视频行为识别相关研究进展

北京理工大学 高广宇 刘子铭 吴桐

一、 引言

行为识别(Action Recognition)是计算机视觉领域的一个重要的研究热点，近年来，随着大量的视频行为识别数据集被公开（如图 1），行为识别相关研究越来越热。文献[1]综述了人体行为识别公开数据集的发展与前瞻，文献[2]对视频行为识别进行了综述。然而，由于光照不确定、气候条件、压缩失真、摄像距离等因素产生的视频图像数据难以满足视觉识别和理解要求。视频行为识别也明显受到上述因素影响，如光照条件，角度变化，复杂背景和同类行为之间的差异，在很长时间没有重大突破。近年来，随着深度学习技术的出现和硬件计算能力的提升，在静态图像和视频数据上的行为识别模型性能得到了显著的提升。

在早期研究工作中，iDT（improved Dense Trajectories)[3]是性能最好的算法之一，很多

图 1 来自 UCF101 数据集的视频动作识别示意图[14]

当时最先进的模型都是基于此算法进行改进。近几年，随着深度学习技术的广泛应用，越来越多的基于深度学习的行为识别方法被提出来并成为主流方法。其中，双流架构（Two Stream）[4]，三维卷积网络（C3D）[5]以及基于循环神经网络的模型（RNN based Model）[6]成为了目前实现行为识别的最主要也是较为成功的三类方法。双流模型基于 2D 卷积网络，输入包括常见的 RGB 图像和光流图像（光流是一种描述视频中运动信息的模态，由视频的相邻帧计算得到，因此双流网络训练之前需要预先得到光流图片）。双流网络的计算量和参数量都是较小的，其较好的性能得益于光流特征的引入。3D 卷积模型是 2D 卷积的扩展，虽然 3D 卷积网络可以实现视频行为识别的最佳结果，但是其三维卷积操作也带来参数量的显著增长，这对计算资源提出来挑战。同时，也有一些工作对三维卷积操作进行了改进以实现更少的参数量更高的性能。最后，也有大量的工作基于循环神经网络(RNN)变体长短时记忆(Long Short-Term Memory,LSTM)网络[7]来设计行为识别模型。基于 LSTM 的模型利用了视频数据天然具有的时序特性，同样可以实现部分数据集上较好的结果。但是，由于图像和视频本质上存在的复杂多样的视觉变化，行为识别问题仍然没有被很好的解决。

事实上，视频行为识别需要考虑的一个重点问题是上下文依赖（Context Dependencies）建模。传统二维卷积操作可以用来建模空间上下文信息，三维卷积操作可以捕捉时空上下文关系，同时，循环神经网络被广泛用于建模序列数据的时序上下文信息。但是，这些操作都仅仅能够建模有限距离内的上下文信息，而长距离上下文信息的建模才是影响行为识别模型的性能和效果的瓶颈所在。对于长距离上下文的建模也已经有

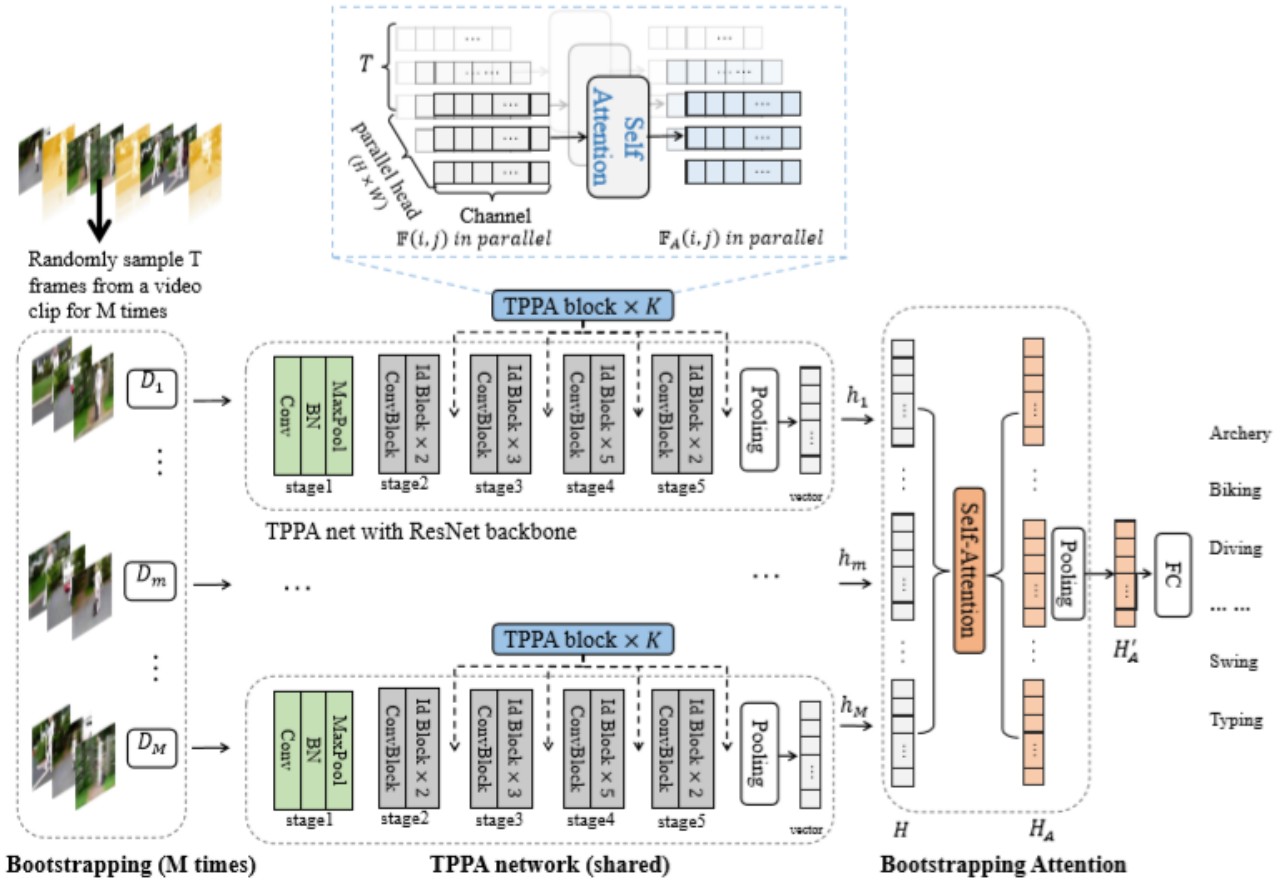


图 2 随机帧采样的集成注意力机制总体架构图

一些工作，例如，非局部网络（Non-local network）证明了在计算机视觉任务（目标检测，视频分类等）上建模长时序信息可以提升模型性能[8]。

此外，视频是高度冗余的数据形式，相邻帧之间可能非常相似。对于大多基于视频数据的行为识别模型（比如 C3D[5]，LSTM[7]），当这些模型在短时序上建模时，容易收缩为在单张图像上的操作。因此，探索长时序建模是另外一个基于视频的行为识别需要重点考虑的理由。

同时，行为通常可以看作是由同一个场景中多个局部特征之间的交互关系决定的，关系网络（Relation Network）证明了同一张图像的不同局部之间有紧密的关系[9]。而且，也有工作提出基于人体不同部位（局部）推理行为结果[10]；基于时空图模型的方法同样证明了局部特征对行为识别的作用[11]。因此，从局部特征角度建模上下文信息是行为识别研究的未来方向之一。

二、 基于长时序行为建模的行为识别模型

下面介绍我们提出的一种新的基于长时序行为建模的行为识别模型，如图 2 所示。该工作通过两个途径来建模长时序上下文：(1)基于自注意力机制（self-attention）[12]的注意力模块。(2)基于随机帧采样的自助法注意力（random frames based bootstrapping attention）行为识别框架[13]。

自注意力机制能够建模序列上任意位置与其他位置关系。简单的自注意力机制[12]可以由以下公式表示：

$$X' = softmax(\frac{X \cdot X}{\sqrt{d_k}}) \cdot X$$

其中，$X = \{x_i\}, i = 1, \ldots, N$, 是一个包含 N 个向量的矩阵；$X'$是注意力操作的输出，形状大小与输入$X$一致；$d_k$ 是其中向量的长度

虽然已经有工作将自注意力机制应用于计算机视觉任务，但是过去的工作仅仅通过注意力

机制建模t时刻图像所有像素与$(t+1)$时刻图像所有像素之间的关系，也就是将整个图像拉伸为一个向量。这种简单的方式不仅带来巨大的计算量而限制了应用范围，同时破坏了图像本身的空间特征。我们的工作充分考虑行为识别中对局部特征的利用，提出了基于时序像素的并行多头注意力模块（temporal pixel based parallel-head attention，TPPA），如图 3 所示。首先，我们定义了时序像素（temporal pixel）概念（所有时刻上在同一空间位置的像素集合，$T \times C$的矩阵），并以此为基本单位进行自注意力的计算。然后，图像的每个空间位置都进行同样的自注意力计算，所有空间位置上是并行的计算。这种在时序像素上操作的自注意力机制可以减少需要学习的参数量，让上下文的建模更容易优化。该模块建模长时序上下文是基于空间上局部特征进行，图像的空间特征得以保持。同时，整个操作被封装为一个模块，可以与任何流行的卷积神经网络结合使用。

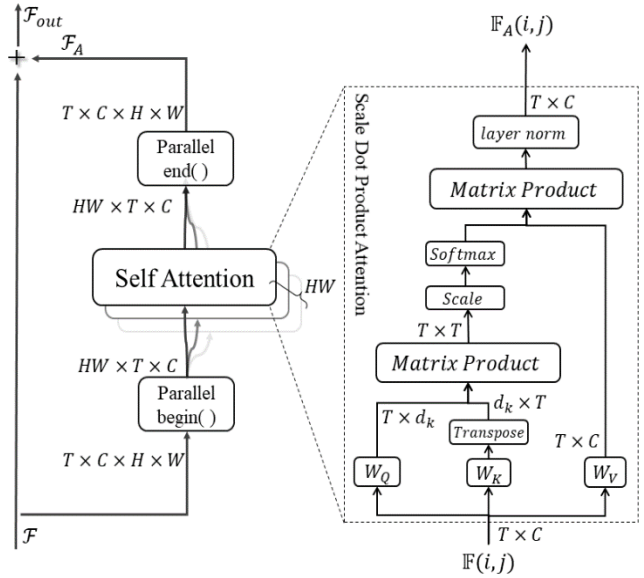


图 3 TPPA 模块的结构

对于大多数的行为类别，与行为识别相关的关键特征往往只是其中的几个关键帧，也就是说，我们可以根据关键帧实现行为识别。为了在一个冗余视频的长时序上下文中捕获这样的关键帧，我们引入了统计学中的自助法（bootstrapping）来提取关键帧。因为关键帧无法被人为先验得到，采用自助法思想，多次采样又放回的方式，每次采样的帧中可能是噪音也可能是有用的关键帧。多组采样结果通过主干网络得到多组特征，再同样采用自注意力的方式对多组特征重新加权，通过这样的方式增强关键帧的特征，弱化噪音的特征。这种架构被称为基于随机帧采样的自助法注意力（Random Frames based Bootstrapping Attention），整体框架见图 2。这个框架在常见行为识别数据集上都取得了最佳的结果，并且和 TPPA 模块一样，RFBA 架构可以与主流的 3D 卷积网络或者双流网络结合使用。

最后，图 4 展示了我们的模型在最常见的行为识别数据集 UCF101[14]上的实验结果。可以看出，我们所提出的基于长时序动作建模的识别模型最终取得了更好的行为识别结果。

Model	modality	acc@1
Two-stream[24]	RGB	83.6
Two-stream[24]	RGB + flow	91.2
LSTM[5]	RGB	81.0
3D-fused[2]	RGB	83.2
3D-fused[2]	RGB + flow	89.3
I3D[2]	RGB	84.5
TPPA net[ours]	RGB	**84.8**
RFBA net[ours]	RGB	**91.7**

图 4 UCF101 数据集上的结果对比

三、 总结

在这篇文章中，我们介绍了关于行为识别研究的最新进展，并且介绍了我们所提出的的最新行为识别模型——基于随机帧采样的自助法注意力框架（RFBA）。在我们的方法中，我们提出了一个新的注意力模块，基于时序像素的并行多头注意力模块（TPPA），这个新的模块可以更有效率地建模长时序上下文信息。

（责任编辑：任桐炜）

参考文献：

[1] 朱红蕾，朱昶胜，徐志刚. 人体行为识别数据集研究进展[J]. 自动化学报, 2018, 44(06):20-46.

[2] 罗会兰，王婵娟，卢飞. 视频行为识别综述[J]. 通信学报, 2018, v.39；No.372(06):173-184.

[3] Wang, Heng, et al. "Dense trajectories and motion boundary descriptors for action recognition." *International journal of computer vision* 103.1 (2013): 60-79.

[4] Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." *Advances in neural information processing systems*. 2014.

[5] Tran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." *Proceedings of the IEEE international conference on computer vision*. 2015.

[6] Du, Wenbin, Yali Wang, and Yu Qiao. "Rpan: An end-to-end recurrent pose-attention network for action recognition in videos." *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

[7] Donahue, Jeffrey, et al. "Long-term recurrent convolutional networks for visual recognition and description." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[8] Wang, Xiaolong, et al. "Non-local neural networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[9] Hu, Han, et al. "Relation networks for object detection." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[10] Zhao, Zhichen, Huimin Ma, and Shaodi You. "Single image action recognition using semantic body part actions." *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

[11] Wang, Xiaolong, and Abhinav Gupta. "Videos as space-time region graphs." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

[12] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

[13] Ziming Liu, Guangyu Gao, A. K. Qin, Tong Wu, and Chi Harold Liu. 2019. Action Recognition with Bootstrapping based Long-range Temporal Context Attention. In Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.

[14] Khurram Soomro, Amir Roshan Zamir and Mubarak Shah, UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild., CRCV-TR-12-01, November, 2012.

高广宇

北京理工大学副教授，数据科学与知识工程研究所副所长，硕士生导师。主要研究方向为深度学习，计算机视觉和多媒体。Email: guangyugao@bit.edu.cn

刘子铭

北京理工大学硕士研究生，软件工程专业。主要研究方向为计算机视觉。Email: liuziming.email@gmail.com

吴桐

北京理工大学硕士研究生，计算机技术专业。主要研究方向为计算机视觉。Email: 3220190896@bit.edu.cn

人体、人脸、人手表面及运动重建

清华大学 软件学院 徐枫 张浩

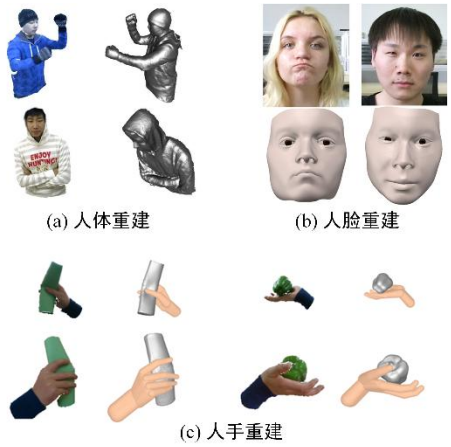
人体及人体局部区域（人脸和人手）的三维表面和运动重建一直以来就是计算机视觉(CV)和计算机图形学(CG)领域的热点研究问题，在教育、电影、游戏、动画等领域有着非常广泛和深远的应用。最近几年，随着 VR/AR，人工智能和 5G 技术的蓬勃发展，人们更是期望借助于技术的发展实现远距离全三维通信、虚拟试衣、实时在线三维直播、甚至实现智能机器人对人体各种运动的理解。因此，人体及其局部区域表面及运动重建技术越发得到了学术界和工业界的重视。

1. 人体重建

人体重建可以使用包含深度信息的相机对人体进行拍摄，并通过拍摄得到的图像序列重建出人体的三维几何模型与运动。根据应用的需求也可以重建出人体的表面颜色反射属性或者周围场景的几何模型。

我们的方法[1]可以使用 RGB-D 相机，即颜色-深度相机实现对人体等动态物体的实时重建，得到物体的几何模型、非刚性运动、表面反射率以及环境光照。在该方法中，对于每一帧的输入，首先使用迭代最近点匹配（Iterative Closest Point, ICP）算法求得相机相对于物体的运动；之后对物体的非刚性运动以及环境光照进行联合求解；最后更新物体的几何模型和表面反射率。物体的非刚性运动表达为表面上稀疏的节点带动空间中的体素进行运动，通过线性展开将对运动的求解化为线性问题，然后使用高斯牛顿法进行求解以达到实时重建的效果。表面反射则使用球谐函数表示的环境光与朗伯表面来进行表达。

我们也能够使用深度相机对动态的物体与其周围的静态场景进行实时重建[2]，可以用于重建室内的动态的人与静态场景。该方法将重建场景和采集的数据分为动态和静态两部分，两部分各自进行运动求解和模型的更新。在该方法中，动态部分的非刚性运动同样使用表面上稀疏的节点进行表达。此外，为了降低重建所占用的空间开销，该方法还使用了哈希体素来管理空间中体素的分配，并进一步使用非刚性运动节点来控制动态物体的体素分配。



(a) 人体重建 (b) 人脸重建 (c) 人手重建

图 1 人体及局部区域重建结果

2. 人脸重建

脸部是人最重要的器官之一，它可以表达人类的身份特征和丰富的情感变化。而在五官之中，眼睛是心灵之窗，它反映了人的关注点、心理活动变化等等。因此人脸的三维重建，尤其是眼部区域的三维重建在计算机图形学和计算机视觉领域都有着重要的意义和广泛的应用。

我们首先利用一套多线性人脸模型和人脸二维特征点进行 RGB 图像中的人脸重建，即将人脸模型上预定义的特征点投影到图像中与二维特征点取得一致。然而，人脸模型对于眼部区域的表达能力十分有限，其中不包含三维眼球模型，且眼皮的形状也与真实图像中相差较大。因此第一步[3]，我们在人脸模型中眼眶的相应位置上

增加了眼球模型来重建眼球的转动。我们根据图像中眼部区域的颜色分布估计眼球模型中虹膜与巩膜的颜色，而后利用基于三维模型表面泰勒展开的光度误差优化方法，通过减小三维眼球投影与二维图像差异的形式重建眼球运动。第二步[4]，我们设计了两组分别表示眼皮的形状特征和运动变化的线性眼皮模型来进行眼皮的三维重建。在图像中，我们利用深度学习的方法提取眼部区域四条主要边界作为二维眼皮特征，包括双眼皮、上眼皮、下眼皮和卧蚕。之后我们通过减小三维眼皮特征点与二维特征投影误差的形式，求解最优的眼皮模型的权重系数，并按其进行线性叠加得到最终的重建结果。最后我们将三维眼球与眼皮的重建结果融合到人脸的重建结果中，使得整体效果更加真实生动。

3. 人手重建

手是人体的重要组成部分，是人与环境进行交互最主要的执行工具。对手和物体的交互过程进行重建对于人体行为的精细化重建，以及机器理解人与环境的交互行为具有重大的意义。我们使用深度相机对交互过程进行拍摄，并通过拍摄得到的图像序列重建出人手的运动，物体的三维几何模型和运动（包含刚性和非刚性运动）[5]。

手与物体交互重建存在许多挑战。这些挑战包括：（1）遮挡，手的自遮挡和手与物体之间的互遮挡；（2）数据特征少，难以通过简单的颜色或深度阈值将手与物体数据分割；（3）手的特征少、数据少、运动复杂，难以进行精准的运动重建；（4）物体的非刚性运动维度高，且没有几何模型。为了解决上述问题，我们首先使用两台相对放置的深度相机对交互过程进行采集，降低遮挡对重建过程的影响；然后训练了一个深度神经网络来对手和物体的深度数据进行识别和分割，为手部运动重建提供指导；其次，借助手的SphereMesh 模型进行手部重建，综合考虑三维点云拟合误差，二维轮廓匹配误差和手运动的先验信息等，可较好的重建手的运动；最后，借助DynamicFusion 进行物体表面和非刚性运动重建，针对被手遮挡的物体区域，我们又引入了交互项，从而获得了较为准确的物体几何和非刚性运动。最终实现了手与物体交互过程的重建。

以上系列工作已发表于国际会议SIGGRAPH2019，SIGGRAPH2017，SIGGRAPH Asia 2017，TVCG2018，TVCG2016。

（责任编辑：邓成）

参考文献

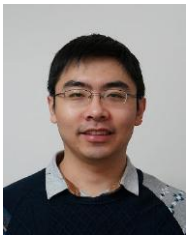
[1] Guo K , Xu F , Yu T , et al. Real-Time Geometry, Albedo, and Motion Reconstruction Using a Single RGB-D Camera[J]. ACM Transactions on Graphics, 2017, 36(3):1-13.

[2] Zhang H , Xu F . MixedFusion: Real-Time Reconstruction of an Indoor Scene with Dynamic Objects[J]. IEEE Transactions on Visualization and Computer Graphics, 2018:1-1.

[3] Wen Q, Xu F, Yong J H. Real-time 3d eye performance reconstruction for rgbd cameras[J]. IEEE transactions on visualization and computer graphics, 2016, 23(12): 2586-2598.

[4] Wen Q, Xu F, Lu M, et al. Real-time 3d eyelids tracking from semantic edges[J]. ACM Transactions on Graphics (TOG), 2017, 36(6): 193.

[5] Zhang, H., Bo, Z. H., Yong, J. H., & Xu, F. (2019). InteractionFusion: real-time reconstruction of hand poses and deformable objects in hand-object interactions. ACM Transactions on Graphics (TOG), 2019, 38(4), 48.



徐枫

清华大学副教授，主要研究方向为人体及场景三维重建。

Email: feng-xu@tsinghua.edu.cn

张浩

清华大学在读博士生，主要研究方向为人体及人手三维运动重建。Email: zhanghao16@mails.tsinghua.edu.cn

可微分网络架构搜索的稳定性研究

华为 诺亚方舟实验室 谢凌曦

目前，网络架构搜索已经成为自动化机器学习（AutoML）领域的重要研究方向。常见的网络架构搜索算法分为两大类，即启发式网络架构搜索和可微分网络架构搜索。相比于启发式搜索算法对计算资源的极度依赖，可微分搜索算法具有搜索速度快、能够快速部署到较大的搜索空间上的优势，因而得到了学界的广泛关注。然而，可微分搜索算法的稳定性一直饱受质疑，这也在一定程度上限制了它们在实际问题中的应用。

在最近的研究中，我们提出了一种假设，将网络架构搜索的不稳定性归结为搜索过程中优化的超网络和搜索结束后保留的子网络之间的结构差异——亦即，最优的超网络并不一定对应于最优的子网络，但是现有的算法大多忽略了这一点。为此，我们提出了一系列的改进方法，以消除这类结构差异带来的负面效果。不失一般性，我们的工作基于 DARTS，一种通用的可微分搜索算法。

在第一个研究课题中，我们注意到了搜索过程中超网络的深度和搜索结束后保留的子网络的深度存在较大差异。也就是说，当前的搜索算法假设不同深度的情况下的最优网络具有相同的局部结构，而这显然是不合理的。为此，我们提出了一种渐进加深的搜索算法，使得搜索阶段临近结束时的超网络逼近最终使用的子网络的深度。同时，为了使得超网络搜索适应不同的深度，我们还提出了几种实用的训练策略以增加搜索过程的稳定性。我们的方法在单张 GPU 上只需要 7 个小时就能够完成搜索，并且在 CIFAR 和 ImageNet 数据集上都取得了稳定的提升。这一工作[1]被 ICCV 2019 接收为口头报告论文。

在后续的研究课题中，我们又注意到了超网络和子网络的拟合性质存在较大差异。超网络中的每个连接都包含有多种不同的基本操作，但是子网络的每个连接只能从这些基本操作中选择一个。这就决定了超网络能够以不同的方式拟合训练数据集，但是这些方式不一定都适用于子网络，因而子网络的最终性能存在较大差异。为此，我们以通道采样的方式来限制超网络的能力，使得过拟合现象得到很大程度的缓解。同时，这种采样方法还降低了搜索过程的计算开销，并且通过增加每个批次内的样本数，提升了搜索的稳定性。我们在单张 GPU 上的运行时间能够进一步下降到 2 个小时以内，并且在 CIFAR 和 ImageNet 上取得很好的效果。这一工作[2]已经提交到 ICLR 2020。

我们最新的一项研究表明，除了上述差异，现有的可微分网络架构搜索还面临一个更加严重的问题，即由分步梯度优化带来的数学近似。我们发现这种近似方式将会带来极大的误差，然而现有的方法都无法直接消除这一误差，只能以早停的方式来减轻误差的伤害。由于搜索过程尚未收敛便要结束，因此随机初始化就会很大程度上影响最后的结果，从而带来搜索的不稳定性质。针对这一问题，我们提出了一种修正误差的方案，以一种巧妙的方式提升了梯度近似的精度。在这一修正的搜索算法下，我们能够确保搜索算法在收敛后仍然取得令人满意的性能。这一工作已经提交到 ICLR 2020 [3]。

总结来说，可微分网络架构搜索是一个迷人而充满前景的研究方向。然而，这套方法存在的问题还有很多，自动化机器学习距离真正的一统天下还有很长的路要走。

（责任编辑：苏航）

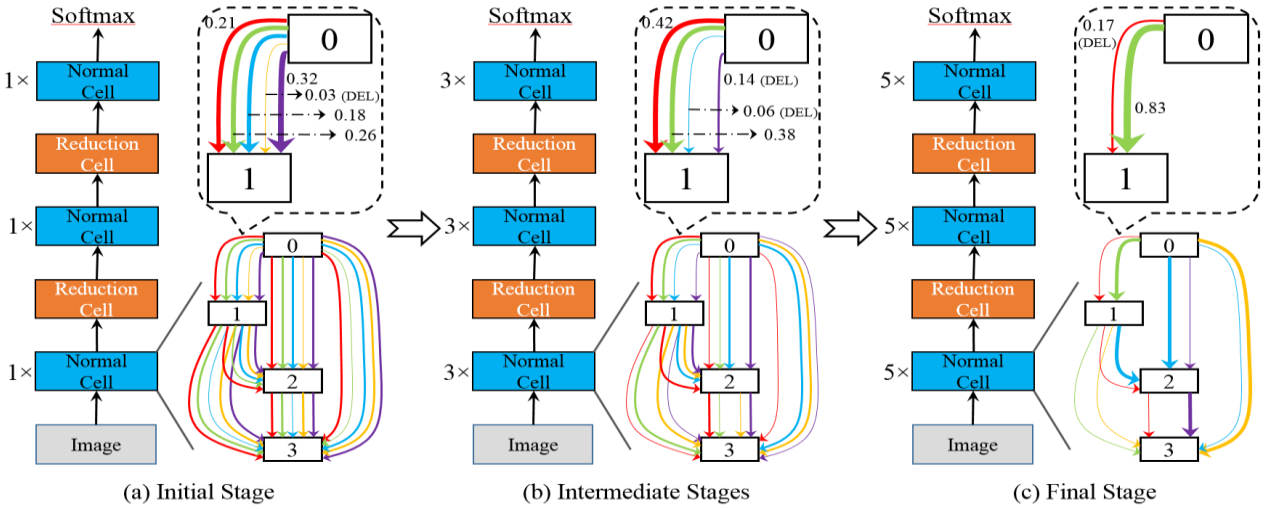


图 1 可微分网络架构搜索总体架构图

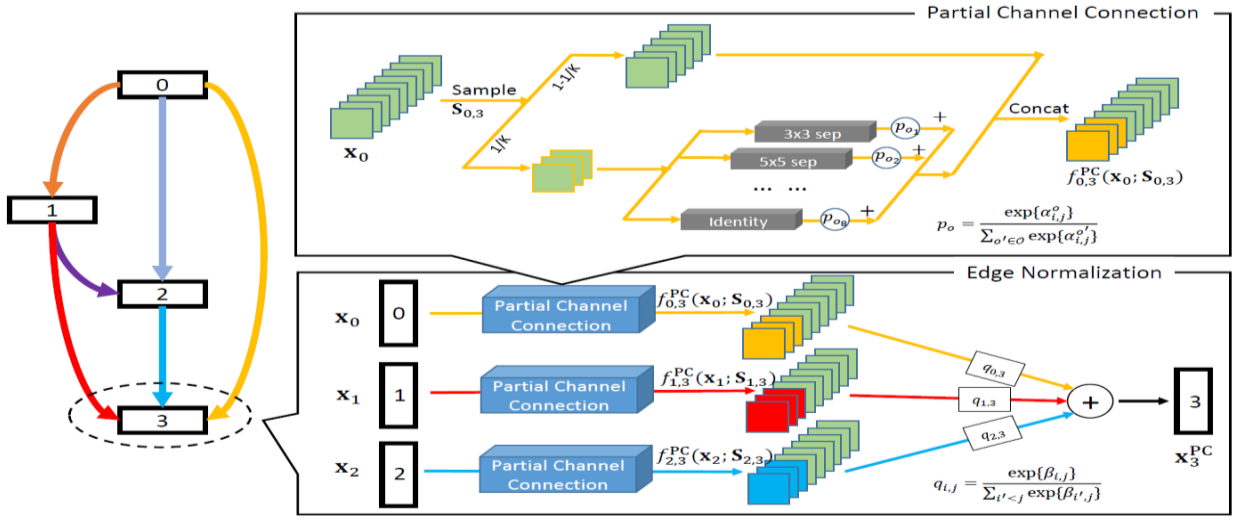


图 2 部分通道连接的可微分网络架构搜索

参考文献

[1] X. Chen et al., Progressive Differentiable Architecture Search: Bridging the Depth Gap between Search and Evaluation, ICCV, 2019.

[2] Y. Xu et al., PC-DARTS: Partial Channel Connections for Memory-Efficient Differentiable Architecture Search, arXiv preprint: 1907.05737, 2019.

[3] K. Bi and C. Hu et al., Stabilizing DARTS with Amended Gradient Estimation on Architectural Parameters, arXiv preprint: 1910.11831, 2019.

谢凌曦

华为诺亚方舟实验室高级研究员，主要研究方向是计算机视觉。

Email：198808xc@gmail.com