Evaluate the Malignancy of Pulmonary Nodules Using the 3D Deep Leaky Noisy-or Network

Fangzhou Liao, Ming Liang, Zhe Li, Xiaolin Hu*, *Senior Member, IEEE* and Sen Song*

摘要：利用计算机断层扫描（**CT**）自动诊断肺癌涉及两个步骤：检测所有可疑病变（肺结节）和评估全肺/肺部恶性肿瘤。目前，关于第一步有许多研究，但第二步的相关研究很少。结节的存在并不能明确指示癌症，结节的形态与癌症之间存在着复杂的关系，肺癌的诊断需要仔细调查每一个可疑的结节和所有结节的信息整合。我们提出了一个三维深度神经网络来解决这个问题。该模型由两个模块组成。第一个是用于结节检测的**3D**区域提议网络，其输出所有可疑的结节。第二个是根据检测的置信度选择前五个结节，评估他们的癌症概率，并将它们与**leaky noisy-or gate**相结合，以获得受试者的肺癌概率。这两个模块共享相同的骨干网络**:**一个修改的**U-net**。训练数据不足导致的过拟合通过交替训练这两个模块来缓解。本文提出的模型赢得了**Data Science Bowl 2017**年比赛的第一名。该代码已公开可用[1]**.**

关键词:

肺结节检测，恶性肿瘤结节评估，深度学习，noisy-or **model, 3D** 卷积神经网络

I. 引言

肺癌是最常见和致命的恶性肿瘤之一。像其他癌症一样，治疗癌症的最佳解决方案是早期诊断和及时治疗。所以定期体检是必要的。立体胸部计算机断层扫描（CT）是肺癌诊断中常见的一种成像工具[1]。它根据对X射线的吸收来显现所有组织。肺部的病变称为肺结节。结节通常与正常组织具有相同的吸收水平，但具有独特的形状.支气管和血管是连续的管道系统，其根部厚,分枝薄，结节通常是球形的和孤立的。通常有经验的医生需要大约10分钟对患者进行彻底检查，因为有些结节很小很难找到。此外，结节有许多亚型，不同亚型对应的癌症概率是不同的。医生可以根据形态来评估结节的恶性程度，但是准确性很大程度上取决于医生的经验，不同的医生可能会给出不同的预测 [2].

计算机辅助诊断（CAD）适用于这个任务，因为计算机视觉模型能够以相同的质量快速扫描，而不会受到疲劳和情绪的影响。 深度学习的最新进展使得计算机视觉模型能够帮助医生诊断各种问题，并且在某些情况下，模型已经向医生发起了挑战 [3, 4, 5, 6, 7].

与一般的计算机视觉问题相比，自动肺癌诊断有几个困难。首先，结节检测是一个比二维物体检测困难的三维物体检测问题。由于有限的GPU内存,二维物体检测方法直接推广到三维面临技术困难。因此，一些方法使用二维区域提议网络（RPN）来获取单个二维图像中的提案(proposal)，然后将它们组合以生成三维提案[8,9]。更重要的是，标注3D数据通常比标注2D数据更困难，这可能使得深度学习模型由于过度拟合而失败。其次，结节形状多样（图1），结节与正常组织之间的差异是模糊的。因此，即使有经验的医生在某些情况下也不能达成共识[10]。第三，结节和癌症之间的关系是复杂的。结节的存在并不一定表明肺癌。对于有多发结节的患者，应考虑所有结节推断癌症的可能性。换句话说，与经典检测任务和经典分类任务不同，在这个任务中，标签对应于若干个对象。这是一个多实例学习（MIL）[11]问题，而MIL问题是计算机视觉中的一个难题。

为了解决这些困难，我们采取以下策略。我们建立了一个3D RPN [12]来直接预测结核的边界框。三维卷积神经网络（CNN）结构使网络能够捕捉复杂的特征。为了处理显存问题，使用基于patch的训练和测试策略。 该模型进行端到端训练，以实现高效优化。应用了多种数据扩增方法以解决过度拟合问题。检测器的阈值被设置的很低，以便包括所有可疑的结节。然后选择前五个可疑结节作为分类器的输入。在分类器中引入leaky noisy-or模型[13]来结合前五个结节的分数。

* Corresponding authors
1https://github.com/lfz/DSB2017

Fangzhou Liao, Zhe Li and Sen Song 清华大学医学院，北京100084
Ming Liang,Xiaolin Hu 清华大学信息科学与技术系国家重点实验室，清华信息技术国家重点实验室（TNList），清华大学计算机科学与技术系，北京100084。（电子邮件：xlhu@tsinghua.edu.cn）。
这项工作部分得到了中国国家基础研究计划（973计划）的资助。 2013CB329403，国家自然科学基金项目批准号。 91420201，61332007，61621136008 , 61620106010.

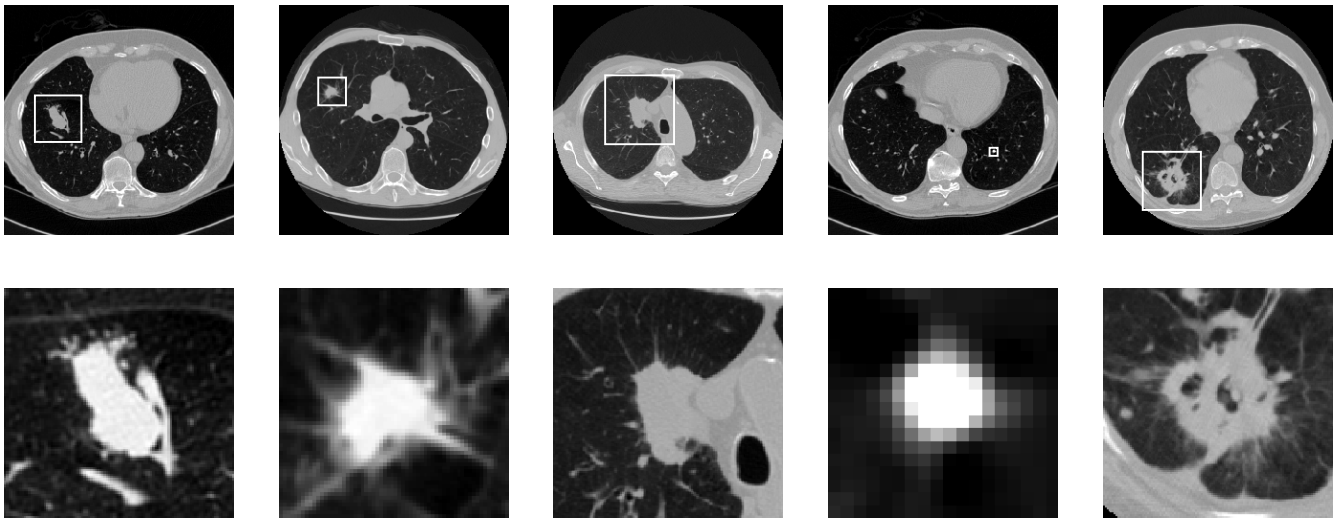


图1: DSB数据集中的结节示意. 上排:整张图. 下排: 放大.

noisy-or模型是在概率图模型中常用的局部因果概率模型[13]。它假设一个事件可能是由不同的因素引起的，任何一个这些因素的发生都可能导致这个事件以独立的概率发生。该模型的一个修改版本被称为leaky noisy-or模型[13]，它假定即使没有任何因素发生，事件也会以"泄漏概率"(leakage probability)发生。leaky noisy-or模型适用于本文的任务。 首先，当一个病例存在多个结节时，所有结节都有助于最终的预测。 其次，高度可疑的结节是增加该病例被判为癌症概率的因素，我们认为这是有理由的。 第三，当没有结节可以解释癌症病例时，癌症以泄露概率发生。

分类网络也是一个3D神经网络。为了防止过度拟合，我们让分类网络共享检测网络的骨干网（两个网络的骨干网的参数是绑定的），并交替地训练这两个网络,而且使用了数据增强。

我们在这项工作中的贡献总结如下:

1) 据我们所知，我们提出了第一个用于三维物体检测的立体的,one-stage的,端到端的CNN。
2) 我们建议将noisy-or gate集成到神经网络中，以解决CAD中的多实例学习任务。

我们在Data Science Bowl 2017[2]上验证了提出的方法，并在1972支队伍中获得第一名。

本文的其余部分安排如下。第二节介绍一些密切相关的工作。本文所提出的方法流程在后面的章节中详细描述。它由三个步骤组成：（1）预处理（第三节）：将肺从其他组织中分出来；（2）检测（第四节）：找出肺内所有可疑的结节；（3）分类（第五节）：对所有结节进行评分，并结合其癌症概率，得出患者的总体癌症概率。 第一步是通过经典的图像预处理技术完成的，另外两个步骤是通过神经网络完成的。 结果见第六节。 第七部分是一些讨论。

II. 有关工作

A. 一般对象检测

已有许多目标检测方法，并且全面的评论这些方法超出了本文的范围。 这些方法大部分都是为二维物体检测而设计的。一些最先进的方法有两个阶段（例如，Faster-RCNN [12]），其中在第一阶段（可能包含一个对象）提出了一些边界框（称为建议）,第二阶段给出分类(框中的对象属于哪一类）。更新的方法只有一个阶段，其中边界框和分类概率是同时给出的（YOLO [14]）或分类概率针对默认框给出的而框不是通过建议给出的（SSD [15]）。一般来说，单级方法更快，但是两级方法更准确。 在单类目标检测的情况下，两阶段方法的第二阶段不再需要，方法简化为单阶段方法。

前沿的2D对象检测方法扩展到3D对象检测任务（例如，视频和三维检测中的动作检测）是有限制的。由于主流GPU的显存约束，一些研究使用2D RPN来提取单个2D图像中的建议，然后使用额外的模块将2D建议合并到3D建议中[8,9]。类似的策略已经被用于三维图像分割[16]。就我们所知，3D RPN尚未用于处理视频或体数据。

[2]https://www.kaggle.com/c/data-science-bowl-2017

B. 结节检测

结节检测是典型的体检测任务。由于其重大的临床意义，近年来越来越受到重视。这个任务通常分为两个子任务[17]：提出建议和减少误报.每个子任务都吸引了很多研究。第一个子任务的模型通常以一个简单而快速的3D描述符开始，然后是一个分类器，以提供许多建议。第二个子任务的模型通常是复杂的分类器。 2010年Van Ginneken等人[17]对六种常规算法进行了综合评估，并在ANODE09数据集上对其进行评估，其中包含55次扫描。在2011 - 2015年期间，开发了一个更大的数据集LIDC [18,19,20]。研究人员开始采用CNN来减少误报的数量。 Setio等人[21]和Dou等人采用多视角CNN。 [22]采用3D CNN来解决这个问题，都取得了比传统方法更好的结果。Ding等人[9]采用2D RPN在每个切片上制作结节提案，并采用3D CNN来减少假阳性样本的数量。一个名为LUng Nodule Analysis 2016（LUNA16）[23]的竞赛是根据LIDC的一个选定的子集进行的。在这场比赛中，大多数参赛者使用了两阶段的方法[23]。

C. 多实例学习(MIL)

在多实例学习任务(MIL)中，输入是一组实例。 如果任何一个实例标记为“正”，则该组被标记为“正”，如果所有实例都被标记为“负”，则该组被标记为“负”。

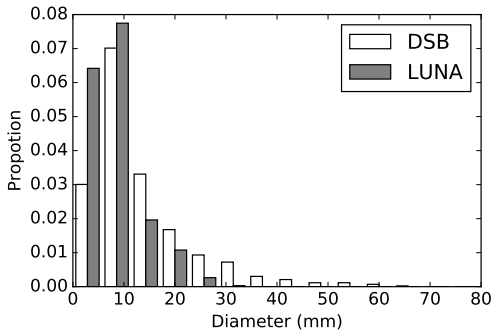
许多医学图像分析任务都是MIL任务，所以在深入学习兴起之前，一些早期的工作已经提出了CAD中的MIL框架。Dundar等人[24]引入凸包代表多实例特征，并将其应用于肺栓塞和结肠癌检测。徐等人[25]从组织检查图像中提取许多patch，并将其作为多个实例来解决结肠癌分类问题。

为了将MIL融合到深度神经网络框架中，关键部分是将来自不同实例的信息组合在一起的一个层，称为MIL Pooling Layer（MPL [26]）,例如：最大池化层[27]，平均池化层[26]，对数求和池化层[28]，广义平均层[25]和noisy-or层[29]。 如果每个样本的实例数量是固定的，那么使用特征拼接作为MPL也是可行的[30]。 MPL可用于在特征级[27,28]或输出级[29]中组合不同的实例。

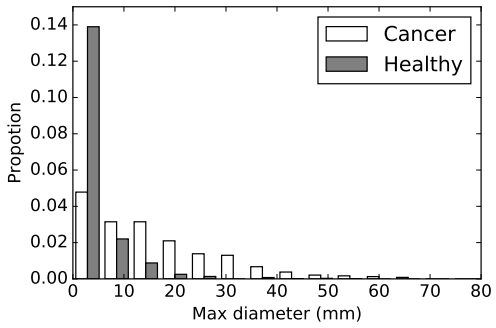
D. Noisy-or 模型

noisy-or贝叶斯模型被广泛用于推断疾病的可能性，如肝脏疾病[31]和哮喘病例[32]。Heckerman[33]基于noisy-or gate建立了一个多特征和多疾病诊断系统。Halpern和Sontag[34]提出了一种基于noisy-or模型的无监督学习方法，并在Quick Medical Reference模型上对其进行了验证。

上面提到的所有研究都将noisy-or模型纳入了贝叶斯模型。然而，noisy-or模型和神经网络的整合却很少见。 Sun等人[29]和Zhang等人在深度神经网络框架中采用它作为MPL来提高图像分类精度。[35]用它作为提高物体检测精度的增强方法。



(a)



(b)

图2: 结节直径的分布。(a) DSB数据集和LUNA数据集 中结节直径的分布。(b)在DSB数据集中癌症患者和健康人的最大结节直径的分布。

III. 数据集和预处理

A. 数据集

使用两个肺扫描数据集来训练模型，LUng Nodule Analysis 2016数据集（缩写为LUNA）和Data Science Bowl 2017（缩写为DSB）的训练集。LUNA数据集包括放射科医师注释的888名病人的1186个结节标签，而DSB数据集仅包括每名受试者的二进制标签，表明该受试者在扫描之后的一年中是否被诊断为肺癌。DSB数据集的训练集有包含1397个病人的数据,验证集是198人,测试集是506人。 我们在训练集中手动标记了754个结节，在验证集中标记了78个结节。

LUNA节点和DSB结节之间有一些显着差异。 LUNA数据集有许多非常小的注释结节，这可能与癌症无关。 根据医生的经验[36]，小于6mm的结节通常并不危险。 然而DSB数据集有很多非常大的结节（大于40mm）（图1中的第五个样本）。DSB数据集的平均结节直径为13.68 mm，LUNA数据集平均结节直径为8.31 mm（图2a）。此外，DSB数据集在主支气管上有许多结节（图1中的第三个样本），这在LUNA数据集中很少见到。 如果仅在LUNA数据集上训练网络，则将很难在DSB数据集中检测到结节。 缺少大结节会导致不正确的癌症预测，因为大结节的存在是癌症患者的特征（图2b）。 为了解决这些问题，我们从LUNA注释中去除小于6mm的结节，并手动标记DSB中的结节。

作者没有肺癌诊断的专业知识，因此结节选择和手工注释可能会产生相当大的噪音。下一阶段的模型（癌症分类）被设计为对错误检测具有鲁棒性，这减轻了对高度可靠的结节标签的需求。

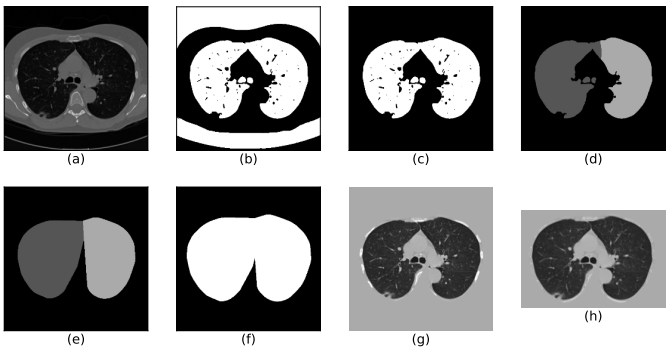


图3:预处理流程。注意粘在肺外壁上的结节。
(a)图像转换到HU,
(b)阈值二值化
(c)选择肺部对应的连通区域
(d)分割左右肺
(e)计算每个肺的凸包
(f) 将另个掩码扩大并组合
(g)将原图与掩码相乘,用组织亮度填充空白区域,并将图像转换为UINT 8
(h) 裁剪图像并对骨组织进行亮度限制

B. 预处理

3. 整个预处理过程如图3所示。所有的原始数据首先被转换成Houns field单位（HU），这是描述放射性强度的一个标准量化标度。每个组织都有其特定的HU范围，这个范围对于不同的人是一样的（图3a）。

1)掩码提取：CT图像不仅包含肺，还包含其他组织，其中一些可能具有球形，看起来像结节。排除这些干扰因素，最方便的方法是提取肺的掩码，忽略检测阶段的所有其他组织。对于每个切片(Slice)，用高斯滤波器（标准偏差=1个像素）对2D图像进行滤波，然后使用-600作为阈值进行二值化（图3b）。所有2D连通分量小于30平方毫米或偏心率大于0.99（某些高亮度径向成像噪声）都被去除。然后计算得到的二值3D矩阵中的所有3D连通分量，并且只保留那些未接触矩阵拐角并具有0.68L至7.5L的体积的那些分量。

这一步之后，通常只剩下一个对应肺部的二元组分，但有时候也有一些分散的区域。与那些分散的区域相比，肺部分总是处于图像的中心位置。对于一个区域的每个切片，我们计算的面积(Area)和到切片中心的距离(MinDist)。如果一个面积大于6000平方毫米的区域的平均MinDist大于62毫米,则移除该区域。 然后将其余的区域联合起来形成肺掩码（图3c）。

肺在某些情况下与外部空间联通，此时为保证预处理程序正常运行,需要将顶部的几张切片预先移除.

2) 凸包和扩张：肺外壁附着有一些结节。它们不包括在上一步获得的掩码中，这是我们不希望看到的。 为了将它们保留在掩码内部，一个方便的方法是计算掩码的凸包。然而，直接计算掩码的凸包将包括太多不相关的组织（如心脏和脊柱）。因此，在使用以下方法进行凸包计算之前，将肺掩码首先分成两部分（大致对应于左肺和右肺）.

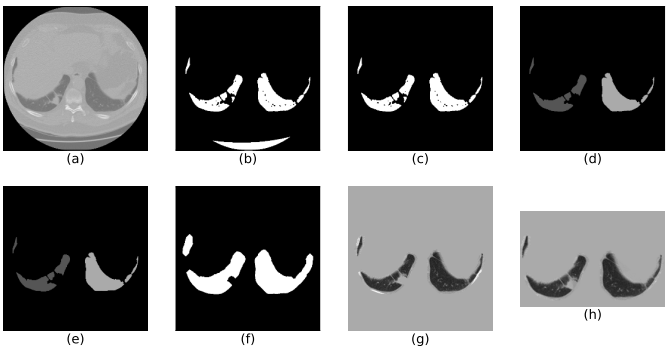


图4: 流程与图3一致, 但是显示的是比较低(靠近横膈膜的方向)的slice,注意步骤(e)没有计算凸包.

掩码被反复腐蚀，直到它被分成两个部分（它们的体积将是相似的），它们是左右肺的中心部分。然后将这两个部件扩大到原来的尺寸。它们与原始掩码的交集现在分别为两侧肺的掩码(图3d)。对于每个掩码，大多数二维切片被替换为凸包，以包括上述结节(图3e)。得到的掩码进一步扩大了10个体素，包括一些周围的空间。通过结合两个肺的掩码获得完整的掩码（图3f）。

然而，肺下部的一些二维切片具有新月形（图4）。他们的凸包可能含有太多不需要的组织。因此，如果2D掩码的凸包面积大于掩码本身的1.5倍，则保留原始掩码（图4e）。

3）强度归一化： 为了准备深度网络的数据，我们将像素值从HU变换到UINT8。 原始数据首先被限制在[-1200,600]内，然后线性变换到[0,255]。 然后乘以上面获得的完整掩码，掩码外面的所有东西都被填充170，这是普通组织的亮度。另外，对于前一步膨胀产生的空间，所有大于210的值也被替换为170。由于周围区域含有一些骨骼（高亮度组织），因此很容易被错误分类为钙化结节。我们用170填充骨头，使它们看起来像正常的组织（图3g）。图像在所有3个维度上被裁剪，使得每边的边缘是10个像素（图3h）。

IV. 用于结节检测的3D CNN

本文设计了一个3D CNN检测可疑结节.它是使用改进的U-net [37]作为骨干模型的RPN的3D版本。由于在这个任务中只有两个类别（结核和非结核），所以预测的提议直接被用作检测结果而没有附加的分类器。 这与一级检测系统YOLO[14]和SSD[15]相似。这个结节检测模型简称为N-Net,其中N代表结节。

A. 基于Patch 输入的训练

对象检测模型通常采用基于图像的训练。在训练阶段，整个图像被用作网络的输入。但是，由于显存限制，这对于我们的3D ConvNet模型来说是不可行的。当肺部扫描的分辨率保持在一个合适的水平时，即使是单个样本，也会轻易耗尽主流GPU的最大显存。

为了克服这个问题，我们从肺部扫描中提取晓得3D的patch作为输入。patch是一个128×128×128×1(高x长x宽x通道数,下文中使用相同的表述方式)的立方体。随机选择两种patch:首先，70%的输入被选择，以便它们至少包含一个结核目标。其次，30%的输入从肺部扫描中随机剪裁，可能不含结节,以确保覆盖足够的负样本。

如果一个patch超出了肺部扫描的范围，那么填充为170，与预处理相同。结节的目标不必位于patch的中心，但是距离边界要大于12个像素（除了一些太大的结节外）。

进行数据扩充以缓解过拟合问题。这些patch随机左右翻转，并以0.8-1.15之间的比例调整大小。还尝试了其他数据扩充方法，例如坐标轴交换和旋转。但它们没有显着的改进.

B. 网络结构

检测器网络由一个U-Net [37]主干网和一个RPN输出层组成，其结构如图5所示。U-Net主干网络能够捕获多尺度信息，这是非常重要的，因为结节的尺寸比较多变。RPN的输出格式允许网络直接生成提议(proposals)。

网络主干具有前馈路径和反馈路径（图5a）。前馈路径上,首先是具有24个通道的两个3×3×3卷积层(蓝色的圆圈"K")。然后是四个3D残差块(红色的圆圈"R")[38]与四个3D最大池化层（池化大小为2×2×2，步长为2）。每个3D残差块（图5b）由三个残差单元组成[38]。图5b示出了残差单元的结构。前馈路径中的所有卷积核大小为3×3×3，延拓值为1。

反馈路径由两个反卷积层和两个组合单元组成。 每个反卷积层的步长为2，核的大小为2.每个组合单元有两个输入，一个前馈路径上的blob和一个反卷积模块输出的blob，并将输出发给一个残差块（图5c）。在左边的合并单元中，我们引入了位置信息作为额外的输入（详见第IV-C部分）。 该组合单元的特征图(feature-map)大小为32×32×32×131。接着是64通道和15通道的两个1×1×1卷积层(第二行蓝色的圆圈"K")，输出的尺寸为32×32×32×15(淡黄色方块)。

4D输出张量调整为32×32×32×3×5。最后的两个维度分别对应锚点和回归系数。 受到RPN的启发，网络在每个地点都有三个不同尺度的锚点，分别对应三个长度分别为10,30和60毫米的边界框。所以总共有32×32×32×3个anchor boxes。五个回归值是$(\hat{o}, \hat{d}_x, \hat{d}_y, \hat{d}_z, \hat{d}_r)$,其中第一个量使用了sigmoid函数:

$$\hat{p} = \frac{1}{1 + \exp(-\hat{o})},$$

其他的四个量没有使用激活函数.

C. 位置信息

建议(proposal)的位置也可能影响结节是否为恶性的判断，因此我们也在网络中引入位置信息。对于每个图像块，我们计算出相应的位置裁剪，它与输出特征图（32×32×32×3）一样大。 位置裁剪有3个特征图，对应X，Y，Z轴。在每个轴中，每个轴的最大值和最小值分别归一化为1和-1，对应分割出来的肺的两端。

D.损失函数

用 (G_x, G_y, G_z, G_r) 表示目标结核的真实边界框(这是从训练数据中来的,因为我们同步训练检测和分类两个网络,所以我们既需要标记了目标位置的数据(LUNA),也许要标记了是否患病的数据(DSB))，用 (A_x, A_y, A_z, A_r),表示锚点(anchor)的边界框(这个框就叫做anchor box,训练的正向传播产生anchor box和概率)，其中前三个元素表示中心点的坐标的最后一个元素表示边长。IoU用于确定每个anchor box的标签。目标结节的IoU大于0.5的认为是正,小于0.02的认为是负.其他情况在训练过程中被忽略。anchor box的预测概率和标签值分别用 $\hat{p}$ 和 p 表示. 注意$p \in \{0, 1\}$ (0 为负,1为正,--DSB数据的标签值是二值的)

[IoU:简单来讲就是模型产生的目标窗口和原来标记窗口的交叠率。具体我们可以简单的理解为：检测结果(Detection Result)与真实值(Ground Truth) 的交集比上它们的并集]

分类误差(Loss Function)定义为:

$$L_{cls} = p\log(\hat{p}) + (1-p)\log(1-\hat{p}). \quad (1)$$

G是真实边框,A是预测值:

$$\begin{aligned} d_x &= (G_x - A_x)/A_r, \\ d_y &= (G_y - A_y)/A_r, \\ d_z &= (G_z - A_z)/A_r, \\ d_r &= \log(G_r/A_r). \end{aligned}$$

定义位置检测误差为:

$$L_{reg} = \sum_{k\in\{x,y,z,r\}} S(d_k, \hat{d_k}) \quad (2)$$

(2)中的S函数是L1-norm function,定义如下:

$$S(d,\hat{d}) = \begin{cases} |d-\hat{d}|, & \text{if } |d_k - \hat{d}| > 1, \\ (d-\hat{d})^2, & \text{else.} \end{cases}$$

每个 anchor box的误差函数定义为:

$$L = L_{cls} + pL_{reg}. \quad (3)$$

这个方程表明，回归损失只适用于正样本，因为只有在这些情况下$p = 1$. 整体损失函数是一些"选定"的anchor box的损失函数的平均值。我们使用正样本平衡和难分样本挖掘来选出这些anchor box（参见下一小节）。

E. 正样本平衡

对于一个大结节，会产生很多判定为正的anchor box。为了减少训练样本之间的相关性，在训练阶段会在这些anchor box中随机选择一个。

尽管我们已经从LUNA中去除了一些非常小的结节，但是结节大小的分布仍然是高度不平衡的。小结节的数量远大于大结节的数量。如果使用均匀采样，网络将学习偏向于小结节。同时牺牲了大结节的准确性。这是我们不愿看到的现象,因为大结节通常比较小的结节更可能是癌症的指标。为了防止这个结果，我们增加了训练集中大结节的采样频率。特别的,大于30mm和40mm的结节的采样频率分别是其他结节的2倍和6倍.

F. 难分样本挖掘

负样本比正样本要多得多，而且他们的分布非常不平衡。虽然他们中的大多数很容易被网络正确分类，但其中也有一些负样本有带有和可疑结节相似的外观,使得网络难以正确决断。作为目标检测的常用手段，通常采用难分样本挖掘来处理这个问题。我们在训练过程中使用一个简单的在线版本的难分样本挖掘。

第一步，我们使用网络处理patches，并获得输出"图"(我们原来把这个叫做输出向量/矩阵,因为到了输出这块,已经看出是什么图形了)，其中每个"像素"(其实就是维度)是分类置信度分数和回归项。第二步，随机选取N个负样本进入候选池。 第三步，把这个池中的负样本根据分类信度得分降序排列，选前n个样本为难分样本。

[难分样本挖掘:当得到检测错误的patch时，会使用这个patch创建一个负样本，并把这个负样本以某种策略添加到训练集中去。不断的训练，分类器会表现的更好，并且会逐渐变得能分清原来难分的负样本.]

G. 测试过程中的图像分裂

网络训练后，整个肺部扫描可以作为输入，以获得所有可疑的结节。 因为网络是完全卷积的，所以做到这一点很简单。但是,即使在测试中需要的的显存比训练时少得多，但需求量仍然超过了GPU的最大显存。为了解决这个问题，我们将肺部扫描分成几部分（每部分208×208×208×1），分别进行处理，然后合并结果。我们保持这些分割大幅度重叠（32像素），以消除卷积计算过程中不必要的边界效应。

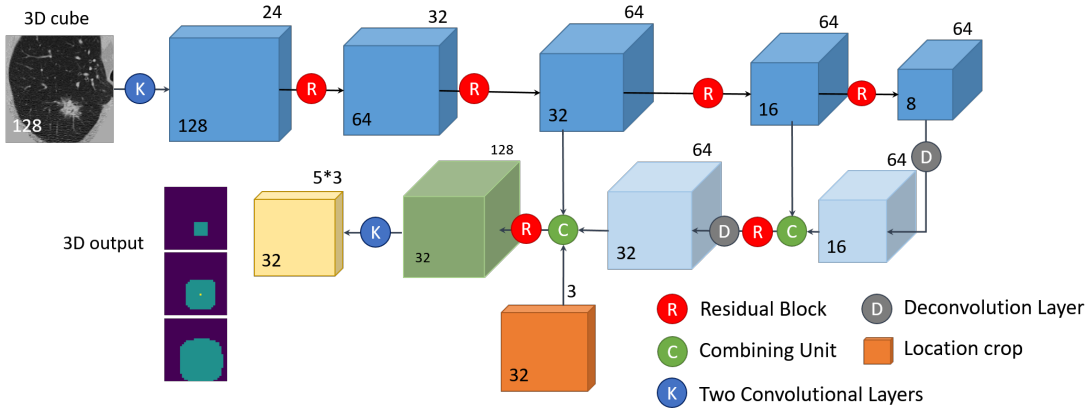
这一步将输出许多结节建议(proposals) $\{x_i,\ y_i,\ z_i,\ r_i,\ p_i\}$, 其中$x_i,\ y_i,\ z_i$为中心坐标， r_i为半径,p_i为置信度. 然后进行非最大抑制(NMS)[39]操作以排除重叠的建议区域。另一个模型会根据这些数据来预测罹患癌症的概率.

V. 癌症分类

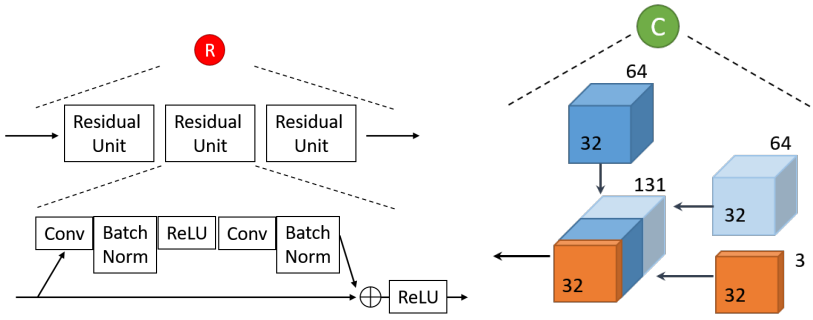
然后我们根据检测到的结节来评估受试者的癌症概率。对于每个受试者，根据他们在N-Net中的置信度得出五个建议。作为一个简单的数据增强方法，在训练期间，建议的选取是随机的,置信度高的建议区域更容易被选中。但在测试时，直接挑选置信度前五的建议区域。如果检测到的建议区域数量少于五个，则填充几个空白图像以使该数目仍然是五。

由于训练样本的数量有限，建立一个独立的神经网络来做这一步是不明智的，那样会发生过拟合。另一种方法是复用在检测阶段训练的N-Net。

对于选定的每个建议区域，我们裁剪一个以结核为中心的96×96×96×1的patch（注意这个patch小于检测阶段的patch），把它送到N-Net，取最后一个卷积层的输出(24×24×24×128)，每个建议区域的中心处2×2×2的体素被提取并进行最大池化，产生一个128维特征向量（图6a）.为了从单个病例的多个结节获得单个分数，我们探索了四种整合方法（参见图6b）。



(a)



(b) (c)

图5: 结节检测网络
(a) 整体网络结构。每个立方体都是一个4D tensor. 图中只显示了两个维度:
立方体内的数字代表空间大小 ($Height = Width = Length$). 立方体外的数字代表通道的数。
(b) 残差块的结构.
(c) （a）中的左侧组合单元的结构，. 右侧的组合单元是类似的,但是没有location crop.

A. 特征组合

首先，将前五个结节的所有特征送到全连接层,输出五个64-D特征。然后将这些特征向量组合起来，通过最大池化输出单个64-D特征。然后特征向量被送到第个二全连接层，其激活函数是Sigmoid 函数，然后输出病例的癌症概率（图6b左）。

如果结节之间存在一些非线性相互作用，这种方法可能是有用的。缺点是在整合步骤中缺乏可解释性，因为每个结节与癌症概率之间没有直接关系。

B. MaxP方法

前五个结节的特征分别被馈送到具有64个隐藏单元和一个输出单元的相同的双层感知器中。 最后一层的激活函数也是sigmoid函数，它输出每个结节的癌症概率。 然后把这些概率的最大值作为该病例的概率。

与特征组合方法相比，该方法为每个结节提供了可解释性。 然而这种方法忽略了结节之间的相互作用。 例如，如果一个病人有两个结节，两个结节都有50%的癌症概率，那么医生就会推断病人患癌症概率远远大于50%，但是该模型仍然会给出50%的预测。

C. Noisy-or 方法

为了克服上面提到的问题，我们假设结节是癌症的独立原因，任何一个结节的恶性都会导致癌症。 像最大概率模型一样，每个结节的特征首先被送到双层感知器以获得概率。 最终的患癌概率是[13]:

$$P = 1 - \prod_i (1 - P_i), \tag{4}$$

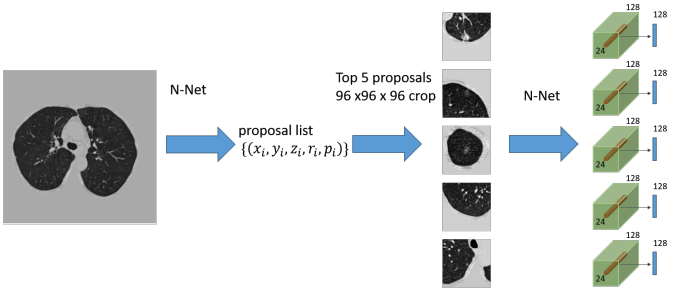
其中P_i 代表第i个节点的癌症概率.

D. Leaky Noisy-or 方法

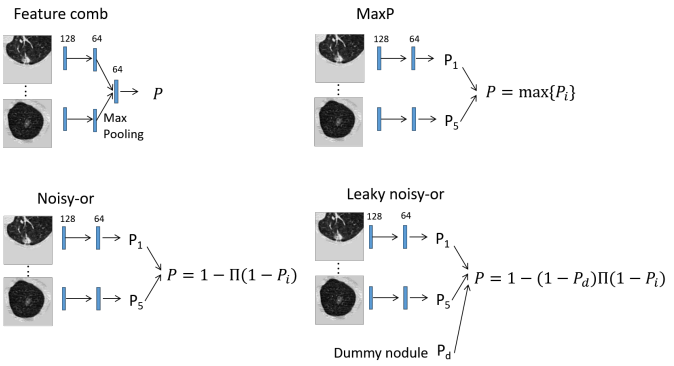
Noisy-or方法和MaxP方法都存在一些问题。 如果一个受试者患有癌症，但检测网络遗漏了一些恶性结节，这些方法会将癌症的病因归因于检测到的良性结节，网络将会认为数据集中其他与之类似的良性结节的患癌概率应该提高一点。显然，这是没有道理的。我们引入一个假设的虚拟结节，定义它的癌症概率为P_d[13].最终的癌症概率为:

$$P = 1 - (1 - P_d) \prod_i (1 - P_i). \tag{5}$$

Pd在训练中自行学习,而不需要进行手动调整,
这个模型就是我们最终使用的方法,我们称之为C-Net(C代表case)



(a)



(b)

图6: 病例分类器
(a) 获得建议(proposal)的流程和建议的特征
(b) 四种多结节信息集成方法。

E. 训练过程

分类时使用标准的交叉熵损失函数。由于内存约束，预先生成每个病例的结节边界框。 然后在这些预先生成的边界框上训练包括共享特征提取层（N-Net部分）和集成层的分类器。由于N-net较深，三维卷积核的参数多于二维卷积核，但分类样本数量有限，模型往往过拟合训练数据。

为了解决这个问题，采取了两种方法：数据增强和交替训练。3D数据增强比2D数据增强更强大。例如，如果我们只考虑翻转和轴交换，则2D情况下有8个变体，3D情况下有48个变体。具体而言，使用以下数据增强方法：

（1）随机翻转3个方向;
（2）在0.75-1.25之间随机调整大小;
（3）以3D中的任意角度旋转;
（4） 以小于半径的15%的距离在3个方向上随机移动。

另一个常用的缓解过拟合的方法是使用一些适当的正则化。 在这个任务中，由于卷积层被检测器和分类器共享，这两个任务自然可以相互正则化。所以我们在检测器和分类器上交替地训练模型。具体来说，在每个training block中，有一个检测器训练时期和一个分类器训练时期。

训练过程相当不稳定，因为每个GPU的batch大小只有2个，训练集中有很多异常值。 因此，在训练的后期阶段使用梯度裁剪，即如果梯度矢量的l_2范数大于1，则将归一化为1。

批量标准化(BN,Batch normalization)[40]在网络中使用。但是在交替训练期间直接应用它是有问题的。在训练阶段，BN的统计数据（平均激活值和方差）在batch内计算，在测试阶段使用存储的统计（运行平均统计）。交替训练方案将使运行平均值不适合分类器和检测器。首先，它们的输入采样是不同的：分类器的patch大小为96，探测器的patch大小为128。其次，分类器的建议始终是patch的中心，但检测器是随机裁剪图像的。因此，这两项任务的平均统计数据会有所不同，运行平均统计数据可能处于中间位置，并且两者的验证阶段的性能会变差。为了解决这个问题，我们首先训练分类器，使BN参数适合分类,切换训练阶段时，这些参数被冻结，即在训练阶段和验证阶段，我们使用存储的BN参数。

总之，训练过程分为三个阶段：（1）从训练过的检测器中传递权值，在标准模式下对分类器进行训练;（2）用梯度限幅训练分类器，然后冻结BN参数;（3）使用梯度裁剪和存储的BN参数交替训练网络进行分类和检测。 这个训练方案对应表1中的A→B→E。

VI. 结论

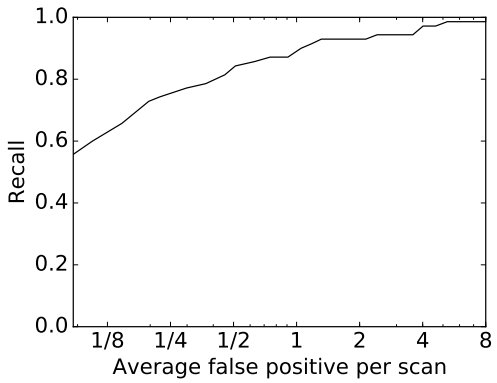
A. 结节检测

由于我们的检测模块被设计为在训练期间忽略非常小的结节，LUNA16评估系统不适合评估其性能。我们评估了网络在DSB验证集上的性能。它包含198例病例的数据，总共有71个（小于6毫米的7个结节被排除）结节。自由响应操作特性（FROC）曲线如图7a所示。召回率 1/8, 1/4, 1/2, 1, 2, 4, 8这几个点的平均 false positive per scan 是0.8562.

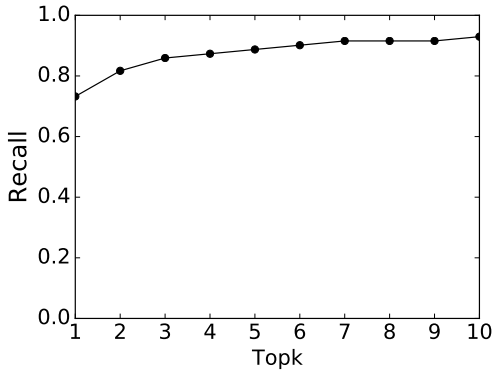
我们还研究了选择不同的top-k个结节时的召回率（图7b）。 结果显示k = 5足以捕获大部分结节。

B. 病例分类

为了选择训练方案，我们重新排列了训练集和验证集，因为我们发现原始训练集和验证集有明显的差异。原始训练集的四分之一被用作新的验证集,其余的与原始的验证集相组合成新的训练集。



(a)



(b)

图7: 检测模块的结果.
(a) FROC曲线.
(b) 不同的top-K值对应的召回率

表I:测试集上不同训练方法的交叉熵损失。
第三部分是前四名参赛队伍的成绩.

Training method	Loss
C-Net	1.2633
(A) C-Net + Aug	0.4173
(B) C-Net +Aug + Clip	0.4157
(C) C-Net +Aug + Alt	0.4060
(D) C-Net +Aug + Alt + Clip	0.4185
(E) C-Net +Aug + Alt + Clip + BN freeze	0.412
A → B	0.4060
A → B → D	0.4024
A → B → E	**0.3989**
grt123	0.3998
Julian de Wit & Daniel Hammack	0.4012
Aidence	0.4013
qfpxfd	0.4018

Aug: 数据增强;
Clip: 梯度裁剪;
Alt: 交替训练;
BN freeze: 冻结batch标准化参数。
grt123 是我们的队名. 比赛中使用的训练方案略有不同。

表II: 四种组合方法对应的测试集上的交叉熵损失值

Name	Loss
Feature comb	0.4286
MaxP	0.4090
Noisy-or	0.4185
Leaky noisy-or	**0.4060**

如第V-E节所述，在训练过程中使用了四种技术：（1）数据增强，（2）梯度限幅，（3）交替训练，（4）冻结BN参数。我们在新的验证集上探索了这些技术的不同组合（在表I中由A，B...E表示）和不同的阶段顺序。发现A→B→E方案表现最好。 比赛结束后，我们仍然可以将结果提交给评估服务器，所以我们对测试集上的训练方案进行了评估。表I显示了测试集的结果（模型在训练集和验证集并集上进行了训练）。发现A→B→E确实是许多方案中最好的一个。

从表1的第一部分我们可以得出几个结论。 首先，没有数据增强，模型会严重过拟合。其次，交替训练显着提高了效果。第三，在这些方案中梯度限幅和BN冻结并不是很有用。

从表1的第二部分可以发现，梯度限幅(Clip)是有用的(A(A→B))。而交替训练对进一步微调模型是有用的（A→B→D）。此外，引入BN冷冻技术进一步改善了结果（A→B→E）。

表一中的第三部分显示了前四名的队伍在比赛中的表现。分数非常接近,但是我们用单一的模型取得了最高分。

表2给出了不同的多结节信息集成模型的结果。所有模型都是使用交替训练方法（表1中的配置C）进行训练的。三种基于概率的方法比特征组合方法好得多。而Leaky Noisy-or模型表现最好。

预测的癌症概率在训练和测试集上的分布如图8a，b所示。通过改变阈值(图8c，d)在每组上获得受试者工作特征（ROC）曲线。训练和测试集的ROC曲线下面积(AUC)分别为0.90和0.87。如果将阈值设置为0.5(如果预测概率高于阈值，则将其分类为癌症)，训练和测试集合的分类准确率分别为85.96%和81.42%。如果将阈值设置为1(所有情况都预测健康），则训练和测试集的分类准确率分别为73.73%和69.76%。

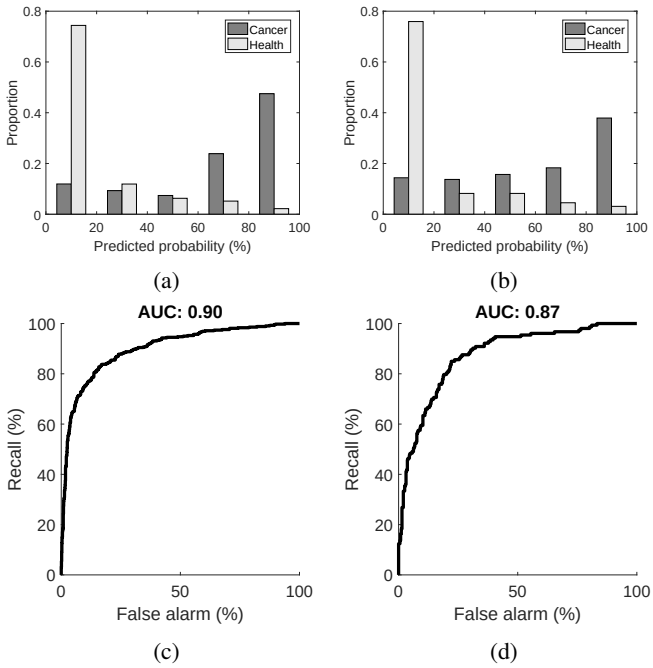


图8: 癌症分类结果.(a)和(b)分别是训练集和测试集上的网络预测出的健康人和患者的癌症概率的分布。(c)和(D)分别是训练集和测试集上的癌症分类任务的ROC曲线 .

几个病例的分类结果如图9所示。对于两个真正的阳性病例（病例1和病例2），该模型正确预测了他们两人的高癌症概率。病例1有一个非常大的肿瘤（结节1-2），这是一个非常高的癌症概率。病例2有几个中等大小的结节，其中三个有明显的癌症发生概率，因此整体概率非常高。此外，该模型不仅基于大小而且还根据形态学来判断恶性肿瘤。结节1-1具有比结节2-1和2-2大的尺寸，但具有较低的癌症可能性。原因如下。结节1-1发亮，圆形，边界清晰，表现为良性。虽然结节2-1的形状不规则，边界不清，结节2-2有不透明的亮度，这些都是恶性的指征。结节2-1称为针状结节，结节2-2称为部分实性毛玻璃状结节[41]，二者都是高度危险的结节。两个假阴性病例（病例3和病例4）没有显着的结节，所以他们的整体概率很低。两个假阳性病例（病例5,6）都有高度可疑的结节，难以正确分类。病例7没有发现结节，病例8只发现2个微小结节，因此预测两个病例都是健康的且预测正确。

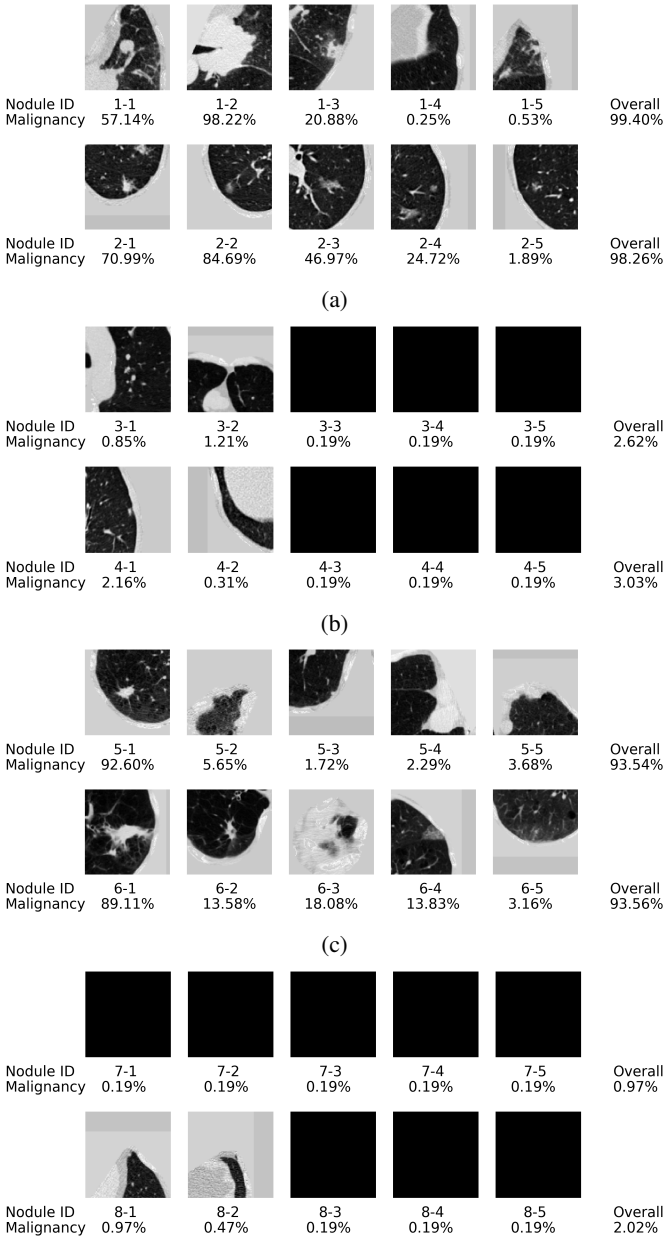


图9: 几个病例在模型中的输出(a)真阳性样本(b)假阴性样本。(c) 假阳性样本。(d)真阴性样本。当检测到的结节数量少于5个时，使用几张空白图像作为输入。

VII. 讨论

本文提出了一种基于神经网络的方法来进行自动肺癌诊断。设计了3D CNN用来检测结节，并且使用leaky noisy-or模型来评估每个检测到的结节的癌症概率并将它们结合在一起。 整个系统在比赛中对癌症分类任务取得了非常好的效果。

所提出的leaky noisy-or网络可能在医学图像分析中找到许多应用。许多疾病诊断从图像扫描开始。图像中显示的病变可能与疾病有关，但是这种关系是不确定的，与在本文研究的癌症预测问题中一样。可以使用leaky noisy-or模型来整合来自不同病变的信息以预测结果。这也减轻了对高精度精细标签的需求。

将3D CNN应用于三维物体检测和分类面临两大难题。首先，模型尺寸越大，模型占用的内存越多，所以运行速度，batch大小和模型深度都是有限的。 我们设计了一个较浅的网络，并使用patch代替整个图像作为输入。其次，3D CNN的参数数量明显大于具有相似架构的2D CNN的数量，因此该模型倾向于过拟合,我们使用数据增强和交替训练来缓解这个问题。

有一些潜在的方法来提高模型的性能。最直接的方法是增加训练样本的数量：1700个病例太少，不能覆盖所有的结节情况，而有一个经验的医生在他的职业生涯中能看到更多的病例。其次，合并结节的分割标签可能是有用的，因为已经表明分割和检测任务的协同训练可以改善两个任务的性能[42]。

虽然很多团队在这个癌症预测竞赛中取得了不错的成绩，但是这个任务本身对于临床却有明显的局限性：不考虑结节的增长速度。事实上，快速生长的结节通常是危险的。为了检测生长速度，需要在一段时间内对患者进行多次扫描，并检测所有结节（不仅是大结节，而且还有小结节），并比较他们随着时间的变化。虽然在这项工作中提出的方法不追求小的结节的高检测精度，但是可以为此对其进行修改。例如可以添加另一个反池化层来合并更精细的信息并减少锚点(anchor)大小。

参考文献

[1] M. Infante, S. Cavuto, F. R. Lutman, G. Brambilla, G. Chiesa, G. Ceresoli, E. Passera *et al.*, "A randomized study of lung cancer screening with spiral computed tomography: three-year results from the dante trial," *American Journal of Respiratory and Critical Care Medicine*, vol. 180, no. 5, pp. 445–453, 2009.

[2] S. Singh, D. S. Gierada, P. Pinsky, C. Sanders, N. Fineberg, Y. Sun, D. Lynch *et al.*, "Reader variability in identifying pulmonary nodules on chest radiographs from the national lung screening trial," *Journal of Thoracic Imaging*, vol. 27, no. 4, p. 249, 2012.

[3] H.-C. Shin, M. R. Orton, D. J. Collins, S. J. Doran, and M. O. Leach, "Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4d patient data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1930–1943, 2013.

[4] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[5] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *The Journal of the American Medical Association*, vol. 316, no. 22, pp. 2402–2410, 2016.

[6] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak *et al.*, "A survey on deep learning in medical image analysis," *arXiv preprint arXiv:1702.05747*, 2017.

[7] J. S. Duncan and N. Ayache, "Medical image analysis: Progress over two decades and the challenges ahead," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 85–106, 2000.

[8] X. Peng and C. Schmid, "Multi-region two-stream R-CNN for action detection," in *European Conference on Computer Vision*. Springer, 2016, pp. 744–759.

[9] J. Ding, A. Li, Z. Hu, and L. Wang, "Accurate Pulmonary Nodule Detection in computed tomography images using deep convolutional neural networks," in *Medical Image Computing and Computer-Assisted Intervention 2017*, ser. Lecture Notes in Computer Science. Springer, Cham, Sep. 2017, pp. 559–567.

[10] S. G. Armato, R. Y. Roberts, M. Kocherginsky, D. R. Aberle, E. A. Kazerooni, H. MacMahon, E. J. van Beek *et al.*, "Assessment of radiologist performance in the detection of lung nodules: dependence on the definition of truth," *Academic Radiology*, vol. 16, no. 1, pp. 28–38, 2009.

[11] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[13] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 2014.

[14] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: single shot multibox detector," in *European Conference on Computer Vision*. Springer, 2016, pp. 21–37.

[16] J. Chen, L. Yang, Y. Zhang, M. Alber, and D. Z. Chen, "Combining fully convolutional and recurrent neural networks for 3D biomedical image segmentation," in *Advances in Neural Information Processing Systems*, 2016, pp. 3036–3044.

[17] B. Van Ginneken, S. G. Armato, B. de Hoop, S. van Amelsvoort-van de Vorst, T. Duindam, M. Niemeijer, K. Murphy *et al.*, "Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the anode09 study," *Medical Image Analysis*, vol. 14, no. 6, pp. 707–722, 2010.

[18] Samuel G., Armato III, M. Geoffrey, B. Luc, M.-G. Michael F., M. Charles R., R. Anthony P., Z. Binsheng *et al.*, "Data From LIDC-IDRI," 2015.

[19] S. G. Armato, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao *et al.*, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed refer-

ence database of lung nodules on CT scans," *Medical Physics*, vol. 38, no. 2, pp. 915–931, 2011.

[20] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of Digital Imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[21] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. van Riel, M. M. W. Wille *et al.*, "Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1160–1169, 2016.

[22] Q. Dou, H. Chen, L. Yu, J. Qin, and P.-A. Heng, "Multilevel contextual 3-d cnns for false positive reduction in pulmonary nodule detection," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1558–1567, 2017.

[23] A. A. A. Setio, A. Traverso, T. de Bel, M. S. Berens, C. v. d. Bogaard, P. Cerello, H. Chen *et al.*, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge," *arXiv preprint arXiv:1612.08012*, 2016.

[24] M. M. Dundar, G. Fung, B. Krishnapuram, and R. B. Rao, "Multiple-instance learning algorithms for computer-aided detection," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 3, pp. 1015–1021, 2008.

[25] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, I. Eric, and C. Chang, "Deep learning of feature representation with multiple instance learning for medical image analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2014, pp. 1626–1630.

[26] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *arXiv preprint arXiv:1610.02501*, 2016.

[27] J. Wu, Y. Yu, C. Huang, and K. Yu, "Deep multiple instance learning for image classification and auto-annotation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3460–3469.

[28] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1713–1721.

[29] M. Sun, T. X. Han, M.-C. Liu, and A. Khodayari-Rostamabad, "Multiple instance learning convolutional neural networks for object recognition," in *International Conference on Pattern Recognition*. IEEE, 2016, pp. 3270–3275.

[30] T. Zeng and S. Ji, "Deep convolutional neural networks for multi-instance multi-task learning," in *Proceedings of the 2015 IEEE International Conference on Data Mining*. IEEE Computer Society, 2015, pp. 579–588.

[31] A. Oniśko, M. J. Druzdzel, and H. Wasyluk, "Learning bayesian network parameters from small data sets: Application of noisy-or gates," *International Journal of Approximate Reasoning*, vol. 27, no. 2, pp. 165–182, 2001.

[32] V. Anand and S. M. Downs, "Probabilistic asthma case finding: a noisy or reformulation," in *AMIA Annual Symposium Proceedings*, vol. 2008. American Medical Informatics Association, 2008, p. 6.

[33] D. Heckerman, "A tractable inference algorithm for diagnosing multiple diseases," in *Proceedings of the Fifth Annual Conference on Uncertainty in Artificial Intelligence*. North-Holland Publishing Co., 1990, pp. 163–172.

[34] Y. Halpern and D. Sontag, "Unsupervised learning of noisy-or bayesian networks," *arXiv preprint arXiv:1309.6834*, 2013.

[35] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in *Advances in Neural Information Processing Systems*, 2006, pp. 1417–1424.

[36] H. MacMahon, D. P. Naidich, J. M. Goo, K. S. Lee, A. N. Leung, J. R. Mayo, A. C. Mehta *et al.*, "Guidelines for management of incidental pulmonary nodules detected on ct images: from the fleischner society 2017," *Radiology*, p. 161659, 2017.

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[39] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[40] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[41] D. M. Ha and P. J. Mazzone, "Pulmonary nodules," *Age*, vol. 30, pp. 0–05.

[42] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *International Conference on Computer Vision*, 2017.