加入复制机制的LSTM用于图像描述

Incorporating Copying Mechanism in Image Captioning for Learning Novel Objects(arXiv:1710.02534)

CVPR 2017

报告人：曹成龙　部门：情境感知计算

主要内 容

01 引 言

02 模 型

03 实 验

04 总 结

01

引言

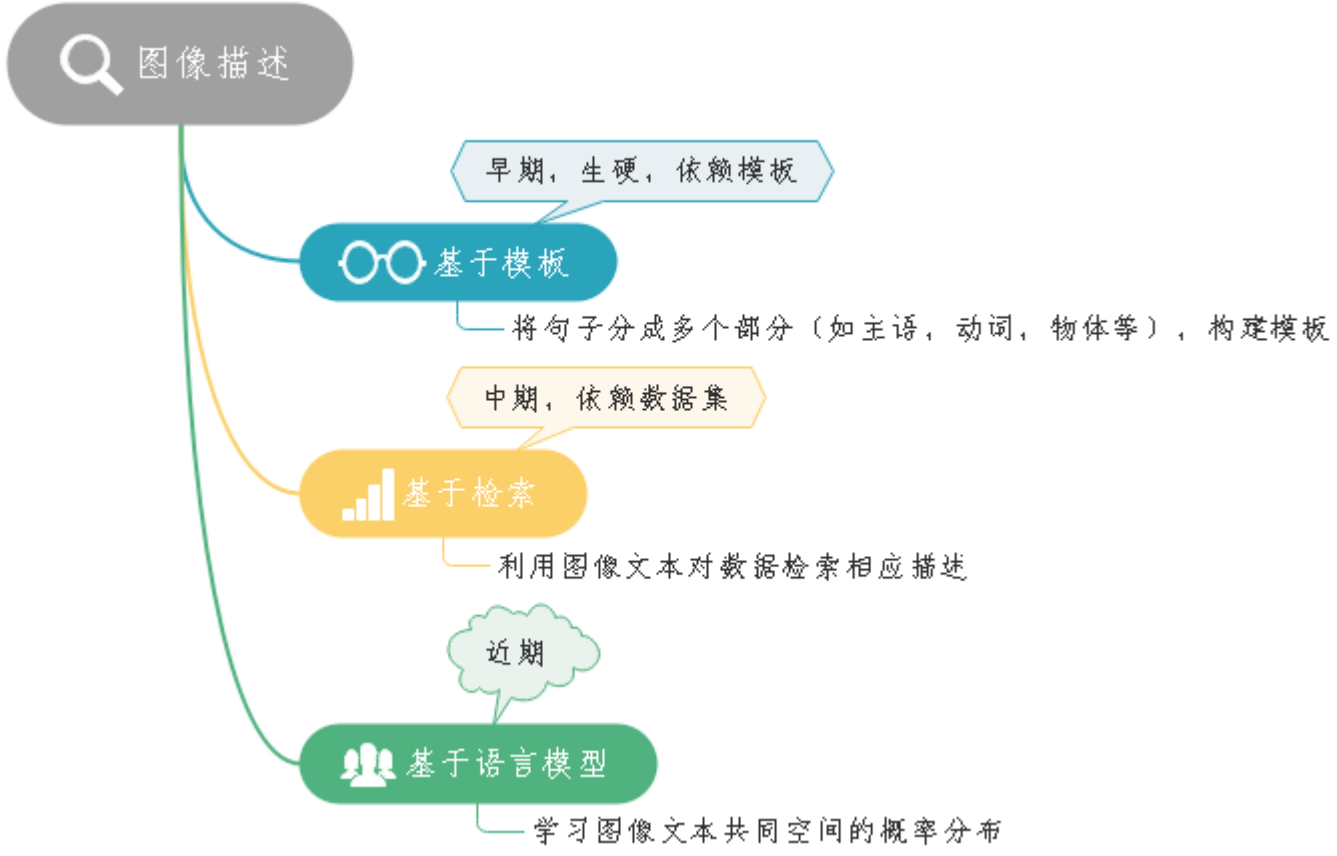
图像描述

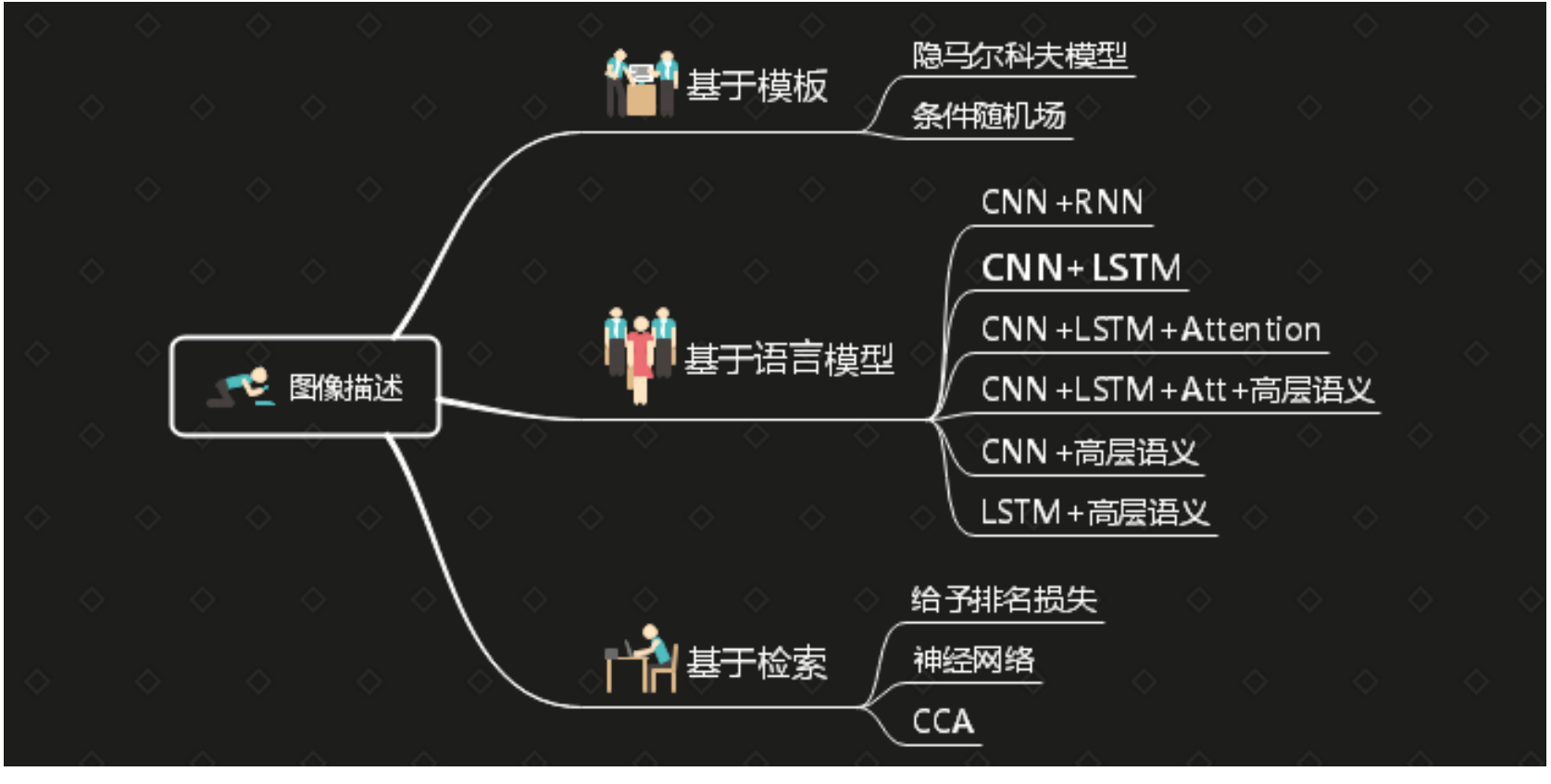
文本

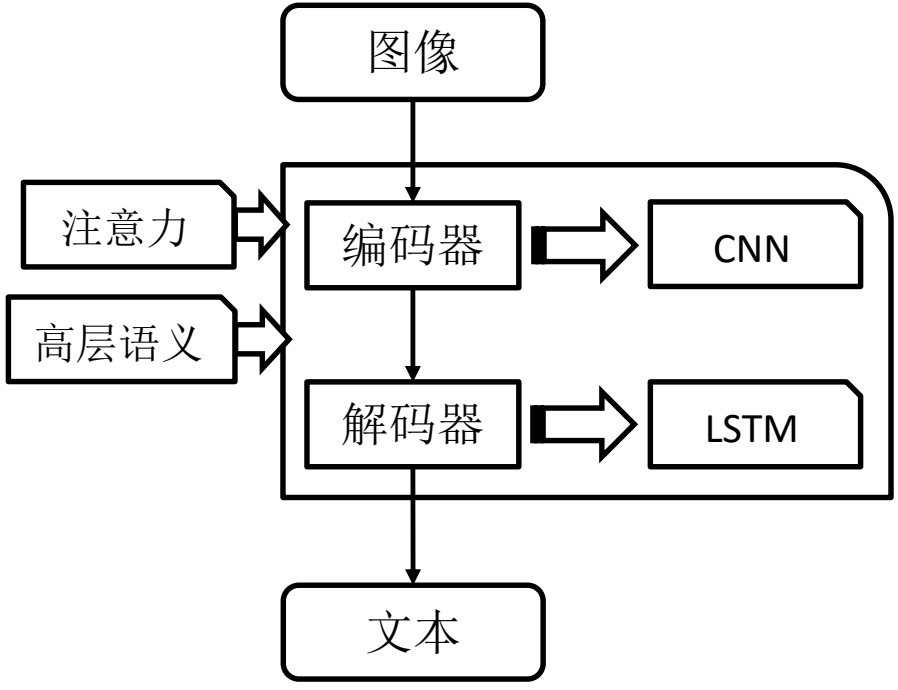
一个金色头发的女子在看书

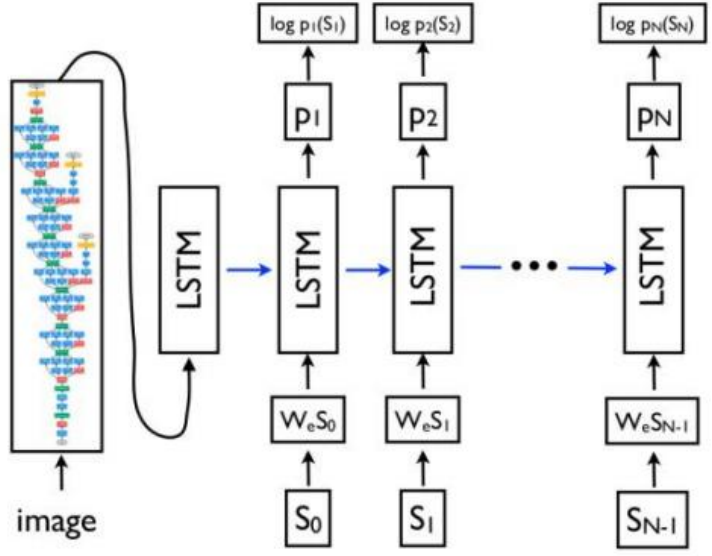
描述

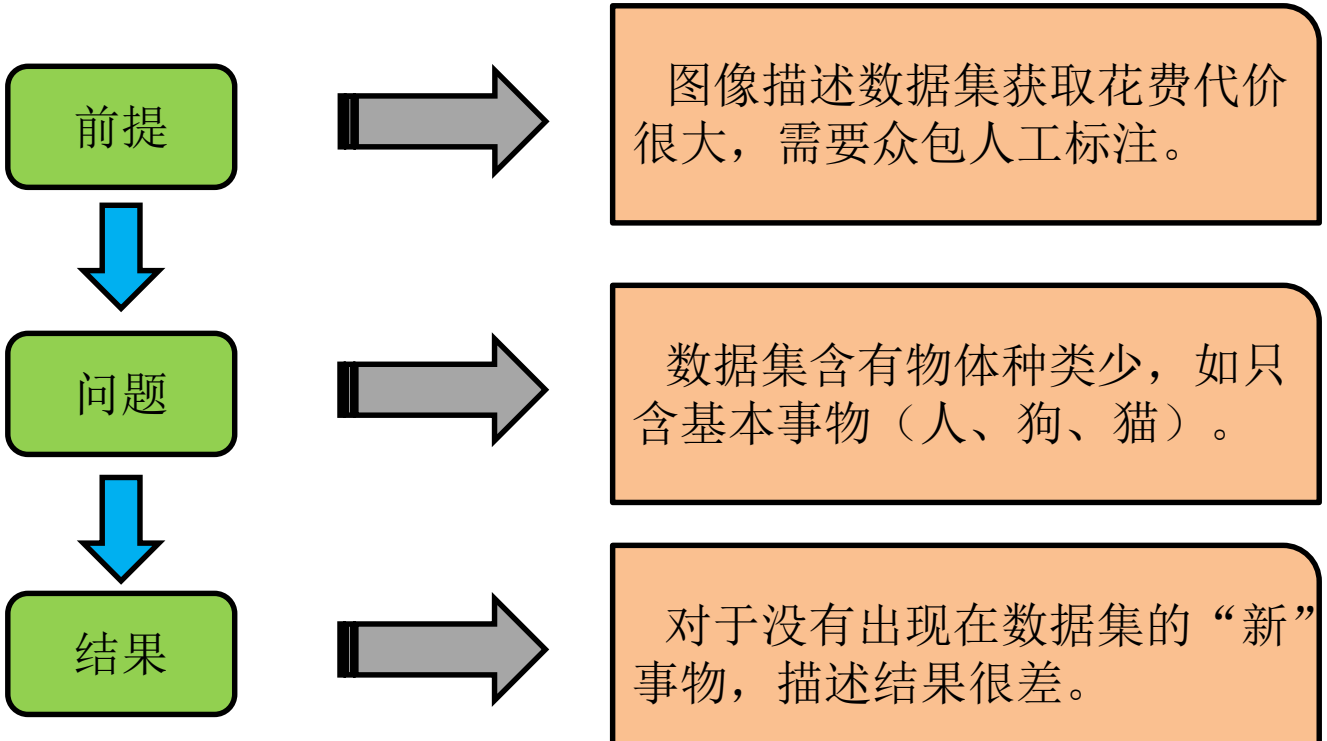
概率 → MAX P(描述|图片)


图像描述
早期，生硬，依赖模板
基于模板
将句子分成多个部分（如主语，动词，物体等），构建模板
中期，依赖数据集
基于检索
利用图像文本对数据检索相应描述
近期
基于语言模型
学习图像文本共同空间的概率分布


图像描述
基于模板
隐马尔科夫模型
条件随机场
基于语言模型
CNN +RNN
CNN+ LSTM
CNN +LSTM +Attention
CNN +LSTM +Att +高层语义
CNN +高层语义
LSTM +高层语义
基于检索
给予排名损失
神经网络
CCA


图像
注意力
高层语义
编码器
解码器
CNN
LSTM
文本


log p1(S1)
log p2(S2)
log pN(SN)
p1
p2
pN
LSTM
LSTM
LSTM
LSTM
WeS0
WeS1
WeSN-1
S0
S1
SN-1
image


前提
问题
结果
图像描述数据集获取花费代价很大，需要众包人工标注。
数据集含有物体种类少，如只含基本事物（人、狗、猫）。
对于没有出现在数据集的“新”事物，描述结果很差。

心路

怎么办

缺啥

描述新物体的图像文本对

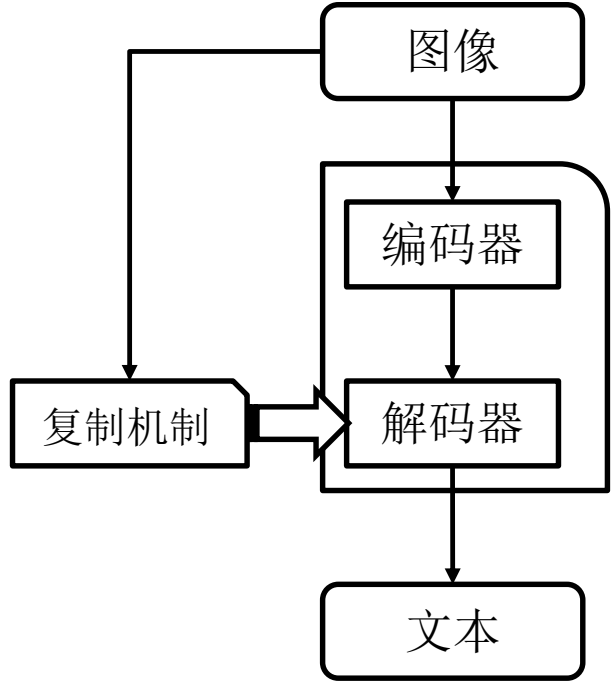
调和互补？

有啥

含有很多物体及其标签的庞大数据集(如Image Net)

文章提出一种带有复制机制的LSTM进行图像描述，它能够从新物体中选择合适的词放在描述的合适位置。从而加强对新物体的描述。

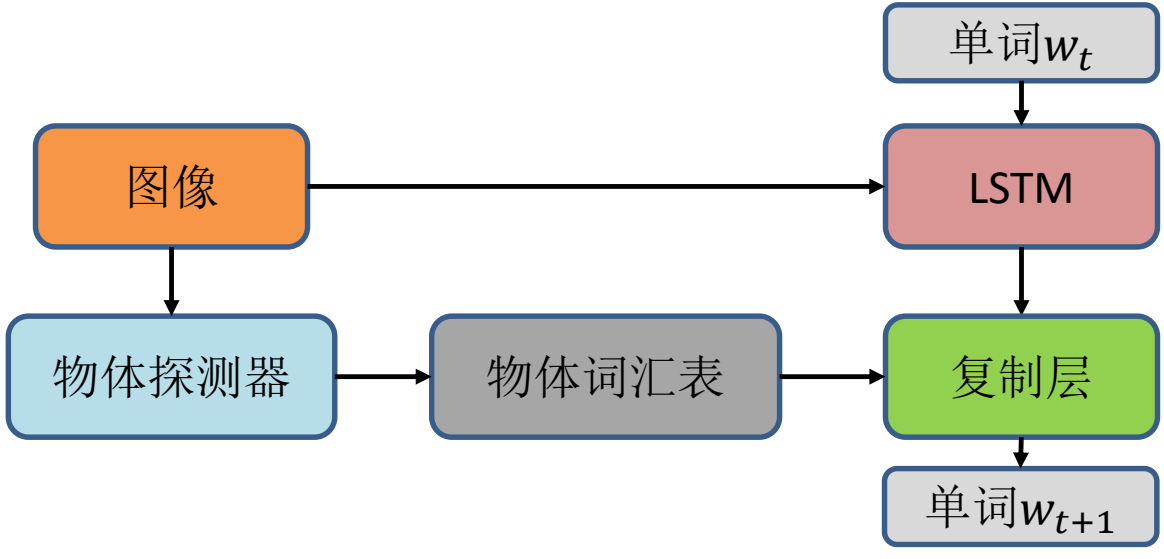
- 使用额外的视觉数据集
- 使用额外数据集训练对新物体的分类器
- 在解码器上加入复制机制进行描述的生成



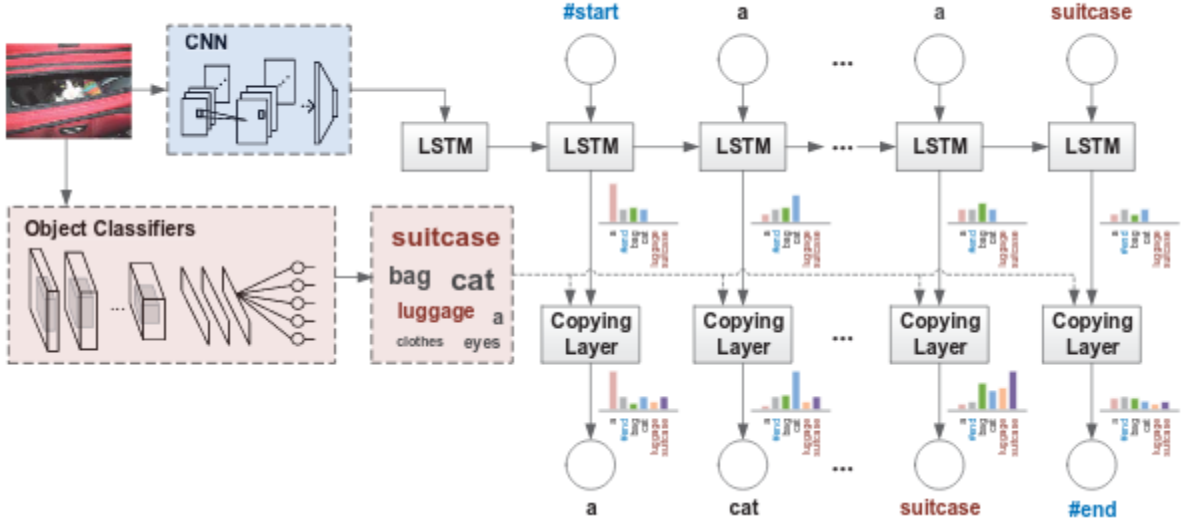
02

模型

模型



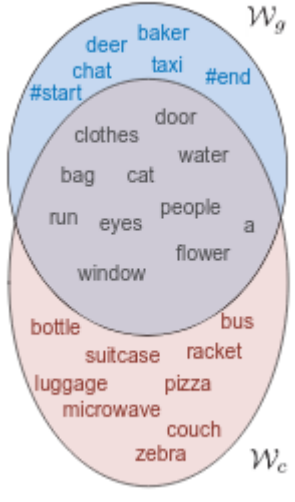
模型展开


CNN
Object Classifiers
suitcase
bag cat
luggage a
clothes eyes
#start
a
a
suitcase
LSTM
LSTM
LSTM
LSTM
LSTM
Copying Layer
Copying Layer
Copying Layer
Copying Layer
a
cat
suitcase
#end

假设图像$I \in R^{D_v}$的描述为:$S = w_1, w_2, \ldots, w_{N_s}, w_t \in \mathrm{R}^{D_w}$,描述包含$N_s$个词，描述被表示为$D_w \times N_s$维度的矩阵。

用W_g表示图像描述数据集的字典，额外的数据集字典用W_c表示。图像I中含有额外数据集中的物体的概率用$\delta(w_i), w_i \in W_c$表示。

W_g和W_c两种字典可能的关系如右图：



文章模型基于encoder-decoder模型

受机器翻译中的encoder-decoder框架的启发，最近的图像描述框架都是基于encoder-decoder的。这种模型首先将图像编码成为固定长度的向量，然后将这个向量解码成目标句子。模型的训练目标是最小化能量函数：

$$E(I,W) = -\log \Pr(\mathrm{W}|\mathrm{I})$$

也就是最大化在给定图像I的条件下生成描述W的概率。

在生成每一个词的时候，通过RNNs会根据已经生成的词信息去生成下一个词，所以概率Pr(W|I)可以被表示成为：

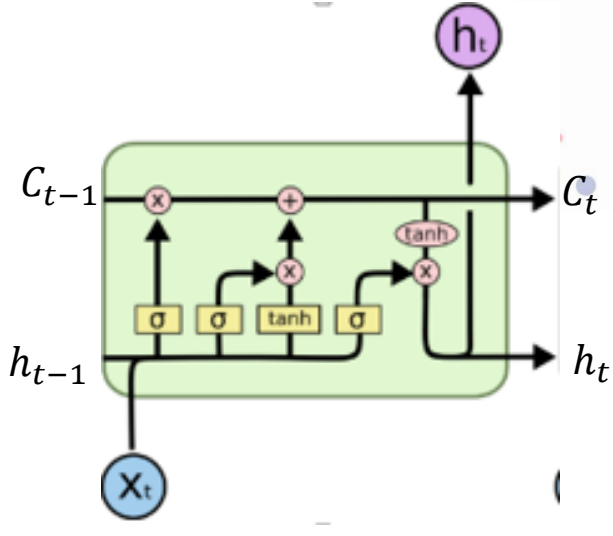
$$\log \Pr(W|I) = \sum_{t=1}^{N_s} \log \Pr(w_t|I, w_0, w_1, \ldots, w_{t-1})$$

两个状态：

隐藏状态h_t和细胞状态C_t

三个门（图中σ符号）：

从左到右分别是遗忘门、输入门和输出门，它们的输出分别记为f_t，i_t和$\widetilde{C_t}$。



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

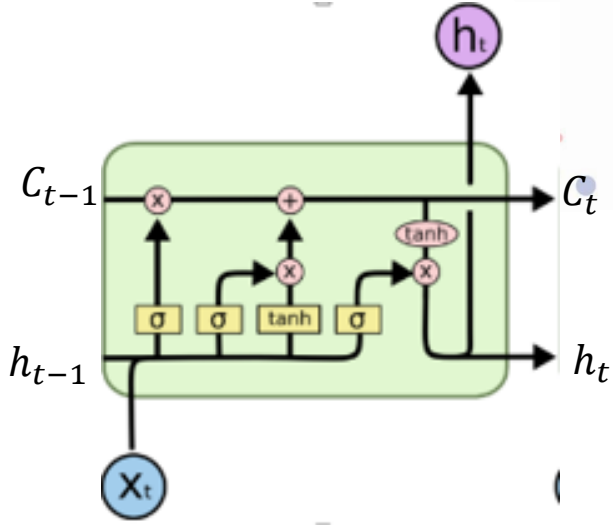
更新状态：

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

输出：

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

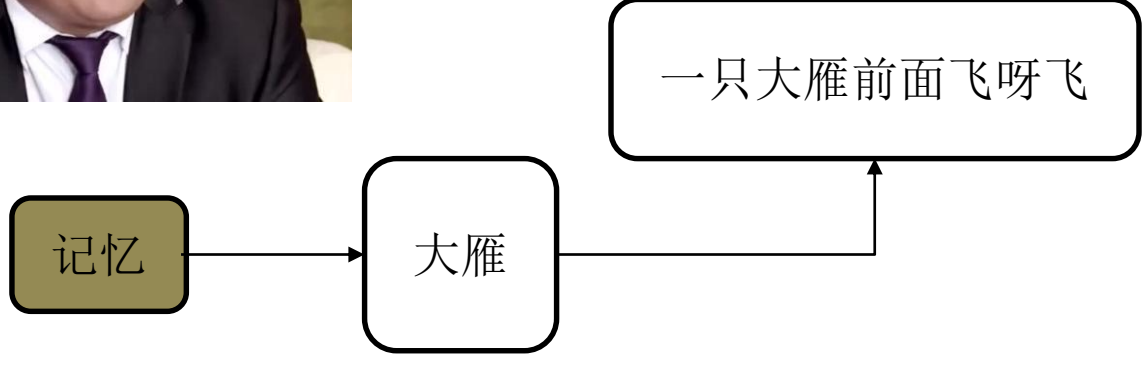
$$h_t = o_t * \tanh\left(C_t\right)$$



- 为了跟后面一致，记$Pr_t^{\mathrm{g}}(\mathrm{w_{t+1}}) = o_{\mathrm{t}}$
- 在时刻-1只输入图像特征

人类在组织语言时通过从记忆中找到某个词直接放入到语句中。

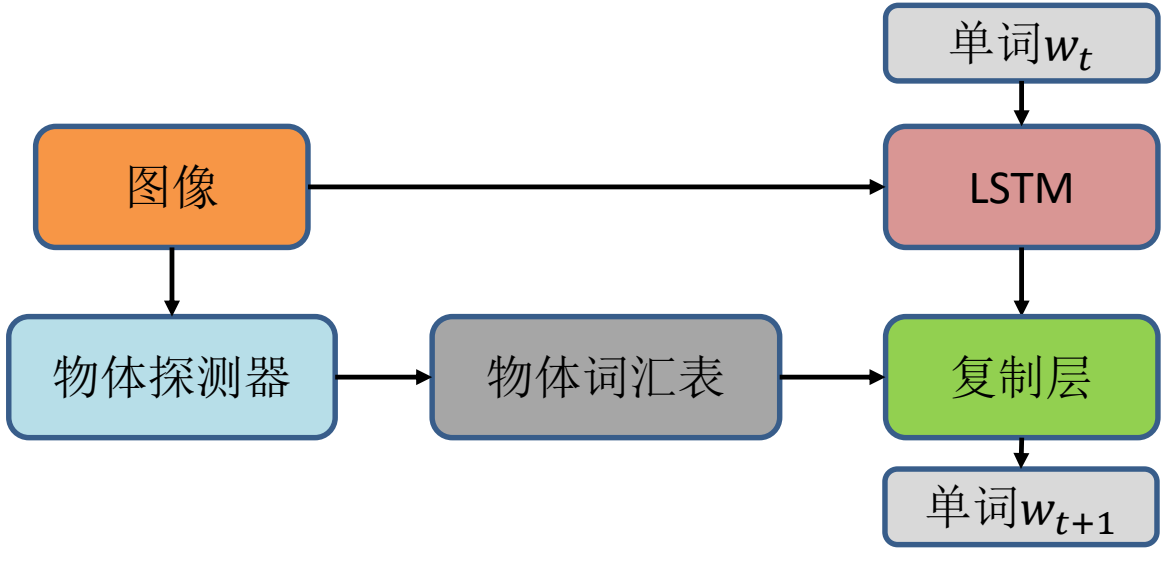
复制机制使得在每次产生下一个词的时候不一定是由前面的词预测得到，也可以直接从外部字典复制过来。



在第t步的解码时，生成的单词 w_{t+1} 直接从图像探测数据集中复制过来的概率为：

$$Pr_t^c(w_{t+1}) = \varphi(w_{t+1}^T M_c) h^t \delta(w_{t+1})$$

其中$M_c \in R^{D_w \times D_h}$代表文本的转换映射矩阵，$\varphi$代表元素级的非线性激活函数，$h^t$代表LSTM上一步解码的输出。$\delta(w_{t+1})$代表词$w_{t+1}$在当前图像中的概率。



对于第t步的解码过程，得到单词w_{t+1}的概率为：

$$Pr_t(w_{t+1}) = \begin{cases} \frac{1}{K} e^{Pr_t^g(w_{t+1})} & , w_{t+1} \in W_g \bigcap \overline{W_c} \\ \frac{\lambda}{K} e^{Pr_t^g(w_{t+1})} + \frac{1-\lambda}{K} e^{Pr_t^c(w_{t+1})} & , w_{t+1} \in W_g \bigcap W_c \\ \frac{1}{K} e^{Pr_t^c(w_{t+1})} & , w_{t+1} \in \overline{W_g} \bigcap W_c \\ 0 & , otherwise \end{cases}$$

其中λ代表复制机制和LSTM预测结果对下一次生成起作用的调节比重，K代表softmax的归一化项，$Pr_t^c(w_{t+1})$ 代表LSTM生成词w_{t+1}的概率。

损失函数表示为：

$$E(I,S) = -\sum_{t=0}^{N_s-1} \log Pr_t(w_{t+1})$$

N表示训练集中的图像文本对数目，则要解决的问题可以总结为以下最优化问题：

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} E\left(I^{(i)}, S^{(i)}\right) + |\theta|^2$$

03

实验

使用两个数据集，一个是图像文本对数据集MSCOCO，另一个是图像分类数据集ImageNet

- Held-out MSCOCO

留下只含bottle，bus，couch，microwave，pizza，racket， suitcase和zebra八种物体的数据。每张图片有五句描述。

- ImageNet

挑选出物体不在MSCOCO数据集中出现的图像进行训练，ImageNet数据集用来训练物体探测器。

F1分数

F_1分数，又称平衡F分数，它被定义为准确率和召回率的调和平均数。

$$F_1 = 2 \cdot \frac{准确率 \times 召回率}{准确率 + 召回率}$$

其中准确率和召回率的计算是使用描述中的新物体和用物体探测器识别出的新物体来计算的。

METEOR评价

机器翻译中的一个评价标准。

对于生成描述W和人工描述U两个句子，首先将他们中的每个词一一对应起来。

对应过程如下：

- 首先写出所有可能的对应。
- 使用下面三种规则得到对应：
  1. 精确对应：两个词完全对应
  2. 转换对应：经过转换可以对应，如”com-puters”和”computers”以及 “comp-uter”的对应
  3. 同义对应：同义词对应

(使用这三种对应的顺序不同会使得结果不同)

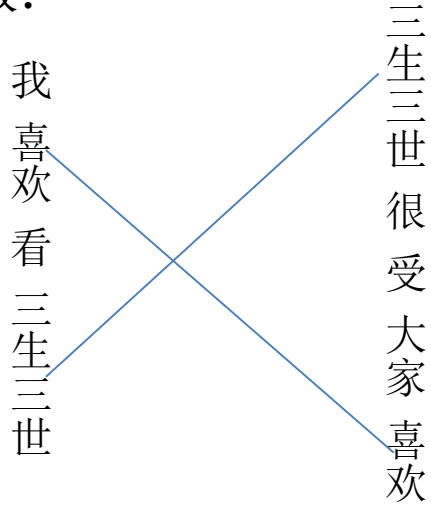
通过上面的步骤可以得到对应的集合，取能够使得W和U每个词都对应起来的最大集合为最终对应集合。

METEOR评价

如果最大对应集合有多个，该怎么办？

取含有“交叉数”最少的集合。

交叉数：



如图，喜欢和三生三世的对应就形成了一个交叉

F_1

METEOR评价

确定了对应组合就可以来计算准确率P和召回率R了：

$$P = \frac{\text{匹配的对应数}}{\text{描述中的单词数}}$$

$$R = \frac{\text{匹配的对应数}}{\text{人工标注中的单词数}}$$

计算其F_{mean}

$$F_{mean} = \frac{10PR}{R + 9P}$$

F_1

为了考虑长匹配的影响，加入一个惩罚值：

$$Penalty = 0.5 * \left(\frac{\text{匹配块数}}{\text{匹配的对应数}}\right)$$

匹配块：

我 喜欢 看 欢乐喜剧人

他 喜欢 看 三生三世

“喜欢看”是两个对应，且两个对应在两个句子中的位置都是相邻的。这就构成了一个匹配块。

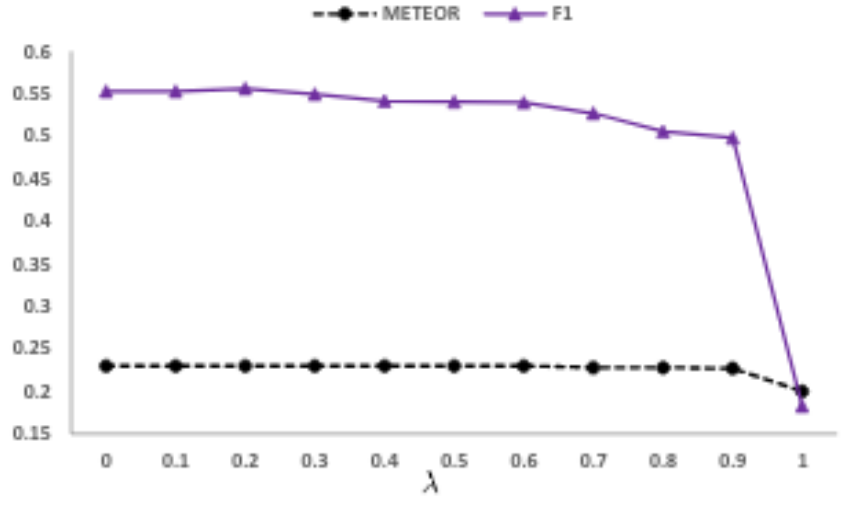
匹配块越长，模型效果好，匹配块数越少，惩罚越小。

最终得到METEOR评价值：

$$Score = F_{mean} \times (1 - Penalty)$$

Model	$F1_{bottle}$	$F1_{bus}$	$F1_{couch}$	$F1_{microwave}$	$F1_{pizza}$	$F1_{racket}$	$F1_{suitcase}$	$F1_{zebra}$	$F1_{average}$	METEOR
LRCN [4]	0	0	0	0	0	0	0	0	0	19.33
DCC [8]	4.63	29.79	**45.87**	28.09	64.59	52.24	13.16	79.88	39.78	21
NOC [22]										
-(One hot)	16.52	68.63	42.57	32.16	67.07	61.22	31.18	88.39	50.97	20.7
-(One hot+Glove)	14.93	68.96	43.82	**37.89**	66.53	65.87	28.13	88.66	51.85	20.7
LSTM-C										
-(One hot)	29.07	64.38	26.01	26.04	**75.57**	66.54	**55.54**	**92.03**	54.40	22
-(One hot+Glove)	**29.68**	**74.42**	38.77	27.81	68.17	**70.27**	44.76	91.4	**55.66**	**23**

分析可以得出，文章提出的模型超过其他模型，除了couch和microwave，因为这两种东西在物体探测器中不容易判别，所以效果不好。



在0~0.6很平稳，在0.2时达到最大，大于0.6时下降很快，说明文章提出的复制机制起到了作用。

GT: black and white cat in a red suitcase

Detected Objects:
cat: 1, suitcase: 0.96, bag: 0.89, luggage: 0.65, black: 0.63

LRCN: a cat sitting on top of a red chair
LSTM-C: a cat laying on a suitcase

GT: a black cat and a bottle of wine

Detected Objects:
cat: 1, bottle: 0.98, wine: 0.84, black: 0.58, standing: 0.58

LRCN: a cat sitting on a desk next to a window
LSTM-C: a cat sitting on a table next to a bottle of wine

GT: the living room has a leather couch near a dining table

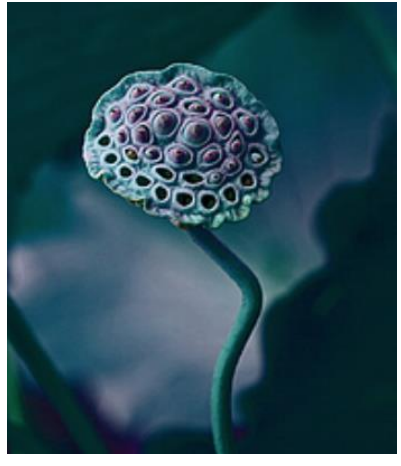
Detected Objects:
room: 1, living: 0.94, television: 0.77, tv: 0.7, couch: 0.62

LRCN: a living room with a laptop computer and a desk
LSTM-C: a living room with a couch and a television

04

总结

提出了加入复制机制的LSTM图像描述模型，很简单的在LSTM上加入了一层复制层，却提高了模型的效果



通过使用外部数据集，训练物体探测器，使得模型对新物体的描述效果增强。

感谢您的观看

CONCISE LITTLE FRESH ACADEMIC REPORT PPT

报告人：曹成龙　　部门：情景感知计算