第 38 卷 第 2 期
2020 年 7 月

天 文 学 进 展
PROGRESS IN ASTRONOMY

Vol. 38, No. 2
July., 2020

doi: 10.3969/j.issn.1000-8349.2020.02.04

深度学习在天文学中的应用与改进

陶一寒[1,2]，崔辰州[1,2]，张彦霞[1]，许允飞[1,2]，樊东卫[1,2]，
韩 叙[1]，韩 军[1,2]，李长华[1,2]，何勃亮[1,2]，
李珊珊[1,2]，米琳莹[1,2]，杨涵溪[1,2]，杨丝丝[1,2]

(1. 中国科学院 国家天文台，北京 100101； 2.国家天文科学数据中心，北京 100101)

摘要： 近年来，深度学习和人工智能技术迅猛发展，在多个学科领域得到了广泛关注和应用。天文学研究也不甘落后，涌现出一大批应用深度学习进行数据分析的工作。总结了深度学习在天文中的应用情况和趋势、天文数据类型和机器学习任务、天文中常用的深度学习网络模型和方法，以及深度学习在天文研究中的代表性应用和进展，并探讨和提出了其未来在天文学领域中的应用和改进建议。

关 键 词： 天文数据分析处理；深度神经网络；机器学习；虚拟天文台

中图分类号： P111.5 **文献标识码：** A

1 深度学习方法在天文学中的应用和发展趋势

机器学习是一种实现人工智能的方法，主要应用于难以用规则描述并显式编程的问题。目标是研究如何让计算机模拟人类的学习行为，通过经验自动提高算法，从数据中学习隐含的模式并建立模型，从而能够对相似的问题做出预测[1]。深度学习是机器学习方法的一种特殊类型，它与传统机器学习的区别主要在于特征表示建立的过程。传统机器学习算法一般需要根据专业领域知识来人工设计特征，用特征集合来表示原始数据，进而训练模型，其中特征设计对算法的效果起着决定性的作用。深度学习由人工神经网络发展而来，通过构建深层神经网络从原始数据中逐层提取抽象特征，称为表示学习，这一过程体现了算法的智能[2, 3]。深度学习具有强大的特征学习能力，特别是在计算机视觉和语音识别领域达到了超

收稿日期： 2019-06-26； **修回日期：** 2019-08-09

资助项目： 国家自然科学基金 (11803055)；国家自然科学基金委员会-中国科学院天文联合基金 (U1731125, U1731243, U1931132)；中国科学院“十三五”信息化建设专项 (XXH13503-03-107)

通讯作者： 陶一寒, y.tao@nao.cas.cn

越人类的水平。随着人工智能技术的发展和普及，如 TensorFlow[①], Keras[②], PyTorch[③] 等深度学习框架发展成熟，构建和训练深度神经网络的技术门槛大大降低。机器学习方法被广泛地应用于医学、生物学、物理学、天文学等众多自然科学领域，为这些学科提供了大数据时代解决问题的新思路。

目前在天文学领域，观测设备和技术飞速发展，望远镜的数据获取率不断提升，数据量持续增长，如泛星计划 (Pan-STARRS)、凌星系外行星巡天望远镜 (Transiting Exoplanet Survey Telescope, TESS)、欧空局盖亚 (Gaia) 全天天体测量干涉仪、建设中的大口径全天巡视望远镜 (Large Synoptic Survey Telescope, LSST) 和平方千米射电望远镜阵 (Square Kilometer Array, SKA) 等项目的数据量都达到 PB 量级。天文学家们迫切需要人工智能方法来分析海量数据，并从中挖掘和获取知识。机器学习方法正好满足了天文大数据分析挖掘的需求：首先，机器学习适合用于靠经验完成的任务，比如图像识别等，可以将天文学家们人工判断的过程自动化建模，同时还能通过自动化的特征学习帮助天文学家从不同维度提取特征，发现他们暂不能明确提炼的特征；其次，机器学习可用于探索复杂高维数据的隐含结构及相关性，能帮助天文学家们挖掘未知的天体及物理特性。

天文学家们从 20 世纪 90 年代起便开始探索使用机器学习方法，2004 年逐步形成规模，2015 年迎来热潮。近年来，深度学习技术带动人工智能的第三次浪潮，基于深度学习的人工智能算法在图像识别、语音识别、无人驾驶等领域不断取得突破性进展。自 2014 年以来，天文学领域也出现了很多应用深度学习进行数据分析处理的论文，并逐年增加。研究表明，在许多特定任务上深度学习优于传统的依靠人工或规则编程的方法，获得了接近甚至超越人类专家的表现，具有广阔的应用前景。

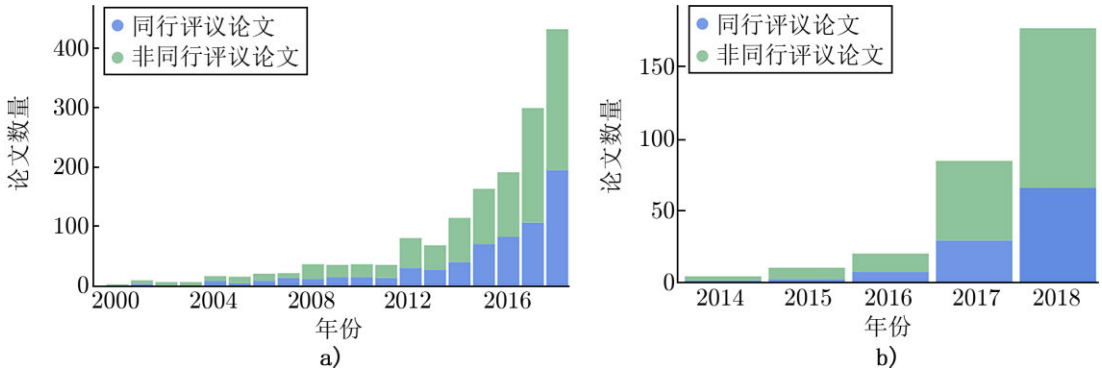
根据天体物理论文数据库[④]的检索数据 (如图 1a) 所示)，天文学领域论文标题、摘要和关键词中提到“机器学习”(包括深度学习)的论文从 2004 年开始出现，在 2012 年左右开始迅速增加，2018 年同行评议论文达到了 195 篇，加上非同行评议论文则达到 400 多篇。其中 2014－2018 年天文学领域中深度学习的论文数量大幅增加。如图 1b) 所示，应用深度学习方法的论文主要从 2014 年开始出现，2017 年以来增长迅速，2018 年发表的同行评议论文超过了 60 篇，加上非同行评议论文超过了 150 篇。图 2 展示了天文学中应用深度学习方法论文的研究主题关键词，揭示了深度学习在天文数据分析处理中的应用方向，包括开展研究较早的星系图像分类、测光红移估计等研究方向，近些年在一些天文学研究的热点内容上应用呈增长趋势，如引力波探测、系外行星搜寻、引力透镜识别、暂现源检测、太阳耀斑预测等。这些论文表明深度学习模型已经被广泛应用于天文学的诸多领域中，并在一些问题上取得了优于传统方法的效果，是一种有效的海量数据分析处理方法。在当今的天文大数据时代，越来越多的天文学家开始尝试运用深度学习方法分析和挖掘数据。

[①] https://www.tensorflow.org/

[②] https://keras.io/

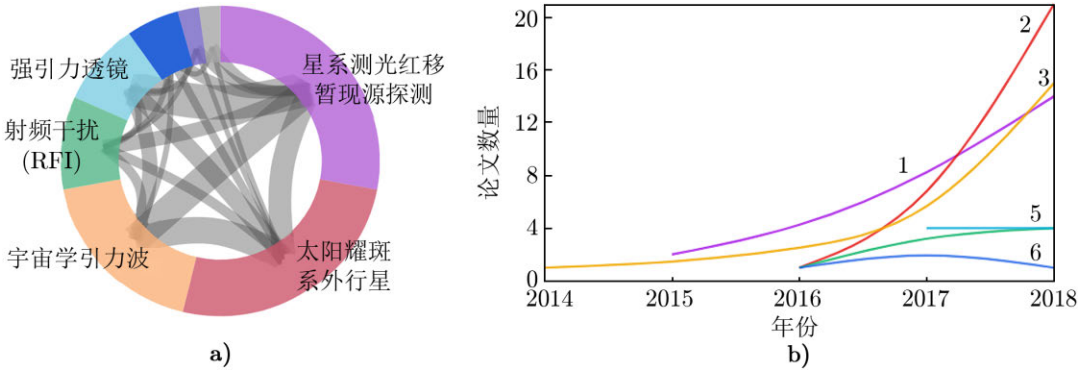
[③] https://pytorch.org/

[④] https://ui.adsabs.harvard.edu/



注：a) 2001－2018 年天文学领域标题、摘要和关键词中包含机器学习/深度学习 (machine learning/deep learning) 的论文数量；b) 2014－2018 年天文学领域标题、摘要和关键词中包含深度学习 (deep learning/deep neural network/convolutional neural network) 的论文数量。

图 1　天文学领域应用机器学习和深度学习的论文数量分布 (来源于 ADS)



注：a) 论文研究主题；b) 论文数量趋势。

图 2　2014—2018 年天文学中应用深度学习的论文研究主题和各主题数量趋势 (来源于 ADS)

2　天文数据类型和机器学习任务

天文学研究中使用的数据可分为观测数据和数值模拟数据两大类。观测数据类型主要包括图像、星表、光谱、时序数据等。图像由测光观测获得，即在望远镜焦面上放置滤光片和探测器，从望远镜拍摄的图像中识别天体并测算光度。光谱是由望远镜焦面上放置光谱仪的光谱观测获得，比图像包含更多天体的物理特性信息，如金属丰度、有效温度、重力加速度和动力学信息等；通过光谱中谱线位移情况可进一步得到红移值，用来估计天体的距离。星表通常也是现代大型巡天项目的科学数据产品之一，一般包含一系列天体的位置、在不同波段上的亮度、类别以及一些其他物理性质等信息。时序数据记录天体在不同时间的性质，由望远镜在一段时间内重复观测相同天区或天体得到，可以用来发现并研究变源。

通常机器学习按建模的形式可分为监督学习、无监督学习和半监督学习。监督学习是根据有标签的数据建模，预测新数据的标签。在机器学习中标签指的是机器学习问题的标准答案，也就是希望机器能够通过分析数据给出的答案。具体来说，监督学习又可分为分类和回归两种任务，两者的区别在于，分类任务数据的标签是离散值，而回归任务标签是连续值。无监督学习则是对无标签的数据建模，发现数据中的隐含特征和规律，聚类和降维算法都属于无监督学习方法。聚类可以根据特征相似性将样本分组；而降维能够将高维数据转化到低维空间中表达，可能更直接地发现数据的联系，但也会丢失数据原本的一些特征。半监督学习是一种将无监督学习与监督学习结合的方法，可以在数据标签不完整时使用，同时使用未标记数据和标记数据建立模型。

根据不同的天文数据类型和研究方向，天文中机器学习任务主要包括以下几种。

(1) 目标检测和分类

目标检测和分类是天文数据分析的一项重要任务。目标检测主要是从望远镜获得的图像和光谱等数据中判断是否包含关注的天体，可以看作是一个二分类问题。而有时还需要标出目标的位置和范围，这就需要计算出目标轮廓的坐标，是一个回归问题。目标分类是对天体具体的类型做出判断，可能是一个二分类或多分类问题。天文学领域的科研人员尝试借鉴在图像识别任务中取得优异效果的深度学习方法——卷积神经网络 (convolutional neural network, CNN)，将其应用于天体的检测和分类，如星系分类[4–9]、引力透镜识别[10–13]、暂现源检测[14, 15]等。经研究测试发现，深度学习方法在很多场景下可以有效替代传统的人工检验方法，比模板匹配等目前常用方法更灵活高效，在精度上通常比传统机器学习方法也有较大提高。

(2) 参数估计

参数估计是机器学习中典型的回归问题。用于研究天体特性的物理量是根据望远镜得到的光谱和测光数据测算出来的，如天体的质量、温度、元素丰度、视向速度、红移等。传统的方法通常是通过模板匹配来得到。近年深度学习方法被成功应用于参数测量[16–20]，并极大提高了效率。

(3) 时序数据分析

时序数据分析是天文学领域近些年广泛关注的数据分析课题。随着大视场快速巡天项目的开展，天文学进入时域天文学时代，高时间分辨率的观测也对时序数据分析提出了新的挑战。如引力波探测[21–23]、系外行星搜寻[24–26]、暂现源识别[27, 28]等都需要对时序数据，如时频图、光变曲线等进行追踪和分析，这些问题广义上也可以算作目标检测和识别，只是时序数据区别于一般的图像和光谱等数据，根据其数据特性在分析方法上有些不同，有时还要求对数据流进行实时分析。

(4) 数据降噪和生成

数据降噪和生成的目标是能在硬件有限的情况下最大还原原始场景的信号。天文学家们利用生成式对抗网络 (generative adversarial networks, GAN)、变分自编码器 (variational auto-encoder, VAE) 等方法对望远镜图像进行超分辨率重建，在望远镜制造成本一定的情况

下获得更好的数据精度或更多的样本。例如，可以利用递归降噪自编码器针对真实 LIGO (Laser Interferometer Gravitational-wave Observatory) 引力波信号中的非高斯噪声进行降噪[29]，而在训练模型时只需用模拟的高斯噪声。这些生成式方法属于无监督表示学习，不以预测标签为目的而是学习数据本身的特征表达。

3 天文数据分析中常用的深度学习方法

最近10年，深度神经网络不断迭代发展，涌现出一些经典的网络结构，被广泛应用于各种机器学习任务，并且取得了很好的效果。深度神经网络从感知机发展而来，以模仿生物学机制的人工神经元为构成单元，在输入层和输出层之间加入多个隐层。包含多隐层的前馈神经网络可以近似任意的连续函数，因此深度神经网络可以用来对复杂的函数建模，输出离散或者连续的值，用于分类和回归任务。

如图 3 所示，原始的全连接神经网络中每个神经元都与下一层的全部神经元相连接，每条连接都有一个权重值，代表此连接对下一层特征表示的贡献，同时每一层通过激活函数决定每个神经元是否激活，从而引入非线性因素。网络可基于梯度下降算法通过反向传播技术来训练，得到网络中每层的参数值。全连接神经网络由于网络参数随着层数激增，训练逐渐变难，需要很长时间和较大内存，因而一些优化的网络结构不断被提出。目前常用的网络结构包括卷积网络、循环和递归网络等。

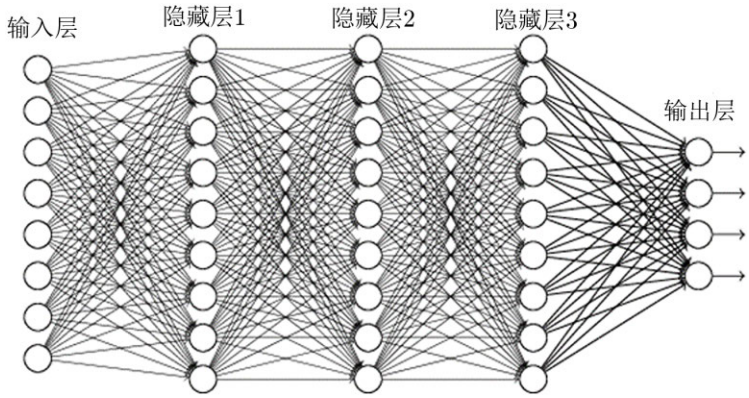
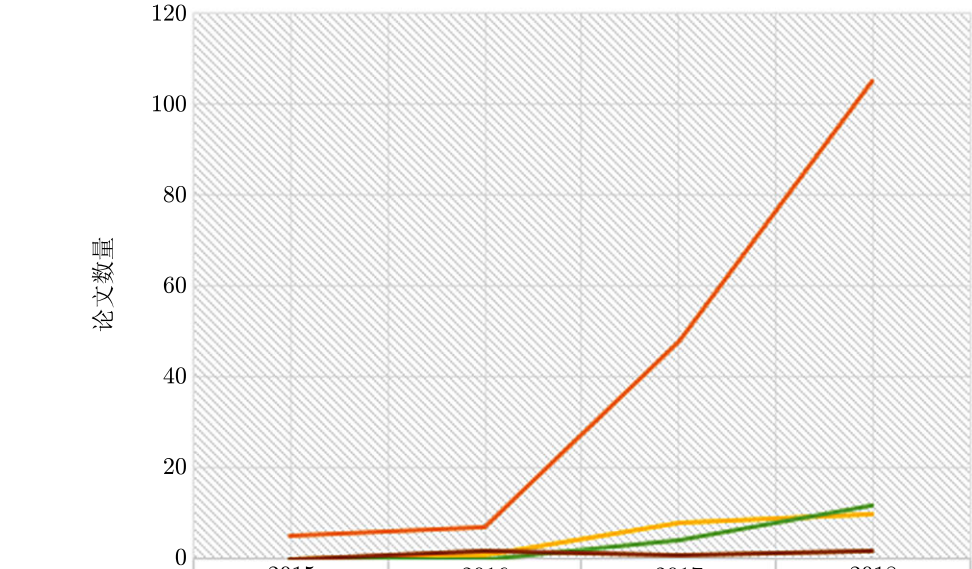


图 3 全连接神经网络模型示意图

深度学习的好处是可以从数据中自动学习特征，减少专家利用领域知识进行人工特征设计的工作量，可灵活地表示任意的复杂函数。图 4 展示了 2015－2018 年天文学领域几种主流深度学习网络模型的应用情况，大量的论文应用了 CNN，远多于其他网络模型；应用GAN 的文章近两年来也呈增长趋势。这两种网络模型在计算机视觉领域经过大量验证和优化，只要稍加修改，便可用于天文图像的分析处理。针对不同的数据和任务，天文数据分析处理中常用的深度学习网络模型主要有以下几种。



	2015	2016	2017	2018
CNN	5	7	48	105
RNN/LSTM	0	1	8	10
GAN	0	0	4	12
VAE	0	2	1	2

图 4　2015—2018 年天文学领域应用主流深度神经网络模型的论文数量

3.1　卷积神经网络

CNN 是一种包含卷积层的前馈神经网络，即用卷积运算代替矩阵乘法运算进行特征提取。它在二维图像数据和一维时间序列数据上应用效果都很优异，也应用于三维视频或图像，如医学影像数据等。在天文学领域应用深度学习方法的论文中，卷积神经网络是应用最多的一类模型，被广泛应用于图像和光谱的分类[30]、参数测量[19, 20]、搜寻系外行星[24, 26]等任务。

卷积神经网络基于三个重要的设计思想：稀疏交互、参数共享以及平移不变[3]。首先，与全连接网络不同，卷积网络的隐层神经元只与上一层的局部数据通过卷积运算相连，因此是稀疏交互；这个局部区域称为此神经元的局部感受野。第二，同一层神经元在一次卷积操作中使用的卷积核是相同的，因此也拥有相同的参数 (如权重和偏差)。参数共享使得模型具有泛化性。每个卷积层的卷积核通常不止一个，每个卷积核对本层输入数据进行卷积运算会得到一张特征图，每张特征图都反映出从原始数据中学习得到的一些空间结构特征。第三，卷积神经网络一般在卷积层后加入池化层，使用某一位置相邻输出的统计特征来代替网络该位置的输出 (最大池化或平均池化)。池化在降维的同时保持局部平移不变性，即当输入少量平移时，池化能使输入的表示近似不变，这在只关心某个特征是否出现而不关心它出现的具体位置的任务中尤为重要。全连接神经网络随着网络加深神经元数量增加，参数增长迅速，而卷积神经网络的稀疏交互、参数共享以及池化的设计使模型参数简化，计算效率

更高，让深层网络的训练变得可能。

卷积神经网络一般由多个卷积层、池化层、全连接层等结构组成，这些基本结构组合形成各种不同的网络结构。随着在计算机视觉领域应用研究逐渐深入，卷积网络的层数逐渐加深，进化出了一代代经典的卷积网络模型，包括 AlexNet[31], GoogLeNet[32], VGGNet[33], ResNet[34] 和 DenseNet[35] 等。在天文数据分析处理的应用中，模型的搭建大多是基于这些计算机视觉领域的经典网络模型。

3.2 循环神经网络

循环神经网络是一类适用于处理序列数据的神经网络，被尝试应用于天文中时序数据的分析，如利用光变曲线进行暂现源、变星分类[36–38]、强引力透镜的参数估计[39]、引力波信号降噪提取[29]等。

循环神经网络 (recurrent neural network, RNN) 的设计思想是使神经网络拥有记忆，在反向传播网络基础上引入基于时间的循环机制，在隐藏单元中加入一个状态向量。如图 5 所示，每个神经元都把更新的参数传递到下一个时刻，每个隐藏层的输入不仅包括上一层的输出，也跟本层上一时刻的输出有关。RNN通过梯度下降法训练，但是随着时间增加，可能会发生梯度消失，导致时间间隔较长的历史信息无法传递。

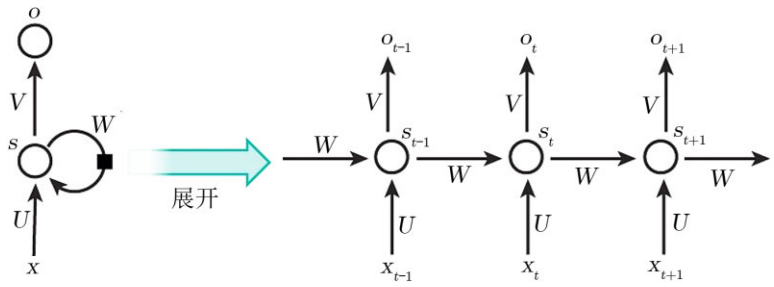


图 5 递归神经网络神经元结构展开示意图[2]

长短期记忆网络 (long short-term memory, LSTM) 是一种基于 RNN 改进的循环神经网络模型，为了改善 RNN 训练中梯度消失或爆炸的情况，它引入了门控自循环单元，在循环节点用多个不同作用的门来控制信息的通过，改变神经元的状态。遗忘门决定前一时间状态中的信息哪些被舍弃，输入门决定什么信息保留在当前状态中。LSTM 能够保留长期记忆，但引入了更多的参数，训练的计算和时间成本也随之增加。

3.3 深度生成模型

机器学习中，判别式模型是根据已知观察变量 x，直接求条件概率分布 $p(y \mid x)$，常用于直接推断所属类别或属性值。而生成式模型是对 $p(x, y)$ 建模，从数据中学习输入输出的联合概率分布，从而得到输入与输出的生成关系，可用于实现图像自动生成、图像信息补全等工作，也可以根据贝叶斯公式 $p(x, y) = p(y \mid x)p(x)$ 求解 $p(y \mid x)$，进而判断类别或属性值。深度学习的网络模型也分为判别模型和生成模型，深度卷积网络以及循环和递归网络即为判别模型。生成模型在无监督深度学习中发挥重要作用，天文中常用的深度生成模型主要

有 VAE 和 GAN 等。

VAE[40] 是一种基于似然的深度生成模型。模型的主要思想是学习数据的低维潜在表示——隐变量 z 的概率分布。假设 z 服从常见的正态分布，训练模型建立 $X = g(z)$ 的概率分布映射。在天文学领域的论文中，VAE 被用来结合高斯混合模型从图像中检测星团[41]、计算星系恒星形成率[42]，以及生成用于暗物质研究的高质量星系图像[43]等。

GAN 是一种基于可微生成器网络和博弈思想的生成式建模方法，由两个神经网络组成——生成器网络和判别器网络。生成器网络试图产生可欺骗判别器的模拟真实样本；判别器网络分别从真实数据和生成器产生的数据中抽取样本并进行区分，判断样本是真实样本而非生成样本的概率。生成式对抗网络于 2014 年诞生，它区别于传统的概率生成模型，不需要经过计算复杂度很高的马尔科夫链式学习，直接采样和推断，学习效率更高。虽然生成式对抗网络依旧存在着多样性缺失和模式崩塌的问题，也没有一个通用的评价标准用来判断模型是否过拟合，但现在仍然广泛应用于图像合成、文本到图像、图像到图像、视频等的生成，以及自然语言处理等领域。在天文学领域生成式对抗网络被广泛应用，主要是因为监督学习通常需要大量的样本，巡天带来了海量数据，但是大多数的数据没有标签，要想利用机器学习方法建立模型在巡天数据中挖宝，需要通过模拟数据构建样本集来训练模型。于是衍生出一些针对不同任务的生成式对抗网络，如生成引力透镜的 CosmoGAN[39]，系外行星大气参数提取的 ExoGAN[44]，星系图像重建的 GalaxyGAN[45]等。

除了以上几类经典网络模型，其他一些深度学习的思想和方法也被应用于天文数据分析中，用来应对标记样本不足等实际问题，例如迁移学习方法[46–50]。迁移学习方法可以用较少的数据对预先训练好的深度神经网络模型进行微调，解决训练数据不足的问题。

4　深度学习在天文研究中的代表性应用

深度学习在天文数据分析处理中的应用，除了最初的目标分类和参数估计等，近些年来也随着深度学习的发展扩展到更多的应用方向，比如引力波探测、系外行星搜寻、暂现源检测等。目前深度学习在天文数据分析中的应用已经相当广泛，以下重点讨论深度学习方法在天文研究中的代表性应用。

4.1　引力波探测

引力波的探测开启了多信使天文学研究的新途径。灵敏的 LIGO 和 Virgo 激光干涉仪引力波探测器能够检测到微弱的引力波信号，同时也会检测到环境和仪器造成的短时脉冲干扰 (glitches)，它们极易与真实引力波信号混淆。短时脉冲干扰是非高斯的，并且有可能是多种情况导致，因而特征复杂。同时引力波事件检测要求实时性，因此天文学家们尝试应用深度学习技术训练模型，以高效实时地区分真实的引力波信号和其他噪声造成的假信号。

George 和 Huerta[21] 率先尝试将深度卷积神经网络应用于引力波时间序列数据，快速检测引力波信号，见图 6。他们用不同参数生成的引力波理论模板波形注入真实 LIGO 噪声作为训练集，训练两个分别用于分类和回归的深度卷积神经网络，从充满噪声的时序信号流中

实时检测引力波信号，同时估计源的质量等参数。测试表明，这种应用深度卷积神经网络的方法在真实 LIGO 数据流上与模板匹配方法的灵敏度相差无几，但误判率更低，且在计算效率上有很大提升，能够实时处理非高斯噪声的微弱时序信号。George 等人[48]应用迁移学习方法，将用真实世界物体识别图像训练好的卷积神经网络应用于 Gravity Spy 项目的引力波频谱图，对脉冲突变信号进行分类和非监督聚类。Razzano 和 Cuoco[23] 采用卷积神经网络对时频图进行建模，对短时脉冲干扰进行分类，并在模拟的短时脉冲干扰上进行了测试。实验显示，此方法能准确快速地对短时脉冲干扰进行分类，效果优于支持向量机、逻辑回归和随机森林等传统机器学习方法，平均准确率超过 99%，可用于开发引力波实时检测工具。

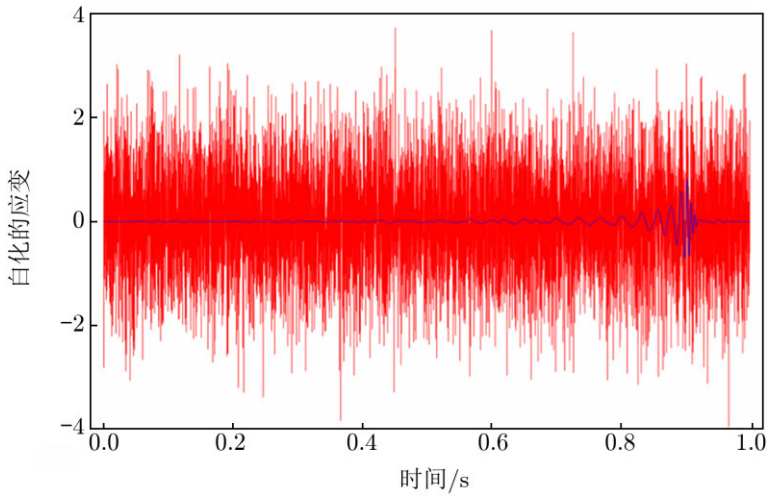


图 6 模拟的引力波信号注入真实 LIGO 噪声[21]

LIGO 和 Virgo 探测器有数十万个辅助信号通道。由于数据量庞大，无法通过人工进行检测。深度学习方法不仅可以帮助我们从这些数据中辨别真实引力波信号和短时脉冲干扰信号，还能对短时脉冲干扰的成因进一步分类分析。随着观测数据的积累，黑洞系统的质量、距离、位置、预计合并时间等参数也将可以通过数据来测量估计，使我们能够将随动望远镜指向对应天区并对整个事件进行观测。在未来引力波数据分析处理中，深度学习方法将有着广阔的应用前景。

4.2 系外行星搜寻

系外行星搜寻是近年来天文学研究的热门方向之一。天文学家发明了视向速度法、凌星法、直接成像法和微引力透镜法等多种方法来搜寻系外行星。其中凌星法是目前找到系外行星最多的方法，其原理是，恒星发光而行星不发光，当行星经过恒星和观测者之间时，会观测到恒星的视亮度有短暂的小幅下降，这种现象称为“凌星”，在光变曲线上大致体现为 U 型。开普勒空间望远镜就是专门为利用凌星法搜寻系外行星而设计的。继它之后，TESS、柏拉图探测器 (Planetary Transits and Oscillations of Stars, PLATO)、LSST 等也在陆续开展系外行星的搜寻。开普勒空间望远镜的数据处理流水线程序根据设定的阈值来发现周期

性凌星事件，超过阈值的事件需要天文学家人工检验每个凌星信号是行星候选体还是其他假信号。在大数据量下这项工作不能完全依靠人工检验，需要自动化的程序来辅助实现。实际上产生凌星现象有多种可能，在光变曲线上可能体现不同的形态，模型需要对噪声有一定的鲁棒性，简单的模板不能完全适应各种情况，于是一些科学家尝试利用深度神经网络学习共性抽象特征来检测系外行星候选体。

谷歌大脑的工程师和天文学家合作验证了利用深度学习方法实现基于凌星法搜寻系外行星的可行性[26]，并发现了 Kepler-80g 和 Kepler-90i。他们利用 NASA 系外行星库①中带标签的凌星事件 (TCEs，即超过阈值事件) 的开普勒光变曲线数据训练了一个卷积神经网络，用来检测一个候选体是真实的凌星系外行星还是其他天文现象或仪器噪声造成的，并根据候选体是真实行星的概率进行排序。用于构建模型的数据集中包括 3 600 个行星候选体和 12 000 多个其他噪声造成的假信号。训练得到的卷积神经网络模型给出的测试集排序可以将 98.8% 的真实行星候选体排在前面。基于深度学习的方法与其他自动化检验方法相比，除了在分类准确率上有所提高，在模型的构建上也免去了人工设计和提取特征的步骤。应用传统机器学习决策树方法的 Autovetter 依赖于开普勒数据处理流水线程序得出的周期、恒星参数 (如有效温度)、信噪比等参数作为特征，而利用卷积神经网络方法可以免于依赖特定流水线程序的产出。Pearson 等人[24]利用如图 7 所示的模拟时序数据训练深度卷积神经网络，从中学习类地系外行星凌星时的测光特征，并用真实的开普勒光变曲线数据验证效果。训练集包含30多万条利用不同参数生成的模拟数据。研究表明，深度卷积神经网络对于未来从大型天文数据集中搜寻系外行星比支持向量机、多层感知机、最小二乘法拟合当现有方法有更高的准确率。

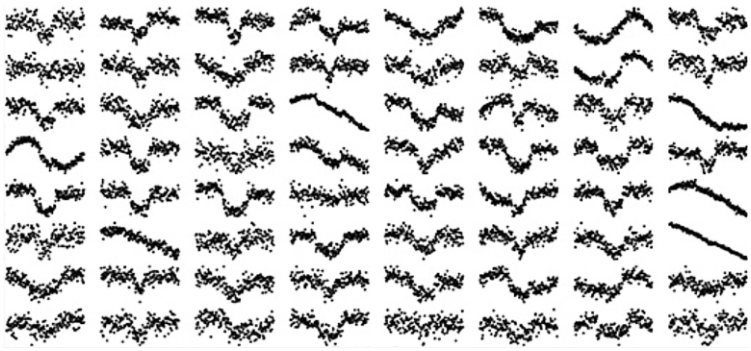


图 7　模拟的时序训练数据[24]

4.3　强引力透镜识别和参数测量

星系尺度的引力透镜系统是研究暗物质分布的重要探针，同时也能提供有价值的宇宙学约束。目前已知的星系尺度强引力透镜很少，传统方法依赖于人工检验发现引力透镜结构。LSST、欧几里得空间望远镜 (Euclid)、大视场红外巡天望远镜 (Wide-Field Infrared Survey Telescope, WFIRST) 等大型巡天项目预计将发现约 10^5 个引力透镜候选体，在巡天图像中

① https://exoplanetarchive.ipac.caltech.edu/

自动检测引力透镜的算法非常重要。自动化方法主要通过从图像中寻找弧状结构，或减去中心星系进行残差分析，也可通过机器学习对强引力透镜的形态参数建模。这些自动化方法各有千秋，受不同数据样本的限制。

Petrillo 等人[10]首次应用卷积神经网络方法进行强引力透镜识别。由于已知的引力透镜真实样本只有几百个，且来源于不同观测项目，没有合适的真实样本集可直接用于训练，作者在千平方度巡天 (KiDS) 真实星系图像基础上合成透镜和非透镜图像 (如图 8 所示)，构建模拟数据集训练卷积神经网络。训练好的模型应用于 255 平方度天区的真实图像数据，开展广泛的引力透镜搜寻。2019 年他们在此基础上对训练样本的复杂度和算法做了进一步改进[11]，在 KiDS 中筛选出的亮红星系测试集上能够找到 3/4 的引力透镜，纯度约为 40%。Lanusse 等人[12]开发了 CMU DeepLens，从图像中自动识别星系-星系强引力透镜系统。利用 20 000 张不同信噪比的 LSST 模拟图像训练深度卷积神经网络模型，拒绝率为 99% 时，能检测出 90% 爱因斯坦环半径大于 1.4″ 且信噪比大于 20 的透镜。Schaefer 等人[13]以 ImageNet 竞赛中取得优异效果的 VGG 网络[33]为基础，加入残差结构和不变性等扩展，提出了基于卷积神经网络的引力透镜识别方法。训练集、验证集和测试集分别包含 17 000, 3 000, 100 000 个模拟图像。在博洛尼亚透镜工厂挑战赛的空间和地面数据集上分别取得 0.94 和 0.977 的 AUC 分数与 0.32 和 0.5 的查全率。除了识别引力透镜，深度卷积神经网络还被用来通过图像估计强引力透镜的参数，并给出参数的不确定性[51, 52]。

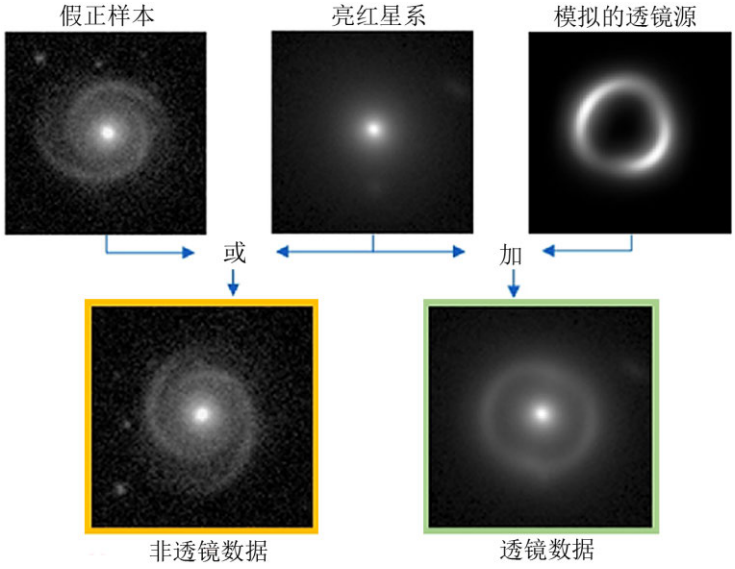


图 8 合成的透镜和非透镜训练数据[10]

4.4 星系形态分类

星系形态分类是应用机器学习方法较早的领域之一。2013 年 Kaggle 数据科学竞赛平台发布了一个星系形态分类算法的竞赛 The Kaggle Galaxy Zoo①，吸引了更多关注。多年来，

①https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge

天文学家们逐步建立并扩充了巡天数据的星系形态星表，数据从初期的光学图像，延伸到了红外和射电等波段，为星系形成和演化的研究提供样本。传统方法通常基于小样本集，需要从原始测光数据中人工提取一系列特征，如椭圆率、聚集度、面亮度等参数，或者对原始数据直接应用主成分分析，然后应用浅层的人工神经网络或支持向量机进行分类。现代大型巡天项目如斯隆数字巡天 (Sloan Digital Sky Survey, SDSS) 提供了大样本集，可用于训练复杂的深层神经网络，直接从原始数据中学习分类特征。

Dieleman 等人[53]利用 60 000 多张带有人工分类标签的 Galaxy Zoo 2 星系图像训练卷积神经网络，建立模型来实现细粒度的星系形态分类，并通过提取旋转不变的特征，增强模型的鲁棒性 (见图 9)。该方法在人工标签共识较高的图像测试集上分类准确率可达 99% 以上。Huertas-Company 等人[4]利用深度卷积神经网络得到在 CANDELS (宇宙近红外超深空组合河外星系遗珍巡天) 5 个观测区域中 50 000 个星系在H波段的形态分类星表。他们利用 GOODS-S 区域约 8 000 个人工分类的星系图像训练网络，然后应用到其他 4 个区域的星系图像。训练得到的模型可以预测每个星系有核球、有星系盘、是不规则星系、致密或点源，以及不可分类的概率。Kim 和 Brunner[6]利用 SDSS 和加拿大-法国-夏威夷望远镜透镜巡天 (the Canada-France-Hawaii Telescope Lensing Survey, CFHTLenS) 的 8 545 个恒星和 57 843 个星系图像组成的数据集，训练深度卷积神经网络，建立恒星-星系图像分类模型。Domínguez Sánchez 等人[9]利用卷积神经网络生成了 SDSS 的 670 000 个星系的形态星表，提供了两种分类方式：哈勃序列 T-type 和星系动物园 2 的形态分类方式。Lukic 和 Brüggen[7]也应用深度神经网络在 60 000 多个 Radio Galaxy Zoo 星系数据集上训练射电星系形态分类模型。Aniyan 和 Thorat[5]用基于 AlexNet 改进的卷积神经网络对展源射电图像按照法纳洛夫-里雷 (Fanaroff-Riley) 类型 (FR Ⅰ, FR Ⅱ) 和弯尾射电星系进行形态分类，在甚大阵 FIRST 巡天数据上分类准确率分别达到 91%、75% 和 95%，查全率分别为 91% (FR Ⅰ, FR Ⅱ) 和 79% (弯尾射电星系)。

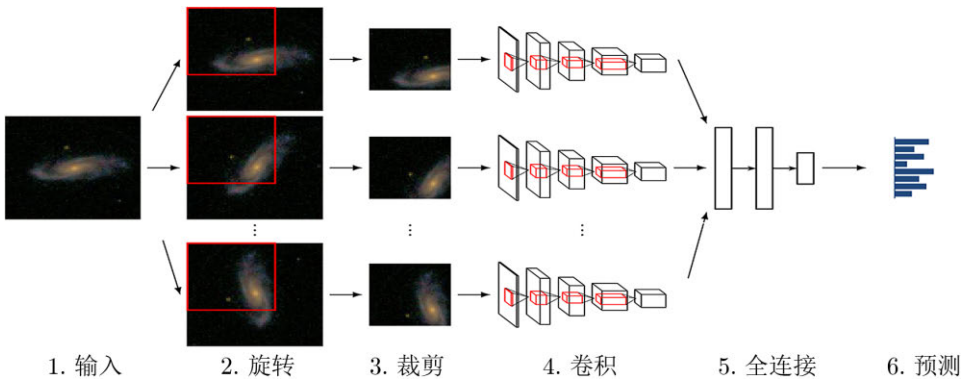


图 9　星系形态分类网络模型[53]

4.5　测光红移估计

在天文学中，红移是天体的电磁辐射由于某些原因导致波长增加的现象，可用来计算天

体的距离。天体的精确红移值需要通过测量光谱中发射线和吸收线的位移得到。但是由于多波段测光数据比光谱更容易获得且成本较低，能够获得更多样本，因此天文学家们也通过测光数据来估计红移值。通常这需要先从测光图像中人工测量出天体的光度、颜色等特征，再应用模板匹配或者传统机器学习方法。近些年，深度神经网络也被用于测光红移估计。不同于传统方法需要从图像中人工提取特征，基于深度卷积神经网络方法的好处是可以完全自动化红移测量流水线程序，避免人工参与，但可能需要更多计算资源。

Hoyle[17]首次直接将多波段星系图像输入一个基于 AlexNet[31]改进的卷积神经网络进行训练，得到模型可预测星系的红移区间。数据集包含 SDSS 的 64 647 个星系的测光参数和图像，该方法预测准确度可媲美效果最好的传统机器学习方法——自适应提升树算法 (AdaBoost)。D'Isanto 和 Polsterer[18]提出了深度卷积网络与混合密度网络相结合的方法，将多波段测光数据直接输入全连接神经网络，直接得到红移的概率密度函数 (PDFs)，而不需要预先进行分类和特征提取。先用卷积层从原始图像中学习特征，然后在全连接层的部分应用混合密度网络得到高斯混合模型参数 (见图 10)。此方法与基于人工特征的随机森林和普通的混合密度网络方法对比，在 SDSS DR9 的星系、类星体和混合数据集上的预测准确度指标均优于其他两种方法，特别是在混合数据集上有较好的表现；也可用于类似的参数估计等场景。Pasquet-Itam 和 Pasquet[16]利用卷积神经网络对 SDSS Stripe 8 中红移已知类星体的光变曲线图像进行训练，然后用于测光红移估计。受样本分布的限制，红移大于 2.5 时，预测效果不如k近邻方法；而红移小于 2.5 时预测效果好于k近邻、支持向量机、随机森林、高斯过程分类器等方法。实验表明，深度卷积神经网络方法将在 LSST 等大样本数据上具有广阔的应用前景。

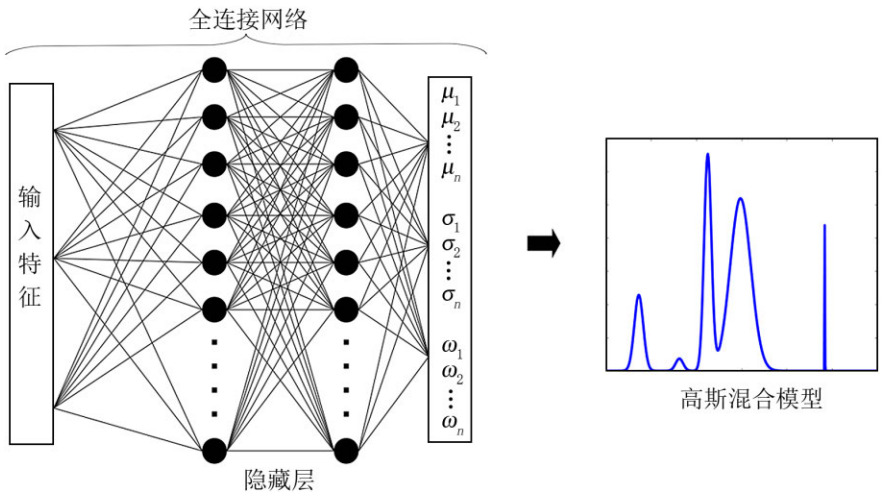


图 10 Deep-HiTS 旋转不变卷积神经网络模型[14]

4.6 暂现源检测

在时域天文学时代，暂现源的研究是一个重要方向。在暂现源的检测中，通常由图像差值方法得到的暂现源候选体中包含了大量的假正样本，且远远高于真实暂现源的比例。由于

真实暂现源的特征复杂多变，难以用一定的规则表示，借助深度神经网络来学习其中的模式成为一种解决方案。

Zhu 等人[27]用卷积神经网络方法进行图像模式识别，寻找脉冲星候选体并将算法集成到 PALFA 巡天数据处理程序里。每个脉冲星候选体由 4 个诊断图表示，可以看作是有上千个像素的图像，由训练好的模型进行预测排序。Cabrera-Vives 等人[14]提出了 Deep-HiTS 模型，应用旋转不变深度卷积神经网络 (见图 11) 检测天文暂现事件。训练好的网络被应用于对高时间分辨率暂现源巡天 (HiTS) 中的暂现源候选体图像进行识别，判断它们是否是真实源，目前流水线程序的准确率为 98.96% ± 0.03%，而 Deep-HiTS 模型的准确率可达 99.45% ± 0.03%，效果优于流水线程序和基于人工特征的随机森林模型。将深度神经网络模型用于新一代巡天如 LSST 等，可能在未知宇宙天体的检测和分类上收获很大。Sedaghat 和 Mahabal[15]用卷积神经网络进行高效的图像差分，用于实时暂现源检测。一个神经网络经过训练可以完成传统的图像处理的全流程，包括图像配准、背景减除、去噪、PSF 匹配减除等步骤。Ackermann 等人[46]用深度卷积神经网络自动检测星系合并，并应用迁移学习方法，用日常物体图像预训练网络，再迁移到星系图像，以此提升小样本集的分类效果。Akeret 等人[28]将卷积神经网络 U-Net 用于识别并减弱射电数据 (时频图) 中的射频干涉信号。Czech 等人[37]利用 CNN 和 LSTM 方法对时域数据中暂现射频干涉 (RFI) 的源进行分类。Connor 和 van Leeuwen[54]集成多个深度神经网络，根据动态频谱和多波束信息进行快速射电暴单脉冲分类，并根据其是否为真实天体物理暂现源的概率进行排序。

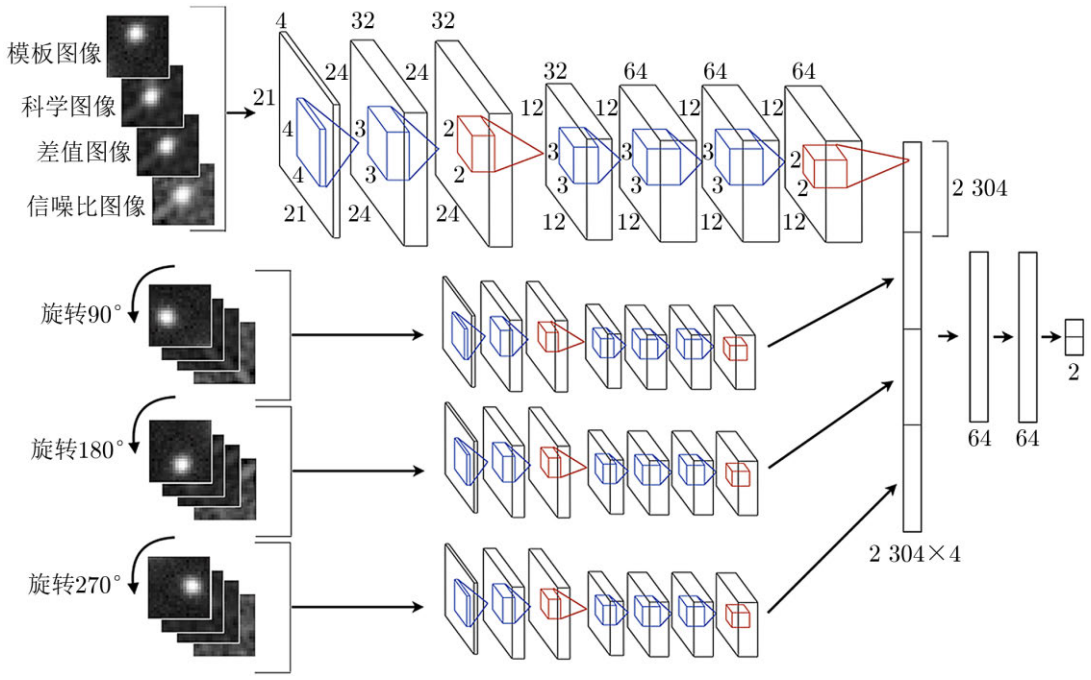


图 11　全连接神经网络和混合密度网络模型[14]

4.7 光谱分类和参数估计

随着大规模巡天项目的开展，获得的光谱数据大量增加。如中国的郭守敬望远镜 (LAMOST)，一次曝光可以获取 4 000 个天体的光谱。随着高光谱获取率设备的使用，准确快速的光谱自动分类和参数估计非常重要。经典的卷积神经网络模型主要是针对二维图像数据设计，从原始图像中学习特定的模式，而光谱数据是一维的，因此需要对数据或模型做一些转换。

Hála[30]将一维光谱转换为二维图像，然后应用 LeNet 卷积神经网络对光谱的类型 (恒星、星系、类星体) 进行自动识别，训练集包含在 60 000 多条光谱数据集上测试，准确率接近 95%，验证了将深度神经网络应用于光谱分类的可行性。Parks 等人[19]应用图 12 中的卷积神经网络搜寻类星体光谱中的阻尼莱曼 α 系统并估计其红移、H Ⅰ 柱密度等参数。由于模型是二维的而光谱是一维的，将其中一维设定为 1。此模型对阻尼莱曼 α 系统的检测获得了 97.4% 的准确率。Waldmann[55]提出基于深度置信网络的系外行星发射谱自动分类算法，能够通过光谱识别行星。Fabbro 等人[20]提出了用于恒星参数估计的卷积神经网络模型 StarNet，用 APOGEE 的恒星光谱作为训练数据，对有效温度、重力加速度、金属丰度 [Fe/H] 等参数进行估计，效果和目前 APOGEE 的数据处理流水线程序类似。

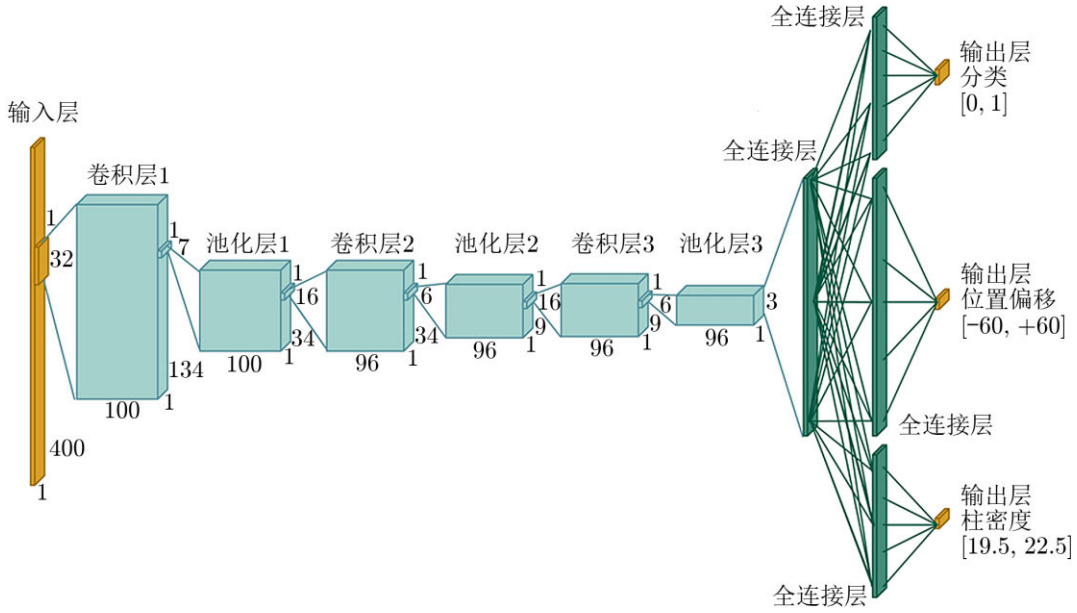


图 12 用于检测类星体光谱中阻尼莱曼 α 系统并进行参数测量的卷积神经网络模型[19]

4.8 太阳耀斑预测

太阳耀斑爆发源自太阳黑子周围大气磁场中存储的能量，然而耀斑的触发机制我们至今还未完全了解。太阳耀斑爆发短时间内释放巨大的能量，可能对航天器、卫星等造成损坏，因此需要及时的太阳耀斑预报，预留充足的时间做出相应的应对措施。传统方法一般

是根据统计数据来预测，如利用不同类型太阳黑子的平均耀斑发生率等数据[56]。近年来，深度学习方法成为一种很有潜力的新方法。近乎实时的太阳观测数据不断积累，为应用深度学习方法提供了数据基础。主要的数据来源包括太阳动力学天文台 (Solar Dynamics Observatory, SDO)、太阳和日球层天文台 (Solar and Heliospheric Observatory, SOHO) 和地球静止环境卫星 (Geostationary Operational Environmental Satellite, GOES)，数据类型包括视向磁图、矢量磁图、各滤光片波段图像、软 X 射线光变曲线等。

Nishizuka等人[57]开发了 Deep Flare Net (DeFN)，一个针对太阳耀斑预测的深度神经网络模型，利用 SDO 和 GOES 2010－2015 年的数据提取的 79 个人工特征来计算在未来 24 h 每个活动区发生耀斑的概率。Huang 等人[58]应用深度卷积神经网络自动提取特征，利用 1996－2015 年 SOHO/MDI 和 SDO/HMI 观测的太阳活动区的视向磁图和对应的 GOES 卫星软 X 射线数据，建立了太阳耀斑预测模型，对未来 6, 12, 24 和 48 h 期间发生 C, M 及 X 级太阳耀斑的概率做出预报，24 h 预测效果与目前最先进的耀斑预报模型[59–61]类似，效率更高。Park 等人[62] 也研究了应用深度卷积神经网络进行太阳耀斑预测，尝试应用计算机视觉领域的经典网络模型 AlexNet 和 GoogLeNet，并提出一个包含多个卷积模块的针对耀斑预报的卷积神经网络结构 (见图 13)。他们提出的模型在准确率、检测概率、临界成功指数、Heidke 技巧评分以及真实技巧统计值等指标上均高于其他代表性耀斑预报模型，包括传统机器学习方法[63]、统计方法[56]、深度学习方法[58]，仅在误报率上略高于 AlexNet 模型。实验说明，相比传统方法深度学习方法能够提升太阳耀斑预报的性能。太阳耀斑预测领域有较统一的数据集和评价指标，不同方法之间可以进行对比，利于深度学习方法的应用和改进。随着数据的积累，预报模型还有较大的改进空间。

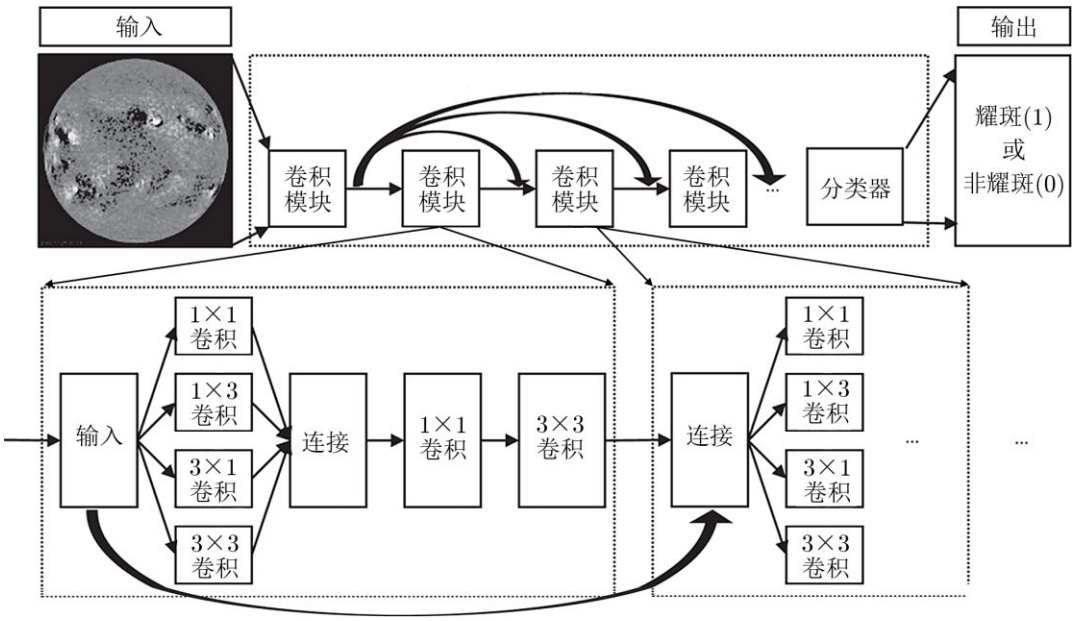


图 13　Park 等人提出的用于太阳耀斑预测的卷积神经网络模型[62]

5 总结、讨论与展望

天文学是典型的数据密集型学科，在很多场景和任务中非常适合使用深度学习。诸多应用实例表明，深度学习的性能可达到甚至超出人们的预期，深度学习方法将在天文学的数据处理中发挥越来越重要的作用。一方面，众多大型巡天计划拥有 TB 甚至 PB 级的数据量，深度学习方法能够在减轻人工负担、提升数据处理效率的同时获得不错的效果。特别是应用于图像、光变曲线等特征复杂的情况。另一方面，随着样本的快速增加，天文研究中还存在大量分类标准和边界并不明确的天体或未知天体。对于未知数据，无监督机器学习是一种重要的工具，而深度学习的自动特征学习比传统的人工特征设计更适合探索未知的领域。

目前在天文学领域深度学习方法被广泛应用于各种数据分析和具体的科研任务，但是还存在着一定的局限性。未来，针对天文科研的应用可以从以下几个方面进行改进。

第一，加强模型的可解释性和严谨性。很多天文学家仍然对深度学习或其他机器学习算法得到的结果持怀疑态度，他们认为深度学习算法是一个黑盒，不具备物理意义上的可解释性，并且对于参数估计应该给出误差分析来描述结果的不确定性。天文学家们较熟悉的基于概率统计的传统建模方法能够根据数据分布的假设求出预测的偏差和标准误差，进而给出置信区间，模型可解释性强。统计方法由于其严谨性，是非常适合用于科学研究的工具，但是它并不能完全适应复杂高维的数据。机器学习建模方法不对数据分布进行假设，无法给出误差条，在一定程度上牺牲了可解释性和科学严谨性，但依靠大量有代表性的数据来训练和验证模型，能更好地拟合复杂非线性关系，因而模型可能有更强的预测能力。尤其是在处理图像数据、时间序列数据等具有复杂特征的数据时，统计方法无法找到有效的数据分布假设构建模型，而机器学习和深度学习能够基于大量训练样本给出高效稳定的预测。特别是深度神经网络，在人们无法总结出明确特征的情况下能够从原始数据中逐层学习特征，进而构建高效可靠的模型。一个未来可能的研究方向是将贝叶斯等概率统计理论与深度学习结合，应用于海量天文数据分析中。贝叶斯深度学习可以对权重和偏置的分布进行多次采样，从而得到多个参数组合，也能够给出结果的不确定性。同时，为加强深度神经网络模型的可解释性，可以利用可视化方法来解释模型学到的特征和模型预测的关注点，而不仅仅是做出精确预测，这样才可能对天文现象的研究做出更大贡献。

第二，更好地利用海量的无标签数据。虽然天文学进入了大数据时代，大型巡天项目的数据产生率可达到TB级，但是已知的天体星表还很有限，与很多其他领域一样，更多的是无标签的数据。因此可以更多地探索无监督学习或半监督学习，减少监督训练需要的标签数据，充分利用无标签的数据，从海量数据中发现新的结构特点。

第三，建立天文学领域机器学习问题标准数据集。目前深度学习天文学领域的深度学习模型通常与科研人员的某一项具体研究任务相关，对于共性问题还没有建立起统一的数据集，提出的方法类似，却不容易互相进行比较。正如ImageNet数据集推动了计算机视觉领域机器学习算法的飞速发展，未来如能总结形成一些经典任务的标准数据集，可更好地促进天文学领域中深度神经网络等机器学习方法的研究和应用。

另外需要注意的是，深度学习有一定的适用条件。首先，深度学习主要适用于不能明确从原始数据中提取特征的情况，或者是人工特征提取过程较难、人工设计特征建模效果不理想的情况。如果已有合适的特征来描述样本，也可以应用其他传统机器学习方法，模型可能有更好的解释性。使用深度学习也应尽量以常用的传统机器学习算法，如随机森林、支持向量机、逻辑回归等方法作为基准进行对比。同时，实际应用中也可以运用多种机器学习方法并进行多模型融合，以提升效果和稳定性。其次，深度神经网络需要训练的参数随着网络层数增多，需要较大的样本量 (通常在 10^5 以上) 来训练模型，在小样本的情况下模型可能会过拟合。深度学习在应用过程中还需要注意训练过程的科学性，避免产生不严谨的科学结论和成果。训练深度学习模型时，数据集应该严格区分训练集、验证集、测试集，保证模型的可靠性。在训练过程中，根据训练集和验证集上的错误率综合判断模型是否过拟合，并合理采取正则化、交叉验证、漏码、提前中断学习等手段避免过拟合，确保模型有一定的泛化能力，即使模型不仅仅符合当前的数据，对新的数据也能够达到类似的预测能力。

随着计算机技术的发展和突破，深度学习已经日渐成熟，形成了较完善的框架。如 TensorFlow, PyTorch 等框架已把网络的基础元件封装好，用户可快速灵活地搭建自己需要的网络结构。与此同时，针对天文数据的科学分析计算平台也在不断发展。虚拟天文台[64]是利用先进的计算和信息技术把世界上各望远镜观测项目获得的数据资源整合到一起的平台，它的目标是让天文学家们能够方便地获取和共享数据资源，并且突破时空限制，协同开展天文数据分析处理和科学研究。大型数据集通过虚拟天文台框架互联互通，构成了一个全球天文数据网络，而在这些海量的数据中探索发现通常需要用到先进的机器学习方法和工具。随着深度学习被广泛应用于天文数据分析处理任务中，虚拟天文台作为天文数据处理的平台，也开始探索如何为天文学家们提供深度学习和机器学习所需的计算资源和环境。例如，智利虚拟天文台基于 astropy，以及 scikit-learn 和 astroML 等 python 机器学习库，开发了一个天文高级计算方法库 (Advanced Computing for Astronomy Library, ACALib)[65]，并计划加入深度学习网络模型。Škoda 等人[66]提出，虚拟天文台支撑的大规模并行机器学习是大规模巡天时代的关键技术之一。中国虚拟天文台的 Zhang 和 Zhao[67]也指出，大数据时代天文学研究需要数据挖掘算法的支持，虚拟天文台有很强的计算平台基础设施和前期技术积累。Xu 等人[68, 69]正在致力于应用深度学习等方法建立太阳大数据分析平台。然而，目前虚拟天文台对于利用深度学习等机器学习算法开展数据分析和挖掘的支持还远远不够。随着云计算及云超算技术的深入应用，基于虚拟天文台技术的科学平台有望能够真正地融合计算和数据资源，把计算带到数据端，天文学家们利用深度学习等方法探索和分析天文大数据将越来越便捷。相信不久的将来，在深度学习这把利器的助力下，天文学家会开辟出一条平坦大道，做出更多更好的科研成果。

致谢

本文得到中国虚拟天文台、国家天文科学数据中心、中科院科学数据中心提供的数据资源和技术支持。感谢国家天文台-阿里云天文大数据联合研究中心对本项工作的支持。

参考文献：

[1] Mitchell T M. Machine learning, International Edition. McGraw-Hill Series in Computer Science. New York: McGraw-Hill, 1997
[2] LeCun Y, Bengio Y, Hinton G E. Nature, 2015, 521(7553): 436
[3] Goodfellow I J, Bengio Y, Courville A C. Deep Learning. Adaptive Computation and Machine Learning Series. Cambridge, Massachusetts: MIT Press, 2016
[4] Huertas-Company M, Gravet R, Cabrera-Vives G, et al. ApJS, 2015, 221(1): 8
[5] Aniyan A K, Thorat K. ApJS, 2017, 230(2): 20
[6] Kim E J, Brunner R J. MNRAS, 2017, 464: 4463
[7] Lukic V, Brüggen M. IAU Symposium, 2017, 325: 25
[8] Lukic V, Brüggen M, Banfield J K, et al. MNRAS, 2018, 476: 246
[9] Domínguez Sánchez H, Huertas-Company M, Bernardi M, et al. MNRAS, 2018, 476: 3661
[10] Petrillo C E, Tortora C, Chatterjee S, et al. MNRAS, 2017, 472: 1129
[11] Petrillo C E, Tortora C, Chatterjee S, et al. MNRAS, 2019, 482: 807
[12] Lanusse F, Ma Q, Li N, et al. MNRAS, 2018, 473: 3895
[13] Schaefer C, Geiger M, Kuntzer T, et al. A&A, 2018, 611: A2
[14] Cabrera-Vives G, Reyes I, Förster F, et al. ApJ, 2017, 836(1): 97
[15] Sedaghat N, Mahabal A. MNRAS, 2018, 476: 5365
[16] Pasquet-Itam J, Pasquet J. A&A, 2018, 611: A97
[17] Hoyle B. Astronomy and Computing, 2016, 16: 34
[18] D'Isanto A, Polsterer K L. A&A, 2018, 609: A111
[19] Parks D, Prochaska J X, Dong S, et al. MNRAS, 2018, 476: 1151
[20] Fabbro S, Venn K A, O'Briain T, et al. MNRAS, 2018, 475: 2978
[21] George D, Huerta E A. Physics Letters B, 2018, 778: 64
[22] George D, Huerta E A. Physical Review D, 2018, 97(4): 044039
[23] Razzano M, Cuoco E. Classical and Quantum Gravity, 2018, 35(9): 095016
[24] Pearson K A, Palafox L, Griffith C A. MNRAS, 2018, 474: 478
[25] Schanche N, Cameron A C, Hbrard G, et al. MNRAS, 2019, 483(4): 5534
[26] Shallue C J, Vanderburg A. AJ, 2018, 155(2): 94
[27] Zhu W W, Berndsen A, Madsen E C, et al. ApJ, 2014, 781(2): 117
[28] Akeret J, Chang C, Lucchi A, et al. Astronomy and Computing, 2017, 18: 35
[29] Shen H, Zhao Z, George D, et al.APS Meeting Abstracts. College Park, MD, US: APS, 2018: S14.008
[30] Hála P. arXiv e-prints, 2014: arXiv:1412.8341
[31] Alex K, Sutskever I, Hinton G E. Commun. ACM, 2012, 60: 84
[32] Szegedy C, Liu W, Jia Y, et al. arXiv e-prints, 2014: arXiv:1409.4842
[33] Simonyan K, Zisserman A. 3rd International Conference on Learning Representations, ICLR ,Conference Track Proceedings. CA, USA: ICLR, 2015
[34] He K, Zhang X, Ren S, et al.IEEE Conference on Computer Vision and Pattern Recognition,CVPR. Piscataway: IEEE Computer Society, 2018: 770
[35] Huang G, Liu Z, Maaten L V D, et al.IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos,CA: IEEE Computer Society, 2017: 2261
[36] Charnock T, Moss A. ApJ, 2017, 837(2): L28
[37] Czech D, Mishra A, Inggs M. Astronomy and Computing, 2018, 25: 52
[38] Naul B, Bloom J S, Pérez F, et al. Nature Astronomy, 2018, 2: 151
[39] Morningstar W R, Hezaveh Y D, Perreault Levasseur L, et al. arXiv:1808.00011, 2018
[40] Kingma D P, Welling M. 2nd International Conference on Learning Representations, Conference Track Proceedings. Banff, AB, Canada: ICLR, 2014

[41] Karmakar A, Mishra D, Tej A. IEEE Recent Advances in Intelligent Computational Systems (RAICS). Piscataway: IEEE, 2018: 122
[42] Schawinski K, Turp D, Zhang C. American Astronomical Society Meeting Abstracts, 2018, 231: 60
[43] Ravanbakhsh S, Lanusse F, Mandelbaum R, et al. arXiv:1609.05796, 2016
[44] Zingales T, Waldmann I P. AJ, 2018, 156(6): 268
[45] Schawinski K, Zhang C, Zhang H, et al. MNRAS, 2017, 467: L110
[46] Ackermann S, Schawinski K, Zhang C, et al. MNRAS, 2018, 479: 415
[47] Vilalta R. arXiv:1812.10403, 2018
[48] George D, Shen H, Huerta E A. Phys. Rev. D, 2018, 97(10): 101501
[49] George D, Shen H, Huerta E A. Phys. Rev. D, 2018, 97: 027
[50] Lim D S S, Warman G L, Gernhardt M L, et al. Planet. Space Sci., 2010, 58(6): 920
[51] Hezaveh Y D, Levasseur L P, Marshall P J. Nature, 2017, 548(7669): 555
[52] Perreault Levasseur L, Hezaveh Y D, Wechsler R H. ApJ, 2017, 850(1): L7
[53] Dieleman S, Willett K W, Dambre J. MNRAS, 2015, 450: 1441
[54] Connor L, van Leeuwen J. AJ, 2018, 156(6): 256
[55] Waldmann I P. ApJ, 2016, 820(2): 107
[56] Bloomfield D S, Higgins P A, McAteer R T J, et al. ApJ, 2012, 747(2): L41
[57] Nishizuka N, Sugiura K, Kubo Y, et al. ApJ, 2018, 858(2): 113
[58] Huang X, Wang H, Xu L, et al. ApJ, 2018, 856(1): 7
[59] Bobra M G, Couvidat S. ApJ, 2015, 798(2): 135
[60] Muranushi T, Shibayama T, Muranushi Y H, et al. Space Weather, 2015, 13: 778
[61] Murray S A, Bingham S, Sharpe M, et al. Space Weather, 2017, 15: 577
[62] Park E, Moon Y J, Shin S, et al. ApJ, 2018, 869(2): 91
[63] Colak T, Qahwaji R. Space Weather, 2009, 7(6): S06001
[64] Cui C, Zhao Y. IAU Symposium, 2008，248: 563
[65] Solar M, Araya M, Mardones D, et al. Astronomical Society of the Pacific Conference Series. San Francisco: ASP, 2017，512: 261
[66] Škoda P. Astronomical Society of India Conference Series. India: Astronomical Society of India, 2017，14: 73
[67] Zhang Y, Zhao Y. Data Science Journal, 2015, 14: 11
[68] Xu L, Yan Y, Cheng J. Solar-Terrestrial Physics, 2017, 3: 9
[69] Xu L, Yan Y. EGU General Assembly Conference Abstracts. Vienna, Austria: European Geosciences Union, 2018，20: 1352

The Application of Deep Learning in Astronomy

TAO Yi-han[1,2], CUI Chen-zhou[1,2], ZHANG Yan-xia[1], XU Yun-fei[1,2], FAN Dong-wei[1,2], HAN Xu[1], HAN Jun[1,2], LI Chang-hua[1,2], HE Bo-liang[1,2], LI Shan-shan[1,2], MI Lin-ying[1,2], YANG Han-xi[1,2], YANG Si-si[1,2]

(1. National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China; 2. National Astronomical Data Center, Beijing 100101, China)

Abstract: When all fields enter big data era, astronomy also steps into its golden age, the

age of big data in astronomy. Astronomy has become a typical data-intensive science. So far astronomers have begun to utilize big data technology to analyse and process large amounts of observational and scientific product data generated by large digital sky survey telescopes as well as simulation data. In recent years, the rapid development of deep learning and artificial intelligence technologies has promoted their applications in various areas, and a large number of papers on the application of deep learning for data analysis and processing have emerged in astronomical research. This paper summarises the status and trend of deep learning application in astronomical data analysis and processing, astronomical data types and machine learning tasks, deep neural network models that are commonly used in astronomical data analysis. Moreover the representative application and progress of deep learning in astronomical scientific research are presented. The possible application directions in the future are discussed and put forward.

Key words: astronomical data analysis; deep neural network; machine learning; virtual observatory