计算机科学
COMPUTER SCIENCE

基于知识蒸馏和高效通道注意力的异常检测

周士金, 邢红杰

引用本文

周士金, 邢红杰. 基于知识蒸馏和高效通道注意力的异常检测[J]. 计算机科学, 2023, 50(11A): 220900034-10.

ZHOU Shijin, XING Hongjie. Novelty Detection Method Based on Knowledge Distillation and Efficient Channel Attention [J]. Computer Science, 2023, 50(11A): 220900034-10.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220900034

基于知识蒸馏和高效通道注意力的异常检测

周士金　邢红杰
河北大学数学与信息科学学院河北省机器学习与计算智能重点实验室　河北 保定 071002
(549409090@qq.com)

摘　要　基于知识蒸馏的异常检测方法通常将经过预训练的网络作为教师网络,并将与该教师网络的模型结构及规模大小相同的网络用作学生网络,对于待测数据,利用教师网络与学生网络之间的差异判定其为正常数据或异常数据。然而,教师网络与学生网络的结构和规模均相同,一方面,会使得基于知识蒸馏的异常检测方法在异常数据上产生的差异过小;另一方面,教师网络的预训练数据集在规模上远大于学生网络的训练集,这会使得学生网络产生大量的冗余信息。为了解决上述问题,将高效通道注意力(Efficient Channel Attention,ECA)模块引入到基于知识蒸馏的异常检测方法中,利用ECA的跨通道交互策略,设计比教师网络结构更简单且规模更小的学生网络,既可以有效地获取正常数据的特征,去除冗余信息,又能增大教师网络与学生网络之间的差异,提高异常检测的性能。在6个图像数据集上的实验结果表明,与其他5种相关方法相比,所提方法取得了更优的检测性能。

关键词:异常检测;知识蒸馏;注意力机制;教师网络;学生网络
中图法分类号　TP391.4

Novelty Detection Method Based on Knowledge Distillation and Efficient Channel Attention

ZHOU Shijin and XING Hongjie
Hebei Key Laboratory of Machine Learning and Computational Intelligence,College of Mathematics and Information Science,Hebei University,Baoding,Hebei 071002,China

Abstract　The knowledge distillation based novelty detection method usually utilizes the pre-trained network as the teacher network. The network that has the same model structure and size as the teacher network is used as the student network. For testing data,the difference between the teacher network and the student network is utilized to discriminate them as normal or novel. However,the teacher network and the student network have the same network structure and size. On the one hand,the knowledge distillation based novelty detection method may produce a small difference in the novel data. On the other hand,because the pre-trained data set of the teacher network is much larger in scale than the training set of the student network,the student network may thus obtain lots of redundant information. To solve this problem,the efficient channel attention(ECA) module is introduced into the knowledge distillation based novelty detection method. Utilizing the cross-channel interaction strategy,the student network with a simpler network structure and smaller size in comparison with the teacher network is designed. Hence,the features of the normal data can be efficiently obtained. The redundant information may be removed. The difference between the teacher network and the student network can also be enlarged. Moreover,the novelty detection performance may be improved. In comparison with 5 related methods,experimental results on the 6 image data sets demonstrate that the proposed method obtains better detection performance.

Keywords　Novelty detection,Knowledge distillation,Attention mechanism,Teacher network,Student network

1　引言

异常检测模型以无监督方式从正常数据中对单类分类器进行学习,并利用所学的模型将待测数据分类为正常数据或异常数据[1]。异常检测模型在网络安全[2]、金融欺诈检测[3]、工业故障和损害检测[4]、医疗诊断[5]等领域得到了广泛应用。与正常数据相比,异常检测问题中的异常数据非常稀少甚至难以获取。当仅由正常数据构成的训练集规模较小时,异常检测模型的学习往往不够充分,以致其不能取得令人满意的检测性能。为了解决该问题,相关学者将知识蒸馏[6-7]引入到

基金项目:国家自然科学基金(61672205);河北省自然科学基金(F2017201020);河北大学高层次人才科研启动项目(521100222002);河北大学附属医院基金项目(2019Q003);复杂能源系统智能计算教育部工程研究中心开放基金(ESIC202101)
This work was supported by the National Natural Science Foundation of China(61672205),Natural Science Foundation of Hebei Province(F2017201020),High-Level Talents Research Start-Up Project of Hebei University(521100222002),Affiliated Hospital Foundation Project of Hebei University(2019Q003) and Open Foundation of Engineering Research Center of Intelligent Computing for Complex Energy Systems(ESIC202101).
通信作者:邢红杰(hjxing@hbu.edu.cn)

异常检测模型中。

知识蒸馏最早由 Hinton 等[8]提出，它首先利用预训练数据集对教师网络进行训练，将教师网络带有温度参数的归一化指数函数的输出用作知识，并在训练集上利用教师网络的知识指导学生网络的训练。对于多类分类问题，相较于一个正类标签，多个负类标签蕴含更多教师网络的推理信息，通过调整温度参数对教师网络在负类标签上的概率值进行调整，使负类标签对应的概率值增大，从而学生网络可从数据中获取更多教师网络的推理信息，有效缩短其训练时间并提高泛化性能。一般情况下，知识蒸馏可以有效地减小学生网络的规模并使其获得较优的分类性能。

为了提高学生网络的收敛速度和测试准确率，Komodakis 等[9]提出了基于激活张量的注意力转移方法，首先提取教师网络中的特征图，然后在其通道维度上进行数据统计得到空间注意力图，最后将所得注意力图用作指导学生网络训练的知识。Huang 等[10]提出了一种网络加速和压缩的方法，即神经元选择性迁移，该方法利用最大均值差异计算教师网络与学生网络中间层激活值分布之间的差异，并使用真实标签和教师网络中间层中的激活值分布对学生网络进行训练。Kim 等[11]提出了因子迁移，将两个卷积模块分别作为释义器和翻译器，首先利用释义器把教师网络的特征图转换为教师因子，使用翻译器将学生网络的特征图转换为学生因子，利用二者之间的差异训练学生网络。Heo 等[12]提出了基于激活边界的知识迁移，而激活边界源于隐含层神经元，在所提方法的蒸馏过程中，学生网络对教师网络中每个神经元所确定的激活区域与非激活区域之间的分离边界进行学习，以使激活迁移损失达到最小。Passalis 等[13]提出了基于信息流建模的异构知识蒸馏，将教师网络中各个层次的信息流建模用作训练学生网络的知识。Passalis 和 Tefas[14]提出了概率知识迁移，利用概率分布对教师网络特征空间中的知识进行建模，并通过最小化教师网络和学生网络概率分布之间的散度实现知识迁移。Jin 等[15]提出了基于路径约束优化的蒸馏方法，在教师网络训练过程的参数空间中选取一些锚点记录教师网络的最优路径，并利用锚点信息对学生网络进行分步训练。Chen 等[16]提出了基于语义校准的跨层知识蒸馏方法，利用注意力机制为学生网络中每一特征层自动分配最合适的教师网络目标层，且学生网络的每一特征层可以同时利用教师网络的多个特征层的知识进行知识蒸馏。

如上所述，知识蒸馏已被成功地用于解决多类分类问题。然而，将知识蒸馏应用于解决异常检测问题仍处于起始阶段。最近，Zhang 等[6]针对异常检测问题提出了基于生成式对抗网络(Generative Adversarial Nets，GAN)的渐进式知识蒸馏方法，将预训练的 GANomaly[17]用作教师网络，使用蒸馏损失训练与其结构完全相同的学生网络，并通过渐进式学习方式对学生网络进行细化训练，有效提高了模型的检测性能和训练阶段的稳定性。Salehi 等[7]提出了用于异常检测的多分辨率知识蒸馏，将 ImageNet 上预训练所得的专家网络用作教师网络，并将教师网络多个层次上的特征蒸馏到一个结构相同但规模较小的学生网络中，有效地解决了异常检测因训练数据不足引起检测性能下降的问题。

众所周知，将注意力机制引入到深度卷积神经网络中可以显著提高其性能[18]。Hu 等[19]为卷积神经网络提出了专注于通道注意力的挤压和激励网络，在挤压阶段，对特征图沿空间维度进行压缩，进一步使用激活函数对压缩后的向量进行缩放得到通道注意力；在激励阶段，将通道注意力与相应的特征图结合在一起产生新的特征图，并通过增强卷积特征通道间的联系提高网络的表示性能。Woo 等[20]提出了基于卷积块的注意力模块，首先求取特征图沿空间维度上的通道注意力，然后将其与对应的特征图相乘得到通道注意力特征图；然后，由通道注意力特征图沿通道维度求取空间注意力，将所得空间注意力图与特征图相乘得到最终的特征图，并利用自适应特征细化提高模型的表示能力。Hu 等[21]为卷积神经网络提出了聚集-激发框架，该方法利用特征间的语境信息增强网络的表示能力。在聚集阶段，利用聚集算子对特征图沿着空间维度聚合局部的语境信息；在激发阶段，将所得语境信息还原回原始尺寸得到最终的特征图。为了校准全卷积神经网络，Roy 等[22]提出了融入空间及通道注意力的挤压和激励块，首先求取特征图沿着空间维度的通道注意力以及沿通道维度的空间注意力，然后将特征图与两种注意力相结合分别得到通道注意力特征图和空间注意力特征图，对两个特征图进行逐像素比较，得到在每个像素上都具有较大值的特征图，利用空间注意力和通道注意力一起对所得特征图进行校准，以提高模型的分割能力。Gao 等[23]提出了全局二阶池化卷积神经网络块，即在卷积神经网络中引入全局二阶池化，并在通道维度及空间维度上计算特征图的全局二阶信息，调整特征图以提高网络的非线性表示学习能力。Fu 等[24]提出了一种具有自注意机制的双注意力网络，利用双注意力网络从通道及位置两个维度自适应整合局部及全局依赖关系，以增强卷积神经网络的特征表示能力，并提高模型在场景分割中的判别能力和分割性能。最近，Wang 等[18]提出了高效通道注意力(Efficient Channel Attention，ECA)，对特征图沿空间维度进行压缩得到通道权重，利用一维快速卷积在通道权重不降维的前提下实现当前通道和相邻若干通道的信息交互得到通道注意力。相较于其他通道注意力块，ECA 能更高效地获取通道注意力，并能显著提高卷积神经网络的表示性能。

当由正常数据构成的训练集规模较小时，传统的异常检测方法难以取得较好的检测性能。为了解决该问题，Salehi 等[7]提出了基于多分辨率知识蒸馏的异常检测(Multiresolution Knowledge Distillation for Anomaly Detection，MKDAD)方法，将预训练的 VGG16 模型中的特征提取网络用作教师网络，并采用知识蒸馏的方式对与其网络结构相同但规模较小的学生网络进行训练。然而，在 MKDAD 中，学生网络与教师网络的模型结构完全相同，使得它们在部分异常数据上产生的差异过小，从而无法正确检测这些异常数据。此外，教师网络的预训练样本来自多个不同的类别，而学生网络的训练集仅由正常数据构成，当正常数据较少时，采用与教师网络相同的复杂结构，会导致学生网络存在很多的冗余信息，从而影响异常检测的性能。为了提高 MKDAD 的检测性能，本文提出了基于知识蒸馏和高效注意机制的异常检测方法。本文的主要贡献如下：

(1)利用 ECA 代替 MKDAD 学生网络中的部分 ReLU 激活层，使得学生网络具有与教师网络不同的模型结构，增大了学生网络和教师网络在异常数据上的差异，提高了异常检测方法的检测性能。

(2)将 ECA 模块融入轻量级学生网络，使学生网络利用局部跨通道交互策略，提高了学生网络对正常数据特征的提取能力也提高了网络的泛化能力，降低了异常检测方法的误报率和漏报率并显著减少了推理时间。

(3)在 MNIST，Fashion-MNIST，CIFAR-10，SVHN，SLT10 及 MVTecAD 工业异常检测数据集上与相关方法进行了实验比较，验证了所提方法的有效性。

2 相关工作

本节将简要回顾基于多分辨率知识蒸馏的异常检测(MKDAD)和高效通道注意力(ECA)。

2.1 用于异常检测的多分辨率知识蒸馏

用于异常检测的多分辨率知识蒸馏(MKDAD)[7]的模型结构如图 1 所示。它由教师网络和学生网络构成。将在 ImageNet 数据集上预训练的 VGG16 模型的特征提取网络用作教师网络。学生网络使用与教师网络完全相同的网络结构，但是其参数量规模小于教师网络的参数量规模。将教师网络多个特征层的激活值作为知识，利用教师网络中不同中间层的信息对学生网络进行训练，以提高学生网络的收敛速度。

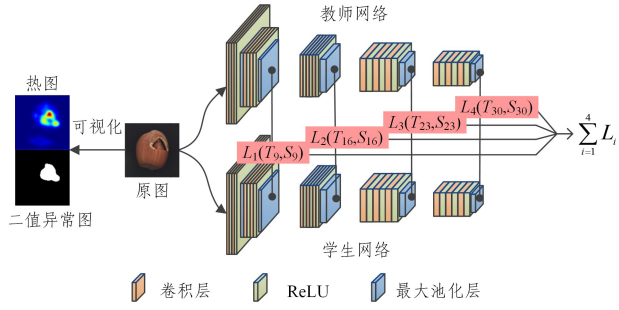


图 1 MKDAD 的模型结构

Fig. 1 Model structure of MKDAD

在训练过程中，训练数据被同时输入学生网络和教师网络，提取二者多个对应特征层的激活值，将两个网络对应特征层中激活值的相似度之和作为损失函数。该损失函数由两部分构成，分别对应特征层中激活值的欧氏距离和余弦相似度。此外，为了便于优化，对余弦相似度进行了修改。欧氏距离、修改后的余弦相似度和损失函数分别表示为：

$$L_{\text{val}}=\sum_{i=1}^{N_{CP}}\frac{1}{N_i}\sum_{j=1}^{N_i}(\boldsymbol{a}_t^{CP_i}(j)-\boldsymbol{a}_s^{CP_i}(j))^2 \tag{1}$$

$$L_{\text{dir}}=\sum 1-\frac{vec(\boldsymbol{a}_t^{CP_i})^{\mathrm{T}}\cdot vec(\boldsymbol{a}_s^{CP_i})}{\|vec(\boldsymbol{a}_t^{CP_i})\|\ \|vec(\boldsymbol{a}_s^{CP_i})\|} \tag{2}$$

和

$$L_{\text{total}}=\sum_{i=1}^{4}L_{\text{val}}^{(i)}+\lambda L_{\text{dir}}^{(i)} \tag{3}$$

其中，$\boldsymbol{a}_t^{CP_i}$ 表示教师网络的第 i 个特征层的激活值；$\boldsymbol{a}_s^{CP_i}$ 表示学生网络的第 i 个特征层的激活值；$vec(\cdot)$表示矢量化函数，它能将任意大小的矩阵转换为一维向量；λ 为超参数；i 表示参与蒸馏的特征层序号。

使学生网络和教师网络对应特征层尽可能相同是 MKDAD 的训练目标，因此，对于正常数据，学生网络可以提取与教师网络相同的特征。由于训练集中没有异常数据，因此学生网络无法有效地提取异常数据的特征，而教师网络的训练集中存在异常数据，它可以有效地提取异常数据的特征，二者对应特征层的激活值会有较大的差异，故使用式(3)所示的相似度作为判定异常的依据。此外，MKDAD 根据 L_{total} 的梯度来寻找对损失函数影响最大的像素，称其为属性图，并获得输入数据的异常定位图。属性图和异常定位图分别表示为：

$$\boldsymbol{\Lambda}=\frac{\partial L_{\text{total}}}{\partial x} \tag{4}$$

$$L_{\text{map}}=(g_\delta(\boldsymbol{\Lambda})\ominus B)\oplus B \tag{5}$$

其中，$\boldsymbol{\Lambda}$ 表示梯度图，$g_\delta(\cdot)$表示标准差为 δ 的高斯滤波，B 是一个二元映射，$\ominus$和$\oplus$表示图像的侵蚀和膨胀。

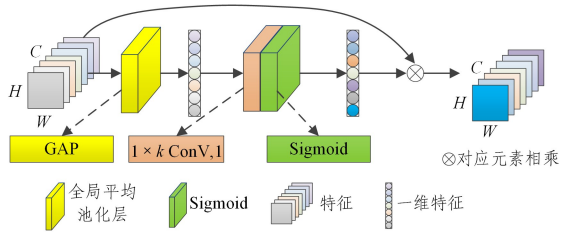


图 2 ECA 的模型结构

Fig. 2 Model structure of ECA

2.2 高效通道注意力

高效通道注意力 ECA[18]是使用了不降维局部跨通道交互策略的高效通道注意力模块，其模型结构如图 2 所示，由全局平均池化层、一维快速卷积层和 Sigmoid 激活层 3 部分组成。该模块首先沿着空间维度获取各通道的权重，利用一维快速卷积对 k 个通道权重进行信息交互，使卷积神经网络的性能取得了明显提升。此外，该模块仅使用了少量的参数。通道权重表示为：

$$w_i=\sigma(\sum_{j=1}^{k}\alpha^j y_i^j),\ y_i^j\in\Omega_i^k \tag{6}$$

其中，$\sigma(\cdot)$表示 Sigmoid 激活函数，α^j 表示参数向量中第 j 个值，Ω_i^k 表示 y_i 的 k 个相邻通道的集合，y_i 表示第 i 个通道的权重，即：

$$y_i=\frac{1}{WH}\sum_{j=1,k=1}^{H,W}x_{j,k} \tag{7}$$

其中，W 和 H 分别表示特征图的宽和高，$x_{j,k}$ 为特征图中的一个像素。另外，式(6)可以通过卷积核大小为 k 的一维快速卷积实现，即：

$$w=\sigma(C1D_k(y)) \tag{8}$$

为了实现局部跨通道的交互能力，需要确定通道交互的覆盖范围 k，其值可使用自适应方式求取，即：

$$k=\phi(C)=\left|\frac{\log_2(C)}{\gamma}+\frac{b}{\gamma}\right|_{\text{odd}} \tag{9}$$

其中，C 表示特征图的通道数，$|t|_{\text{odd}}$ 表示最接近 t 的奇数，γ 和 b 表示超参数。

3 基于知识蒸馏和高效通道注意力的异常检测

本节将从模型结构和损失函数两方面详细介绍所提方法，为了方便描述，将所提方法简写为 ECA_MKDAD。

3.1 模型结构

ECA_MKDAD 的模型如图 3 和图 4 所示。ECA_MKDAD 由教师网络 T 和学生网络 S 构成。将在 ImageNet 数据集上预训练的 VGG16 的特征提取网络用作教师网络 T。为了便于知识蒸馏的实现，学生网络 S 与教师网络 T 采用相似的网络结构，由于教师网络的预训练数据来自于多个不同的类别，而学生网络的训练集仅由正常数据构成。当正常数据较少时，采用与教师网络相同的复杂结构，会导致学生网络

存在很多冗余信息，从而影响异常检测的性能。由图 3 和图 4 可知，VGG16 网络的特征提取网络采用 ReLU 激活函数[25]。由于 VGG16 的预训练集 ImageNet 规模非常大且其网络结构较为复杂，因此，相较于 Sigmoid 激活函数，采用 ReLU 激活函数可以有效提高模型的训练速度且防止出现过拟合问题。然而在基于知识蒸馏的异常检测方法中，学生网络 T 的训练集仅由正常数据构成，且学生网络需要模仿教师网络的输出。

由图 3 和图 4 可知，教师网络 T 和学生网络 S 均由卷积神经网络构成，实验结果表明注意力机制可以显著提高卷积神经网络的性能[18]，而通道注意力可以对特征图的不同通道进行加权，为有重要特征的通道赋予更高的权重，从而更好地提取特征，因此在学生网络 S 中加入 ECA 可以提高其特征提取的能力。批归一化[26]（Batch normalization）可将一个批量中的数据整合到统一的区间中，减少数据的发散并且降低模型的训练难度，并且在一定程度上保持原有数据的分布。使用批归一化层对数据分布进行整合，可以使得 ECA 更加快速地求取通道注意力。因此，在学生网络 S 中使用批归一化层和高效通道注意力 ECA 的组合代替部分 ReLU 层，可以有效地提高学生网络 S 对正常数据特征的提取能力，增大学生网络 S 所提取的异常数据的特征与正常数据的特征之间的差异。此外，由于教师网络的预训练集远大于学生网络的训练集，因此使用与教师网络相同参数规模的学生网络，必然会使其学习到较多的冗余信息，从而影响异常检测的性能，故学生网络 S 的参数规模应小于教师网络 T 的参数规模，如图 4 所示。

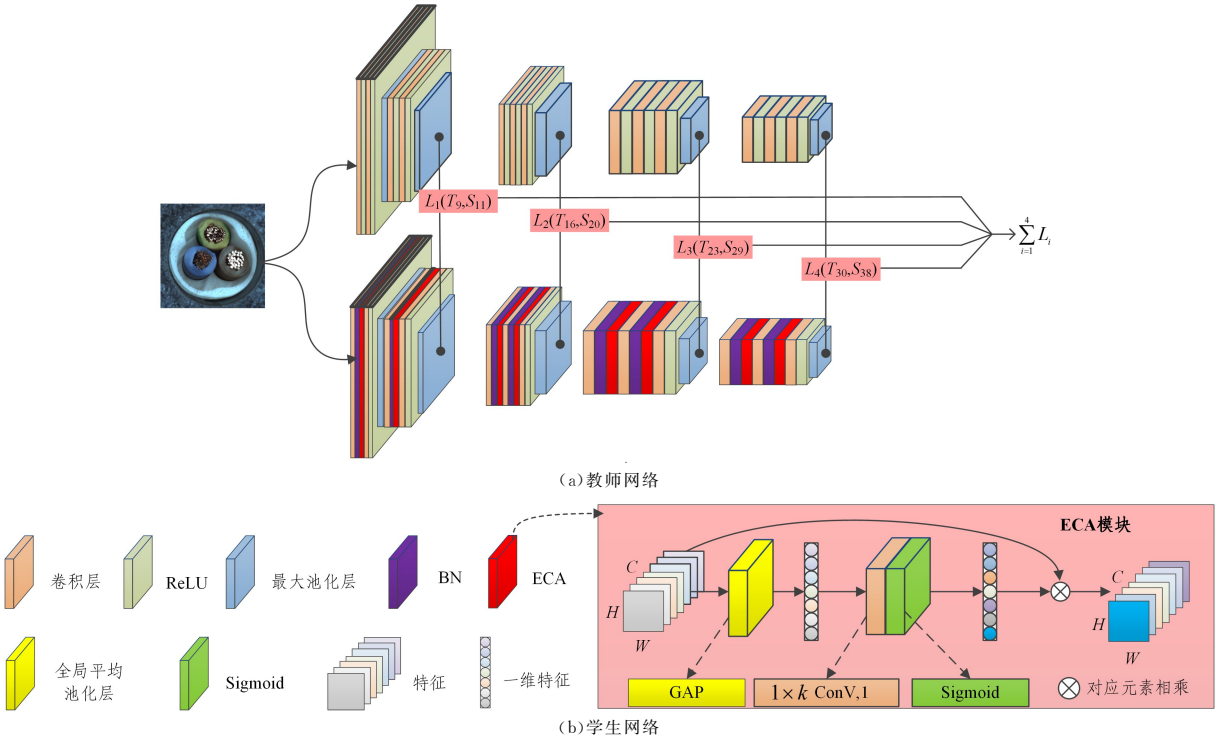


图 3 ECA_MKDAD 的模型结构

Fig. 3 Model structure of ECA_MKDAD

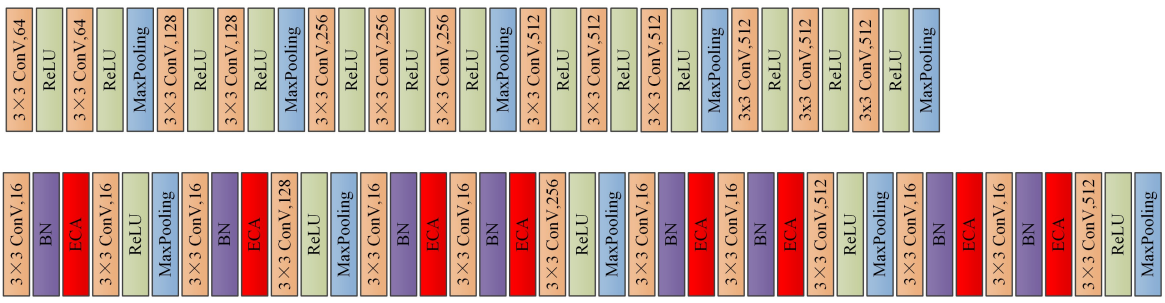


图 4 ECA_MKDAD 的网络架构

Fig. 4 Network architecture of ECA_MKDAD

ECA[18]是使用了不降维局部跨通道交互策略的高效通道注意力模块，给定包含 n 个通道的特征图 $\boldsymbol{x}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，ECA 的输出特征图为 $\boldsymbol{y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\}$，表示为：

$$\boldsymbol{y}_c=\sigma\left(C1D_k\left(\frac{1}{WH}\sum_{i=1,j=1}^{H,W}\boldsymbol{x}_c^{i,j}\right)\right)\boldsymbol{x}_c \tag{10}$$

其中，$\boldsymbol{y}_c$ 和 $\boldsymbol{x}_c$ 分别表示输出特征图 $\boldsymbol{y}$ 和输入特征图 $\boldsymbol{x}$ 在第 c 个通道的特征矩阵，$\boldsymbol{x}_c^{i,j}$ 表示 $\boldsymbol{x}_c$ 中的一个像素点。下面将 ECA 与动态 ReLU（DY-ReLU）[27]及 SENet[19]进行比较，以展示 ECA 的优越性。

DY-ReLU 在特征图 $\boldsymbol{x}$ 上的输出特征图 $\boldsymbol{y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\}$ 表示为：

$$\boldsymbol{y}_c=f_{\theta(\boldsymbol{x})}\boldsymbol{x}_c=\max_{1\leqslant m\leqslant M}\{a_c^m(\boldsymbol{x})\boldsymbol{x}_c+b_c^m(\boldsymbol{x})\} \tag{11}$$

其中，$\theta(\cdot)$ 表示超函数，用于计算激活函数的参数；$f_{\theta(\boldsymbol{x})}(\cdot)$

是激活函数，表示用 $\theta(\boldsymbol{x})$ 生成通道的激活值；M 表示表达式个数；a_c^m 和 b_c^m 表示斜率和偏置。

SENet 是 DY-ReLU 的一个特例，它在特征图 $\boldsymbol{x}$ 上的输出特征图 $\boldsymbol{y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\}$ 表示为：

$$\boldsymbol{y}_c=\sigma\left(\boldsymbol{W}_2\delta\left(\boldsymbol{W}_1\frac{1}{WH}\sum_{i=1,j=1}^{H,W}\boldsymbol{x}_c^{i,j}\right)\right)\boldsymbol{x}_c \tag{12}$$

其中，$\boldsymbol{W}_1\in\mathbb{R}^{\frac{C}{r}\times C}$，$\boldsymbol{W}_2\in\mathbb{R}^{C\times\frac{C}{r}}$，$\delta(\cdot)$ 为 ReLU 激活函数。由式(12)可知 SENet 是使用了降维跨通道交互策略的通道注意力模块，其为有重要特征的通道赋予更高的权重，更好地提取特征，但该方法破坏了通道与其权值的直接对应关系[18]，ECA 完美地克服了上述问题。

利用 ResNet-50 作为骨干模型在 ImageNet 上比较 DY-ReLU，SENet 和 ECA 模块的参数量，DY-ReLU 为 $2KC^2/r$[27]，SENet 为 $2\times C^2/r$，而 ECA 则只需要卷积核大小为 K 的一维卷积[18]的参数量。

综上，利用 ECA 代替 ReLU 层能使卷积神经网络学习到有效的通道注意力，此外，使用一维快速卷积实现不降维局部跨通道交互策略，即可用少量参数实现 k 个通道间的信息交互，能有效加快学生网络 S 的收敛速度，降低 ECA_MKDAD 的模型复杂度并提高其性能。

此外，卷积神经网络中不同层次的特征图包含不同的语义特征[12-13,15-16]，因此，我们利用知识蒸馏的思想，将教师网络 T 中多个不同层次中间层的特征图作为知识，利用不同的中间层信息对学生网络 S 进行训练，从而使学生网络 S 可以充分地学习教师网络 T 中的知识，提高异常检测的性能。为了使学生网络学习到更重要的知识且减少计算量，我们将教师网络 T 中最大池化层的特征图作为知识进行蒸馏，将该层称为教师网络 T 的蒸馏层，将学生网络 S 中与教师网络 T 的蒸馏层交互的网络层称为学生网络 S 的蒸馏层。为了简化计算，两个网络对应蒸馏层及其输入的卷积层的参数设置完全相同，如图 4 所示。

在测试阶段，当待测数据为正常数据时，教师网络 T 和学生网络 S 提取的特征相同；当待测数据为异常数据时，预训练的教师网络 T 由于使用了性能优异的 VGG16 网络并使用 ImageNet 数据集进行预训练，可以较为准确地提取异常数据的特征，然而学生网络 S 规模小于教师网络 T，且训练集中无异常数据，故学生网络 S 不能准确地提取异常数据特征。又因为学生网络融入了高效通道注意力，进一步扩大了两个网络对应蒸馏层特征图的差异，因此学生网络 S 和教师网络 T 对应蒸馏层特征图的差异较大，以此将该待测数据判定为异常数据，从而提高了异常检测的性能。

3.2 损失函数

将蒸馏过程中的蒸馏层特征图称为 $\boldsymbol{DL}$，$\boldsymbol{DL}_T^{(i)}$ 和 $\boldsymbol{DL}_S^{(i)}$ 分别表示教师网络 T 中第 i 个蒸馏层的特征图和学生网络 S 中第 i 个蒸馏层的特征图。将 $\boldsymbol{DL}_T^{(i)}$ 作为知识蒸馏到 $\boldsymbol{DL}_S^{(i)}$ 中，参照文献[17]的设置，将 $\boldsymbol{DL}_T^{(i)}$ 和 $\boldsymbol{DL}_S^{(i)}$ 间的欧氏距离和改进的余弦相似性度量综合起来作为蒸馏损失函数，定义如下：

$$L_{\mathrm{EKAD}}=\sum_{i=1}^{N_{DL}}\left(\frac{1}{N_i}\sum_{j=1}^{N_i}(\boldsymbol{DL}_T^{(i)}(j)-w_E^{(i)}\boldsymbol{DL}_S^{(i)}(j))^2+\lambda\left(1-\frac{flat(\boldsymbol{DL}_T^{(i)})^{\mathrm{T}}\cdot flat(w_E^{(i)}\boldsymbol{DL}_S^{(i)})}{\|flat(\boldsymbol{DL}_T^{(i)})\|\,\|flat(w_E^{(i)}\boldsymbol{DL}_S^{(i)})\|}\right)\right) \tag{13}$$

其中，N_{DL} 表示蒸馏层的个数，N_i 表示特征层的特征数。$flat(\cdot)$ 表示矢量化函数，将任意大小的矩阵依次排序后组成一维向量。$\boldsymbol{w}_E^{(i)}$ 表示第 i 个蒸馏层使用 ECA 求得的注意力权重，注意力权重定义如下：

$$w_E=\sigma(C1D_k(\frac{1}{WH}\sum_{i=1,j=1}^{H,W}x_{i,j})) \tag{14}$$

其中，H 和 W 分别表示特征图的高和宽。$\sigma(\cdot)$ 表示 Sigmoid 激活函数，$C1D_k$ 表示卷积核大小为 k 的一维高速卷积。

在所提模型的训练阶段仅使用正常数据进行训练，目标是使学生网络 S 和教师网络 T 对应蒸馏层的特征图相同，因此可最小化损失函数(13)直至其值为零[7]。在测试阶段，当待测数据为异常数据时，学生网络 S 和教师网络 T 对应蒸馏层的特征图会有较大的差异，因此可将式(13)中的损失函数用作异常得分并设置阈值进行异常检测。

4 实验

为了验证所提 ECA_MKDAD 的有效性，将它与 6 种相关方法在 6 个基准数据集上进行了实验比较。

4.1 数据集及其参数设置

实验中所使用的 6 个图像数据集分别为 MNIST[28]，Fashion-MNIST[29]，CIFAR-10[30]，SVHN[31]，STL-10[32] 和 MVTecAD[33]。

以下实验中，针对每个数据集，依次选取数据集中的单类图像作为正常数据，其余类图像作为异常数据，组成多个异常检测数据集。将原训练集中的正常数据作为训练数据。将原测试集的所有数据作为测试数据。以 MNIST 数据集为例，MNIST(0)表示由 MNIST 原训练集中类别为 0 的图像数据作为正常数据所构成的数据集。

对所提 ECA_MKDAD，使用在 ImageNet 数据集中预训练的 VGG16 网络并仅使用特征提取网络作为教师网络，MNIST，Fashion-MNIST，CIFAR-10 和 SVHN 的图像大小均设置为 32×32 像素，STL-10 的图像大小设置为 96×96 像素，MVTecAD 的图像大小设置为 128×128 像素。根据图像数据的不同复杂程度，设置了不同的迭代次数，MNIST 和 Fashion-MNIST 迭代 100 次，CIFAR-10，SVHN 和 STL-10 迭代 300 次，MVTecAD 迭代 500 次。训练过程使用学习率为 0.1 的 Adam 优化器，batchsize 的大小为 32，因为 MVTecAD 较为复杂，故其超参数 λ 为 0.1，其他数据集的超参数 λ 为 0.01。ECA_MKDAD 模型中所包含的网络均在 PyTorch 框架下搭建，编程语言 Python 的版本为 3.6.10，GPU 型号为 NVIDIA GeForce GTX TITAN X。

此外，为了衡量各种异常检测模型的性能，本文使用 ROC 曲线下的面积(AUROC)作为性能度量指标。AUROC 是判断预测模型优劣的标准，其特性是无论数据集正常数据和异常数据是否存在不平衡，随机猜测的基线始终是 0.5[34]。AUROC 是最常用的性能度量之一[35]，但是当测试集高度不平衡时，AUROC 可能产生过于乐观的结果[40]。因此，本文还使用几何均值(gmean)、误报率(FPR)、漏报率(FNR)作为度量指标来比较了模型的性能，其表达式如下：

$$gmean=\sqrt{\frac{TP*TN}{(TP+FN)*(TN+FP)}} \tag{15}$$

$$FPR=1-\frac{TP}{TP+FN} \tag{16}$$

$$FNR=\frac{FP}{TN+FP} \tag{17}$$

其中，TP 表示正常数据被正确预测的数量，TN 表示正常数据被误判的数量，FP 表示异常数据被正确预测的数量，FN 表示异常数据被误判的数量。

分类的阈值通过使用约登指数寻找最佳 ROC 的阈值来确定。

4.2 实验结果

为了比较不同异常检测方法的性能，将 AUROC 作为主要评价指标，同时使用几何均值、漏报率、误报率作为评价指标，6 种相关方法分别为 MKDAD[7]，GANomaly[17]，MemAE[36]，f-AnoGAN[37]，DSVDD[38] 和 RKDAD[39]。所有异常检测方法在 MNIST，Fashion-MNIST，CIFAR-10，SVHN，STL-10 和 MVTecAD 上的 AUROC、几何均值、误报率和漏报率如表 1－表 5 所列，最优结果加粗显示。

表 1 6 种不同方法在 6 种图像数据集上的 AUROC 测试性能

Table 1 AUROC testing results of six different methods on six different Image datasets

数据集	MKDAD	GANomaly	MemAE	f-AnoGAN	DSVDD	ECA_MKDAD
MVTecAD	0.8908	0.7003	0.5548	0.7427	0.8338	**0.9107**
MNIST	0.9803	0.9603	0.8825	0.8906	0.9359	**0.9916**
Fashion-MNIST	0.9464	0.9012	0.8690	0.9195	0.9116	**0.9495**
CIFAR-10	0.8460	0.7415	0.6525	0.6311	0.6106	**0.8489**
SVHN	0.7330	0.5482	0.5210	0.5597	0.6068	**0.7444**
STL-10	0.9312	0.5689	0.6089	0.5603	0.6081	**0.9439**

表 2 6 种不同方法在 6 种不同图像数据集上的几何均值平均测试性能

Table 2 Gmean average test results of six different methods on six different image datasets

数据集	MKDAD	GANomaly	MemAE	f-AnoGAN	DSVDD	ECA_MKDAD
MNIST	0.9576	0.9066	0.8145	0.8301	0.8754	**0.9614**
FashionMNIST	0.8873	0.8353	0.8022	0.8593	0.8539	**0.8917**
CIFAR-10	0.7735	0.6914	0.5909	0.5568	0.5315	**0.7834**
SVHN	0.6674	0.4953	0.4053	0.5151	0.5750	**0.6812**
STL-10	0.8600	0.4749	0.5514	0.5113	0.5296	**0.8711**
MVTecAD	0.8245	0.6538	0.5266	0.5391	0.7289	**0.8538**

表 3 6 种不同方法在 6 种不同图像数据集上的误报率平均测试性能

Table 3 FPR average test results of six different methods on six different image datasets

数据集	MKDAD	GANomaly	MemAE	f-AnoGAN	DSVDD	ECA_MKDAD
MNIST	0.0538	0.0923	0.1742	0.1905	0.1197	**0.0448**
FashionMNIST	0.0823	0.2038	0.2285	0.1616	0.1582	**0.0700**
CIFAR-10	0.2099	0.3261	0.4866	0.5184	0.3671	**0.2018**
SVHN	0.4066	0.2721	0.2915	0.2821	0.3530	**0.3978**
STL-10	0.1225	0.5623	0.5647	0.2757	0.4427	**0.1052**
MVTecAD	0.1366	0.3157	0.4981	0.3652	0.2773	**0.1067**

表 4 6 种不同方法在 6 种不同图像数据集上的漏报率平均测试性能

Table 4 FNR average test results of six different methods on six different image datasets

数据集	MKDAD	GANomaly	MemAE	f-AnoGAN	DSVDD	ECA_MKDAD
MNIST	**0.0307**	0.0935	0.1949	0.1459	0.1288	0.0324
FashionMNIST	0.1369	0.1848	0.1649	**0.1174**	0.1294	0.1279
CIFAR-10	**0.2417**	0.2719	0.2575	0.2486	0.4557	0.2492
SVHN	0.2407	0.6452	0.6628	0.6199	0.4840	**0.2245**
STL-10	0.1556	0.2834	0.2356	0.6268	0.3676	**0.1511**
MVTecAD	0.2005	0.3399	0.2413	0.6123	0.2455	**0.1771**

表 5 6 种不同方法在 MVTecAD 图像数据集上的 AUROC 测试性能

Table 5 AUROC test results of six different methods on MVTecAD

数据集	MKDAD	GANomaly	MemAE	DSVDD	RKDAD	ECA_MKDAD
Bottle	0.9952	0.8000	0.4250	0.8810	0.9905	**0.9984**
Cable	0.8951	0.7371	0.5400	0.8255	**0.9271**	0.9040
Capsule	0.8109	0.7180	0.6210	0.8085	0.7308	**0.8113**
Carpet	0.8411	0.6376	0.2010	0.8535	0.7771	**0.8640**
Grid	0.7970	0.8028	**0.9040**	0.7845	0.6642	0.8764
Hazelnut	0.9904	0.7271	0.6030	0.7957	0.9682	**0.9932**
Leather	0.9507	0.7867	0.3630	0.7921	0.6090	**0.9817**
Metal_Nut	0.7483	0.5528	0.3390	0.6403	**0.8279**	0.7801
Pill	0.8494	0.6328	0.7050	0.8118	0.7870	**0.8550**
Screw	0.8418	0.6852	**1.0000**	0.9367	0.9467	0.9502
Tile	0.9271	0.6851	0.4190	0.7893	0.7879	**0.9426**
Toothbrush	**0.9306**	0.4028	0.4060	0.9861	0.8833	0.9167
Transistor	0.8608	0.7817	0.7260	0.8421	**0.8775**	0.8758
Wood	**0.9719**	0.9123	0.5950	0.9386	0.9237	0.9614
Zipper	**0.9522**	0.6429	0.4750	0.8217	0.8997	0.9496

(1)MKDAD是基于知识蒸馏的异常检测方法中最有代表性的方法之一。其使用教师网络和学生网络之间的差异做异常检测和异常定位，较先前的方法相比，有计算成本低和训练稳定的特点[7]。与MKDAD相比，ECA_MKDAD使用BN和ECA的组合代替学生网络中的部分ReLU激活函数，仅增加了极少的计算复杂度，使学生网络和教师网络间网络结构的差异增加，并增强了学生网络对正常数据特征的提取能力，故ECA_MKDAD取得了更优的结果，由表2、表3可知，在gmean和误报率评价标准下，ECA_MKDAD取得了最优的结果。

(2)f-AnoGAN和GANomaly是使用生成式对抗网络做异常检测的具有代表性的方法。前者提出了一种编码器将测试数据快速映射到GAN的潜在空间的某个点，并使用WGAN做异常检测，相较于之前方法，其推理速度更快。后者使用了编码器-解码器-编码器的结构作为生成器，极大地提高了异常检测的速度，并取得了很好的性能。与f-AnoGAN和GANomaly相比，ECA_MKDAD使用预训练的VGG16网络作为教师网络，当训练数据较少的时候，ECA_MKDAD可以取得非常好的性能，如表1中STL-10和MVTecAD实验结果所示。由表3和表4可知，ECA_MKDAD有极低的误报率和漏报率，在工业生产中有明显优势。

(3)MemAE是一种使用带有存储功能的自编码器结构进行异常检测的方法，在训练阶段将有代表性的瓶颈特征存储到记忆模块中，在测试阶段将测试数据的瓶颈特征在记忆模块中寻找最接近的特征组合。与传统基于自编码器做异常检测的方法相比，MemAE克服了由于自编码器优秀的泛化性能而导致无法检测异常的缺点。与MemAE相比，ECA_MKDAD则使用教师网络和学生网络之间的差异作为异常检测依据。由表3可知，MemAE的存储功能极大地增加了误报的风险，在工业生产中存在较大的问题。

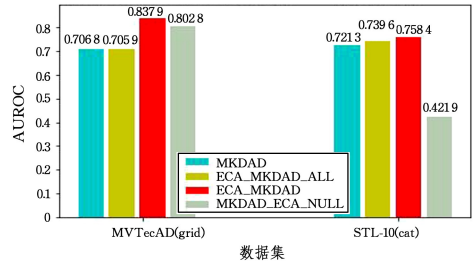
由表1可知，ECA_MKDAD在MVTecAD，MNIST，Fashion-MNIST，CIFAR-10，SVHN以及STL-10这6种不同的图像数据集中均取得了最优的ROAUC结果；由表2可知，ECA_MKDAD在6个数据集上均取了最优的gmean平均测试结果；工业领域较为重视误报率和漏报率，由表3可知，ECA_MKDAD在6个数据集上均取得了最优的误报率平均测试结果；由表4可知，除了MNIST，FashionMNIST和CIFAR-10外，ECA_MKDAD在3个数据集上均取得了最优的误报率平均测试结果；表5列出了6种不同方法在无监督异常检测的综合真实世界数据集MVTecAD上的AUROC测试结果，由表5可知，ECA_MKDAD在7个类别的数据集中表现出了最优的性能，在其余类别的数据集均表现出了次优的性能。

(4)DSVDD是一种将深度神经网络和传统SVDD相结合的异常检测方法，其利用深度神经网络对数据做特征提取，同时将输出空间优化为包含全部输入数据的超球来判定异常，表现出了非常优异的性能。与DSVDD相比，ECA_MKDAD使用了预训练的VGG16网络，并考虑不同特征层的输出，更加全面地使用了数据的特征，从而取得了更好的成绩。

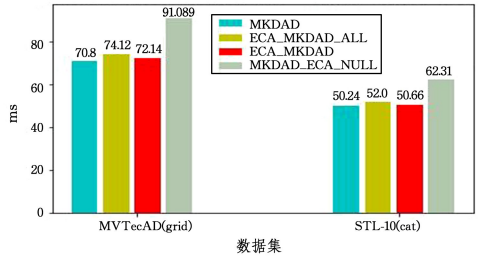
(5)RKDAD是基于知识蒸馏的异常检测方法，该方法将两个特征层之间的内积定义为“FSP”矩阵，利用教师网络和学生网络各自“FSP”矩阵的差异来判定异常，表现出了优异的性能。与RKDAD相比，ECA_MKDAD将高效通道注意力ECA添加到学生网络中，增加了异常数据在学生网络和教师网络之间的差异，有效地提高了异常检测的性能。由表5可知，在工业异常检查数据集MVTecAD中，ECA_MKDAD在13个类别的数据集中均优于RKDAD。

4.3 消融实验

为了探讨使用注意力机制替换学生网络ReLU激活函数的有效性以及对推理时间的影响，本文分别在MVTecAD(grid)和STL-10(cat)数据集上做消融实验进行验证。MKDAD学生网络中一共只有13个ReLU激活函数，所提方法保留最大池化层前的ReLU激活函数，经过大量实验验证，当仅替换8个ReLU激活函数，保留蒸馏层前的ReLU激活函数时得到最优效果，替换率为61.5%。为了展现注意力机制代替学生网络部分ReLU激活函数的合理性，对照了不替换学生网络ReLU的方法(ReLU的替换率为0%)，将其命名为MKDAD，将使用ECA替换MKDAD学生网络全部ReLU激活函数的方法(ReLU的替换率为100%)命名为ECA_MKDAD_ALL，将ECA模块放在MKDAD学生网络的部分ReLU激活函数层之后(ReLU的替换率为0%)的方法命名为MKDAD_ECA_NULL，实验结果如图5所示。



(a)4种方法在两个数据集实验的AUROC



(b)4种方法在两个数据集实验的推理时间

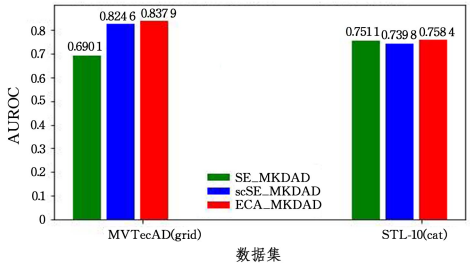
图5 使用注意力机制代替学生网络不同数量的ReLU激活函数在不同数据集上的AUROC和推理时间

Fig. 5 Attentional mechanism is used to replace AUROC and inference time of different numbers of ReLU activation functions on different datasets

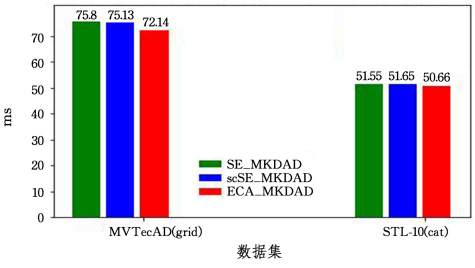
为了验证使用ECA注意力机制替换学生网络部分ReLU激活函数的高效性，对照了使用SENet[23]替换学生网络部分ReLU激活函数的方法(将其命名为SE_MKDAD)和使用空间与通道注意力模块scSE[26]替换学生网络部分ReLU激活函数的方法(将其命名为scSE_MKDAD)。分别求取了SE_MKDAD，scSE_MKDAD和ECA_MKDAD在两个数据集上的AUROC和推理时间，实验结果如图6所示。

由图5(a)中MKDAD，ECA_MKDAD_ALL，ECA_MKDAD以及MKDAD_ECA_NULL的AUROC值可知，ECA_MKDAD在两个不同数据集上均表现出了最佳性能。相较于

MKDAD，使用 ECA 代替学生网络部分 ReLU 激活函数，扩大了学生网络和教师网络网络结构的差异，并提高了学生网络提取正常数据的能力，因此获得了更好的性能。相较于 MKDAD_ECA_NULL，ECA_MKDAD 去除了部分 ReLU 结构，使得 ECA 可以提取更丰富的特征信息，从而可以更好地求得注意力，同时，模型更加轻量，只需更短的推理时间即可得到较高的异常检测性能。相较于 ECA_MKDAD_ALL，ECA_MKDAD 根据学生网络和教师网络对应蒸馏层的特征图的差异来判定异常，所以需要保证蒸馏层最相邻的激活函数保持一致，这样可以保证两个网络对正常数据识别的准确率，故 ECA_MKDAD 表现出了最佳性能。由图 5(b)中 MKDAD，ECA_MKDAD_ALL 和 ECA_MKDAD 的推理时间可知，与 MKDAD 相比，ECA_MKDAD 在 MVTecAD(grid)数据集上增加了 1.4 ms，在 STL-10(cat)数据集上仅增加了 0.4 ms，性能得到了较大的提升，证明了使用 ECA 代替学生网络部分 ReLU 激活函数的方法的合理性。



(a)3 种方法在两个数据集实验的 AUROC



(b)3 种方法在两个数据集实验的推理时间

图 6 使用不同注意力机制代替学生网络部分 ReLU 激活函数在不同数据集上的 AUROC 和推理时间

Fig. 6 Different attention mechanisms are used to replace AUROC and inference time of partial ReLU activation functions on different datasets

由图 6 中 ECA_MKDAD，SE_MKDAD 和 scSE_MKDAD 的 AUROC 值和推理时间可知，ECA_MKDAD 在两个不同数据集上均表现出了最佳性能。在 MVTecAD(grid)数据集上 scSE_MKDAD 与 ECA_MKDAD 的 AUROC 值相近，但是其推理时间却多出了 3 ms；在 STL-10(cat)数据集上 SE_MKDAD 与 ECA_MKDAD 的 AUROC 值相近，但是其推理时间也多出了 0.9 ms。验证了 ECA 相较于 SE 和 scSE 的高效性，也验证了使用 ECA 注意力机制替换学生网络部分 ReLU 激活函数的高效性。

所提模型的学生网络参数规模小于教师网络，使用参数规模较小的学生网络可以去除冗余信息，并且可以显著地减少推理时间，从而提高异常检测性能。为此，对照方法使用与教师网络同等参数规模的学生网络，在 STL-10 数据集上进行验证，在相同的环境下，均训练迭代 300 次，记录其 AUROC 和推理时间，实验结果如图 7、图 8 所示。

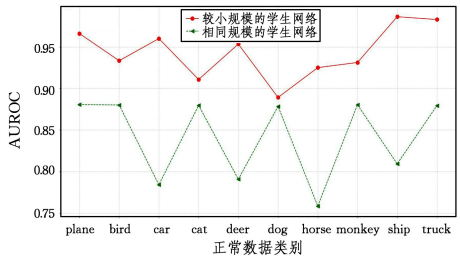


图 7 使用不同规模的学生网络在 STL-10 数据集上的 AUROC

Fig. 7 AUROC on STL-10 dataset using student networks of different network sizes

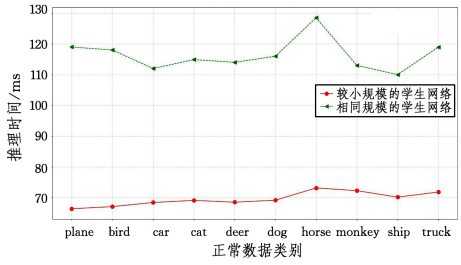


图 8 使用不同规模的学生网络在 STL-10 数据集上的推理时间

Fig. 8 Inference time on STL-10 dataset using student networks of different network sizes

如图 7 所示，在 STL-10 数据集上选取不同的类别作为正常数据，和对照方法相比，所提方法完全超越了对照方法，表现出更优的异常检测性能。因此验证了使用较小参数规模的学生网络可以有效地去除冗余信息，从而提高异常检测性能。如图 8 所示，所提方法的推理时间远低于对照方法的推理时间，对照方法的推理时间甚至是所提方法的 2 倍。综上，使用较小参数规模的学生网络可以有效地去除教师网络中的冗余信息，并且可以显著地减少推理时间，有效地提高了异常检测的性能。

为了更加直观地比较教师网络 T 和学生网络 S 的参数规模，使用 fvcore 库分别计算了二者在 MVTecAD 数据集上的参数量和浮点运算数(FLOPs)，结果如表 6 所列。由表 6 可知，学生网络的参数量和 FLOPs 远低于教师网络，由此可知学生网络的参数规模小于教师网络的参数规模。

表 6 教师网络和学生网络在 MVTecAD 数据集上的参数规模

Table 6 Parameter scale of teacher network and student network on MVTecAD dataset

模型	骨干网络	参数量	浮点运算数/G
教师网络 T	VGG16	14.71×10^6	160.35
学生网络 S	VGG16(轻量级)+ECA	0.34×10^6	7.38

为了更加直观地比较异常数据在学生网络 S 和教师网络 T 对应蒸馏层的特征图的差异，我们选取 MVTecAD(bottle)数据集，分别对学生网络 S 和教师网络 T 多个蒸馏层的特征及二者对应蒸馏层的特征值之差进行了可视化，如图 9 所示。图中颜色越深说明该区域的数值越大，最后一列是二者对应蒸馏层的特征差异，颜色越深差异越明显。从图 9 可以直观地看出，当异常数据出现时，学生网络 S 未能提取与教师网络 T 相同的特征，且存在明显差异。因此使用二者对应蒸馏层特征差异作为判定异常的依据是可行的。

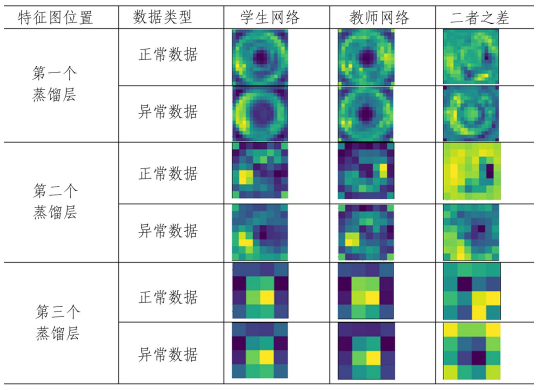


图 9　在 MVTecAD 数据集上学生网络和教师网络各蒸馏层的特征及二者对应蒸馏层特征值差异可视化

Fig. 9　Visualization of distillation layer characteristics in student network and teacher network on MVTecAD dataset

结束语　在基于知识蒸馏的异常检测方法中，以 MKDAD 为例，由于教师网络与学生网络有相同的网络类型和结构，因此学生网络和教师网络对测试数据中部分异常数据给出了相似的特征图，从而产生较小的差异，导致这些数据被错分为正常数据，影响了异常检测的性能。针对上述问题，本文提出了基于 ECA 和知识蒸馏的异常检测方法。该方法通过引入高效的注意机制模块 ECA，在提高学生网络性能的同时，增加学生与教师网络类型和结构的差异，当异常数据出现时，二者对应蒸馏层的特征图会出现较大的差异，从而提高了异常检测的性能。此外还证明了 ECA 代替 ReLU 的合理性。

在未来工作中，将会从两个方面对所提方法进行探索：(1)当训练数据与预训练教师网络的数据有明显差异时，可能会影响知识的有效性，因此需要探索更好的教师网络或者更好的优化策略，如首先对教师网络微调优化，再进行蒸馏；(2)寻找更好的异常检测评价方法，充分利用蒸馏的知识来提高异常检测的性能。

参 考 文 献

[1] RUFF L, KAUFFMANN J R, VANDERMEULEN R A, et al. A Unifying Review of Deep and Shallow Anomaly Detection[J]. Proceedings of the IEEE, 2021, 109(5): 756-795.

[2] MALAIYA R K, KWON D, KIM J, et al. An Empirical Evaluation of Deep Learning for Network Anomaly Detection[C]// 2018 International Conference on Computing, Networking and Communications(ICNC). IEEE, 2018.

[3] ZHENG Y J, ZHOU X H, SHENG W G, et al. Generative adversarial network based telecom fraud detection at the receiving bank[J]. Neural Networks, 2018, 102: 78-86.

[4] ZHAO R, YAN R, CHEN Z, et al. Deep learning and its applications to machine health monitoring[J]. Mechanical Systems and Signal Processing, 2019, 115(15): 213-237.

[5] GUO P, XUE Z, MTEMA Z, et al. Ensemble Deep Learning for Cervix Image Selection toward Improving Reliability in Automated Cervical Precancer Screening[J]. Diagnostics(Basel), 2020, 10(7): 451.

[6] ZHANG Z, CHEN S, SUN L. P-KDGAN: Progressive Knowledge Distillation with GANs for One-class Novelty Detection[C]// Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. 2020.

[7] SALEHI M, SADJADI N, BASELIZADEH S, et al. Multiresolution Knowledge Distillation for Anomaly Detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Nashville, TN, USA, 2021.

[8] HINTON G, VINYALS O, DEAN J. Distilling the Knowledge in a Neural Network[J]. arXiv: 1503.02531, 2015.

[9] ZAGORUYKO S, KOMODAKIS N. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer[J]. arXiv: 1612.03928, 2016.

[10] HUANG Z, WANG N. Like What You Like: Knowledge Distill via Neuron Selectivity Transfer[J]. arXiv: 1707.01219, 2017.

[11] KIM J, PARK S, KWAK N. Paraphrasing complex network: Network compression via factor transfer[J]. Advances in Neural Information Processing Systems, 2018, 31: 2765-2774.

[12] HEO B, LEE M, YUN S, et al. Knowledge Transfer via Distillation of Activation Boundaries Formed by Hidden Neurons[C]// Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, USA, 2019.

[13] PASSALIS N, TZELEPI M, TEFAS A. Heterogeneous Knowledge Distillation Using Information Flow Modeling[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, Seattle, WA, USA, 2020.

[14] PASSALIS N, TEFAS A. Learning Deep Representations with Probabilistic Knowledge Transfer[C]// Proceedings of the European Conference on Computer Vision(ECCV). Cham: Springer, 2018.

[15] JIN X, PENG B, WU Y, et al. Knowledge Distillation via Route Constrained Optimization[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV). 2019.

[16] CHEN D, MEI J P, ZHANG Y, et al. Cross-Layer Distillation with Semantic Calibration[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021.

[17] AKCAY S, ATAPOUR-ABARGHOUEI A, BRECKON T P. GANomaly: Semi-supervised Anomaly Detection via Adversarial Training[C]// Asian Conference on Computer Vision. Springer, 2018.

[18] WANG Q, WU B, ZHU P, et al. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks[C]// IEEE/CVF Conference on Computer Vision Pattern Recognition, Seattle, WA, USA, 2020.

[19] HU J, SHEN L, SUN G. Squeeze-and-Excitation Networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Salt Lake City, UT, USA, 2018.

[20] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional Block Attention Module[C]// Proceedings of the European Conference on Computer Vision(ECCV). Cham: Springer, 2018.

[21] HU J, SHEN L, ALBANIE S, et al. Gather-Excite: Exploiting Feature Context in Convolutional Neural Networks[C]// Advances in Neural Information Processing Systems 31(NeurIPS 2018). 2018.

[22] ROY A G, NAVAB N, WACHINGER C. Recalibrating Fully Convolutional Networks With Spatial and Channel "Squeeze and Excitation" Blocks[J]. IEEE Transactions on Medical Imaging, 2019, 38(2): 540-549.

[23] GAO Z, XIE J, WANG Q, et al. Global Second-Order Pooling

Convolutional Networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019.
[24] FU J, LIU J, TIAN H, et al. Dual Attention Network for Scene Segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2019.
[25] NAIR V, HINTON G E. Rectified Linear Units Improve Restricted Boltzmann Machines[C]// Proceedings of the 27th International Conference on International Conference on Machine Learning. Haifa, Israel, 2010: 807-814.
[26] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]// Proceedings of the International Conference on Machine Learning, PMLR, 2015.
[27] CHEN Y, DAI X, LIU M, et al. Dynamic ReLU [C]// Proceedings of the Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020.
[28] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
[29] XIAO H, RASUL K, VOLLGRAF R. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms [J]. arXiv: 1708. 07747, 2017.
[30] KRIZHEVSKY A, HINTON G. Learning Multiple Layers of Features from Tiny Images[J/OL]. https://citeseerx. ist. psu. edu/viewdoc/download; jsessionid = 0D60E5DD558A91470E0EA1725FF36E0A?doi=10. 1. 1. 222. 9220&rep=rep1&type=pdf.
[31] NETZER Y, WANG T, COATES A, et al. Reading digits in natural images with unsupervised feature learning[J/OL]. http://ufldl. stanford. edu/housenumbers/nips2011_housenumbers. pdf.
[32] COATES A, NG A, LEE H. An Analysis of Single-Layer Networks in Unsupervised Feature Learning[C]// Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. 2011.
[33] BERGMANN P, FAUSER M, SATTLEGGER D, et al. MVTec AD— A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Los Alamitos, CA, USA, 2019.
[34] FAWCETT T. An introduction to ROC analysis [J]. Pattern Recognition Letters, 2006, 27(8): 861-874.
[35] CAMPOS G O, ZIMEK A, SANDER J, et al. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study [J]. Data Mining Knowledge Discovery, 2016, 30(4): 891-927.
[36] GONG D, LIU L, LE V, et al. Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Long Beach, CA, USA, 2019.
[37] SCHLEGL T, SEEBOCK P, WALDSTEIN S M, et al. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks[J]. Med Image Anal, 2019, 54: 30-44.
[38] RUFF L, VANDERMEULEN R, GOERNITZ N, et al. Deep One-Class Classification[C]// Proceedings of the 35th International Conference on Machine Learning, Stockholm. PMLR, 2018.
[39] CHENG H, YANG L, LIU Z. Relation-Based Knowledge Distillation for Anomaly Detection[C]// Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Springer, 2021.

ZHOU Shijin, born in 1997, postgraduate. His main research interests include novelty detection and generative adversarial network.

XING Hongjie, born in 1976, Ph.D, professor, master supervisor. His main research interests include kernel methods, neural networks, novelty detection, and ensemble learning.