**非本地神经网络**

罗斯Girshick2

枭龙Wang1 2∗

卡内基梅隆大学

1

**摘要**

*卷积运算和循环运算都是每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一个通用族的构建块捕获长期依赖。受计算机视觉中经典的非局部均值方法[4]的启发，我们的非局部操作以所有位置上特征的加权和计算位置上的响应。这个构建块可以插入到许多计算机视觉体系结构中。在视频分类任务上，即使没有任何附加功能，我们的非局部模型也可以在Kinetics和Charades数据集上竞争或超过当前的竞争对手。在静态图像识别中，我们的非局部模型改进了COCO系列任务的目标检测/分割和姿态估计。代码可在https://github.com/ facebookresearch/video-nonlocal-net获得。*

**1.介绍**

在深度神经网络中，获取长期依赖关系是非常重要的。对于顺序数据(例如，语音、语言)，循环操作[38,23]是长期依赖建模的主要解决方案。对于图像数据，远距离依赖是通过深度卷积运算堆栈形成的大接收域来建模的[14,30]。

卷积运算和递归运算都处理一个局部邻域，无论是在空间上还是在时间上;因此，只有在重复应用这些操作，并在数据中逐步传播信号时，才能捕获长期依赖关系。重复本地操作有几个限制。首先，它的计算效率很低。二是造成优化困难，需要谨慎处理[23,21]。最后，这些挑战使多跳依赖建模变得困难，例如，当消息需要在遥远的位置之间来回传递时。

在本文中，我们提出非局部操作作为一个有效的，简单的，通用的组件，以捕获长期依赖与深度神经网络。我们所提出的非局部运算是对计算机视觉中经典的非局部均值运算[4]的推广。直观地说,

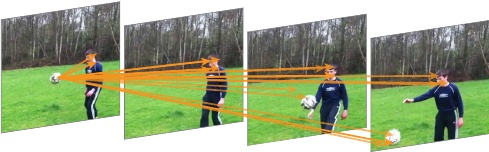
∗在Facebook人工智能研究实习期间完成的工作。

阿古普塔1

开明,他2

Facebook的人工智能研究

2



**x***i*

**x***j*

图1所示。在我们的网络中一个时空非局部操作训练视频分类动力学。一个位置xi的响应由所有位置x的特征的加权平均值计算j (这里只显示权重最高的)。在这个由我们的模型计算的例子中，请注意它是如何将第一帧中的球与最后两帧中的球联系起来的。图3中有更多的例子。

非局部操作以输入特征映射中所有位置特征的加权和计算位置上的响应(图1)。位置集合可以是空间、时间或时空上的，这意味着我们的操作适用于图像、序列和视频问题。

使用非局部运算有几个优点:(a)与递归运算和卷积运算的渐进行为相比，非局部运算直接通过计算任意两个位置之间的相互作用获取长期依赖关系，而不考虑它们的位置距离;(b)正如我们在实验中所显示的，非局部操作是有效的，即使只有几个层(例如，5)也能取得最好的结果;(c)最后，我们的非局部操作维护可变的输入大小，可以很容易地与其他操作组合(例如，我们将使用卷积)。

我们展示了非局部操作在视频分类应用中的有效性。在视频中，远距离的相互作用发生在遥远的像素之间的空间和时间。一个单独的非局部块，它是我们的基本单元，可以以前馈的方式直接捕获这些时空依赖性。通过一些非局部块，我们称为非局部神经网络的架构在视频分类方面比2D和3D卷积网络[48](包括膨胀的变种[7])更准确。此外，非局部神经网络比3D卷积网络在计算上更经济。在Kinetics[27]和Charades[44]数据集上进行了全面的消融研究。*只使用RGB，没有任何附加功能(例如，光流，多尺度测试)，我们的方法在两个数据集上取得了与最新竞赛获奖者相同或更好的结果。*

arXiv: 1711.07971 v3 [cs。CV] 2018年4月13日

1

为了证明非局部操作的通用性，我们进一步在COCO数据集[33]上进行了目标检测/分割和姿态估计实验。在强大的Mask R-CNN基线[19]之上，我们的非局部块可以以较小的额外计算成本提高所有三个任务的准确性。这些图像实验和视频上的证据表明，非局部操作通常是有用的，可以成为设计深度神经网络的基本构件。

**2.相关工作**

**非本地图像处理。**非局部意味着[4]是一种经典的滤波算法，它计算图像中所有像素的加权平均值。它允许远处的像素在基于补丁外观相似性的位置对过滤响应作出贡献。这种非局部过滤思想后来发展为BM3D (block-matching 3D)[10]，它对一组相似但非局部的补丁进行过滤。与深度神经网络[5]相比，BM3D是一个坚实的图像去噪基线。采用块匹配结合神经网络进行图像去噪[6,31]。非局部匹配也是纹理合成[12]、超分辨率[16]和修复[1]算法成功的关键。

**图形模型。**长期依赖可以通过图形模型来建模，如条件随机场(CRF)[29,28]。在深度神经网络中，CRF可以用于网络[9]的后处理语义切分预测。CRF的迭代平均场推理可以转化为一个循环网络并训练[56,42,8,18,34]。相比之下，我们的方法是一个简单的前馈块计算非局部滤波。与这些用于分割的方法不同，我们的通用组件用于分类和检测。这些方法和我们的方法还涉及到一个更抽象的模型，称为图神经网络[41]。

**序列的前馈建模。**最近出现了一种使用前馈的趋势。，非循环)网络建模语音和语言序列[36,54,15]。在这些方法中，长期依赖被非常深的一维卷积所产生的巨大的接受域所捕获。这些前馈模型适用于并行实现，并且比广泛使用的循环模型更有效。

**Self-attention。**我们的工作是关于最近的自我注意[49]机器翻译方法。自注意模块通过关注所有位置并在嵌入空间中取其加权平均值来计算序列中某个位置(如句子)的响应。在接下来,我们将讨论self-attention可以被视为一种非本地的意思是[4],在这个意义上讲我们的工作桥self-attention为机器翻译的更一般的类非本地过滤操作,适用于计算机视觉图像和视频问题。

**交互网络。***交互网络(Interaction Networks, IN)[2,52]是最近提出的用于物理系统建模的方法。*它们作用于一对对相互作用的物体的图形。Hoshen[24]在多智能体预测建模中呈现了更高效的顶点注意IN (VAIN)。另一种变体叫做关系网络[40]，它计算输入中所有位置的特征嵌入函数。我们的方法还处理所有对，我们将解释(f(x)i, xj )在Eq。(1))。当我们的非局部性网络与这些方法相连接时，我们的实验表明，模型的非局部性与注意/交互/关系(例如，网络可以关注一个局部区域)的思想是其经验成功的关键。非局部建模是图像处理的一个长期的关键元素(例如，[12,4])，但在最近的计算机视觉神经网络中很大程度上被忽视了。

**视频分类架构。**一种自然的视频分类解决方案是将成功的用于图像的神经网络和用于序列的神经网络结合起来[55,11]。相比之下，前馈模型是通过时空上的3D卷积(C3D)[26,48]来实现的，3D滤波器可以通过对预训练的2D滤波器进行“充气”[13,7]来形成。除了对原始视频输入进行端到端建模之外，已经发现光流[45]和轨迹[50,51]也有帮助。流动和轨迹都是现成的模块，可能会发现长期、非本地依赖。在[7]中可以找到对视频架构的系统比较。

**3.非本地神经网络**

我们首先给出非局部操作的一般定义，然后提供它的几个具体实例。

**3．1．配方**

继非局部均值操作[4]之后，我们将深度神经网络中的一种通用非局部操作定义为:



(1）

这里i是要计算响应的输出位置(空间、时间或时空)的索引，j是枚举所有可能位置的索引。X为输入信号(图像、序列、视频;通常他们的功能)和y是相同大小的输出信号为x。成对函数f计算一个标量(代表关系如亲和力)i和j。一元函数g之间计算输入信号的表示位置j。响应是C (x)规范化的因素。

式(1)中的非局部行为是由于在操作中考虑了所有位置(∀j)。作为比较，卷积运算将一个局部邻域内的加权输入求和(例如，核大小为3的一维情况下，i−1≤j≤i + 1)，并在时刻进行递归运算

I通常只基于当前和最新的时间步长

FORMULA

非本地操作也不同于全连接(fc)层。式(1)根据不同位置之间的关系计算响应，而fc使用学习到的权值。也就是说，x之间的关系j 和xi 与非本地层不同，在fc中不是输入数据的函数。此外，式(1)中的公式支持可变大小的输入，并在输出中保持相应的大小。相反，一个fc层需要一个固定大小的输入/输出，并且丢失了位置对应(例如，xi yi 在i)位置。

非局部操作是一种灵活的构建块，可以很容易地与卷积/循环层一起使用。它可以被添加到深度神经网络的早期部分，而不像fc层通常被使用在最后。这使我们能够构建一个结合非本地和本地信息的更丰富的层次结构。

**3.2。实例化**

接下来，我们将描述f和g的几个版本。有趣的是，我们将通过实验(表2a)表明，我们的非局部模型对这些选择并不敏感，这表明通用的非局部行为是观察到的改进的主要原因。

为简单起见，我们只考虑g的线性嵌入形式:g(xj ) = Wgxj , Wg 是一个待学习的权重矩阵。这可以实现为，例如1×1空间卷积或1×1×1时空卷积。

接下来我们讨论成对函数f的选择。

**高斯函数。**在非局部均值[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:



（2）

xTi xj 是点积相似。[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易于实现。归一化因子设为C(x) = P ∀jf (xi, xj )．

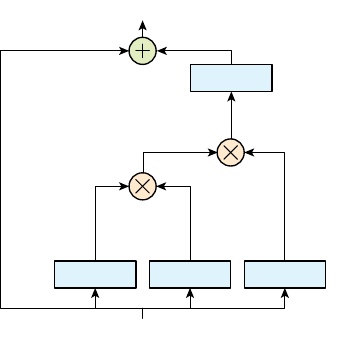
**嵌入式高斯。**高斯函数的一个简单扩展是计算嵌入空间中的相似度。在本文中，我们考虑:



（3)

θ(xi) = Wθxi和φ(xj)P= Wφxj为两嵌件。如上所述，我们令C(x) = ∀j f (xi, xj )．

我们注意到最近为机器翻译提出的自注意模块[49]是嵌入式高斯版本中非局部操作的一种特殊情况。这可以从一个事实看出对于给定的i, C(x)1f(xi, xj)成为沿维j的softmax计算



**z**

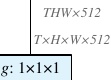
*W T H×××1024 1×1×1*



*失业的×失业*



*θ: 1×1×1 φ: 1×1×1*



**x T W H×××1024**

图2。时空非局部块。特征映射显示为其张量的形状，例如，T×H×W×1024为1024个通道(适当的重塑在注意时执行)。“⊗”表示矩阵乘法，“⊕”表示元素级和。对每一行执行softmax操作。蓝色的方框表示1×1×1个卷积。这里我们展示了嵌入式高斯版本，具有512通道的瓶颈。普通高斯版本可以通过去掉θ和φ来实现，点积版本可以通过1/N的缩放来代替softmax。

y = softmax(xT WθT Wφx)g(x)，即[49]中的自注意形式。因此，我们的工作通过将这个最新的自注意模型与经典的非局部means[4]计算机视觉方法联系起来，并将[49]中的顺序自注意网络扩展为用于计算机视觉图像/视频识别的一般空间/时空非局部网络。

尽管与[49]有关系，但我们发现注意行为(由于softmax)在我们研究的应用中并不是必不可少的。为了说明这一点，我们接下来描述非本地操作的两个替代版本。

**点积。**F可以定义为点积相似度:



（4）

这里我们采用嵌入式版本。在这种情况下，我们将归一化因子设为C(x) = N，其中N是x中的位置数，而不是f的和，因为它简化了梯度计算。这样的规范化是必要的，因为输入的大小可能是可变的。

点积和内嵌高斯版本的主要区别在于softmax的存在，它扮演着激活函数的角色。

**连接。**连接被关系网络[40]中的成对函数用于视觉推理。我们还计算了f的连接形式:



(5）

这里[·，·]表示连接和wf 是将连接向量投影到标量的权重向量。如上所述，我们设C(x) = n。本例中，f中采用ReLU[35]。

上面的几个变体展示了通用非本地操作的灵活性。我们相信替代版本是可能的，可能会改善结果。

**３．３．非本地块**

我们将Eq.(1)中的非局部操作包装成一个非局部块，该块可以被合并到许多现有的体系结构中。非局部块的定义如下:

(6）



在yi 式(1)和“+x ?i表示剩余连接[21]。剩余连接允许我们将一个新的非局部块插入到任何预先训练的模型中，而不破坏它的初始行为(例如，如果Wz 初始化为零)。非局部块示例如图2所示。Eq.(2)、(3)、(4)中的两两计算可以简单地通过矩阵乘法来完成，如图2所示;(5)中的连接版本很简单。

非局部块的两两计算是轻量级的，当它被用于高级的，sub-采样特征映射。例如，图2中的典型值是T = 4, H = W = 14或7。通过矩阵乘法所做的两两运算可与标准网络中的典型卷积层相媲美。我们进一步采用以下实现，使其更有效。

**非局部块的实现。**我们设置以W表示的通道数gWθ,和Wφ 这遵循了[21]的瓶颈设计，并将块的计算量减少了大约一半。权值矩阵Wz 式(6)计算沿位置嵌入yi，将通道数与x的通道数匹配，见图2。

一个次级抽样的技巧可以用来进一步重新

ducePcomputation。我们将Eq.(1)修改为:yi = 1 f(x，ˆx)g(ˆx)，其中xˆ为asub采样版本



j j

我

(例如，通过池化)。我们在空间域执行这个，可以减少1/4的两两计算量。这个技巧不会改变非局部行为，只是使计算变得稀疏。这可以通过在图2中的φ和g之后添加一个最大池化层来实现。

我们对本文研究的所有非局部块都使用了这些有效的修正。

**4.视频分类模型**

为了理解非局部网络的行为，我们对视频分类任务进行了综合消融实验。首先，我们描述了这项任务的基本网络架构，然后将其扩展为3D卷积网络[48,7]和我们提出的非本地网络。

**2D ConvNet基线(C2D)。**为了隔离非本地网络与3D卷积网络的时间效应，我们构建了

|  |  |  |
| --- | --- | --- |
|  | 层 | 输出的大小 |
| conv1 | 7×7, 64，步幅2,2,2 | 16×112×112 |
| 池1 | 3×3×3 max，步幅2,2,2 | 56 8××56 |
| res2 | 1×1, 64   3×3,64 ×3  1×1,256 | 56 8××56 |
| 池2 | 3×1×1 max，步幅2,1,1 | 56 4××56 |
| res3 | 1×1, 128   3×3,128 ×4  1×1,512 | 4×28×28 |
| res4 |   1×1,256     3×3,256 ×6  1×1,1024 | 4×14×14 |
| res5 |   1×1,512     3×3,512 ×3  1×1,2048 | 4×7×7 |
| 全球平均池，fc | | 1×1×1 |

表1。我们的基线ResNet-50 C2D视频模型。3D输出映射和过滤器内核的尺寸在T×H×W中(2D内核在H×W中)，通道的数量如下。输入为32×224×224。括号中显示了剩余块。

一个简单的2D基线架构，其中时间维度是微不足道的。，只有通过共用)。

表1显示了ResNet-50骨干下的C2D基线。输入视频剪辑有32帧，每个帧都是224×224像素。表1中的所有卷积本质上都是逐帧处理输入的2D内核(实现为1×k×k内核)。这个模型可以直接从ImageNet上预先训练的ResNet权重初始化。ResNet-101的对等体也是以同样的方式构建的。

唯一涉及时态域的操作是池化层。换句话说，这个基线只是汇总了时间信息。

**充气三维ConvNet (I3D)。**正如在[13,7]中所做的那样，可以通过“膨胀”内核将表1中的C2D模型转换为3D卷积模型。例如，一个2D k×k内核可以膨胀为一个3D t×k×k内核，它跨越t个框架。这个内核可以从2D模型(在ImageNet上进行预训练)初始化:t×k×k内核中的每个t平面都由预训练的k×k权重初始化，并按1/t重新调整。如果一个视频包含一个重复的静态帧，这个初始化产生的结果与

2D预训练模型在静态框架上运行。

我们研究两个通货膨胀的例子:我们要么使通货膨胀

3×3内核中的残留块到3×3×3(类似于[7])，或者第一个1×1内核中的残留块到3×1×1(类似于[13])。我们将其称为I3D3×3×3和I3D3×1×1。由于三维卷积是计算密集型的，我们每两个残块只膨胀一个核;膨胀更多层显示收益递减。我们充气conv1 5×7×7。

[7]的作者已经证明，I3D模型比CNN+LSTM模型更准确。

**非本地网络。**我们将非本地块插入到C2D或I3D中，从而将它们转换为非本地网络。我们研究添加1、5或10个非局部块;实现细节将在下一节的上下文中描述。



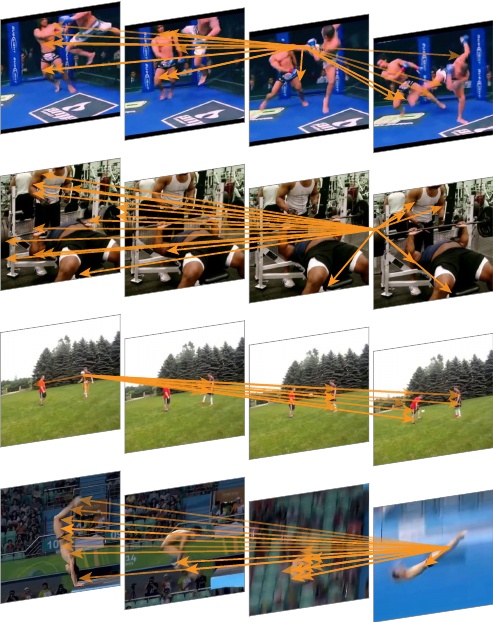


图3。非局部块在res中的行为示例3 由动力学训练的5块非局部模型计算。这些例子都来自于验证视频。箭头的起点代表一个xi，终点代表xj ．每个x的权重最高的20个箭头i 可视化。这4帧来自32帧的输入，显示为8帧的步幅。这些可视化显示了模型如何找到相关线索来支持其预测。

**4．1.实现细节**

**培训。**我们的模型是在ImageNet[39]上预先训练的。除非指定，我们使用32帧的输入剪辑来微调我们的模型。这些剪辑是由从原始全长视频中随机裁剪出64帧连续的帧，然后每隔一帧删除。空间大小为224×224像素，从缩放后的视频中随机裁剪，其较短的边按照[46]随机抽样[256,320]像素。我们在一台8-GPU的机器上训练，每个GPU在一个小批处理中有8个片段(所以总共有64个片段的小批处理)。我们总共为400k次迭代训练我们的模型，开始时学习率为0.01，每150k次迭代将学习率降低10倍(参见图4)。我们使用动量为0.9，权重衰减为0.0001。我们在全局池化层后采用dropout [22]， dropout ratio为0.5。我们在应用BatchNorm (BN)[25]时启用了它来微调我们的模型。这与常规的微调ResNets[21]形成对比，其中BN被冻结。我们发现，在我们的应用中启用BN可以减少过拟合。

我们采用[20]中的方法初始化非局部块中引入的权重层。我们在代表W的最后一个1×1×1层之后添加一个BN层z；我们不加

BN到非局部块中的其他层。该BN层的scale参数初始化为0，紧随[17]。这确保了整个非局部块的初始状态是一个身份映射，所以它可以插入到任何预先训练的网络，同时保持其初始行为。

**推理。**在[46]之后，我们对短边重调为256的视频执行空间上的全卷积推理。对于时域，在我们的实践中，我们从一个完整的视频中平均抽取10个剪辑，并分别计算它们的softmax分数。最终的预测是所有剪辑的平均softmax分数。

**5.视频分类实验**

我们对具有挑战性的动力学数据集[27]进行了全面研究。我们还报告了Charades数据集[44]的结果，以显示我们模型的通用性。

**5.1。动力学实验**

Kinetics[27]包含了约246k的训练视频和20k的验证视频。它是一个包含400个人工操作类别的分类任务。我们在训练集上训练所有模型，在验证集上测试。

|  |  |  |
| --- | --- | --- |
| 模型,R50 | 排名前 | 前5名 |
| C2D的基线 | 71.8 | 89.7 |
| 高斯 | 72.5 | 90.2 |
| 高斯,嵌入 | 72.7 | **90.5** |
| 点积 | **72.9** | 90.3 |
| 连接 | 72.8 | **90.5** |

(a)实例化:将1个不同类型的非本地块添加到C2D基线中。所有条目都使用ResNet- 50。

|  |  |  |
| --- | --- | --- |
| 模型,R50排名前 | | 前5名 |
| 基线 | 71.8 | 89.7 |
| 它 | 72.7 | 90.3 |
| res3 | **72.9** | 90.4 |
| res4 | 72.7 | **90.5** |
| res5 | 72.3 | 90.1 |

(b)阶段:将1个非本地区块加入不同阶段。所有条目都是ResNet-50。

模型(前5名

基线71.8 - 89.7

1-block 72.7 90.5 R50

5街区73.8 91.0 10街区74.3 91.2

基线73.1 - 91.0

1-block 74.3 91.3 R101

5-block 75.1 - 91.7

10块75.1 91.6 (c)更深的非局部模型:我们比较添加到C2D基线中的1、5和10个非局部块。我们显示了ResNet-50(顶部)和ResNet-101(底部)的结果。

|  |  |  |  |
| --- | --- | --- | --- |
|  | 模型 | 排名前 | 前5名 |
|  | 基线 | 71.8 | 89.7 |
| R50 | 空间仅  来讲 | 72.9  73.1 | 90.8  90.5 |
|  | 时空 | **73.8** | **91.0** |
|  | 基线 | 73.1 | 91.0 |
| R101 | 空间仅  来讲 | 74.4  74.4 | 91.3  90.5 |
|  | 时空 | **75.1** | **91.7** |

(d)空间、时间和时空:我们分别沿着空间、时间和时空维度比较应用的非局部操作。使用了5个非本地块。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型,R101 | 参数失败 | | 排名前 | 前5名 |
| C2D的基线 | 1× | 1× | 73.1 | 91.0 |
| I3D3×3×3 | 1.5× | 1.8× | 74.1 | 91.2 |
| I3D3×1×1 | **1.2×** | 1.5× | 74.4 | 91.1 |
| 问汇集,5-block | **1.2×** | **1.2×** | **75.1** | **91.7** |

(e) Non-local vs. 3D convv: 5块非本地C2D vs.膨胀的3D ConvNet (I3D)[7]。所有条目都使用ResNet-101。参数和FLOPs的数量相对于C2D基线(43.2M和34.2B)。

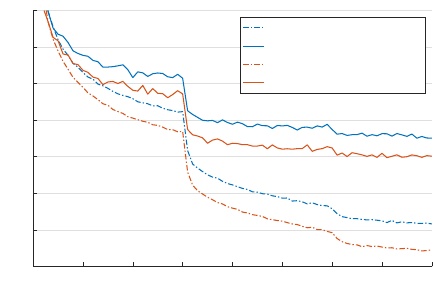
|  |  |  |  |
| --- | --- | --- | --- |
|  | 模型 | 排名前 | 前5名 |
|  | C2D的基线 | 71.8 | 89.7 |
| R50 | I3D | 73.3 | 90.7 |
|  | 问I3D | **74.9** | **91.6** |
|  | C2D的基线 | 73.1 | 91.0 |
| R101 | I3D | 74.4 | 91.1 |
|  | 问I3D | **76.0** | **92.1** |

(f) Non-local 3D ConvNet:在我们最好的I3D模型上添加了5个Non-local block。这些结果表明，非局部运算与三维卷积是互补的。

|  |  |  |  |
| --- | --- | --- | --- |
|  | 模型 | 排名前 | 前5名 |
|  | C2D的基线 | 73.8 | 91.2 |
| R50 | I3D | 74.9 | 91.7 |
|  | 问I3D | **76.5** | **92.6** |
|  | C2D的基线 | 75.3 | 91.8 |
| R101 | I3D | 76.4 | 92.7 |
|  | 问I3D | **77.7** | **93.3** |

(g)更长的剪辑:我们微调和测试表2f中的模型在128帧剪辑。我们非本地业务的收益是一致的。

表2。**烧蚀对动力学作用的分类。**我们显示了top-1和top-5分类准确率(%)。



60

C2D基线(train) C2D基线(val) NL C2D, 5-block (train) NL C2D, 5-block (val)

误差(%)

25

0

50

One hundred.

150

200次迭代(K)

250

300

350

400

55 50

45 40

35 30

图4。ResNet-50 C2D基线(蓝色)与非局部C2D 5块(红色)动力学训练程序曲线。我们展示了前1个训练错误(破折号)和验证错误(实线)。验证错误的计算方法与训练错误的计算方法相同(所以在训练时间使用相同的随机抖动进行1剪辑测试);最终结果如表2c (R50, 5-block)所示。

图4显示了ResNet-50 C2D基线与非局部C2D 5块训练过程的曲线(更多细节见下文)。在整个训练过程中，无论是训练误差还是验证误差，我们的非局部C2D模型始终优于C2D基线。

图1和图3显示了由我们的模型计算的非局部块的行为的几个例子。我们的网络可以学会寻找有意义的关系线索，而不管空间和时间的距离。

表2为烧蚀结果，分析如下:

**实例化。**表2a比较了添加到C2D基线的单个非局部块的不同类型(就在res的最后残余块之前)4)．即使添加一个非本地块，也可以比基线提高约1%。

有趣的是，嵌入式的高斯、点积和级联版本执行起来类似，但有一些随机变化(72.7到72.9)。如第3.2节所述，高斯内核的非局部操作变得类似于self-attention模块[49]。然而，我们的实验表明，该模块的注意力(softmax)行为并不是我们应用程序改进的关键;相反，非本地行为更可能是重要的，而且它对实例化不敏感。

在本文的其余部分，我们默认使用嵌入式高斯版本。这个版本更容易可视化，因为它的softmax分数在[0,1]的范围内。

**在哪个阶段添加非本地块?**表2b比较了添加到ResNet不同阶段的单个非本地块。区块被添加到一个阶段的最后一个剩余区块之前。对res上非局部块的改进2, res3或物4 是相似的吗5 是略小。一种可能的解释是5 空间规模小(7×7)，无法提供精确的空间信息。更多利用空间信息的非局部块的证据将在表2d中调查。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 模型 | 骨干 |  | 形态 | (瓦尔 | 前5名瓦尔 | (测试 | 前5测试 | | avg测试成果 | |
| I3D在[7] | 《盗梦空间》 | RGB |  | 72.1 | 90.3 | 71.1 | 89.3 | | 80.2 | |
| 2-流I3D在[7] | 《盗梦空间》 | RGB +流 | | 75.7 | 92.0 | 74.2 | 91.3 | | 82.8 | |
| [3]中的RGB基线 | Inception-ResNet-v2 | RGB |  | 73．0 | 90.9 | - |  | - |  | - |
| 三流晚期融合[3] | Inception-ResNet-v2 | RGB +流+音频 | | 74.9 | 91.6 | - |  | - |  | - |
| 3-stream LSTM [3] | Inception-ResNet-v2 | RGB +流+音频 | | 77.1 | 93.2 | - |  | - |  | - |
| 3-stream SATT [3] | Inception-ResNet-v2 | RGB +流+音频 | | 77.7 | 93.2 | - |  | - |  | - |
| 问I3D[我们] | ResNet-50  resnet - 101 | RGB  RGB |  | 76.5  **77.7** | 92.6  **93.3** | -  - |  | -  - | -  **83.8** | |

表3。与动力学的最新结果的比较，报告val和测试集。我们包括了2017年Kinetics竞赛的获胜者的结果[3]，但是他们的最佳结果利用了音频信号(灰色标记)，所以不仅仅是视觉解决方案。 †:“avg”为top-1和top-5准确率的平均值;在提交此稿件时，个人排名前1或前5的数字无法从测试服务器获得。

**更深入地研究非本地块。**表2c显示了更多非本地块的结果。我们添加1块(到res4)、5个街区(3到res4 2到res3，每隔一个剩余块)和10个块(每一个剩余块以res表示)3 和物4在ResNet-50);在ResNet-101中，我们将它们添加到相应的剩余块中。表2c显示了更多的非本地块通常会带来更好的结果。我们认为多个非本地块可以实现远程多跳通信。信息可以在时空中遥远的位置之间来回传递，这很难通过局部模型实现。

值得注意的是，非局部块的改进不仅仅是因为它们增加了基线模型的深度。为了看到这一点，我们注意到在表2c非本地5块ResNet-50模型有73.8的精度，比更深的ResNet-101基线的73.1要高。然而，与ResNet-101相比，5区块的ResNet- 50只有约70%的参数和约80%的FLOPs，而且更浅。这个比较表明，非局部块带来的改进与以标准方式进行更深层次的改进是互补的。

我们还尝试在基线模型中添加标准剩余块，而不是非局部块。准确性没有提高。这再次表明，非局部块的改进不仅仅是因为它们增加了深度。

**非本地时空。**我们的方法可以自然地处理时空信号。这是一个很好的属性:视频中的相关对象可以在遥远的空间和长期的时间间隔中呈现，它们的依赖性可以通过我们的模型捕获。

在表2d中，我们研究了沿空间、时间或时空应用的非局部块的影响。例如，在只有空间的版本中，非本地依赖只发生在相同的框架中:即。，式(1)中，它只对索引i的同一坐标系下的索引j进行求和，同样可以建立only time版本。表2d显示，仅空间和仅时间版本都比C2D基线有所改善，但不如时空版本。

**非本地网络vs. 3D ConvNet。**表2e比较了我们的非本地C2D版本和膨胀的3D ConvNets。非局部操作和三维卷积可以看作是将C2D扩展到时间维度的两种方法。

表2e还比较了相对于基线的参数和FLOPs的数量。我们的非本地C2D模型比I3D模型更准确(例如，75.1 vs 74.4)，而FLOPs的数量更少(1.2× vs. 1.5×)。对比表明，我们的方法比单独使用三维卷积更有效。

**非本地3 d事先。**尽管有上述比较，非局部操作和3D卷积可以对问题的不同方面进行建模:3D卷积可以捕获局部依赖性。表2f显示了在I3D中插入5个非局部块的结果3×1×1 模型。这些非局部I3D (NL I3D)模型比对应的I3D模型改进(+1.6点精度)，表明非局部操作和3D卷积是互补的。

**时间序列。**最后，我们研究了我们的模型在较长的视频输入上的通用性。我们使用128帧连续的没有子采样的输入剪辑。因此，网络中所有层的序列比32帧的对应序列长4倍。为了将这个模型放入内存中，我们将每个GPU的小批量大小减少到2个剪辑。由于使用小批量，在这种情况下，我们冻结了所有BN层。我们用32帧的输入对相应的模型进行初始化。我们使用与32帧相同的迭代次数对128帧的输入进行微调(尽管现在小批处理的大小更小)，从0.0025的学习速率开始。其他实现细节与之前相同。

表2g显示了128帧剪辑的结果。与表2f中的32帧相比较，所有模型在较长的输入时都有更好的结果。我们还发现，我们的NL I3D可以保持其增益相对于I3D的同行，表明我们的模型在较长的序列工作良好。

**与最先进的结果进行比较。**表3显示了来自I3D作者[7]和2017年Kinetics比赛冠军[3]的结果。我们注意到这些是系统的比较，它们在许多方面可以不同。然而，我们的方法超过了所有现有的基于RGB或RGB +流的方法。*不使用光流，也没有任何花哨的设计，我们的方法与2017年竞赛获胜者的设计成果相当。*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 模型 | 形态 | *火车/瓦尔* | *trainval /测试* | |
| 2-Stream [43] | | RGB +流 | 18.6 |  | - |
| 2-Stream + LSTM [43] | | RGB +流 | 17.8 |  | - |
| Asyn-TF [43] | | RGB +流 | 22.4 |  | - |
| I3D [7] |  | RGB | 32.9 | 34.4 | |
| I3D[我们] | | RGB | 35.5 | 37.2 | |
| 问I3D[我们] | | RGB | **37.5** | **39.5** | |

表4。分类mAP(%)在Charades数据集[44]上，在train/val分割和trainval/test分割。我们的结果基于ResNet-101。我们的NL I3D使用了5个非本地块。

**5．2.实验猜谜游戏**

Charades[44]是一个视频数据集，包含了约8k的训练、约1.8k的验证和约2k的测试视频。它是一个多标签分类任务，具有157个动作类别。我们使用每个类别的sigmoid输出来处理多标签属性。

我们在动力学(128帧)上初始化预先训练的模型。迷你批处理大小设置为每个GPU 1剪辑。我们训练我们的模型进行200,000次迭代，从0.00125的学习率开始，每75k次迭代减少10次。我们使用一个抖动策略类似于动力学确定224×224裁剪窗口的位置，但我们重新缩放视频，使这个裁剪窗口输出288×288像素，我们微调我们的网络。我们在320像素的单一尺度上进行测试。

表4显示了与之前在字谜游戏上的结果的比较。[7]的结果是2017年猜谜游戏的冠军，猜谜游戏也是根据动力学训练的模型进行了微调。我们的I3D基线比以前的结果更高。作为对照，我们的非本地网络在测试集上比我们的I3D基线提高了2.3%。

**6.扩展:COCO实验**

我们还研究了静态图像识别模型。我们在Mask R-CNN基线[19]上进行实验，用于COCO[33]目标检测/分割和人体姿态估计(关键点检测)。模型是在COCO train2017(即。2014年，trainval35k)，并在val2017上测试。2014年迷你版)。

**对象检测和实例分割。**我们通过添加一个非局部块来修改Mask R-CNN骨干

(就在最后一块保留地之前4)．所有模型都从ImageNet训练前微调。我们以ResNet-50/101为标准基线，以ResNeXt-152 (X152)[53]为高基线进行评估。与采用RPN阶段训练的原始论文[19]不同，我们使用了与[37]类似的端到端联合训练的改进实现，这导致了比[19]更高的基线。

表5显示了COCO上的框和掩码AP。我们看到一个单独的非本地块改善了所有R50/101和X152基线，包括所有涉及检测和分割的指标。美联社box 在所有情况下都增加了约1点(例如，R101中的+1.3点)。我们的非本地块是对增加模型容量的补充，即使模型是up-

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 方法 | | APbox | APbox 50 | APbox 75 | APmask APmask 50 | | APmask 75 |
| R50 | 基线  + 1问 | 38.0  **39.0** | 59.6  **61.1** | 41.0  **41.9** | 34.6  **35.5** | 56.4  **58.0** | 36.5  **37.4** |
| R101 | 基线  + 1问 | 39.5  **40.8** | 61.4  **63.1** | 42.9  **44.5** | 36.0  **37.1** | 58.1  **59.9** | 38.3  **39.2** |
| X152 | 基线  + 1问 | 44.1  **45．0** | 66.4  **67.8** | 48.4  **48.9** | 39.7  **40.3** | 63.2  **64.4** | 42.2  **42.8** |

表5所示。在Mask R-CNN中添加1个非局部块用于COCO对象检测和实例分割。主干是ResNet-50/101或ResNeXt-152[53]，均采用FPN[32]。

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | APkp | AP50 kp | AP75 kp |
| R101基线 | 65.1 | 86.8 | 70.4 |
| NL，头部+4 | 66.0 | 87.1 | 71.7 |
| NL，头部+4，脊柱+1 | **66.5** | **87.3** | **72.8** |

表6所示。添加非局部块到Mask R-CNN用于COCO关键点检测。骨干是ResNet-101, FPN为[32]。

从R50/101到X152。这一比较表明，尽管现有模型的深度/容量有所增加，但非本地依赖尚未被充分捕获。

此外，上述收益的代价非常小。单个非局部块只给基线模型增加了<5%的计算量。我们也尝试在骨干中使用更多的非本地块，但发现收益递减。

**关键点检测。**接下来我们评估Mask R-CNN中的非局部块以进行关键点检测。在[19]中，Mask R-CNN使用8个卷积层的堆栈来预测关键点为1-hot masks。这些层是本地操作，可能会忽略长距离的关键点之间的依赖关系。基于此，我们在关键点头部插入4个非局部块(每2个卷积层之后)。

表6显示了关于COCO的结果。R101强有力的基线,将4非本地块添加到关键点头导致∼1点增加关键点美联社。如果我们添加一个额外的非本地块骨干作为对象检测完成,我们观察一个总计1.4点增加关键点美联社的基线。特别是，我们看到AP的标准更加严格75 提升了2.4分，这意味着更强的本地化性能。

**7.结论**

我们提出了一种新的神经网络，它通过非局部操作捕获长期依赖关系。我们的非本地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、目标检测和分割以及姿态估计等任务的重要性。在所有任务中，简单地添加非本地块可以提供对基线的可靠改进。我们希望非本地层将成为未来网络架构的重要组成部分。

**感谢:这项工作得到了ONR MURI N000141612007、Sloan、Okawa Fellowship to AG和NVIDIA Fellowship to XW的部分支持。**我们还要感谢樊浩奇、杜川、王恒、Georgia Gkioxari和Piotr Dollar进行的许多有益的讨论。

**参考文献**

C. Barnes, E. Shechtman, A. Finkelstein, D. Goldman。Patchmatch:用于结构化图像编辑的随机对应算法。《计算机图形学报》，2009。2

[2] P. Battaglia, R. Pascanu, M. Lai, D. J. Rezende，等。用于学习物体、关系和物理的交互网络。神经信息处理系统(NIPS)， 2016。2

[10]卞勇，甘超，刘旭东，李峰，龙晓龙，李艳，齐华，周建军，文胜，林勇。一种基于时间建模的大规模视频分类方法。*arXiv: 1708.03805, 2017。*7

A. Buades, B. Coll，和j - m。莫雷尔。一种非局部图像去噪算法。计算机视觉与模式识别(CVPR)， 2005。1、2、3

[5] H. C.伯格，C. J.舒勒，S.哈梅林。图像去噪:纯神经网络能与BM3D竞争吗?计算机视觉与模式识别，2012。2

[6] H. C.伯格，C. J.舒勒，S.哈梅林。用多层感知器去噪图像，第2部分:训练的权衡和对其机制的分析。*arXiv: 1211.1552, 2012。*2

J.卡雷拉和A. Zisserman。Quo vadis，动作识别?一个新的模型和动力学数据集。计算机视觉与模式识别，2017。1 2 4 6 7 8

S. Chandra, N. Usunier和I. Kokkinos。使用深度嵌入的高密度低秩高斯crf。计算机视觉(ICCV)， 2017。2

[9] L.-C。陈，G.帕潘德里欧，I. Kokkinos, K. Murphy和A. L. Yuille。基于深度卷积网络和完全连接crf的语义图像分割。*arXiv: 1412.7062, 2014。*2

K. Dabov, A. Foi, V. Katkovnik和K. Egiazarian。图像去噪的稀疏三维变换域协同滤波。*图像处理学报，2007。*2

[11] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko和T. Darrell。用于视觉识别和描述的长期循环卷积网络。计算机视觉与模式识别，2015。2

A. A. Efros和T. K. Leung。基于非参数采样的纹理合成。计算机视觉(ICCV)， 1999。2

C. Feichtenhofer, A. Pinz和R. Wildes。用于视频动作识别的时空残差网络。神经信息处理系统(NIPS)， 2016。2、4

福岛和三宅。新认知:一种用于视觉模式识别机制的自组织神经网络模型。神经网络中的竞争与合作。施普林格,1982年。1

[15] J. Gehring, M. Auli, D. Grangier, D. Yarats和Y. N. Dauphin。卷积序列到序列学习。机器学习(ICML)， 2017。2

[16] D. Glasner, S. Bagon和M. Irani。从单一图像超分辨率。计算机视觉与模式识别(CVPR)， 2009。2

[17] P. Goyal, P. Doll´ar, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He。准确的，大的小批量sgd: 1小时的训练图像网。*arXiv: 1706.02677, 2017。*5

A. Harley, K. Derpanis和I. Kokkinos。使用局部注意掩码的分割感知卷积网络。计算机视觉(ICCV)， 2017。

2

K. He, G. Gkioxari, P. Doll´ar, R. Girshick。面具R-CNN。计算机视觉(ICCV)， 2017。2、8

何凯，张旭东，任士生，孙建平。深入研究整流器:在图像网分类上超越人类水平的性能。计算机视觉(ICCV)， 2015。5

何凯，张旭东，任士生，孙建平。用于图像识别的深度残差学习。计算机视觉与模式识别(CVPR)， 2016。1、4、5

G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. R. Salakhutdinov。通过防止特征检测器的共同适应改进神经网络。*arXiv: 1207.0580, 2012。*5

S. Hochreiter和J. Schmidhuber。短期记忆。*神经计算,1997年。*1

[24] y Hoshen。带注意评论的多智能体预测建模。神经信息处理系统(NIPS)， 2017。2

[25] S. Ioffe和C. Szegedy。批归一化:通过减少内部协变量偏移来加速深度网络训练。机器学习(ICML)， 2015。5

关键词:三维卷积神经网络，人的动作识别，卷积神经网络在机器学习国际会议(ICML)， 2010。2

[2]张建平，张建平，张建平，张建平，等。动态人体动作视频数据集。*arXiv: 1705.06950, 2017。*1、5

P. Kr¨ahenb¨uhl和V. Koltun。具有高斯边势的全连通crfs的有效推理。神经信息处理系统，2011。2

J. Lafferty, A. McCallum和F. C. Pereira。条件随机场:分割和标记序列数据的概率模型。在机器学习(ICML)国际会议，2001。2

B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel。反向传播在手写邮政编码识别中的应用。*神经计算,1989年。*1

[31] Lefkimmiatis。基于卷积神经网络的非局部彩色图像去噪。计算机视觉与模式识别，2017。2

[32] T.-Y。Lin, P. Doll´ar, R. Girshick, K. He, B. Hariharan, S. Belongie。用于目标检测的特征金字塔网络。计算机视觉与模式识别，2017。8

[33] T.-Y。Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ra- manan, P. Doll´ar, C. L. Zitnick。Microsoft COCO:上下文中的常见对象。在欧洲计算机视觉会议(ECCV)上。2014.2、8

Liu s, S. De Mello, Gu j, Zhong g, m.h。杨和j·考茨。通过空间传播网络学习亲和。神经信息处理系统(NIPS)， 2017。2

V. Nair和G. E. Hinton。修正线性单元改进了受限玻尔兹曼机。在机器学习国际会议(ICML)， 2010。3.

A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu。小波网:原始音频的生成模型。*arXiv: 1609.03499, 2016。*2

[37] Ren S.， He K.， R. Girshick, and J. Sun。更快的R-CNN:使用区域提议网络实现实时目标检测。*模式分析与机器智能，2017。*8

D. E. Rumelhart, G. E. Hinton, R. J. Williams。通过反向传播错误来学习表征。*自然,1986年。*1

A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei. A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei。ImageNet大型视觉识别挑战。*计算机视觉学报，2015。*5

A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pas- canu, P. Battaglia，和T. Lillicrap。一个用于关系推理的简单神经网络模块。神经信息处理系统(NIPS)， 2017。2、3

F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner和G. Monfardini。图神经网络模型。*神经网络学报，2009。*2

A. G.施维因和R. Urtasun。完全连接的深层结构网络。*arXiv:1503.02351, 2015。*2

G. A. Sigurdsson, S. Divvala, A. Farhadi和A. Gupta。用于动作识别的异步时间域。计算机视觉与模式识别，2017。8

G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, A. Gupta。家庭中的好莱坞:为了解活动而众包数据收集。2016年欧洲计算机视觉会议(ECCV)。1、5、8

[45] K. Simonyan和A. Zisserman。用于视频动作识别的双流卷积网络。神经信息处理系统，2014。2

[46] K. Simonyan和A. Zisserman。用于大规模图像识别的深度卷积网络。《学习表征国际会议》，2015。5

C. Tomasi和R. Manduchi。灰度和彩色图像的双边滤波。计算机视觉(ICCV)， 1998。3.

D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri。利用三维卷积网络学习时空特征。计算机视觉(ICCV)， 2015。1、2、4

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin。注意力是你所需要的。神经信息处理系统(NIPS)， 2017。2、3、6

Wang H.和C. Schmid。改进轨迹的动作识别。计算机视觉(ICCV)， 2013。2

王磊，乔勇，唐晓。基于轨迹集合深度卷积描述符的动作识别。计算机视觉与模式识别，2015。2

[52] N. Watters, A. Tacchetti, T. Weber, R. Pascanu, P. Battaglia和D. Zoran。视觉交互网络。神经信息处理系统(NIPS)， 2017。2

[53]谢绍华，R. Girshick, P. Doll´ar, Z. Tu, K. He。深度神经网络的聚合残差变换。计算机视觉与模式识别，2017。8

熊文伟，黄旭东，黄旭东，黄旭东。微软2016会话语音识别系统。声学、语音和信号处理(ICASSP)国际会议，2017。2

[55] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga和G. Toderici。除了短片段宠物:用于视频分类的深度网络。计算机视觉与模式识别，2015。2

郑s, S. Jayasumana, B. Romera-Paredes, V. Vineet, Su Z.， Du D.， Huang C.， P. H. Torr。作为递归神经网络的条件随机场。计算机视觉(ICCV)， 2015。2