基于自注意力机制的云服务器业务流量异常检测方法

干淇任

摘要—随着云技术的迅猛发展，云服务器应用场景广泛，但伴随而来的问题是业务流量异常检测任务加重。传统的异常检测方法存在准确率差和效率低等问题，无法满足当今市场的需求。为了解决深度学习在云服务器业务流量异常检测应用领域面临的数据集不平衡、动态数据等挑战，我们提出了一种创新的方法。首先，我们将时序数据转化为曲线图像，并采用区域填充技术生成曲线填充图像。然后，我们采用卷积神经网络（CNN）模型，并引入自注意力机制（Self-Attention）来提高图像样本特征的提取能力。实验测试结果显示，本文所提方法优于其他相关异常检测方法，在私有数据集上的三分类任务中 F1 分数优于 CNN-LSTM 方法 1.85%，在 ElectricDevices 数据集上的六分类任务中 F1 分数优于 LSTM 方法 8.36%。

Index Terms—业务流量；异常检测；卷积神经网络；自注意力机制；曲线图像

1 引　言

随着云计算 [1] 的快速发展，越来越多的企业专注于提供云服务器，为企业和个人提供租用服务。作为基于云计算技术的虚拟服务器，云服务器通过互联网提供计算资源，允许用户在云端运行应用程序和存储数据。与传统的物理服务器相比，云服务器具有更高的灵活性、可扩展性和可靠性。

随着网络技术和硬件设备的迅猛发展，云服务器的应用场景日益广泛。目前，云服务器已广泛应用于汽车 [2]、医疗 [3]、音影视频 [4]、游戏 [5] 等多个领域。

国际数据公司（IDC）最新发布的《中国公有云服务市场（2023 上半年）跟踪》报告显示，2023 年上半年中国公有云服务整体市场规模（IaaS/PaaS/SaaS）达到 190.1 亿美元。其中，IaaS 市场规模为 112.9 亿美元，同比增速 13.2%；Pass 市场规模为 32.9 亿美元，同比增速为 26.3% [6]。

在这庞大的云服务市场中，对于以提供云服务器为主要盈利方式的企业来说，实时监控云服务器中的业务是否出现异常并及时解决问题成为关注的焦点。因此，为保障云服务器的稳定运行，降低因异常情况带来的损失，准确进行业务流量的异常检测至关重要。

在云服务器中，时间序列数据通常是最具有意义且值得分析的主要数据类型 [7]，具有周期性、趋势、非平稳性等特点，可用于判断业务流量是否出现异常。传统异常检测方法依赖于专业技术人员的知识和经验，存在主观因素，难以适应市场需求。基于深度学习的异常检测方法显著提高了异常检测的准确率和效率，但这些方法时常面临数据集不平衡 [8]、难解释性等挑战。在缺少标签的情况下，正常数据和异常数据之间的类不平衡会阻碍模型的训练 [9]，导致检测准确率降低。

本文提出了一种基于自注意力机制的云服务器业务流量异常检测方法，借助流量分类，实现异常检测。本文将一维时序数据转化为二维曲线图像，并采用区域填充技术生成曲线填充图像作为模型输入。我们根据卷积神经网络在处理图像数据上的优势和自注意力机制可以捕捉数据中长期依赖关系的能力，提出了 CNN2-SA 异常检测模型，这样的组合有效提高了检测效率。

2 相关研究

2.1 传统流量异常检测方法

流量异常检测是异常检测领域中常见的一种问题。Kotenko I 等人 [10] 提出了一种通过统计方法评估网络流量自相似性的技术，从而识别网络流量中的异常情况。这项技术验证了早期流量异常检测方法的良好性能。

Li Y 等人 [11] 提出了一种基于 TCM-KNN 数据挖掘算法的无监督网络异常检测方案，采用基于遗传算法的实例选择方法限制训练集规模，同时使用基于过滤器的特征选择方法提高了 TCM-KNN 的性能并降低了计

算成本，确保了其方法在异常检测中的有效性。

Teoh Teik-Toe 等人 [12] 通过使用四种常见的机器学习算法（决策树、朴素贝叶斯算法、支持向量机和多层感知器）进行网络流量异常行为检测，证明了这些监督机器学习算法对网络流量异常行为完成异常检测和分类的可行性。

2.2 基于深度学习的流量异常检测方法

随着大数据驱动的人工智能技术的快速发展，深度学习凭借其强大的学习能力和特征提取能力，在异常检测领域取得了显著进展。

Loukas 等人 [13] 基于循环神经网络（RNN）的深度学习，使用长短期记忆网络（LSTM）进行车辆云物理入侵检测，相较于机器学习方法，其方法表现出更高的准确率。

Ullah 等人 [14] 提出了一种混合深度学习模型，结合了卷积神经网络和循环神经网络，用于物联网网络中的异常检测。他们使用多个数据集进行实验，该模型在二元分类任务和多分类任务方面均取得了较高的准确率，相较于其他深度学习方法而言表现出色。

2.3 时间序列数据

时间序列被定义为按时间顺序索引的测量值的集合 [15]。其中，单变量时间序列是一组数据点的序列，数据通常在均匀时间间隔的连续时间点上捕获，每个连续测量数据之间存在时间相关性或依赖性。单变量时间序列可以表示为 $T = \{t_1, t_2, ..., t_n\}$，$n$ 是 T 的长度 [16]。

我们可以通过检查时间序列数据中是否有偏离正常行为模式的行为模式来识别数据异常 [17]。基于时序数据的异常检测技术已广泛应用于网络入侵检测、医疗异常检测、工业物联网等领域。

Al-Ghuwairi 等人 [18] 提出了一种基于时间序列异常的云计算入侵检测技术，它将 Facebook Prophet 模型、时间序列分析技术、异常检测和因果关系测试相结合，解决了时间序列异常和攻击之间产生误导性联系的问题，并在 CSE-CIC-IDS2018 数据集上验证了模型性能。

Vitor Cerqueira 等人 [19] 提出了一种用于预测关键健康事件的分层方法，通过对时间序列中的异常进行早期检测，完成关键健康事件预测。他们使用 Multi-parameter Intelligent Monitoring for Intensive Care (MIMIC) II 数据库验证了他们所提出的方法相较于最先进的关键健康事件预测方法，具有更好的性能。

Woong Hyun Suh 等人 [20] 提出了一种基于元启发式的时间序列聚类技术，结合时间序列聚类技术和元启发式算法，通过采用扩展紧凑遗传算法（ECGA）证明了数据分析结果的逻辑结果。该方法在制造业时间序列数据的异常检测中得到了有效的应用，与其他时间序列分类技术相比，该方法显示出更有意义的性能。

2.4 图像分类

图像分类是计算机视觉中最重要的研究任务之一，现已应用于人们日常生活的方方面面，如人脸识别、车牌识别、故障检测等 [21]。CNN 是专为图像数据而设计的神经网络，其中包括卷积层、池化层和全连接层，具有稀疏连接和参数共享等特点。CNN 架构在 ImageNet 数据集上的表现展示了它在处理图像数据方面上的惊人力量 [22]。目前，卷积神经网络是图像分类任务中应用最广泛、最有效的方法。

Xiangdong 等人 [23] 提出了一种基于混合深度学习的图像识别算法，使用改进后的 MobileNet 模型进行图像识别。他们的改进使传统的 MobileNet 模型性能得到很大提升，并在 Fruit-360 数据集上验证了他们所提出的改进措施的优越性。

Banerjee 等人 [24] 提出了一种将二维卷积神经网络与批处理归一化相结合的 Resnet ConvLSTM 模型，该模型有助于最大限度地降低计算复杂度并从高光谱图像中提取特征。他们在多种不同类型的高光谱数据集和时间序列数据集上验证了模型的性能。

P.Pravin Sironmani 等人 [25] 提出了一种具有高效通道化组织病理学医学图像分类的新型 CNN 架构，引入通道化层代替第一卷积层，使用自适应滞后阈值技术将图像数据通道化为特征图，以帮助 CNN 模型提取更精细的特征。他们在 APTOS2019,COVID-19, COVID-19(2021)，Cancer 四个医学影像数据集上验证了该模型在组织病理学医学图像分类任务中的高效性能。

2.5 自注意力机制

注意力机制是一种模仿人类注意力过程的方法，它被引入神经网络模型中以提升对输入数据的处理能力。Vaswani 等人 [26] 提出，注意力机制旨在有选择性地关注特定的输入，从而使模型能够权衡并提取与给定任务最相关的信息。

自注意力机制是注意力机制的一种演变形式，它利用数据之间的相关性进行建模，并通过为输入数据中的每个元素分配不同的权重来更好地捕捉它们之间的关系。自注意力机制能够有效地捕捉序列中不同位置之间的长距离依赖关系，具有更全面的序列建模能力。因此，自注意力机制通常被嵌入到卷积模块之后，以增强网络对于短期和长期依赖关系的处理能力 [27]。同时，其并

行计算特性使得模型在处理大规模数据时能够更高效地进行训练。这些优势令自注意力机制在需要处理各种大规模、复杂数据集的深度学习任务中表现出色。

基于自注意力机制的神经网络已成功应用于机器翻译和图像生成等领域。在自然语言处理领域，Ziyu Zhou 等人 [28] 提出了一种基于位置和自注意力机制的方面级情感分类方法，在捕获到上下文的全局和局部信息后，利用自注意力机制获取该方面的关键词，最后生成特定方面的上下文表示以进行分类。

在图像分类领域，Cui Y 等人 [29] 提出了一种基于双注意力机制的高光谱图像分类特征融合网络模型，主要用于捕获更准确的全局-局部上下文注意力特征。该模型使用自注意力机制（SA）提取全局上下文注意力特征，并使用交叉注意力机制提取局部上下文注意力特征。

在流量预测领域，Xuesen Ma 等人 [30] 提出了一种基于相关性的 ConvLSTM 和自注意力机制的网络用于预测复杂的蜂窝网络流量，采用自注意力机制来聚合提取外部因素特征与网络流量特征之间依赖关系。他们对真实世界蜂窝网络流量数据集的实验评估证明了该模型的有效性，其性能优于最先进的（SOTA）方法。

3 基于自注意力机制的异常检测模型结构

本文提出的云服务器业务流量异常检测模型结构如图 1 所示：

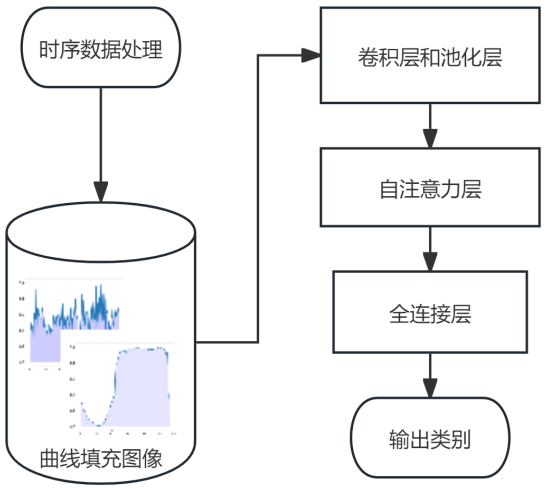


图 1　异常检测模型结构

本文所提的方法利用图像进行流量分类。首先，我们进行数据预处理，将原始时间序列数据处理为单变量时间序列数据。然后将单变量时间序列数据转化为曲线填充图像，作为模型的输入图像。最后，经过训练的 CNN2-SA 模型对图像进行分类。

3.1 数据预处理

由于我们采集得到的原始时间序列数据属于多变量时序数据，其中包括采集时间、总流量和云服务器位置等复杂信息，因此我们需先对数据进行预处理。首先，我们对原始数据进行清洗，删除了空值，并根据采集时间将数据划分为单独的单变量时序数据文件。在此过程中，我们只保留了“总流量”数据信息，并对数据进行了归一化处理，确保数据的一致性和准确性。

3.2 图像分析

3.2.1 数据转化

数据可视化通过将数据压缩成易于理解的视觉元素，如图形或图表，使信息更易于理解和做出决策 [31]。本文利用数据可视化技术，将包含业务流量的单变量时间序列数据转换为曲线图像，并对图像进行了区域填充操作，得到了曲线填充图像。利用可视化技术，深度学习算法能够更有效地分析流量数据中的复杂模式和特征，从而做出更明智和更精确的分类决策 [32]。

根据生成曲线图像类型的不同，可将其分类为灰度图像、黑白图像、彩色图像。其中，由于填充方式的不同，曲线图像又可以分类为无填充曲线图像、上填充曲线图像、下填充曲线图像、全填充曲线图像，具体样例如图 2 所示。

3.2.2 异常检测模型

传统的机器学习技术在处理原始形式的自然数据（例如图像的像素值）方面受到一定的限制。相比之下，深度学习允许建立由多个处理层组成的计算模型，从而学习到更多抽象层次的原始数据表示 [33]。CNN 是深度学习的重要代表之一，自注意力机制则是被视为一种更为复杂且灵活的卷积神经网络。因此，在图像任务中，自注意力机制需要使用庞大的数据集进行训练，才能超越 CNN 在性能上的表现 [34]。

本文基于 CNN 的深度学习，引入自注意力机制构建神经网络模型 CNN2-SA，对流量图像进行分析。CNN2-SA 模型结构如图 3 所示。

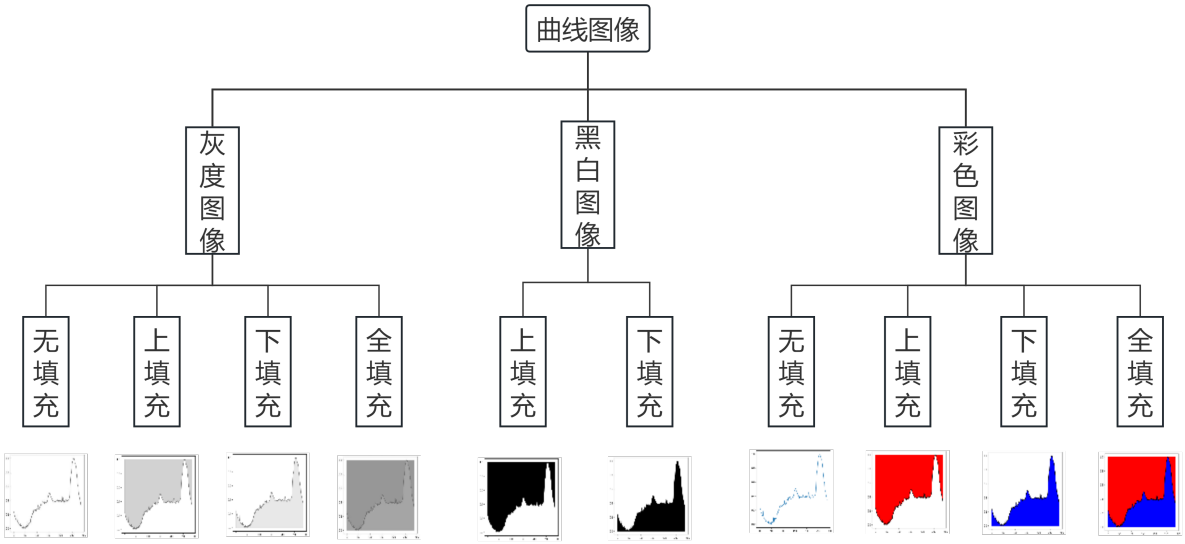


图 2　不同图像类型样例

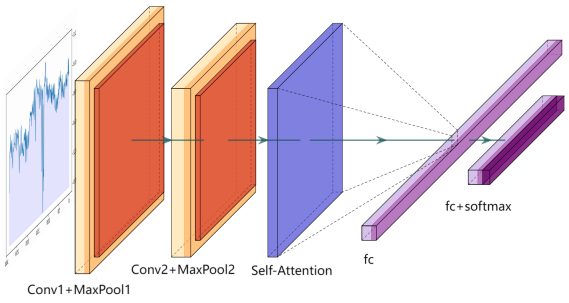


图 3　CNN2-SA 模型

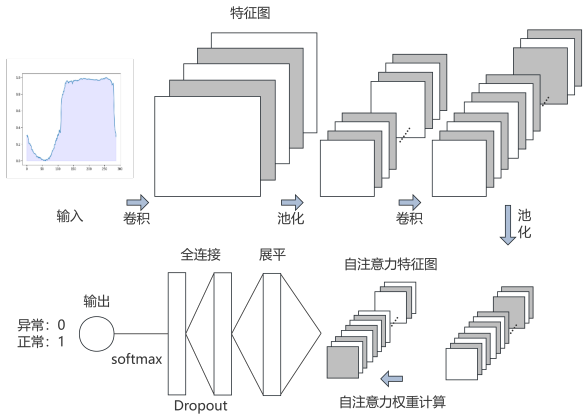


图 4　图像分析架构

在 CNN 模型中嵌入自注意力机制的主要目的是利用两者的优势互补，使整体模型更好地适应真实云服务器环境中复杂、动态的网络流量模式。CNN 通过卷积操作捕捉图像局部特征，自注意力机制则在整个图像空间上进行全局关联，从而使模型能够准确地捕捉图像中的全局信息。同时，相比 VGG、GoogLeNet 和 ResNet 等动辄几十或上百层深度的神经网络模型，本文构建的 CNN2-SA 模型大大减少了训练过程中的计算资源消耗。

在 CNN2-SA 模型中，每个作为输入的曲线填充图像都需要经过两个卷积层、两个池化层、一个自注意力层和两个全连接层。图像分析架构如图 4 所示。

在 CNN2-SA 模型中，第一个卷积层包括 32 个大小为 3*3 的卷积核，第二个卷积层包括 64 个大小为 3*3 的卷积核，均采用激活函数 Relu 用于非线性映射。Relu 函数由式 (1) 表示，其中 $Output$ 表示该层的输出值，W 是权重矩阵，X 是输入向量，B 是偏置项。Relu 函数通过对输入的线性组合进行修正，去除小于等于 0 的部分，从而得到非线性激活后的输出。

$$Output = max(0, W^T X + B) \tag{1}$$

在子采样过程中，采用窗口大小为 2×2 的最大池化来减少训练参数。最大池化完成后，我们将特征图输入到自注意力层，计算自注意力权重分数，输出自注意力

特征图。然后，将自注意力特征图展平，并将其输入两个全连接层。第一个全连接层采用激活函数 Relu，并在该层之后添加一个概率为 0.5 的 Dropout 层，以避免过拟合。第二个全连接层采用 Softmax 函数，输出类别。Softmax 函数由式 (2) 表示，其中 $f(z_j)$ 表示第 j 类的输出概率，z_j 是第 j 类对应的输入值，n 是类别总数。Softmax 函数将每个输入值转换为一个介于 0 和 1 之间的概率，并且所有输出概率的总和为 1。这种归一化处理确保了输出的每个值都是一个有效的概率，能够直观地表示每个类别的预测概率。

$$f(z_j) = \frac{e^{z_j}}{\sum_{i=1}^{n} e^{z_i}} \tag{2}$$

在自注意力权重分数计算过程中，将卷积特征图输入自注意力层，利用训练过程中自主学习到的参数进行权重计算，最终得到自注意力特征图，具体过程如图 5 所示。

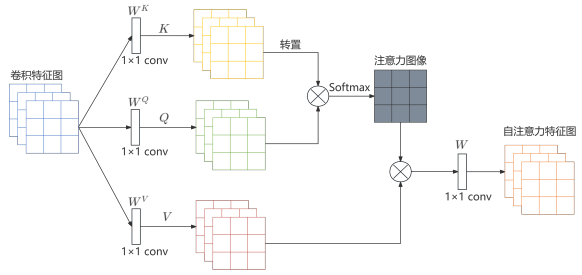


图 5　自注意力权重计算

其中，卷积特征图作为输入 X，分别与权重矩阵 W^K、W^Q、W^V 进行卷积操作生成新的特征图 K、Q、V。W 通常代表网络中其他线性变换的权重矩阵。Q 用于计算当前时间步输入与其他时间步输入的相关性。K 和 Q 一起用于计算相关性分数，并通过 Softmax 函数归一化得到注意力权重图像。V 则是注意力机制最后输出的加权求和结果。将注意力权重图像与 V 相乘并求和，再与权重矩阵 W 进行卷积操作，得到最终的自注意力特征图。

4　实　验

4.1　实验数据和实验装置

在云服务器业务流量异常检测领域公开的数据集很少，本文通过某个云服务提供商获取了该企业云计算服务器的网络访问流量数据集 D:

$$D = \{X_0, X_1, \ldots, X_{N-1}, X_N\} \tag{3}$$

式 (3) 中：向量 $X_i = \{x_{i,0}, x_{i,1}, ..., x_{i,j-1}, x_{i,j}\}$ 表示第 i 天从 00:00 至 23:55，每隔 5 分钟采集的数据，$x_{i,j}$ 表示第 i 天第 j 个采样时刻采集的数据。

我们使用随机划分法，将数据集按照 8:1:1 的比例将所有样本随机分为训练集、验证集和测试集。表 1 表示数据集中异常样本、业务模式一样本和业务模式二样本的数量分布情况。

表 1　三类样本分布

类别	异常样本	业务模式一	业务模式二
数量	2124	1006	493

本文的实验程序是使用 Python 编写的，硬件环境是 16GB 内存、Intel Core i5-12400F 处理器以及 NVIDIA GeForce RTX 3060 显卡。

4.2　实验评价指标

实验样本分为正类别（P）、负类别（N）,TP 表示模型正确分类的正类别样本计数，TN 表示模型正确分类的负类别样本计数，EP 表示模型分类不正确的正类别样本计数，EN 表示模型分类不正确的负类别样本计数，分类混淆矩阵如表 2 所示：

表 2　分类结果混淆矩阵

预测情况 \ 真实情况	正类别	负类别
正类别	TP	FP
负类别	FN	TN

为了评估模型性能，本文实验使用的评价指标包括 Accuracy(准确率)、Precision(精确度)、Recall(召回率)、F1-Score(F1 分数)，其计算方法如下：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{7}$$

4.3 实验结果与分析

4.3.1 消融实验

我们使用灰色曲线图像、黑白曲线图像、彩色曲线图像进行三分类任务消融实验，分别将每种不同类型的图像作为输入图像进行 50 次模型训练，图 6 表示了本文所提方法使用不同类型图像训练的模型准确率。其中，每个柱子上的数字代表该类型图像进行 50 次模型训练的平均模型准确率。误差条表示了这些准确率的变化范围，其中黑色部分代表 50 次模型训练中的最高准确率和最低准确率之间的误差范围。表 3 表示 CNN2-SA 模型使用不同类型图像训练的模型评估指标分数。

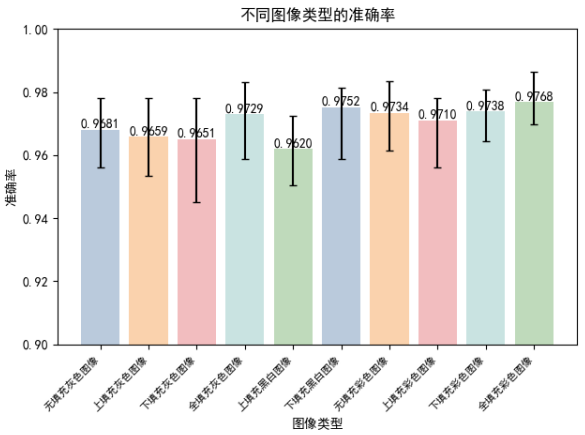


图 6　不同图像类型的模型训练准确率

从图 6 中我们可以看到，全填充彩色图像的平均准确率比其他图像类型的平均准确率较高，而黑白曲线图像类型则是所有图像类型中平均准确率最低的。同时，彩色图像的模型训练准确率误差也比其他图像类型小，而灰色曲线图像类型的模型训练准确率误差较大。

表 3　不同图像类型的三分类任务

方法	Accuracy↑	Precision↑	Recall↑	F1-Score↑
无填充灰色图像	0.9681	0.9688	0.9681	0.9683
上填充灰色图像	0.9659	0.9664	0.9659	0.9660
下填充灰色图像	0.9651	0.9652	0.9651	0.9650
全填充灰色图像	0.9729	0.9721	0.9730	0.9716
上填充黑白图像	0.9620	0.9623	0.9620	0.9620
下填充黑白图像	0.9752	0.9755	0.9753	0.9753
无填充彩色图像	0.9734	0.9742	0.9734	0.9736
上填充彩色图像	0.9710	0.9721	0.9710	0.9715
无填充彩色图像	0.9738	0.9738	0.9738	0.9738
全填充彩色图像	0.9768	0.9771	0.9770	0.9770

从表 3 中我们可以看出，在模型评估指标方面，全填充彩色图像的各项分数均高于其他图像类型，下填充黑白图像的各项分数则居于第二。结合各项实验数据，我们认为全填充彩色图像为 CNN2-SA 模型的最佳输入图像类型。因此，本文在后续实验中均使用全填充彩色曲线图像作为 CNN2-SA 模型的输入图像。

4.3.2 分类任务实验

为评估本文提出的基于自注意力机制的异常检测方法的效果，我们进行了分类任务实验，并与文献中提出的一些常见异常检测方法进行实验对比，包括 SVM [12]、MLP [12]、RNN [13]、LSTM [13]、CNN-LSTM [14] 等。此外，本文还将一种基于自注意力机制的一维卷积神经网络（SACNN ） [35] 进行实验对比。这些对比实验旨在全面评估我们方法的性能表现，并为我们的研究提供更具说服力的结果。

为体现数据不平衡对模型产生的影响，本文使用不同算法以及本文算法对现有数据集进行多种分类任务，包括异常流量与正常流量的二分类任务、异常流量与业务模式一的二分类任务、异常流量与业务模式二的二分类任务、业务模式一与业务模式二的二分类任务以及异常流量与业务模式一、业务模式二的三分类任务，具体实验测试结果如表 4-表 8 所示。

表 4　异常流量与正常流量的二分类任务

方法	Accuracy↑	Precision↑	Recall↑	F1-Score↑
SVM [12]	0.9419	0.9421	0.9421	0.9419
MLP [12]	0.9442	0.9462	0.9442	0.9443
RNN [13]	0.7233	0.6854	0.7233	0.6861
LSTM [13]	0.9325	0.9349	0.9325	0.9310
CNN-LSTM [14]	0.9746	0.9750	0.9746	0.9747
SACNN [35]	0.9738	0.9738	0.9738	0.9738
CNN2-SA^{ours}	0.9766	0.9766	0.9766	0.9766

表 5　异常流量与业务模式一的二分类任务

方法	Accuracy↑	Precision↑	Recall↑	F1-Score↑
SVM [12]	0.9625	0.9637	0.9625	0.9625
MLP [12]	0.9694	0.9709	0.9694	0.9693
RNN [13]	0.8569	0.8487	0.8569	0.8458
LSTM [13]	0.8410	0.8617	0.8410	0.8342
CNN-LSTM [14]	0.9767	0.9773	0.9766	0.9767
SACNN [35]	0.9710	0.9718	0.9710	0.9710
CNN2-SA^{ours}	0.9906	0.9907	0.9907	0.9907

表 6　异常流量与业务模式二的二分类任务

方法	Accuracy↑	Precision↑	Recall↑	F1-Score↑
SVM [12]	0.9312	0.9329	0.9312	0.9312
MLP [12]	0.9477	0.9499	0.9477	0.9477
RNN [13]	0.8562	0.8563	0.8563	0.8562
LSTM [13]	0.8625	0.8662	0.8625	0.8622
CNN-LSTM [14]	0.9735	0.9742	0.9735	0.9735
SACNN [35]	0.9225	0.9282	0.9225	0.9222
CNN2-SA^{ours}	0.9785	0.9788	0.9785	0.9786

表 7　业务模式一与业务模式二的二分类任务

方法	Accuracy↑	Precision↑	Recall↑	F1-Score↑
SVM [12]	0.9125	0.9190	0.9125	0.9122
MLP [12]	0.9653	0.9680	0.9653	0.9652
RNN [13]	0.8152	0.8237	0.8152	0.8141
LSTM [13]	0.7952	0.8149	0.7952	0.7853
CNN-LSTM [14]	0.9813	0.9818	0.9813	0.9812
SACNN [35]	0.9612	0.9622	0.9613	0.9612
CNN2-SA^{ours}	0.9826	0.9827	0.9826	0.9825

表 8　三分类任务

方法	Accuracy↑	Precision↑	Recall↑	F1-Score↑
SVM [12]	0.9159	0.9173	0.9159	0.9141
MLP [12]	0.9297	0.9316	0.9297	0.9296
RNN [13]	0.6745	0.6334	0.6745	0.6274
LSTM [13]	0.9200	0.9208	0.9200	0.9180
CNN-LSTM [14]	0.9700	0.9703	0.9700	0.9700
SACNN [35]	0.9540	0.9551	0.9540	0.9542
CNN2-SA^{ours}	0.9768	0.9771	0.9770	0.9770

从实验结果中，我们可以直接得到以下几点：

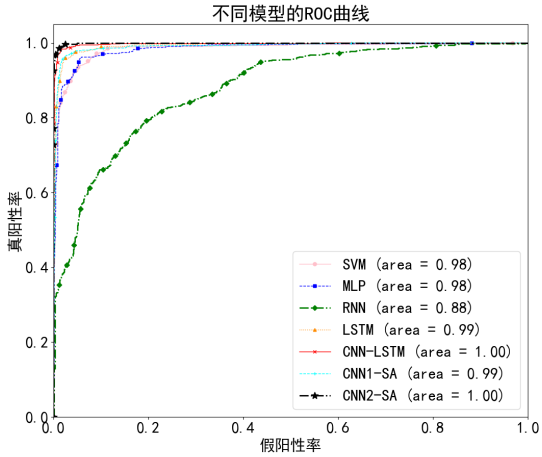
(1) 在所有的二分类任务和三分类任务中，本文提出的 CNN2—SA 模型的准确率、精确度、召回率和 F1 分数等指标均高于其他模型。

(2) 在所有的二分类任务和三分类任务中，CNN-LSTM 模型的准确率、精确度、召回率和 F1 分数等指标均高于除 CNN2—SA 模型外的其他模型。

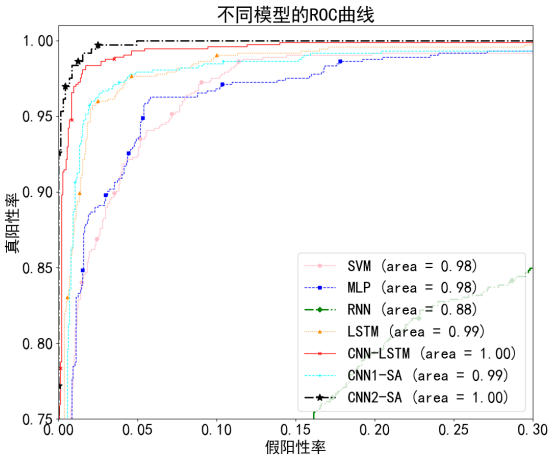
(3) 在不同的二分类任务中，SVM、RNN、LSTM、SACNN 模型的准确率、精确度、召回率和 F1 分数等指标均出现较为明显的变化，而 MLP、CNN-LSTM、CNN2-SA 模型的准确率、精确度、召回率和 F1 分数等指标变化较小。

根据这些信息，我们认为本文所提出的方法相比其他方法，在私有流量数据集上的多分类任务中表现更佳。

除了上述的各项模型评估指标，我们还绘制了各模型的 ROC 曲线图以及各模型训练过程中的训练准确率曲线图，用于更加准确地比较这些模型的性能。图 7 表示流量数据集上不同模型在三分类任务中的 ROC 曲线。图 8 表示流量数据集上不同模型在三分类任务中的模型训练准确率随训练轮次的变化。



(a) 各模型 ROC 曲线图



(b) 图像局部

图 7　各模型在三分类任务中的 ROC 曲线

图 7 中的虚线表示该方法总体分类性能的 ROC 曲线，实线则分别表示其他几种方法总体分类性能的 ROC 曲线。根据 ROC 曲线图像可以看出，CNN2—SA 模型的 ROC 曲线下方面积最大，且该曲线呈现出更加明显的向左上方倾斜，表明了 CNN2—SA 模型在三分类任务中具有非常好的分类性能。图 7 中 CNN-LSTM 模型和 CNN2-SA 模型的 AUC 值均为 1.00，表明这两种模型在测试集上具有非常优秀的分类性能。

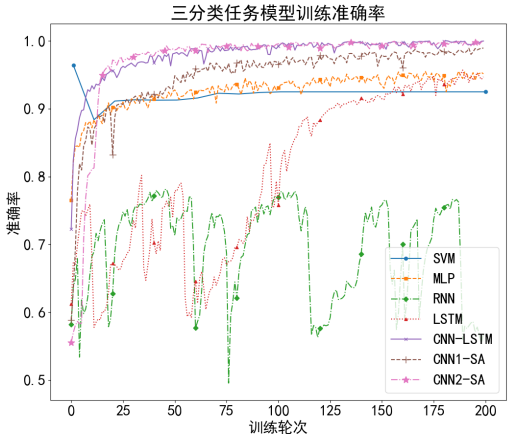


图 8　各模型在三分类任务中的训练准确率

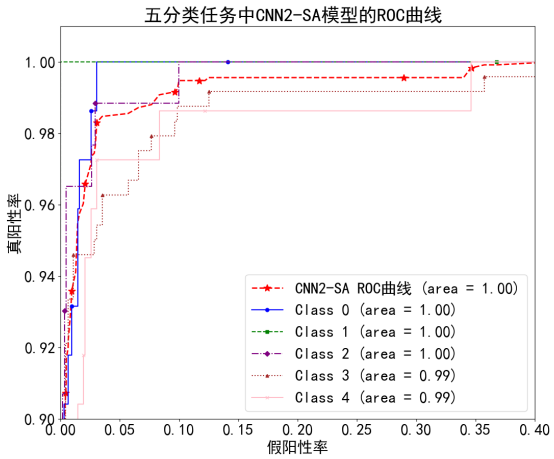
图 8 中带有“*”标记的粉色点划线代表 CNN2—SA 模型在训练过程中准确率随训练轮次的变化。根据图像可以看出，CNN2-SA 模型在训练轮次到达第 50 次时，准确率已基本达到收敛。与其他模型相比，CNN2—SA 模型的训练准确率较高，收敛速度快，且训练过程中模型的稳定性也明显较优。

综上所述，在私有流量数据集上，本文提出的 CNN2—SA 模型在二分类任务和三分类任务中的分类性能较好且稳定，受数据样本分布数量的影响较小，实验结果优于与本文比较的其他同类研究方法。

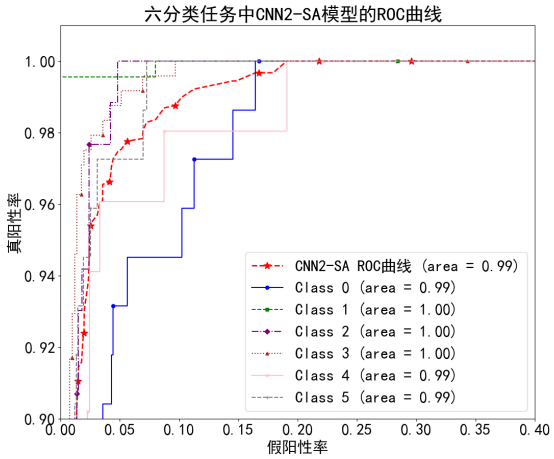
4.4　模型性能验证

为了进一步验证本文提出的 CNN2-SA 模型在多分类任务上的有效性，本文使用 UCR 上公开的 ElectricDevices 数据集对该模型性能进行评估，并与前面提到的其他方法进行了实验对比。

图 9(a)、9(b) 分别表示 ElectricDevices 数据集上五分类任务实验和六分类任务实验中 CNN2—SA 模型的总 ROC 曲线以及各个类别的 ROC 曲线。图 10(a)、10(b) 分别表示 ElectricDevices 数据集上五分类任务实验和六分类实验中的 CNN2—SA 模型与其他模型的训练准确率随训练轮次的变化。表 9 和表 10 分别表示 ElectricDevices 数据集上几种不同方法在五分类任务和六分类任务实验中的模型评估指标分数。



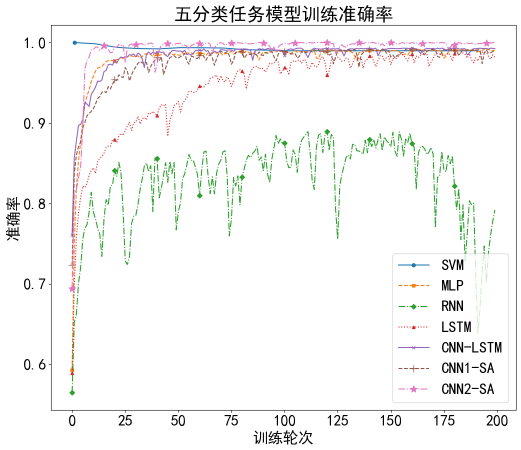
(a) 五分类任务



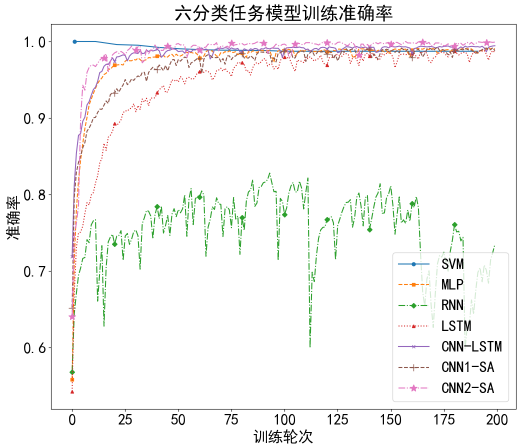
(b) 六分类任务

图 9　ElectricDevices 数据集上多分类任务中 CNN2—SA 模型的 ROC 曲线

从图 9(a) 中我们可以看到，CNN2-SA 模型的整体 ROC 曲线和各类别的 ROC 曲线均明显呈现向左上方倾斜，且曲线下方面积较大，这表明了 CNN2-SA 模型在五分类任务中具有较好的分类性能。与图 9(a) 相比，图 9(b) 中 CNN2-SA 模型的整体 ROC 曲线和各类别的 ROC 曲线均出现了下降和较为明显的波动，表明 CNN2-SA 模型在六分类任务中的分类性能有所下降，但整体仍具有较好的分类性能。



(a) 五分类任务



(b) 六分类任务

图 10　各模型在 ElectricDevices 数据集上多分类任务中的训练准确率

从图 10(a)、10(b) 中我们可以看到，CNN2-SA 模型在进行五分类任务和六分类任务的训练过程中，模型的训练准确率均在训练轮次到第 25 次时达到基本收敛。相比其他模型，CNN2-SA 模型的准确率较高，准确率收敛所需训练轮次较少，且训练过程中准确率更稳定。

表 9　五分类任务

方法	Accuracy↑	Precision↑	Recall↑	F1-Score↑
SVM [12]	0.6098	0.7554	0.6098	0.5826
MLP [12]	0.8072	0.8077	0.8072	0.8062
RNN [13]	0.7867	0.7940	0.7867	0.7802
LSTM [13]	0.9050	0.9068	0.9050	0.9050
CNN-LSTM [14]	0.9082	0.9101	0.9082	0.9080
SACNN [35]	0.8877	0.8899	0.8877	0.8877
CNN2-SA^{ours}	0.9816	0.9835	0.9816	0.9811

表 10　六分类任务

方法	Accuracy↑	Precision↑	Recall↑	F1-Score↑
SVM [12]	0.5942	0.7499	0.5942	0.5633
MLP [12]	0.7820	0.7808	0.7820	0.7802
RNN [13]	0.7619	0.7659	0.7619	0.7552
LSTM [13]	0.9072	0.9071	0.9072	0.9067
CNN-LSTM [14]	0.9047	0.9045	0.9047	0.9036
SACNN [35]	0.8795	0.8798	0.8795	0.8788
CNN2-SA^{ours}	0.9685	0.9698	0.9685	0.9677

从表 9 和表 10 中我们可以看到，在五分类任务和六分类任务实验中，CNN2-SA 模型的多项评估指标均高于其他模型。同时，将表 9-表 10 中的实验结果与表 4-表 8 中的实验结果进行对比，我们可以看到 CNN2-SA 模型的评估指标变化幅度相比其他模型的评估指标变化幅度要小得多。

综上所述，相较于其他方法，本文提出的方法不仅在私有流数据集上的多分类任务中具有较好的检测性能，还能在面临更加复杂的应用场景时，同样拥有较好的性能。

5　总　结

在云服务器广泛应用的互联网环境下，及时有效地检测出云服务器中的异常情况是对云服务提供商的重大挑战。本文提出了一种基于自注意力机制的云服务器业务流量异常检测方法。该方法首先对云服务器上的业务流量进行数据采集，将一维时序数据转化为二维曲线图像，并对曲线图像进行区域填充处理得到图像数据样本，作为模型输入。使用 CNN2—SA 模型对流量数据中的阶段性波动特征进行捕捉，同时分析数据中各阶段之间的依赖关系。该算法以云服务器上的不同业务模式流量作为研究对象，验证了本文提出的方法在云服务器业务流量异常检测方面取得了较好的效果。

虽然本文提出的异常检测方法具有较高的准确率，可以较好地满足目前市场的检测需求，但在未来的研究中，仍需要考虑加强模型对真实云环境变化的适应性，以便更好地应对未来云服务市场的需求变化。

参考文献

[1] J. Weinman, “The future of cloud computing,” in *2011 IEEE Technology Time Machine Symposium on Technologies Beyond 2020*. IEEE, 2011, pp. 1–2.

[2] K. Ansari, "Cloud computing on cooperative cars (c4s): An architecture to support navigation-as-a-service," in *2018 IEEE 11th International Conference on Cloud Computing (CLOUD)*. IEEE, 2018, pp. 794–801.

[3] Y. Zhou, C. Qian, Y. Guo, Z. Wang, J. Wang, B. Qu, D. Guo, Y. You, and X. Qu, "Xcloud-pfista: A medical intelligence cloud for accelerated mri," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 3289–3292.

[4] M. Darwich, E. Beyazit, M. A. Salehi, and M. Bayoumi, "Cost efficient repository management for cloud-based on-demand video streaming," in *2017 5th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud)*. IEEE, 2017, pp. 39–44.

[5] S. S. Sabet, S. Schmidt, S. Zadtootaghaj, C. Griwodz, and S. Moller, "Towards the impact of gamers strategy and user inputs on the delay sensitivity of cloud games," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2020, pp. 1–3.

[6] 谭伦, "中国公有云市场增速放缓迈入平稳发展期," 中国经营报, Tech. Rep., 2023-11-13.

[7] J. Shi, F. Lai, W. Li, H. Wang, X. Zhang, and Y. Li, "Anomaly detection of cloud network resource state based on deep learning," in *International Conference on Big Data and Security*. Springer, 2021, pp. 518–526.

[8] M. Zhao, A. Sadhu, and M. Capretz, "Multiclass anomaly detection in imbalanced structural health monitoring data using convolutional neural network," *Journal of Infrastructure Preservation and Resilience*, vol. 3, no. 1, p. 10, 2022.

[9] K. Choi, J. Yi, C. Park, and S. Yoon, "Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines," *IEEE Access*, vol. 9, pp. 120 043–120 065, 2021.

[10] I. Kotenko, I. Saenko, A. Kribel, and O. Lauta, "A technique for early detection of cyberattacks using the traffic self-similarity property and a statistical approach," in *2021 29th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*. IEEE, 2021, pp. 281–284.

[11] Y. Li and B.-X. Fang, "A lightweight online network anomaly detection scheme based on data mining methods," in *2007 IEEE International Conference on Network Protocols*. IEEE, 2007, pp. 340–341.

[12] T. Teik-Toe, Y. E. Jaddoo, and N. Y. Yen, "Machine learning based detection and categorization of anomalous behavior in enterprise network traffic," in *2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*. IEEE, 2019, pp. 750–754.

[13] G. Loukas, T. Vuong, R. Heartfield, G. Sakellari, Y. Yoon, and D. Gan, "Cloud-based cyber-physical intrusion detection for vehicles using deep learning," *IEEE Access*, vol. 6, pp. 3491–3508, 2018.

[14] I. Ullah and Q. H. Mahmoud, "Design and development of rnn anomaly detection model for iot networks," *IEEE Access*, vol. 10, pp. 62 722–62 750, 2022.

[15] J. D. Hamilton, *Time series analysis*. Princeton university press, 2020.

[16] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao, "Time series classification using multi-channels deep convolutional neural networks," in *International Conference on Web-age Information Management*. Springer, 2014, pp. 298–310.

[17] S. Oswal, S. Shinde, and M. Vijayalakshmi, "A survey of statistical, machine learning, and deep learning-based anomaly detection techniques for time series," in *International Advanced Computing Conference*. Springer, 2022, pp. 221–234.

[18] A.-R. Al-Ghuwairi, Y. Sharrab, D. Al-Fraihat, M. AlElaimat, A. Alsarhan, and A. Algarni, "Intrusion detection in cloud computing based on time series anomalies utilizing machine learning," *Journal of Cloud Computing*, vol. 12, no. 1, p. 127, 2023.

[19] V. Cerqueira, L. Torgo, and C. Soares, "Early anomaly detection in time series: a hierarchical approach for predicting critical health episodes," *Machine Learning*, pp. 1–22, 2023.

[20] W. H. Suh, S. Oh, and C. W. Ahn, "Metaheuristic-based time series clustering for anomaly detection in manufacturing industry," *Applied Intelligence*, pp. 1–20, 2023.

[21] H. Yang and J. Li, "Label contrastive learning for image classification," *Soft Computing*, pp. 1–10, 2023.

[22] S. Chavda and M. Goyani, "Scene level image classification: a literature review," *Neural Processing Letters*, vol. 55, no. 3, pp. 2471–2520, 2023.

[23] T. Xiangdong, "Image recognition algorithm based on hybrid deep learning," *International Journal of System Assurance Engineering and Management*, pp. 1–11, 2023.

[24] A. Banerjee and D. Banik, "Resnet based hybrid convolution lstm for hyperspectral image classification," *Multimedia Tools and Applications*, pp. 1–12, 2023.

[25] P. P. Sironmani and M. G. Augasta, "A novel cnn architecture with an efficient channelization for histopathological medical image classification," *Multimedia Tools and Applications*, pp. 1–21, 2023.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[27] H. Ge, L. Wang, M. Liu, X. Zhao, Y. Zhu, H. Pan, and Y. Liu, "Pyramidal multiscale convolutional network with polarized self-attention for pixel-wise hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–18, 2023.

[28] Z. Zhou, F. Liu, and Q. Wang, "R-transformer network based on position and self-attention mechanism for aspect-level sentiment classification," *IEEE Access*, vol. 7, pp. 127 754–127 764, 2019.

[29] Y. Cui, W. Li, L. Chen, L. Wang, J. Jiang, and S. Gao, "Feature fusion network model based on dual attention mechanism for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[30] X. Ma, B. Zheng, G. Jiang, and L. Liu, "Cellular network traffic prediction based on correlation convlstm and self-attention network," *IEEE Communications Letters*, vol. 27, no. 7, pp. 1909–1912, 2023.

[31] F. Bajić and J. Job, "Review of chart image detection and classification," *International Journal on Document Analysis and Recognition (IJDAR)*, pp. 1–22, 2023.

[32] F. A. Demmese, A. Neupane, S. Khorsandroo, M. Wang, K. Roy, and Y. Fu, "Machine learning based fileless malware traffic classification using image visualization," *Cybersecurity*, vol. 6, no. 1, p. 32, 2023.

[33] S. Ni, Q. Qian, and R. Zhang, "Malware identification using visualization images and deep learning," *Computers & Security*, vol. 77, pp. 871–885, 2018.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[35] W. Lu, Y. Duan, and Y. Song, "Self-attention-based convolutional neural networks for sentence classification," in *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*. IEEE, 2020, pp. 2065–2069.