给定一副图像，描述它的内容

概要：

用机器来自动的描述一幅图像的内容，这项技术结合了机器视觉和自然语言处理。

本文要创建的模型使用了deep recurrent architecture。该框架结合了机器视觉和机器翻译的最新成果。

模型的训练方式：给定训练集，maximize the likelihood of the target description sentence

我们的模型，不论是在质量上还是数量上，相当精确。比如说，当前在Pascal数据集上最高的分数（BLEU-1分数）是25分，我们的模型能达到59分。一个人正常情况下的得分是69。这些是在Pascal数据集上的表现。在其他数据集上的表现如何呢？在flickr30k数据集上，我们的模型把BLEU-1分数从56提升到66。在SBU数据集上，我们的模型把BLEU-1分数从19提升到28。在最新的COCO数据集上，该模型可以达到27.7的BLEU-4的分数，是目前最高的分数。

1. 介绍

让机器拥有自动描述一幅图像的能力，是一个很艰巨的任务。可是一旦完成，它的用处很广泛。比如帮助盲人理解网站上的图像内容。

要描述一幅图像，模型不仅要识别图像中的物体，还要表达出这些物体之间关系，以及他们的特征和所参与的活动。更重要的是，还要用英语表达出来。也就是说，还需要一个语言模型。

先前的方法是把图像描述这个大问题分成多个子问题。正相反，本文是直接它作为一整个模型来训练，即输入图像I，输出一系列的单词来描述图像。其中每个单词都来自一个给定的单词字典。模型is trained to maximize the likelihood *p*(*S|I*) of producing a target sequence of words *S* = *{S*1*, S*2*, . . .}*

我们的灵感来自于最近在机器翻译上取得的成果。where the task is to transform a sentence S written in a source language, into its translation *T* in the target language, by maximizing *p*(*T|S*).这些年以来，机器翻译是被分割成一系列的子问题来完成。比如逐个翻译单词，连成句子，排序等。最近在机器翻译上取得的成果显示，机器翻译用一种简单的方法，卷积神经网络RNN，就能完成，而且效果很好。原理是，用一个编码器encoder RNN读取一个句子，将它转变成一个固定长度的向量。将这个向量再输入下一个RNN 解码器decoder，最后输出一个句子。

我们顺着他的思路，把编码器RNN替换成一个深度卷积神经网络（CNN）。卷积神经网络经过这几年的发展，已经可以用一个固定长度的向量表示输入图像。因此，我们可以把卷积神经网络CNN作为编码器。这个编码器CNN需要在图像分类任务中预先训练好，将CNN输出的固定长度的向量作为解码器RNN的输入，最后由RNN生成句子。我们把这个模型叫做Neural Image Caption, or NIC。

我们的成果如下。1，First, we present an end-to-end system for the problem. It is a neural net which is fully trainable using stochastic gradient descent。第二，我们的模型结合了目前表现最好的图像分类卷积神经子网络，和语言模型子网络。这些子网络可以用大的语料库来预先训练好。最终我们的网络表现得很好。

1. 相关工作

在机器视觉领域已经开始关注用自然语言来描述视觉信息这一问题，但主要是研究视频。这些模型很复杂，手动设计成分高，很脆弱。因此应用范围有限，例如交通场景和体育赛事。

最近越来越多人关注静态图像的自然语言描述。静态图像就是单幅图像，视频就是动态图像。最近物体识别的发展，也推动了自然语言生成系统的进步。有人检测出图像中的三种元素，然后用模板转换成文本。有人检测图像中的所有对象，用包含这些对象的词组和关系词组成一个句子。还有更多其他的基于语言分析的模型。这些模型手动设计成分高，生成文本时很脆弱，鲁棒性差。

目前，大量的工作用于对生成的图像描述进行排序。这些方法是基于一个idea，就是把在同一向量空间中表示图像和文本。要找到一幅图像，就在向量空间中图像描述附近找。目前这中模型还不能描述图像中的物体看不见的那一部分，甚至图像中包含的单个对象。而且，模型不对生成的描述进行打分。

最近某某发表的一篇文章是在一个神经网络中用图像和前面预测的单词来预测下一个单词。还有人是用同样的原理训练一个卷积神经网络RNN。这和我们的工作很相似，不同的是，我们用了一个更强大的RNN模型，向RNN模型输入图像信息。这些不同看起来不怎么重要，却使得我们的模型取得了明显更好的结果。最近，某某又提出新的方法，他把机器视觉模型和LSTM模型组合在一起，视觉模型编码图像，LSTM模型编码文本，使得结果处于同一嵌入空间。和本文的方法相比，他是用两个独立的路线（一个是处理图像，一个是处理文本）结合在一起得到一个嵌入层。

1. 模型

本文创建的模型结合了神经网络和概率论来为图像生成描述语句。最近在统计机器翻译上取得的进步表明，给定一个输入句子，通过最大化正确的翻译的可能性，来训练模型。