基于卷积神经网络算法的X射线晶体衍射实验数据筛选

惠 子[1] 余 立[2,3] 周 欢[4] 唐 琳[1] 何建华[1]

1（武汉大学 高等研究院 武汉 430072）

2（中国科学院上海应用物理研究所 上海 201800）

3（中国科学院大学 北京 100049）

4（中国科学院上海高等研究院 上海 201204）

摘要 X射线串行晶体学作为一种解析蛋白质晶体结构的新方法，因为拥有室温采集、辐射损伤低、时间分辨等优势而得到迅速的发展。利用串行晶体学方法解析蛋白质结构需要在整合大量晶体衍射图的基础上筛选出有效的衍射数据，然而常规的数据筛选方法在处理衍射图时存在准确度不高且效率低的问题。基于卷积神经网络的数据筛选方法具有流程自动化的优势，并且已经被证明相对于传统的“找点法”具有更高的分类准确度。因此在比较5种不同卷积神经网络筛选晶体学衍射图的准确度和效率的基础上，选择并构建一个准确率高且运行速率快的卷积神经网络数据筛选工具，用于不同蛋白质晶体样品衍射图的筛选。结果显示：MobileNets不仅具有ResNet、GoogleNet-Inception等大型网络相似的准确度，而且运行速率更快，为串行晶体学实验提供了一个有效便捷的数据筛选工具。

关键词 串行晶体学，卷积神经网络，机器学习，MobileNets

中图分类号 TL99

DOI: 10.11889/j.0253-3219.2023.hjs.46.030101

X-ray crystallography experimental data screening based on convolutional neural network algorithms

HUI Zi[1] YU Li[2,3] ZHOU Huan[4] TANG Lin[1] HE Jianhua[1]

1(*The Institute for Advanced Studies*, *Wuhan University*, *Wuhan 430072*, *China*)

2(*Shanghai Institute of Applied Physics*, *Chinese Academy of Sciences*, *Shanghai 201800*, *China*)

3(*University of Chinese Academy of Sciences*, *Beijing 100049*, *China*)

4(*Shanghai Advanced Research Institute*, *Chinese Academy of Sciences*, *Shanghai 201204*, *China*)

Abstract **[Background]** Serial X-ray crystallography has developed rapidly due to its advantages of data collection at room temperature, low radiation damage and time resolution. To solve protein structures by using the serial X-ray crystallography, a large amount of produced diffraction data needs to be screened for finding the effective diffraction patterns. The use of convolutional neural networks (CNN) can not only automate the data screening process, but also improve the accuracy of data classification comparing with the traditional "point finding

武汉大学人才科研启动项目(No.420541310049)资助

第一作者：惠子，女，1997年出生，2019年毕业于西安交通大学，现为硕士研究生，研究领域为机器学习、串行晶体学

通信作者：何建华，E-mail：hejianhua@whu.edu.cn

收稿日期：2022-10-28，修回日期：2023-02-06

Supported by the Talent Research Start-up Program of Wuhan University (No.420541310049)

First author: HUI Zi, female, born in 1997, graduated from Xi'an Jiao Tong University in 2019, master student, focusing on machine learning and serial crystallography

Corresponding author: HE Jianhua, E-mail: hejianhua@whu.edu.cn

Received date: 2022-10-28, revised date: 2023-02-06

method". **[Purpose]** This study aims to explore five types of popular convolutional neural networks, i.e., AlexNet, GoogleNet, MobileNets, Vgg16, ResNet, for screening crystallographic diffraction patterns, and compare the accuracy and efficiency of them to build up a fast and accurate convolutional neural network tool for screening the diffraction patterns of different protein crystal samples. **[Methods]** Firstly, the primitive data for model training extracted from the coherent X-ray image database, collected by Linac Coherent Light Source (LCLS) and Spring-8 Angstrom Compact free electron laser (SACLA), were pre-processed by gray level equalization and data enhancement. The deep learning models were trained by iteration of the preprocessed data. Then, the selected convolutional neural network through the comparison of accuracy and efficiency was used to process further the experimental data of protein crystals diffractions. **[Results]** The results show that MobileNets not only has the accuracy similar to large networks such as ResNet, GoogleNet-Inception, but also runs faster. **[Conclusions]** MobileNets provides an effective and convenient screening tool for serial X-ray crystallography experimental data.

Key words Serial X-ray crystallography, Convolutional neural network, Machine learning, MobileNets

同步辐射因其具有高亮度、高准直性及波长可调等特点，对尺寸小、衍射强度低的生物大分子晶体也可以得到清晰的衍射图，为蛋白质晶体结构测定和分子动力学实验提供了强有力的工具，促进了结构生物学的发展。目前，蛋白质数据库所测定的蛋白质结构中有90%是利用X射线晶体衍射方法解析的。然而，基于同步辐射的X射线晶体学方法存在辐射损伤，并且会受到蛋白质晶体尺寸的限制，Chapman等[1]首次利用基于X射线自由电子激光（X-ray Free Electron Laser，XFEL）的串行飞秒晶体学（Serial Femtosecond crystallography，SFX）方法实现了Photosystem I蛋白质复合物200 nm~2 μm的微小晶体在室温下的衍射数据收集，这一方法不仅降低了对于晶体尺寸的要求，而且在辐射损伤之前收集到足够信号避免了辐射损伤的问题。串行晶体学[2]作为一种解析蛋白质晶体结构的新方法，相比于传统的蛋白质晶体结构测定方法，拥有室温采集、辐射损伤低、时间分辨等优势。目前，同步辐射串行晶体学[3]的发展主要面临两点挑战：一是X射线对高速飞行晶体的命中率比较低，传统固定靶式的上样方法与新的约束流动式的上样方法应用于串行晶体学实验都需要实践与优化；二是实验会产生大量的衍射数据，传统的数据筛选方法难以准确、高效地从巨量衍射图中选取出有效的衍射图用于蛋白质结构的解析。Stellato等[4]采用毛细血管上样方式在室温下解析了溶菌酶的结构，而Przemyslaw Nogly[5]采用脂立方相喷射（Lipidic Cubic Phase jet）上样方式在室温下解析了细菌视紫红质的结构。两位学者在实验过程中均得到了上百万张衍射图，而利用传统的“找点法”经过CrystFEL[6]软件处理后得到的有效命中衍射图仅有上千张。由此可见，串行晶体学实验会给传统的数据处理方法带来大量无效的衍射图，极大地影响实验效率。因此，同步辐射串行晶体学需要一种高效准确的数据预处理方法。

卷积神经网络（Convolutional Neural Network，CNN）是一种包含卷积计算且具有深度结构的前馈神经网络，使用CNN提取图像特征并利用结构特征进行图像分类、图像识别是批量处理大量图像数据的有效方法之一。Ke等[7]实现了基于CNN串行晶体学数据筛选，且证明了CNN预测的准确度高于传统的“找点算法”。Zimmermann等[8]使用深度神经网络作为氦纳米液滴广角衍射图像的特征提取器，证明了深度神经网络在复杂衍射图形分类中的优势。以上研究都证明了CNN应用于晶体学实验数据筛选的可行性，但并未解决应用CNN进行筛选的效率问题。

MobileNets[9]是基于流线型的深度可分离深层CNN架构，与当前比较流行的AlexNet[10]、GoogleNet[11−12]、Vgg16[13]、ResNet[14]等网络相比，在保证准确率的前提下通过网络结构的优化提高了运行速率。因此，我们尝试基于MobileNets实现串行晶体学衍射图的筛选，同时将结果与使用AlexNet、GoogleNet、Vgg16、ResNet等网络的分类结果进行比较，尝试证明MobileNets在准确率与运行速率上都有不错的表现。

1 数据

为了构建可以用于蛋白质结构解析的准确且高效的CNN模型，本文选取X射线成像数据库[15]的衍射数据进行模型的构建（数据源自Linac Coherent Light Source（LCLS）和SPring-8 Angstrom Compact free electron Laser（SACLA）装置的串行飞秒晶体学实验，网址：http://cxidb.org/id-76.html，登录号为76[16]），数据具体情况如表1所示。

表1 实验数据
Table 1 Experimental data

数据 Dataset	蛋白质 Protein	入射能量 Incident energy / keV	仪器 Instrument	探测器 Detector
LN83	氢化酶蛋白质晶体 Hydrogenase	9.498	MFX	Rayonix
LN84	光系统 II Photosystem II	9.516	MFX	Rayonix
LO19	辛环素 Cyclophilin A	9.442	MFX	Rayonix
L498	嗜热菌蛋白酶 Thermolysin	9.773	CXI	CSPAD

1.1 数据分类

衍射图像中低于水环分辨率的布拉格衍射点数量的多少可以直接表示衍射图是否命中，因此，基于衍射图像中低于水环分辨率的布拉格衍射点数量[7]，将获得数据分为三类：“命中”“未命中”和“也许命中”。“命中”表示具有明显的衍射图样，包含多个有效衍射点；“未命中”表示不具有衍射图特征，没有有效衍射点，考虑到衍射点识别存在误差，设置衍射点最低阈值作为判据；“也许命中”介于二者之间，其主要目的是避免有效数据因“命中”判断误差而丢失。分类标准如表2所示。

表2 数据分类
Table 2 Data classification

数据类型 Data type	布拉格点的数量 Number of Bragg points	有效信息含量 Effective information content
命中 Hit	$X\geqslant 10$	较多有效信息 More valid information
也许命中 Maybe	$10>X\geqslant 4$	较少有效信息 Less valid information
未命中 Miss	$X\leqslant 3$	缺失有效信息 Loss valid information

注：表中采用了与文献[7]相同的分类阈值设定以便于结果比较，对于不同的数据类型可通过大量实践进一步优化阈值设定以实现更高的分类准确度

Notes: The same classification threshold setting as Ref.[7] is adopted in the table to facilitate the comparison of results. For different data types, the threshold setting can be further optimized through practice to achieve higher classification accuracy.

1.2 数据预处理

本文主要采用翻转、旋转、裁剪、平移等方法增强数据来避免过拟合的问题，同时采用灰度值均衡化的方法增强衍射点以达到提高CNN准确度的目的。

1.2.1 灰度值均衡化

现代探测器由于记录图像的像素强度动态范围大、代表不良伪影的像素强度高而很难直接从原始图像强度中识别布拉格点。灰度值均衡化[17]是以累积分布函数变换为基础的直方图修正方法，它可以产生灰度级分布概率均匀的图像，使图像具有更大的信息量。对于图像中的每一个像素点，灰度图均衡化的基本公式为：

$$f(a)=\frac{H[a](K-1)}{MN} \tag{1}$$

式中：$H[a]$表示累积直方图；K表示亮度的取值，这里K的取值是0~255，MN表示图像的像素总数。图1为LN83衍射图数据通过灰度值均衡化前后的对比，4种蛋白质晶体在灰度值均衡化后的衍射图都可以更加清晰地分辨出布拉格衍射点，如图2所示。

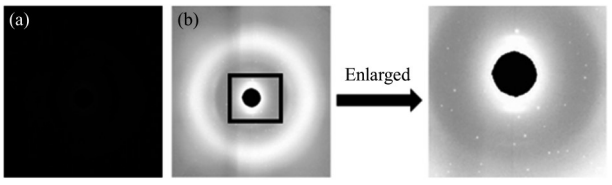

图1 LN83衍射图灰度值均衡前(a)与灰度值均衡化后(b)的对比图
Fig.1 Comparison of LN83 diffraction pattern before (a) and after (b) gray value equalization

1.2.2 数据增强

将裁剪图像的中心与原始图像的中心重合，并结合所使用的神经网络决定裁剪大小进行图像裁剪（例如：Vgg16需要将图像裁剪为224×224像素）。此外，还通过将图像在水平或垂直方向移动几个像素，并进行随机的镜像翻转、旋转以及缩放来生成更多的训练样本，如图3所示。最终使用2 000个原始图像获得42 000个训练样本。

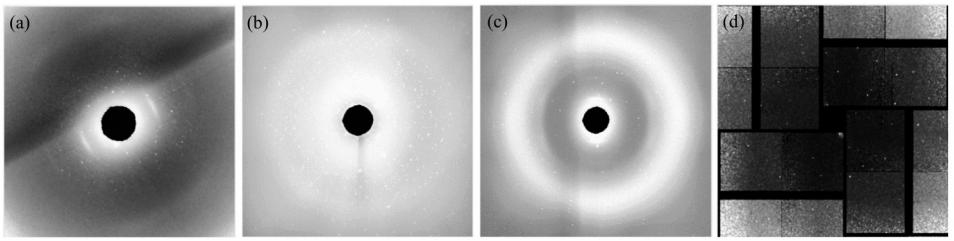

图2 蛋白质晶体灰度值均衡化后的衍射图
(a) LN84-0015[命中],(b) LO19-0134[命中],(c) LN83-0002[命中],(d) L498-0032[命中]
Fig.2 Diffraction pattern of protein crystal after gray value equalization
(a) LN84-0015 [Hit], (b) LO19-0134 [Hit], (c) LN83-0002 [Hit], (d) L498-0032 [Hit]

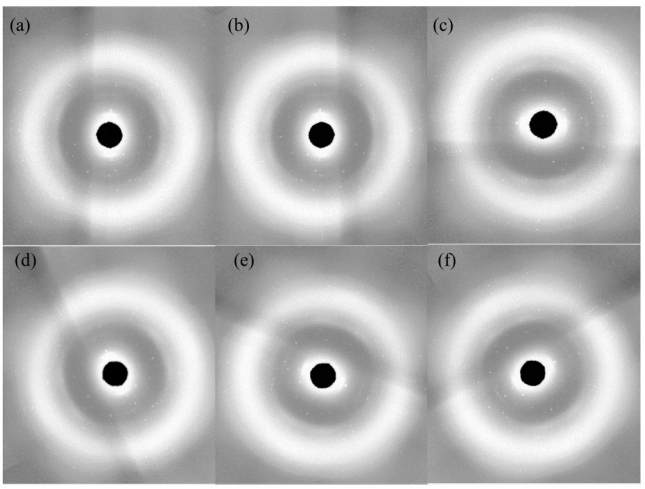

图3 LN83衍射图图像增强结果 (a) 原始衍射图,(b) 左右翻转,(c) 逆时针旋转90°,(d) 逆时针旋转25°,并右移10个像素,(e) 顺时针旋转110°,并右移5个像素,下移5个像素,(f) 顺时针旋转60°
Fig.3 LN83 diffraction pattern image enhancement results (a) Original image, (b) Flip left and right, (c) Rotate 90° counterclockwise, (d) Rotate 25° counterclockwise and move 10 pixels to the right, (e) Rotate 110° clockwise, move 5 pixels to the right and 5 pixels to the down, (f) Rotate 60° clockwise

2 训练方法

2.1 卷积神经网络

深度学习是人工智能研究领域的一个分支学科,用于图像分类识别的CNN[18]是深度神经网络领域研究的重点之一。CNN由输入层、若干隐层与输出层构成,相邻神经元通过权重相互连接,并通过模仿人脑机制自动学习数据特征以进行图像分类。在这里,我们使用的CNN[19]架构分别为AlexNet、Vgg16、GoogleNet-Inception-V1、GoogleNet-Inception-V3、ResNet101和Mobilenets-V1,这些网络在结构与性能方面不断完善,拥有的网络层数与特点如表3所示。

表3 5种卷积神经网络
Table 3 Five convolutional neural networks

网络 Net	网络深度 / 层 Depth / layer	特点 Characteristic
AlexNet	8	网络层数少,采用ReLu激活函数 Less layer, use ReLu activation function
Vgg16	16	采用小卷积核,收敛速度加快 Small convolution kernels to speed up convergence
Inception-V1	22	并行计算,去除全连接层 Parallel computing, remove the full connection layer
Inception-V3	46	并行计算,将卷积拆分,减少数据规模 Parallel computing, split convolution
ResNet101	101	采用残差网络优化学习目标 Optimize learning objectives using residual network
MobileNets-V1	28	卷积可分离,引入全局超参数 Separate the convolution depth, use global hyperparameters

MobileNets是一种轻量级深度CNN,此网络的核心为深度可分离卷积,本质上是将卷积分解为深层卷积和点对点卷积,这种结构实现了在不降低网络性能的前提下减少网络参数和计算量。同时,

MobileNets引入两个轻量的全局超参数在网络延迟和准确率之间做权衡，并且根据实际情况调整出合适的模型大小。串行晶体学产生的实验数据由于数量很多而对运行速率的要求很高，因此，将MobileNets应用于衍射图的分类，希望可以在不降低准确度的基础之上提高运行速率。

2.2 训练与测试

按照7∶2∶1的比例将从蛋白质晶体库下载并通过预处理后得到的42 000个衍射图分别作为训练集、验证集以及测试集，并且设定训练集batch size、初始化学习率、总训练次数分别为16、0.001和100 000代，其中CNN采用Tensorflow[20]框架实现。同时出于均衡样本的考虑，对训练集与验证集之中“也许命中”的样本以及“未命中”的样本分别进行过采样和欠采样的处理。运行流程图如图4所示。

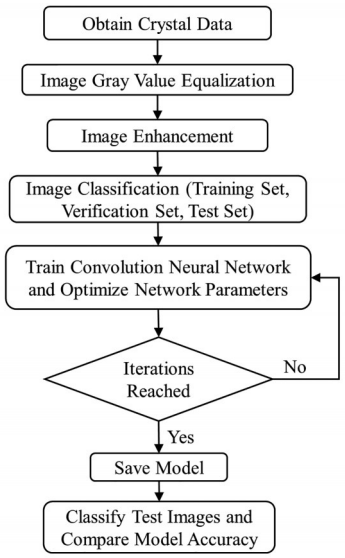

图4 卷积神经网络训练及预测流程图
Fig.4 Flow chart of convolutional neural network for training and prediction

3 训练结果分析与评价

首先，为了验证MobileNets是否适合用于串行晶体学衍射图的筛选，对4种不同的蛋白质晶体样品使用MobileNets进行训练及测试，结果如表4所示。进行测试的10个样本是从测试集中随机抽样得到的，其中标签为“命中”“也许命中”以及“未命中”的样本分别为3个、3个、4个。

从表4可知，对验证集及测试集样本，MobileNets所得到的分类准确度普遍比较高，我们初步判定MobileNets可以用于衍射图的筛选。为了进一步对比MobileNets与其他CNN的准确度以及运行速率，在GPU、Tensorflow1.4.0的运行环境下，我们分别使用AlexNet、GoogleNet-Inception-v1、GoogleNet-Inception-v3、Vgg16、ResNet和MobileNets对所有样本进行图像筛选，运行结果如图5所示。其中，运行速率是指网络完成一次衍射图筛选分类所需要的时间。

分析图5结果可知，对于大多数蛋白质晶体，AlexNet模型训练的准确度保持在0.5左右。事实上，AlexNet由于网络层数较少导致其对衍射点特征的甄别存在困难，无法提取到图像特征，因此AlexNet不适合运用于串行晶体衍射图的分类。

此外，GoogleNet-Inception-v1、GoogleNet-Inception-v3、Vgg16、ResNet、MobileNets都可以用来进行衍射图的筛选。然而不同的神经网络在面对不同的蛋白质晶体时，表现出不同的分类准确率和运行速率。其中。Inception-v1与ResNet与其他网络相比，对大部分蛋白质晶体进行数据筛选时表现出较高的准确率，但两者在测试时的运行速率没有MobileNets表现好。综合比较各个网络面对不同蛋白质晶体在验证集与测试集上的准确度及速率，MobileNets面对各类蛋白质在准确度和运行速率上都有较为不错的表现。

图5给出了不同的CNN在面对不同的蛋白质晶体时的分类准确度，我们为了进一步对比MobileNets相较于其他网络的分类准确度，以LN83（Hydrogenase，氢化酶蛋白质晶体）为例，将全部样本数据按照“命中”“也许命中”和“未命中”的标准将测试集样本的分类准确度进行统计，结果如表5所示。

表4 各个样品使用MobileNets的验证集及测试集准确度
Table 4 Verification set and test set accuracy of each samples based on MobileNets

样品 Samples	验证集准确度 Accurancy / %	测试集准确度 Accurancy
L498-氢化酶蛋白质晶体 Thermolysin	62.2	7/10
LN84-光系统 II Photosystem II	82.3	8/10
LN83-嗜热菌蛋白酶 Hydrogenase	81.8	8/10
LO19-辛环素 Cyclophilin A	78.0	9/10

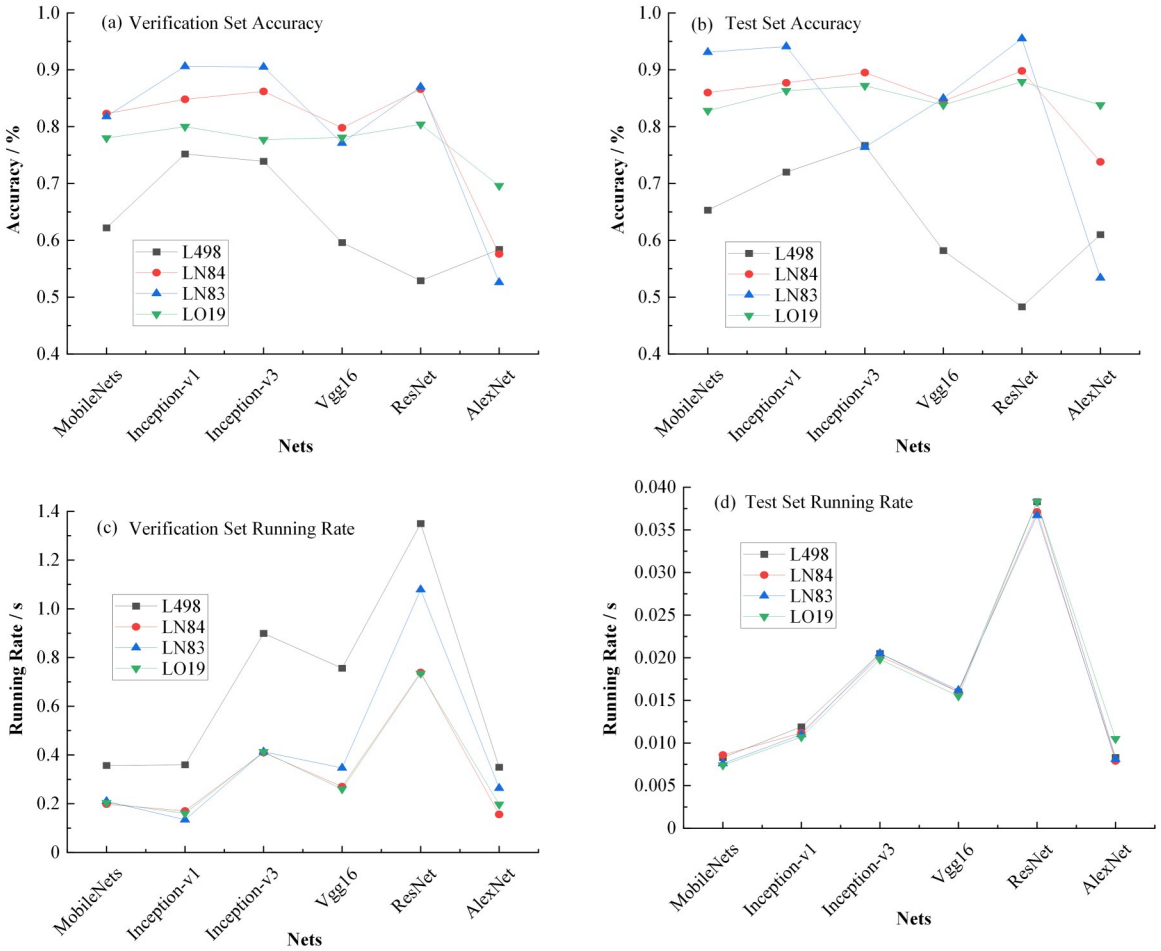

图5 使用不同网络训练的验证集及测试集准确度和运行速率
(a) 验证集准确度,(b) 测试集准确度,(c) 验证集运行速率,(d) 测试集运行速率

Fig.5 Accuracy and operation rate of verification set and test set based on different networks
(a) Verification set accuracy, (b) Test set accuracy, (c) Verification set running rate, (d) Test setverification set running rate

由于“也许命中”的样本衍射图中布拉格点的数量处于其他两种样本之间,因此,这类样本的分类对CNN特征提取能力的要求最高。从表5可以看到,AlexNet、Inception-v1、Inception-v3、ResNet对“也许命中”样本的分类结果都不理想,Vgg16与MobileNets对这类样本的分类准确度可以达到0.7以上,可以认为,Vgg16和MobileNets对LN83蛋白质晶体衍射图特征的提取能力比较强。对于“未命中”与“命中”的样本,MobileNets与Inception-v1的分类准确度都高于0.91,是6种CNN中表现最好的。

综上所述,MobileNets不论是在处理“也许命中”的晶体衍射图特征时,还是在处理“未命中”与“命中”的晶体衍射图特征时都有良好的表现。

接下来,通过可视化分类结果的方式来分析CNN分类原理。以LN84(光系统II,Photosystem II)为例,随机抽取1 000个预测样本,分别对6种CNN使用*t*-SNE[21]流形学习方法将分类结果降至二维平面。结果如图6所示。

由图6可见,不同标签的样本以聚集的方式出现,CNN提取特征的过程是一个聚类的过程。除了MobileNets以外,ResNet、Inception-v1、Vgg16、Inception-v3和AlexNet在分类结果降维后都产生一些小的独立集合,这些小聚类的产生说明这些网络在对衍射图特征选择时存在偏好,但MobileNets降维后比较明显地分为“也许命中”“未命中”和“命中”三类。以上事实说明,除MobileNets以外,其他网络在进行蛋白质衍射图筛选时偏向于基于衍射图某一类的特征进行评判,但MobileNets综合衍射图各个方面的特征,给出一个无偏好的分类结果,从侧面证明了MobileNets提取的衍射图特征更加准确。

根据CNN的模型理论,MobileNets将卷积分解为深层卷积和点对点卷积,运行速率应该有明显的提升,然而,在GPU、tensorflow1.4.0环境下,MobileNets模型的速率并没有明显比Inception-V1和AlexNet这种网络层数比较少的网络快。为了研究这一现象产生的原因,我们以LN83为例分别在

表5 LN83使用不同网络的验证集及测试集准确度
Table 5 Accuracy of verification set and test set using different networks based on LN83

网络 Nets	标签 Label	LN83-氢化酶蛋白质晶体 Hydrogenase		
		命中 Hit	也许命中 Maybe	未命中 Miss
MobileNets	命中 Hit	0.919	0.070	0.011
	也许命中 Maybe	0.168	0.701	0.131
	未命中 Miss	0.014	0.043	0.943
Inception-v1	命中 Hit	0.935	0.043	0.022
	也许命中 Maybe	0.350	0.416	0.234
	未命中 Miss	0.008	0.028	0.964
Inception-v3	命中 Hit	0.958	0.029	0.013
	也许命中 Maybe	0.547	0.343	0.109
	未命中 Miss	0.058	0.202	0.740
Vgg16	命中 Hit	0.893	0.086	0.021
	也许命中 Maybe	0.073	0.876	0.051
	未命中 Miss	0.020	0.141	0.840
ResNet	命中 Hit	0.854	0.084	0.063
	也许命中 Maybe	0.015	0.518	0.467
	未命中 Miss	0.001	0.004	0.995
AlexNet	命中 Hit	0.907	0.014	0.079
	也许命中 Maybe	0.927	0.022	0.051
	未命中 Miss	0.509	0.016	0.475

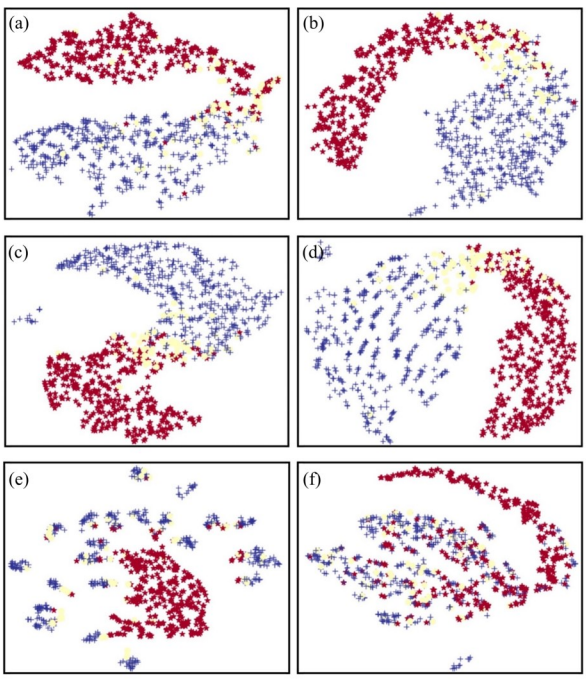

图6 6种卷积神经网络*t*-SNE降维结果（圆形为“也许命中”，十字为“未命中”，五角星为“命中”）
(a) MobileNets，(b) ResNet，(c) Inception-v1，(d) Inception-v3，(e) Vgg16，(f) AlexNet
Fig.6 *t*-SNE dimensionality reduction results of six convolutional neural networks (the circle is the "maybe " sample, the cross is the "Miss" sample, and the pentagram is the "hit" sample)
(a) MobileNets, (b) ResNet, (c) Inception-v1, (d) Inception-v3, (e) Vgg16, (f) AlexNet

GPU、tensorflow1.4.0和CPU、tensorflow1.4.0两种环境下进行实验，结果如图7所示。

由图7可知，MobileNets在GPU上的运行速率并不突出，然而，在CPU上的运行速率优于其他CNN。这是因为GPU是并行处理大规模数据的运算平台，总运算时间的主导因素是网络层数；CPU缺

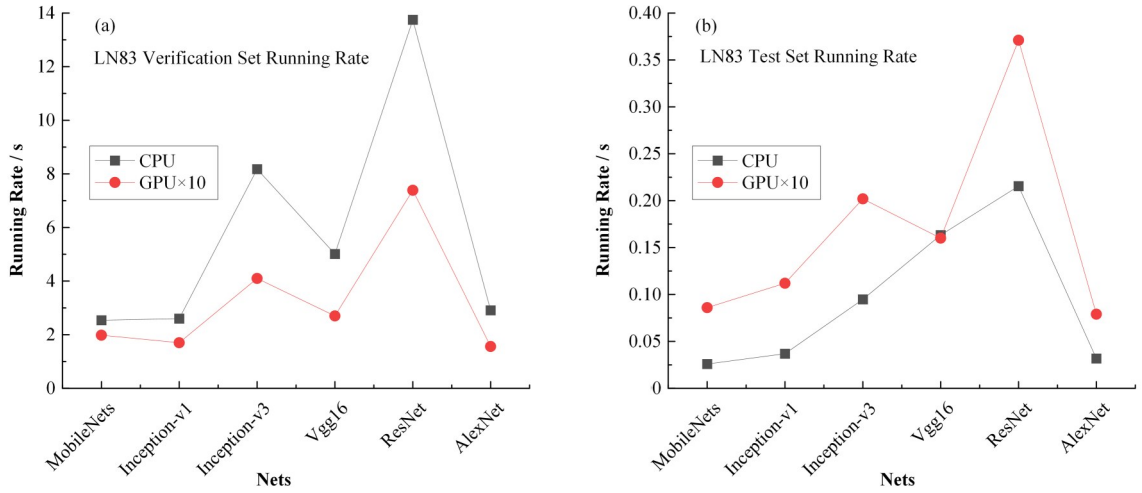

图7 LN83在GPU和CPU上的运行速率
Fig.7 Running rate of LN83 on GPU and CPU

乏并行处理的能力，总运算时间的主导因素是总计算量。MobileNets一方面将一个标准卷积分割为一个深度卷积与一个逐点卷积，减少了模型参数量与总计算量，另一方面为了提高网络准确度，增加了网络层数。因此，MobileNets在运行速率主要由网络层数决定的GPU平台上表现不突出，但是在运行速率主要由总计算量决定的CPU平台上的运行速率提升明显。

事实上，串行晶体学实验中样本的命中率很低，为了保证有效的实验数据尽可能不被遗漏，在进行实验数据筛选后，可以汇总“命中”与“也许命中”的样本进行进一步分析和蛋白质晶体结构解析。因此，在进行数据筛选的过程中，要尽可能保证“命中”与“也许命中”的样本没有被分类为“不命中”，而不过多考虑两者之间是否能被准确分类，即只要“命中”与“也许命中”的样本没有被分类为“不命中”，无论“命中”的样本被分类为“命中”或者“也许命中”，还是“也许命中”的样本被分类为“命中”或“也许命中”都可以保证实验数据的高效利用，因此，以LN83为例，使用6个模型对测试集样本进行二分类（“命中”与“也许命中”为一类、“未命中”一类），结果如表6所示。

由表6可见，MobileNets、Inception-V3和Vgg16对“命中/也许命中”样本的分类准确度达到0.97，明显优于其他神经网络。这意味着使用MobileNets、Inception-V3和Vgg16进行蛋白质衍射图筛选可以有效避免实验数据的浪费。

综上所述，MobileNets对蛋白质晶体衍射数据进行筛选时，在准确率和运行速率上相比于其他CNN都表现优秀。因此，最后为了确认MobileNets

表6 LN83样本测试集二分类的准确率
Table 6 Accuracy of two classification based on Ln83 sample

网络 Nets	命中/也许命中 Hit/maybe	未命中 Miss
MobileNets	0.970	0.943
Inception-V1	0.944	0.964
Inception-V3	0.972	0.740
Vgg16	0.974	0.840
ResNet	0.873	0.955
AlexNet	0.925	0.475

分类结果的可靠性，以LN83的测试集中“命中/也许命中”样本和“未命中”样本分类结果的可靠性为指标作图，结果如图8所示。

衍射图通过CNN分类输出的结果是一个1×3维的数组，分别代表着“命中”的概率、“也许命中”的概率和“未命中”的概率。若“命中”和“也许命中”的概率相加不小于0.5，则这个数据被分类为“命中/也许命中”；否则，数据会被分类为“未命中”。对于被分类为“命中/也许命中”的样本，其标签为“命中/也许命中”的概率越高，分类结果越可靠（我们对于可靠的定义是：CNN进行数据筛选时分类出现错误的可能性的高低，其中越可靠代表分类出现错误的可能性越低）。图8(a)中点的纵坐标越靠近1，图8(b)中点的纵坐标越靠近0表示CNN模型越可靠。观察图8(a)中点的分布，发现图8(a)中点的纵坐标普遍靠近或等于1，因此我们认为MobileNets可以得到可靠的分类结果。

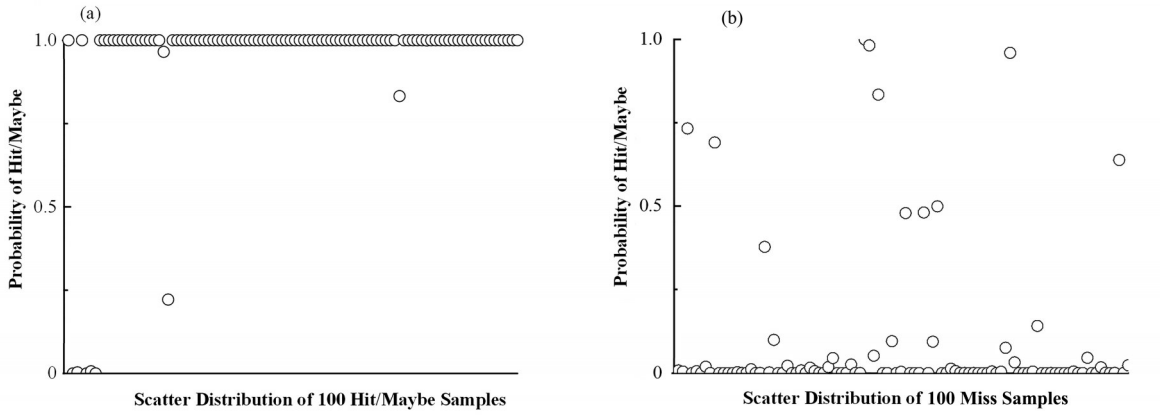

图8 MobileNets命中/也许命中(a)与未命中样本(b)可靠性分布
Fig.8 MobileNets hit /maybe (a) and miss sample (b) reliability distribution

4 测试

最后，为了验证所构建的MobileNets是否可以实际应用于串行晶体学实验，模拟了一组串行晶体学实验数据（实验晶体：溶菌酶，数据来源：上海同步辐射光源生物大分子晶体学光束线站[22–23]，探测器：Eiger X 16M），按照上述方法对图像进行预处理后，使用经过LN84数据集预训练的MobileNets对该组数据进行筛选，最终在2 000个数据中有155个Hit样本、11个Maybe样本和1 834个Miss样本（图9），为了证明这些样本是否有衍射点，是否可以用来解析蛋白质晶体结构，我们利用CrystFEL选择合适的寻峰方法及参数，将数据进行index指标化处理，凡是能够进行指标化处理的衍射数据均为有效的衍射数据，不能进行指标化处理的为无效衍射数据，其中包括有衍射点但不能进行指标化处理和没有衍射点二种情形。对2 000个测试数据进行index指标化处理，共有138个衍射数据可以index（均为被MobileNets分类为Hit的数据）、1 862个衍射数据无法index（被MobileNets分类为Hit的数据17个，被MobileNets分类为Maybe的数据11个、被MobileNets分类为Miss的数据1 834个），具体情况如下：

1）被MobileNets分类为Hit的数据：155个图像中有17个图像无法index，观察这些无法index的图像，可以发现有少部分图像中的衍射点很不明显，和Miss图像相似，这可能是CNN进行数据筛选时模型本身存在误差导致的。对于肉眼可以很明显看到衍射点的图却无法index的图像，原因可能有如下几点：一是衍射点数量少，分辨率低；二是多颗晶体的衍射点在一张图上混合；三是样品并非蛋白质晶体，但因其晶胞很大，在有效衍射点数量较少的情形下与蛋白晶体衍射图看起来十分相似，导致肉眼可见的衍射点并不一定是蛋白质晶体衍射点等。

2）被MobileNets分类为的Maybe数据：11个图像均无法index，但肉眼可以看到这些图像中有一部分衍射点，并不好直接判断这部分是否需要舍弃。因此可以先保留这部分数据，后续再进行分析。

3）被MobileNets分类为Miss的数据：1 834个图像因为衍射点很少或没有衍射点均无法index，为无效的衍射数据，CNN模型没有遗漏对结构解析有帮助的数据。

从测试数据集筛选的准确率来看，CNN模型已经做到了将所有“命中”以及“也许命中”的数据准确筛选出来，实现了初期目标。对于一部分CNN认为“命中”，但无法index的数据：一方面可能是CNN主

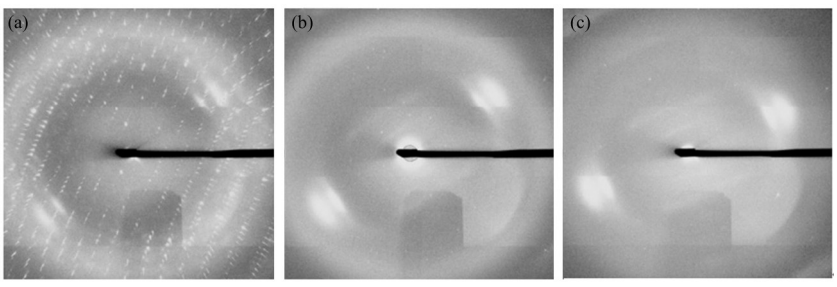

图9 MobileNets筛选 (a) 命中样本，(b) 也许命中样本，(c) 不命中样本
Fig.9 Sample selected by MobileNets (a) Hit, (b) Maybe, (c) Miss

要考虑的是衍射点的数量，而这些数据本身并不是单一的蛋白质晶体衍射图或者衍射强度太低，无法直接通过index解析这些图像的结构，对于这些数据，我们可以先通过CNN的筛选将这些数据保留，再进行人工分析；另一方面可能受到了探测器坏点的影响，坏点会对阈值的设置产生一定影响（阈值过大，会遗漏一些衍射点从而无法index；阈值过小，会将坏点认为是衍射点，影响index的结果），而阈值的设置会影响寻峰结果导致index的参考指标存在偏差。从测试集数据筛选的效率来看，CrystFEL对一张图像进行index需要几十秒甚至几分钟（根据图像中峰值的多少，即运算量决定），并且还需要尝试调整寻峰方法及寻峰参数才能保证结果的正确性，需要消耗一定的时间，而CNN虽然在模型建立的时候需要花费时间，但一旦模型建立成功，在GPU和CPU上进行图像筛选时每次分别只需要10^{-3} s、10^{-2} s，运行速率的提升十分明显。因此，经过测试发现，将MobileNets应用于同步辐射串行晶体学实验数据的预筛选，是一种准确而高效的方法。

5 结语

本文比较了AlexNet、Vgg16、GoogleNet-Inception、ResNet和MobileNets卷积神经网络用于晶体学衍射图筛选的准确度以及运行速率，试图找到一个准确度高且运行速率快的CNN用于衍射数据的筛选。实验结果表明：

1）在比较不同的CNN面对不同的蛋白质晶体时表现出的分类准确度和运行速率后，发现MobileNets在准确度和运行速率上都有较为不错的表现。

2）通过研究“命中”“也许命中”和“未命中”样本的具体分类准确度，发现MobileNets不论是在处理特征比较明显的“未命中”与“命中”的蛋白质晶体样本衍射图特征时，还是在处理特征较不明显的“也许命中”的蛋白质晶体样本衍射图特征时都有良好的表现。

3）通过对不同CNN分类结果的降维，发现MobileNets更倾向于综合衍射图各个方面的特征，给出一个无偏好的分类结果，特征提取更准确。

4）通过比较不同CNN在GPU与CPU两种环境下的运行速率，发现MobileNets在CPU平台上相比于其他CNN有较好的运行速率。

5）通过对蛋白质晶体衍射图的二分类，发现MobileNets可以有效避免实验数据的浪费。

6）通过对分类结果中样本被正确分类的概率分析，发现MobileNets可以得到可靠的分类结果。

因此，基于MobileNets可以构建一个准确高效的、用于晶体学实验数据筛选的自动化数据筛选工具。今后，随着CNN不断优化，这一实用的数据处理工具可以真正应用于晶体学实验，从而提高实验准确度和效率。

致谢 感谢上海光源生物大分子晶体学线站提供蛋白质晶体衍射实验数据用于测试，感谢上海光源郁峰研究员在蛋白质晶体结构数据分析方面提供的指导与帮助。

作者贡献声明 惠子：负责研究的提出、方法设计、数据收集、模型建立及模型对比选择、文章的起草与修订；余立：负责模型在上海光源的测试及对比软件上的数据处理；周欢：负责实验数据的提供；唐琳：负责文章的修订与实验方法及思路指导；何建华：负责最终版本的修订、项目的监督与管理。

参考文献

1 Chapman H N, Fromme P, Barty A, *et al*. Femtosecond X-ray protein nanocrystallography[J]. Nature, 2011, **470**(7332): 73 – 77. DOI: 10.1038/nature09750.

2 Emma P, Akre R, Arthur J, *et al*. First lasing and operation of an ångstrom-wavelength free-electron laser [J]. Nature Photonics, 2010, **4**(9): 641 – 647. DOI: 10.1038/nphoton.2010.176.

3 Rossmann M G. Serial crystallography using synchrotron radiation[J]. IUCrJ, 2014, **1**(Pt 2): 84 – 86. DOI: 10.1107/S2052252514000499.

4 Stellato F, Oberthür D, Liang M N, *et al*. Room-temperature macromolecular serial crystallography using synchrotron radiation[J]. IUCrJ, 2014, **1**(Pt 4): 204 – 212. DOI: 10.1107/S2052252514010070.

5 Nogly P, James D, Wang D J, *et al*. Lipidic cubic phase serial millisecond crystallography using synchrotron radiation[J]. IUCrJ, 2015, **2**(Pt 2): 168 – 176. DOI: 10.1107/S2052252514026487.

6 White T A, Kirian R A, Martin A V, *et al*. CrystFEL: a software suite for snapshot serial crystallography[J]. Journal of Applied Crystallography, 2012, **45**(2): 335 – 341. DOI: 10.1107/s0021889812002312.

7 Ke T W, Brewster A S, Yu S X, *et al*. A convolutional neural network-based screening tool for X-ray serial crystallography[J]. Journal of Synchrotron Radiation, 2018, **25**(Pt 3): 655 – 670. DOI: 10.1107/S1600577518004873.

8 Zimmermann J, Langbehn B, Cucini R, *et al*. Deep neural

networks for classifying complex features in diffraction images[J]. Physical Review E, 2019, **99**(6 – 1): 063309. DOI: 10.1103/PhysRevE.99.063309.

9 Howard A G, Zhu M, Chen B, *et al*. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL]. 2017: arXiv: 1704.04861. https://arxiv.org/abs/1704.04861.

10 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, **60**(6): 84 – 90. DOI: 10.1145/3065386.

11 Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 27-30, 2016, Las Vegas, NV, USA. IEEE, 2016: 2818 – 2826. DOI: 10.1109/CVPR.2016.308.

12 Szegedy C, Liu W, Jia Y Q, *et al*. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 7-12, 2015, Boston, MA, USA. IEEE, 2015: 1 – 9. DOI: 10.1109/CVPR.2015.7298594.

13 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. 2014: arXiv: 1409.1556. https://arxiv.org/abs/1409.1556.

14 He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 27-30, 2016, Las Vegas, NV, USA. IEEE, 2016: 770 – 778. DOI: 10.1109/CVPR.2016.90.

15 Maia F R N C. The coherent X-ray imaging data bank[J]. Nature Methods, 2012, **9**(9): 854 – 855. DOI: 10.1038/nmeth.2110.

16 Goodfellow I, Bengio Y, Courville A. Deep learning[M]. MIT Press , 2016.

17 简丽琼 . 基于二维直方图均衡化的图像增强算法[J]. 信息与电脑(理论版), 2015(22): 49 – 54. DOI: 10.3969/j.issn.1003-9767.2015.22.021.
JIAN Liqiong. Image enhancement algorithm based on two-dimensional histogram equalization[J]. China Computer & Communication, 2015(22): 49 – 54. DOI: 10.3969/j.issn.1003-9767.2015.22.021.

18 Reid C R. From functional architecture to functional connectomics[J]. Neuron, 2012, **75**(2): 209 – 217. DOI: 10.1016/j.neuron.2012.06.031.

19 LeCun Y, Boser B, Denker J S, *et al*. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, **1**(4): 541 – 551. DOI: 10.1162/neco.1989.1.4.541.

20 Abadi M, Agarwal A, Barham P, *et al*. TensorFlow: large-scale machine learning on heterogeneous distributed systems[ED/OL]. [2016-05-16]. https://arxiv. org/abs/1603.04467.

21 Der Maaten L V, Hinton G E. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, **9**(86): 2579 – 2605.

22 Wang Q S, Zhang K H, Cui Y, *et al*. Upgrade of macromolecular crystallography beamline BL17U1 at SSRF[J]. Nuclear Science and Techniques, 2018, **29**(5): 68. DOI: 10.1007/s41365-018-0398-9.

23 Zhang W Z, Tang J C, Wang S S, *et al*. The protein complex crystallography beamline (BL19U1) at the Shanghai Synchrotron Radiation Facility[J]. Nuclear Science and Techniques, 2019, **30**(11): 170. DOI: 10.1007/s41365-019-0683-2.