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

**展示、(上下文)参与和识别：采用视觉注意力机制的**

**神经网络图像标题生成**

**Kelvin Xu***y* [KELVIN.XU@UMONTREAL.CA](mailto:KELVIN.XU@UMONTREAL.CA)

**Jimmy Lei Ba**† [JIMMY@PSI.UTORONTO.CA](mailto:JIMMY@PSI.UTORONTO.CA)

**Ryan Kiros**† [RKIROS@CS.TORONTO.EDU](mailto:RKIROS@CS.TORONTO.EDU)

**Kyunghyun Cho***y* [KYUNGHYUN.CHO@UMONTREAL.CA](mailto:KYUNGHYUN.CHO@UMONTREAL.CA)

**Aaron Courville***y* [AARON.COURVILLE@UMONTREAL.CA](mailto:AARON.COURVILLE@UMONTREAL.CA)

**Ruslan Salakhutdinov**†∗ [RSALAKHU@CS.TORONTO.EDU](mailto:RSALAKHU@CS.TORONTO.EDU)

**Richard S. Zemel**†∗ [ZEMEL@CS.TORONTO.EDU](mailto:ZEMEL@CS.TORONTO.EDU)

**Yoshua Bengio***y*∗ [YOSHUA.BENGIO@UMONTREAL.CA](mailto:YOSHUA.BENGIO@UMONTREAL.CA)

*>* Universite´ de Montre´al, † University of Toronto, ∗ CIFAR

**摘要**

通过对机器翻译和对象检测的最新研究，我们提出了一种基于注意力的模型，该模型可以自动地对图像的内容进行描述。我们描述了如何使用标准的反向传播(优化)的确定性方法和通过最大化一个变分下限变量的随机化方法来训练这个模型。我们还通过可视化展示了该模型如何能够自动地学习对突出对象的注视，同时生成输出序列中的对应文字。我们在三个基准数据集上验证了注意力的使用:包括flickr、Flickr30k和COCO

# 介绍

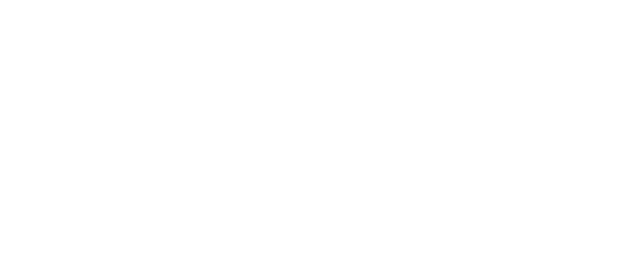
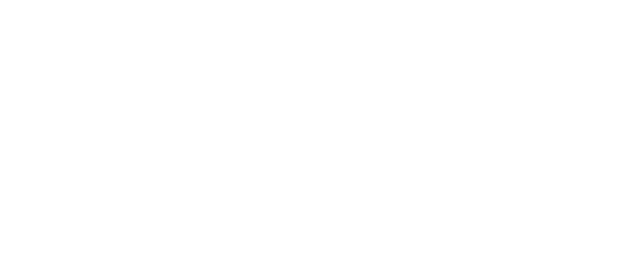
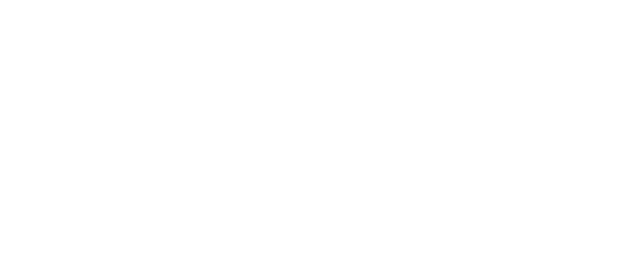
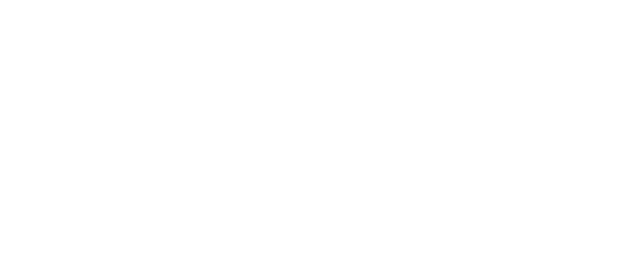
# 自动为图像生成标题是一项接近场景理解中心的任务，这是计算机视觉系统的目标之一。不仅必须要有标题生成模型能够解决计算机视觉上的挑战，以确定图像中的对象是什么。而且它们也必须足够强大，能够捕捉和表达它们在自然语言中的关系。出于这个原因，标题生成一直被视为一个难题。它相当于模仿人类将大量重要的视觉信息压缩成描述性语言的非凡能力，这是机器学习和人工智能研究的一个重要挑战

2015年法国里尔，国际机器学习会议的第32次会议。JMLR:W&CP卷37。Copyright 2015 by the author(s).

*Figure 1.* 图1所示。我们的模型学习了一个单词/图像的对齐方式。视觉化的注意力图(3)在章节1和5.4中解释

|  |
| --- |
| **A** |
| **bird** |
| ﬂying over |
| a  body of |
| **water** |

然而，尽管这项任务的难度非常大，但最近有一项研究对“图像说明生成问题”的研究兴趣大增。在火车-深层神经网络的发展([Krizhevsky 等人.](#_bookmark32),[2012)](#_bookmark32) 以及大型分类数据集的可用性([Russakovsky](#_bookmark58) [等人.](#_bookmark58),[2014)](#_bookmark58)的帮助下，最近的研究通过卷积神经网络(convnets)的结合来显著提高了标题生成的质量，以获得图像的矢量表示和神经网络的重复使用，从而将这些表示编码成自然语言的句子(参见sec2)。人类视觉系统最令人好奇的一个方面就是注意力的存在([Rensink](#_bookmark55), [2000](#_bookmark55);[Corbetta & Shulman,](#_bookmark31)2002[).](#_bookmark31)相比于将整个图像压缩到静态表示中，**注意力(机制)允许突出的特性在需要时动态地保留到最前面（突出重要性）**。当图像中有很多杂乱无关成分的时候，这一点尤其重要。使用表示形式(如从最顶层的一个卷积层)来将图像中的信息分解为最突出的对象是一种有效的解决方案，在以前的工作中被广泛采用。



14x14 Feature Map

**LSTM**

1. Input

Image

2. Convolutional

3. RNN with attention 4. Word by

Feature Extraction over the image

word generation

不幸的是，这有一个潜在的缺点，那就是可能丢失有用的更富有描述性的标题(信息)。

使用低级别的表示可以帮助保存这些信息。然而，使用这些特性需要一种强大的机制来引导模型对当前任务的重要信息进行引导，并且我们展示了如何在不同的位置上学习以生成标题可以实现这一点。**我们提出了两种变体:一种“硬”随机注意机制和一种“软”确定性的注意机制。**我们还展示了包含注意力的一个优势是(可以)通过可视化模型“看到”的东西来获得洞察力。最近在标题生成方面取得的进展以及最近在机器翻译方面([Bahdanau等人.](#_bookmark25),[2014)](#_bookmark25) 和物体识别 ([Ba 等人.](#_bookmark24),[2014;](#_bookmark24)Mnih等人,2014[)的成功所取得的成就鼓舞了人们,](#_bookmark51)  我们研究的模型可以在生成标题的同时，对图像的重要部分进行处理。

这篇论文的贡献如下:

我们在一个共同的框架下(Sec.3.1):1)引入了两种基于注意力的图像标题生成框架，一个“软”的确定性注意力机制，由标准的反向传播方法和2)一个“硬”的随机注意力机制，由最大化的，通过强化REINFORCE ([Williams](#_bookmark69),[1992](#_bookmark69)).，得到一个近似的变化下界或相等的约束机制。

•

我们展示了如何通过将注意力集中在“哪里”和“什么”上(见sec5.4.)来理解这个框架的结果并解释这个框架的结果。

•

最后，我们用最先进的性能(Sec.5.3)在3个基准数据集上，定量验证了注意力在标题生成上的有效性，数据集如下：Flickr8k ([Hodosh 等人.](#_bookmark47),[2013),](#_bookmark47) Flickr30k (Y[oung](#_bookmark72) [等人.](#_bookmark72),[2014)](#_bookmark72) 和 MS COCO 数据集([Lin 等人.,](#_bookmark44) [2014](#_bookmark44)).

•

# 相关工作

在本节中，我们将提供有关图像标题生成和注意力的先前工作的相关背景知识。最近，对于生成图像标题，已经提出了几种方法。这些方法中的许多都是基于循环神经网络的，并且是通过使用神经网络的序列-序列训练([Cho et al.](#_bookmark29),[2014;](#_bookmark29)[Bahdanau](#_bookmark25) [et al.](#_bookmark25),[2014;](#_bookmark25)[Sutskever et al.,](#_bookmark63)2014[;](#_bookmark63)[Kalchbrenner & Blun-](#_bookmark49) [som](#_bookmark49),[2013](#_bookmark49))来获得灵感的。机器翻译的编码-解码器框架(Cho et al，2014)非常适合(这项工作)，因为它类似于“将”图像翻译成句子。

Kiros等人(2014a)提出了使用神经网络的第一个方法，他使用了一种多模式的对数-双线性模型，该模型带有图像的特征。

这一工作后来被Kiros等人(2014b)所采用，他的方法被设计成明确的低水平，以一种自然的方式来进行排名和生成。 Mao等人(2014)使用了一种类似的方法，但用一种重复的方法代替了前馈神经语言模型。Vinyals等人(2014)和donahue 等人(2014)的模型都使用了基于长短期记忆(LSTM)单([Hochreiter &](#_bookmark46) [Schmidhuber](#_bookmark46),[1997)](#_bookmark46) 的循环神经网络(RNN)。与Kiros等人(2014a)和Mao等人(2014)的模型在每一次输出词序列的每一步中都看到了图像不同的是，Vinyals等人(2014)只在开始时将图像展示给RNN。除了图像，Donahue等人(2014)和Yao等人(2015年)也将LSTM应用到视频中，允许他们的模型生成视频描述。.

大多数这些模型都将图像展示为从预先训练过的卷积网络的顶层的一个单一的特征矢量。[Karpathy & Li(](#_bookmark52)2014[)](#_bookmark52) 建议学习一个联合嵌入空间，用于排名和生成，它的模型学会给句子和图像的相似性评分，用双向RNN的输出作为R-CNN对象探测的函数。Fang等人(2014)提出了一种通过结合物体探测来生成的3级管道。他们的模型首先基于多实例学习框架来学习几个视觉概念的检测器。一种经过文字说明训练的语言模型随后被应用到探测器的输出上，然后再从一个联合图像-文本嵌入空间中重新进行。

**与这些模型不同的是，我们提出的注意力框架并没有明确地使用对象探测器，而是从一开始就从已有的(隐含标准)学习潜在的比对。这使得我们的模型超越了“对象”，并学习了抽象的概念。**

在使用神经网络生成标题之前，有**两种**主要方法占主导地位。第一个涉及的生成标题模板是基于物体探测和属性发现的结果(Kulkarni等人(2013)、李等人(2011)、杨等(2011)、米歇尔等人(2012)、埃利奥特和凯勒(2013))。第二种方法是基于从一个大型数据库中检索类似的标题，然后修改这些检索的标题，以适应查询(Kuznetsova等人，2012年，2014年)。这些方法通常涉及到一个中间的“泛化”步骤，以删除仅与检索到的图像相关的标题的细节，例如一个城市的名称。这两种方法后来都落后(淘汰)于现在占主导地位的神经网络方法。

在把注意力机制的想法融合到神经网络中的工作已经进行了有很长的一段时间。一些人与我们的工作有同样的想法，包括Hinton和Denil(2010);Denil等人(2012);唐等人(2014)，以及最近的格莱格等人(2015)。特别是，我们的工作直接扩展了Bahdanau等人(2014)的工作;Mnih等(2014);巴等(2014);格拉夫(2013)。

# 使用注意力机制的图像标题生成

## 模型细节

在这一节中，我们描述了两种不同的注意力模式模型。首先描述它们的共同框架。**关键的不同之处在于*φ*函数的定义，我们将在Sec.4部分进行详细描述。**

*Figure 2.***LSTM单元，带有粗体的方格表示有一个学习的权向量的投影。每个单元学习如何衡量其输入组件(输入门)，同时学习如何调节对记忆的贡献(输入调制器)。它还学习了消除记忆单元(忘记门)的权重，以及控制如何释放记忆的权重(输出门)**

图1显示了所提出的模型的图示。我们用粗体的字体表示向量和用带有大写字母的字母表示矩阵。 为了可读性我们在接下来的描述中将忽略偏置项。

**ht-1**

**zt**

**ht-1**



***i***

**zt**

**Eyt-1 ht-1**

***o***

**Eyt-1**

**输入门 *输出门***

**zt *c* ht**

* + 1. **编码器:卷积得到的特征**

我们的模型输入一个单一的未经处理的图像，生成一个用一组1-k编码的单词编码的标题(描述) **y**

***y* = {y1*, . . . ,* y*C*} *,* y*i* ∈ R*K***

**K**是词汇表的大小 **C**是标题的长度.

**Eyt-1**

**输入编制器**

***f***

**ht-1**

**zt**

**记忆单元**

**忘记门**

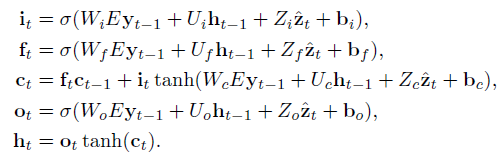
**Eyt-1**

我们使用卷积神经网络来提取一组特征向量，我们称之为**注释向量**。提取器产生L个向量，每一个都是与图像的一部分对应的D维表示

***a* = {a1*, . . . ,* a*L*} *,* a*i* ∈ R*D***

为了获得特征向量和二维图像的部分之间的对应关系，我们从一个**较低的卷积层中**提取特征，这与之前使用一个完全连接的层不一样。这允许解码器通过加权所有特征向量的子集，选择性地关注图像的某些部分。

* + 1. **译码器:长的短期记忆(LSTM)网络**

我们使用一个长的短期记忆(LSTM)网络(Hochreiter&施密特huber，1997)，它在每次的步骤中产生一个单词后，基于一个上下文向量和之前的隐藏状态和之前生成的单词的条件下产生一个标题。我们的LSTMs的实现，如图2所示，贴近Zaremba等人(2014)使用的模型

学习到的矩阵权重和偏置。是嵌入矩阵。

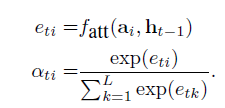
让**m**和**n**分别表示嵌入空间和LSTM的维数，***σ*** 表示逻辑s(sigmoid)型激活

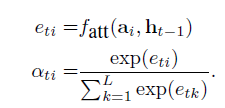
简单地说，上下文向量 **z**ˆ*t* 是在**t** 时对图像输入的相关部分一个动态的表示。我们定义一个计算机制***φ***从注释向量**a*i, i* = 1*, . . . , L ,L***与在不同图像位置提取的特征相对应。中计算 **zˆ*t* .**

对于每一个位置，这个机制产生一个正的权重 ***i*** 可以被理解为是i是正确的来产生下一个单词的位置的可能性(随机注意机制)， **或者作为将位置i和a*i*（图像所在部分）嵌合的相对重要性** (一种确定性的注意力机制)

每个注释向量***αi***的权重由一个注意力模型*f*att计算

在这里我们使用了一个多层感知器，基于在之前的隐藏状态**h***t*−1。要强调的是，我们注意到隐藏状态随着**RNN**在它的输出序列上的进行而变化:

下一步网络走向 “哪里”取决于已经生成的单词序列。



**这里** **分别是LSTM输入、遗忘、记忆、**

一旦权重(相加总和为1)计算完毕，上下文向量***z*ˆ*t*** 将由下面公式计算：

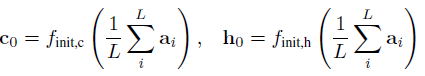
**zˆ*t* = *φ* ({a*i*} *,* {*αi*}) *,***(1)

**输出和隐藏状态**.**。 *W*•, *U*•, *Z*•** 和 **b 是**

***φ*** 是一个当输入注释向量集合和它们对应的权重将输出一个单一的向量的函数

***φ***函数的细节将在Sec.4讨论

LSTM的初始记忆状态和隐藏状态是通过两个独立的MLPs(init,c和init,h)过程，以注释向量的平均值为输入来预测获得的。



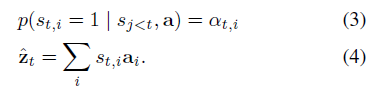
在这个工作中，我们使用一个深的输出层(Pascanu 等人，2014)来计算输出词的概率。它的输入是来自图像(上下文矢量)、前面生成的单词和解码器(*ht*)的状态。



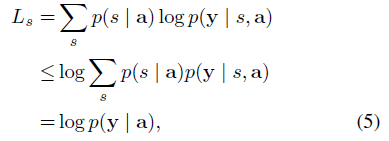
其中，**E是被随机初始化的学习参数**

**4.1.随机“硬”注意力机制**

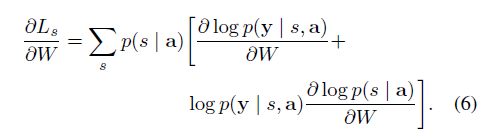
我们表示在生成第t个单词时，模型决定集中注意力在哪个部分的位置变量***st***。***st,I***是一个独热编码的指示器，如果第i个位置)是用于提取视觉特征的一个变量，那么它将被设置为1。通过将注意力位置作为中间的潜在变量，我们可以分配一个由**{*αi*}**确定的多项分布参数，并将**z**ˆ*t*看作一个随机变量:

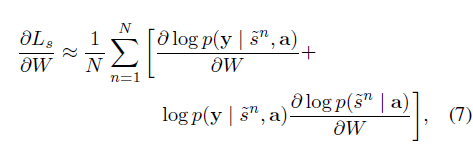


我们定义了一个新的目标函数***Ls***，它是一个在边界对数似然log*p*(**y**|**a**)上的一个变分界，它观察了给定的图像特征**a**的单词序列**y**。类似于生成的深层生成模型(Kingma&Welling，2014;Rezende等人，2014年)，模型的参数*W*的学习算法可以直接由下面进行优化：



还有它的梯度：



我们通过蒙特卡洛方法估计***Ls***的梯度，因此：  


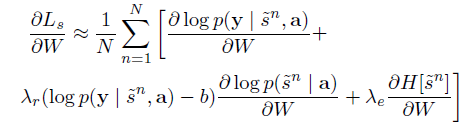
这里，是抽样得到的注意力位置的序列。我们从等式Eq.(3)定义得到的多项分布抽样得到***st***位置：



我们用移动的平均基线技术([Weaver & Tao](#_bookmark68),[2001).](#_bookmark68)来减少这个估计量的方差。当看到第k个小批量数据时，移动的平均基准线被估计为前一个指数似然的累积和指数衰减：

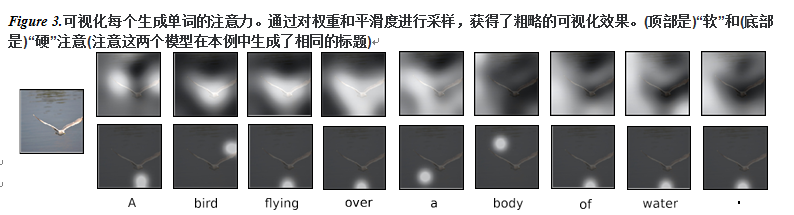


为了进一步减少估计量的方差，多项的熵*H*[*s*]的梯度被增加到等式Eq(7)的RHS中，这个模型的最终学习规则是：



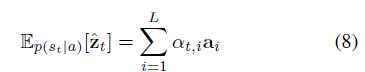
***λr*** 和***λe***是通过交叉验证设置的两个超参数。正如Ba[等人(](#_bookmark24)2014[)](#_bookmark24) 和[Mnih 等人.(](#_bookmark51)2014[),](#_bookmark51)所指出的，这一公式相当于强化学习规则(Williams，1992)，在这种情况下，注意力机制选择一系列行为的奖励(增强因子)是与样本注意力轨迹下目标句子的对数似然性成比例的真实值。

为了进一步提高学习规则的健壮性，在给定图像的概率为0.5的情况下，我们将取样的注意力位置*s*˜设置为它的期望值 *α*(相当于Sec.4.2的确定性注意力)。



## 确定性“软”注意力机制,

学习随机的注意力需要每次都对注意力***st***位置进行采样，而我们可以直接采用上下文向量**z**ˆ*t*的期望



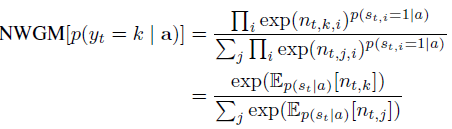
和通过计算一个用软注意力赋予权重的注释向量来公式化一个确定性注意力模型，正如Bahdanau等人所提议的那样。

This corresponds to feeding in a soft α weighted context into the system. The whole model is smooth and differentiable under the deterministic attention, so learning end-to-end is trivial by using standard back-propagation.

这对应于将一个“软”*α*加权的上下文输入到系统。整个模型在确定性的注意力是平滑和可微的，因此，通过使用标准的反向传播学习端到端是无关紧要的。

学习确定性的注意力也可以理解为从sec1.4.1中随机变量***st*** 注意力位置上近似优化等式Eq.(5)的边缘似然。LSTM的隐含激活变量**h***t*是随tanh非线性计算的随机上下文矢量**z**ˆ*t*的线性投影。对于一阶泰勒近似，期望值等价于使用期望的上下文向量来进行前向计算而得到的**h***t*。

让我们用**n*t,i***代替**n**表示在等式Eq(2)中通过**a***i*.推出的**z**ˆ*t*。然后，我们可以写出预测第k个词最大可能性的标准加权几何平均值(NWGM)



这意味着，预测词可能性的NWGM可以通过期望的上下文向量E [**z**ˆ*t*],来进行估算，而不是采样的上下文向量**a***i*. 此外，从baldi和Sadowski(2014)的结果中，在Eq中(9) 可以通过一种单一的前馈计算得到的NWGM估计了基于随机变量*st*.诱导生成的所有可能的注意力位置输出的E[*p*(*yt*=*k|***a**)]期望值

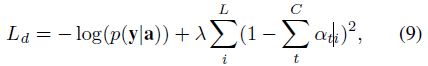
4.2.1. 双重随机注意力机制

在训练我们模型的确定性版本时，我们介绍这是一种双重随机的规则，鼓励这个模型对图像的每个部分都给予相同的关注。然而通过构造在每一个时间点的注意力相加为1，注意力无论如何都不会受到限制。这使得对于解码器忽略输入图像的某些部分成为可能。

为了减缓这种影响，我们提议这里 。在我们的实验中，我们观察到这个处罚量化地改善整体性能并定性地引导更多的描述性标题。

此外，“软”注意力模型在每一个时刻t步骤时从之前的隐含状态预测一个门(阈值)的标量，因此 此处. **这里的门(阈值)变量让编码器在每一个时刻步骤决定是否将重点放在语言建模上或是上下文。**定性地,我们注意，门控变量大于描述图像中的一个对象的编码器。

“软”注意力模型通过最小化以下的惩罚负对数似然来训练端到端：



这里简单地把***τ***设置为1

**4.3 训练过程**

我们的注意力模型的变异体都是通过自适应学习率的随机梯度下降训练的。对于Flickr8k数据集，我们发现，RMSProp（y优化方法）(Tiele-man&Hinton，2012)是最好的，而对于Flickr30k/MS COCO 数据集，Adam 优化算法(Kingma，2014)是非常有效的。

***Table 1.* BLEU-1,2,3,4/METEOR 度量标准与其他方法相比，**†**表示一个不同的分割，(-)表示一个未知的度量，o表明作者通过个人通信提供了缺少的度量，Σ表示一个整体，*a*表示使用AlexNet表示**

*† ◦*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | **BLEU** | | | |  |
| Dataset | Model | | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR |
| Flickr8k | Google NIC([Vinyals et al.](#_bookmark67),[2014)](#_bookmark67) †Σ  Log Bilinear ([Kiros et al.](#_bookmark57),[2014a)](#_bookmark57) ◦  Soft-Attention Hard-Attention |  | 63  65.6  **67**  **67** | 41  42.4  44.8  **45.7** | 27  27.7  29.9  **31.4** | —  17.7  19.5  **21.3** | —  17.31  18.93  **20.30** |
| Flickr30k | Google NIC†◦Σ  Log Bilinear Soft-Attention Hard-Attention |  | 66.3  60.0  66.7  **66.9** | 42.3  38  43.4  **43.9** | 27.7  25.4  28.8  **29.6** | 18.3  17.1  19.1  **19.9** | —  16.88  **18.49**  18.46 |
| COCO | CMU/MS Research ([Chen & Zitnick](#_bookmark30),[2014)](#_bookmark30)  MS Research ([Fang et al.](#_bookmark40),[2014)](#_bookmark40) †*a* BRNN ([Karpathy & Li](#_bookmark52),[2014)](#_bookmark52) ◦ Google NIC†◦Σ  Log Bilinear◦ Soft-Attention Hard-Attention | *a* | —  — 64.2  66.6  70.8  70.7  **71.8** | —  — 45.1  46.1  48.9  49.2  **50.4** | —  — 30.4  32.9  34.4  34.4  **35.7** | —  — 20.3  24.6  24.3  24.3  **25.0** | 20.41  20.71  —  — 20.03  **23.90**  23.04 |

为了创建被我们编码器使用的注释向量ai

我们使用牛津大学的VGGnet(simon yan&Zisserman，2014)在ImageNet上进行了培训，但没有进行最后的微调。在我们的实验中，我们使用了在特征各维最大合并(max pooling)前的第4个卷积层的14\*14\*512特征图。这意味着我们的解码器在扁平的196\*512(i.e L D)编码。**然而，原则上，任何编码函数都可以使用**。此外，有了足够的数据，编码器也可以从模型的其他部分(或微调)中得到训练。

由于我们的实现需要时间与每次更新最长的句子的长度成比例，我们发现在一个随机的标题组上进行训练是非常浪费的。为了减轻这个问题，在预处理过程中，我们构建了一个字典，将一个句子的长度映射到相应的标题的子集。然后，在训练过程中，我们随机抽取一个长度，并取一个长度为64的小批量。我们发现，这大大提高了“收敛速度”，而且在“性能”方面没有明显的减少。在我们最大的数据集(MS COCO)中，我们的“软”模型花了不到3天的时间在NVIDIA的“泰坦”黑GPU训练。.

除了dropout (Srivastava等人，2014年)，我们使用的唯一的规则化策略是在BLEU得分上提前停止。我们观察到在实验的后期训练中，验证集的可能性与BLEU之间的相关性。因为BLEU是最常见的度量指标，所以我们在模型选择的验证集上使用BLEU

在我们的“软”注意力机制实验中，我们使用了Whet-lab1(Snoek等人，2012年，2014年)在我们的Flickr8k实验中。我们从它所探索的超参数区域获得的一些直觉在我们的Flickr30k和COCO中尤为重要。我们将这些模型的代码公开用于鼓励未来在这个领域的研究[2](#_bookmark16).

# 实验

我们描述了我们的实验方法和定量结果，验证了我们模型的有效性。

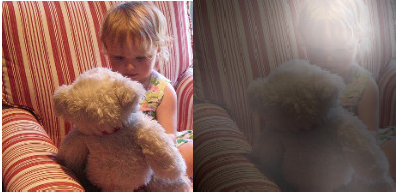
## 数据

我们报告了广泛使用的Flickr8k和Flickr30k数据集的结果，以及更加公开地引入了MS-COCO数据集。在flickr8k/30k数据集中的每个图像都有5个参考标题描述。在对我们的COCO数据集进行预处理时，我们在数据集之间保留了相同数量的引用，通过丢弃超过5的标题。我们只对COCO进行了基本的标记化处理，因此它与Flickr8k和Flickr30k的标记化是一致的。在我们所有的实验中，我们使用了一种固定的词汇量，即10000.

在表1中报告了我们的基于注意力的架构的结果。我们使用常用的BLEU metric3来报告结果这是图像标题生成的标准

*******Figure 4.* 参与正确对象的示例(白色表示参与的区域，突出显示对应的单词)**





我们验证了我们的BLEU评估代码与au-research是一致的。我们从BLEU 1到4的报告中没有一个简短的惩罚。然而，有对BLEU标准的批判，所以我们报告了另一种常见的度量标准METEOR (Denkowski和Denkowski，2014)，并在可能的情况下进行比较

## 评估过程

在结果对比中存在一些挑战。第一个挑战是卷积特性提取器的不同选择。对于相同的解码器架构，使用最近的架构，如GoogLeNet (Szegedy等人，2014)或牛津VGG(Si-monyan&Zisserman，2014)可以在使用AlexNet (Krizhevsky等人，2012)的性能上获得提升。在我们的评估中，我们只与使用GoogLeNet/Oxford VGG特征的结果进行比较，但是对于METEOR比较，我们有一些使用AlexNet的结果。

第二个挑战是一个单一的模型与整体的比较。虽然其他方法通过使用组合方式报告了性能提升，**但在我们的结果中，我们报告了单个模型的性能**

最后，由于数据集的不同存在一个挑战。在我们报告的结果中，我们使用了Flickr8k的预先定义的分割。**然而，对于Flickr30k和COCO数据集来说，没有标准化的分割结果报告**。因此，我们报告了V[inyals et al.(](#_bookmark67) (2014年)、Karpathy和李(2014)和kiros等人(2014年)中使用的公开可用的分割。为了公平起见，我们只对我们的BLEU评估代码是相同的结果进行了比较。

**(BLEU是n-gram精度的几何平均值)。**.

然而，我们注意到，数据集分割的差异并不能在总体表现上造成实质性的差异。

## 定量分析

在表1中，我们提供了一项实验的总结，即对注意力的量化有效性进行测定。我们在Flickr8k、Flickr30k和COCO上获得了最高水准的性能。此外，我们还注意到，在我们的实验中，我们能够在MS COCO数据集上显著地提高最高水准的性能。我们推测这与我们使用的一些规则化技术我们的(图像信息)低级表示有关(参见sec4.2.1) 。

## 定性分析: 学习参与

通过可视化模型所获得的注意力，我们能够为模型的输出添加一个额外的解释层(见图1)。其他已经完成这一任务的系统依赖于对象检测系统来生成候选目标[Karpath](#_bookmark52)[y & Li,](http://cs.stanford.edu/people/karpathy/deepimagesent/)i，2014)。我们的方法更加灵活，因为模型可以处理“非-对象”的突出区域。

19层的OxfordNet使用的是3x3过滤器的堆栈，这意味着只有在最大的池层上，特征映射的缩小才会出现。输入图像的大小调整使最短的边为256维，保持纵横比。卷积网络的输入是中间裁剪的224x224图像。因此，有了四个最大的共享层，我们得到了14x14的顶部卷积层的输出维度。因此，为了使软模型的注意权重可视化，我们对权重进行了采样，使权重为2^4=16，并应用一个高斯滤波器来模拟大的接受域大小。

5

*Figure 5.* 在错误的例子中我们可以利用注意力来获得模型所见的直觉.



.



正如我们在图3和图4中所看到的，模型学习了与人类直觉非常一致的排列。特别是在图5中错误的例子中，我们发现利用这样的可视化来理解为什么会犯错误是有可能的。我们提供了一个更广泛的可视化列表作为读者的补充材料。

# 结论

我们提出一种基于注意力的方法，利用BLEU和METEOR的度量标准，在三个基准数据集的基础上给出了最高水准的表现。我们还展示了如何利用习得的注意力来为模型生成过程提供更多的解释能力，并证明了学习的校准与人类的直觉是非常一致的。我们希望这篇论文的结果能够鼓励未来使用视觉注意力的工作。我们还期望编码-解码器方法的模块化与注意力机制在其他领域能够开创更有用的应用程序。

# 鸣谢

作者们想感谢Theano的开发人员(伯格斯特拉等人，2010年，Bastien等人，2012)。我们感谢以下组织的支持:NSERC，三星，英伟达，计算bec，计算加拿大，加拿大研究协会和CIFAR.作者还想感谢Nitish Srivastava所开发的ConvNet包和准备牛津的卷积网络的帮助，以及帮助解决大量与基础设施相关的问题的Relu Patrascu

.



# 引用文献

Ba, Jimmy Lei, Mnih, Volodymyr, and Kavukcuoglu, Ko- ray. Multiple object recognition with visual attention. *arXiv:*1412.7755 [cs.LG], December 2014.

Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neu- ral machine translation by jointly learning to align and trans- late. *arXiv:*1409.0473 [cs.CL], September 2014.

Baldi, Pierre and Sadowski, Peter. The dropout learning algo- rithm. *Artificial intelligence*, 210:78–122, 2014.

Bastien, Frederic, Lamblin, Pascal, Pascanu, Razvan, Bergstra, James, Goodfellow, Ian, Bergeron, Arnaud, Bouchard, Nico- las, Warde-Farley, David, and Bengio, Yoshua. Theano: new features and speed improvements. Submited to the Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

Bergstra, James, Breuleux, Olivier, Bastien, Fre´de´ric, Lam- blin, Pascal, Pascanu, Razvan, Desjardins, Guillaume, Turian, Joseph, Warde-Farley, David, and Bengio, Yoshua. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, 2010.

Chen, Xinlei and Zitnick, C Lawrence. Learning a recurrent visual representation for image caption generation. *arXiv preprint arXiv:1411.5654*, 2014.

Cho, Kyunghyun, van Merrienboer, Bart, Gulcehre, Caglar, Bougares, Fethi, Schwenk, Holger, and Bengio, Yoshua. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, October 2014.

Corbetta, Maurizio and Shulman, Gordon L. Control of goal- directed and stimulus-driven attention in the brain. *Nature re- views neuroscience*, 3(3):201–215, 2002.

Denil, Misha, Bazzani, Loris, Larochelle, Hugo, and de Freitas, Nando. Learning where to attend with deep architectures for image tracking. *Neural Computation*, 2012.

Denkowski, Michael and Lavie, Alon. Meteor universal: Lan- guage specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Ma- chine Translation*, 2014.

Donahue, Jeff, Hendrikcs, Lisa Anne, Guadarrama, Se-  gio, Rohrbach, Marcus, Venugopalan, Subhashini, Saenko, Kate, and Darrell, Trevor. Long-term recurrent convo- lutional networks for visual recognition and description. *arXiv:*1411.4389v2 [cs.CV], November 2014.

Elliott, Desmond and Keller, Frank. Image description using vi- sual dependency representations. In *EMNLP*, pp. 1292–1302, 2013.

Fang, Hao, Gupta, Saurabh, Iandola, Forrest, Srivastava, Rupesh, Deng, Li, Dolla´r, Piotr, Gao, Jianfeng, He, Xiaodong, Mitchell, Margaret, Platt, John, et al. From captions to visual concepts and back. *arXiv:*1411.4952 [cs.CV], November 2014.

Graves, Alex. Generating sequences with recurrent neural net- works. Technical report, arXiv preprint arXiv:1308.0850, 2013.

Gregor, Karol, Danihelka, Ivo, Graves, Alex, and Wierstra, Daan. Draw: A recurrent neural network for image generation. *arXiv* *preprint arXiv:1502.04623*, 2015.

Hochreiter, S. and Schmidhuber, J. Long short-term memory.

*Neural Computation*, 9(8):1735–1780, 1997.

Hodosh, Micah, Young, Peter, and Hockenmaier, Julia. Framing image description as a ranking task: Data, models and evalu- ation metrics. *Journal of Artificial Intelligence Research*, pp. 853–899, 2013.

Kalchbrenner, Nal and Blunsom, Phil. Recurrent continuous translation models. In *Proceedings of the ACL Confer- ence on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1700–1709. Association for Computational Lin- guistics, 2013.

Karpathy, Andrej and Li, Fei-Fei. Deep visual-semantic align- ments for generating image descriptions. *arXiv:*1412.2306 [cs.CV], December 2014.

Kingma, Diederik P. and Ba, Jimmy. Adam: A Method for Stochastic Optimization. *arXiv:*1412.6980 [cs.LG], De- cember 2014.

Kingma, Durk P. and Welling, Max. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

Kiros, Ryan, Salahutdinov, Ruslan, and Zemel, Richard. Multi- modal neural language models. In *International Conference on Machine Learning*, pp. 595–603, 2014a.

Kiros, Ryan, Salakhutdinov, Ruslan, and Zemel, Richard. Unify- ing visual-semantic embeddings with multimodal neural lan- guage models. *arXiv:*1411.2539 [cs.LG], November 2014b.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey. Ima- geNet classification with deep convolutional neural networks. In *NIPS*. 2012.

Kulkarni, Girish, Premraj, Visruth, Ordonez, Vicente, Dhar, Sag- nik, Li, Siming, Choi, Yejin, Berg, Alexander C, and Berg, Tamara L. Babytalk: Understanding and generating simple im- age descriptions. *PAMI, IEEE Transactions on*, 35(12):2891– 2903, 2013.

Kuznetsova, Polina, Ordonez, Vicente, Berg, Alexander C, Berg, Tamara L, and Choi, Yejin. Collective generation of natural image descriptions. In *Association for Computational Lin- guistics: Long Papers*, pp. 359–368. Association for Computa- tional Linguistics, 2012.

Kuznetsova, Polina, Ordonez, Vicente, Berg, Tamara L, and Choi, Yejin. Treetalk: Composition and compression of trees for im- age descriptions. *TACL*, 2(10):351–362, 2014.

Larochelle, Hugo and Hinton, Geoffrey E. Learning to com- bine foveal glimpses with a third-order boltzmann machine. In *NIPS*, pp. 1243–1251, 2010.

Li, Siming, Kulkarni, Girish, Berg, Tamara L, Berg, Alexander C, and Choi, Yejin. Composing simple image descriptions us- ing web-scale n-grams. In *Computational Natural Language Learning*, pp. 220–228. Association for Computational Lin- guistics, 2011.

Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dolla´r, Piotr, and Zitnick, C Lawrence. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755. 2014.

Mao, Junhua, Xu, Wei, Yang, Yi, Wang, Jiang, and Yuille, Alan. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv:*1412.6632 *[cs.CV]*, December 2014.

Mitchell, Margaret, Han, Xufeng, Dodge, Jesse, Mensch, Alyssa, Goyal, Amit, Berg, Alex, Yamaguchi, Kota, Berg, Tamara, Stratos, Karl, and Daume´ III, Hal. Midge: Generating im- age descriptions from computer vision detections. In *European Chapter of the Association for Computational Linguistics*, pp. 747–756. Association for Computational Linguistics, 2012.

Mnih, Volodymyr, Hees, Nicolas, Graves, Alex, and Kavukcuoglu, Koray. Recurrent models of visual atten- tion. In *NIPS*, 2014.

Pascanu, Razvan, Gulcehre, Caglar, Cho, Kyunghyun, and Ben- gio, Yoshua. How to construct deep recurrent neural networks. In *ICLR*, 2014.

Rensink, Ronald A. The dynamic representation of scenes. *Visual cognition*, 7(1-3):17–42, 2000.

Rezende, Danilo J., Mohamed, Shakir, and Wierstra, Daan. Stochastic backpropagation and approximate inference in deep generative models. Technical report, arXiv:1401.4082, 2014.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, An- drej, Khosla, Aditya, Bernstein, Michael, Berg, Alexander C., and Fei-Fei, Li. ImageNet Large Scale Visual Recognition Challenge, 2014.

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

Snoek, Jasper, Larochelle, Hugo, and Adams, Ryan P. Practi- cal bayesian optimization of machine learning algorithms. In *NIPS*, pp. 2951–2959, 2012.

Snoek, Jasper, Swersky, Kevin, Zemel, Richard S, and Adams, Ryan P. Input warping for bayesian optimization of non- stationary functions. *arXiv preprint arXiv:1402.0929*, 2014.

Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 15:1929– 1958, 2014.

Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc VV. Sequence to sequence learning with neural networks. In *NIPS*, pp. 3104– 3112, 2014.

Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Van- houcke, Vincent, and Rabinovich, Andrew. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.

Tang, Yichuan, Srivastava, Nitish, and Salakhutdinov, Ruslan R. Learning generative models with visual attention. In *NIPS*, pp. 1808–1816, 2014.

Tieleman, Tijmen and Hinton, Geoffrey. Lecture 6.5 - RMSProp. Technical report, 2012.

Vinyals, Oriol, Toshev, Alexander, Bengio, Samy, and Erhan, Dumitru. Show and tell: A neural image caption generator. *arXiv:*1411.4555 [cs.CV], November 2014.

Weaver, Lex and Tao, Nigel. The optimal reward baseline for gradient-based reinforcement learning. In *Proc. UAI’2001*, pp. 538–545, 2001.

Williams, Ronald J. Simple statistical gradient-following al- gorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Yang, Yezhou, Teo, Ching Lik, Daume´ III, Hal, and Aloimonos, Yiannis. Corpus-guided sentence generation of natural images. In *EMNLP*, pp. 444–454. Association for Computational Lin- guistics, 2011.

Yao, Li, Torabi, Atousa, Cho, Kyunghyun, Ballas, Nicolas, Pal, Christopher, Larochelle, Hugo, and Courville, Aaron. Describ- ing videos by exploiting temporal structure. *arXiv preprint arXiv:1502.08029*, April 2015.

Young, Peter, Lai, Alice, Hodosh, Micah, and Hockenmaier, Julia. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.

Zaremba, Wojciech, Sutskever, Ilya, and Vinyals, Oriol. Re- current neural network regularization. *arXiv preprint arXiv:1409.2329*, September 2014.