计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(7):106−112 [doi: 10.15888/j.cnki.csa.008593]

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

基于访问控制模块与原始信息注入的图像描述①

李　阳, 路　静, 郝宇钦, 韦学艳, 吴春雷

(中国石油大学 (华东) 计算机科学与技术学院, 青岛 266580)
通信作者: 吴春雷, E-mail: wuchunlei@upc.edu.cn

摘　要: 近年来在图像描述领域对于应用场景图生成描述的研究越来越广泛. 然而, 当前基于场景图的图像描述模型并未考虑到长短期记忆神经网络 (LSTM) 对于先前输入的细节信息的保留, 这可能会导致细节信息的丢失. 针对这个问题, 本文提出基于原始信息注入的图像描述网络, 该网络对基线模型中语言 LSTM 的输入变量做了改进, 目的是尽可能多地保留原始输入信息, 减少输入信息在计算过程中的损失. 另外, 本文还认为当前的场景图更新机制中存在结点更新程度过大的问题, 因此本文设计了一个访问控制模块更新已访问过的结点权重, 避免引起结点信息丢失的问题. 同时, 本文设计一个图更新系数 (GUF) 来指导图更新, 以确定更新程度的大小. 本文在官方数据集 MSCOCO 上进行了实验, 各种评估机制的实验结果表明, 基于访问控制模块与原始信息注入的图像描述模型与基线模型对比, 取得了更有竞争力的结果, 表现出明显的优越性.

关键词: 图像描述; 场景图; 访问控制; 长短期记忆网络; 原始信息注入

引用格式: 李阳,路静,郝宇钦,韦学艳,吴春雷.基于访问控制模块与原始信息注入的图像描述.计算机系统应用,2022,31(7):106–112. http://www.c-s-a.org.cn/1003-3254/8593.html

Image Captioning Based on Visiting Control Module and Original Information Injection

LI Yang, LU Jing, HAO Yu-Qin, WEI Xue-Yan, WU Chun-Lei

(College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)

Abstract: In recent years, the application of scene graphs in image captioning has been increasingly researched. However, the current image captioning models based on scene graphs cannot take into account the previous input retained in long short-term memory (LSTM) networks, which may lead to missed information. In this study, we firstly propose the image captioning network based on original information injection, which keeps the original input information as much as possible and reduces the missed information. Secondly, we consider that the degree of the current graph updating mechanism is too large, which may lead to the missing of node information. Thus, we propose a visit control module to update the weights of visited nodes, avoiding such missing. Finally, we design a graph update factor (GUF) to determine the update level. We conduct experiments on the official dataset: MSCOCO. The mechanism evaluation shows that our model has achieved more competitive results compared with the baseline model.

Key words: image captioning; scene graph; visiting control; LSTM network; original information injection

1　引言

计算机根据给定的图像自动生成简短的描述图像的句子, 这个任务被称为图像描述[1]. 在当前的计算机视觉领域中, 图像描述融合了机器学习、计算机视觉等多个不同领域, 是一项具有挑战性的任务. 主流的图像字幕模型大多数采用卷积神经网络 (CNN) 获取图像视觉特征, 并对显著区域和对象施加注意力, 通过递归模型生成描述. 随着对图像描述任务的研究逐渐增多,

① 基金项目: 山东省自然科学基金 (ZR2020MF136); 中央高校自主创新科研计划 (20CX05018A)
收稿时间: 2021-10-21; 修改时间: 2021-11-18; 采用时间: 2021-11-30; csa 在线出版时间: 2022-03-09

图像的场景图被用来增强图像描述模型, 从而利用场景图的结构语义, 如对象、关系和属性. 然而当前基于场景图的图像描述模型并未考虑到长短期记忆神经网络 (LSTM)[2] 对于先前输入信息的保留, 这可能会导致丢失细节信息. 原始输入信息中的细节能够指导句子的生成, 因为对于模型生成的句子, 其中每个单词的生成都要依赖于输入信息, 假如丢失了先前的原始输入信息, 则很难生成准确的句子. 此外, 当前的场景图更新机制中存在结点更新程度过大的问题, 导致生成句子的准确度降低.

为了在一定程度上解决丢失原始信息和图更新程度过大的问题, 本文提出了基于访问控制模块与原始信息注入的图像描述网络, 该网络改进了基线模型的图更新机制及语言 LSTM 中的输入信息, 目的是使图更新程度的大小更合理, 并减少原始信息的细节损失. 首先, 每张图像对应一个场景图信息, 网络对场景图进行编码, 对编码后的场景图特征施加注意力, 网络将得到的上下文特征传递给双层 LSTM 进行解码, 其中将原始信息注入到语言 LSTM 中, 最后通过访问控制模块将已访问过的结点权重降低, 既可以使网络关注未关注过的结点, 又尽可能保留结点的内容信息.

本文中, 创新点可以总结归纳为如下 3 点:

(1) 本文对基线模型中语言 LSTM 的输入变量做了改进, 将原始特征与经过注意力 LSTM 所得的特征拼接后得到新特征作为语言 LSTM 的输入, 以充分利用全局图像信息和嵌入信息来生成句子.

(2) 本文设计了一种新的访问控制模块 (VCM) 来实现图更新机制, 改进了现有的基于场景图的图更新方法, 它可以使网络关注重要信息的同时尽可能保留原始结点的信息, 我们设计了图更新系数 (GUF) 来指导图更新, 以确定更新程度的大小.

(3) 通过大量实验对提出的模型进行了分析与验证. MSCOCO 数据集上的实验结果表明了所提出的基于访问控制模块与原始信息注入的图像描述方法的有效性.

2 相关工作

2.1 图像描述

随着深度学习技术的发展, 在图像描述领域中, 对于神经网络的编码器-解码器框架的研究越来越多, 近年来已经取得了显著的改进. Vinyals 等人[3] 采用卷积神经网络 (CNN) 将图像视觉信息编码为固定长度向量, 递归神经网络 (RNN) 作为解码器, 依次生成单词. 为了更有效地关注图像中重要的区域, 注意力机制在图像描述模型中被广泛使用[4], 在生成描述过程中, 模型生成的所有的单词都和图像的某一特定区域一一对应. 由于传统的注意力机制存在强制将每个单词都对应到图像某一区域的问题, Lu 等人[5] 提出了一种自适应注意力机制, 在模型生成单词时判断是否需要关注图像信息及关注的程度. 此外, 为了减少顺序训练[6] 中的暴露偏差问题, Rennie 等人[7] 使用强化学习减少累计误差和优化不可微函数. 目前大部分图像描述任务都是基于编码器-解码器框架结构, 但解码器对于输入到 LSTM 中的信息经过多次计算后可能会丢失部分原始输入信息, 那么如何在 LSTM 中充分利用原始输入信息, 是一个值得思考的问题.

2.2 场景图

当前, 最流行的图像特征提取方法是使用 Faster R-CNN[8] 获取特征, 在自下而上的注意力模型[9] 中采用的就是此方法. 根据观察习惯, 人类视觉往往不是将图像分割为多个区域来观察的, 而是针对图像中较明显的物体来获取信息, 但仅关注物体信息会忽略多个物体之间的关联. 因此, 有研究进一步探索了一种在图像描述研究中更结构化的图像表示, 即场景图[10–12], 场景图的引入有效地促进了图像描述的发展, 它可以在图像描述任务中利用检测到的对象及其关系, 获得对图像更有条理的表述.

场景图显式地描述了图像中物体以及它们互相具有的关系. 场景图生成任务是建立在目标检测[13] 的基础上, 当前的一些研究介绍了场景图生成 (SGG)[14,15]. 大多数模型使用预先训练的 Faster R-CNN 或类似的体系结构来预测对象, 在此基础上加入一个额外的组件来预测对象之间的关系. Zellers 等人[16] 提出完成场景图生成任务可以利用很多先验知识. 在使用场景图生成描述的研究中, Chen 等人[17] 提出了一个图更新模块, 在每一步解码后更新当前的图, 改变图结点的权重以保证结点不被重复使用, 但是改变权重的方式容易丢失有效的信息, 那么如何在更新过程中保持删除信息和保留信息的平衡, 是一个值得思考的问题. 因此, 本文设计了一个访问控制模块更新已访问过的结点权重, 有效解决了结点内容丢失的问题.

3 基于访问控制模块与原始信息注入的图像描述网络

3.1 整体框架

给定一个输入图像 I, 本文采用文献 [17] 的方法来获得场景图特征 $G=(V, E)$. G 表示图像的场景图特征, 是一个有向图, V 表示图像 I 中检测到的对象对应的结点集, 包含物体结点和关系结点, E 表示对应于结点之间连接的边集, 表示两个结点之间有连接, 模型最终生成一组句子. 整个网络架构如图 1 所示.

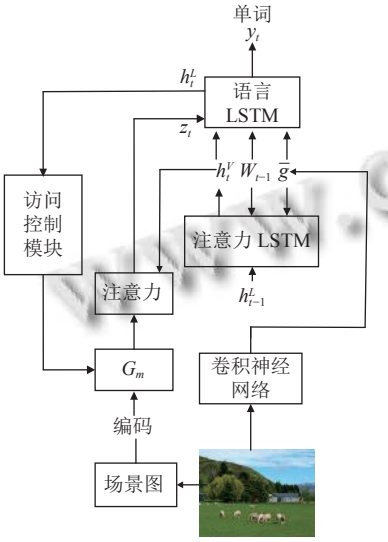


图 1 网络架构

具体来说, 本文的模型首先使用图卷积网络集成 G 中的信息得到 G_m. 本文采用文献 [17] 的图注意力方法将图内容注意力和图流向注意力融合得到集成上下文信息. 然后将得到的集成上下文信息馈入解码器进行字幕生成. 解码器包括两层 LSTM 结构, 分别用来处理注意力信息和单词信息. 并且本文对基线模型中语言 LSTM 的输入变量做了改进, 以充分利用全局图像信息和嵌入信息, 在第 3.2 节将详细介绍. 最后, 在生成单词 y_t 后, 本文通过访问控制模块将结点嵌入 X_t 的权重更新, 并根据本文提出的图更新系数作为调整结点权重的依据, 使下一时间步的结点 X_{t+1} 权重更为合理, 本文将在第 3.3 节详细介绍.

3.2 原始信息注入

本文解码器采用两层 LSTM 结构, 如图 2 所示. 其中, 注意力 LSTM 表示视觉注意 LSTM, 作用是整合视觉信息以及隐藏层信息, 并将自身计算得到的隐藏层信息作为模型注意力机制的一部分输入; 语言 LSTM 表示用来生成语言的 LSTM, 实现顺序地预测单词生成的功能.

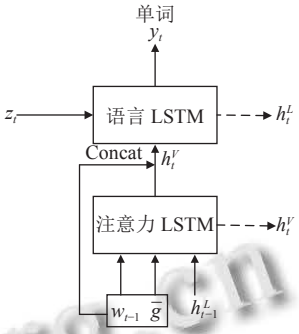


图 2 LSTM 模型图

本文认为全局图像编码嵌入和单词嵌入不仅可以指导注意力 LSTM 整合当前信息, 而且对于指导语言 LSTM 生成单词也是有价值的, 因此将全局图像编码嵌入、已生成的单词嵌入 w_{t-1} 注入到语言 LSTM 中, 充分利用视觉信息和单词嵌入信息指导句子的生成, 模型如图 2 所示.

注意力 LSTM 在每个时间步中会接收输入图像的特征编码嵌入 $\bar{g}$、词嵌入向量以及之前时间步的信息, 注意力 LSTM 将以上输入的信息进行整合得到 LSTM 的隐藏状态, 然后将输出的隐藏层信息作为注意力机制输入的一部分, 计算得到上下文特征. 最后, 计算得出的上下文信息和注意力 LSTM 的隐藏状态一起作为模型语言 LSTM 的输入. 另外, 本文模型为了充分利用原始信息, 将全局图像编码嵌入 $\bar{g}$、已生成的单词嵌入 w_{t-1} 与经过注意力 LSTM 所得的特征拼接后得到新特征作为语言 LSTM 的输入, 得到语言 LSTM 的输出. 最后, 在 t 时刻要生成的单词 y_t 由模型利用语言 LSTM 的隐藏状态预测得到, 具体公式如式 (1)–式 (3):

$$h_t^V = LSTM^V([\bar{g}; w_{t-1}; h_{t-1}^L], h_{t-1}^V) \tag{1}$$

$$Z_t = Attn(W_u I, h_t^V) \tag{2}$$

$$h_t^L = LSTM^L([Z_t; w_{t-1}; h_t^V; \bar{g}], h_{t-1}^L) \tag{3}$$

其中, h_{t-1}^L是语言 LSTM 前一时刻的输出, h_{t-1}^V是注意力 LSTM 前一时刻的输出, $Attn$ 为注意力操作, 上下文向量 z_t 经过 $Attn$ 操作后得到. w_{t-1} 是已生成单词的嵌入, $\bar{g}$是全局编码嵌入, W_u 是参数. 在时间步长 t 处单词分布的概率如下:

$$P(y_t \mid y_{1:t-1}) = Softmax\left(W_p h_t^L + b_p\right) \tag{4}$$

其中, W_p 是学习权重, b_p 是偏差. 句子概率分布计算公式如下:

$$P(y_{1:T}) = \prod_{t=1}^{T} p(y_t|y_{1:t-1}) \tag{5}$$

3.3 访问控制模块

为了充分表达场景图中的信息, 必须在不丢失且不重复的情况下表达场景图 G 中的所有结点, 本文结合了文献 [17] 的图更新机制, 在每一时间步生成单词 y_t 后, 将 t 时刻的结点嵌入 X_t 重新赋予权重, 更新为下一时刻使用的 X_{t+1}, 即更新结点的访问状态, 如图 3 所示.

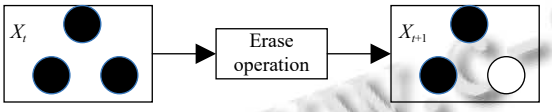


图 3 图更新机制

本文在此基础上改进了结点更新过程, 提出了访问控制模块, 如图 4 所示.

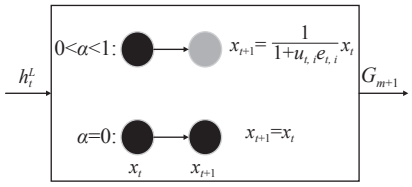


图 4 访问控制模块结点更新

注意力分数 α_t 表示每个结点的访问强度, 当 t 时刻某一个结点注意力高时, 代表已经使用过当前结点, 为了不重复使用, 当前结点的权重应该被更新为较低的权重, 注意力分数越高的结点下一时刻权重被降低的幅度应越大.

在生成一些特殊单词, 如“a”和“this”时, 尽管访问了图结点, 但这些单词是非可视的, 此时不更新结点. 因此, 本文采用了文献 [5] 的自适应更新机制, 用来修改注意力强度, 如式 (6) 所示:

$$u_t = Sigmoid(f_{vs}(h_t^L;\theta_{vs}))\alpha_t \tag{6}$$

其中, f_{vs} 是一个全连接网络, θ_{vs} 是参数, 该网络输出一个标量来表示当前注意的结点是否由已生成的单词表达的.

本文认为现有的图更新机制在更新结点的过程中, 结点权重有可能会直接被设置为 0 即完全被删除, 虽然 t 时刻对注意力分数高的结点 $X_{\max}$ 在 t+1 时刻关注程度应降低, 但 $X_{\max}$ 中的信息仍是有价值的, 如果直接删除会导致结点保存的信息不能对后续生成单词起到任何指导作用. 本文设计了图更新系数 GUF 来指导图更新, 并不会完全删除结点, 仅使结点权重降低, 计算公式如式 (7) 和式 (8):

$$e_{t,i} = Sigmoid(f_{ers}([h_t^l;x_{t,i}];\theta_{ers})) \tag{7}$$

$$GUF = \frac{1}{1+u_{t,i}e_{t,i}} \tag{8}$$

其中, $e_{t,i}$ 代表 t 时刻对第 i 个结点的注意力强度, 取值在 0–1 之间, 如果 $e_{t,i}$ 取值为 0, 代表结点在 t 时刻未被使用, 因此不应被更新, 如果 $e_{t,i}$ 取值为 1, 代表结点需要被更新的程度最大. $u_{t,i}$ 是视觉哨门, 控制更新的程度, 在 0–1 之间, $u_{t,i}$ 的值越高代表更新的程度越大. 根据对变量取值的分析, GUF 的取值在 0.5–1 之间.

使用 GUF 来指导图更新, GUF 的取值决定了结点下一时刻被访问的程度, 从而实现访问控制. 通过式 (9) 来更新图结点:

$$x_{t+1,i} = GUF \cdot x_{t,i} \tag{9}$$

根据本文对图更新系数 GUF 的分析, GUF 的取值在 0.5–1 之间, 即使更新程度最大, X_{t+1} 也将更新为 $0.5X_t$, 并不会被设为 0, 即并不会被完全删除. 因此本文模型更新的幅度比完全删除更小, 结点保存的信息仍能在一定程度上起到指导句子生成的作用.

通过这种方式, 我们将结点嵌入 X_t 更新为下一个解码步骤使用的 X_{t+1}.

3.4 目标函数

本文在训练模型时使用的损失为标准的交叉熵损失. 在训练过程中, 对于给定标签序列 $y_{1:T}$、场景图 G 和图像 I 的描述模型, 采用最小化交叉熵损失:

$$L = -\log \sum_{t=1}^{T} p(y_t|y_{1:t-1},G,I) \tag{10}$$

经过训练, 本文的模型可以通过给定的图像生成图像描述.

4 实验

4.1 数据集

本文实验使用的数据集为 MSCOCO[18], 它在图像描述领域是被广泛应用的官方数据集. 数据集中的图

像数量超过 12 万张, 其中每一张图像都有大约 5 个注释. 本文采用了 MSCOCO 数据集的图像及注释, 并采用“Karpathy”划分设置. 当前的图像描述评测标准分别有 BLEU1–BLEU4[19] , ROUGE[20], METEOR[21] , CIDEr[22], 本文使用以上评测标准来评估模型的性能, 通过比较模型生成的句子描述和参考句子的相似程度来评估生成的图像文本描述语句的得分.

在上述评测标准中, BLEU 是一种来源于机器翻译中计算精度的评估方法, 是用于计算模型所生成的句子和参考句子差异的方法, 重点考虑了生成句子中单词的准确性, 计算的结果在 0.0–1.0 之间, 结果越接近 1 代表句子的匹配程度越高. BLEU 方法的缺点是极少关注召回率. ROUGE 也是计算精度常用的方法之一, 基于查全率的相似度来计算模型生成描述的准确率, 与上文的 BLEU 具有相似的计算方法. METEOR 在机器翻译评估中也是常用的方法之一, 计算时对齐模型生成的描述与图像的正确描述, 这种自动评估标准对生成句子的准确率和召回率都进行计算. CIDEr 引入了“共识”的概念, 是用于衡量图像描述的一致性的标准, 它将句子表示成向量, 根据余弦相似度的标准来判断, 该评价方法对生成描述句子的语义考虑较多.

4.2 实验结果

4.2.1 模型消融测试

为了分析模型在引入访问控制模块和原始信息注入对于图像描述生成的作用, 本文进行了消融实验测试, 测试了模型在 3 种方法作用下的模型性能: ① 仅引入访问控制模块; ② 仅引入原始信息注入; ③ 同时采用原始信息注入+访问控制模块, 即本文模型所采用的实验情况. 测试结果如表 1 所示.

表 1 在数据集 MSCOCO 上的模型消融测试结果 (%)

模型	BLEU4	METEOR	ROUGE	CIDEr
访问控制模块	23.0	24.6	50.3	204.9
原始信息注入	23.2	24.9	50.6	207.8
访问控制模块+原始信息注入	**23.4**	**25.0**	**50.8**	**209.8**

从表 1 可以看出, 在上述 3 种情况中, 采用“访问控制模块+原始信息注入”对于图像描述具有最好的性能, 在生成语句的准确度、流畅度上的表现都得到了最高的指标值. 其主要原因是原始信息注入充分利用了原始输入信息, 另外, 本文设计的访问控制模块可以使更新的程度大小更合理, 从而生成与图像内容更相符的描述.

4.2.2 与其他模型的比较

图 5 是本文提出的方法训练的模型与基线模型 (ASG2Caption 模型)[17] 在官方数据集 MSCOCO 上的结果对比, 可以清楚地看出, 本文模型生成的句子与图像实际内容更加契合, 语言更加准确.

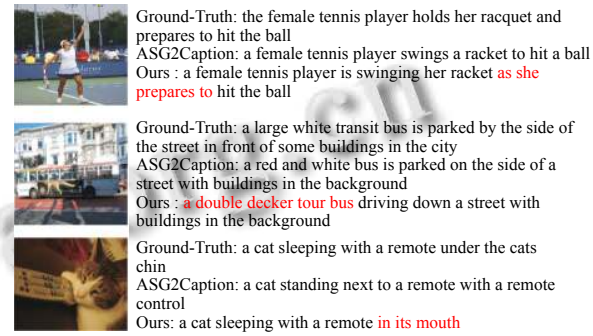


图 5 实验结果对比

如表 2 所示, 对于本文提出的模型, 本文在官方数据集 MSCOCO 上测试句子得分, 以评估模型的有效性. 表 2 中的数据表明, 与基线模型 (ASG2Caption 模型)[17] 和其他方法相比, 本文训练的模型具有更高的评分. 本文模型在 CIDEr 评分中提升最为明显, 约提高了 5.6 个百分点. 本文训练的模型通过 *GUF* 来指导图更新, 可以使更新的程度大小更合理, 从而使用更丰富的细节信息生成更高质量的图像描述.

表 2 在数据集 MSCOCO 上的实验结果对比 (%)

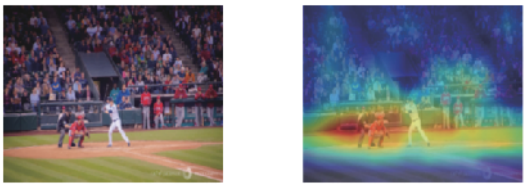
方法	BLEU4	METEOR	ROUGE	CIDEr
Mind’s Eye[23]	18.8	19.6	—	—
LRCN-AlexNet[24]	21	—	—	—
Multimodal RNN[25]	23	19.5	—	66
ASG2Caption[17]	23	24.5	50.1	204.2
基于访问控制模块与原始信息注入的图像描述网络	**23.4**	**25.0**	**50.8**	**209.8**

4.3 实验分析

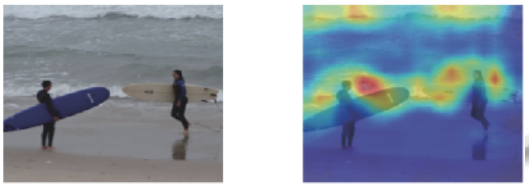
在 MSCOCO 数据集上, 本文模型的注意力权重可视化图如图 6 所示. 3 组图的左边为在 MSCOCO 中选取的图像, 图像右边是句中对于划线单词将注意力权重表示在图中的结果. 图中的 3 张可视化图对应了 3 张不同的图像, 图像下方的句子为本文提出的基于访问控制与原始信息注入模型生成的图像描述.

在图 6(a) 中, “a baseball player” 在整体图像中重要程度较大, 可以看出, 本文模型能够很好地关注棒球

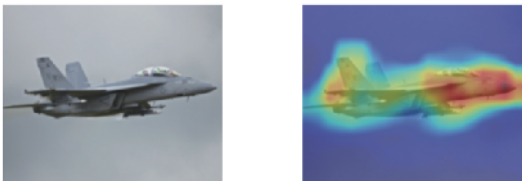
运动员和球场地面, 其中对球场的关注度最大, 从而准确地判断出棒球运动员击球的信息. 因此, 在理解图像时, 本文模型能够关注相应的图像区域, 为句子的生成提供准确的依据.



(a) A baseball player is getting ready to hit the ball during a game.



(b) Two women walking on a beach holding surfboards.



(c) A large airplane is flying in the air.

图 6 模型注意力权重可视化图

在图 6(b) 中, “beach” 在图像中占有较大的区域, 是较为重要的信息. 当生成描述时, 本文模型对图像的上半部分的沙滩和海浪关注程度最大, 且注重沙滩和冲浪板的关联性.

在图 6(c) 中, 当生成 “airplane” 时, 本文模型对飞机轮廓的判断能够达到较高的准确度, 且对于飞机前半部分和机翼关注度较高, 对包含信息较少的部分关注度较低.

由上述分析可知: 本文模型能够准确地判断重要信息所在的位置, 关注图像目标之间的关系, 实现准确表达图像中有效信息的功能.

5 结论与展望

本文提出了一种基于访问控制模块与原始信息注入的图像描述网络, 该网络对基线模型中语言 LSTM 的输入变量做了改进, 以充分利用全局图像信息和嵌入信息来生成句子. 此外, 提出访问控制模块的概念, 用于实现图更新机制. 同时, 本文设计了图更新系数, 用于指导图更新来确定更新的程度大小, 可以在一定程度上优化结点的更新程度. 本文进行了充分的实验证明该方法的有效性. 在未来的工作中, 本团队会继续研究场景图及模型框架的改进方式, 并考虑研究立体场景下的图像描述模型来进一步提升应用价值.

参考文献

1 Chen XL, Fang H, Lin TY, et al. Microsoft COCO captions: Data collection and evaluation server. arXiv: 1504.00325, 2015.

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

3 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: IEEE, 2015. 3156–3164. [doi: 10.1109/CVPR.2015.7298935]

4 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 International Conference on Machine Learning. Lille: JMLR.org, 2015. 2048–2057.

5 Lu JS, Xiong CM, Parikh D, *et al*. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 3242–3250. [doi: 10.1109/CVPR.2017.345]

6 Ranzato MA, Chopra S, Auli M, et al. Sequence level training with recurrent neural networks. arXiv: 1511.06732, 2015.

7 Rennie SJ, Marcheret E, Mroueh Y, *et al*. Self-critical sequence training for image captioning. Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1179–1195. [doi: 10.1109/CVPR.2017.131]

8 Ren SQ, He KM, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

9 Anderson P, He XD, Buehler C, *et al*. Bottom-up and top-down attention for image captioning and visual question answering. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City:

IEEE, 2018. 6077–6086. [doi: 10.1109/CVPR.2018.00636]

10 Yang X, Tang KH, Zhang HW, *et al*. Auto-encoding scene graphs for image captioning. Proceedings of 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 10677–10686. [doi: 10.1109/CVPR.2019.01094]

11 Johnson J, Krishna R, Stark M, *et al*. Image retrieval using scene graphs. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 3668–3678. [doi: 10.1109/CVPR.2015.7298990]

12 Li XY, Jiang SQ. Know more say less: Image captioning based on scene graphs. IEEE Transactions on Multimedia, 2019, 21(8): 2117–2130. [doi: 10.1109/TMM.2019.2896516]

13 Felzenszwalb PF, Girshick RB, Mcallester D, et al. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627–1645. [doi: 10.1109/TPAMI.2009.167]

14 Li YK, Ouyang WL, Zhou BL, et al. Factorizable net: An efficient subgraph-based framework for scene graph generation. Proceedings of 15th European Conference on Computer Vision. Munich: Springer, 2018. 346–363. [doi: 10.1007/978-3-030-01246-5_21]

15 Xu DF, Zhu YK, Choy C B, *et al*. Scene graph generation by iterative message passing. Proceedings of 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 3097–3106. [doi: 10.1109/CVPR.2017.330]

16 Zellers R, Yatskar M, Thomson S, *et al*. Neural motifs: Scene graph parsing with global context. Proceedings of 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2017. 5831–5840. [doi: 10.1109/CVPR.2018.00611]

17 Chen SZ, Jin Q, Wang P, *et al*. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9959–9968. [doi: 10.1109/CVPR42600.2020.00998]

18 Lin TY, Maire M, Belongie S, *et al*. Microsoft COCO: Common objects in context. Proceedings of 13th European Conference on Computer Vision. Zurich: IEEE, 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]

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

20 Lin CY. Rouge: A package for automatic evaluation of summaries. Text summarization branches out. Barcelona: Association for Computational Linguistics, 2004. 74–81.

21 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. Ann Arbor: Association for Computational Linguistics, 2005. 65–72.

22 Vedantam R, Zitnick CL, Parikh D. CIDEr: Consensus-based image description evaluation. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 4566–4575. [doi: 10.1109/CVPR.2015.7299087]

23 Chen XL, Zitnick CL. Mind's eye: A recurrent visual representation for image caption generation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 2422–2431. [doi: 10.1109/CVPR.2015.7298856]

24 Donahue J, Hendricks LA, Guadarrama S, *et al*. Long-term recurrent convolutional networks for visual recognition and description. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 2625–2634. [doi: 10.1109/CVPR.2015.7298878]

25 Mao JH, Xu W, Yang Y, et al. Deep captioning with multimodal recurrent neural networks (m-RNN). arXiv: 1412.6632, 2014.

(校对责编: 孙君艳)