东北大学自然语言处理作业

基于Transformer的声学事件检测方法研究

院（系）名　称：计算机科学与工程学院

专　业　名　称：计算机科学与技术

学　生　姓　名：孙润泽

2022年12月

摘 要

音频数据具有易采集、信息含量丰富等特点，近年来，随着人工智能的发展，计算机识别领域的声学事件检测方法逐渐被应用到各种场景。然而大部分声学事件检测方法的检测性能会随着任务量的增加而降低，无法对上下文信息既考虑全局又聚焦重点。本文引进注意力因子，改进特征提取网络，提出自注意力机制的声学事件检测模型，提高识别准确率。

针对卷积神经网络（Convolutional Neural Network, CNN）提取高层特征时，无法结合数据内部信息重要程度的缺点，本文引进卷积注意力机制模块（[Convolutional Block Attention Module](https://arxiv.org/abs/1807.06521), CBAM），在模型处理声学数据时关注有用信息，忽略无关信息，实现内部重要性的建模，提取相关特征。在此基础上，基于特征中的关系，使用自注意力机制计算信息的注意力因子，完成对信息之间相关性的整合。同时利用转换编码器（Transformer Encoder）中实现信息的整合、对输入特征做出类别判断的符号分类标记（Class Token, CLS）作为事件预测结果，提出CAM-Transformer Encoder模型完成声学事件检测任务。

通过与CNN、多层感知机（Multilayer Perceptron, MLP）、深度神经网络（Deep Neural Networks, DNN）、残差网络（Residual Network, ResNet）、VGGNet等优秀检测模型的实验结果对比，本文提出的CAM-Transformer Encoder模型在声学数据集上取得了更优的检测性能，验证了自注意力机制确实能够有效提升声学事件的检测性能。

关键词：卷积神经网络；声学事件检测；Transformer

# 基于Transformer的声学事件检测

本章首先构建CAM网络来完成高层抽象声学特征提取，再使用Transformer Encoder模型完成检测功能。CAM网络主要使用与注意力机制结合的CNN模型，加强了CNN在进行特征映射时，对重要信息的敏感性与关注度，忽略无用信息带来的影响。最后使用检测模型做进一步的数据建模，通过多头自注意力机制，寻找输入信息之间的相互性，得到兼容上下文信息的特征向量，并完成最终的检测任务。

## 基于卷积神经网络的特征提取

对于音频数据来说，由于信号源的差异、信道的衰减、噪声的干扰等因素，导致数据的信息不纯。在经过数据预处理后提取出Fbank特征，能够在保证信息最大化保留的同时尽可能的去除噪声。但是FBank特征从某种程度来说，本身还存在着相当大一部分的无用信息，例如汽车声中可能混淆着人声等其他外来因素。本章提出使用CAM网络进行FBank的深层特征提取，而由于CNN在计算机视觉领域取得的巨大成功，在NLP领域也有很好的应用，CAM网络将选取传统CNN作为主线网络。但是在声学事件检测任务中，声音数据中往往包含多类别事件，且每种事件都具有不确定性，无法对其进行分割，导致CNN在处理多类别音频数据特征的时候无法发挥其提取抽象特征并泛化模型的作用。所以，在CNN的基础上结合注意力机制模块，关注重要区域信息，弱化无用信息带来的影响，共同构建完整的CAM网络，完成特征提取操作。



图3.1 卷积注意力机制模块

本章使用的注意力模块为卷积注意力机制模块（[Convolutional Block Attention Module](https://arxiv.org/abs/1807.06521)，CBAM），作为常用的小型信息处理单元，可以直接应用到传统卷积神经网络架构中，与传统卷积神经网络完成模型训练，并且其中网络层数的增多不会增加模型复杂度。CBAM如图3.1所示。

CBAM由空间注意力模块与通道注意力模块串行连接构成，在多个领域的检测任务中，注意力块串行连接的方式取得的效果超过了并行连接[30]。CBAM分别从通道和空间两个维度计算输入向量图的注意力分布图，然后进行图与图逐项相乘完成特征自适应学习。在了解CBAM之前，先对通道注意力与空间注意力进行介绍。

所谓通道注意力，指的是对于不同通道的特征，具有的意义并不等同。通过自适应的方法，调整模型对于各个通道特征的关注程度，实现对于高影响力通道的聚焦，减少无关通道对当前任务的影响。同时为了避免空间维度的影响，使用压缩的方式平化特征，使之变成一维向量。在空间维度上进行压缩时，使用平均值池化（Average Pooling），而不使用最大值池化（Max Pooling）。由平均池化聚合空间信息得到的特性，被共享网络处理，使用元素求和的方法生成一个通道向量，代表通道数。此外，为了减少参数开销，将通道向量压缩至，其中代表压缩比。在全连接层之后，进行数据标准化操作以归一化向量的比例。通道注意力机制的具体表示为：

 (3-1)

式中：，，，。

与通道注意力不同，空间注意力机制关注不同空间位置的特征，对通道注意力模块处理后的特征沿着通道轴进行卷积计算，并将它们连接起来以生成一个有效的特征符号，是对通道注意力的补充。模块结构上，使用多层卷积操作沿着通道轴聚合特征图的信息，使用降维卷积减少参数和计算开销，获取降维后的特性之后，使用归一化完成特征标准化操作。空间注意力机制的表示为：

 (3-2)

式中：代表卷积操作，表示卷积核的大小，代表归一化操作。

特征图依次通过通道注意力与空间注意力模块分别获得通道及空间注意力图和后，由于两个注意力分布图形状不同，需要对其进行维度扩展，在进行计算前将两个模块得出的注意力图的大小都调整为，再采用逐项求和的方式进行合并,生成中间注意力图。再使用Sigmoid函数对中间注意力图进行取值范围为的区间映射，得到最终的注意力图，公式如(3-3)所示。最后将注意力图和特征图进行乘积，分配各个特征的权重，对重要部分进行关注，忽略无用部分，最后与特征图相加，得到新的特征图为后续分类工作提供输入，公式见(3-4)。

 (3-3)

 (3-4)

式中：表示Sigmoid函数，表示逐元素相乘操作。

CNN模型在许多领域表现良好，不仅解决了许多全神经连接网络无法解决的问题，如处理特征时丢失空间信息、参数过多时训练困难等，还拥有着解决过拟合的Dropout技术。与后面衍生出的众多CNN模型相比，传统CNN因为复杂度低、模型收敛速度快等优势保持着自身的地位。例如，VGGNet模型通过使用较小的卷积核，可以加深网络层次，通常能够达到十层，虽然有效提升了模型效果，但是网络的复杂度和计算量大大增加，导致训练困难性也会相应增加。AlexNet模型中存在局部归一化的问题，容易陷入局部最优的陷阱，造成计算资源的浪费。

为了使模型拥有最优效果，CNN模型使用的卷积核大小为。采用或等更大的卷积核容易造成网络参数过多，模型收敛速度下降等问题，还有提取到的特征可能信息量不足。从感受野的角度来看，使用多个尺寸较小卷积核的感受野和一个尺寸较大卷积核的感受野是相同的，因为感受野总是与输入的深度相符。同时采用多个小尺寸卷积核的网络模型，可以在层与层之间加入ReLU等激活函数使模型避免过拟合问题，输出的结果更优，计算量相对来说也不会多。多层非线性层能够实现拓深网络的作用，由此能够增强模型的泛化能力。因此CAM网络采用传统CNN模型结合所提出的通道及空间注意力机制来完成高层特征的映射及提取。

针对声学事件检测任务，设计包含二维卷积、池化等一系列操作的三层CNN来完成音频数据中特征的提取的检测，每层CNN除输入输出维度即卷积核数量递增外，其余结构相同。具体实现为使用1个卷积核大小为的步长和padding都为1的卷积操作，每个卷积操作结束后经过批量归一化处理与激活函数，使之拥有更加丰富的函数形式。再通过池化完成对特征信息的局部总结，方便对重要特征的关注。之后每个池化操作后面的Dropout层用以减少网络对神经元的依赖性，避免过拟合，并提升网络的鲁棒性。特征经过三层CNN传入两次全连接层，第一次全连接层任务是将高维的特征空间映射到隐藏空间，主要用于检测模型的输入，另一层需要将其进行类别映射，对输出进行可视化操作。最后将中间分类结果放入Sigmoid函数，完成到取值范围的概率映射，得到各个类别的预测概率，用以计算模型损失。3层CNN的卷积核数量分别为32、64、128，下图为3层CNN模型的结构。

对CNN添加通道及空间注意力机制，以增大模型对数据中声学事件特征相关信息的关注程度并为其分配较高注意力权重，弥补CNN在进行低维空间映射时，造成的信息丢失问题。完成对输入数据的特征提取的同时，提高对声学事件特征重要部分的敏感性，忽略声学特征中无用的信息。具体的添加方法是在每层经过卷积、激活、池化、Dropout操作的CNN后，加入CBAM模块。模块中输入的大小为每层卷积核的数量。经过实验测试，发现通道及空间注意力模块被添加的位置对结果来说是有一定影响的。网络层数越多，模型越复杂，注意力机制产生的效果越差。实验数据表明，CBAM添加在CNN的卷积层或者池化层后面，并未取得好的效果。传统CNN与CNN+CBAM构建的CAM网络的具体结构如图3.2、3.3所示。



图3.2 传统CNN结构图

在使用CAM网络进行特征提取前，首先对原始音频进行预处理，提取声学特征FBank。同时需要对FBank进行归一化处理和分段操作，按照音频数据中的文本进行分别标注，作为CAM网络的输入。具体的方法为，使用给定的文本标注信息对应其中各个数据的开始时间和结束时间，依次标注对应声学事件的种类。将数据与标签输入到模型中，模型采用二元交叉熵作为损失函数，Adam优化器完成训练工作。其中二元交叉熵函数常常会被用于计算二分类问题的loss，其本身属于凸优化问题。好处是在使用梯度下降方法求解时，凸优化具有良好的收敛性，不仅可以帮助模型加速训练，也可以避免模型过拟合问题。模型训练完毕后，下载已经训练好的模型参数进行特征提取与保存，以供后续检测模型的输入使用。二元交叉熵函数公式见式(3-5)，其中，为输入大小，为标签，代表输出结果为的概率。

 (3-5)



图3.3 CAM网络结构图

## 基于Transformer的检测模型

第二章介绍了基于自注意力机制的Transformer模型的相关结构与数学原理，成功应用在NLP领域与图像识别领域。对于多分类的声学事件检测任务而言，声学特征输入的维度与NLP领域相似，都是二维的向量。其中第一维度代表输入的批量，第二维度为一个声学特征的长度。NLP任务中，词向量提及到了字典一说，通过输入项与字典中的单词进行对应标号，并在隐藏空间中拓展了输入项的维度，即得到了每个输入的嵌入表示，此嵌入表示整句的上下文。但是声学事件无法拥有类似NLP任务的字典，也就说明不能训练出Embedding隐藏空间。本文选择直接将特征输入到检测模型，但是还需要做一些预处理。CLS标记作为经过上下文左右信息的整合，对输入内容做出含义判断的重要符号，需要被添加在特征的开头，与输入特征并列，作为最后进行声学事件检测任务分类的依据。大部分使用的Transformer Encoder模型都会添加位置信息，但是声学事件检测任务从某种程度上来说，属于对数据帧的检测，由于特征维度并不大，位置信息的添加会在一定程度上加大了模型训练难度。上一章提及到相对位置对于声学事件检测模型来说效果并不显著，而可训练的绝对位置信息，虽然有希望强化信息之间的相关性，增强模型对不同位置信息之间的关注，使网络最终的输出向量表达能力更强，但是更大可能会造成模型参数大幅度的增加，使模型无法收敛，最终影响整个实验。所以本节提出的Transformer Encoder检测模型放弃了添加位置信息。同时为了使模型更好的应用在分类任务中，本文使用MLP模块代替了前馈网络，并且为了保持前馈网络中的ReLU激活函数非线性变换的作用，在MLP模块中添加高斯误差线性单元激活函数GELU。GELU早已成功应用在Bert模型上，相比于ReLU激活函数，其不仅保留了不重要信息被归0的概率性，同时也保留了对输入的依赖性。特征在经过多个重叠的Encoder Block形成的Transformer Encoder处理后，需要进行层标准化与CLS符号提取，然后通过全连接层与Softmax函数，输出为声学检测任务的类别概率。基于Transformer的声学事件检测模型结构如图3.4。



图3.4 基于Transformer的检测模型结构图

CAM网络提取到的特征作为输入，在经过CLS记号拼接后，使用Dropout进行信息拟合。然后放入Encoder Block中，其中的层标准化是针对NLP领域提出的，类似于RNN循环神经网络，解决了输入的每项重新调整在合适的范围，避免了梯度消失、梯度爆炸问题，使隐藏状态的传递更加稳定。经过线性变换的多头自注意力机制，可以得到注意力矩阵，其中包含着特征内部信息的相关性，使用Dropout方法随机使部分神经元不产生作用，减少自注意力机制对某些神经元的依赖，进而防止过拟合。到目前为止，所有的操作都是线性变换，如果神经网络只使用线性变换是毫无意义的，无法凭空产生新的信息，而随后使用的MLP Block主要用来解决这个问题。MLP Block通过全连接层、GELU激活函数、Dropout等方法使被处理后的输入拥有更强的表达能力。最终多层的Encoder Block在不断的参数共享与迭代中，整合特征中重要信息，并作为最终的特征向量放进CLS符号中，提取出CLS符号并放进分类器，获取最终声学事件检测任务的分类结果。由于使用作为头部的CLS符号作为分类输入，并使用MLP分类方法，分类器叫做MLP Head。最终的CAM-Transformer Encoder模型如图3.5所示。



图3.5 CAM-Transformer Encoder模型

由于任务属性为多分类问题，在分类阶段，采用阈值机制来判断任务事件的发生。主要实现方式为，设置一个取值范围为的阈值，对于类别概率超过此阈值的声学事件，判断其发生，将其标注为1，否则为0。阈值大小会对检测结果产生影响，太大会造成声学事件致盲性，太小会造成声学事件溢出，在评价指标中，两种情况都会导致错误率ER过大、F1过小，最终对实验产生影响。

# 实验结果与分析

## 声学数据集

TUT声学场景2017数据集由15小时的双耳音频组成，音频数据由各个场景的记录组成，拥有不同的记录位置。场景涵盖车辆、室内外多个方面。其中室内记录由餐厅、杂货店、图书馆与地铁站等组成，室外包括森林小径、住宅区、湖畔沙滩，车辆音频数据来源于公共汽车、汽车驾驶、火车与电车。音频录取采用44.1kHz的采样率和24位分辨率的收音机，使采集到的声音尽量靠近人体听觉系统。

本文使用的数据集为TUT声学场景2017数据集的子集，TUT Sound Events 2017数据集。子集的开发数据集由室外场景的24个录音组成，总共有92分钟的音频数据，其中包含了6类声学事件。分别为刹车声（brakes squeaking）、汽车（car）、孩子（children）、大型车辆（large vehicle）、人声（people speaking）、人走路的声音（people walking）。表4.1是不同声学事件在数据集中的案例分布。

表4.1 声学事件在数据集中的分布

|  |  |  |
| --- | --- | --- |
| 声学事件标签 | 验证集 | 未验证 |
| 刹车声 | 52 | 59 |
| 汽车 | 304 | 304 |
| 孩子 | 44 | 58 |
| 大型车辆 | 61 | 61 |
| 人说话 | 89 | 117 |
| 走路 | 109 | 130 |
| 总共 | 659 | 729 |

根据每个声学事件类的数量，将数据划分为开发数据集与评估数据集。同时划分时还需要考虑记录位置。理想情况下，两个数据集对于每个声学事件类具有相同的相对量，但是由于不同类的事件在记录中分布不均，大部分的事件都在开发数据集中。与此同时，为了使数据集拥有统一性，提供交叉验证设置，共分为四个部分。

数据集的标注文本中包含不同录音文件中每个声学事件的开始与结束时间，以及对应的类别信息。

同时，由于数据集中含有空白部分，也就是无标签数据，会大大降低实验效果。需要在进行音频数据预处理时剔除，也可以在提取声学特征后，对特征内容进行标注时将无标注的特征删除。

## 不同阈值对比实验结果与分析

上一章在基于Transformer Encoder的声学事件检测模型部分提及过阈值的重要性，取值范围在的阈值是在计算模型损失时的关键参数。在模型分类阶段，由于检测模型最后的可视化输出为声学事件案例对应在各个类别中的预测概率，并且声学事件检测为多分类属性，不能单独将此案例划分为预测结果中最大概率的类别。采用阈值机制，将预测概率超过阈值的类别判断为发生，对其标注为1，否则为0。不同的阈值会对结果产生影响，过大时容易造成声学事件的漏判，过小又会造成未发生的声学事件判定为发生，最终会造成模型损失与评价体系的负面效果。通过调整阈值的大小，查看CNN-Trasnformer Encoder模型在开发数据集与评估数据集上的评价体系优劣，来确定最优的阈值设置。没有使用本文提出的模型的原因是避免模型不具备鲁棒性，声学事件检测模型在评估数据集上最终结果如下。

通过实验结果表4.2和图4.1比较，可以看出随着阈值的增大，ER逐渐降低，F1逐渐升高，说明检测性能正逐步增强，并在阈值等于0.8时达到最高点，此时ER为0.496，F1为0.650，模型检测效果最优。随后检测性能急剧降低，最终模型效果达到最差，ER为1，F1为0，说明此时模型不再工作。因此后续实验均使用0.8的阈值作为模型参数。

表4.2 不同阈值下的检测结果

|  |  |  |
| --- | --- | --- |
| 阈值 | 评估数据集 | |
| ER | F1 |
| 0.1 | 2.439 | 0.444 |
| 0.2 | 2.033 | 0.480 |
| 0.3 | 1.566 | 0.527 |
| 0.4 | 1.308 | 0.536 |
| 0.5 | 1.090 | 0.551 |
| 0.6 | 0.841 | 0.584 |
| 0.7 | 0.672 | 0.612 |
| 0.8 | 0.496 | 0.650 |
| 0.9 | 0.780 | 0.344 |
| 1.0 | 1.000 | 0 |



图4.1 不同阈值下评价指标ER与F1对比

## Transformer模型对比实验结果与分析

基于Transformer的检测模型由多个Encoder Block堆叠在一起，再与MLP Head分类器共同组成检测模型。其中检测模型的输入是否添加CLS标记，会直接影响多头自注意力机制的效果。伴随着CLS符号的增加，神经网络系数也会相应增长，模型的表达能力也会随之加强，但同时也有可能会出现网络的收敛速度慢、梯度消失等问题。本节通过使用0.8阈值并已经预训练好的CAM-Transformer Encoder实验模型完成声学事件检测任务，除添加CLS符号外，每一个模型的相关超参数如学习率等均保持一致。

表3.3 是否添加CLS符号的检测结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 训练次数 | 是否添加CLS符号 | 评估数据集 | | |
| 损失 | ER | F1 |
| 2 | 是 | 0.368 | 0.413 | 0.674 |
| 否 | 0.395 | 0.481 | 0.655 |
| 4 | 是 | 0.369 | 0.411 | 0.676 |
| 否 | 0.390 | 0.475 | 0.661 |
| 6 | 是 | 0.368 | 0.409 | 0.677 |
| 否 | 0.390 | 0.473 | 0.662 |
| 8 | 是 | 0.368 | 0.408 | 0.678 |
| 否 | 0.390 | 0.473 | 0.662 |
| 10 | 是 | 0.369 | 0.412 | 0.674 |
| 否 | 0.393 | 0.489 | 0.663 |



图4.2 有无CLS符号的检测结果对比

对表3.3与图4.2中实验结果进行分析，在对特征添加CLS标记以后进入模型训练，虽然模型训练速度略微下降，但是在评估数据集上的损失与评价指标均有提升。其中，F1代表的F-Score提升了1.6％，ER代表的三种错误下降了7％左右，证明模型添加CLS标记后拥有更好的鲁棒性，同时能对整个输入起到关注全局的作用。

## 不同检测方法对比实验结果与分析

为了验证本文提出的自注意力机制的有效性，我们对CAM-Transformer Encoder检测模型与一系列检测模型，其中如传统检测方法中的MLP检测模型，深度学习检测方法中的CNN检测模型、ResNet检测模型、VGGNet检测模型进行对比。同时，为了证明CAM网络添加的注意力机制发挥了作用，还与未添加注意力机制的CNN-Transformer Encoder的检测模型进行对比。实验过程中由于模型参数数量存在巨大差异，为使实验正常运行，ResNet34与ResNet50检测模型与其他检测模型的批处理大小不同。ResNet34检测模型的批处理大小为128，ResNet50检测模型的批处理大小为64，其余检测模型的批处理大小均为264.模型参数设置阈值选用0.8，训练次数为100，其中基于Transformer Encoder的检测模型均添加CLS符号。通过对比，查看任务模型参数数量以及模型在评估数据集上的效果，实验结果如下图所示。

表4.4 不同检测方法下的检测结果

|  |  |  |  |
| --- | --- | --- | --- |
| 检测方法 | 参数数量 | 评估数据集 | |
| ER | F1 |
| MLP | 125K | 0.676 | 0.565 |
| DNN | 290K | 0.642 | 0.623 |
| CNN | 390K | 0.609 | 0.611 |
| ResNet18 | 8.8M | 0.512 | 0.608 |
| ResNet34 | 17.85M | 0.548 | 0.577 |
| ResNet50 | 17.26M | 0.778 | 0.522 |
| VGGNet | 4.73M | 0.517 | 0.603 |
| CNN-Transformer Encoder | 9.47M | 0.496 | 0.650 |
| CAM- Transformer Encoder | 9.55M | 0.407 | 0.678 |



图4.3 不同检测模型评价指标对比

表4.4与图4.3中给出了本文提出的模型与一系列模型的对比数据。从表中看出，ResNet50检测模型性能最差，可能是批处理大小不一样所导致，ER为0.778，F1为0.522。然后是MLP检测模型，也反映了深度学习检测模型在声学事件检测任务上的性能是超过传统检测方法的，MLP在评估数据集上ER为0676，F1为0.565。基于CNN的检测模型ER是0.609，F1是0.611，而基于DNN的检测模型评价指标ER为0.642，F1为0.623，综合来看，CNN在声学事件检测任务中略优于DNN，证明卷积对特征的高层提取是有帮助的。在所有基于深度学习方法构建的检测模型中，ResNet检测模型检测性能最优，ER指标达到了0.512，F1指标达到了0.608。此外，基于Transformer的检测模型，在声学事件检测任务中性能都有一定提高。相比于CNN检测模型，CNN-Transformer Encoder检测模型在ER与F1上均有所提升，ER指标为0.496，F1指标为0.650。说明了自注意力机制中关注上下文信息建模的方法确实对分类任务有作用。最后，在CNN基础上结合通道及空间注意力形成的CAM特征提取网络，搭配Transformer Encoder检测模型，在参数数量上有些许增加，但是检测性能也是获得了良好的提升，ER指标为0.407，F1指标为0.678。证明了注意力机制也能帮助模型更好的进行特征提取工作。

# 结 论

通过对声学事件检测方法与自注意力机制的研究现状进行分析，本文在自注意力机制的基础上，使用注意力与深度学习方法相结合提取特征，对上下文信息建模，完成声学事件检测任务。提出了CAM-Transformer检测模型，通过各种测试以及与MLP、DNN、CNN、ResNet、VGGNet等模型对比得到结果并分析了其具有的优势，验证在声学事件检测领域具有良好的性能表现。本文主要工作内容如下。

（1）将CNN具有的提取高层特征的优点与注意力机制结合组建CAM网络，完成对重要信息的提取。通过对输入的声学特征进行处理，关注特征中重要数据，忽略无关部分，生成更加具有代表性的高维特征作为后续检测模型的输入。通过模型对比，提出的CAM网络在特征提取网络中相对拥有更少的参数、更优的性能，说明注意力机制能够有效提升模型检测性能。。

（2）针对使用多头自注意力机制的Transformer Encoder模型对于输入的不合理要求，经过处理的声学特征无法拥有文本输入的词空间，也无法像图像数据一样保持自身的特征方阵。本文选择直接在特征上添加分类标记CLS，并且选择抛弃使模型难以收敛的位置信息与前馈网络，换成了拥有GELU激活函数的MLP模块。最后将经过上下文左右信息的整合，对输入内容做出含义判断的CLS符号作为最终的分类依据。通过实验与分析，验证特征输入添加CLS标记后，检测模型在性能上得到提升。

（3）通过对模型参数阈值的对比，选择对本文提出的模型以及对比模型都相对友好的阈值作为实验参数。经过与不同检测模型对比，实验结果表明，本文提出的基于自注意力机制的声学事件检测模型在数据集上的性能有一定优势。

但是模型仍有待改进，经过注意力模块提取的特征，在一定程度上有了信息集成，想要通过此特征在基于自注意力机制的检测模型上有所突破，单独训练是比较困难的。由于特征的信息密度增加，模型训练起来更加困难，过拟合的现象出现的也比较频繁，想要突破这种限制，可能需要在训练方式上做出调整。本文使用的方法为，将训练好的CNN-Transformer Encoder检测模型参数应用在CAM-Transformer Encoder模型上，这样做虽然性能有所提升，但是有几率使模型陷入局部最优。由于精力有限，始终未能找到更好的方法解决此问题。