[文章编号] 1003－4684(2021)04-0017-05

基于文本层级结构的图像描述生成算法

吴 禹，靳华中

（湖北工业大学计算机学院，湖北 武汉 430068）

[摘 要] 针对现有图像描述生成算法在解码阶段由于语言模型结构简单，解码表达能力较弱，容易造成语义缺失的问题，引入有序长短时记忆网络（ON－LSTM），改进现有模型解码器，构建双层 LSTM 架构，显式的提取描述文本层级结构，解码出更丰富的语义特征。在 MSCOCO 数据集上进行训练和测试，实验结果表明，改进的算法能够生成更加符合自然语言习惯的描述语句。

[关键词] 图像描述生成；语言模型；有序长短时记忆网络；文本层级结构

[中图分类号] TP3-0 [文献标识码] A

图像描述生成是一个融合计算机视觉、自然语言处理和机器学习的综合问题。图像描述生成方法使用符合人类语言习惯的句子描述图像。算法模型在检测图像中的目标的同时，还要对目标的视觉元素，如目标的动作和属性有一定的认知。在此基础上，通过理解目标之间的相互关系，构建图像的场景，目的是生成具有语义关系的、符合自然语言习惯的描述句子。

目前图像描述生成模型普遍采用编码器－解码器框架。编码器利用卷积神经网络（CNN）从图像中提取图像特征[1]，解码器使用循环神经网络作为语言模型来预测文本，引入注意力机制，有效地选择视觉特征向量来初始化语言模型隐藏状态[2]，提高视觉信息处理效率，在客观指标上展现出明显优势。但在语言模型的构建上存在不足，使得语义信息不能充分表达。文献[3]将图像特征向量与每个单词的嵌入连接起来，以便为以后生成的单词保留视觉信息，但难以解决 RNN 梯度消散问题。文献[4]提出通过与自动重构网络（ARnet）耦合来增加相邻隐藏状态之间的相关性。并嵌入上一隐藏层状态解码更多语义特征信息，然而使用欧几里得距离的正则化方法可能会直接减少每个隐藏状态的 L2 范数，使得评价指标没有获得较大改善。文献[5]在自下而上和自上而下的组合注意力机制的基础上融入图文匹配模型（Stacked Cross Attention Network，SCAN）[6]对注意力机制的训练过程进行弱监督，增强了注意力机制对单词和图像区域的对应能力，但难以表征图像目标之间语义关系。这些方法的语言模型普遍只将当前单词隐藏状态作为输入，并仅针对一种输出状态计算结果，忽略了相邻单词之间的文本层级结构，容易在最终生成的文本中带来累积的错误。

而在自然语言处理领域，已有文献利用文本层级结构进行语言建模。文献[7]引入了句法距离这一概念来引导语言模型完成句法解析任务，但算法实现的复杂度较高，很难在实际情况中使用。文献[8]使用具有不同时间尺度的递归模型获取层次结构，更新 RNN 的隐藏状态，但需要施加预定义的层次结构。受此启发，本文在解码器阶段构建双层 LSTM 网络，第一层视觉选择 LSTM 融合注意力机制，从整体上得到图像中目标之间的语义信息，同时能够从细节得到图像特征信息。第二层语言模型 LSTM 使用有序长短时记忆网络[9]，在训练过程利用文本层级结构预测描述，增强语言模型表达能力，从而生成更符合自然语言习惯的描述。

1 自然语言的文本层级结构分析

1.1 文本层级结构

在自然语言处理领域中，语言的表现形式遵循一定的层级结构[10]，组成语句的各个单位处于语义层面，构成树状的文本层级结构[11]，即自然语言是由处在不同层级结构的单位要素组成的层级装置。如图 1 所示，在英文句子中，单词可以认为是最低层级的结构，词组次之。

[收稿日期] 2021－03－15

[第一作者] 吴 禹（1996－），男，湖北咸宁人，湖北工业大学硕士研究生，研究方向为图像描述生成

[通信作者] 靳华中（1973－），男，湖北洪湖人，湖北工业大学副教授，研究方向为计算机视觉

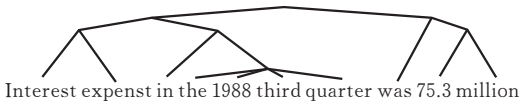

图 1 文本层级结构示例

语言与其他数据一个显著的区别就在于其本身拥有一定的层级结构，因为语言的组成受到语法规则限制，低层级的语义单元组成高层级的语义单元，而最高层级的语义单元就代表了整个句子的含义。人们曾经试图对语言的这种结构进行建模，利用语法规则进行语义解析，建立语义分析树，再根据解析的结果从下而上递归获得句子的表征。单位结构层级越高，在句子中的跨度就越大。这意味着编码时能区分高低层级的信息；其次，高层级的信息意味着它要在高层级对应的编码区间保留更久，而低层级的信息则意味着它在对应的区间更容易被遗忘。

1.2 文本层级结构的提取

针对语言的层级结构，文献[9]提出了有序长短时记忆网络(Ordered Neurons Long Short-Term Memory，ON-LSTM)。传统 LSTM 网络中，神经元通常都是无序的，运算过程中涉及到的所有向量的位置按照相同方式重新打乱，权重的顺序也将相应地打乱，输出结果可以只是原来向量的重新排序，信息量不变。有序长短时记忆网络则把神经元的序信息利用起来，按排序分区间更新状态，使其表示一些特定的结构。用这种结构来表征文本层级信息，使 ON-LSTM 在训练中自然地学习到文本的层级结构，从而增强语言模型表达能力。算法流程如图 2 所示。

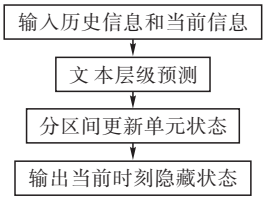

图 2 有序长短时记忆网络算法流程

具体过程为：ON-LSTM 单元状态 c_t 按照向量索引值进行排序，语法层次较高的信息储存在 c_t 上面的维度中，较低的则储存在下面的维度。定义两个 one -hot 向量代表历史信息最低层级 l_f 和当前信息最高层级 l_i，分区间更新规则为：1) $l_f < l_i$，历史信息层级和当前信息层级有重合部分，低于历史信息 l_f 层级部分，c_t 只保留当前信息；高于当前信息 l_i 层级部分，c_t 只保留历史信息；重合部分则按原始 LSTM 算法更新。2) $l_f \geq l_i$，历史信息层级和当前信息层级没有重合部分，c_t 分别保留各自层级部分，未重合部分置为 0。

经过以上规则分区间更新 c_t，文本高层信息更新频率较低，在模型循环过程中能保留较长距离，文本底层信息在每一个时间步内都可能更新。从而通过定序嵌入层级结构，即按信息跨越幅度分组更新输入文本序列的层级结构。如图 3 所示，对于给定语言序列 $[x_1, x_2, x_3]$ 及其句法树，ON-LSTM 通过上述算法流程，动态分配其隐藏状态向量的维数，用以对应表示给定文本 $[x_1, x_2, x_3]$ 的层级结构。

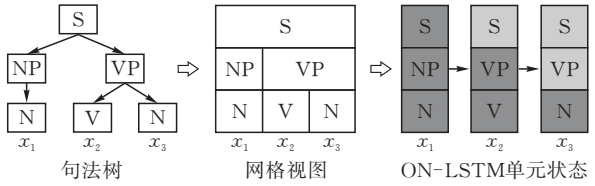

图 3 文本层级结构表征

2 基于文本层级结构的图像描述生成算法

2.1 本文模型框架

本节提出了基于文本层级结构的图像描述生成算法，如图 3 所示，模型采用编码器-解码器架构，编码器提取图像特征，解码器接收特征进行解码，生成图像的最终描述。在编码器阶段，用 CNN 对图像进行特征提取，并根据神经网络卷积层特性分别获取图像对应的全局特征和局部特征。然后，在第二个阶段使用注意力机制和双层有序长短时记忆网络将初始图像全局特征和局部特征信息相融合，将融合后的信息特征输入 ON-LSTM 进行解码。

在编码器阶段应用了两种不同尺度的图像特征，分别为局部特征和全局特征，全局特征包含图像目标语义关系，引导第一层视觉 LSTM 关注特定目标；局部信息包含目标具体特征，引导第二层语言 LSTM 解码准确信息。这两种不同尺度的图像特征全部由解码器经过预先训练好的卷积神经网络提取得到。在本文中用 f 表示局部特征，则有：

$$f = \{f_1, f_2, \cdots, f_k\}, f_k \in \mathbb{R}^{1*r}$$

其中，$\{f_1, f_2, \cdots, f_k\}$ 表示 k 个局部特征，$f_k \in \mathbb{R}^{1*r}$ 表示每个图像区域的特征维度为 $1*r$。局部特征通过全局平均池化得到图像的全局特征 $\overline{f}$

$$\overline{f} \in \mathbb{R}^{1*d}$$

其中，$\overline{f} \in \mathbb{R}^{1*d}$ 为全局特征的维度是 $1*d$，最后将图像全局特征和局部特征馈入解码器。

2.2 融合文本层级结构的解码过程

第一层 LSTM 输入为全局图像特征 $\overline{f}$、上一时间步输出的词向量 $W_e y_t$ 以及第二层 LSTM 的输出 h_{t-1}^2，其中语言模型隐藏层 h_{t-1}^2 在生成首个单词时没有输入第一层 LSTM 中。

在 attend 部分中输入为所有的图像特征 $f =$

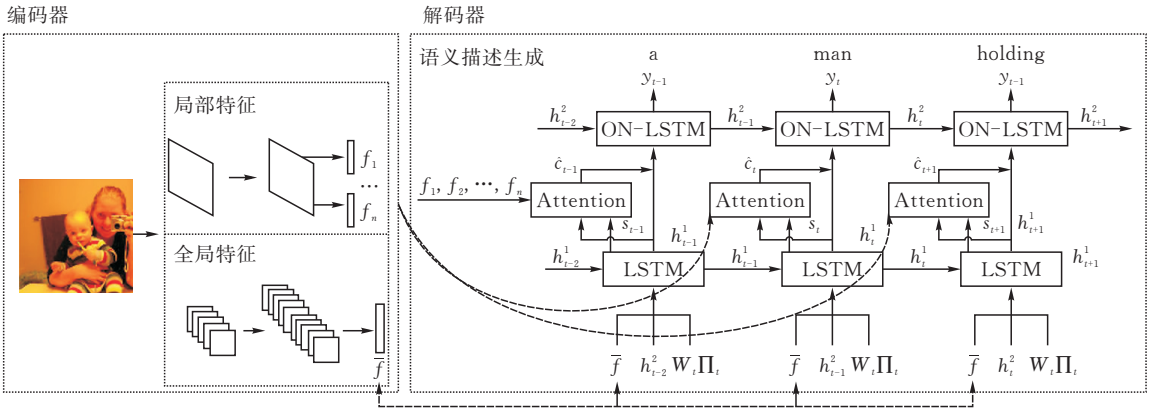

图 4　本文模型框架结构

$\{f_1, f_2, \cdots, f_k\}$，$f_k \in \mathbb{R}^{1*r}$ 和视觉选择 LSTM 的输出 h_t^1。具体使用的是两层全连接层，attend 模块的输出 c_t 的计算过程如下式所示：

$$z_t = \boldsymbol{w}_h^{\mathrm{T}} \tanh(\boldsymbol{W}_a a + (\boldsymbol{W}_g h_t^1)) \tag{1}$$

$$\alpha_t = \mathrm{softmax}(z_t) \tag{2}$$

$$c_t = \sum_{i'=1}^{k} \alpha^{<t,i'>} f^{i'} \tag{3}$$

式(1)中，W_a、W_g 分别表示两个全连接层的映射矩阵，输出结果如经过式(2)一层 softmax 得到对某一区域关注度 α_t。在(3)式中，关注度用 $\alpha^{<t,t'>}$ 表示，其中 t 表示生成第 t 个单词，t' 表示图像的第 t' 个区域。

语言模型 ON-LSTM 的输入为：attend 模块处理后的包含图像上下文信息 c_t，第一层视觉选择 LSTM 的隐藏层状态 h_t^1，两向量连接为当前输入 x_t。文本层级结构的具体机制在于，定义两个 one-hot 向量，根据向量索引值区分当前信息 x_t 和历史信息 h_{t-1}^2 层级，分区间更新第二层 ON-LSTM 单元状态 c_t^2，在单元状态进行分层，促进每个神经元内部存储的信息生命周期的区分：单元状态 c_t^2 较高维度将存储长期信息，这些信息包含了生成描述的高层语义信息。而排名较低的维度将存储可以迅速被忘记的短期信息。分别记为主遗忘门 $\tilde{f}_t$ 和主输入门 $\tilde{i}_t$：计算公式为

$$\tilde{f}_t = \mathrm{cusum}(\mathrm{softmax}(W_{\tilde{f}} x_t + U_{\tilde{f}} h_{t-1}^2 + b_{\tilde{f}}))$$

$$\tilde{i}_t = 1 - \mathrm{cusum}(\mathrm{softmax}(\widetilde{W}_i x_t + \widetilde{U}_i h_{t-1}^2 + b_{\tilde{i}}))$$

其中 cusum 为累积函数，以主动分配维度来存储长期或短期信息，避免在高级维度和低级维度之间进行严格划分。将单元状态的维度动态地重新分配给每个节点，迫使神经元在不同的时间尺度上代表信息。给定任意序列 $[y_1, y_2, \cdots, y_n]$，cusum 计算公式如下：

$$\mathrm{cusum}([y_1, y_2, \cdots, y_n]) =$$
$$[y_1, y_1 + y_2, \cdots, y_1 + y_2 + \cdots y_n]$$

通过主遗忘门 $\tilde{f}_t$ 和主输入门 $\tilde{i}_t$ 对单元状态分区间更新，强制更新神经元的顺序，使每个神经元的门都依赖于其他神经元，将树状结构显式编码进语言描述生成阶段，使图像特征语义信息和语言模型句法结构融合交互，进一步增强了解码器的语言解码能力，基于文本层级结构的门控单元状态更新规则为：

$$\omega_t = \tilde{f}_t \cdot \tilde{i}_t$$

$$\hat{f}_t = f_t \cdot \omega_t + (\tilde{f}_t - \omega_t)$$

$$\hat{i}_t = i_t \cdot \omega_t + (\tilde{i}_t - \omega_t)$$

$$c_t^2 = \hat{f}_t \cdot c_{t-1}^2 + \tilde{i}_t \cdot \hat{c}_t^2$$

$$h_t^2 = o_t \cdot \tanh(c_t^2)$$

通过门控单元，获得语言模型 LSTM 的隐藏层状态 h_t^2。语言模型 LSTM 隐藏层 h_t^2 通过 softmax 层，输出对应词汇表中单词的概率分布，其词向量维度与词汇表向量大小 V 相同，取其中最大概率值的索引，该索引值返回词汇表中搜索单词，即为模型在时刻 t 所输出的单词。生成第 t 个单词计算方式为：

$$p(y_t \mid y_{t-1}) = \mathrm{softmax}(W_p h_t^2 + b_p)$$

3　实验结果和分析

3.1　数据准备和预处理

本文采用的数据集为微软 COCO2014，包含三部分内容，训练集、验证集和测试集。各部分数据集由图像和 json 文件组成，json 文件包含对每幅图像的 5 个英文描述。数据集包含的图像总共 82 783 张，对应的英文描述为 413 915 个。

描述文本的预处理阶段过程为：1)图像描述中的特殊符号“&”用“and”代替，标点符号用空格代替；2)使用图像 id、图像文件名和图像描述建立描述句库，通过检索图像信息来查找图像描述；3)使用数据集描述出现单词建立词汇表，词汇表向量每一维度对应数据集中单词，语言模型通过检索词汇表索引值生成描述单词。

本文图像描述生成方法在 tensorflow 平台上建

立。采用小批量梯度下降法对损失函数进行优化，提高模型训练的收敛速度。学习速率为 0.01，迭代次数为 100 次，最小批次为 128 次。

3.2 评价指标和实验结果

目前图像自动标注领域常用的评价标准主要分为 5 类，分别是 BLUE、METEOR、ROUGE、CIDEr 和 SPICE。这 5 类标准能对模型生成的图像描述进行量化标准的客观评价。在本文实验中采用 BLUE、METEOR 和 CIDEr 对生成描述进行评分。

表 1 MSCOCO 数据集实验对比

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
NIC	64.2	45.1	30.4	24.6	—	—
mRNN	67.0	47.6	32.3	25.8	23.7	—
Log Biliner	65.2	42.9	27.1	17.5	17.1	—
Soft-Attend	69.5	47.8	33.4	25.9	23.4	0.875
Our model	70.2	49.9	34.8	26.4	23.9	0.909

实验结果证明，本文模型在 BLUE、METEOR 和 CIDEr 评价指标上要优于 NIC、mRNN、Log Biliner[12] 和 Soft-Attend 模型。

3.3 实验结果分析

在实验结果可视化对比中，图 5 中本文模型对比 mRNN 模型，mRNN 模型生成描述句法树高度为 6，叶子结点数为 7，本文模型生成描述句法树高度为 7，叶子结点数为 11。生成描述将“field”生成为“hillside”，并生成了“lush green”加以修饰，即提取到了更为复杂的语义特征，使描述更加生动。图 5 中本文模型对比 soft-attend 模型，soft-attend 模型生成描述句法树高度为 5，叶子结点数为 8，本文模型生成描述句法树高度为 6，叶子结点数为 11。预测了“in the ocean”这一空间背景信息，从整体上提取到更丰富的语义信息，模型语义表达能力较 soft-attend 更强。

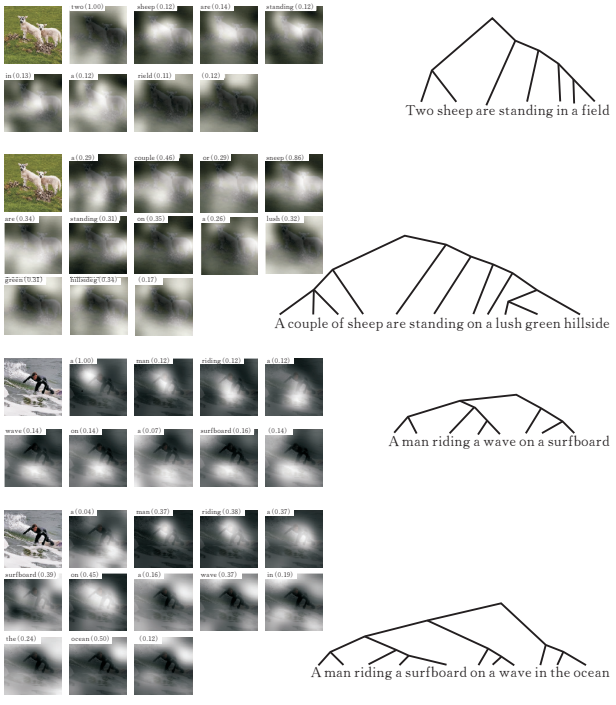

图 5 实验结果可视化对比

4 结论

针对现有采用编码器—解码器框架的图像描述生成算法，在解码阶段由于语言模型结构简单，解码表达能力较弱，容易造成语义缺失的问题。本文方法通过引入有序长短时记忆网络构建双层 LSTM 架构，来改进现有模型解码器，使模型能够显式的提取描述文本层级结构，解码出更丰富的语义特征。本文改进的方法在 MSCOCO 数据集上进行训练和测试，实验结果表明，改进的算法能够有效提取文本层级结构，充分利用图像空间信息与内容语义对齐来改善语言模型解码表达能力，最终提高了图像描述实验效果，生成更加符合自然语言习惯的描述语句。

[参 考 文 献]

[1] ORIOL V, ALEXANDER T, SAMY B, et al. Show and tell: A neural image caption generator [C]// CVPR2015: Proceedings of the 2015 International Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015, 3156-3164.

[2] XU K, BA J, KIROS R, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[J]. Computer Science, 2015:2048-2057.

[3] MAO J, XU W, YANG Y, et al. Deep captioning with multimodal recurrent neural networks (m-RNN) [C]//3rd International Conference on Learning Representations (ICLR 2015). San Diego, CA, United States, 2015.

[4] CHEN X, MA L, JIANG W, et al. Regularizing rnns for caption generation by reconstructing the past with the present[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2018, 7995-8003.

[5] ZHOU Y, WANG M, LIU D, et al. More grounded image captioning by distilling image-text matching model[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[6] LEE K H, XI C, GANG H, et al. Stacked cross attention for image-text matching[C]// 15th European Conference on Computer Vision (ECCV 2018). Munich, Germany: Springer Verlag, 2018,212-228.

[7] YIKANG S, ZHOUHAN L, CHIN-WEI H, et al. Neural language modeling by jointly learning syntax and lexicon [C]//6rd International Conference on Learning Representations (ICLR 2018). Vancouver, Canada,2018.

[8] JAN K, KLAUS G, FAUSTINO G, et al. A clockwork RNN[C]//In Proceedings of the 31st International Conference on International Conference on Machine Learning-Volume 32 (ICML 2014). JMLR.org, II-1863-II-1871.

[9] SHEN Y, TAN S, SORDONI A, et al. Ordered neurons: integrating tree structures into recurrent neural networks[C]//7rd International Conference on Learning Representations(ICLR 2019). New Orleans, LA, United States,2019.

[10] 万齐智,万常选,胡蓉,等.基于句法语义依存分析的中文金融事件抽取[J].计算机学报,2021,44(3):508-530.

[11] 宗成庆.统计自然语言处理[M]. 北京:清华大学出版社,2013:16-17

[12] KIROS R, SALAKHUTDINOV R, ZEMEL R. Multimodal neural language models [C]//Proceedings of the International Conference on International Conference on Machine Learning. Beijing: JMLR,2014: 595-603.

Image Caption Based on Text Hierarchical Structure

WU Yu,JIN Huazhong

(*School of Computer Science, Hubei Univ. of Tech., Wuhan* 430068,*China*)

Abstract: Aiming at the existing image description generation algorithm, in the decoding stage, the language model is simple in structure and weak in decoding expression ability, which can easily cause the problem of lack of semantics. The ordered neurons Long Short-Term Memory network (ON-LSTM) is introduced to construct a two-layer LSTM architecture to improve the decoder of the existing model, so that it can explicitly extract the text hierarchical structure of the description to decode richer semantic features. Training and testing on the MSCOCO data set, the experimental results show that the improved algorithm can generate description sentences that are more in line with natural language habits. ordered neurons Long Short-Term Memory network

Keywords: image caption; language model; ON-LSTM; text hierarchical structure

[责任编校: 张岩芳]