EIDETIC 3D LSTM: A MODEL FOR VIDEO PREDICTION AND BEYOND

**Abstract**

时空预测学习虽然长期被认为是一种很有前途的自监督特征学习方法，但其有效性很少能超越未来的视频预测。其原因是短期框架依赖和长期高层关系的良好表征难以学习。我们提出了一种新的模型，Eidetic 3D LSTM (E3D-LSTM)，它将3D卷积集到RNNs中。封装的3D-Conv使局部感知器感知RNNs的运动，使记忆单元存储更好的短期特征。为了长期的关系，我们通过一个门控的自我注意模块使当前的记忆状态与它的历史记录相互作用。我们将这种记忆转换机制描述为清晰的，因为它能够跨多个时间戳有效地回忆存储的记忆，即使是在长时间的干扰之后。我们首先在广泛使用的未来视频预测数据集上对E3D-LSTM网络进行了评估，并获得了最先进的性能。然后我们证明E3D-LSTM网络在早期活动识别方面也有很好的表现，通过观察有限的视频帧来推断正在发生什么或将要发生什么。

**Introduction**

时空预测学习的一个基本问题是如何有效地学习视频推理或推理的良好表征。目前，递归神经网络(RNNs)仍然是该领域最有前途的模型，并在未来的视频预测基准测试中取得了最先进的结果。然而，除了帧预测之外，基于RNN的模型在学习高级视频表示或捕获长期关系方面的效率较低。另一方面，最近的研究表明3D卷积神经网络(3D-CNNs)在学习更好的动作分类表示方面超过了RNNs

基于3D-CNN的成功，本文提出了一种基于递归建模(针对时间依赖)和前馈3D-Conv建模(针对局部动态)的时空预测学习新模型。

**Related work**

**Spatiotemporal Predictive Learning Models**：

近年来，RNNs在序列预测和未来帧预测中得到了广泛的应用。Srivastava等(2015)将基于lstm的序列扩展为序列模型，Shi等人(2015)提出了卷积LSTM，将卷积集成到循环状态转换中进行高维序列预测。卷积LSTM模型是由Finn等(2016)扩展来预测机器人环境的未来状态。

上述递归模型主要基于顺序更新的记忆状态来预测未来帧。当内存单元刷新时，旧的内存将立即被丢弃。相比之下，所提出的E3D-LSTM模型维护历史内存记录列表，并在必要时撤销它们，从而促进了远程视频推理

不完全模型的未来预测误差可分为两个方面:一是由于对确定性变量缺乏建模能力而导致的“系统误差”;b)未来内在的随机不确定性。我们的目标是使这项工作中的第一个因素最小化。

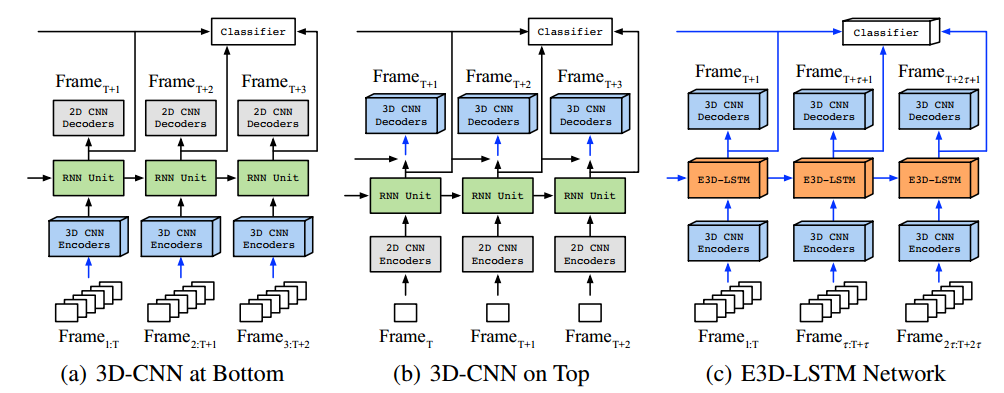
**Convolutional Recurrent Networks**：我们的模型与卷积递归网络密切相关。在ConvLSTM网络中(Shi et al.， 2015)，所有状态转换都用2D卷积实现。

**3. EIDETIC 3D LSTM**

本节首先介绍了在视频中感知和记忆短期和长期表象的Eidetic 3D LSTM。然后，我们讨论了一个预定的多任务学习策略，使用预测学习作为一个辅助的自我监督任务的活动识别。

**3.1 3D CONVOLUTIONS IN RECURRENT NETWORK**

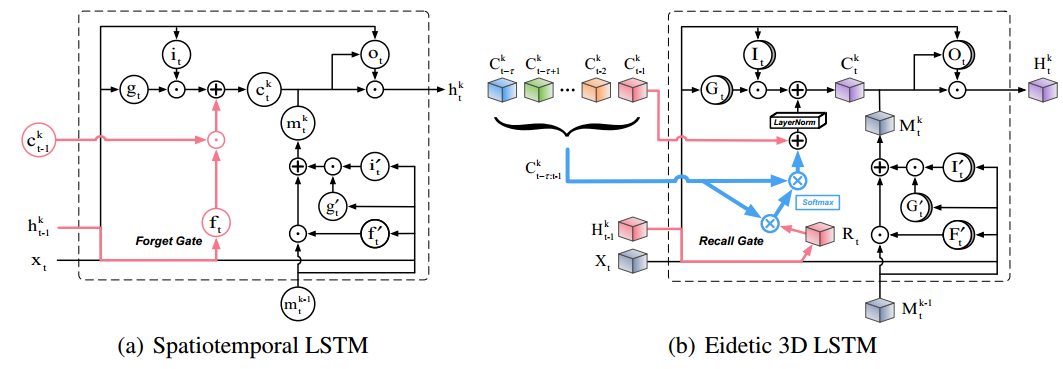
一个理想的预测模型依赖于对视频表示的有效学习，RNNs和3D-CNNs是不同时空数据建模机制的网络结构。在这项工作中，我们的目标是在一个统一的架构中利用每一个的力量，并开始讨论两个合理的叠加3D-Convs和RNN单元的扩展。图1(a)和图1(b)展示了在叠加时空LSTMs之前或之后添加3D-CNNs的两种混合基线网络。



然而，我们发现在LSTM单元之外集成3D-Convs的性能明显低于基线RNN模型。为此，我们建议在LSTM单元内对3D-Convs进行“更深入”的集成，以便将卷积特性纳入周期性状态转换。图1(c)显示了整个编码器-解码器架构。在该模型中，首先用若干层3D-Convs对连续的T输入帧进行编码，得到高维特征映射。将三维conv地形图直接输入到一种新的E3D-LSTM中，对其进行长期时空交互作用的建模。最后，通过若干个叠加的3D-Conv层对E3D-LSTM隐状态进行解码，得到预测的视频帧。对于分类任务，隐藏状态可以直接作为学习的视频表示。

**3.2 EIDETIC MEMORY TRANSITION**

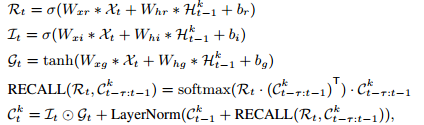
所提出的Eidetic 3D LSTM的架构如图2所示，其中红色箭头表示短期信息流，蓝色箭头表示长期信息流



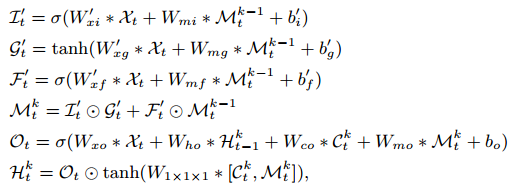
输入有四个：Xt，在以前的E3D-LSTM层从编码器或隐藏状态获得的3D-Conv特征映射。Ht-1k以前时间戳中的隐藏状态, Ct-1k内存状态来自以前的时间戳; Mt-1k-1先前所描述的时空记忆状态.

我们使用周期性的3D-Convs作为运动感知感知器，提取连续时空场中的短期表象和局部运动，并将其存储在一个小时空体中。因此，视频表象和短期运动可以用RT×H×W×C张量进行编码，其中每个维度分别表示时间深度、空间大小和地物图通道数量。通过在时间维度上对存储状态的扩展，我们发现该方法是可行的E3D-LSTM变得更有能力描述和记忆局部或短期动作。

为了捕获长期的帧交互，我们提出了一种新的记忆回忆机制，改进了记忆状态的递归过渡函数:



与传统的记忆转换函数不同，回忆函数学习时间交互的大小。我们的工作部分是由自我关注机制驱动的(Lin et al.， 2017;Vaswani等人，2017)。然而，在我们的模型中，注意机制不是应用于输出状态，而是应用于内存转换期间。t是用来唤起过去的记忆，从遥远的时间戳，以记忆和提取有用的信息，从什么是感知。我们发现学习注意过去的记忆状态有助于回忆长期的历史背景。Ctk内存张量被命名为eidetic 3D memory，整个单元被命名为E3D-LSTM，我们还利用相同的回忆方法将M1: tk与垂直内存转换流关联起来，但是结果证明这种方法的帮助不大。使用更新后的内存状态Ctk，输出隐藏状态为:



3.3 SELF-SUPERVISED AUXILIARY LEARNING

对于许多有监督的任务，如视频动作识别，通常没有足够的监督或注释来训练一个令人满意的RNN。作为解决这一问题的一种辅助措施，未来视频预测被认为是一种很有前途的表示学习方法，随着时间的推移，它的监督更严密，并可能提取有用的特征来帮助视频理解。我们考虑两个任务:像素级的未来帧预测和另一个视频级的分类任务(在我们的例子中是早期活动识别)。对于帧预测，目标函数为:



其中X^和X分别是预测的未来帧和实际的帧。

对于早期的活动识别，我们使用一个多任务学习目标，使这两个任务的模型在端到端训练中共享相同的网络骨干网:



其中y^和y分别是高级预测的类和相应的真值类。

尽管改进这两项任务需要适当的长时间、短期的上下文表示，但不能保证使用像素级监督学习的功能将完全符合任何高级目标。因此，我们提出一种课程学习策略，目标函数逐渐从一个任务倾斜到另一个任务。具体来说,我们应用一个线性衰减λ/迭代次数i:



5. conclusion

时空预测学习在天气预报、交通流预测和物理交互仿真等诸多应用领域都取得了显著的进步。虽然它被认为是一种很有前途的自监督特征学习范式，但它的有效性很少超出视频预测。在本文中，我们提出了基于3D卷积递归单元的E3D-LSTM模型。在该模型中，我们将3D-Convs集成到状态转换中来感知短期运动，并设计了一个由递归门控制的记忆关注模块来捕获长期的视频帧交互。