NJU_NLP_SummerCamp_2019_report_week6

> 2019.08.05-2019.08.11
> @author Eric ZHU

本周学习内容

1. 论文阅读：Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning
2. 论文源码阅读：Adaptive Attention
3. Django + Vue 数据可视化 初步

附录

1. "Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning" 阅读报告（见下页）

论文信息

类别	内容
论文名称	Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning
发表时间	CVPR 2017
作者	Jiasen Lu, Caiming Xiong, Devi Parikh, Richard Socher
关键词	Image Captioning, Attention Mechanism

核心信息

文章领域及方向

CV & NLP，Image Captioning 问题

文章概要

本文提出了Spatial Attention与Adaptive Attention这两种机制，分别解决了图片空间特征提取的问题与Caption中非视觉单词无需参考图片特征的问题。这两种机制的引入大幅提升了模型表现，在多个数据集上达到了State-of-the-art的水平。

引言

当前主流方案

当前主要使用的方案是基于视觉Attention的**Encoder-Decoder**模型，其中Attention机制会指出在每个词语生成过程中**语义对应的图像区域**。

当前方案存在的问题

并不是Caption中的**所有**单词都有对应的视觉特征。例如，在句子 "A white bird perched on top of a red stop sign." 中，单词 "A", "of" 就**不存在**对应的视觉特征。此外，语言本身的一些特征用法，如 "perched" 后紧跟的 "on" 和 "top", 和 "a red stop" 后紧跟的 "sign", 使得在生成部分单词时是**不需要参考**对应的视觉特征的。

事实上，**非视觉单词**的梯度会误导模型，并且在Caption过程中降低整个视觉特征的有效性。

本文贡献

本文提出了一种基于**自适应性（Adaptive）Attention**的Encoder-Decoder框架结构。它可以自动地决定**何时应当参考视觉特征**，而**何时应当仅依赖于语言模型**。当然，在模型参考视觉特征时，它也会决定应当关注图片的哪个区域。

首先，我们提出了一种创新的**spatial Attention**模型来提取图片中的空间特征。此外，在提出**自适应性Attention机制**的同时，我们引入了一种**新的LSTM结构**，即LSTM层在生成hidden state的同时，会生成一个额外的 visual sentinel（**视觉哨兵**）。视觉哨兵保存了Decoder的记忆。在生成新词时，模型会通过一个**sentinel gate**来决定从图片中获得多少信息，即生成该词是应当参考视觉特征亦或是依赖于语言模型。

总的来说，本文贡献如下：

1. 提出了一种基于**自适应性（Adaptive）Attention**的Encoder-Decoder框架结构。它可以自动地决定**何时应当参考视觉特征**，而**何时应当仅依赖于语言模型。**
2. 提出了创新的**Spatial Attention**机制来提取图片中的空间特征。
3. 提出了LSTM的一种新的拓展，即通过在hidden state外新增一个**visual sentinel**来解决单词的生成过程对图像特征的依赖程度不固定的问题。

具体方法

Encoder-Decoder结构

对于给定的图像与对应的Caption文本，Encoder-Decoder结构直接优化如下的目标函数：

$$\theta^* = arg\ max\ \theta \sum_{(I,y)} log\ p(y|I;\theta) \tag{8}$$

其中 θ 为模型的参数，I 为图像，$y = \{y_1, \ldots, y_n\}$ 为对应的Caption文本。

根据链式法则，联合概率分布的对数似然可以被分解成如下的有序条件概率：

$$log\ p(y) = \sum_{t=1}^{T} log\ p(y_t|y_1, \ldots, y_{t-1}, I) \tag{1}$$

为了方便起见，我们此处暂时不考虑对模型参数的依赖关系。

在Encoder-Decoder结构中，每个词语的条件概率可以被表示为：

$$log\ p(y_t|y_1, \ldots, y_{t-1}, I) = f(h_t, c_t) \tag{2}$$

其中 f 是输出 y_t 概率的一个非线性函数，c_t 是在时刻 t 从图像 I 中提取出的视觉特征向量，h_t 是时刻 t RNN的hidden-state.

我们在本文中采用LSTM作为RNN的实际模型。对于LSTM，h_t 可以被如下表示：

$$h_t = LSTM(x_t, h_{t-1}, m_{t-1}) \tag{3}$$

其中 x_t 为输入向量，m_{t-1} 是在 $t-1$ 时刻的记忆单元。

通常的，c_t 对模型的表现会产生较大的影响。而生成 c_t 的方法在不同的架构中被分为两类：

1. 在原始的Encoder-Decoder架构中，c_t 即**为Encoder的直接输出。**在生成Caption的不同阶段，c_t 总是保持恒定。
2. 在基于Attention的架构中，c_t **同时依赖于Encoder与Decoder。**在时刻 t, 基于Decoder的隐藏层状态，模型将会关注于图像的不同部分，并以此计算出 c_t.

Spatial Attention 模型

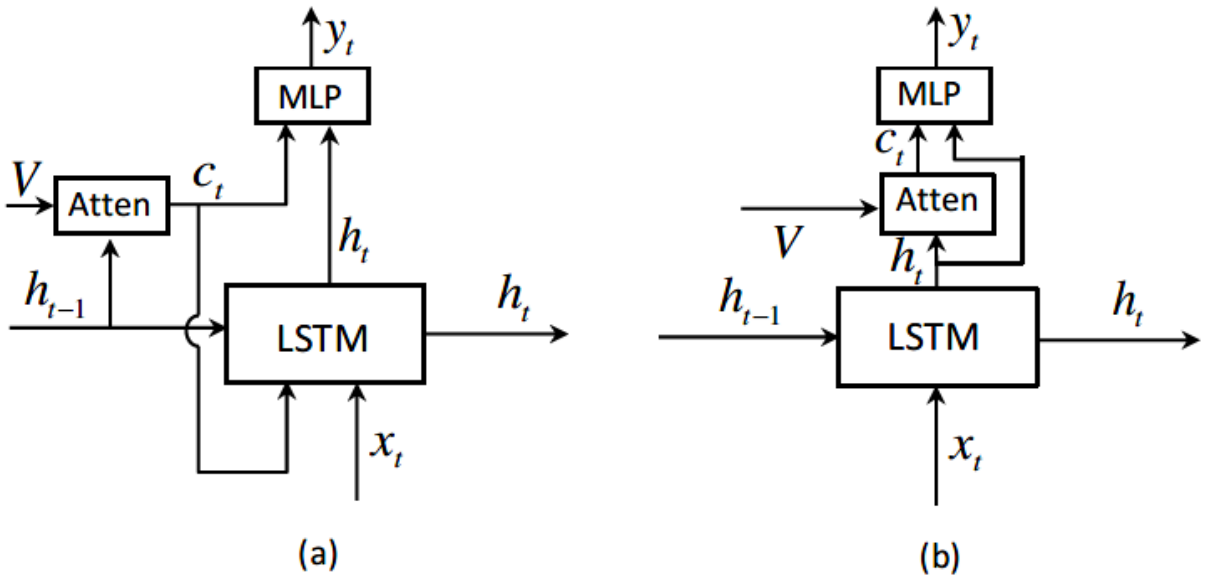
在该模型中，视觉特征向量 c_t 的计算被如下定义：

$$c_t = g(V, h_t) \tag{4}$$

其中，g 是Attention函数，$V = [v_1, \ldots, v_k], v_i \in R^d$ 是图像内 k 个区域的特征，h_t 是RNN在时刻 t 的隐藏层状态。

对于给定的空间图像特征 $V \in R^{d*k}$ 和LSTM的隐藏层状态 $h_t \in R^d$，我们把它们送进一个单层的全连接网络中，并用softmax函数获得对应的K个区域上的attention分布。

下图为Soft Attention模型的图示 (a), 与本文提出的Spatial Attention模型的图示 (b).



Adaptive Attention 模型

尽管基于Spatial Attention的Decoder在Image Captioning任务中表现出众，它无法决定何时应当依赖于视觉特征，而何时应当依赖于语言模型。因此，我们提出了visual sentinel的概念来进一步拓展上述模型。

什么是 visual sentinel ? Decoder的记忆储存了**视觉**和**语言模型**的**长短期**信息。我们的模型从Decoder的记忆中提取出**新的信息**，并在模型决定不参考图像的时候使用该信息来生成单词。这种新的信息就是visual sentinel.

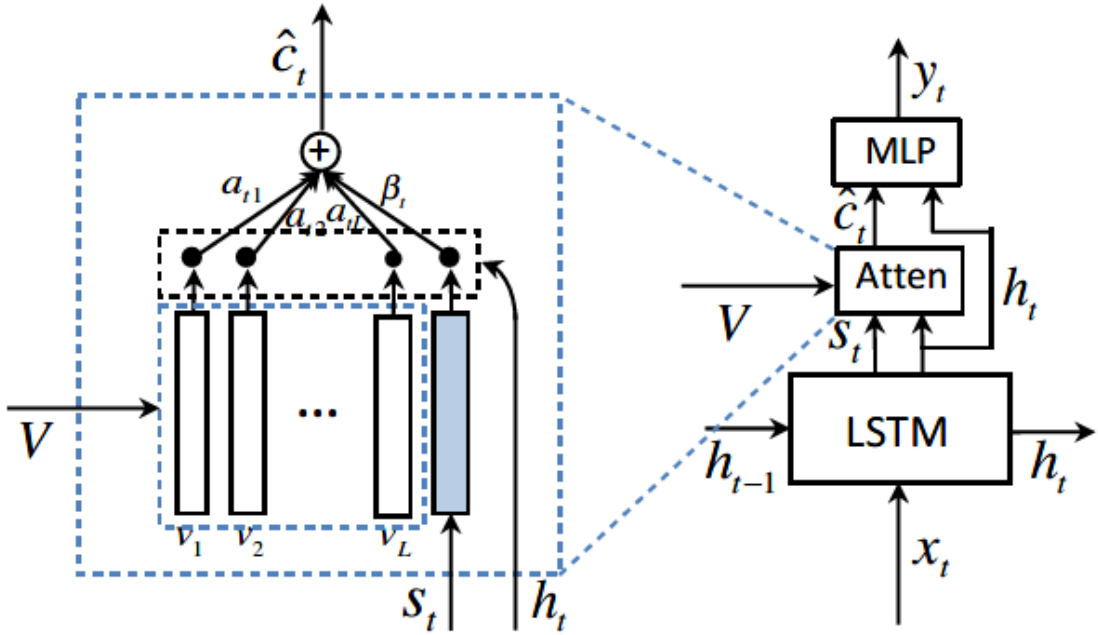
具体来说，我们对LSTM进行如下改进来生成 visual sentinel向量 s_t：

$$g_t = \sigma(W_x x_t + W_h h_{t-1}) \tag{5}$$

$$s_t = g_t \cdot tanh(m_t) \tag{6}$$

其中 W_x 和 W_h 是需要学习的权重向量，x_t 是在时刻 t 的LSTM输出，g_t 是对记忆单元 m_t 所英勇的sentinel gate，σ 是对数sigmoid激活函数。

下图展示了第 t 个单词 y_t 的生成过程。



在Adaptive Attention模型中，新的视觉特征向量 $\hat{c}_t$ 的计算方法如下：

$$\hat{c}_t = \beta_t s_t + (1 - \beta_t) c_t \tag{7}$$

即$\hat{c}_t$ 是visual sentinel s_t 与原始视觉特征向量 c_t 的组合，其中 β_t 是时刻 t 的新sentinel gate，其取值范围为 $[0, 1]$.

实现细节

Encoder-CNN

本模型采用ResNet的最后一个卷积层输出作为图像特征，维度为2048 x 7 x 7.

Decoder-RNN

本模型将word embedding向量 w_t 与全局的图像特征向量 v^g 进行组合来获得输入向量 $x_t = [w_t; v^g]$. 我们用单层的全连接网络来把visual sentinel向量 s_t 与LSTM输出 h_t 转化为d维的新向量。

训练细节

1. RNN使用深度为512的单层LSTM
2. 使用Adam优化器，语言模型的初始学习率为5e-4，CNN的初始学习率为1e-5
3. momentum与weight-decay被分别设为0.8和0.999
4. 在20个epochs后开始fine-tune CNN
5. Batch size为80，训练最大次数为50 epchs，在CIDEr的验证分数连续6轮无提升时early-stoppping.
6. 在采样caption时采用了beam size为3的beam search.

模型评价

模型表现

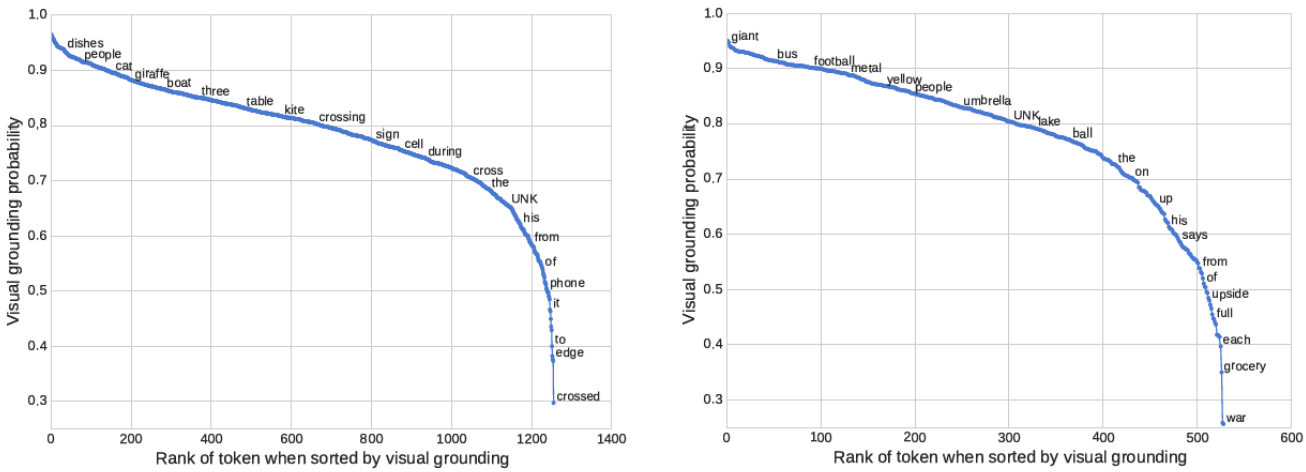
以下为模型在Flicker30k与MS-COCO数据集上的训练表现。

Method	Flickr30k						MS-COCO					
	B-1	B-2	B-3	B-4	METEOR	CIDEr	B-1	B-2	B-3	B-4	METEOR	CIDEr
DeepVS [11]	0.573	0.369	0.240	0.157	0.153	0.247	0.625	0.450	0.321	0.230	0.195	0.660
Hard-Attention [30]	0.669	0.439	0.296	0.199	0.185	-	0.718	0.504	0.357	0.250	0.230	-
ATT-FCN† [34]	0.647	0.460	0.324	0.230	0.189	-	0.709	0.537	0.402	0.304	0.243	-
ERD [32]	-	-	-	-	-	-	-	-	-	0.298	0.240	0.895
MSM† [33]	-	-	-	-	-	-	0.730	0.565	0.429	0.325	0.251	0.986
Ours-Spatial	0.644	0.462	0.327	0.231	0.202	0.493	0.734	0.566	0.418	0.304	0.257	1.029
Ours-Adaptive	**0.677**	**0.494**	**0.354**	**0.251**	**0.204**	**0.531**	**0.742**	**0.580**	**0.439**	**0.332**	**0.266**	**1.085**

可以看出，在只使用Spatial Attention的情况下，模型已经取得了不错的成绩；引入Adaptive Attention后，模型的表现有了进一步的提升。

评估细节

下图为Adaptive Attention中sentinel gate $1-\beta$ 的可视化结果。



可以看出，对于视觉词，模型给出的概率较大，即更倾向于关注图像特征 c_t，对于非视觉词的概率则比较小。同时，同一个词在不同的上下文中的概率也是不一样的。如"a"，在一开始的概率较高，因为**开始时没有任何的语义信息可以依赖**，并且需要确定句子的单复数。