论文译文 ECA-Net- Efficient Channel Attention for Deep Convolutional Neural Networks

论文译文 ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks

Abstract

近年来，通道注意力机制在提高深度卷积神经网络(CNN)的性能方面显示出巨大的潜力。然而，现有的大多数方法致力于开发更复杂的注意力模块以获得更好的性能，这不可避免地增加了模型复杂度。为了克服性能和复杂度权衡的悖论，本文提出了一种高效的信道注意力(ECA)模块，该模块在带来明显性能增益的同时只涉及少数参数。通过剖析 SENet 中的通道注意力模块，我们通过经验表明，避免降维对于学习通道注意力非常重要，而适当的跨通道交互可以在保持性能的同时显著降低模型的复杂性。因此，我们提出了一种无需降维的局部跨信道交互策略，可通过一维卷积有效实现。此外，我们还开发了一种自适应选择一维卷积核大小的方法，以确定局部跨信道交互的覆盖范围。例如，与 ResNet50 骨干网相比，我们的模块的参数和计算量分别为 80 vs. 24.37M 和 4.7e-4 GFLOPs vs. 3.86 GFLOPs，在 Top-1 准确率方面的性能提升超过 2%。我们利用 ResNets 和 MobileNetV2 骨干网对 ECA 模块的图像分类、对象检测和实例分割进行了广泛评估。实验结果表明，我们的模块效率更高，同时性能也优于同类模块。

1 Introduction

深度卷积神经网络(CNNs)在计算机视觉领域得到了广泛的应用，在图像分类、目标检测和语义分割等广泛的任务中取得了很大的进展。从开创性的AlexNet [17]开始，为了进一步提升深度卷积神经网络[29、30、11、15、19、20、32]的性能，不断有研究进行探索。深度卷积神经网络(CNNs)在计算机视觉领域得到了广泛的应用，在图像分类、目标检测和语义分割等任务中取得了很大的进展。从开创性的AlexNet [17]开始，为了

进一步提升深度卷积神经网络[29、30、11、15、19、20、32]的性能，不断有研究进行探索。最近，将通道注意力融入卷积块引起了很多研究兴趣，在性能提升[14、33、13、4、9、18、7]方面表现出巨大潜力。其中一种代表性的方法是压缩激励网络(SENet) [14]，它为每个卷积块学习通道注意力，为各种深度CNN架构带来明显的性能增益。

根据 SENet [14] 中的挤压（即特征聚合）和激励（即特征重新校准）设置，一些研究通过捕捉更复杂的信道依赖关系 [33, 4, 9, 7] 或结合额外的空间注意力 [33, 13, 7] 来改进 SE 区块。这些方法虽然获得了更高的精度，但往往带来了更高的模型复杂度和更沉重的计算负担。与上述以较高的模型复杂度为代价获得较好性能的方法不同，本文将重点放在一个问题上：能否以更高效的方式学习有效的通道注意力？

为了回答这个问题，我们首先重访SENet中的通道注意力模块。具体来说，给定输入特征，SE块首先对每个通道进行独立的全局平均池化，然后使用两个具有非线性的全连接(FC)层，然后使用Sigmoid函数生成通道权重。这两个 FC 层旨在捕捉非线性跨信道交互作用，其中涉及降维以控制模型的复杂性。虽然这一策略在后续的通道注意力模块中被广泛使用[33, 13, 9]，但我们的实证研究表明，降维会对通道注意力预测产生副作用，而且捕捉所有通道的依赖关系既低效又没有必要。

因此，本文提出了深度 CNN 的高效通道关注（ECA）模块，它避免了降维，并能有效捕捉跨通道交互。如图 2 所示，在不降维的情况下进行通道全局平均池化后，我们的 ECA 通过考虑每个通道及其 k 个邻近通道来捕捉局部跨通道交互。这种方法被证明可以同时保证效率和效果。注意，我们的ECA可以通过大小为k的快速一维卷积有效实现，其中核大小k表示局部跨通道交互的覆盖率，即一个通道有多少邻居参与注意力预测。为了避免通过交叉验证手动调整 k，我们开发了一种自适应确定 k 的方法，其中交互的覆盖范围（即内核大小 k）与信道维度成正比。如图 1 和表 3 所示，与骨干模型[11]相比，带有我们的 ECA 模块（称为 ECA-Net）的深度 CNN 只需引入很少的额外参数和可忽略不计的计算量，就能带来显著的性能提升。例如，对于拥有 24.37M 个参数和 3.86 GFLOPs 的 ResNet-50，ECA-Net50 的额外参数和计算量分别为 80 和 4.7e4 GFLOPs；同时，ECA-Net50 的 Top-1 准确率比 ResNet-50 高 2.28%。

表 1 总结了现有的注意力模块在通道降维（DR）、跨通道交互和轻量级模型方面的情况，我们可以看到，我们的 ECA 模块通过避免通道降维来学习有效的通道注意力，同时以极其轻量级的方式捕捉跨通道交互。为了评估我们的方法，我们使用不同的深度 CNN 架构在 ImageNet-1K [6] 和 MS COCO [23] 的各种任务中进行了实验。

本文的贡献概述如下。(1) 我们剖析了 SE 模块，并通过实证证明了避免降维和适当的跨通道交互对于学习有效和高效通道注意分别非常重要。(2) 基于上述分析，我们尝试为深度 CNN 开发一种极其轻量级的通道注意力模块，提出了一种高效通道注意力（ECA），它几乎不增加模型复杂度，却能带来明显的改进。(3) 在 ImageNet-1K 和 MS COCO 上的实验结果表明，我们的方法在获得极具竞争力的性能的同时，模型复杂度也低于同行。

2 Related Work

.....

3 Proposed Method

在本节中，我们首先回顾了SENet [14] (即, SE块)中的通道注意力模块。然后，我们通过分析降维和跨通道交互的影响对SE块进行了实证诊断。这促使我们提出我们的ECA模块。此外，我们开发了一种自适应确定ECA参数的方法，最后展示了如何将其用于深度CNN。

1. Revisiting Channel Attention in SE Block

设一个卷积块的输出为X∈RW × H × C，其中W、H、C分别为宽度、高度和通道维数(即滤波器的个数)。相应地，SE块中信道的权重可以计算为

其中g (X) = 1 WH∑W，H i = 1，j = 1 Xij为通道级全局平均池化(GAP)，σ为Sigmoid函数。设y = g (X)，f { W1，W2 }取形式

其中ReLU表示整流线性单元[25]。为避免过高的模型复杂度，W1和W2的大小分别设置为C × (Cr)和(Cr) × C。可以看出，f { W1，W2 }涉及到通道注意力块的所有参数。而Eq . (2)可以降低模型复杂度，破坏信道与其权重的直接对应关系。例如，单个FC层使用所有通道的线性组合来预测

每个通道的权重。但Eq .(2)首先将通道特征投影到低维空间，然后将其映射回来，使得通道与其权重之间的对应是间接的。

2. Efficient Channel Attention (ECA) Module

在重新审视 SE 块之后，我们进行了实证比较，以分析通道降维和跨通道交互对通道注意力学习的影响。根据这些分析，我们提出了高效通道注意力（ECA）模块。

2.1. Avoiding Dimensionality Reduction

如上所述，公式（2）中的降维使得通道与其权重之间的对应关系是间接的。为了验证降维的效果，我们比较了原始 SE 块和它的三个变体（即 SE-Var1、SE-Var2 和 SEVar3），所有变体都没有进行降维。如表 2 所示，不含任何参数的 SE-Var1 仍然优于原始网络，这表明通道注意力有能力提高深度 CNN 的性能。同时，SE-Var2 可以独立学习每个通道的权重，这比 SE block 略胜一筹，但涉及的参数较少。这可能表明，信道及其权重需要直接对应，而避免降低二维性比考虑非线性信道依赖性更重要。此外，采用单个 FC 层的 SEVar3 在 SE 块中的降维效果优于采用两个 FC 层的 SEVar3。上述结果都清楚地表明，避免降维有助于学习有效的信道注意力。因此，我们开发的 ECA 模块不需要通道降维。

2.2. Local Cross-Channel Interaction

在不进行降维的情况下，给定聚合特征 y∈ RC，可以通过以下方法学习信道注意力

其中 W 是一个 C × C 参数矩阵。特别是，对于 SE-Var2 和 SE-Var3，我们有

其中，SE-Var2 的 Wvar2 是一个对角矩阵，涉及 C 个参数；SE-Var3 的 Wvar3 是一个全矩阵，涉及 C×C 个参数。如式（4）所示，主要区别在于 SE-Var3 考虑了跨信道交互，而 SEVar2 没有，因此 SE-Var3 取得了更好的性能。这一结果表明，跨信道交互有利于学习信道注意力。然而，SEVar3 需要大量参数，导致模型复杂度较高，尤其是在通道数量较多的情况下。

SE-Var2 和 SE-Var3 之间的一个折中方案是将 Wvar2 扩展为一个对角矩阵，即

其中公式 (5) 将信道分为 G 组，每组包括 C/G 信道，并在每组中独立学习信道注意力，从而以局部方式捕捉跨信道交互。因此，它涉及 C2/G 参数。从卷积的角度来看，SE-Var2、SEVar3 和公式（5）可分别视为深度可分离卷积、FC 层和组卷积。这里，带有群卷积（SE-GC）的 SE 块用 σ(GCG(y)) = σ(WGy)表示。然而，如 [24] 所示，过多的分组卷积会增加内存访问成本，从而降低计算效率。与 SE-Var2 相比，SE-GC 没有带来任何增益，这表明它不是捕捉局部跨信道交互作用的有效方案。原因可能是 SE-GC 完全摒弃了不同组之间的依赖关系。

在本文中，我们探索了另一种捕捉本地跨渠道互动的方法，旨在保证效率和效果。具体来说，我们采用带状矩阵 Wk 来学习信道注意力，Wk 有 显然，公式（6）中的 Wk 涉及 k × C 个参数，通常比公式（5）中的参数少。此外，公式 (6) 避免了公式 (5) 中不同组之间的完全独立性。如表 2 所示，公式（6）中的方法（即 ECA-NS）优于公式（5）中的 SE-GC。在公式（6）中，计算 yi 的权重时只考虑 yi 与其 k 个邻居之间的交互作用，即

其中，Ωk i 表示 yi 的 k 个相邻信道集合。

更有效的方法是让所有通道共享相同的学习参数，即

需要注意的是，这种策略可以通过快速的一维卷积来轻松实现，其内核大小为 k，即

其中 C1D 表示一维卷积。这里，式 (9) 中的方法称为高效信道关注（ECA）模块，它只涉及 k 个参数。如表 2 所示，我们的 ECA 模块在 k = 3 时取得了与 SE-var3 相似的结果，而模型复杂度却低得多，通过适当捕捉局部跨信道交互，保证了效率和效果。

2.3. Coverage of Local Cross-Channel Interaction

由于我们的 ECA 模块（9）旨在适当捕捉局部跨信道交互，因此需要确定交互的覆盖范围（即一维卷积的核大小 k）。对于不同 CNN 架构中具有不同通道数的卷积块，可以手动调整交互的优化覆盖范围。不过，通过交叉

验证进行手动调整将耗费大量计算资源。组卷积已被成功用于改进 CNN 架构 [37, 34, 16]，其中高维（低维）信道涉及固定组数的长程（短程）卷积。同理，相互作用(即1D卷积的核尺寸k)的覆盖率与通道维数C成正比是合理的。换句话说，k和C之间可能存在一个映射φ：

最简单的映射是一个线性函数，即φ（k）= γ * k - b。然而，以线性函数表征的关系过于局限。另一方面，众所周知，通道维度C (即滤波器的个数)通常设置为2。因此，我们通过将线性函数φ（k）= γ * k - b扩展为非线性函数来引入一个可能的解，即

然后，给定信道维数C，核尺寸k可以自适应地确定

式中：| t | odd表示t的最近奇数。在本文中，我们将γ和b分别设置为2和1。显然，通过映射ψ，高维通道具有更长的距离相互作用，而低维通道通过非线性映射具有更短的距离相互作用。

3. ECA Module for Deep CNNs

图2展示了我们的ECA模块的概况。ECA模块在不降维的情况下使用GAP聚合卷积特征后，首先自适应地确定核大小k，然后执行1D卷积和Sigmoid函数学习通道注意力。为了将我们的ECA应用到深度卷积神经网络中，我们使用与[14]相同的ECA模块替换SE模块。得到的网络命名为ECA - Net。图3给出了我们的ECA的PyTorch代码。

4 Experiments

在这一部分中，我们使用ImageNet [6]和MS COCO [23]分别在大规模图像分类、目标检测和实例分割上对所提出的方法进行了评估。具体来说，我们首先评估了内核大小对我们的ECA模块的影响，并与ImageNet上最先进的同行进行了比较。然后，我们使用Faster R- CNN [26]、Mask R- CNN [10]和Retina Net [22]验证了我们的ECA - Net在MS COCO上的有效性。

1. Implementation Details

为了评估我们的ECA - Net对ImageNet的分类效果，我们使用了4个广泛使用的CNN作为骨干模型，包括ResNet - 50 [11]、ResNet - 101 [11

]、ResNet - 512 [11]和MobileNetV2 [28]。为了使用我们的ECA训练深度残差网络，我们在[11、14]中采用了完全相同的数据增强和超参数设置。具体来说，输入图像通过随机水平翻转随机裁剪到224 × 224。采用随机梯度下降法(SGD)对网络参数进行优化，权重衰减为1e - 4，动量项为0.9，小批量为256。通过将初始学习率设置为0.1，每隔30个周期减少10倍的学习率，所有模型在100个周期内进行训练。为了使用我们的ECA训练MobileNetV2，我们遵循[28]中的设置，其中网络在400个周期内使用SGD进行训练，权重衰减为4e - 5，动量为0.9，最小批大小为96。初始学习率被设置为0.045，并以0.98的线性衰减率下降。为了在验证集上进行测试，首先将输入图像的较短边尺寸调整为256，并使用224 × 224的中心裁剪进行评估。所有模型均由PyTorch工具包1实现。

我们进一步使用Faster R- CNN [26]、Mask R- CNN [10]和Retina Net [22]在MS COCO上评估我们的方法，其中Res Net - 50和Res Net - 101以及FPN [21]被用作主干模型。我们使用MMDetect工具包[3]实现所有检测器，并使用默认设置。具体来说，将输入图像的短边调整为800，然后使用SGD对所有模型进行优化，权重衰减为1e - 4，动量为0.9，小批量为8 (4个GPU ,每个GPU有2幅图像)。学习率初始化为0.01，分别在8和11个历元后下降10倍。我们在COCO的train2017上训练了12个历元内的所有探测器，并在val2017上报告了结果以供对比。所有程序都在一台装有4块RTX 2080Ti GPU和Intel (R)至强4112 CPU @ 2.60 GHz的PC机上运行。

2. Image Classification on ImageNet-1K

在这里，我们首先评估了内核大小对我们的ECA模块的影响，验证了我们自适应确定内核大小的方法的有效性，然后我们使用ResNet - 50、ResNet - 101、ResNet - 152和MobileNetV2与最先进的同行和CNN模型进行了比较。

2.1. Effect of Kernel Size (k) on ECA Module

如公式所示。(9)式中，我们的ECA模块涉及一个参数k，即一维卷积的核大小。在这一部分中，我们评估了其对我们的ECA模块的影响，并验证了我们的内核大小自适应选择方法的有效性。为此，我们使用ResNet - 50和

ResNet - 101作为主干模型，并使用我们的ECA模块进行训练，设置k从3到9。结果如图4所示，从中我们有以下观察。

首先，当所有卷积块中k固定时，ECA模块在ResNet - 50和ResNet - 101中分别在k = 9和k = 5时获得最佳结果。由于ResNet101拥有更多的中间层，主导了ResNet - 101的性能，因此可能更倾向于使用较小的内核尺寸。此外，这些结果表明，不同的深度CNN具有不同的最优k，并且k对ECA - Net的性能有明显的影响。此外，ResNet - 101的精度波动(0 . 5 %)大于ResNet - 50的精度波动(0 . 15 %)，我们猜测原因是深层网络对固定核尺寸的敏感度高于浅层网络。此外，由式(1)自适应确定的核大小。(12)通常优于固定参数，但它可以通过交叉验证来避免参数k的手动调整。以上结果证明了我们的自适应核大小选择在获得更好和稳定的结果方面的有效性。最后，不同数量k的ECA模块一致优于SE block，验证了避免降维和局部跨通道交互对学习通道注意力有积极影响。

2.2. Comparisons Using Different Deep CNNs

Res Net - 50我们在Image Net上使用Res Net - 50将ECA模块与几种最先进的注意力方法进行了比较，包括SENet [14]、CBAM [33]、A2 - Nets [4]、AA - Net [1]、GSo P-Net1 [9]和GCNet [2]。评价指标包括效率(即,网络参数,每秒浮点运算(FLOPs))和训练/推理速度)和有效性(即Top - 1 / Top5精度)。为了进行比较，我们复制了[14]中ResNet和SENet的结果，并在他们的原始论文中报告了其他比较方法的结果。为了测试各种模型的训练/推理速度，我们使用了被比较的CNN的公开可用模型，并将它们运行在相同的计算平台上。结果如表3所示，我们可以看到我们的ECA - Net与原始的ResNet - 50具有几乎相同的模型复杂度(即网络参数、FLOPs和速度)，同时在Top - 1精度上获得了2.28 %的提升。与当前最先进的(即SENet、CBAM、A2 - Nets、AA - Net、GSoP - Net1和GCNet)相比，ECA - Net在获得更好或更有竞争力的结果的同时，还获得了更低的模型复杂度。

ResNet - 101使用ResNet - 101作为骨干模型，我们将ECA - Net与SENet [14]，CBAM [33]和AA - Net [1]进行了比较。从表3可以看出，在模型复杂度几乎相同的情况下，ECA - Net比原始ResNet - 101提高了1.8 %。在ResNet - 50上，ECA - Net表现出相同的趋势，优于SENet和CBAM，而在模型复杂度较低的AA - Net上，ECA - Net具有很强的竞争

力。注意到AA - Net是通过Inception数据增强和设置不同的学习率进行训练的。

ResNet - 152使用ResNet - 152作为骨干模型，我们将ECA - Net与SENet [14]进行了比较。从表3可以看出，在模型复杂度几乎相同的情况下，ECA - Net在Top - 1准确率方面比原始的ResNet - 152提高了约1.3 %。与SENet相比，ECANet在Top - 1上获得了0.5 %的增益，且模型复杂度更低。Res Net - 50、Res Net101和Res Net - 152的实验结果证明了我们的ECA模块在广泛使用的Res Net架构上的有效性。

MobileNetV2除了ResNet架构外，我们还验证了我们的ECA模块在轻量级CNN架构上的有效性。为此，我们采用MobileNetV2 [28]作为骨干模型，并将我们的ECA模块与SE模块进行比较。特别地，在残差连接位于MobileNetV2的每个"瓶颈"之前，在卷积层中集成SE block和ECA模块，SE block的参数r设置为8。所有模型均使用完全相同的设置进行训练。表3的结果表明，我们的ECA - Net在Top - 1准确率方面比原始的MobileNetV2和SENet分别提高了约0.9 %和0.14 %。此外，我们的ECA - Net比SENet具有更小的模型尺寸和更快的训练/推理速度。上述结果再次验证了我们的ECA模块的高效性和有效性。