arXiv:2304.00487v1 [eess.IV] 2 Apr 2023

反事实的影响：胸部X光片阅片

JosephPaulCohen[1, 2, A] Rupert Brooks[3] Sovann En[3] Evan Zucker[1,2] Anuj Pareek[1,2]MatthewP Lungren[1,4,5] Akshay Chaudhari[1,2]1 斯坦福大学医学人工智能与成像中心，[2] 斯坦福大学放射科，[3] Nuance Communications，[4] 加州大学旧金山分校放射学与生物医学成像系，[5] 微软公司

摘要

本研究评估了反事实解释对胸部 *X* 光片解读的影响。我们开展了一项阅片研究，由两名放射科医师评估 *240* 个胸部 *X* 光片预测结果，并使用 *5* 分制对模型预测正确的置信度进行评分。其中一半预测为假阳性。每个预测结果均使用两种方法分别解释一次：一次采用传统归因方法，另一次采用反事实解释。总体结果表明，与传统方法相比，反事实解释能够使放射科医生对真阳性预测更有置信度（*0.15±0.95*，*p=0.01*），而假阳性预测仅小幅增加（*0.04±1.06*，*p=0.57*）。我们观察到，与其他任务相比，肿块和肺不张这两项具体预测任务似乎受益最大。

引言 我们研究反事实解释如何影响放射科医生解读神经网络模型对胸部 X 光片（CXR）所作预测的能力，并将其与传统归因方法进行比较。具体而言，我们关注假阳性预测问题，以及能否利用反事实来缓解这一问题。反事实解释是指将模型用于作出阳性预测的图像特征替换为良性特征，从而得到阴性预测。模型在面对分布外数据时，或在学到伪相关时，可能会作出错误预测。即使在部署前经过严格测试，这些问题仍可能出现在罕见输入中。为了确保模型因正确的理由而正确 [3], ，还有谁比领域专家更适合作出这一判断? 通过让医生直观确认模型用于作出预测的特征，他们就能发现与病理类型不一致之处，从而不采纳该预测。

[A]在加入 Amazon 之前完成的工作

方案与材料 本研究使用三种传统方法。方法 $input\ gradients$（称为 $grad$）计算输入相对于模型对所有阳性类别图像所作预测的绝对梯度 $\left|\frac{\partial \hat{y}_1}{\partial \mathbf{x}}\right|$。方法 引导反向传播 [4]（称为引导式梯度）仅反向传播正梯度，以尝试忽略彼此抵消的梯度。方法积分梯度 [5]（称为积分梯度）通过对输入图像 x_i 与全零基线图像之间的梯度进行积分来实现。

使用的反事实方法是潜变量偏移方法 [1]。该方法利用自编码器对像素变化进行正则化，以降低分类器的预测结果。这解决了生成对抗图像时的一个常见问题：图像中的变化难以察觉。该方法会生成一段简短的循环视频，展示模型用于作出预测的特征如何被移除。每一帧都呈现移除分类器用于作出预测的特征这一变换的一个步骤。

所用的分类器是 TorchXRayVision 库中的 DenseNet-121（称为 XRV-all） [2]，该模型在 7 个最大的胸部X光数据集（美国国立卫生研究院、PC、CheX、MIMIC-CXR、Google、OpenI、北美放射学会）上联合训练。在所测试病理类型上的曲线下面积均高于 0.77。

研究详情 本研究招募了两位放射科医生（A.J. 和 E.Z.，分别有 2 年和 12 年从业经验）。他们两次查看了 240 张图像（来自美国国立卫生研究院 ChestXray14 数据集 [6]），每张图像均由 XRV-all 模型预测为至少患有 6 种病理类型之一（肺不张、心脏扩大、积液、肺部阴影、肿块、气胸）。每种病理类型选取 40 张图像，其中 50% 被模型错误预测为阳性（假阳性）。如果图像带有阴性标签，而模型给出的预测为 >50%，则该预测被定义为错误预测；模型经过校准，使得验证数据上的曲线下面积曲线以 50% 预测值为判定阈值。

每个样本都通过界面使用传统归因方法和反事实方法进行查看

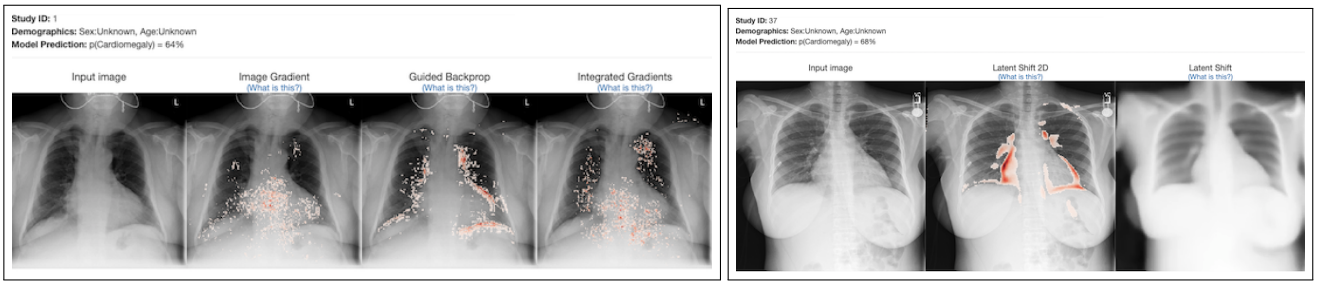


图1. 阅片研究中使用的逐图像界面截图。左侧是使用传统方法的界面，右侧是使用反事实方法的界面。反事实图像以动画形式显示，呈现输入图像与反事实版本之间的变换。界面还提供了热力图，帮助用户识别变化。

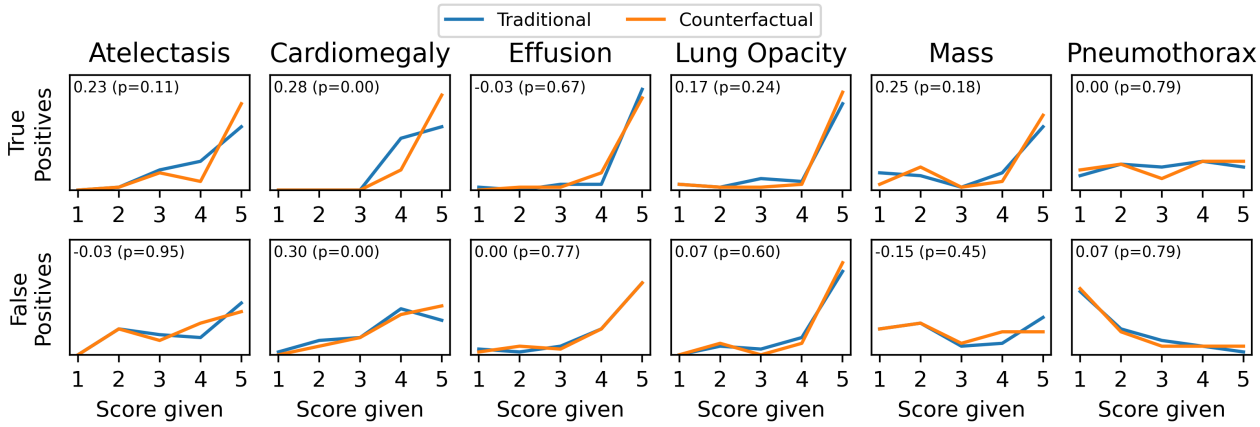


图2. 对调查问题“您对模型的预测有多大信心？”的回答直方图，按病理类型分别呈现。图表已按每种病理类型的回答数上限（40）进行归一化。理想的方法应使所有真阳性的得分为5，所有假阳性的得分为1。每张图左上角显示了方法之间的差值（反事实方法–传统方法）及p值。总体而言，反事实方法提高了真阳性预测的置信度，对假阳性则没有显著影响。

如图 1所示。使用传统方法时，放射科医生可以同时查看所有方法（梯度、引导式梯度和积分梯度）。使用反事实方法时，放射科医生可以查看一段动画图像，展示输入图像与反事实图像之间的变换，并查看标示变化位置的热力图。我们请放射科医生使用 5 点李克特量表回答以下问题：“您对模型的预测有多大信心？（1-5）”

主要研究结果见图2。总体而言，对于真阳性预测，使用反事实方法后置信度提高了0。15±0.95（采用威尔科克森符号秩检验，p=0.01）。总体而言，对于假阳性预测，置信度提高了0。04±1.06，但这一变化并不显著（p=0.57）。

肿块和肺不张预测任务的改善最为显著。对于肿块预测，真阳性的置信度提高了 0.25，而假阳性的置信度下降了 -0.15。对于肺不张预测，真阳性的置信度提高了 0.23，而假阳性的置信度下降了 -0.03。这些结果令人鼓舞，因为这两项任务中真阳性与假阳性之间的变化幅度最大，而这一趋势表明阅片表现有所改善。这一改善的一个解释是，分类器可能会将端口误认为肿块；反事实图像会通过缩小端口的尺寸清楚地表明，分类器并未识别出肿块。

在放射科医生的反馈中，他们认为反事实方法更加直观，并且觉得这增强了他们的置信度，让他们相信模型关注的是正确特征。他们观察到，这种方法关注的是异常区域的边界。一位放射科医生认为，模型使用了胸管这一虚假相关因素来预测气胸，而不是关注正确的区域。

我们发现，与传统方法相比，反事实解释能让放射科医生对真正例预测更有置信度。然而，我们也发现，识别假正例预测颇具挑战，这凸显了加强放射科医生与算法协同的必要性。

这份扩展摘要概述了该作品中进行的一项实验 [1]。详情请参阅该作品。

伦理

本作品仅使用了此前已公开的患者数据。本研究开展的研究已获得斯坦福大学伦理审查委员会的批准。

参考文献

[1] Joseph Paul Cohen、Rupert Brooks、Sovann En、Evan Zucker、Anuj Pareek、Matthew P. Lungren 和 Akshay Chaudhari。基于潜在偏移的 Gifsplanation：一种用于胸部 X 光片反事实生成的简单自编码器方法。深度学习医学影像，2021 年。1，2[2] Joseph Paul Cohen、Joseph D. Viviano、Paul Bertin、Paul Morrison、Parsa Torabian、Matteo Guarrera、Matthew P Lungren、Akshay Chaudhari、Rupert Brooks、Mohammad Hashir 和 Hadrien Bertrand。TorchXRayVision：胸部 X 光数据集和模型库。发表于深度学习医学影像，2022 年。1[3] Andrew Ross、Michael C Hughes 和 Finale Doshi-Velez。因正确的理由而正确：通过约束解释来训练可微模型。发表于国际人工智能联合会议，2017 年。1[4] Jost Tobias Springenberg、Alexey Dosovitskiy、Thomas Brox 和 Martin Riedmiller。追求简洁：全卷积网络。发表于国际学习表征会议研讨会，2015 年。1[5] Mukund Sundararajan、Ankur Taly 和 Qiqi Yan。深度网络的公理归因。发表于国际机器学习会议，2017 年。1[6] Xiaosong Wang、Yifan Peng、Le Lu、Zhiyong Lu、Mohammadadi Bagheri 和 Ronald M. Summers。ChestX-ray8：医院规模的胸部 X 光数据库，以及常见胸部疾病弱监督分类和定位的基准测试。发表于计算机视觉与模式识别会议，2017 年。1