

基于精简卷积神经网络的低分辨率乳腺癌识别

王兵锐^{1,2}, 张新刚¹, 杨晓非²

(1. 南阳师范学院 河南省智能应急研究中心, 河南 南阳 473007;

2. 华中科技大学 光学与电子信息学院, 湖北 武汉 430074)

摘要:乳腺癌严重威胁女性健康,应用人工智能进行及时诊断是应对乳腺癌的重要方法。卷积神经网络(convolutional neural network, CNN)是人工智能中最经典的处理方法之一。通常健康人数量(称作多数类数据)远大于癌症患者数量(称作少数类数据),学习后的网络模型严重倾向于多数类导致失败。针对这种数据集不平衡问题,对多数类健康数据集采用随机下采样减少数据,对少数类癌症数据采用数据增强扩充处理,控制网络模型的权重比例,同时融合这三种方法应对数据集不平衡。针对采用的50×50像素癌症数据集分辨率过低的问题,调整到100×100像素以便提取更多细节。提出一种4卷积层CNN网络,分别针对两种像素进行训练测试,并与经典的16层VGG16网络进行对比。精度损失曲线和混淆矩阵的实验结果表明,提出的CNN的乳腺癌识别精度优于VGG16多达4个百分点。

关键词:卷积神经网络;乳腺癌;低分辨率;精度损失;不平衡

中图分类号: TP18

文献标识码: A

文章编号: 1673-629X(2020)09-0200-05

doi: 10.3969/j.issn.1673-629X.2020.09.036

Low Resolution Breast Cancer Recognition Based on Simplified Convolutional Neural Network

WANG Bing-rui^{1,2}, ZHANG Xin-gang¹, YANG Xiao-fei²

(1. Henan Intelligent Emergency Research Center, Nanyang Normal University, Nanyang 473007, China;

2. School of Optics and Electronics, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: Breast cancer is a serious threat to women's health. It is an important method to diagnose breast cancer in time by using artificial intelligence. Convolutional neural network (CNN) is one of the most classical processing methods in artificial intelligence. Generally, the number of healthy people (called as majority data) is far greater than that of cancer patients (called as minority data), and the learned network model tends to fail in most cases. For this kind of data set imbalance, the healthy data are sub-sampled randomly. A few kinds of cancer data are augmented, and the weight proportion of network model is controlled. At the same time, these three methods are combined to deal with the imbalance of data. In order to solve the problem of low resolution of 50×50 pixel cancer data set, we adjust the data set to 100×100 pixel to extract more details. A 4-convolution layer CNN network is proposed and compared with the classical 16-layer VGG16 network. Two kinds of pixels are trained and tested. The experimental results of accuracy loss curve and confusion matrix show that the recognition accuracy of breast cancer based on the proposed CNN is better than VGG16 by up to 4 percentage points.

Key words: convolutional neural network; breast cancer; low resolution; accuracy loss; imbalance

0 引言

乳腺是由腺体、导管和脂肪组织等构成,乳腺癌发生在腺体或导管的上皮组织。乳腺癌是危害女性健康的重要恶性肿瘤,位居女性恶性肿瘤第一位。乳腺癌手术之后,依然有转移的可能性,需要及时诊疗^[1]。乳腺癌如果不及时诊疗,或者诊断失误认为健康,容易发生乳腺癌转移^[2-3]。转移之后,乳腺癌细胞之间连接

松散易脱落,游离的癌细胞可以随血液或淋巴液散播,乳腺癌继续发展转移到肺时,引起胸腔积液、呼吸困难。转移到骨骼时^[4],入侵骨髓,出现不规则的骨质破坏,引起骨折甚至瘫痪。转移到大脑时,引起脑肿胀导致颅内压增高,影响中枢神经,产生各种疼痛。许多乳腺癌患者早期疼痛并不明显,容易忽视诊断,往往会造成更加严重的后果。积极的进行早期诊断^[5],尤其及

收稿日期: 2019-10-16

修回日期: 2020-02-20

基金项目: 河南省科技攻关项目(182102210114); 河南省高等学校重点项目(18A520044); 南阳师范学院青年项目(501-17323)

作者简介: 王兵锐(1986-),男,博士,研究方向为卫星通信、人工智能; 张新刚,副教授,研究方向为人工智能; 杨晓非,博士,教授,研究方向为图像传感与数据处理。

时治疗,一定程度上抑制癌细胞生长,使病情得到控制。传统疾病治疗需要医生通过辅助机器检测乳腺癌并对病情做出判断。人工智能应用在医学肿瘤上是当前的研究热点^[6-7]。

应用卷积神经网络进行乳腺癌诊断^[8-9],可以将病情数据交给训练过的学习模型,然后由模型去判断病情,判断结果更加精确,人工判断一般有主观倾向,并且带有经验倾向,容易判断失误,CNN 可以快速处理乳腺癌数据,效率远远高于人工,更加节省时间。CNN 对于诊断乳腺癌,及早发现乳腺癌,保障人类生命安全等方面都会产生巨大的影响。

1 数据不平衡处理

通常健康人的数量远大于癌症患者的数量,获得的癌症图片远小于健康图片的数量,导致数据类别分布不平衡。应用深度学习时,这种不平衡会导致学习模型偏向于健康类,测试时容易将癌症类判别为健康类,学习模型在测试数据集上的泛化性不好^[10]。所以,需要得到一个性能不错的学习模型,对健康和癌症图片都能提供较好的分类准确率。解决数据不平衡的方法通常分为两类^[11],校正学习模型和调整数据集。

校正学习模型常采用损失函数加权的方法,具体过程如下所述。

$$\text{loss} = - \sum_k t_k \log_e y_k \quad (1)$$

式(1)为交叉熵损失函数, t_k 是正确的标签,并采用独热码表示, y_k 是学习模型的输出。数据不平衡时,采用式(1)计算出来的损失值是相近的,对少数类容易判为多数类。此时需要在式(1)右边乘以一个加权系数,从而加大少数类的损失值,使学习模型更倾向于预测少数类,这样能达到平衡预测数据。式(1)改进为:

$$\text{loss} = - \sum_k \lambda t_k \log_e y_k \begin{cases} \lambda = 1, k = \text{多数类} \\ \lambda > 1, k = \text{少数类} \end{cases} \quad (2)$$

其中, λ 为加权系数,训练多数类时, λ 的值为 1,训练少数类时, λ 的值大于 1,从而加大少数类的损失,更多关注少数类的样本。

调整数据集通常分为三种方法,合成少数类数据、减小多数类以及增大少数类^[12-13]。合成少数类数据最经典的方法是 SMOTE (synthetic minority over-sampling technique)^[14-16]。该算法合成新的少数类数据,对每个少数类数据,根据欧氏距离从它的最近邻中随机选取另一个少数类数据,然后在这两个少数类数据的连线上随机选择一个位置,把该位置作为一个新的少数类数据。SMOTE 算法常用来处理股票、表格等数据且效果较好,用来处理图像的情况较少。

减小多数类方法是从多数类样本中随机移除一些数据,从而达到数据平衡。但有可能丢失包含重要特征的信息。但可以随机选择等量的不同的多数类数据,和少数类数据一起学习训练,重复 n 次学习并得到多个学习结果,最后将多个结果进行大数判决,从而决定最终的分类结果。增大少数类,最简单的方法是随机复制一些少数类数据,扩充到和多数类一致的数据量。如此的增大策略可以达到所需的任意平衡,而且该方法比较简单,易于理解和实现。但这样增大少数类数据,并没有增加数据的多样性,容易导致过拟合。

比较有效的增大少数类的方法是,做数据增强处理。常用的操作有旋转、剪切、尺度变换、反转、平移、噪声扰动、色彩抖动。旋转是对输入图像进行 0 度到 360 度随机旋转。裁剪是对图像随机摘取出一部分,并放大到和输入图像一样大小,容易丢失主要特征。尺度变化是按照一定比例对图像随机进行整体缩小或放大,和剪切有一定区别。反转是对图像做垂直或水平反转。有些图像比如猫,做垂直翻转后,物体彻底改变,特征变化巨大。平移是随机对图像左右移动一定的比例。噪声扰动是指在图像中随机加入少量的高斯噪声,可以有效防止过拟合,让 CNN 不能学习图像的全部特征。色彩抖动指在颜色空间中,每个通道随机抖动一定的程度,比如改变图像的对比度、亮度、饱和度等,容易产生不符合实际的图像,使得学习效果变差。

2 提出的 CNN 网络

CNN 是一类具有卷积计算的前馈神经网络^[17],卷积神经网络具有局部感知、特征降维等特点。因其优异的算法性能,卷积神经网络多被用于深度学习中,用来进行图像识别,自然语言处理等。CNN 是以梯度下降迭代的方法来进行学习的,要对输入的信息进行预处理归一化。因为不同的特征信息对应的值域不同,为了提高之后的训练效率,要把各个特征信息的值域尽量控制在相同值域上,一般采用除以 255 进行归一化。CNN 包括卷积层、激活层、池化层、全连接层等。

卷积层最重要的作用就是用来进行特征信息提取。卷积层不能一步处理一图片的全部信息,需要把图片等分成多份,对每一小份进行处理。卷积层关键的运算是卷积运算,如式(3)所示。卷积运算会提取出每一小份所蕴含的特征,输入数据是个矩阵,需要和权重系数构成的矩阵进行卷积。权重系数的初始值是随机生成的,通常采用正态分布函数。在后期的训练学习中,权重系数会不断调整。权重矩阵也称为卷积核。图片的每个部分与权重矩阵至少进行一次运

算,然后产生一个完整的卷积层输出。

$$\begin{aligned} u^i &= w^i x^{i-1} + b^i \\ x^i &= f(u^i) \end{aligned} \quad (3)$$

其中, x^{i-1} 是第 i 层的输入数据, w^i 是第 i 层的权重系数, b^i 是一个常数参数, $f(u^i)$ 是一个激活函数, x^i 是第 i 层的输出值。

卷积运算之后会得到不同的值。卷积层的输出越大,说明当前卷积矩阵与输入的图片越匹配。换言之,通过卷积层输出可以判断哪个卷积核更能描述待识别图片的具体特征。特定的某个卷积矩阵能够识别相同或相似的特征,如果要识别其他特征则需要多个卷积矩阵。卷积层的特点就是不断更新卷积矩阵,确定有哪些卷积矩阵更能揭示待识别的输入图片。

激活层的特点表现为非线性。如果不采用激活层,各层之间的关系将是一种线性映射。激活函数的非线性特征经过深层次的反复叠加使得神经网络可以趋于任意的函数,激活层使得 CNN 拥有更好的适用性。池化层去除卷积层提取到的无用信息,降低矩阵维度,减少训练参数以及减少噪声向下一层传递。最大池化是指在相应区域内选择一个最大的数代表整个区域,如图 1 所示,图中采用的是 2×2 大小的矩阵池化,也就是 4 个元素中选取 1 个最大值。均值池化是指用相应区域内的平均值去代表整个数据区域。最大池化和均值池化是经常采用的方式。

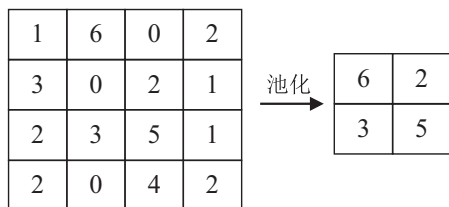


图 1 最大池化

全连接层通常被放置在最后几层,其作用主要表现在分类上。卷积层每一次获取到的都是局部特征,而想要起到分类识别的作用,局部特征是远远不够用的。全连接层是把所有的局部特征给组合起来,成为完整的特征图,以便完成正确的分类识别。

提出精简的 4 个卷积层的 CNN 来进行快速运算,第一层为卷积层 1、激活层,第二层为卷积层 2、激活层、池化层,第三层为卷积层 3、激活层,第四层为卷积层 4、激活层、池化层,按照层层递进的关系如图 2 所示,最后一层为全连接层。激活层采用的激活函数为 Relu,池化层采用最大池化方法。4 个卷积层采用的权重矩阵都是 3×3 矩阵,每个卷积层含 64 个以上的权重矩阵。4 个卷积层,也就是卷积层 1、2、3、4 采用的权重矩阵个数为 64、64、128、128,如图 2 中用括号进行标注。

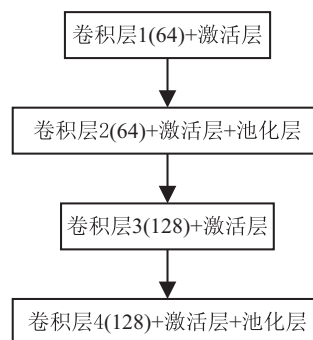


图 2 提出的 4 层 CNN

3 实验测试

采用的是 Kaggle 公开提供的 50×50 的乳腺癌数据集,健康图片约为 19.8 万张,癌症图片约为 7.8 万张。把癌症图片中包含信息较少的小图片去掉,同时也方便快速运算,抽取得到癌症图片为 4.6 万张,采用减小多数类的随机采样方法抽取 6 万张健康图片。同时采用增大少数类的数据增强方法,把癌症图片增加到 5 万张,数据增强扩大产生 4 千张图片。如果在程序运行时,一面扩增数据一面运行模型,处理速度较慢,预先把数据扩增好会提高运行速度。数据增强的具体方法为水平移动范围为 $50 * 0.08 = 4$,即图片水平偏移的幅度为 4 个像素。垂直移动范围为 $50 * 0.1 = 5$,即图片垂直偏移的幅度为 5 个像素。图片随机转动的角度设置为 12 度,设置水平随机翻转也就是做左右对称变换。当进行变换时超出边界的点根据就近插值原则进行处理。采用标准化进行增强,将输入的图片除以数据集的标准差完成标准化。

3.1 精度损失曲线测试

除了针对数据集进行控制,在训练学习模型时,采用校正学习模型的损失函数加权方法,来平衡健康类和癌症类的学习影响。把数据集拆分为训练集、验证集和测试集,分别为 8.8 万张、1.21 万张、9.9 千张。由于采用的数据集的分辨率较低,把 50×50 像素图片增大到 100×100 大小,从而更好地提取细节特征。在人工智能中,迁移学习容易理解,且速度较快备受青睐。迁移学习是把一个场景学到的知识用来帮助应对新场景的学习任务,比如有编写歌词的学习经验,就有助于唱歌跳舞。

迁移学习最经典的应用是 VGG16,包含 13 个卷积层和 3 个全连接层,如图 3 所示。最大输入为 $224 \times 224 \times 3$ 的图片,最小输入为 $48 \times 48 \times 3$ 的图片。经过两次卷积层,每层包含 64 个卷积核,进行一次池化操作。之后又经过两次卷积层,每层包含 128 个卷积核,进行一次池化操作。再经过三次卷积层,每层包含 256 个卷积核,进行一次池化操作。然后进行三次卷积层,每

层包含 512 个卷积核,进行一次池化,这个操作重复两遍。最后经过三个全连接层,前两个全连接层有 4 096 个神经元节点,第三层有 1 000 个神经元节点。前 13 层采用的卷积核个数,依次为 64、64、128、128、256、256、512、512、512、512、512、512。文中应用 VGG16 来识别判断乳腺癌,训练倒数 4 个卷积层和全连接层,其余层进行冻结。

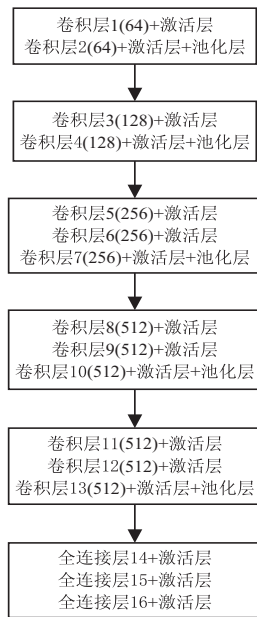
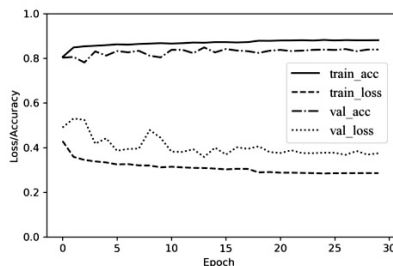
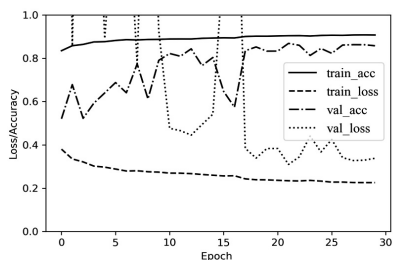


图3 VGG16 结构

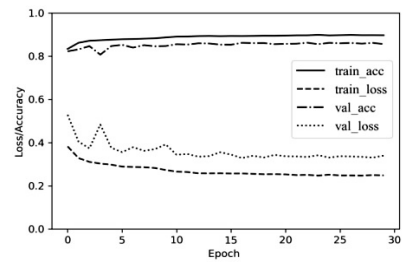
下面从精度损失曲线上对比提出的 CNN 与 VGG16 的性能,如图 4 所示,其中 train_acc 为训练精度,train_loss 为训练损失, val_acc 为验证精度, val_loss 为验证损失。从图中可以发现,乳腺癌图片为 50×50 时,精度损失曲线波动比较大。把图片扩大为 100×100 时,曲线比较平缓,尤其采用提出的 CNN, train_acc 和 val_acc 结合的比较紧密,说明训练效果较好。



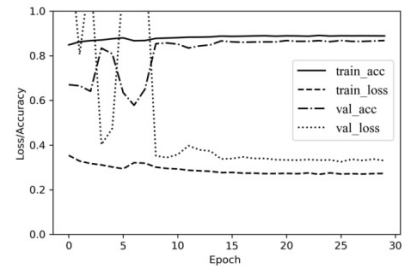
(a) 50×50 时 VGG16



(b) 50×50 时提出的 CNN



(c) 100×100 时 VGG16



(d) 100×100 时提出的 CNN

图4 精度损失曲线

3.2 混淆矩阵测试

上面的精度损失曲线是针对训练集和验证集。还需要对测试集进行考查。采用混淆矩阵针对测试集做进一步的探讨,混淆矩阵容易看到深度学习模型是否将癌症的类别混淆。混淆矩阵是对深度学习模型预测结果的一种总结分析表格,列出测试集图片的真实类别与模型预测类别,以矩阵形式呈现出来。基于提出的 CNN 和 VGG16,面向乳腺癌和健康图片测试集,分别针对 50×50 像素和 100×100 像素进行实验测试,得到混淆矩阵,如图 5 和图 6 所示。图中矩阵的行表示真实值,矩阵的列表示预测值。

		健康		癌症	
		健康	癌症	健康	癌症
健康	a VGG16	2960.00	1940.00	3046.00	1854.00
	b 提出的CNN	119.00	4881.00	53.00	4947.00

图5 50×50 时的混淆矩阵

		健康		癌症	
		健康	癌症	健康	癌症
健康	a VGG16	3210.00	1690.00	3286.00	1614.00
	b 提出的CNN	121.00	4879.00	59.00	4941.00

图6 100×100 时的混淆矩阵

图 5 和图 6 中,左上角表示真实为健康且训练模型判别为健康的数量,称作真阴性 (true negative, TN)。右上角表示真实为健康但训练模型判别为乳腺

癌的数量,称作假阳性(false positive,FP)。左下角表示真实为乳腺癌但训练模型判别为健康的数量,称作假阴性(false negative,FN)。右下角表示真实为乳腺癌且训练模型判别为乳腺癌的数量,称作真阳性(true positive,TP)。混淆矩阵里面列出的是数量,为了更加直观衡量模型的优劣,在混淆矩阵的统计结果上给出进一步的评价指标,正确率(accuracy)、灵敏度(sensitivity)、特异度(specificity)。正确率是指被网络模型识别正确的乳腺癌和健康数除以所有的图片数,一般正确率越高,分类器越好。

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4)$$

4 结束语

采用 CNN 进行诊断识别来及早发现乳腺癌。但乳腺癌数据集远小于健康人的数据集,数据不平衡。采用随机下采样的方法减少多数类健康数据集,采用数据增强的方法增大少数类癌症数据,同时控制学习模型的癌症与健康学习权重比例,融合三种方法应对数据不平衡。通过混淆矩阵,对原始 50×50 像素和调整后的 100×100 像素数据集进行测试评估,根据式(4),50×50 像素时,经典 16 层的 VGG16 网络和提出的 4 卷积层 CNN 网络的准确率分别为 79.2% 和 80.7%。100×100 像素时,VGG16 网络和提出的 CNN 网络的准确率分别为 81.7% 和 83.1%。提出的 CNN 的乳腺癌识别精度在 100×100 像素时优于 VGG16 在 50×50 像素时多达 4 个百分点。采用的癌症数据集分辨率较低,纹理特征不易提取。同时为了方便快速计算、节约硬件资源,提出了 4 卷积层的 CNN,不易过拟合,性能优于 16 层的 VGG16 网络。因为 VGG16 网络容易产生过拟合,从而导致精度较低。

参考文献:

- [1] CHANGSRI C, PRAKASH S, SANDWEISS L, et al. Prediction of additional axillary metastasis of breast cancer following sentinel lymph node surgery[J]. Breast Journal, 2015, 10(5): 392-397.
- [2] HOSSEINI H, OBRADOVIC M M S, HOFFMANN M, et al. Early dissemination seeds metastasis in breast cancer[J]. Nature, 2017, 540(7634): 552-558.
- [3] XIE Hongyan, SHAO Zhimin, LI Daqiang. Tumor microenvironment: driving forces and potential therapeutic targets for breast cancer metastasis[J]. Chinese Journal of Cancer, 2017, 36(3): 121-130.
- [4] SAVCI-HEIJINK C D, HALFWERK H, KOSTER J, et al. A novel gene expression signature for bone metastasis in breast carcinomas[J]. Breast Cancer Research & Treatment, 2016, 156(2): 249-259.
- [5] HA R, CHIN C, KARCICH J, et al. Prior to initiation of chemotherapy, can we predict breast tumor response? deep learning convolutional neural networks approach using a breast mri tumor dataset[J]. Journal of Digital Imaging, 2019, 32(5): 693-701.
- [6] 陈龙, 郑焜, 沈云明, 等. 基于深度学习的神经母细胞瘤计算机辅助分级系统的研发初探[J]. 中国医疗器械杂志, 2019, 43(4): 255-258.
- [7] 祁亮, 沈洁. 机器学习在肝癌诊疗领域的应用进展[J]. 癌症进展, 2019, 17(5): 519-525.
- [8] CHOUGRAD H, ZOUAKI H, ALHEYANE O. Multi-label transfer learning for the early diagnosis of breast cancer[J]. Neurocomputing, 2019, 5(3): 7-25.
- [9] BAKKOURI I, AFDEL K. Multi-scale CNN based on region proposals for efficient breast abnormality recognition[J]. Multimedia Tools & Applications, 2019, 78(10): 12939-12960.
- [10] 史作婷, 吴迪, 荆晓远, 等. 类不平衡稀疏重构造度学习软件缺陷预测[J]. 计算机技术与发展, 2018, 28(6): 125-128.
- [11] BUDA M, MAKI A, MAZUROWSKI M A. A systematic study of the class imbalance problem in convolutional neural networks[J]. Neural Networks, 2018, 106: 249-259.
- [12] STURM B L. Classification accuracy is not enough[J]. Journal of Intelligent Information Systems, 2013, 41(3): 371-406.
- [13] POUYANFAR S, CHEN S C. Automatic video event detection for imbalance data using enhanced ensemble deep learning[J]. International Journal of Semantic Computing, 2017, 11(1): 85-109.
- [14] FERNÁNDEZ A, GARCIA S, HERRERA F, et al. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary[J]. Journal of Artificial Intelligence Research, 2018, 61: 863-905.
- [15] GUTIERREZ P D, LASTRA M, BENITEZ J M, et al. SMOTE-GPU: big data preprocessing on commodity hardware for imbalanced classification[J]. Progress in Artificial Intelligence, 2017, 6(4): 347-354.
- [16] JIANG Kun, LU Jing, XIA Kuiliang. A novel algorithm for imbalance data classification based on genetic algorithm improved SMOTE[J]. Arabian Journal for Science & Engineering, 2016, 41(8): 3255-3266.
- [17] CHEN Hu, ZHANG Yi, ZHANG Weihua, et al. Low-dose CT via convolutional neural network[J]. Biomedical Optics Express, 2017, 8(2): 679-694.