文章编号：2095-2163(2022)03-0092-05 中图分类号：TP391.4 文献标志码：A

基于 CNN 的乳腺癌病理图像分类研究

易才键，陈 俊，王师玮
（福州大学 物理与信息工程学院，福州 350108）

摘 要：乳腺癌已经成为全球第一大癌症，乳腺癌的早期发现及良恶性诊断对于治疗具有重要的意义。针对传统机器学习方法在乳腺癌病理图像分类任务中性能不足和准确率低的问题，本文提出了基于 CNN（卷积神经网络）的乳腺癌病理图像分类模型，将乳腺癌病理图像分为良性与恶性。该模型以 VGG 网络为基础，对网络结构进行调整，在公开的 BreakHis 数据集上实验。针对数据集存在的样本不均衡问题，采用焦点损失函数进行优化，并在网络训练过程结合了迁移学习和数据增强策略。实验结果表明，该模型在 4 种放大倍数下的平均识别率达到 96.96%，分类准确率较先前的模型有了大幅提升，能够为乳腺癌病理图像的分类提供有意义的参考。

关键词：乳腺癌病理图像分类；卷积神经网络；样本不均衡；迁移学习；数据增强

Classification of breast cancer histopathological images based on CNN

YI Caijian, CHEN Jun, WANG Shiwei
(College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China)

【Abstract】 Breast cancer has become the largest cancer over the world. Early detection and diagnosis of benign and malignant breast cancer are of great significance for the treatment. Aim at the problem of insufficient neutrality and low accuracy of traditional machine learning in breast cancer histopathological image classification task, this paper proposes a breast cancer image classification model based on CNN (Convolution Neural Network), which divides breast cancer histopathological images into benign and malignant ones. Based on VGG network, the model adjusts the network structure, experiments on the public data set BreakHis. Aiming at the sample imbalance problem in the data set, we use the focal loss function for optimization, and the training process combines transfer learning and data enhancement strategies. The experimental results show that the average recognition rate of our model under 4 magnifications reaches 96.96%, and the classification accuracy rate has been greatly improved compared with the previous models, which can provide a meaningful reference for the classification of breast cancer histopathological images.

【Key words】 breast cancer histopathological image classification; convolutional neural network; sample imbalance; transfer learning; data enhancement

0 引 言

据世界卫生组织国际癌症研究机构（IARC）2020 年发布的研究数据显示，乳腺癌正式取代肺癌，成为全球第一大癌症[1]。其中，女性癌症患者中乳腺癌的占比最高，远超其他癌症类型。目前对乳腺癌的诊断主要是依靠组织病理学分析，乳腺癌的最终诊断，包括分级和分期，大都由病理学家对组织病理图像进行分析得到，因此这是诊断乳腺癌的金标准[2]。

随着计算机技术的发展，已有许多学者尝试将计算机辅助诊断（CAD）应用在乳腺癌病理图像的自动分类中，并取得了一系列的研究进展。在传统机器学习领域中，自动诊断的方法主要是基于人工的特征提取，结合分类器实现的。Roy 等人[3]设计了特征提取器，提取了纹理和统计特征，将这些特征组合起来，生成一个包含 782 个特征的数据集，通过使用多种分类器进行训练和分类，得到的最优识别率为 92.55%；Spanhol 等人[2]公开了 BreakHis 数据集，并基于此数据集，使用了 6 种不同的特征提取器，并为每个特征提取器结合了 4 种分类器，最终的识别准确率为 80%-85%。但基于人工的特征提取不仅需要耗费大量的时间和精力，还要求特征提取人员具有相应的专业领域知识。此外，特征提取人员的经验和精神状态都会影响到特征提取的质量，严重影响了计算机辅助诊断技术在实际中的应用。

近年来，随着计算机运算能力和人工智能的快速发展，深度学习技术在许多领域得以应用，尤其在图像处理方面取得了很大的进展[4]。利用深度学习技术可以自动的从图像中提取特征，避免了传统

作者简介：易才键（1998-），男，硕士研究生，主要研究方向：医学图像处理、深度学习；陈 俊（1978-），男，硕士，副教授，主要研究方向：物联网通信；王师玮（1998-），女，硕士研究生，主要研究方向：计算机视觉。

收稿日期：2021-11-23

机器学习中人工提取特征的局限性，节省了人力。如今已有很多的学者将深度学习技术应用在乳腺癌诊断中，在一定程度上提高了乳腺癌诊断的准确性。Spanhol 等人[5]在 BreaKHis 数据集上应用 AlexNet 网络，得到的识别率比传统机器学习算法高出 6%；Nawaz 等人[6]使用 DenseNet CNN 模型对乳腺肿瘤的亚型进行预测，准确率达到 95.4%；邹文凯等人[7]对 GoogleNet 中的 Inception 结构进行调整，并采用所有放大倍数统一训练、独立测试的方法，以患者级别作为评价标准，其准确率为 87%－90%。上述方法虽然已经具有一定的准确率，但还需进一步提高识别的准确率和模型的鲁棒性。

针对上述问题，本文以 VGG16 网络为基础，对网络结构进行调整，同时结合数据增强和迁移学习策略，在公开的 BreakHis 数据集上进行训练，训练得到的模型将用作于乳腺癌病理图像的良恶性分类；为解决数据集存在的样本不均衡问题，本文使用焦点损失函数（Focal Loss）作为实验的损失函数，能在一定程度上缓解样本不均衡问题；对 4 种不同放大倍数的图像统一训练，让网络能够学习到更深层次、更复杂的特征，提高模型的鲁棒性，在测试时则对不同放大倍数的图像进行独立测试，更好地模拟实际应用场景中的乳腺癌病理图像分类。

1　本文方法

1.1　卷积神经网络

在 2012 年的 ImageNet 图像分类竞赛上，AlexNet 网络强势夺冠，该网络的分类效果远超当时的其他模型，深度学习技术从此受到广泛的关注。与传统的机器学习方法相比，深度学习的优势在于不需要人为的提取特征，而是依靠神经网络本身去学习样本的特征，提高了特征提取的便利和准确性。

卷积神经网络（Convolutional Neural Network，CNN）作为最常用的深度学习模型之一，在图像处理领域表现优异，本文使用 CNN 来构造图像分类模型。CNN 通常由输入层、卷积层、池化层和全连接层组成，如图 1 所示。将 2D 或 3D 图像输入，由卷积层提取图像的特征，池化层对提取到的特征进行降维、压缩数据和参数的数量。经过一系列的卷积和池化操作，CNN 可以同时学习到数据的低层特征和高层特征，在全连接层得到易被网络区分的特征，便于后续的分类。

相较于传统的神经网络，CNN 具有两大优势：局部连接和权值共享。局部连接是相对于全连接而言的，全连接是指网络中的每个结点都相连，而局部连接则是部分结点相连。实际处理过程中，图像的像素点通常与临近的像素点关联较大，与远处的像素点关联较小，局部连接可以形成具有高区分性的局部特征。权值共享是指使用同一卷积核对整幅图像进行卷积运算，可以减少运算时的参数量，加快运算速度。

图 1　卷积神经网络典型结构

Fig. 1　Typical structure of convolutional neural network

1.2　迁移学习

迁移学习是将从一个任务训练得到的模型移植到其他任务上。目前，迁移学习方法主要有实例迁移、特征迁移、共享参数迁移和关系知识迁移[8]。本文采用参数迁移方法，用已经在其他数据集（源域）上训练好的模型来初始化本文的网络，之后在本文使用的数据集（目标域）上重新训练，对网络的参数进行调整。卷积神经网络在开始训练时，是随机初始化每个参数的，如果此时训练的数据量较小，容易导致模型无法学习到数据的规律，进而影响模型的性能。借助迁移学习技术，可以在一定程度上缩短训练时间，有效的抑制欠拟合和过拟合现象，提高模型的泛化性能。

ImageNet 数据集是一个用于计算机视觉的大型可视化数据集，该数据集有超过 1 000 万幅的自然图像，共 1 000 个类别的手动标注[9]。本文将 ImageNet 数据集作为源域，先将网络模型在该数据集上训练，训练得到的模型参数用作本文数据集训练时网络的初始化。考虑到自然图像和医学图像存在的差异，本文仅将源域的模型参数用作网络初始化，且构造新的全连接层，在 BreakHis 数据集上对网络层的所有参数进行新的训练和调整。

1.3　VGG16 网络

VGG 网络是由牛津大学计算机视觉组（Visual Geometry Group）和 Google DeepMind 公司的研究员一起研发的，该网络取得了 ILSVRC2014 比赛分类项目的第二名，具有良好的特征提取能力[10]。本文以经典的 VGG16 网络为基础，对网络的全连接层进行调整，调整后的网络结构如图 2 所示。

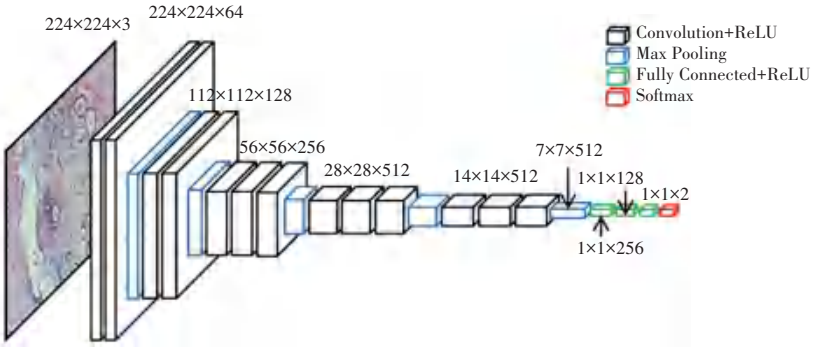

图 2 调整后的 VGG 网络结构

Fig. 2 Adjusted structure of VGG network

网络的输入采用 224×224 的 RGB 彩色图像,共包含 13 个卷积层,5 个最大池化层以及 3 个全连接层。3 个全连接层对应的神经元节点个数调整为 256,128,2,原网络的全连接层神经元节点个数为 4 096,4 096,1 000。调整后的 VGG16 网络具有以下特点:

(1)使用小尺寸的卷积核,以 3×3 大小的卷积核为主。相较于 5×5 或 7×7 的大尺寸卷积核,小尺寸的卷积核不但计算量小,而且更能提取到图像的细节信息;

(2)全连接层神经元的个数较少,由于卷积神经网络的大部分参数量都集中在全连接层,对全连接层的维度进行压缩,可以轻量化模型,降低过拟合的风险。

深度学习算法的缺点是网络训练困难,通常要消耗较多的时间,且利用梯度下降法容易陷入到局部最优解。为了解决这些问题,本文将批量归一化(BN)算法加入到网络中,来缩小每个训练批次间的分布差距,加快网络训练速度。BN 算法的公式(1)和公式(2):

$$\hat{x}^i = \frac{x^i - E[x^i]}{\sqrt{Var(x^i)}} \tag{1}$$

其中,$\hat{x}^i$ 表示第 i 层的所有神经元的输出数据经过归一化之后的结果;x^i 表示第 i 层神经元的输出数据;$E[x^i]$,$\sqrt{Var(x^i)}$ 分别表示第 i 层输出数据的均值和方差。

$$\hat{y}^i = \gamma^i \hat{x}^i + \beta^i \tag{2}$$

其中,$\hat{y}^i$ 为第 i 层的 BN 层输出结果,γ^i,β^i 为 BN 算法引入的可学习参数,通过这两个参数为神经元增加一个线性变换,用于调节神经元的激活值。

综上所述,本文使用网络参数量少,训练速度快,分类性能优秀,用该网络对 BreakHis 乳腺癌组织病理图像数据集进行训练和分类,取得了良好的效果。

2 数据集

2.1 数据集来源

本文采用公开的数据集 BreakHis,该数据集包含来自于 82 位患者的 7 909 幅已标注的乳腺癌组织病理图像,其中良性肿瘤图像 2 480 幅,恶性肿瘤图像 5 429 幅。每幅病理图像均采用 4 种不同的放大倍数(40X、100X、200X、400X),大小均为 700×460 的 R、G、B 三通道图像。BreakHis 数据集的部分图像如图 3 所示;该数据集的具体分布情况见表 1。

(a) 良性肿瘤

(b) 恶性肿瘤

图 3 数据集部分图像

Fig. 3 Partial image of data set

表 1 不同放大倍数的良、恶性肿瘤图像分布

Tab. 1 Image distribution of benign and malignant tumors with different magnification

放大倍数	良性	恶性	总计
40X	625	1 370	1 995
100X	644	1 437	2 081
200X	623	1 390	2 013
400X	588	1 232	1 820
总计	2 480	5 429	7 909

2.2　数据增强

BreakHis 数据集仅有 7 909 幅乳腺癌病理图像，这对于神经网络的训练来说是远远不够的，因此需要利用数据增强来增加训练数据，降低模型过拟合的风险，提高模型的泛化性能。常用的数据增强方法包括：翻转、旋转、裁剪、平移、高斯噪声，模糊等。

本文按照 7∶3 的比例将原数据集划分为训练集和测试集，且仅对训练集的数据进行 6 种方式的数据增强。首先，将训练集数据进行水平翻转、垂直翻转、逆时针旋转 90°、180°、270°共 5 种操作，将数据扩充至原来的 6 倍；再对上述图像按照 0.8 的比例缩放。经过这 6 种方式的变换，训练集数据扩充至原来的 12 倍，其中训练集图像 66 444 张，测试集图像 2 372 张。扩充后的数据集的分布情况见表 2。

表 2　数据增强后的图像分布情况

Tab. 2　Image distribution after data enhancement

放大倍数	训练集	测试集
40X	16 488	621
100X	17 616	613
200X	16 848	609
400X	15 492	529
总计	66 444	2 372

3　实验及结果分析

本文的实验基于开源的深度学习框架 Pytorch，CPU 型号为 IntelCore i7-9000K，内存为 16 GB，显卡型号为 NVIDIA GeForce RTX 2080 Ti。

3.1　训练策略

为了更好地训练分类模型，本文模型的参数通过迁移学习策略进行初始化。在实验过程中，将所有训练数据的尺寸统一为 224×224×3，然后分为小批次训练，每个小批次包含 32 幅图像。采用 Adam 作为本次实验的优化器，在训练过程中自动调整学习率，提高模型分类的准确率，本次 Adam 优化器的参数均采用默认参数，使用 *ReLU* 函数作为激活函数。

3.2　焦点损失函数

通常在分类任务中，会使用交叉熵函数作为损失函数，以二分类为例，二分类交叉熵（Binary CrossEntropy，BCE）的公式（3）为：

$$Loss = -y\log\hat{y} - (1-y)\log(1-\hat{y}) \qquad (3)$$

其中，*Loss* 代表损失值；y 为病理标签，$y=0$ 代表良性，$y=1$ 为恶性；$\hat{y} \in (0,1)$ 为神经网络输出的预测值。

交叉熵函数虽然有着广泛的应用，但也存在明显的缺陷，即交叉熵函数会受到简单易分类样本的影响，导致训练过程中偏离正确的优化方向，对分类效果产生一定的影响。从表 1 可知，BreakHis 数据集存在样本类别不均衡问题，经过数据增强后，训练集中的良、恶性肿瘤图像数量分别为 20 856 和 45 588张，两种类别的图像数量差距明显，故采用焦点损失函数代替二分类交叉熵函数，其公式（4）为：

$$L_{fl} = -\alpha(1-\hat{y})^{\beta}y\log\hat{y} - (1-\alpha)\hat{y}^{\beta}(1-y)\log(1-\hat{y}) \qquad (4)$$

其中，L_{fl} 代表焦点损失函数的损失值，α 与 β 为引入的超参数，通常将 α 设置为 0.25，β 设置为 2。

实验结果表明，引入焦点损失函数能够在一定程度上缓解类别不均衡问题，提高模型的分类效果。

3.3　评价标准

医学图像的分类通常可以从两个方面评价模型的分类性能：患者级别和图像级别。

本文不考虑患者级别，仅从图像级别来计算识别准确率，则图像级别的识别率可表示为公式（5）：

$$Image\ Recognition\ Rate = \frac{N_r}{N_{all}} \qquad (5)$$

其中，N_{all} 代表测试集中病理图像总的数量，N_r 代表被正确分类的图像数量。

3.4　实验对比分析

3.4.1　不同损失函数下的准确率对比

本次实验将焦点损失函数（Focal Loss）与分类任务中应用广泛的二分类交叉熵（BCE）对比，分别使用这两种函数作为训练过程中的损失函数，实验结果见表 3。从表 3 可以看出：

（1）Focal Loss 作为损失函数时，良恶性肿瘤的分类准确率仅相差 0.29%；而使用 BCE 的情况下，相差 3.44%，此时模型对于较多样本（恶性肿瘤）产生了倾向性，不利于对肿瘤的诊断；

（2）使用 Focal Loss 时，虽然对恶性肿瘤的分类准确率略低于使用 BCE 的情况，但对于良性肿瘤的分类准确率却得到了很大的提升，这样的模型更接近实际生活，具有更强的鲁棒性；

（3）模型的平均准确率有所提高。

表 3　不同损失函数下的准确率对比

Tab. 3　Comparison of accuracy with different loss functions

损失函数	良性肿瘤	恶性肿瘤	平均准确率
BCE	93.80	97.24	96.16
Focal Loss	96.77	97.06	96.96

3.4.2 不同训练策略下的准确率对比

使用不同的训练策略,共进行 4 次实验,实验均采用 Focal Loss 作为损失函数。这 4 种策略分别是数据增强结合迁移学习策略、数据增强策略、迁移学习策略、无数据增强和迁移学习策略,结果为网络迭代10 000次过程中的最佳模型在测试集上的准确率,如图 4 所示。

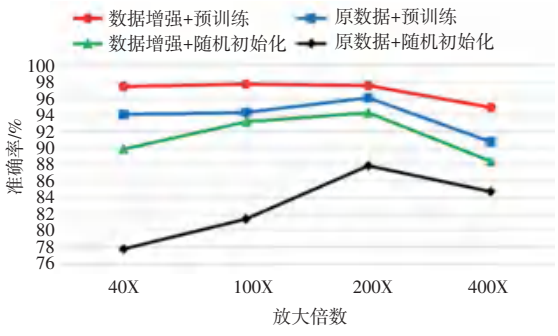

图 4 4 种训练策略下的准确率

Fig. 4 Accuracy under four training strategies

由图 4 可知,采用迁移学习策略,无论是否进行数据增强,准确率都得到了大幅度的提升(图 4 中红色和蓝色曲线对比),证实了迁移学习策略的有效性;采用数据增强策略后,无论是否使用迁移学习对网络进行初始化,训练的准确率都得到了一些提升(见图 4 中红色和绿色曲线对比),证实了数据增强策略的有效性。实验表明,本文采用有效的训练策略防止了训练过程中过拟合的现象,并大大的提高了模型的泛化能力,在 BreakHis 数据集上的识别率为 94%-98%。

3.4.3 与其他的分类方法对比

为了更好的评价本文的模型,本文选择与应用在同一数据集 BreakHis 上的其他分类方法进行对比,这些方法采用与本文相同的评价标准,即以图像级别的识别率作为评价标准,见表 4。通过与其他分类方法的对比可知,本文方法在 4 种不同放大倍数下的识别准确率均高于其他的分类方法,表明了本文训练策略的有效性及本文深度学习模型的鲁棒性。

表 4 不同放大倍数下各方法识别准确率的对比

Tab. 4 Comparison of recognition accuracy of various methods with different magnifications

方法	40X	100X	200X	400X
AlexNet 网络	85.60	83.50	83.10	80.80
VGG 网络+SVM	87.00	86.20	85.20	82.90
GooleNet+迁移学习	90.89	90.99	91.00	90.97
本文方法	**97.42**	**97.72**	**97.54**	**94.90**

4 结束语

为解决传统机器学习在病理图像分类任务中存在的不足,提高乳腺癌病理图像的分类准确率,本文提出了基于 CNN 的乳腺癌病理图像分类模型。在公开的 BreakHis 数据集上进行训练与参数优化,最终在 4 种放大倍数下的平均识别率达到 96.96%,其中 40X、100X 和 200X 倍数下的识别率均超过 97%,展现出了优秀的分类能力;为解决医学图像数据集较少的问题,本文采用迁移学习和数据增强策略,利用迁移学习初始化网络,同时将数据集扩充至原有的 12 倍,避免了过拟合现象的发生;为解决 BreakHis 数据集存在的类别不均衡问题,本文采用焦点损失函数代替传统的交叉熵函数。通过多个对比实验,验证了本文模型的优异性和训练策略的有效性,能够为早期发现和诊断乳腺癌提供有力指导。

参考文献

[1] WILD C P, WEIDERPASS E, STEWART B W. World Cancer Report: Cancer research for cancer prevention [M]. Lyon: International Agency for Research on Cancer, 2020.

[2] Spanhol Fabio A, Oliveira Luiz S, Petitjean Caroline, et al. A Dataset for Breast Cancer Histopathological Image Classification. [J]. IEEE transactions on bio-medical engineering, 2016, 63(7): 1455-1462.

[3] Roy Soumya Deep, Das Soham, KarDevroop, et al. Computer Aided Breast Cancer Detection Using Ensembling of Texture and Statistical Image Features. [J]. Sensors (Basel, Switzerland), 2021, 21(11): 3628.

[4] YANN LeCun, Yoshua Bengio, Geoffrey Hinton. Deep learning [J]. Nature: International weekly journal of science, 2015, 521(7553): 436-444.

[5] SPANHOL F A, OLIVEIRA L S, PETITJEAN C, et al. Breast cancer histopathological image classification using Convolutional Neural Networks [C] //2016 International Joint Conference on Neural Networks (IJCNN), 2016: 2560-2567.

[6] NAWAZ M, ADEL A, HASSAN T. Soliman. Multi-Class Breast Cancer Classification using Deep Learning Convolutional Neural Network[J]. International Journal of Advanced Computer Science and Applications, 2018, 9(6):316-332.

[7] 邹文凯, 陆慧娟, 叶敏超, 等. 基于卷积神经网络的乳腺癌组织病理图像分类[J]. 计算机工程与设计, 2020, 41(6): 1749-1754.

[8] Karl Weiss,Taghi M. Khoshgoftaar, DingDing Wang. A survey of transfer learning[J]. Journal of Big Data, 2016, 3(1): 9.

[9] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet Large Scale Visual Recognition Challenge [J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[10] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv: 1409.1556, 2014.