Computer Science and Application 计算机科学与应用, 2020, 10(6), 1087-1097
Published Online June 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.106113

Hans汉斯

Image Description Generation in Chinese Based on Keywords Guidance

Xiucong Shi

Faculty of Computer, Guangdong University of Technology, Guangzhou Guangdong
Email: xiuc_shi@163.com

Received: May 15th, 2020; accepted: May 28th, 2020; published: Jun. 4th, 2020

Abstract

Technology for generating image description has a wide range of applications as it can speed up the production of graphic content. To meet the practical requirements, we propose a new method based on the encoder-decoder framework. Our model guides the generation of image description by fusing image and text features as input. The text features contain the semantics of keywords of images, as a supplement to image information. The experimental results show that keyword information enhances the mapping from image to image description. The model in this paper has better performance than the model without fusing keyword information, and different keyword information has certain control over the generation of image description.

Keywords

Image Description Generation, Encoder-Decoder Framework, Keyword Information, Multimodal Fusion

基于关键词指导的图像中文描述生成

史秀聪

广东工业大学计算机学院，广东 广州
Email: xiuc_shi@163.com

收稿日期：2020年5月15日；录用日期：2020年5月28日；发布日期：2020年6月4日

摘 要

图像描述生成技术可以加速图文内容的生产，因而有着广泛的应用前景。为了满足实际需要，我们提出

文章引用：史秀聪. 基于关键词指导的图像中文描述生成[J]. 计算机科学与应用, 2020, 10(6): 1087-1097.
DOI: 10.12677/csa.2020.106113

了一种基于编码－解码框架的新方法。我们的模型通过融合图像和文本特征作为输入来指导图像描述的生成，文本特征包含图像的关键词信息作为图像特征的补充。实验结果表明，关键词信息加强了图像到图像描述的映射，本文模型比未融合关键词信息的模型具有更好的性能，并且不同的关键词信息对图像描述的生成有一定的控制作用。

关键词

图像描述生成，编码－解码框架，关键词信息，多模态融合

1. 引言

随着互联网的发展，图文内容已经成为互联网上的主要表现形式。图文内容的呈现形式可以提高网民的阅读体验。汽车门户网站通常会发布汽车产品信息来吸引网民的关注，从而促进产品的推广。这些内容通常由许多图像和相应的描述组成，而目前这些图像描述通常由人工进行撰写，然后发布到互联网上。对于大量图文信息需求量的情况，手动撰写描述是一项耗时且枯燥的工作。因此，图像描述自动生成技术具有很大的实用价值。它可以加速图文内容的生产和发布，从而吸引人们的关注。近年来，由于在神经机器翻译的成功应用[1]，编码－解码框架被应用于图像描述生成技术中。

与基于 MSCOCO [2]和 Flickr30k [3]等数据集的研究不同，本文基于实际应用为汽车图像生成相应的中文描述句子。在我们的应用需求中，输入一组图像，然后为这组图像生成描述语句。对于相似的图像，我们希望生成倾向不同侧重点的描述句子，以实现内容的多样性。因此对于同一幅图像，我们期望控制其描述的侧重点。

考虑到我们数据集的特点，许多相似的图像可能具有不同侧重点的图像描述。例如，两张相似的图片(a)和(b)，图片(a)的描述是“发动机并没有显得多么暴躁，风格沉稳舒适。”，图片(b)的描述是“变速箱很聪明，为新车的行驶质感添色不少。”。它们具有完全不一样的侧重点的描述。

针对上述情况，我们需要研究为同一张图像生成不同侧重点的图像描述句子的方法。在实际应用中面临的挑战是如何确定图像的可能存在的侧重点和如何引导描述生成过程从而在相应的侧重点生成描述句子。

为了解决这个问题，我们提出了一种新的方法，该方法利用多模态融合的思想，将文本信息和图像信息融合作为方法的输入。文本信息是图像信息的补充，用于指导模型往不同的侧重点生成描述句子。本文的主要贡献如下：

1) 提出了一种新的图像描述生成方法，该方法将关键词文本信息和图像信息相融合作为图像描述的输入，通过关键词信息来指导模型往不同的侧重点生成图像描述。

2) 我们使用自行开发的爬虫采集的真实数据集进行了实验。该数据集由 2100 个图像－关键词－描述对组成，每个图像对应一个中文描述和关键词列表。

2. 相关工作

近年来，随着深度学习的发展，编码－解码框架被广泛应用于图像描述生成领域，它将图像编码成

视觉特征图，然后再由解码器将提取到的图像特征解码成自然语言句子。然而，基础的编码－解码框架仅捕获整个图像的全局特征，而在解码端每个步骤生成的单词通常与图像的特定区域相关[4] [5]。为了解决图像特征和词语的对齐问题，学者在编码－解码框架中引入了注意力机制[6]-[14]。例如，文献[7]提出了“软注意机制”和“硬注意机制”两种类型的注意力机制，使得解码器能够在每个时间步注意到图像不同的区域特征。其他一些研究也集中在视觉注意力机制上。然而，这些视觉注意机制通常没有考虑图像和文本之间的任何相关性。文献[12]提出了一种基于文本的视觉注意模型，该模型通过消除由先前生成的文本表示的显著对象来消除无关信息。注意力机制能够使得模型在每个时间步关注不同区域的特征，以便在解码器每个时间步对齐视觉和文本特征。

除了引入注意力机制外，很多研究基于编码器或解码器进行了改进。例如，文献[14]使用多实例学习(Multiple instance learning, MIL)来训练一个视觉单词检测器，其输出被视为最大熵语言模型的条件输入。文献[15]提出了一种新的特征提取方法，通过目标检测算法将图像特征提取为低层特征，由图像特征训练的属性检测器将属性特征提取为高层特征。文献[16]提出了一种由粗到细的方法，将原始图像描述分解为骨架句子及其属性，然后分别生成骨架句子和属性短语。

一些方法直接针对模型的评价指标对模型进行优化。由于这些指标通常是不可微的，因此几乎所有这些方法都是通过使用强化学习进行优化的[17] [18]。

受到多模态数据融合的启发，我们的模型使用图像和文本特征作为输入，通过文本特征作为图像特征的补充可以加强图像到图像描述的映射，根据不同的文本特征指导图像描述的侧重点。

3. 关键词信息指导的图像描述生成模型

3.1. 模型总体结构

图 1 和图 2 展示了我们提出的模型的结构。图 1 是训练阶段的模型结构，图 2 是测试阶段的模型结构。在训练阶段，模型的输入包括图像和关键词信息，其中图像和关键词信息都是从训练集中直接获得。而测试阶段仅仅输入图像，关键词信息由关键词模块根据输入图像预测获得。因此总体模型大致可分为图像编码器、关键词预测模块、文本编码器和解码器四个模块。

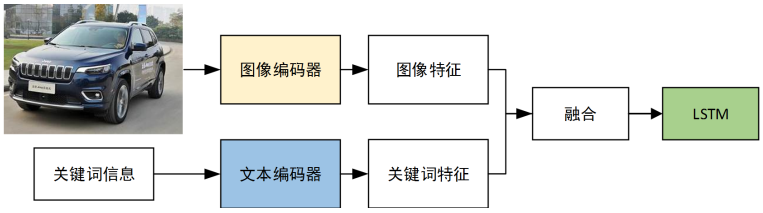

Figure 1. The structure of the model during training
图 1. 训练阶段的模型结构

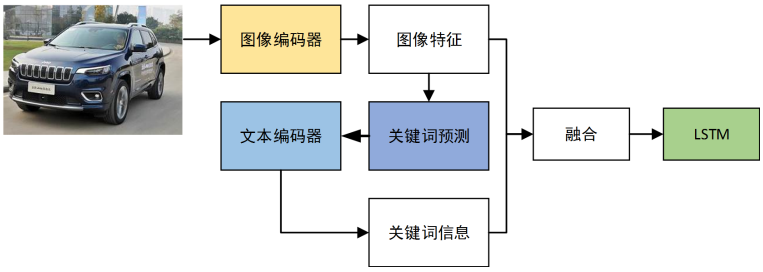

Figure 2. The structure of the model during testing
图 2. 测试阶段的模型结构

图像编码器：图像编码器主要对输入的图像进行特征提取。得益于深度卷积神经网络强大的特征提取能力，我们使用深度卷积神经网络作为图像编码器。我们选择了 VGG-16 [19]模型的最后一个卷积层的输出特征图作为图像的特征，其维度为 $W \times H \times D$，然后将其转化成 $L \times D$。L 是图像特征的区域个数，每个区域的特征由一个维度为 D 的向量表示。

关键词预测器：关键词预测器用于预测输入图像的关键词信息，从而在测试阶段判断输入图像可能的关键词信息。在测试阶段，对于关键词信息的输入，如果人为地给出关键词信息，可能会造成关键词信息与图像不相关的情况。因此，有必要确定图像中可能存在的关键词信息。

文本编码器：文本编码器主要用于提取关键词文本信息的特征。关键词信息是由 1~4 个中文词语组成的文本信息。关键词信息包含了图像描述语句中的一些关键信息。每一个词语由文本编码器编码成一个 D 维词向量。然后，我们以按位相加的方式将这些词语的词向量相加作为关键词信息的特征表示。

解码器：解码器用于将编码器获得的特征向量解码成自然语言句子。解码器由一个注意力机制(Attention mechanism)和循环神经网络(Recurrent neural network, RNN)组成。其中，注意力机制由一个单层的全连接神经网络实现。RNN 被广泛应用于自然语言处理领域，它比较擅长处理时间序列问题。然而，传统的 RNN 模型会出现梯度消失的问题。因此，我们部署了长短时记忆神经网络(Long-short term memory, LSTM) [20]作为将特征解码成自然语言句子。LSTM 引入了一个核心单元 Cell，Cell 的状态由三个门控制，分别为遗忘门、输入门和输出门。通过三个门控制历史信息的遗忘和保留，其公式如(1)~(5)所示。

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right), \tag{1}$$

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right), \tag{2}$$

$$C_t = f_t * C_{t-1} + i_t * \tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right), \tag{3}$$

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right), \tag{4}$$

$$h_t = o_t * \tanh\left(C_t\right), \tag{5}$$

其中，f_t，i_t，o_t 分别表示 t 时刻遗忘门、输入门和输出门的状态；h_t 表示 t 时刻模型的隐藏状态；c_t 表示 t 时刻 Cell 状态的更新；$\sigma(\cdot)$ 是 Sigmoid 激活函数；W 是可训练的模型参数矩阵；b 是偏置值。

由于引入了注意力机制，在 LSTM 的每个时间步都会产生不同的上下文向量。给定一张图像和它的关键词信息，图像由卷积神经网络提取特征，记为 $I\{I_1, \cdots, I_L\}, I_i \in R^D$。关键词特征由文本编码器提取，记为 T。

关键词特征作为图像信息的补充，我们将关键词特征拼接到图像的每个区域特征后面，然后将其输入到解码端，如图 3 所示。

注意力机制被用于在每个时间步结合 LSTM 的上一个隐藏状态计算输入特征的不同区域的关注度，公式如(6)~(8)所示。

$$e_{ti} = f_{att}\left(\left[I, T\right]_i, h_{t-1}\right), \tag{6}$$

$$\alpha_{ti} = \mathrm{softmax}\left(e_{ti}\right), \tag{7}$$

$$z_t = \sum_{i=1}^{L} \alpha_{ti} \cdot \left[I, T\right]_i, \tag{8}$$

其中 e_{ti} 是 t 时刻注意力机制的输出；α_{ti} 是 t 时刻输入特征的区域 i 的权重；z_t 是 t 时刻输入到 LSTM 模型的上下文向量。上下文向量将结合当前时刻的词语嵌入向量一同输入到 LSTM 模型。在每个时间步，LSTM 模型的隐藏状态 h_t 会被输入到一个 softmax 函数中，然后产生一个所有词语的概率分布，概率最高的词语则为当前时间步生成的词语。

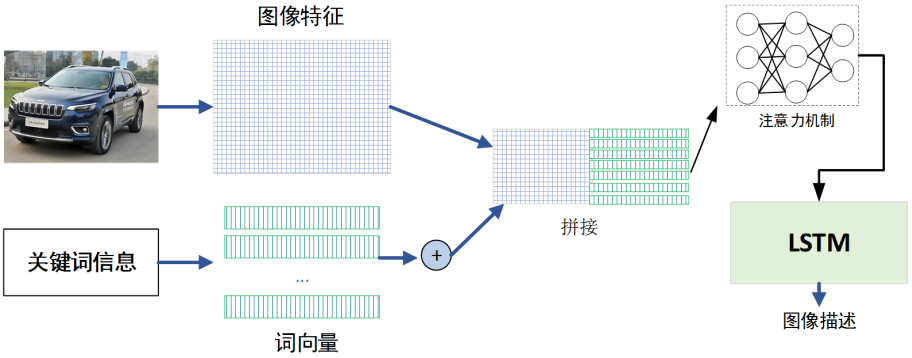

Figure 3. Image and text feature fusion
图 3. 图像和文本特征融合

3.2. 训练和测试过程

在训练阶段，我们采用了分步的训练方式。图像编码器和关键词编码器都是单独训练，然后再通过编码器对输入数据进行特征提取来训练 LSTM 模型。总体大致流程分为图像预处理、特征提取、图像－文本特征融合和梯度下降优化模型参数。

图像预处理：图像预处理包括图像裁剪和归一化处理。图像编码器要求输入图像的大小一致，因此需要对图像进行尺寸的统一。此外，归一化可以加快模型的收敛速度。

特征提取：特征提取包括图像特征提取和关键词特征提取，我们通过图像编码器对图像特征进行提取，文本编码器对关键词信息进行特征提取。

图像－文本特征融合：由于输入特征包含图像特征和关键词文本特征。在对图像特征和关键词特征进行提取后，我们将图像特征和关键词特征进行拼接融合作为解码器的输入。

梯度下降：为了训练模型，我们采用了有监督的学习方法。在 LSTM 生成下一个词语的时候，模型会根据当前输入 x 和历史状态 S_{t-1} 做出动作 a_t 的概率分布 $p(a_t|x, s_{t-1})$。在训练过程中，梯度下降算法通过最小化交叉熵损失函数对模型参数进行更新，公式如(9)所示。

$$\log p\left(S|x\right)=\sum_{t=0}^{N}\log p\left(S_t|x, S_0,\cdots,S_{T-1};\theta\right) \tag{9}$$

其中 x 是模型的输入；S 是图像描述句子；θ 是模型的参数。我们使用梯度下降算法对模型参数进行更新。

在测试阶段，关键词信息不是直接从训练集中获取，这不同于训练阶段。在测试阶段，输入数据仅有图像，关键词信息通过关键词预测模块确定测试图像可能存在的关键词信息。产生的候选关键词分别结合图像信息输入到模型生成描述句子。

4. 实验结果与分析

4.1. 评价指标

实验结果通过双语互评辅助工具(Bilingual Evaluation Understudy, BLEU) [21]、ROUGE [22]和 CIDEr [23]进行评估。BLEU 的分数通过计算生成句子和参考句子之间的 n-gram 匹配精度得到。我们使用 uni-gram、bi-gram、tri-gram 和 4-gram 计算 BLEU 评分，分别标记为 BLEU-1、BLEU-2、BLEU-3 和 BLEU-4。ROUGE 是一种基于 n-gram 共现信息的面向召回的评价方法。它有一系列的评价方法，包括 ROUGE-N (N = 1, 2, 3, 4)、ROUGE-L、ROUGE-S 等。我们选择 ROUGE-L 来评估我们的结果，因为它基于最长的公共子序列，并且适合于短句。此外，CIDEr 是专门为图像标记问题而设计的。该指标将每个句子视为一个“文档”，并将其表示为一个术语频率逆文档频率(Term Frequency Inverse Document Frequency, TF-IDF)向量。

通过计算每个 n-gram 的权重，计算参考句子与模型生成的句子的余弦相似度来度量图像标签的一致性。

4.2. 数据集

我们在一个汽车图像数据集上对提出的方法进行了实验。我们开发了一个爬虫程序采集了一些图像数据，原始图像格式如图 4 所示。

Figure 4. Raw data format
图 4. 原始数据格式

每张图像上带有对图像的描述，我们对这些图像进行了重写以避免太长和口语化表述，然后对图像进行截取以使得图像不再附有文本描述。经过处理后，我们获得了一个“图像－描述句子”映射关系的数据集。对于关键词信息的提取，我们将数据集中所有的描述句子合并成一个文档，然后使用 Jieba [24] 工具库中的关键词提取接口提取了该文档的关键词。关键词按照权重由高到低排列，然后我们遍历了每个描述句子中的词语以确定词语是否是关键词。通过处理，我们为图像确定了其描述的关键词，关键词是一个集合，由 1~4 个词语组成。

对数据进行处理后，我们构建了一个“图像－关键词－描述句子”映射关系的数据集。数据集的大小为 2100 条数据，我们随机选择 2000 条作为训练集，100 条作为测试集。由于图像到描述的映射关系较弱，即使生成的描述质量很好，通顺流畅，如果和测试图像的参考描述差别很大，那模型性能很难在评估结果得到体现，因此本文在与测试图像最相似的 20 张图像的描述句子中随机选择 5 条作为测试图像的参考描述。这些信息写入了一个 JSON 文件中，其结构如下所示：

{
images: [
{"file_name": [file name], "height": [height of image], "width": [width of image], "id": [image ID], "keywords": [keywords of image]}, ...
],
annotations: [
{"image_id": [image ID], "id": [description ID], “description ": [Chinese description]}
], ...
}

4.3. 图像编码器

在本文中，我们使用了在 ImageNet [25]图像数据集上进行了预训练的 VGG-16 作为图像编码器。我

们选择了最后一个卷积层的输出作为图像的特征。其维度为 14 × 14 × 512。512 是特征图的通道数目，14×14 是特征图的尺寸的大小，分别表示特征图的高度和宽度。

4.4. 关键词预测

对于测试图像关键词预测，我们通过图像检索的方式实现。通过 VGG-16 模型提取了测试图像和训练数据集中的图像的特征。其中，VGG-16 和图像编码器是同一个模型。然后将图像的特征转化成向量表示，通过计算测试图像的特征向量和训练数据集中图像的特征向量的余弦相似度，找到训练数据集中前 10 个和测试图像最相似的图像，将这 10 个图像的关键词信息作为测试图像的候选关键词信息。余弦相似度计算公式如(10)所示。

$$\cos(\theta)=\frac{I_1 I_2}{\|I_1\|\|I_2\|}=\frac{\sum_{k=1}^{n} i_{1k}\cdot i_{2k}}{\sqrt{\sum_{k=1}^{n} i_{1k}^2}\sqrt{\sum_{k=1}^{n} i_{2k}^2}} \tag{10}$$

其中 I_1 和 I_2 是图像特征的向量表示，记为 $I_1=\{i_{11},\cdots,i_{1n}\}$ 和 $I_2=\{i_{21},\cdots,i_{2n}\}$。

图 5 展示了图像#(183)和#(174)的检索结果。最上方的一张图像是测试图像，中间区域的 10 张图像是检索得到的与测试图像最相似的图像，最下方是 10 张图像对应的关键词信息。结果表明，VGG-16 模型能够准确提取图像的特征，通过余弦相似度计算的方式能够有效检索出相似的图像从而确定测试图像的关键词信息。

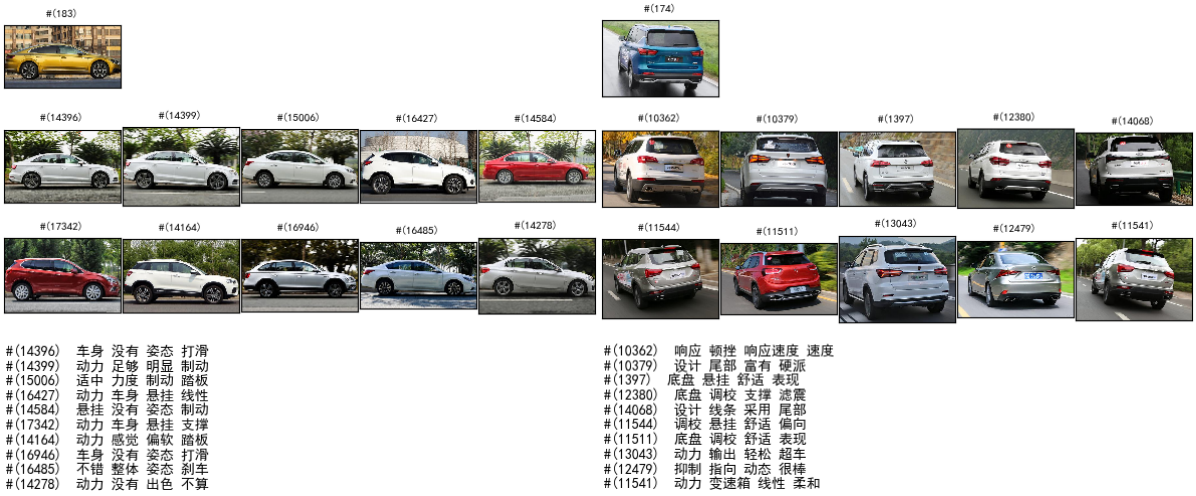

Figure. 5. Retrieval results of image #(183) and #(174)
图 5. 图像#(183)和#(174)的检索结果

4.5. 文本编码器

对于词语的语义表示，一种有效的方式是将词语映射到高维度的词向量。所有这些词向量构成一个词向量空间，词向量之间的余弦距离可以反映词语之间的语义相似度。因此，连续词袋模型(Continuous Bag-of-Words, CBOW) [26]可以满足要求。通过在大规模的语料库中训练词向量模型，可以使得词向量模型包含丰富的语义信息从而可以充分表示文本信息。

我们利用 Genism [27]工具库部署了一个词向量模型用做文本编码器。我们设置了窗口大小为 5，词向量维度为 512。然后在维基百科中文数据集上对词向量模型进行了训练。通过训练后，我们获得了一个词向量模型，该模型可以将词语表示为一个 512 维的向量。为了观察模型的性能，我们测试了与词语

“座位”最相似的词语，如图 6 所示。

```
model.most_similar('座椅')
('座垫', 0.6821202039718628)
('椅背', 0.6746518015861511)
('椅套', 0.6726085543632507)
('椅垫', 0.6525329947471619)
('排座位', 0.6487257480621338)
('坐椅', 0.6456358432769775)
('双前座', 0.6343128681182861)
('仿皮', 0.6216391921043396)
('驾驶座', 0.6194170713424683)
('躺椅', 0.6186153888702393)
```

Figure. 6. Word distribution similar to “seat”
图 6. 与“座椅”语义相似的词语分布

此外，为了观察词向量的分布情况，我们随机选择 100 个词语的词向量表示通过 TSNE 工具降维可视化，如图 7 所示。

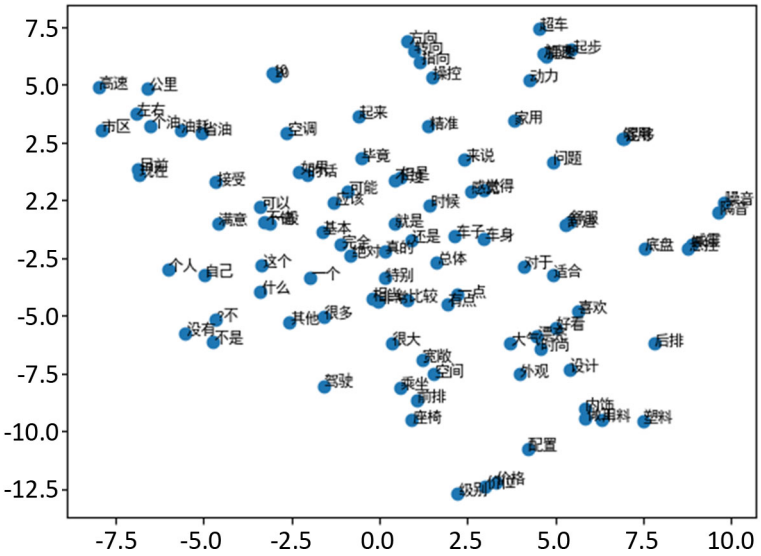

Figure. 7. Word vector visualization
图 7. 词向量可视化

图 6 显示词向量模型可以有效找出与“座椅”相似的词语，图 7 中语义相近的词语，如“方向”、“指向”、“转向”等词语集中在很近的位置。结果显示，词向量模型可以很好地对词语进行表征。

4.6. 图像描述生成

训练完编码器后，再结合解码器进行训练。解码器的主要模型是 LSTM，其目的是将编码器得到的上下文向量解码成自然语言句子。对于 LSTM 模型的训练，我们做出了如下处理。首先，将所有的描述句子进行分词处理，构建了一个包含 2561 个词语的词典，每个词都映射到一个整型的数字，表示该词在词典中的位置。在模型中，对于词语的表示，采用 2561 × 512 维的嵌入矩阵表示，每个词语由

一个 512 维的词向量表示，该嵌入矩阵使用均匀分布初始化器初始化，然后在模型训练的过程中不断优化。LSTM 的隐藏单元的维度设置为 512，隐藏层层数为 1。初始学习率设置为 0.001，采用 Adam 梯度下降优化算法对模型参数进行更新。我们在初始学习率为 0.001 时迭代了 5000 次，然后使用学习率 0.0005 迭代了 4000 次。我们在 NIC 和 Soft Attention 方法上做了对比实验。本文模型采用了 10 组关键词分别结合图像作为输入，然后使用 BLEU、Rouge-L 和 CIDEr 评价方法对模型进行评价。NIC 方法是基础的编码－解码模型，Soft Attention 方法是在 NIC 方法的基础上引入了软注意力机制。评估结果如表 1 所示。

Table 1. Evaluation of experiments on BLEU-n (n = 1, 2, 3, 4), ROUGE-L and CIDEr
表 1. 模型在 BLEU-n (n = 1, 2, 3, 4)，ROUGE-L 和 CIDEr 上的评估结果

模型	B@1	B@2	B@3	B@4	ROUGE-L	CIDEr
NIC	0.311	0.212	0.161	0.075	0.076	0.337
Soft attention	0.330	0.214	0.153	0.079	0.092	0.342
本文模型	**0.418**	**0.287**	**0.165**	**0.107**	**0.153**	**0.394**

由表 1 结果可以看出，本文模型在各个评估指标上的评估结果比 NIC、Soft Attention 的性能好。通过引入关键词信息，可以加强图像到图像描述的映射。图 8 展示了同一张图像在不同关键词信息下的描述情况，结果显示不同的关键词对图像描述的侧重点产生了一定的作用。

Figure 8. Effect of different keyword information on image description
图 8. 不同关键词信息对图像描述的影响

5. 总结

本文提出了一种将图像和关键词信息一起输入从而生成图像描述句子的新方法。根据实验结果，本文模型的性能比 NIC 和 Soft Attention 模型要好，能够生成流畅通顺的图像描述句子，并且同一张图像结合不同的关键词信息可以控制描述的侧重点，一定程度上增加了图像描述的多样性。虽然我们取得了一定的进展，但还存在一些问题，数据集不够大，生成的描述句子偏短。未来我们会扩充数据集，同时对模型进行优化以获取更好的性能从而满足实际的应用需求。

参考文献

[1] Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., *et al.* (2017) Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, **5**, 339-351. https://doi.org/10.1162/tacl_a_00065

[2] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., *et al.* (2014) Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B. and Tuytelaars, T., Eds., *European Conference on Computer Vision*, Springer, Cham, 740-755. https://doi.org/10.1007/978-3-319-10602-1_48

[3] Flickr Image Dataset. Kaggle.com. https://www.kaggle.com/hsankesara/flickr-image-dataset

[4] Vinyals, O., Toshev, A., Bengio, S. and Erhan, D. (2015) Show and Tell: A Neural Image Caption Generator. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 7-12 June 2015, 3156-3164. https://doi.org/10.1109/CVPR.2015.7298935

[5] Karpathy, A. and Li, F.-F. (2015) Deep Visual-Semantic Alignments for Generating Image Descriptions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 7-12 June 2015, 3128-3137. https://doi.org/10.1109/CVPR.2015.7298932

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., *et al.* (2017) Attention Is All You Need. In: *Advances in Neural Information Processing Systems*, 5998-6008.

[7] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., *et al.* (2015) Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *International Conference on Machine Learning*, June 2015, 2048-2057.

[8] You, Q., Jin, H., Wang, Z., Fang, C. and Luo, J. (2016) Image Captioning with Semantic Attention. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 27-30 June 2016, 4651-4659. https://doi.org/10.1109/CVPR.2016.503

[9] Lu, J., Xiong, C., Parikh, D. and Socher, R. (2017) Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 21-26 July 2017, 375-383. https://doi.org/10.1109/CVPR.2017.345

[10] Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W. and Chua, T.S. (2017) SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 21-26 July 2017, 5659-5667. https://doi.org/10.1109/CVPR.2017.667

[11] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S. and Zhang, L. (2018) Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 18-23 June 2018, 6077-6086. https://doi.org/10.1109/CVPR.2018.00636

[12] He, C. and Hu, H. (2019) Image Captioning with Text-Based Visual Attention. *Neural Processing Letters*, **49**, 177-185. https://doi.org/10.1007/s11063-018-9807-7

[13] He, X., Yang, Y., Shi, B. and Bai, X. (2019) VD-SAN: Visual-Densely Semantic Attention Network for Image Caption Generation. *Neurocomputing*, **328**, 48-55. https://doi.org/10.1016/j.neucom.2018.02.106

[14] Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., *et al.* (2015) From Captions to Visual Concepts and Back. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 7-12 June 2015, 1473-1482. https://doi.org/10.1109/CVPR.2015.7298754

[15] Li, N. and Chen, Z. (2018) Image Cationing with Visual-Semantic LSTM. *IJCAI*, July 2018, 793-799. https://doi.org/10.24963/ijcai.2018/110

[16] Wang, Y., Lin, Z., Shen, X., Cohen, S. and Cottrell, G.W. (2017) Skeleton Key: Image Captioning by Skeleton-Attribute Decomposition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 21-26 July 2017, 7272-7281. https://doi.org/10.1109/CVPR.2017.780

[17] Ren, Z., Wang, X., Zhang, N., Lv, X. and Li, L.J. (2017) Deep Reinforcement Learning-Based Image Captioning with Embedding Reward. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 21-26 July 2017, 290-298. https://doi.org/10.1109/CVPR.2017.128

[18] Zhang, L., Sung, F., Liu, F., Xiang, T., Gong, S., Yang, Y. and Hospedales, T.M. (2017) Actor-Critic Sequence Training for Image Captioning. arXiv preprint arXiv:1706.09601

[19] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556

[20] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[21] Papineni, K., Roukos, S., Ward, T. and Zhu, W.J. (2002) BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the* 40*th Annual Meeting on Association for Computational Linguistics*, July 2002, 311-318. https://doi.org/10.3115/1073083.1073135

[22] Lin, C.Y. and Och, F.J. (2004) Looking for a Few Good Metrics: ROUGE and Its Evaluation. *NTCIR Workshop*, Tokyo, 2-4 June 2004.

[23] Vedantam, R., Lawrence Zitnick, C. and Parikh, D. (2015) Cider: Consensus-Based Image Description Evaluation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 7-12 June 2015, 4566-4575. https://doi.org/10.1109/CVPR.2015.7299087

[24] Sun, J. (2012) Jieba Chinese Word Segmentation Tool. https://github.com/fxsjy/jieba

[25] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Li, F.-F. (2009) ImageNet: A Large-Scale Hierarchical Image Database. 2009 *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 20-25 June 2009, 248-255. https://doi.org/10.1109/CVPR.2009.5206848

[26] Ling, W., Dyer, C., Black, A.W. and Trancoso, I. (2015) Two/Too Simple Adaptations of Word2Vec for Syntax Problems. *Proceedings of the* 2015 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Denver, Co, May-June 2015, 1299-1304. https://doi.org/10.3115/v1/N15-1142

[27] gensim: Topic Modelling for Humans. Radimrehurek.com. https://radimrehurek.com/gensim/models/word2vec.html