Non-local Neural Networks

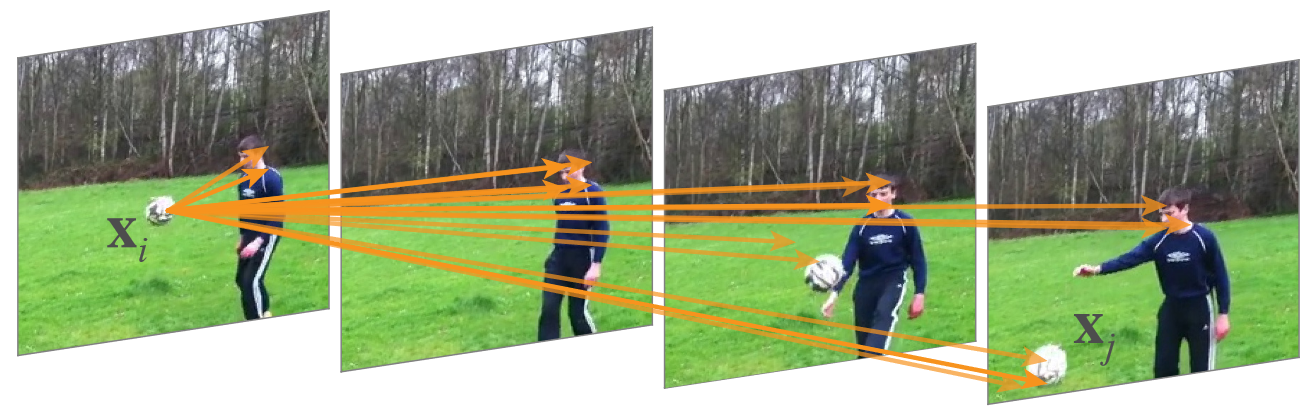
摘要

卷积和递归操作都是一次处理一个局部邻域的构建块。在本文中，我们提出了非局部操作，作为捕捉长距离依赖关系的通用构件系列。受计算机视觉中经典的非局部手段方法[4]的启发，我们的非局部操作将一个位置的响应计算为所有位置的特征的加权和。这个构件可以插入许多计算机视觉架构中。在视频分类的任务上，即使没有任何花哨的东西，我们的非本地模型也能在Kinetics和Charades数据集上竞争或超过目前的竞争者。在静态图像识别中，我们的非本地模型改善了COCO系列任务中的物体检测/分割和姿态估计。

代码可在https://github.com/ facebookresearch/video-nonlocal-net找到。

1. 介绍

捕捉长距离的依赖关系在深度神经网络中具有核心意义。对于序列数据（如语音、语言），递归操作[38, 23]是长距离依赖关系建模的主要解决方案。对于图像数据，长距离的依赖性是由卷积运算的深层堆叠形成的大接收场来建模的[14, 30]。卷积和递归操作都是在空间或时间上处理一个局部邻域；因此，只有当这些操作被反复应用，通过数据逐步传播信号时，才能捕捉到长距离的依赖关系。重复进行局部操作有几个局限性。首先，它1.在计算上是低效的。其次，它2.导致优化困难，需要仔细解决[23, 21]。最后，这些挑战使多跳依赖性建模变得困难，例如，当消息需要在遥远的位置之间来回传递时。在本文中，我们提出了非本地操作，作为一种高效、简单、通用的组件，用于用深度神经网络捕捉长距离的依赖关系。我们提出的非本地操作是计算机视觉中经典的非本地平均操作[4]的一般化。直观地说，非本地操作将某个位置的响应计算为输入特征图中所有位置的特征的加权和（图1）。



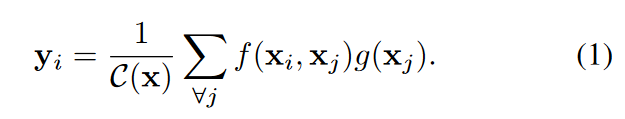
位置的集合可以是空间、时间或时空，这意味着我们的操作适用于图像、序列和视频问题。使用非局部操作有几个优点：（a）与递归和卷积操作的渐进行为相反，非局部操作通过计算任何两个位置之间的相互作用直接捕捉长距离的依赖关系，而不考虑它们的位置距离；（b）正如我们在实验中所显示的，非局部操作是高效的，即使只有几层（如5层）也能达到最佳效果；（c）最后，我们的非局部操作保持可变的输入大小，可以轻松地与其他操作相结合（如我们将使用卷积）。我们在视频分类的应用中展示了非本地操作的有效性。在视频中，远距离的像素在空间和时间上都会发生长距离的相互作用。一个单一的非本地块，也就是我们的基本单元，可以直接以前馈的方式捕捉这些时空依赖性。有了几个非本地块，我们称为非本地神经网络的架构比二维和三维卷积网络[48]（包括膨胀的变体[7]）更准确地进行视频分类。此外，非局部神经网络在计算上比其三维卷积网络更经济。对Kinetics[27]和Charades[44]数据集进行了综合消融研究。为了证明非局部操作的通用性，我们进一步介绍了COCO数据集[33]上的物体检测/分割和姿态估计实验。在强大的Mask R-CNN基线[19]的基础上，我们的非本地块能够以很小的额外计算成本提高所有三个任务的准确性。加上视频上的证据，这些图像实验表明，非局部操作通常是有用的，可以成为设计深度神经网络的一个基本构件。

3. 网络

我们首先给出了非本地操作的一般定义，然后我们提供了它的几个具体实例。

3.1. 公式

按照非局部平均操作[4]，我们将深度神经网络中的一般非局部操作定义为

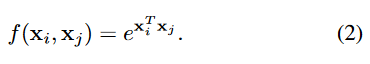


这里i是要计算其响应的输出位置（在空间、时间或时空）的索引，j是列举所有可能位置的索引。x是输入信号（图像、序列、视频；通常是它们的特征），y是与x相同大小的输出信号。一个配对函数f计算i和所有j之间的标量（代表关系，如亲和力）。公式（1）中的非局部行为是由于所有的位置（∀j）在操作中都被考虑到了。作为比较，卷积运算将局部邻域的加权输入相加（例如，在核大小为3的一维情况下，i - 1 ≤ j ≤ i + 1），而时间i的递归运算通常只基于当前和最近的时间步骤（例如，j = i或i - 1）。非局部操作也与全连接（fc）层不同。公式（1）根据不同位置之间的关系计算响应，而fc使用学习的权重。换句话说，在fc中，xj和xi之间的关系不是输入数据的函数，与非局部层不同。此外，我们在公式（1）中的表述支持不同大小的输入，并在输出中保持相应的大小。相反，fc层需要一个固定大小的输入/输出，并且失去了位置上的对应关系（例如，从xi到yi的位置i）。非局部操作是一个灵活的构建模块，可以很容易地与卷积/递归层一起使用。它可以被添加到深度神经网络的早期部分，而不像fc层那样经常被用在最后。这使我们能够建立一个更丰富的层次，结合非本地和本地信息。

3.2. 实例化

有趣的是，我们将通过实验表明（表2a），我们的非局部模型对这些选择并不敏感，表明通用的非局部行为是观察到的改进的主要原因。为了简单起见，我们只考虑线性嵌入形式的g：g(xj)=Wgxj，其中Wg是一个待学习的权重矩阵。这可以实现为，例如，空间中的1×1卷积或时空中的1×1×1卷积。接下来我们讨论配对函数f的选择。

**高斯。**在非局部平均数[4]和双边滤波器[47]之后，f的一个自然选择是高斯函数。在本文中，我们考虑

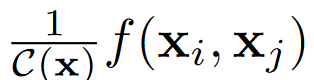


这里xT i xj是点积相似度。在[4，47]中使用的欧氏距离也适用，但点积在现代深度学习平台中更便于实施。

归一化系数设为

**嵌入的高斯。**高斯函数的一个简单扩展是计算嵌入空间中的相似性。在本文中，我们考虑



这里θ(xi)=Wθxi和φ(xj)=Wφxj是两个嵌入物。如上所述，我们设定。我们注意到，最近提出的用于机器翻译的自我注意模块[49]是嵌入式高斯版本中非局部操作的一个特例。这可以从以下事实看出来：对于给定的i，成为沿维度j的softmax计算，所以我们有，这就是[49]中的自我注意形式。因此，我们的工作通过将这种最新的自我注意模型与非局部手段的经典计算机视觉方法[4]联系起来，并将[49]中的顺序自我注意网络扩展到计算机视觉中用于图像/视频识别的通用空间/时空非局部网络，从而提供了深刻的见解。尽管与[49]有关系，但我们表明，在我们研究的应用中，注意力行为（由于softmax）并不重要。为了说明这一点，我们接下来描述两个非局部操作的替代版本。

**点积。**f可以被定义为点积的相似性。  


这里我们采用嵌入式版本。在这种情况下，我们将归一化系数设定为C(x)=N，其中N是x中的位置数，而不是f的总和，因为它简化了梯度计算。像这样的归一化是必要的，因为输入可以有可变的大小。点积和嵌入式高斯版本的主要区别在于softmax的存在，它起到了激活函数的作用。

**串联。**关系网络[40]中的成对函数使用串联法进行视觉推理。我们也评估了f的串联形式。



这里[-, -]表示串联，wf是一个权重向量，将串联的向量投射到标量上。在这种情况下，我们在f中采用ReLU[35].以上几种变体显示了我们的通用非局部操作的灵活性。我们相信其他的版本是可能的，并可能改善结果。

3.3非局部块

我们将公式（1）中的非本地操作包装成一个非本地块，可以被纳入许多现有的架构中。我们将一个非本地块定义为



其中yi在公式(1)中给出，"+xi "表示一个残余连接[21]。残余连接允许我们在任何预训练的模型中插入一个新的非局部区块，而不破坏其初始行为（例如，如果Wz被初始化为零）。图2展示了一个非本地块的例子。

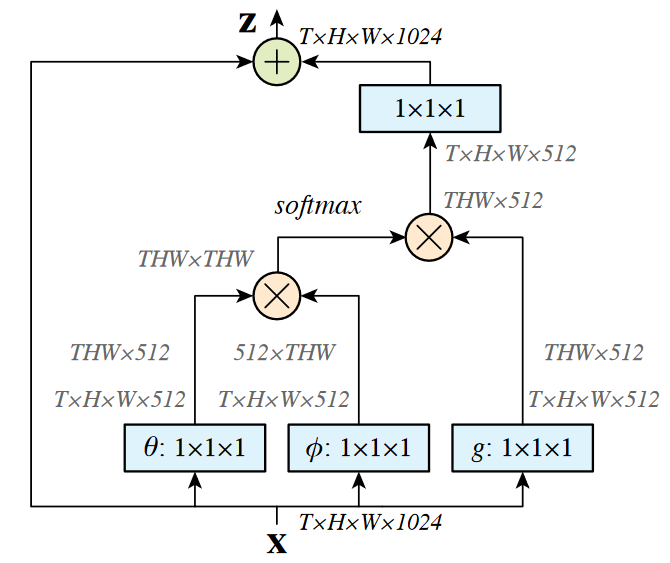


图2. 一个时空非局域块。特征图显示为其张量的形状，例如，T×H×W×1024为1024个通道（注意时进行了适当的重塑）。"⊗"表示矩阵乘法，"⊕"表示元素相加。对每一行都进行了softmax操作。蓝色方框表示1×1×1的卷积。这里我们展示的是嵌入式高斯版本，其瓶颈是512个通道。虚构的高斯版本可以通过去掉θ和φ来完成，而点积版本可以通过用1/N的缩放来代替softmax。

如图2所示，公式（2）、（3）或（4）中的成对计算可以通过矩阵乘法简单完成；（5）中的连接版本是直接的。当非局部区块的成对计算用于高层次、子采样的特征图时，它是轻量级的。例如，图2中的典型值是T=4，H=W=14或7。通过矩阵乘法进行的成对计算与标准网络中的典型卷积层相当。我们进一步采用了以下的实现方式，使其更有效率。  
**非局部区块的实现。**我们将Wg、Wθ和Wφ所代表的通道数量设定为x中通道数量的一半。这遵循了[21]的瓶颈设计，将一个块的计算量减少了一半左右。式(6)中的权重矩阵Wz在yi上计算出一个位置导向的嵌入，与x的通道数相匹配，见图2。一个子采样技巧可以用来进一步减少计算量。我们将公式(1)修改为：

yi = ，其中xˆ是x的一个子采样版本（例如，通过集合）。我们在空间域中进行，这样可以减少1/4的成对计算量。这一招并不改变非局部行为，而只是使计算更加稀疏。这可以通过在图2中的φ和g之后添加一个最大池层来实现。我们对本文研究的所有非本地块都使用了这些有效的修改。

7. 结论

我们提出了一类新的神经网络，它通过非局部操作捕捉长距离的依赖性。我们的非局部区块可以与任何现有的架构相结合。我们展示了非局部建模对于视频分类、物体检测和分割以及姿势估计等任务的重要性。在所有的任务中，简单地增加非本地块就能提供比基线更坚实的改进。我们希望非本地层将成为未来网络架构的一个重要组成部分。