指导教师：杨涛　　提交时间：2016/3/18

CVPR2015 Paper Translation

No: 01
姓名：惠东
学号：2013302548
班号：10011304

长期反复回旋的视觉识别和描述网络

Jeff Donahue Lisa Anne Hendricks Sergio Guadarrama Marcus Rohrbach
Subhashini Venugopalan† Kate Saenko‡ Trevor Darrell
†UT Austin ‡UMass Lowell UC Berkeley, ICS
Austin, TX Lowell, MA Berkeley, CA
vsub@cs.utexas.edu saenko@cs.uml.edu {jdonahue, lisa anne, sguada, rohrbach, trevor}@eecs.berkeley.edu

摘要

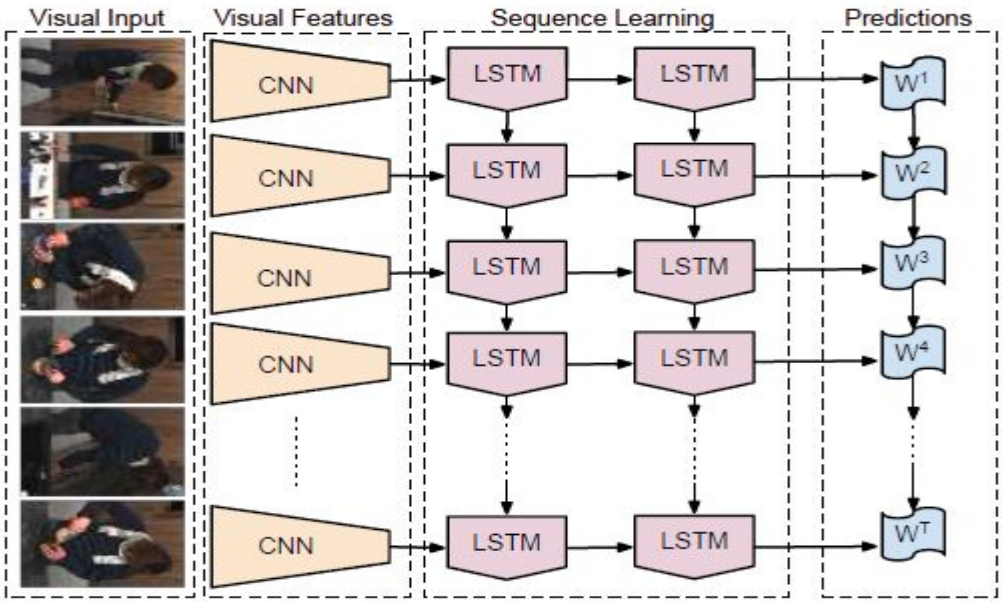
基于深卷积模型的网络已经在近年来的图像解译工作中占主导地位；我们研究一种经常反复或“暂时深”的气象模型对于涉及序列的，视觉的，相反的任务是有效的。我们开发一种新型适合大型视觉学习的周期性卷积架构模型，它是端至端的可训练的，并且证明这些模型在卷积视频识别任务，图像描述和恢复的问题和视频记录挑战方面的价值。对比于顺序处理一假定的固定时空接受域或简单是件平均域的电流模型，反复卷积模型是“双重深”的，因为它可以在空间和时间“层”组成。这样的模型具有优势，当目标概念是复杂的或（和）训练数据是有限的。长期学习依赖可能是非线性的时候被纳入网络状态更新。长期递归神经网络模型是吸引人的，因为它们直接可以将可变长度输入（例如，视频帧），转为可变长度的输出（例如，自然语言文本），并且在复杂的时间动态模型，它们可以用反向传播来优化。我们的长期反复模型直接连接到现代视觉模型中，并且可以共同训练同时学习时间动态和卷积视觉的表示。我们的结果表明这种模式有超过最先进的识别或个别的明确或（和）优化的新一代模型的明显优点。

1、简介

图像和视频的识别与描述 计算机视觉的一个基本挑战。戏剧性的视觉特征序列视频输入学习预测图 1：我们提出长期反复卷积网络（LRCNs）是一类利用快速进步的力量在细胞神经网络的视觉识别的问题而且应用随时间变化的输入和输出的模型发展愿望的架构，长期反复卷积网络处理（可能）可变长度视觉输入（左）与反复神经网络的（中间偏左），其输出送入堆叠反复序列模型（LSTMs，中间偏右），这最终产生的可变长度预测 （右）。

发展已经通过监督卷积模型使图像识别任务来实现，并且近来已经提出很多扩展处理的视频。理想的情况下，视频模型应该允许处理可变长度的输入序列，并且应该提供可变长度的输出，包括在常规以外的全长句子的描述预测任务。在本文中，我们提出了长期反复卷积网络(LRCNs)，一种新兴的结合卷积层和长期时间递归并且是端到端可训练的用于视觉识别和描述的架构（见图 1）。我们举例说明我们的为特定的视频活动识别，图像，字幕生成和视频描述任务架构，如下所述。

迄今为止，反复神经网络模型视频处理已经成功地考虑通过原始序列数据学习三维时空滤光器[13，2]，并且通过固定窗口或视频尽头片段学习瞬时光流或基于轨迹模型合并的从帧到帧的表示[16，33]。这种模式探索感知的时间序列表示学习的两个极值：无论是学习一种完全通用的随时间变化的权重，或者应用于简单的时空池。接下来，与激发当前深度卷积模型相同的灵感，我们主张视频识别和描述模型也是深度超过时间维度，即，具



有潜变量的时间循环。递归神经网络模型是在“时间深度”众所周知的;例如，此时明确地展开，并形成在时域中含蓄表示组成。这样的“深”模式是早期深度的空间卷积模型并在文献

发展已经通过监督卷积模型使图像识别任务来实现，并且近来已经提出很多扩展处理的视频。理想的情况下，视频模型应该允许处理可变长度的输入序列，并且应该提供可变长度的输出，包括在常规以外的全长句子的描述预测任务。在本文中，我们提出了长期反复卷积网络(LRCNs),一种新兴的结合卷积层和长期时间递归并且是端到端可训练的用于视觉识别和描述的架构（见图 1）。我们举例说明我们的为特定的视频活动识别，图像，字幕生成和视频描述任务架构，如下所述。

迄今为止，反复神经网络模型视频处理已经成功地考虑通过原始序列数据学习三维时空滤光器[13，2]，并且通过固定窗口或视频尽头片段学习瞬时光流或基于轨迹模型合并的从帧到帧的表示[16，33]。这种模式探索感知的时间序列表示学习的两个极值：无论是学习一种完全通用的随时间变化的权重，或者应用于简单的时空池。

接下来，与激发当前深度卷积模型相同的灵感，我们主张视频识别和描述模型也是深度超过时间维度，即，具有潜变量的时间循环。递归神经网络模型是在“时间深度”众所周知的;例如，此时明确地展开，并形成在时域中含蓄表示组成。这样的“深”模式是早期深度的空间卷积模型并在文献[31，45]。

递归神经网络几十年以来一直长期在探索感知应用，并具有不同的结果。一种简单 RNN 模型严格对时间的整合状态信息的显著限制称为“梯度消失”的效果：通过长距离的时间间隔反向传播的误差信号的能力在实践中变得越来越不可能。一类启用远程学习的模型在长短期记忆中首次被提出，并且增加隐藏的非线性机制状态导致不加修改，更新，或复位而传播，使用简单的存储单元类似神经大门。虽然该模型证明了几个任务是有用的，但在最近的研究结果它的效用在报告语音识别[10]和语言翻译模式的大型学习[39，5]中变的很明显。

在这里，我们表明长期反复卷积模型一般适用于视觉的时间序列模型。

我们认为，先前在静态或平面时空模型已采用视觉任务，当具有充足的

测试数据时，长期递归神经网络模型可以提供重要丰富的学习或提炼的表示。具体来说，我们将展示 LSTM 型模型提供的改进了常规视频活动挑战的识别和一种新兴端到端的从图像像素到语句级的自然语言描述的优化的图像显示。我们还表明，这些模型提高了从传统的可视化模型衍生出中间视觉表现的描述。

我们在三个实验设置中提出的实例架构（见图 3）。首先，我们显示，直接连接深度 LSTM 网络的可视化卷积模型，我们能够训练捕获复杂的时空状态依赖的视频识别模型（图 3 左;第 4 节）。尽管现有的标记视频活动数据集可能不会有极其复杂的动态时的行为或活动，但我们仍然看到了在传统基准的数量级上的改进了 4%。

其次，我们探索了直接终端到终端的可训练的图像判别映射。最近报道的机器翻译任务的优秀业绩[39，5]；这样的模型是基于 LSTM 网络编码器/解码器对的。我们提出该模型的多模类似物，并描述了使用一个视觉编码深度状态向量的体系结构，和一个 LSTM 解码矢量为一个自然语言串（图 3 中部;第 5 部分）。所得到的模型可以训练关于端至端的大型图象和文本数据集，甚至适度训练相比现有的方法提供了竞争力。最后，我们表明 LSTM 解码器可以直接从现有的计算机视觉方法其中预测更高级别的有差别的标签，如在《以自然语言描述翻译的视频内容》[30]中预测驱动的语义视频角色元组（图 3 右;第 6 条）。虽然没有端到端可训练的，但是这种模式提供了比以前的统计机器翻译基础的办法在架构和性能上具有优势，如下报道。

我们已经了解了广泛采用开源的深度学习框架和嵌入卷积架构快速功能的广义“LSTM”式的 RNN 模型[14]，结合具体 LSTM 单位[47，39，5]。

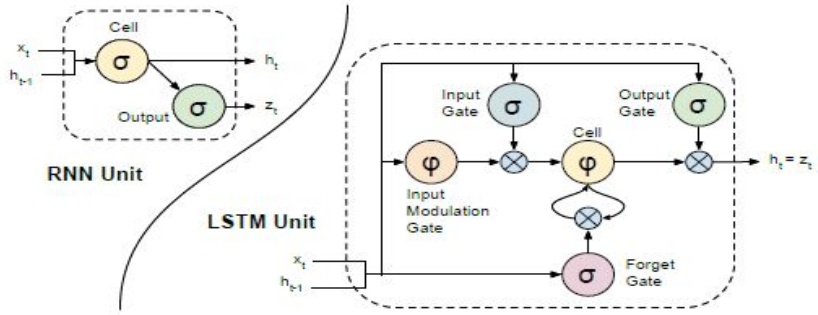
2. 背景：回归神经网络 （人工神经网络）

传统 RNNs（图 2 左）可以通过映射输入序列学习复杂时空动力学隐状态序列，和通过以下公式复发输出隐藏的状态（图 2 左）：

$$h_t = g(W_{xh}x_t + W_{hh}h_{t-1} + b_h)$$
$$z_t = g(W_{hz}h_t + b_z)$$

这里 g 是一个逐元素非线性,如一个 S 型或双曲正切，x_t 是输入，$h_t \in R^N$ 是用 N-隐藏单元的隐藏状态，y_t 是在时间 t 上的输出，对于一个长度为 T 的输入序列$< x_1, x_2, ..., x_T >$，用h_1（让$h_0 = 0$），y_1，h_2，y_2，. . . . .，h_T，y_T。

虽然 RNNs 已证明了成功的任务，如语音识别[43]和文本生成[38]，但是培养他们学习长期动力学是很困难的，可能由于部分消失和爆炸梯度问题[12]，导致通过许多层中的反复性网络的向下传播梯度，每一个对应于一个特定的时间步长。LSTMs 提供了通过存储单元使当网络学习的时候忘了先前的隐藏状态以及何时更新了隐藏状态时的新的信息解决方案。在 LSTMs 研究上已经取得进展，已经提出了与存储器单元内不同的连接隐藏单元。我们使用 LSTM 单元如学习执行[46]（图 2 描述的， 右），这是所描述的一个的轻微简化在对反复发作的神经网络终端到终端的语音识别[10]。让$\sigma(x) = (1+e^{-x})^{-1}$是非线性压缩实值，输入为[0,1]范围，并且让$\phi(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = 2\sigma(2x) - 1$是双曲正切非线性，同样压缩输入为[-1,1]范



围内。对于时间步长为 t 的 LSTM 更新给输入 x_t， h_t 和 c_{t-1} 是：

$$
\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\
g_t &= \phi(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \varphi\ (c_t)
\end{aligned}
$$

图 2：一个基本 RNN 结构（左）和一个 LSTM 存储单元的图（右）中使用在本文中（从学习执行[46]，产生与复发性神经网络序列[9]中描述的体系结构，其从所述 LSTM 衍生最初的轻微简化在长短期记忆[12]。）

除了一个隐藏单元 $h_t \in R^N$，所述 LSTM 包括输入门 $i_t \in R^N$，忘记门 $f_t \in R^N$，输出门 $o_t \in R^N$，输入调制门 $g_t \in R^N$，和存储单元 $c_t \in R^N$。存储单元组 ct 是一个求和的两件事情：先前的存储单元组 ct-1 是通过 f_t 和 g_t 调制的，当前输入和以前的隐蔽状态的函数，是通过输入门 i_t 调制的，因为 i_t 和 f_t 是 S 形的，其值位于范围[0，1]，并且 i_t 和 f_t 可以被认为是在 LSTM 学会选择性地忘记了它以前的记忆或考虑其电流输入的旋钮。同样地，输出门 o_t 得知多少存储器单元转移到隐藏状态。这些额外的细胞启用 LSTM 学习极其复杂和长期的时空动态而 RNN 模型不能够学习。额外的深度可以通过堆叠它们在彼此的顶部加入到 LSTMs，使用 LSTM 的 $l-1$ 层的隐藏状态作为输入到 LSTM 的 l 层。

近日，LSTMs 在语言任务方面都取得了不俗的业绩，如语音识别[10]和机器翻译[39，5]。类似 CNN 的，LSTMs 是有吸引力，因为它们允许端至端微调。例如，通过训练映射频谱输入到文本的深度双向的 LSTM，反复的神经网络终端到终端的语音识别[10]消除了在复杂的多步骤的语音识别管道的必要性。即使没有语言模型和发音词典，该模型也提供了令人信服的文字翻译。序列测序与神经网络学习[39]和神经机器翻译的属性：编码器，解码器的方法[5]的翻译是从英语到法语的具有多层 LSTM 编码器和解码器。在源语言中的句子所使用的编码映射到一个隐藏的状态 LSTM，然后一个解码 LSTM 隐藏状态映射到目标语言的序列。这种编码器的解码器方案允许不同长度的序列被映射到彼此。如反复的神经网络终端到终端的语音识别[10]所述序列到序列结构机器翻译避开对语言模型的需要。

LSTMs 的建模序列数据的优势在视

力问题上是双重的。首先，与目前的视觉系统集成，LSTM 模型简单微调终端到终端。第二，LSTMs 并不局限于固定长度的输入或输出，允许不同长度的连续的数据，如文本或视频简单的建模。我们接下来描述了一个统一的框架结合深卷积网络 LSTMs 创建一个模型，它在空间和时间上深度。

3.长期反复卷积网络（LRCN）模型

这项工作提出了一个与长期反复发作卷积网络（LRCN）模型相结合的深层次的视觉特征提取器（如 CNN）和可以学会识别和合成涉及连续数据（输入或输出）任务时间动态模型，视觉，语言或其他方式。图 1 描述了我们的方法的核心。我们 LRCN 模型通过传递每个视觉输入v_t（从视频中隔离的图像或帧）通过特征变换ϕV（v_t）参数通过 V 产生一个固定长度的向量表示$\phi_t \in R^d$。那么序列模型接管具有计算的视觉输入序列的特征空间表示$\langle \phi_1, \phi_2, ... \phi_T \rangle$。

在其最一般的形式，序列模型参数通过 W 将输入x_t和以前的时间步隐藏状态h_{t-1}为一个输出z_t和更新隐藏状态h_t。因此，推理必须按顺序（在序列学习图 1 的盒即从上到下），通过为了计算运行：$h_1 = fw(x_1, h_0) = fw(x_1, 0)$,，然后$h_2 = fw(x_2, h_1)$等，最高$h_T$。我们的一些模型堆叠多个 LSTMs 顶上彼此的第 2 节。

在预测分配的最后一步P（y_t）在时间步长 t 输出z_t连续模式，产生了分配（在这种情况下，有限的和离散）可能每个时间步输出空间 C：

$$P(y_t = c) = \frac{\exp(W_{zc} z_{t,c} + b_c)}{\sum_{c' \in C} \exp(W_{zc} z_{t,c'} + b_c)}。$$

最近很深模型的物体识别深卷积神经网络分类，非常深刻的卷积网络的大型图像识别，回旋不断深入[22，34，40]成功表明，战略性组成许多“层”的非线性函数会导致非常强大的模型，感性的问题。对于大 T，上述复发表示从一个经常性的网络以 T 时间步的最后几个预测是由一个非常“深”（T-层状）的非线性函数计算，这表明所得到的复发模型可具有相似表达能力与 T-层神经网络。关键的是，然而，顺序模型的加权 W 是在每一个时间步长重复使用，迫使模型学习通用时间步太时间步动力学（相对于动力学直接调节对 T，序列号），并防止所述参数大小在比例生长到时间步的最大数目。

在大多数我们的实验中，视觉特征变换ϕ对应于一深 CNN 的一些层的激活。使用可视化改造$\phi v(.)$，这是时不变的，独立的在每一个时间都有制作成本卷积推断并在所有的输入时间步训练并行，方便使用快速当代 CNN 实现的效率依赖的重要优势视觉和顺序模型的独立批量处理，并且结束对终端优化参数 V 和 W。

我们考虑三个视力问题（行为识别，图像描述和视频介绍），它实例如下几大类别的顺序学习任务之一：

1.顺序输入，固定输出（图 3，左）：$\langle x_1, x_2, ..., x_T \rangle \mapsto y$。视觉活动识别问题可以落入本伞下，以任意长度 T 作为输入的视频，但预测像运行或从固定词汇跳跃绘制单个标签的目标。

2.固定输入，输出顺序（图 3 中）：$x \mapsto \langle y_1, y_2, ..., y_T \rangle$。该类别中的

图像描述问题配合，具有作为输入的非时变的图像，但更大的和更丰富的标签中心包括任何长度的句子。

3.连续输入和输出（图3右）：$\langle x_1, x_2, ... x_T \rangle \mapsto \langle y_1, y_2, ... y_{T'} \rangle$.

最后，很容易想象为其中两个视觉输入和输出是随时间变化的任务，并且在一般的输入和输出的时间步的数量可以有所不同（即，我们可以具有$T \neq T'$）。在影片说明任务，例如，输入和输出都是连续的，并且帧的视频中的数量应不限制的（数字）的自然语言描述的长度。

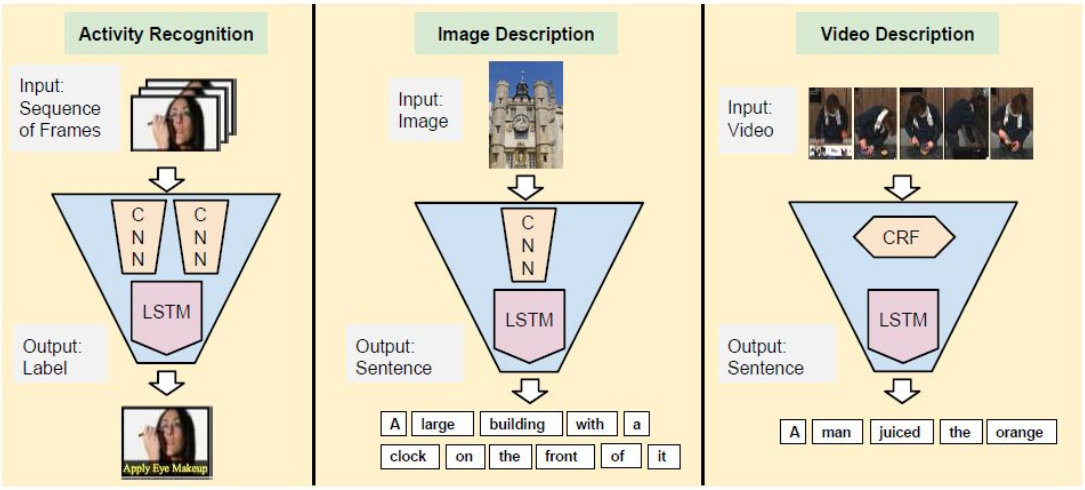
在前面描述的制剂，每一个实例是T的输入$\langle x_1, x_2, ..., x_T \rangle$和T输出$\langle y_1, y_2, ..., y_T \rangle$。我们描述我们如何在我们的混合模式，以解决各个上述三个问题的设置适应这一提法。随着连续输入和输出标量，我们就一晚融合的方法来每个时间步长的预测融入了全序列$\langle y_1, y_2, ..., y_T \rangle$的单一预测。随着固定大小的输入和输出的顺序，我们简单重复的输入X在所有时间步长$x_t := x$（注意这是可以做到廉价由于时间不变的视觉特征提取）。最后，对于一个序列到序列问题（一般）不同的输入和输出的长度，我们取一个“编码器 – 解码”的方法由回归神经网络正规化[47]的启发。在这种方法中，一个序列模型，编码器，用于映射的输入顺序，以固定长度的向量，则另一序列模型中，解码器，用于将展开该载体来任意长度的连续的输出。在这种模式下，该系统作为一个整体可以被认为是具有输入和输出的$T + T_0$时间步，其中，所述输入处理和解码器输出的第一T时间步忽略，并且预测被制成和“虚拟”输入将被忽略后者T_0时间步长。根据所提出的系统，该模型的视觉和顺序成分的重量（V，W）可以被共同通过最大化地面实况输出条件的输入数据y_t的可能性了解到和标签到这一点尤其是$(x_{1:t}, y_{1:t-1})$，对于特定的训练序列$(x_t, y_t)_{t=1}^{T}$，我们最小化负对数似然$\zeta(V,W) = -\sum_{t=1}^{T} \log P_{V,W}(y_t \mid x_{1:t}, y_{1:t-1})$。

一项的所述系统的最吸引人的方面是学习的参数的能力“端到端”，以使得视觉特征提取的参数V学习挑选出是相关的顺序的视觉输入的各方面分类问题。我们培训使用具有动量随机梯度下降LRCN模型，与反向传播用于计算目标L的梯度$\nabla\zeta(V,W)$相对于所有参数（V，W）。

接下来，我们展示的模型这两者都是通过探索三种应用在太空深深时间的力量：活动识别，图像描述，和视频的说明。

4.活动表示

活动识别是上述的第一个顺序学习任务的一个例子;T单独的帧是输入入T卷积网络然后将其连接到具有256隐藏单元单层LSTM。庞大的身躯最近的工作提出了行为识别深层结构的大型视频分类与卷积神经网络和二流卷积网络在视频行为识别以及3D卷积神经网络对人类行为的认可和连续深度学习人类行为识别和分类行动中长短期记忆回归神经网络的足球视频（[16，33，13，2，1]）。二流卷积网络在视频行为识别和大型视频分类与卷积神经网络[33，16]既提出一种学习基于输入N帧的堆叠上的过滤器的卷积网络。虽然我们分析在此工作16



帧的短片，我们注意到，LRCN 系统更比柔性二流卷积网络在视频行为识别和大型视频分类与卷积神经网络[33，16]，因为它不约束于分析的固定长度的输入和有可能学习识别复杂的视频序列（例如，如在 6 中烹饪序列）。分类行动中长短期记忆回归神经网络的足球视频和连续深度学习人类行为识别[1，2]用递归神经网络学习的任何传统的视觉特征（分类行动中长短期记忆回归神经网络的足球视频[1]）或深部特征（连续深度学习人类行为识别。[2]），但不要对培养他们的模型终端到终端的，不预先火车时间动态较大的物体识别数据库中重要的性能提升。

我们探索 LRCN 架构的两个变种：一个在其中 LSTM 被放置在美国有线电视新闻网（LRCN-fc6）的第一个完全连接层后，另一个在其中 LSTM 放置在 CNN 的第二完全连接层后（LRCN-fc7）。我们训练 LRCN 网络，16 帧的视频剪辑。该 LRCN 预测在每个时间步长视频类和我们平均这些预测进行最后的分类。在测试时，我们提取 16 帧的短片，从每个视频 8 帧和平均剪辑跨了一大步。

我们也考虑 RGB 和流输入。流计算与神经机器翻译的属性：编码器，解码器的方法[4]，并转化成“流动图象”由约 128 定心 x 和 y 流量值和由一个标量相乘，使得流量值 0 和 255 之间。通过计算流量大小建立为流图像的第三信道。该 LRCN 的 CNN 的分别是来自卷积架构快速功能嵌入[14]参考模型的混合，深卷积神经网络分类[22]的一个较小的变体，和通过蔡勒＆弗格斯[48]所用的网络。在 1.2M 图像 ILSVRC-2012 网是预训练[32]所述一个大型的分层图像数据库[7]数据集的分类训练集，给予该网络的强初始化，以促进更快的训练和防止过拟合到相对小的视频数据集。当进行分类中心作物，顶高 1 分类精度分别为 60.2％和用于混合动力车和来自卷积架构快速功能嵌入参考模型 57.4％。在我们的基准模型，T 视频帧分别由 CNN 分类。正如 LSTM 模型中，整个视频分类是通过平均在所有的视频帧的分数完成。

4.1。评估

我们评估我们在数据集 101 人的动作类从野外视频。[37]它由分为 101 人的动作类 12000 视频架构。数据集时，为每个分割设定的训练下，8000 的视频分为三劈叉，一点点。

表 1，列 2-3，比较我们提出的模型（LRCN-FC6，LRCN-FC7）对基准架构 RGB 和流输入的视频分类。每个 LRCN 网络进行训练端到端。该 LRCN-FC6 网的工作产生了两个 RGB 最好的结果和流和 0.49％和 5.27％，分别在基线网络提高。

Model	Single Input Type		Weighted Average	
	RGB	Flow	1/2, 1/2	1/3, 2/3
Single frame (split-1)	69.00	72.20	75.71	79.04
LRCN-fc_6 (split-1)	71.12	76.95	81.97	82.92
LRCN-fc_7 (split-1)	70.68	69.36	79.01	80.51
Single frame (all splits)	67.70	72.19	75.87	78.84
LRCN-fc_6 (all splits)	**68.19**	**77.46**	80.62	**82.66**

表 1：活动识别：比较单一框架模型 LRCN 的行为识别网络中的 UCF-101[37]的数据集，具有 RGB 和流输入。为分割 1 以及跨越所有三个分割的平均值被示出。我们 LRCN 模式始终极力优于基于单从底层的卷积网络体系结构的预测模型。在分 1，我们证明了放置 LSTM 上 fc6 性能比 fc7 更好。

RGB 和流网络可以通过如在二流卷积网络在视频行为识别[33]中提出的计算网络的分数的加权平均进行组合。像二流卷积网络在视频行为识别[33]，我们报告从 RGB 的预测和流量网络表 1 中（右）两加权平均值。由于流网络优于 RGB 网络工作，加权流网络更高的意料之中带来更好的精度。在这种情况下，LRCN 由 3.82%优于基线单帧模型。

该 LRCN 显示了基线单框系统明显改善和办法，由等深车型达到的精度。二流卷积网络在视频行为识别[33]通过计算（所有平均分割 86.4%为 1 分和 87.6%，）流量和 RGB 网络之间的加权平均报告 UCF-101 的结果。大型视频分类与卷积神经网络[16]报道在 UCF-101，它基本上比我们 LRCN 模式降低了 65.4%的准确率。

5.图片说明

相反，行为识别，静态图像描述的任务只需要一个网络卷积由于输入由一个单一的形象。各种深和多模态模型的深视觉语义嵌入模式和接地寻找和描述的句子成分的图像语义 多式联运神经语言模型 深片段嵌入双向图像映射判决 解释与复发性多神经网络的图像 多式联运神经语言模型统一视觉语义嵌入多式联运神经语言模型[8，36，19，15，25，20，18]已经提出了用于图象描述;特别是多式联运神经语言模型 统一视觉语义嵌入多式联运神经语言模型[20，18]结合代深型号卷积表示。多式联运神经语言模型[20]，利用“香草”作为 RNN 第 2 节，有可能使学习长期的时间相关困难。同期以和最相似的我们的工作是统一视觉语义嵌入多式联运神经语言模型[18]，其中提出一种使用 LSTM 编码器的隐藏状态在时间 T 作为长度 T 输入序列的编码表示一个不同的体系结构。然后将其映射该序列表示，与来自修道院的视觉表示相结合，变成从其中一个单独的解码器预测字的关节间隙。这是我们可以说简单的架构，这需要根据时间步输入静态输入图像的副本，与之前的字一起明显的。我们提出表明我们的集成 LRCN 架构优于这些现有的方法，其中没有一个包括端部到端可优化系统通过视觉和时间参数的层次结构经验结果。

现在我们描述了我们对 LRCN 架构的图像描述的任务实例。在每个时间步长，无论是图像的特征和前一个单词被作为输入提供给顺序模式，在这种情况下，堆叠 LSTMs（每 1000 隐单元），这是用来学习所述时变输出序列的动态，自然语言。在时间步长 t，输入到最底 LSTM 是从先前时间步长 w_{t-1} 的嵌入地面实况字。对于句子的产生，输入 w_{t-1} 变为样本从模型的在先前预

测分布时间步长。

堆栈中的第二LSTM最底下LSTM的输出与图像表示，$\phi V(x)$产生的视觉和语言输入的联合代表最多时间 t。（可视化模型$\phi V(x)$在本实验中使用的是分别来自卷积架构快速功能嵌入[14]参考模型，非常相似的公知深卷积神经网络分类[22]，预训练上大型视觉识别挑战[32]如第 4 节）栈中的任何进一步 LSTM 变换低于 LSTM 的输出，以及第四LSTM输出是输入到产生过字的分布$p(w_t \mid w_{1t-1}, \phi V(x))$。

以下[19]，我们指的是使用最底LSTM专门处理语言输入（无视觉输入）作为模型的因子分解的版本。我们研究的这个附录中的重要性??通过比较一个无分裂变体。参见图 6 上我们研究了变异的详细信息。

没有任何露骨的语言建模或定义的语法结构，描述 LRCN 系统学习从像素强度值自然语言描述的往往是语义的描述和语法正确的映射。

5.1。评估

我们评估我们的检索和生成任务图像描述模型。我们首先通过定量评价其对取景图像描述作为一个排名的任务：数据，模型和评价指标[26]提出并在解释与复发性多神经网络的图像 深片段嵌入双向图像句话映射 接地寻找和描述的句子成分的图像语义深视觉语义模型嵌入 统一视觉语义

	R@1	R@5	R@10	Medr
Caption to Image (Flickr30k)				
DeViSE [8]	6.7	21.9	32.7	25
SDT-RNN [36]	8.9	29.8	41.1	16
DeFrag [15]	10.3	31.4	44.5	13
m-RNN [25]	12.6	31.2	41.5	16
ConvNet [18]	11.8	34.0	46.3	13
LRCN_{2f} (ours)	**17.5**	**40.3**	**50.8**	**9**
Image to Caption (Flickr30k)				
DeViSE [8]	4.5	18.1	29.2	26
SDT-RNN [36]	9.6	29.8	41.1	16
DeFrag [15]	16.4	40.2	54.7	8
m-RNN [25]	18.4	40.2	50.9	10
ConvNet [18]	14.8	39.2	50.9	10
LRCN_{2f} (ours)	**23.6**	**46.6**	**58.3**	**7**
Caption to Image (COCO)				
LRCN_{2f} (ours)	29.0	61.6	74.8	3
Image to Caption (COCO)				
LRCN_{2f} (ours)	39.1	69.0	80.9	2

表 2：图片说明：Flickr 的 30K[28]检索结果和 2014 年的 COCO[24]的数据集。 -[R@K 是在等级 K（高是好的），平均召回。 Medr 是平均等级（低为好）。请注意，[18]达到使用更强的 CNN 结构见文更好的检索性能。

嵌入多式联运神经语言模型[25，15，36，8，18]中看到的像和标题检索任务显示我们的模型的有效性。我们在从形象描述视觉外延：新的相似度超过事件描述语义推理上 30K[28]报告结果，以及新近发布的 2014 年在上下文中常见的对象[24]的数据集，都与每个图像 5 句话的注解。

检索结果被记录在表 2 中我们报告中位数秩，Medr，第一检索地面实况图像或字幕和召回@ K，对于其中一个正确标题或图像相关的前 k 结果中检索到的图像或字幕的数量。我们的模型一致地优于强基线从最近的工作[18，25，15，36，8]如可从表 2 可以看出。在这里，我们注意到，在新牛津网模型[18]优于上检索任务我们的模型。然而，牛津大学网[18]采用了性能更好的卷积网[35]，并得到了基地统一视觉语义嵌入多式联运神经语言模型[18]造成的额外优势。我们的时域模型（以及时间和视觉模型的整合）的强度可以对统一视觉语义嵌入多式联运神经语言模型[18]结果，使用相同的基 CNN 结构[22]预训练的对相同的数据进行更直接测量

为了评估句话一代，我们主要使用的 BLEU（机器翻译的自动评估的方法）[27]指标被设计为统计机器翻译的自动评估。BLEU 是精密改良形式的 N-gram 片段比较假设翻译与多个参考翻译。我们使用 BLEU 作为描述的相似性的量度。单字组分数（B-1）占的充分性（或保留的信息）由翻译，而较长的 N-gram 占流畅分数（B-2，B-3）。我们比较我们的结果解释与复发性多神经网络图片[25]（Flickr 上 30K），并使用 AlexNet FC7 和 FC8 层输出计算两强近邻基线。我们使用近邻检索在训练数据库最相似的图像和平均 BLEU 得分超过了字幕。在 Flickr30k 的结果列于表 3 中报道。此外，我们提出的新 COCO2014 在上下文中常见的对象[24]的数据集其中拥有 8 个训练图像和 40,000 验证图像的结果。类似 Flickr30k，每个图像标注有 5 个或更多的图像注释。我们隔离来自验证组 5000 的图像，用于测试的目的，结果列在表 3 中报道。

	Flickr30k [28]			
	B-1	B-2	B-3	B-4
m-RNN [25]	54.79	23.92	19.52	-
INN fc_8 base (ours)	37.34	18.66	9.39	4.88
INN fc_7 base (ours)	38.81	20.16	10.37	5.54
LRCN (ours)	**58.72**	**39.06**	**25.12**	**16.46**
	COCO 2014 [24]			
	B-1	B-2	B-3	B-4
INN fc_7 base (c5)	46.23	26.39	15.07	08.73
LRCN (ours)	**66.86**	**48.92**	**34.89**	**24.92**

表 3：图片说明：句子生成的结果（BLEU 分数（%） - 我们都与简洁罚调整）为 Flickr 的 30k 的[28]和 COCO2014 年[24]测试集。

表 3 基于 B-1 的分数，使用 LRCN 一代中的描述中所传达的信息方面与间 RNN[25]同等进行。此外，LRCN 显著优于基线和该男子有关发电的流畅性（B-2，B-3），说明 LRCN 保留了较多的双字母组，并从人的注解说明卦。

除了标准定量评价之外，我们还采用亚马逊机械零工（AMT）所生成的句子评价。给定一个图像和一组不同模型的描述，我们请零工的基础上的正确性，语法和相关性排序的句子。我们比较了我们的模型的句子来取得由统一视觉语义嵌入多式联运神经语言模型[18]可公开获得的人。如表 4 所示，我们的微调（FT）LRCN 模式看齐

执行与最近邻居（NN）上的正确性和实用性，以及语法更好。我们在图 7 例句中。

	Correctness	Grammar	Relevance
TreeTalk [23]	4.08	4.35	3.98
OxfordNet [18]	3.71	3.46	3.70
NN [18]	**3.44**	3.20	**3.49**
LRCN fc_8 (ours)	3.74	3.19	3.72
LRCN FT (ours)	**3.47**	**3.01**	**3.50**
Captions	2.55	3.72	2.59

表 4：图片说明：1-6 人工评估排名（低是好的）平均为每个方法和标准。我们评估由统一视觉语义嵌入多式联运神经语言模型[18]的作者反对这种类似的方法当代比较的目的选择 785 Flickr 图片。

6.视频介绍

在影片说明我们要生成的话，类似于第 5[11，30，17，3，6，17，41，42]的一个可变长度流提出的方法，用于产生用于视频句子描述，但就我们所知，我们目前深车型首次应用到视频描述的任务。

该 LSTM 框架允许我们采取了不同的路径，如第 3。然而讨论，由于可用的视频描述数据集的限制我们的视频作为一个可变长度的输入流建模。我们依靠更多的“传统”活动，视频识别处理输入和使用茎产生一个句子。

我们首先区分视频说明（参见图 10）以下的架构。对于每一个架构，我们假设我们有基于完整的视频输入对象，主题和动词出现在视频由 CRF 的预测。以这种方式，我们观察到视频作为整体在每个时间步骤中，由帧不递增帧。

（一）LSTM 编码器和解码器，CRF 最大。（图 10（a））的第一种结构是由在[30]提出的视频描述的方法的动机。他们首先认识到使用的 CRF 采取视频功能的最大后验估计（MAP）的视频语义表示。这表示，鸡蛋的人，切，切板，然后串联到输入句子（人砍切板），这是翻译使用基于短语的统计机器翻译（SMT）的自然句（对船上人员的削减）[21]。我们与 LSTM，这表明国家的最先进的替代 SMT 的语言之间的翻译机器性能[39，5]。该架构（在图 10（a）所示）具有的编码器 LSTM（橙色），如[39]进行编码（在一个词汇二进制索引向量）的一热矢量输入句子的。这允许可变长度的输入。(请注意，输入句子可能具有比语义表示的元素的不同数目的单词。）在编码器阶段结束时，最后的隐蔽单元必须被输入到解码器级（粉红色）之前记住所有必需的信息，其中隐藏的表示被解码成一个句子，在每个时间步一个字。我们使用用于编码和解码的两个相同的层 LSTM。

（二）LSTM 解码器 CRF 最大。（图 10（b））的在该变型，我们利用了语义表示可以被编码为单个的固定长度向量。我们提供在每个时间步到 LSTM，类似于整个图像是如何作为输入提供给在图像描述的 LSTM 提供整个视觉输入表示。

（三）LSTM 解码器 CRF 概率。（图 10（c））的使用机器翻译 LSTMs 相比基于短语的 SMT 的益处[21]是它自然可以在训练和测试时间，它允许 LSTM 学习在视觉代的不确定性，而不是依赖合并概率矢量在 MAP 估计。该架构是相同的（b），但我们替换概率分布

最高的预测。

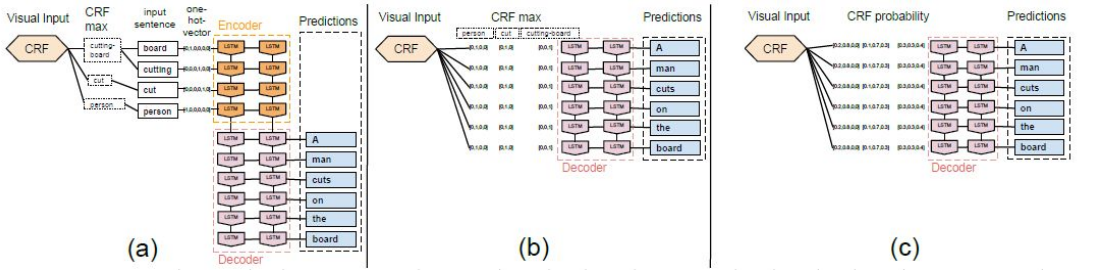


图 4：我们对视频的描述方式。（一）LSTM 编码器和解码器，最大 CRF（B）LSTM 解码器最大 CRF（三）LSTM 解码器 CRF 概率。 （对于较大的数字变焦或见补充）。

6.1。评估

我们评估的玉米饼多我们的方法取景图像描述作为一个排名的任务：数据，模型和评价指标[29]数据集，其中有 44762 视频/句对（约 40,000 培训/认证）。我们比较翻译的视频内容，以自然语言描述[30]谁使用最大预测以及在取景图像描述作为一个排名的任务：数据，模型和评价指标[29]提出了一种变体，它需要的 CRF 概率在测试时间，并使用一个字格找到一个最佳句子预测。由于我们使用了最大的预测以及由取景图像描述作为一个排名的任务：数据，模型和评价指标[29]所提供的概率得分，我们具有相同的视觉表示。取景图像描述作为一个排名的任务：数据，模型和评价指标[29]采用密集的轨迹密集的轨迹和运动边界描述符的行为识别[44]和 SIFT 特征以及时间上下文推理的 CRF 建模。

表 5 示出了 BLEU-4 的得分。结果表明：（1）LSTM 优于基于 SMT 的方法现有的视觉识别管道的难易程度使得来视频说明；（2）在更简单的解码器体系结构（b）和（c）达到比更好的性能（a）中，有可能因为输入并不需要被记忆；（3）我们的方法达到 28.8%，由[29]明显优于上的玉米饼多的 26.9%的最好的报号。

更广泛地说，这些结果表明，我们的体系结构不局限于深神经网络的输入，但可以与来自其他视觉系统其它固定或可变长度的输入来完全集成。

Architecture	Input	BLEU
SMT [30]	CRF max	24.9
SMT [29]	CRF prob	26.9
(a) LSTM Encoder-Decoder (ours)	CRF max	25.3
(b) LSTM Decoder (ours)	CRF max	27.4
(c) LSTM Decoder (ours)	CRF prob	**28.8**

表 5：视频描述：对玉米饼多层次取景图像描述作为一个排名的任务：数据，模型和评价指标[29]，在%详细说明结果，请参见 C.3 了解详情。

7. 结论

呈现 LRCN，一类的模型是在空间和时间深，并且具有适用于多种方法包括连续输入和输出视觉任务的灵活性。我们的研究结果一致表明，通过学习顺序动态带着深深的序列模型，我们可以提高其学习的参数深层次仅在视觉领域，并在其上搭输入一个固定的可视化表示，仅学习方法与以往的方法输出序列的动态。

随着计算机视觉领域的成熟超越静态输入和预测的任务，我们设想，“双深”的序列建模工具，如 LRCN 将很快成为最视觉系统的核心部分，卷积架构最近才有。与这些工具可并入它们对于感知问题的自然选择与随时

间变化的视觉输入或顺序输出，这些方法能够产生与小输入预处理和无手设计的功能。

致谢

作者感谢奥里奥尔 Vinyals 对这项整个工作提出宝贵的意见和有益的讨论。这项工作是由 DARPA 的电子工程硕士学位和 SMISC 计划的部分资助，NS 奖项 IIS-1427425 和 IIS-1212798，和伯克利愿景和学习中心。用于该研究的GPU是由NVIDIA 的捐赠是由德国学术交流中心（DAAD）的 FIT 世界各地的程序中的奖学金支持。

参考文献

[1] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Action classification in soccer videos with long shortterm memory recurrent neural networks. In *ICANN*. 2010. 4, 5

[2] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In *Human Behavior Understanding*. 2011. 2,4, 5

[3] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, L. Schmidt, J. Shangguan, J. M. Siskind, J. Waggoner, S. Wang, J. Wei, Y. Yin, and Z. Zhang. Video in sentences out. In *UAI*, 2012. 7

[4] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*. 2004. 5

[5] K. Cho, B. van Merri ¨enboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoderdecoder approaches. *arXiv preprint arXiv: 1409.1259*, 2014. 2, 3, 7

[6] P. Das, C. Xu, R. Doell, and J. Corso. Thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *CVPR*, 2013. 7

[7] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. FeiFei. ImageNet: A largescale hierarchical image database. In *CVPR*, 2009. 5

[8] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. DeViSE: A deep visual-semantic embedding model. In *NIPS*, 2013. 6, 7

[9] A. Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013. 3

[10] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *ICML*, 2014. 2,3

[11] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. YouTube2Text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shoot recognition. In *ICCV*, 2013. 7

[12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*,

1997. 2, 3

[13] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):221 - 231, 2013. 2, 4

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014. 2, 5, 6

[15] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *NIPS*, 2014. 6, 7

[16] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 2, 4, 5, 12

[17] M. U. G. Khan, L. Zhang, and Y. Gotoh. Human focused video description. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011. 7

[18] R. Kiros, R. Salakhuditnov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 6, 7

[19] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In *ICML*, 2014. 6, 15

[20] R. Kiros, R. Zemel, and R. Salakhutdinov. Multimodal neural language models. In *Proc. NIPS Deep Learning WorkShop*, 2013. 6

[21] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *ACL*, 2007. 7, 8

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. 4, 5, 6

[23] P. Kuznetsova, V. Ordonez, T. L. Berg, U. C. Hill, and Y. Choi. Treetalk: Composition and compression of trees for image descriptions. *Transactions of the Association for Computational Linguistics*, 2(10):351 - 362, 2014. 7

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick. Microsoft coco: Common objects in context. *arXiv preprint arXiv: 1405.0312*, 2014. 6, 7, 13, 16

[25] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. *arXivpreprint arXiv: 1410.1090*, 2014. 6, 7

[26] P. Y. Micah Hodosh and J. Hockenmaier. Framing image description as a ranking task:

Data, models and evaluation metrics. *JAIR*, 47:853 899, 2013. 6

[27] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2002. 6

[28] M. H. Peter Young, Alice Lai and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67 – 78, 2014. 6, 7

[29] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. Coherent multi-sentence video description with variable level of detail. In *GCPR*, 2014. 8, 18, 20

[30] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In *ICCV*, 2013. 2, 7, 8

[31] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. Technical report, DTIC Document, 1985. 2

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, 2014. 5, 6

[33] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv: 1406.2199*, 2014. 2, 4, 5, 12

[34] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv: 1409.1556*, 2014. 4

[35] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.6

[36] R. Socher, Q. Le, C. Manning, and A. Ng. Grounded compositional semantics for finding and describing images with sentences. In *NIPS Deep Learning Workshop*, 2013. 6, 7

[37] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5, 6, 14

[38] I. Sutskever, J. Martens, and G. E. Hinton. Generating text with recurrent neural networks. In *ICML*, 2011. 2

[39] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014. 2, 3, 7

[40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014. 4

[41] C. C. Tan, Y.-G. Jiang, and C.W. Ngo. Towards textually describing complex video contents with audio-visual concept classifiers. In *MM*, 2011. 7

[42] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. J. Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *COLING*, 2014. 7

[43] O. Vinyals, S. V. Ravuri, and D. Povey. Revisiting recurrent neural networks for robust ASR. In *ICASSP*, 2012. 2

[44] H. Wang, A. Kl ¨aser, C. Schmid, and C. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 2013. 8

[45] R. J. Williams and D. Zipser. A learning algorithm for con Tinually running fully recurrent neural networks. *Neural Computation*, 1989. 2

[46] W. Zaremba and I. Sutskever. Learning to execute. *arXiv preprint arXiv:1410.4615*, 2014. 3

[47] W. Zaremba, I. Sutskever, and O. Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv: 1409.2329*, 2014. 2, 4

[48] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*. 2014. 5