Milestone report

李诗懿[*1] and 陈旭鹏[†2]

[1] 计算机科学与技术系, 清华大学
[2] 生命科学学院，清华大学

2017 年 12 月 3 日

摘要

我们希望建立一个使用胸片图像来诊断心脏疾病的模型，我们从医院及 openi 的公开数据集中收集到了一定数量的胸片。通过一系列的清洗和预处理工作，共得到 3026 张包括 VSD、ASD、TOF 等心脏病类型的病人胸片以及 2017 张正常人胸片。我们要解决的最主要问题是使网络精确地学到心脏的特征，保证网络的可靠性的同时保证准确率。我们尝试了 VGG16 网络进行分类，以及模型可视化方法分析模型学习的特征的问题。通过学习医生的诊断经验，标注了部分图像，并使用 opencv 对图像进行一系列处理，对心脏的轮廓进行多边形拟合。

关键词：胸片、心脏疾病、VGG-net、opencv

1 选题背景

1.1 医疗大数据与精准医疗的火热

医疗健康问题是人们长久关心的大问题，近年来十分火热的深度学习技术在图像处理领域应用广泛，取得很好的效果，深度学习技术也已经在医学图像的部分问题上取得了较好的结果，人工智能技术解决大数据、精准化问题的能力也被人们广泛看好。

我们团队选择了应用胸片诊断心脏疾病的课题。我们的理想是利用深度学习技术，实现更快捷迅速的疾病诊断, 包括独立诊断与辅助医生诊断。我们相信未来基于深度学习的医学诊断和健康服务会广泛地应用到各个医院，为病人提供更好的康复和治疗机会。

[*] shiyi-li15@mails.tsinghua.edu.cn

[†] xp-chen14@mails.tsinghua.edu.cn

1.2 为什么要选择诊断心脏病的问题

心脏疾病每年夺去数以百万计的民众的生命，能够更好的发现、诊断心脏病可以帮助疾病的预防与治疗。胸片作为最便宜、快捷、方便的检测心脏与肺部疾病的材料，对于偏远地区和先天性心脏病的诊断具有很大的意义。我们选择课题的另一个原因是考虑到可以通过医生获得胸片的数据，这样的数据有一定的价值，在付出了一些努力之后，我们拿到了原始的，一手的病人样本。我们还通过学习医生的诊断更好地认识需要解决的问题，我们认识到用胸片诊断心脏疾病是一个有意义且有挑战的问题，而深度学习技术是有希望更好地解决这个问题的。

2 问题分析与解决

我们认为医疗影像类问题，最重要的不仅是模型的诊断准确率，而且还要保证模型的可靠性。简单地套用图像识别的模型和方法并不是终点，我们分析了一些已有的研究，使用了一些可视化的方法分析了模型学习到的特征，并且发现直接适用图像分类模型的问题，即无法控制其学习心脏的特征，导致模型的诊断不具备可靠性。

为了让模型真正学到心脏区域的特征，我们想要模型关注图像的更多细节，但是胸片数据过于相似，且缺少对具体区域的标注。因此我们尝试了使用人工简单标注加上 opencv 一系列处理的方法，拟合出心脏的多边形轮廓坐标，接下来我们希望进一步利用我们的标注针对性地对心脏区域进行重点学习。

3 技术

3.1 问题描述

在深度学习应用到医学影像方面，已经有相当多的研究了，尤其是癌症检测等领域。但是关于心脏病的研究并不是很多，尤其是只利用胸片诊断心脏病，和 CT、MRI 这样的技术产生的数据量相比，胸片所能提供的信息量少了很多的，在数据量上面更加有挑战。

我们选择用胸片数据，是考虑到胸片是最简单有效而且成本低的诊断方式，对便宜和快捷的诊断很有意义。但是分析难度大，也给我们带来了很大的挑战。我们不能抛开生理意义，直接套用模型就试图解决这个问题，我们发现已有的一些用胸片诊断心脏病研究，大多数是直接使用一些比较经典的深度学习图像处理网络，得出一些“比较好”的结果，而没有深究这背后的问题。

大多数的论文大多使用了 VGG、Resnet、Densenet 等模型，基本没有考虑过分割问题，也没有考虑网络的可靠性，是否学到心脏相关的特征，并且利用这些特征进行诊断，是没有探讨的。

而我们认为可靠性与透明性是医学影像问题最重要的需要解决的问题。我们认为只是收集数据，直接送进网络里，是远远不够的。如果模型的整个处理过程都处于无法分析和掌控的黑箱状态，那么这样的技术也就没法真正推广。

一个有趣的例子可以说明准确率与可靠性是多么重要的事情，假设模型对于心脏病的检出率为 0.8，即患病者检验结果为阳性以及未患病者检验结果为阴性的概率为 0.8，人群中患有先心病的概

率为 0.0074，随机从总体中抽取一个人进行检测，检查结果为阳性时此人患病的概率可以用 Bayes 公式计算如下：

$$\begin{aligned}\mathbb{P}(A|B) &= \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(A)\mathbb{P}(B|A)+\mathbb{P}(A^c)\mathbb{P}(B|A^c)} \\ &= \frac{0.0074\times 0.8}{0.0074\times 0.8+(1-0.0074)\times 0.2} \\ &= 0.02874903 \approx 2.87\%\end{aligned} \tag{1}$$

可以看到实际上一个准确率 0.8 的模型对于真阳性的检出率是相当低的，因为发病率总体较低，除非诊断准确率非常高，才会提高真阳性的概率，而一个不够可靠的检测方法，就更容易使得参与检测的患者怀疑，反对和排斥。

我们一开始就对这个问题非常重视，一开始我们就尝试了区分正常样本和心脏病的其中一个主要类别（VSD），在 VGG16 模型经过 fine-tune 之后可以获得 0.99 的测试准确度，我们很难知道这样好的结果是模型通过学习心脏的轮廓特征得到的，还是通过学习其他无关区域得到的。已有的一些研究大多数满足于使用已有模型，获得了一个较高的的准确率，我们认为这还没有解决完问题。

我们使用了一些可视化的方法，分析出了网络学习各个区域的权重，结果是意料之中的：神经网络其实是在使用肺部的一些特征在分类的，其实它并没有使用心脏的特征在预测心脏病，实际上这是一个看起来不错，但是不够可靠的模型。

为了解决让模型学习到真正的心脏特征的问题，我们咨询了医生诊断的一些经验，通过人工标注，结合 opencv 进行一系列预处理等方法，目前已经标注了部分心脏的轮廓，下一步我们希望通过 U-net 网络，利用我们找到的轮廓对心脏区域进行分割，进一步重点学习心脏部分的特征。

3.2 解决方案

为了实现利用胸片诊断心脏疾病的目标，我们从与医院医生沟通获得胸片数据开始，经历了处理图片，分析问题，建立模型等过程。

3.2.1 预处理

首先我们拿到的部分图片是 DICOM 格式，为了不侵犯患者的隐私，我们需要处理掉病人的信息，并且转成 png 或者 jpg 格式的文件。由于胸片的规格并不统一，我们还将图片处理成了几种不同的大小和分辨率。在之后具体训练的时候，因为机器的内存所限。我们根据网络的规模继续调整了图片的大小。

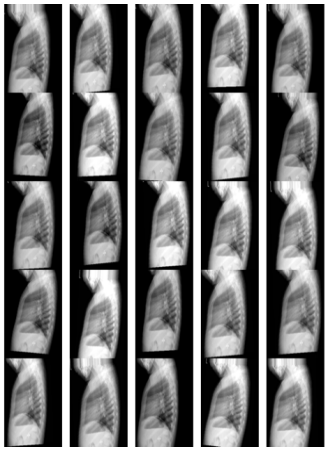
3.2.2 VGG 网络进行分类

VGG 网络是由 Alex-net 发展而来的网络，牛津大学 visual geometry group（VGG）的 Karen Simonyan 和 Andrew Zisserman 于 14 年撰写了论文探讨了深度对于网络的重要性；并建立了一个 19 层的深度网络，在 ILSVRC 上取得过定位第一，分类第二的好结果。

VGG 的目标是 go deeper，同时也继承了 LeNet 以及 AlexNet 的一些框架，尤其是跟 AlexNet 框架非常像，VGG 也是 5 个 group 的卷积、2 层 fc 图像特征、一层 fc 分类特征。根据前 5 个卷积 group，每个 group 中的不同配置，VGG 与 AlexNe 的主要不同在于：在第一个卷基层层使用更小的 filter 尺寸和间隔，VGG(kernel size=3, stride=1), AlexNet(kernel size=11, stride=4)，以及 VGG 是在整个图片和 multi-scale 上训练和测试图片

我们使用的 VGG16 的网络基本结构如下：

(a) VGG 网络



(b) 同一样本经过仿射变换产生的样本

图 1

我们一开始就选择先使用 VGG 网络来试一试我们所要面临的问题。我们用 keras 定义了 VGG 的网络结构，先做了一个简单的区分胸片正面与侧面的分类来测试流程。发现准确率大概有 0.96，我们觉得这个准确率不算太高，毕竟只是区分了正面侧面，应该有很明显的区分特征才对。

后续我们又尝试了和其他研究一样使用 VGG 来区分正常人、VSD、ASD 患者，我们没有加入更多种类的疾病样本是考虑到我们收集到的胸片样本数量有限，因此在早期的尝试中并没有加入这些病人的样本。在解决学习可靠性问题后，可以考虑解决小样本学习的问题。

3.2.2.1　图像增广　我们也使用了图像变换的方法，因为收集的样本量有限，经过层层筛选和清洗后，所能利用的图片数比较少，我们使用了各种常用的仿射变换增加了样本量。以一个病人的侧面胸片为例，我们通过旋转、缩放、对称、平移等变换认为增加了一些样本。这是在向医生咨询后，了

解到心脏的相对位置信息并非是诊断的一个需要考虑的（重要）因素后，我们确定可以使用这种常用的图像扩增方法。

我们选择使用了一个在 imagenet 数据集上 pretrained 过的 VGG 模型用来分类，对 VGG 网络的最后三层进行了 fine tune。我们也找到了其他的常用的图像识别网络，但是通过我们的分析，此问题的关键不在于简单的使用常见的网络提升准确率，因此我们并没有停留在换用各种网络，寻找‘最高的准确率’上。

在结果展示部分可以看到，单纯的使用预训练的 VGG16 网络，经过 fine tune 之后，就取得了“不错的结果”。但是我们还需要考虑到模型学习的究竟是什么特征的问题。

3.2.2.2 分析模型学习效果 在问题的难度与难点部分，我们对比过一些文章的工作，几乎没有文章分析了神经网络学到了什么，而我们认为这种分析是有意义而且必须的。

我们开始时试着查看每一层网络都输出了什么，比如上图所示同一批次图片在 block5_conv1 层的输出。（图 6）但是显然这种方法并不能告诉我们网络究竟在学什么。因此我们尝试了几种分析网络究竟学到了什么特征的方法，首先是 sensitivity analysis，这是一种基于梯度的算法，计算图像的 sensitive relevance，可以看到图像不同像素的重要程度。如图 7 所示。这种方法可以让我们可视化地分析究竟那些特征成为了网络学习和分类的关键特征。图中颜色亮的区域为网络重点学习的特征。我们发现非常有趣的一点，网络更多地关注了肺部的区域，而忽视了心脏区域的特征，这与我们希望的恰恰相反，也许这个网络更适合做肺部疾病的检测，而不是心脏疾病的检测。通过这样的分析，我们可以大概得出结论，单纯地适用图像分类网络实现胸片的诊断是不可靠的。

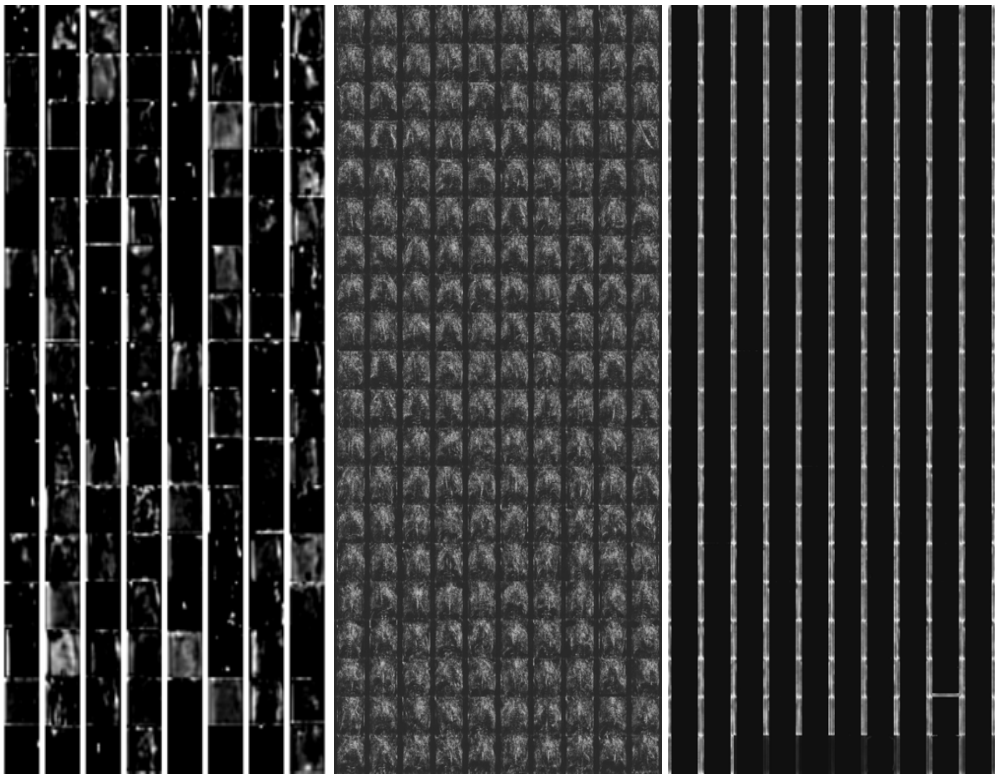
我们也尝试了 deep taylor decomposition 的可视化分析方法，deep taylor decomposition 是 layer-wise relevance propagation 的一种，但是我们得到了一个很奇怪的结果。如图 8 所示。由于时间原因，我们已经通过 sensitivity analysis 发现了神经网络确实学习有误的事实，也就没有在这里耽误时间寻找 deep taylor 分析的问题，而是想办法解决学习特征有误的问题。

3.2.3 进行人工标注

我们想要解决的问题是更精准、可靠地通过胸片诊断病人是否得病，因此我们认为必须想办法让网络主要关注心脏部位的特征。在和心脏病方面的专家交流的过程中，我们已经知道心脏几个部位的特征在诊断中占据了绝大多数的权重（图 3），因此我们希望能够让模型关注到心脏的特征。

我们的图片是缺少人工标注的，而如果希望使用 U-net 分割出心脏的轮廓再针对性学习，就需要我们进行一些人工的标注。但是如果像图 3 中对每个图片进行几个部位的圈画，就需要很多的时间花费，因此我们想到用 opencv 帮助我们解决标注的部分问题。从拟合心脏轮廓的思路出发，我们希望能够标出来心脏一圈的轮廓。我们寻找了一系列方法来将一张胸片图处理得只剩下心脏的轮廓。因为将胸片图二值化后，有大量的干扰组织和骨骼，噪声很大，而且大量的图片的心脏轮廓并不闭合，不符合要求。而如果人工标注出心脏的轮廓需要很长时间才能完成，也不现实。

我们最终找到了合适的处理方法，因为心脏上下部位并没有明显的轮廓，因此我们首先人工标注了胸片的上下两个边界（目前已标注约两千张），剩下的都用 opencv 处理，获得心脏的轮廓，并且用多边形拟合出轮廓。



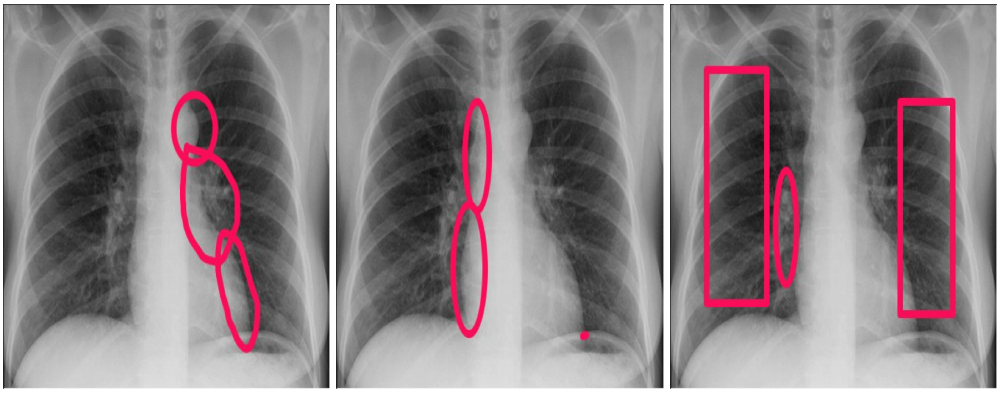
(a) 神经网络某层输出示意图　(b) sensitivity analysis　(c) deep taylor decomposition

图 2

3.2.3.1 使用 opencv 获得心脏轮廓线的具体步骤：

- **二值化**：轮廓检测有时能更好的反映图像的内容。而要对图像进行轮廓检测，则必须要先对图像进行二值化，图像的二值化就是将图像上的像素点的灰度值设置为 0 或 255，这样将使整个图像呈现出明显的黑白效果。
- **膨胀与腐蚀**：腐蚀和膨胀是图像的形态学处理中最基本的操作，之后遇见的开操作和闭操作都是腐蚀和膨胀操作的结合运算。腐蚀和膨胀的应用非常广泛，而且效果还很好：腐蚀可以分割独立的图像元素，膨胀用于连接相邻的元素，这也是腐蚀和膨胀后图像最直观的展现。
- **填充与区域选择**：漫水填充法是一种用特定的颜色填充联通区域，通过设置可连通像素的上下限以及连通方式来达到不同的填充效果的方法。漫水填充经常被用来标记或分离图像的一部分以便对其进行进一步处理或分析，也可以用来从输入图像获取掩码区域，掩码会加速处理过程，操作的结果是某个连续的区域。
- **多边形拟合**：使用 opencv 算法将一段连续光滑曲线折线化，最终达到向多边形逼近的目的。

经过这样的处理后，每张图都会生成一个对应的多边形各顶点的 txt 文件，用于接下来的训练使用，在具体的训练时需要将轮廓的坐标重新转化为区域（traces to masks）。通过训练 U-net，我们发现神经网络实现了我们的要求，自己也可以学会很精确地标注心脏的轮廓。（图 8）



(a) 心脏左侧重点关注区域　(b) 心脏右侧重点关注区域　(c) 肺部重点关注区域

图 3: 诊断时重点关注的区域

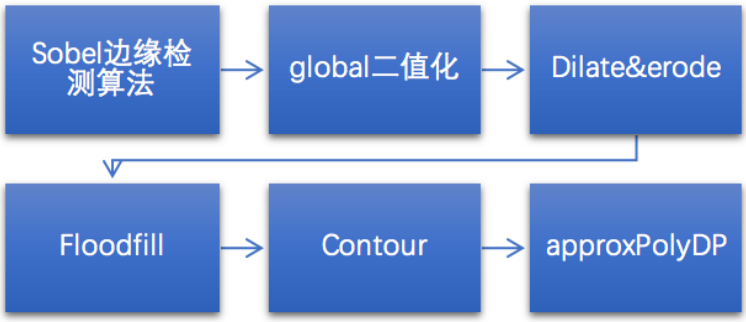


图 4: opencv 拟合心脏轮廓的步骤

多边形的轮廓示意图如图 5：

由于用正面的轮廓训练 U-net 来划分侧面心脏的轮廓，效果当然不好，我们又手工标注了一批侧面的心脏轮廓，用来提高网络分割侧面胸片心脏区域的准确性。

3.2.3.2　进一步数据处理　因为我们收集了三种不同来源的数据，每种数据都整理出了正常和得病的一些种类，做过多种形式的处理（详见预处理和标注部分），包括调整分辨率的图片，加分割线的图片，二值化的图片，多边形拟合出的心脏轮廓坐标值 txt 文件（traces）。为了方便之后的训练，我们又进行了一些处理。

在图片的 label 方面，我们直接将图片的名称加上疾病名称方便对应。因为机器内存的限制，在每批训练送入图片数量和分辨率的权衡中，我们将图片进一步处理成 320*320 分辨率大小。并且使用 python 的 h5py 库将数据集转化为 HDF5 格式的文件，实现更快的存取，并且根据每次训练的需要，将不同的数据集合并在一起。

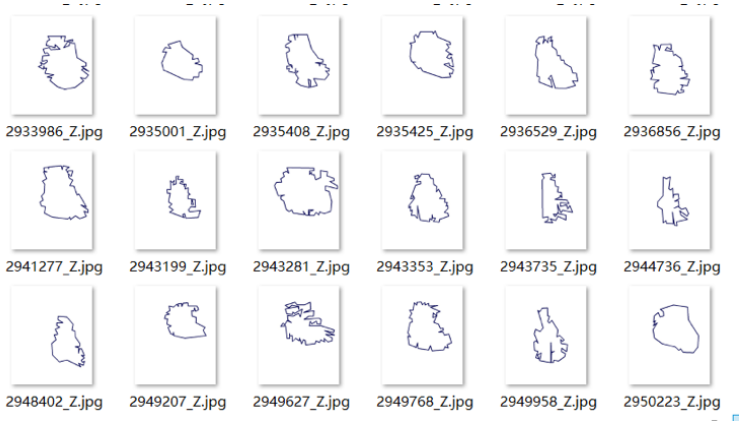


图 5: opencv 拟合出的多边形轮廓

3.3 结果展示

3.3.1 以下是我们用 VGG 分类的一些结果：

3.3.1.1 网络：VGG，数据集：Normal_ny VSD 十折平均结果：

accuracy	variance	auroc	variance
0.940	0.00019629	0.988	1.94358E-05

3.3.1.2 网络 VGG，数据集：Normal_ny VSD ASD 十折平均结果：

type	accuracy	variance
总体	0.696	0.010905912
Normal	0.945	0.000448034
VSD	0.709	0.010017944
ASD	0.738	0.012352246

3.3.1.3 网络：VGG，数据集：Normal_indiana VSD 十折平均结果：

accuracy	variance	auroc	variance
0.975	0.000156748	0.999	1.15406E-06）

3.3.1.4 网络：VGG，数据集：Normal_indiana VSD ASD 十折平均结果：

type	accuracy	variance
总体	0.852	0.001002353
Normal	0.979	0.000124253
VSD	0.863	0.001000045
ASD	0.862	0.001117035

前面已经分析过，虽然 VGG 和其他研究一样，都可以取得很高的 accuracy 和 auc_roc, 但是我们认为其结果是不可靠的，因此我们接下来希望利用 U-net 网络，使用我们拟合出的部分样本的心脏边缘轮廓，转化为 Mask，实现对心脏区域的分割，帮助我们获得每个样本的心脏部分区域，再进行更针对性的学习。

4 后续工作安排

接下来我们将完成以下几项工作：

- **完成标注**：完成剩余的图像的标注，由 opencv 进行预处理完成对轮廓的拟合。
- **将轮廓转化为 Mask**：将 traces 转化为 mask，训练一个 U-net 网络对心脏区域进行分割。
- **重点学习心脏区域的特征**：训练分类模型重点学习心脏区域的特征，获得更可靠的结果。

5 参考文献

[1]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.

[2]J. M. Carrillo-de Gea, G. Garc´ıa-Mateos, J. L. Fern´andez-Alem´an, and J. L. Hern´andez-Hern´andez, "A computer-aided detection system for digital chest radiographs," Journal of Healthcare Engineering, vol.

[3]S. Candemir, S. Jaeger,W. Lin, Z. Xue, S. Antani, and G. Thoma, "Automatic heart localization and radiographic index computation in chest x-rays," in SPIE Medical Imaging. International Society for Optics and Photonics, 2016, pp. 978 517–978 517. 2016, 2016.

[4]A. Kumar, Y.-Y. Wang, K.-C. Liu, I.-C. Tsai, C.- C. Huang, and N. Hung, "Distinguishing normal and pulmonary edema chest x-ray using gabor filter and svm," in Bioelectronics and Bioinformatics (ISBB), 2014 IEEE International Symposium on. IEEE, 2014, pp. 1–4.

[5]G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. S´anchez, "A survey on deep learning in medical image analysis," arXiv preprint arXiv:1702.05747, 2017.

[6]Mohammad Tariqul Islam, Md Abdul Aowal, Ahmed Tahseen Minhaz, Khalid Ashraf, "Abnormality Detection and Localization in Chest X-Rays using Deep Convolutional Neural Networks", arXiv preprint arXiv:1705.09850v3, 2017

[7]Yuxi Dong, Yuchao Pan, Jun Zhang and Wei Xu, ”Learning to Read Chest X-Ray Images from 16000+ Examples Using CNN”, Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2017 IEEE/ACM International Conference on.

[8]Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[9]Matthew D. Zeiler Adaptive Deconvolutional Networks for Mid and High Level Feature Learning. International Conference on Computer Vision, 2011.

[10]Jimmy Lei Ba MULTIPLE OBJECT RECOGNITION WITH VISUAL ATTENTION, ICLR 2015.

[11]Ciresan D, Giusti A, Gambardella L M, et al. Deep neural networks segment neuronal membranes in electron microscopy images[C]//Advances in neural information processing systems. 2012: 2843-2851.

[12]Grégoire Montavon, Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognition 65 (2017) 211–222