北京航空航天大学学报
Journal of Beijing University of Aeronautics and Astronautics
ISSN 1001-5965,CN 11-2625/V

《北京航空航天大学学报》网络首发论文

题目：基于依存句法的图像描述文本生成
作者：刘茂福，毕健旗，胡慧君，代建华
DOI：10.13700/j.bh.1001-5965.2020.0443
收稿日期：2020-08-21
网络首发日期：2020-09-29
引用格式：刘茂福，毕健旗，胡慧君，代建华．基于依存句法的图像描述文本生成．北京航空航天大学学报. https://doi.org/10.13700/j.bh.1001-5965.2020.0443

基于依存句法的图像描述文本生成

刘茂福[1,2]✉，毕健旗[1,2]，胡慧君[1,2]，代建华[3]

（1. 武汉科技大学 计算机科学与技术学院，武汉 430065；2. 武汉科技大学 智能信息处理与实时工业系统湖北省重点实验室，武汉 430081；3. 湖南师范大学 智能计算与语言信息处理湖南省重点实验室，长沙 410081）

*通信作者 E-mail：liumaofu@wust.edu.cn

摘要 图像描述文本生成是计算机视觉与自然语言处理交叉领域的研究热点。现有深度学习模型能够应用词性序列和句法树使生成的文本更符合语法规则。然而，上述模型生成的文本多为简单句，在多样性和句法复杂度方面，尚未取得突破；在语言模型促进深度学习模型的可解释性方面，当前研究甚少。将依存句法信息融合到深度学习模型以监督图像描述文本生成的同时，可使深度学习模型更具可解释性。图像结构注意力机制基于依存句法和图像视觉信息，用于计算图像区域间关系并得到图像区域关系特征；融合图像区域关系特征和图像区域特征，与文本词向量通过长短期记忆网络(LSTM)，用于生成图像描述文本。在测试阶段，通过测试图像与训练图像集的内容关键词，计算两幅图像的内容重合度，间接提取与测试图像对应的依存句法模板；模型基于依存句法模板，生成多样的描述文本。实验结果验证了模型在改善生成文本多样性和句法复杂度方面的能力；表明模型中的依存句法信息增强了深度学习模型的可解释性。

关键词 图像描述生成；依存句法；图像结构注意力；内容重合度；深度模型可解释性

中图分类号 TP37；V19 文献标识码：A DOI: 10.13700/j.bh.1001-5965.2020.0443

Image Captioning Based on Dependency Syntax

LIU Maofu[1,2]✉, BI Jianqi[1,2], HU Huijun[1,2], DAI Jianhua[3]

(1. School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430081, China; 2. Hubei Provincial Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan 430081, China; 3. Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Changsha 410081, China)

*Tel.: +86-18971374322 E-mail: liumaofu@wust.edu.cn

Abstract Image captioning is one of the research hotspots in cross-domain of computer vision and natural language processing. Current deep learning model can automatically apply the part-of-speech sequences and syntactic trees to make the generated text in line with grammar. However, the above-mentioned models generally generate the simple sentences, and have not performed well in terms of the text diversity and syntax complexity. There is no groundbreaking work in language models promoting the interpretability of deep learning models. Integrating the dependency syntax into the deep learning model to supervise the image captioning, which can make deep learning models more interpretable. An image structure attention mechanism, which recognizes the relationship between image regions based on the dependency syntax, is applied to compute the visual relations and obtains the features. The fusion of image region relation features and image region features, and the word embedding are employed into Long Short-Term Memory (LSTM) to generate the captions. In testing, the content keywords of the testing and training image datasets are produced due to the content overlap of two images, and the dependency syntax template corresponding to the test image can be indirectly extracted. According to the dependency syntax template, the diverse descriptions can be generated. Experiment results have verified the capacity of the proposed model to improve the diversity of generated captions and syntax complexity and indicate that the dependency syntax can enhance the

收稿日期：2020-08-21

基金项目：国家社科基金重大研究计划（11&ZD189）；全军共用信息系统装备预先研究项目（31502030502）

作者简介：刘茂福 ,男，博士，教授，博士生导师。主要研究方向：自然语言处理，图像分析与理解。毕健旗， 男，硕士研究生。主要研究方向：自然语言处理。胡慧君， 女，博士，副教授，硕士生导师。主要研究方向：智能信息处理，图像分析与理解。代建华 ，男，博士，教授，博士生导师。主要研究方向：人工智能，智能信息处理

Fund：Major Projects of National Social Science Foundation of China (11&ZD189); Pre-research Foundation of Whole Army Shared Information System Equipment (31502030502).

网络首发时间：2020-09-29 16:13:35 网络首发地址：https://kns.cnki.net/kcms/detail/11.2625.V.20200929.1427.003.html

interpretability of deep learning model.
Key words Image Captioning; Dependency Syntax; Image Structure Attention; Content Overlap; Interpretability of Deep Learning Model

近年来，面向跨模态的图像描述文本生成因其在计算机视觉和自然语言处理交叉领域的重要性，已引起越来越多的研究者的关注。给定一幅图像，图像描述文本生成旨在生成语法正确且可精准描述图像内容的文本。基于注意力机制的编解码框架已被广泛应用于图像描述文本生成[1]，其提取有价值的视觉特征，从而使生成的文本与图像的内容紧密相关。基于视觉文本注意力机制的编解码框架虽能较精确地发现图像视觉区域信息，但仍存在两个问题：（1）在编码阶段生成当前时刻图像的视觉区域信息时，其仅考虑当前时刻前的文本信息，忽略了图像本身蕴含的视觉结构信息；然而此类信息难以在输入文本中发现，在缺少图像结构信息的情况下，模型无法很好地发现图像中存在的潜在语义信息，也无法发现图像区域与文本的对应关系，进而不能生成准确的描述文本。（2）在解码阶段的长短期记忆网络（LSTM）中，生成文本所需的语法规则只能从其自编码机制中获得，而由于记忆局限性，LSTM 更易于捕获图像特征不能很好地学习到长距离的词与词间的依赖关系，使生成的文本多以简单句的形式存在。

为了使模型更好地发现图像结构信息，Lu 等[2]使用 Faster R-CNN[3,4]来识别图像实体，再进而根据文本信息来建立图像中实体间关系；为了使模型更好地学习语法规则，Aditya 等[5]和 Wang 等[6]将词性序列和句法树等语法信息作为模型的先验知识，并将之与词向量共同输入到解码端参与文本生成。然而，词性序列和句法树包含的语法规则只蕴含每个词的标注信息，虽然序列或树的表示形式能提供一定的结构信息，但仍不能精准地体现文本中词与词之间的关系，即图像中实体间应存在的语法关系，如依存句法关系。因此，本文拟为模型引入由依存句法构建的语言模型，增强模型理解图像实体间的语法规则的能力，同时在一定程度上提高基于深度学习的图像描述文本生成模型的可解释性。

由于文本单词间可能存在长距离的依存关系，本文选择可以捕获序列内部长距离依赖的自注意力机制(Self-Attention)[7]来对每个依存句法单元进行编码。编码阶段，在模型得到基于视觉注意力机制计算出的图像区域特征后，为了得到图像区域间的结构语法关系表示，本文提出了一种图像结构注意力方法，从依存句法中提取词与词之间的语法关系，并将其关联在与图像对应的实体上；最后，本文将两者进行融合并与当前时刻的文本特征向量共同作为 LSTM 的输入，解码生成文本。

针对图像与依存句法的对应关系，在训练阶段，每幅图像与依存句法以描述文本为纽带，呈现直接对应关系；在测试阶段，为得到与测试图像匹配的依存句法模板，本文提出了一种基于内容重合度的提取方法，该方法将训练一个图像标签分类器，旨在输出给定图像相应的内容关键词，根据测试图像与训练集图像的最大内容重合度建立起测试图像与训练图像的对应关系，从而间接建立测试图像与依存句法模板的映射关系。

本文提出的基于依存句法的图像描述文本生成模型训练阶段将依存句法应用在图像结构注意力中，测试阶段根据图像内容重合度建立测试图像与依存句法模板的映射关系。此外，模型生成的图像局部间结构语法关系与依存句法单元对应，生成文本的句法与提取的依存句法相吻合。

1 相关工作

编解码方法最初被广泛应用于机器翻译[8]，Vinyals 等[9]将其迁移至图像描述生成领域，取得了显著的效果。Xu 等[1]提出了一种基于注意力机制的方法来优化长短期记忆网络(LSTM)的存储能力，使模型在不同时刻关注图像不同区域，生成更准确的描述句子；Zhu 等[10]提出了一种基于主题词的图像描述方法，使生成的句子可以包含更多的图像焦点信息；Wang 等[11]提出了一种基于内源性和外源性视觉信息来生成不同角度的描述句子的方法；为了使图像信息与文本信息更加契合，Liu 等[12]提出了一种多头注意力机制来使图像中的信息与视觉文本信息对齐，使解码端可以更加准确地解析模型提供的图像与文本信息。

上述方法均致力于从图像中挖掘更多的视觉信息来填充句子，而在句子结构与句法层次的优化上

稍有欠缺。为了使生成模型学习到更多的句子结构信息，Wang 等[6]提出了一种基于句子骨架与属性词的图像描述生成方法，将从图像中得到的信息按照属性词与骨架的排序生成句子；Aditya 等[5]提出了一种基于词性标注序列的图像描述生成方法，加快句子生成速度的同时，也增加了句子描述事物的多样性。

依存句法是一种以谓语动词为结构中心的句法分析方法，在自然语言处理领域被广泛应用。Falenska 等[13]基于 BiLSTM 结构验证了依存句法内含结构信息的有效性；Li 等[14]将半监督学习方法应用在跨领域依存句法分析中，得到了更好的中文依存句法分析效果；Wang 等[15]按照依存句法单元构建句法树，验证了按此顺序定位自注意力机制中每个词，可以优化原模型，达到了提升模型效果的目的。综上所述，依存句法在提取句子结构信息方面显得十分有效。但目前，在图像描述生成任务当中，图像内容分析与生成文本的过程缺乏结构信息。因此，本文将研究基于依存句法的图像描述生成方法。

2　基于依存句法的图像描述方法

本文提出的基于依存句法的图像描述文本生成方法的训练阶段框架图如图 1 所示。首先，输入图像经过 ResNet-101 网络[16]得到图像的特征向量；接着，图像特征向量与词向量一同输入至视觉注意力以计算图像的局部区域特征向量表示，与经过 Self-Attention 的依存句法向量一同输入至图像结构注意力来计算图像的局部区域关系特征向量表示；最后，将图像局部区域特征向量与局部区域关系特征向量融合，与当前时刻的词嵌入向量一同输入到 LSTM 中生成描述文本。图 1 所示视觉注意力来自于基于注意力的 NIC+ATT 模型[1]中的视觉注意力。

2.1　依存句法库的构建

本文使用 Stanford CoreNLP[17]对全部训练集文本进行依存句法分析，构成依存句法库。本文将每个依存关系与存在此依存关系的两个单词位置信息作为一个依存句法单元，其表示形式为由依存关

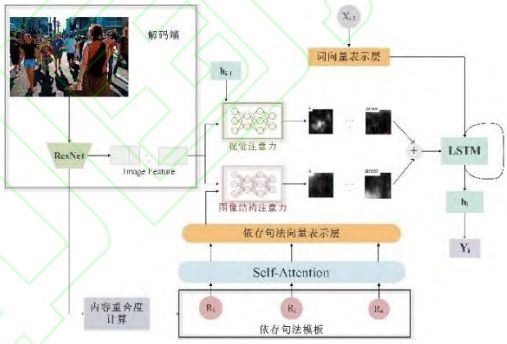


图 1　模型框架图
Fig.1 The framework of our model

系类型、根节点位置和自身节点位置三个元素构成的三元组，如公式（1）所示。

$$U=(r_i,l_r,i),i\in[1,n] \tag{1}$$

式中:U 为依存句法单元三元组，l_r 代表与文本中第 i 个单词存在依存关系的节点，在文本中对应的单词下标，r_i 代表文本中第 l_r 个单词与第 i 个单词之间的依存关系，n 表示文本长度。

数据集文本中的每个单词，依存句法单元序列使其都存在唯一的依存关系三元组与之对应，且每个句子的长度与依存句法单元序列的长度相等。

如图 2 所示，在依存句法分析中，名词“牛仔”是动词“骑马”的名词主语，因此三者构成三元组（“nsubj”，6，1）。此外，中文分词的结果中包含 7 个词，对应的依存句法单元有 7 个。

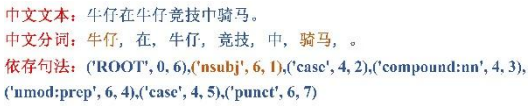


图 2 图像描述文本依存句法示例
Fig.2 A dependency Syntax example of an image caption

2.2 依存句法单元向量表示

自回归机制可以提取文本的向量表示，但不能很好地解决长距离依赖问题。依存句法体现的是句子中的词与根、词与词之间关系，包含长距离的语法规则。为了使计算机更好地理解依存句法单元三元组序列，本文使用自注意力机制来解决长短期记忆网络的自回归机制不能捕获长距离关系的问题。自注意力机制摒弃自回归机制逐个输入单词计算向量表示的方法而一次性处理整个文本，因此自注意力机制更适合用来处理具有长距离关系的文本。

$$R=soft\max(\frac{QK^T}{\sqrt{d_k}})\times V \tag{2}$$

式中:Q、K 和 V 为输入的依存句法单元序列随机嵌入向量，$\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{N\times E}$，$N$ 是依存句法长度，E 是词嵌入向量的维度，d_k为 K 的第三维维度，R 则为依存句法单元序列的向量表示，$softmax()$为归一化函数。

2.3 图像结构注意力

一幅图像包含实体、背景等信息，且在对应的描述文本中可找到与图像中的实体对应的单词。从构成句子的句法规则角度，为了生成句法正确的文本，掌握单词与单词之间的关系十分重要，也就是需要发现图像中实体与实体之间的关系。图像结构注意力的目的是通过文本的依存句法中表示的词与词之间的关系，使模型学习到图像中对应每个实体之间的关系特征向量表示。具体的计算公式如公式（3）所示。

$$\begin{aligned}\alpha &= soft\max(relu(W_i I^{'} + W_r R))\\ g &= sigmoid(W_g R)\\ I &= g\times mean(\alpha \times I^{'})\end{aligned} \tag{3}$$

式中: $\boldsymbol{I'} \in \mathbb{R}^{K\times F}$与$\boldsymbol{R} \in \mathbb{R}^{N\times E}$表示图像特征向量与依存句法单元序列向量，$K$ 与 F 分别表示图像特征点数和维度，$\boldsymbol{\alpha}$表示依存句法单元向量在图像特征点上关注的权重系数，$\boldsymbol{g}$表示依存句法单元向量在图像特征维度上关注的权重系数，$\boldsymbol{W}_{*}, * \in \{i, r, g\}$是模型参数，$softmax$、$relu$和$mean$分别是归一化函数、激活函数和均值函数，$I$ 则为最终的图像结构特征向量。

2.4 描述文本生成

图像的描述文本由 LSTM 生成。t 时刻输入至 LSTM 的元素由 LSTM 在 t-1 时刻生成的单词 Y_{i-1} 与隐藏状态 h_{i-1}、Y_{i-1} 在 t 时刻关注的图像视觉区域特征 I_x、依存句法单元 R 在 t-1 时刻关注的图像结构特征 I_r。具体的计算公式如公式（4）所示。

$$\begin{aligned}I_{add} &= I_x + I_r\\ h_t &= LSTM(X_{t-1}, h_{t-1}, I_{add})\\ Y_t &= soft\max(h_t)\end{aligned} \tag{4}$$

本文提出的模型训练过程的目标为：通过最小化损失函数来学习模型的最优参数，损失函数的具体计算公式如下：

$$L = -\frac{1}{N}\sum_{1}^{N}\sum_{i=1}^{len(Y)}\log p(Y_i) \tag{5}$$

式中:N 为训练集描述文本的数量，$len(Y)$表示描述文本的长度。

2.5　基于内容重合度的依存句法模板提取

2.5.1　内容关键词分类器

内容关键词分类器用来对测试图像进行分类，以建立起测试图像与依存句法模板的映射关系。为得到描述文本依存句法尽可能相似的图像，首先，内容关键词标签来自于训练集描述文本中的单词，且该单词在该文本对应的依存句法中带有“ROOT”关系。此外，该单词是训练集文本中出现的高频词。本文根据上述规则选择 300 个关键词作为内容关键词分类器的标签，且将该分类器视为一种图像多分类问题。具体计算公式如下：

$$labels = soft\max(MLP_2(I)) \tag{6}$$

式中:I 代表图像特征向量，MLP_2 是多层感知机模型，此处层数为 2，激活函数为 $ReLU$。

2.5.2　内容重合度

内容重合度体现了测试图像与训练图像集中的每幅图像的相似程度。由于每个内容关键词代表图像中可能出现的实体或关系，基于关键词的内容重合度可以很好地提取主题类似的两幅图像。

测试图像经过内容关键词分类器得到 Top-K 的标签集 C_t，同样地，训练集中的每幅图像也得到 Top-K 的标签集 C_{tr}；由于每个标签的可信度均不同，导致每个标签都会被赋予不同的权重；最后，根据测试图像与训练及图像的内容重合度 S 来确定测试图像与训练集图像之间的相似关系。内容重合度的计算公式如下：

$$S_i^{\,x} = \frac{(K - ind_{ti}) + (K - ind_{tr})}{2}$$

$$S^x = \sum_i^M S_i^{\,x} \tag{7}$$

式中:ind_t^i 表示测试图像的 Top-K 标签集 C_t 中的第 i 个，同时在 C_{tr} 中出现的标签 c_i 在 C_t 中对应的下标值。同理，ind_{tr} 表示标签 c_i 在训练图像的 Top-K 标签集 C_{tr} 中的下标值。在所有的标签集合 C_t 和 C_{tr} 中，标签的下标值越大，其对应的权重越小。$S_i^{\,x}$代表第 x 幅训练图像与测试图像相同的第 i 个标签的得分，M 为相同标签的个数，S^x则为第 x 幅训练图像与测试图像的内容重合度。本文最后通过比较S^x的数值大小来确定训练图像集中与测试图像最相似的图像，进而建立测试图像与依存句法模板的映射关系。以 K 为 3 举例，假设 C_t 为[16，7，8]，C_{tr}^1[1，7，6] ，C_{tr}^2[1，7，8]，则 S^1 为 2，S^2 为 3。

算法 1. 内容重合度计算。

输入：测试图像 $T=(t_1, t_2, \ldots, t_n)$;

训练图像 $X=(x_1, x_2, \ldots, x_m)$;

输出：测试图像与训练图像的重合度 S。

1. 测试图像标签集 $C_t \leftarrow labels(T)$　　式（6）
2. 训练图像标签集 $C_{tr} \leftarrow labels(X)$　　式（6）
3. 相同标签集 $L \leftarrow (C_t, C_{tr})$
4. 对应下标 $Ind_t, Ind_{tr} \leftarrow (L, T, X)$
5. 计算内容重合度S^x　　式（7）

算法 1 最耗时操作集中在标签识别上，即步骤 1 和步骤 2 得到测试与训练图像标签的过程，其次是计算内容重合度S^x。假设测试图像数量为 n，训练图像数量为 m，则算法 1 的时间复杂度为 $O(mn)$。

3　实验

3.1　实验设置

本文实验数据集采用 Flickr30K、Flickr8K 和 Flickr8K-CN。Flickr30K 是一个被广泛使用的公开

英文数据集，共有图像 31784 幅，每幅图像对应 5 个英文描述句子。Flickr8K 中包含图像 8091 幅，同样每幅图像对应 5 个英文描述句子，Flickr8K-CN 则为 Flickr8K 对应的中文版本。在测试阶段，本文使用了七种常见的评价指标和句子长度来验证模型生成文本的质量。七种指标分别为：BLEU-1 至 BLEU-4[18]、METEOR[19]、ROUGE-L[20]、CIDEr[21]，其中前四个指标使用了 n-gram 算法统计生成文本与参考文本之间的覆盖率；METEOR 对 BLEU 算法进行了一定的改进，使其更好地体现出句中单词的召回率与准确率；ROUGE-L 基于最长公共子串来计算准确性与召回率；CIDEr 基于 TF-IDF 计算生成文本与参考文本的余弦相似度来衡量文本的一致性；此外，本文将模型生成的描述文本平均长度作为第八种评测指标 Len。

表1 超参数设置

Table 1 The hyperparameters setting

图像特征向量	14×14×2048 维
词向量	512 维
依存句法向量	512 维
LSTM 隐向量	512 维
自注意力机制头数	8
批处理大小	32

本文提出的模型选择 ResNet-101 网络结构来提取图像视觉特征；词嵌入与依存句法嵌入向量均为随机初始化。训练优化器是 Adam，编码器的学习率为 0.0001，解码器的学习率为 0.0004，每 8 轮次损失未下降则使用 0.8 的学习率衰减系数。损失函数为交叉熵函数，模型的超参数设置如表 1 所示。

3.2 实验结果分析

表 2 至表 4 分别表示本文提出的模型在 Flickr30K、Flickr8K 与 Flickr8K-CN 数据集上的实验结果，其中 NIC+ATT 是基于注意力机制的神经图像描述生成模型，AdaptAtt[22]是基于自适应的图像描述生成模型，NIC+WC+WA+RL[23]是基于图像特征动态建立与图像相关词汇表的模型，MLO/MLPF-LSTM+(BS)[24]是基于深度 LSTM 的模型，CACNN-GAN(ResNet-152)[25]是基于 GAN[26]的模型。

NIC+DS 指在神经图像描述生成模型的基础上，编码阶段 LSTM 增加输入依存句法向量，而不添加图像结构注意力。

NIC+DSSA 指在神经图像描述生成模型的基础上，增加图像结构注意力。Top-5、Top-10 分别指的是内容关键词分类器按概率大小取降序排名前 5 或前 10 的标签来计算内容重合度。

表 2 Flickr30K数据集的实验结果

Table 2 The experiment results on Flickr30K dataset

Model	Flickr30k							
	B1	B2	B3	B4	M	R	C	Len
NIC+ATT(Baseline)	62.84	39.00	25.07	17.52	17.98	44.57	30.18	11.06
AdaptAtt	60.69	41.80	25.92	18.63	19.71	45.61	33.36	-
NIC+WC+WA+RL	-	-	-	24.50	21.50	51.60	58.40	-
MLO/MLPF-LSTM+(BS)	66.20	47.20	33.10	23.00	19.60	-	-	-
CACNN-GAN	69.30	49.90	35.80	25.90	22.30	-	-	-
NIC+DS(Top-5)	57.09	39.35	28.66	20.73	20.81	48.24	49.78	17.58
NIC+DSSA(Top-5)	58.62	40.46	29.81	22.62	20.96	49.98	51.74	17.56
NIC+DS(Top-10)	59.76	44.53	31.48	24.75	21.31	51.36	50.91	18.43
NIC+DSSA(Top-10)	61.81	47.33	33.97	26.06	23.57	52.81	52.48	18.62

表 3 Flickr8K数据集的实验结果

Table 3　The experiment results on Flickr8K dataset

Model	Flickr8k							
	B1	B2	B3	B4	M	R	C	Len
NIC+ATT(Baseline)	60.32	37.88	24.66	16.33	18.48	46.16	34.99	11.17
NIC+DS(Top-10)	57.76	41.16	30.70	27.78	19.54	48.81	36.69	14.47
NIC+DSSA(Top-10)	59.45	45.86	36.05	29.36	21.92	50.06	40.24	15.72

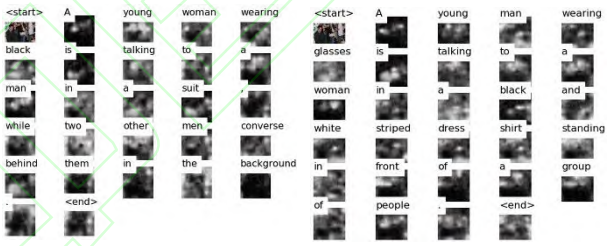
表 4 Flickr8K-CN数据集的实验结果

Table 4　The experiment results on Flickr8K-CN dataset

Model	Flickr8k-CN							
	B1	B2	B3	B4	M	R	C	Len
NIC+ATT(Baseline)	59.16	36.30	22.73	16.02	16.87	43.59	31.09	10.82
NIC+DS(Top-10)	56.28	40.03	29.42	25.61	17.48	46.21	34.16	13.45
NIC+DSSA(Top-10)	58.72	46.86	33.05	28.16	20.57	49.10	38.48	14.36

由表 2 可知，本文模型 NIC+DSSA 除 BLEU-1 指标外，其它七种指标均优于 Baseline 模型，尤其在 BLEU-4 与 CIDEr 的指标上分类改善了 8.54%和 22.30%。由此可见，依存句法的使用可以提高模型对文本与图像的结构信息的捕获能力。在 CIDEr 指标上，NIC+DSSA 略低于 NIC+WC+WA+RL 模型，由于 CIDEr 计算得分时考虑到参考文本与模型生成文本的长度差，由文本平均长度 Len 可知，本文模型 NIC+DSSA 生成的句子较长，所以在 CIDEr 指标上不占优势。CACNN-GAN 使用了 ResNet-152 网络提取图像特征，但由于在生成阶段不能很好地考虑到长距离的依赖关系，在 BLEU-4 指标上仍低于本文提出的模型。此外，由表 3 与表 4 的数据可知本文提出的模型在中英文数据集上具有一定的泛化能力。

在三个实验数据集中，NIC+DS 模型与 NIC+DSSA 模型在生成句子的平均长度接近的情况下，NIC+DSSA 在各个指标上均好于 NIC+DS，在 BLEU-4 指标上提高近 1.5%。针对图像结构注意力，本文选取 Flickr30K 数据集中的一幅测试图像构建出两种模型对应的图像注意力在测试图像上的分布图，如图 3 所示。



(a)　NIC+DS 模型　　(b)　NIC+DSSA 模型

图 3　生成文本效果对比图

Fig.3 Comparison of the generated captions

由 NIC+DSSA 模型生成的图像注意力分布图与描述文本可知，除主句中的实体外，添加了图像结构注意力的 NIC+DSSA 模型捕捉到了其它图像实体间关系（<<“young woman”, “talk”, “man”>, “converse”, <“two other men”>>），且使用“while”将主句与从句连接起来；而 NIC+DS 模型生成的文本更注重图像中某个实体细节描述，如“black and white striped dress”，而忽略了图像实体间的关系，不能很好地在句式上体现出依存句法添加至模型后的优势。为了控制其它变量的影响，两者均输入相同的依存句法模板。

给定测试图像，本文提出的模型应尽可能地提供符合图像内容实体之间实际句法关系的依存句法模板。在内容关键词分类器中，若 K 值不同，模型为测试图像选择的依存句法模板可能不同。因此本文研究了 K 取值对 BLEU-4 指标的影响，如图 4 所示。

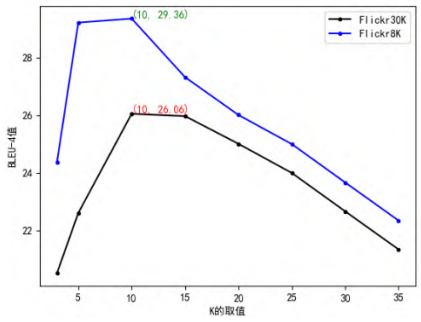


图 4 不同 K 值选取对实验结果的影响
Fig.4 The experiment results affected by different K

由图 4 可知，在 K 值为 10 时，模型在实验中得到了最好的描述文本生成效果，BLEU-4 值最高。K 取值较小时，分类器选择的图像描述标签较少，不能完整地概括一幅图像；K 取值较大时，标签会同图像主题偏离而降低生成描述文本的准确性。无论 K 值过大还是过小，都会导致“假重合”情况发生，故本文 K 值为 10。

当 K 值为 10 时，每个关键词标签可以较准确地描绘图像的不同侧面，当两幅图像的内容重合度较高时，描述文本中实体或事件高度相似。下面给出了当 K 值为 10 时，计算两幅图像内容重合度的例子（此例内容重合度最大为 55），如图 5 所示。

根据内容重合度，测试图像可以在训练图像集中匹配到与自身内容最相似的图像，由于在训练集中，每幅图像都有五个参考文本与之对应，因此本文提出的模型可以将对应的五个依存句法模板依次输入到模型中，指导模型生成描述文本，在一定程度上增加了生成句子的多样性。

图 5 K 取 10 时分类效果图
Fig.5 The results of classification when K is 10

如图 6 所示，本文提出的 NIC+DSSA 模型可以根据不同的依存句法模板生成不同的描述文本，且生成的文本（5）与参考文本（5）在句式上保持一致。

图 6 模型生成的文本多样性示例
Fig.6 The example of the diversity of captions generated by the model

3.3 深度模型可解释性分析

本节将进一步细粒度地分析本文提出的模型如何根据输入图像与依存句法信息生成描述文本，使深度学习模型更具可解释性。

由图 7（a）可知，NIC+ATT 模型在生成文本句的主体时，图像关注的区域较分散，不能很好地区分每个单词或词组对应图像区域的异同。而 NIC+DSSA 模型由于加入了依存句法信息中的结构关系，生成的每个单词或词组关注的图像区域可以清晰地分辨出差异，并具有一定的连续性，如图 7（b）所示。

(a) NIC+ATT 模型　　(b) NIC+DSSA 模型

图 7 图像注意力对比图
Fig.7 The comparison of the image attention

在图 7 所示的两个生成文本中，图 7(a)的 NIC+ATT 模型在生成图像注意力时，特别在词组“are gathered in”对应的图像区域中，关注的图像区域居中、分布较广且高度相似，因此不能很好地捕获图像区域内部实体间关系，即在此图中仅识别出“parade”；而图 7(b)的 NIC+DSSA 模型，其在生成图像注意力时，不同的图像区域间存在一定的差异性，而仅在描述事物的名词词组“brown shirt”、动词词组“standing up”或介词词组“the middle of”对应的图像区域分布较类似，符合人类分析图像的过程。

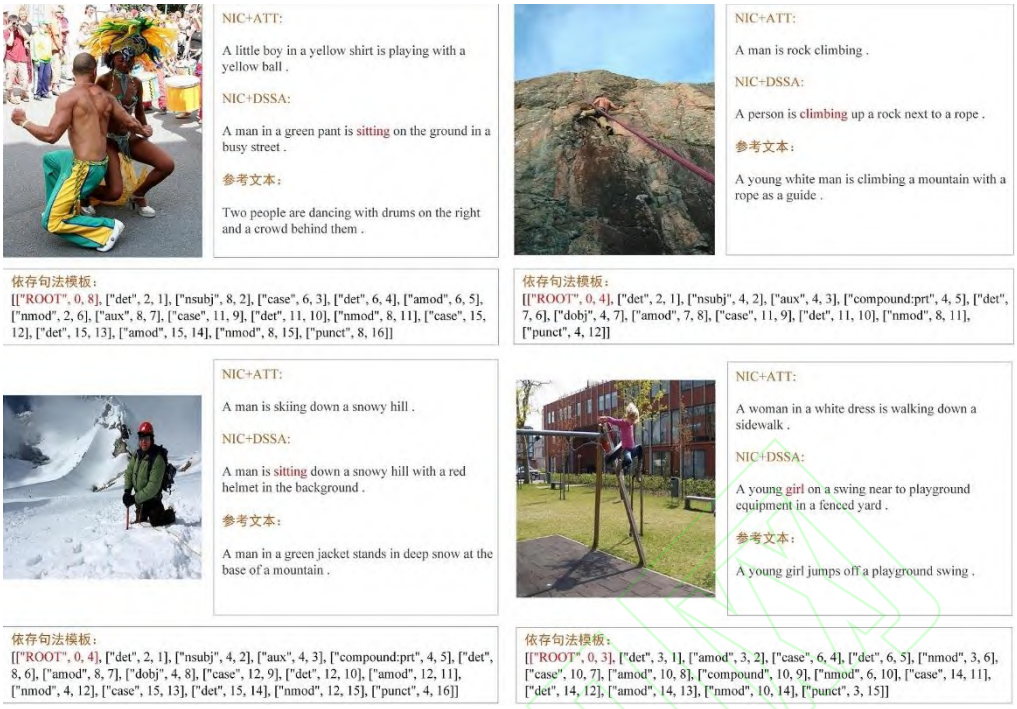


图 9 基于依存句法的图像描述文本生成示例图

Fig.9 The examples of image captioning based on dependency syntax

如图 8 所示，根据相似图像，模型将得到五个依存句法模板。在图 8 显示的图 7 文本对应的依存句法模板中，第一个依存句法单元代表生成的文本在位置 8 上的单词“standing”应为该文本句的词根“ROOT”；第二个依存句法单元代表生成的文本在位置 1 上的单词“A”应为在位置 2 上的单词“man”的限定词“det”；第三个依存句法单元则代表生成的句子在位置 2 上的单词“man”应为在位置 8 上的单词“standing”的名词主语“nsubj”。由图 8 可知，输入的依存句法通过图像结构注意力间接指导了描述文本生成，且将图像的区域间结构关系与文本中的依存关系对应。图 9 给出了本文模型 NIC+DSSA 在 Flickr8K 部分图像上的实验结果，体现了依存句法在模型生成描述文本中的指导作用。

图 8 相似图像和依存句法模板

Fig.8 The similar image and dependency syntactic template

为分析模型生成文本时子句的使用情况，本文统计了测试阶段模型在 Flickr30K 数据集上生成描述文本连接词的使用情况，如表 5 所示，本文提出的模型由于依存句法信息的加入，在生成连接词方面表现能力较突出，很好地解释了其生成文本较长这一现象。

表 5 描述文本中连接词数量统计

Table 5 The statistic of conjunctions in captions

模型	“6w 1h 1t”数量
NIC	1
NIC+DSSA	66
参考文本	58

4　结束语

本文提出了一种基于依存句法的图像描述文本生成模型，将依存句法作为语法规则输入至图像结构注意力中，使模型在生成描述文本的过程中考虑到依存句法提供的词与词之间的依赖关系。此外，本文验证了通过使用图像结构注意力加强模型对图像中实体间关系的学习理解能力。本文提出了一种通过提取图像内容关键词来计算图像内容重合度的方法，建立测试图像与训练图像集的映射关系，进而将测试图像与依存句法模板匹配。在 Flickr30K、Flickr8K 和 Flickr8K-CN 三个数据集上进行的大量实验表明，本文提出的模型在增加文本多样性和句法复杂度两个方面均有较好的表现。此外，本文还通过模型生成描述文本的实例分析，解释了依存句法在本文提出的模型中如何指导文本生成及关注图像结构特征，增强了深度学习模型的可解释性。

在未来的工作中，仍有两个问题需要持续关注：（1）依存句法表示问题，目前使用的自注意力机制虽然可以较好地得到长距离的依赖关系表示，但其可解释性不高；（2）依存句法模板获取问题，本文提出的内容重合度计算方法旨在根据“ROOT”中心词识别内容相似的图像，但忽略了图像内容的深层含义如图像主题[27]等，有时无法得到最优的依存句法模板。

参 考 文 献（References）

[1] XU K, BA J, KIROS R, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[C]//Proceedings of the 32nd International Conference on Machine Learning. 2015:2048-2057.

[2] LU C, KRISHNA R, BERNSTEIN M S, et al. Visual Relationship Detection with Language Priors[C]//Proceedings of the 14th European Conference on Computer Vision. 2016: 852-869.

[3] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Proceedings of Advances in Neural Information Processing Systems 28. 2015: 91−99.

[4] GUO Y, CHENG Z, NIE L, et al. Quantifying and alleviating the language prior problem in visual question answering[C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019: 75-84.

[5] ADIYA D, JYOTI A, LIWEI W, et al. Fast, Diverse and Accurate Image Captioning Guided By Part-of-Speech[C]//Proceedings of the 32nd IEEE Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE Press, 2019: 10695-10704.

[6] WANG, Y, LIN Z, S X, et al. Skeleton key: Image captioning by skeleton-attribute decomposition[C]//Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE Press, 2017: 7272–7281.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the Advances in Neural Information Processing Systems 30. 2017: 5998–6008.

[8] SUTSKEVER I, VINYALS O, LE Q V. Sequence to Sequence Learning with Neural Networks[C]//Proceedings of the Advances in Neural Information Processing Systems 27. 2014: 3104-3112.

[9] VINYALS O, TOSHEV A, BENGIO S, et al. Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 39(4): 652-663.

[10] ZHU Z, XUE Z, YUAN Z. Topic-Guided Attention for Image Captioning [C]//Proceedings of the 25th IEEE International Conference on Image Processing. 2018: 2615-2619.

[11] WANG T, HU H, HE C. Image Caption with Endogenous–Exogenous Attention[J]. Neural Processing Letters, 2019, 50(1): 431-443.

[12] LIU F, LIU Y, REN X, et al. Aligning Visual Regions and Textual Concepts: Learning Fine-Grained Image Representations for Image Captioning[J]. arXiv preprint arXiv:1905.06139, 2019.

[13] FALENSKA A, KUHN J. The (Non-)Utility of Structural Features in BiLSTM-based Dependency Parsers[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, 2019: 117-128.

[14] LI Z, PENG X, ZHANG M, et al. Semi-supervised Domain Adaptation for Dependency Parsing[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, 2019: 2386-2395.

[15] WANG X, TU Z, WANG L, et al. Self-Attention with Structural Position Representations[J]. arXiv preprint arXiv:1909.00383, 2019.

[16] HE K , ZHANG X , REN S , et al. Deep Residual Learning for Image Recognition[C]//Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[17] CHRISTOPER D. M, MIHAI S, JOHN B, et al. The Stanford CoreNLP Natural Language Processing Toolkit[C]//Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, 2014: 55–60.

[18] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, 2002: 311–318.

[19] CHOUEIRI T K, ESCUDIER B, POWLES T, et al. Cabozantinib versus everolimus in advanced renal cell carcinoma (METEOR): final results from a randomised, open-label, phase 3 trial[J]. The Lancet Oncology, 2016, 17(7): 917-927.

[20] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]//Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, 2004: 74-81.

[21] VEDANTAM R, ZITNICK C L, PARIKH D. Cider: Consensus-based image description evaluation[C]//Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition. 2015:4566–4575.

[22] LU J, XIONG C, PARIKH D, et al. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning[C]//Proceedings of the 30th Conference on Computer Vision and Pattern Recognition. 2017: 3242-3250.

[23] FAN Z, WEI Z, HUANG X, et al. Bridging by Word: Image Grounded Vocabulary Construction for Visual Captioning[C]//Proceedings of the International Conference on the Association for the Advance of Artificial Intelligence. New York, USA. 2019: 6514-6524.

[24] 汤鹏杰，王瀚漓，许恺晟． LSTM 逐层多目标优化及多层概率融合的图像描述． 自动化学报， 2018， 44（7）： 1237-1249．
TANG P J, WANG H L, XU K S. Multi-objective Layer-wise Optimization and Multi-level Probability Fusion for Image Description Generation Using LSTM[J]. Acta Automatica Sinica, 2018, 44(7): 1237-1249 (in Chinese).

[25] 薛子育，郭沛宇，祝晓斌，张乃光． 一种基于生成式对抗网络的图像描述方法． 软件学报， 2018， 29（2）： 30-43．
XUE Z Y, GUO P Y, ZHU X B, et al. Image Description Method Based on Generative Adversarial Networks[J]. Journal of Chinese Information Processing, 2018, 29(2): 30-43(in Chinese).
[26] SALVARIS M, DEAN D, TOK W H, et al. Generative Adversarial Networks[J]. arXiv: Machine Learning, 2018: 187-208.
[27] LIU M, HU H, LI L, et al. Chinese Image Caption Generation via Visual Attention and Topic Modeling[J]. IEEE Transactions on Cybernetics, 2020.