深度学习模型在人体行为识别领域方法综述

王艺博 201714403

摘要: 理解视觉数据中的人为行为与互补研究领域的进展密切相关，包括对象识别，人体动力学，领域适应和语义分割。在过去的十年中，人类行为分析从早期的计划发展而来，这些计划通常仅限于受控环境，现在可以从数百万视频中学习并应用于几乎所有日常活动的高级解决方案。从视频监控到人机交互，计算机互动，动作识别中的科学里程碑得以更快速地实现，近年来，随着深度学习领域的飞速发展，图片中物品的识别得到了长足的发展，因此，深度学习是否能够应用到人体行为识别领域还有待研究。本文着重总结了目前在视频中的人体行为领域取得不错效果的深度学习模型，希望能够为未来的研究打下坚实的基础。

0 介绍

由于深度和数据驱动架构，我们目睹了无数任务的重大进步。深度神经网络如卷积神经网络（CNN）[1]已经成为学习图像内容的首选方法[2,3,4,5]。一般来说，学习的问题是从可用数据中确定一个复杂的决策函数。

在深层架构中，这是通过组合多层次的非线性操作来实现的。考虑到决策表面的非凸性，搜索深层架构的参数空间并不容易。基于梯度下降方法的学习算法具有新硬件的计算能力当大量的注释数据可用时，已证明它是成功的[6,7,8]。本节我们的目的是讨论已经使用（或可能用于）解决视频学习行为问题的深层模型。从分类学的角度来看，我们可以确定四类适用于行为识别的架构

下面，我们详细讨论每个类别，并介绍开放问题和可能改进。

1 人体行为识别领域深度网络模型

1.1 时空神经网络

卷积体系结构有效地利用图像结构通过“池化”和“权重共享”来减少网络的搜索空间（见图 1.1(左)为概念图)。池化和权值共享也有助于实现模型对尺度和空间变化的鲁棒性。分析由 CNN 架构学习的卷积核表明，第一层学习低级特征（例如，类似 Gabor 的卷积核），而顶层学习高级语义[9]。 这进一步扩展了卷积网络作为通用特征提取器的用途。

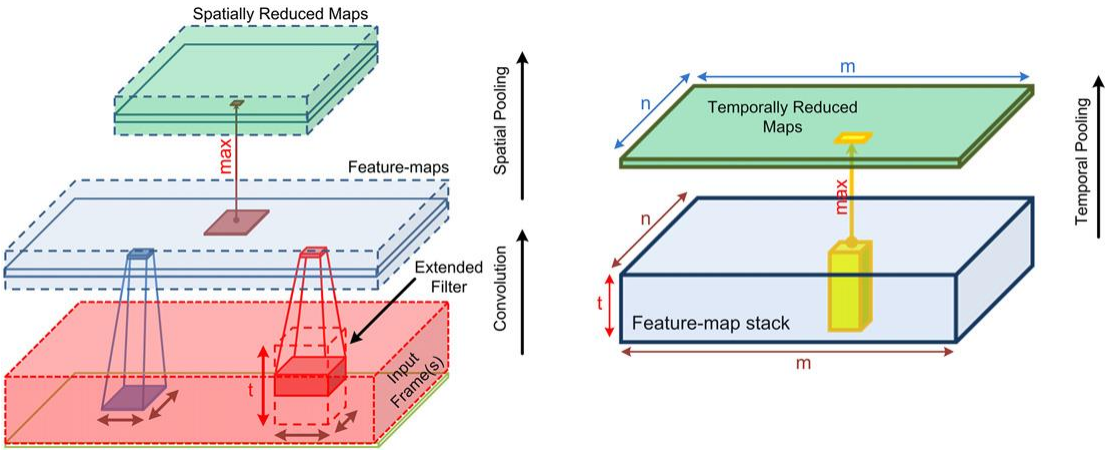

图 1.1 时空操作：2D 卷积（蓝色），3D 帧卷积（红色），如 Ji 等人[11]，传统的空间最大池化（棕色）和时间最大池化（黄色），如 Ng 等人[10]。

一种直接采用深度网络进行行为识别方法是利用时间信息优化卷积运算。为实现这一点，在 Ji 等人的文章中引入了 3D 卷积网络[11]。 顾名思义，3D 卷积网络使用 3D 内核（沿时间轴延伸的卷积核）从中提取特征.因此预计将捕获在相邻帧中编码的时空信息和运动（对于概念图参见图 1.1）。 实际上，向网络提供补充信息（例如光流）以促进训练是重要的。经验上，Ji 等人[11]表明，3D 卷积网络优于基于二维框架的相应结构，并具有明显的差距。

一般来说，三维卷积网络具有非常严格的时间结构。 网络接受预定数量的帧作为输入（例如，在 Ji 等人[11]中，输入仅由 7 个帧组成）。 尽管具有固定的空间维度在某种程度上是可以提高繁华能力的（池化往往会提供不同规模的健壮性），但是为什么在时间域上应该进行类似的假设还不清楚。 更不明确的是时间跨度的正确选择，因为不同动作中的宏运动具有不同的速度，因此具有不同的跨度。

为了回答如何将时间信息馈送到卷积网络，研究了各种融合方案。 Ng 等人 [10]探讨了时域池化，并得出结论认为，时间域中的最大池化是最优的。 Karpathy 等人 [12]提出了慢融合的概念来增加卷积网络的时间意识。 在缓慢融合中，卷积网络接受视频中的几个但是连续的部分，并通过同一组层来处理它们以在时域上产生响应。 然后通过完全连接的层来处理这些响应以产生视频描述符（详情参见图 1.2）。

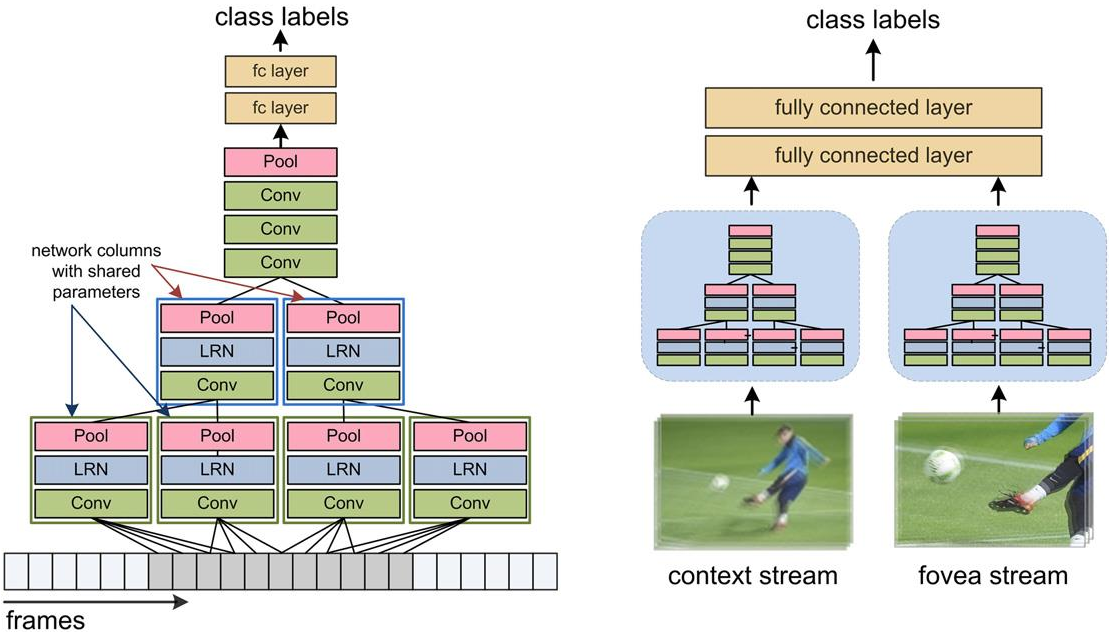

图 1.2 Karpathy 等人的网络结构[12]。 用绿色表示，红色和蓝色分别表示归一化，空间池化和卷积层。

其他形式的融合包括早期融合（例如 3D 卷积网络[11]），其中网络被馈送一组相邻帧和后期融合，其中逐帧特征融合在最后一层[12]。 Karpathy 等人 [12]也表明，使用两个独立网络的多分辨率方法不仅提高了准确性，而且减少了学习参数的数量。 这是由于网络的每个支路（即图 1.2 中的中央区域和上下文流）接受较小的输入的事实。 我们注意到，中央接收帧的中心区域以利用许多视频中存在的照相机偏差，因为感兴趣的对象经常占据中心区域。

与使用 VGG [2]和 Decaf [13]网络作为图像的一般描述符相似，Tran 等人[14]试图找到基于 3D 卷积网络的通用视频描述符。特征提取网络在 Sports-1 M [12]数据集上进行训练。经验上，作者表明，具有 3×3×3 均匀卷积核（每层恒定深度）的网络比改变卷积核上的时间深度更好。 包含 3D 池化层可以获得时间范围的灵活性。 然后通过平均 C3D 网络的第一个全连接层的输出来获得名为 C3D 的通用描述符。

Varol 等人 [15]探索在输入层的更长时间持续时间上执行 3D 卷积的效果。 通过扩展输入的时间深度以及在输入处结合具有不同时间注意力的网络的决定来观察改进。

将空间卷积核扩展到 3D 卷积核虽然被主流化，但不可避免地会增加网络参数的数量。在改善 3D 卷积核的不利影响方面，Sun 等人 [16]建议将 3D 卷积核分解成 2D 和 1D 卷积核的组合。随着参数的减少，他们获得与 Simonyan 和 Zisserman [17]相当的性能，而在训练时几个视频数据集之间没有任何知识转移。

为了利用时间信息，一些研究诉诸于反馈结构的使用。 Baccouche 等人的作品 [4]和 Donahue 等人。 [18]通过级联的卷积网络和一类递归神经网络（RNN）称为长期短期记忆（LSTM）[19]网络来解决动作识别问题[20]。 正如“循环”这个词所表明的那样，一个 RNN（见图 1.3）使用反馈回路对动力学进行建模。 RNN 块的典型形式接受外部信号 $x^{(t)} \in R_d$，并基于其隐藏状态 $h^{(t)} \in R_r$ 产生输出 $z^{(t)} \in R_m$ 。

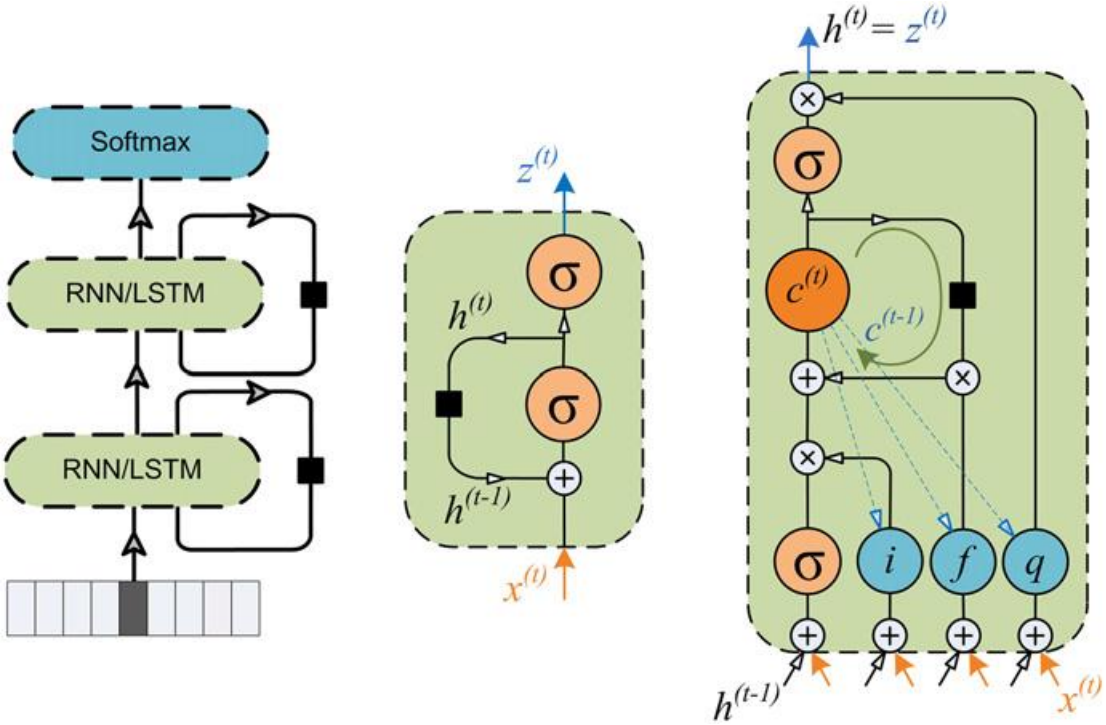

图 1.3 左图：2 层 RNN / LSTM 网络的循环结构。 中心：复制线性动力系统的 RNN 单元结构。 右图：包含附加门控的 LSTM 单元。 时间延迟以黑色方块表示。

一般来说，由于梯度消失（或爆炸）的问题，对 RNN 进行训练并不容易[5]。 为了讨论，假设 RNN 小区的递归表达式具有 $h^{(t)} = w_h h^{(t-1)}$ 的形式，其中 x，h，z∈R。这个递归形式可以展开为 $h^{(t)} = w_h h^{(t-1)} = w_h w_h h^{(t-2)} = \cdots = w_h^t h^{(0)}$。 因此，网络要么学习短期依赖关系（如果 $w_h < 1$）或非常长期依赖关系（如果 $w_h > 1$），这是不可取的[5]。 LSTM 单元（如图 1.3 所示）通过控制门限制 RNN 单元的状态和输出来解决这个问题。

为了对行为进行分类，Baccouche 等人 [4]建议给 LSTM 网络提供从三维卷积网络提取的特征。两个网络，即三维卷积网络和 LSTM 网络分别进行训练。 也就是说，首先使用注释的动作数据来训练 3D 卷积网络。一旦获得了三维卷积网络，卷积特征就被用来训练 LSTM 网络（网络结构见图 1.4）。

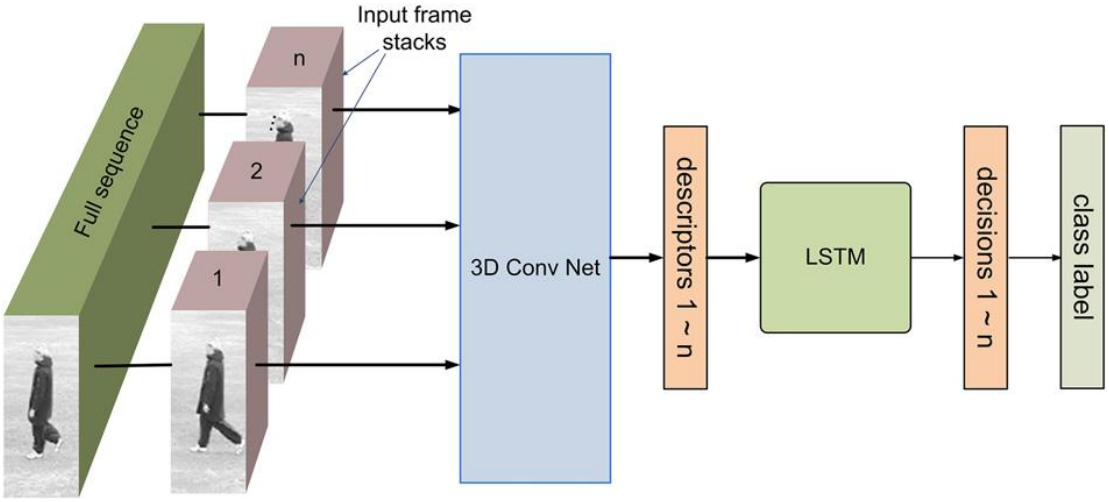

图 1.4 Baccouche 等[4]的网络结构。

Donahue 等人提出了另一种基于 LSTM 的体系结构。 [18]利用复合网络上的端到端训练，如图 1.5 所示。名为长期递归卷积网络（LRCN）的最终结构不仅在识别动作方面而且在图片和视频的字幕标注任务中也获得了成功。 通过端到端学习和 CNN-LSTM 卷积，时空接收卷积核参数以数据驱动的方式进行计算。

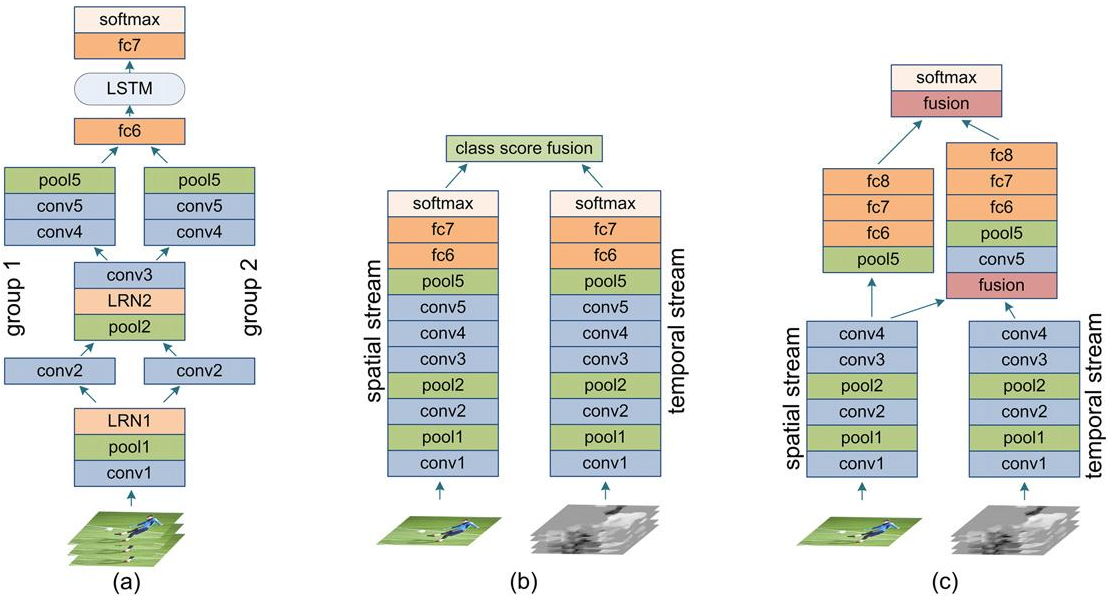

图 1.5（a）Donahue 等人的 LRCN 网络结构[18]。 一个组是一组卷积滤波器，仅对来自上一层的一组特征图进行操作。为了清楚起见，我们用不同的卷积块表示每个组。（b）Simonyan 和 Zisserman [17]将 RGB 流和堆叠光流帧作为输入的双流网络。（c）

Feichtenhofer 等人的双流融合网络的一个例子[29]。

1.2 多流神经网络

在视觉感知中，我们视觉皮层的腹侧流处理对象属性，例如外观，颜色和身份。 物体的运动及其位置通过 Dosaral Stream [21]单独处理。 一类深度神经网络被设计为将基于外观的信息与运动相关的信息分开来识别人体行为[17]。
Simonyan 和 Zisserman [17]介绍了其中的第一个用于行为识别的多流深度卷积网络，其中有两个并行网络用于动作识别（见图 1.5）。 所谓的空间流网络接受原始视频帧，而时间流网络将光流场作为输入。 Simonyan 和 Zisserman [17]提出了以下观点：

1. 预训练空间流网络。 从头开始训练空间流网络并不是最佳实践。 根据经验，在 ILSVRC-2012 图像数据集[22]上微调预训练网络会导致更高的准确性。

2. 早期的时间流网络融合。 在时间流网络的输入处堆叠光流场（即，前期融合）是有益的。

3. 时间流网络的多任务学习。 时间流网络需要纯粹从可用视频数据进行训练。 对于深度网络中的中小型数据集，这被认为是具有挑战性的。 为了避开这个困难，时间流网络被修改为具有多于一个的分类层。 每个分类层在特定数据集上运行（例如，一个在 HMDB-51 上运行，另一个在 UCF-37 数据集上运行），并且仅响应来自相应数据集的视频。 这种架构是多任务学习的实现，旨在学习一种表示，它不仅适用于所讨论的任务，而且适用于其他任务。

使用 softmax 分数将两条流融合在一起。 Feichtenhofer 等[29]的工作表明，中间层的融合不仅提高了性能，而且显着减少了参数的数量（参见图 1.5）。它证明了当最后的卷积层之后进行融合时达到了最好的精度。 有趣的是，在卷积层之后进行融合将消除两个流中昂贵的完全连接层的要求。 与原始网络 Simonyan 和 Zisserman [17]相比，融合网络的性能与仅使用一半参数相当。

两个流网络的扩展包括 Wang 等[23]的工作，其中用双向流网络的卷积特征映射跟踪的稠密轨迹[18]使用 Fisher 向量聚合，[31]使用音频的第三个流信号被添加到网络。

光流帧是两个流网络中唯一使用的运动相关信息。 这将引发两个流网络是否可以捕捉细微而长期的运动动态（这种运动不能用光流模拟）的问题。 深层架构和手工解决方案的有效结合所带来的改进表明，在深层解决方案中，行动仍然遥遥无期[29]。

1.3 深度生成网络

考虑到网络上广泛且不断增加的视频，设计需要很少监督或根本不监督的深层模型的潜在回报超乎想象。一个好的生成模型就是可以准确地了解数据的底层分布的模型。序列分析的生成模型[32,4]主要用于预测序列的未来。也就是说，给定序列{x_1，x_2，...，x_t}，可以认为学习了一个模型来预测它的未来（例如下一个实例 x_{t+1}）。然而，如果序列的内容和动态（例如，运动基元）可以很好地被模型捕获，则可以实现准确的预测。深层生成架构[33,19,34]旨在实现这一目标，即从无监督问题中的时态数据中学习。在视频分析中，注释数据成本高昂，无监督技术优于监督技术。设想可能的潜力，在这一部分，我们将回顾深层生成架构的着名例子，而不会将自己放在直接应用于行动识别的研究上。

1.3.1 Dynencoder

Yan 等人[35]的启发是受 LDS 建模的启发[8]。Dynencoder 是一类深度自动编码器，可捕捉视频动态。 在最基本的形式中，Dynencoder 由三层构成。第一层将输入 x_t 映射到隐藏状态 h_t。 第二层是预测层，其使用当前的预测层（即，h_t）预测下一个隐藏状态 h_{t+1}。 最后一层是从预测隐藏状态 h_{t+1} 到产生估计输入帧 x_{t+1} 的映射。为了减少训练的复杂性，网络的参数分两个阶段学习。 在训练阶段，每一层都是分开训练的。 预训练完成后，进行端到端的微调被执行。

Dynencoder 被证明能够成功地合成动态纹理。 人们可以将 Dynencoder 视为表示视频时空信息的紧凑方式。 因此，给定 Dynencoder 的视频的重建误差可以用作分类的均值。

1.3.2 LSTM 自动编码模型

动作识别的生成模型预计会发现长期的线索，使 LSTM 细胞的深层模型成为自然选择。为此，Srivastava 等人 [32]介绍了如图 1.6 所示的 LSTM 自编码器模型.LSTM 自编码器由两个 RNN 组成，即编码器 LSTM 和解码器 LSTM。该编码器 LSTM 接受一个序列（作为输入）并学习相应的紧凑表示。 编码器 LSTM 的状态包含序列的外观和动态。 如此，序列的紧凑表示被选择为编码器 LSTM 的状态。 解码器 LSTM 接收学习表示以重建输入序列。 有关更多详细信息，请参见图 1.6。

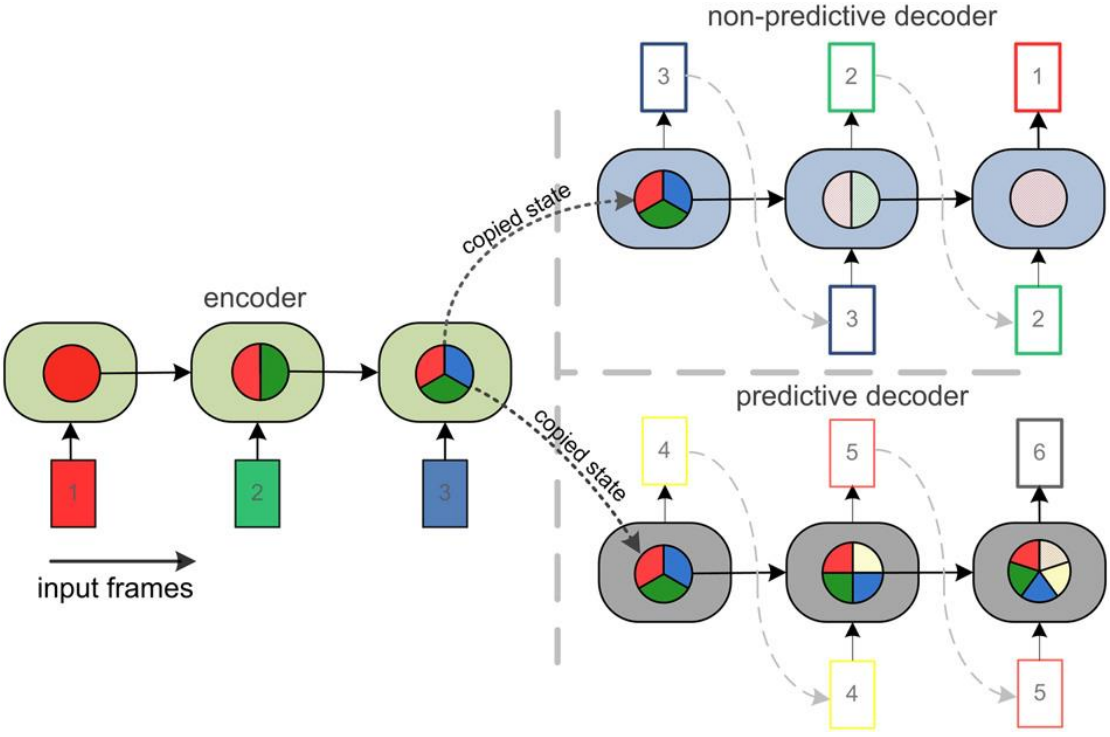

图 1.6 Srivastava 等人的复合生成 LSTM 模型[32]。 编码器 LSTM 的内部状态（由内部的圆圈表示）捕获压缩版本（例如，帧 1,2 和 3）。 之后的状态被复制到两个解码器模型中，这是重建和预测的。 重建解码器尝试以相反的顺序重建原始帧。 对预测模型进行预测未来帧 4,5 和 6 的训练。状态标记上的颜色表示存在来自特定帧的信息。

LSTM 自编码器也可用于预测序列的未来。 在实践中，一个既能重建输入序列又能预测其未来的复合模型能够提供最准确的响应。

1.3.3 对抗模型

为了避免培养深度生成模型的各种困难，Goodfellow 等人 [33]引入敌对网络，其中一个生成模型与一个被称为敌手的区分模型竞争。判别模型学习确定样本是来自生成模型还是数据本身。在训练过程中，生成模型学习生成与原始数据有更多相似性的样本，而对抗模型则改进对给定样本是否真实的判断或不。 Mathieu 等人 [36]采用对抗性方法来训练用于视频预测的多尺度卷积网络。 他们利用对抗性训练来使卷积网络避免池化层。 他们还提供了关于池化对生成模型的优势的讨论。

1.4 时域一致性网络

在结束本部分之前，我们希望将时间一致性的概念纳入观点。 时间一致性是弱监督的一种形式，并表明连续的视频帧在语义上和动态上都是相关的(即突发运动的可能性较小)。对于行动，甚至有更强大的空间和时间线索[37]。 如果一个序列的帧按照正确的时间顺序，则该序列称为一致序列。 如果模型分别由有序和无序序列馈送为正样本和负样本，则可以通过深度模型学习时间相关性。这个概念已被 Goroshin 等人使用。 [38] Wang 和 Gupta [26]从未标记的视频中学习鲁棒性的视觉表现。

Misra 等[84]研究了时间一致性如何用于训练动作识别和姿态估计的深度模型。 具体来说，一个连体网络[15,74,135]（见图 1.7）用元组进行训练，以确定给定序列是否一致。实际上，已经表明：

与其他受监督的预训练方法相比，例如，ImageNet [22]通过元组学习给予了更多的关注人的姿势。

在具有丰富运动的帧中选择元组将避免正元组和负元组之间的不确定性。

与从零开始训练的网络相比，基于时间一致性的预训练网络有可能提高准确性。

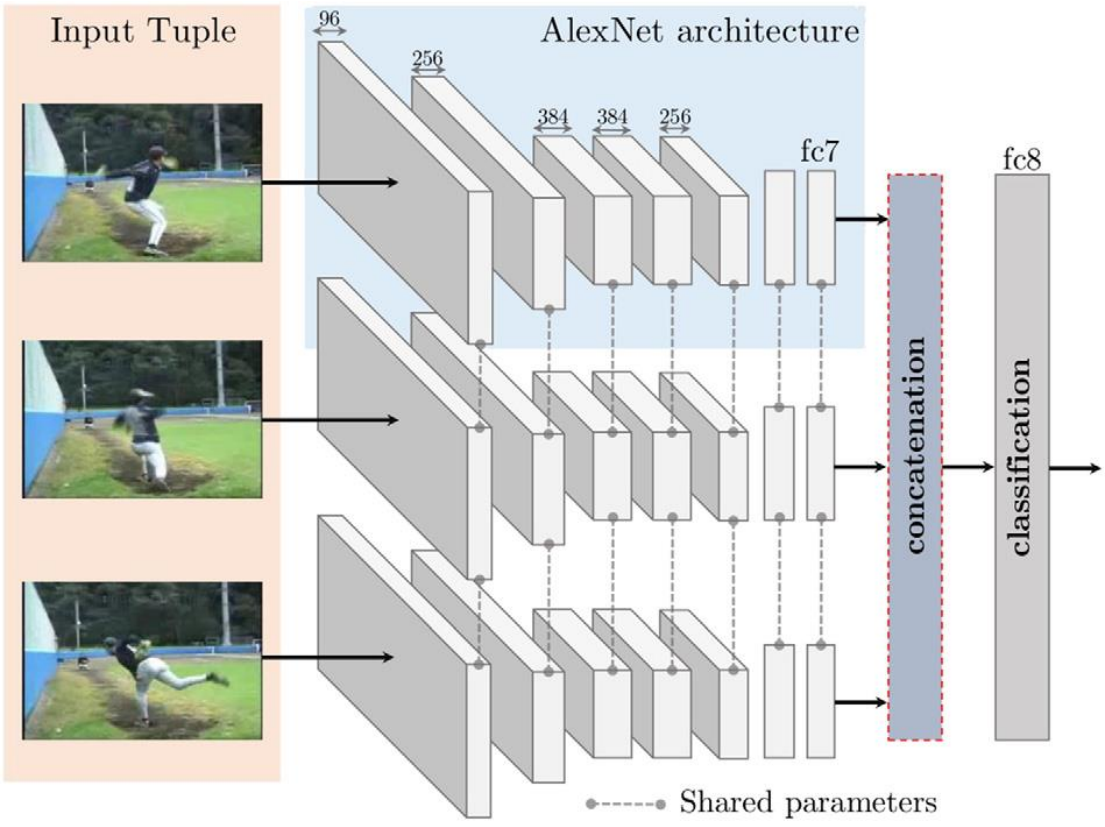

图 1.7 Misra 等人使用的 Siamese Triplet 网络[84]。预计每个网络都会捕捉到动作和姿势。

我们注意到时间一致性并不总是一个强有力的假设依靠。 例如，在运动事件期间（例如，在 SPORTS-1M 数据中）显示的突然的场景变化（诸如广告）可容易地违反时间相干性[71]。

对时间一致性的另一个相关研究是 Wang 等人[25]的工作，其中一个动作分为两个阶段进行分类。 更具体地说，将具有帧{x_1，x_2，…，x_n}的视频分成前置条件集 X_p = {x_1，x_2，…，x_p}和效果集 X_e = {x_e，x_{e+1}，…，x_n}。 两组的基数都是由深度模型学习的。 然后通过所需的转换来识别动作将从 X_p 提取的高级描述符映射到从 X_e 提取的高级描述符。 特别是，使用 Siamese 网络学习高级描述符和转换（详情参见图 1.8）。

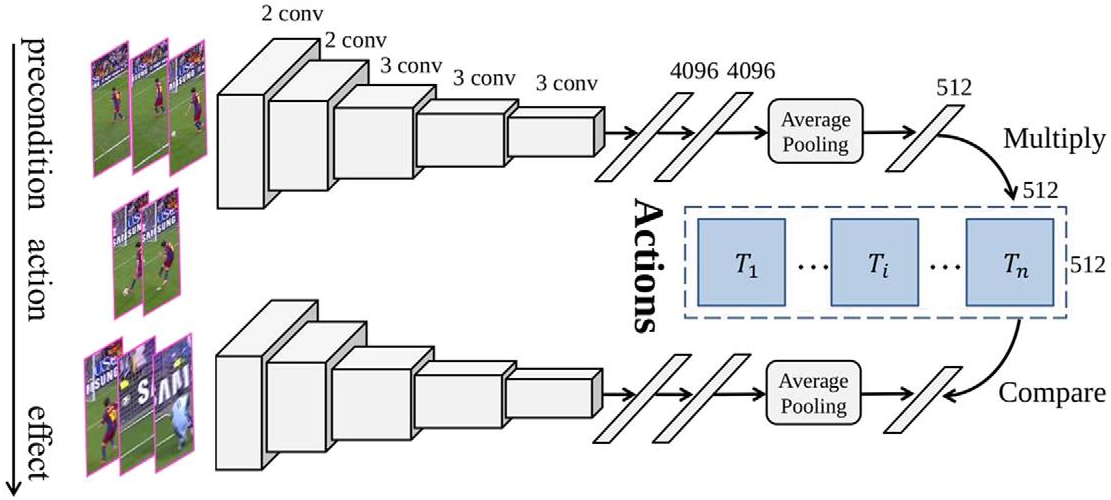

图 1.8 平行卷积结构用于提取前置和后置特征。

排序池化[31]是捕获序列时间演化的有效解决方案。 在其原始形式中，视频表示（通过排名）和行动分类的学习是分开进行的。这是由于这样一个事实：与其他池化操作（如最大池化）不同，用于等级池化操作的封闭表单解决方案并不容易获得。最近，费尔南多和古尔德[32]提出了一种端到端的学习方案，可以学习池式操作和带有反向传播的分类器。 一个相关的工作，虽然不是一个基于深度学习的解决方案，但是层次化的等级池[30]旨在通过迭代应用等级池化操作来对视频中的多个动态粒度进行编码。

为了完整起见，我们通过讨论 Ranzato 等人的工作来总结本节。[103]。 Ranzato 等人。[103]指出，通过 RNN 进行语言建模的成功是离散信息空间的结果。 基于此，他们通过用代表性的图像块集合来量化它们，为视频帧引入了离散结构。 毫不奇怪，自然视频似乎缺乏动态词序列，这暗示了为什么语言模型优于他们的视频对应物。根据他们的观察，Ranzato 等人 [103]建议训练一个循环卷积网络来预测长序列可能会导致更好的视频建模鲁棒性。

2 相关实验结果

在表 2.1 中，我们提供了 31 个必须知道的方法以及七个具有挑战性的行动数据集的准确性的完整列表。准确度直接来自原始作品。 我们不是单独考虑每个案例，而是选择对各种解决方案进行高层比较。

文献	方法	类型	数据集						
			HMDB51	UCF37	UCF50	UCF-Sportsa	Hollyw-ood2	Olympic Sportsa	Sports-1 M
Wang et al.	Dense Traj (Traj + HoG + HoF + MBH)	R				88.2	58.3		
Kliper-Gross et al.	Motion interchange patterns	R	29.2		68.5				
	General		26.9						
Sadanand and Corso	Video wise	R			76.4				
	Group wise				57.9				
Oneata et al.	MBH + SIFT + Sqrt + L2 normalization	R	54.8		90		63.3	82.1	
	Without human detector	R	55.9		90.5		63	90.2	
Wang and Schmid	With human detector		57.2		91.2		64.3	91.1	
Jain et al.	Traj + HoG + HoF + MBH + DCS on w-flow	R	52.1				62.5		
Peng et al.	Stacked FVs + FV	R	66.8						
Peng et al.	Hybrid-BoW	R	61.1	87.9	92.3				
Kantorov and Laptev	MPEG-flow: VLAD encodings of	R	46.3						
Gaidon et al.	SDT tree ATEP	R	41.3				54.4	85.5	
Simonyan and Zisserman	Two-stream (CNN-M-2048)	D	59.4	88.0					
	Transfer learning on Sports-1 M			65.4					
Karpathy et al.	Clip hit @ 1 – slow fusion	D							41.9
	Video hit @ 1 – slow fusion								60.9
Sun et al.	Factorized spatiotemporal conv. nets	D	59.1	88.1					
	Two-stream (ClarifaiNet)			88.0					
Wang et al	Two-stream (GoogLeNet)	D		89. 3					
	Two-stream (VGG-16)			91.4					
Wang et al.	TDD + Wang and Schmid	F	65.9	91.5					
	TDD (only)	F	63.2	90.3					
	Conv pooling hit @ 1 (best)								72.4
Ng et al.	LSTM hit @ 1 (best)	D							73.1
	Conv pooling (image + opt flow)			88.2					
	LSTM (image + opt flow)			88.6					
Fernando et al.	Rank pooling	R	63.7				73.7		
Donahue et al	LRCN-weighted average of RBG + flow	R		82.9					
Wu et al.	Adaptive multi-stream fusion	D		92.6					
Jiang et al.	TrajShape + TrajMF	R	48.4	78.5			55.2	80.6	
	TrajShape + TrajMF + Wang and Schmid		57.3	87.2			65.4	91	
Lan et al.	Multi-skip feat. stacking	R	65.1	89.1	94.4		68.0	91.4	
Hoai and Zisserman	Proposed SSD + RCS	R	62.2				72.7		
Tran et al.	C3D on SVM	D		85.2					
	C3D + Wang and Schmid [141] on SVM	F		90.4					
Misra et al.	ImageNet pretrain + tuple verification	D	29.9						
	HMDB + UCF101 labels only		30.6						
Wang et al.	Proposed only (RBG + opt flow network.	D	62	92.4					
Fernando and Gould	End to end rank-pooling	D				87	40.6		
Fernando et al.	Hierarchical rank-pooling (CNN features)	D	47.5	78.8			56.8		
	Hierarchical RP on CNN + Fernando et al.	F	65.0	90.7			74.1		
Li et al.	VLAD^3	F		84.7				90.8	
	VLAD^3 + Wang and Schmid	F		92.2				96.6	
Varol et al.	$LTC_{flow+RGB}$	D	64.8	91.7					
	$LTC_{flow+RGB}$ + Wang and Schmid	F	67.2	92.7					
Feichtenhofer et al.	Two stream fusion (VGG-16)	D	65.4	92.5					
	Two stream fusion (VGG-16) + Wang and Schmid	F	69.2	93.5					
de Souza et al.	Hybrid fusion of Wang and Schmid [141] + Deep-nets	F	70.4	92.5			72.6		

表 2.1 行为识别技术的准确性（数字是以百分比给出的真实识别准确度）。 列类型表示方法是纯粹的基于深网的（D），基于表示的（R）还是融合的解决方案（F）。

3 总结

尽管与静态图像分析有相似之处，但视频数据分析要复杂得多。 一个成功的视频分析解决方案不仅需要克服规模，类内多样性和噪音等变化，还必须分析视频中的运动线索。

由于其广泛的应用和由关节式身体运动产生的运动模式的复杂性，人类行为识别可被视为视频分析问题的关键。 在这次调查中，我们调查了现有行动识别解决方案的几个方面。

我们首先审查方法，基于手工表示，然后专注于从深层架构中受益的解决方案。 我们对这两种流行的研究路线进行了比较分析。

参考文献

[1] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 10 (11) (1998) 1338–238.

[2] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: delving deep into convolutional nets, British Machine Vision Conference, 202.

[3] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Proc. Advances in Neural Information Processing Systems (NIPS), 2012. pp. 1097–225.

[4] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, Proc. Advances in Neural Information Processing Systems (NIPS), 202.pp. 3104–3112.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. pp. 1–9.

[6] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition,Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016.

[7] R.K. Srivastava, K. Greff, J. Schmidhuber, Training very deep networks,Proc. Advances in Neural Information Processing Systems (NIPS), 2015. pp.2337–2385.

[8] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, Towards Good Practices for Very Deep Two-Stream ConvNets, 2015. CoRR abs/1507.01859

[9] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, Proc. European Conference on Computer Vision (ECCV), 202. pp. 818–833.

[10] J.Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G.Toderici, Beyond short snippets: deep networks for video classification, Proc.IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. pp.6–4702.

[11] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 131–231.

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Largescale video classification with convolutional neural networks, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 202. pp.1725–1732.

[13] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, DeCAF: a deep convolutional activation feature for generic visual recognition, International Conference in Machine Learning (ICML), 202.

[14] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, Proc. Int. Conference on Computer Vision (ICCV), 2015. pp. 1989–1997.

[15] G. Varol, I. Laptev, C. Schmid, Long-term temporal convolutions for action recognition, 2016. arXiv:1604.01994

[16] L. Sun, K. Jia, D.Y. Yeung, B.E. Shi, Human action recognition using factorized spatio-temporal convolutional networks, Proc. Int. Conference on Computer Vision (ICCV), 2015. pp.

4597–4605.
[17] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Proc. Advances in Neural Information Processing Systems (NIPS), 202. pp. 128–576.
[18] J. Donahue, L.A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K.Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. pp. 235–2634.
[19] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1360.
[20] A.J. Robinson, F. Fallside, Static and dynamic error propagation networks with application to speech coding, Proc. Advances in Neural Information Processing Systems (NIPS), 1988. pp. 632–641.
[21] M.A. Goodale, A.D. Milner, 1 2 Separate Visual Pathways for Perception and Action. Essential Sources in the Scientific Study of Consciousness, 2003, 175.
[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A.Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 17 (3) (2015) 181–181.
[29] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. pp. 1933–1941.
[23] L. Wang, Y. Qiao, X. Tang, Action recognition with trajectory-pooled deep-convolutional descriptors, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. pp. 4305–432.
[31] Z. Wu, Y. Jiang, X. Wang, H. Ye, X. Xue, J. Wang, Fusing Multi-Stream Deep Networks for Video Classification, CoRR. 2015.
[32] N. Srivastava, E. Mansimov, R. Salakhutdinov, Unsupervised Learning of Video Representations Using LSTMs, CoRR. 2015.
[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair,A. Courville, Y. Bengio, Generative adversarial nets, Proc. Advances in Neural Information Processing Systems (NIPS), 202. pp. 2672–2680.
[34] P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, Extracting and composing robust features with denoising autoencoders, Proc. Int. Conference on Machine Learning (ICML), 2008. pp. 1096–223.
[35] X. Yan, H. Chang, S. Shan, X. Chen, Modeling video dynamics with deep dynencoder, Proc. European Conference on Computer Vision (ECCV), 202. pp. 185–230.
[36] M. Mathieu, C. Couprie, Y. LeCun, Deep Multi-Scale Video Prediction Beyond Mean Square Error, CoRR. 2015.
[37] H. Rahmani, A. Mian, Learning a non-linear knowledge transfer model for cross-view action recognition, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015. pp. 858–20.
[38] R. Goroshin, J. Bruna, J. Tompson, D. Eigen, Y. LeCun, Unsupervised learning of spatiotemporally coherent metrics, Proc. Int. Conference on Computer Vision (ICCV), 2015. pp. 4010–4093.
[39] X. Wang, A. Gupta, Unsupervised learning of visual representations using videos, Proc.

Int. Conference on Computer Vision (ICCV), 2015. pp. 2794–2802.