计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(8):194−200 [doi: 10.15888/j.cnki.csa.008043]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于卷积块注意力模块的图像描述生成模型①

余海波[1,2], 陈金广[1,2]

[1](西安工程大学 计算机科学学院, 西安 710600)
[2](河南省电子商务大数据处理与分析重点实验室, 洛阳 471934)
通讯作者: 陈金广, E-mail: xacjg@163.com

摘　要: 图像描述生成模型是使用自然语言描述图片的内容及其属性之间关系的算法模型. 对现有模型描述质量不高、图片重要部分特征提取不足和模型过于复杂的问题进行了研究, 提出了一种基于卷积块注意力机制模块(CBAM) 的图像描述生成模型. 该模型采用编码器-解码器结构, 在特征提取网络 Inception-v4 中加入 CBAM, 并作为编码器提取图片的重要特征信息, 将其送入解码器长短期记忆网络 (LSTM) 中, 生成对应图片的描述语句. 采用 MSCOCO2014 数据集中训练集和验证集进行训练和测试, 使用多个评价准则评估模型的准确性. 实验结果表明, 改进后模型的评价准则得分优于其他模型, 其中 Model2 实验能够更好地提取到图像特征, 生成更加准确的描述.

关键词: 图像描述生成; 卷积块注意力模块; 卷积神经网络; 长短期记忆网络

引用格式: 余海波,陈金广.基于卷积块注意力模块的图像描述生成模型.计算机系统应用,2021,30(8):194–200. http://www.c-s-a.org.cn/1003-3254/8043.html

Image Caption Generation Model Based on Convolutional Block Attention Module

YU Hai-Bo[1,2], CHEN Jin-Guang[1,2]

[1](School of Computer Science, Xi'an Polytechnic University, Xi'an 710600, China)
[2](Henan Key Laboratory for Big Data Processing & Analytics of Electronic Commerce, Luoyang 471934, China)

Abstract: The image caption generation model uses natural language to describe the content of images and the relationship between attributes. In the existing models, there are problems of low description quality, insufficient feature extraction of important parts of images, and high complexity. Therefore, this study proposes an image caption generation model based on a Convolutional Block Attention Module (CBAM), which has an encoder-decoder structure. CBAM is added into the feature extraction network Inception-v4 and as an encoder, extracts the important feature information of the images. The information is then sent into the Long Short-Term Memory (LSTM) of the decoder to generate the caption of the corresponding pictures. The MSCOCO2014 data set is applied to training and testing, and multiple evaluation criteria are used to evaluate the accuracy of the model. The experimental results show that the improved model has a higher evaluation criterion score than other models, and Model2 can better extract image features and generate a more accurate description.

Key words: image caption generation; Convolutional Block Attention Module (CBAM); Convolution Neural Network (CNN); Long Short-Term Memory (LSTM)

图像描述涉及了计算机视觉和自然语言处理两部分的内容, 它是利用模型把一张图片转化成与之对应的自然语言描述. 图像描述在多个领域都具有重要的作用, 例如导盲、自动生成图片标签等, 这不仅方便了

① 基金项目: 河南省电子商务大数据处理与分析重点实验室开放课题 (2020-KF-7); 陕西省教育厅科研计划 (21JP049)
Foundation item: Open Fund of Henan Key Laboratory for Big Data Processing & Analytics of Electronic Commerce (2020-KF-7); Scientific Research Program of Education Bureau, Shaanxi Province (21JP049)
收稿时间: 2020-11-24; 修改时间: 2020-12-22; 采用时间: 2021-01-07; csa 在线出版时间: 2021-07-31

© 中国科学院软件研究所 http://www.c-s-a.org.cn

人们的生活, 也为大数据时代随之而来的海量图片标注减少了大量的人力.

图像描述生成[1] 主要经历了 3 个发展阶段: 基于模板的图像描述生成[2–4], 该方法通过检测得到物体及物体属性之间的关系, 之后将单词填入固定的句子模板, 但该模型过于死板; 基于检索的图像描述生成[5], 该方法先检索与当前图像相似的图像作为模板, 在检索图像关系前需要调整, 这个步骤增加了算法的复杂度; 基于深度学习的图像描述生成[6,7], 通过构建编码器-解码器框架, 采用端到端的方法对模型进行训练. 相对前两种方法, 后者在图像描述的准确性上有较大的提升. Vinyals 等提出 NIC (Neural Image Caption) 模型[8], 其思路来源于机器翻译通过最大化源语言 S 转化成目标语言 T 的概率 $p(T/S)$, 将第一个循环神经网络 (Recurrent Neural Networks, RNN) 替换成卷积神经网络 (Convolutional Neural Networks, CNN)[9], 用于提取图片的特征. Xu 等[10] 在 NIC 模型的基础上引入注意力机制, 提取图片的重要信息, 提升了模型的准确率. 大多数的视觉注意力机制只建模空间注意力机制 (spatial attention). Chen 等提出了 SCA-CNN 模型[11], 该模型同时建模空间注意力机制和通道注意力机制 (channel-wise attention), 较大的提升了模型的性能, 但该模型不够轻便、灵活. Woo 等在 SCA-CNN 的基础上提出了一种轻量级通用卷积块注意力机制模块 (Convolutional Block Attention Module, CBAM)[12]. 该注意力机制结合空间注意力机制和通道注意力机制, 并且两种注意力机制都使用平均池化和最大池化技术, 使模型的效果更好.

考虑到注意力机制在图像描述生成中的有效性, 在文献 [12] 的基础上提出了一种基于 CBAM 的图像描述生成模型. 该模型将 CBAM 模块应用到 Inception-v4[13] 网络中, 用于提取图片特征, 并送入长短期记忆网络 (Long Short-Term Memory, LSTM)[14], 生成符合图像内容描述的自然语言. 模型使用 dropout 和正则化技术防止过拟合, 利用 Word2Vec[15] 技术对自然语言进行编码处理, 以避免维度灾难等问题.

1 模型架构

1.1 Inception-v4 网络结构

卷积神经网络是深度学习的核心技术, 目前性能接近的两个网络为 Inception-v4 和 Inception-ResNet-v2. 本文选取 Inception-v4 网络作为基准网络, 结构如图 1.

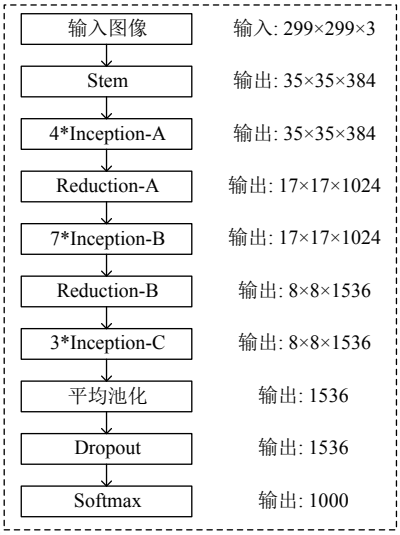

图 1 Inception-v4 网络

Inception-v4 网络与 Inception-v3 网络相比具有更多的 Inception 模块, 可以弥补 Inception-v3 网络的缺点. 在 Inception 模块和 ResNet 模块不混合的情况下, Inception-v4 网络达到了较好的性能. 针对数据集中图片大小不一致的问题, 首先对图片的尺寸进行规范化, 规范化后图片的大小为 299×299. 在 Stem 模块中使用并行结构, 在保证损失最小的情况下, 使得模型计算量最小. 网络中使用 4 层 Inception-A、7 层 Inception-B 和 3 层 Inception-C 结构, 使得网络结构加深, 分别在最后一个 Inception-A 和 Inception-B 模块后面添加 Reduction 模块, 用来降低计算量, 从而降低模型的复杂度. 为了降低特征维度, 更好地提取图像特征信息, 加入平均池化模块. 引入 dropout 模块, 防止模型在训练中过拟合, 提升模型的泛化能力.

1.2 长短期记忆网络

模型采用 LSTM 网络作为解码器处理时序信息, 生成对应图片的描述语句. LSTM 能够在一定程度上解决梯度消失和难以捕捉远距离时间信号的问题. LSTM 网络主要包括了 4 个模块, 分别是遗忘门 (f_t)、输入门 (i_t)、输出门 (O_t) 和细胞状态 (C_t), 如图 2 所示. 从细胞状态 C_{t-1} 到 C_t 的信息传输线中完成了 C_t 的更新. 遗忘门、输入门和输出门用 Sigmoid 层表示, tanh 层分别表示细胞状态的输入与输出. LSTM 首先通过 Sigmoid 层控制遗忘层, 对上一时刻的输出结果选择性的通过. 更新公式为:

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f) \quad (1)$$

© 中国科学院软件研究所 http://www.c-s-a.org.cn

其中, σ 表示 Sigmoid 函数, h_{t-1} 表示上一个 LSTM 的输出, x_t 表示此刻 LSTM 的信息输入, W_f 为权重矩阵, b_f 是偏置向量, $[h_{t-1}, x_t]$ 表示两个矩阵的拼接.

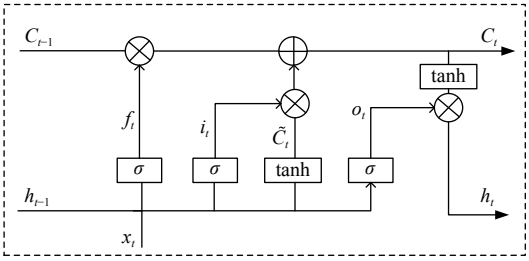

图 2 LSTM 网络

接下来决定在细胞中保存哪些重要信息, 包括两部分, 一部分是通过 i_t 更新数值, 另一部分是通过 tanh 层得到新的候选值. 给上一时刻的状态乘 f_t, 遗忘掉之前不重要的信息, 再用 $i_{t\times t}$ 加上前者得到 C_t. 公式如下:

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

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{3}$$

其中, W_i 表示权重矩阵, b_i 表示偏置向量, $\tilde{C}_t$表示细胞状态的候选值向量.

最后一步先计算得到 O_t, 然后使用 tanh 函数对细胞状态 C_t 进行处理, 乘上 O_t 的值得到 LSTM 单元的输出 h_t. 公式如下所示:

$$O_t = \sigma(W_o * [h_{t-1}, x_t] + b_o) \tag{4}$$

$$h_t = O_t * \tanh(C_t) \tag{5}$$

其中, W_o 表示权重矩阵, b_o 表示偏置向量.

1.3 CBAM 模块

为了提高模型提取图像特征的准确性, 在 Inception-v4 网络的基础上加入 CBAM 模块, 从空间注意力机制和通道注意力机制两方面获取图像更多的关键信息. 与文献 [11] 中的 SCA-CNN 模型相比, CBAM 模块可移植性强, 轻便灵活, CBAM 模块如图 3 所示.

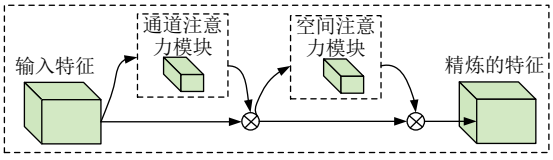

图 3 CBAM 结构

CBAM 模块包括两部分内容, 分别是空间注意力模块和通道注意力模块. CBAM 的输入是特征矩阵, 首先经过通道注意力机制生成新的特征矩阵, 再和保留的特征矩阵进行卷积操作, 所得矩阵作为空间注意力机制模型的输入, 通过空间注意力机制模块的特征再和未通过的特征卷积, 就得到了新的特征矩阵. 在卷积网络的每个卷积模块上, 通过 CBAM 自适应地调整特征矩阵. 为了提高网络的表示能力, 在通道注意力机制模块和空间注意力机制模块中加入最大池化和平均池化操作. 两种注意力机制模块分别如图 4 和图 5 所示.

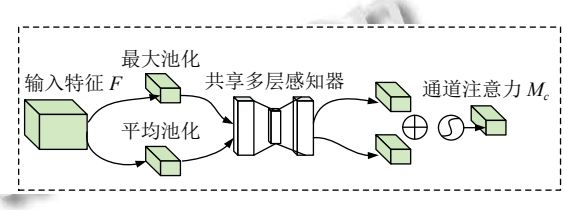

图 4 通道注意力机制

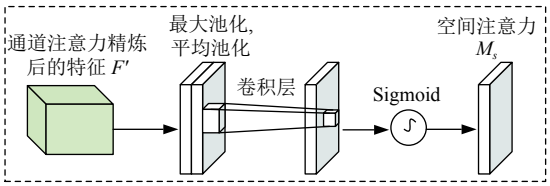

图 5 空间注意力机制

对于输入的特征 F, 分别经过最大池化和平均池化, 接着经过共享多层感知器, 将得到的两个特征相加, 再经过 Sigmoid 函数, 最终生成通道注意力特征映射 M_c, 公式如下:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F^c_{\text{avg}})) + W_1(W_0(F^c_{\max}))) \end{aligned} \tag{6}$$

式中, σ 表示 Sigmoid 函数, MLP 表示共享多层感知器, $AvgPool$ 表示平均池化, $MaxPool$ 表示最大池化. W_0 和 W_1 都表示权重矩阵, r 表示缩减率. 首先使用平均池化和最大池化聚合通道特征信息, 生成F^c_{avg}和$F^c_{\max}$, 分别表示平均池化特征和最大池化特征. 然后转发到由多层感知器构成的共享网络中, 生成注意力特征映射 M_c.

将通道注意力特征和输入特征进行一个基于对应元素逐个相乘的乘法操作, 生成空间注意力机制的输入特征F, 对于输入的特征F, 分别经过最大池化和平均池化操作, 然后进行卷积操作, 通过 Sigmoid 函数生成空间注意力特征. 公式如下:

$$\begin{aligned} M_s(F) &= \sigma\left(f^{7\times7}([AvgPool(F); MaxPool(F)])\right) \\ &= \sigma\left(f^{7\times7}([F^s_{\text{avg}}; F^s_{\max}])\right) \end{aligned} \tag{7}$$

© 中国科学院软件研究所 http://www.c-s-a.org.cn

式中, $f^{7\times7}$表示 7×7 的卷积核. 利用平均池化和最大池化操作, 聚合空间信息, 生成平均池化特征F^s_{avg}和最大池化特征F^s_{max}. 空间注意力机制集中于图片中某个感兴趣区域, 弥补通道注意力机制的不足.

1.4 模型框架

本文模型采用编码器-解码器 (encoder-decoder) 框架. 编码器部分选用 Inception-v4 网络作为特征提取的基准网络. 解码器部分选取 LSTM 网络, 该网络能够较好的处理序列类型的结构数据, 解决梯度消失等问题. 通过将 CBAM 融入 Inception-v4 网络提取图像特征, 并和 LSTM 网络共同构建图像描述模型, 生成与图像对应的自然语言描述. 通过最大化 $p(T/S)$ 完成图片 T 到目标句子 S 的转化任务. 模型框架如图 6 所示. 输入图像经过预处理之后尺寸大小为 299×299, 并作为模型的输入. 在 Inception-v4 网络中加入 CBAM 注意力机制作为模型提取图片特征的网络, 分别在每个 Inception-A、Inception-B、Inception-C 模块后面加入 CBAM, 共加入 14 个 CBAM 模块. 改变 Inception-v4 的原有结构, 去掉 Softmax 层, 在最底层加入全连接层, 目的是将 1536 维特征向量转化为 512 维, 便于图片特征向量与词向量映射到同一向量空间. 同时, 为了避免维度灾难问题, 对标注语句 S 用 Word2Vec 进行编码, 将编码后的矩阵 W_e 与上一时刻 LSTM 单元生成的单词 S_{t-1} 相乘, 并将乘积送入此刻 LSTM 单元, 按时序逐步得到与目标图片内容相符的句子 S. 使用 Adam 优化模型, 使模型概率之和达到最优. 模型使用 LSTM 作为解码模块, 可以较好的处理时序问题, 提升整体模型的准确性.

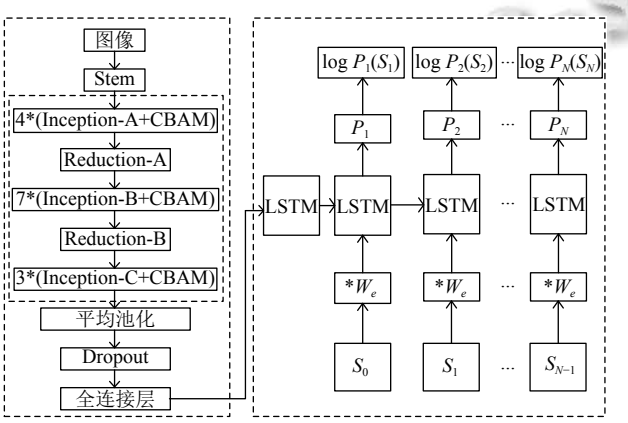

图 6 模型框架

对于图像描述生成模型, 模型中所有参数训练更新可以概括如下:

$$\theta^* = \arg\max_{\theta} \sum_{(I,S)} \log p(S|I;\theta) \tag{8}$$

其中, θ代表模型的所有参数, I 为输入的训练集图片, S 是模型生成的相应图片的描述语句. S 的长度可以用 $S_0, S_1,\cdots, S_N$ 表示, N 表示生成语句的长度.

LSTM 网络按照时间序列处理数据, 每个时刻生成一个单词, 选取概率最大的单词加入句子, 逐步生成对应的描述语句. 生成句子概率公式如下:

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \cdots, S_{t-1}) \tag{9}$$

使用随机梯度下降算法对式 (9) 进行优化. 用固定长度的隐藏层状态 h_t 表示 $S_0,S_1,\cdots, S_N$, 当输入 x_t 时, h_t 可以按照如下公式进行更新:

$$h_{t-1} = f(h_t, x_t) \tag{10}$$

其中, f 是一个函数, 为了更好地解决问题, 选取 LSTM 网络作为 f.

2 实验过程

2.1 实验环境

采用微软 MSCOCO 2014 版本的数据集. 该数据集实际共有 80 类, 例如 Bottle、Sofa、Car 类等, 包括训练集、验证集和测试集, 分别放在 train2014、val2014 和 test2014 文件夹中, 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, 每张图片共有 5 句标注, 并分别存放在相应的 JSON 文件中. 模型训练过程中并没有沿用划分测试集加入训练集的重新构造的方式, 而是把原训练集所有类型的所有图片全部用于训练整个模型, 验证集所有类型的所有图片全部用于模型评价准则的评估. 测试集可以选取少部分验证模型的有效性. 该数据集能够较好地完成图像描述生成模型的实验. 采用 Tensorflow 框架, 使用 GPU (TITAN XP) 进行训练.

2.2 评价指标

实验采用 Bleu-1[16]、Bleu-4[16]、METEOR[17] 和 CIDEr[18] 作为模型的评价指标. Bleu 主要是用来测试两个句子之间的相似程度, 最初, Bleu 通过一个句子出现在另一个句子中单词的数量来判定两个句子的相似度, 之后, 经过几次不断的改进, 引入惩罚值和最佳匹配长度计算语句之间的精度. METEOR 测试精度主要是考虑准确率和召回率, 它的出现是为了弥补 Bleu 中的

© 中国科学院软件研究所 http://www.c-s-a.org.cn

不足. Bleu 和向量空间模型结合产生了 CIDEr, 可以用来评价图像描述生成模型是否提取到图片的关键信息.

2.3 实验设置

本文模型对不同参数设置了两个实验, 分别称为 Model1 和 Model2. Model1 首先对标注语句进行处理, 限定句子的长度为 20, 不足的位置补 0. 语句开始标志为<S>, 结束标志为</S>. 设置 batch_size=27, LSTM 随机失活因子 lstm_dropout_keep_prob=0.5, Inception-v4 模型参数的学习率 train_inception_learn_rate=0.0003, 梯度裁剪 clip_gradients=5. 模型训练时设置迭代次数为 60 万步. 初始化 learn_rate 值为 2, 使用 tf.train.exponential_decay(其为 Tensorflow 中的方法) 创建训练步数衰减的学习速率, 设置 staircase 为 true, 表示阶梯衰减, 如图 7 所示. 使用集束搜索 (beamsearch) 方法逐步生成描述语句, 每个时间序列保留概率 p 最大的几个句子, 迭代操作这个步骤, 将 beam 大小设置为 3. 初始化后图像尺寸为 299×299, 长短期记忆网络输入输出均为 512 维. 将词汇字典尺寸大小设置为 12 000, 将频率出现 4 次以上的单词存入词汇表. 利用 Adam 计算并得到频率最高的单词.

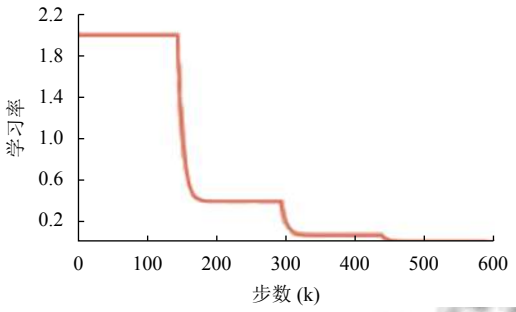

图 7 Model1 学习率衰减

由图 7 可以看出, 学习率在 450 k 步的时候受超参数设置影响而趋于稳定, 考虑到这些超参数影响因子的问题, 重新设置了个别超参数的值并进行实验, 称为 Model2. 设置 batch_size=32, 向上调整每次批处理数据的大小, 对 Inception-v4 模型参数的学习率重新设置, 即 train_inception_learn_rate=0.0005, 将梯度裁剪 clip_gradients 设置为 8. Learn_rate 的值仍然初始化为 2, 变化如图 8 所示. 其他超参数及方法不变.

由图 8 可见, 使超参数值都向上增加之后的模型 Model2 的学习率在 900 k 的时候趋于稳定. 说明 Model1 过早拟合, Model2 的学习结果较好.

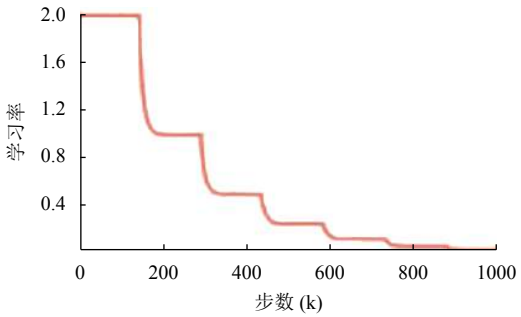

图 8 Model2 学习率衰减

2.4 实验办法

Model1 和 Model2 均采用以下实验方法: 为了提高模型的准确性和训练速度, 预先加载在 ImageNet 上预训练好的 Inception-v4 网络参数文件. 训练集图像初始化尺寸为 299×299, 并作为模型的输入, 经过改进后特征提取网络的各层之后, 图片特征为 512 维. 将图像特征和采用 Word2Vec 技术编码的词向量矩阵作为 LSTM 网络的输入, 每个时刻, LSTM 单元都会生成单词. 按照 beamsearch 方法保留概率最大的 3 个, 逐步生成描述语句 S.

3 实验结果与分析

3.1 模型的损失率

Model1 和 Model2 均采用本文模型进行实验, 模型的损失细化为每个步骤产生的正确单词的概率之和的负数, 公式如下所示:

$$L(I,S)=-\sum_{t=1}^{N}\log p_t(S_t) \tag{11}$$

其中, I 表示输入模型的训练集图像, S_t 为每个时刻生成的单词, S 表示图片的标注语句.

为了提高模型的准确率, 在 Inception-v4 网络中融合 CBAM 注意力机制, 更加精确地提取图片重要部分的信息. 使用 Word2Vec 技术对标注语句进行编码, 相对于 one-hot 编码而言, Word2Vec 可以指定维度, 对特征矩阵进行压缩, 减少了特征矩阵的存储空间, 提升了模型的准确性. 图 9 给出了 Model1 的损失图, 可以看出 Model1 的损失稳定在 2.2 左右. 图 10 给出了 Model2 的损失图, 可以看出, 重新设置超参数之后的模型的损失稳定在 2 到 2.1 之间, Model2 比 Model1 损失更少, 算法性能更好. 为了更好的降低误差, 采用随机梯度下降算法更新参数, 优化模型. 模型的损失度明显呈下降趋势, 最终趋于稳定.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

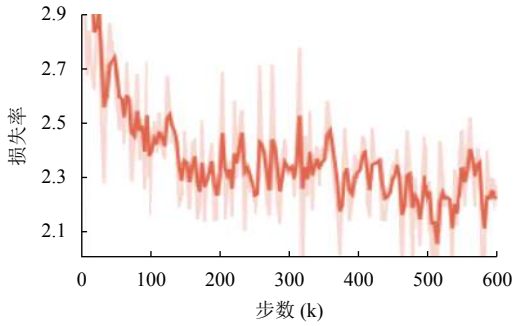

图 9 Model1 损失率

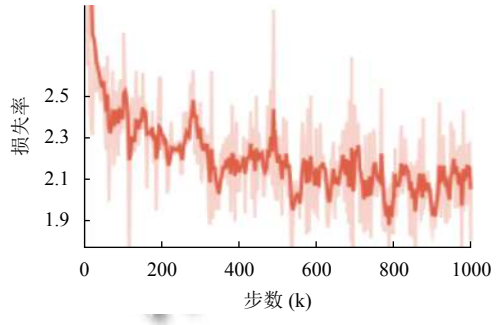

图 10 Model2 损失率

3.2 模型的实验效果

为了展示 Model1 和 Model2 的实验效果, 在验证集中选取了 4 张图片, 如图 11 所示. Moedl1 在对图 11(a) 生成的描述是: a row of motorcycles parked next to each other (一排并排停放着的摩托车), Model2 对图 11(a) 生成的描述是: A motorcycle parked in front of a garage (停在车库前的摩托车), 可以发现 Model2 的概率高于 Model1, 并且根据人工标注 (a motorcycle parked in front of a building) 和图片本身内容发现 Model2 描述更加准确. 图 11(b) 中, Model1 生成的描述语句是: A black and white dog sitting on a bench (一只黑白相间的狗坐在长凳上), Model2 生成的描述语句是: A dog sitting on a sidewalk next to a bike (一只狗坐在人行道上, 旁边是一辆自行车), 人工标注语句是: A dog sitting on a sidewalk next to a bicycle (一只狗坐在人行道上, 旁边是一辆自行车), 可以看出 Model2 比 Model1 描述效果更好. 图 11(c) 中 Model1 的描述语句为: A baseball player holding a bat on top of a field (球场上棒球手拿着球棒), Model2 和人工标注语句是一致的: A baseball player swinging a bat at a ball (球场上挥击棒球的运动员), 结合图片内容, Model1、Model2 和人工标注均能准确描述图片内容并且语句基本相同. 图 11(d) 中描述差异较大, Model1 的描述是: A brown bear standing on top of a rock (一只熊站在岩石上), 而图片内容中并没有出现岩石, 描述不准确, Moedl2 (a brown bear standing on top of a grass covered field) 和人工标注 (a brown bear is sitting in a field) 均表达正确. 综合以上所诉, Model2 能够很好地表达出图片的属性以及属性之间的关系.

Model1: a row of motorcycles parked next to each other. (p=0.000 157)

Model2: a motorcycle parked in front of a garage. (p=0.000 182)

人工标注: a motorcycle parked in front of a building.

(a) 摩托车

Model1: a black and white dog sitting on a bench. (p=0.000 138)

Model2: a dog sitting on a sidewalk next to a bike. (p=0.000 163)

人工标注: a dog sitting on a sidewalk next to a bicycle.

(b) 狗

Model1: a baseball player holding a bat on top of a field. (p=0.003 023)

Model2: a baseball player swinging a bat at a ball. (p=0.003 162)

人工标注: a baseball player swinging a bat at a ball.

(c) 棒球运动

Model1: a brown bear standing on top of a rock. (p=0.001 271)

Model2: a brown bear standing on top of a grass covered field. (p=0.001 310)

人工标注: a brown bear is sitting in a field.

(d) 熊

图 11 Model1、Model2 和人工标注

3.3 客观评价准则对比

为了进一步验证 Model1 和 Model2 两个实验的有效性, 采用 Bleu-1、Bleu-4、METEOR 和 CIDEr 这 4 个评价准则进行评估, 并与 Google NIC[8]、Multimodal RNN[19]、Hard-Attention[10] 和 SCA-CNN-ResNet[11] 比较, 结果如表 1 所示.

可以看出, Model2 的性能明显优于其他模型. Model2 在 Bleu-4 的数值与 SCA-CNN-ResNet 模型相同, 在 METEOR 上的分数高出 0.009. 在 CIDEr 上比 Multimodel RNN 高 0.266, 该值说明了 Model2 更好的提取到了图像重要部分信息. Mdel2 在 Bleu-1 上的分数接近 SCA-CNN-ResBet 模型, 比 Multimodal RNN

© 中国科学院软件研究所 http://www.c-s-a.org.cn

高 0.091. Model1 的性能在四个评价指标上略低于 Model2, 分别低于 Model1 0.007、0.009、0.009 和 0.005, Model2 的性能优于 Model1. 从模型评价准则得分表可以得出, Model2 的综合性能优于其他模型.

表 1 模型评价准则得分表

模型	Bleu-1	Bleu-4	METEOR	CIDEr
Google NIC	0.666	0.246	—	—
Multimodal RNN	0.625	0.230	0.195	0.660
Hard-Attention	0.718	0.250	0.230	—
SCA-CNN-ResNet	**0.719**	**0.311**	0.250	—
Model1	0.709	0.302	0.250	0.921
Model2	0.716	**0.311**	**0.259**	**0.926**

注: 模型得分较高者加粗体表示.

4 结论与展望

模型采用 Inception-v4 网络作为基准网络. 为了进一步增强模型提取特征的能力, 在每个 Inception 模块之后加入 CBAM 模块. CBAM 是一个轻量级的模块, 它的计算量可以忽略不计, 它可以嵌入到任何一个卷积神经网络中, 提升卷积神经网络的准确性, 更好地提取图片重要信息. SCA-CNN 模型中空间注意力机制和通道注意力机制的应用与 CBAM 相比较为复杂. 采用 LSTM 网络弥补循环神经网络的缺点, 更好地处理远距离信号问题. 在图像的关键信息提取方面仍有很大的进步空间, 需要进一步研究.

参考文献

1 楼佳珍. 基于深度学习的图像描述生成 [硕士学位论文]. 西安: 西安电子科技大学, 2018.

2 Farhadi A, Hejrati M, Sadeghi MA, *et al*. Every picture tells a story: generating sentences from images. 11th European Conference on Computer Vision. Heraklion, Crete, Greece. 2010. 15−29.

3 Kuznetsova P, Ordonez V, Berg AC, *et al*. Collective generation of natural image descriptions. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers. Stroudsburg, PA, USA. 2012. 359−368.

4 Mitchell M, Han XF, Dodge J, *et al*. Midge: Generating image descriptions from computer vision detections. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France. 2012. 747−756.

5 Socher R, Karpathy A, Le QV, *et al*. Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association for Computational Linguistics, 2014, 2(1): 207−218.

6 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. 计算机应用, 2020, 40(1): 23–27.

7 孔锐, 谢玮, 雷泰. 基于神经网络的图像描述方法研究. 系统仿真学报, 2020, 32(4): 601–611.

8 Vinyals O, Toshev A, Bengio S, *et al*. Show and tell: A neural image caption generator. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 3156−3164.

9 李彦冬, 郝宗波, 雷航. 卷积神经网络研究综述. 计算机应用, 2016, 36(9): 2508–2515, 2565. [doi: 10.11772/j.issn.1001-9081.2016.09.2508]

10 Xu K, Ba JL, Kiros R, *et al*. Show, attend and tell: Neural image caption generation with visual attention. Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lile, France. 2015. 2048−2057.

11 Chen L, Zhang HW, Xiao J, *et al*. 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, USA. 2017. 6298−6306.

12 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. 15th European Conference on Computer Vision. Munich, Germany. 2018. 3−19.

13 Szegedy C, Ioffe S, Vanhoucke V, *et al*. Inception-v4, inception-ResNet and the impact of residual connections on learning. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, CA, USA. 2017. 4278−4284.

14 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735−1780. [doi: 10.1162/neco.1997.9.8.1735]

15 Liu Y, Liu ZY, Chua TS, *et al*. Topical word embeddings. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Austin, TX, USA. 2015. 2418−2424.

16 You QZ, Jin HL, Wang ZW, *et al*. Image captioning with semantic attention. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 4651−4659.

17 Lin CY. Rouge: A package for automatic evaluation of summaries. Proceedings of the Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004. Barcelona, Spain. 2004. 74−81.

18 Vedantam R, Lawrence Zitnick C, Parikh D. CIDEr: Consensus-based image description evaluation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 4566−4575.

19 Karpathy A, Li FF. Deep visual-semantic alignments for generating image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 664–676. [doi: 10.1109/TPAMI.2016.2598339]

© 中国科学院软件研究所 http://www.c-s-a.org.cn