arXiv:2303.01258v1 [cs.CL] 1 Mar 2023

用于核医学报告分类的领域适配大型语言模型

Zachary Huemann威斯康星大学麦迪逊分校，美国威斯康星州麦迪逊 zhuemann@wisc.edu

Changhee Lee威斯康星大学麦迪逊分校，美国威斯康星州麦迪逊 changhee.lee@wisc.edu

Junjie Hu威斯康星大学麦迪逊分校，美国威斯康星州麦迪逊 junjie.hu@wisc.edu

Steve Y. Cho威斯康星大学麦迪逊分校，美国威斯康星州麦迪逊 cho85@wisc.edu

Tyler Bradshaw威斯康星大学麦迪逊分校，美国威斯康星州麦迪逊 tbradshaw@wisc.edu

2023年3月3日

摘要

随着基于 Transformer 的语言模型在医学领域中的应用日益广泛，尚不清楚这些模型对核医学的泛化能力如何，因为核医学具有领域专属词汇和独特的报告风格。本研究基于临床18F-氟脱氧葡萄糖（FDG）PET/CT报告，通过针对5分制Deauville评分预测任务适配语言模型，评估了领域适配在核医学中的价值。我们从临床影像数据库中回顾性检索了2008年至2018年间的4542份文本报告和1664幅图像，这些资料来自FDG PET/CT淋巴瘤检查。我们从报告中删除Deauville评分，并将报告中的剩余文本用作模型输入。我们采用多个通用Transformer语言模型，将报告分类为Deauville评分1至5。随后，我们使用掩码语言建模将这些模型适配至核医学领域，并评估其对分类性能的影响。在七折蒙特卡洛交叉验证中，我们将这些语言模型与视觉模型、多模态视觉语言模型及一名核医学医师进行了比较，结果报告为均值和标准差。领域适配提升了所有语言模型的性能。例如，BERT的五分类准确率从61.3 ± 2.9%提高至领域适配后的65.7 ± 2.2%。性能最佳的模型（领域适配的RoBERTa）取得了77.4 ± 3.4%的五分类准确率，高于医师的表现（66%）和最佳视觉模型的表现（48.1 ± 3.5%），并与多模态模型的表现（77.2 ± 3.2%）相近。领域适配提升了大语言模型解读核医学文本报告的性能。

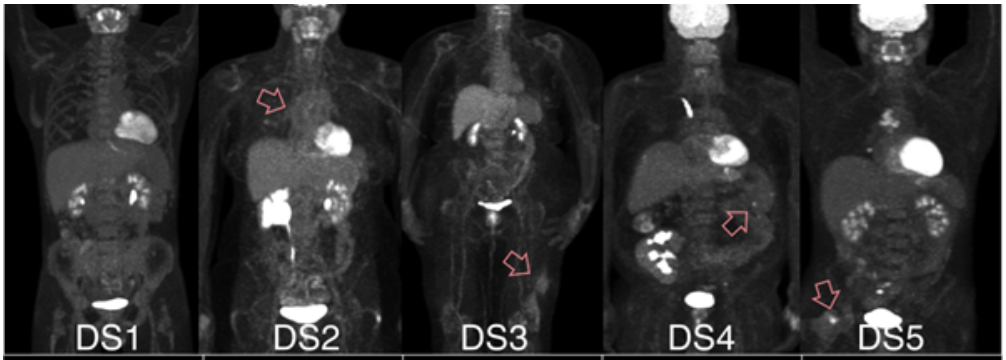
1 引言

一份放射学报告是放射学检查的最终成果。它代表对患者影像的正式解读，旨在为后续护理提供可操作的信息。临床影像数据库中存储着大量信息丰富的放射学报告，因此，基于人工智能（AI）的语言模型开始在放射学领域得到越来越多的应用，也就不足为奇了。语言模型可以从报告中提取训练标签 [1][2][3], 总结报告 [4], 根据图像生成报告 [5], 等等 [6]。尽管语言模型在各个领域表现出色，但在大型通用文本语料库上训练的语言模型，应用于放射学和核医学时，表现可能并不理想。大型预训练语言模型，例如基于Transformer的双向编码器表示（BERT） [7], , 通常通过自监督训练开发，其中包括

[1]本文已提交至《放射学：人工智能》，待考虑发表。版权可能在未另行通知的情况下转让，届时此版本可能不再可访问。

预测训练语料库中被遮蔽的词语或相邻句子。研究表明，通过使用特定领域的文本进行额外的自监督训练，可以将通用语言模型适配到特定领域，从而提升某些任务的表现 [8]。例如，BioBERT是在PubMed的生物医学语料库上进行额外预训练，从BERT适配而来 [8]。经过领域适配的BioBERT在应用于下游生物医学任务时，表现优于BERT。其他研究也证实了领域适配的益处 [9] [10]。

然而，将语言模型适配至核医学领域的研究仍不充分。核医学报告包含一些在放射学其他领域并不常用的独特术语。例如，“高代谢”、“局灶性摄取”和“SUVmax”都是较为专业的术语，但对于解读核医学报告至关重要。目前尚不清楚通用语言模型或经过生物医学领域适配的语言模型能否理解核医学词汇的深层含义。此外，领域适配能带来多大收益也尚不确定。在这项可行性研究中，我们评估了大语言模型解读和分类 PET/CT 文本报告的性能。我们研究的预测任务是根据 Deauville标准进行五分制评分。Deauville评分（DS）常规用于淋巴瘤 PET 检查的临床报告，以作出诊断并评估疗效（图 1）。我们探讨了 DS 预测任务，因为它可代表诊断提取等其他类似的解读任务，并要求语言模型学习自由文本报告的整体含义。我们将语言模型的分类性能与人类专家、视觉模型和多模态模型进行了比较。我们假设，针对核医学进行领域适配将提升模型在下游 DS 预测任务中的性能。



No uptake	Uptake less than mediastinum	Uptake between mediastinum and liver	Uptake moderately more than liver	Uptake markedly increased; new disease
“FDG PET-CT findings of complete metabolic treatment response with no abnormal residual FDG uptake. Deauville 1.”	“There is marked interval decreased metabolic activity and size of mediastinal, bilateral internal mammary, and subpectoral lymphadenopathy. Remaining FDG activity in these regions are less than or equal to blood pool compatible with Deauville 2.”	“There is interval decrease in the size of the metabolic activity of a soft tissue mass along the anterolateral aspect of the distal left thigh. This mass shows mild residual hypermetabolic activity. Deauville score 3.”	“Several hypermetabolic foci within the spleen are likely due to lymphomatous involvement. These findings are compatible with a Deauville score of 4.”	“There is a new intense hypermetabolic lesion within the right femoral head. There is also interval increased activity in multiple presacral and paratracheal lymph nodes. Deauville 5.”

图1：预测任务是依据 Deauville标准对淋巴瘤 PET 治疗反应进行临床报告时所采用的五分制。图中展示了每种 Deauville（DS）类别对应的放射学报告文本描述示例及最大强度投影图像。

2 方法

2.1 数据收集

我们采用经机构审查委员会批准、豁免知情同意的回顾性方案，检索了各大学的图像存档与通信系统（PACS），筛选检查指征或印象中包含“淋巴瘤”一词的临床 18F-氟脱氧葡萄糖（FDG）PET/CT检查。我们于2008至2018年间共找到4,542项检查，这些检查由至少44位不同的医师（12位放射学教员和32位住院医师）撰写报告。我们使用 ClinicalTrialProcessor 对图像进行匿名化处理，并将图像及其放射学报告一并下载。

2.2 Deauville评分提取

通过在报告文本中搜索字符串“Deauville”及其常见拼写错误，找出相应报告中含有医生指定的DS的PET/CT检查。随后进行N元语法分析，以识别所有

Deauville评分	定义	频率
1	无高于背景水平的摄取	313
2	摄取低于纵隔	355
3	摄取高于纵隔，但低于或等于肝脏	155
4	任何部位的摄取均较肝脏中度增加	221
5	任何部位的摄取均较肝脏显著增加	620
总计	–	1664

表1：数据集中各Deauville评分的频率。

医师报告DS的方式（例如“Deauville评分为1”、“Deauville量表上为1”等）。随后根据DS为每项检查赋予1-5的分类标签。如果报告中包含两个或更多DS，例如进行病灶特异性报告时，则使用数值最高的DS作为检查级标签。随后从报告中删去DS，并将剩余文本作为语言模型的输入。

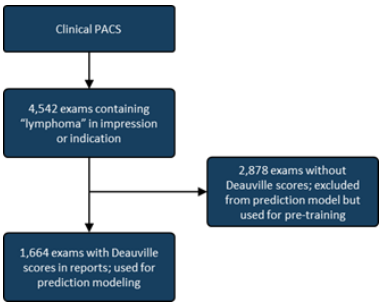


图2：患者纳入/排除流程图。

2.3 语言模型

我们评估了四种基于 Transformer 的语言模型（图 3）。其中包括 BERT，这是一种拥有 1.1 亿个参数的 Transformer 模型，包含 12 个 Transformer 模块和 12 个自注意力头，使用 BooksCorpus 和英文维基百科进行预训练，采用了两项预训练任务：掩码语言建模（MLM）和下一句预测（NSP） [7]。我们还评估了 BioClinicalBERT [11], , 该模型以 BioBERT [8] 为初始模型，随后又使用临床笔记（8.8 亿词）进行预训练[12]。RadBERT 是一种拥有 1.23 亿个参数的模型，以 RoBERTa-Base 为初始模型，并使用 400 万份放射学报告进行预训练 [4]。最后，我们评估了 RoBERTa [13], , 这是一种拥有 3.55 亿个参数的模型，包含 24 个 Transformer 模块和 16 个自注意力头，并使用动态掩码语言模型进行预训练。所有预训练语言模型均用于评估其将放射学报告分类为 DS 类别 1–5 的能力。脱敏后的 PET/CT 文本报告经过标点符号去除、日期剔除和数值取整等预处理。我们使用自定义同义词列表进行同义词替换，以统一不同医师的词汇用法（例如，“SUV” = 、“SUVmax” = 和“标准摄取值”），但后来发现这对模型性能没有可测量的影响。由于语言模型对输入文本有 512 个词元的限制，我们优先输入印象部分，然后尽可能纳入所见部分。这是因为印象部分通常会概述所见部分。输入文本通过子词分词转换为词嵌入，再输入预训练语言模型。每种语言模型的输出均作为一个三层分类器的输入，该分类器由两个全连接层（BERT 为 768 个节点，RoBERTa 为 1024 个节点）和一个最终的 Softmax 层组成。分类器使用 Adam 优化器和交叉熵损失进行训练。所有模型均通过七轮随机抽样交叉验证进行训练和评估，每轮划分为 80% 的训练集、10% 的测试集和 10% 的验证集（用于早停）。所有语言模型均从 Hugging Face 库 [14] 导入，并在 PyTorch 中实现。

2.4 领域适配

我们评估了领域适配对语言模型分类性能的影响。我们使用 4,542 份 PET/CT 报告（共 200 万词）作为预训练语料库，进行了额外的掩码语言模型（MLM）预训练。MLM 预训练通过随机遮蔽 15% 的词元，并预测被遮蔽的词元来完成。我们使用 1e-6 的学习率对模型进行了 3 个周期的预训练，以防止过拟合。随后，按照上述相同的训练流程，对采用领域适配的语言模型的分类模型进行了微调。

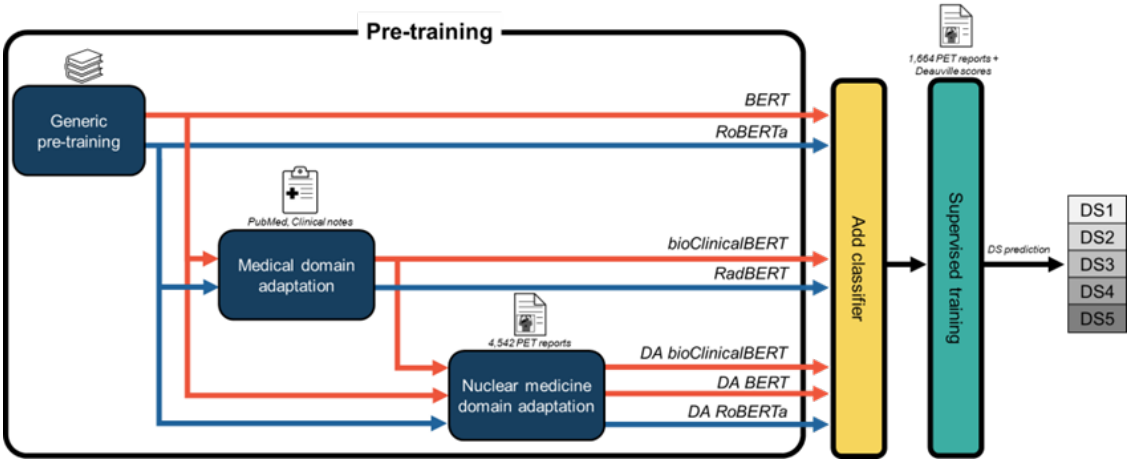


图3：展示了各种语言模型的预训练和微调过程。将基于通用文本训练的Transformer模型（BERT和RoBERTa）与在医学语料库上进行额外预训练的模型（bioClinicalBERT），以及专门针对核医学进行领域适配（DA）的模型（DA bioClinicalBERT、DA BERT和DA RoBERTa）进行了比较。DA包括使用来自PET报告的文本进行掩码语言建模的进一步预训练。每个模型后接一个分类器，并进行监督训练，以预测5分类Deauville评分（DS）。

2.5 对照方法

我们训练了以 PET图像为输入的视觉模型，以进行比较。我们训练了视觉Transformer（ViT） [15] 和 EfficientNet-B7 卷积神经网络 [16]，以对应的冠状位最大强度投影（MIP）PET图像为输入来预测 DS。ViT 模型由 12 个 Transformer层组成，每层有 12 个注意力头。我们使用 ImageNet-21k [17]对模型进行了预训练。图像裁剪至大腿部位，并调整为 384x384，同时进行像素归一化。我们采用了标准的数据增强：水平翻转、垂直翻转、随机旋转和随机平移。

我们还将语言模型和视觉模型的性能与一名人类专家进行了比较。一名具有4年核医学经验的核医学医师获得了50个随机病例的隐去身份信息报告和MIP PET图像，并被要求预测阅片医师最初判定的DS。

最后，我们评估了一个可同时处理配对图像和文本的多模态模型的性能。我们旨在确定文本报告和图像在预测任务中包含的是互补信息还是冗余信息。我们的多模态模型结合了 ViT 和领域适配的RoBERTa。将这两个模型生成的嵌入拼接后，输入一个由两层各含 1024 个节点的 3 层分类器中，见图 4。

3 结果

在 4,542 次淋巴瘤 PET/CT检查中，共有 1,664 次检查包含 DS，表 1 显示了不同 DS 类别的频率。未包含 DS 的检查是在我们诊所采用 Deauville标准之前进行的（尤其是基线扫描），或很可能是因淋巴瘤以外的适应证而进行的。患者年龄为 18 至 95 岁（平均年龄：53 岁），女性和男性分别占 42% 和 58%。

图 5 展示了采用和未采用领域适配的语言模型对 DS 分类的结果。加权 Cohen's κ 见表 2。BERT 和 RoBERTa 模型在 5 类分类中的准确率分别为 61.3 ± 2.9% 和 73.7± 4.0%。BioClinicalBERT 的表现与 BERT 相近：63.0 ± 3.6%。领域适配提升了每个模型的预测性能：BERT 提升至 65.7 ± 2.2%，BioClinicalBERT 提升至 66.4 ± 3.2%，RoBERTa 提升至 77.4 ± 3.4%（见图 5）。RadBERT 的准确率为 73.0 ± 1.9%，与较大的 RoBERTa 模型相当。如图 5 所示，较大的 RoBERTa 模型优于 BERT 模型，甚至超过了人类专家（66%）。

视觉模型的表现显著逊于语言模型。ViT 和 EfficientNet 的准确率分别为 46.5± 3.1% 和 48.1 ± 3.5%。图 5 展示了视觉模型与语言模型的比较结果。加权 Cohen's κ显示于表 2。

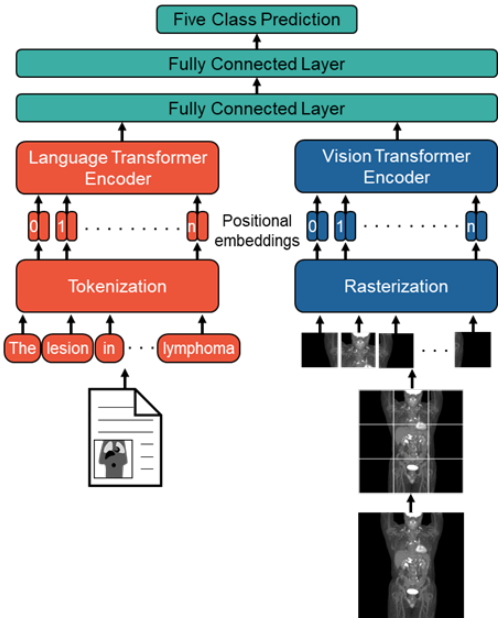


图 4：该多模态模型由两条针对不同模态的路径组成：一条基于 RoBERTa 语言Transformer，另一条基于视觉Transformer（ViT）。两条路径的嵌入经过拼接后，输入一个 3 层分类器。

方法	加权 Cohen's κ	5 类准确率（%）
BERT	0.65	61.3 ± 2.9
RoBERTa	0.81	73.7 ± 4.0
bioClinicalBERT	0.68	63.0 ± 3.6
RadBERT	0.81	73.0 ± 1.9
DA BERT	0.70	65.7 ± 2.2
DA RoBERTa	0.83	77.4 ± 3.4
DA bioClinicalBERT	0.74	66.4 ± 3.2
视觉Transformer	0.39	46.5 ± 3.1
EfficientNet	0.53	47.1 ± 3.5
多模态	0.83	77.2 ± 3.2
人类专家	0.79	66

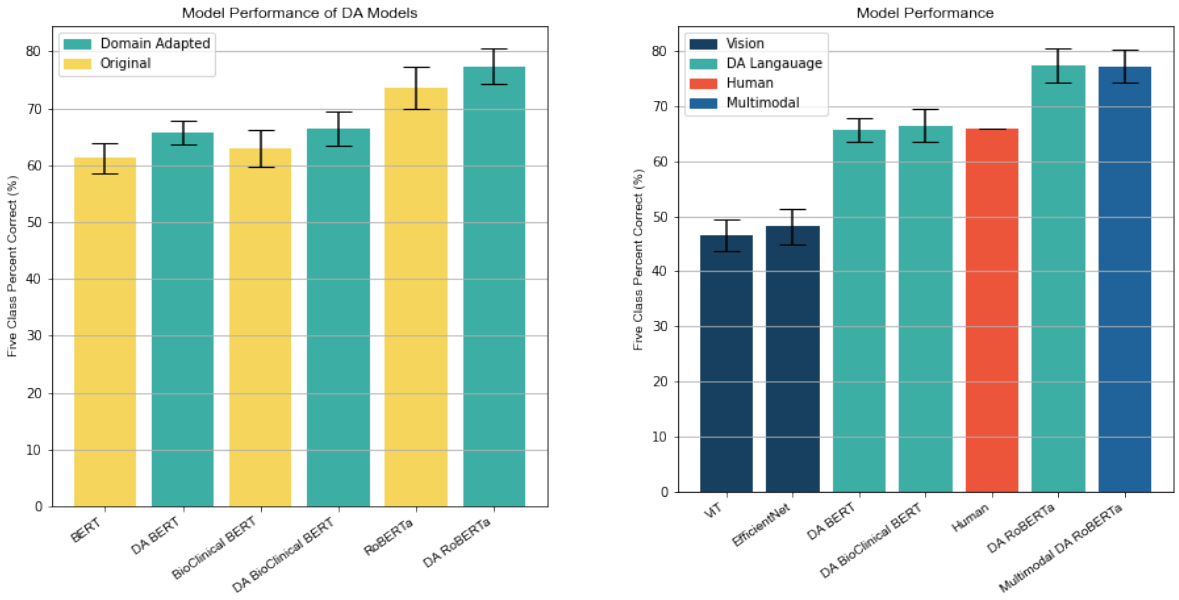
表 2：不同方法在预测 Deauville评分方面的表现。

多模态模型的表现未能超过领域适配的RoBERTa语言模型，后者构成了多模态模型的语言通路。多模态模型的准确率为 77.2 ± 3.2%，而领域适配的RoBERTa模型的准确率为 77.4 ± 3.4%，这表明，在该特定预测任务中，从图像中学习到的表征并未补充从文本中学习到的表征。两种模型的表现均优于人类专家。

4 讨论

我们发现，在根据核医学报告预测 DS 方面，领域适配的语言模型表现优于通用语言模型。五分类准确率提升了 3.4–4.4%。我们的 DA BERT 模型在 BERT 的基础上额外使用了一个包含 200 万词的语料库，使准确率提升了 4.4%。BioClincialBERT 是一种标准的 BERT 模型，在此基础上额外使用了一个包含 8.8 亿词的生物医学文本语料库，使准确率提升了 1.7%。与通用生物医学领域预训练（例如 BioClincialBERT）相比，这凸显了在核医学领域进行领域特定预训练的优势。RadBERT 借助领域特定预训练，达到了与规模更大的 RoBERTa 模型相近的表现；后者的参数量几乎是 RadBERT 的 3 倍。鉴于这些结果，以及领域适配相对容易实现——它不要求对数据进行标注，且借助现有的 Python库，相关技术流程已得到便捷的简化——我们认为，旨在处理核医学数据的语言模型应进行领域适配。此外，适配于通用生物医学领域的语言模型（例如 BioBERT）并未明显优于通用 BERT 模型，且二者的表现都不如核医学领域适配的模型。这表明，在用于将语言模型适配到放射学领域的任何语料库中，纳入核医学文本将十分重要。

根据核医学报告预测DS，可以说是语言模型面临的一项困难任务。虽然依据DS对放射学报告进行分类本身并不是一项具有临床实用价值的任务，但它对语言模型而言构成了复杂的解读挑战，而且这一问题有现成的数据标签（即医生指定的DS），便于开展大规模分析。我们探索这项任务，是因为预期它能够代表其他类似的解读任务（例如诊断提取、报告摘要生成等），而那些任务并没有方便获取的标签。与一些允许模型针对文本输入进行较局部处理的语言任务不同，例如命名实体识别或拼写/语法检查，DS预测需要对报告进行更全面的解读。模型必须能够权衡疾病阴性句子（例如“既往脾脏摄取已消退”）与疾病阳性句子（例如“新出现的



(a) (b)

图 5: （a）：领域适配（DA）语言模型与非领域适配模型的比较。（b）：领域适配模型、dision 模型、多模态模型和人类专家在预测 Deauville评分方面的比较。图中展示了各模型相对于报告的 Deauville评分的五分类准确率 ±SD。

纵隔淋巴结高代谢”），因为报告通常兼有这两类句子。此外，相邻两个Deauville类别之间的差异可能很小。总体而言，语言模型在这项任务中表现出色，部分模型的表现甚至超过人类专家，这为其未来在核医学领域的应用带来了极大希望。

我们发现，仅基于图像进行判断的模型在 DS 预测任务中的准确率低于语言模型。考虑到该任务是预测原阅片医师给出的 DS，而这位医师也撰写了报告，因此这一结果并不令人意外；这两者都反映了对 PET图像的主观解读，而非某种客观的真实标准。我们使用多模态模型来判断图像提供的信息能否补充报告中的语言信息。但我们发现，语言信息在预测任务中占据主导地位，以至于图像信息并未带来额外收益。

本研究存在一些值得注意的局限性。首先，本研究所使用的报告和图像均来源于单一机构，报告风格的多样性可能不足。尽管我们的文本语料库中的报告由44名不同的医师撰写，其中包括主治医师、受训医师、核医学专家和双重培训放射科医师，但单一机构的报告实践可能比不同机构之间更加统一。我们的研究还仅限于单一预测任务。因此，我们的结果在多大程度上能够推广到不同机构和不同预测任务，尚不确定。此外，我们仅以一名人类专家作为基准，因此没有可用于比较该特定任务的观察者间一致率。最后，我们将PET图像压缩为二维最大密度投影图，以便使用现有的图像分类变换器和卷积神经网络。研究发现，最大密度投影图对人工智能算法具有预后价值[18]，但尽管其在临床阅片中被广泛使用，仍存在一些固有局限性。

总之，我们发现大语言模型在解读核医学文本报告方面能力出众。通过自监督学习适配核医学领域的模型，表现优于通用模型，其中一些模型的预测性能甚至超过了人类专家。

参考文献

[1] Akshay Smit、Saahil Jain、Pranav Rajpurkar、Anuj Pareek、Andrew Ng 和 Matthew Lungren。结合自动标注器与专家标注，使用 BERT 实现准确的放射学报告标注。发表于《.....论文集》

*2020*年自然语言处理实证方法会议（*EMNLP*），第1500–1519页，线上，2020年11月。计算语言学协会。[2] Ali S. Tejani、Yee S. Ng、Yin Xi、Julia R. Fielding、Travis G. Browning 和 Jesse C. Rayan。多种预训练 BERT 模型在大型数据集数据标注自动化与加速中的性能。《放射学：人工智能》，4(4):e220007，2022年7月。[3] Matthias A. Fink、Klaus Kades、Arved Bischoff、Martin Moll、Merle Schnell、Maike Küchler、Gregor Köhler、Jan Sellner、Claus Peter Heussel、Hans-Ulrich Kauczor、Heinz-Peter Schlemmer、Klaus Maier-Hein、Tim F. Weber 和 Jens Kleesiek。基于深度学习和自然语言处理的结构化放射学报告肿瘤学结局评估。《放射学：人工智能》，4(5):e220055，2022年。[4] An Yan、Julian McAuley、Xing Lu、Jiang Du、Eric Y. Chang、Amilcare Gentili 和 Chun-Nan Hsu。RadBERT：将基于 Transformer 的语言模型适配于放射学。《放射学：人工智能》，4(4):e210258，2022年7月。[5] An Yan、Zexue He、Xing Lu、Jiang Du、Eric Chang、Amilcare Gentili、Julian McAuley 和 Chun-Nan Hsu。用于胸部X线报告生成的弱监督对比学习，2021年。[6] Vera Sorin、Yiftach Barash、Eli Konen 和 Eyal Klang。放射学自然语言处理的深度学习——基础与系统综述。《美国放射学会杂志》，17(5):639– 648，2020年5月。[7] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向 Transformers 预训练。[8] Jinhyuk Lee、Wonjin Yoon、Sungdong Kim、Donghyeon Kim、Sunkyu Kim、Chan Ho So 和 Jaewoo Kang。BioBERT：用于生物医学文本挖掘的预训练生物医学语言表征模型。《生物信息学》，36(4):1234–1240，2019年09月。[9] Gunvant R. Chaudhari、Tengxiao Liu、Timothy L. Chen、Gabby B. Joseph、Maya Vella、Yoo Jin Lee、Thienkhai H. Vu、Youngho Seo、Andreas M. Rauschecker、Charles E. McCulloch 和 Jae Ho Sohn。领域特定 BERT 在放射学报告语音识别错误检测中的应用。《放射学：人工智能》，4(4):e210185，2022年7月。[10] Yash Khare、Viraj Bagal、Minesh Mathew、Adithi Devi、U Deva Priyakumar 和 CV Jawahar。Mmbert：用于改进医学视觉问答的多模态 BERT 预训练，2021年。[11] Emily Alsentzer、John R. Murphy、Willie Boag、Wei-Hung Weng、Di Jin、Tristan Naumann 和 Matthew B. A. McDermott。公开可用的临床 BERT 嵌入，2019年。[12] Alistair E.W. Johnson、Tom J. Pollard、Lu Shen、Li-wei H. Lehman、Mengling Feng、Mohammad Ghassemi、Benjamin Moody、Peter Szolovits、Leo Anthony Celi 和 Roger G. Mark。MIMIC-III：一个免费开放的重症监护数据库。《科学数据》，3(1):160035，2016年5月。[13] Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar J oshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。Roberta：一种稳健优化的 BERT 预训练方法，2019年。[14] Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Remi Louf、Morgan Funtowicz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander Rush。Transformers：最先进的自然语言处理。在*2020*年自然语言处理实证方法会议：系统演示论文集中，第38–45页，线上，2020年。计算语言学协会。[15] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、Jakob Uszkoreit 和 Neil Houlsby。一张图像胜过16×16个词：用于大规模图像识别的 Transformers，2020年。[16] Mingxing Tan 和 Quoc V. Le。Efficientnet：重新思考卷积神经网络的模型缩放，2019年。[17] Tal Ridnik、Emanuel Ben-Baruch、Asaf Noy 和 Lihi Zelnik-Manor。面向大众的 ImageNet-21k 预训练，2021 年。[18] Kibrom Berihu Girum 、 Louis Rebaud 、 Anne-Ségolène Cottereau、Michel Meignan、Jérôme Clerc、Laetitia Vercellino、Olivier Casasnovas、Franck Morschhauser、Catherine Thieblemont 和 Irène Buvat。 ^{18} F-FDG PET 最大强度投影与人工智能：轻松测量弥漫性大B细胞淋巴瘤患者预后生物标志物的双赢组合。《核医学杂志》，第jnumed.121.263501页，2022年6月。