**A Robust IoT Device Identification Method with Unknown Traffic Detection**

发表会议：WASA 2021(CCF-C)

Motivation

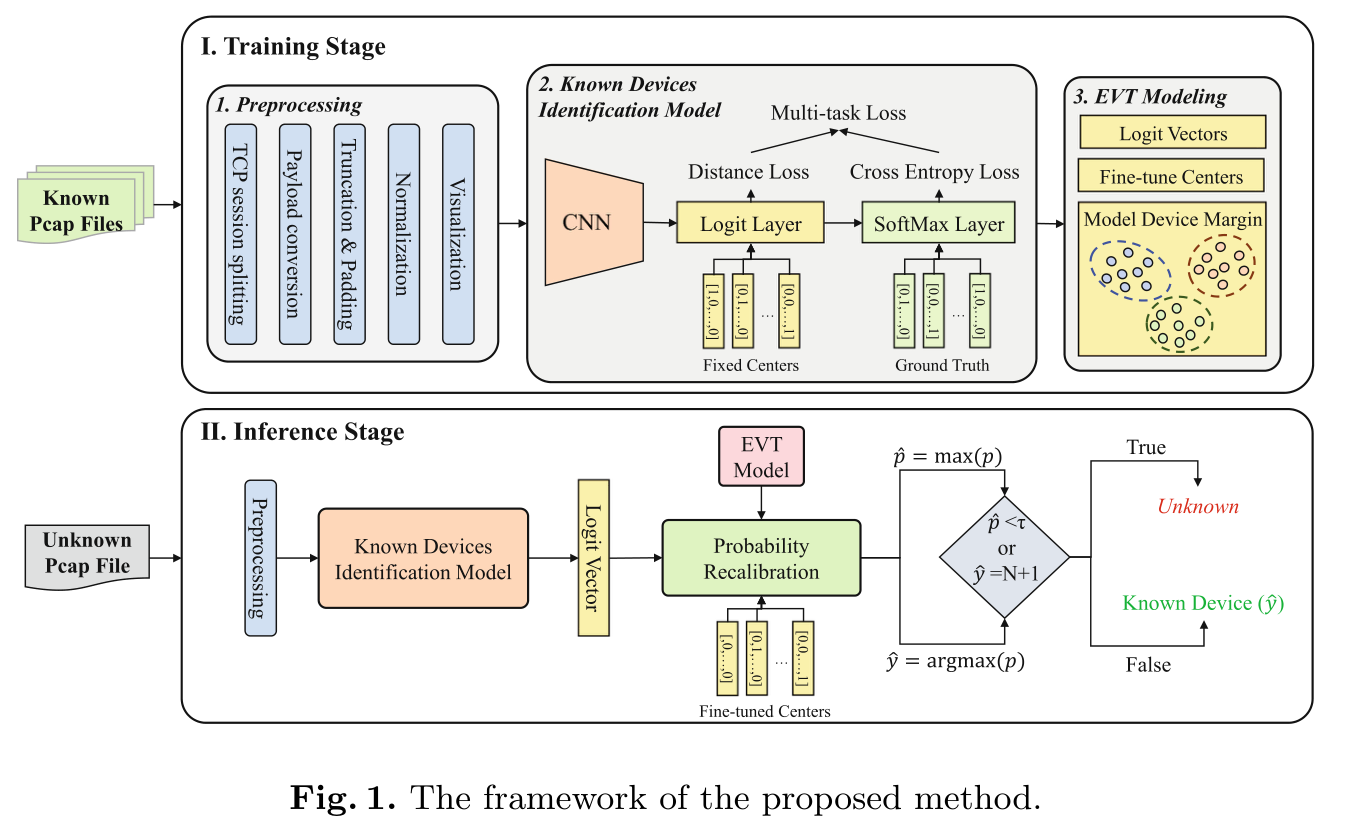
首先，本文研究的任务是物联网设备识别，当前针对该任务主要有两大类方法，主动检测和被动检测，由于前者占用网络资源较大，且后者得到的信息足以完成任务且成本低，本文也是从后者进行切入的

现有方法的局限性：①先前的检测方法不考虑如何处理数据集之外新的或不可见的流量，这将迫使未知流量被错误地分类到预定义的已知类之一；②很多方法运用的是手工提取的流量特征，泛化性差。

本文的应对方法：①针对已知流量分类的问题，本文使用深度神经网络自动提取特征，优化了先前手工提取特征的方式；引入多任务损耗，在保证分类性能的基础上，使得已知设备流量在特征空间上形成了紧密的集群，为后文判别未知流量奠定基础；②针对未知流量，使用极值理论EVT估计流量被分为未知流量的可能性。

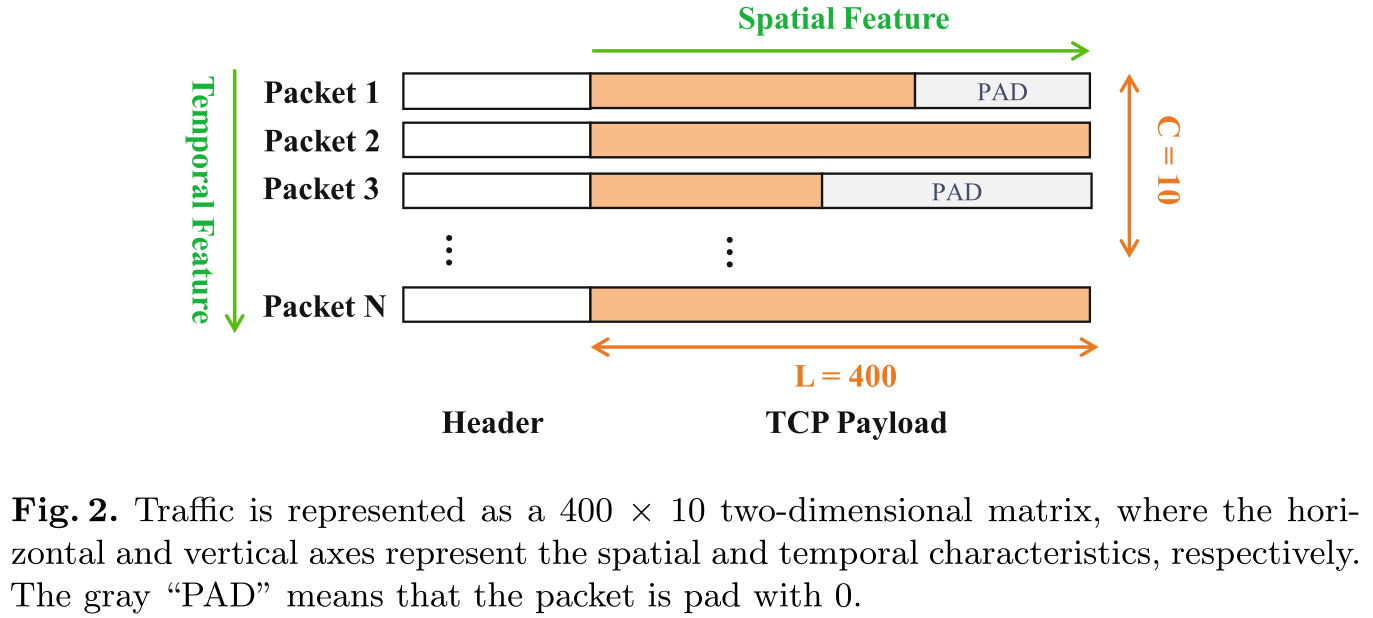
模型设计

本文提出的模型框架如图1所示，它包括训练阶段(general training)和新推理阶段(novel inference stage)。训练阶段的目标是利用多任务损失对所有已知设备进行准确分类，并为每个设备建立EVT模型；在推理阶段，使用构建的EVT模型来重新校准流量被分为未知类的概率。



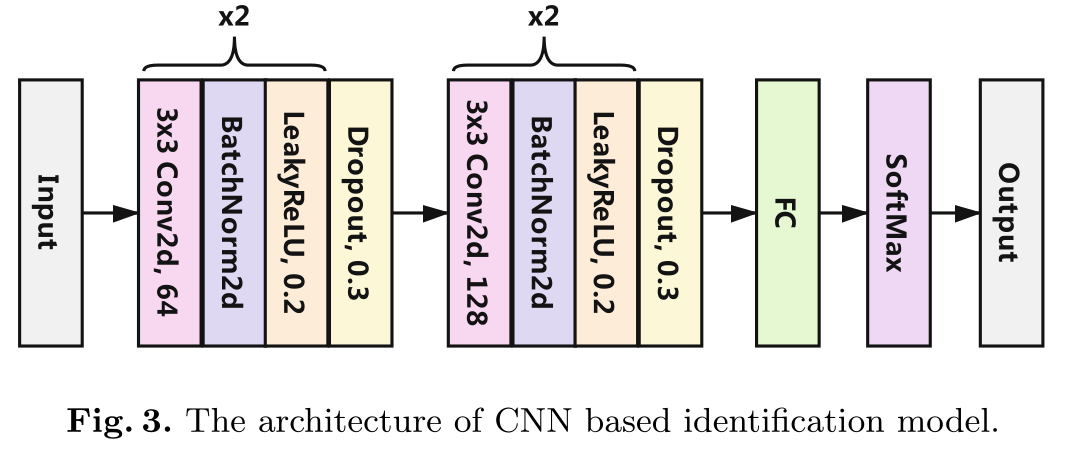
**训练阶段**

原始数据预处理：①只使用TCP会话的有效载荷来表示流量，因为基于UDP的流量很难区分不同的物联网设备；②去掉报头，防止模型过拟合IP和MAC地址；③将有效载荷转换为一个二维矩阵：流量首先从二进制转换为以字节为单位的十进制，然后根据有效载荷数据包的数量和负载长度的经验分布进行截断和填充。‘



从图2可以看出，流量数据被表示为了一个10×400的二维矩阵，超出固定长度的部分被阶段，小于固定长度的部分用0填充。

训练设备识别模型：该任务的目标是根据输入的流量数据识别相应的设备标签。这里的模型是由4个2D-CNN组成的神经网络模型，每个卷积层后面都跟着一个batch normalization和LeakyReLU激活函数。每两个卷积层之间都由一个dropout连接。我们选择倒数第二层[14]，一般称为logit层作为特征空间(softmax的输入)。

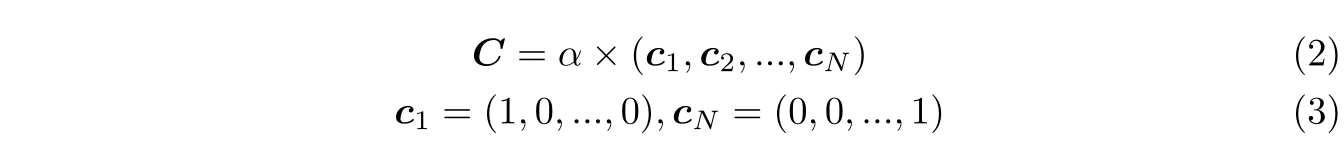


对于第一个任务，使用一下交叉熵损失训练模型：

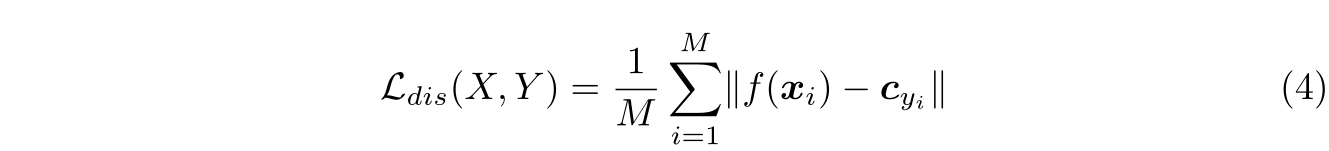


其中当时，，其余情况等于0。是softmax函数，表示第i个流量被划分为第j类的概率。

为了显式构建紧密集群，我们为每个已知设备设置固定的类中心，定义为：



其中*α*是标度标准基向量的超参数。我们引入了一个距离损失，以确保从输入到其正确的类中心的绝对距离被强制减小。利用距离损失对任务2的模型进行优化，其表达式为：

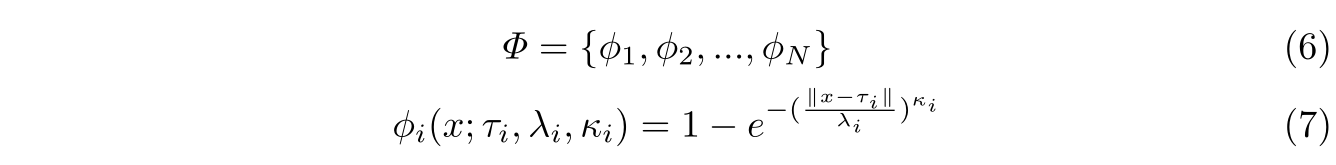


我们将交叉熵损失和距离损失相结合，得到最终的损失函数：



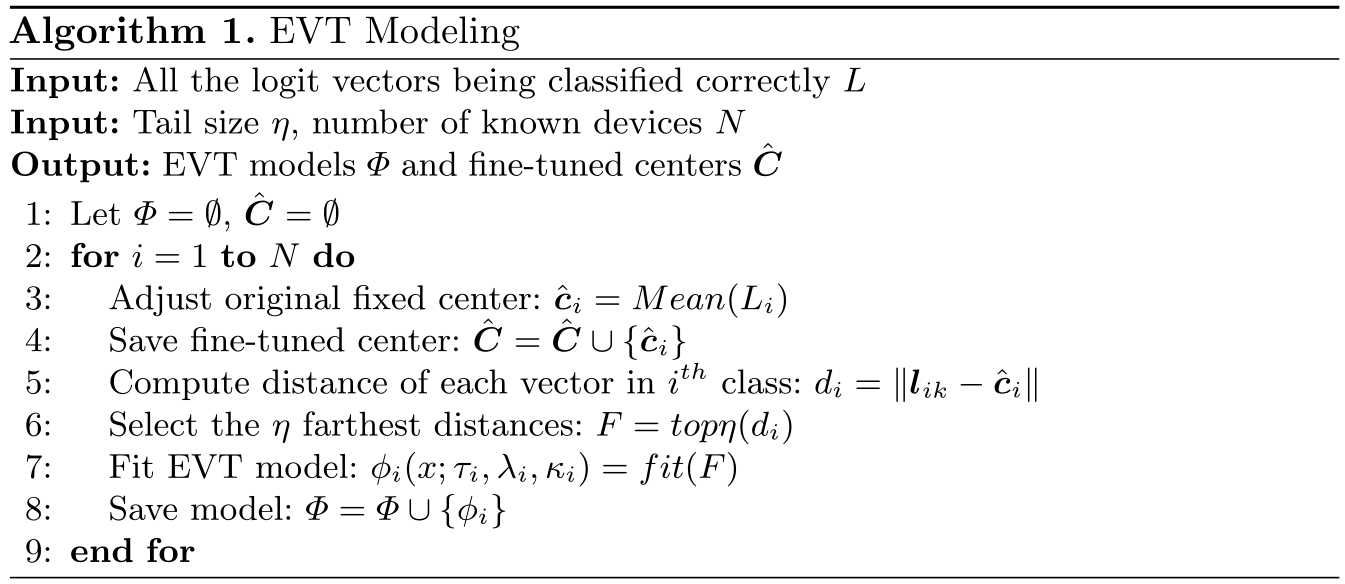
其中是调整两种损失比例的超参数。

EVT建模：威布尔累积分布函数(Weibull cumulative distribution function, CDF)对输入落入极值区域的概率进行建模(尾部建模)，这意味着如果输出概率越高，则输入属于相应类别的可能性越小。已知器件的累计概率分布函数表示为：



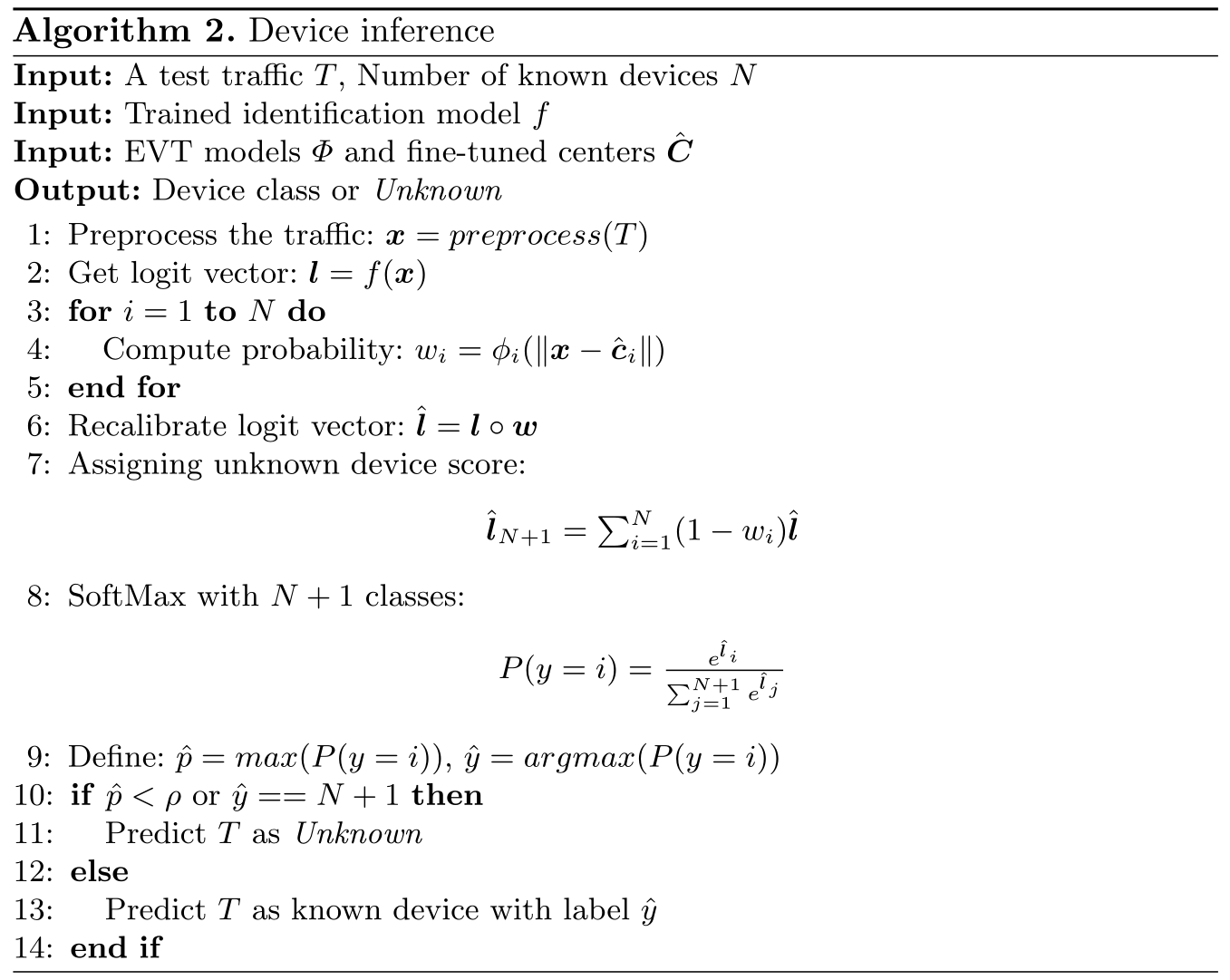
其中*x*是logit向量到其中一个中心的距离。、、分别为位移参数、尺度参数和形状参数。为了估计这些参数，我们引入了一个叫做尾部尺寸的超参数*η*(这个*η*对每一类来说是相同的，是否应该因类而异？)。它表示了远离类别中心的样本数量。

给定训练过程中所有正确分类样本的logit空间向量：，，通过计算同一类向量的均值，我们调整原来的固定中心。通过上述准备，对每个已知设备的威布尔分布进行拟合。算法1总结了EVT整体建模的过程：



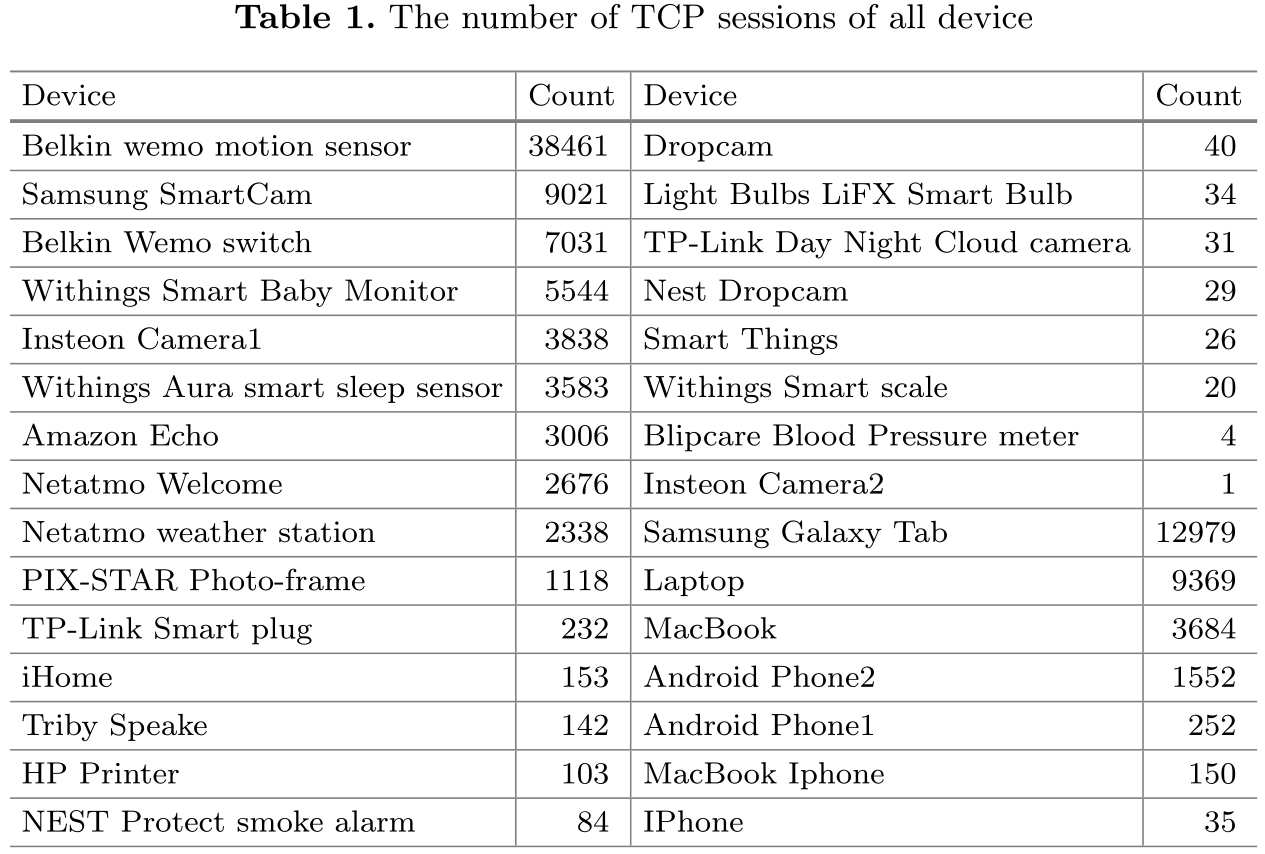
**推断阶段**

算法2阐述了推理过程(给一个新的样本判断它属于已知类还是未知类)。按照3.1节对测试流量进行预处理，并将其输入训练后的识别模型以获得其logit向量。然后计算到logit空间中每个聚类中心的距离，然后得到每个EVT模型的输出概率。随后我们重新校准logit向量(类似OpenMax)，并为未知类分配一个分数()。最后，定义最大概率及其对应的类*id*。如果小于*ρ*或，则测试流量属于未知类，否则返回标记为对应的已知类。这里*ρ*是一个阈值超参数，将在4.3节讨论。



**讲解实验**

**数据集**：使用了[6]的公开可用物联网流量数据集。数据集收集自一个仪器化的实验环境，包括30个商业设备(23个物联网设备和7个非物联网设备)。如3.1节所述，我们提取了流量中的所有TCP会话，并通过MAC地址将它们分开，因为在这个数据集中，一个MAC地址对应一个设备(物联网流量特色？流很短)。所有设备的TCP会话数如表1所示。



**实验1：识别已知设备**

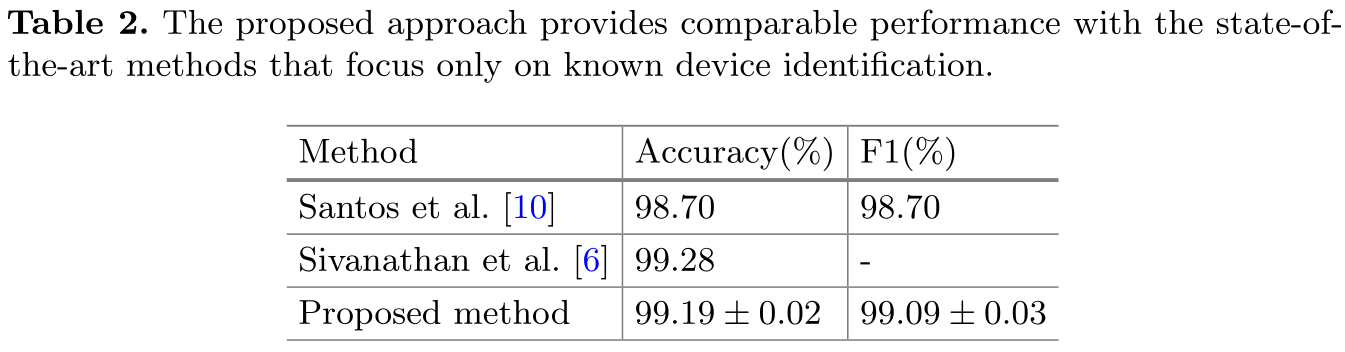
·Baseline：两种最先进的物联网设备识别方法

·为了防止数据不平衡降低模型性能，选择了20个设备(包括14个物联网设备和6个非物联网设备)来评估方法。

·设置降采样的次数为3000；

·训练集与测试集设为8：2；

·实验结果：表2显示了5次运行的平均分数。我们的模型精度高于[10]，与[6]相当。请注意，他们的方法过度使用人工特征，使其难以推广。我们的方法可以通过自动学习特征来避免这类问题。综上所述，该方法保证了已知类的分类性能。(实验1，在闭集条件下，本文方法的性能优于使用手工特征的另外两种方法，说明特征提取模块的优越性)



**实验2：未知流量检测**

·Baseline：

**SoftMax：**训练过程与本文方法相同，**推理过程直接比较阈值和最大输出概率**，**缺少了使用威布尔分布得分去校正概率的过程**。

**CE：训练阶段只使用交叉熵损失函数**，推理过程与本文方法一样。

**CNN：只是用了两个CNN层**(减少了层数)，其他和本文方法一样。

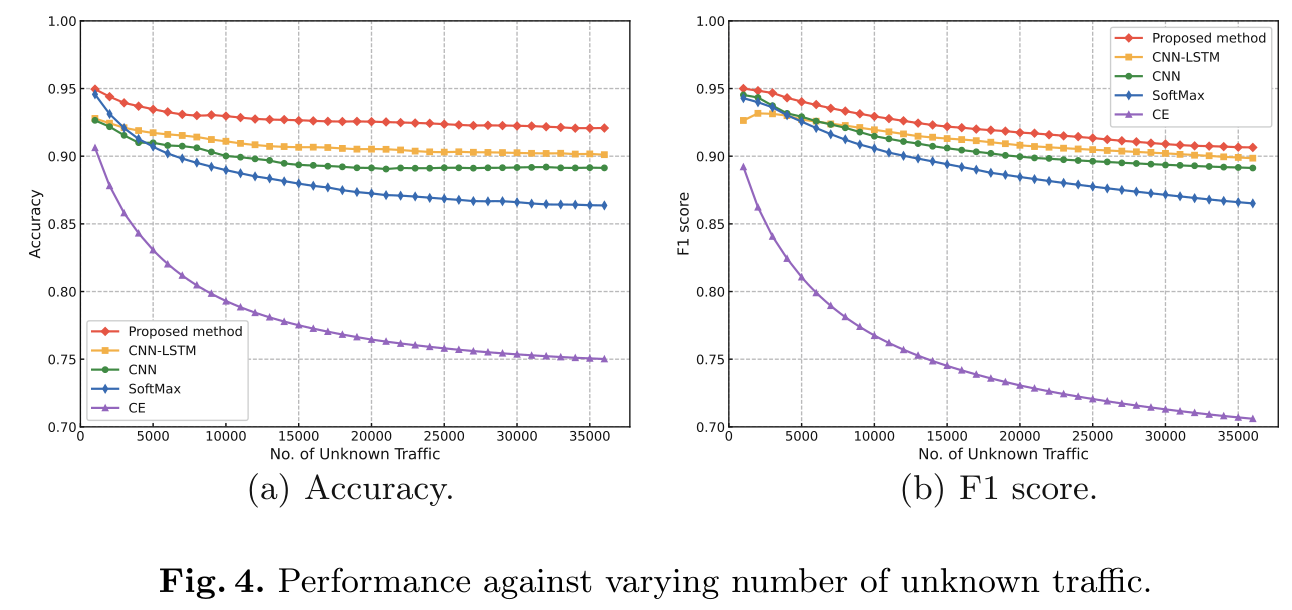
**CNN-LSTM：该基线使用两个1D-CNN层提取空间特征，并将其输入一个两层LSTM中捕捉时间特征。**训练和推理过程与所提出的方法相同。

为了模拟未知流量，我们从拥有超过1000条TCP会话的设备中随机选择8台物联网设备和2台非物联网设备作为已知流量，其他设备的流量作为未知流量。然后我们对已知的TCP会话进行采样并进行分割。使用训练集对模型进行优化，所有未知流量经过洗牌后加入到测试集中。通过11个类别(10个已知设备和1个未知设备)的准确性和F1得分来评估性能。

**实验结果1：未知流量数量的影响**

一般情况下，未知流量的数量是影响模型性能的一个重要因素。未知流量越多，检测难度越大。我们为每增加1000个未知TCP会话计算一次度量。图4显示，未知流量包含了来自20个不同设备的36,000多个TCP会话，不同程度地影响了所有方法的准确性和F1得分。值得注意的是，F1分数的下降幅度大于准确率，因为每一个类的分类比整体准确率更难。

只有交叉熵损失训练的CE基线表现最差。这一结果说明了距离损失模型对类别紧密性的有效性和优越性。CNN的基线略低于所提出的模型，这意味着更深层的神经网络结构有助于捕获更高级的特征。对于CNN-LSTM模型，它是次优模型，因为它学习了一个有意义的流量表示。对于SoftMax基线，准确度和F1得分下降了近5%和8%，而我们的方法分别可以达到低于4%和5%的下降。此外，我们的方法在所有测试步骤中都优于SoftMax基线。两者之间的差距随着未知流量的增加而扩大。结果表明，与基线相比，该方法在性能和鲁棒性方面更有效。(**这里没说阈值*ρ*选的多少**)



**实验结果2：阈值变化的影响**

我们还研究了不同阈值的影响。给定一个阈值*ρ*，所有的方法都使用同一个阈值对测试集中的所有流量进行分类。图5报告了所有方法的性能。

当阈值非常小时，我们的方法的性能明显高于SoftMax(准确率：40%以上，F1评分:20%左右)。说明阈值越小，SoftMax将测试流量归类为未知流量的可能性越小。但我们的方法可以通过EVT估计未知流量的概率。与其他消融基线相比，该方法具有更高的消融性能。(说明EVT建模的性能)

随着阈值的逐渐增大，所有方法的性能都有所提高。但是我们方法的性能始终在基线之上，即使是在阈值接近于1的时候。值得注意的是，即使阈值很高，CE模型也很难获得令人满意的分数，因为它不能在没有距离损失的情况下很好地建模类边界。综上所述，该方法能有效检测未知流量，且对已知设备的识别性能下降较小。(说明距离损失的重要性，多任务损失是合理的)

