基于语义加权的双层 LSTM 图像描述生成方法研究

邵景晨 柴玉梅* 王黎明
(郑州大学信息工程学院 河南 郑州 450001)

摘 要 为了克服当前一些模型对图像语义信息使用不充分以及没有特定场划分景的问题,提出 SW-2LSTM 图像描述方法。构建基于 ResNet-LSTM 网络的模型,加入线性层和 BN 层,并预处理图像描述得到相应标签。提取图像标签生成向量直接作用于权重矩阵,将原权重矩阵扩展为一个与标签相关的权重矩阵集合,采用张量分解思想将其分解,并添加集束搜索算法。最后将 MS COCO 数据集在基本类别上进行场景分类。实验结果表明提出的模型可以有效地提高生成描述的质量。

关键词 图像描述 深度学习 长短时记忆网络 图像特征 标签

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.10.024

IMAGE CAPTION GENERATION METHOD OF A TWO-LAYER LSTM BASED ON SEMANTIC WEIGHTING

Shao Jingchen Chai Yumei* Wang Liming
(*School of Information Engineering, Zhengzhou University, Zhengzhou 450001, Henan, China*)

Abstract In order to overcome the problems that some current models do not fully use the semantic information of images and do not have specific scene division, an image caption method named SW-2LSTM is proposed. A model based on the ResNet-LSTM network was constructed, and the linear layer and the BN layer were added. And image caption was processed to get corresponding tags. Image tags were extracted to generate tag vectors, which were directly applied to the weight matrix, and the original weight matrix was extended to a set of weight matrices related to tags. The weight set was decomposed by using tensor decomposition, and the bean search algorithm was added. The MS COCO data set was classified on its basic categories. Experimental results show that the model can effectively improve the quality of generating caption.

Keywords Image caption Deep learning LSTM Image feature Tag

0 引 言

随着互联网的快速发展,每天产生的图像规模都以亿计,呈爆炸式增长,这些图像具有很高的社会价值与商业价值。将图像与自然语言相结合在当前具有广泛的应用前景,图像描述生成任务就是对输入的图像生成准确描述图像的句子,它融合了计算机视觉和自然语言处理技术。近年来,基于深度学习的图像描述生成[1]打破了传统方法的枷锁,取得了一些突破性成果,如 Liu 等[2]通过视觉注意力和主题建模设计了 NICVATP2L 模型,提升了生成中文描述的质量,Sur 等[3]首次引入将区域图像特征和抽象交互似然嵌入相结合的概念,得到了较优的表现。但现有方法存在对图像语义信息使用不充分、缺少特定场景处理等问题,为此,本文对基于深度学习的图像描述生成方法展开研究,主要工作分为以下三个方面:

(1) 提出基于 ResNet-LSTM 网络的语义标签提取模型。使用 ResNet 处理图像得到图像特征,并使用 LSTM 网络提取 top-k 的标签。针对网络之间直接传

收稿日期:2021 - 05 - 03。NSFC-通用技术基础研究联合基金项目(U1636111)。**邵景晨**,硕士,主研领域:自然语言处理,图像处理。**柴玉梅**,教授。**王黎明**,教授。

输数据会影响本层学习到的特征以及标签需要与描述之间相关联的问题,首先,在 ResNet 中添加线性层以及 batch normalization 层,对提取的图像特征进行转化并作为 LSTM 模型的输入。其次,预处理图像描述句子得到对应的标签,对模型进行有监督的学习,以生成与图像描述句子更加相关的标签。

(2) 提出基于语义加权的双层 LSTM 模型(A two-layer LSTM model based on semantic weighting,SW-2LSTM)。该模型将提取的标签向量直接作用于权重矩阵,并采用张量分解思想和集束搜索算法提高生成描述质量。针对语义信息使用不充分以及矩阵参数过多的问题,首先,将提取的语义标签向量全局作用于双层 LSTM 的权重矩阵,训练时每个标签对应一个权重矩阵,将原始的一个权重矩阵扩展为与标签相关的权重矩阵集合,这将形成一个具有大量参数的权重张量,为了减少参数数量,采用张量分解方式将权重张量分解为三向矩阵乘积,提取出公共参数。其次,测试时加入集束搜索算法,算法通过加入概率模糊增加预测单词的不确定性,从而生成更优的图像描述。

(3) 给出基于场景分类的描述生成。针对现有算法一般采用多场景训练,没有考虑到特定场景的问题,将 MS COCO 数据集在其基本类别上划分场景,并使用 SW-2LSTM 模型在不同场景划分的训练集和测试集上完成描述生成任务。

1 相关工作

随着深度学习的快速发展和计算设备不断地更新升级,生成的图像描述句子质量得到了较高的提升。

Mao 等[4]首次将图像描述生成任务分解为两大部分,即提取图像特征和生成图像描述,提出了多模态循环神经网络(Multimodal RNN,MRNN)。Vinyals 等[5]提出的 NIC 算法使用 LSTM 替代了 RNN,并且 NIC 采用效果更好的 GoogleNet[6]。

Huo 等[7]通过结合混合深度学习阶段和描述生成阶段的两阶段框架,提高了识别人-物之间的交互能力及其与场景之间的语义关系的能力。汤鹏杰等[8]对训练过程进行改变,提出了基于逐层优化的多目标优化及多层概率融合的 LSTM 模型。在训练模型过程中,先训练浅层 LSTM 至收敛,在保留现有的分类层和目标函数的基础上再添加 LSTM 层进行训练,同时对模型参数进行微调,模型性能得到了较大的提高。陈龙杰等[9]提出了多注意力多尺度特征融合的图像描述生成模型,提取图像在不同层上的特征作为多注意力结构的输入,并在多层循环网络中加入残差连接,使得生成的图像描述有更好表现。

2 基于 ResNet-LSTM 网络的标签提取

本文提出的图像描述生成模型的基础之一为提取语义标签及概率。为了准确提取标签,在 ResNet 网络[10]和 LSTM 网络[11]的基础上,融合两者的优势提出基于 ResNet-LSTM 的模型,以标签的形式解释机器对于图像的理解,且尽量使生成的标签更贴近图像描述。

2.1 模型的构建

基于 ResNet-LSTM 的标签提取网络,使用 ResNet 来获取图像的浅层语义信息,并将其作为 LSTM 的输入,再结合预处理过程中将图像描述转化得到的图像标签训练模型,使模型进行有监督的学习。

2.1.1 语义标签提取任务

为了提高模型的提取标签的效果,采用词嵌入编码方式,将每个标签表示为连续且等长的向量。通过该方式使得网络模型不仅能够预测和提取多个标签,还能够有效地学习标签与图像之间的联合依赖关系。本节将标签提取作为多分类任务,因此训练网络模型时,采用 categorical_crossentropy 计算 *loss* 值,式为:

$$loss = -\sum_{i=1}^{n}(y_{i1}\log t_{i1} + y_{i2}\log t_{i2} + \cdots + y_{im}\log t_{im}) \quad (1)$$

式中:m 表示词嵌入编码的长度;y 表示预测值;t 表示真实值;使用交叉熵损失的方法判断训练的模型对图像预测结果的好坏。

2.1.2 构建 ResNet-LSTM 标签提取网络

标签提取模型共分为两部分,ResNet 提取图像特征,LSTM 生成标签。为了避免 ResNet 提取的特征输入到 LSTM 时有一定的损失,在基本 ResNet 网络后添加线性层和 BN 层对特征进行转化。LSTM 网络负责处理输入的图像特征的同时也负责对标签之间的关联关系进行建模,最终生成图像语义标签。该模型训练时式表示为:

$$f = \mathrm{BN}(\mathrm{liner}(y_n)) \quad (2)$$

$$h_i = \mathrm{lstm}(\boldsymbol{c}, f, h_{i-1}) \quad (3)$$

$$o_i = \mathrm{liner}(h_i) \quad (4)$$

式中:y_n 表示提取的图像语义特征;$\boldsymbol{c}$ 表示图像的真实标签的词嵌入向量;h_i 表示 LSTM 网络当前时刻的输

出;o_i 表示 LSTM 网络对图像标签的预测。其中,模型预测标签时式表示为:

$$tag_i = \text{liner}(\text{lstm}([o_{i-1}, f])) \quad (5)$$

式中:tag_i 表示模型提取的图像标签。图 1 为模型结构及训练过程,在训练阶段,网络结合图像特征以及预处理描述句子得到的图像标签一起作为 LSTM 网络的输入,计算输出标签。与真实标签对比,通过交叉熵损失函数计算模型的 loss 值,再通过反向传播修正模型参数,达到训练网络模型的目的。模型标签提取时模型直接采用训练时所用的词汇集合对图像标签进行逐单词的预测并计算相应的概率值。

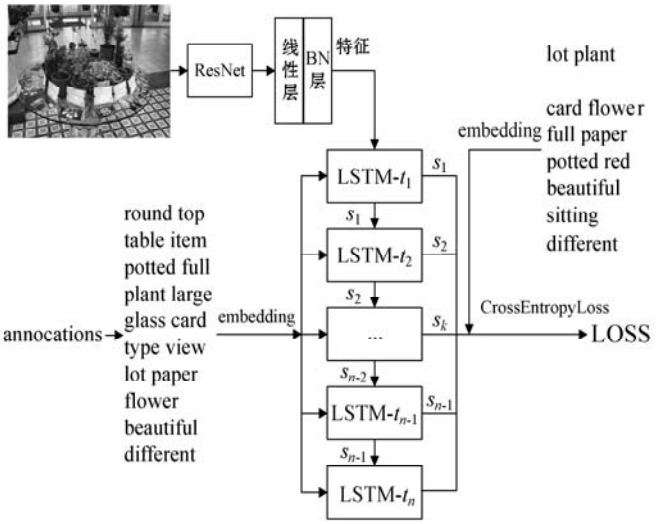

图 1　模型训练图

2.2　数据预处理

数据预处理部分主要是对每幅图像五句人工标注的描述以及标签的处理,共分为四个步骤,分别为:获取图像描述、提取高频词汇、图像描述转单词标签、过滤图像标签。

获取图像描述是将对乱序分布的图像描述句子转换为有序分布,并将每幅图像对应的描述句子形成一个列表。提取高频词汇是将整个数据集图像描述句子中出现的所有单词进行统计,使用描述句子中出现频率最高的前 1 000 个单词来确定词汇表。

图像描述转单词标签过程将人工标注的描述句子转化为单个语义标签的形式,使得语义的表示更加精炼、简洁。预处理最后一步过滤图像标签,将提取的语义标签词汇进一步简化,去除掉前 15 名被频繁使用的无意义词汇,比如“on”、“a”等。在训练时,每个单词都对应唯一的数字编码,这使整个训练过程只需处理图像对应的数字标签而不用单独处理冗长的单词信息。数据的预处理过程中,将图像描述转化为相应的数字标签过程如图 2 所示。

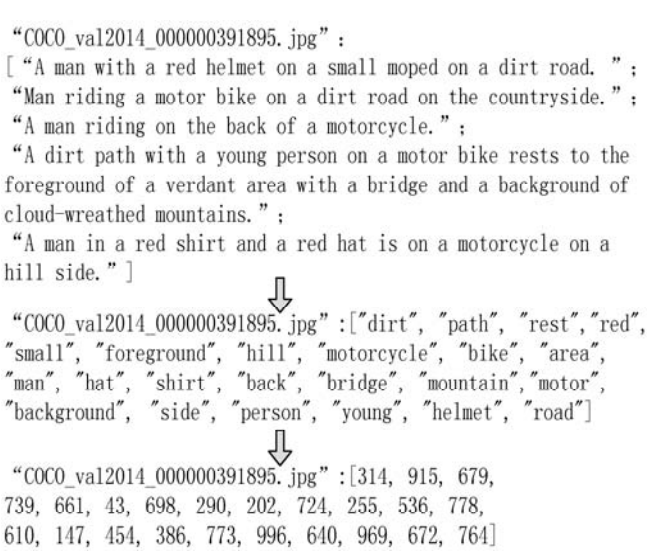

图 2　图像描述转换为数字标签过程

2.3　标签提取结果展示

图 3 展示了构建的模型在测试集上标签提取的结果,通过与真实结果进行对比,可以得到模型提取的标签能够很好地反映图像的内容。

测试图像	
提取标签	train(0.85) track(0.92) forest(0.98), blue(0.98), tree(0.99), area(0.66), wooded(0.67), red(0.64), engine(0.91), bridge(0.73), locomotive(0.83), car(0.96), green(0.67), passenger(0.86), narrow(0.62), old(0.70), white(0.82)
真实标签	car, train, blue, bridge, track, go, commuter

图 3　标签提取结果展示

3　图像描述生成模型 SW-2LSTM

将提取的标签及概率值生成语义标签向量,并与每层 LSTM 的权重矩阵相作用,将 LSTM 的基本权重矩阵扩展为与图像语义标签信息相关的权重矩阵集合。集合中每个标签的权重矩阵参与生成图像描述的程度与本身概率相关。这形成一个具有大量参数的权重张量,因此采用因式分解的思想分解权重张量减少模型参数。在测试时加入集束搜索算法,提高生成描述的质量。最后给出基于场景分类的图像描述生成。

3.1　SW-2LSTM 模型的构建

在图像描述任务中,编码端部分通过 ResNet 将图像转变为对应的编码向量。解码端主要是通过双层

LSTM 解码生成图像描述。为了解决编码端 ResNet 提取并传递至解码端的特征向量不足以概括图像本身包含的大量特征信息,以及特征信息作为中间编码在解码端随着 LSTM 时间步不断向前过程中所占比例越来越小的问题,提出了 SW-2LSTM 模型,结构如图 4 所示。在编码端部分,在原来提取图像特征向量的基础上,额外提取图像标签及概率值。在解码端,提取图像的特征向量作为双层 LSTM 的初始输入,提取的标签向量与双层 LSTM 的权重矩阵进行加权操作(图中三角形位置)。

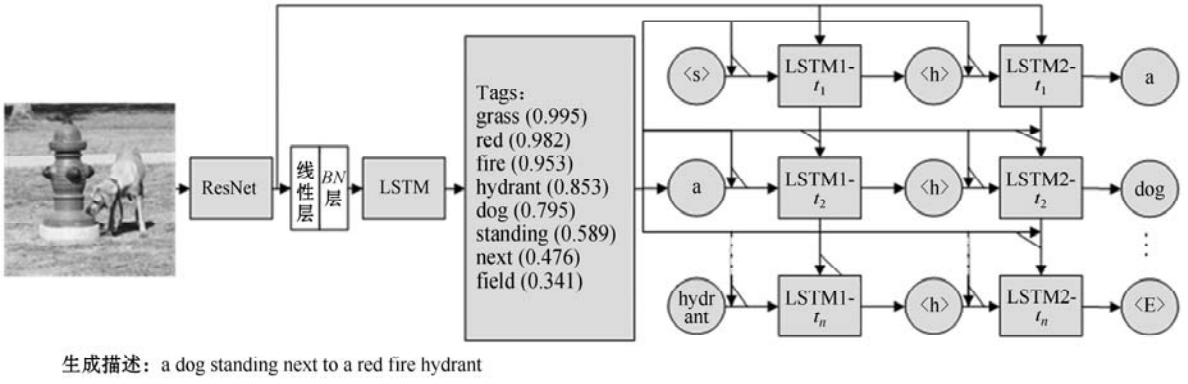

图 4 SW-2LSTM 模型

SW-2LSTM 模型基于 TensorFlow 深度学习框架,图像特征提取过程采用在 ImageNet 上预训练好的 ResNet-152 网络,将 pool 5 层输出作为双层 LSTM 的初始输入。双层 LSTM 生成图像描述时,每个时间步中都将标签向量与原权重矩阵进行加权操作,得到与图像标签相关的权重张量。并在测试时加入集束搜索算法,模糊每个时间步中可选状态集的概率,增加预测单词的不确定性,从而生成质量更好的描述。

SW-2LSTM 模型整体流程如下:

步骤 1 预处理图像,对图像进行归一化操作,处理结果送入预训练好的 ResNet-152 网络中,获得 2 048 维的图像特征表达。

步骤 2 将图像特征降维到单词的嵌入空间,得到大小为 300 维的图像特征向量 $\boldsymbol{v}$。

步骤 3 将图像送入基于 ResNet-LSTM 网络的标签提取模型中,提取图像的标签及其概率值,并生成标签向量 $\boldsymbol{s}$。

步骤 4 图像特征向量 $\boldsymbol{v}$ 作为初始输入送到第一层 LSTM 中。

步骤 5 标签向量 $\boldsymbol{s}$ 作为额外信息对权重矩阵进行加权;第一层 LSTM 的输出作为第二层 LSTM 的输入,此时 $\boldsymbol{s}$ 像同样对第二层 LSTM 的权重矩阵进行加权。

步骤 6 第二层 LSTM 的输出经过 softmax 得到当前时刻在描述词典上的概率分布。

步骤 7 若当前时刻第二层 LSTM 的输出为结束标志,跳至步骤 8;否则作为下一时刻第一层 LSTM 的输入,跳至步骤 5。

步骤 8 算法结束,完成生成图像描述任务。

3.2 用语义标签向量对权重矩阵加权

基础 LSTM 中遗忘门 $\boldsymbol{f}$、输入门 $\boldsymbol{i}$、输出门 $\boldsymbol{o}$ 以及前时刻细胞记忆 $\hat{\boldsymbol{c}}_t$ 权重矩阵同时包含 $\boldsymbol{W}$、$\boldsymbol{U}$,暂时用 $\boldsymbol{W}$、$\boldsymbol{U}$ 统一代表。初始权重矩阵 $\boldsymbol{W}$、$\boldsymbol{U}$ 经过与标签向量 $\boldsymbol{s}$ 发生作用之后,变为与标签相关的权重矩阵集合 $\boldsymbol{W}(\boldsymbol{s})$、$\boldsymbol{U}(\boldsymbol{s})$,并且该集合的成员受到标签概率的限制。对于 $\boldsymbol{W}(\boldsymbol{s})$、$\boldsymbol{U}(\boldsymbol{s})$ 对其定义如下:

$$\boldsymbol{W}(\boldsymbol{s}) = \sum_{k=1}^{K} s_k \boldsymbol{W}_{\boldsymbol{T}}[k] \tag{6}$$

$$\boldsymbol{U}(\boldsymbol{s}) = \sum_{k=1}^{K} s_k \boldsymbol{U}_{\boldsymbol{T}}[k] \tag{7}$$

若 $\boldsymbol{s} \in \mathbf{R}^K$,则定义 $\boldsymbol{W}_{\boldsymbol{T}} \in \mathbf{R}^P$,$\boldsymbol{U}_{\boldsymbol{T}} \in \mathbf{R}^Q$,其中 P 为 $n_h \times n_x \times K$,Q 为 $n_h \times n_h \times K$,n_h 为隐藏单元的数量,n_x 为单词嵌入的维数。而经过语义加权的 LSTM 中每个标签 s_k 都有相对应的权重矩阵 $\boldsymbol{W}_{\boldsymbol{T}}[k]$、$\boldsymbol{U}_{\boldsymbol{T}}[k]$,这些权重矩阵就构成了与标签相关的权重矩阵的集合。则加权后 LSTM 的计算式如下:

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{f,s}\boldsymbol{x}_{f,t-1} + \boldsymbol{U}_{f,s}\boldsymbol{h}_{f,t-1}) \tag{8}$$

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{i,s}\boldsymbol{x}_{i,t-1} + \boldsymbol{U}_{i,s}\boldsymbol{h}_{i,t-1}) \tag{9}$$

$$\hat{\boldsymbol{c}}_t = \tanh(\boldsymbol{W}_{c,s}\boldsymbol{x}_{c,t-1} + \boldsymbol{U}_{c,s}\boldsymbol{h}_{c,t-1}) \tag{10}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \circ \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \circ \hat{\boldsymbol{c}}_t \tag{11}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{o,s}\boldsymbol{x}_{o,t-1} + \boldsymbol{U}_{o,s}\boldsymbol{h}_{o,t-1}) \tag{12}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \circ \tanh(\boldsymbol{c}_t) \tag{13}$$

式中:$\circ$ 表示点乘运算,权重矩阵 $\boldsymbol{W}$、$\boldsymbol{U}$ 展开后与式(6)、式(7)相同。

3.3 SW-2LSTM 网络权重矩阵分解与集束搜索

3.3.1 权重矩阵的分解

加权后的 LSTM 网络权重矩阵的参数数量与标签数目 K 成正比,当 K 取值很大时,如本文的 K 设为

1 000,会由于参数数量太大而影响网络的训练速度。为了解决此问题,受 Memisevic 等[12]提到分解思想的启发,采用权重矩阵分解的思想分解 $\boldsymbol{W}(\boldsymbol{s})$、$\boldsymbol{U}(\boldsymbol{s})$。分解后,式(6)、式(7)变为:

$$\boldsymbol{W}(\boldsymbol{s}) = \boldsymbol{W}_a \cdot \mathrm{diag}(\boldsymbol{W}_b s) \cdot \boldsymbol{W}_c \tag{14}$$

$$\boldsymbol{U}(\boldsymbol{s}) = \boldsymbol{U}_a \cdot \mathrm{diag}(\boldsymbol{U}_b s) \cdot \boldsymbol{U}_c \tag{15}$$

式中:$\boldsymbol{W}_a \in \mathbf{R}^E$,$\boldsymbol{W}_b \in \mathbf{R}^G$,$\boldsymbol{W}_c \in \mathbf{R}^J$,$E$ 为 $n_h \times n_f$,G 为 $n_f \times K$,J 为 $n_f \times n_x$,n_f 是分解时因子数量,与此类似,$\boldsymbol{U}_a \in \mathrm{R}^E$,$\boldsymbol{U}_b \in \mathrm{R}^G$,$\boldsymbol{U}_c \in \mathrm{R}^J$。经过分解,将权重矩阵变为了与标签相关的三向矩阵乘积。为了进一步分析式(14)、式(15),让 $\boldsymbol{w}_{bk}$ 表示 $\boldsymbol{W}_b$ 的第 k 列,则可以得到:

$$\boldsymbol{W}(\boldsymbol{s}) = \sum_{k=1}^{K} s_k [\boldsymbol{W}_a \cdot \mathrm{diag}(\boldsymbol{w}_{bk}) \cdot \boldsymbol{W}_c] \tag{16}$$

与此类似,让 $\boldsymbol{u}_{bk}$ 表示 $\boldsymbol{U}_b$ 的第 k 列,则可以得到:

$$\boldsymbol{U}(\boldsymbol{s}) = \sum_{k=1}^{K} s_k [\boldsymbol{U}_a \cdot \mathrm{diag}(\boldsymbol{u}_{bk}) \cdot \boldsymbol{U}_c] \tag{17}$$

式(16)中的 $\boldsymbol{W}_a \cdot \mathrm{diag}(\boldsymbol{w}_{bk}) \cdot \boldsymbol{W}_c$ 可以解释为权重张量的第 k 个切片,与第 k 个标签相对应。因此,通过式(14)、式(15)可以成功学习到 K 个一般 LSTM 集合的参数,每个参数集合对应一个标签,对于每个标签,共享 $\boldsymbol{W}_a$、$\boldsymbol{W}_c$、$\boldsymbol{U}_a$、$\boldsymbol{U}_c$。则模型的加权与分解算法如下:

算法 1　权重矩阵的加权与分解

输入:初始权重矩阵 $\boldsymbol{W}$、$\boldsymbol{U}$,标签向量 $\boldsymbol{s}$。

输出:权重矩阵 $\boldsymbol{W}(\boldsymbol{s})$、$\boldsymbol{U}(\boldsymbol{s})$。

1. 输入 $\boldsymbol{W}$、$\boldsymbol{U}$、$\boldsymbol{s}$
2. for each $k \in K$ do　　//K 为标签数量
3. 将初始权重矩阵 $\boldsymbol{W}$ 输入式(6)得到标签相关的权重矩阵 $\boldsymbol{W}_k$
4. 将 $\boldsymbol{W}_k$ 作为列向量加入权重矩阵集合 $\boldsymbol{W}[KS]_1$
5. 将初始权重矩阵 $\boldsymbol{U}$ 输入式(7)得到标签相关的权重矩阵 $\boldsymbol{U}_k$
6. 将 $\boldsymbol{U}_k$ 作为列向量加入权重矩阵集合 $\boldsymbol{U}[KS]_1$
7. end for　　//得到经过标签向量加权的权重矩阵集合 $\boldsymbol{W}[KS]_1$、$\boldsymbol{U}[KS]_1$
8. 将 $\boldsymbol{W}[KS]_1$ 输入到式(14)进行初步因式分解为三向矩阵乘积 $\boldsymbol{W}[KS]_2$
9. 将 $\boldsymbol{U}[KS]_1$ 输入到式(15)进行初步因式分解为三向矩阵乘积 $\boldsymbol{U}[KS]_2$
10. 将 $\boldsymbol{W}[KS]_2$、$\boldsymbol{U}[KS]_2$ 输入到式(16)、式(17)进一步分解,提取出公共参数,将张量分解为每个标签相对应的形式,得到最终分解结果 $\boldsymbol{W}(\boldsymbol{s})$、$\boldsymbol{U}(\boldsymbol{s})$
11. return $\boldsymbol{W}(\boldsymbol{s})$、$\boldsymbol{U}(\boldsymbol{s})$

经过加权和分解之后,得到最终 SW-2LSTM 网络权重计算方法:

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{fa}\boldsymbol{x}'_{f,t-1} + \boldsymbol{U}_{fa}\boldsymbol{h}'_{f,t-1}) \tag{18}$$

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{ia}\boldsymbol{x}'_{i,t-1} + \boldsymbol{U}_{ia}\boldsymbol{h}'_{i,t-1}) \tag{19}$$

$$\hat{\boldsymbol{c}}_t = \tanh(\boldsymbol{W}_{ca}\boldsymbol{x}_{c,t-1} + \boldsymbol{U}_{ca}\boldsymbol{h}_{c,t-1}) \tag{20}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \circ \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \circ \hat{\boldsymbol{c}}_t \tag{21}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{oa}\boldsymbol{x}'_{o,t-1} + \boldsymbol{U}_{oa}\boldsymbol{h}'_{o,t-1}) \tag{22}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \circ \tanh(\boldsymbol{c}_t) \tag{23}$$

$$\boldsymbol{x}'_{f,t-1} = \boldsymbol{W}_{fb}\boldsymbol{s} \circ \boldsymbol{W}_{fc}\boldsymbol{x}_{t-1} \tag{24}$$

$$\boldsymbol{h}'_t = \boldsymbol{U}_{fb}\boldsymbol{s} \circ \boldsymbol{U}_{fc}\boldsymbol{h}_{t-1} \tag{25}$$

式(19)、式(20)、式(22)中的 $\boldsymbol{x}'$、$\boldsymbol{h}'$的计算方法与式(24)、式(25)一样。与普通 LSTM 相比,模型中三个门都多出式(24)、式(25)两式,正是这两个公式表达了标签特征如何与基础 LSTM 的权重矩阵相作用。

本文提出的模型以独特的方式使用从输入图像 I 中提取的标签特征向量 $\boldsymbol{s}$,其作为额外补充的图像语义信息,用于 LSTM 权重矩阵加权。给定输入图像 I,最大化正确描述的概率计算式如下:

$$\theta^* = \sum_{(I,D)} \log p(D \mid I;\theta) \tag{26}$$

式中:θ 表示模型的参数,I 表示输入的图像,D 表示对应的描述,N 表示句子的长度。则根据链式法式(26)可以得到:

$$\log p(D \mid I) = \sum_{t=0}^{N} \log p(D_t \mid I, D_0, D_1, \cdots, D_{t-1}) \tag{27}$$

这里省略了 θ,式(27)可以将其理解为在已知输入图像 I 和若干个已知单词 $D_0, D_1, \cdots, D_{t-1}$,要使得下一个单词 D_t 的概率最大。训练时,(D, I) 是一组训练样本,通过训练大量样本并使用随机梯度下降方法优化 log 的概率和,使权重参数 θ 调整到最大化正确描述概率。在加入标签特征向量 $\boldsymbol{s}$ 后,式(27)等价变为:

$$\log p(D \mid I) = \sum_{t=0}^{N} \log p(D_t \mid I, D_0, D_1, \cdots, D_{t-1}, v, s) \tag{28}$$

3.3.2　集束搜索

网络经过训练已学习了单词之间的关系,预测下个单词时,最大概率的单词可能是固定的,从而生成的描述与图像可能严重不符合。需要加入模糊概率,增加预测单词的不确定性,因此在双层 LSTM 中加入集束搜索算法。

在测试生成描述时,使用集束搜索来获取前 *top-k* 个最优描述。本文模型集束搜索的步骤为:

步骤 1　将图像描述生成模型第一次输出作为初始候选词,设置集束宽度为 3。

步骤 2　判断当前最新候选词是否为结束标志,若为结束标志跳至步骤 4;否则跳至步骤 3。

步骤 3　将当前候选词作为输入,生成新的候选词,对当前序列进行扩展,计算当前输出概率和前 3 名的序列,回到步骤 2。

步骤 4　选取当前概率和最大的序列为最终生成的图像描述句子,输出结果,算法结束。

在进行测试时,加入集束搜索算法的双层 LSTM 能更好地处理图像中各物体之间以及各单词之间的依赖关系,避免预测单词时直接出现固定搭配的情况。

3.4 基于场景分类的图像描述生成

现有模型通常直接在数据集上进行多场景训练和测试。随着图像数据爆炸式增长,在图像检索时,人们对结果有更高的期待和要求,图像含有关键字信息的同时,还要伴随其他相关信息的图像供其选择。盲人通常生活在特定的生活场景中,盲人导航则需要在生活的特定场景产生更安全高效的导航,所以有必要对模型进行单独场景训练,检验模型在特定场景生成描述的能力。因此,本文将数据集根据特定场景进行分类。数据集中“categories”共有 80 类,“supercategory”共有 12 类,每幅图像根据图像内容可能同时属于不同的类别。由于“categories”划分过细,每个类别包含的图像数量过少,则将数据集按“supercategory”划分,其 12 个类别及图像数量统计情况如表 1 所示。

表 1 数据集类别分布情况统计表

supercategory	训练集	验证集	总计
person	45 174	21 634	66 808
vehicle	27 247	13 088	40 335
outdoor	9 637	4 840	14 477
animal	17 970	8 859	26 829
accessory	15 832	7 661	23 493
sports	20 411	9 676	30 087
kitchen	27 412	13 014	40 426
food	12 357	5 899	18 256
furniture	28 438	13 829	42 267
electronic	13 929	6 969	20 898
appliance	8 205	3 939	12 144
indoor	10 934	5 451	16 385

按照 12 个大类划分后,每个类别包含的图像数量相比之前数量减少较多,因此在每个类别只取出 2 000 幅图像用于测试,剩余的图像用于训练。

4 实 验

4.1 实验设计

为验证 SW-2LSTM 网络的有效性及性能表现,本节设计了三组实验,第一组为消融实验,验证标签语义加权以及不同加权方式对生成描述结果的影响;第二组为同类型模型对比实验,将本文模型与其他基于深度学习的描述生成方法进行比较,验证本文所提模型的有效性;第三组为特定场景对比实验,将本文模型在经过分类后得到的特定场景上分别进行训练,并将结果与整体数据集上的结果进行对比,验证模型对特定场景生成的描述有更好的结果。

4.1.1 参数设置

为了避免数据稀疏等问题,采用 Word2vec 向量表达,若词典中的单词不在该向量表达表中,则随机初始化该单词。使用在 ImageNet 预训练好的 ResNet-152 的 pool5 层输出表示图像向量,长度为 2 048 维。

训练时,初始化所有参数均匀分布在[-0.01, 0.01]之间。语义加权网络的隐藏单元数和分解因子数设置为 512。根据硬件条件,batch 设置为 64,训练 20 轮。为了避免训练过程中出现过拟合现象,采用 Dropout[13]优化方式;同时在验证集上采用提前停止方法,方便选择最佳的权重参数保存点。Adam 优化算法学习率设置为 0.01,集束搜索算法的集束宽度 k 设为 3。

4.1.2 数据集与评价指标

为保证实验结果的可靠性,使用大型公开数据集 MS COCO 2014。选择在描述中出现次数不小于 5 次的单词,构建出大小为 8 791 的单词字典,统一句子的向量长度为 20,句子长度不够 20 时用 0 填充。并且选取 5 000 幅带有人工标注描述句子的图像作为测试集。

在本章实验中,将模型生成的描述在评分标准 BLEU-N[14]、METEOR[15]、CIDEr-D[16] 上进行评分,其中,N 一般取 1 至 4,表示由句子中一个或者多个连续单词组成 n 元组。评分越高说明生成的描述与人工标注的参考句子越接近,生成的描述质量越高。

4.2 实验结果与分析

这部分内容主要围绕设计的三组实验展开。

4.2.1 消融实验结果及分析

该组实验以双层 LSTM 模型为基础,根据是否添加标签向量以及标签向量的加权方式来验证标签对描述生成的影响。对比模型如下:

2LSTM-v:该模型只使用图像特征作为初始输入。

2LSTM-s:该模型只使用标签向量作为初始输入。

2LSTM-vs:该模型同时使用图像特征以及标签向量作为初始输入。

SW(1)-2LSTM:该模型使用图像特征作为初始输入,同时标签向量只在第一个时间步作用于双层 LSTM 的权重矩阵。

SW-2LSTM:该模型使用图像特征向量作为初始输入,同时将标签向量作用于双层 LSTM 每一个时间步的权重矩阵。

实验结果如表 2 所示,其中 B 表示 BLEU,M 表示 METEOR,C 表示 CIDEr-D。比较 2LSTM-v 模型和 2LSTM-s 模型,可以得到只输入标签向量要比只输入图像特征的结果稍好,结合 2LSTM-vs 模型,得到联合图像特征和标签向量的输入会有更好的结果。SW(1)-2LSTM 模型和 2LSTM-vs 模型相比较,得分只在 B-1 上略低,在其他标准上均有提高,说明标签向量作用于权重矩阵对结果有一定的提升。最后比较 SW(1)-2LSTM 和 SW-2LSTM,会发现 SW-2LSTM 在各个标准上均有一定的提升,特别是在针对图像描述任务提出的 CIDEr-D 上提升了 1%,这表明本文提出的最终模型能够有效提升生成图像描述的质量。

表 2　消融实验结果

模型	B-1	B-2	B-3	B-4	M	C
2LSTM-v	0.698	0.525	0.390	0.292	0.238	0.889
2LSTM-s	0.714	0.539	0.406	0.308	0.244	0.922
2LSTM-vs	0.722	0.548	0.411	0.311	0.246	0.960
SW(1)-2LSTM	0.721	0.551	0.414	0.315	0.251	0.965
SW-2LSTM(ours)	0.725	0.557	0.419	0.319	0.254	0.975

为了再次验证标签对模型生成描述的影响,设计了手动更改标签的实验,实验结果如图 5 所示,选取输入图像进行定性分析,首先检测出图像标签,生成原始描述,再手动更改标签,对比生成的图像描述。检测到标签生成原始描述后,将“boy”替换为“girl”,描述变为“a little girl is holding a teddy bear”;去掉“teddy”、“stuffed”后,描述变为“a small child is sitting on a bed”,不再含有“teddy bear”。对比更改标签前后生成的描述,证明了标签对生成描述有较大的影响。

检测到的标签
bear (0.983), indoor (0.977), teddy (0.96), stuffed (0.937), child (0.913), sitting (0.889), boy (0.809), person (0.805), bed (0.759), baby (0.685), small (0.684), little (0.633)
原始描述:a little boy is holding a teddy bear
将“boy”替换成“girl”:a little girl is holding a teddy bear
去掉“teddy”、“stuffed”:a small child is sitting on a bed

图 5　手动更改标签实验结果

4.2.2　同类模型对比实验结果及分析

与同类模型对比,验证本文模型的有效性,对比模型为:NIC、m-RNN、ATT、MLO/MLPF-LSTM + (BS)、Att-CNN-LSTM。

对比模型结果如表 3 所示,对比典型模型 NIC 和 m-RNN,本文提出的模型在 BLEU-N 标准上均有所提高。与 NIC 相比,本文模型从 B-1 到 B-4 分别提高了 8.9%、23.5%、37.8%、57.1%。与 m-RNN 相比,本文模型从 B-1 到 B-4 分别提高了 8.2%、13.7%、19.7%、27.6%。与 ATT 相比,本文模型从 B-1 到 B-4 分别提高了 2.3%、3.7%、4.2%、4.9%,在 M 上提高 7.0%,评分都有一定的提升。与 MLO/MLPF-LSTM + (BS)相比,本文模型从 B-1 到 B-4 分别提高了 1.4%、2.2%、2.2%、2.9%,在 C 上提高了 3.5%。Att-CNN + LSTM 是现有高水平的模型,与其相比,本文模型与其从 B-1 到 B-3 和 M 上虽然分别有 2.1%、0.5%、0.2%、2.4% 的差距,但是在 B-4 和 C 上分别有 2.9%、3.5% 的提升。因为在 B-1 至 B-3 和 M 上更注重生成描述中与标记语句存在的相同的 N-gram,而在 B-4 和 C 上对模型的性能要求更高,特别是 M 更注重句子本身含义以及与输入图像的关联,结果说明本模型更注重语句整体水平。

表 3　同类模型对比实验结果

模型	B-1	B-2	B-3	B-4	M	C
NIC	0.666	0.451	0.304	0.203	—	—
m-RNN	0.670	0.490	0.350	0.250	—	—
ATT	0.709	0.537	0.402	0.304	0.243	—
MLO/MLPF-LSTM + (BS)	0.715	0.545	0.410	0.310	—	0.942
Att-CNN-LSTM	0.740	0.560	0.420	0.310	0.260	0.940
SW-2LSTM(ours)	0.725	0.557	0.419	0.319	0.254	0.975

通过结果对比,证明本文提出的 SW-2LSTM 网络模型确实能够提升描述生成的水平,且在要求更高的评价标准上模型有更好的表现。图 6 展示了模型生成的图像描述示例,可以看出描述句子比较贴合图像内容,并且具有较好的可读性。

(a) a baseball player swinging a bat at a ball

(b) a group of people sitting aroud a table

(c) a man holding a tennis racquet on a tennis court

(d) a dog sitting in the drivers seat of a car

图 6　描述生成结果示例

4.2.3 特定场景实验结果与分析

该组实验将 SW-2LSTM 网络模型在完整数据集和分类后的特定场景的结果进行对比,验证模型对特定场景的描述生成有更好的结果。模型在训练时先完整数据集上训练 1 轮,再分别在每个特定场景上训练 15 轮,模型基本完成收敛。此时实验结果如表 4 所示,表中列出了六类包含较多图像场景的评分。根据表 4 中得分可以看出,在特定场景,经过训练的模型生成的图像描述质量均有一定的提升。特别是在 person、animal 场景的得分,均有较高的提升,在要求更高的评价标准的 B-4 和 C 上,person 场景更是分别有 1.2%、1.4% 的提升,animal 场景分别有 1.2%、1.3% 的提升。则可以得到 SW-2LSTM 模型能够有效地完成基于场景分类的描述生成任务,并且生成质量还有一定提升。

表 4 特定场景对比实验结果

场景	B-1	B-2	B-3	B-4	M	C
person	0.728	0.564	0.429	0.323	0.257	0.989
vehicle	0.727	0.561	0.425	0.321	0.255	0.986
sports	0.728	0.562	0.427	0.322	0.256	0.988
kitchen	0.726	0.559	0.423	0.320	0.255	0.983
furniture	0.726	0.560	0.424	0.320	0.255	0.982
animal	0.727	0.563	0.427	0.323	0.257	0.988

5 结 语

本文分析了图像描述生成的研究现状,提出了基于语义加权的双层 LSTM 模型 SW-2LSTM,用语义标签向量对权重矩阵进行加权。本文首先构建标签提取网络,提取图像标签并作用于生成描述网络的权重矩阵,形成权重矩阵集合,采用张量分解的方法减少参数,并加入集束搜索算法,最后通过实验证明本文提出模型的有效性。本文在图像描述生成任务上未来进一步工作为考虑采用多个 CNN 网络联合提取图像的表达特征,提取更加全面的图像信息。同时可以在减少模型参数上再做些尝试,提高模型训练速度。

参 考 文 献

[1] Li X R, Ye Z L, Zhao Z, et al. Clothes image caption generation with attribute detection and visual attention model [J]. Pattern Recognition Letters, 2021, 141: 68 - 74.

[2] Liu M F, Hu H J, Li L J, et al. Chinese image caption generation via visual attention and topic modeling [J]. IEEE Transactions on Cybernetics, 2020, 52(2): 1247 - 1257.

[3] Sur C. aiTPR: Attribute interaction-tensor product representation for image caption[J]. Neural Processing Letters, 2021, 53: 1229 - 1251.

[4] Mao J H, Xu W, Yang Y, et al. Explain images with multimodal recurrent neural networks [EB]. arXiv: 1410.1090, 2014.

[5] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator [C]//International Conference on Computer Vision and Pattern Recognition, 2015: 3156 - 3164.

[6] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions [C]//IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1 - 9.

[7] Huo L, Bai L, Zhou S M. Automatically generating natural language descriptions of images by a deep hierarchical framework[J]. IEEE Transactions on Cybernetics, 2021, 52(8): 7441 - 7452.

[8] 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[J]. 自动化学报,2018,44(7):1237 - 1249.

[9] 陈龙杰,张钰,张玉梅,等. 基于多注意力多尺度特征融合的图像描述生成算法[J]. 计算机应用,2019,39(2):354 - 359.

[10] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770 - 778.

[11] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735 - 1780.

[12] Memisevic R, Hinton G E. Learning to represent spatial transformations with factored higher-order Boltzmann machines[J]. Neural Computation, 2010, 22(6): 1473 - 1492.

[13] Watt N, Plessis M C. Dropout for recurrent neural networks [C]//INNS Big Data and Deep Learning Conference, 2019: 38 - 47.

[14] Reiter E. A structured review of the validity of BLEU[J]. Computational Linguistics, 2018, 44(3): 393 - 401.

[15] Satanjeev B, Lavie A. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments[J]. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and Summarization, 2005: 65 - 72.

[16] Vedantam R, Zitnick C L, Parikh D. CIDEr: Consensus-based Image Description Evaluation[EB]. arXiv: 1411.5726, 2015.