计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(8):209−213 [doi: 10.15888/j.cnki.csa.006501]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于用户注意力与视觉注意力的社交图像描述①

褚晓亮, 朱连章, 吴春雷

(中国石油大学 (华东) 计算机与通信工程学院, 青岛 266000)
通讯作者: 吴春雷, E-mail: wuchunlei06@163.com

摘　要: 图像描述是机器学习和计算机视觉的重要研究领域, 但现有方法对于视觉特征和模型架构之间存在的语义信息关联性探索还存在不足. 本文提出了一种基于用户标签、视觉特征的注意力模型架构, 能够有效地结合社交图像特征和图像中用户标签生成更加准确的描述. 我们在 MSCOCO 数据集上进行了实验来验证算法性能, 实验结果表明本文提出的基于用户标签、视觉特征的注意力模型与传统方法相比具有明显的优越性.

关键词: 社交图像描述; 用户注意力; 视觉注意力; 用户标签; 长短时记忆网络

引用格式: 褚晓亮,朱连章,吴春雷.基于用户注意力与视觉注意力的社交图像描述.计算机系统应用,2018,27(8):209–213. http://www.c-s-a.org.cn/1003-3254/6501.html

Social Image Caption with Visual Attention and User Attention

CHU Xiao-Liang, ZHU Lian-Zhang, WU Chun-Lei

(College of Computer & Communication Engineering, China University of Petroleum, Qingdao 266000, China)

Abstract: Image captioning has attracted much attention in the field of machine learning and computer vision. It is not only an important practical application, but also a challenge for image understanding in the field of computer vision. Nevertheless, existing methods are simply rely on several different visual features and model architectures, the correlation between visual features and user tags has not been fully explored. This study proposes a multifaced attention model based on user tags and visual features. This model can automatically choose more significant image features or contain the user semantic information. The experiments are conducted on MSCOCO dataset, and the results show that the proposed algorithm outperforms the previous methods.

Key words: social image captioning; user attention; visual attention; user tags; LSTM

1　引言

随着深度学习的兴起, 图像描述[1]已成为计算机视觉和机器学习领域的热门研究, 它的具体任务是给定一张图像产生针对该幅图像的描述. 目前主流算法是在生成的每个词与图像区域间建立对应关系来生成描述. Facebook,Twitter 等社交网站的兴起让社交图片已成为人们展示自我的一种重要方式, 社交图片与用户的个人喜好、习惯等紧密相连, 与传统的图片相比, 它更加个性化, 用户可以通过标签对社交图片进行标记来表明自己的关注点与个人喜好. 然而现有的方法并没有直接针对于社交图像来产生描述. 因此, 本文提出一种基于图片视觉特征与用户标签的社交图像描述方法, 该方法利用图片的视觉特征与用户标签这两种模态进行分析, 然后利用注意力机制将图像特征和用户标签的语义信息相结合来生成更加准确的描述.

2　相关工作

2.1　基于注意力的图像描述模型

Vinyals 等人采用了 encoder-decoder 架构进行图

① 基金项目: 国家科技部创新方法工作专项 (2015IM010300)
Foundation item: Special Project for Innovative Work Method of Ministry of Science and Technology (2015IM010300)
收稿时间: 2018-01-02; 修改时间: 2018-02-01; 采用时间: 2018-02-06; csa 在线出版时间: 2018-07-28

像描述, 他们将 CNN 提取的图片特征作为 encoder 传入 LSTM 来解码生成图像描述[2]. 但是对于一幅图像, 人类所关注的并不是全部的内容, 即对于图像的每个像素点的关注度是不一样的. 为了让机器最大限度地模仿人类的学习机制, Toshev 等人提出了在图像上引入了注意力机制, 将上下文信息引入到 encoder-decoder 框架中. 在 encoder 阶段, 作者使用了保留图像空间信息的较低层的卷积层作为图像特征, 然后结合注意力机制将其用于 decoder 阶段, 该方法有效地提取了图像的视觉信息来生成更加准确的描述[3]. Xu 等人采用了三种不同的语义信息来指导描述的生成. 其中的指导分别为: 基于检索的指导, 语义嵌入指导、图像指导[4]. Zhou 等人考虑到该方法的指导采用了时间不变性, 忽略了不同时刻的指导的信息不同, 因此提出了将生成的词与图像特征结合的方法, 该方法能够根据当前生成的词来选取图像的部分特征来生成描述[5]. 腾讯人工智能实验室提出 SCA-CNN 新方法[6], 该方法首先肯定了视觉注意力机制对于图像描述的发展的重要意义, 并指出目前的注意力机制只是针对空间上的, 在图像卷积的过程中并没有进行注意力的操作. 基于这一问题, 他们提出一种新的注意力机制, 具体来讲这是一种将空间和多通道结合的注意力机制. 这种机制学习的是多层 3D-feature map 中的每一个 feature 与隐藏层之间的联系, 也就是在 CNN 中引入注意力机制, 而不是仅仅使用 CNN 部分的输出. 我们的方法同样基于注意力机制, 不同的是用户的标签在生成描述时应该被考虑进来.

2.2 基于属性的图像描述模型

只将图片特征作为 encoder-decoder 框架的输入有时候并不能反应图像的高级语义信息. Wu 提出利用多标签来取代图像特征作为 LSTM 的输入[7]. 该方法首先利用 VggNet 模型进行多标签的预训练, 然后通过 CNN 产生多标签的预测结果, 将预测结果经过 maxpooling 处理后, 输入到 LSTM 产生描述. Yao 等人探究了图像的标签对于描述效果的影响[8], 作者利用多实例学习的方法来产生图像的标签, 并且尝试了不同的组合形式. Lu 等人考虑到某些单词的生成并不依赖于图像特征而是依据当前的语言状态首次提出了‘哨兵’的概念, 让模型自动选择利用图像特征或者语言模型[9]. You 等人考虑到生成的单词有时往往不准确, 提出了在模型的输入输出阶段加入图像的标签作为引导[10]. 即将模型输出的单词与标签进行注意力机制的融合来产生更加准确的描述. 当前图像描述的模型中使用的语言模型都是逐个单词生成[11]. 但是从生物学的角度, 特别是人类, 在观察一幅图片的时候, 首先确定图像中存在哪些物体, 他们之间有哪些关系, 然后将他们之间的联系用自然语言清楚地描述出来. 因此 Wang 提出了一种由粗到细的方法[12], 将图片描述的任务分成两个部分, 一个主干句和各种物体的特征即标签, 同样在生成描述的时候也分为这两个部分进行. 然而这些方法都没有基于用户的标签来对社交图像进行引导, 本文我们提出了基于用户标签的注意力社交图像方法. 该方法首先将图像特征与用户的标签经过注意力机制的处理, 然后将处理后的特征作为 encoder 传入 LSTM 来生成描述.

3 社交图像描述模型

3.1 LSTM 网络

RNN 网络又称为循环神经网络, 它在原有的神经网络的基础上添加了反馈调节的功能, 因此可以做用于序列模型的生成, 该网络的更新不再像传统的神经网络一样只依赖于输入, 隐藏层的状态也是更新的一个重要依据. RNN 可以依据输入$(a_1,a_2,\cdots,a_n)$更新网络的隐藏状态$(h_1,h_2,\cdots,h_n)$, 其具体公式如下:

$$h_n = \psi(w_h a_{n-1} + p_h h_{n-1} + b_h) \tag{1}$$

其中, W, p, b是需要学习的参数, $\psi()$ 是激活函数. 但是 RNN 在训练较长的时间序列上信息容易丢失, 因此长短时记忆网络 LSTM 被引入来解决这一问题. LSTM 网络在 RNN 基础上引入了门的机制来改变 RNN 的细胞状态 (添加或修改信息)[13].

在已知输入序列$(a_1,a_2...a_n)$的情况下, LSTM 单元可以通过如下公式来计算隐藏状态h和细胞状态c:

$$i_m = \sigma(W_i a_m + P_i h_{m-1} + b_i) \tag{2}$$

$$f_m = \sigma(W_f a_m + P_f h_{m-1} + b_f) \tag{3}$$

$$o_m = \sigma(W_o a_m + P_o h_{m-1} + b_o) \tag{4}$$

$$g_m = \psi(W_g a_m + P_g h_{m-1} + b_g) \tag{5}$$

$$c_m = f_m \cdot c_{m-1} + i_m \cdot g_m \tag{6}$$

$$h_m = o_m \cdot c_{m-1} \tag{7}$$

其中, m代表第m时刻, $\sigma()$ 是 sigmoid 的激活函数, $\cdot$代

表两个向量的点乘.

3.2 整体框架

给定一张社交图像$s \in S$, 和一系列的用户标签$T_i(i=1,2,\cdots,m)$, 社交图像描述的任务就是产生m个基于用户标签的描述$c_i(i=1,2,\cdots,m)$. 更简单的理解是我们利用社交图像和用户的标签(s,T)来生成描述. 卷积神经网络 CNN 提取图片全局的视觉特征$V=\{V_1,V_2,V_3,\cdots,V_L\}$即将图片划分为 L 块区域, 每个区域都是一个 D 维的向量, 此外我们还获取了能够反映用户关注度的标签$T \in R^{N*D}$, $T=\{T_1,T_2,\cdots,T_n\}$, 其中n 代表标签的长度. 视觉特征V经过视觉注意力处理后得到特征V_{att}, 用户标签经过用户注意力处理后得到T_{att}, 然后将V_{att}和T_{att}一块传入 LSTM 生成t时刻的单词W_t, 其流程图如图 1所示. 不同于以往的图像描述模型, 我们的算法考虑到了用户的因素, 同时用户的标签还可以纠正因部分视觉特征导致的描述偏差, 其工作流程可以概括为以下公式:

$$V_{att}=\begin{cases} f_{vatt}(V), t=0 \\ f_{vatt}(V,W_{t-1}), t>0 \end{cases} \tag{8}$$

$$T_{att}=f_{tatt}(T) \tag{9}$$

$$T_{lstm}=W_{tag}T_{att}+b_{tag} \tag{10}$$

$$W_t=LSTM(V_{att},T_{lstm},h_{t-1}) \tag{11}$$

公式 (8)、(9) 分别用视觉注意力模型和用户注意力模型来对图像特征V和用户标签T进行注意力的权重分配得到处理后的特征V_{att}、T_{att}, 公式 (10) 对加权后的用户标签进行线性化处理使得与V_{att}处于同一维度. 公式 (11) 将V_{att}、T_{lstm}传入 LSTM生成当前时刻的单词W_t. $f_{vatt}()$、$f_{tatt}()$具体细节将在 3.3 节和 3.4 节介绍.

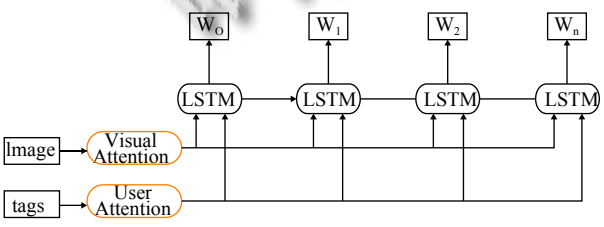

图 1 模型框架图

3.3 视觉注意力模型

使用卷积神经网络提取的图像特征V是一个LxD维的向量, 即将图像划分为L个区域, 每个区域用D维的向量表示:

$$V=\{V_1,V_2,V_3...V_L\}, V_i \in R^D$$

其中, R^D表示属于 D 维度; V_i表示第i个图像区域; 对于图像的每个区域, 注意力分配函数G_{att}[14]根据图像特征V和语义注意力模型在$t-1$时刻的生成的单词W_{t-1}产生一个权重α_i^t:

$$\alpha_i^t=G_{att}(V,W_{t-1}) \tag{12}$$

归一化处理:

$$\alpha_k^t=\frac{\exp(\alpha_i^t)}{\sum_{k=1}^{L}\exp(\alpha_k^t)} \tag{13}$$

其中, α_i^t表示视觉注意力模型中第i个图像区域在t时刻的权重; α_k^t表示视觉注意力模型中第k个图像区域在t时刻的权重.

经过视觉注意力模型处理以后的图像特征V_{att}:

$$V_{att}=\sum_{k=1}^{L}V_i\alpha_i^t \tag{14}$$

3.4 用户注意力模型

在社交图像中, 用户的标签可以反映用户的关注点, 对于用户标签中的每个单词, 注意力分配函数G_{att}根据用户的标签产生一个权重β_i^t:

$$\beta_i^t=G_{att}(T) \tag{15}$$

对β进行归一化处理:

$$\beta_k^t=\frac{\exp(\beta_i^t)}{\sum_{k=1}^{z}\exp(\beta_i^t)} \tag{16}$$

式中, β_k^t表示语义注意力模型中第 k 个词在t时刻的权重, z表示标签里的词的个数.

生成对当前标签的语义的状态T_{att}:

$$T_{att}=\sum_{k=1}^{z}T_k\beta_k^t \tag{17}$$

将用户的标签T_{att}进行维度转换为T_{lstm}, 与经过注意力模型处理后的视觉特征V_{att}一起传入 LSTM 生成当前时刻的单词:

$$W=LSTM(V_{att},T_{lstm},h_{t-1}) \tag{18}$$

用户注意力模型更加注重于生成的句子的语义结构, 因为对于句子的分析单凭视觉概念往往导致语义存在偏差, 因此将注意力模型产生的视觉特征V_{att}与标

签T_{lstm}一起传入 LSTM 中进行语义的完善. 对于标签 T 及生成的句子中的单词W, 本文采用维度为 D 的 one-hot 向量来表示, 用户标签用维度为ZxD的向量 T 来表示:

$$T = \{T_1, T_2, \cdots, T_n\},\ T_i \in R^D$$

其中, D 表示词典的大小, Z 表示标签的长度. 图像生成的句子用维度为CxD的向量W来表示:

$$W = \{w_1, w_2, \cdots, w_c\},\ w^i \in R^D$$

其中, D 表示词典的大小, C 表示产生的句子的长度.

4 实验

4.1 数据集和评估方法

算法在 MS COCO[15] 数据集上验证了其性能. COCO[12]分为训练集、验证集、测试集. 其中训练集包含 82 783 张图片, 验证集包含 40 504 张图片, 测试集包含 40 775 张图片. 每一张图片对应于 5 个人类标注的描述. 社交图像的用户的标签是本实验的关键部分, 考虑到现在没有对于社交图像描述的数据集, 我们针对图像描述中的每个句子随机提取一到两个关键字 (除去介词和名词) 即一幅图像对应 5 个标签和 5 个描述. 在社交图像中用户的标签有时候往往存在一定的噪声, 为了接近于更加真实的社交图像场景, 我们在标签的提取过程中随机添加了 7% 的噪声 (来自于其他图片的单词).

4.2 实验对比方法介绍

Soft[3] 利用空间注意力机制来处理卷积后的图像特征, 图像的每个区域被分配不同的权重来表示上下文的信息, 然后将这些信息输入到编码-解码框架中.

gLSTM[4] 采用三种不同的指导信息 (基于摘要指导、基于语义指导、基于图像指导) 来生成单词.

Sem-ATT[10] 对属性进行了注意力机制的处理并且与图像特征 (只在 t=0 时) 传入 LSTM 来生成单词.

Att+cnn+lstm[7] 利用预训练模型提取属性作为图像的高级语义信息, 然后将它们传入 CNN-RNN 框架中生成单词.

BIC+ATT[8] 探究了图像特征与属性对于图像描述的影响, 作者采用了 5 种不同的组合形式进了对比.

4.3 实验分析

我们采用了图像描述评测指标 Bleu[16]、Meteor[17]、Rouge-L[18]、CIDEr[19] 来评测我们的模型[20], 如表 1 所示, 我们的算法表现出较好的优越性, 这表明用户的标签能够纠正视觉偏差并且能够与视觉特征相互作用来生成更加准确的描述. 此外具有先验知识 (属性或标签) 的方法 (Sem-ATT[10]、Att+cnn+lstm[7]、BIC+ATT[8]) 明显要优于单纯的视觉描述方法 (Soft[3]、gLSTM[4]). 考虑到社交图片提取标签的方法与传统图片的不同以及为了证明我们算法的优越性不是依赖于我们提取的标签, 我们采用相同的标签实现了一些经典的算法 (Sem-ATT[10]、Att+cnn+lstm[7]、BIC+ATT[8]), 实验结果如表 2 所示, 考虑到不同的模型对于标签的抗噪性的影响不同, 实验结果会存在差异. 在表 2 中, 我们的算法依然保持着优越性, 该结果同时也表明我们算法的抗噪性要优于其他算法 (Sem-ATT[10]、Att+cnn+lstm[7]、BIC+ATT[8]). 综合表 1 与表 2, 我们可以得出如下结论: 在社交图像描述中用户的关注度 (标签) 可以纠正视觉特征的偏差, 于此同时视觉特征与用户的标签又能够相互影响并且可以有选择地参与 LSTM 中生成更加真实的描述.

表 1 模型与经典算法的比较

方法	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Soft[3]	0.707	0.492	0.344	0.243	0.239	0.960
gLSTM[4]	0.67	0.491	0.358	0.264	0.227	-
Sem-ATT[10]	0.709	0.537	0.402	0.304	0.243	-
Att+cnn+lstm[7]	0.74	0.56	0.42	0.31	0.26	0.94
BIC+ATT[8]	0.73	0.565	0.429	0.325	0.251	0.986
我们的算法	0.753	0.582	0.436	0.330	0.256	0.997

表 2 模型与带有标签的方法 (用户标签取代原有标签) 比较

方法	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Sem-ATT[10]	0.710	0.540	0.401	0.298	0.261	-
Att+cnn+lstm[7]	0.689	0.524	0.387	0.285	0.249	0.883
BIC+ATT[8]	0.696	0.526	0.386	0.285	0.248	0.871
我们的算法	0.753	0.582	0.436	0.330	0.256	0.997

5 结论与展望

本文提出了基于视觉注意力与用户注意力的社交图像描述方法, 并且在 MS COCO 数据集上表现优异. 该算法的核心思想是利用用户的关注度能够自适应地融合全局与局部的信息来生成更加准确而真实的描述. 相比于前人的工作, 我们在图像描述算法中考虑到了图像的视觉特征与用户的关注度 (用户标签) 之间的内在联系. 针对下一步的工作, 我们将利用多种不同的模型架构来探索用户的注意力机制对于社交图像描述任务的影响.

参考文献

1 Karpathy A, Li FF. Deep visual-semantic alignments for generating image descriptions. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 3128–3137.

2 Vinyals O, Toshev A, Bengio S, *et al*. Show and tell: A neural image caption generator. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 3156–3164.

3 Xu K, Ba JL, Kiros R, *et al*. Show, attend and tell: Neural image caption generation with visual attention. Proceedings of the 32nd International Conference on Machine Learning. Lille, France. 2015. 2048–2057.

4 Xu J, Gavves E, Fernando B, *et al*. Guiding long-short term memory for image caption generation. arXiv:1509.04942, 2015.

5 Zhou LW, Xu CL, Koch P, *et al*. Image caption generation with text-conditional semantic attention. arXiv:1606.04621, 2016.

6 Chen L, Zhang HW, Xiao J, *et al*. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. HI, USA. 2017. 6298–6306.

7 Wu Q, Shen CH, Liu LQ, *et al*. What value do explicit high level concepts have in vision to language problems? Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 203–212.

8 Yao T, Pan YW, Li YH, *et al*. Boosting image captioning with attributes. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy. 2016. 4904–4912.

9 Liu JS, Xiong CM, Parikh D, *et al*. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. HI, USA. 2017. 3242–3250.

10 You QZ, Jin HL, Wang ZW, *et al*. Image captioning with semantic attention. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 4651–4659.

11 Fang H, Gupta S, Iandola F, *et al*. From captions to visual concepts and back. arXiv: 1411. 4952, 2014.

12 Wang YF, Lin Z, Shen XH, *et al*. Skeleton key: Image captioning by skeleton-attribute decomposition. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. HI, USA. 2017. 7378–7387.

13 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

14 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv: 1409.0473, 2014.

15 Lin TY, Maire M, Belongie S, *et al*. Microsoft COCO: Common objects in context. In: Fleet D, Pajdla T, Schiele B, *et al*. eds. Computer Vision-ECCV 2014. Zurich: Springer, 2014: 740–755.

16 Papineni K, Roukos S, Ward T, *et al*. Bleu: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational. Philadelphia, PA, USA. 2002. 311–318.

17 Banerjee S, Lavie A. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. MI, USA. 2005. 65–72.

18 Lin CY. ROUGE: A package for automatic evaluation of summaries. Proceedings of the Workshop on Text Summarization Branches out. Barcelona, Spain. 2004. 10.

19 Vedantam R, Zitnick CL, Parikh D. CIDEr: Consensus-based image description evaluation. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 4566–4575.

20 Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 2013, 47(1): 853–899.