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。