文章编号：2095-2163(2022)02-0058-06 中图分类号：TP305 文献标志码：A

基于改进注意力机制的图像描述算法

周宇辉，何志琴
（贵州大学 电气工程学院，贵阳 550025）

摘 要：图像描述的任务是根据输入图像自动生成描述该图像的语句，属于计算机视觉与自然语言处理的交叉领域。针对传统注意力机制提取特征能力不足、模型复杂且训练困难等问题，本文提出了一种改进注意力机制的图像描述模型。在传统注意力机制的基础上引入高效通道注意模块，在提升特征提取效果的同时降低模型复杂度，在保证性能的同时提高模型效率，更好的提取图像重要部分特征，生成更为准确的自然语言描述。模型在 MSCOCO 数据集上进行了验证，实验结果表明，相较于传统的注意力机制，模型在生成描述语句准确性方面有较大提升，在 BLEU-1、BLEU-3、BLEU-4 上分别有 0.2%、0.3%、0.6%的提高。

关键词：图像描述；注意力机制；自然语言处理；通道注意模块

Image description algorithm based on improved attention mechanism

ZHOU Yuhui, HE Zhiqin
(School of Electrical Engineering, Guizhou University, Guiyang 550025, China)

【Abstract】 Image description task uses machine to automatically generate the description of the image according to the input image, which belongs to the intersection of computer vision and natural language processing. In order to solve the problems of traditional attention mechanism, such as insufficient feature extraction ability, complex model and difficult training, this paper proposes an improved image description model of attention mechanism. Based on the traditional attention mechanism, the efficient channel attention module is introduced to improve the feature extraction effect while reducing the model complexity, improve the model efficiency and ensuring the performance results, better extraction of the features of important parts of the image and more accurate generation of natural language description. The model is validated on MSCOCO data set. Experimental results show that compared with the traditional attention mechanism, the model has a significant improvement in the accuracy of description generation, with an improvement of 0.2% in BleU-1, 0.3% in BleU-3 and 0.6% in BLEU-4, respectively.

【Key words】 image description; attention mechanism; natural language processing; channel attention module

0 引 言

近年来，深度学习的兴起以及计算机硬件的快速升级，使得图像处理和自然语言处理领域发展迅速[1]。取得的大量成果使跨领域研究成为可能，促进多个跨领域研究任务发展，如：图片对应文本、视觉问答、视频讲故事以及图像描述[2]。

图像描述是指让计算机自动根据输入图像生成该图像的自然语言描述句子，要求设计模型算法去理解并建立视觉与文本之间的联系[3]。如今互联网存在大量的图像资源，有效的利用这些资源，图像描述有着重要的应用。如：购物软件的商品图片搜索、搜索引擎中的图片检索、视频中多事物目标的识别、系统对话框以及帮助视觉障碍的人群等。因此，作为图像处理和自然语言处理的交叉领域，受到了广泛的关注，但同时也面临巨大的挑战，需要准确识别图中目标物、场景，并且理解其之间的相对关系。

图像描述生成的传统方法主要有两大类：基于模板的生成方法，该类方法通过识别图像中的信息并与固定的句子模板相匹配，该方法操作简单，但生成的局式结构固定，与实际图片相差过大；基于检索的生成方法，该类方法将需要生成描述图像与图像数据库中的图像进行相似排序，选择相似度最高的图像标注生成描述，但该方法生成的描述依赖图像数据库中提前标注好的描述，难以生成新颖的描述，缺乏灵活性[4]。随着深度学习的不断发展，使得图像描述生成方法有了突破性的进展，受机器翻译的启发，将机器翻译中编码源文字的循环神经网络替换为卷积神经网络来编码图像，并将编码转换为文字描述输出，Vinyals[5] 等人提出了一个端到端的图像描述模型（Neural Image Caption，NIC），该模型联合了卷积神经网络（Convolution Neural Network，

作者简介：周宇辉（1996-），男，硕士研究生，主要研究方向：计算机控制；何志琴（1971-），女，硕士，教授，硕士生导师，主要研究方向：计算机控制。
通讯作者：何志琴 Email：641443416@qq.com
收稿日期：2021-10-12

CNN)和循环神经网络(Recurrent Neural Network，RNN)，使用长短期记忆网络(Long Short Term Memory，LSTM)作为解码器生成图像描述，在图像描述领域取得巨大突破。为了进一步增强模型对图像重要区域的信息捕捉，研究人员将注意力机制融入到图像描述生成中，并提出了两种不同的注意力机制，分别是软注意力机制(soft-attendtion)和硬注意力机制(hard-attention)，使得图像描述生成的网络能够捕捉图像的局部信息[6]。深度学习的图像描述生成算法相较于早期的方法有了很大改进，但仍然存在一些缺陷[7]。当前存在的主要问题：

(1)传统的卷积神经网络的特征提取能力不足，提取的特征丢失了很多关键信息；

(2)缺失的信息将会导致生成的描述质量低下；

(3)当前的注意力机制模型复杂并且训练困难。

为解决这些问题，本文改进传统的注意力机制方法，在注意力机制的基础上引入高效通道注意(ECA)模块，在简化模型复杂度，方便训练的同时提高了提取图像特征的能力，该方法可以生成准确丰富的图像描述语句。本文使用 MSCOCO 数据集对所提出的模型进行评估，效果优于传统模型。

1　相关工作

1.1　循环神经网络

RNN 是深度学习领域中一类特殊的神经网络，可以学习复杂的矢量到矢量的映射，目前已被广泛运用于各种与时间序列相关的任务中。循环神经网络的输出不仅能够往下一层传播，也能够传递给同层下一时刻，RNN 的网络结构如图 1 所示。

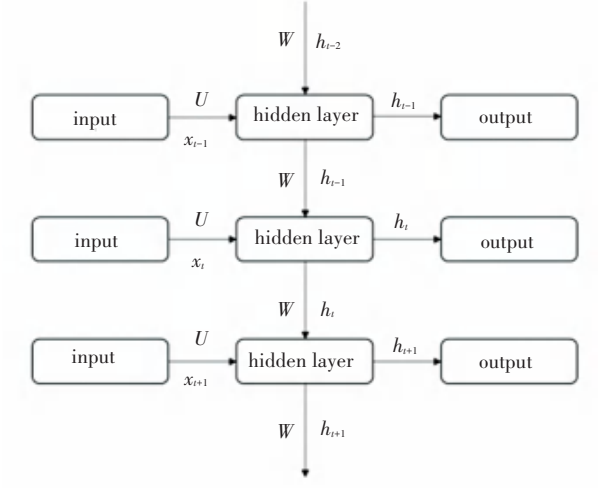

图 1　循环神经网络结构

Fig. 1　Recurrent neural network architecture

其中，U 和 W 分别是 x 和 h 的权值；X 为一个完整的时间序列；x_t 指的是 t 时刻的输入；h_t 是第 t 个时刻的输出，也就是 x_t 和上一时刻的输出 h_t 在隐藏层运算后的结果。隐藏层节点的运算如式(1)和式(2)所示：

$$h_t = f(U \times x_t + W \times h_{t-1} + b) \tag{1}$$

$$h_{t-1} = f(U \times x_{t-1} + W \times h_{t-2} + b) \tag{2}$$

其中，f 是激活函数，b 是偏置。计算第 t 时刻输出 h_t 时，带入上一时刻的输出状态 h_{t-1}。

1.2　编码-解码结构

深度学习的图像描述方法受到机器翻译领域的启发，采用在该领域效果良好的编码-解码结构作为图像描述的模型结构，解决图像描述问题，对视觉信息进行编码。编码-解码模型在图像描述领域被广泛使用，使用 CNN 作为编码器，LSTM 作为解码器。其中图像编码器用于提取图像的视觉特征，解码器基于视觉特征生成描述语句。

1.3　注意力机制

作为解码器的 LSTM 在生成图像描述时，生成的描述语句靠后的单词生成依赖靠前的单词，生成的描述依赖于语言模型，生成的描述语句准确率不够高。为了解决这个问题，研究人员将注意力机制引入了图像描述领域，在生成每个单词时，先将图像划分为若干区域，然后对不同区域的视觉特征都加入权重，通过该权重计算出图像新的视觉特征，引导单词的生成。实验表明，基于注意力机制的方式能够有效生成图像描述。

2　改进注意力机制的图像描述模型

本模型采用 encoder-decoder 图像描述框架，主要包括两个部分，Encoder 编码器部分与 Decoder 解码器部分。编码器部分通过卷积神经网络 ResNet50 提取输入图像的特征，并将该特征送入解码器当中，改进传统注意力机制，增加高效通道注意(ECA)模块，在提升模型特征提取效果的同时具有更低的模型复杂度，这种捕捉跨通道信息交互的方法在保证性能结果的同时提高了模型效率，更好的捕捉图像重要部分的特征，得到图像的注意力表征，选择所有特征向量的子集来选择性的聚焦于图像的某些部分。解码器部分选取长短期记忆网络，该网络能够很好的解决长序列训练过程中存在的梯度消失和梯度爆炸问题，在长序列的训练中有表现良好。输入的图像在经过预处理后大小为 224×224 像素，作为编码器的输入，在传统注意力机制基础上引入

ECA 模块,改变注意力机制的原有结构,在解码的每个时刻输入改进注意力机制,计算出的编码向量。使用 Adam 优化模型,使模型概率之和达到最优。模型总体框架如图 2 所示。

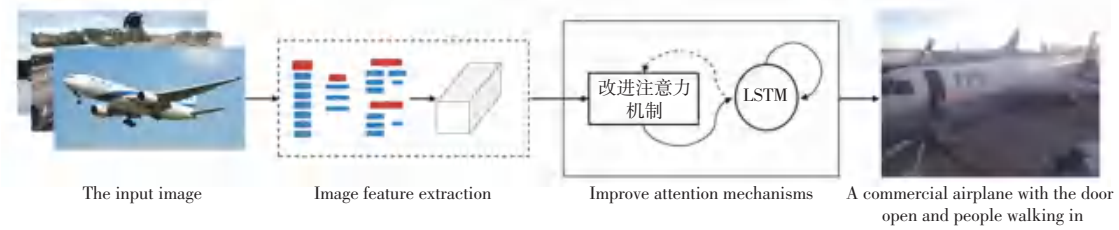

图 2 模型总体架构

Fig. 2 Overall model architecture

2.1 图像特征提取

卷积神经网络是深度学习的代表算法之一,是一类含有卷积计算并且具有深度结构的前馈神经网络,被广泛应用于计算机视觉、自然语言处理等领域。本文选用 ResNet50 作为图像特征提取的卷积神经网络,结构如图 3 所示。

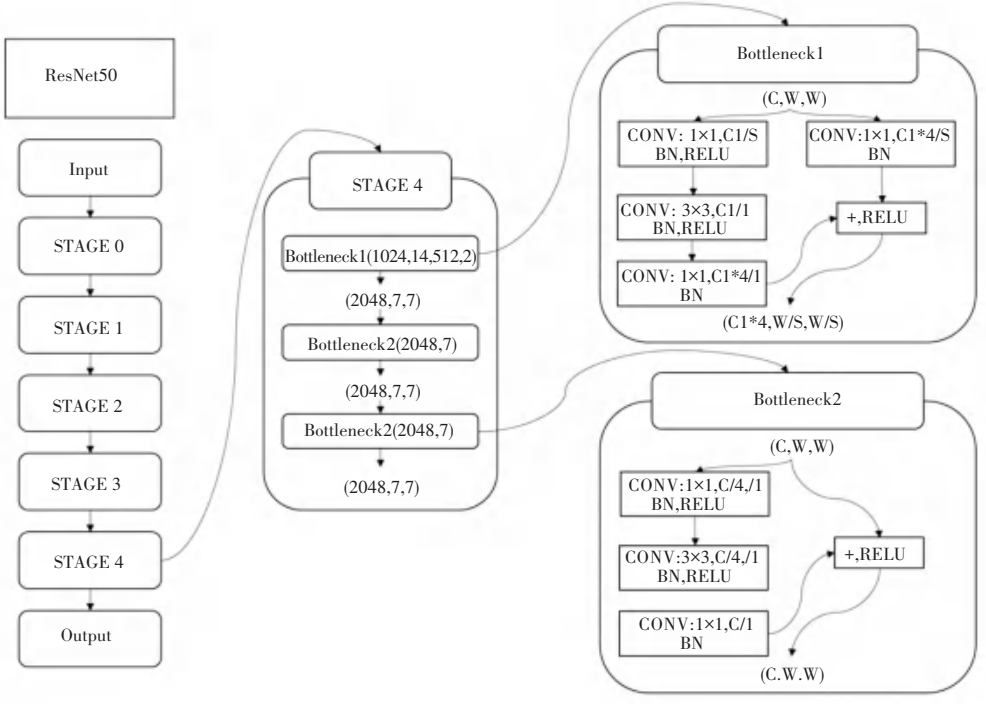

图 3 ResNet 结构图

Fig. 3 ResNet structure

ResNet50 与其他卷积神经网络如 AlexNet、VGG19 等相比,通过残差学习解决了深度网络的退化问题,可以训练出更深的网络,捕捉图像更深层次的特征。在训练之前,首先对数据集进行预处理,将参与训练的图片处理为 224×224 大小,输入图片后,提取网络训练输出向量作为图像的特征。

2.2 改进注意力机制

在注意力机制方面,采用 ECANet 方法来对图像特征进行注意力权重计算,这是一种不降维的局部跨信道交互策略和自适应选择一维卷积核大小的方法,避免降维,采取适当的跨信道交互,在显著降低模型复杂度的同时保持性能,从而实现性能上的提升。高效通道注意(ECA)模块结构如图 4 所示。

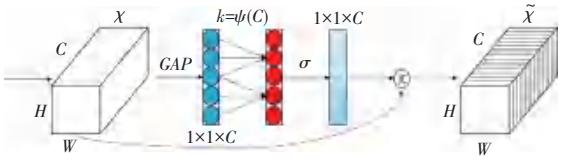

图 4　ECA 模块结构图

Fig. 4　ECA module structure diagram

一个卷积块的输出为$X \in R^{W\times H\times C}$，其中 W、H、C 分别为宽度、高度和通道尺寸，模块中的权值可以用式(3)计算：

$$f_{\{W_1,W_2\}}(y) = W_2 ReLU(W_1 y) \tag{3}$$

其中，$ReLU$ 为线性单元，W_1 和 W_2 的大小设置为 $C\times\left(\frac{C}{r}\right)$ 和 $\left(\frac{C}{r}\right)\times C$。

对于权重，只考虑 y_i 与附近 k 个值之间的信息交互，计算式(4)如下：

$$\omega_i = \sigma\left(\sum_{j=1}^{k} w_i^j y_i^j\right), y_i^j \in \Omega_i^k \tag{4}$$

还可以通过让所有通道共享权重信息的方法，进一步提高性能，计算式(5)如下：

$$\omega_i = \sigma\left(\sum_{j=1}^{k} w^j y_i^j\right), y_i^j \in \Omega_i^k \tag{5}$$

根据上述分析，利用一种新的方法，通过卷积核大小为 K 的一维卷积来实现通道之间的信息交互，提高模型的准确性，模块式(6)如下：

$$\omega = \sigma(C1D_k(y)) \tag{6}$$

其中，$C1D$ 代表了一维卷积。

这种方法称之为 ECA 模块，其只涉及 K 个参数信息，在提升效果的同时具有更低的模型复杂度，这种捕捉跨通道信息交互的方法在保证性能结果的同时提高了模型效率。

由于 ECA 模块的作用是适当捕获局部跨通道信息交互，需要确定通道交互信息的大致范围，也就是卷积核的大小 K。针对不同的卷积神经网络架构可以手动优化设置信息交互的最佳范围，但是手动设置调整会花费大量的计算资源，由于分组卷积成功的改善 CNN 架构，通道与卷积成正比，k 与 C 之间存在映射关系(7)：

$$C = \Phi(k) \tag{7}$$

用简单的线性映射表达 k 与 C 之间的关系有很大的局限性，由于通道维数通常是 2 的指数倍，采用以 2 为底的指数函数来表示非线性映射关系，式(8)：

$$C = \Phi(k) = 2^{(\gamma * k - b)} \tag{8}$$

由此可以推出，给定通道数，卷积核大小可以由式(9)计算得出：

$$k = \psi(C) = \left|\frac{\log_2(C)}{\gamma} + \frac{b}{\gamma}\right|_{odd} \tag{9}$$

2.3　长短期记忆网络

模型采用循环神经网络中的长短期记忆网络(LSTM)作为解码器生成图像描述语句。LSTM 是一种特殊的循环神经网络，可以很好的解决长序列训练过程中存在的梯度消失和梯度爆炸问题，相比于普通的循环神经网络，LSTM 能够在长序列的训练中有更好的表现。长短期记忆网络结构如图 5 所示。

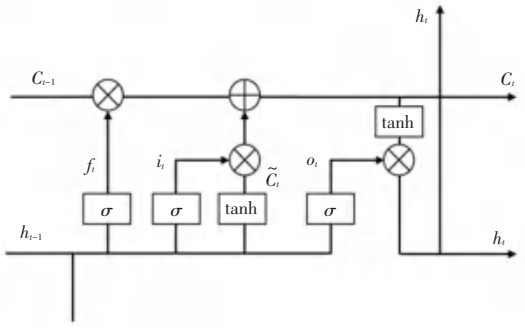

图 5　长短期记忆网络结构图

Fig. 5　Long and short memory network structure diagram

LSTM 主要有 4 个部分组成，分别是遗忘门、输入门、输出门和状态门。LSTM 首先要计算从细胞状态丢失的信息，这个计算由遗忘门完成，读取 h_{t-1} 与 x_t，输出一个范围在 0~1 之间的数值给 C_{t-1}，式(10)：

$$f_t = \sigma(\boldsymbol{W}_f \cdot [h_{t-1}, x_t] + b_f) \tag{10}$$

其中，σ 表示 $sigmoid$ 函数；h_{t-1} 表示上一个网络的输出；x_t 表示当前的输入；$\boldsymbol{W}_f$ 为权重矩阵；b_f 为偏置。

接下来通过 $sigmoid$ 层即输入门层，决定什么值要被更新，tanh 层创建了一个新的候选向量 $\widetilde{\boldsymbol{C}}_t$ 加入到状态更新中，式(11)~式(13)：

$$i_t = \sigma(\boldsymbol{W}_i \cdot [h_{t-1}, x_t] + \boldsymbol{b}_i) \tag{11}$$

$$\widetilde{\boldsymbol{C}}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{12}$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{\boldsymbol{C}}_t \tag{13}$$

其中，$\boldsymbol{W}_i$ 表示权重矩阵；$\boldsymbol{b}_i$ 表示偏置向量；$\widetilde{\boldsymbol{C}}_t$ 表示细胞状态的候选值向量。

最后，通过计算得到 $\boldsymbol{o}_t$，式(14)；利用 tanh 函数对细胞状态进行处理，得到 LSTM 的输出 h_t，式(15)：

$$o_t = \sigma(\boldsymbol{W}_o [h_{t-1}, x_t] + b_o) \tag{14}$$

$$h_t = o_t * \tanh(C_t) \quad (15)$$

其中，$\boldsymbol{W}_o$ 表示权重矩阵，b_o 表示偏置。

3 结果与分析

3.1 数据集及实验环境

本实验在 MSCOCO 数据集上进行训练测试，该数据集包括训练集、验证集和测试集，其中训练集共有 82 783 张图片，验证集有 40 504 张，测试集有 40 775 张，每张图片都有对应好的人工标注，存放在对应的 JSON 文件中。80 000 张图像用于训练，20 000 张用于测试评估。

实验环境为 windows10 操作系统下的 Tensorflow 深度学习框架，使用 GPU 进行训练。

3.2 评价指标

图像描述生成的评价有主观评价与客观量化评价。当前的客观量化评价方法主要有：BLEU、Meteor、ROUGE、CIDEr 和 SPICE。本文使用针对机器翻译常用指标：BLEU、Meteor 和针对图像描述的评价指标 CIDEr 对实验结果进行评价。

3.3 实验设置

实验中，首先将图像预处理为 ResNet 的输入格式及大小，将图像调整为 224×224 像素，标准化图像，使其包含-1~1 范围内的像素；将数据集中的标记字幕进行分割，建立词汇表，创建词到索引和索引到词的映射；最后，将所有序列填充为与最长序列相同的长度。

在训练时，提取存储在相应文件中的特征，通过编码器传递这些特征。编码器输出、隐藏状态和解码器输入被传递给解码器，解码器返回预测和解码器隐藏状态；将解码器隐藏状态传回模型，并使用预测来计算损失，使用 Teacher Forcing 来决定解码器的下一个输入；最后，计算梯度并将其应用于优化器和反向传播。

3.4 实验和结果分析

为了验证本文算法对图像描述生成的效果，使用 DeepVS 以及 Google NIC 等模型的结果和本文模型的结果进行对比，DeepVS 是一种利用深度网络来实现图像区域和文本内容匹配的多模态 RNN 图像描述生成模型；Google NIC 是代表性的编码-解码结构的图像描述模型；Soft-Attention 和 Hard-Attention 是引入了注意力机制的两种图像生成模型，前者通过确定性的得分计算编码隐状态，梯度可以经过注意力机制，反向传播到模型中，后者依概率来采样输入端的隐状态进行计算，采用蒙特卡洛采样的方法来估计模块的梯度。使用的数据集为 MSCOCO 数据集。评价结果见表 1，加粗数值表示当前最高。

从实验结果可以看出，本文提出的模型在评价生成语句通顺性和准确性的 BLEU 评价指标上相较于其他模型有较大提升，在 BLEU-1 指标上比此前效果最好的 Hard-Attention 模型还要提高 0.002，虽然在 BLEU-2 指标上相较于 Hard-Attention 落后，但是在 BLEU-3 与 BLEU-4 指标上相较于此前提出的模型，都有很大的提高，BLEU-3 提高了 0.003，BLEU-4 提高了 0.006，表明本文模型更好地提取到了图像重要部分的信息。模型在评价指标 Meteor 和 CIDEr 上不如效果最好的 Google NIC 模型和 Soft-Attention模型，但是差距并不明显，综合各指标来看，本文提出的模型优于其他模型。本文模型生成的图像描述如图 6 所示。

表 1 MS COCO 数据集评价结果对比

Tabl. 1 Comparison of evaluation results of MSCOCO data set

模型名称	评价指标					
	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Meteor	CIDEr
Multimodal RNN	0.670	0.490	0.350	0.250	—	—
DeepVS	0.625	0.450	0.321	0.230	0.195	0.660
Google NIC	0.666	0.461	0.329	0.246	0.237	**0.813**
Soft-Attention	0.707	0.492	0.344	0.243	**0.239**	0.802
Hard-Attention	0.718	**0.504**	0.357	0.250	0.230	0.780
Ours	**0.720**	0.491	**0.360**	**0.256**	0.235	0.785

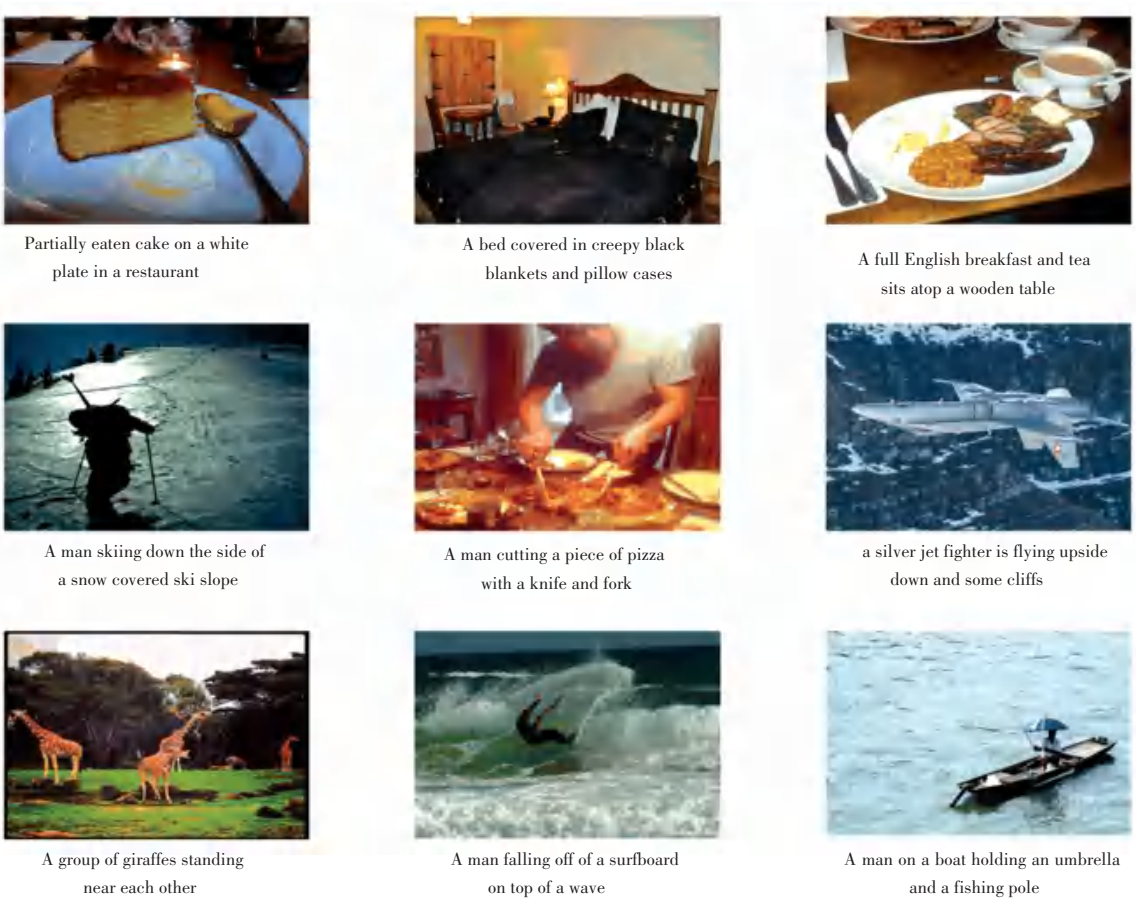

图 6　模型生成图像

Fig. 6　Generated images

4　结束语

本文采用 ResNet50 作为图像特征提取的网络，为了进一步增强模型提取特征的能力，改进传统的注意力机制，增加 ECA 模块，在提升效果的同时具有更低的模型复杂度，这种捕捉跨通道信息交互的方法在保证性能的同时提高了模型效率，更好的捕捉图像重要部分的特征。如何进一步的生成多样化的图像描述，生成更准确的描述语句，在改善图像的关键信息提取能力与构建高效的模型的方面有很大的进步空间，需要进一步研究。

参考文献

[1] 马倩霞，李频捷，宋靖雁，等. 图像描述问题发展趋势及应用[J]. 无人系统技术，2020，3(6)：25-35.
[2] 许昊，张凯，田英杰，等. 深度神经网络图像描述综述[J]. 计算机工程与应用，2021，57(9)：9-22.
[3] 李欣晔，张承强，周雄图，等. 多场景融合的细粒度图像描述生成算法[J]. 计算机与现代化，2021(9)：1-6.
[4] 黄友文，游亚东，赵朋.融合卷积注意力机制的图像描述生成模型[J].计算机应用，2020，40(1)：23-27.
[5] 王志平，郑宝友，刘仪伟. 一种改进的 LSTM 模型在图像标题生成中的应用[J]. 计算机与现代化，2020(4)：37-41.
[6] 郭淑涛，赵德新. 一种基于深度学习的中文图像描述模型[J]. 天津理工大学学报，2020，36(3)：30-35.
[7] 廖南星，周世斌，张国鹏，等. 基于类激活映射-注意力机制的图像描述方法[J]. 山东大学学报(工学版)，2020，50(4)：28-34.

(上接第 57 页)

参考文献

[1] 高鹏，王秀英，杨德贺，等. “张衡一号”卫星电场数据存储实验[J]. 地球物理学进展，2021，36(4)：1386-1392.
[2] 林子雨. 大数据技术原理与应用[M]. 北京：人民邮电出版社，2017：70.
[3] 蒋叶林. 基于 HBase 数据库的时空大数据存储与索引研究[D]. 昆明：昆明理工大学，2021.