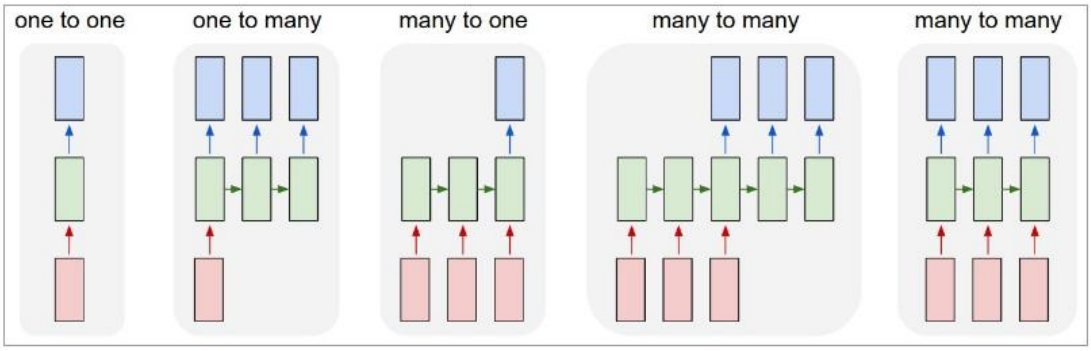
RNN Review : RNN 网络在不同领域的应用

王忠 2016/11/27

（一）常用结构框架



（图中，红色代表输入，蓝色代表输出，绿色代表隐状态）

1） 一对一，DNN 的模式，定长输入产生定长的输出，典型应用如图像分类等任务；

2） 一对多，变长输出，DNN 不能处理，可应用于图像标注等任务；

3） 多对一，变长输入，定长输出，可应用于文本情感分析等；

4） 多对多，RNN 里面的 encoder-decoder 架构，在图像标注，语言翻译等领域应用广泛；

5） 多对多，可用于视频逐帧标注等；

以上结构中，在时间方向，长度是任意可变的，使之能够有效地处理序列数据。

（二）RNN 在自然语言处理（NLP）中的应用

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

一般前馈深度网络在处理序列数据时存在很大的局限性，无法有效地建模前后数据间的关系。利用 RNN 对序列数据处理的独特优势来建模机器翻译。

创新点

提出了 RNN 里的 encoder-decoder 的架构。为了跨越不同语言之间的鸿沟，首先利用编码器将源语言编码，然后利用解码器从编码后的数据里翻译出目标语言。

方法

当输入序列和输出序列存在不同的长度，而且他们之间不存在直接性的关系时，利用 RNN 来建模是一件很困难的事情。一个简单的策略就是先利用一个 RNN 将输入进行编码，然后再利用另一个 RNN 将编码后的数据进行解码。

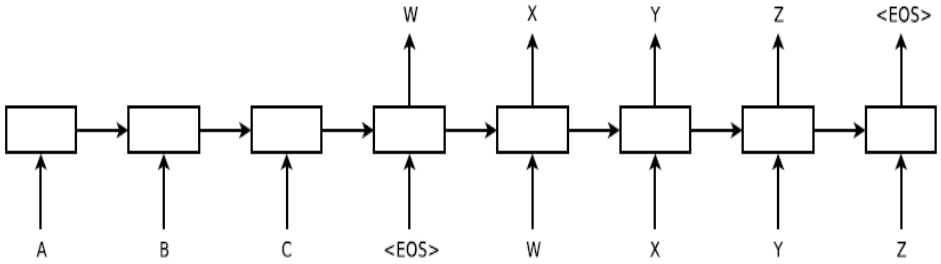


Figure 1: Our model reads an input sentence "ABC" and produces "WXYZ" as the output sentence. The model stops making predictions after outputting the end-of-sentence token. Note that the LSTM reads the input sentence in reverse, because doing so introduces many short term dependencies in the data that make the optimization problem much easier.

1） 解码和编码使用不同的 LSTM 结构，参数不共享；

2） 试验中，采用深层 LSTM（4 层）比单层 LSTM 效果好；

3） 训练时，发现将源语言的顺序翻转，加入到训练集中，能够更快地收敛。（个人理解，类似于图像处理中镜像，缩放等 data augmentation 方法）

评价

学术价值在于 encoder-decoder 结构的提出，工业价值在于将语言翻译推进到一个新的台阶(state-of-art)。类似的结构后来被用于语音识别（Towards End-to-End Speech Recognition with Recurrent Neural Networks——Google Deepmind, Alex Graves），同样达到了很好的效果。

（三）RNN 在图像以及视频分析领域的应用

ReNet: A Recurrent Neural Network Based Alternative to Convolutional Networks

作者机构

Francesco Visin, Kyle Kastner, Kyunghyun Cho, Matteo Matteucci, Aaron Courville, Yoshua Bengio，3 May 2015

针对的问题和创新点

用 RNN 代替 CNN 来抽取图像特征。

方法

Model：

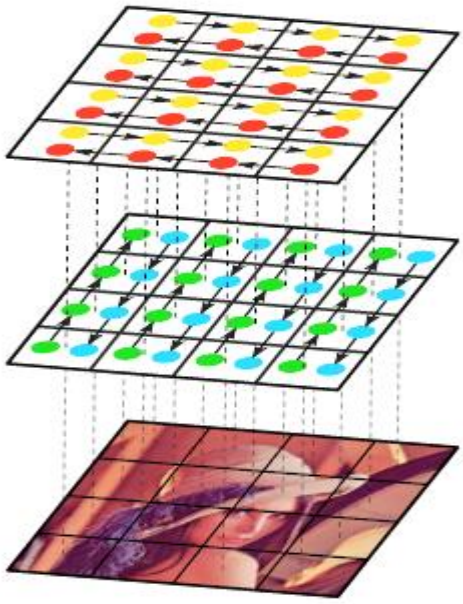


Figure 1: A one-layer ReNet

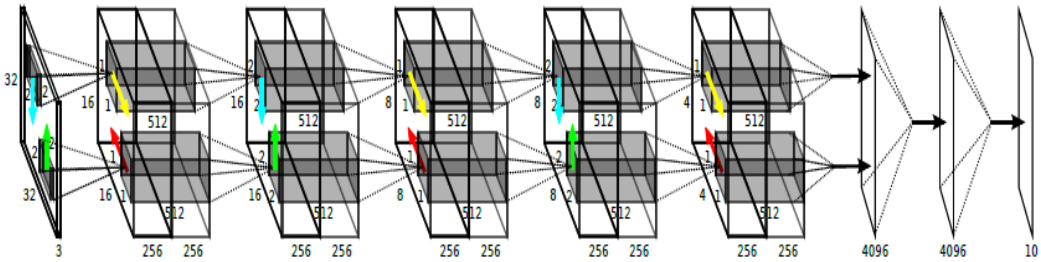


Figure 2: The ReNet network used for SVHN classification

模型说明：

1) Bidirectional LSTM

$$v_{i,j}^{F} = f_{\text{VFWD}}(z_{i,j-1}^{F}, p_{i,j}), \text{for } j = 1, \cdots, J$$
$$v_{i,j}^{R} = f_{\text{VREV}}(z_{i,j+1}^{R}, p_{i,j}), \text{for } j = J, \cdots, 1$$

按竖直方向自下而上、自上而下各扫描一遍建立两个方向的 RNN，每个 LSTM 单元以前一个 patch 的隐变量输出和当前 patch 作为输入，计算激活值（新的特征，类似于 CNN 中的 Feature Map）。同理在水平方向也建立两个方向的 LSTM。

2) 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）

3) 通过加深 RNN 的层数可以获得更加抽象的特征映射，也类似 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 来抽取图像特征的方法（有一定的初步效果 Mnist 测试集正确率 97%(LSTM)，99%（BLSTM），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迁移到图像像素的空间连续性上。

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

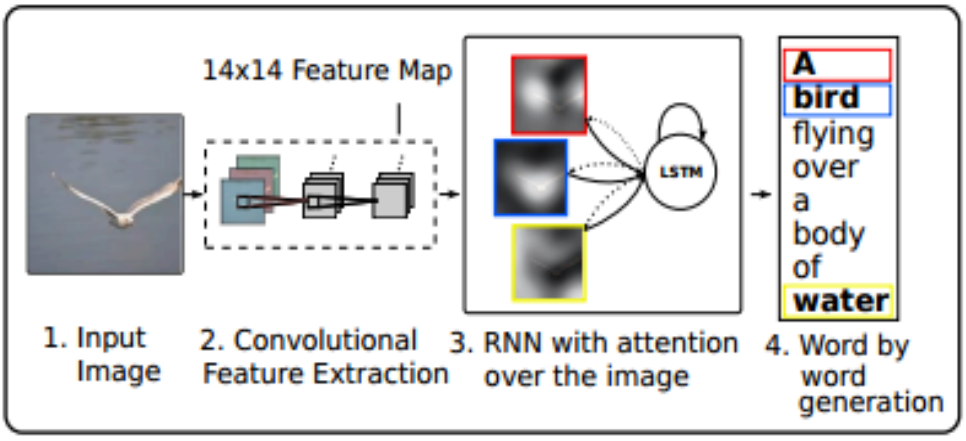
作者机构

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio. 19 Apr 2016.

针对问题和创新点

给定图像，生成图像的描述。引入 attention mechanism（聚焦机制）用于图像描述生成，每次由图像的焦点 patch 生成新的描述词。

方法



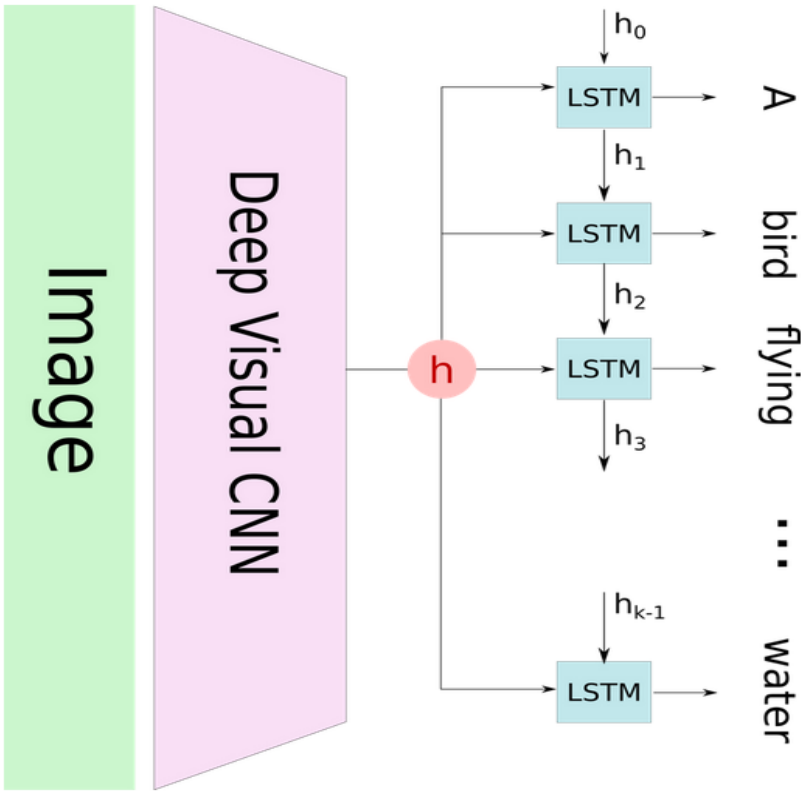
Attention mechanism

https://blog.heuritech.com/2016/01/20/attention-mechanism/

A wealth of results in the past few year suggest that visual structure can be better captured by a sequence of partial glimpses, or foveations, than by a single sweep through the entire image (Larochelle & Hinton, 2010; Denil et al., 2012; Tang et al., 2013; Ranzato, 2014 Zheng et al., 2014; Mnih et al., 2014; Ba et al., 2014; Sermanet et al., 2014).

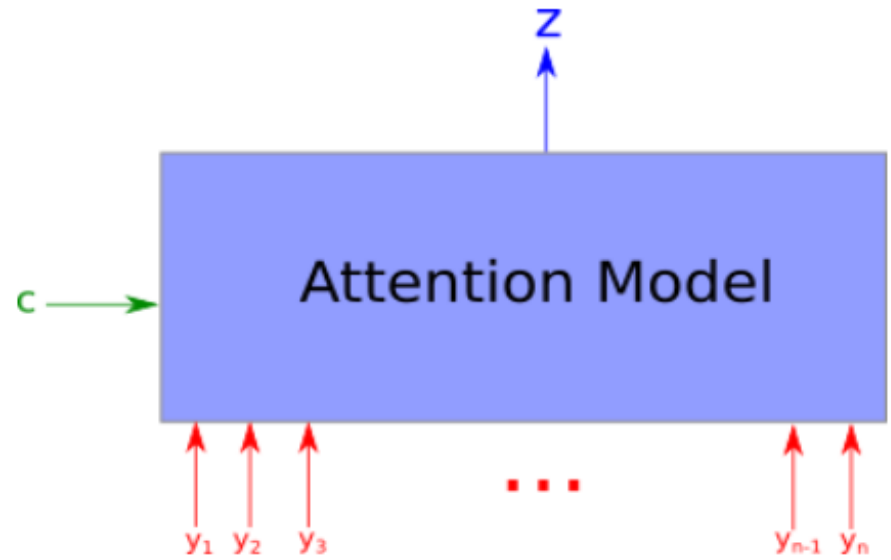
动物视觉系统观察物体时倾向于集中观察主要的部分以获取足够的观察对象信息。

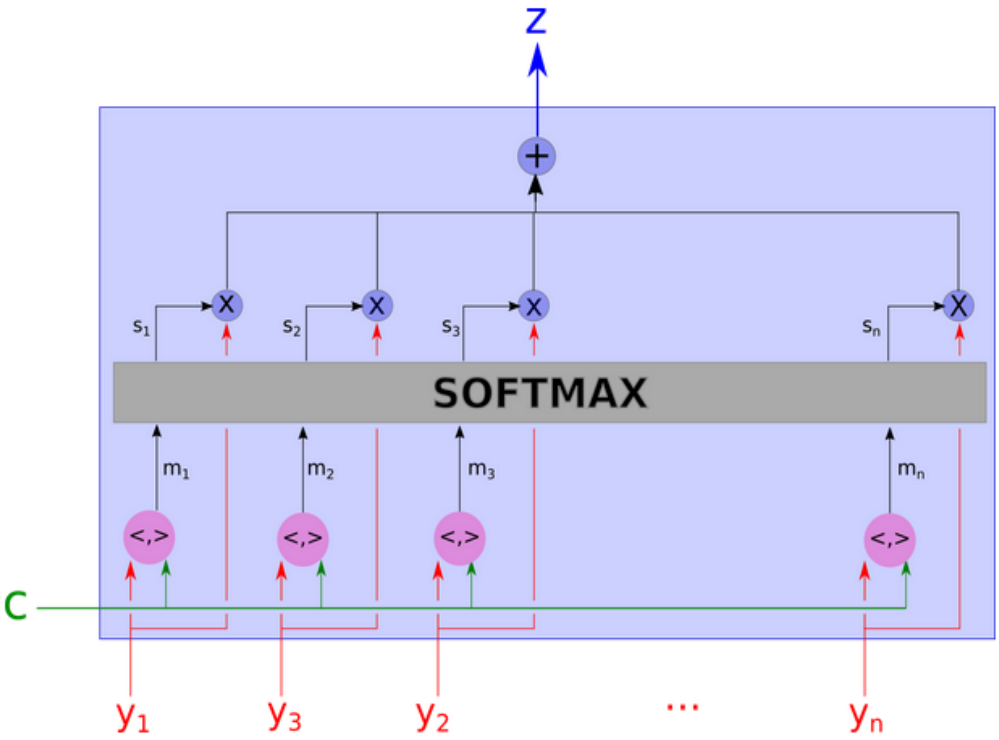
聚焦机制用于图像描述



上图是通常的解决图像描述生成问题的一般模型架构，首先由 CNN 提取图像信息，将图像进行 encode，然后将图像特征（通常由一个向量表示）送入 LSTM，由 LSTM 逐个产生描述单词。

以上模型存在的一个缺陷是，每次生成新的描述词时，一个词应该集中于描述图像的一个 object、patch 或者 region，所以在每个节点上将整个图作为输入是存在信息冗余的，应该找到一种机制使得可以有效地降低这种冗余，使得生成的描述更精确。



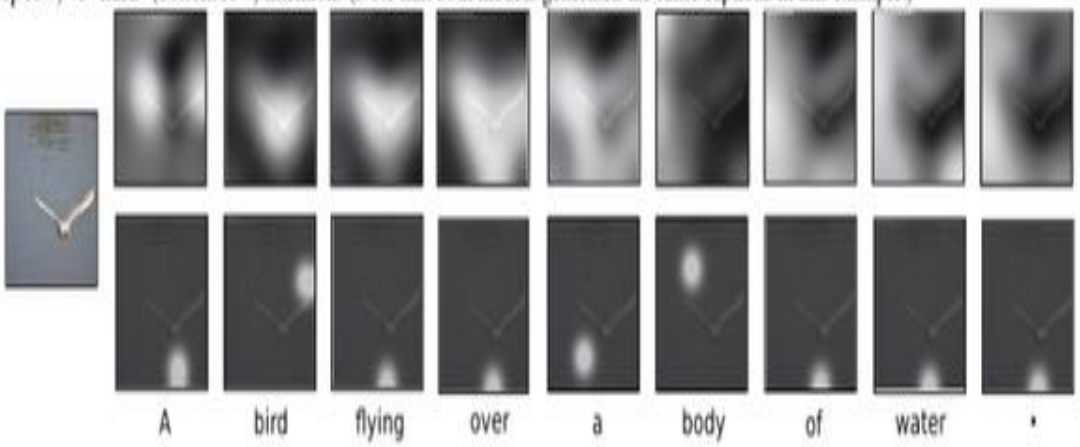


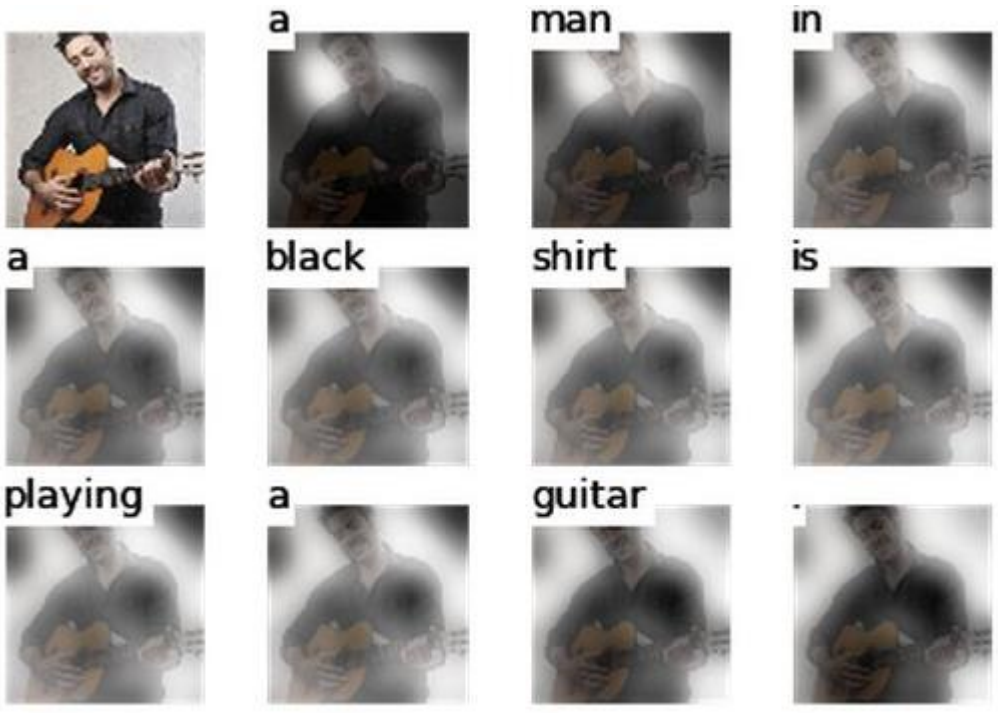
Attention 机制在图像描述上的实现：

将输入图像分成不同的 patch（上图中的 y），对每个 patch 提取 CNN 特征，然后将其与上一时刻的隐状态输出 C 一起作一定运算后(比如内积)激活，再经过 softmax 层（当只有一个变量比较大的时候，

就相当于在作 argmax 操作）后作为新的 attention 图像输出。

Figure 2. Attention over time. As the model generates each word, its attention changes to reflect the relevant parts of the image. "soft" (top row) vs "hard" (bottom row) attention. (Note that both models generated the same captions in this example.)





评价

该方法将 attention mechanism 和 encoder-decoder 架构运用到图像描述的生成上来。相比每一步直接以整图输入产生描述单词更精确，能有效地去除冗余信息，提高描述的准确性，在几个常用数据集上都达到了 state-of-art 的效果。但是同时一定程度地增加了模型的复杂性，有更多的参数需要训练。需要更多的训练技巧（DropOut，选择不同梯度下降算法等）。

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 recognition）
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

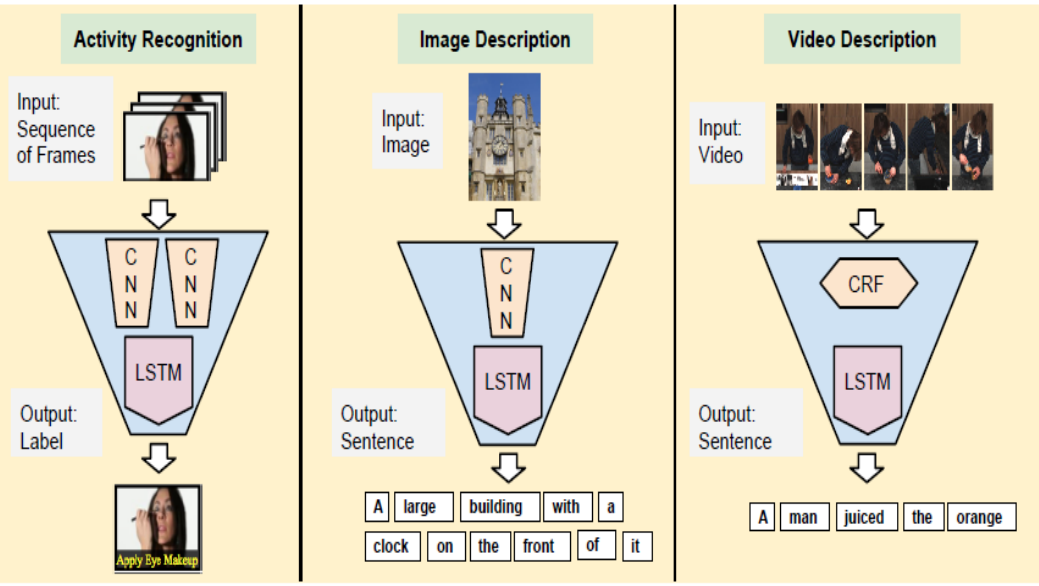


Figure 3: Task-specific instantiations of our LRCN model for activity recognition, image description, and video description.

1. Activity recognition（sequence-to-scalar）

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

数据集：UCF-101

结果：

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**

Table 1: Activity recognition: Comparing single frame models to LRCN networks for activity recognition in the UCF-101 [37] dataset, with both RGB and flow inputs. Values for split-1 as well as the average across all three splits are shown. Our LRCN model consistently and strongly outperforms a model based on predictions from the underlying convolutional network architecture alone. On split-1, we show that placing the LSTM on fc_6 performs better than fc_7.

2. Image description

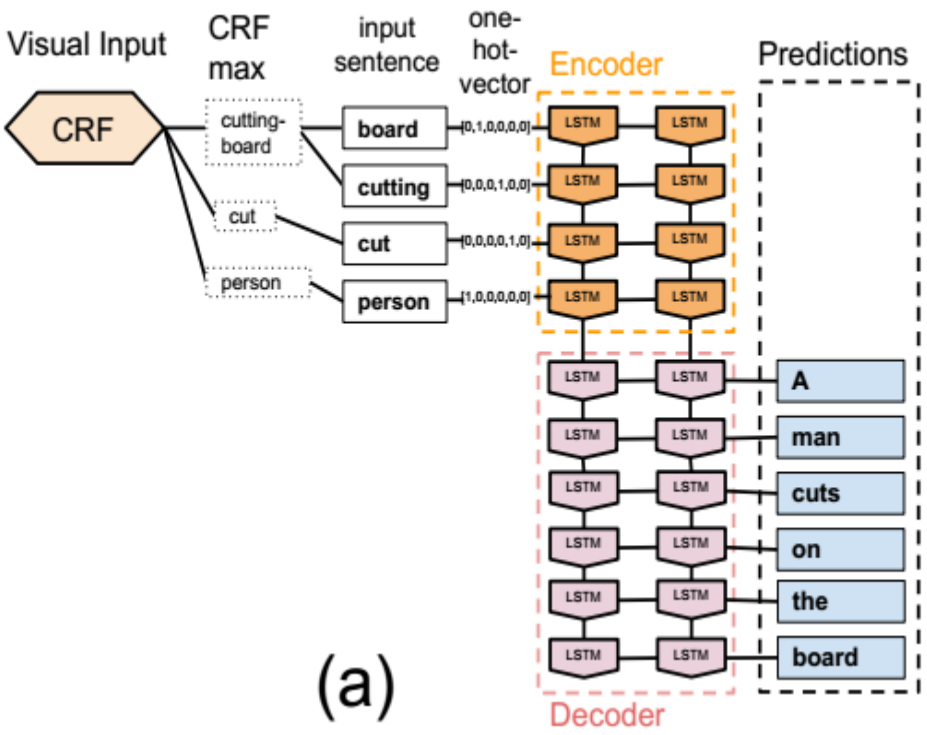
数据集：Flickr30k、COCO2014

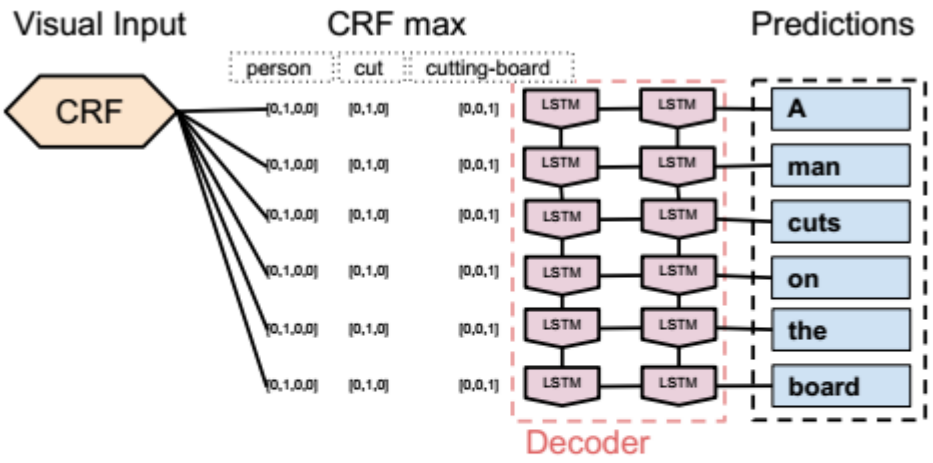
评价指标：BLEU(top 5 descriotion coverage)

结果：

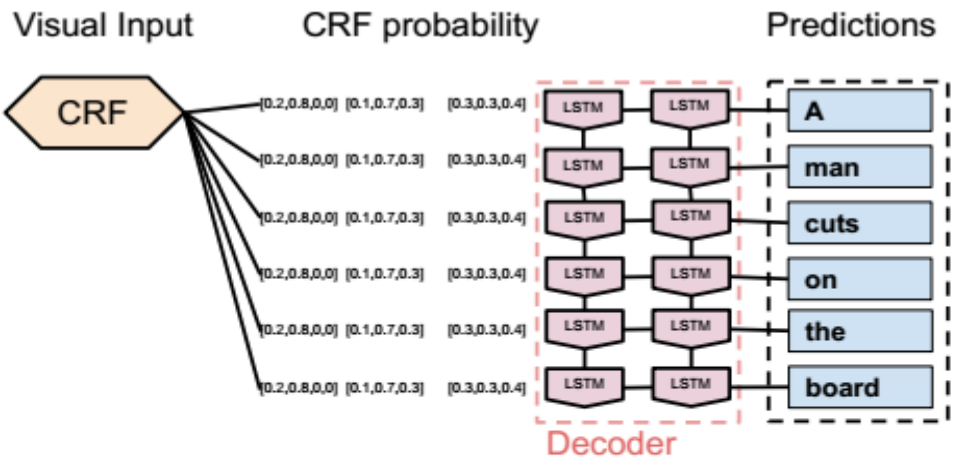
	R@1	R@5	R@10	Med*r*
Caption to Image (Flickr30k)				
DeViSE [8]	6.7	21.9	32.7	25
SDT-RNN [36]	8.9	29.8	41.1	16
DeFrag [15]	10.3	31.4	44.5	13
m-RNN [25]	12.6	31.2	41.5	16
ConvNet [18]	11.8	34.0	46.3	13
LRCN_{2f} (ours)	**17.5**	**40.3**	**50.8**	**9**

3. Video description





(b)



(c)

说明：

1) CRF+encoder-decoder

2) CRF->one_hot_vector->LSTM

3) CRF->probability->LSTM

（四）RNN 在目标行进轨迹预测、目标跟踪上的应用

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, CVPR2015

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 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:

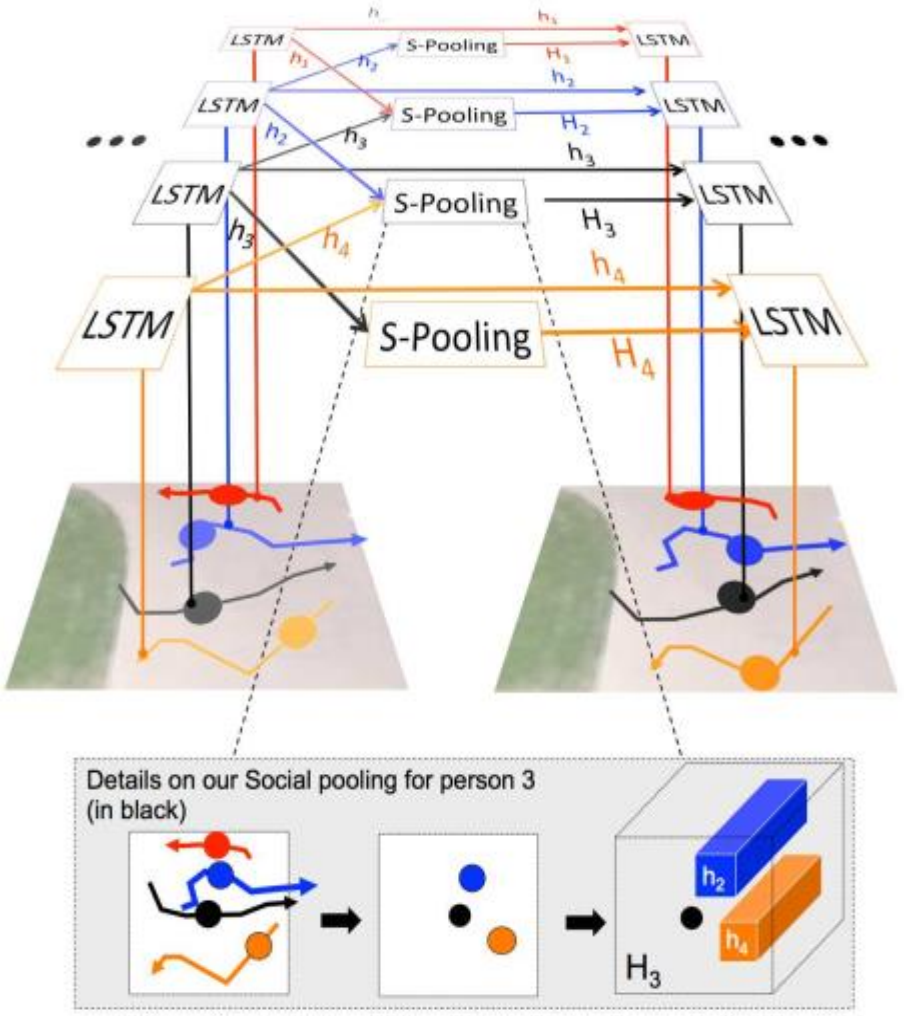
   1) 利用手工设计的特征函数来建模目标间的“相互影响力”，

而不是让数据驱动来进行推断。

2) 集中于建模目标间的相互影响力来避免行进的冲突，然而没有对将来会发生的冲突进行建模。

2. 本文方法

1) Model(Social LSTM):



2) 模型说明：

1）对于每一个在场景中出现的人都维持一个 LSTM 网络，该 LSTM 负责学习该人的状态和预测他们的位置，但是所

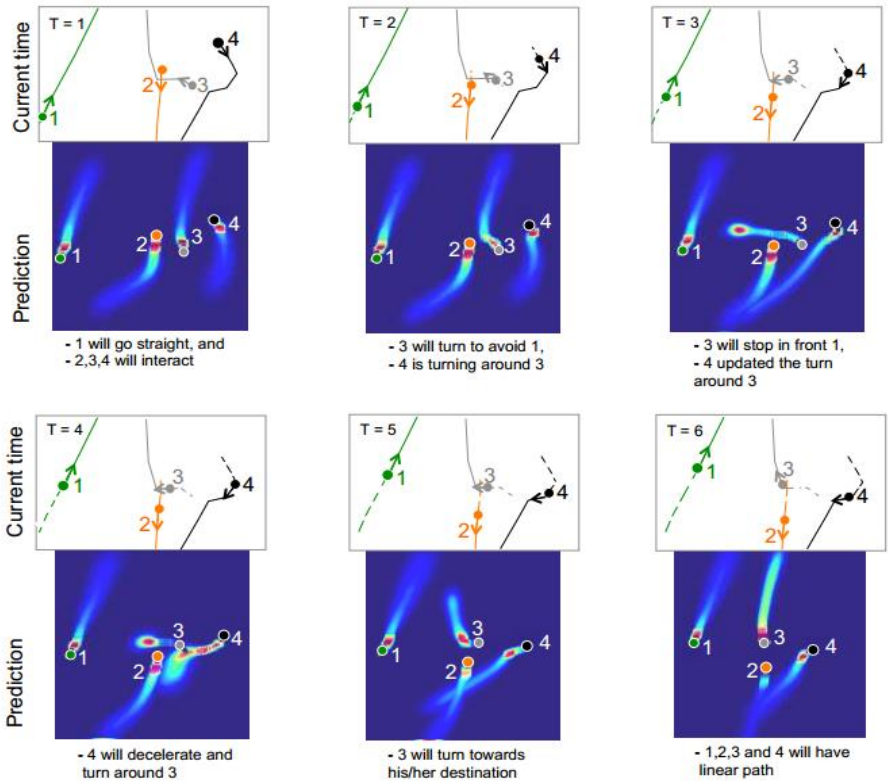
有的 LSTM 共享参数。

2）希望 LSTM 的隐状态能够捕获目标随时间变化的运动模式，本文通过 social-pooling 层来建模相互之间的影响。将当前目标一定范围（grid）内的其他临近目标的隐状态加入到此目标的隐状态中来。

$$H_t^i(m,n,:) = \sum_{j \in \mathcal{N}_i} \mathbf{1}_{mn}[x_t^j - x_t^i, y_t^j - y_t^i] h_{t-1}^j,$$

$$\begin{aligned} e_t^i &= \phi(x_t^i, y_t^i; W_e) \\ a_i^t &= \phi(H_t^i;\ W_a), \\ h_i^t &= \text{LSTM}\left(h_i^{t-1}, e_i^t, a_t^i;\ W_l\right) \end{aligned}$$

3）模型输出是目标下一时刻坐标的高斯分布参数。



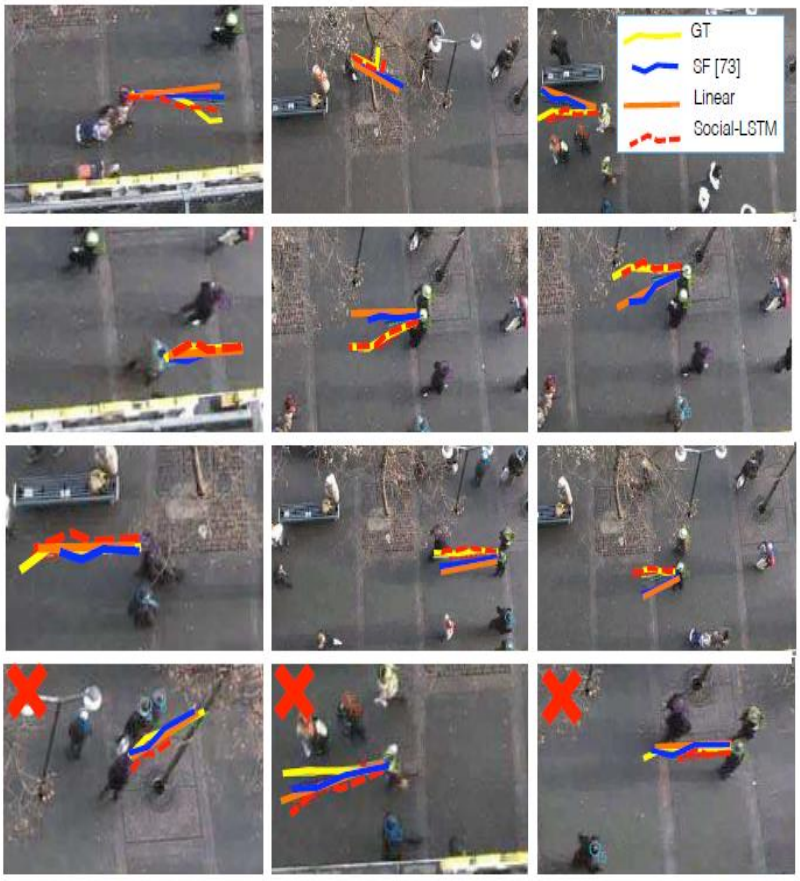


Figure 5. Illustration of our Social-LSTM method predicting trajectories. On the first 3 rows, we show examples where our model successfully predicts the trajectories with small errors (in terms of position and speed). We also show other methods such as Social Forces [73] and linear method. The last row represents failure cases, e.g., person slowed down or took a linear path. Nevertheless, our Social-LSTM method predicts a plausible path. The results are shown on ETH dataset [49].

评价

在两个数据集上达到了 state-of-the-art 的预测效果，对物体之间存在冲突、避让的非线性运动能很好地预测，但是反而对直线运动的预测效果欠佳（比不上一般的线性模型）。

Deep Tracking: Seeing Beyond Seeing Using Recurrent Neural Networks

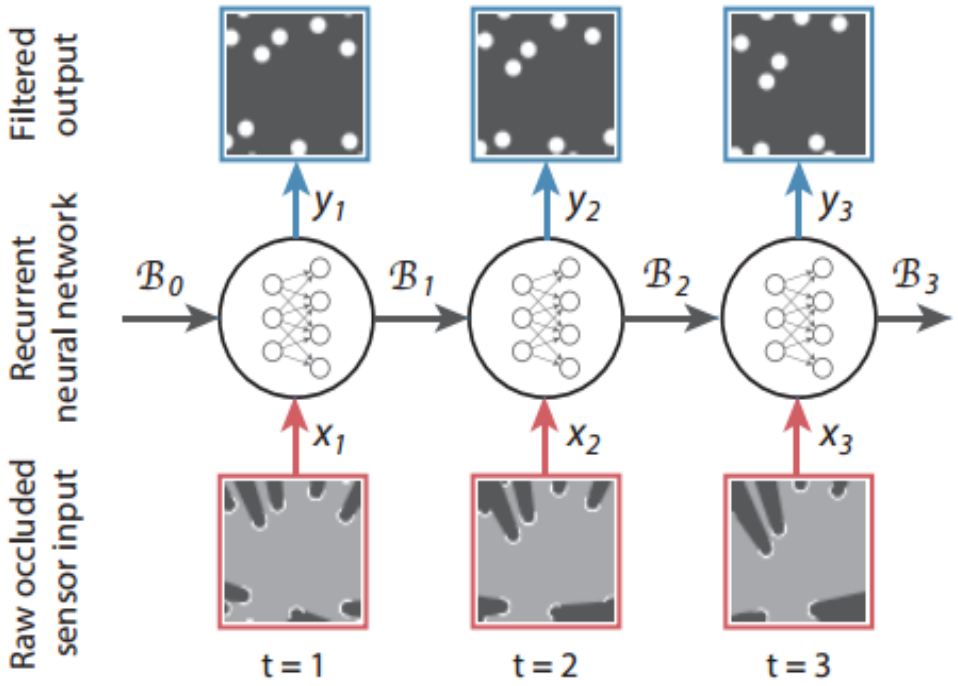
作者机构

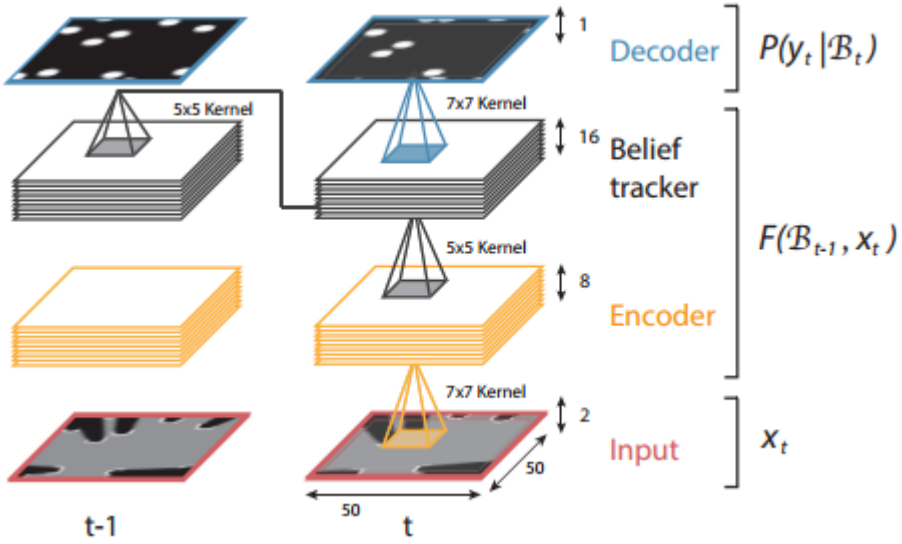
Peter Ondruska, Ingmar Posner，Mobile Robotics Group, University of Oxford, United Kingdom，AAAI-16 conference, February 12-17, 2016

针对问题

利用 RNN 建立端到端的目标跟踪方法，做到实时地从原始的传感器信号输入产生整个环境状态的预测，能有效地处理遮挡问题。

方法





训练：

1) 监督学习方式：

有每个时刻环境的真实状态 y_i（通过增设传感器）。

目标函数：

$$\mathcal{L} = -\sum_{t=1}^{N} \log P(y_t|x_{1:t})$$

在监督学习框架下，将当前的传感器观测和上一时刻的隐状态作为输入，目标是预测当前的真实环境状态。这是此类问题的一般解决思路，然而在实际中，真实环境状态难以获得，通过增设传感器的方式虽然可以解决，但是代价巨大，不实际。

因此，转而寻求非监督的学习方式，即只从一系列的传感器观测中学习预测状态的模型。

2) 非监督学习方式

与监督学习方式的不同点在于：

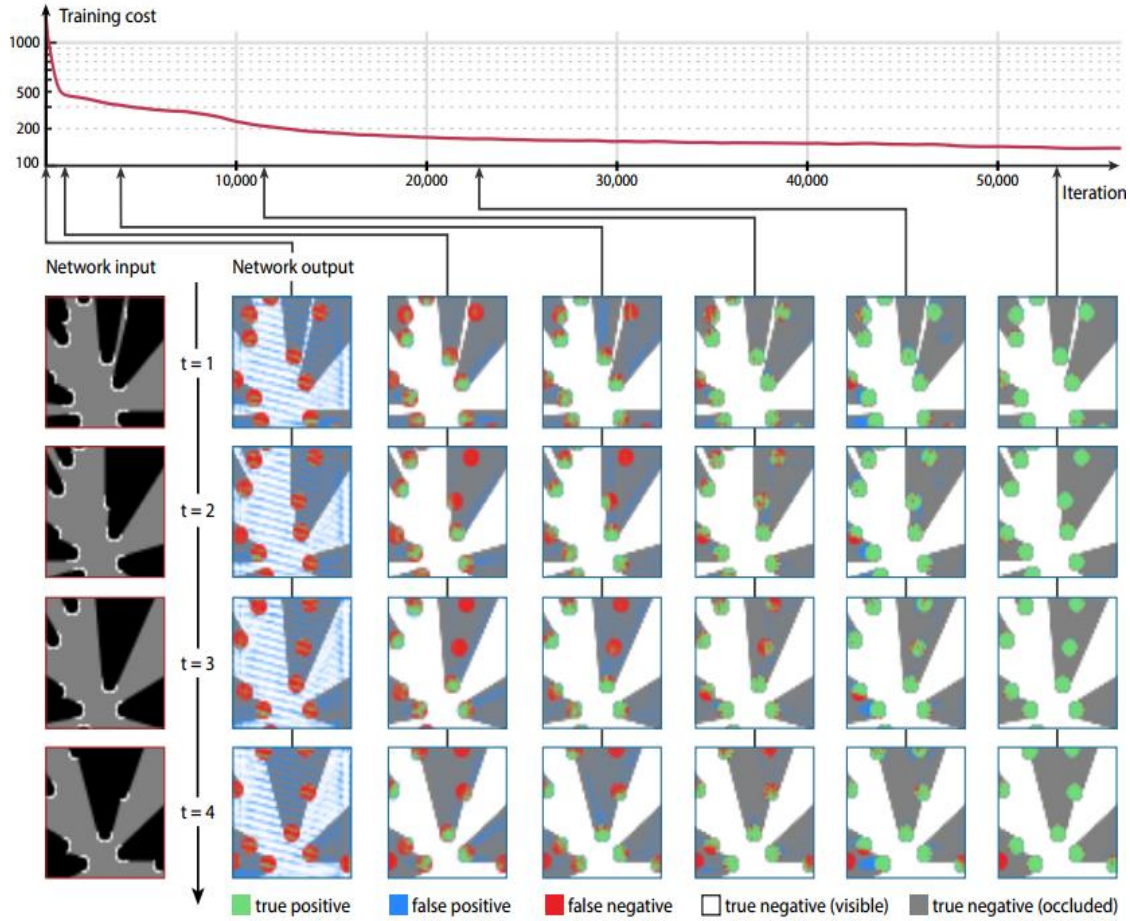
1. 输入数据只有传感器观测 x，没有真实状态 y;

2. 预测输出不再是当前的真实环境状态，变为若干步之后的

观测状态；

通过这样的设定，可以使得模型区学习到预测若干步后的观测，从而具备轨迹推测和遮挡推测的能力。

结果：



（五）RNN 用于图像生成

DRAW: A Recurrent Neural Network For Image Generation

作者机构

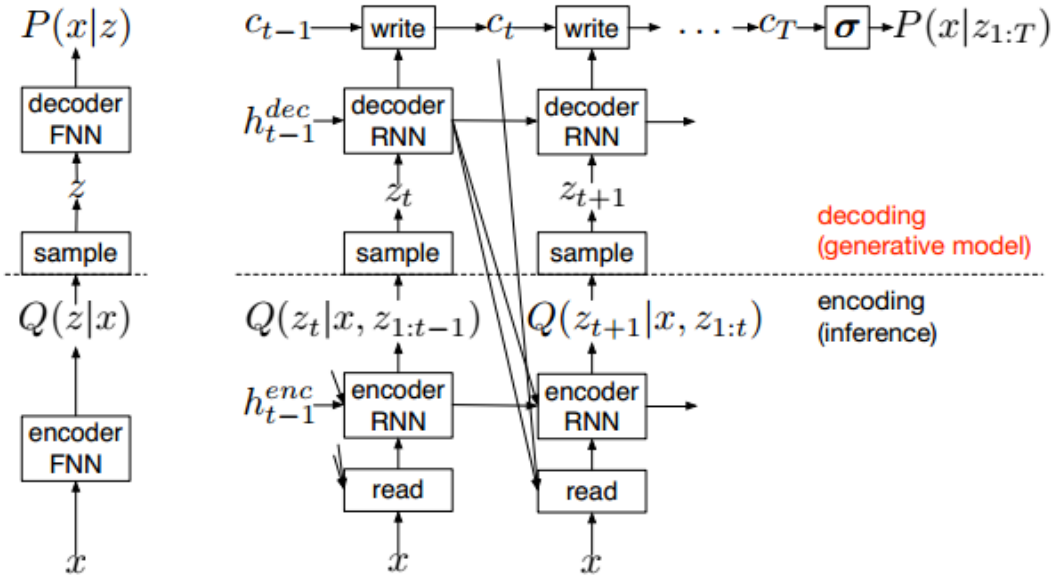
Karol Gregor，Ivo Danihelka，Alex Graves，Google DeepMind，2015.

针对问题和创新点

RNN 用作图像生成；将 attention mechanism 和 VAE(variational auto-encoder)作结合。

方法

GAN and VAE



损失函数包括两个部分：

$$\mathcal{L}^x = -\log D(x|c_T)$$

$$\mathcal{L}^z = \sum_{t=1}^{T} KL\big(Q(Z_t|h_t^{enc})||P(Z_t)\big)$$

Attention 和 Without Attention 的比较:

其 attention 机制的运用跟前面文章中所述稍有区别，本文中采用的方法类似于 RCNN 中使用的 BoundingBox 回归，它从上一时刻的隐变量中直接学习控制下一时刻焦点窗口的五个参数。

$$(\tilde{g}_X, \tilde{g}_Y, \log\sigma^2, \log\tilde{\delta}, \log\gamma) = W(h^{dec}) \quad (21)$$

$$g_X = \frac{A+1}{2}(\tilde{g}_X + 1) \quad (22)$$

$$g_Y = \frac{B+1}{2}(\tilde{g}_Y + 1) \quad (23)$$

$$\delta = \frac{\max(A,B)-1}{N-1}\tilde{\delta}$$

上一步注意力集中的范围编码于解码后的隐状态中，类似objectdetection里面的bboxRegression

实验结果证明，增加 Attention 机制后效果有明显的提升。

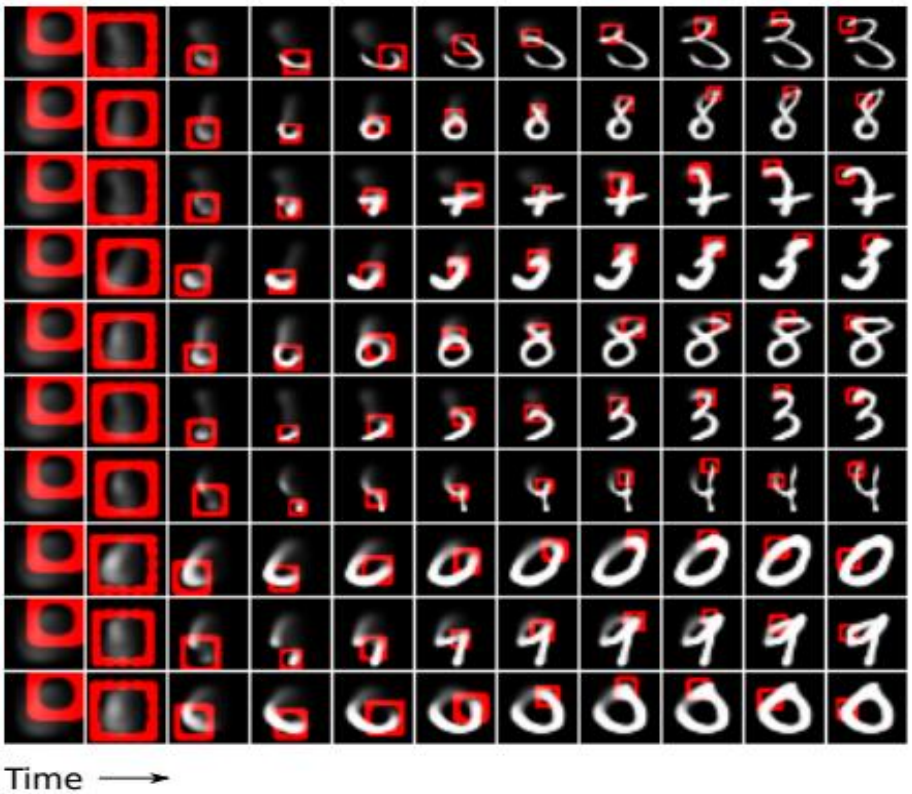
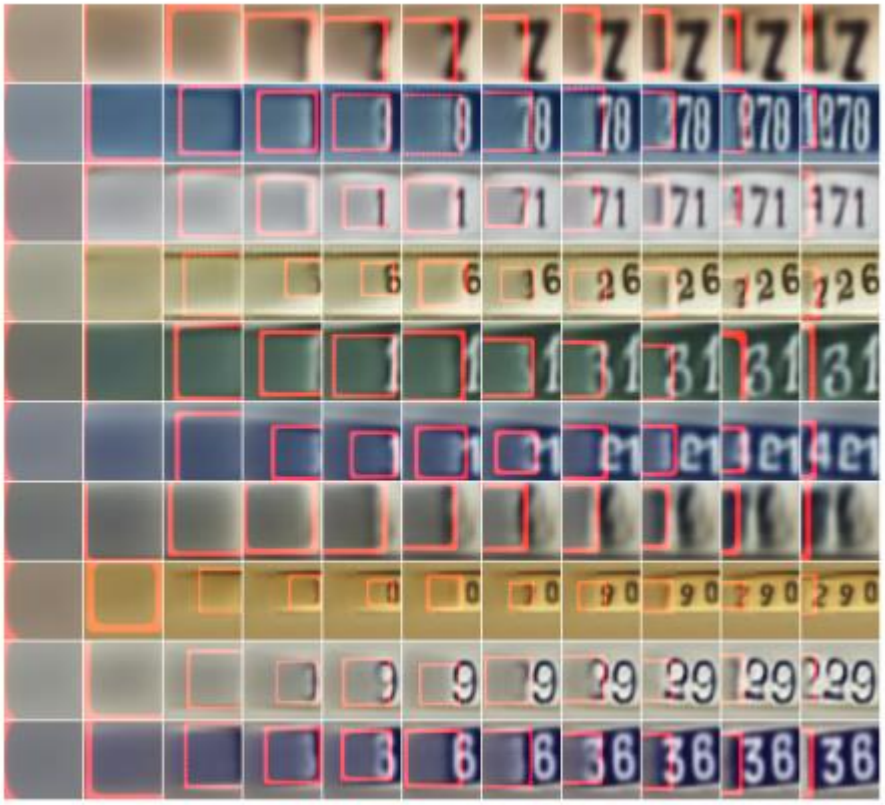
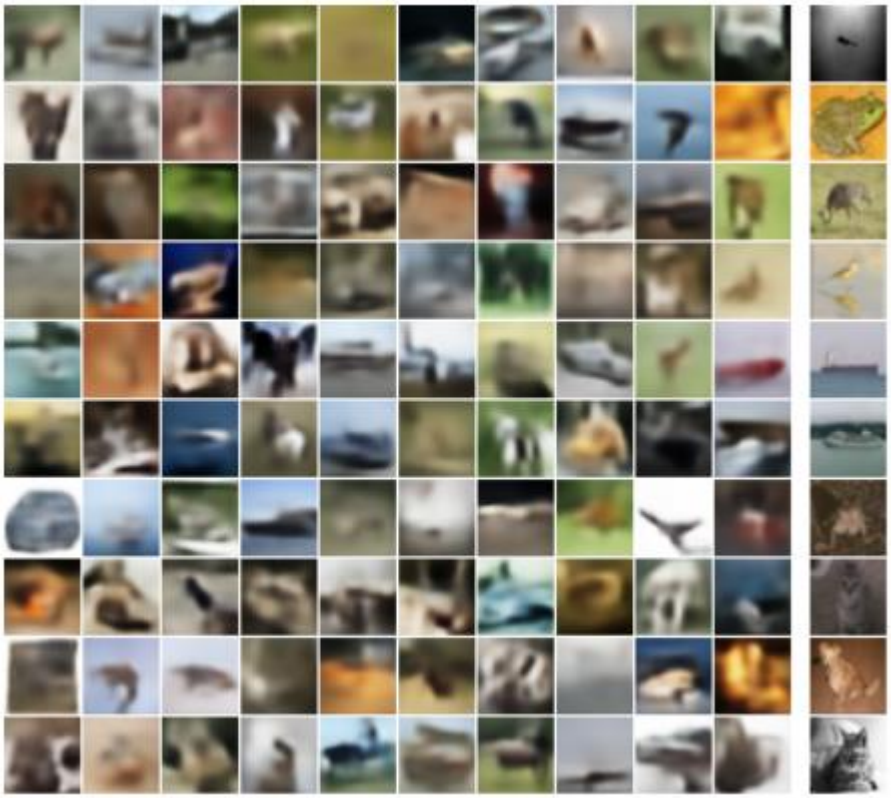


Figure 9. **Generated SVHN images.** The rightmost column shows the training images closest (in L^2 distance) to the generated images beside them. Note that the two columns are visually similar, but the numbers are generally different.



Time ⟶



评价

利用 VAN 对隐变量的分布加以限制可以获得更为有效的隐变量特征。在其他针对特定的学习任务时，也可以考虑增加先验知识。通过改变模型结构或者添加目标函数限制得到更好的描述模型。