**展示说明:神经图像标题生成器**

## Oriol Vinyals谷歌

[vinyals@google.com](mailto:vinyals@google.com)

## 亚历山大Toshev谷歌

[toshev@google.com](mailto:toshev@google.com)

## 萨米Bengio谷歌

[bengio@google.com](mailto:bengio@google.com)

## Dumitru Erhan谷歌

[dumitru@google.com](mailto:dumitru@google.com)

# 摘要

*自动描述图像的内容是人工智能的一个基本问题,连接计算机视觉和自然语言处理。
在本文中,我们提出一个基于深重新生成模型当前的体系结构,结合最新进展在com -把视觉和机器翻译和可用于生成自然句子描述图片。
模型训练的可能性最大化目标de -从普通句子训练图像。
几个实验数据集显示模型的准确性和流利的语言学习仅仅从图片descrip——一代长大成人。
我们的模型往往是很准确,我们验证定性和定量。
例如,当前最先进的BLEU-1得分(越高越好)帕斯卡数据集是25,我们的方法产生59,相比人类性能约69。
我们也显示BLEU-1分数Flickr30k的改进,从56到66年,在全局,从19岁到28岁。
最后,新发布的可可数据集,我们实现BLEU-4 27.7,这是当前最先进的。*

# 介绍

能够自动使用正确描述图像的内容形成的英语句子是一个非常具有挑战性的任务,但它可以有很大的影响,例如通过帮助视力受损的人更好地了解图像在网络上的内容。
这个任务明显困难,例如,比研究图像classifi -阳离子或对象识别任务,这一直是一个主要集中在计算机视觉界[27]。
事实上,不仅必须捕捉对象的描述中包含一个图像,但它也必须表达这些对象之间的关系以及它们的属性和它们参与的活动。
此外,上述语义知识——边缘必须表达自然语言(如英语),这意味着需要一个语言模型除了视觉的理解。

大多数先前提出了缝合在一起



愿景深CNN

语言生成RNN

**一群人在露天市场购物。**

**有许多蔬菜水果店。**

图1所示。
网卡,我们的模型,端到端基于神经网络组成的——工作愿景CNN语言总的来说——操作RNN紧随其后。
在自然语言生成完整的句子从输入图像,在上面的示例中所示。

现有解决方案上面的子问题,为了从一个图像到它的描述(6、16)。
相比之下,我们想在这工作一个联合模型,我需要一个图像作为输入,和训练有素的最大化可能p (S I)生[产目标](#_bookmark28)[的单词序](#_bookmark38)列S = S1, S2,……
每个单词圣从哪里来给定的字典,充分描述图像。

*{ }*

*|*

我们工作的主要灵感来源于最近的广告——万斯在机器翻译的任务是把一个句子用源语言编写,到其transla, T在目标语言,通过最大化p (T S)。多年来,机器翻译也通过单独任务的se -里斯(单独翻译单词,对齐——荷兰*国际*集团(ing)话说,重新排序,等等),但是最近的研究表明,翻译可以通过一个简单得多的方式使用Re -当前*神经*网络(RNNs)(3 2、30),同时达到最先进的性能。
“编码器”RNN*读*取*源句子*转换成一个丰富的固定长度的向量表示,进而作为初始隐藏状态的“译码”RNN生成目标句子。

*|*

在这里,我们建议遵循这个优雅的配方,replac——荷兰国际集团(ing)编码器RNN深卷积神经网络(CNN)。
在过去的几年里已经令人信服地表明,cnn可以产生丰富的表示输入图像的嵌入到一个固定长度的向量,这样这个表达式可以用于各种各样的愿景

arXiv: 1411.4555 v2 (cs。
简历)2015年4月20日

1

任务[28][。
因此](#_bookmark50),自然使用CNN作为“编码器”形象,首先训练为一个图像分类任务,使用最后一个隐层作为输入RNN解码器生成句子(见图1)。我们称之为模型神经图像标题,或者网卡。

我们的贡献如下。
首先,我们提出一个端到端系统的问题。
它是一个完全可训练的神经网络使用随机梯度下降法。
第二,我们的模型结合技术发展水平子网为视觉和语言模型。
这些可以pre-trained大全集,因此可以利用额外的数据。
Fi -基因表达,它的收益率明显更好的性能比最先进的方法;
例如,帕斯卡数据集,网卡了蓝色的59岁,是当前最先进的25岁,而人类性能达到69。
Flickr30k,我们改善从56到66年,单位,从19岁到28岁。

# 相关工作

从视觉数据生成自然语言描述的问题,长期以来一直研究在计算机视觉中,但主要用于视频[7,32]。
这导致了复杂系统——显微镜由视觉原始识别器结合结构化的形式语言,例如与或图或逻辑系统,它进一步转化为自然局域网,通过基于规则的系统规[。](#_bookmark29)[这样的系](#_bookmark54)统是上帝——y hand-designed,相对脆弱,被恶魔——strat只在有限的领域,如交通场景或运动。
静态图像与自然文本描述的问题最近得到了利益。
利用最近的广告——万斯在识别的对象,它们的属性和轨迹规划设计,使我们能够驱动自然语言生成系统——显微镜,尽管这些在他们的表达能力是有限的。
哈蒂等。[6]使用检测推断三联体的场景元素转换为文本使用模板。
同样,李et al。[ 19]开始检测和拼凑fi - nal描述使用短语包含检测对象和关系。
更复杂的图之外的检测使用三胞胎Kulkani et al。[16],但基于模板的文本生成。
更强大的语言模型基于语言解析被使用([5]23日1,17日,18日。
上述方法能够描述图像“野生的”,但他们是严重的手

设计和严格的文本生成时。

大量的工作解决问题的等级- - -荷兰国际集团(ing)对于一个给定的图像描述[11日8日24]。
这样的美联社——针对的想法是基于co-embedding相同[图像和文](#_bookmark33)[本的](#_bookmark30)[向量空间](#_bookmark46)。
对查询图像,从普通检索这谎言接近图像在嵌入空间中。
最密切,神经网络用于co-embed图片和句子一起[29]甚至形象作物和subsentences[13]但不要试图产生新颖的描述。
一般而言,上述方法不能

描述以前看不见的对象的组合,尽管单个对象可能被观察到在训练数据。
此外,他们避免解决问题——lem评估生成的描述有多好。

在这个工作我们结合深卷积网im -年龄分类序列建模与复发性网络[12][10],创建一个网络,生成de - - -从普通的图像[。
R](#_bookmark34)NN训练在这个单一的背景下,“端到端”网络。
序列生成的模型是受最近的成功机反[式-副调制](#_bookmark32)(3 2、30)的差异,而不是以一个句子开始,我们提供一个图像处理con - volutional净。
最近的作品Kiros et al。[15]使用神经网络,但是一个前馈,预测下一个单词形象和以前的单词。
最近的一份工作毛et al。[21]使用相同的复发性神经网络预测任务。
这非常类似于目前的提议,但有一些重要的区别:我们使用一个更强大的RNN模型,并提供直接RNN的视觉输入模型,这使得它可能RNN跟踪的对象已经被解释为文[本。](#_bookmark25)[由于这](#_bookmark23)[些看似无关](#_bookmark52)紧要的dif -费伦茨,我们的系统在建立基准达到更好的结果。
最后,Kiros et al。[14],对建立一个联合多通道嵌入空间通过使用强大的计算机视觉模型和一个LSTM编码文本。
与我们的方法相比,他们用两个国家环保总局——利率途径(一个用于图像,一个用于文本)来定义一个联合嵌入,即使他们可以生成文[本,他们](#_bookmark37)的方法是高度调整排名。

# 模型

在本文中,我们提出一个神经和概率框架来从图像生成的描述。
最近统计机器翻译的发展表明,给定一个强大的序列模型,可以实现先进的结果通过直接最大化proba——能正确的翻译提供一个输入句子的“端到端”的方式——培训和推理。
这些模型使用的递归神经网络编码可变长度输入一个固定的维向量,并使用这种表示“解码”德——句子——输出。
因此,人们很自然地使用相同的美联社——友善,给一个图像(而不是源语言中的一个输入句子),适用同样的原则之一“翻译”成它的描述。

因此,我们建议直接给出的正确描述图像的概率最大化利用下- ing配方:

Σ *|*

*θ*٨= arg m axl*o*g  *p (*S*我*;
θ)(1)

θ

(我的)

我们的模型*的*参数θ,我是一个图像,然后呢

*其正*确的转录。
因为年代代表任何句子,其长度是无限的。
因此,它是*常见*的链式法则应用于模型联合概率S0,。
SN, N是这个例子的长度

词的预测



太

LSTM

内存块

σ

输出门f

ct-1

忘记门f

c

σ

σ

输入门我

ct

h

更新的术语

x

softmax

N

Σ

对数p  *(*S  *|我)*= p (S*t*  |*我*S*0。。。**、圣−1)*(*2*)

t = 0

为方便我们把依赖θ。
在训练时,(S I)是一个训练的例子,我们优化*日志*概率之和(2)中描述在整个训练集使用随机梯度下降(在第四节给出进一步培*训的细节*)。

自然模型p (St我S0。
,圣−1)重新当*前*神*经*网*络(RN*N*)的变量数目的字我们条*件*到*t  1所表达的是一个固定长度或内存ht隐藏状态。
这个记忆是更新后看到一个新的输入xt通过非线性函数:

*|*

*−*

输入

*h*t + 1  *=*  f  *(*h*t,*  xt*)* 。(3)

使上述RNN更具体的两个关键设计选择是:什么是f的具体形式以及美联储作为输入xt的图片和文字。
f我们使用多空词记忆(LSTM)净,这表明先进的性能在序列翻译等任务。
这个模型是在下一节中列出。

表示的图像,我们使用一个卷积神经网络(CNN)。
他们被广泛应用和研究图像任务,和目前先进的目标识别和检测。
我们的特定选择的CNN使用一种新颖的方式批量标准化和收益率目前的最佳性能[12]2014年ILSVRC分类竞争。
此外,他们可以推广到其他任务,比如场景一堂课——sification通过转移学习[4]。
用一个嵌入模型表示。

### LSTM-based句子发电机

f的选择(3)是由其处理能力*消失*和爆炸[梯度[1](#_bookmark3)0],最RNNs com——我的挑战设计和培训。
应对这一挑战,一个特定形式的经常性的网,叫LSTM,介绍了[10]和应用取得了巨大的[成功翻译(](#_bookmark32)3、30)和序列生成[9]。

LSTM模型的核心是一个存储单元c encod - ing知识在每一个时间步的输入已经观*察到*这一步(参见图2)。
细胞的行为是由“盖茨”——控制层的应用——tiplicatively和因此可以保持一个值从封闭层如果门1或0这个值,如果门是0。
在

特别地,有三门控制是否使用

忘记当前单元格的值(忘记门f),如果它应该

图2。
LSTM:内存块包含一个细胞c是由三个门。
在蓝色我们显示经常性connec -设计*- m*在时间t 1输出反馈的记忆在时间t通过三个门;
细胞值反馈通过忘记门;
预测词在时间t 1反馈除了内存输出m t时刻为词将Softmax预测。

读它的输入(输入门我)以及是否输出新的细胞值(输出门o*)*。盖茨和细胞更新和输出的定义如下:

*−*

*−*

*它*=σ(W*i*x*x*t +  *W*im*m*t−*1*)  *(*4)

*英*国《金融*时*报*》*=σ*(*Wf*x*xt  *+*  W*f*mm t−1) (5)

*不*=σ(W*o*x*x*t +  *W*om*m*t−*1*)  *(*6) ct  *=*英尺Ⓢct*−*1  *+Ⓢ*h  *(*Wcx*x*t  *+ W*c*m*mt*−*1)(*7*)太*=*  o*t*Ⓢct (*8*)

*p*t +  1 = Softm*a*x  ( mt) (9)

与门值,代表产品,各种W矩阵训练参数。
这样多plicative盖茨能够训练LSTM强劲作为交易这些门爆炸和消失的草——d*ie*nts [10]。
非线性是乙状结肠σ()和过度bolic切线h ()。
最后方程太Softmax是用来喂养,这将产生一个概率distri - bution pt /所有单词。

*Ⓢ*

*·*

*·*

**培训LSTM模型**训练预测句子的每个单词后看到的图像以及所有前面的单词所定义的p (St我S0。
、圣−1)。
为了这个目的,它是指导把LSTM联合国-滚形式的副本创建LSTM内存

*|*

p1

日志p1 (S1)

p2

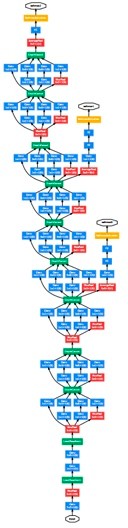
日志p2 (S2)

pN

日志pN (SN)

图像

...



S1

SN-1

S0

WeSN-1

LSTM

LSTM

|  |  |
| --- | --- |
| LSTM | |
|  |  |
| WeS0 | |

|  |  |
| --- | --- |
| LSTM | |
|  |  |
| WeS1 | |

图3。
LSTM模型结合CNN图像嵌入([12]中定义)和嵌入。
的展开connec LSTM之间,记忆是蓝色的和他们来往,spond[反复连接如图](#_bookmark34)2所示。
所有LSTMs共享相同的参数。

形象和每个句子的词,这样所有LSTMs共享相同的参数和输出太−1在时间t的LSTM 1是美联储LSTM在时间t(参见图3)。所有反复连接转换为前馈con - necti*o*n*s*展开版本。
在更多的细节,如果*我*们表示我输入图像和S = (S0,。。
,*S*N)真正的森-对于描[述这张](#_bookmark5)图片,展开过程写着:

*−*

*x−*1  = CNN  *(*I ) (10)

*x*t  =西*方*,*t*∈*{ 0。*
*N−1} (*1*1* )

*p*t +  1 = LS*T*M  (*x t),*  t*∈{0。*
*N* −1} (12)

我们代表了每个单词的一个炎热的向量圣维等于字典的大小。
注意,我们表示S0特殊启动词和SN特殊的*停*止词指定句子的开始和结束。
特别是通过发射停止词LSTM信号,生成一个完整的句子。
图像和文字被映射到相同的空间,图像通过使用视觉CNN,我*们*通过使用嵌入’这个词。
图片我只输入一次,在t = 1,*通*知LSTM图像内容。
我们实际验证,喂养的图像在每个时间步作为额外输入收益率差的结果,随着网络可以显式地利用图像中的噪声和overfits更容易。

*−*

我们的损失负对数似然的和正确的词在每个步骤如下:

N

Σ

*L*(*我的)*=−*日* 志pt  *(*St*)*。(*1* 3)

t = 1

上述损失最小化w.r.t. LSTM的所有参数,图像的顶层嵌入CNN和嵌入的我们。

**推理有多种方法可以**用来生成一个句子给一个图像,网卡。
第一个是抽样,我们只是示例第一个单词ac - p1绳索,然后提供相应的嵌入作为输入样本p2,继续像这样直到我们样品或者一些最大长度的特殊的句子结束标记。
第二个是BeamSear**ch:反复考虑的**k最好的句子到时间t的候选人来生成句子大小t + 1,最好只保留结果k。
这更好的近似*S*  =参数最大值的p (S′)。我们使用了BeamSearch方法的力量在以下试验,用一束大小20。
使用1(即光束的大*小*。
、贪婪搜索)并降低我们的结果平均2蓝色点。

# 实验

*|*

我们进行了大量的实验来评估我们的模型使用几个指标的有效性,数据源,和模型架构,为了比较现有技术。

### 评价指标

尽管有时候不清楚描述应该被视为成功与否给一个图像,现有技术提出了若干评价指标。
最re -责任(但耗时)要求评级机构给主观分数每个描述图像的有效性。
在本文中,我们使用的加强,一些自动指标确实与这个”——有效得分,在[11]指导方针提出后,要求学生评价每个生成的句子与规模从1到41岁。

对于这个指标,我们建立一个亚马逊土耳其机器人实验。
每个图像被评为2工人。
typ - ical的工人之间的协议是65%。
在分歧的情况下我们只是平均分数和记录的平均得分。
方差分析,我们执行

引导(重采样的结果替换和计算方法/标准偏差在重新取样re -结果)。
像[11]我们报告分数的分数或等于比一组预定义的阈值。

其余的指标可以自动计算假设访问groundtruth,即人类创——害死描述。
图像中迄今为止最常用的度量描述文学一直是蓝色分数[25],这是一种精确的词字格生成和引用的句子之间2。
尽管

这个评级机构被问到描述的图像没有任何er - ror,与小错误描述,有些相关的描述,或与一个不相关的描述,4分是最好的,1是最坏的打算。

2在这文学,BLEU-1大多数以前的工作报告。
,他们只计算精度在unigram层面,而BLEU-n是精密的几何平均1 -字格。

这个指标有一些明显的缺点,它已被证明与人类相关的评价。
在这项工作中,我们证实这,正如我们在4.3节。
一个广泛的评估协议,以及生成的输出的系统,可以在http://nic上找到。
droppages.com/。

除了蓝色,一个可以使用模型的困惑

对于一个给定的转录(这是我们的目标函数(1)密切相关)。
困惑是几何平均数的倒数为每个预测词概率。
我们用这个指标来执行选择模型selec伸出组,和hype[rpara](#_bookmark1)meter调优,但我们不报告由于蓝色总是优先3。
可以找到更详细讨论有关指标[31],和关于这一主题的研究小组工作报告的其他指标被认为更适合评估说明。
我们报告两个这样的指标——我——TEOR和苹果酒——希望出现更多的讨论和研究,对指标的选择。

最后,当前文献图像描述也被使用的代理任务排名一组效果,可以描述对一个给定的图像(见的立场[14])。
这样做的优点是可以使用已知的像recall@k排名指标。
另一方面,将描述一代任务到一个排名是不满意:随着图像复杂度的de -文士的增长,连同它的字典,可能的句子的数量呈指数级增长的大小字典,和一个预定义的句子的可能性将适合新形象的数量会下降,除非这样[的句子也呈指](#_bookmark36)数级增长,这是不现实的;
更不用说底层的计算复杂度评估有效的这样一个大型语料库存储森——却常常为每个图像。
同样的论点已经用于语音识别,其中一个生产相对应的句子给声序列;
在早期,诱使集中在孤立的音素的分类或单词,最先进的方法这个任务已经生成,可以产生从大型dictio——连句子。

现在,我们的模型可以生成rea - sonable质量的描述,尽管评估图像描述的歧义(可能有多个有效的描述不是groundtruth)我们认为我们应该专注于评估指标生成任务,而不是排名。

### 数据集

评估我们使用大量的数据集,包括图片和句子用英语描述这些图片。

3即使它,更可取的方法应该是优化的蓝色得分收益率离散优化问题。
一般来说,困惑和蓝色分数是非常相关的。

数据集的统计如下:

|  |  |  |  |
| --- | --- | --- | --- |
| 数据集名称 | 大小 | | |
| 火车 | 有效的。 | 测试 |
| 2008年帕斯卡VOC [[6]](#_bookmark28) | - | - | 1000 |
| Flickr8k  [[26]](#_bookmark48) | 6000 | 1000 | 1000 |
| Flickr30k  [[33]](#_bookmark55) | 28000 | 1000 | 1000 |
| MSCOCO  [[20]](#_bookmark42) | 82783 | 40504 | 40775 |
| 业务单[位[24]](#_bookmark46) | 1米 | - | - |

除了业务单位,每个图像已经被贴标签机注释5句子相对视觉和公正的。
单位由图像所描述的所有者时将它们上传到Flickr。
因此他们不能保证视觉或公正,因此这个数据集有更多的噪音。

帕斯卡数据集通常用于测试只有af - ter系统一直在训练不同的数据,如其他四个数据集。
在单位的情况下,我们坚持1000图像进行测试和训练所使用的其他[18]。
Sim - ilarly,我们保留4 k随机从MSCOCO验证设置为测试图像,称为COCO-4k,报告结果在下一节中使用它。

### 结果

自我们的模型是数据驱动的,训练有素的端到端,并给予丰富的数据集,我们想要一个- sw这样的问题“数据集大小如何影响一般——ization”,“什么样的转移学习它能够实现“,和”如何处理弱标记的例子”。
因此,我们在五个不同的数据集进行实验,4.2节中解释的那样,使我们了解我们的模型在深度。

#### 培训细节

我们面临的许多挑战,当训练我们的模型过度拟合。
事实上,纯粹的监督方法需要大量的数据,但数据集,是高质量的不到100000的图像。
分配的任务是严格描述比对象分类和数据驱动方法最近才成为主要由于数据集一样大ImageNet(十倍数据比我们本文中描述的数据集,除了全局。
因此,我们相信,即使我们得到的结果都不错,我们的方法的优势与最新的工程化方法只会增加在未来几年内随着训练集规模将增长。

尽管如此,我们探索了一些技术处理过度拟合。
最明显的方法不是overfit ini——tialize CNN重量组件的系统pretrained模型(例如,在ImageNet)。
我们做了这个实验(类似于[8]),它确实帮助很大

很多方面的概括。
另一组的权重可以明智地初始化我们,“嵌入”这个词。
我们初始化它们从一个大新闻语料库[22],但没有观察到显著的收益,为了简单起见,我们决定把他们未初始化。
最*后*,我们做了一些模型层级overfitting-avoiding技术。
我们尝试辍学[34]和整体模型,以及探索(即大小。
,容量)模型的交易数量的[隐藏单元与](#_bookmark44)深度。
辍学和合奏给几个蓝色点改进,这就是我们在整个论文报告。

我们训练有素的权重集合使用随机gradi - ent下降和固定的学习速率和动量。
所有重量都是随机初始化除了CNN权重,我们保持不变,因为改变他们带来负面的影响。
我们使用em - 512维箱包LSTM内存的大小。

描述与基本预处理标记,将所有单词出现至少5次火车——荷兰国际集团(ing)。

#### 生成结果

我们报告我们的主要结果在所有相关的数据集在Ta -提单1和2。
帕斯卡以来没有一个训练集,我们使用了系统训练使用MSCOCO(可以说[是这](#_bookmark13)个任务的[最大](#_bookmark14)和最高质量的数据集)。
帕斯卡和业务单位的状态——艺术的结果没有使用基于深度学习的图像特征,因此可以说是一大提高,这些成绩来自于变化。
最近使用Flickr数据集[14]11日,21日,但主要是评估在一个检索框架。
著名的前任——ception[21],他们做了检索和属,在Flickr和得到最好的性能数据。

人类分数在表2的计算是通过比较一个人类对其他[四个](#_bookmark14)标题。
我们为每个5评级机构,和平均他们蓝色的分数。
因为这给了一个微弱的优势,我们的系统,考虑到蓝色的分数计算对5引用的句子,而不是四个,我们添加回人类分数的平均差异有五个引用,而不是四个。

鉴于该领域已经出现了明显进步在过去的几年,我们认为更有意义的报告BLEU-4,这是机器翻译的标准mov——荷兰国际集团(ing)前进。
此外,我们报告指标显示与人类共同与更好的评估表14。
尽管最近努力更好的评价指标[31],我们的模型票价强烈和评委。
然而,当国际化程度——操作我们的标题使用人工评分(参见4.3.6节),我们的模型表现更差,建议更多的工作

4我们使用这些指标的实现请提供

[http://www.mscoco.org](http://www.mscoco.org/)。

|  |  |  |  |
| --- | --- | --- | --- |
| 度规 | BLEU-4 | 流星 | 苹果酒 |
| 网卡 | **27.7** | **23.7** | **85.5** |
| 随机 | 4.6 | 9.0 | 5.1 |
| 最近的邻居 | 9.9 | 15.7 | 36.5 |
| 人类 | 21.7 | 25.2 | 85.4 |

表1。
分数MSCOCO发展集。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 方法 | 帕斯卡  (转移) | Flickr  30 k | Flickr  8 k | 在布 |
| Im2Text  [[24]](#_bookmark46)  TreeTalk  [[18]](#_bookmark41)  话题[16]  Tri5Sem  [[11]](#_bookmark33)  m-RNN  [[21]](#_bookmark43)  MNLM  [[14](#_bookmark36)[] 5](#_bookmark15) | 25 | 55  56 | 48  58  51 | 11  19 |
| SOTA | 25 | 56 | 58 | 19 |
| 网卡 | **59** | **66** | **63** | **28** |
| 人类 | 69 | 68 | 70 |  |

表2。
BLEU-1分数。
以前我们只报告工作结果时可用。
SOTA代表当前最先进的。

需要向更好的指标。
在官方测试集的标签只能通过官方网站,我们的模型有一个27.2 BLEU-4。

#### 学习,数据传输质量大小和标签

因为我们有训练有素的许多模型和我们有几个测试集,我们想要研究是否可以转移到一个不同的数据集模型,多少不匹配的域将补偿如更高质量的标签或更多的培训数据。

最明显的案例学习和数据传输大小Flickr30k和Flickr8k之间。
这两个数据集也同样贴上他们由同一组。
实际上,当培训Flickr30k(大约4倍的训练数据),获得的结果是4蓝色点更好。
很明显,在这种情况下,我们看到收益通过添加更多的训练数据以来的整个过程是数据驱动的,容易过度拟合。
MSCOCO是更大的比Flickr30k训练数据(5倍),但由于收集过程完成不同,可能会有更多的不同- ence词汇和一个更大的不匹配。
事实上,所有的蓝色分数降低10分。
尽管如此,de -从普通仍然是合理的。

因为帕斯卡没有正式训练集和Flickr和MSCOCO独立收集,我们重新港口转移学习MSCOCO(表2)。做学习从Flickr30k转移产生了更糟糕的结果与BLEU-1 53 (cf 59)。

最后,尽管全局弱标记(即。
,洛杉矶-贝尔字幕并不是人类生成的描述),

5我们计算这些蓝色的作者的输出

[[14]](#_bookmark36)请提供OxfordNet系统。

的任务难度大得多,且签证官——cabulary。
然而,更多的数据用于训练- - -荷兰国际集团(ing)。
死伤MSCOCO模型在运行时,我们每,以至于降低表现从28到16。

#### 一代多样性的讨论

拥有训练有素的生成模型,给出了p (S I), ob - vious问题是模型生成小说帽,*,*以*及生成*的标题是否多样化和高质量。
表3显示了一些样品,重新把N-best列表从我们定向搜索译码器,代替最好的假说。
注意样品di -诗,从相同的图像可能显示不同的方面。
之间的协议在蓝色得分前15名生成的句子是58,这类似于人类。
这表明我们模型的多样性gen -害死。
粗体的句子在训练[集](#_bookmark16)不存在。如果我们把最好的候选人,这句话出现在训练集时间的80%。
这是并不值得大惊小怪,因为训练数据量非常小,所以相对比较容易为模型选择“exem plar”句子和使用它们来生成描述。
如果我们分析句子生成的前15名,大约一半的时候我们看到一个全新的描述,但仍然拥有相似的蓝色得分,表明他们有足够的质量,然而,他们提供一个健康的多样性。

*|*

|  |
| --- |
| 一个男人在公园里扔飞盘。  **一个男人手里拿着飞盘。**  **一个男人站在草地上的飞盘。** |
| 密切的三明治板。  密切的一盘食物和炸薯条。  一个白色板顶部设有一个三明治减少一半。 |
| 展示柜装满很多甜甜圈。  **一个充满很多的蛋糕展示柜。**  **一家面包店展示柜装满很多甜甜圈。** |

表3。
N-best MSCOCO测试集的例子。粗线表示一种新颖的句子没有出现在训练集。

#### 排名结果

虽然我们认为排名是国际化程度的一个不令人满意的方法,吃了从图像描述的一代,许多论文报告分数排名,使用测试标题设置为候选材料-日期排列给定一个测试图像。
这些指标的方法,效果最好(MNLM),具体实现了一个ranking-aware损失。
然而,网卡上都做得非常好排名的任务(排名描述给我,年龄,和排名图片给出的描述),如表4和表5中可以看到。
注意,对于图像注释任务,我们归一化分数类似[21]。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 方法 | 图像注释  地中海R@1 R@10  *r* | | | 图片搜索  R@1 R@10地中海 | | | *r* |
| 整理磁盘碎片[[13]](#_bookmark35) | 13 | 44 | 14 | 10 | 43 | 15 | |
| m-RNN  [[21]](#_bookmark43) | 15 | 49 | 11 | 12 | 42 | 15 | |
| MNLM  [[14]](#_bookmark36) | 18 | 55 | 8 | 13 | 52 | 10 | |
| 网卡 | **20** | **61** | **6** | **19** | **64** | **5** | |

表4。
Recall@k Flickr8k中间等级。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 方法 | 图像注释  地中海R@1 R@10  *r* | | | 图片搜索  R@1 R@10地中海 | | | *r* |
| 整理磁盘碎片[[13]](#_bookmark35) | 16 | 55 | 8 | 10 | 45 | 13 | |
| m-RNN  [[21]](#_bookmark43) | 18 | 51 | 10 | 13 | 42 | 16 | |
| MNLM  [[14]](#_bookmark36) | **23** | **63** | **5** | **17** | **57** | **8** | |
| 网卡 | 17 | 56 | 7 | **17** | **57** | **7** | |

表5所示。
Recall@k Flickr30k中间等级。

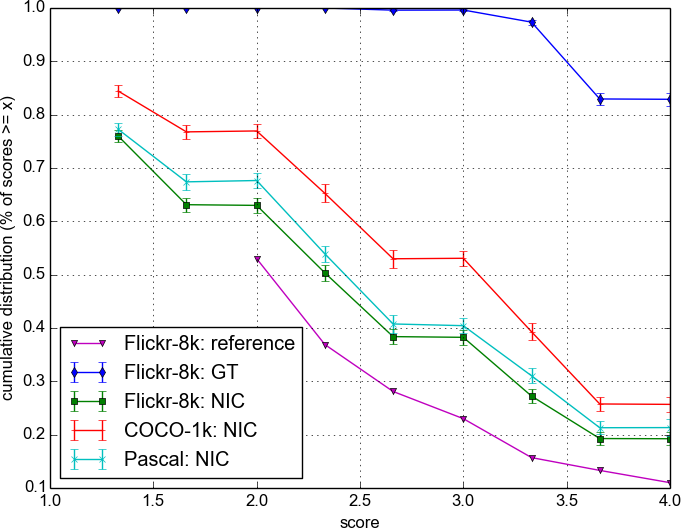


图4。
Flic*kr-8k:网卡:预测由N*IC Flickr8k测试集(平均分:2.37);
帕斯卡:网卡:(平均分:2.45);
COCO-1k:网卡:1000张照片的一个子集从MSCOC*O测试集描述由NIC*(平均分:2.72);
Flickr-8k:裁*判:这些结果从[11]F*lickr8k额定使用相同的协议,作为基准(平均分:2.08);
Flickr-8k: GT:我们从Flickr8k额定groundtruth标签——荷兰国际集团(ing)相同的协议。
这为我们提供了一个*“校准”的分数(平均分:3*.89)

#### 人的评价

图4显示了人类[的结](#_bookmark19)果评估提供的描述网卡,以及系统和groundtruth各种数据集的引用。
我们可以看到网卡比参考系统,但显然比groundtruth,如预期。
这表明蓝色并不是一个完美的指标,因为它没有很好的捕捉NIC和人类之间的差异描述由评级机构评估。
评价图像的例子可以看到在图5中。
有趣的是,例如第二图像第一列,该模型是如何能够注意到飞盘给它的大小。



图5。
的评估结果,分组由人类评级。

#### 分析嵌入

为了代表前面的词圣−1作为输入的解码LSTM生产圣,我们使用字嵌入向量[22],*它*的*优*点是份措辞尖锐的字典的大小(与一个简单- hot-encoding方法)。
此*外*,这些字嵌入——丁氏可以共同训练模型的其余部分。
这一点值得注[意学习表示已](#_bookmark44)抓获的一些语义语言的统计数据。
表4.3.7显示了几例的话,最近的句话说学嵌入空间中找到。

注意的一些关系学习模型将有助于视觉组件。
的确,有“马”,“小马”,“驴”接近彼此鼓励CNN提取horse-looking动物相关的特性。
我们假设,在极端的情况下,我们看到很少的例子一个类(例如,“独角兽”),其接近其他字嵌入的(例如,“马”)应该提供更多的信息,就会完全失去了与更传统的基于bag-of-words方法。

# 结论

我们提出了网卡,一个端到端的神经网络系统,可以自动查看图像和生成

|  |  |
| --- | --- |
| 词 | 邻居 |
| 车 | 车、出租车、suv、vehicule,吉普车 |
| 男孩 | 蹒跚学步,绅士,女儿,儿子 |
| 街 | 道路、街道、公路、高速公路 |
| 马 | 矮种马、驴、猪、山羊、骡子 |
| 电脑 | 电脑、个人电脑、crt、芯片、计算 |

表6所示。
最近的邻居的几个例子的话

一个合理的用简单的英语描述。
网卡是基于卷积神经网络编码图像到一个紧凑的表示,其次是复发性神经网络——工作,生成一个相应的句子。
模型训练句子给定图像的可能性最大化。
几个实验数据集显示网卡的健壮——洛克的定性结果(生成的句子是很合理)和定量评估,使用排名指标或蓝色,一个度量使用ma -中国翻译评估sen -对于生成的质量。
从这些实验,很明显,作为图像描述的可用数据集的大小增加,所以会像网卡的性能的方法。
此外,这将是有趣的,看看一个可以使用无监督数据,从图像和文本,以改善im -年龄描述方法。

# 确认

我们要感谢杰弗里•辛顿Ilya Sutskever, Quoc勒,文森特•Vanhoucke有用和杰夫院长说——cussions背后的思想,和写。

# 参考文献

1. 答:阿克尔和r . Gaizauskas。
   生成图像描述使用依赖关系模式。
   2010年ACL。
2. d . Bahdanau k .赵,y Bengio。
   马的神经——中国翻译共同学习和翻译一致。
   arXiv: 1409.0473, 2014。
3. k .赵b . van Merrienboer c . Gulcehre f . Bougares

h . Schwenk和y Bengio。
为统计机器学习短语表示使用RNN encoder-decoder transla,。
2014年EMNLP。

1. j·多纳休,y, o . Vinyals j·霍夫曼,n .张

大肠Tzeng, t·达雷尔。
脱咖啡因:深卷积活动- vation为通用的视觉识别特性。
2014年ICML。

1. d·艾略特和f·凯勒。
   图像描述使用视觉de -悬垂表示。
   2013年EMNLP。
2. 哈蒂,m . Hejrati m . a . Sadeghi p .年轻的时候,

c . Rashtchian j . Hockenmaier和d·福赛斯。
每一个图片——真正的讲了一个故事:从图像生成的句子。
2010年大会。

1. r·嘉宝和h。
   内格尔。
   知识表示生成量化的自然语言描述的车辆交通图像序列。
   在ICIP。
   IEEE 1996。
2. y锣,l . Wang m . Hodosh j . Hockenmaier和

美国Lazebnik。
改善image-sentence嵌入使用大弱带注释的图片集。
2014年大会。

1. 答:坟墓。
   生成序列与复发性神经网络——工作。
   arXiv: 1308.0850, 2013。
2. 美国。的Hochreiter和j。施密德胡贝尔表示
   短期记忆。

*神经计算,9 (8),1997。*

1. m . Hodosh p .年轻,j . Hockenmaier。
   帧图像描述作为一个任务:排名数据、模型和评价指标。
   睚珥,47岁,2013年。
2. 美国约飞和c Szegedy。
   批规范化:加速深层网络培训通过减少内部协变量的转变。
   在arXiv: 1502.03167, 2015。
3. a . Karpathy a Joulin和l .菲菲。
   深句子片段em -舍双向图像映射。
   少量的,2014年。
4. r . Kiros r . Salakhutdinov, r·s·泽梅尔。
   统一visual-semantic嵌入与多通道神经lan -规模型。
   在arXiv: 1411.2539, 2014。
5. r . Kiros和r . z . r . Salakhutdinov。
   多通道神经lan -规模型。
   在少量的深度学习研讨会,*2013。*
6. g . Kulkarni诉Premraj s Dhar崔y s . Li a·c·伯格和t·l·伯格。
   宝宝说:理解和生成简单的图像描述。
   2011年CVPR。
7. p .库兹涅佐娃诉Ordonez a·c·伯格,t·l·伯格和

崔y。
集体一代的自然形象的描述。
2012年ACL。

1. p .库兹涅佐娃诉Ordonez t·伯格和崔y。
   Treetalk:成分和压缩的树木图像descrip——一代长大成人。
   ACL, 2 (10), 2014。
2. s, g . Kulkarni t . l . Berg a·c·伯格和崔y。
   Com -使用网络级字格构成简单的图像描述。
   在会议上计算自然语言学习- - -荷兰国际集团(ing), 2011年。
3. T.-Y。
   林,m . Maire s Belongie j·海斯,p . Perona一起,d . Ra - manan p Dolla´r和c l . Zitnick。
   可可:微软Com -我的对象上下文。
   arXiv: 1405.03*12, 2014。*
4. j .毛泽东,w .徐,y, j . Wang和a . Yuille。
   除纯与递归神经网络的多通道图像。
   在arXiv: 1410.1090, 2014。
5. t . Mikolov拉和j . k . Chen院长。
   词表示向量空间的有效估计。
   2013年ICLR。
6. m·米切尔x汉,j·道奇,高洁的人,a . Goyal a·c·伯格k .山口t . l . Berg k . Stratos和h . d . III。
   蚊:从计算机视觉检测生成图像的描述。
   2012年EACL。
7. 诉Ordonez g . Kulkarni, t·l·伯格。
   Im2text:堤防- ing图像使用100万标题下的照片。
   在少量的酒,2011。
8. k . Papineni s Roukos、t·沃德和w·j·朱。
   蓝色:自动机器翻译评价的方法。
   2002年ACL。
9. c . Rashtchian p .年轻的时候,m . Hodosh和j . Hockenmaier。
   使用亚马逊的土耳其机器人收集图像注释。
   在NAACL停止创建研讨会演讲和语言数据与亚马逊的土耳其机器人,139 - 147页,*2010年。*
10. o . Russakovsky j .邓h·苏j . Krause s Satheesh

黄马,z, a . Karpathy a·斯拉·m·伯恩斯坦

a·c·伯格和l .菲菲。
ImageNet大规模视觉识别的挑战,2014年。

1. p . Sermanet d特征,x, m·马蒂厄勒存r·费格斯和y。
   Overfeat:综合识别、定位和检测使用卷积网络。
   arXiv预印本arXiv: 1312.6229, 2013。
2. r . Socher a . Karpathy诉勒,c·曼宁,a . y . Ng。
   接地成分语义寻找和堤防,荷兰国际集团(ing)图片与句子。
   2014年ACL。
3. Sutskever, o . Vinyals,问:诉勒。
   序列,序列与神经网络学习。
   在少量的酒,2014。
4. r . Vedantam和d . c . l . Zitnick帕里克说。
   苹果酒:一致同意的形象描述评价。
   在arXiv: 1411.5726, 2015。
5. 林b z姚明,x, l, m . w . Lee和研究所。
   朱。
   I2t:图像解析文本描述。
   IEEE学报,98(8),2010年。
6. p .年轻,a .赖m . Hodosh和j . Hockenmaier。
   im -年龄描述视觉外延:新的语义相似度满足——rics推理对事件的描述。
   2014年ACL。
7. w·扎,即Sutskever, o . Vinyals。
   复发性神经网络正规化。
   在arXiv: 1409.2329, 2014。