第42卷，第3期
2022年3月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 42, No. 3, pp699-703
March, 2022

利用高性能混合深度学习网络提升光谱分类性能研究

刘忠宝[1]，王 杰[2*]

1. 北京语言大学信息科学学院，北京 100083
2. 中国科学院新疆天文台，新疆 乌鲁木齐 830011

摘 要 随着观测设备的不断完善，人们获得的光谱数量持续上升，如何进一步提高光谱自动分类的性能引起广泛关注。为此，以恒星光谱为研究对象，在近年来新出现的BERT和CNN等深度学习模型的基础上，试图融合了BERT模型和CNN模型在特征提取和智能分类方面的优势，提出高性能混合深度学习网络BERT-CNN，用以探讨该模型在提升光谱分类性能方面的有效性。该模型首先将恒星光谱数据输入BERT模型；然后，利用BERT模型中的Transformer进行特征提取，得到特征向量；最后，将特征向量输入CNN模型，通过softmax分类器获得分类结果。该实验的编程语言为Python3.7，引入TensorFlow1.14作为深度学习模型框架，并以SDSS DR10中的K型、F型、G型的恒星光谱数据作为实验数据集。使用min-max方法对恒星光谱数据做归一化处理，通过与SVM、CNN等分类模型的比较来验证BERT-CNN混合模型在恒星光谱分类中的有效性。引入网格搜索和10折交叉验证来获得模型的实验参数。实验包括两部分：一是利用精准率P、召回率R、调和平均值$F1$等指标对BERT-CNN模型的恒星光谱分类性能进行评价。当训练数据集占比实验数据集的30%～70%时，BERT-CNN模型处理K，F和G型恒星光谱数据集的精准率P、召回率R、调和平均值$F1$随训练样本数的增加而提升。在相同规模的训练样本条件下，BERT-CNN模型在K型恒星光谱数据集上的P，R和$F1$值均最高，其次是G型恒星光谱数据集，F型恒星光谱数据集上的分类效果较差。二是利用准确率对SVM，CNN和BERT-CNN等模型的对比实验结果进行评价。对K，F和G型恒星光谱数据集上，BERT-CNN模型分类效果最优，其次是CNN模型，SVM模型分类效果较差。表明，BERT-CNN模型有助于提升光谱分类性能。

关键词 光谱分类；深度学习网络；BERT模型；CNN模型

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)03-0699-05

引 言

随着海量光谱数据的不断涌入，如何进一步提高光谱分类性能引入广泛关注。目前，国内外有关恒星光谱分类的研究已有不少成果。Daniel等探讨了降维技术在恒星光谱分类中的有效性问题，他们引入局部线性嵌入技术，通过保持高维光谱数据在低维空间的局部结构，进而实现恒星光谱的自动分类[1]。Navarro等利用人工神经网络对低信噪比的恒星光谱进行分类[2]。Sanchez等试图利用k-均值聚类算法对SDSS SEGUE和SEGUE-2恒星光谱进行无监督分类[3]。鉴于传统分类方法具有较高的时间复杂度问题，Liu等受协同管理思想启发，提出非线性集成学习机，并将该模型应用于恒星光谱分类[4]。Huertas-Company等在支持向量机的基础上提出一种确定星系形态的非参数方法[5]；Peng等利用支持向量机从SDSS、UKIDSS等巡天项目获得的光谱中搜寻类星体候选体[6]；Malek等在VIPERS数据集上利用SVM来将恒星、活动星系核和星系区分开来[7]；Brice等在SDSS数据集上利用K近邻算法和随机森林算法进行对恒星光谱进行分类[8]。

此外，越来越多的研究人员将深度学习模型用于解决恒星光谱分类问题。Liu等研究了基于一维卷积神经网络的恒星光谱分类方法[9]。王楠楠等探讨了卷积神经网络应用于恒星光谱分类的可行性问题[10]。尽管实验结果表明上述模型较之传统机器学习算法具有更优的分类效率，然而受其工作机理限制，该模型在特征提取以及特征理解方面仍存在一定

收稿日期：2021-03-16，**修订日期**：2021-04-22

基金项目：国家自然科学基金项目(11803080)资助

作者简介：刘忠宝，1981年生，北京语言大学信息科学学院教授 e-mail：zbliu@blcu.edu.cn

*通讯作者 e-mail：wangjie@xao.ac.cn

差距，严重影响了该模型分类效率的进一步提升。幸运的是，BERT(bidirectional encoder representation from transformers)模型的出现为解决上述问题提供了可能。鉴于此，本工作提出高性能混合深度学习网络 BERT-CNN，试图充分利用 BERT 模型和 CNN 模型在特征提取和自动分类方面的优势，以期进一步提高光谱分类性能。

1 高性能混合深度学习网络 BERT-CNN

BERT-CNN 模型如图 1 所示。该模型的工作流程为：首先，将恒星光谱数据依次输入 BERT 模型；然后，利用 BERT 模型中的 Transformer(图 1 简写为 Trm)进行特征提取，得到特征向量 T_1—T_N；最后，在 CNN 模型中输入上述特征向量并自动分类，进而得到恒星光谱的分类结果。

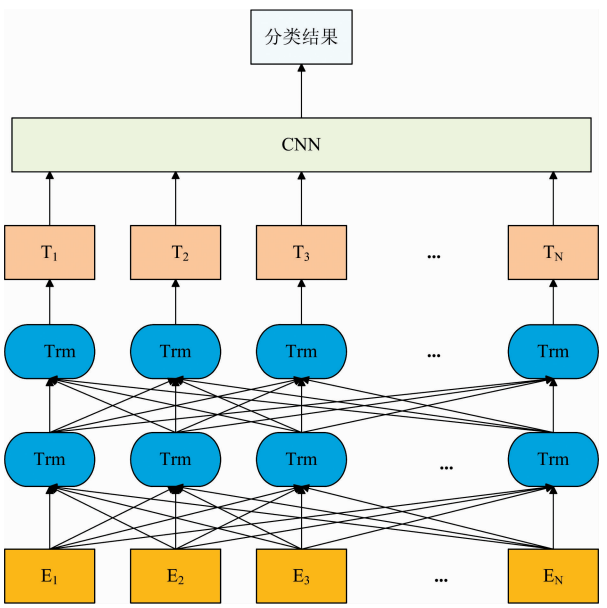

图 1 BERT-CNN 模型

Fig. 1 The structure of BERT-CNN

(1)BERT 模型

BERT 模型采用了多层双向 Transformer 编码器，能够更好地提取恒星光谱数据的深层次特征。Transformer 编码器(以下简称 Transformer)是 BERT 模型最重要的部分，其主要由多头自注意力机制和全连接前馈神经网络层两个子层组成。为了解决随着网络的加深而产生的性能退化等问题，Transformer 在两个子层间加入了残差网络，并在每个子层后添加归一化层来加速模型收敛。

Transformer 基于自注意力机制，该机制更易捕获光谱特征之间的内在关系。其计算过程见式(1)，其中 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 为维度为 d_k 的 Query 矩阵和 Key 矩阵，$\boldsymbol{V}$ 为维度为 d_v 的 Value 矩阵。

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

Transformer 采用多头自注意力机制可以将多个自注意力机制进行横向拼接，以增强模型的关注能力。其原理见式(2)和式(3)，其中，head 表示注意力的头，h 为头的个数，$\boldsymbol{W}_i^{\boldsymbol{Q}}$，$\boldsymbol{W}_i^{\boldsymbol{K}}$，$\boldsymbol{W}_i^{\boldsymbol{V}}$ 为第 i 个 head 的 Query，Key 和 Value 的权重矩阵，$\boldsymbol{W}^{\boldsymbol{O}}$ 为附加权重矩阵，Concat()表示拼接函数。

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}(\text{head}_1,\cdots,\text{head}_h)\boldsymbol{W}^{\boldsymbol{O}} \tag{2}$$

$$head_i=Attention(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}}) \tag{3}$$

层归一化与前馈神经网络的计算过程见式(4)和式(5)。

$$LN(x_i)=\alpha\times\frac{x_i-\mu_L}{\sqrt{\sigma_L^2+\varepsilon}}+\beta \tag{4}$$

$$FFN(x)=\max(0,xW_1+b_1)W_2+b_2 \tag{5}$$

式(4)中，μ 和 σ 为输入层的均值与方差，α 和 β 为待学习的参数，ε 的取值很小；式(5)中，前馈神经网络层以修正线性单元 ReLU 作为激活函数，x 表示网络的输入，W 和 b 为待训练的参数。

(2)卷积神经网络

CNN 模型由输入层、卷积层、池化层以及全连接层组成。输入层为恒星光谱矩阵，矩阵中的每一行向量对应一条恒星光谱。卷积层对输入向量进行卷积操作，进而生成特征向量。卷积计算见式(6)和式(7)，其中 l 为 CNN 的网络层数，j 为特征图，k 为卷积核，b_C 为偏置，N_j 为特征向量集合，ReLU 为激活函数。

$$x_j^{(l)}=\sum_{i\in N_j}a_i^{(l-1)}k_{ij}^{(l)}+b_C^{(l)} \tag{6}$$

$$a_i^{(l)}=\text{ReLU}(x_j^{(l)})=\max(0,x_j^{(l)}) \tag{7}$$

池化层的作用是压缩特征向量的规模，以期达到降低特征向量维度、减少参数规模的目的。该层经过最大池化方法保存局部信息，以期得到池化后的特征向量。在全连接层，将池化后的特征向量进行整合，最后通过 softmax 分类器获得分类结果。softmax 分类器的表达式见式(8)。

$$\text{softmax}(x_j^{(l)})=\frac{\mathrm{e}^{x_j^{(l)}}}{\sum_i\mathrm{e}^{x_i^{(l)}}} \tag{8}$$

2 实验部分

将 Python3.7 作为的编程语言，并使用 TensorFlow1.14 作为深度学习模型框架。实验数据集为 SDSS DR10 中的 K 型、F 型、G 型恒星光谱数据，如表 1(a)—(c)所示。K 型恒星包含 K1，K3，K5 和 K7 次型，而这四种次型光谱的信噪比(signal noise ratio，SNR)区间均是(60，65)；F 型光谱包括 F2，F5 和 F9 次型，其各次型光谱的信噪比区间分别为(50，65)，(65，70)，(75，80)；G 型光谱包括 G0，G2 和 G5 次型，其各次型信噪比区间为(55，65)，(60，65)，(40，70)。

表 1(a) K 型恒星光谱数据集

Table 1(a) The dataset of K stars

Stellar Subclass Type	K1	K3	K5	K7
SNRs	(60，65)	(60，65)	(60，65)	(60，65)
Number	1115	959	850	317

表 1(b)　F 型恒星光谱数据集

Table 1(b)　The dataset of F stars

Stellar Subclass Type	F2	F5	F9
SNRs	(50, 65)	(65, 70)	(75, 80)
Number	1 915	1 671	1 535

表 1(c)　G 型恒星光谱数据集

Table 1(c)　The dataset of G stars

Stellar Subclass Type	G0	G2	G5
SNRs	(55, 65)	(60, 65)	(40, 70)
Number	949	992	600

采用 min-max 标准化方法对恒星光谱数据进行归一化处理，该方法通过对恒星光谱数据进行线性变换，使原始光谱数据保持在[0，1]区间。其计算公式为

$$x_{\text{Norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

其中，x_{Norm}表示恒星光谱数据归一化后的特征值，x 表示原始恒星光谱数据，$x_{\max}$和 $x_{\min}$分别表示每条恒星光谱数据的最大值和最小值。

通过与 SVM、CNN 等分类模型的比较来验证所提模型的有效性。引入网格搜索以及 10 折交叉验证来得到模型的实验参数。在 SVM 模型中，在网格{0.01，0.05，0.1，0.5，1，5，10}中搜索惩罚因子的最优取值，多次实验表明，当惩罚因子等于 0.1 时，模型的性能最优。在 CNN 模型和 BERT+CNN 混合模型中，batch_size 表示一次训练选取的样本数，learning_rate 表示模型的学习率大小，两者均在网格{1×10^{-2}，1×10^{-3}，5×10^{-4}，1×10^{-4}，5×10^{-5}，2×10^{-5}，1×10^{-5}}中选取；hidden_units 表示隐藏层神经元数，在网格{64，128，256，512，1 024}中选取；dropout 为丢弃率，在网格{0.1，0.2，0.4，0.5，0.6，0.8}中选取。表 2 给出了 CNN、BERT-CNN 等模型的实验参数设置。

表 2　CNN，BERT-CNN 模型参数设置表

Figure 2　The parameters of CNN and BERT-CNN

参数	CNN	BERT-CNN
batch_size	128	32
learning_rate	1×10^{-3}	1×10^{-3}
hidden_units	128	256
dropout	0.5	0.5

利用精准率 P、召回率 R、调和平均值 $F1$ 等指标对模型的分类性能进行评价，其中 P 表示正确分为 K(或 F、G)型的光谱数与上述模型分类结果中该类型光谱总数的比值，R 表示正确分为 K(或 F、G)型的光谱数与测试集中该类型光谱总数的比值，$F1=\frac{2\times P\times R}{P+R}$。

当训练数据集占比实验数据集的 30%～70%且剩余数据集为测试数据集时，BERT-CNN 模型的实验结果如表 3(a)—(c)所示，其中括号前的值表示实验数据规模，括号中的值表示占比。

由表 3(a)—(c)可以看出，BERT-CNN 模型的精准率 P、召回率 R、调和平均值 $F1$ 随训练样本数的增加而提升。在相同规模的训练样本条件下，BERT-CNN 模型在 K 型数据集上的 P，R 和 $F1$ 值均最高，其次是 G 型数据集，F 型数据集上的分类效果较差。当训练样本数占比大于等于 50%时，三类数据集上的 P，R 和 $F1$ 值均超过 0.91，这表明 BERT-CNN 模型适用于解决恒星光谱分类问题。

表 3(a)　BERT-CNN 模型在 K 型恒星数据集上的实验结果

Table 3(a)　The experimental results of BERT-CNN on the K-type dataset

Training Size	Test Size	P	R	$F1$
30%(972)	70%(2 269)	0.881 4	0.870 5	0.875 9
40%(1 296)	60%(1 945)	0.906 4	0.897 8	0.902 1
50%(1 621)	50%(1 620)	0.935 1	0.927 1	0.931 1
60%(1 945)	40%(1 296)	0.943 6	0.952 7	0.948 1
70%(2 269)	30%(972)	0.969 6	0.980 9	0.975 2

表 3(b)　BERT-CNN 模型在 F 型恒星数据集上的实验结果

Table 3(b)　The experimental results of BERT-CNN on the F-type dataset

Training Size	Test Size	P	R	$F1$
30%(1 536)	70%(3 585)	0.847 1	0.866 5	0.856 7
40%(2 048)	60%(3 073)	0.875 6	0.890 0	0.882 7
50%(2 561)	50%(2 560)	0.910 1	0.927 9	0.918 9
60%(3 073)	40%(2 048)	0.947 0	0.926 6	0.936 7
70%(3 585)	30%(1 536)	0.956 5	0.970 8	0.965 6

表 3(c)　BERT-CNN 模型在 G 型恒星数据集上的实验结果

Table 3(c)　The experimental results of BERT-CNN on the G-type dataset

Training Size	Test Size	P	R	$F1$
30%(762)	70%(1 779)	0.842 2	0.873 1	0.857 3
40%(1 016)	60%(1 525)	0.892 5	0.905 5	0.899 0
50%(1 271)	50%(1 270)	0.916 1	0.933 3	0.924 6
60%(1 525)	40%(1 016)	0.937 8	0.945 6	0.941 7
70%(1 779)	30%(762)	0.962 0	0.965 9	0.963 9

三类模型的对比实验结果由准确率 A 来评价，准确率是正确分类光谱数与总体测试光谱数的比值。实验数据集的 70%作为训练数据集，剩余数据集作为测试数据集，实验结果如表 4 所示。

表 4　实验结果比较

Table 4　Comparison of experimental results

Stellar Type	Training Size	Test Size	SVM	CNN	BERT-CNN
K	70%(2 269)	30%(972)	0.849 8	0.880 7	0.931 1
F	70%(3 585)	30%(1 536)	0.831 4	0.889 3	0.910 8
G	70%(1 779)	30%(762)	0.871 4	0.904 2	0.923 9
Average classification accuracy			0.850 9	0.891 4	0.921 9

由表 4 可以看出，BERT-CNN 模型分类效果最优，其次是 CNN 模型，最后是 SVM 模型。具体而言，在 K 型数据集上，BERT-CNN 模型比 SVM 模型的准确率高 0.081 3，比 CNN 模型高 0.050 4；在 F 型数据集上，BERT-CNN 模型比 SVM 模型的准确率高 0.079 4，比 CNN 模型高 0.021 5；在 G 型数据集上，BERT-CNN 模型比 SVM 模型的准确率高 0.052 5，比 CNN 模型高 0.019 7。此外，BERT-CNN 模型的平均准确率均最高。这表明，BERT-CNN 模型具有更优的光谱分类性能。

3 结 论

为了进一步提高以 CNN 模型为代表的深度学习模型恒星光谱分类效率，以恒星光谱为研究对象，充分利用 BERT 模型和 CNN 模型在特征提取和自动分类方面的优势，提出高性能混合深度学习网络 BERT-CNN。SDSS 数据集上的实验结果表明，所提模型有助于提升恒星光谱分类性能。上述结论在其他类型光谱上是否成立有待于进一步研究。

References

[1] Daniel S F, Connolly A, Schneider J, et al. Astronomical Journal, 2011, 142(6): 203.
[2] Navarro S G, Corradi R L M, Mampaso A. Astronomy & Astrophysics, 2012, 538: A76.
[3] Sanchez A J, Allende P C. Astrophysical Journal, 2013, 763(1): 50.
[4] Liu Z B, Song L P, Zhao W J. Monthly Notices of the Royal Astronomical Society, 2016, 455(4): 4289.
[5] Huertas-Company M, Rouan D, Tasca L, et al. Astronomy & Astrophysics, 2008, 478(3): 971.
[6] Peng N B, Zhang Y X, Zhao Y H, et al. Monthly Notices of the Royal Astronomical Society, 2012, 425(4): 2599.
[7] Malek K, Solarz A, Pollo A, et al. Astronomy & Astrophysics, 2013, 557: A16.
[8] Brice M J, Andonie R. Astronomical Journal, 2019, 158(8): 188.
[9] Liu W, Zhu M, Dai C, et al. Monthly Notices of the Royal Astronomical Society, 2019, 483(4): 4774.
[10] WANG Nan-nan, QIU Bo, MA Jie, et al(王楠楠，邱 波，马 杰，等). Spetroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(10): 3297.

Research on the Improvement of Spectra Classification Performance With the High-Performance Hybrid Deep Learning Network

LIU Zhong-bao[1], WANG Jie[2*]

1. School of Information Science, Beijing Language and Culture University, Beijing 100083, China
2. Xinjiang Astronomical Observatory, Chinese Academy of Sciences, Urumqi 830011, China

Abstract With the development of observation apparatus, the spectra number rises constantly. How to further improve the classification performance deserves to be research. Because of this, the stellar spectra are taken as the research object, the high-performance hybrid deep learning network is proposed based on integrating the advantages of the BERT model in feature extraction and the CNN model in automatic classification, to verify the effectiveness of improving the spectra classification performance. Firstly, the stellar spectra are input into the the BERT model; And then the part in BERT model named Transformers are used to extract the features and based on which, the feature vectors are formed; Finally, the above feature vectors are input into the CNN model, and the stellar spectra classification results can be obtained with the help of softmax classifier. Python3.7 writes the models used in the experiment and the deep learning framework named TensorFlow is introduced. The K-, F-, G-type stellar spectra in SDSS DR10 are used for the experimental dataset, normalized by the min-max normalization method. The effectiveness of the BERT-CNN model is verified by comparing with the support vector machine models (SVM) and CNN. The performances of the above models are related to the parameters, and therefore, the ten cross-validation and the grid search method are used to obtain the optimal experimental parameters. There are two parts to the experiment. One is to evaluate the classification performances of BERT-CNN with precision P, recall R and $F1$ values. The proportion from 30% to 70% of the experimental dataset is respectively used for the training dataset, and the remainder is used for the test dataset. P, R and $F1$ values rise with the training size on the K-, F-, G-type stellar datasets. In the case of the same training size, the values of P, R and $F1$ arrive at the highest, followed by the performance on the G-type stellar dataset, the classification results on the F-type stellar dataset are much poorer. The other experiment is to evaluate the classification performances of SVM, CNN and BERT-CNN with accuracy. The classification performances of BERT-CNN on the K-, F-, G-

type stellar datasets are all best, followed by CNN. The classification accuracies of SVM are much lower than the other two models. It indicates that the BERT-CNN model contributes to improving the spectra classification performance.

Keywords Spectra classification; Deep learning network; BERT model; CNN model

(Received Mar. 16, 2021; accepted Apr. 22, 2021)

* Corresponding author

(上接 691 页)

论文摘要提交具体步骤

1. 请您在光谱网上(http://www.sinospectroscopy.org.cn)用真实姓名注册，注册系统已经设置认证功能，请用手机号码或邮箱注册。

2. 点击光谱网会议会展栏目，选择《第 22 届全国分子光谱学学术会议暨 2022 年光谱年会》。

3. 点击会议基本情况下的会议快捷通道中的“会议投稿”。

4. 输入用户名和密码登陆，在页面下选择稿件提交。

5. 按照提示提交稿件。

报告形式

为充分提高会议学术交流的效率，会议将采用“口头报告”和“墙报展示”两种方式进行学术交流。无论是口头报告还是墙报展示，均属大会同等学术交流。为尊重个人意见和便于组委会的安排，请大家在会议注册时，提交“口头报告”或“墙报”的题目。为了鼓励博士、硕士研究生积极参与学术交流活动，本次会议将继续设立“优秀青年论文奖”和“优秀墙报奖”，表彰那些研究水平高、能突出研究内容要点、条理清晰的“口头报告”和“墙报”，大会将给获奖作者颁发优秀论文证书和奖金。同时会议还将邀请国内外知名专家学者就光谱有关学术领域的前沿热点问题作大会报告和主题报告。

主要报告形式有：

1. 大会邀请报告：主要邀请国内外知名专家学者报告光谱分析的前沿技术在各个领域的最新研究进展。

2. 主题邀请报告：本次会议将选择光谱技术的热点应用领域，开设多个专题论坛，邀请在该领域的知名专家作论坛主题报告。

3. 口头报告：由参会代表申请、组委会审核方式确定报告人选。

4. 青年论坛报告：为博士、硕士研究生开设交流平台，并评选“优秀青年论文奖”。

5. 墙报展示：作为本次会议的主要交流和展示形式。会议统一安排墙报讲解时间，希望作者按时到位讲解。

重要时间

开通会议注册系统：2022 年 5 月 10 日

论文截稿日期：2022 年 6 月 30 日

第二轮会议通知：2022 年 5 月

第三轮会议通知：2022 年 10 月

会议召开期：2022 年 11 月 11 日—14 日

会议组织机构、注册费及缴纳方式、宾馆住宿介绍及住房预定等信息将在 2022 年 5 月初在会议主页上发布，请您经常浏览光谱网上会议主页，了解会议筹备情况和会议具体安排。网址：http://www.sinospectroscopy.org.cn

产品展示

会议热忱邀请国内外仪器厂商参会及展示仪器设备，大会组委会将在本次会议的网站和会议现场提供展出场所，希望各仪器厂商充分利用这次机会展示自己的最新产品。

(下转 712 页)