成更加一致和准确的报告。Li 等人(2019) 证明了基于检索的模型具有优越性，其性能超过了许多基于编码器-解码器的模型。随后，Endo 等人(2021) 提出了一种名为 CXR-RePaiR 的基于检索的模型，该模型引入了对比语言-图像预训练（CLIP）（Radford 等人，2021），用于计算文本和图像嵌入之间的相似度。CXR-RePaiR 根据给定的特定X光图像，从大型报告语料库中选择匹配度最高的医学报告来生成预测结果。该模型在其新开发的指标上取得了最先进性能。

以往的方法主要通过改进模型架构来提升模型性能，而相对较少关注放射学报告的独特特征，尤其是涉及比较的内容。不过，也有少数值得注意的例外，例如Ramesh 等人（2022）的研究，该研究专门探讨了报告中先验表达的影响。该研究的作者推出了一个名为 MIMIC-PRO 的新数据集，其中他们识别并修改了包含对不存在的既往检查进行幻觉性引用的报告。这些幻觉性引用可视为与本文讨论的先验表达类似的概念。Ramesh 等人（2022）提出了一种基于BioBERT的模型，用于释义改写或删除涉及既往报告或图像的句子，他们认为这些表达会使模型混淆

报告	标签
1. 可见心脏扩大，且与之前的检查相比**com-无变化**，检查时间为 XXXX。	1
2. **再次发现**左肺舌段区域有边界不清的阴影。	1
3. 肺容积偏低。肺其余部分清晰。	0
4. 左下叶已清晰，**在间隔期内**。	1

表 2：标注器对从 IU X射线数据集中抽样的报告生成的输出。加粗短语表示标注器识别出的先验表达。

数据集	阴性	阳性	总计
IU X-ray	3,426	529	3,955
MIMIC-CXR	106,628	99,935	206,563

表 3：我们的基于规则的标注器分类为阴性（0）或阳性（1）的研究数量。

并导致句子错误地引用了相关内容。他们与专家合作创建了一个“干净”的 MIMIC-CXR 测试数据集，并比较了在 MIMIC-CXR 和 MIMIC-PRO 上训练的模型。

与 Ramesh 等人 (2022不同，我们采用了一种不同的方法来解决比较先验信息的问题：我们将先验信息纳入模型，使其能够以端到端的方式生成更全面的报告，而不是完全移除比较先验信息。在实际医疗领域，使用先验表达在放射学报告中撰写比较内容是不可避免的，而且从真实报告中构建干净、准确的数据集也是一项繁重的工作。因此，我们的工作重点是将比较先验直接应用于现有模型，例如 R2Gen 和 M^2Tr.

3 方法

如表1所示，即使是迄今为止表现最佳的模型，也难以处理意外的先验表达。为解决这一问题，我们提出一种两步方法：（1）构建一个基于规则的标注器，以区分包含和不包含先验表达的报告（第3.1节）；（2）扩展基于Transformer的架构（R2Gen 和M^2Tr），使其将比较先验作为输入（第3.2节）。在第一步中，我们引入一个基于规则的标注器，用于检测特定的比较先验表达，并根据检测结果将每份报告分类为首次检查（阴性）或后续检查（阳性）。我们从中获得启发——借鉴CheXpert标注器的否定与分类原则（Irvin等人，2019），以设计我们新颖的标注器。在后续阶段，我们将先验标签作为输入整合到最先进的架构中，以生成更实用、更全面的报告，并将结果与基线模型进行比较。通过为我们的新模型提供比较先验（在实际诊断场景中通常会传达给放射科医生），我们能够生成更加全面且一致的医学报告。

3.1 基于规则的标注器

我们的基于规则的标注器遵循 CheXpert 标注器的基本结构（Irvin 等人，2019 年），该标注器依据专家制定的固定规则，检测放射学报告中 14 种观察结果是否存在。因此，我们的标注器由三个不同阶段组成：提及抽取、提及分类和提及聚合。标注器以放射学报告的 Findings 部分为输入，并生成二元输出（0 或 1）。负标签（0）表示报告中没有先验表达，而正标签（1）表示报告中存在先验表达。

提及抽取 提及是指可能出现在先验表达中的特定关键词，例如“previous”（之前）、“prior”（先前）、“preceding”（此前）、“previously”（此前）、“again”（再次）、“comparison”（比较）、“interval”（间隔）、“increase”（增加）、“decrease”（减少）、“enlarge”（扩大）等。在此阶段，标注器从每份报告中抽取提及，并在每个句子中标记这些提及。不过需要注意的是，即使某些句子包含指定关键词，在这一步也无法确认是否存在先验表达，因为这些关键词可能用于其他语境。例如，“comparison”（比较）一词可以用于“with no comparison studies”（没有可供比较的检查）这样的句子中，表示不存在先验表达。

提及分类 在第一阶段提取提及内容后，我们的标注器会判断每个提及是否对应预定义的先验表达。由于报告中经常使用类似表达来指代与既往检查的比较，我们可以将这些先验表达的模式归纳为若干经验丰富的放射科医生熟悉的关键词短语。例如，“与 {men- tion} 相比 / 类似 ” 这一短语确认了既往报告的存在，其中

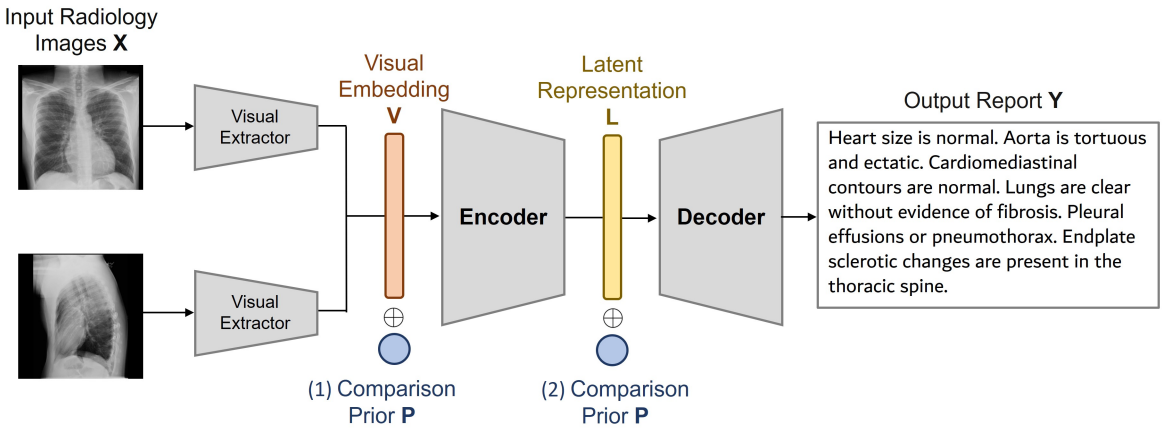

图 2：我们方法的概念示意图。报告生成模型（R2Gen 和 M^2 Tr）由视觉提取器、编码器和解码器组成。我们的核心思路是将基于规则的标注器生成的比较先验信息注入（1）视觉嵌入 V 和（2）潜在表示 L。

“{mention}”代表“以往”“之前”和“既往”等关键词。类似地，当“{mention}”涉及“再次”和“先前”等关键词时，“见到 / 发现 / 可视 / ..... / 注意到 {men- tion}” 便构成先验表达。

提及聚合 在最后阶段，标注器会汇总已分类的提及，并生成负标签（0）或正标签（1）：负标签表示报告中没有先验表达，正标签则表示报告中存在先验表达。标注器输出示例见表 2，IU X-ray 和 MIMIC-CXR 数据集中阴性和阳性检查的数量见表3。

3.2 扩展模型

本节介绍如何将比较先验整合到 R2Gen 和 M^2Tr 等现有模型中，以生成信息更丰富、更全面的报告。

生成过程R2Gen 和 M^2Tr 的生成过程如图 2 所示。其流程如下：输入放射影像 $X \rightarrow$ 视觉嵌入 $V \rightarrow$ 潜在表示 $L \rightarrow$ 输出报告 Y。首先，将胸部X光图像 X 作为输入提供给视觉提取器，其中 X 由正位图像 X_f 和侧位图像 X_l组成，表示为 $X = \{X_f, X_l\}$。视觉提取器生成视觉嵌入 $V = \{v_1, v_2, ..., v_S\}$，其中包含图像块特征 $v_s \in \mathbb{R}^d$，d 表示特征向量的大小。随后，V经过多个 Transformer 层在编码器中获得潜在表示$L = \{l_1, l_2, ..., l_T\}$，其中潜在特征向量表示为 $l_t \in \mathbb{R}^f$。最后，解码器利用L 生成最终输出报告 Y。

融入比较先验 比较先验 $P \in \mathbb{R}$ 由我们的基于规则的标注器生成，并表示阴性（0）或阳性（1）标签。我们旨在将比较先验融入现有的数据流程，同时确保添加先验不会改变模型架构，也不会增加任何需要训练的权重。否则，将难以衡量比较先验对生成模型的影响。因此，我们在图 P2所示的生成模型中，将先验 V 同时加入视觉嵌入 L 和潜在表示。编码器需要获取先验信息，才能生成恰当的中间表示。此外，我们还在 P 上添加了 L，因为经过编码器中的多个 Transformer 层后，P所包含的信息可能会被弱化。解码器将基于以 P为条件的潜在表示生成报告输出。整个过程模拟了放射科医生结合既往检查进行诊断的过程。因此，新的视觉嵌入 V_{new} 和新的潜在表示 L_{new} 可按如下方式计算：

$$V_{new} = V \oplus P, \quad L_{new} = L \oplus P \tag{1}$$

其中 $\oplus$表示逐元素求和。我们的方法的优势在于，它适用于大多数现有的基于Transformer的模型，且不需要额外的数据集或信息。

数据集	模型	自然语言生成指标					
		BL-1	BL-2	BL-3	BL-4	CIDEr	RG-L
IU X-ray	R2Gen (Chen 等人，2020)	0.421	0.262	0.183	0.137	0.480	0.337
	含先验（我们的方法）	**0.438**	**0.280**	**0.201**	**0.155**	**0.631**	**0.351**
	M^2Tr (Cornia 等人, 2020)	0.400	0.240	0.159	0.112	0.300	0.324
	含先验（我们的方法）	0.406	0.249	0.167	0.120	0.323	0.330
MIMIC-CXR	R2Gen (Chen 等人, 2020)	0.335	0.206	0.138	0.100	0.148	0.278
	含先验（我们的方法）	0.342	0.222	**0.152**	**0.110**	**0.166**	**0.301**
	M^2Tr (Cornia 等人, 2020)	0.353	0.211	0.137	0.094	0.089	0.262
	使用先验信息（我们的方法）	**0.357**	**0.224**	0.151	0.108	0.101	0.293

表4：基线模型和融入先验信息的模型的训练结果。我们的方法的结果以灰色行显示，最佳指标以粗体标出。所有指标均为3次运行的平均值。包含标准差的完整表格见表6

4 实验

架构 为了提取视觉特征，我们使用预训练的卷积神经网络（CNN），例如 DenseNet121（Huang 等人，2017）和 ResNet121（He 等人，2016）。通过实证评估，我们发现 DenseNet 在我们的生成任务中表现更好，因此选择它作为基础视觉特征提取器。我们采用网格记忆变换器（M^2Tr）（Cornia 等人，2020）和关系记忆驱动变换器（R2Gen）（Chen 等人，2020）的结构来构建我们的编码器和解码器。

数据集 我们在两个广泛使用的医学报告生成英语数据集上评估所提出的方法：IU X-ray（Demner-Fushman 等人，2016）和 MIMIC-CXR（Johnsonetal.,2019）。IU X射线数据集是一个公开可用的放射学数据集，包含7,470张胸部X光图像和3,955份放射学报告。每份报告都配有一张正位图像，并可选配一张侧位图像。MIMIC-CXR是一个大型胸部X线片数据库，包含473,057张胸部X光图像和206,563份报告。我们使用完整的数据配对训练模型，每组数据包括两张图像（正位和侧位）以及一份报告（所见部分）。我们按照Chen 等人（2020）中描述的数据划分方式，将数据集划分为训练集、验证集和测试集。

训练细节 我们首先使用基于规则的标注器为每份报告生成比较先验。然后，以两张图像和比较先验作为输入、医学报告作为输出，训练模型。我们使用 Adam优化器，视觉特征提取器的初始学习率设为 0.00005，编码器-解码器的初始学习率设为 0.0001 模型。学习率在预先设定的步骤逐渐下降。所有实验均在一块“NVIDIA GeForce RTX 1080 Ti”GPU 上进行，使用 3 个不同的随机种子和 16 的批次大小。我们的代码实现基于以下公开代码： Chen 等人 (2020) 和 Nooralahzadeh 等人 (2021)。

评估指标我们报告了通用自然语言生成（自然语言生成）指标，包括 BLEU（Papineni 等人，2002年）、CIDEr（Vedantam 等人，2015）和 ROUGE-L（Lin，2004）。这些指标常用于评估生成文本的质量。BLEU 衡量生成文本与参考文本之间的 n元语法重叠，而 CIDEr 基于词嵌入之间的余弦相似度，并同时考虑一元语法和多词短语。ROUGE-L 评估生成文本与参考文本之间的最长公共子序列。纳入这些指标，可以定量比较生成的报告与真实标注及先前模型的结果，从而深入了解所提出方法的性能。

结果在表4中，我们展示了所提出方法在两个医学图像报告生成数据集 IU X-Ray 和 MIMIC-CXR 上的结果。该方法将先验信息融入最先进的自然语言生成模型，并在所有自然语言生成指标上始终优于基线方法，证明其能够提升自然语言生成模型所生成医学图像报告的质量。

在 IU X射线数据集上，与基线模型相比，我们的方法在 R2Gen 和 M^2Tr 的所有自然语言生成指标上分别实现了 11.58% 和 4.49% 的平均改善，

尤其值得注意的是，CIDEr 指标的改善幅度最大，R2Gen 提高了 31.46%， M^2Tr 提高了 7.00%。这表明，根据 CIDEr 的评估结果，我们的方法生成的描述更多样且与上下文更相关，相较于其他指标，这些结果与人类对质量的判断更加吻合。

对于 MIMIC-CXR 数据集，与之前的模型相比，我们的方法在所有自然语言生成指标上分别使 R2Gen 和 M^2Tr 模型提升了 8.40% 和 9.62%。最显著的改善体现在 ROUGE-L 上，R2Gen 提升了 8.27%， M^2Tr 提升了 11.83%。这表明我们的方法能够生成语法更加正确的描述，这一点在医学报告中尤为重要，因为语言错误可能会造成严重后果。

我们发现，最高阶的 n 元语法（即 n=3, 4）表现出最显著的改善。这表明，引入外部先验信息对于生成流畅且信息丰富的句子尤为有益，因为这类句子通常包含更长的短语和更复杂的结构。

总体而言，我们的研究结果表明，整合外部先验信息能够提升现有自然语言生成模型在医学影像报告任务中的表现，从而生成信息更丰富、更加准确的医学报告。通过将额外的领域特定知识融入模型，我们能够生成更精确、信息更丰富的报告，同时最大限度地降低计算开销和训练数据需求。

5 分析

在本节中，我们比较真实标注、由我们提出的模型生成的合成报告，以及此前发表的两个模型 R2Gen 和M^2Tr，以评估我们的模型在生成简洁、准确且不包含无关或错误先验信息的报告方面的有效性。

表 5 （见附录）提供了各模型生成的合成报告示例，以及同一放射影像对应的真实标注。前两行比较了 R2Gen 和我们采用先验注入的模型所生成的报告。我们观察到，R2Gen 会生成错误的既往检查表述，例如“与既往检查相比”“与既往相比无变化”以及“再次无变化”，这些表述均指向实际上并不存在的既往检查。相比之下，我们的模型生成的报告更加简洁、准确，且……不包含任何先验表达的报告，因此在自然语言生成指标上表现更佳。

同样，附录中表5 的最后两行比较了由 M^2Tr 和我们提出的模型生成的报告。 M^2Tr 生成的报告包含错误的先验表达，例如“在既往检查中出现”和“再次观察到”，而我们的模型则避免包含任何比较短语。此外，由于比较需要额外说明，包含先验表达的报告往往更长。然而，报告生成模型实际上无法获取既往检查进行比较，因此加入先验表达无关紧要，甚至会造成误导。因此，我们的模型可以通过先验条件直接控制这些短语的生成。

总体而言，与 R2Gen 和 M^2Tr 生成的报告相比，我们提出的模型生成的合成报告更加简洁准确，这一点可由其在自然语言生成指标上的更高表现得到证明。我们的模型通过基于规则的标注器避免无关或错误的先验表达，并生成仅包含相关且准确的信息的报告，从而实现了这一点。模型生成的这些简洁而精确的报告将有效协助放射科医生开展工作。

6 局限性与伦理考量

我们提出的方法存在一些值得讨论的局限性和伦理考量。我们的方法在很大程度上依赖于基于规则的标注器。然而，需要认识到，该标注器可能无法捕捉未见过的模式或变化，从而可能限制各项评估指标的改进。此外，由于资源有限，本研究未能对基于规则的标注器进行全面的人工评估。因此，未来的工作应开展详细评估，以衡量其性能并解决任何潜在的局限性。

与京都大学三位放射科医生合作是我们工作中的一个关键环节。基于规则的标注器中设计的正则表达式，经过计算机科学家和放射科医生的共同确认，得到了验证。不过，需要注意的是，参与合作的放射科医生主要在日本的医院环境中工作。这可能会引入特定于当地情境的潜在偏差或模式。因此，有必要与来自不同地区的放射科医生交叉核验基于规则的标注器的性能

和医疗保健系统，以确保其适用范围更广，并尽量减少潜在偏差。

关于本研究所使用的数据集，我们仅使用了经过适当匿名化和去标识化处理的公开数据集，以解决隐私方面的顾虑。然而，必须强调的是，如果未来有包含对比检查的数据集可供使用，则必须采取额外的预防措施，以确保不会无意中泄露任何个人身份信息，也不会以可能识别个体患者身份的方式使用这些信息。

通过认识到这些局限性和伦理考量，我们旨在鼓励该领域未来的研究与讨论，推动放射学报告生成的发展，同时优先保障患者隐私、准确性和公平性。这些考量将有助于开发稳健且符合伦理规范的放射学报告生成方法。

7 结论

在本研究中，我们提出了一种从胸部X光图像生成医学报告的新方法，旨在弥合放射科医生的知识与生成模型缺乏先验信息之间的差距。为此，我们开发了一种基于规则的标注器，能够从 IU X-ray 和 MIMIC-CXR 数据集的放射学报告中提取比较先验信息。随后，我们将这些先验信息整合到用于条件报告生成的最先进模型中，使我们的方法能够模拟放射科医生掌握患者先验信息时的真实诊断流程。

我们的实验结果表明，与以往最先进的模型相比，我们的方法具有优越性，其表现为自然语言生成指标的性能提升，以及错误引用既往检查的情况显著减少。通过分析，我们证明，融入比较先验信息能够生成更加准确、简洁的报告，因此在提升胸部 X 射线图像医学报告生成的质量和效率方面具有巨大潜力。最终，这一进展将通过提供更加可靠、信息更加丰富的报告，使医疗专业人员和患者受益。

此外，我们的工作凸显了生成医疗报告的未来潜力，从端到端的角度来看……

如果能够获得包含所有既往检查的数据集，便可采用端到端的方式。利用全面的先验信息将进一步提升医学报告生成的准确性和有效性，为改善医疗成效铺平道路。

参考文献

Ronald Arenson 和 N Reed Dunnick。2006 年。培养更优秀的放射科医生。《美国放射学会杂志》，3(6)：389–393。

Shuang Bai 和 Shan An。2018。《自动图像描述生成综述》。《*Neurocomputing*》，311：291–304。

Adrian P Brady. 2017. 放射学中的错误与差异：不可避免还是可以避免？*Insights into I maging*, 8:171–182. Zhihong Chen, Yan Song, Tsung-Hui Chang, 和 Xiang Wan. 2020. 通过记忆驱动型变换器生成放射学报告。*arXiv* 预印本 *arXiv:2010.16056*. Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, 和 Rita Cucchiara. 2020. 用于图像描述的网状记忆变换器。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第 10578–10587 页。Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, 和 Clement J McDonald. 2016. 为分发和检索整理放射学检查数据集。《美国医学信息学会杂志》，23(2):304–310。Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, 和 Pranav Rajpurkar. 2021. 使用预训练的对比式语言图像模型，通过检索生成胸部X线报告。载于健康机器学习，第 209–219 页。PMLR。Kaiming He, Xiangyu Zhang, Shaoqing Ren, 和 Jian Sun. 2016. 用于图像识别的深度残差学习。载于*IEEE* 计算机视觉与模式识别会议论文集，第 770–778 页。Sepp Hochreiter 和 Jürgen Schmidhuber. 1997. 长短期记忆。*Neural Computation*, 9(8):1735–1780。Gao Huang, Zhuang Liu, Laurens Van Der Maaten, 和 Kilian Q Weinberger. 2017. 密集连接卷积网络。载于*IEEE* 计算机视觉与模式识别会议论文集，第 4700–4708 页。

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya 等. 2019. CheXpert：一个包含不确定性标签并提供专家对比的大型胸部X线影像数据集。载于*AAAI* 人工智能会议论文集，第 33 卷，第 590–597 页。

Baoyu Jing、Pengtao Xie 和 Eric Xing。2017。关于医学影像报告的自动生成。*arXiv* 预印本 *arXiv:1711.08195*。

Alistair EW Johnson、Tom J Pollard、Nathaniel R Greenbaum、Matthew P Lungren、Chih-ying Deng、Yifan Peng、Zhiyong Lu、Roger G Mark、Seth J Berkowitz 和 Steven Horng。2019年。MIMIC-CXR-JPG：一个大规模公开可用的带标签胸部X线影像数据库。*arXiv* 预印本 *arXiv:1901.07042*。

Elizabeth A Krupinski。2010。医学图像感知的当前观点。*Attention, Perception, &Psychophysics*，72(5):1205–1217。

Yann LeCun、Yoshua Bengio 和 Geoffrey Hinton。2015 年。深度学习。*Nature*，521(7553):436–444。

Christy Y Li、Xiaodan Liang、Zhiting Hu 和 Eric P Xing。2019。《知识驱动的编码、检索与释义改写：用于医学图像报告生成》。载于*AAAI* 人工智能会议论文集，第 33 卷，第 6666–6673 页。

Chin-Yew Lin。2004年。ROUGE：一种用于摘要自动评估的工具包。载于文本摘要的拓展方向，第74–81页。

Guanxiong Liu、Tzu-Ming Harry Hsu、Matthew McDer- mott、Willie Boag、Wei-Hung Weng、Peter Szolovits 和 Marzyeh Ghassemi。2019a。临床准确的胸部X线报告生成。载于医疗保健机器学习会议，第 249–269 页。PMLR。

Xiaoxiao Liu、Qingyang Xu 和 Ning Wang。2019b。基于深度神经网络的图像描述生成综述。*TheVisual Computer*，35(3):445–470。

Justin Lovelace 和 Bobak Mortazavi。2020 年。学习生成临床连贯的胸部X线报告。载于 计算语言学协会会议论文集：*EMNLP 2020*，第 1235–1243 页。

Yasuhide Miura、Yuhao Zhang、Emily Bao Tsai、Cur- tis P Langlotz 和 Dan Jurafsky。2020 年。改进图像到文本放射学报告生成的事实完整性和一致性。*arXiv* 预印本 *arXiv:2010.10042*。

Farhad Nooralahzadeh、Nicolas Perez Gonzalez、Thomas Frauenfelder、Koji Fujimoto 和 Michael Krauthammer。2021 年。基于渐进式 Transformer 的放射学报告生成。*arXiv* 预印本 *arXiv:2102.09777*。

欧洲放射学会（ESR）http://www. myESR. org communications@ myESR. org。2011。《放射学报告撰写规范》。指南。

来自欧洲放射学会（ESR）。*InsightsInto Imaging*，2(2):93–96。Kishore Papineni、Salim Roukos、Todd Ward 和 Wei-Jing Zhu。2002年。BLEU：一种机器翻译自动评估方法。载于计算语言学协会第*40*届年会论文集，第311–318页。Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021年。通过自然语言监督学习可迁移的视觉模型。载于国际机器学习会议，第8748–8763页。PMLR。

Vignav Ramesh、Nathan A Chi 和 Pranav Rajpurkar。2022 年。通过移除对不存在的既往影像的幻觉性引用，改进放射学报告生成系统。载于 *MachineLearning forHealth*，第 456–473 页。PMLR。

Graciela Ramirez-Alonso、Olanda Prieto-Ordaz、Roberto López-Santillan 和 Manuel Montes-Y-Gómez。2022。基于放射影像的医学报告生成：综述。*IEEE* 拉丁美洲汇刊，20(6):986–999。

Paul Suetens。2017。医学成像基础。剑桥大学出版社。

Ramakrishna Vedantam、C Lawrence Zitnick 和 Devi Parikh。2015。Cider：基于共识的图像描述评估。载于*IEEE* 计算机视觉与模式识别会议论文集，第4566–4575页。

Xiaosong Wang、Yifan Peng、Le Lu、Zhiyong Lu 和 Ronald M Summers。2018。Tienet：用于胸部X光片常见胸部疾病分类与报告的文本-图像嵌入网络。载于*IEEE* 计算机视觉与模式识别会议论文集，第9049–9058页。

Yuan Xue、Tao Xu、L Rodney Long、Zhiyun Xue、Sameer Antani、George R Thoma 和 Xiaolei Huang。2018。采用注意力机制的多模态循环模型，用于自动放射学报告生成。发表于国际医学图像计算与计算机辅助介入大会，第 457–466 页。Springer。

Changchang Yin、Buyue Qian、Jishang Wei、Xiaoyu Li、Xianli Zhang、Yang Li 和 Qinghua Zheng。2019。基于分层循环神经网络的医学影像诊断报告自动生成。发表于*2019 IEEE* 国际数据挖掘大会（*ICDM*），第 728–737 页。IEEE。

Jianbo Yuan、Haofu Liao、Rui Luo 和 Jiebo Luo。2019。基于多视图图像融合与医学概念丰富化的自动放射学报告生成。发表于国际医学图像计算与计算机辅助介入大会，第 721–729 页。Springer。

Zizhao Zhang、Yuanpu Xie、Fuyong Xing、Mason Mc-Gough 和 Lin Yang。2017。Mdnet：一种语义和视觉上均可解释的医学图像诊断网络。发表于 *IEEE* 计算机视觉与模式识别会议论文集，第 6428–6436 页。

真实标注	R2Gen（Chen 等人，2020）	带先验信息的 R2Gen（我们的方法）
心纵隔轮廓大小和形态正- 常。肺容积偏低，未见局灶性气- 肺容积无局灶性气胸或 大量胸腔积液。正常 xxxx。	气管位于中线。 心脏纵隔 轮廓正常，且 未变，**与之前的检查相比** **次**。心脏上方可见密度影 xxxx 是 xxxx 冠状动脉支架。 双侧肺叶内可见小圆形钙化密度影， 这些密度影位于双侧肺叶，**未** **较既往检查**和 xxxx**无变化** 提示陈旧性肉芽 肿性疾病的后遗改变。	心纵隔轮廓 外观在正常范围内。 胸主动脉迂曲。未见局灶性肺- 实变区域。 未见气胸。未见大量胸腔 积液。轻度退行性改变 以及胸椎骨质减少。
心纵隔轮廓在正常范围内。 正常范围内。未见局灶性实变。 可见右肺下叶瘢痕。无其他异常。 气胸或大量胸腔积液。 存在肉芽肿。无急性骨骼异常。 异常。	气管位于正中。 心血管- 纵隔轮廓正常。该 上胸椎**再次不** **与先前相比发生了变化**。 透亮的肺部- 肺实质表现一致，符合..... 影像表现伴有肺气肿，且呈现 **与此前相比无变化 xxxx**。未发现任何- 气胸的证据。 无局灶性病灶 肺泡疾病或胸腔积液。 右肺内侧模糊密度影 顶端最多是 xxxx。	心纵隔轮廓正常。 大小和轮廓正常。 无局灶性病灶 无气胸或大量胸腔积液。 未见急性骨异常。
真实标注	M^2Tr（Cornia 等人，2020）	M^2Tr 含既往影像（我们的方法）
肺部无变化 xxxx。 xxxx 乳腺不透光影- 右肺下叶存在异常。 未见局灶性浸润。心脏及纵隔—— astinum 未见明显异常。主动脉正常- mal.	心纵隔轮廓稳定 纵隔钙化，心脏正常。 影像提示既往肉芽肿性病变， 肺尖不规则影亦 **见于既往检查**，提示 瘢痕形成。未见局灶性肺泡影， 未见明确胸腔积液。未见典型 肺水肿表现。未见 气胸。	心脏大小和轮廓正常。 纵隔未增宽。 肺过度充气。散在 肉芽肿。未见局灶性肺泡疾病。 未见大量胸腔积液或气胸- 胸廓。xxxx 完整。
气管居中。未见 胸膜异常或局灶性肺实变。 心脏大小正常。	心脏大小处于正常上限。 纵隔未见异常。 The 肺充气不足，但整体清晰。 明显的退行性改变累及 xxxx **再次可见**双侧。	心脏大小正常。纵 隔未见异常。肺 清晰。

表 5：IU X-ray 的真实报告（第一列）、R2Gen 和 M^2Tr 生成的报告（第二列），以及我们的模型生成的报告。先验表达以粗体显示。

模型	自然语言生成指标（IU X-ray）					
	BL-1	BL-2	BL-3	BL-4	CIDEr	RG-L
R2Gen	$0.421_{\pm 0.001}$	$0.262_{\pm 0.003}$	$0.183_{\pm 0.004}$	$0.137_{\pm 0.005}$	$0.480_{\pm 0.046}$	$0.337_{\pm 0.005}$
有先前影像（本文方法）	$0.438_{\pm 0.003}$	$0.280_{\pm 0.002}$	$0.201_{\pm 0.002}$	$0.155_{\pm 0.002}$	$0.631_{\pm 0.028}$	$0.351_{\pm 0.001}$
M^2Tr	$0.400_{\pm 0.003}$	$0.240_{\pm 0.002}$	$0.159_{\pm 0.002}$	$0.112_{\pm 0.002}$	$0.300_{\pm 0.004}$	$0.324_{\pm 0.002}$
有先前影像（本文方法）	$0.406_{\pm 0.003}$	$0.249_{\pm 0.002}$	$0.167_{\pm 0.001}$	$0.120_{\pm 0.001}$	$0.323_{\pm 0.001}$	$0.330_{\pm 0.001}$
模型	自然语言生成指标（MIMIC-CXR）					
	BL-1	BL-2	BL-3	BL-4	CIDEr	RG-L
R2Gen	$0.335_{\pm 0.001}$	$0.206_{\pm 0.003}$	$0.138_{\pm 0.005}$	$0.100_{\pm 0.002}$	$0.148_{\pm 0.006}$	$0.278_{\pm 0.003}$
有先前影像（本文方法）	$0.342_{\pm 0.002}$	$0.222_{\pm 0.002}$	$0.152_{\pm 0.002}$	$0.110_{\pm 0.001}$	$0.166_{\pm 0.004}$	$0.301_{\pm 0.005}$
M^2Tr	$0.353_{\pm 0.001}$	$0.211_{\pm 0.003}$	$0.137_{\pm 0.003}$	$0.094_{\pm 0.002}$	$0.089_{\pm 0.002}$	$0.262_{\pm 0.002}$
有先前影像（本文方法）	$0.357_{\pm 0.001}$	$0.224_{\pm 0.001}$	$0.151_{\pm 0.002}$	$0.108_{\pm 0.003}$	$0.101_{\pm 0.001}$	$0.293_{\pm 0.003}$

表6：基线模型和融入先验信息的模型的训练结果。上表为 IU X-ray 的结果，下表为 MIMIC-CXR 的结果。我们的方法的结果以灰色行显示。所有指标均为 3 次运行的平均值（均值 ± 标准差）。