基于深度学习的太赫兹时域光谱识别研究

胡其枫，蔡 健

博微太赫兹信息科技有限公司，安徽 合肥 230088

摘 要 太赫兹时域光谱技术，由于其具有物质“指纹谱”特性，是一种可以快速无损地鉴别物质的重要手段，在毒品和爆炸物的无损检测等方面有广阔的应用前景。其中，光谱识别是太赫兹时域光谱技术应用研究的重要方向之一。现有的光谱识别方法多是依靠手工选取特征后进行机器学习分类，或是通过设置吸收峰阈值门限进行判断。由于一些物质在太赫兹波段内并没有明显的吸收峰特征，同时样品浓度、空气湿度、各类噪声等会对太赫兹时域光谱造成干扰从而使信噪比下降，这些方法并不能很好地适应，并且物质类别和数量的增加也会导致计算量不断增加。近年来，随着深度学习技术兴起，以卷积神经网络(CNN)和循环神经网络(RNN)为代表的方法在计算机视觉和自然语言处理等领域得到广泛应用，相比于传统的机器学习方法其效果有了很大的提升。由于深度学习技术强大的非线性分类能力，基于RNN和CNN设计了两个网络用于光谱识别：基于RNN的一维谱线分类网络和基于CNN的二维谱图分类网络。模拟实际应用场景，在非真空环境下采集了12种物质的两万多个光谱数据作为训练集和测试集。在分析了样品浓度、空气湿度对光谱特征的影响后，使用S-G(Savitzky-Golay)滤波对光谱进行降噪。实验结果表明，对比未处理和经过S-G预处理的数据，处理后的光谱特征更加明显，识别准确率更高；与传统的机器学习算法k最近邻(k-NN)方法相比，RNN和CNN方法在测试集上有更好的准确率，且算法速度更快；对于光谱识别，CNN方法比RNN方法能够更好地克服噪声的影响。因此，深度学习技术可以对太赫兹时域光谱进行快速有效的识别，能够为新型无损安全检查技术提供理论和实验基础。

关键词 太赫兹时域光谱；光谱识别；卷积神经网络；循环神经网络；预处理

中图分类号：TP391.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)01-0094-06

引 言

太赫兹波介于远红外和微波之间，频率在0.1～10 THz。在太赫兹光学技术中，太赫兹时域光谱(terahertz time-domain spectroscopy，THz-TDS)技术是目前使用最广泛的技术之一。THz-TDS技术是一种相干探测技术，不同的物质分子被一定频宽的太赫兹波透射过后，会吸收不同频率的太赫兹光波能量，从而产生特征吸收峰，对应的光谱又被称为“太赫兹指纹光谱”。通过对物质“指纹谱”的识别可以实现对毒品和爆炸物等生化危险品进行非接触式无损检测，因此THz-TDS技术受到了警方、海关、安保反恐等部门的高度重视[1]。

总结近年来国内外关于太赫兹时域光谱识别方法的研究，主要集中在一些光谱分析法和机器学习方法相结合的技术[2-4]。马帅等提出一种采用两层受限玻尔兹曼机(restricted Boltzmann machine，RBM)构建深层信念网络模型自动提取太赫兹光谱特征，使用k最近邻(k-nearest neighbor，k-NN)分类器对不同物质进行识别。Yin等[5]提出一种利用遗传算法和偏最小二乘判别分析相结合的方法来鉴别食用油。Mumtaz等[6]通过主成分分析(principal component analysis，PCA)区分了对太赫兹辐射是透明的聚合物。

这些方法往往需要经验丰富的工程师手工设计特征提取器，对于变化的自然数据具有局限性。深度学习方法目前已经成功运用在图像分类、语音识别等领域，不需要人工设计特征提取器，通过一些非线性的结构把原始数据转变成更加抽象的表达，自动提取特征，特别适合自然数据，并且算法性能会随着数据的丰富而提升。太赫兹时域光谱的识别，本质上是一个非线性分类问题，深度学习方法由激活函数引入非线性，更加适合非线性分类问题。作为深度学习的代表方法，卷积神经网络(convolutional neural network，CNN)在太赫兹时域光谱识别上应用的相关文献资料很少，循环神经网络

收稿日期：2019-11-15，**修订日期**：2020-03-12

基金项目：安徽省重点研究和开发计划项目(201904e01020005)资助

作者简介：胡其枫，1991年生，博微太赫兹信息科技有限公司算法工程师 e-mail：fengmaomao1991@126.com

(recurrent neural network，RNN)的应用暂无相关文献报道。

本文设计了两种深度学习识别光谱的网络：基于 RNN 的一维谱线分类网络和基于 CNN 的二维谱图分类网络。由于毒品、爆炸物样本难以获得，以 12 种物质(包含 11 种有机物和空气)为研究对象，通过 Matlab 和 Python 编码实现对物质类别的判定。相较于传统方法，深度学习识别方法能够更好地克服噪声等干扰、耗时不受数据量的影响，准确率更高、速度更快，为太赫兹时域光谱的毒品、爆炸物识别提供参考。

1　实验部分

用于探测物质的 THz-TDS 系统构成如下：由飞秒激光器产生激光脉冲通过分束镜，分为泵浦光路和探测光路；泵浦光入射光电导天线激发出 THz 脉冲，经过一组抛物面镜，对准射向测量样品；探测光与透射样品的 THz 波共同射入探测天线，通过控制时间延迟系统来改变 THz 脉冲和探测光脉冲之间的时间延迟，获得完整的时域光谱。经过傅里叶变换得到频域谱，从中进一步能够获取吸收谱、折射率、透射率等光学参数。

已有研究[7]发现常见毒品在 1.0～2.5 THz 具有不同的特征吸收峰，单质炸药在 0.6～2.3 THz 具有不同的特征吸收峰。由于毒品、爆炸物的样本难以获得，本文以五种酸类物质(抗坏血酸、L-谷氨酸、L-组氨酸、L-苏氨酸、L-酪氨酸)为例，获取其太赫兹吸收谱如图 1 所示。可以看到在 0.5～2.5 THz 频段内，五种物质吸收峰的位置各不相同。因此，根据太赫兹时域光谱来对不同物质进行识别是可行的。由于 2.0 THz 频段之后包含较多的无效信号，识别时需要对光谱进行截取。

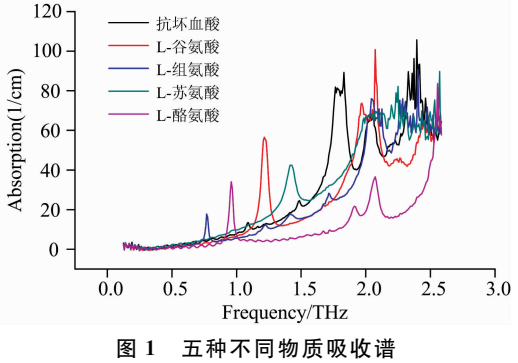

图 1　五种不同物质吸收谱

Fig. 1　Absorption spectra of five acids

太赫兹时域光谱会受到各种噪声干扰[8]：光源漂移带来的本底噪声；平台、器械振动带入的机械噪声；空间电磁辐射带来的电子噪声；光路准直、光学元件带来的衍射噪声等等。另外，样品浓度、空气湿度也对太赫兹时域光谱的“指纹”造成干扰，本文对这两种因素的影响进行分析。

选用在太赫兹波段无吸收特征、基本透明的聚乙烯粉末作为稀释混合剂，将维生素 B2 与聚乙烯粉末按照 1∶1，1∶3，1∶5，1∶7 的质量比(即浓度)进行混合，用压片机将粉末压成片剂。在空气湿度 4%下测试不同质量比的维生素 B2 的太赫兹吸收谱，结果如图 2(a)所示；将维生素 B2 在空气湿度 15%和 70%下分别测试其吸收谱，结果如图 2(b)所示。

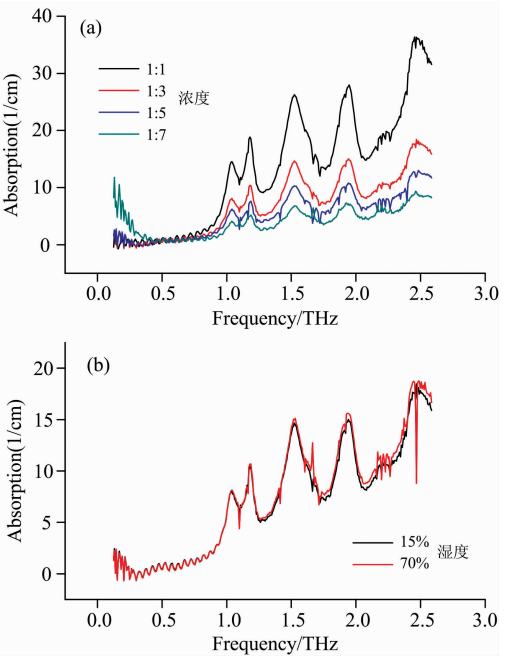

图 2　(a)不同浓度的维生素 B2 吸收谱；(b)不同湿度下的维生素 B2 吸收谱

Fig. 2　(a) Absorption spectra of Vitamin B2 of different concentrations; (b) Absorption spectra of Vitamin B2 in different humidity

从图 2 中可以看出，在 0.5～2.5 THz 频段，吸收峰的位置不随物质浓度的变化而变化，浓度越大、信噪比越好；空气湿度不会消除物质原本的吸收峰[9]，但会引入额外的吸收峰，湿度越大、毛刺(噪声)越多，识别的难度越大。

因此，有必要对光谱数据进行平滑除噪。S-G(Savitzky-Golay)滤波器[10]是一种广义移动平均滤波器，在时域内基于局域多项式最小二乘法拟合的线性滤波器，被广泛地运用于光谱数据平滑除噪。使用 S-G 滤波器对维生素 B2 的吸收谱进行处理，如图 3 所示，黑色曲线是原始数据、红色曲线是处理后数据。可以看出，S-G 滤波后的数据能够保留信号的峰值等重要特征，提高了光谱的平滑性同时降低了噪声干扰。

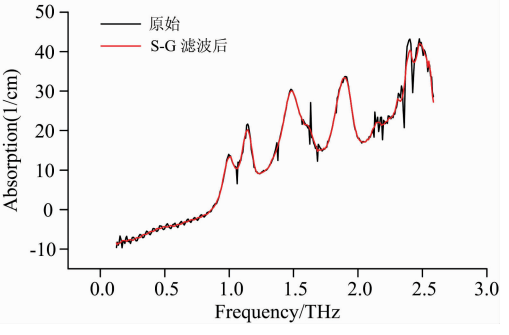

图 3　S-G 滤波结果

Fig. 3　Result of S-G filtering

2 光谱识别

2.1 算法结构

CNN 和 RNN 都是在传统的深度神经网络(deep neural network)基础上发展起来的，是深度学习中最具代表性的两类方法。深度神经网络一般由输入层、隐藏层、输出层构成。输入层一般是原始数据或图像，也可以有一些预处理操作。

CNN 中常见的隐藏层有：(1)卷积层：前一层的特征图与卷积核进行运算，经过激活函数构成该层特征图，过滤有用的特征而抑制无用的特征；(2)池化层：对特征图通过下采样来降低网络的空间分辨率，在保留主要特征的同时减少参数计算量；(3)全连接层：卷积和池化层的输出经过全连接运算，将特征加权映射到样本标记空间。CNN 模拟的是动物大脑的视觉皮层机制，隐藏层的内部神经元之间是无连接的。CNN 在图像分类等视觉任务上取得了较好的性能[11]，可以将光谱数据转化为图像来进行处理。

而 RNN 中的隐藏层，内部神经元之间是有连接的，输入来自上一层的输出和上一时刻本层的输出，通过记忆信息保留序列依赖性。因此，RNN 非常适合研究序列和时间数据，太赫兹时域光谱数据就是有时序关系的数据。

输出层和网络的任务有关，根据需求在特征的基础上增加一层网络用于分类或回归。对不同类别的图像内容进行描述的图像分类[12]任务，使用的就是 softmax 层。softmax 层将多个标量映射为一个概率分布的归一化。假如有 10 种类别，softmax 层的输出就是一个 10 维的向量(n_1，n_2，…，n_{10})，n_1 是属于第一类的概率值、n_2 是属于第二类的概率值……所有类别的概率之和等于 1，即 $\sum_{k=1}^{10} n_k = 1$，概率最高的类别即是输出类别。

一个完整的深度学习算法流程有如下几个步骤：(1)网络结构设计：这是算法的核心部分。根据问题的输入和输出，同时参考经典的网络结构(如 LeNet、VGG 等)，来设计网络的输入层、隐藏层、输出层。(2)数据准备：数据集是由形如(数据，标签值)的向量对构成，数据可能是图像、信号、特征等等，标签值是输入数据的类别。按照互斥同分布的原则和一定的比例，将数据分为训练集和测试集。(3)网络训练和测试：训练集用于训练模型，数据通过网络进行前向传播，损失函数计算前向传播结果与标签之间的误差；反向传播确定梯度向量，进而调整每一个隐藏层的权值；重复前向传播和反向传播，直到损失不再显著下降，将这时网络的权值保存下来，称为模型。最后，用测试集来整体评估模型的性能。

(1)基于 RNN 的一维谱线分类网络

网络结构如图 4 所示，第一层是数据输入层，输入数据是一维数据。第二层是一个长短时记忆(long-short term memory，LSTM)单元，LSTM[13]是传统 RNN 的变体，解决了传统 RNN 训练时间比较长会出现梯度弥散的问题。第三层是全连接层，起到连接所有的特征、将输出值送给分类器的作用，因为我们要解决的是分类问题。第四层是 softmax 层，输出类别和置信度。

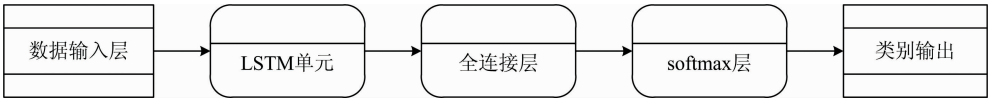

图 4 RNN 分类网络结构

Fig. 4 RNN classification network structure

为了识别多种物质光谱，本文采集了 12 种物质在非真空环境下的太赫兹光谱，每种物质的样本数量需要保持相当，否则算法会朝着样本数量多的物质种类过拟合(overfitting)，过拟合是神经网络常见问题，即模型对训练数据拟合非常好，对测试数据拟合比较差。同时，增加不同浓度的样本以提高识别算法的鲁棒性。

数据格式为一维的太赫兹光谱频域信号，数据取前 256 维，包含主要特征信息。

(2)基于 CNN 的二维谱图分类网络

网络结构如图 5 所示：第一层是数据输入层，输入数据是二维图像。接下来是若干个卷积层、池化层、全连接层、BN(batch normalization)层。使用多个卷积层是为了得到更深层次的特征图。BN 层把所有训练样本的统计分布标准化，降低了不同样本的差异性，使得网络的训练速度加快、效果

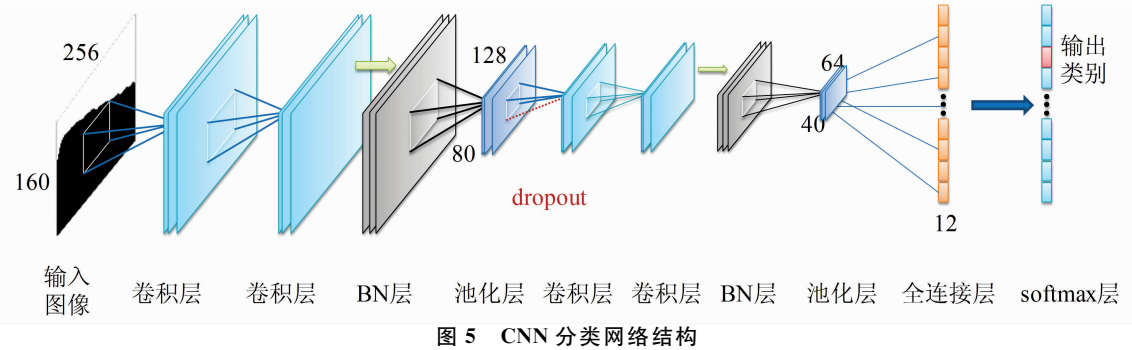

图 5 CNN 分类网络结构

Fig. 5 CNN classification network structure

提升。随机失活(dropout)在训练过程中，随机将部分隐藏层节点的权值归零，能够克服过拟合。最后一层是 softmax 层，输出类别和置信度。

数据格式为二维图像，是由太赫兹光谱频域信号前 256 维转化而来，每一张图像大小固定为 256×160。如图 6 所示，每张图像代表一条光谱信号。

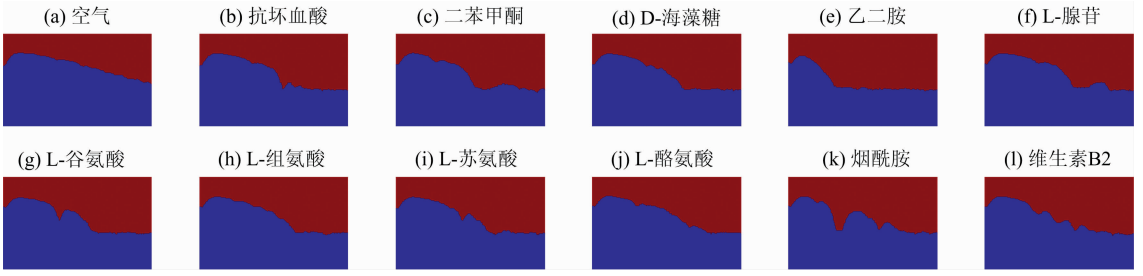

图 6　12 种物质的光谱图像

Fig. 6　Spectral images of 12 materials

2.2　方法

第一步是数据采集和处理：首先进行数据采集，实验设备采用德国 Menlo Systems 公司的 Tera K15 型太赫兹时域光谱仪，动态范围为 60 dB，泵浦激光的重复频率为 100 MHz，波长为 1 560 nm。数据采集时，设定脉冲的时域范围为 80 ps，每 8 组原始数据进行一次平均，此时频率分辨率为 6.25 GHz。在非真空环境下采集 22 491 组 3 种浓度(1∶0，1∶1，1∶2)的如图 6 所示 12 种物质的光谱。数据处理首先去除异常数据，对原始数据进行 S-G 滤波、滤波窗口设为 9；然后将光谱数据截断，取前 256 维的数据作为 RNN 网络输入，变换成 256×160 的图像作为 CNN 网络输入；最后采用“留出法”将数据集划分为训练集和测试集：将数据集中所有类别的数据均按照 4∶1 的比例划分为两个互斥的集合，较多的数据集合作为训练集、较少的数据集合作为测试集。

第二步是模型训练：将训练集数据通过前向传播和反向传播，观察损失变化，保存模型。通过改变数据集和网络超参数等方式，训练多个模型，从中择优。

(1)基于 RNN 的一维谱线分类程序使用 MATLAB 语言编码，调用 Deep Learning 工具箱。学习率设为 0.001。如图 7 所示，训练到 2 000 次 iteration 左右，损失已经不再显著下降，将此时的权值保存作为模型。训练耗时约 20 min。

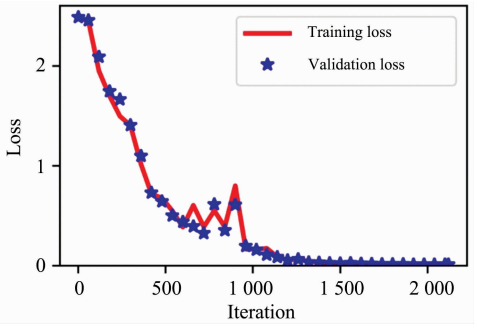

图 7　RNN 网络训练过程

Fig. 7　The training process of RNN network

(2)基于 CNN 的二维谱图分类程序使用 Python 语言编码，基于 TensorFlow 框架。学习率设为 0.000 01，batch 不宜设置过大，否则泛化性不好，本文设为 8。如图 8 所示，训练到 50 个 epoch 左右，损失已经不再显著下降，将此时的权值保存作为模型。训练耗时约 80 min。

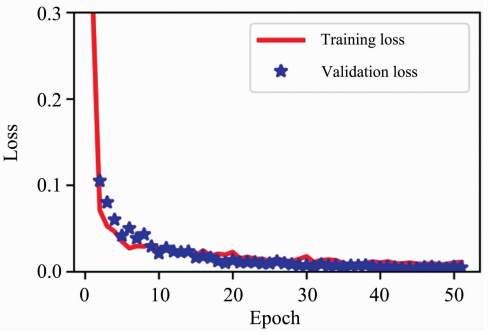

图 8　CNN 网络训练过程

Fig. 8　The training process of CNN network

第三步是模型测试：载入上一步中保存的模型，将测试集数据通过前向传播得到测试结果。

3　结果与讨论

使用经过 S-G 滤波后的 3 种浓度(1∶0，1∶1，1∶2)的测试数据对上述两种算法以及 k-NN 算法进行比较分析。k-NN算法的基本原理是计算测试数据与已知类别的训练数据之间的距离，找到与测试数据距离最近的 k 个邻居，根据邻居所属的类别来判断测试数据的类别。测试结果如表 1 所示。由表可知，k-NN 算法的缺陷在于：测试数据每一次都需要和训练数据逐一进行比较，算法测试耗时与训练集大小有关，训练集越大、测试耗时越长；算法对所有数据逐点进行距离计算，对噪声敏感、泛化性差。而基于 RNN 和 CNN 的算法，训练集只影响训练耗时，测试耗时仅和网络复杂度即层数有关；算法可以克服一定程度的噪声干扰。

为了分析数据预处理对于深度学习方法的影响，分别使用不经预处理的数据集和 S-G 滤波后的数据集，在同样的超参数设置下，使用本文方法进行对比实验，测试结果如表 2 所示。由表可知，数据预处理对于深度学习类方法是有必要的，可以消除噪声对模型的干扰，提升模型泛化性。

表 1 三种算法的性能比较

Table 1 Performance comparison of three algorithms

算法	训练准确率/%	测试准确率/%	测试耗时/ms	时间复杂度	测试环境
RNN	97.6	97.5	<1	$O(1)$	GPU：Nvidia GeForce RTX2080Ti
CNN	99.9	99.6	<1	$O(1)$	GPU：Nvidia GeForce RTX2080Ti
k-NN	100	87.6	>100	$O(n)$	CPU

表 2 数据预处理对于 RNN 和 CNN 性能的影响

Table 2 The effect of data preprocessing on performance of RNN and CNN

数据预处理	RNN/%	CNN/%
无	78.7	83.4
S-G 滤波	95.1	99.6

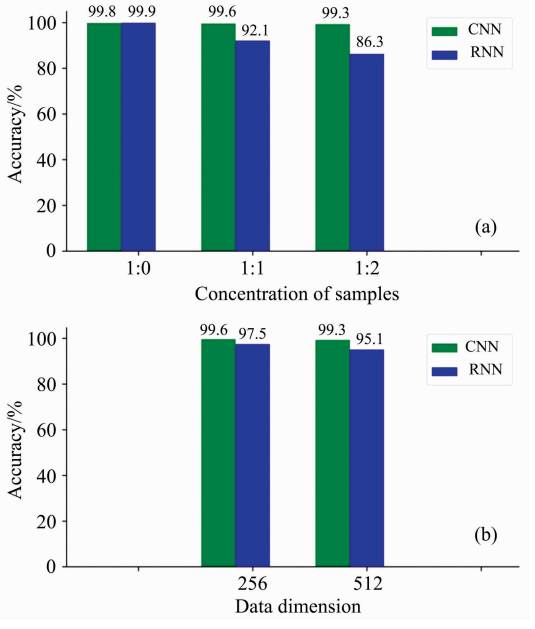

图 9 (a)样品浓度的比较实验；(b)数据维度的比较实验

Fig. 9 (a) Comparative experiment of concentrations of samples; (b) Comparative experiment of data dimensions

为了对比 RNN 和 CNN 方法的效果，在同样的超参数设置下，分别使用不同浓度和不同维度的数据进行对比实验。样品浓度越大，光谱信噪比越好；前 256 维光谱频域信号包含主要特征信息，维度越大，包含的冗余信息越多。测试结果如图 9 所示，对于不同浓度和不同维度的数据，CNN 方法的准确率普遍比 RNN 更高，且更加稳定。因此，在光谱的识别上，CNN 方法要优于 RNN 方法，其泛化能力更好、能更好地克服低信噪比和噪声。

4 结 论

对非真空环境下 12 种物质的光谱进行分析，提出基于 RNN 的一维谱线分类网络和基于 CNN 的二维谱图分类网络。测试结果表明，两种算法均能够实现光谱识别，在测试集上分别能达到 97.5% 和 99.6% 的准确率，算法耗时均小于 10 ms，与传统 k-NN 方法相比准确率更高、速度更快。由于自然环境下的空气湿度和其他噪声干扰，我们对光谱数据进行 S-G 滤波处理，发现处理后的光谱数据特征更加明显，算法的准确率得到提高。进一步对 RNN 和 CNN 方法进行对比分析，发现 CNN 方法能够更好地克服样品浓度和数据维度的影响，比 RNN 方法的鲁棒性更强。

本文探索了两种深度学习算法在光谱识别上的应用，克服了空气湿度和样品浓度对信噪比的干扰，解决了 k-NN 算法速度慢的问题，为太赫兹技术在无损安全检查领域的应用提供了算法基础。

References

[1] Trofimov V A, Varentsova S A. PLOS ONE, 2018, 13(8): e0201297.
[2] Liu J, Li Z, Hu F, et al. Optical and Quantum Electronics, 2015, 47(2): 313.
[3] Yin X, Mo W, Wang Q, et al. Advances in Condensed Matter Physics, 2018, 2018: 1618750.
[4] Liang J, Guo Q, Chang T, et al. Optik, 2018, 174: 7.
[5] Yin M, Tang S, Tong M. Analytical Methods, 2016, 8(13): 2794.
[6] Mumtaz M, Mahmood A, Khan S D, et al. Applied Spectroscopy, 2017, 71(3): 456.
[7] XIE Qi, YANG Hong-ru, LI Hong-guang, et al(解 琪, 杨鸿儒, 李宏光, 等). Optics and Precision Engineering(光学精密工程), 2016, 24(10): 2392.
[8] Naftaly M. IEEE Sensors Journal, 2013, 13(1): 8.
[9] LI Jin, LIU Quan-cheng, XIONG Liang(李 进, 刘泉澄, 熊 亮). Laser & Optoelectronics Progress(激光与光电子学进展), 2018, 55(9): 70.
[10] Liu H, Zhang Z, Yang Y, et al. Optik, 2018, 172: 668.
[11] Yu S, Jia S, Xu C. Neurocomputing, 2017, 219: 88.

[12] Maggiori E, Tarabalka Y, Charpiat G, et al. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(2): 645.
[13] Karim F, Majumdar S, Darabi H, et al. IEEE Access, 2018, 6: 1662.

Research of Terahertz Time-Domain Spectral Identification Based on Deep Learning

HU Qi-feng, CAI Jian
Brainware Terahertz Information Technology Co., Ltd., Hefei 230088, China

Abstract Terahertz time-domain spectroscopy (THz-TDS) is an important method for rapid and nondestructive material identification due to its spectral fingerprint properties, which has broad application exploitation in the nondestructive inspection of drugs and explosives. Spectral identification is one of the most important aspects of the applied research of THz-TDS. Most existing spectral identification methods are machine-learning based classification of manually selected features or thresholding classification of absorption spectral peak. Those methods are not adapt well to low signal-to-noise ratio, because some materials have few or no spectral absorption peaks features in the terahertz waveband and spectra are affected easily by concentrations of samples, air humidity and noises. Meanwhile computational cost increases with data quantity and category. In recent years, with the rise of deep learning technology, the methods represented by CNN (Convolutional Neural Network) and RNN (Recurrent Neural Network) have been widely applied to fields such as computer vision and natural language processing where they have been shown to produce better results than traditional machine learning methods. Due to the strong nonlinear classification capability of deep learning technology, two networks respectively were designed based on RNN and CNN for spectral identification in this paper: one-dimensional spectral line classifier based on RNN and two-dimensional spectral image classifier based on CNN. To simulate the practical application scenario, over 20 000 terahertz time-domain spectra of 12 materials were measured in a non-vacuum environment as training-set and test-set. After analyzing the effects of concentrations of samples and air humidity on spectra, S-G(Savitzky-Golay) filter was introduced to reduce noises of spectra. Experimental results show that S-G filter could improve the identification accuracy, because processed spectra have more obvious feature compared with the unprocessed spectra; the proposed methods based on RNN and CNN are more accurate and faster on the test-set, compared with traditional machine learning algorithm k-NN (k-Nearest Neighbor); CNN demonstrated better robustness to noises than RNN on spectral identification task. Therefore, deep learning technology could be utilized for quick and effective identification terahertz time-domain spectra, which provide a theoretical and experimental basis for new nondestructive safety inspection techniques.

Keywords Terahertz time-domain spectroscopy; Spectral identification; Convolutional neural network; Recurrent neural network; Preprocessing

(Received Nov. 15, 2019; accepted Mar. 12, 2020)