

文章编号: 2095-4980(2021)01-0022-07

一种面向电磁识别模型的分散计算方法

陆鹏威^{1a}, 颜子彦^{1a}, 张伟², 曾歆^{1b}, 史清江^{*1a}

(1.同济大学 a.软件学院; b.电子与信息工程学院, 上海 201804; 2.电子信息控制重点实验室, 四川 成都 610036)

摘要: 基于张量分裂技术, 设计了一种面向电磁目标识别的神经网络模型分散计算方法。该方法根据不同的隐藏层选择特定的张量分裂方法, 将权重无损地分散到多个分布式节点上, 以分散、聚合的方式完成协同推理计算。在树莓派设备上进行的仿真实验表明, 该方法可以对集中式电磁识别模型进行无损拆分并分布式部署, 可以保持与原始模型完全相同的准确率。并且当原始模型由于参数量过大而无法加载到内存中进行处理时, 该方法仍可以正常完成计算。

关键词: 边缘计算; 分布式计算; 神经网络; 张量分裂

中图分类号: TP391.4

文献标志码: A

doi: 10.11805/TKYDA2021153

Decentralized calculation of neural network model for electromagnetic object detection

LU Pengwei^{1a}, YAN Ziyang^{1a}, ZHANG Wei², ZENG Xin^{1b}, SHI Qingjiang^{*1a}

(1a.School of Software Engineering; 1b.School of Electronic and Information Engineering, Tongji University, Shanghai 201804, China; 2.Key Laboratory of Electronic Information Control, Chengdu Sichuan 610036, China)

Abstract: Based on tensor splitting technique, a decentralized computing method of neural network model for electromagnetic object detection is introduced. In this method, different tensor splitting techniques are selected according to different hidden layers, and the weights are distributed to multiple distributed nodes losslessly. The simulation results on Raspberry PI show that this method can decompose and deploy the centralized detection model losslessly, and ensure the same accuracy as the original model. And when the original model is too heavy to be loaded into memory for calculation, this method can still complete the calculation properly.

Keywords: edge computing; distributed computing; neural networks; tensor splitting

基于神经网络的电磁目标识别对于边缘设备一直是很大的挑战, 其主要原因在于边缘设备的计算资源和存储资源有限, 而神经网络模型又往往需要很高的计算量和存储空间。如果使用云服务器与边缘设备相结合^[1-2]的方案来部署神经网络, 会引入新的通信方面的问题与挑战。虽然云边结合的方式可以满足部分场景的需要, 但并不适用于所有的场景, 如云端节点与智能体之间距离远大于智能体节点相互之间的距离, 会导致云边通信有较长的时延且通信极易受到干扰, 无法保证实时性与可靠性。不仅如此, 云边结合方案十分依赖云端节点, 一旦云端计算机受到干扰或因其他原因而断开通信, 整个系统将处于瘫痪状态, 鲁棒性较差。

为解决上述问题带来的挑战, 可通过优化边缘设备端的算法与协作方式, 减小神经网络模型的计算量与存储空间^[3-5]。

本文根据电磁目标识别的需求特征, 设计了一种基于张量分裂的神经网络模型分散计算方法, 将神经网络的数据张量与权重张量以特定的方式进行不同维度上的分裂, 并部署于多个智能体节点中, 通过节点之间的数据传递与整合, 实现与集中式模型完全相同的感知结果, 完成边缘端对电磁目标的分布式协同感知, 将深度神经网络计算从云边结合的方式推向纯边缘。仿真结果表明, 本文提出的方法降低了电磁目标识别任务中边缘计算对云资源 and 高质量网络的依赖性, 提供了一种处理边缘设备采集得到的原始数据的解决方案。

收稿日期: 2021-04-15; 修回日期: 2021-06-14

基金项目: 基于物联网和大数据的智能建筑管理关键技术研究资助项目(2017YFE0119300); 上海市青年科技英才扬帆计划资助项目(19YF1451500)

*通信作者: 史清江 email:shiqj@tongji.edu.cn

1 全连接层张量分裂计算

1.1 全连接层矩阵计算

全连接层的计算如式(1)所示：

$$a = Wx + b \tag{1}$$

式中： W 为权重； x 为输入数据； b 为偏移量； a 为输出数据。

在电磁目标识别模型中，全连接层的张量计算和张量在边缘设备内存中的存储对于单个设备而言是较大的负担。在实际场景中，由于数据分布式存储在多个边缘设备节点上，如果使用集中式计算模型，需要将不同设备的数据进行通信合并，作为一个张量输入全连接层中进行计算，这将在数据处理阶段就带来大量的数据通信开销。而且尽管集中式模型的识别性能好，但单个边缘设备由于计算资源的限制往往无法承载这种集中式模型的计算量。对于一个计算量和存储量要求较高的集中式电磁目标识别模型，本文通过将模型全连接层计算量分散到不同的边缘设备，辅以一定的信息交互，协同地完成整个计算。本文将全连接层的张量分裂计算方法根据全连接层的拆分维度和输入数据的情况分为输入张量分裂计算方法和输出张量分裂计算方法。

1.2 输入张量分裂

在全连接层的输入张量分裂中，基本思想是将权重 W 和输入数据 x 进行相应的分块，分成 w_i 和 x_i ，节点使用分割后的子矩阵进行计算，每个边缘设备进行 $w_i x_i$ 计算，当每个节点完成计算后，再将所有节点对应的部分相加并合并结果。对于 n 个节点，最后的结果如式(2)所示：

$$a = \sum_{i=1}^n w_i x_i + b \tag{2}$$

输入张量分裂计算方法如图 1 所示。对于一个输入，将输入数据与全连接层全部进行拆分，分别完成计算后，再将所有的计算结果汇总求和，得到最终的识别结果。在全连接层输入张量分裂中，一个节点需运行该层所有神经元的部分矩阵计算。图 1 所示的计算示例中，一个节点计算所有神经元所需矩阵乘法的 1/4，图 1 中虚线框内为一个节点的计算内容。

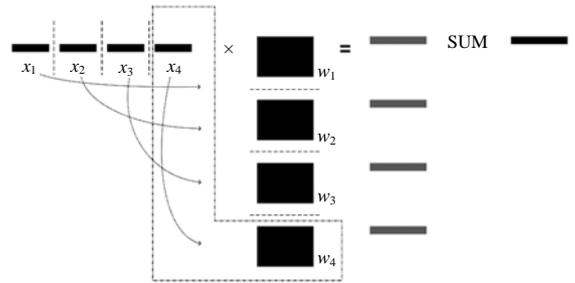


Fig.1 Input splitting
图 1 输入张量分裂

1.3 输出张量分裂

在电磁目标识别模型中，除了对输入张量进行分裂计算外，还可将全连接层的输出进行张量分裂，同样减少单个设备上的计算和存储量。在实际场景中，给定一个识别性能较好但计算量较高的集中式电磁目标识别模型，通过将模型全连接层以一种与输入张量分裂不同的方式进行分裂并分散到不同的边缘设备，同样辅以一定的信息交互协同完成整个计算，最后将不同边缘设备的计算结果进行整合得出最终感知结果。

与输入张量分裂计算不同的是，在全连接层的输出张量分裂中，只将 W 进行相应的分块表示，每个边缘设备的输入 x 相同。每个边缘设备进行 $w_i x$ 计算，当每个节点完成计算后，再将所有节点对应的部分通过直接相连的方式合并得到结果。对于 n 个节点，最终结果如式(3)所示：

$$a = (w_1 x, w_2 x, \dots, w_n x) + b \tag{3}$$

输出张量分裂计算方法具体如图 2 所示。在全连接层输出张量分裂中，并行化该层每个神经元的计算，每个节点进行该层的一个或多个神经元的完整计算。在这种方法中需要将完整的输入数据传输给每一个节点。图 2 所示的计算示例中，每个节点拥有完整的输入数据 x ，并且拥有对应于输出数据的权重。在每个节点完成计算后，结果数据将被传输至汇总节点，使用拼接的方式得到最终结果。在这一方法中，可以选择在每个节点上应用或者是合并之后再应用激活函数。图 2 中虚线框内为一个节点的计算内容。

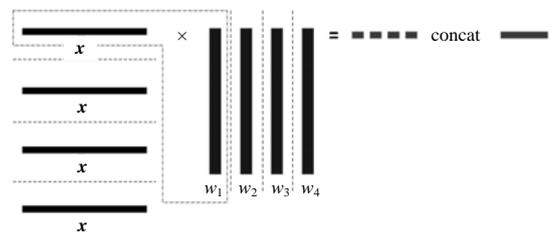


Fig.2 Output splitting
图 2 输出张量分裂

1.4 全连接层张量分裂方法对比

全连接层不同张量分裂方法的性能比较如表 1 所示。其中 d_i 和 d_o 分别为全连接层的输入数据维度和输出数据维度。

表 1 全连接层的输入张量分裂与输出张量分裂的对比
Table 1 Comparison of input splitting and output splitting of fully connected layers

name	#node	calculation(per node)	input communication	output communication	merge operation	distributed activation
baseline	1	$d_1 d_o$	d_i	d_o	N/A	N/A
output splitting	n	$d_i(d_o/n)$	nd_i	d_o	Concat	Yes
input splitting	n	$(d_i/n)d_o$	d_i	nd_o	Sum	No

两种全连接层张量分裂方法相对集中式计算方法都减少了对单个节点的内存占用和计算量。如何选择两类分裂方法的一个重要根据是输入和输出数据的维度，输出分裂的总通信量为 nd_i+d_o ，输入分裂的总通信量为 d_i+nd_o ，若输入数据维度较输出数据维度小，则输出分裂效果更好，反之则输入分裂效果更好。值得注意的是，输出张量分裂方法可以在节点内直接调用激活函数，而输入张量分裂方法则必须先汇总计算结果后再进行激活，因此不能在节点内完成激活^[6]。

2 卷积层张量分裂计算

2.1 卷积层矩阵计算

卷积层由一系列滤波器成，也称为卷积核。滤波器会按照一定的步长从输入张量的宽度以及高度两个方向进行滑动，并将自己的参数与该部分对应的数值进行点乘运算，将所有的乘积相加得到一个输出。每个滤波器遍历整个输入张量后，会得到输出特征图中的一个通道。因此卷积层的输出通道维度仅取决于该层滤波器的个数。

在电磁目标识别模型中，卷积层的张量计算会消耗设备大量计算资源。实际的卷积层计算中，滤波器在输入张量的宽高方向滑动进行卷积运算，提取到上一层张量的局部特征后，还需要与偏置值相加，再通过一个激活函数才能得到当前层的特征图。卷积层的计算中输入数据的维度为 $H_i \times W_i \times C_i$ ，滤波器的维度为 $H_f \times W_f \times C_f \times k$ ， k 为滤波器的个数，输出数据的维度为 $H_o \times W_o \times C_o$ ，并且根据卷积层定义可知 $C_f=C_i$ ， $C_o=k$ 。卷积层的计算首先将一个滤波器置于输入数据的左上角进行内积操作，得到结果后按照一定的步长从输入张量的宽度以及高度两个方向进行滑动，每次滑动都伴随着一次新的内积操作并得到一个新的值。将所有遍历得到的计算结果构成矩阵并与偏置值 b 相加得到一个输出特征矩阵。每个滤波器都对应着一个输出特征矩阵，将这些特征矩阵叠在一起便得到该卷积层最终的卷积运算输出矩阵。为了减少单个设备上的计算量，降低设备负担，提升整体性能，可以通过卷积层的张量分裂方法对卷积层进行拆分。卷积层的分布式运行可从滤波器、输出数据和输入数据等角度对张量进行分裂，通过滤波器张量分裂、通道张量分裂和空间张量分裂的方法来提高设备的计算性能。

2.2 滤波器张量分裂

本节将卷积层的滤波器和输入数据进行张量分裂，减少单个设备上的计算量，提升整体计算性能。卷积层滤波器张量分裂方法具体如图 3 所示。其中，输入数据的维度为 $H_i \times W_i \times C_i$ 。在图 3 中，以维度为 $H_f \times W_f \times C_f \times 1$ 的滤波器为例，有 3 个节点，在滤波器拆分方法中输入数据和滤波器都在第三维被等分为 3 个部分，各部分的输入数据和滤波器第三维长度都分别被标为 C_1, C_2 和 C_3 ， $C_i=C_f=C_1+C_2+C_3$ ，每个部分分别由不同节点进行卷积计算。由于输入数据和滤波器元素之间存在一对一的对应关系，因此每个节点都会计算部分输出。最后，为了得到最终输出结果，需要对所有对应的张量元素对应位置求和来合并结果。图中相同颜色的块为各节点的计算内容。

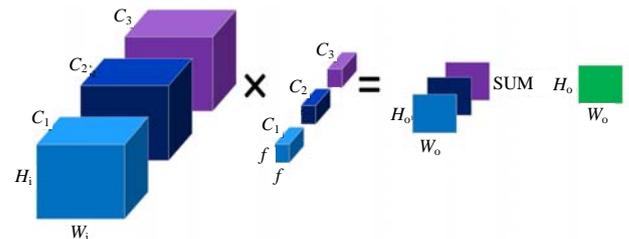


Fig.3 Filter splitting
图 3 滤波器张量分裂

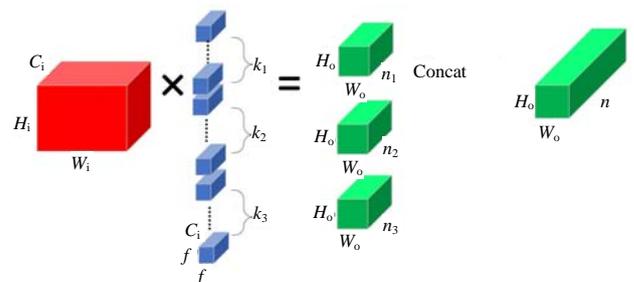


Fig.4 Channel splitting
图 4 通道张量分裂

2.3 通道张量分裂

在卷积层的计算中，计算结果的通道数等于滤波器的个数。通道张量分裂仅对卷积层的通道数量进行拆分。在通道张量分裂计算方法中，每个节点用相同的输入数据计算卷积层输出数据中不重叠的一组通道，并利用不同的滤波器进行计算。这一方法具体如图 4 所示。图 4 中，原始卷积层一共拥有 k 个滤波器，拆分之后，每

个边缘设备节点分别拥有 k_1, k_2 和 k_3 个滤波器，其中 $k=k_1+k_2+k_3$ 。各节点拥有一份完整的输入数据，利用节点拥有的滤波器分别计算各自的输出，得到的输出数据维度为 $H_o \times W_o \times k_i$ 。最后，为了得到最终输出结果，需要将不同节点对应的张量元素拼接来合并结果。

2.4 空间张量分裂

与上述两种张量分裂方法不同，空间分裂仅对输入数据进行拆分。在空间张量分裂计算方法中，每个节点计算卷积层输入数据中的一个部分。在这一方法中，节点存储的权重与原始集中式模型完全一致，仅有输入数据被拆分，具体如图 5 所示。在图 5 中，将卷积层输入数据的长和宽两个维度分别分割成 2 份，输入数据一共被分割为等量的 4 份数据。每个节点将其中的一份输入数据作为卷积层的输入，对其进行卷积计算。最后，需要将不同节点对应的张量元素拼接来得到结果。

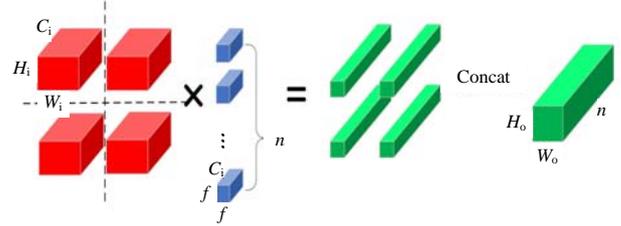


Fig.5 Spatial splitting
图 5 空间张量分裂

值得注意的是，为了保持输入数据的完整性，在分割输入数据时会根据卷积核的大小来设置重叠区域，在重叠区域内的数据会被分配到多个节点进行计算，其计算量会有增加。如图 5 所示，当输入数据被分为 4 块时，除了原本 $H_i \times W_i \times C_i$ 的数据维度以外，还需要增加的重叠部分维度为： $[(H_i + W_i) \times f - f^2] \times C_i$ 。

2.5 卷积层张量分裂方法对比

卷积层不同张量分裂方法的具体对比如表 2 所示。其中，输入数据的维度是 $H_i \times W_i \times C_i$ ，滤波器的维度是 $H_f \times W_f \times C_f \times k$ ，输出数据的维度是 $H_o \times W_o \times C_o$ 。

表 2 卷积层的通道张量分裂、空间张量分裂与滤波器张量分裂的对比
Table 2 Comparison of channel splitting, spatial splitting and filter splitting of convolutional layers

name	#node	weights(per node)	input communication	output communication	merge operation	distributed activation
baseline	1	$k \times W_f \times H_f \times C_i$	$W_i \times H_i \times C_i$	$W_o \times H_o \times C_o$	N/A	N/A
channel splitting	n	$(k/n) \times W_f \times H_f \times C_i$	$n(W_i + H_i + C_i)$	$W_o \times H_o \times C_o$	Concat	Yes
spatial splitting	n	$k \times W_f \times H_f \times C_i$	$W_i \times H_i \times C_i + [W_i \times H_f + H_i \times W_f - W_f \times H_f] \times C_i$	$(1/d^2) \times W_o \times H_o \times C_o$	Concat	Yes
filter splitting	n	$k \times (C_i/n) \times H_f \times W_f$	$W_i \times H_i \times C_i$	$n W_o \times H_o \times C_o$	Sum	No

总体而言，不同的卷积层分裂方法相对集中式计算，都减少了单节点的计算量，另外，通道分裂和滤波器分裂也减少了对内存的占用。不同张量分裂方法的差异对性能的影响强度由特定卷积层的特性决定。具体而言，通道分裂的劣势在于复制输入数据的开销，每个节点都需获得一份原始卷积层中的完整输入数据。空间分裂的劣势在于无法降低内存开销，节点需要存储一份与原始卷积层相同的权重数据。滤波器分裂的劣势在于每个节点需要传输一份与原始卷积层输出数据量相同的结果数据。在拆分方式的选择中，需要根据不同卷积层的权重数、输入维度、输出维度，计算得出具体通信量、计算量和内存占用量，以此作为标准做合理拆分。值得注意的是，通道张量分裂和空间分裂方法可以在节点内直接调用激活函数，而滤波器张量分裂方法则必须先汇总计算结果后再进行激活，因此不能在节点内完成激活。

3 张量分裂性能分析

3.1 仿真平台搭建

本节通过在树莓派 3b+^[7]设备上进行仿真实验，成功模拟分布式节点上的分布式张量分裂计算方法。利用 4 台树莓派 3b+搭建的测试平台如图 6 所示。该实验模拟了星型网络拓扑结构，将 4 个树莓派节点分别标记为 A、B、C 和 D。其中，B、C、D 三个节点可以与 A 节点通过 WiFi 直接通信，进行数据交换。4 个节点可同时进行分布式计算，节点 A 同时负责汇总整合数据。该通信网络的拓扑示意图如图 7 所示。

3.2 数据集与数据预处理

仿真实验采用的数据是调制信号数据集 RadioML2016.10a^[8]，该数据集包含 220 000 个数据样本，共 11 个调制类别。数据集借助 GNU Radio 开源软件无线电平台产生，在产生数据集时采用了 GNU Radio 动态信道模型^[9]的分块电路模块，模拟了大量的信道中的其他影响因素，如中心频移、采样率偏差、多径、衰落、加性高斯白噪

声等，将真实信号通过严格未知的信号模型后，再通过分片和矩形滑窗处理 128 个采样，对仿真产生的数据，随机挑选时间段进行采样，再将采样的结果保存到输出向量中。

通常主流的深度学习框架均采用 4 维的 32 位实数数据存储方式，即 $N_{\text{examples}} \times N_{\text{channels}} \times D_{\text{im1}} \times D_{\text{im2}}$ 的数据存取。其中 N_{examples} 代表样本个数，每个样本包含 128 个复实数点的时间采样，采样通道数 $N_{\text{channels}}=1$ ， $D_{\text{im1}}=2$ 代表有 I,Q 两路数据，采样后数据长度 $D_{\text{im2}}=128$ 。无线通信中通常采用复数的表达方式，但是目前许多机器学习的框架工具并不支持复数表示，为了尽量保存数据信息，数据集利用 D_{im1} 的 I/Q 两路数据将每次采样数据的同相和正交分量组成一个 2×128 的数据。

在此基础上，通过将同一条数据复制若干份进行叠加的方式来模拟信号输入。实际情况下输入信号不可能完全相同而且会受到一定程度的干扰，故再加入方差为 0.05 的高斯噪声用于模拟实际情况下多个节点分别采集到的单通道数据叠加成多通道数据作为数据的输入。



Fig.6 Raspberry Pi testbed
图 6 树莓派测试平台

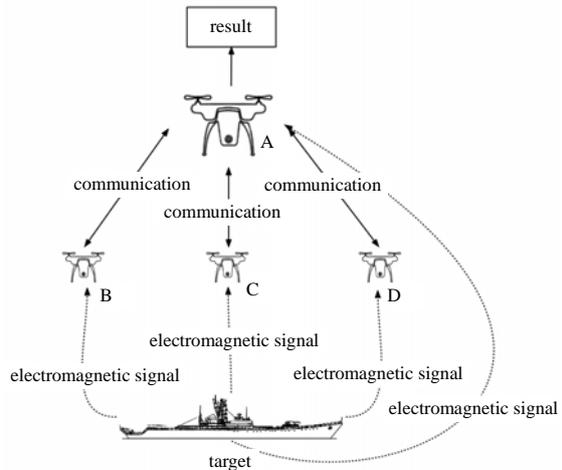


Fig.7 Topology of node network in simulation experiment
图 7 仿真实验节点网络拓扑

3.3 实验结果

实验采用的集中式神经网络如图 8 所示。图 8 中 CN 为卷积层，BN 为批归一化层，MP 为最大池化层，DP 为 dropout 层，FC 为全连接层。该网络的输入数据为 $1 \times 128 \times 8$ 维的数据，该数据由 4 条维度为 $1 \times 128 \times 2$ 维的数据在第三维叠加后得到。

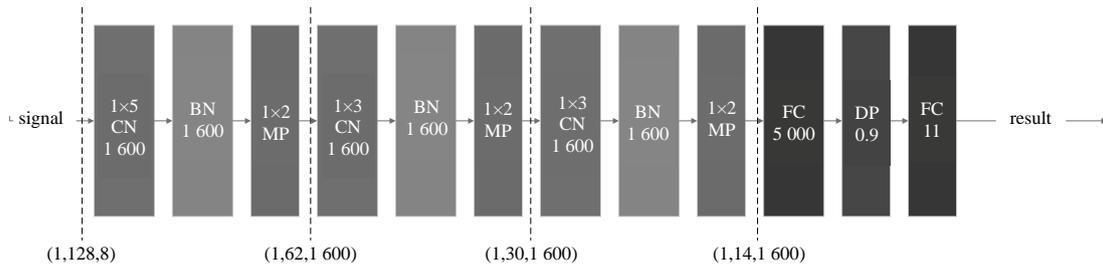


Fig.8 Centralized neural network
图 8 集中式神经网络

将集中式神经网络进行分布式拆分，具体张量分裂方式如图 9 所示。图 9 中的 Sum 与 Concat 分别为求和与拼接。该神经网络中对计算性能及存储性能要求较低的池化层等未做分布式计算，而是在汇总节点进行集中式计算。该分布式网络四个节点分别都接收一条维度为 $1 \times 128 \times 2$ 维的原始信号数据输入，在计算过程中，将第一层卷积层做滤波器张量分裂，在汇总节点进行求和汇总得到结果。在第二和第三层卷积层中，使用通道张量分裂，在汇总节点进行拼接汇总得到结果。在第一层全连接层中使用输出张量分裂，在汇总节点进行拼接汇总得到结果。最后再在汇总节点进行最后一次集中式全连接层计算后得到输出结果。

实验结果如表 3 所示。在实验过程中，可以发现集中式模型由于参数量过大，无法加载于内存中，当处理器使用率达到 98%，内存使用率达到 96%时，程序报错，无法继续进行计算。分布式计算方式在计算结果与原始模型完全一致的情况下，实现了大型模型在测试平台上的计算。在相同的准确率条件下，分布式计算中单节点的

- [9] O'SHEA T, KARRA K. GNU radio signal processing models for dynamic multi-user burst modems[J/OL]. arXiv:1604.08397, 2016.
- [10] HAN S, MAO H, DALLY W J. Deep compression: compressing deep neural networks with pruning, trained quantization and huffman coding[J/OL]. arXiv:1510.00149v5, 2016.
- [11] YU J, LUKEFAHR A, PALFRAMAN D, et al. Scalpel: customizing DNN pruning to the underlying hardware parallelism[C]// 2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture. Toronto, ON, Canada: IEEE, 2017: 548–560.
- [12] KIM Y D, PARK E, YOO S, et al. Compression of deep Convolutional Neural Networks for fast and low power mobile applications[J]. Computer Science, 2015, 71(2): 576–584.
- [13] ZHOU L, XIA Y, ZANG H, et al. An edge-set based large scale graph processing system[C]// IEEE International Conference on Big Data. Washington, DC, USA: IEEE, 2017: 1664–1669.

作者简介:

陆鹏威(1997–), 男, 在读硕士研究生, 主要研究方向为通信网络与人工智能. email: 1931544@tongji.edu.cn.

张伟(1985–), 男, 硕士, 高级工程师, 主要研究方向为阵列信号处理、电子对抗等.

颜子彦(1996–), 男, 硕士, 主要研究方向为通信网络与人工智能.

曾 歆(1987–), 男, 博士, 讲师, 主要研究方向为通信网络与人工智能.

史清江(1980–), 男, 博士, 教授, 主要研究方向为网络优化与分布式感知和计算.

(上接第 21 页)

- [6] MA J, QIU T. Automatic modulation classification using cyclic correlation spectrum in impulsive noise[J]. IEEE Wireless Communications Letters, 2019, 8(2): 440–443.
- [7] WANG Y, WANG J, ZHANG W, et al. Deep learning-based cooperative automatic modulation classification method for MIMO systems[J]. IEEE Transactions on Vehicular Technology, 2020, 69(4): 4575–4579.
- [8] ZHANG Z, GUO X, LIN Y. Trust management method of D2D communication based on RF fingerprint identification[J]. IEEE Access, 2018(6): 66082–66087.
- [9] SHAH M H, DANG X. Robust approach for AMC in frequency selective fading scenarios using unsupervised sparse-autoencoder-based deep neural network[J]. IET Communications, 2019, 13(4): 423–432.
- [10] LIN Y, TU Y, DOU Z. An improved neural network pruning technology for automatic modulation classification in edge devices[J]. IEEE Transactions on Vehicular Technology, 2020, 69(5): 5703–5706.
- [11] MA M, LI Z, LIN Y, et al. Modulation classification method based on deep learning under non-Gaussian noise[C]// 2020 IEEE 91st Vehicular Technology Conference(VTC2020-Spring). Antwerp, Belgium: IEEE, 2020: 1–5.

作者简介:

师长立(1984–), 男, 工程师, 在读博士研究生, 主要研究方向为卫星电源、卫星通信关键技术. email: shichangli@mail.iee.ac.cn.

吴理心(1984–), 男, 助理研究员, 硕士, 主要研究方向为卫星电源、卫星用储能关键技术.

韦统振(1976–), 男, 研究员, 博士生导师, 主要研究方向为电力电子装备、卫星电源关键技术.

叶泽雨(1995–), 男, 在读博士研究生, 主要研究方向为电力电子变换技术、卫星用储能关键技术.

尹靖元(1987–), 男, 副研究员, 在读硕士研究生, 主要研究方向为卫星电源、卫星通信关键技术.