DOI: 10.11992/tis.201707032
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180121.1457.002.html

基于深度学习的视频预测研究综述

莫凌飞, 蒋红亮, 李煊鹏
（东南大学 仪器科学与工程学院, 江苏 南京 210096）

摘　要: 近年来, 深度学习算法在众多有监督学习问题上取得了卓越的成果, 其在精度、效率和智能化等方面的性能远超传统机器学习算法, 部分甚至超越了人类水平。当前, 深度学习研究者的研究兴趣逐渐从监督学习转移到强化学习、半监督学习以及无监督学习领域。视频预测算法, 因其可以利用海量无标注自然数据去学习视频的内在表征, 且在机器人决策、无人驾驶和视频理解等领域具有广泛的应用价值, 近两年来得到快速发展。本文论述了视频预测算法的发展背景和深度学习的发展历史, 简要介绍了人体动作、物体运动和移动轨迹的预测, 重点介绍了基于深度学习的视频预测的主流方法和模型, 最后总结了当前该领域存在的问题和发展前景。

关键词: 视频预测; 深度学习; 无监督学习; 运动预测; 动作识别; 卷积神经网络; 递归神经网络; 自编码器
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2018)01−0085−12

中文引用格式: 莫凌飞, 蒋红亮, 李煊鹏. 基于深度学习的视频预测研究综述[J]. 智能系统学报, 2018, 13(1): 85–96.
英文引用格式: MO Lingfei, JIANG Hongliang, LI Xuanpeng. Review of deep learning-based video prediction[J]. CAAI transactions on intelligent systems, 2018, 13(1): 85–96.

Review of deep learning-based video prediction

MO Lingfei, JIANG Hongliang, LI Xuanpeng
(College of Instrument Science and Engineering, Southeast University, Nanjing 210096, China)

Abstract: In recent years, deep learning algorithms have made significant achievements on various supervised learning problems, with their accuracy, efficiency, and intelligence outperforming traditional machine learning algorithms, in some instances even beyond human capability. Currently, deep learning researchers are gradually turning their interests from supervised learning to the areas of reinforcement learning, weakly supervised learning, and unsupervised learning. Video prediction algorithms have developed rapidly in the last two years due to its capability of using a large amount of unlabeled and naturalistic data to construct the forthcoming video as well as its widespread application value in decision making, autonomous driving, video comprehension, and other fields. In this paper, we review the development background of the video prediction algorithms and the history of deep learning. Then, we briefly introduce the human activity, object movement, and trajectory prediction algorithms, with a focus on mainstream video prediction methods that are based on deep learning. We summarize current problems related to this research and consider the future prospects of this field.
Keywords: video prediction; deep learning; unsupervised learning; motion prediction; action recognition; convolution neural network; recurrent neural network; auto encoder

“我们缺乏的一个关键要素是预测 (或无监督) 学习: 机器具有模拟环境, 预测未来的可能性, 以及通过观察和参与理解世界如何运作的能力。”[1]

近年来, 深度学习在学术界和工业界得到了广泛的发展和应用, 其在计算机视觉[2-6]、语音识别[7]、自然语言处理[8-9]以及游戏策略[10-11]等众多领域取得丰硕成果, 在某些领域甚至取得了超越人类的表现。但当前的深度学习算法模型大部分都是以有监督的方式训练, 模型严重依赖于大量的标注数据和长时间的训练。以知名的 ImageNet 数据集[12]为例, 其包含 1 500 万张人工标注的图片, 超过 2.2 万个类

收稿日期: 2017−07−19.　**网络出版日期:** 2018−01−22.
基金项目: 国家十二五科技支撑计划重点项目 (2015BAG09B01).
通信作者: 莫凌飞. E-mail: lfmo@seu.edu.cn.

别, 创建和标注一个如此大规模的数据集需要耗费许多人数月的时间才能完成。另外, 依赖大量的标记数据来获取概念和知识与人类的学习机制不符, 人类依赖很少的样本就可以获取一个新的概念。当儿童第一次观察到“猫”并被告知这种动物是“猫”以后, 儿童并不需要长期被重复告知这是“猫”, 但监督学习的方式需要大量的样本以及多次重复训练, 才能掌握“猫”的概念。以类似人类的方式, 通过有限样本或者无监督的方式获取知识和表征, 成为当前人工智能领域的热点研究问题。

另外, 人类与其他动物的一个重要区别是人类有很强的预测能力。尽管一些动物也有一些预测能力, 例如在围捕猎物、躲避天敌和预测天气变化上等; 但人类显然有更强的推理和预测能力, 例如, 人类驾驶汽车时可以推理其他汽车的运行轨迹, 提前决策。当前计算机视觉领域的研究, 也逐渐开始借鉴人类这种“预测编码”能力。

在这种背景下, 视频预测因其可以用海量的无标注自然视频数据来训练, 而且具有广泛的应用场景, 成为了当前深度学习研究领域的一个热点研究方向, 并且已经取得了一定的研究成果。

给出一个视频序列, 预测未来视频, 这需要构建一个可以精准建模视频内容和动态变化的内部表征模型, 这也是视频预测被视为无监督表征学习的一个很有前景的研究方向的原因。视频预测模型学习到的表征可以迁移到监督学习任务中。例如, 文献[13]通过实验证明, 通过无监督视频预测模型学习到的表征可以在动作识别数据集上提升分类结果, 因此建模视频动态是一种有效的无监督表征学习方法。另外, 在视频中推断未来的场景可以使机器人、自动驾驶汽车和无人机提前决策, 因此有广泛的应用价值。

1　深度学习概述

机器学习算法是一种可以自动从数据中发现规律, 并利用此规律对未知数据进行预测的算法, 机器学习在数据挖掘、计算机视觉、自然语言处理、搜索、推荐系统以及策略游戏等众多领域得到了广泛的应用, 取得了突出的成果。然而, 自然界的原始数据, 例如图像、视频和传感器测量数据等一般具有高维度、高复杂性和高冗余性的特点, 人工提取特征需要依赖专家知识, 费时费力且提取到的特征通常不太好。而传统机器学习算法往往依赖人工提取特征, 导致实际的机器学习问题退化为数据预处理和特征工程[2], 成为机器学习应用和发展的一大障碍。

深度学习是人工神经网络 (artificial neural network, ANN) 的一个分支。最早的人工神经网络研究可以追溯到 Mcculloch 和 Pitts[14]在 1943 年提出的阈值逻辑单元, 他们从原理上证明了人工神经网络可以计算任何算术和逻辑函数。随后 Hebb 学习规则[15]、感知机[16]、反向传播算法[17]等概念先后被提出, 并得到了一定的应用, 例如手写数字识别[18]和语音识别[7]。然而, 由于当时人们对神经网络认识有限, 计算机的计算能力也有限, 神经网络并未得到过多关注。2006 年, Hinton 等提出以无监督限制玻尔兹曼机 (restricted Boltzmann machine, RBM) 进行逐层预训练的方法来高效地训练多层神经网络[19], 深度学习的概念开始进入公众视野。2012 年 Krizhevsky 等使用深度卷积神经网络 (convolutional neural network, CNN)[18]构建的 AlexNet 模型[3]以绝对优势赢得了 ImageNet 大规模图像识别竞赛 (ILSVRC2012) 的冠军, AlexNet 的成功成为了计算机视觉发展史上的转折点, 自此深度学习得到了飞速发展。卷积神经网络 (convolutional neural network, CNN)、递归神经网络 (recurrent neural network, RNN)[20]、自编码网络 (auto encoder)[21]和生成对抗网络 (generative adversarial networks, GANs)[22]及其各种变种得到了广泛的发展和应用。

表征学习 (或特征学习, representation learning)[23]旨在利用机器自动从原始高维数据中获得可以被机器学习算法高效利用的特征[21]。深度学习可看作一种通过简单、非线性映射方式获取多层特征的表征学习方法, 它把原始输入数据通过逐层映射, 转变为高阶的、更为抽象的特征。以分类问题为例, 高层的表征放大了那些更有区分度的特征, 而抑制了那些无关变量。深度学习被证明非常擅长发现高维度数据中的复杂特征, 因此在科学界和工业界得到广泛应用, 并打破了图像识别、语音识别和机器翻译的记录。

2　深度学习主要模型

近些年来, 有越来越多的深度学习模型被提出, 其中最基础、最重要的模型主要有卷积神经网络、递归神经网络、自编码器以及生成对抗网络, 这几种模型构成了视频预测模型的基础, 下面我们简要介绍这 4 种主流模型。

2.1　卷积神经网络

卷积神经网络是前馈神经网络的一种, 这种神经元连接模式受动物视觉皮层检测光学信号原理的启发[24]。1980 年 Fukushima 等[25]提出了 CNN 的前身——NeoCognitron, 20 世纪 90 年代, Lecun 等[18]

发表论文, 确立了 CNN 的现代结构, 这是一种多层的人工神经网络, 取名为 LeNet-5。自 2012 年起, 研究人员又不断提出更深、性能更强的卷积神经网络模型: AlexNet[3]、VGGNet[5]和 ResNet[6]等。卷积神经网络一般是由多个卷积层和全连接层组成, 卷积操作、局部连接性和权值共享是卷积神经网络最显著的特点。卷积神经网络通常用来处理 2-D 结构的数据, 其在图像领域和语音识别上都得到了广泛的应用。

2010 年, Zeiler 等[26]首次提出了反卷积 (卷积转置或小数步进卷积, Deconvolution) 的概念, 用于卷积神经网络的特征可视化以及图像无监督特征学习。反卷积网络被越来越多的模型所采用, 例如图像语义分割[27]、生成模型[28]等。另外, 为处理序列图像, Ji 等[29]使用 3-D 卷积去提取数据的空间和时间特征, 从而可以使卷积神经网络能很好地处理序列信息, 3-D 卷积在人体动作识别等领域取得了显著的结果。

2.2　递归神经网络

递归神经网络[20]是一种处理序列数据的神经网络, 它把状态在自身网络中循环传递, 能够处理任意长度的序列, 递归神经网络比前馈神经网络更加符合生物神经网络的结构。

因为 RNN 容易受到梯度消失或者梯度爆炸的影响, Schmidhuber 等[30]在 1997 年提出了长短期记忆 (long short term memory, LSTM) 神经网络, 该模型增加了“遗忘门”和“更新门”。实验表明, LSTM 模型能有效避免梯度消失或者梯度爆炸的问题, 很好地解决了长期依赖问题。随后学者提出了很多 LSTM 模型的变体。Gers 等[31]于 2001 年提出了窥视孔 LSTM(peephole LSTM), 该模型增加了一个窥视孔连接, 意味着可以让门限层监视神经元状态。Cho 等[32]于 2014 年提出了门递归单元 (gated recurrent unit, GRU), 它组合遗忘门和输入门为一个“更新门”, 合并了神经元状态和隐层状态, 这个模型比标准的 LSTM 模型更简单。Shi 等[33]在 2015 年提出了卷积 LSTM(convolutional LSTM), 把卷积层和递归层做了很好的结合, 卷积 LSTM 与常规 LSTM 的区别是把部分矩阵乘积操作换成了卷积操作。因为卷积 LSTM 可以很好地处理图像的空间信息和时间动态信息, 它在图像生成模型和视频处理等领域得到了广泛应用。

2.3　自编码器

自编码器是一种以无监督的方式来学习数据表征的神经网络, 通常用来做数据降维[21]。自编码器通常分为编码器和解码器两部分, 编码器将数据编码为潜在变量, 解码器将潜在变量重建为原数据。

自编码器有很多变体, 例如降噪自编码器[34]、稀疏自编码器[35]、变分自编码器 (VAE)[36-37]。因为自编码器可以高效地进行数据降维, 相当一部分视频预测模型采用了自编码器架构。

2.4　生成对抗网络

Goodfellow 等[22]在 2014 年提出了生成对抗网络的概念, 其为生成模型提供了一种全新的高效训练模式, 近两年来生成对抗网络成为了机器学习领域最热门的研究方向之一。LeCun 认为“生成对抗网络是过去十年来机器学习领域最有趣的想法”, 很多 GAN 的衍生模型, 如条件 GAN(condition GAN)[38]、InfoGAN[39]、DCGAN[28]相继被提出。

生成对抗网络由一个生成器 (generator, G) 和一个判别器 (discriminator, D) 组成。生成器输入一个潜在编码, 其输出需无限逼近真实样本; 判别器的输入为真实样本和生成器的输出, 并识别出真实样本和生成样本。两个网络以零和博弈的方式交替训练, 训练鉴别器时最小化鉴别误差, 训练生成器时最大化鉴别误差, 最终目的是使鉴别器无法鉴别出生成样本和真实样本, 生成器的输出与真实样本分布一致。生成对抗网络的架构如图 1 所示。

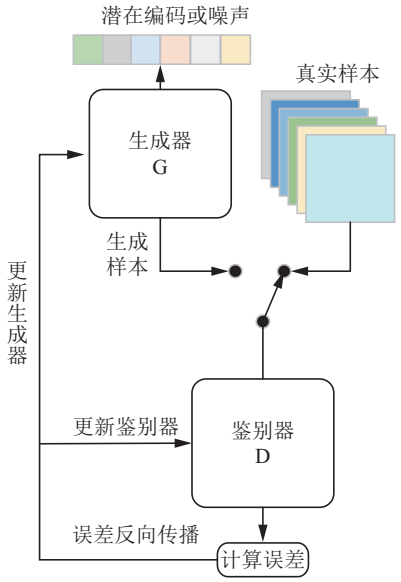


图 1　生成对抗网络架构

Fig. 1　Architecture of generative adversarial nets

生成对抗网络的目标函数可以用式 (1) 描述:

$$\min_G \max_D V(D,G) = E_{x\sim p_{\text{data}}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{1}$$

生成对抗网络在生成逼真的自然样本[28]、图像超分辨率[40]、三维建模[41]、图像风格迁移[42]和视频预测领域[43]得到了广泛应用。

3 运动预测研究

给出一张静态图片或者一段场景视频，人类不仅可以迅速地获取图像中的即时内容，还可以推断出图像中的场景动态。然而，对于计算机来说，推演出图像中的场景动态是一个比较困难的任务，因为它依赖计算机利用自然界大量难以参数化的知识来建模[44]。

在视频预测研究兴起之前，学术界比较关注的是运动预测。运动预测一般是指从静态图像或视频前几帧中推断出人体动作、物体移动轨迹等动态信息；而视频预测是从静态图片或视频前几帧中直接预测未来图像。本节我们对动作、运动和物体移动轨迹预测算法进行简要回顾。

3.1 动作和运动预测

从静态图像或有限帧视频中预测人类动作和行为是一个比较基础也比较重要的任务。在动作预测方面，研究人员主要使用统计学习方法和传统的机器学习方法来建模。Lan 等[45]和 Hoai 等[46]使用最大化边界框架来推测动作场景；Ryoo[47]把动作预测问题概率化，使用时空特征积分直方图来建模特征分布如何随时间变化；Vu 等[48]提出了一种使用动作和场景之间的关联信息，从静态场景中预测人类动作的方法；Pei 等[49]提出了一种基于随机场景感知语法的事件解析、推断事件目标和预测可信动作的算法，与 Vu 的方法类似，该方法使用事件的层次组成和子事件间的时态关系来鉴别不同事件以及预测动作；Fouhey 等[50]和 Koppula 等[51]通过使用条件随机场来建模人的可能动作从而来做未来场景的预测。

Huang 等[52]提出了一种基于双实体交互的方式来理解一个实体的动作如何影响另外一个实体的动作。本文把双实体交互模型看作一种最优控制问题，该模型使用一种基于核以及增强学习的近似软最大值函数去处理高维度的自然人体运动，另外还使用了连续代价函数的均值转移方法来平滑动作序列。

Pickup 等[53]、Lampert 等[54]和 Pintea 等[55]分别用统计流方法、向量值回归和随机森林回归算法回归物体移动方向；Pintea 等还论证了运动预测在动作识别、运动显著性检测等方面有很大的应用价值。也有学者使用深度学习进行动作预测。Vondrick 等[44]提出一种用深度回归网络的方法来学习视频表征，结合动作识别模型，能够很好地根据静态图像来推测未来动作。

3.2 物体移动轨迹预测

除人体动作和运动预测外，物体轨迹预测也具有广泛的应用价值。Kitani 等[56]提出了一种基于马尔可夫决策过程和反转最优控制的动作理解和轨迹预测方法，并在运动分析 (包括运动平滑、路径和目的地预测) 以及场景迁移学习上做了定量和定性的评估。Kitani 等[56]和 Gong 等[57]都提出用行人轨迹预测来辅助多目标追踪，并取得了高效的结果。

Kooij 等[58]提出了一种动态贝叶斯网络来做行人路径预测；Walker 等[59]使用条件变分自编码器来预测静态图像中每个像素的运动轨迹；Walker 等[60]使用光流算法来标记视频，进而训练一个光流预测模型，该模型可以预测每个像素的运动；Walker 等[61]还尝试了通过奖赏函数选择最优目标的方式建模汽车运动的轨迹。

Yuen 等[62]提出一种基于大数据的方法，通过检索大数据中与被检索图片或视频相似场景的方式来预测物体可能的位置，该方法类似于 k 近邻算法，不需要训练模型，在数据量足够大的情况下可以取得比较好的效果；Mottaghi 等[63]使用两个 CNN 和一个 RNN 来建模物体移动动态，从而预测可能移动的物体。

运动预测模型一般从建模移动物体的运动轨迹出发，能较好地预测前景物体的瞬时运动轨迹，其处理的数据维度低于视频预测，但不能预测图像的结构信息，且其学习到的特征无法迁移到有监督学习领域，因而其应用范围和价值有限。

4 视频预测模型架构

“不是我创造的，我就不能理解。”著名物理学家 Feynman 这句话背后的内涵是：通过构建验证过的概念来理解事物。在人工智能领域，可以理解为：如果一个机器能够生成高度真实的数据，那么它就发展出了对自然数据的理解能力。

视频预测是指给出一段连续视频帧 $X_1, X_2, \cdots, X_n$，构造一个模型可以精准地生成随后的帧 $X_{n+1}, X_{n+2}, \cdots, X_{n+t}$ (t 是需要预测的帧的数量)。或者，给出一段序列 $X_1, X_2, \cdots, X_N$，其中 $X_n (1 < n < N)$ 是缺失的，模型可以推断缺失的帧 (插值)。视频预测不需要额外的标注信息，因此属于无监督学习的范畴。

一般常用于评估视频质量的指标有均方误差 (mean square error, MSE)、峰值信噪比 (peak signal to noise ratio, PSNR) 和结构相似性 (structural similarity index, SSIM)。用 Y 来表示真实帧，$\hat{Y}$ 表示预测

帧, MSE、PSNR 和 SSIM 的定义如式 (2)~(4):

$$\mathrm{MSE}(Y,\hat{Y})=\frac{1}{N}\sum_{i=0}^{N}(Y_i-\hat{Y}_i)^2 \tag{2}$$

$$\mathrm{PSNR}(Y,\hat{Y})=10\lg\frac{\max^2_{\hat{Y}}}{\frac{1}{N}\sum_{i=0}^{N}(Y_i-\hat{Y}_i)^2}=10\lg\frac{\max^2_{\hat{Y}}}{\mathrm{MSE}(Y,\hat{Y})} \tag{3}$$

式中$\max^2_{\hat{Y}}$是像素的最大值, 例如 8 位的像素表示法, 其像素最大值是 255。PSNR 的值越大, 代表失真越小。

$$\mathrm{SSIM}(Y,\hat{Y})=\frac{(2\mu_Y\mu_{\hat{Y}}+c_1)(2\sigma_{Y\hat{Y}}+c_2)}{(\mu_Y^2+\mu_{\hat{Y}}^2+c_1)(\sigma_Y^2+\sigma_{\hat{Y}}^2+c_2)} \tag{4}$$

式中: μ_Y是 Y 的均值; $\mu_{\hat{Y}}$是$\hat{Y}$的均值; σ_Y^2是Y的方差; $\sigma_{\hat{Y}}^2$是$\hat{Y}$ 的方差; $\sigma_{Y\hat{Y}}$是$Y\hat{Y}$ 的协方差; $c_1=(k_1L)^2$ 和 $c_2=(k_2L)^2$是用于维持稳定的常数; L 是像素值的动态范围, k_1=0.01, k_2=0.03; SSIM 的范围是 −1~1, 值越大表示相似度越大。

视频预测为一个较新的研究领域, 目前尚未有专用于视频预测的数据集, 学者一般使用视频动作数据集进行训练和测试。表 1 给出了部分常用数据集和使用该数据集的部分文献。

视频预测模型一般基于自编码器架构、递归神经网络架构和生成对抗网络架构, 表 2 为部分基于以上 3 类架构的视频预测文献概览。下面我们按照这 3 类进行介绍。

表 1　视频预测算法常用数据集

Table 1　Common datasets used by video prediction algorithms

公开数据集	使用该数据集的视频预测文献
KTH[64]	文献[65]
Human3.6M [66]	文献[67-68]
UCF-101[69]	文献[13, 43, 70-73]
THUMOS-15[74]	文献[73]
KITTI[75]	文献[73, 76]
HMDB-51[77]	文献[13]
CityScape[78]	文献[79]

4.1　自编码器架构

自编码器因其可以进行高效的压缩编码, 因而很多视频预测模型采用自编码器来进行视频的降维和生成。基于自编码器的视频预测常用架构如图 2 所示。

表 2　视频预测算法概览

Table 2　Overview of video prediction algorithms

架构基础	算法	初始化模型帧数	单次输出帧数/有效预测帧数	备注
自编码器	文献[65]	1	1/15+	在生成第15帧处仍未模糊
	文献[73]	1	1/1	插值或预测图像较清晰, 未给出多帧预测结果
	文献[80]	1	1/1	侧重可以生成多种可信的动作图像
	文献[81]	120	1/1200+	合成纹理图片, 可以生成接近无限张图片
递归神经网络	文献[13]	10	1/10	可以同时重构、预测图像
	文献[68]	10	1/128	基于骨架结构信息
	文献[72]	2	1/30	基于运动差分
	文献[76]	1	1/9	可以在KITTI数据上高效预测视频
	文献[82]	2	1/4	图像容易模糊
	文献[83]	1	1/100+	基于动作, 能够生成大于100帧有效视频
	文献[84]	1	1/8	基于动作和快捷连接
生成对抗网络	文献[43]	1	32/32	从单幅图像上直接预测32帧图像
	文献[67]	1+1(前景+骨架图)	1/10+	使用骨架作为辅助信息
	文献[70]	4	4/8	预测图像锐利性较好
	文献[80]	10	1/8	在合成数据集上验证, 难度稍低
	文献[85]	4	1/8	视频预测与语义分割解析结合在一起
	文献[86]	2	1/14	视频差值模型, 双向输入

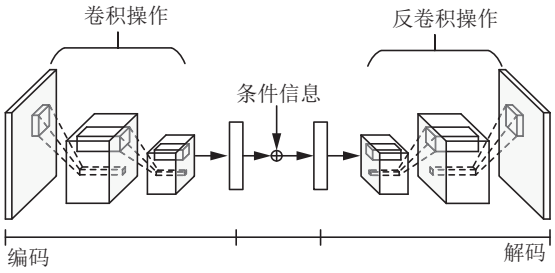


图 2　基于自编码器的视频预测模型架构

Fig. 2　Architecture of video prediction based on auto encoder

Yan 等[81]提出了一种深度动态编码器模型(deep DynEncoder), 该模型输入原始像素图像, 经编码器编码成隐状态变量, 然后使用动态预测器(DynPredictor) 将时序动态编码。使用合适的堆叠策略、逐层预训练和联合微调, 可以构建多层深度动态编码器。实验表明, 文献[81]提到的方法可以描绘复杂的视频动态, 合成高质量的纹理序列视频。作者还构造了基于深度动态编码器模型的分类和聚类方法, 在交通场景分类和运动分割上取得了接近甚至优于之前最好的模型的效果。

Vukoti 等[65]提出基于时间差Δt的卷积自编码器模型。编码器有两个分支, 一个接收输入图像, 另外一个接收期望预测的时间差Δt, 解码器根据编码器输出的潜在变量生成可信的图像。以没有时间差输入的常规卷积自编码器模型为基准, 作者提出的方法在 KTH 数据集上生成的图像有更高的语义性, 均方误差也更低。然而, 该模型存在诸多不足, 例如生成的人体动作具有歧义, 不能很好地建模快速移动的物体, 不能充分地处理前景和背景信息等。

Liu 等[73]提出一种深度体元流模型, 该模型是一种全卷积自编码器架构, 由 3 个卷积层、3 个反卷积层和一个瓶颈层组成。为更好地保留空间信息, 在每个卷积层和反卷积层之间有跳跃连接。在 UCF-101 和 THUMOS-15 数据集上的内插和外推视频实验上的结果表明, 该模型比文献[70]中提到的多尺度对抗训练架构和光流法的结果要更优。

Xue 等[87]提出一种基于变分自编码器和交叉卷积网络的模型, 该模型可以从一张图片生成可能的未来帧。该模型通过条件变分自编码器来建模未来帧的复杂条件分布。另外, 该模型利用了图像差分(欧拉运动) 原理, 因为图像差分是稀疏的, 并且比原始图像更容易建模。Xue 等还在合成数据集与自然图像上验证了模型的有效性, 另外, 作者还通过实验证明了该模型在无监督、零样本类比学习上取得了很好的结果。

4.2　递归神经网络 (RNN) 架构

递归神经网络可以很好地进行序列数据建模, 视频预测本身也是一种序列学习问题, 很多研究人员采用递归神经网络来解决视频预测问题。基于编解码的递归神经网络架构如图 3(a) 所示。

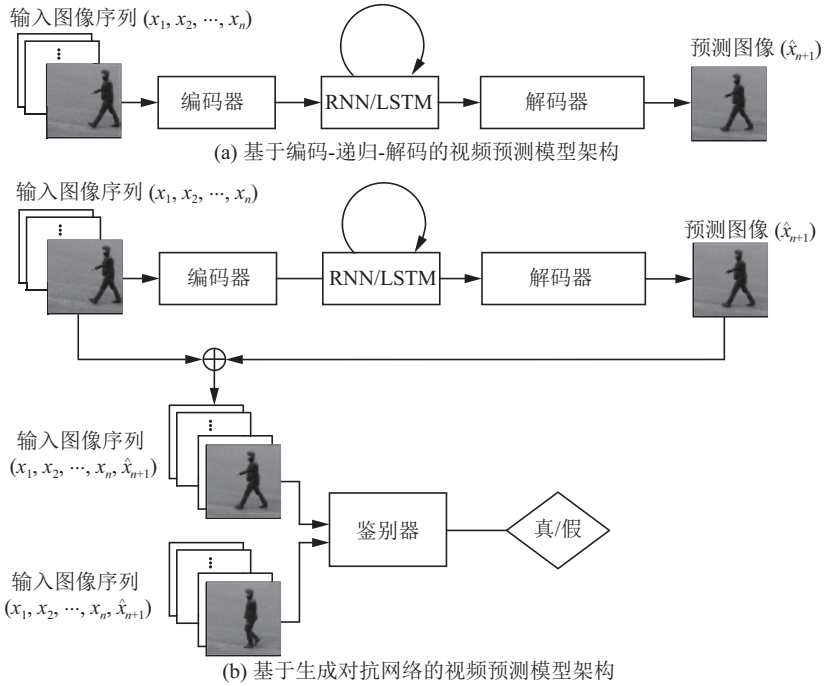


图 3　视频预测模型的抽象结构

Fig. 3　Abstract architecture of video prediction model

Ranzato 等[82]从自然语言处理领域借鉴了经典的 n-grams 算法, 将之与 CNN 和 RNN 结合起来, 给出了一个视频预测和视频插值的基准。Ranzato 还在 RNN 架构基础上提出了递归卷积神经网络

(recurrent convolution neural network, RCNN) 架构, RCNN 是在 RNN 输入和输出端连接卷积层, 使其能够更好地处理图像结构信息。

Srivastava 等[13]提出了一种使用 LSTM 架构的无监督视频表征学习模型。该模型将图像经过编码器编码后送入 LSTM 网络, 解码器可以重建原视频, 或者预测未来视频。然而, 一个高容量的自编码器网络倾向于记忆输入数据, 预测模型倾向于仅仅存储最近几帧, 因此本文提出了一个复合模型, 复合模型可以同时重构原图像、预测未来图像, 强迫模型来更好地学习视频表征。Srivastava 最后把无监督学习过程学习到的表征应用到有监督学习——动作分类中, 实验结果表明, 在训练样本很少的情况下, 无监督视频预测学习到的特征显著提升了分类结果。

Lotter 等[76]从神经科学的“预测编码”概念获得启发, 提出了一种视频预测架构——PredNet, 该架构的每一层只做局部预测, 向后面的层传递残差。PredNet 在 KITTI 数据集上的结果表明其可以统一建模背景和移动物体 (车辆、行人) 的运动。

Oh 等[83]受 DeepMind 使用雅利达 (Atari) 游戏进行增强学习研究的启发, 提出未来图像不仅与过去的图像有关, 还与当前的操作行为有关。Oh 因此提出一种由编码器、操作变换和基于 CNN 和 RNN 的解码器组成的模型。实验结果表明, 基于操作信息的条件模型可以生成视觉上较真实的、可用于游戏控制的大约 100 帧预测视频。Finn 等[84]随后也提出了基于动作的视频预测模型, 该模型可以根据不同的动作预测不同的视频, 该模型主要由卷积 LSTM 构成, 通过跳跃连接 (skip connection) 保存图形背景信息, 最后通过掩膜 (mask) 把背景和转变图像拼接起来。作者提出 3 个不同的架构: 动态神经平流、卷积动态神经平流和空间变换预测器。这 3 个模型在视频预测上都取得了不错的结果。

以上提到的方法都是直接预测高阶的视频, 由于误差累积和放大, 预测多帧视频是一个非常困难的任务。Villegas 等[68]用高阶结构信息辅助进行视频预测。他们提出的算法先从输入图像中提取人体骨架结构, 然后预测骨架结构的变化, 与参考图片联结在一起生成动作视频。实验表明, 这种以高阶结构信息为条件的视频生成策略有效减小了误差传播和累积, 在 Human3.6M 等数据集上取得了较好的效果, 且可以预测多达 128 帧的视频。但是该方法仅能预测一种可能的运动, 而且背景信息保持不变, 不能建模背景的变化, 因此有一定的局限性。

有些研究人员试图将背景和运动分开建模。Villegas 等[72]提出一种基于自编码器、CNN 和卷积 LSTM 架构的模型, 该模型有两个编码器输入, 其中一个编码器接收图像序列差分作为运动输入, 使用 LSTM 建模运动动态, 另一个编码器接收最后一帧静态图像, 然后将 LSTM 的输出与静态图像的编码输出组合起来, 经由解码器解码为预测图像。作者还提出多尺度残差版本, 将编码器各个池化层的输出通过快捷连接接入到解码器, 以更好地保存图像的结构信息。

4.3　生成对抗训练架构

生成对抗网络为机器学习领域引入了一种新的训练模式, 其优越的性能引起了众多学者的关注, 也有很多学者采用对抗训练的方式来进行视频预测。一种常用的基于编解码与生成对抗网络的视频预测架构如图 3(b) 所示。

Lotter 等[80]提出了基于编码器、LSTM 和解码器的预测生成模型, 通过对抗训练的方式, 在“弹球”数据集和计算机生成的旋转人脸数据集上取得了很好的结果, 作者还论证了无监督预测学习是一种有力的表征学习方法。

在度量生成样本和真实样本的距离上, 学者通常使用 l_1 或者 l_2 距离, 然而, 实验表明, 仅使用 l_1 或者 l_2 距离作为损失函数会导致生成图像较为模糊, 当向前预测更多帧的时候, 该问题更为严重。Mathieu 等[70]为解决预测图像模糊的问题, 提出 3 个互补的解决策略: 多尺度架构、对抗训练方法和图像梯度差分损失函数。

受限于卷积核的大小问题, 卷积操作仅能处理短范围的依赖; 另外, 使用池化还会导致分辨率降低, 文献[70]使用多尺度网络, 通过在多个不同尺度的图像进行上采样和线性组合操作来更好的保持高分辨率。

为解决使用 l_1 或者 l_2 损失函数导致的图像模糊问题, 文献[70]使用对抗训练方法。使用对抗训练方法, 模型生成的图像更锐利。然而仅优化对抗损失函数会产生训练不稳定问题, 生成器生成的图像通常可以生成“迷惑”鉴别器的样本, 然而却与真实样本 Y 并不相似。为解决这个问题, 作者使生成器采用对抗损失和 l_p 组合损失函数。通过加入损失函数迫使预测图像的分布与真实图像的分布保持一致。

Mathieu 等[70]还提出一种图像梯度差分损失, 通过引入近邻图像强度差异来惩罚预测样本和真实样本之间的梯度不一致性。最终生成器损失函数为对抗损失、l_2 损失和图像梯度差分损失的加权和。

Mathieu 的实验结果表明，使用对抗损失函数和梯度差分损失函数，性能要超过仅使用 l_2 损失函数，并且在图像锐利度上要远好于 l_2 损失函数。Hintz[71]受文献[70]的启发，将生成器替换为储蓄池计算，鉴别器结构以及训练方法与文献[70]保持相同。作者在 UCF-101 数据集上的实验结果表明，虽然其在 PSNR 和 SSIM 评测上结果略低于文献[70]，但其收敛时间明显快于前者，也取得了相当好的结果。

图像语义分割具有广泛的应用价值。Luc 等[85]在文献[70]的基础上，使用多尺度架构和对抗训练方法来预测语义分割图像。实验结果表明，预测语义分割图像的精度要好于直接预测 RGB 图像，且预测分割图像的平均 IoU 达到了真实图像分割结果的 2/3。

Vondrick 等[43]提出使用时空卷积生成对抗网络的视频预测模型 VGNN，该模型利用时空卷积网络将前景和背景解耦。本文使用生成对抗网络从潜在编码向量生成高维视频，分别提出了由时空卷积和反卷积组成的单流架构，以及可以建模静态的背景和动态的前景的双流架构。该模型在超过 200 万条视频上训练后可以自己“创作”视频内容。作者以自编码器架构作为基准，经“亚马逊土耳其机器人”测试。结果表明，双流对抗网络性能优于对应的单流对抗网络，远优于自编码器网络，甚至有 20% 的人认为模型生成的视频比自然视频更“真实”。在预测未来帧问题上，Vondrick 等在生成器前加入一个编码器，将静态图片编码为潜在编码向量，作为双流生成对抗网络的输入，生成模型可以生成 32 帧的视频 (一般视频是 25 帧/秒，因此模型可以生成约 1.5 s 的视频)。结果表明，生成器生成的视频虽然不是严格意义上的正确视频，但在语义上是可接受的。

Vondrick 等最终把通过无监督方式学习到的鉴别模型参数用在监督学习任务上 (例如动作分类)，将鉴别器最后一层替换为 Softmax 分类器。实验结果表明，使用无监督学习到的参数初始化分类器，在同样样本量大小情况下，其分类性能高于随机初始化的网络，对比效果图见图 4。Jin 等[79]使用基于生成对抗网络的时空特征学习方法，结合预测转向解析模型，可以增强现有的场景解析模型。其实验结果表明，其在 Cityscapes 视频分割数据集上取得了较好的结果。

Denton 等[88]也提出将视频背景内容和运动前景分开编码的视频表征分解模型，与文献[80]不同的是，文献[88]是以生成对抗网络的方式训练背景内容编码器、运动姿势编码器以及解码器。在 KTH 数据集上的实验结果表明，文献[88]的视频预测在准确性和图像锐利性方面要好于文献[80]。作者还提出，背景内容编码器可以构建图像分类模型，运动前景编码器可以构建视频动作分类模型。

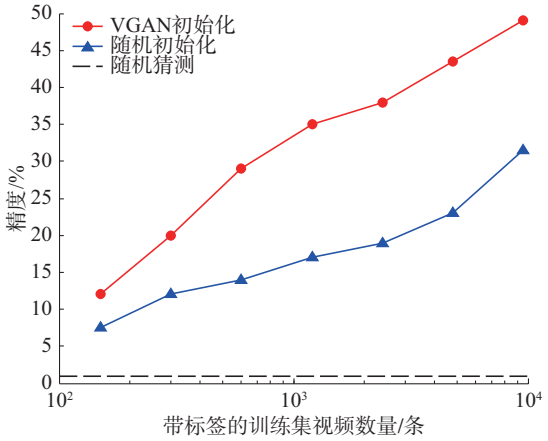


图 4　在 UCF101 数据集上，VGAN 鉴别器参数初始化分类器、随机值初始化分类器以及随机猜测类别的性能对比

Fig. 4　Performance comparison of classifier initialized by VGAN discriminator's parameters, classifier initialized by random value and random classification on UCF101

与文献[68]类似，Yan 等[67]基于条件 GAN 架构，用人体骨骼作为辅助信息，可以生成多帧栩栩如生的运动视频。

Chen 等[86]提出一种双向预测网络来进行视频插值，该模型采用编码器—解码器架构，通过两个编码器分别编码起始帧和结尾帧，从而产生一个潜在表征，解码器以潜在表征作为输入来生成多帧插值视频。该模型采用多尺度架构，其损失函数为 l_2 重建损失、特征空间损失 (以 AlexNet 最后一个卷积层提取到的特征作为基准) 与对抗损失的加权和。该模型在合成 2D 数据集和 UCF101 数据集上的结果表明，其比基于光流场的模型的效果要更好。

5　结束语

当前深度有监督学习在计算机视觉、自然语言处理和机器翻译等领域取得了远超传统方法的性能，但这些成就多属于深度学习在感知层面的工作，这属于人工智能的第一步；下一步就是让机器能够理解自然界变化的规律，对自然界动态进行建模，使其能够对现实世界中将要发生的事情进行预测，要达到这一步，需要借助于无监督学习。无监督学习因其可以在自然界海量的无标注数据上进行训练，且应用范围广泛，因而被誉为“深度学习的圣杯”。

视频预测作为无监督学习的一个最新的也是最有前景的研究方向之一，其意义不仅在于能够很好

地建模视频场景来推测未来视频, 从而帮助机器能够更好地决策, 还在于其以无监督方式学习到的内部视觉表征可以加速或提升弱监督学习和有监督学习的性能, 因此得到了越来越多学者的关注, 也取得了非常多的进展。但是, 现有的方法仍旧存在许多不足:

1) 当前提出的各种模型, 结构比较单一, 多数是基于自编码器、递归神经网络 (包括 LSTM) 和生成对抗网络, 虽然这些架构取得了不错的效果, 但是仍无法高效建模自然界复杂的动态结构, 导致当前的模型仅能预测有限的几帧或者几十帧图像, 且在预测的后期画面会变模糊或者失去语义信息。

2) 目前学术界使用的视频预测损失函数比较单一, 常使用的损失函数是均方误差损失、对抗损失函数和图像梯度差分损失函数。因为图像具有高维复杂结构信息, 当前常用损失函数没有充分考虑结构信息, 导致模型预测的图像缺乏语义信息。另外, 使用峰值信噪比、结构相似性作为图像评价标准, 与人眼的视觉感知并不完全一致, 人眼的视觉对于误差的敏感度并不是绝对的, 其感知结果会受到许多因素的影响而产生变化, 因此在图形评价指标上仍有待研究。

3) 理论上, 预测视频动态在机器人决策、无人驾驶和视频监控系统等领域具有广泛的应用价值, 但当前视频预测的研究多数在学术界, 且研究处于早期阶段, 具体在工业界的应用还未起步。

视频预测学习是理解和建模自然界场景动态的有力手段, 也是无监督学习的一个新的、重要的突破点, 尽管该领域的研究面临着不少挑战和未解决的问题, 但当前认知科学和深度学习领域发展非常迅速, 尤其是在增强学习、半监督学习和无监督学习方向, 且当前的计算机计算能力越来越强, 这些有利因素定会加速视频预测研究的进展。

参考文献:

[1] LECUN Y. Predictive Learning[R]//Proceedings of the 30th Annual Conference on Neural Information Processing Systems. Barcelona, Spain, 2016

[2] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436–444.

[3] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//Proceedings of the 26th Annual Conference on Neural Information Processing Systems 2012. South Lake Tahoe, NV, USA, 2012: 1097–1105.

[4] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1026–1034.

[5] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[Z]. arXiv preprint arXiv: 1409.1556, 2014.

[6] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA, 2016: 770–778.

[7] HINTON G, DENG Li, YU Dong, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE signal processing magazine, 2012, 29(6): 82–97.

[8] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Quebec, Canada, 2014: 3104–3112.

[9] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model[J]. Journal of machine learning research, 2003, 3: 1137–1155.

[10] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[Z]. arXiv preprint arXiv: 1312.5602, 2013.

[11] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484–489.

[12] DENG Jia, DONG Wei, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA, 2009: 248–255.

[13] SRIVASTAVA N, MANSIMOV E, SALAKHUDINOV R. Unsupervised learning of video representations using LSTMs[C]//Proceedings of the 32nd International Conference on Machine Learning. Lille, France, 2015: 843–852.

[14] MCCULLOCH W S, PITTS W. A logical calculus of the ideas immanent in nervous activity[J]. The bulletin of mathematical biophysics, 1943, 5(4): 115–133.

[15] HEBB D O. The organization of behavior: A neuropsychological theory[M]. New York: Chapman & Hall, 1949.

[16] MINSKY M L, PAPERT S A. Perceptrons: an introduction to computational geometry[M]. 2nd ed. Cambridge, UK: MIT Press, 1988.

[17] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533–536.

[18] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings

of the IEEE, 1998, 86(11): 2278–2324.

[19] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7): 1527–1554.

[20] JORDAN M I. Serial order: A parallel distributed processing approach[J]. Advances in psychology, 1997, 121: 471–495.

[21] BENGIO Y. Learning deep architectures for AI[J]. Foundations and trends in machine learning, 2009, 2(1): 1–127.

[22] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Quebec, Canada, 2014: 2672–2680.

[23] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: a review and new perspectives[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8): 1798–1828.

[24] HUBEL D H, WIESEL T N. Receptive fields and functional architecture of monkey striate cortex[J]. The journal of physiology, 1968, 195(1): 215–243.

[25] FUKUSHIMA K, MIYAKE S. Neocognitron: a self-organizing neural network model for a mechanism of visual pattern recognition[M]//AMARI S I, ARBIB M A. Competition and Cooperation in Neural Nets. Berlin Heidelberg: Springer, 1982: 267–285.

[26] ZEILER M D, KRISHNAN D, TAYLOR G W, et al. Deconvolutional networks[C]//Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA, 2010: 2528–2535.

[27] NOH H, HONG S, HAN B. Learning deconvolution network for semantic segmentation[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1520–1528.

[28] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[Z]. arXiv preprint arXiv: 1511.06434, 2015.

[29] JI Shuiwang, XU Wei, YANG Ming, et al. 3D convolutional neural networks for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(1): 221–231.

[30] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735–1780.

[31] GERS F A, SCHMIDHUBER J. Recurrent nets that time and count[C]//Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. Como, Italy, 2000, 3: 189–194.

[32] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[Z]. arXiv preprint arXiv: 1406.1078, 2014.

[33] SHI Xingjian, CHEN Zhourong, WANG Hao, et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Quebec, Canada, 2015: 802–810.

[34] VINCENT P, LAROCHELLE H, LAJOIE I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion[J]. Journal of machine learning research, 2010, 11: 3371–3408.

[35] NG A. Sparse autoencoder[R]. CS294A Lecture Notes, 2011: 72.

[36] KINGMA D P, WELLING M. Auto-encoding variational bayes[Z]. arXiv preprint arXiv: 1312.6114, 2013.

[37] REZENDE D J, MOHAMED S, WIERSTRA D. Stochastic backpropagation and approximate inference in deep generative models[Z]. arXiv preprint arXiv: 1401.4082, 2014.

[38] MIRZA M, OSINDERO S. Conditional generative adversarial nets[Z]. arXiv preprint arXiv: 1411.1784, 2014.

[39] CHEN Xi, DUAN Yan, HOUTHOOFT R, et al. InfoGAN: interpretable representation learning by information maximizing generative adversarial nets[C]//Proceedings of the 30th Annual Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 2172–2180.

[40] LEDIG C, THEIS L, HUSZáR F, et al. Photo-realistic single image super-resolution using a generative adversarial network[Z]. arXiv preprint arXiv: 1609.04802, 2016.

[41] WU Jiajun, ZHANG Chengkai, XUE Tianfan, et al. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling[C]//Proceedings of the 30th Annual Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 82–90.

[42] ISOLA P, ZHU Junyan, ZHOU Tinghui, et al. Image-to-image translation with conditional adversarial networks[Z]. arXiv preprint arXiv: 1611.07004, 2016.

[43] VONDRICK C, PIRSIAVASH H, TORRALBA A. Generating videos with scene dynamics[C]//Proceedings of the 30th Annual Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 613–621.

[44] VONDRICK C, PIRSIAVASH H, TORRALBA A. Anticipating visual representations from unlabeled video[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA, 2016: 98–106.

[45] LAN Tian, CHEN T C, SAVARESE S. A hierarchical representation for future action prediction[C]//Proceedings of

the 13th European Conference on Computer Vision. Zürich, Switzerland, 2014: 689–704.

[46] HOAI M, DE LA TORRE F. Max-margin early event detectors[J]. International journal of computer vision, 2014, 107(2): 191–202.

[47] RYOO M S. Human activity prediction: Early recognition of ongoing activities from streaming videos[C]//Proceedings of the 2011 IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 1036–1043.

[48] VU T H, OLSSON C, LAPTEV I, et al. Predicting actions from static scenes[C]//Proceedings of the 13th European Conference on Computer Vision. Zürich, Switzerland, 2014: 421–436.

[49] PEI Mingtao, JIA Yunde, ZHU Songchun. Parsing video events with goal inference and intent prediction[C]//Proceedings of the 2011 IEEE International Conference on Computer vision. Barcelona, Spain, 2011: 487–494.

[50] FOUHEY D F, ZITNICK C L. Predicting object dynamics in scenes[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA, 2014: 2027–2034.

[51] KOPPULA H S, SAXENA A. Anticipating human activities using object affordances for reactive robotic response[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 38(1): 14–29.

[52] HUANG Dean, KITANI K M. Action-reaction: Forecasting the dynamics of human interaction[C]//Proceedings of the 13th European Conference on Computer Vision. Zürich, Switzerland, 2014: 489–504.

[53] PICKUP L C, PAN Zheng, WEI Donglai, et al. Seeing the arrow of time[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA, 2014: 2043–2050.

[54] LAMPERT C H. Predicting the future behavior of a time-varying probability distribution[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA, 2015: 942–950.

[55] PINTEA S L, VAN GEMERT J C, SMEULDERS A W M. Déja vu: Motion prediction in static images[C]//Proceedings of the 13th European Conference on Computer Vision. Zürich, Switzerland, 2014: 172–187.

[56] KITANI K M, ZIEBART B D, BAGNELL J A, et al. Activity forecasting[C]//Proceedings of the 12th European Conference on Computer Vision. Florence, Italy, 2012: 201–214.

[57] GONG Haifeng, SIM J, LIKHACHEV M, et al. Multi-hypothesis motion planning for visual object tracking[C]// Proceedings of the 2011 IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 619–626.

[58] KOOIJ J F P, SCHNEIDER N, FLOHR F, et al. Context-based pedestrian path prediction[C]//Proceedings of the 13th European Conference on Computer Vision. Zürich, Switzerland, 2014: 618–633.

[59] WALKER J, DOERSCH C, GUPTA A, et al. An uncertain future: Forecasting from static images using variational autoencoders[C]//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, Netherlands, 2016: 835–851.

[60] WALKER J, GUPTA A, HEBERT M. Dense optical flow prediction from a static image[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 2443–2451.

[61] WALKER J, GUPTA A, HEBERT M. Patch to the future: Unsupervised visual prediction[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA, 2014: 3302–3309.

[62] YUEN J, TORRALBA A. A data-driven approach for event prediction[C]//Proceedings of the 11th European Conference on Computer Vision. Heraklion, Crete, Greece, 2010: 707–720.

[63] MOTTAGHI R, RASTEGARI M, GUPTA A, et al. “What happens if...” learning to predict the effect of forces in images[C]//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, Netherlands, 2016: 269–285.

[64] SCHUKDT C, LAPTEV I, CAPUTO B. Recognizing human actions: a local SVM approach[C]//Proceedings of the 17th International Conference on Pattern Recognition. Cambridge, UK, 2004, 3: 32–36.

[65] VUKOTI V, PINTEA S L, RAYMOND C, et al. One-step time-dependent future video frame prediction with a convolutional encoder-decoder neural network[C]//Proceedings of the 19th International Conference on Image Analysis and Processing. Catania, Italy, 2017: 140–151.

[66] IONESCU C, PAPAVA D, OLARU V, et al. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 36(7): 1325–1339.

[67] YAN Yichao, XU Jingwei, NI Bingbing, et al. Skeleton-aided articulated motion generation[Z]. arXiv preprint arXiv: 1707.01058, 2017.

[68] VILLEGAS R, YANG Jimei, ZOU Yuliang, et al. Learning to generate long-term future via hierarchical prediction[Z]. arXiv preprint arXiv: 1704.05831, 2017.

[69] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[Z]. arXiv preprint axXiv:1202.0402, 2012

[70] MATHIEU M, COUPRIE C, LECUN Y. Deep multi-scale video prediction beyond mean square error[Z]. arXiv preprint arXiv: 1511.05440, 2015.

[71] HINTZ J J. Generative adversarial reservoirs for natural video prediction[D]. Austin, USA: The University of Texas.

[72] VILLEGAS R, YANG Jimei, HONG S, et al. Decomposing motion and content for natural video sequence prediction[C]//Proceedings of the 2017 International Conference on Learning Representations. Toulon, France, 2017.

[73] LIU Ziwei, et al. Video frame synthesis using deep voxel flow[C]//Proceeding of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA, 2017:4463–4471

[74] GORBAN A, IDREES H, JIANG Yugang, et al. THUMOS challenge: Action recognition with a large number of classes[EB/OL]. (2015–05). http://www.thumos.info.

[75] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: the KITTI dataset[J]. The international journal of robotics research, 2013, 32(11): 1231–1237.

[76] LOTTER W, KREIMAN G, COX D. Deep predictive coding networks for video prediction and unsupervised learning[Z]. arXiv preprint arXiv: 1605.08104, 2016.

[77] Kuehne H, Jhuang H, Garrote E, et al. HMDB: A large video database for human motion recognition[C]//Proceeding of the 2011 IEEE International Conference on Computer Vision, ICCV. Barcelona, Spain, 2011:2556–2563.

[78] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA, 2016: 3213–3223.

[79] JIN Xiaojie, LI Xin, XIAO Huaxin, et al. Video scene parsing with predictive feature learning[Z]. arXiv preprint arXiv: 1612.00119, 2016.

[80] LOTTER W, KREIMAN G, COX D. Unsupervised learning of visual structure using predictive generative networks[Z]. arXiv preprint arXiv: 1511.06380, 2015.

[81] YAN Xing, CHANG Hong, SHAN Shiguang, et al. Modeling video dynamics with deep dynencoder[C]//Proceedings of the 13th European Conference on Computer Vision. Zürich, Switzerland, 2014: 215–230.

[82] RANZATO M, SZLAM A, BRUNA J, et al. Video (language) modeling: a baseline for generative models of natural videos[Z]. arXiv preprint arXiv: 1412.6604, 2014.

[83] OH J, GUO Xiaoxiao, LEE H, et al. Action-conditional video prediction using deep networks in atari games[C]// Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Quebec, Canada, 2015: 2863–2871.

[84] FINN C, GOODFELLOW I, LEVINE S. Unsupervised learning for physical interaction through video prediction[C]// Proceedings of the 30th Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 64–72.

[85] LUC P, NEVEROVA N, COUPRIE C, et al. Predicting deeper into the future of semantic segmentation[Z]. arXiv preprint arXiv: 1703.07684, 2017.

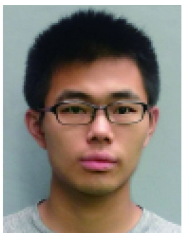
[86] CHEN Xiongtao, WANG Wenmin, WANG Jinzhou, et al. Long-term video interpolation with bidirectional predictive network[Z]. arXiv preprint arXiv: 1706.03947, 2017.

[87] XUE Tianfan, WU Jiajun, BOUMAN K, et al. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks[C]//Proceedings of the 30th Annual Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 91–99.

[88] DENTON E, BIRODKAR V. Unsupervised learning of disentangled representations from video[Z]. arXiv preprint arXiv: 1705.10915, 2017.

作者简介：

莫凌飞，男，1981 年生，副教授，博士，主要研究方向为机器学习与人工智能、物联网与边缘计算、智能机器人。发表学术论文多篇，其中被 SCI、EI 检索 40 余篇。



蒋红亮，男，1993 年生，硕士研究生，主要研究方向为深度无监督学习和计算机视觉。

李煊鹏，男，1985 年生，讲师，博士，主要研究方向为机器视觉、驾驶辅助系统、环境感知与信息融合。