# 1 Sequence to Sequence Learning with Neural Networks

## 作者机构

Ilya Sutskever，Oriol Vinyals，Quoc V. Le. Google. 2014

## 针对的问题

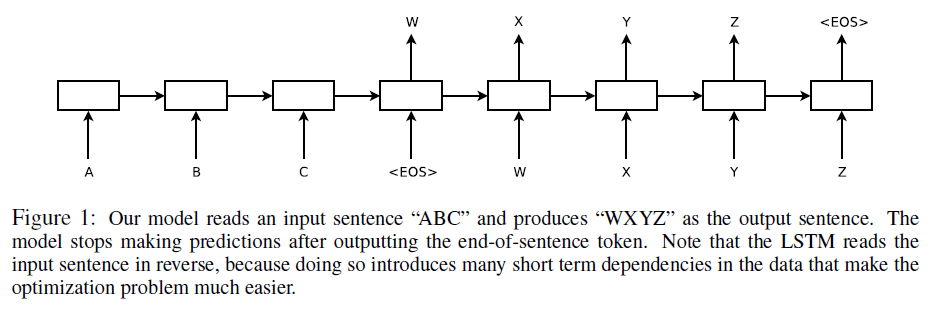
1. DNNs are powerful because they can perform arbitrary parallel computation for a modest number of steps. Despite their flexibility and power, DNNs can only be applied to problems whose inputs and targets can be sensibly encoded with vectors of fixed dimensionality.
2. Use RNN to build a end2end model for seq2seq learning in machine translation.

## 创新点

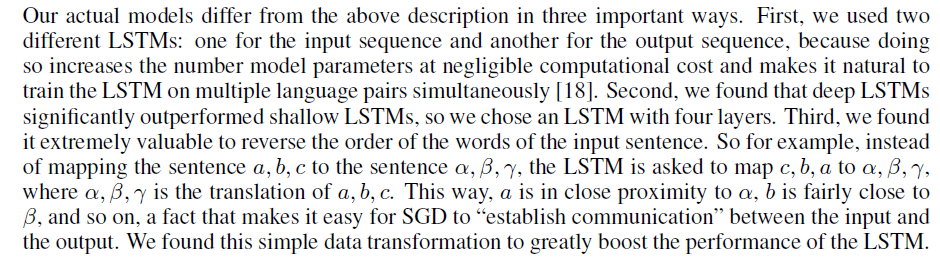
1. Encoder-decoder model.
2. Map the entire input sentence to a fixed length vector.

## 方法

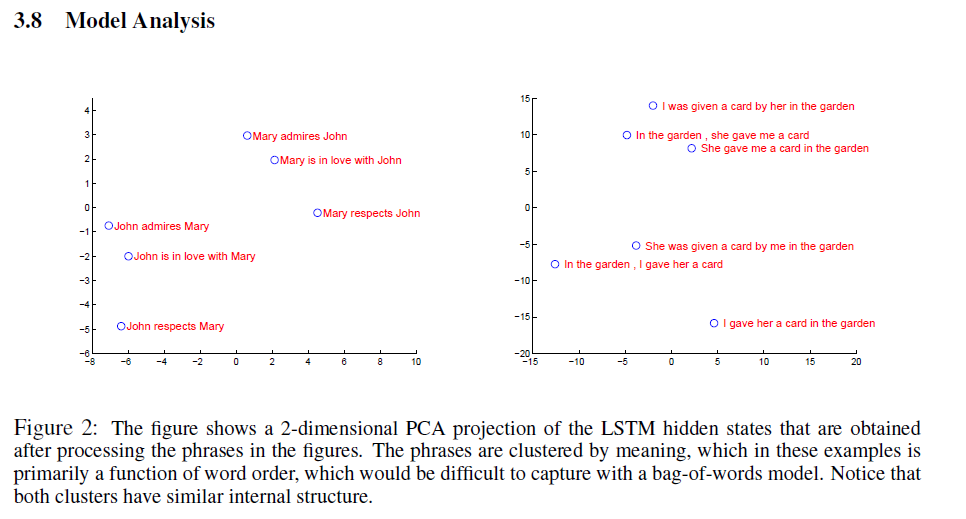
1. 模型



1. However, it is not clear how to apply an RNN to problems whose input and the output sequences have different lengths with complicated and non-monotonic relationships，A simple strategy for general sequence learning is to map the input sequence to a fixed-sized vector using one RNN, and then to map the vector to the target sequence with another RNN。
2. 模型特点



## 评价



# 2 Long-term Recurrent Convolutional Networks for Visual Recognition and Description

## 作者机构

Jeff Donahue ，Austin, Lowell ，Berkeley. 2015

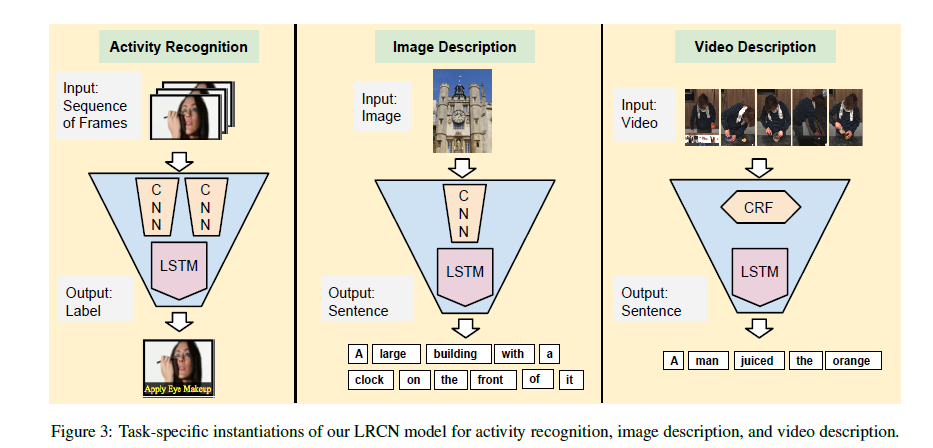
## 针对的问题

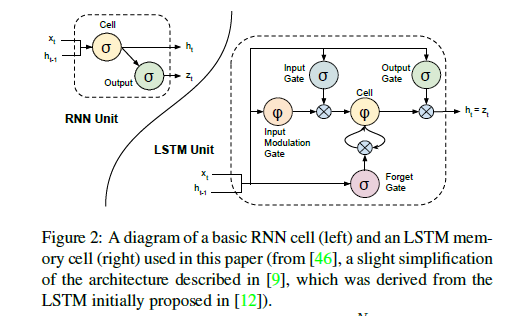
1. RNN（LSTM）用于动作识别（activity recognion）
2. RNN（LSTM）用于图像描述（image description）
3. RNN（LSTM）用于视频描述（video description）

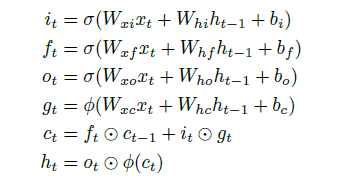
## 创新点

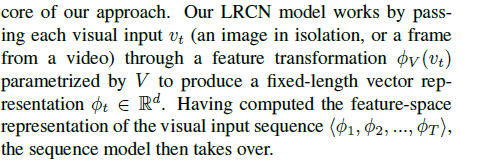
将CNN与RNN作了有效结合，利用CNN生成图像的representation vector，然后用RNN建模生成图片描述。解决了三类序列数据（图像）的具体问题：seq2scalar、one2seq、seq2seq。

## 方法







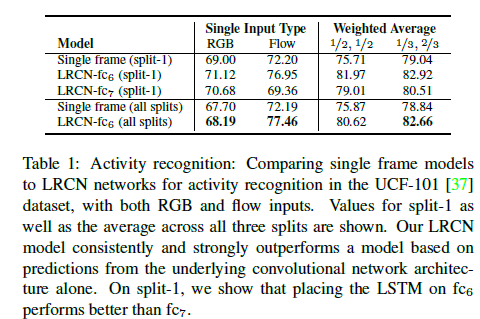


1. Activity recognition

特征：CNN feature: LRCN-fc6/LRCN-fc7.(AlexNet), RGB+flow image

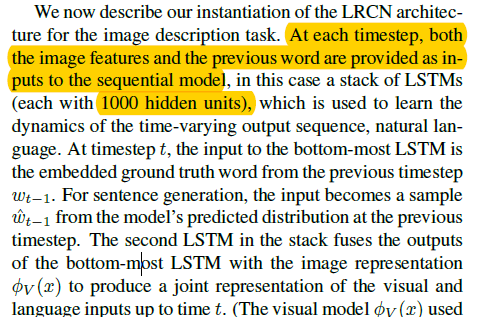
数据集：UCF-101

结果：



1. Image description

模型说明：

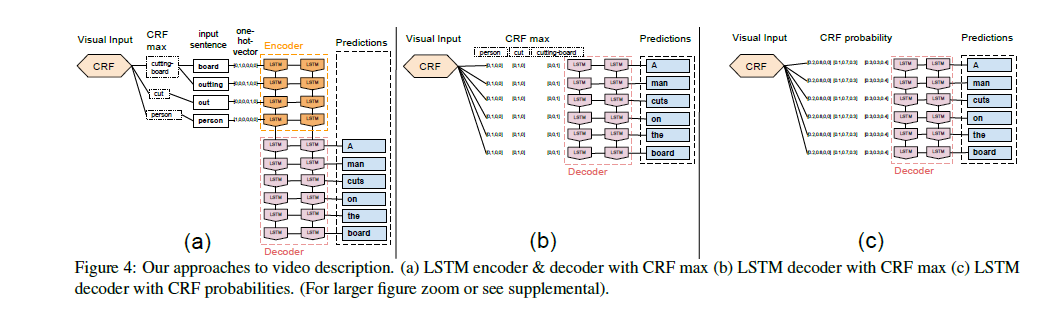


数据集：Flickr30k、COCO2014

评价指标：BLEU(top 5 descriotion coverage)

结果：见原论文。（better than listed）

1. Video description



## 评价

## 文中要点

1. Such models may have advantages when target concepts are complex and/or training data are limited. Learning long-term dependencies is possible when nonlinearities are incorporated into the network state updates.
2. Ideally，a video model should allow processing of variable length input sequences, and also provide for variable length outputs, including generation of full-length sentence descriptions that go beyond conventional one-versus-all prediction tasks.
3. timestep. LSTMs provide a solution by incorporating memory units that allow the network to learn when to forget previous hidden states and when to update hidden states given new information.

# 3 Social LSTM: Human Trajectory Prediction in Crowded Spaces

## 作者机构

Alexandre Alahi\_, Kratarth Goel\_, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, Silvio Savarese, Stanford University

## 针对的问题

依据目标过去的位置预测其行动轨迹

## 创新点

we propose an LSTM model which can learn general human movement and predict their future trajectories. This is in contrast to traditional approaches which use hand-crafted functions such as Social forces。

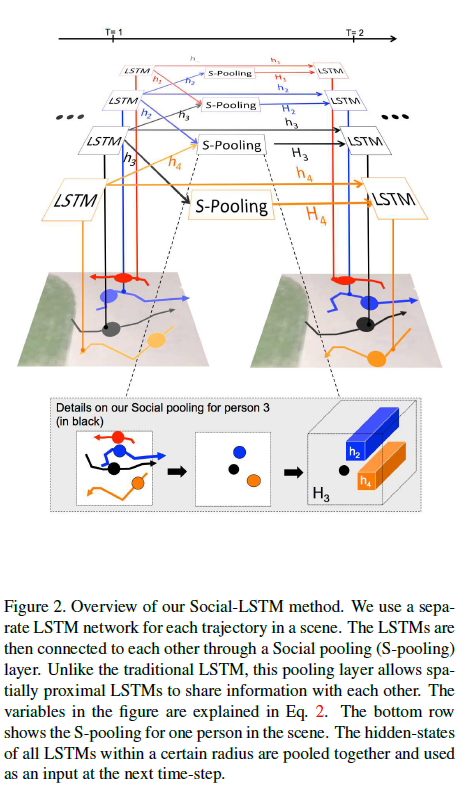
（利用RNN(LSTM)建模运动目标轨迹预测问题）

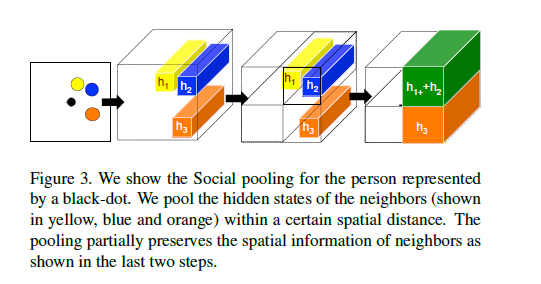
## 方法

1. 传统方法： Social forces(背景建模，考虑目标之间的相互影响)

Shortcoming:

1. They use hand-crafted functions to model ”interactions” for specific settings rather than inferring them in a data-driven fashion.
2. They focus on modeling interactions among people in close proximity to each other (to avoid immediate collisions). However, they do not anticipate interactions that could occur in the more distant future.
3. 本文方法
4. Model(Social LSTM):



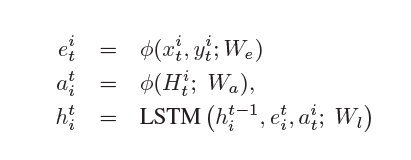


1. 模型说明：

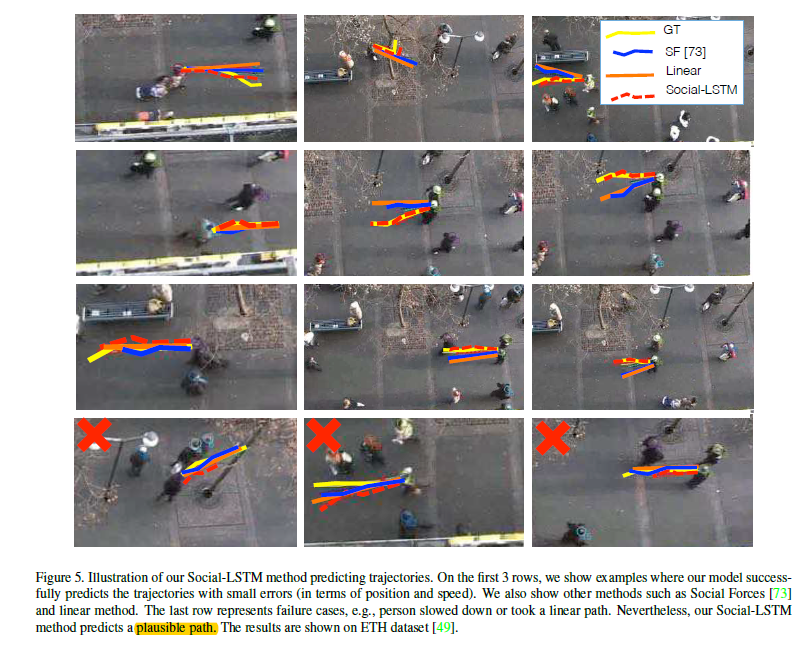
1）In particular, we have one LSTM for each person in a scene. This LSTM learns the state of the person and predicts their future positions as shown in Fig. 2. The LSTM weights are shared across all the sequences.

2）time. We expect the hidden states of an LSTM to capture these time varying motion-properties. We handle this by introducing “Social” pooling layers。While pooling the information, we try to preserve the spatial information through grid based pooling as explained below.

3）



1. 结果：



## 评价

1. The ability to model these rules and use them to understand and predict human motion in complex real world environments is extremely valuable for a wide range of applications - from the deployment of socially-aware robots to the design of intelligent tracking systems in smart environments.
2. Social-LSTM successfully predicts various non-linear behaviors arising from social interactions, such as a group of individuals moving together.
3. 对于直线行进的目标预测效果欠佳。

# 4 ReNet: A Recurrent Neural Network Based Alternative to Convolutional Networks

## 作者机构

Francesco Visin，Kyle Kastner, Yoshua Bengio. University of Montreal. Politecnico di Milano

## 针对的问题

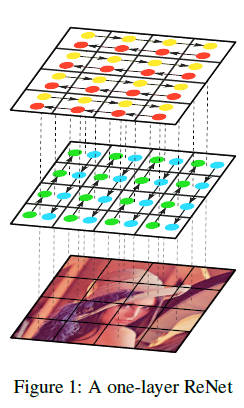
寻找CNN的替代品。

## 创新点

用RNN来干CNN干的事情。物体检测。

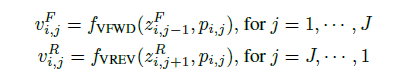
## 方法

Model：



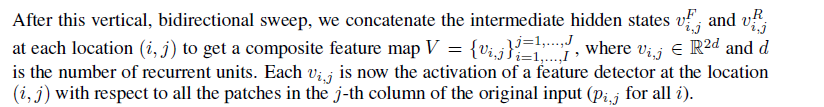
模型说明：

1. **Bidirectional LSTM**



按竖直方向上下方向各自扫描一遍建立两个方向的RNN，每个LSTMD单元以前一个patch的因变量输出和当前patch作为输入，计算激活值。同理在水平方向也建立两个方向的RNN。

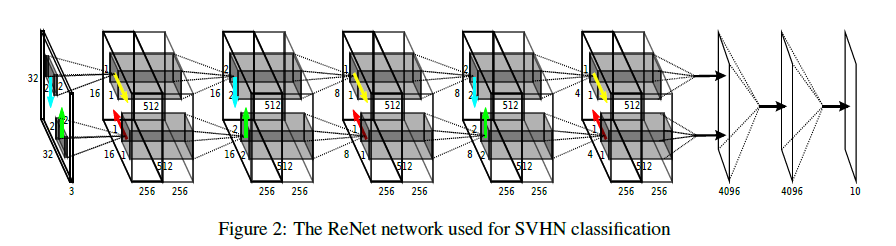
1. **串接起正向和反向的隐层输出作为patch的feature map.**



经过RNN后的输出h包含了patch的图像背景信息（周围像素的关联信息）

1. 通过加深RNN的层数可以获得更加抽象的特征映射，类似CNN.

we can stack multiple ‑’s to make the proposed ReNet deeper and capture increasingly complex features of the input image. After any number of recurrent layers are applied to an input image, the activation at the last recurrent layer may be flattened and fed into a differentiable classifier.



## 评价

找到了一种用RNN替代CNN的可能方法（有一定的初步效果，但还需深入研究）

利用RNN能有效获得图像在空间上的依赖性，有效的建模像素间的相互关系。与CNN的pooling+convolution有异曲同工之妙。

有效的将RNN对序列数据的强处理能力应用到图像像素的空间连续性上。

# 5 Deep Visual-Semantic Alignments for Generating Image Descriptions

## 作者机构

Andrej Karpathy Li Fei-Fei ，Department of Computer Science, Stanford University

## 针对的问题

由图像生成图像描述，由图像描述生成区域描述

## 所作工作

1. We develop a deep neural network model that infers the latent alignment between segments of sentences and the region of the image that they describe。
2. We introduce a multimodal Recurrent Neural Network architecture that takes an input image and generates its description in text.

## 创新点

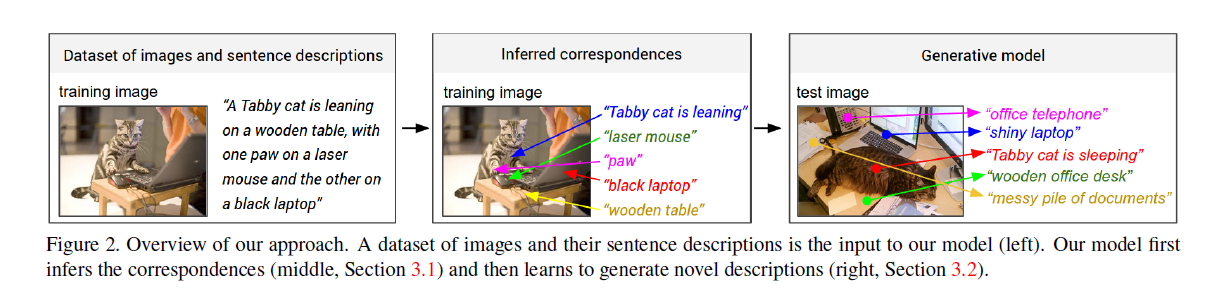
1. The focus of these works is on correctly labeling scenes, objects and regions with a fixed set of categories, while our focus is on richer and higher-level descriptions of regions.

(其他方法集中在描述scene type, objects and their spatial support in the image is inferred)

1. While our Recurrent Neural Network (RNN) model conditions the probability distribution over the next word in a sentence on all previously generated words.

(其他方法captions based on fixed templates，or generative grammars)

## 方法

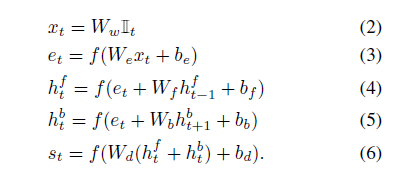


1. **从Descriptions到Image regions对应关系的建立（CNN+BRNN+MRF）**

Image representation(CNN):



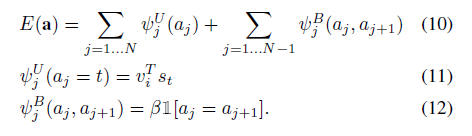
Sentences representation（BRNN）:



相似性度量：



MRF生成 region描述：



1. **从Image regions到全图Descriptions的建立**

