基于深度学习的癌细胞图像识别技术

陈书华

（昆明市第一中学，云南昆明，650031）

摘要：当今时代，癌症是人类可怕的敌人。人们很难发现它的存在，现在几乎没有任何办法能治愈它，而且治疗费用昂贵。为解决这一问题，本文提出了“基于深度学习的癌细胞图像识别技术”。本文通过在网上搜集大量数据集，设计好深度神经网络(DNN)模型和卷积神经网络(CNN)模型并调整它们的参数，对含有癌细胞和正常细胞的数据集进行训练。为求证DNN与CNN哪个模型准确率更高及扩大数据集是否能提高模型准确率，我们用含有2000张图片和含有4000张图片的两组数据集分别对两个模型进行训练。结果是DNN模型的准确率分别是72%和73%，准确率提高了1%。CNN模型的准确率分别是75%和78%，准确率提高了3%。我们看到，CNN模型识别癌细胞的准确率更高，并且扩大数据集能提高模型的准确率。

关键词：癌细胞；深度学习；卷积神经网络；深度神经网络；图像识别

DOI:10.16589/j.cnki.cn11-3571/tn.2020.20.013

0 引言

当今时代，癌症是我们的大敌。2019 年 1 月，最新报告显示，我国癌症发病率每年约为 392.9 万人，死亡人数达 233.8 万人 [1]。而在世界上，每年则有近 800 万人死于癌症 [2]。有很多不好的行为习惯都会提高癌症的发病率，如吸烟 [3]。世界吸烟人口高达 11 亿，中国就有近 3.5 亿人。《中华肿瘤杂志》显示，每年癌症所致的医疗花费超过 2200 亿元。而现阶段检查癌症，大多是需要在有意识地进行身体全面检查，通过化验，核磁共振，CT 等方式才能查出癌症。这样就造成人们即使得了癌症但是难以发现的情况。

近年来 , 随着深度神经网络的发展，深度神经网络开始被应用到图像识别、自然语言处理等领域。而将神经网络用于识别癌细胞的前人也有不少。厉谨 , 康涛等人提出通过图像分割技术来进行癌细胞识别 [4]。这种方法是在分割后的图像中，通过观察细胞核边缘光滑程度，形状来区分癌细胞与正常细胞。这需要将图片分割处理为细胞核，过程较为复杂。细胞癌变最大的特征是细胞外形发生改变，而不仅仅是细胞核，而这种方法仅通过细胞核进行确定，形式比较单一，有一定局限性。而本文的神经网络模型，只需要病人的一张图片，便能初步确定其是否患有癌症，并判断是否需要进行治疗，为癌症的早期发现和尽早治疗提供了可能。

为了避免判断条件单一，操作复杂，我们采用整体判断的方法来识别癌细胞。我们搜集了大量数据集，这些数据集都是未经处理的癌细胞和正常细胞原图。建立 DNN 和 CNN 两种模型。DNN 模型由一个输入层，逐次连接四个隐藏层，再连接一个输出层，得到分类结果。整个 DNN 网络模型是全链接结构的。CNN 模型先有一个输入层，然后进行先卷积后池化的过程三次，然后将得到的张量转化为向量，接上一个全链接神经网络，得到最终的分类结果。为探究 DNN 与 CNN 哪个模型准确率更高，我们用含有 2000 张图片和 4000 张图片的两组数据集分别对两个模型进行训练。结果是 DNN 模型的准确率分别是 72% 和 73%，CNN 模型的准确率分别是 75% 和 78%。当数据集从 2000 张图片增加到 4000 张图片以后，DNN 模型识别癌细胞的准确率提高了 1%，CNN 模型识别癌细胞的准确率提高了 3%。由此我们得到了 CNN 模型识别癌细胞的准确率更高，并且扩大数据集能提高模型的准确率的结论。

1 方法

■ 1.1 识别癌细胞的深度神经网络

深度神经网络是深度学习的基础。深度神经网络英文名为 Deep Neural Networks，简称 DNN。DNN 一般由输入层，数个隐藏层，一个输出层构成，是深度学习的主要算法之一。DNN 模型的最大优点，在于它能具有足够多的隐藏层和权值。理论上，只要拥有足够多的隐藏层和权值，DNN 就能够模仿出任何方程。也就是说，DNN 有可能很好地还原人的思考方式。但 DNN 模型也有明显的缺点，就是计算太过复杂。在 DNN 中，BP 算法是反向传播，使模型更新参数，达到学习目的的学习算法。

BP 算法全称为反向传播算法，即 Backpropagation 算法。这是把误差信号按原来正向传播的通路反向传回，并对每个隐藏层的各个神经元的权系数进行修改，以使误差信号趋向最小的算法。我们使用了随机梯度下降的方法来让模型进行学习。这个方法就是，我们设 W 为第一隐藏层到第二隐藏层的一个参数，首先对 w 进行赋值，可以是随机的，也可以为全零向量，然后求 w 的变化值 Δw，使得目标函数 J(w) 按梯度下降的方向进行减少。梯度方向由 J(w) 对 w 的偏导数确定。然后更新 w 的参数，计算损失函数：

$$w = w - 学习率 \times \Delta w \tag{1}$$

$$J(w) = \frac{1}{2}(d-p)^2 \tag{2}$$

公式 (1) 中，Δw 为 w 的变化值。公式 (2) 中，d 为真实值，p 为预测值，J(w) 是定义目标函数。BP 算法的目的就是使 J(w) 最小化。

如图 1 所示，是我们用于训练的 DNN 模型图。此模型

为全连接神经网络。我们将分别包含 2000 张图片和 4000 张图片的两个数据集输入，训练过程中，每张图片都裁剪为 64*64*3 的张量，再转化为向量，连上有 218 个单元的第一隐藏层，用激活函数 Relu 进行激活。然后连上有 126 个单元的第二隐藏层，并用激活函数 Relu 进行激活。然后再连上有 32 个单元的第三隐藏层，继续用激活函数 Relu 进行激活。接着再连上有 8 个单元的第四隐藏层，用激活函数 Sigmoid 进行激活。最后输出时做二分类，识别是否是癌细胞。这样就完成了一张图片的训练。我们每轮要训练 2000 或 4000 张图片，这样训练 500 轮。最终 DNN 模型的损失有 0.7，准确率有 73%。

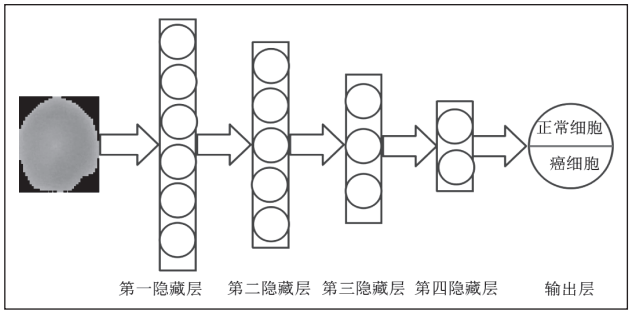

图 1　我们的 DNN 模型图

激活函数 Relu 计算公式为：

$$f(x)=\max(0,x) \tag{3}$$

x 为上一层输出结果。

Sigmoid 激活函数数学表达式为：

$$f(x)=1/(1+e^{-x}) \tag{4}$$

公式 (3) 中，x 为上一层输出结果。

损失函数（LOSS）是计算预测值和真实值之间误差的函数，是我们评估模型是否准确的重要依据。LOSS 越小，说明输出量和实际量相差很小，模型准确。LOSS 很大，说明模型没有很好地学习，效果欠佳。而损失函数计算出来的值即为损失。模型识别正确的时候，损失小；模型识别错误的时候，损失大。

本文使用交叉熵来计算损失函数：

$$LOSS=J(w,b)=-\frac{1}{m}\sum_{i=1}^{m}[y\times\log\bar{y}+(1-y)\log(1-\bar{y})] \tag{5}$$

$$\bar{y}=f(w\times x+b) \tag{6}$$

公式 (5) 中，w 为参数，b 为偏置，f 为激活函数 Sigmoid，m 为样本数量，y 为真实值，$\bar{y}$ 为预测值。

■ 1.2 识别癌细胞的卷积神经网络

卷积神经网络，英文名为 Convolutional Neural Networks，简称 CNN[5]。CNN 是一类包含卷积计算且具有深度结构的神经网络，是一种深度学习的算法。CNN 的特点在于它是局部链接而不是全链接，还能进行权值共享。使用局部链接是因为图像中的像素并不是孤立存在的，每一个像素与它周围的像素都有着相互关联，而并不是与整幅图像的像素点相关。这样对于图像特征提取有很好的效果。权值共享的优点在于在对图像进行卷积操作时，并不需要对每一个卷积核新建立参数，滑动过程中的卷积核参数都是共享的。与 DNN 相比较，CNN 能很大程度上减少计算量。

卷积神经网络包括输入层，卷积层，激活层，池化层，全连接层，归一化指数层。输入层是将图片输入的一层；卷积层是用卷积核提取特征的层；激活层是用激活函数将数据从线性转化为非线性的层；池化层是降低模型的参数量的层，本文使用的均是最大池化层；全连接层是进行全链接的层；归一化指数层是将数据做二分类的层。我们输入数据集，通过卷积计算提取图片特征，以此来让模型学习癌细胞和正常细胞的特点并进行识别。

如图 2 所示，是我们的卷积神经网络模型图。

我们先随机对图片进行裁剪，得到 32*32*3 的张量，用 16 个步长为 1，大小为 7*7 的卷积核，进行卷积，得到 26*26*16 的张量，再用 2*2 的最大池化层进行池化，池化是分别作用于每个输入的特征并减小其大小，降低模型的参数量并一定程度上抑制过拟合的一种计算方式，然后变为 13*13*16 的张量，并用 Relu 激活函数进行激活。然后再用 32 个 6*6 的卷积核进行第二次卷积，得到 8*8*32 的张量，用 Relu 激活函数进行激活，再用同样的池化层进行池化，

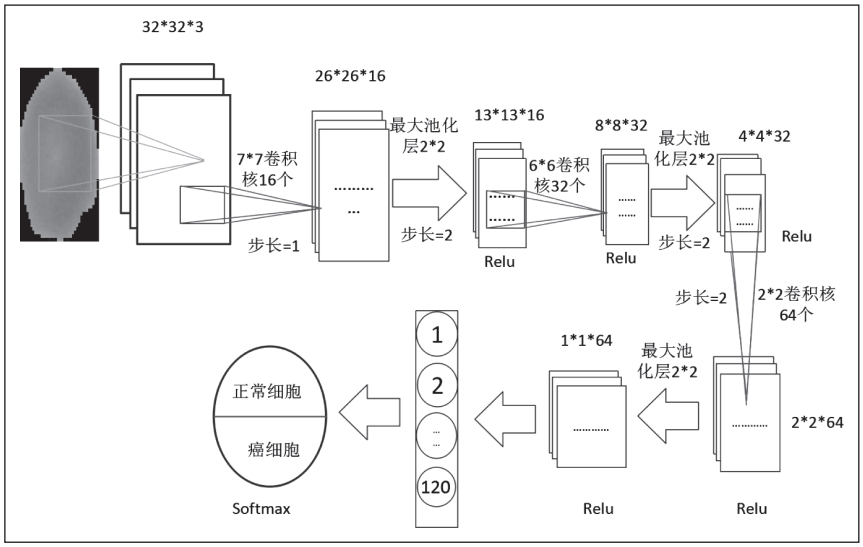

图 2　我们的卷积神经网络模型图

得到 4*4*32 的张量，再用 Relu 激活函数进行激活。我们再用 64 个 2*2 的卷积核进行第三次卷积，得到 2*2*64 的张量，用 Relu 激活函数进行激活，再用同样的池化层进行池化，得到 1*1*64 的张量，再用 Relu 激活函数进行激活。然后进行归一化，即转化为一个向量。接上一个有 120 个单元的全连接神经网络，我们分别用包含 2000 张图片的数据集和包含 4000 张图片的数据集进行训练，最后得到输出结果。CNN 模型最终的损失有 0.51，准确率有 78%。

2 实验

■ 2.1 实验设置

我们用数据集进行试验。这个数据集作者是 Tschandl P，其中包含了 13779 张正常细胞图片，如图 3，与 13779 张癌细胞图片，如图 4。图片大小在 100*100*3 到 120*120*3 左右。我们从数据集中挑选了 4000 张图片进行实验。我们总共进行了两次试验。第一次我们用了 2000 张图片进行试验。癌细胞与正常细胞各 1000 张。其中 900 张为训练集，100 张为测试集。第二次我们用了 4000 张图片进行试验，癌细胞与正常细胞各 2000 张。其中 1800 张为训练集，200 张为测试集。

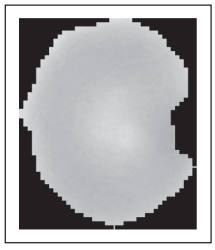

图 3　正常细胞

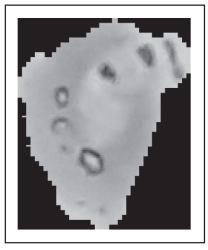

图 4　癌细胞

■ 2.2 实验结果

（1）DNN 与 CNN 癌细胞识别二分类实验

我们这里来说明两种模型的试验参数，准确率的计算方法和训练结果。

在 DNN 模型的训练中，我们每次以 4 张图片为一组进行学习。我们将学习率设置为 0.001，迭代次数即 epoch 设为 500 次，裁剪大小为 32*32*3。

在 CNN 模型的训练中，我们以 32 张图片为一组进行学习。学习率为 0.001，迭代次数为 500 次，图片翻转概率为 0.5, 图片大小缩小为 64*64*3。

准确率是判断模型识别癌细胞准确程度的具体数值。

$$准确率 = \frac{分类正确的样本数}{测试样本总数} \tag{7}$$

训练结果如下。共 2000 张图片的实验结果，四幅图横坐标均为迭代次数，即训练次数。图 5、图 6 的纵坐标为损失，损失定义如公式 5 所示；图 7、图 8 的纵坐标为准确率。

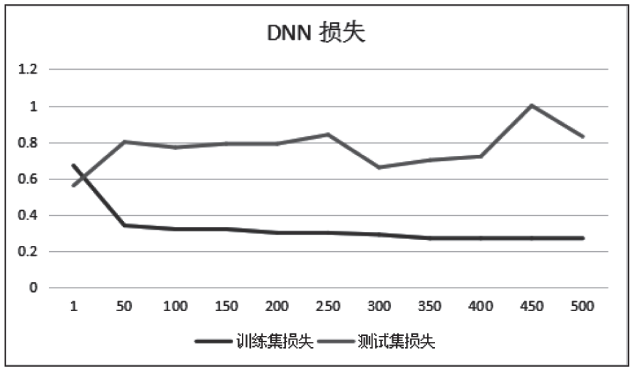

图 5　DNN 训练集和测试集的损失

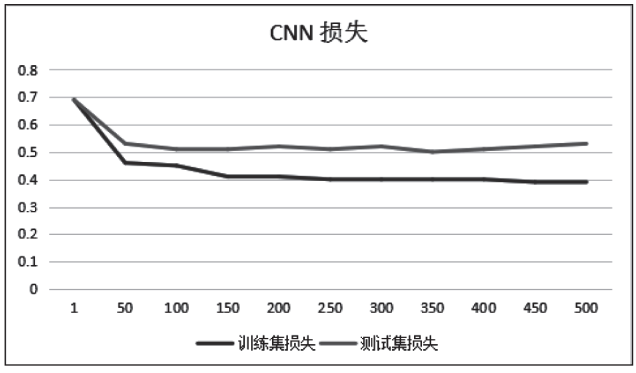

图 6　CNN 训练集和测试集的损失

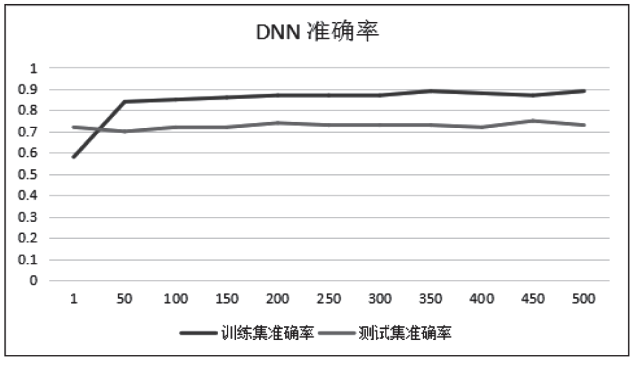

图 7　DNN 训练集和测试集的准确率

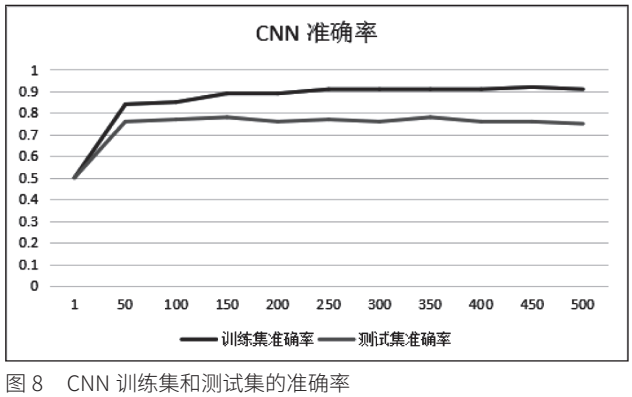

图 8　CNN 训练集和测试集的准确率

从损失来看，在 DNN 中，训练集和测试集一开始的损失都在 0.6 左右，训练集的损失最终降到了 0.27 左右。而

测试集最终反而在 0.8 上下波动。这是因为学习率设计过大，学习没有循序渐进，一下就学成，导致模型对于细胞的区分不是很准确，所以出现了较大的波动。在 CNN 中，一开始训练集和测试集的损失都在 0.7 左右，训练集最终降到了 0.4 左右。而在测试集降到了 0.5 左右。这是由于训练集和测试集的图片之间有差异的原因，模型没有完全学会如何区分癌细胞与正常细胞，出现了过拟合，所以在测试集中的损失较多。CNN 模型在测试集中，损失下降情况很理想，但 DNN 模型在测试集中的表现不是很好。这是因为 CNN 模型能更好地提取数据集特征进行学习，能抓住重点的原因。

从准确率来看，在 DNN 中，训练集的准确率由 50% 提高到了 86%。测试集一开始表现很好，准确率有 70%，但在后面的学习过程中，没有太大进步，准确率也一直保持在 73% 左右。这主要是因为数据集比较少，导致学习效果不佳，训练不到位。在 CNN 中，训练集和测试集的准确率开始在 50% 左右，随着训练的增多，训练集准确率达到了 90%，测试集准确率达到了 78% 左右。这是因为模型在训练集的拟合程度较高，而在测试集的拟合程度稍低的原因。我们可以看出，最终 CNN 模型的准确率比 DNN 模型高 5%，说明 CNN 模型学习效果更佳。这是因为 CNN 模型具有卷积神经网络提取特征能力强的特点，同时包含了全链接全面学习的优势，比起 DNN 单纯地使用全链接神经网络更有优势。

（2）数据集大小对癌细胞识别的影响

在 DNN 和 CNN 两个模型中，对于训练集，数据集大小影响程度不大，损失和准确率都十分接近。而对于测试集，在数据集较大的情况下，训练的准确率有提升，损失也有减小。

表1 数据集大小对DNN和CNN的测试集损失和准确率的影响

训练集大小	模型	损失	准确率
2000	DNN	0.8	0.72
4000	DNN	0.7	0.73
2000	CNN	0.5	0.75
4000	CNN	0.51	0.78

我们可以看到，数据集由 2000 增加到 4000 时，在 DNN 中准确率有 1% 的提升，而损失减少了 0.1。在 CNN 中的准确率有 3% 的提升，损失较为接近。这可以说明，随着数据集的增大，模型学习图片的样式、形态、环境都有所增加。模型能更好地学习癌细胞和正常细胞各自的特点，减小了细胞样式、形态、环境等外界因素对识别造成的影响，所以对癌细胞识别的准确率会有一定的提升，而 CNN 提升多于 DNN 是因为数据集中，癌细胞也有部分与正常细胞相似，而且 CNN 具有提取图片特征的能力，更适合此类有特征的数据集的学习。

3 总结与展望

当今社会，癌症的隐蔽性和难以治愈性给人们宝贵的生命造成极大威胁，使得癌症的早期诊断和治疗成为一个重要的亟待研究解决的问题。本文主要通过训练 DNN 和 CNN 两种模型，并且在训练中不断调整学习率、迭代次数等参数，以实现两种模型更好的识别效果。最终得出了识别癌细胞的有效方法。DNN 识别癌细胞的准确率达到 73%，损失为 0.7。CNN 识别癌细胞的准确率达到 78%，损失为 0.51。CNN 在癌细胞的识别上更有优势。而通过扩大数据集，DNN 和 CNN 的准确率分别有 1% 和 3% 的提升，说明了扩大数据集对提高模型的准确率有好处。

由于客观因素，癌细胞涉及个人隐私问题，每个癌症患者被拍摄的癌细胞图片都受到了严格的保护，所以我们搜集到的数据集比较单一，目前我们只找到了一种癌细胞，所以训练得不够全面。如果今后我们能找到有更多癌细胞种类的数据集，还可用以上方法，训练出更为全面的神经网络模型，识别更多种类的癌细胞。我们也可以再更改参数，使用控制变量的方法，细分实验组别，逐一进行分析，看每个参数对模型的影响，再做更多的试验，找到识别癌细胞正确率最高的办法。这样便于这项技术从实验进一步走向社会，为社会提供服务，成为给人民带来幸福的实用的技术。

参考文献

- [1] 郑荣寿 , 孙可欣 , 张思维 , 等 .2015 年中国恶性肿瘤流行情况分析 [J] . 中华肿瘤杂志 ,2019,41(1): 19-28. DOI: 10.3760/cma.j.issn.0 253-3766.2019.01.005.
- [2] 潘钢火 , 鲁晓明 . 中国癌症分布以及影响因素的研究进展 [J]. 首都师范大学学报（自然科学版）,2016,37(1):56-60. DOI:10.3969/j.issn.1004-9398.2016.01.012.
- [3]50th Anniversary Report: Even More Known About Smoking, Cancer Connections，OncLive，Tuesday, June 24,2014.
- [4] 厉谨 , 康涛 , 李力 . 基于 PCNN 分割的癌细胞图像识别方法研究 [J]. 咸阳师范学院学报 ,2010,25(02):49-52.
- [5]LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
- [6] https://www.kaggle.com/kmader/skin-cancer-mnist-ham10000#hmnist_8_8_RGB.csv