北京航空航天大學

本科生《计算机视觉》
图像描述

王田

wangtian@buaa.edu.cn

202309

图片描述与关系识别

图片描述与关系识别是一类结合计算机视觉与自然语言处理相关领域的任务。

将首要介绍单词、句子在深度学习中模型中的表示，并引入Encoder-Decoder结构与注意力机制，最后对目前流行的图片描述与关系识别模型进行了介绍。

背 景 介 绍

如何理解一张图片?

我们现在已经学习了对图片、图片中目标乃至像素的分类。

背 景 介 绍

如何理解一张图片?

我们现在已经学习了对图片、图片中目标乃至像素的分类。

比较这两幅图片，从目标检测的角度，检测的结果都是一个人和一条狗。这能否说明这两幅图片的内容是一样的?

背 景 介 绍

如何理解一张图片?

我们现在已经学习了对图片、图片中目标乃至像素的分类。

比较这两幅图片，从目标检测的角度，检测的结果都是一个人和一条狗。这能否说明这两幅图片的内容是一样的?
我们需要更多、更深层次的语义信息来描述图片。

图片描述

一个人在**追逐**一条狗

一个人在**帮助**一条狗

图片描述

一个人在**追逐**一条狗

一个人在**帮助**一条狗

图片描述指的是对于一张输入的图片，输出一句话来描述图片的内容。

我们通过图片描述和关系识别的任务，捕获图片中不同对象之间的关系，获取更高层次的语义信息。


综述
图片描述与关系识别是一类结合计算机视觉与自然语言处理相关领域的任务，由图像生成相应的文本
介绍文本在深度学习中模型中的表示，并引入Encoder-Decoder结构与注意力机制
单词、句子在深度学习模型中的表示
One-Hot 表示
分布式词向量表示
综述
特点
基于于矩阵的分布表示
基于神经网络的分布表示
基于聚类的分布表示
Encoder-Decoder模型
Encoder-Decoder结构
综述
网络结构
编码
存储
解码
特点
Attention机制
结构
特点
对输入进行选择性的学习
对齐模型
突破
打破了传统编码器-解码器结构在编解码时都依赖于内部一个固定长度向量的限制
图片描述与关系识别模型
综述
图片描述是一种动态的目标检测，由全局信息生成图像摘要
基于Encoder-Decoder的模型
NIC
CNN编码图像
生成中间隐藏层变量
RNN生成目标语言文字
Attention机制
改进
多标签分类
输入局部赋予权重
高等级语义概念向量

概 述


图片描述与关系识别
单词、句子的表示
One-hot
词向量
Encoder-decoder
Encoder-decoder模型
Attention机制
图片描述与关系识别
NIC
Attention
改进

单词、句子的向量表示

神经网络的输入是一个向量或矩阵。

图像处理任务中，图像可以自然地被表示成一个(C, H, W)的矩阵。
如何将单词和句子表示为向量或矩阵？

One-hot表示

最简单的方法：One-hot

例如词库里一共只有三个单词：人、猫、狗

- 人：(1,0,0)
- 猫：(0,1,0)
- 狗：(0,0,1)

One-hot表示

最简单的方法：One-hot

例如词库里一共只有三个单词：人、猫、狗

- 人：(1,0,0)
- 猫：(0,1,0)
- 狗：(0,0,1)

思考：如果词库里有五个单词：人、猫、狗、兔、鸟，one-hot编码应该是怎样的?

One-hot表示

最简单的方法：One-hot

例如词库里一共只有三个单词：人、猫、狗

- 人：(1,0,0)
- 猫：(0,1,0)
- 狗：(0,0,1)

思考：如果词库里有五个单词：人、猫、狗、兔、鸟，one-hot编码应该是怎样的?
思考：如果词库里有3000个词，每个词向量维度是多少？10000个呢？100000个呢?

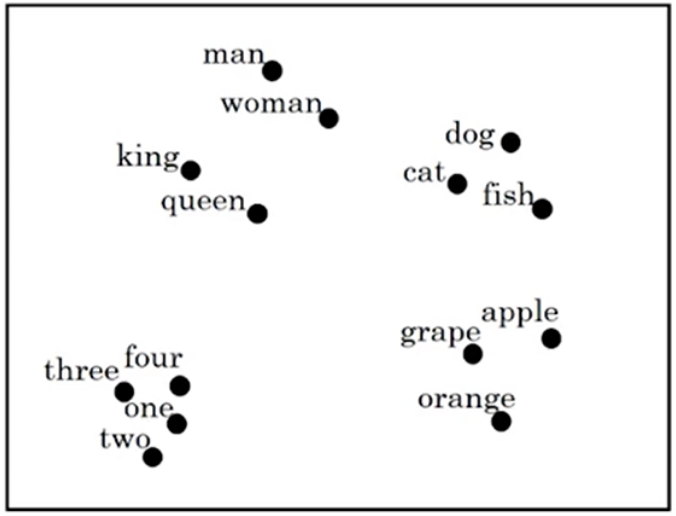
词嵌入表示

One hot编码的问题：

- 在词表很大的情况下向量太长
- 词向量不包含语义信息

词的分布式表示：通过训练，将每个词都映射到一个较短的词向量上来

- 所有的这些词向量就构成了向量空间
- 具有相近意义的词距离近



词嵌入表示

三种词向量建模方式：

- 基于矩阵的分布表示
  - 通常称为分布语义模型
  - 矩阵中的一行为对应词的表示，这种表示描述了该词的上下文的分布
  - 代表性方法：GloVe （Global Vector）
- 基于聚类的分布表示
  - 通过聚类手段构建词与其上下文之间的关系
  - 代表性方法：布朗聚类
- 基于神经网络的分布表示
  - 核心依然是上下文的表示以及上下文与目标词之间的关系的建模
  - 代表性方法：Word2Vec

Encoder-decoder模型

Encoder-Decoder是深度学习中常见的一个模型框架。

- 解决Sequence to sequence问题
  - 输入一个序列，输出一个序列

Encoder-decoder模型

Encoder-Decoder是深度学习中常见的一个模型框架。

- 解决Sequence to sequence问题
  - 输入一个序列，输出一个序列

每个box代表一个RNN单元，通常是LSTM或GRU

- 当前时刻的隐藏层的状态由上一时刻的隐藏层的状态和当前时刻的输入来决定

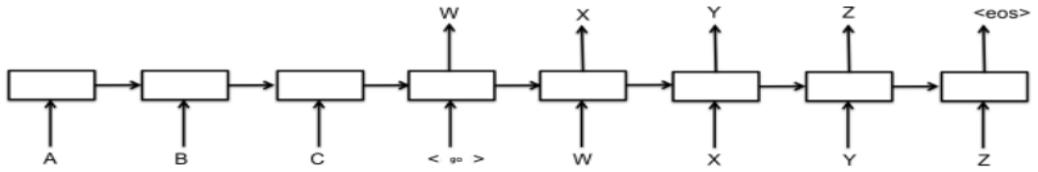
$$h_t = f(h_{t-1}, \mathrm{x}_t)$$

- 获得了各个时间段的隐藏层以后，再将隐藏层的信息汇总，生成最后的语义向量

$$\mathrm{C} = q(h_1, h_2, h_3 \ldots h_{T_x})$$

- 一种简单的方法是将最后的隐藏层作为语义向量C，即

$$\mathrm{C} = q(h_1, h_2, h_3 \ldots h_{T_x}) = h_{T_x}$$



Encoder-decoder模型

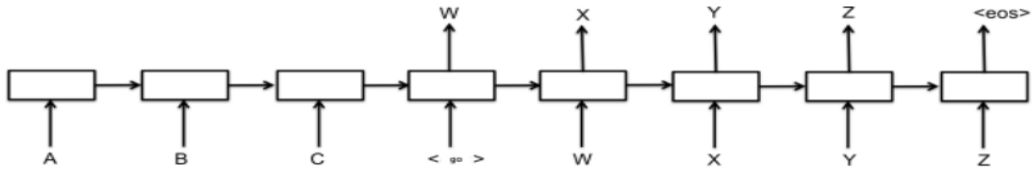
Encoder-Decoder是深度学习中常见的一个模型框架。

- 解决Sequence to sequence问题
  - 输入一个序列，输出一个序列

整个流程可以分为编码、存储、解码这三个过程：

- 编码

Encoder通过学习输入，将其编码为一个固定大小的状态向量S，接着将S传入Decoder。Decoder通过对状态向量S的学习来进行输出



Encoder-decoder模型

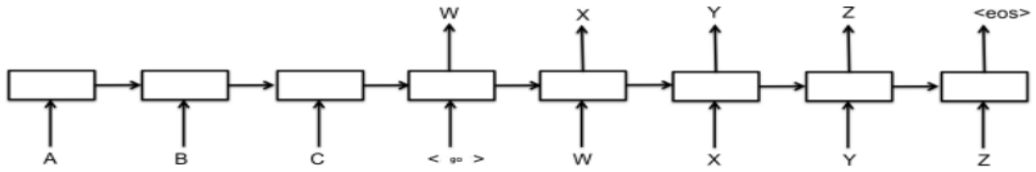
Encoder-Decoder是深度学习中常见的一个模型框架。

- 解决Sequence to sequence问题
  - 输入一个序列，输出一个序列

整个流程可以分为编码、存储、解码这三个过程：

- 编码
- 存储

通过RNN存储中间语义向量



Encoder-decoder模型

Encoder-Decoder是深度学习中常见的一个模型框架。

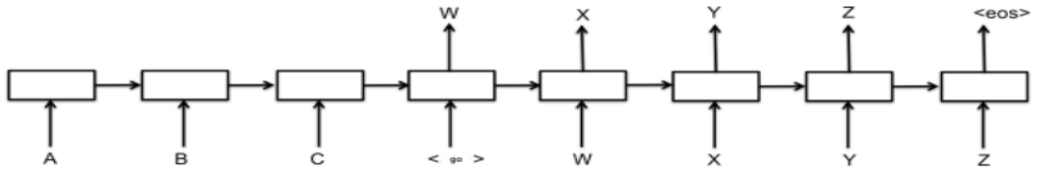
- 解决Sequence to sequence问题
  - 输入一个序列，输出一个序列

整个流程可以分为编码、存储、解码这三个过程：

- 编码
- 存储
- 解码

解码阶段可以看成编码的逆过程。这个阶段，我们要根据给定的语义向量C与之前已经生成的输出序列来预测下一个输出的单词，即

$$y_t = argmaxP(y_t) = \prod^{T} P(y_t|\{y_1, y_2 \dots y_{t-1}\}, C)\, h_{T_x}$$



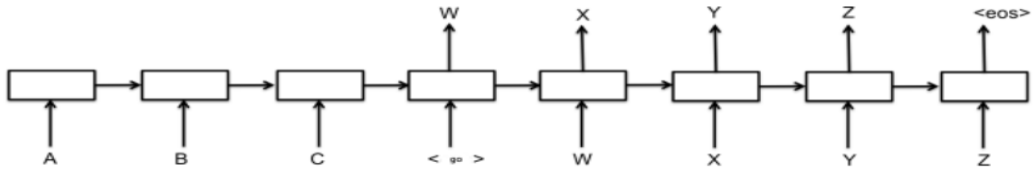
Encoder-decoder模型

Encoder-decoder的缺陷

- 编码和解码之间的唯一联系就是一个固定长度的语义向量C，编码器要将整个序列的信息压缩进一个固定长度的向量中去

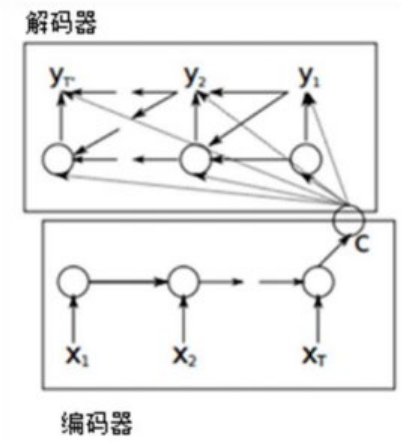
不足：

- 语义向量无法完全表示整个序列的信息
- 先输入的内容携带的信息会被后输入的信息稀释掉



Attention机制

- **Encoder-Decoder结构**： Encoder只将最后一个输出递给了Decoder。Decoder无法得到位置信息等输入的具体细节
- **Attention**：保留LSTM编码器对输入序列的中间输出结果，然后训练一个模型来对这些输入进行选择性的学习并且在模型输出时将输出序列与之进行关联



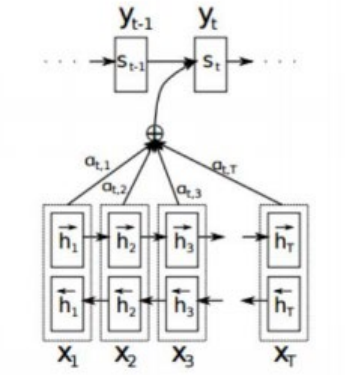


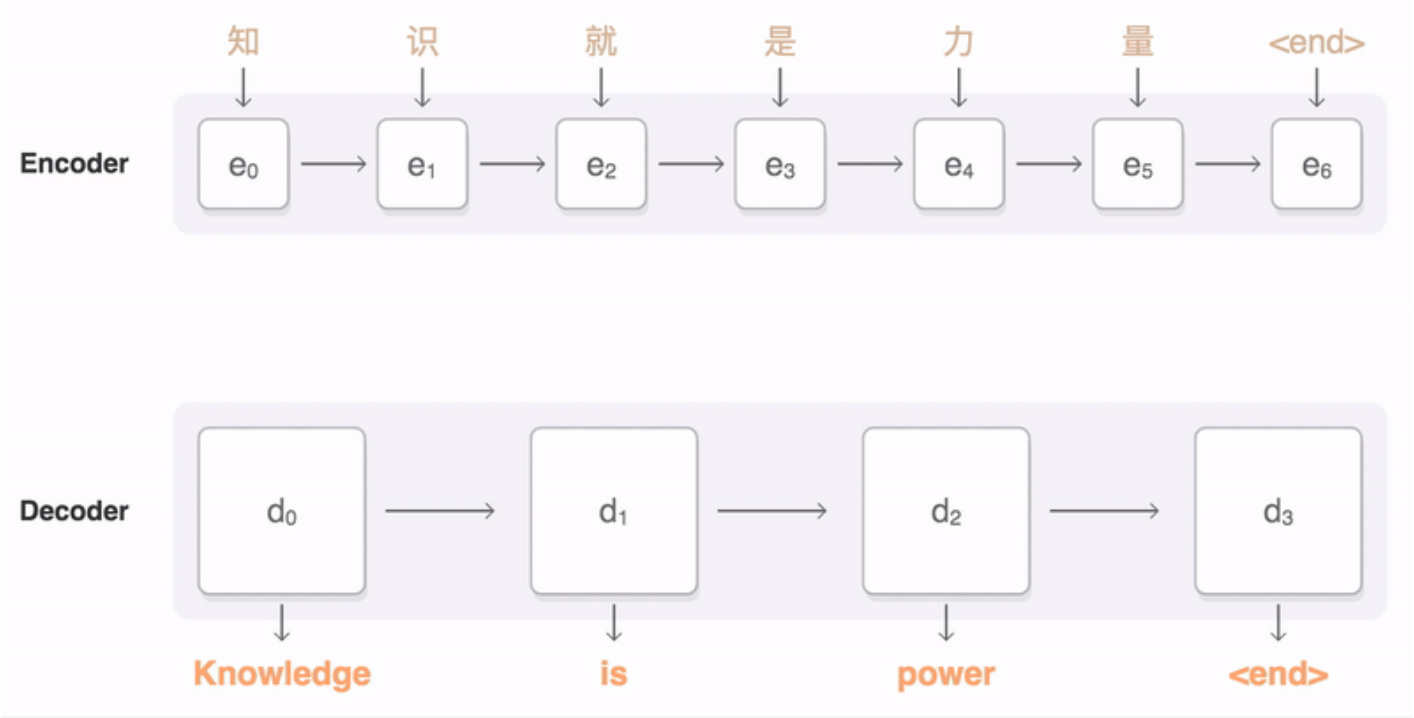
图 7-6 左图：原始 Encoder-Decoder 结构　　右图：基于 Attention 的 Encoder-Decoder 结构

Attention机制

- Attention：Captioning任务中输出不同的词时候关注图片的不同位置

Attention机制

- Attention：NMT任务中输出不同的词关注不同的输入词



Attention机制

定义每个时刻Decoder的输出

$$S_t = f(S_{t-1}, y_{t-1}, C_t)h_{T_x}$$

输出的条件概率

$$P(y_i|\{y_1, y_2 \dots y_{i-1}\}, X) = \mathrm{g}(y_{t-1}, s_t, \mathrm{C})h_{T_x}$$

此处条件概率与每个目标输出y_i对应的内容向量c_i有关。

- 在传统方式中，只有一个内容向量C
- Attention中，c_i与不同时刻的h_i有关

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \, h_{T_x}$$

α_{ij}定义为

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j=1}^{T_x} \exp(e_{ik})} h_{T_x}$$

$$e_{ij} = \alpha(s_{i-1}, h_j)h_{T_x}$$

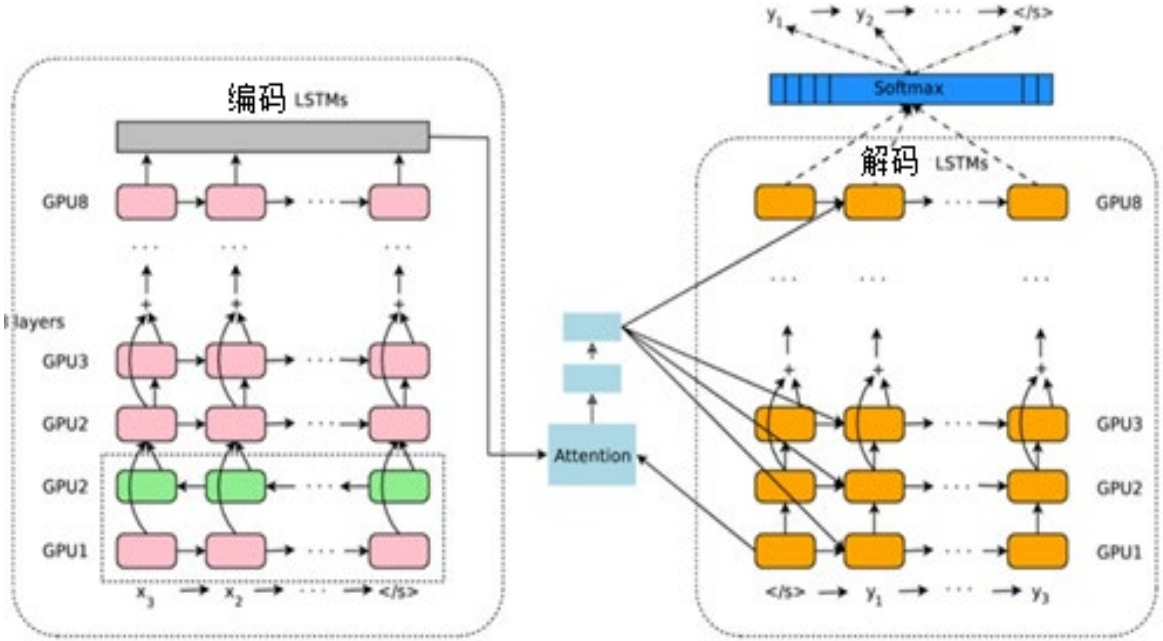
α_{ij}的值越高，表示第i个输出在第j个输入上分配的注意力越多，生成第i个输出的时候受到第j个输入的影响也就越大

Attention的本质抽象

$Source$中的构成元素是由一系列的<Key,$Value$>数据对构成。
Attention的计算步骤如下：

1. 给定$Target$中的某个元素$Query$
2. 计算$Query$和各个Key的相似性或者相关性，得到每个Key对应$Value$的权重系数
3. 对$Value$进行加权求和，即得到了最终的$Attention$数值。

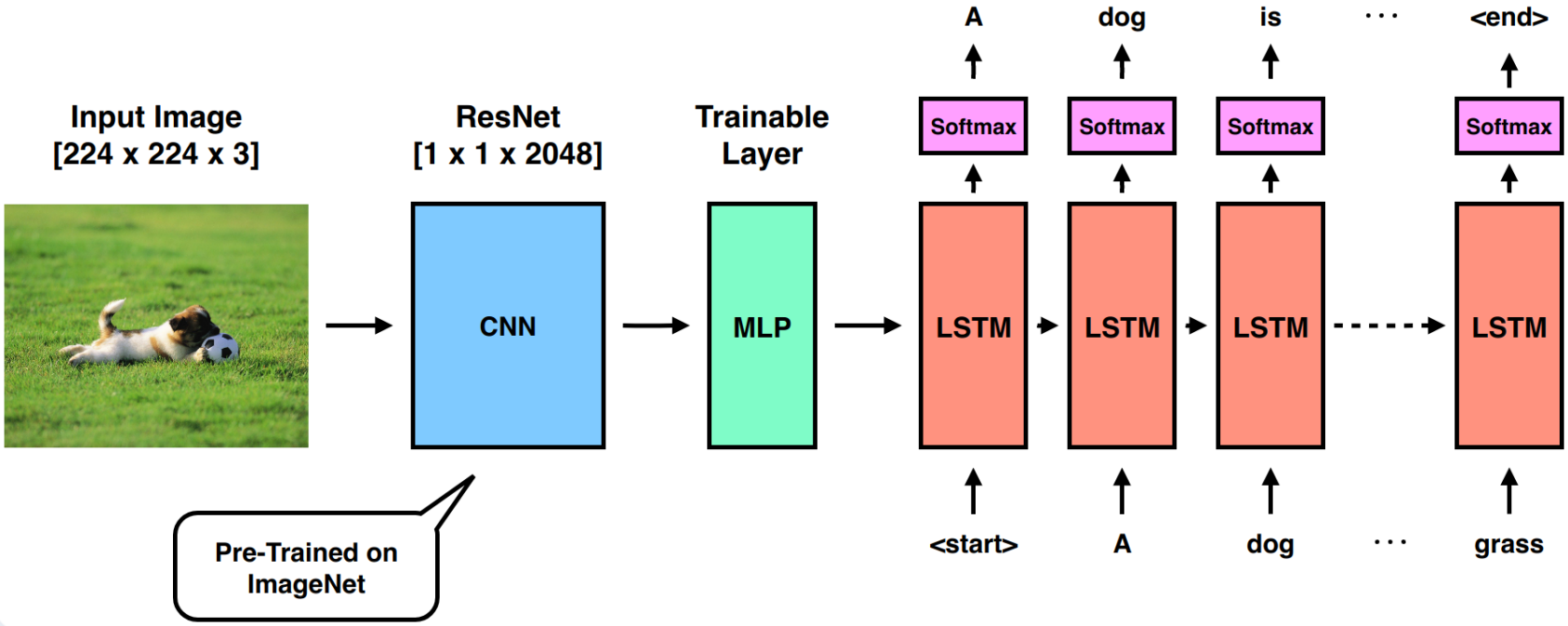
本质上$Attention$机制是对输入中元素的$Value$值进行加权求和，而$Query$和Key用来计算对应$Value$的权重系数。



基于Encoder-Decoder的图片描述与关系识别模型

图像描述与关系识别问题其本质是视觉到语言的问题。
基于Encoder-decoder的图片描述模型通常遵循以下框架：

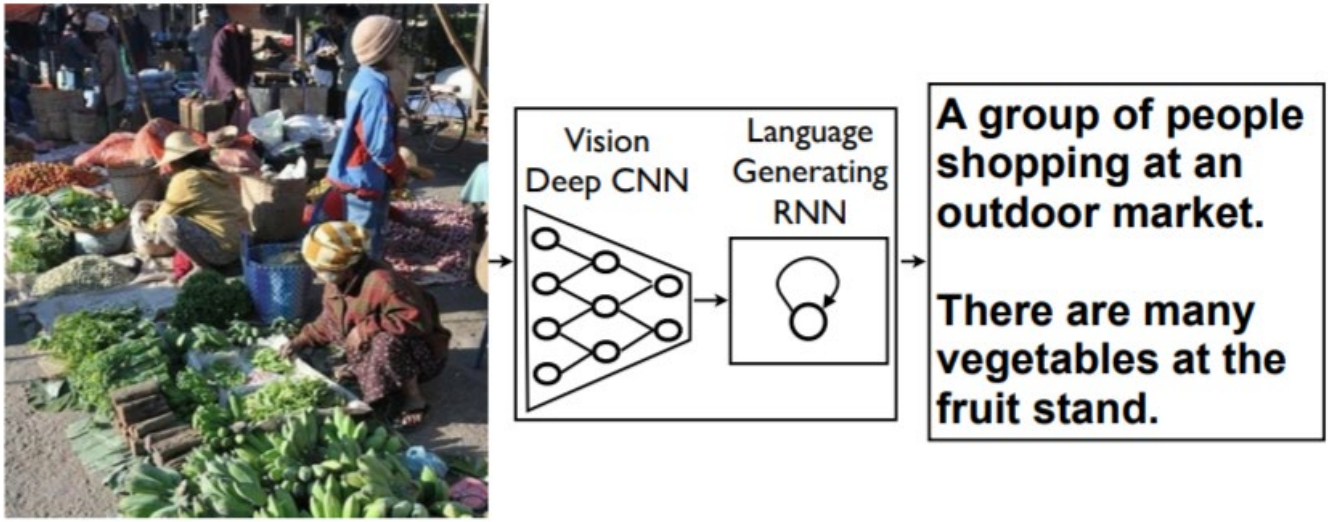
- Encoder：基于CNN提取图片特征
- Decoder：基于RNN生成文本



NIC

Google将机器翻译中编码源输入的RNN替换成CNN来编码图像，提出了NIC模型。
(**Show and Tell: A Neural Image Caption Generator**)

- **源输入**：图像
- **目标文字**：生成的描述
- **标签**：给定的图片描述向量



基于Attention的图片描述

在图像描述任务中，可以引入在Encoder-Decoder框架中的Attention机制，对输入信息的各个局部赋予权重。使用了Attention机制后，我们可以根据权重系数的大小，得知在生成每个词时模型关注到了图片的哪个区域。

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

高层语义特征

把这个高层语义理解为一个多标签分类问题。在Image Caption任务中，需要知道图片里面有哪些物体，由于一张图片中的物体数目会有很多，因此图片和物体标签就是一个一对多的关系，而不是通常的一对一的关系。
在一对多关系中，假设我们要找出c类物体，那么就分别使用c个Softmax层。

假设有N个训练样例，$y_i = [y_{i1}, y_{i2}, \dots, y_{ic}]$是第i个图像对应的标签向量，如果$y_{ij} = 1$，表示图像中有该标签，反之则没有。$p_i = [p_{i1}, p_{i2}, \dots, p_{ic}]$是对应的预测概率向量，则最终的损失函数为：

$$\mathrm{J} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c}\log\left(1+\exp\left(-y_{ij}p_{ij}\right)\right)h_{T_x}$$

在训练时，首先在所有描述中提取出现最频繁的c个单词作为总标签数，每个图像的训练数据直接从其描述单词中取得。训练完成后，针对每张图片提取高层的语义表达向量。

总结

- 句子和词的向量表示
  - One hot
  - Word embedding
- Encoder-decoder模型
  - Attention机制
- 基于Encoder-decoder模型的图片描述方法