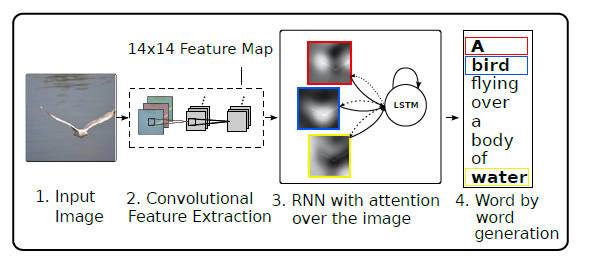
# 问题描述

基于神经网络产生图片描述：既要描述图片中的物品，还要讲述他们的联系

流程图：

重点是**注意力机制**，注意力机制保证了输入图片只有一部分放入神经网络。

分为hard和soft两种，soft由标准反向传播方法训练，hard由强化学习实现

好处：避免噪音干扰，专注于object

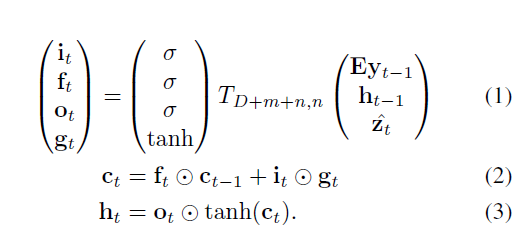
坏处：可能会丢失信息

# 编码器

输出文字，K是词汇量，C是句子长度。

注释向量，ai图片部分的一个D维的描述，共有L个

# 解码器

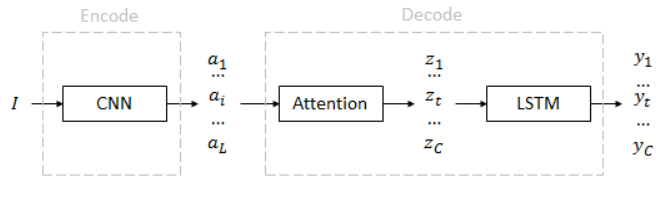
使用LSTM[[1]](#footnote-1)：每次基于内容向量产生一个单词。经典结构如下图所示：

I表示输入，f表示遗忘，o表示输出，h表示隐藏层，c表示存储，z是内容向量，包括信息和输入的位置。E是embedding矩阵：embedding是一种矩阵，其中每列是与词汇表中的item对应的向量。 要获取单个词汇item的稠密向量，就检索与该item对应的列。令m,n表示embedding，和是logistic sigmoid激活函数。定义用于从a生成z， a是诸注释向量。模型使用注意力模型对于每个都产生一个权重。

然后用上面的参数计算:

初始memory state用以下公式计算：

输出单词的可能性由以下公式界定：

总体输入输出流程如下图所示：

# 随机Hard Attention

表示在t时刻（产生第t个单词）时注意力的位置，是one-hot编码，表示如果第i个位置被使用，则置1。

定义待优化函数

# 确定性 soft Attention

直接取内容向量的期望：

对于第k个单词的预测，定义正则化的几何平均如下图所示：

# 训练过程

首先，利用VGGNet训练出各个ai作为注释向量。然后训练z ，使用Flicker8k作为训练集。测试时使用了Flicker8k, Flicker30k, 以及微软COCO

# 网络的选择

更新的网络(如VGG, GoogleNet)相对于AlexNet更好。

# 评价标准

BLEU, METEOR等

1. https://www.open-open.com/lib/view/open1440843534638.html [↑](#footnote-ref-1)