NIH Chest X-ray Dataset of 14 Common Thorax Disease Categories

美国国立卫生研究院胸部x射线数据集的14种常见的胸部疾病类别:

(1, Atelectasis; 2, Cardiomegaly; 3, Effusion; 4, Infiltration; 5, Mass; 6, Nodule; 7, Pneumonia; 8, Pneumothorax; 9, Consolidation; 10, Edema; 11, Emphysema; 12, Fibrosis; 13, Pleural\_Thickening; 14 Hernia)

**背景与意义**:胸部x光检查是最常见和最具成本效益的医学影像学检查之一。然而，胸部x线的临床诊断是具有挑战性的，有时被认为比通过胸部CT成像诊断更困难。即使最近一些有前景的工作已经开始研究，尤其是关于结核病分类的深度学习工作中。在真实世界的医学站点中，如果可以的话，只有几千张图像被用于研究，所以在所有的胸部x射线数据设置上实现临床相关的计算机辅助检测和诊断(CAD)仍然是非常困难的。这从[2]中可以明显看出，由于只有4143张正面图像[3](Openi是迄今为止最大的公开胸透数据集)的可用性，用于胸部疾病识别的深度神经网络的性能受到严重限制。

在这个数据库中，我们提供了在最近的工作[1]中使用的数据集的一个增强版本(还有6个疾病类别和更多的图像)，它大约是[3]中x线胸片图像数量的27倍。我们的数据集是从美国国立卫生研究院临床中心的临床PACS数据库中提取的，包括医院数据占60%。因此，我们期望该数据集比以前的任何胸部x线数据集更能代表真实的患者群体分布和现实的临床诊断挑战。当然，我们的数据集的大小，就图像的总数和胸腔疾病的频率而言，将更好地促进深度神经网络训练[2]。有关如何通过自然语言处理(NLP)提取数据集和挖掘图像标签的详细信息，请参阅[1]。

**详细信息**:X光胸片数据集包含了112,120张来自30,805位独特患者的正面x射线图像，带有文本挖掘的14个疾病图像标签(每个图像可以有多个标签)，使用自然语言处理从相关的放射学报告中挖掘。肺不张、实变、浸润、气胸、水肿、肺气肿、纤维化、积液、肺炎、胸膜增厚、心脏增大、结节、肿块和疝是我们CVPR 2017论文中列出的8种常见疾病类型的延伸。请注意，最初的放射学报告(与这些胸透研究相关)由于许多原因不能公开共享。文本挖掘的疾病标签的准确率预计为>90%。请参阅我们的arxiv论文:1705.02315，了解更多关于基于14种疾病标签的训练模型的细节和基准性能。

**内容:**

1、112,120正面胸片1024\*1024分辨率的PNG图像(图像文件夹下)

2、所有图像的元数据(Data\_Entry\_2017.csv):图像索引、查找标签、随访编号、患者ID、患者年龄、患者性别、查看位置、原始图像大小和原始图像像素间距。

3、1000张图片的包围框(BBox\_List\_2017.csv):图片索引，查找标签，Bbox[x, y, w, h]。[x,y]表示每个框左上角的坐标。[w h]表示每个框的宽度和高度。

4、提供了两个数据分割文件(train\_val\_list.txt和test\_list.txt)。在病人层面上，胸部x光数据集中的图像分为这两组。来自同一患者的所有研究将只出现在训练/验证或测试集。

如果你发现数据集对你的研究项目有用，请引用我们的CVPR 2017论文:

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald M. Summers. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly- Supervised Classification and Localization of Common Thorax Diseases, IEEE CVPR, pp. 3462-3471,2017

**局限性**:

1)图像标签是提取的NLP，因此会有一些错误的标签，但NLP标签的精度约为>90%。

2)疾病区域包围框数量非常有限。

3)胸部x线片报告不对外公开。鼓励使用这个公共数据集的团体共享他们“更新的”图像标签和/或在他们研究中手工注释的新的边界框。

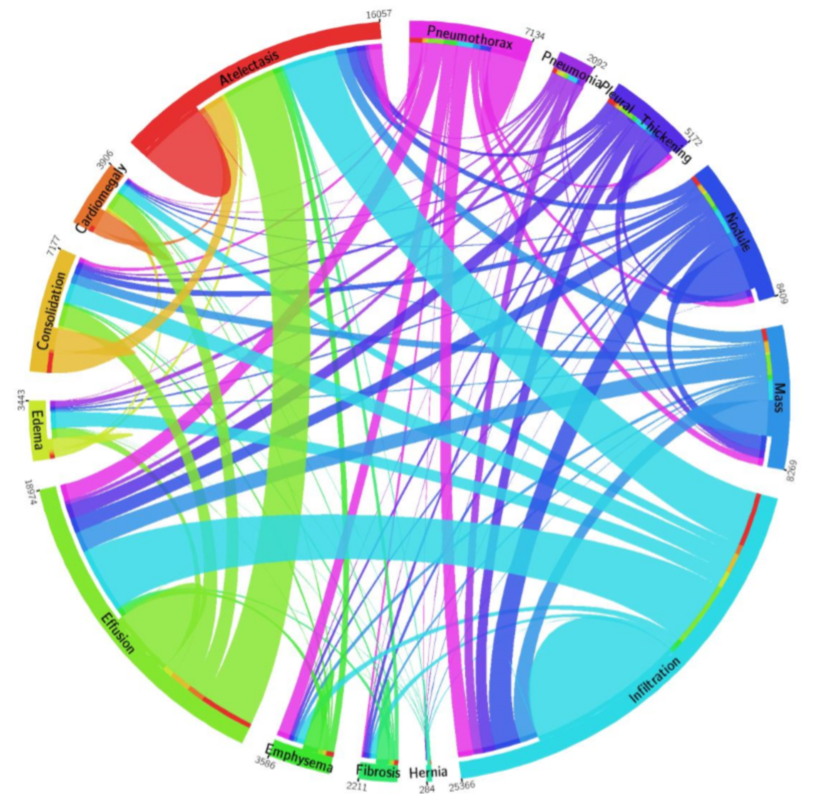
**致谢**:这项工作由国立卫生研究院临床中心(clinicalcenter.nih.gov)和国家医学图书馆(www.nlm.nih.gov)的校内研究项目资助。我们感谢NVIDIA公司对GPU的捐赠。

**Reference:** [1] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald Summers, ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, IEEE CVPR, pp. 3462-3471, 2017

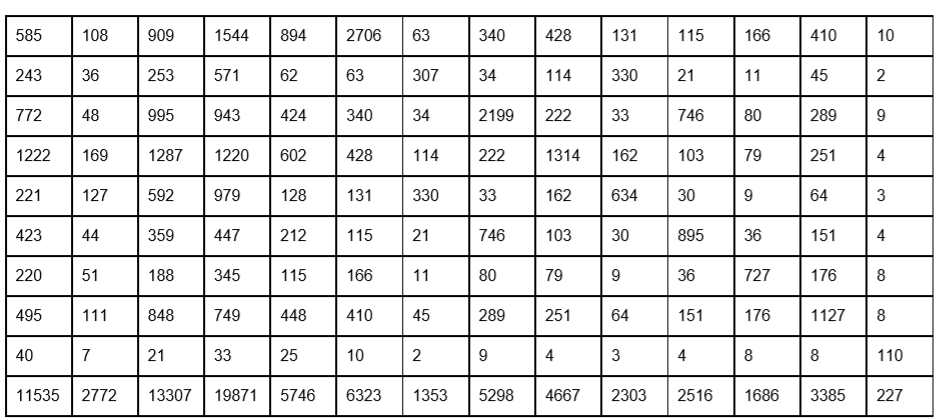
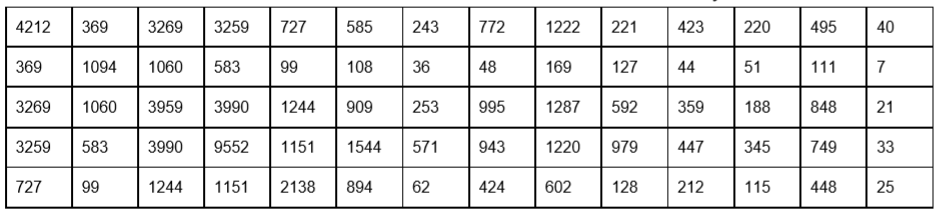
[2] Hoo-chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, Ronald M. Summers, Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation, IEEE CVPR, pp. 2497-2506, 2016

[3] Open-i: An open access biomedical search engine. https: //openi.nlm.nih.gov

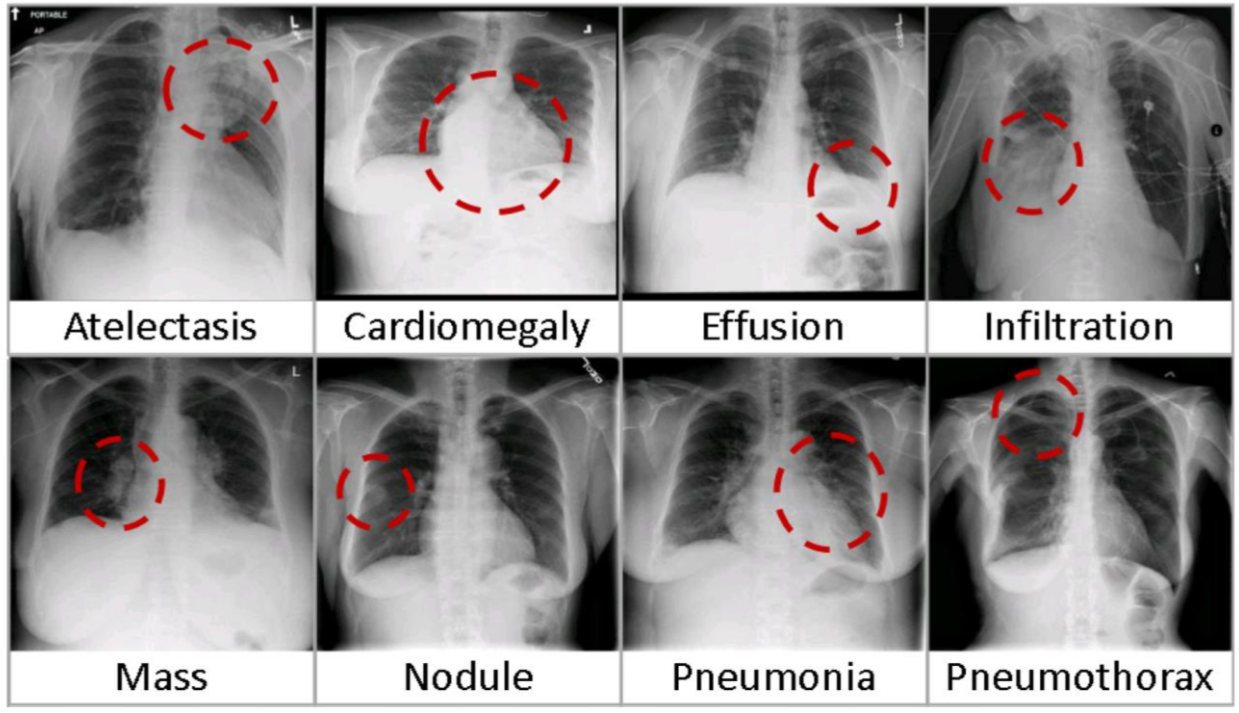
A、14种疾病的分布及共现统计



A.2在这个胸部x光数据集中的14种胸部疾病的共现矩阵



B、八个常见的胸部疾病的视觉例子



C、使用弱监督深度神经网络进行疾病定位的两个例子

