指导教师：杨涛　　提交时间：2016/3/17

CVPR2015 Paper Translation

No: 01

姓名：丛晓岚

学号：2013302466

班号：10011301

心灵之眼:周期性可视化的图像标题生成技术

陈鑫磊
卡内基梅隆大学
xinleic@cs.cmu.edu

C. Lawrence Zitnick
微软研究院，雷德蒙
larryz@microsoft.com

摘要

在本文中，我们探讨的是图像及其基于句子描述间的双向映射。关键是建立一个动态递归神经网络，其试图动态地建立一个可视化场景并用其表示一个正在生成或阅读的标题。此表示方法会自动学会记忆长期的视觉概念。我们的模型不仅能够基于给定的图像生成新的标题，而且能在给定图像描述前提下重建视觉特征。我们在多个任务中评估我们的方法，其中包括句子生成、句子检索和图像检索等。其中，最先进的结果显示在生成新图象描述的任务上。相比人类生成标题的时间，我们自动生成标题等于或快于人类的 21.0%；在图像和句子检索任务中，结果优于或可比于类似的使用视觉特性方法的最先进的结果。

1. 绪论

常言道“良好的形象描述如同在你的脑海中画一幅画。”精神形象的创建可能在帮助人类理解句子方面扮演了一个重要的角色[18]。事实上，通常人们记住的是其精神形象，而确切的句子往往被人们遗忘[34，26]。举个有说明力的例子，图 1 显示了当进行阅读描述时，精神形象可能会发生变化并增强记忆。那么计算机视觉算法，理解和生成图像标题能否利用类似的方法提取视觉特征？

图 1. 内部可视化表示对于产生和理解场景的语义描述客观重要。而视觉表示可以由于一个描述歧义而变化，一个好的描述能够传达场景的显著方面。

最近，一些论文已经在学习图像联合特征空间及其描述方面进行探索研究[16，38，20]。这些方法将形象特征和句子特征映射到一个共同的空间，这可能用于图像搜索或图像标题排名。各种方法均被用来测试这个映射，包括核典型相关分析(KCCA)[16]，递归神经网络[38]，或深层神经网络[20]。虽然这些方法都能够将语义和视觉特性映射到一起，但他们无法进行逆映射。也就是说，他们不能从映射关系中产生句子或视觉描绘。

在本文中，我们提出一种双向表示方法，即能够从新的描述和生成的图像和视觉表现来进行。这些任务的

关键是动态捕获已经描述场景的视觉方面的新颖表示。也就是说，当产生或读出一个字时，它的视觉表示被更新以反映包含在字中的新的信息。我们做到这一点要使用递归神经网络（RNNs）[9，29，32]。RNNs 长期存在的问题之一是他们在形成记忆递归的概念时需要经过一系列的迭代。例如 RNN 语言模型常常在没有专门的控制单元[15]时学习长距离关系[3、29]上遇到困难。在句子的产生过程中，我们的新颖的动态更新可视化表示作为一个长期记忆的概念已经被提到。这使得网络自动选择突出概念来传达尚未口头表达的句子。我们证明，相同的表示，可以用来创建一个书面描述的可视化表示。

我们在大量的数据集中证明我们的方法。其中包括 PASCAL 数据集[36]，Flickr 8K [36]，Flickr 30K [36]，Microsoft COCO 数据集[27，4]共包含超过 400000 个句子。当生成新的形象描述时，我们将演示结果以 BLEU [35]， METEOR [1]和 CIDEr [40]来表示。定性结果显示为新的生成图像标题。我们也在图像和句子检索的任务中评估算法的双向能力。因为我们不需要像许多以前的论文那样评估生成新的句子的能力，我们的结果显示是优于或可比于以往使用类似的视觉特征方法得到的最先进的成果。

2. 相关工作

构建视觉记忆的任务核心就在于两个长期存在的 AI 困难问题：自然语言符号和物理世界间的联系与理解图像内容的语义。而学习图像修补程序和单独的文本标签之间的映射关系仍是计算机视觉[23，12，13]中的一个热门话题，而使用整个句子描述并联合像素来学习共同嵌入[16，38， 20， 14]更是成为与日俱增的话题热点。查看相关的文本和图像，KCCA[16]是发现共享特征空间的一个自然选择。然而，鉴于两者之间的高度非线性映射，找到一个通用的基于浅显表示的距离度量可以说是极其困难的。最新的论文都在寻求更好的目标函数用来直接优化排名[16]，或直接采用 pre-trained 表示[38]来简化学习，或对二者进行结合[20，14]。

具有良好的距离度量时，我们有可能实现像双向图像句子检索那样执行任务。然而，在许多情况下，还期望基于给定的句子描述可以产生新的图象描述并幻想出一个场景。多篇论文已经探索出生成图像的区域描述[11， 45， 24， 46， 33， 25，21]。这些论文使用各种方法来生成文本，如使用基于模板的句子生成 pre-trained 对象探测器[45， 11，24]。检索到的句子可以组合成新的描述[25]。最近， 基于抽样[21]或周期性神经网络[28]的纯粹统计模

型被用来生成句子。而[28]也使用一个 RNN 模型，但明显不同于我们的模型。特别是他们 RNN 并不试图重建视觉特性，并且更类似于上下文内容相关 RNN[32]。对于从句子中合成图像，Zitnicket al[47] 在最近的论文中使用抽象的剪贴画图像来进行句子的视觉诠释。从句子中提取出关系元组并使用条件随机字段用于视觉场景模型。

有许多论文使用递归神经网络进行语言建模[2， 29， 32， 21]。我们直接地建立在[2，29，32]之上使用人工神经网络来学习单词的上下文。模型使用其他来源的上下文信息以帮助形成语言模型[32，21]。尽管它成功了，RNNs 仍然很难捕获在远距离关系上的顺序建模[3]。一种解决方案是使用长短期存储器（LSTM）网络[15，39，21]，它使用“门”，以明确地控制梯度反向传播，并允许长期相互作用的学习。然而，本文的重点是要表明，通过多模态之间的“翻译”学会了的隐藏层已经可以在大数据中发现丰富的数据结构和自动地学习数据驱动方式下的长距离关系。

有几个同生论文[7， 10， 19，22，42] 探索新的生成图像标题使用 LSTMs [7， 22， 42]，人工神经网络 RNNs [19]和传统的最大熵模型。与这些模型不同，我们的模型动态地构建一个可视化表示正在生成场景的标题。正如我们所演示的那样，这可以改进结果。

3. 方法

在本节中，使用递归神经网络描述我们的方法。我们的目标是两个方面：首先，我们希望能够基于给定一组视觉观察或特性来生成句子。具体来说，我们要计算一个词的概率 w_t 产生在时间 t 给定的集合之前生成的单词 $W_{t-1} = w_1, \cdots, w_{t-1}$ 和观察到的视觉特性 V；其次，我们希望使计算的视觉特征 V 可能性的能力式给出一套口头或读字 W_t 生成场景的视觉表示或用于进行图像搜索。为了完成这两项任务，我们引进了一组潜在变量 U_{t-1} 编码的视觉解释先前生成的或读到的单词 W_{t-1}。如我们稍后部分说明，潜变量 U 在充当先前已经产生或读出的字的长期视觉记忆的方面发挥出关键作用。

利用潜在变量 U，我们的目的在于计算

$$P(w_t | V, W_{t-1}, U_{t-1}) \qquad P(V | W_{t-1}, U_{t-1})$$

这两种可能性相结合我们的最终目标是取得下式最大化：

$$\begin{aligned} & P(w_t, V | W_{t-1}, U_{t-1}) \qquad (1) \\ & = P(w_t | V, W_{t-1}, U_{t-1}) P(V | W_{t-1}, U_{t-1}). \end{aligned}$$

也就是说，我们希望基于以前的单词和视觉诠释词使得词 Wt 的可能性和观察到的视觉特性 V 最大化，需要注意的是在以前的文献[32，28]，其目的仅仅计算 $P(w_t | V, W_{t-1})$，而不是

$P(\mathrm{V}|\mathrm{W}_{t-1})$。

3.1. 模型结构

我们提出的递归神经网络模型结构建立在之前提出的模型[29， 32]上。Mikolov[29]提出了一个 RNN 语言模型中的绿色框所示图 2(a)。时间 t 这个词是由一个向量W_t使用“热”

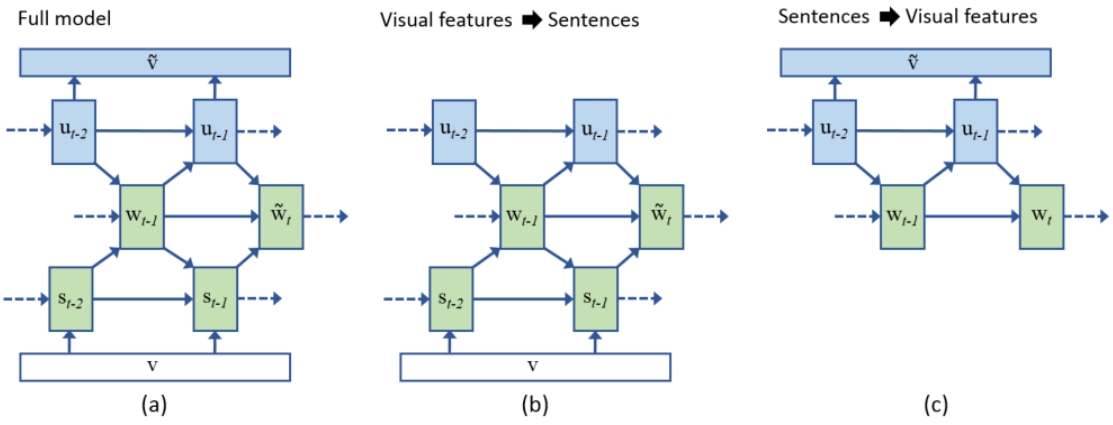

图 2.说明我们的模型：(a)显示了完整的用于培训的模型。(b)和(c)显示模型分别显示所需的部分从视觉特性生成的句子和从句子生成的视觉特性。

来表示。即W_t与单词词汇同样大小，且每个条目都有一个 0 或 1 的值取决于是否使用了这个词。输出W_t包含每个单词生成的可能性。周期隐藏状态 s 提供基于前面单词的上下文。不过，s 通常只有由于梯度消失问题引起的短距离相互作用模型 [3，29]。这个简单，但有效的语言模型被证明提供一个有用的连续词嵌入到各种不同的应用程序中[30]。

继[29]后，Mikolov[32]等添加了一个输入层 **v** 到 RNN 模型，如图 2 中的白盒所示。这一层可能代表不同的信息，如主题模型或词性[32]。在我们的应用程序中，v 代表观察到的视觉特征的集合。我们假设 v 视觉特性常数。这些视觉特性帮助我们选择单词。例如，如果检测到一只猫，“猫”这个词更可能是口头的。注意，与[32]不同，我们没有必要直接连接 **v** 和 **w**，因为 **v** 对于应用程序而言是静态的。在[32]中，**v** 代表的是动态信息，如需要直接访问的 **w** 等词类。我们也发现，只有连接一半 s 单位的 **v** 才能提供更好的结果，因为它在文本或视觉特性的建模中允许不同的单位。

本文的主要贡献是提出了周期性视觉隐藏层的 **u**，如图 2(a)中蓝色框所示。周期性隐藏层 **u** 试图从之前的单词中重建视觉特性，即$\tilde{V} \approx V$。使用的视觉隐层也是在$\tilde{w}_t$的帮助下预测下一个单词。也就是说，网络可以比较其视觉记忆 **u**，与目前所观察到的 **v** 预测接下来要说什么，其中 **u** 代表它已经说了什么。在句子的开头，**u** 代表视觉特性的先验概率。随着越来越多单词被观察，视觉特性得以更新，以反映词的目视判读。例如，如

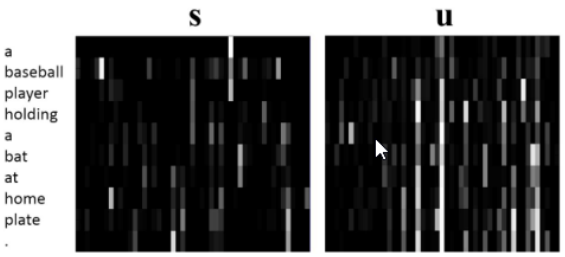

图 3. 说明隐藏单位 s 和 u 激活时间(垂直轴)。注意到视觉隐藏单位 u 通过一些单位的时间来展示长期的稳定的记忆，在每个时间隐藏单位年代均发生重大变化。

果生成词“下沉”，相对应的视觉特性应该增加。其他特性，对应于火炉或冰箱可能增加，因为他们与下沉高度相关。

周期性视觉特征的一个关键属性 **u** 是他们长期记忆视觉概念的能力，这个属性来源于模型结构。直观地说，我们应直到完成句子才去估计视觉特性，。也就是说，**u** 要等到W_t生成句子的结束标记再来估计 **v** 。然而我们的模型中，强迫 **u** 在每一个时间步都去估计 **v**。这有助于长期视觉概念的学习。例如，如果生成“猫”这个词， 对应视觉特性的u_t的可能性将增加。假设 “猫”的视觉特性 **v** 是活跃的，网络将强化 **u** 关于“猫”的记忆并从某一个到另一个实例进行传播。图 3 显示了一个说明性的例子和隐藏的单位 s 和 u。可以观察到，一些视觉隐藏单位 u 显示出长时间的稳定性。

注意，相同的网络结构可以从句子或从视觉特生成句子来预测视觉特性。生成的句子(图 2(b))，**v** 是已知的而$\tilde{v}$可能被忽略。为了从句子中预测视觉特性(图 2(c))，**w** 是已知的，**s** 和 **v** 可能被忽略。这个属性源自于这样一个事实:词单位 w 将模型划分为两个部分来分别预测单词或视觉特性。另外，如果隐藏的单位 s 是直接连接到 u，这个属性将会丢失，并且网络作为一个正常的自编码[41]。

3. 2. 实现细节

这一部分将详细描述我们的语言模型并展示如何学习我们的网络。

3. 3. 语言模型

我们的语言模型通常有 3000 到 20000 个单词。虽然每个单词可能独立预测，这种方法仍需要大量的计算。相反，我们采用的概念词分级[29]和映像分配到一个产品两个方面：

$$P(w_t|\cdot) = P(c_t|\cdot) \times P(w_t|c_t,\cdot). \quad (2)$$

$P(w_t|\cdot)$是这个单词的可能性，$P(c_t|\cdot)$是这个类的可能性。这个词的类标签用一种无监督的方式来计算，频率相近的单词的分在一组。一般来说，这种方法极大地加快了学习过程减小困惑。预测的词可能使用标准 soft-max 函数计算。每个实验后，独立验证困惑，如果困惑不减少则评估学习减少(我们的实验减半)。为了进一步减少困惑，我们将 RNN 模型的输出和最大熵模型的输出结合 [31]，对于所有的实验，当用最大熵模型来预测下一个单词时，我们修正需要回顾单词的数量。

对于任何自然语言的处理任务，预处理对最终的性能是至关重要的。对于所有的句子，在送入 RNN 模型前我们做了以下三个步骤。 1）使用斯坦福 CoreNLP 工具来标记句子； 2）小写所有字母； 3）更换出现在词汇表（OOV）外的低于 5 倍的单词。

3.4. 学习

对于学习，我们使用反向传播通过时间算法（BPTT）[43]。具体地，网络是展开的几个单词和应用标准的反向传播。需要注意的是，请注意，我们会遇到重置模型结束的句子（EOS），所以预测不会跨越边界的句子。如图所示，在[29]是有利的，我们使用在线学习的权重周期性单元输出单词。权重的网络一次批量更新每个句子。所有单元的激活使用乙状结肠函数计算 $\sigma(z) = 1/(1 + \exp(-z))$ 与剪裁，除非这个词使用 soft-max 的预测。我们发现解决线性单元(ReLUs)[23]与无限激活数值不稳定，通常用于复发网络时“爆炸”。

	PASCAL		
	PPL	BLEU	METEOR
Midge [33]	-	2.9	8.8
Baby Talk [24]	-	0.5	9.7
Our Approach	25.3	9.8	16.0
Our Approach+FT	24.6	10.4	16.3
Our Approach+VGG	23.8	12.0	17.6
Human	-	20.1	25.0

表 1. PASCAL 1K 新句子生成的结果。结果是使用困惑（PPL），BLEU（%）[35]和 METEOR（METR，%）测定值[1]。结果由 Midge[33]和 BabyTalk[24] 提供。人类协议分数显示在最后一行。详情请参照文本。

我们使用了开源 RNN 代码[29]和[17]的咖啡框架来实现我们的模型。两者结合的一大优势是，我们可以共同学习文字和图像表示法：从预测的话，可以直接回传播到影像级特征的错误。然而，深的卷积神经网络需要大量的数据来训练，但最大的句子图像数据组只有 80 k 图像[27]。因此，我们不是从头开始训练，而是选择从前期培训的 1000 类 ImageNet 模型以微调[6]，以避免潜在的过度学习。在所有的实验中，我们使用 BVLC 参考网[17]或牛津大学 VGG-Net 的[37]。

4. 结果

在本节中，我们在多个任务中评估双向 RNN 模型的有效性。我们首先描述用于训练和测试的数据集，其次是我们的基线。我们的第一组评估衡量我们的模型生成新的描述图像的能力。由于我们的模型是双向的，我们在句子检索和图像检索任务中评估其性能。其他结果请参阅[5]。

	Flickr 8K			Flickr 30K			MS COCO Val			MS COCO Test		
	PPL	BLEU	METEOR	PPL	BLEU	METEOR	PPL	BLEU	METEOR	BLEU	METEOR	CIDEr
RNN	17.5	4.5	10.3	23.0	6.3	10.7	16.9	4.7	9.8	-	-	-
RNN+IF	16.5	11.9	16.2	20.8	11.3	14.3	13.3	16.3	17.7	-	-	-
RNN+IF+FT	16.0	12.0	16.3	20.5	11.6	14.6	12.9	17.0	18.0	-	-	-
RNN+VGG	15.2	12.4	16.7	20.0	11.9	15.0	12.6	18.4	19.3	18.0	19.1	51.5
Our Approach	16.1	12.2	16.6	20.0	11.3	14.6	12.6	16.3	17.8	-	-	-
Our Approach+FT	15.8	12.4	16.7	19.5	11.6	14.7	12.0	16.8	18.1	16.5	18.0	44.8
Our Approach+VGG	15.1	13.1	16.9	19.1	12.0	15.2	11.6	18.8	19.6	18.4	19.5	53.1
Human	-	20.6	25.5	-	18.9	22.9	-	19.2	24.1	21.7	25.2	85.4

表 2.对 Flickr 8K，Flickr 30 K 生成的新句子结果，MS COCO 验证和 MS COCO 测试。结果用困惑(PPL)，BLEU(%)[35]，METEOR (%)[1]和 CIDEr-D(%)[40]来测量。人类协议成绩如最后一行所示。详情请参照文本。

4.1. 数据集

为了评价，我们在一些标准数据集上进行生成句子和句子图像检索任务的实验：

PASCAL 1K [36] 图像的数据集包含一个PASCAL VOC挑战的子集。20 个类别中的每一个都有 50 张由亚马逊的 Mechanical Turk（AMT）提供 5 描述的随机样本的影像。

Flickr 8K and 30K [36] 这些数据集分别包含了 8000 和 31783 张从 Flickr 收集的图片。大多数的图像描绘人类参与各种活动。每个图像也搭配 5 个句子。这些数据集有一个标准的培训、验证和测试分裂。

MS COCO [4，27] 微软可可数据集包含 82783 个训练图像和 40504 张验证图像，每张图像含 5 人生成的描述。Flickr 收集的图片均是通过搜索公共对象类别找到的，通常包含多个对象和重要的上下文信息。我们使用训练集和验证集在实验中训练我们的模型，并上传生成标题测试集(40775 张照片)到可可服务器[4]来进行评价。

报告用 5 参考标题来显示结果。

4.2. RNN 基线

为了深入了解我们模型中的各个部件，我们比较了有三个 RNN 基线的最终模型。公平起见，所有实验中随机种子的初始化是固定的。隐藏层 s 和 u 的大小固定为 100。我们试着增加隐藏单元的数量，但结果没有改善。对于小型数据集，更多的单位会导致过度拟合。

RNN based Language Model (RNN) 这是基本 RNN[29]模型的语言，没有输入的视觉特性。

RNN with Image Features (RNN+IF) 这是一个图像特征送入由[32] 启发的隐藏层的 RNN 模型。如在第 3 部分所述，v 仅连接至 s 而未连接 $\tilde{w}$。对于视觉特性 v，我们使用 ReLUs 后的 BVLC 参考网[17]的 4096D 第七层输出。以下[23]，我们平均从裁剪 4 个角和中心计算出的五个表示。这个网络是预训练的 ImageNet 1000 单

向分类任务[6]。我们尝试其他层（第5和第6），但他们的效果不好。

RNN with Image Features Fine-Tuned (RNN+FT) 这个模型和RNN +IF有相同的架构，但误差反向传播到卷积神经网络[13]。CNN从BVLC参考网的权重初始化。该RNN初始化是用预先训练的RNN语言模型。也就是说，唯一的随机初始权重是从视觉特征v到隐藏层s。如果RNN没有经过预训练，我们发现在初始梯度成为CNN时会出现大量噪声。如果权重从v到隐藏层s为预先训练好的，那么搜索空间变得很有限。我们目前的实施约需5秒来学习特斯拉K40 GPU的128大小批量。特斯拉K40 GPU。它对于跟踪验证错误至关重要，并避免过度拟合。MS COCO帮助我们观察到这种微调策略，但在Flickr数据集并没有太多的性能提升。Flickr的数据集可能无法提供足够的训练数据，以避免过度拟合。

RNN with Oxford VGG-Net Features (RNN+VGG) 代替BVLC参考网络功能特点，我们还尝试牛津VGG-Net[37]的特点。最近的许多文献[28，19]报道比此表示更好的性能。我们使用ReLU后的最后一层来送入RNN模型。

4. 3. 句子生成

我们的第一套实验在模型产生新句子来描述图片的能力方面进行评估。我们先前描述的所有图像句子数据集实验和RNN基线及其他先前的论文[33，24]进行比较。因为PASCAL 1 K是数量有限的训练数据，我们报告结果是在MS COCO上训练，在PASCAL 1K上进行测试。我们在Flickr 8 K和30 K的数据集上使用标准的训练-测试分裂。

用MS COCO检验我们训练和验证的训练集（~37K/~3K）并比较我们方法的变型。最后，我们报告的测试结果是在MS COCO测试集上并使用MS COCO评估服务器[4]。为了产生一个句子，我们首先从长度的多项分布和训练数据了解到的目标句子长度来进行采样，则对于该固定长度，我们随机采样了100个句子，并且使用具有最低损失（负面的可能性，在我们的模型的情况下，重建误差）作为输出。我们选择以下四种自动指标用COCO标题评估工具来评估产生的句子的质量，有困惑度，BLEU[35]，METEOR [1]和CIDEr[40] [4]。

困惑测量生成测试基于需要编码的比特数句子的可能性，值越低越好。BLEU和METEOR最初设计被用于机器自动翻译，他们的评定方式是用几个给定的句子来判定翻译句子的质量。我们可以把翻译任务当作将图片生成句子的一种“翻译”。对于BLEU，我们采取了从1克至4克的几何平均得分，并用于最靠近真实情况的长度所生成的句子来惩罚简

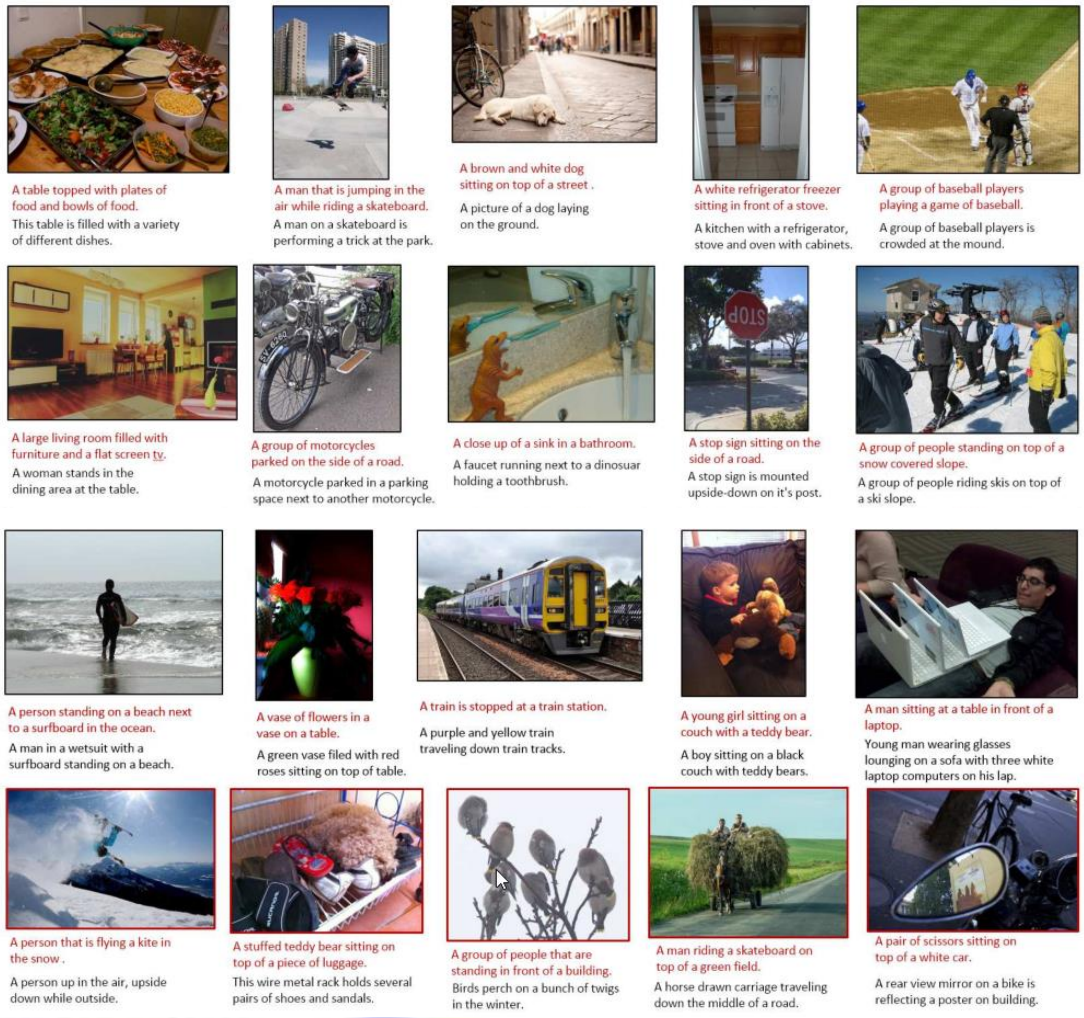

图 4. MS COCO 句子生成的数据集的定性结果。生成的句子(红色)使用(我们的方法+英尺)，人工生成标题(黑)所上所示。最后一行显示了几个具有代表性的失败案例。

洁。对于 METEOR，我们使用了最新版本。CIDEr [40]是专门用于评估图片说明制定的指标。我们用 CIDEr 的变体 CIDEr-D。对于 BLEU，METEOR 和 CIDEr 分数均是越高越好。供参考，我们还原报告人和注释者之间的一致性(用 1 句话作为查询，其余为所有测试参考，但 MS COCO 测试除外)。

PASCAL 1K 结果如表 1 所示。由 BLEU 和 METEOR 测量，我们的方法显著改进了 Midge [33]和 BabyTalk [24]。我们的方法通常能够产生更自然的描述语句，如图像是黑白的，或公共汽车是“双层”的。Midge 的描述往往比较短且细节更少但 BabyTalk 描述的比较长，但经常有多余的描述。在表 2 中还提供了在 Flickr 8K 和 Flickr 30K 中的结果。

MS COCO 数据集中包含更多比较复杂的照片，我们提供 BLEU，METEOR 和 CIDEr 的得分。令人惊讶的是，我们的 BLEU 和 METEOR 得分（18.5 和 19.4），只是比人的得

分（21.7 和 25.2）略低。我们的 CIDEr 的结果（52.1）比人类（85.4）的结果明显低。利用图像特征（RNN+ IF）显著提高了只是使用一个 RNN 语言模型的性能。微调（FT）在所有数据集中都比我们的做法提供更多的改进。使用 VGG-NET[37]（我们的方法+ VGG）的结果表现出一定的改善。但是，我们相信，微调可能会产生更好的效果。MS COCO 数据集的定性结果如图 4.所示。关于本文和其它论文在 MS COCO 测试集上的最新结果请访问 MS COCO 标题评价排行榜。

众所周知，自动措施只是大致等同于人类的判断[8， 40， 16]，所以使用人类研究的方法来评估生成的句子尤为重要。我们在 MS COCO 验证上评估了 1000 个产生的句子，要求人类受试者评判生成的标题是否有更好的、更糟糕的还是同等的质量。 5 受试者要对每幅图像进行评估，并记录了多数票。有平局的情况（2-2-1），两个获奖者均得到了一半的票。我们发现 5.1％的标题（我们的方法+ VGG）是优于人类生成的标题及 15.9％被判定为与人类生成标题质量相等。我们在 MS COCO 上仅仅使用图像级别来处理复杂图像的视觉特性，得到了令人如此印象深刻的结果。

4. 4. 双向检索

我们的 RNN 模型是双向的。也就是说，它可以从句子生成从图像特征，也可以从图像特征生成句子。为了评估双向的能力，我们衡量其处理两个检索任务的性能，基于句子描述的图像检索和基于图像的句子描述检索。因为大多数先前的方法只能够完成检索任务，这也有助于提供实验间的比较。

继其他方法，使用多个图像的描述时我们采用了两种协议。第一种是每 5 个句子单独处理。在这种情况下，检索到真实情况句子的等级被用于评价。在第二种情况下，我们把所有的句子作为一个单一的注释，并将它们连接起来一起进行检索。

对于每个检索任务我们有两种排名的方法。首先，我们基于句子给定图像的可能性(T)进行排名。因为短的句子自然生成概率更高，我们遵循[28]并在整个检索集上对归一化概率除以总概率求和。第二，我们基于该图像的视觉特征 V 和其重构视觉特征 $\tilde{V}$ （I）之间的重构误差进行排序。为了得到更好的性能，我们在所有时间步骤都使用平均重建误差，而不是仅仅用在句末。在表 3 中，我们报告上仅使用（I）的文本可能性术语及其视觉特征重构误差（T + I）的组合来检索结果。所有的结果均使用 VGG-NET[37]所产生的视觉特征。

从先前论文的句子检索和图像检索任务中采纳的评价指标。他们用 R@K(K = 1，5，10)作为测量，这是

（第一）真实情况句子（句子检索任务）或图像（图像检索任务）的召回率。高 R@K 对应于更好的检索性能。我们报告了中位数/平均（第一）获取真实情况句子或图像的等级（中/平均性能。Flickr8K 和 30K 已经提出几个不同的评价。我们对应于所提出的方法，Flickr 8K 三个报告分数分别是[38]，[16]和[28]；Flickr 30K 的[12]和[28]。

如表 3 和 4 所示，Flickr 8 K 和 30K 方法比除了最近提出的 DeepFE[20]以外的其他方法有类似或更好的结果。然而，DeepFE 使用一组基于较小图像区域的不同的特性。如果使用类似的特性(DeepFE + 脱咖啡因)作为我们的方法，我们能产生更好的结果。我们相信这些贡献是互补的，通过使用更好的特性我们的方法也可能有进一步改善。在基于文本和视觉特征（T + I）的总排名中只用了文本（T）。关于 PASCAL 和 MS COCO 的检索结果请参见[5]。

	Sentence Retrieval				Image Retrieval			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
SDT-RNN [38]	4.5	18.0	28.6	32	6.1	18.5	29.0	29
DeViSE [12]	4.8	16.5	27.3	28	5.9	20.1	29.6	29
DeepFE [20]	12.6	32.9	44.0	14	9.7	29.6	42.5	15
DeepFE+DECAF [20]	5.9	19.2	27.3	34	5.2	17.6	26.5	32
RNN+VGG	8.9	25.7	38.7	20.5	6.5	17.3	28.4	25
Our Approach (T)	9.6	29.1	41.6	17	7.0	23.6	33.6	23
Our Approach (T+I)	9.9	29.2	42.4	16	7.3	24.6	36.0	20
[16]	8.3	21.6	30.3	34	7.6	20.7	30.1	38
RNN+VGG	7.7	23.0	37.2	21	6.8	24.0	33.9	23.5
Our Approach (T)	8.1	24.4	39.1	19	7.4	25.0	37.5	21
Our Approach (T+I)	8.6	25.9	40.1	17	7.6	24.9	37.8	20
M-RNN [28]	14.5	37.2	48.5	11	11.5	31.0	42.4	15
RNN+VGG	14.4	37.9	48.2	10	15.6	38.4	50.6	10
Our Approach (T)	15.2	39.8	49.3	8.5	16.4	40.9	54.8	9
Our Approach (T+I)	15.4	40.6	50.1	8	17.3	42.5	57.4	7

表 3. Flickr 8K 的检索实验。 分别使用[38]，[16]和[28]每一行中的协议。详情请参阅文本。

	Sentence Retrieval				Image Retrieval			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
DeViSE [12]	4.5	18.1	29.2	26	6.7	21.9	32.7	25
DeepFE+FT [20]	16.4	40.2	54.7	8	10.3	31.4	44.5	13
RNN+VGG	10.2	26.9	36.7	22	7.6	21.3	31.4	27
Our Approach (T)	11.3	30.1	43.2	16	8.2	24.7	37.0	22
Our Approach (T+I)	11.9	32.9	45.1	14	8.4	25.7	36.8	21
M-RNN [28]	18.4	40.2	50.9	10	12.6	31.2	41.5	16
RNN+VGG	14.9	36.7	52.1	11	15.1	41.1	54.1	9
Our Approach (T)	15.8	42.0	57.4	9	17.7	44.9	57.2	7.5
Our Approach (T+I)	16.6	42.5	58.9	8	18.5	45.7	58.1	7

表 4. Flickr 30K 的检索实验。 分别使用[12]和[28]每一行中的协议。详情请参阅文本。

5. 讨论

图片标题描述图像中的对象和它们之间的关系。未来工作的区域是检查图像的顺序探索和它与图象描述间的关系。许多单词对应于空间关系，但我们当前的模型难以检测。最近的一篇论文表明[20]，更好的在图像中进行功能定位可以大大提高检索任务的性能，类似的改善可能出现在下一代描述任务中[10， 44]。

在我们的论文中不探讨使用LSTM模型来显示学习长期概念的能力 [15]。其他的论文[7，42]使用LSTMs 产生令人印象深刻的效果。在未来的工作中，用 LSTM 模型代替我们的RNNs模型来学习与本文类似的双向模型将会很有趣。

总之，我们描述的第一个双向模型--既能够产生新的图像描述又能产生视觉特征。不同于以往的许多使用RNNs 的方法，我们的模型能够长期互动学习。这源自其使用周期性视觉记忆，当读或生成新单词时，他能够学会重构视觉特性。我们基于大量的数据集，在句子生成、图像检索和句子检索任务中展示了最先进的结果。

致谢

感谢 Hao Fang， Saurabh Gupta，Meg Mitchell，Xiaodong He，Geoff Zweig，John Platt 和 Piotr Dolla 的深思熟虑和有见地的讨论，也感谢NVIDIA 公司捐赠的 VC 系列的 GPU 来帮助我们完成此次试验。

参考文献

[1] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, 2005. 2, 4, 5, 6

[2] Y. Bengio, H. Schwenk, J.-S.Sen ́cal,F. Morin,and J.-L.Gauvain.Neural probabilistic language models. In Innovations in Machine Learning, pages 137–186. Springer, 2006.2

[3] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. Neural Networks, IEEE Transactions on, 5(2):157–166, 1994. 1, 2, 3

[4] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollr,and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325,2015. 2, 5, 6, 7

[5] X. Chen and C. L. Zitnick. Learning a recurrent visual representation for image caption generation. arXiv preprint arXiv: 1411.5654, 2014. 4, 8

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 248–255. IEEE, 2009. 4, 5

[7] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. CVPR, 2015. 2, 8

[8] D. Elliott and F. Keller. Comparing automatic evaluation measures for image description. 2014. 7

[9] J. L. Elman. Finding structure in time. Cognitive science,14(2):179–211, 1990. 1

[10] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. CVPR, 2015. 2, 8

[11] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In ECCV, pages 15–29. Springer, 2010. 2

[12] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In Advances in Neural Information Processing Systems, pages 2121–2129, 2013. 2, 7, 8

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. 2014. 2, 5

[14] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In ECCV, pages 529–545, 2014. 2

[15] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997. 1, 2, 8

[16] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluationmetrics. J. Artif. Intell. Res.(JAIR), 47:853–899, 2013. 1, 2, 7, 8

[17]Y.Jia, E.Shelhamer, J.Donahue, S.Karayev, J.Long, R.Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. arXiv preprintarXiv:1408.5093, 2014. 4, 5

[18] M. A. Just, S. D. Newman, T. A. Keller, A. McEleney, and

P.A.Carpenter. Imageryinsentencecomprehension: anfmri study. Neuroimage, 21(1):112–124, 2004. 1

[19] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. CVPR, 2015. 2, 5

[20] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. arXivpreprint arXiv:1406.5679, 2014. 1, 2, 8

[21] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In ICML, 2014. 2

[22] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539, 2014. 2

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012. 2, 4, 5

[24] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In CVPR, pages 1601–1608. IEEE, 2011. 2, 4, 6, 7

[25] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. 2012. 2

[26] L. R. Lieberman and J. T. Culpepper. Words versus objects: Comparison of free verbal recall. Psychological Reports, 17(3):983–988, 1965. 1

[27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In ECCV, 2014. 2, 4, 5

[28] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. arXivpreprint arXiv:1410.1090, 2014. 2, 5, 7, 8

[29] T.Mikolov. Recurrent neural network based language model. 1, 2, 3, 4, 5

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. International Conference on Learning Representations: Workshops Track, 2013. 3

[31] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. Cernocky. Strategies for training large scale neural network language models. In Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on, pages 196–201. IEEE, 2011. 4

[32] T. Mikolov and G. Zweig. Context dependent recurrent neural network language model. In SLT, pages 234–239, 2012. 1, 2, 3, 5

[33] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daumé III. Midge: Generating image descriptions from computer vision detections. In EACL, pages 747–756. Association for Computational Linguistics, 2012. 2, 4, 6, 7

[34] A. Paivio, T. B. Rogers, and P. C. Smythe. Why are pictures easier to recall than words? Psychonomic Science, 11(4):137–138, 1968. 1

[35] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: amethod for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics, 2002. 2, 4, 5, 6

[36] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using Amazon's mechanical turk. In NAACL HLT Workshop Creating Speech and Language Data with Amazon's Mechanical Turk, 2010. 1, 5

[37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014. 4, 5, 7

[38] 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. 1, 2, 7, 8

[39] I. Sutskever, J. Martens, and G. E. Hinton. Generating text with recurrent neural networks. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), pages 1017–1024, 2011. 2

[40] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. CVPR, 2015. 2, 5, 6, 7

[41] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning, pages 1096–1103. ACM, 2008. 4

[42] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. CVPR, 2015. 2, 8

[43] R. J. Williams and D. Zipser. Experimental analysis of the real-time recurrent learning algorithm. Connection Science, 1(1):87–111,

1989. 4

[44] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044, 2015. 8

[45] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In EMNLP, 2011. 2

[46] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu. I2T: Image parsing to text description. Proceedings of the IEEE, 98(8):1485–1508, 2010. 2

[47] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pages 3009–3016. IEEE, 2013. 2