【何恺明最新论文】非局部神经网络，打造未来神经网络基本组件

2017-11-23 新智元

新智元AI World 2017世界人工智能大会开场视频

中国人工智能资讯智库社交主平台新智元主办的 AI WORLD 2017 世界人工智能大会 11月8日在北京国家会议中心举行，大会以“AI 新万象，中国智能+”为主题，上百位AI领袖作了覆盖技术、学术和产业最前沿的报告和讨论，2000多名业内人士参会。新智元创始人兼CEO杨静在会上发布全球首个AI专家互动资讯平台“新智元V享圈”。

全程回顾新智元AI World 2017世界人工智能大会盛况：

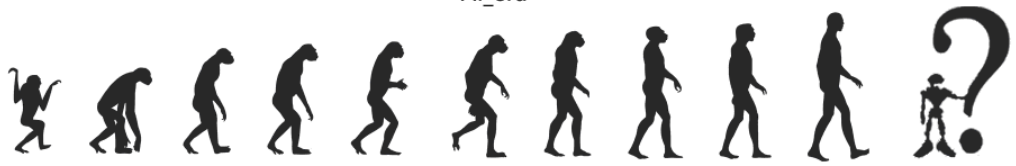
新华网图文回顾
http://www.xinhuanet.com/money/jrzb20171108/index.htm
爱奇艺
上午：http://www.iqiyi.com/v_19rrdp002w.html
下午：http://www.iqiyi.com/v_19rrdozo4c.html
阿里云云栖社区
https://yq.aliyun.com/webinar/play/316?spm=5176.8067841.wnnow.14.ZrBcrm

新智元报道

来源：arXiv；知乎

作者：费欣欣

【新智元导读】何恺明又出论文了，这次与CMU和FAIR的几位作者合作，提出了“非局部神经网络”。受计算机视觉中的经典非局部均值方法的启发而来，非局部网络可以作为一个简单高效的通用模块，嵌入现有视觉模型中，实验证明能够提高图像及视频分类精度，用作者的话说，在视频分类任务上，即使没有任何花里胡哨的处理，我们的非局部模型也能在 Kinetics 和 Charades 数据集上获得与一些当前视觉竞赛的冠军模型相当乃至更好的效果。



大神 Kaiming He 日前在 arXiv 上新挂出来一篇论文，标题延续了一贯的简洁风格，叫做《非局部神经网络》（Non-local Neural Networks）。

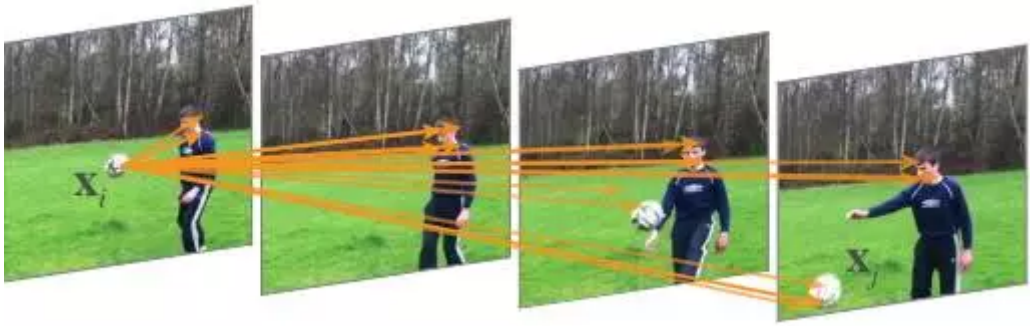
这是一篇 CMU 与 FAIR 合作的论文，第一作者是 CMU 的 Xiaolong Wang，其他两位作者是 Ross Girshick（DenseNet 作者之一）和 Abhinav Gupta（CMU 教授，他今年 CVPR + ICCV 一共发表了 15 篇论文）。

Non-local Neural Networks

Xiaolong Wang[1,2*] Ross Girshick[2] Abhinav Gupta[1] Kaiming He[2]
[1]Carnegie Mellon University [2]Facebook AI Research

摘要

卷积和递归运算都是一次处理一个局部邻域的基本计算组件（building block）。在本文中，我们将非局部运算（non-local operation）作为获取长时记忆（long-range dependency）的一类通用组件。受计算机视觉中的经典非局部均值方法的启发，我们的非局部运算将一处位置的响应计算为所有位置的特征的加权和。这个组件可以插入到许多计算机视觉结构中。在视频分类任务上，即使没有任何花里胡哨的处理，我们的非局部模型也能在 Kinetics 和 Charades 数据集上获得与一些当前视觉竞赛的冠军模型相当乃至更好的效果。在静态图像识别中，我们的非局部模型提高了 COCO 物体检测/分割和人体姿态估计这些任务的性能。代码今后公开。



作者训练好用于视频分类的一个网络，在时空上进行非局部运算的示例。计算时，位置 Xi 的响应等于时空中所有位置的特征的加权和。注意第一帧里球的位置与后两帧球的关联。来源：论文（下同）

$$\mathbf{y}_i = \frac{1}{\mathcal{C}(\mathbf{x})} \sum_{\forall j} f(\mathbf{x}_i, \mathbf{x}_j) g(\mathbf{x}_j).$$

论文中给出的深度神经网络非局部计算公式，其中 i 是（空间、时间或时空上的）输出位置的序号，计算的就是位置 i 的响应，j 是所有可能位置的序号。x 是输入信号（一般是特征，可以是图像数据、序列数据或视频数据），y 是与 x 相同大小的输出信号。函数 f 计算 i 和所有 j 之间的标量（代表两者间的关系）。一元函数 g 计算位置 j 处的输入信号的表征。最终响应通过因子C（x）归一化。

将非局部计算作为获取长时记忆的通用模块，提高神经网络性能

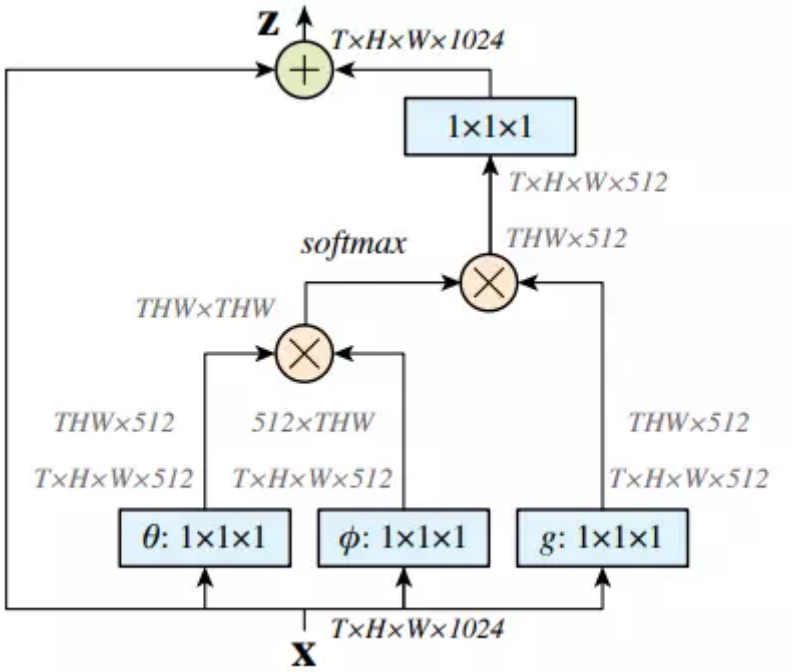
在深度神经网络中，获取长时记忆（long-range dependency）至关重要。对于序列数据（例如语音、语言），递归运算（recurrent operation）是长时记忆建模的主要解决方案。对于图像数据，长时记忆建模则依靠大型感受野，后者是多层卷积运算堆叠的结果。

卷积和递归运算处理的都是一个局部邻域，可以是空间局部邻域，也可以是时间局部邻域，因此只有不断重复这些运算，逐步在数据中传播信号，才能获取长时记忆。而不断重复局部计算有几个限制。首先，计算效率低下。其次，会产生一些优化问题，需要仔细解决。最后，这些问题使 multihop dependency 建模十分困难，multihop dependency 建模就是在很长的时间/空间位置之间来回传送信息。

非局部运算是计算机视觉中经典的非局部均值运算的一种泛化结果。直观地说，非局部运算将某一处位置的响应作为输入特征映射中所有位置的特征的加权和来进行计算。

在本文中，我们将非局部运算作为一个高效、简单和通用的模块，用于获取深度神经网络的长时记忆。我们提出的非局部运算是计算机视觉中经典的非局部均值运算的一种泛化结果。直观地说，非局部运算将某一处位置的响应作为输入特征映射中所有位置的特征的加权和来进行计算。这些位置可以是空间位置，也可以是时间位置，还可以是时空位置，这意味着我们的计算适用于图像、序列和视频问题。

作者在论文中写道，使用非局部运算有几大好处：（a）与递归和卷积运算的渐进的操作相比，非本局部运算直接通过计算任意两个位置之间的交互来获取长时记忆，可以不用管其间的距离；（b）正如他们在实验中所显示的那样，非局部运算效率很高，即使只有几层（比如实验中的5层）也能达到最好的效果；（c）最后，他们的非局部运算能够维持可变输入的大小，并且能很方便地与其他运算（比如实验中使用的卷积运算）相组合。



一个时空非局部组件。特征映射被表示为张量，⊗表示矩阵乘法，⊕表示单元和。每一行进行softmax。蓝框表示1×1×1的卷积。图中显示的是嵌入式高斯版本，具有512个通道的瓶颈。

“我们展示了非局部运算在视频分类应用中的有效性。在视频中，分隔开的像素在空间和时间上都会发生长时交互（long-range interaction）。我们的基本单元，也即单一的一个非局部模块，可以以前向传播的方式直接获取这些时空记忆。增加了几个非局部模块后，我们的“非局部神经网络”结构能比二维和三维卷积网络在视频分类中取得更准确的结果。另外，非局部神经网络在计算上也比三维卷积神经网络更加经济。我们在 Kinetics 和 Charades 数据集上做了全面的对比研究。我们的方法仅使用 RGB 数据，不使用任何高级处理（例如光流、多尺度测试），就取得了与这两个数据集上竞赛冠军方法相当乃至更好的结果。”

为了证明非局部运算的通用性，作者在 COCO 数据集上进行了物体检测、实例分割和人体姿态关键点检测的实验。**他们将非局部运算模块与 Mask R-CNN 结合，新模型在计算成本稍有增加的情况下，在所有三个任务中都取得了最高的精度。**由此表明非局部模块可以作为一种比较通用的基本组件，在设计深度神经网络时使用。

实验及结果

在这一节我们简单介绍论文中描述的实验及结果。

layer		output size
$conv_1$	7×7, 64, stride 2, 2, 2	16×112×112
$pool_1$	3×3×3 max, stride 2, 2, 2	8×56×56
res_2	[1×1, 64; 3×3, 64; 1×1, 256] ×3	8×56×56
$pool_2$	3×1×1 max, stride 2, 1, 1	4×56×56
res_3	[1×1, 128; 3×3, 128; 1×1, 512] ×4	4×28×28
res_4	[1×1, 256; 3×3, 256; 1×1, 1024] ×6	4×14×14
res_5	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	4×7×7
global average pool, fc		1×1×1

视频的基线模型是 ResNet-50 C2D。三维输出映射和滤波核的尺寸用T×H×W 表示（二维核则为H×W），后面的数字代表通道数。输入是32×224×224。方括号里的是残差模块。

model		top-1	top-5
R50	baseline	71.8	89.7
	1-block	72.7	90.5
	5-block	73.8	91.0
	10-block	**74.3**	**91.2**
R101	baseline	73.1	91.0
	1-block	74.3	91.3
	5-block	**75.1**	**91.7**
	10-block	**75.1**	91.6

(c) **Deeper non-local models**: we compare 1, 5, and 10 non-local blocks added to the C2D baseline. We show ResNet-50 (top) and ResNet-101 (bottom) results.

model		top-1	top-5
R50	baseline	71.8	89.7
	space-only	72.9	90.8
	time-only	73.1	90.5
	spacetime	**73.8**	**91.0**
R101	baseline	73.1	91.0
	space-only	74.4	91.3
	time-only	74.4	90.5
	spacetime	**75.1**	**91.7**

(d) **Space *vs.* time *vs.* spacetime**: we compare non-local operations applied along space, time, and spacetime dimensions respectively. 5 non-local blocks are used.

（c）展示了将非局部模块加入 C2D 基线后的结果，实验中用到了50层和101层的ResNet，可以看出，总体而言，增加的非局部模块越多，最后的精度越高。

（d）展示了时间、空间和时空同时非局部的效果，时空一起的效果最好。

（e）对比了非局部模块和三维卷积神经网络，增加了非局部模块（5个）的效果要好一点点。

model, R101	params	FLOPs	top-1	top-5
C2D baseline	1×	1×	73.1	91.0
$I3D_{3\times3\times3}$	1.5×	1.8×	74.1	91.2
$I3D_{3\times1\times1}$	**1.2**×	1.5×	74.4	91.1
NL C2D, 5-block	**1.2**×	**1.2**×	**75.1**	**91.7**

(e) **Non-local *vs.* 3D Conv**: A 5-block non-local C2D *vs.* inflated 3D ConvNet (I3D) [6]. All entries are with ResNet-101. The numbers of parameters and FLOPs are relative to the C2D baseline (43.2M and 34.2B).

（f）将非局部与三维卷积相结合的效果，结合了比单纯的三维卷积更好。

（g）检验了在128帧的视频中（f）中的模型的效果，发现能够保持比较稳定。

	model	top-1	top-5
R50	C2D baseline	71.8	89.7
	I3D	73.3	90.7
	NL I3D	**74.9**	**91.6**
R101	C2D baseline	73.1	91.0
	I3D	74.4	91.1
	NL I3D	**76.0**	**92.1**

(f) **Non-local 3D ConvNet**: 5 non-local blocks are added on top of our best I3D models. These results show that non-local operations are complementary to 3D convolutions.

	model	top-1	top-5
R50	C2D baseline	73.8	91.2
	I3D	74.9	91.7
	NL I3D	**76.5**	**92.6**
R101	C2D baseline	75.3	91.8
	I3D	76.4	92.7
	NL I3D	**77.7**	**93.3**

(g) **Longer clips**: we fine-tune and test the models in Table 2f on the 128-frame clips. The gains of our non-local operations are consistent.

最后，下面这张图展示了将非局部模块与 Mask R-CNN 结合后，在 COCO 物体检测、实例分割以及人体关键点检测任务中性能均有所提升，使用了50和100层的ResNet，以及152层的ResNeXt。

method		AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
R50	baseline	38.0	59.6	41.0	34.6	56.4	36.5
	+1 NL	**39.0**	**61.1**	**41.9**	**35.5**	**58.0**	**37.4**
R101	baseline	39.5	61.4	42.9	36.0	58.1	38.3
	+1 NL	**40.8**	**63.1**	**44.5**	**37.1**	**59.9**	**39.2**
X152	baseline	44.1	66.4	48.4	39.7	63.2	42.2
	+1 NL	**45.0**	**67.8**	**48.9**	**40.3**	**64.4**	**42.8**

Table 5. Adding 1 non-local block to Mask R-CNN for COCO **object detection** and **instance segmentation**. The backbone is ResNet-50/101 or ResNeXt-152 [51], both with FPN [30].

model	AP^{kp}	AP^{kp}_{50}	AP^{kp}_{75}
R101 baseline	65.1	86.8	70.4
NL, +4 in head	66.0	87.1	71.7
NL, +4 in head, +1 in backbone	**66.5**	**87.3**	**72.8**

Table 6. Adding non-local blocks to Mask R-CNN for COCO **keypoint detection**. The backbone is ResNet-101 with FPN [30].

知乎讨论：将传统视觉处理里的好方法整合进入深度神经网络，值得关注

知乎用户 Dr.Frankenstein 认为：这篇论文动机直接，但也需要一些想象力才能提出，将传统视觉处理里的好操作想法融合到深度网络里做成一个组件，延续了一直以来发paper的思路，值得关注的是实验结果。以下内容经授权转自知乎，详见文末链接。

> Non-local mean的传统工作，这篇文章里做了简要介绍。在这里应用到CNN里面，说得暴力一点，就是在做一个网络的时候把卷积核搞的跟整张图片一样大，那为啥要这么暴力呢？
>
> 我们要回归CNN一开始的设计思想。CNN一开始是面向目标实体识别的任务的。它就是要模拟人的认知方式，达到一个从局部到宏观的层次化认知流程。所以每一层的卷积核就不该设计的太大，底层的去捕捉轮廓信息，中层的组合轮廓信息，高层的组合全局信息。
>
> 但对于序列化的任务，这种思路就不一定能学到充分的需要的信息。比如一个人跳水的视频，每一帧中CNN可以很好的识别他的脚在哪，一个卷积核能覆盖的位置也就是脚及其四周。而要识别跳水的动作，我们要看到他的脚跟他的膝盖、他的大腿、胳膊，发生了一系列的相对位移关系，这些信息是将跳水区别于相扑运动的重要信息，因为运动员静态地看都是只穿小裤衩而已。而这些信息难以被关注于局部的卷积核收集到。
>
> 要注意的是，卷积核真的只关注于局部吗？如果只看一层，那答案就是“是的”。但纵观整个网络，不同的全局信息最终被综合，但由于sampling损失了大量信息，就没有这篇文章这种暴力做法来得效果明显。所以传统CNN不是很local，但是信息逐层传递丢失太多以致于不能有期待的效果。

> 到这里你会发现这篇文章的想法是很简单直接的，但是有趣的其实是实验的结果：
>
> 1. 单一的non-local block加在较浅层次效果显著。Reasonable。高层次丢失的信息太多了，找不到细小的远距离的联系，太模糊了。
>
> 2. 多个non-local block加入，也就是加深non-local特性，有一定效果提升但不会很明显。Reasonable。既然容易起作用的是在低层加，那么使劲加深其实意义不大，加多了这种东西就要考虑梯度消失和引入噪声。毕竟你把背景全都扔进来算。
>
> 3. 时空同时non-local比单一时间维度或单一空间维度效果都要好。这不是废话吗。
>
> 4. Non-local比三维CNN要好。也是废话。这是有人会问，non-local这么吊怎么不把卷积层全都替换掉？肯定不行的！你要依赖小卷积核去捕捉主体信息，同时用他的block捕捉全局信息，两者相辅相成才有好的效果。
>
> 值得注意的是，在视频变长以后，non-local的trick的提升变小了。Reasonable。因为在时间维度上，这些短视频帧数太短，时间维度上的小卷积得到的信息不足，劣势明显。时间变长了，non-local也不能handle这么大的信息量了，损失一些信息的小卷积反而不那么差劲了。
>
> 总结来说，insight可以的，不算很灌水。实验有些有趣的结论，但不属于极其优秀的那一类。调参是玄学，能不能有很好效果且看公布代码。

知乎用户 2prime 也做了想法类似的工作，我们取得授权后将部分评论节选如下：

> 文章里面提出来attention就是nonlocal算子的特例是非常有见识的，在视频的时间一档做了一个nonlocal attention，巧妙地将kernel做成两个函数的内积，就是\phi和\theta，然后矩阵内积做出来了kernel，最后内积上提取出来的特征1x1卷积核用来降维。

非局部神经网络究竟如何？至少，作者在结论中写道，所有任务，只需简单增加一个非局部模块，就能得到稳定的性能提升。我们希望非局部神经网络能够成为未来网络结构的基本组成部分（essential component）。

了解更多：

- 论文地址：https://arxiv.org/pdf/1711.07971.pdf
- 知乎用户 Dr.Frankenstein的回答：https://www.zhihu.com/question/68473183/answer/263743198
- 知乎用户 2prime的回答：https://www.zhihu.com/question/68473183/answer/263728472
- 更多知乎讨论：https://www.zhihu.com/question/68473183

阅读原文