论文译文 Coordinate Attention for Efficient Mobile Network Design

论文译文 Coordinate Attention for Efficient Mobile Network Design

Abstract

最近关于移动网络设计的研究已经证明了通道注意力(例如, Squeeze - and - Excite注意)对于提升模型性能的显著效果，但它们普遍忽略了位置信息，而位置信息对于生成空间选择性注意力图很重要。在本文中，我们通过将位置信息嵌入到信道注意力中，提出了一种新的移动网络注意力机制，我们称之为"协同注意力"。不同于通道注意力通过2D全局池化将特征张量转化为单个特征向量，坐标注意力将通道注意力分解为两个1D特征编码过程，分别沿两个空间方向聚合特征。通过这种方式，可以沿一个空间方向捕捉远距离依赖关系，同时沿另一个空间方向保留精确的位置信息。然后，将得到的特征图分别编码成一对方向感知和位置感知的注意力图，并将其补充应用到输入特征图中，以增强感兴趣对象的表征。我们的坐标注意力非常简单，可以灵活地插入经典的移动网络，如 MobileNetV2、MobileNeXt 和 EfficientNet，几乎没有计算开销。大量实验证明，我们的坐标注意力不仅有利于 ImageNet 分类，更有趣的是，它在物体检测和语义分割等下游任务中表现更佳。代码见 https://github.com/Andrew-Qibin/CoordAttention。

1 Introduction

注意力机制用于告诉模型 "关注什么 "和 "关注哪里"，已被广泛研究[47, 29]，并被广泛应用于提升现代深度神经网络的性能[18, 44, 3, 25, 10, 14]。然而，它们在移动网络（模型规模有限）中的应用明显落后于大型网络 [36, 13, 46]。这主要是因为大多数关注机制带来的计算开销不是移动网络所能承受的。

考虑到移动网络的计算能力有限，迄今为止，移动网络中最流行的注意力机制仍然是挤压-激发（SE）注意力[18]。它借助二维全局池计算通道注意力，以相当低的计算成本显著提高了性能。然而，SE 注意力只考虑了通道

间信息的编码，却忽略了位置信息的重要性，而位置信息对于捕捉视觉任务中的物体结构至关重要 [42]。后来的工作，如 BAM [30] 和 CBAM [44]，试图通过减少输入张量的通道维度来利用位置信息，然后使用卷积计算空间注意力，如图 2(b) 所示。然而，卷积只能捕捉局部关系，却无法模拟对视觉任务至关重要的长程依赖关系[48, 14]。

在本文中，在第一个工作的基础上，我们提出了一种新的高效的注意力机制，通过将位置信息嵌入到信道注意力中，使移动网络能够在大范围内进行关注，同时避免了巨大的计算开销。为了减少2D全局池化带来的位置信息损失，我们将通道注意力分解为两个并行的1D特征编码过程，以有效地将空间坐标信息整合到生成的注意力图中。具体来说，我们的方法利用两个1D全局池化操作分别将垂直和水平方向的输入特征聚合成两个独立的方向感知特征图。然后，这两个包含特定方向信息的特征图被分别编码成两个注意力图，每个注意力图捕捉输入特征图沿一个空间方向的长距离依赖关系。因此，位置信息可以保留在生成的注意力图中。然后通过乘法将这两种注意力图应用于输入特征图，以强调感兴趣的表征。我们将所提出的注意方法命名为坐标注意，因为它的操作可以区分空间方向（即坐标）并生成坐标感知的注意图。

我们的坐标注意力具有以下优势。首先，它不仅能捕获跨信道信息，还能捕获方向感知和位置敏感信息，这有助于模型更准确地定位和识别感兴趣的对象。其次，我们的方法灵活轻便，可以轻松插入移动网络的经典构建模块，如 MobileNetV2 [34] 中提出的倒残差模块和 MobileNeXt [49] 中提出的沙漏模块，通过强调信息表征来增强特征。第三，作为一个预训练模型，我们的协调注意力可以为移动网络的下游任务带来显著的性能提升，尤其是那些具有密集预测的任务（如语义分割），我们将在实验部分展示这一点。

为了证明所提出的方法相对于以前的移动网络注意力方法的优势，我们在 ImageNet 分类 [33] 和流行的下游任务（包括物体检测和语义分割）中进行了广泛的实验。在可学习参数和计算量相当的情况下，我们的网络在 ImageNet 上的前 1 级分类准确率提高了 0.8%。如图 1 所示，在物体检测和语义分割方面，与采用其他注意力机制的模型相比，我们也观察到了显著的改进。我们希望我们简单高效的设计能促进未来移动网络注意力机制的发展。

2 Related work

.....

3 Coordinate Attention

坐标注意力块可以看作一个计算单元，旨在增强移动网络学习到的特征的表达能力。它可以将任意中间特征张量X = [x1 , x2 , ... , xC]∈RC × H × W作为输入，并输出与X具有相同大小的增广表示的变换张量Y = [y1 , y2 , ... , yC]。为了提供对所提出的坐标注意力的清晰描述，我们首先重新考虑了在移动网络中广泛使用的SE注意力。

1. Revisit Squeeze-and-Excitation Attention

正如文献[18]所证明的，标准卷积本身很难建模信道关系。显式构建通道间依赖关系可以增加模型对信息通道的敏感度，这些信息通道对最终分类决策的贡献更大。此外，使用全局平均池化也可以帮助模型捕获全局信息，这是卷积所缺乏的。

在结构上，SE块可以分解为挤压和激励两个步骤，分别用于全局信息嵌入和信道关系的自适应重新校准。给定输入X，第c通道的挤压步可以表示为如下形式：

其中zc是与第c通道相关的输出。输入X直接来自具有固定内核大小的卷积层，因此可以看作是局部描述符的集合。挤压操作使得收集全局信息成为可能。

第二步，激励，旨在完全捕获通道依赖关系，可以表述为

其中·是通道乘法，σ是Sigmoid函数，（ z是一个变换函数生成的结果，表示如下：

这里，T1 和 T2 是两个线性变换，可以通过学习来捕捉每个通道的重要性。

SE 块已被广泛应用于最近的移动网络中[18, 4, 38]，并被证明是实现最先进性能的关键组成部分。然而，它只考虑了通过对信道关系建模来重新权衡每个信道的重要性，却忽略了位置信息，而我们将在第 4 节中通过实验

证明位置信息对于生成空间选择性注意图非常重要。在下文中，我们将介绍一种新的注意力模块，它同时考虑了信道间关系和位置信息。

2. Coordinate Attention Blocks

我们的坐标注意力通过两个步骤：坐标信息嵌入和坐标注意力生成，用精确的位置信息对通道关系和长程依赖关系进行编码。拟议的坐标注意模块图见图 2 右侧部分。下面我们将对其进行详细介绍。

2.1. Coordinate Information Embedding

全局池化常用于通道注意力中对空间信息进行全局编码，但它将全局空间信息压缩到通道描述符中，因此难以保留位置信息，而位置信息对于捕获视觉任务中的空间结构至关重要。为了鼓励注意力区块利用精确的位置信息捕捉空间上的长程相互作用，我们将公式（1）中的全局池化分解为一对一维特征编码操作。具体来说，在给定输入 X 的情况下，我们使用池化核的两个空间范围（H，1）或（1，W）来分别沿横坐标和纵坐标对每个通道进行编码。因此，高度h处第c通道的输出可以表示为

类似地，宽度为w的第c个通道的输出可以写为

上述两种变换分别沿两个空间方向聚合特征，生成一对方向感知特征图。这与通道注意方法中产生单一特征向量的挤压操作（公式 (1)）截然不同。这两种变换还能让我们的注意力模块捕捉到一个空间方向上的长程依赖性，并保留另一个空间方向上的精确位置信息，从而帮助网络更准确地定位感兴趣的对象。

2.2. Coordinate Attention Generation

如上所述，公式 (4) 和公式 (5) 可实现全局感受野并编码精确的位置信息。为了利用由此产生的具有表现力的表征，我们提出了第二种转换，即坐标注意力生成。我们的设计参考了以下三个标准。首先，就移动环境中的应用而言，新变换应尽可能简单、廉价。其次，它可以充分利用捕捉到的位置信息，从而准确地突出感兴趣的区域。最后但并非最不重要的一点是，它还应能有效捕捉频道间的关系，这在现有研究中已被证明是至关重要的[18, 44]。

具体来说，我们首先将公式 4 和公式 5 生成的汇总特征图连接起来，然后将它们发送到共享的 1 × 1 卷积变换函数 F1，得到

其中，[-, -]表示沿空间维度的连接操作，δ 是非线性激活函数，f∈RC/r×(H+W) 是中间特征图，用于编码水平方向和垂直方向的空间信息。这里，r 是缩减率，用于控制 SE 区块的大小。然后，我们将f沿空间维度拆分成两个独立的张量f h∈RC / r × H和f w∈RC / r × W .利用另外两个1 × 1卷积变换Fh和Fw，分别将f h和f w变换为与输入X具有相同通道数的张量，得到

记σ为sigmoid函数。为了降低开销模型的复杂度，我们常以适当的减少比例r(例如, 32)来减少f的信道数。我们将在实验部分讨论不同缩减率对性能的影响。然后，输出结果 gh 和 gw 将分别展开并用作注意力权重。最后，我们的坐标注意力模块 Y 的输出可以写成

Discussion.与只关注重新权衡不同通道重要性的通道注意力不同，我们的坐标注意力模块还考虑了空间信息的编码。如上所述，水平和垂直方向的注意力同时作用于输入张量。两个注意力图中的每个元素都反映了感兴趣的对象是否存在于相应的行和列中。这一编码过程能让我们的协调注意力更准确地定位感兴趣物体的确切位置，从而帮助整个模型更好地进行识别。我们将在实验部分详尽展示这一点。

3. Implementation

由于本文的目标是研究一种更好的方法来增强移动网络的卷积特征，因此我们在这里以两种具有不同残差块类型的经典轻量级架构（即 MobileNetV2 [34] 和 MobileNeXt [49]）为例，来展示所提出的坐标注意力块与其他著名轻量级注意力块相比的优势。图 3 显示了我们如何将注意力区块插入 MobileNetV2 中的倒残差区块和 MobileNeXt 中的沙漏区块。

4 Experiments

在本节中，我们首先介绍了实验设置，然后进行了一系列消融实验，以证明拟议的协调注意力中的每个部分对性能的贡献。接下来，我们将我们的方法与一些基于注意力的方法进行比较。最后，我们报告了在物体检测和

语义分割方面，与其他基于注意力的方法相比，我们提出的方法所取得的结果。

1. Experiment Setup

我们使用PyTorch工具箱[31]来实现所有的实验。在训练过程中，我们使用衰减和动量为0.9的标准SGD优化器来训练所有的模型。权重衰减始终设置为4 × 10 - 5。采用初始学习率为0.05的余弦学习调度。我们使用4个NVIDIA GPU进行训练，批量大小设置为256。在没有额外声明的情况下，我们以MobileNetV2作为基线，训练了200个历元的所有模型。对于数据增强，我们使用与MobileNetV2相同的方法。我们在ImageNet数据集上[33]报告了分类结果。

2. Ablation Studies

协调注意力的重要性。为了证明提出的协同注意力的性能，我们进行了一系列的消融实验，相应的结果都列在表1中。我们将水平注意力或垂直注意力从坐标注意力中移除，以看出编码坐标信息的重要性。如表1所示，在两个方向上都有注意力的模型具有与SE注意力模型相当的性能。然而，当同时考虑水平注意力和垂直注意力时，我们得到了最好的结果，如表1所示。这些实验反映了在可学习参数和计算成本相当的情况下，坐标信息嵌入对图像分类更有帮助。

Different weight multipliers.在此，我们以两个经典的移动网络（包括带有反转残差块的 MobileNetV2 [34] 和带有沙漏瓶颈块的 MobileNeXt [49]）为基准，观察在不同权重乘数下，与 SE attention [18] 和 CBAM [44] 相比，所提出方法的性能。在本实验中，我们采用了三种典型的权重乘数，包括{1.0, 0.75, 0.5}。如表 2 所示，以 MobileNetV2 网络为基准，使用 CBAM 的模型与使用 SE 注意力的模型结果相似。不过，在每种情况下，采用所建议的坐标注意力的模型都能获得最佳结果。在使用 MobileNeXt 网络时也可以观察到类似的现象，如表 3 所示。这表明，无论考虑沙漏瓶颈区块还是倒置残差区块，也无论选择哪种权重乘数，我们的坐标注意力都表现最佳，因为它采用了先进的方式同时对位置和信道间信息进行编码。

The impact of reduction ratio r.为了研究不同的注意力区块缩减率对模型性能的影响，我们尝试减小缩减率的大小并观察性能变化。如表 4 所示，

当我们将 r 减少到原始大小的一半时，模型的大小会增加，但性能会更好。这表明，通过减小缩减比来增加参数对改善模型性能非常重要。更重要的是，在本实验中，我们的协调注意力仍然优于 SE 注意力和 CBAM，这反映了我们提出的协调注意力对缩减率的稳健性。

4.3 Comparison with Other Methods

Attention for Mobile Networks.在表 2 中，我们将协调注意力与其他用于移动网络的轻量级注意力方法进行了比较，包括广泛采用的 SE 注意力 [18] 和 CBAM [44]。可以看出，加入 SE 注意力后，分类性能提高了 1% 以上。就 CBAM 而言，与 SE 注意力相比，图 2(b) 所示的空间注意力模块在移动网络中的作用似乎不大。然而，当考虑到所提出的坐标注意力时，我们取得了最好的结果。我们还在图 4 中直观地展示了采用不同注意力方法的模型所生成的特征图。显然，与 SE 注意和 CBAM 相比，我们的坐标注意能更好地帮助定位感兴趣的对象。

我们认为，与 CBAM 相比，所提出的位置信息编码方式具有两方面的优势。首先，CBAM 中的空间注意力模块会将通道维度压缩为 1，从而导致信息丢失。然而，我们的坐标注意力采用了适当的缩减比来降低瓶颈处的通道维度，从而避免了过多的信息损失。其次，CBAM 利用核大小为 7 × 7 的卷积层来编码局部空间信息，而我们的坐标注意力则通过两个互补的一维全局池化操作来编码全局信息。这使得我们的坐标注意力能够捕捉空间位置之间的长程依赖关系，而这对于视觉任务来说是至关重要的。

……