基于扩张卷积自编码器的时序数据增强方法

邓文文[1] 徐海洋[1] 申 艺[1]
DENG Wenwen XU Haiyang SHEN Yi

摘 要 优秀且智能的识别模型建立于大量的数据基础之上，而优质数据的来源却很稀少。样本数量不足都会造成识别模型的类别分布不均衡、模型泛化能力差的问题。为了解决这一难题，研究者们提出了几何变换、加白噪音、神经网络等诸多数据增强方法以丰富数据的多样性。然而对于时序数据这种有明确时间先后的序列数据，传统的增强手段会破坏信息在时间上的联系，因此，文章提出了一种基于自定义损失函数的扩张卷积自编码器的半监督数据生成方案。实验结果显示，采用本方法 92% 的生成数据和原始数据的平均相关性大于 0.9，时间差不超过 0.03 s。这表明该方案可以在保留时序信息的前提下，实现增强数据以增加时序数据的多样性。

关键词 神经网络；扩张卷积；自编码器；时序数据；数据增强

doi：10.3969/j.issn.1672-9528.2025.01.030

0 引言

此前，深度神经网络处理时序数据在很大程度上依赖于训练使用的数据集大小和信息表达的一致性，然而这类数据在解决现实问题时往往并不丰富，通常带有限制因素和约束条件，数据样本获取难度大[1]。因此，增加数据量的有效方法是使用数据增强技术，通过添加噪声或重新排列以生成新的数据[2]。

时序数据的收集需人工操作，成本高且耗时长。例如，在利用声音信号进行距离感知[3]、耳道信息认证[4]的系统中，因需用户长时间配合，不仅降低了其使用系统的满意度，还会增加隐私泄露的风险。而用户参与时间短时，将限制数据的获取，导致数据样本较少，影响识别效果。另一方面在数据的采集处理过程中，受采集方式、类别、次数不同，有效的样本的量也不同，这会导致识别时类别间的不平衡，也会降低识别精度。结合以上两点，本文提出了基于扩张卷积自编码网络的数据扩充方法，通过该网络生成大量不同的数据来解决数据样本不足的问题。

1 数据增强方法及类型

数据增强不仅能够帮助模型学习更健壮和更具代表性的特征，减少过拟合，还能提高其处理现实世界和场景变化的能力。传统的数据增强方式是在原有数据的基础上进行各种转换或操作来完成的，例如几何变换、裁剪以及添加噪声等实现的。这种实现方式算法简单，易于实现，但对于时序数据容易造成数据失真问题。传统的数据扩充不能直接运用到时序数据的原因主要有两点。一是时序数据的每个时间点刻画了对象的瞬时状态，其值随时间变化。例如在距离感知的应用中代表该时刻离探测器的距离，在耳道型形变的感知中代表在该时刻的耳道形变状况[5]。如果通过简单的翻转或旋转方式增加样本的多样性，则会破坏样本本身的时序特点，产生错误的数据会降低识别的准确率。二是时序信号的增强方法通常采用增加噪声方式，但其主要难点在于增加噪声的量难以确定。增加的量过多，会导致信号失真，无法代表原始样本。如果增加的量过少，又不能提供足够的样本多样性。通过人工的方式确定增加噪声的量需要大量的时间成本。

综合以上两点，本文采用基于扩张卷积神经网络的方式生成数据样本。该方式不仅可以丰富生成样本的多样性，而且还保证生成的样本仍然可以被分类网络正确分类。

2 扩张卷积简介

为了解决语义分割等密集预测问题，英特尔实验室等受到小波分解算法中带有扩张滤波器卷积操作的启发，提出了一种新的卷积网络模块，其主要部分是扩张卷积[6]对普通卷积的泛化，如图 1，它通过扩展因子控制核之间的间隙，增加了卷积核感受输入的范围。扩张卷积公式为：

$$(F*_l k)(p)=\sum_{s+lt}F(s)k(t) \quad (1)$$

式中：l 表示扩张系数；F 和 k 分别表示输入和卷积核大小。当 l =1 时，该式就表示普通卷积；当 l >1 时，则表示不同程度的扩张。

1. 中国航空工业集团公司西安航空计算技术研究所
陕西西安 710065

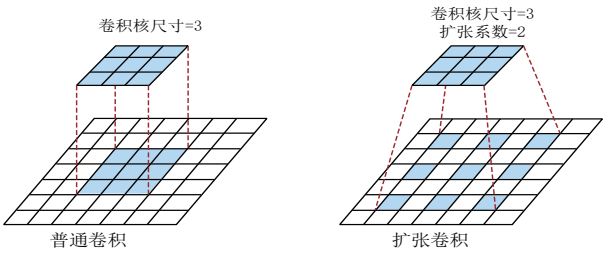


图 1 扩张卷积示意图

因此，扩张卷积可以通过不同的扩张系数，使相同配置的卷积操作感受不同的输入范围，实现多尺度信息的提取，这表明扩张卷积可以从多尺度系统聚合上下文的信息，使特征图上的单个特征点能够表示的输入图像的更大区域。因此，扩张卷积常被用于上下文信息密集的预测任务，例如语音处理（如 WaveNet）[7]，以提高其准确性。

综上所述，扩张卷积可以捕获与常规卷积相同级别的细节，但参数更少，这可以使模型更高效，更容易训练。其次不同扩张率的卷积核对输入进行多尺度分析，可以使得模型能够捕捉从细粒度到粗粒度的特征。

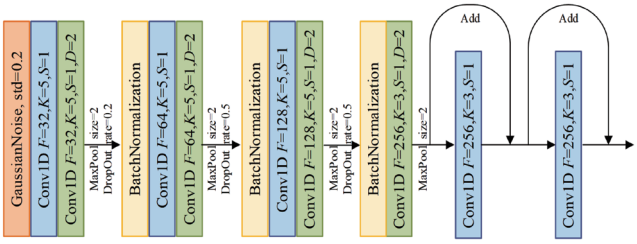
3 基于扩张卷积自编码器的数据扩充模型

3.1 模型的结构

受到对抗生成网络（GAN 网络）[8] 以及自编码器（AE 网络）[9] 的启发，本文将利用扩张卷积网络形成具有对称结构的自编码器以生成数据，并且将生成数据的识别结果作为损失函数的一部分，保证生成的数据能够保留原本的数据特点。

（1）编码器的结构

基于自编码的生成网络由编码器和解码器两部分组成，其中编码器的结构如图 2 所示，其中 F 表示通道数，K 表示卷积核大小，S 代表卷积操作的步长，D 代表系数卷积的扩张率。



S= 步长　K= 卷积核尺寸　D= 膨胀卷积系数

图 2 AE 生成网络编码器结构

该编码器的设计主要考虑了三点：

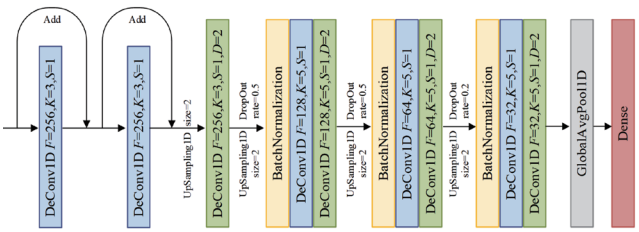
首先，编码器包含了普通卷积和稀疏卷积两种卷积模式，稀疏卷积的扩张率决定了卷积核中值之间的间距，这种扩张操作允许卷积层拥有更大的接受域，这意味着它可以在输入数据中捕获更多的上下文和更大的特征，尤其在处理输入时序数据中的长期依赖关系时会凸显其优势。使用两种卷积操作，有助于多尺度地提取时序序列中的信息。

其次，本文为所有的卷积操作增加了系数为 0.1 的 L2 正则化项。添加 L2 正则化可以使模型总体上拥有更小的权重，同时在训练数据上仍然表现良好。这可以防止模型对训练数据的过度拟合，并提高其泛化到新的、未曾见过的数据的能力。此外还在增加卷积通道数的同时，采用最大池化操作减少进入下一层的输入维度，这将有助于减少模型中后续层的计算成本，并通过减少参数数量来防止过拟合。

最后，为了避免在存量过程中存在的梯度消失问题，为编码器增加了残差结构，其结构能够使得梯度传播更快地到达高层，使得高层的权重能被及时更新。

（2）解码器的结构

编码器将输入的时序数据变换成高度抽象的信息，而解码器则是利用高度抽象的信息形成新的数据。解码器与编码器有着对称的结构，其结构如图 3。



S= 步长　K= 卷积核尺寸　D= 膨胀卷积系数

图 3 生成网络解码器结构

与编码器相反，权重层采用与普通卷积相反的操作即解卷积操作，将高度抽象的信息重构成低维信息。并使用上采样层（up sampling layer）来完成与池化层相反的操作，增加数据参数量，为生成提供更多的信息。同时，在解卷积的过程中自动添加 0 值补齐卷积操作，会导致生成数据维度多于原始数据，因此增加了稠密连接层（dense layer）减少数据维度，目的是产生与输入数据具有同维的数据。

3.2 loss 函数的设计

本文使用基于自编码器来生成数据，然而网络的训练离不开损失函数，它是模型的优化目标，模型训练的过程就是在不断调整参数以最小化损失函数的过程，损失函数决定了模型参数的更新方向和大小。因此准确的损失函数设计将有助于模型生成更准确的数据。

为了定义生成数据的准确性，本文考虑了生成数据与原始数据的差异、生成数据平滑程度、生成数据识别结果三方面因素。

（1）生成数据与原始数据都是序列数据，衡量序列数据相似度的方法包括欧式距离、余弦相似度等。余弦相似度通过将序列特征视为高维向量，计算之间的夹角来表示其相

似性的，其最主要的缺陷在于平等地对待向量的所有维度，而不考虑数据的意义，然而时序特征在相对时刻上有特殊含义。为了简化运算，本文采用欧式距离来表示生成数据与原始数据的差异。计算 A 和 B 序列的欧式距离公式为：

$$\text{EuclideanDist}(A,B)=\sqrt{\sum_{i=1}^{m}(A_i-B_i)^2} \tag{2}$$

可以通过每个生成值与对应的实际值之间的差值取平方，对所有数据点的差值的平方取平方根得到，该方法计算了每个时间点的差异，欧式距离的值越小生成数据越相似。

（2）如果不对生成该数据的平滑度进行度量，会导致生成的数据具有较大的波动，会造成生成信号的失真。因此为了避免产生与原始信号差异较大或者完全相同的数据，需控制生成数据的信噪比，本文在模型的损失函数中引入了平滑因素。在计算数据的平滑程度上，本文使用了二阶差分，其计算公式为：

$$\text{DA}=\text{diff}(A) \tag{3}$$

$$\text{Smoothness}(\text{DA})=\sum_{i=1}^{n}\text{diff}\left(\frac{\text{DA}_i-\text{mean}(\text{DA}_i)}{\text{std}(\text{DA}_i)}\right)^2 \tag{4}$$

式中：A 代表原始的序列数据；diff、mean、std 分别表示序列数据的一阶差分、均值、标准差。数据平滑程度是对生成数据中噪声的度量，该值越小生成数据越光滑，越大代表越粗糙。

（3）增加了识别模型的结果作为损失函数的一部分，这使得模型在数据生成过程中能够自动判断是否出现失真。具体而言，识别模型将生成数据作为输入数据，并根据生成数据的特点产生对应的所述类别的概率分布，计算该概率分布和真实标签的分类交叉熵（categorical cross entropy）[10]，即可得到识别误差。分类交叉熵也被称为 Softmax 损失，它是 Softmax 函数和交叉熵的结合，如图 4。该损失越小，表明生成的数据越准确，越大表明数据越失真。

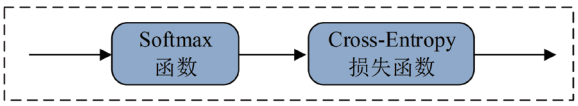


图 4 分类交叉熵计算过程示意图

其中，Softmax 和 CategoricalCrossEntropy 计算公式分别为：

$$\text{Softmax}(t)_i=\frac{e^{t_i}}{\sum_j^C e^{t_j}} \tag{5}$$

$$\text{CategoricalCrossEntropy}=-\sum_i^C g_i\log(\text{Softmax}(t)_i) \tag{6}$$

式中：g_i 和 t_i 分别表示真实和识别模型在每个类上的概率。将该损失作为优化目标之一，保证模型在不降低现有识别精度的情况下生成数据，增加数据多样性。

综上所述，基于 AE 的生成网络的损失函数表示为：

$$\begin{aligned}\text{Total Loss}=&\ \text{CategoricalCrossentropy}(A_T,B_T)\\&+\text{Smoothness}(B)+\text{EuclideanDist}(A,B)\end{aligned} \tag{7}$$

式中：A、B 表示原始特征和生成特征；A_T、B_T 分别表示原始特征和生成特征的真实标签。由上式可知，本文使用的生成网络的输入数据是原始特征和其真实的类别标签，而其输出则是带有分类标签的生成数据。

4 模型训练和数据生成的结果

本文以声音信号感知形变的应用数据为样本评估数据生成的效果，本小节从数据标准化、自定义损失函数、时序影响三个方面介绍生成模型的生成效果。

（1）数据标准化前后对比

在数据生成过程中，数据是否标准化不仅会影响模型的性能，还会对数据生成产生很大影响。不同于归一化将数据归到 0 到 1 之间，标准化是在不改变数据趋势的情况下，使数据点符合均值为 0，方差为 1 的分布，标准化更符合数学的统计特征。图 5 是数据标准化前后生成数据的对比，在使用数据训练生成网络之前，如果没有进行标准化，其生成的数据中间部分的趋势与原始数据很接近，但在首末两端会出现失真现象。与真值相比，其误差在 0.05 到 0.35 之间，有很大波动。而对数据进行标准化后，其生成的数据两侧的失真现象减弱，生成的数据更接近于真实数据。产生这种现象的原因是由于数据本身的分布区间窄，模型生成的细微不同就会形成很大差异，而标准化后的数据区间和网络中的正则化后的表示区间相对应，减少了数据在网络层间传递的损失，因此能够很好的生成数据。

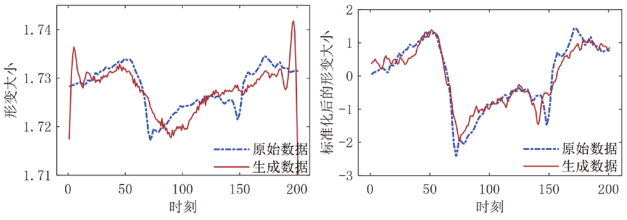


图 5 标准化前后数据生成对比

（2）loss 函数对比

为了比较自定义 loss 函数在生成网络中的作用，在相同的训练参数以及相同的模型结构下，使用不同的损失函数对同一集中的多个类别数据分别进行生成，并使用生成数据的平均分类精度作为比较的标准。对比了对数均方误差（MSLE）、余弦相似度（Cos）、绝对平均误差（MAE）、均方误差（MSE）以及自定义的 loss 函数，其结果如图 6 所示。

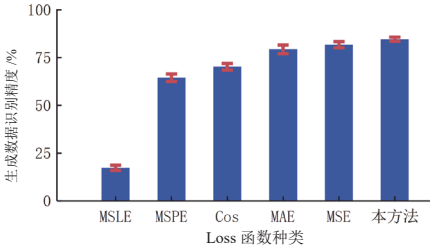


图 6 自定义损失函数与其他方法的对比

从图 6 中可以得到，自定义 loss 函数的效果和 MAE 以及 MSE 相近，但要比余弦相似度要好。总体上自定义 loss 函数识别精度在 85% 左右，比其他 loss 函数效果更好。这说明自定义的 loss 函数下训练的模型生成的数据更准确。

（3）生成的数据对时序序列的影响

本文使用了互相关和相关系数评估生成模型对时序的影响。互相关是常用于信号处理，用于测量两个信号之间的相似度。其思想是一个信号相对于另一个信号在时间上进行平移，并计算平移后两个信号之间的相关系数。因此时间偏移越接近 0，对时序的影响越小。此外 Pearson 相关系数可以衡量两个信号的线性相关性，相关系数越接近 1，两端信号越相关。因此，本文利用互相关的平移量和 Pearson 相关系数来表示生成数据和原始数据在时间上的相似度，其结果如图 7 所示。

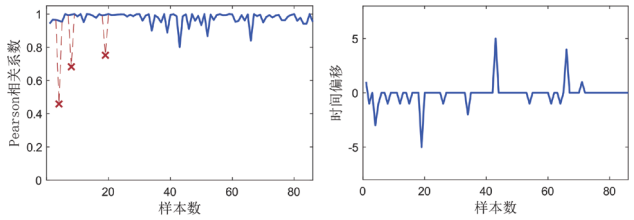


图 7 生成数据与原始数据的相关系数和偏移

从图 7 中可以看到，除了少数样本外，92% 的生成数据和原始数据的平均相关性大于 0.9，这表明生成的数据和原始数据具有较强的相关性。其次从时间平移上看，所有的时间差均在 5 个单位时间以内，考虑到插值的影响，实际上的时间差不超过 0.03 s。综上所述，本文基于 AE 的生成网络能够很好的保留原本数据的时间特点。

5 结语

以上分析可知，基于 AE 的生成网络在自定的损失函数下可以达到准确生成数据的目的，并且生成的数据兼顾了识别准确率和原本时序特点。引入残差结构和扩张卷积这两种机制提高了 CNN 在处理时序数据方面的能力，提高了模型的精度，也加快了模型的训练速度，并且在处理复杂任务时表现得更加出色。最后，还需要说明的是由于在损失函数中引入了识别结果的交叉熵，因此数据的生成会受到识别模型的影响。识别模型的精度越高，其生成的数据越准确。

参考文献：

[1] ALOMAR K, AYSEL H L, CAI X H. Data augmentation in classification and segmentation: a survey and new strategies[J]. Journal of imaging, 2023, 9(2): 46.

[2] GUILLERMO I, EDGAR T, ANGEL G P, et al. Data augmentation techniques in time series domain: a survey and taxonomy[J]. Neural computing & applications, 2023,35(14):10123-10145.

[3] WANG T B, ZHANG D Q, ZHENG Y Q, et al. C-FMCW based contactless respiration detection using acoustic signal[J]. Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies, 2018,1(4):1-20.

[4] GAO Y, WANG W, PHOHA V, et al. EarEcho: using ear canal echo for wearable authentication[C]. Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies. NewYork: ACM, 2019: 1-24.

[5] WANG Z, TAN S, ZHANG L H, et al. An ear canal deformation based user authentication using in-ear wearable devices[C]//Proceedings of the 27th Annual International Conference on Mobile Computing and Networking.NewYork: ACM, 2021: 819-821.

[6] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[DB/OL].(2015-11-23)[2024-06-14]. https://doi.org/10.48550/arXiv.1511.07122.

[7] DIELEMAN S, ZEN H, SIMONYAN K, et al. Wavenet: agenerative model for raw audio[DB/OL].(2016-09-12)[2024-05-22]. https://doi.org/10.48550/arXiv.1609.03499.

[8] 王坤峰，苟超，段艳杰，等．生成式对抗网络 GAN 的研究进展与展望 [J]. 自动化学报，2017, 43(3): 321-332.

[9] 来杰，王晓丹，向前，等．自编码器及其应用综述 [J]. 通信学报，2021, 42(9): 218-230.

[10]ZHANG Z L, SABUNCU M. Generalized cross entropy loss for training deep neural networks with noisy labels[C]. NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems. NewYork: CAI, 2018: 8792-8802.

【作者简介】

邓文文（1998—），男，四川资阳人，硕士，助理工程师，研究方向：共用构建模块管控技术，email: redpanda.brilliant@gmail.com。

徐海洋（1998—），男，陕西西安人，硕士，助理工程师，研究方向：深度学习、神经网络。

申艺（1996—），男，陕西西安人，硕士，研究方向：通用技术。

（收稿日期：2024-09-26）