计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(7):232−238 [doi: 10.15888/j.cnki.csa.007996]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于类时序注意力机制的图像描述方法①

段海龙, 吴春雷, 王雷全

(中国石油大学 (华东) 计算机科学与技术学院, 青岛 266580)
通讯作者: 吴春雷, E-mail: wuchunlei@upc.edu.cn

摘　要: 近年来, 注意力机制已经广泛应用于计算机视觉领域, 图像描述常用的编码器-解码器框架也不例外. 然而, 当前的解码框架并未较清楚地分析图像特征与长短期记忆神经网络 (LSTM) 隐藏状态之间的相关性, 这也是引起累积误差的原因之一. 基于该问题, 本文提出一个类时序注意力网络 (Similar Temporal Attention Network, STAN), 该网络扩展了传统的注意力机制, 目的是加强注意力结果与隐藏状态在不同时刻的相关性. STAN 首先对当前时刻的隐藏状态和特征向量施加注意力, 然后通过注意力融合槽 (AFS) 将两个相邻 LSTM 片段的注意力结果引入到下一时刻的网络循环中, 以增强注意力结果与隐藏状态之间的相关性. 同时, 本文设计一个隐藏状态开关 (HSS) 来指导单词的生成, 将其与 AFS 结合起来可以在一定程度上解决累积误差的问题. 在官方数据集 Microsoft COCO 上的大量实验和各种评估机制的结果表明, 本文提出的模型与基线模型相比, 具有明显的优越性, 取得了更有竞争力的结果.

关键词: 图像描述; 注意力机制; 类时序注意力; 长短期记忆网络

引用格式: 段海龙,吴春雷,王雷全.基于类时序注意力机制的图像描述方法.计算机系统应用,2021,30(7):232–238. http://www.c-s-a.org.cn/1003-3254/7996.html

Image Captioning with Similar Temporal Attention Mechanism

DUAN Hai-Long, WU Chun-Lei, WANG Lei-Quan

(College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)

Abstract: Recently, attention mechanisms have been widely used in computer vision in such aspects as the common encoder/decoder framework for image captioning. However, the current decoding framework does not clearly analyze the correlation between image features and the hidden states of the Long Short-Term Memory (LSTM) network, leading to cumulative errors. In this study, we propose a Similar Temporal Attention Network (STAN) that extends conventional attention mechanisms to strengthen the correlation between attention results and hidden states at different moments. STAN first applies attention to the hidden state and feature vector at the current moment, and then introduces the attention result of two adjacent LSTM segments into the recurrent LSTM network at the next moment through an Attention Fusion Slot (AFS) to enhance the correlation between attention results and hidden states. Also, we design a Hidden State Switch (HSS) to guide the generation of words, which is combined with the AFS to reduce cumulative errors. According to the extensive experiments on the public benchmark dataset Microsoft COCO and various evaluation mechanisms, our algorithm is superior to the baseline model and can get more competitive attention results.

Key words: image captioning; attention mechanism; similar temporal attention; LSTM network

① 基金项目: 山东省自然科学基金 (ZR2020MF136); 中石油重大科技项目 (ZD2019-183-001); 中央高校基本科研业务费专项资金 (20CX05018A)
Foundation item: Natural Science Foundation of Shandong Province (ZR2020MF136); Major Science and Technology Projects of CNPC (ZD2019-183-001); the Fundamental Research Funds for the Central Universities of China (20CX05018A)
收稿时间: 2020-11-01; 修改时间: 2020-12-02; 采用时间: 2020-12-18; csa 在线出版时间: 2021-06-30

1 引言

图像描述是计算机视觉的主要任务之一, 其主要目的是为计算机提供图像, 计算机可以将图片与图片中各对象之间的关系结合起来自动生成相应的自然语言描述. 这是一项非常具有挑战性的任务[1–4]. 随着深度学习的发展, 注意力机制已经广泛应用于图像描述, 在该领域常用的编码器-解码器框架中起着举足轻重的作用. 然而, 当前的解码框架并未较清楚地分析图像特征与长短期记忆神经网络 (Long Short-Term Memory, LSTM) 隐藏状态之间的相关性, 这也可能导致累积误差. 众所周知, 单词是由 LSTM 的隐藏状态直接指导生成, 如果隐藏状态与特征向量之间的相关性不够清晰, 则很难指导生成正确的单词. 目前的注意力机制, 往往忽略了前一时刻和后一时刻注意力结果对当前时刻的影响, 导致生成的句子不是很理想, 因为对于一个句子, 单词与单词之间具有一定的相关性, 当前时刻生成的单词会受到前后时刻生成单词的影响.

为了在一定程度上解决该问题, 本文提出了类时序注意力网络 (Similar Temporal Attention Network, STAN), 该网络扩展了传统的注意力机制, 目的是加强注意力结果与隐藏状态在不同时刻的相关性. 首先, STAN 对图像进行编码并提取其自下而上的特征, 然后将编码后的图像特征传递给 LSTM 进行解码, 同时对 LSTM 的隐藏状态和图像特征施加注意力, 最后通过注意力融合槽 (AFS) 将两个相邻 LSTM片段的注意力结果引入到下一时刻的网络循环中, 以增强注意力结果与隐藏状态之间的相关性.

本文中, 创新点可以总结归纳为如下 3 点:

1) 本文设计一种新的类时序注意力网络来进行图像描述, 该网络扩展了传统的注意力机制, 以增强注意力在不同时刻与隐藏状态之间的相关性.

2) 本文提出注意力融合槽 (Attention Fusion Slot, AFS) 的概念, 它可以用于实现不同时刻注意力结果之间的跳跃连接. 我们设计了隐藏状态开关 (Hidden State Switch, HSS) 来指导生成单词, 将其与AFS 相结合, 在一定程度上可以解决累积误差的问题.

3) 通过大量的实验对提出的模型进行了分析与验证. MSCOCO 数据集上的实验结果表明了所提出的基于类时序注意力机制的图像描述方法的有效性.

2 相关工作

2.1 图像描述

近年来, 随着深度学习技术的发展, 有关图像描述的文献越来越多. 早期的图像描述方法基于规则-模板[5,6], 是一种经典方法. 该方法通过目标检测技术[7–9] 将视觉概念, 对象和属性转换为单词和短语, 然后将它们组合成具有固定语言模型的句子. 另一种比较主流的方法是基于神经网络的编码器-解码器框架, 受机器翻译的启发发展而来[10]. 如何通过改进网络架构来提高模型性能已逐渐成为图像描述领域的主流研究方向.

当前, 最流行的图像特征提取工具是自下而上的注意力模型[11], 该模型已在许多文章中被使用, 本文也是如此. 相信随着技术的进一步发展, 更加有效的图像特征提取方法会被提出. 另外, 近年来出现了许多有关场景图的文章. Yang 等人[12] 通过图卷积将每个对象及其自身属性与其他对象之间的关系集成在一起, 搭建出场景图并规范化网络模型的输入. 同时, 提出了词典 D 的概念, 经过文本语料库训练之后, 再用来初始化描述模型, 目的是在语料库中引入一些先验知识. 场景图和先验知识的引入有效地促进了图像描述的发展. 当然, 图像描述领域最常见的文章是关于注意力机制的改进和网络结构的创新. 尽管场景图是一个非常热门的话题, 但由于发展刚起步不久, 相对而言, 此类论文较少. 此外, 强化学习已逐渐发展成为一种有效的模型性能改进方法. Rennie 等人[13] 使用强化学习来优化图像描述模型, 并提出了自关键序列训练 (SCST) 方法, 该方法使用测试阶段模型的输出对奖励进行归一化处理, 而不是评估基准模型的归一化奖励.

2.2 注意力机制

注意力模型 (Attention Mechanism, AM)[14,15] 最初用于机器翻译, 现已成为神经网络领域的重要概念. 如今, 注意力机制已成为深度学习神经网络的重要组成部分, 并且在自然语言处理, 统计学习, 语音翻译和计算机视觉领域具有大量的应用. 注意力机制源自人类的视觉直觉, 人类视觉快速扫描图像全局以获得需要关注的目标区域, 即所谓的关注焦点, 也即是目标区域具有更多的关注资源, 在抑制其他无用信息的同时, 更多地关注目标的详细信息. 注意力机制首先计算每个候选向量的重要性得分, 然后通过 Softmax 函数将其标准化为权重, 最后将这些权重应用于候选向量以生成注意力结果, 即加权平均向量[16]. 注意力机制有许多

扩展的变体. Yang 等人[17]提出了堆叠式注意力网络, 该网络通过多次迭代来实现对图像的区域关注. Lu 等人[18]提出了一种带有视觉标记的自适应注意力模型, 在每个时间步长, 模型都会决定是更依赖图像还是更依赖视觉标记. 此外, 视觉哨兵会存储解码器已经知道的信息. Chen 等人[19] 基于编码器-解码器模型层设计了空间和通道注意力卷积神经网络 (CNN), 该网络使得原始的 CNN 多层特征图能够自适应句子上下文. Vaswani 等人[20] 放弃了基于卷积神经网络 (CNN) 或循环神经网络 (RNN) 的传统编码器-解码器模型, 通过单独使用注意力, 在不影响最终实验结果的前提下达到减少计算量、提高并行效率的目的. Huang 等人 [21] 提出了一个“双重注意力”(AoA)模块, 该模块扩展了常规的注意力机制来进一步确定注意力结果和查询之间的相关性. 但是, 网络框架的创新和注意力机制的改进都相对比较简单, 同时, 注意力机制和循环神经网络结合不够紧密. 注意力本身没有时序性, 但是将其嵌入神经网络后, 我们可以认为该注意力具有时序性, 那么如何使注意力机制更有效地集成到神经网络中, 是一个值得思考的问题.

3 类时序注意力网络

3.1 整体框架

本文使用自下而上的注意力模型[11](由目标检测区域特征提取框架 Faster RCNN 和 ResNet-101 [22] CNN 组合而成) 来提取图像特征 V, 然后将所有视觉特征馈入 LSTM 进行字幕生成. 其中, 解码框架采用两个连续的 LSTM 作为循环单元, 并且对每一时刻的隐藏状态和图像特征施加注意力, 以增强它们之间的相关性. 由于单词是由隐藏状态来指导生成, 因此单词与图像特征之间的相关性也需要增强. 整个网络架构如图 1 所示.

给定一组图像特征V, 本文提出的描述模型仍使用传统的软注意力方法, 在生成自然语言的过程中给每个图像特征施加权重. 该模型主要由两个 LSTM 层组成. 本文将在 3. 2 节详细介绍注意力融合槽如何与两个 LSTM 层组合生成单词. 首先, 本文通过以下公式表示当前 LSTM 的隐藏状态:

$$h_t = f_{\mathrm{LSTM}}(x_t, h_{t-1}, \hat{v}_{t-1}) \tag{1}$$

其中, x_t是 LSTM 在时刻t的输入向量, h_{t-1}是 LSTM 在时刻t的输出向量, $\hat{v}_{t-1}$表示上一时刻的注意力结果, 初始化为 0. 为了方便表示, 本文对于 LSTM 存储单元的单元状态忽略不计, 统一使用式 (1) 表示在时刻t处每一层 LSTM 的输入和输出向量.

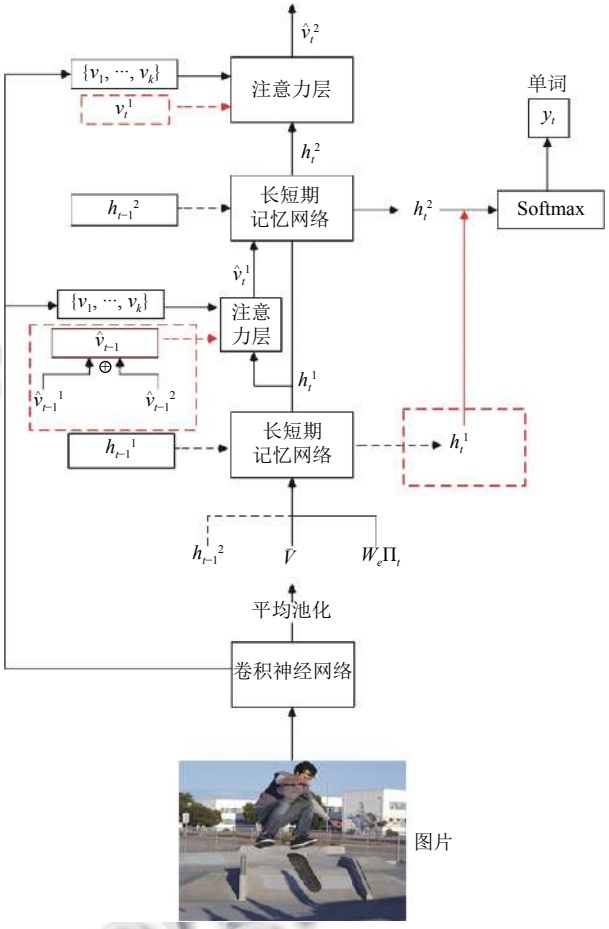

图 1 网络架构

3.2 类时序注意力层

对于描述模型, 本文将第 1 个 LSTM 层称为类时序注意力层, 将第 2 个 LSTM 层称为语言注意力层, 使用V表示图像特征. 类时序注意力模型的输入如 3.1 节所示, 是通过前一时刻语言注意力模型的输出向量与均值特征级联运算得到, 表示为$\bar{v} = \frac{1}{k}\sum_i v_i$. 输入如下式所示:

$$x_t^1 = \left[h_{t-1}^2, \bar{v}, W_e \prod_t\right] \tag{2}$$

其中, $W_e \in \Re^{E\times|\Sigma|}$是词典$\sum$的词嵌入矩阵, $\prod_t$是时刻t的独热编码. 这些输入为当前时刻类时序注意力层提供特征信息和上一时刻语言注意力层的隐藏状态信息. 另外, 词嵌入矩阵是从随机初始化中学习的, 无需预先

训练.

当在时刻t获得类时序注意力层的输出h_t^1时, 对相应的k个图像特征v_i施加注意力权重$\alpha_{i,t}^1$. 同时, 为了加强隐藏状态与图像特征之间的相关性, 我们通过注意力融合槽 (AFS) 将前一时刻类时序注意力层和语言注意力层的输出引入到当前时刻. 如图 2 所示, 其具体公式如下:

$$\hat{v}_{t-1} = \lambda_1 \hat{v}_{t-1}^1 + \lambda_2 \hat{v}_{t-1}^2 \tag{3}$$

$$a_{i,t} = w_a^T \tanh\left(W_v v_i + W_h h_t^1 + W_h \hat{v}_{t-1}\right) \tag{4}$$

$$\alpha_t = Softmax(a_t) \tag{5}$$

其中, $W_v \in \Re^{H\times V}$, $W_h \in \Re^{H\times M}$和$W_a \in \Re^H$分别是学习参数. 对图像特征和隐藏状态施加注意力之后, 用$\hat{v}_t^1$表示类时序注意力层的输出结果, 用$\hat{v}_t^2$表示语言注意力层的输出结果. λ_1和λ_2是超参数, 设置为 0. 5. 另外, $\hat{v}_t$是由下述公式计算得到:

$$\hat{v}_t = \sum_{i-1}^{k} \alpha_{i,t} v_i \tag{6}$$

3.3 语言注意力层

语言注意力层的输入由施加注意力之后的图像特征和类时序注意力层的输出级联而成, 用下式表示:

$$x_t^2 = \left[\hat{v}_t^1, h_t^1\right] \tag{7}$$

本文认为前一时刻 LSTM 隐藏状态中包含的信息对当前时刻单词的生成具有促进作用. 为了充分利用 LSTM 隐藏状态之间的关系, 本文设计了隐藏状态开关 (HSS), 如图 3 所示. 计算公式如下.

$$h = \begin{cases} h_t^2, S_h = 0 \\ h_t^2 + \lambda_h h_t^1, S_h = 1 \end{cases} \tag{8}$$

其中, λ_h是学习参数, S_h表示 HSS 的状态, $S_h = 0$ 表示 HSS 的状态为“OFF”, $S_h = 1$ 表示 HSS 的状态为“ON”. 对于单词序列 $(y_1, \cdots, y_T)$, 本文使用符号$y_{1:T}$统一进行表示. 通过以下公式来表示在时间步长t处单词分布的概率:

$$P(y_t|y_{1:t-1}) = Softmax(W_y h + b_y) \tag{9}$$

其中, $W_y \in \Re^{|\Sigma|\times M}$和$b_h \in \Re^{|\Sigma|}$是学习权重和偏差. 句子概率分布计算公式如下:

$$P(y_{1:T}) = \prod_{t=1}^{T} p(y_t|y_{1:t-1}) \tag{10}$$

3.4 目标函数

在训练过程中, 对于给定的标签序列$y_{1:T}^*$和带有参数η的字幕模型, 本文仍然使用最小化交叉熵损失:

$$L_{XE}(\eta) = -\sum_{t=1}^{T} \log\left(p_\eta\left(y_t^* | y_{1:t-1}^*\right)\right) \tag{11}$$

交叉熵训练结束后, 本文将采用目前比较流行的强化学习方法来训练和优化最终模型. 为了尽量减少负面期望得分, 对交叉熵训练得到的最终模型进行重新训练和初始化. 计算公式如下:

$$L_R(\eta) = -E_{y_{1:T}\sim p_\eta}[S_r(y_{1:T})] \tag{12}$$

其中, S_r是得分函数 (例如 CIDEr). 这种损耗的梯度可以近似为:

$$\nabla_\eta L_R(\eta) \approx -\left(S_r\left(y_{1:T}^s - S_r(\hat{y})\right)\right)\nabla_\eta \log\left(p_\eta\left(y_{1:T}^s\right)\right) \tag{13}$$

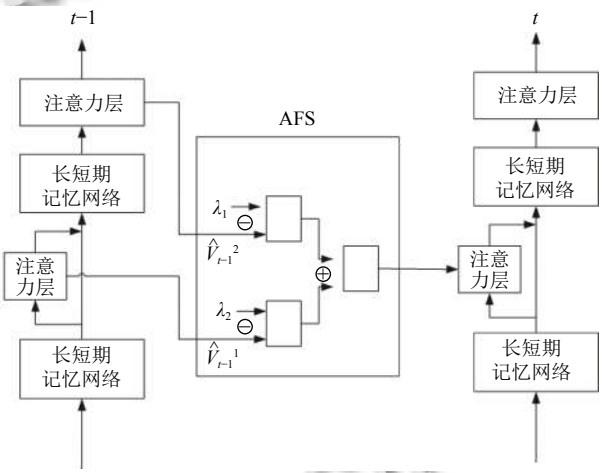

图 2 类时序注意力模型

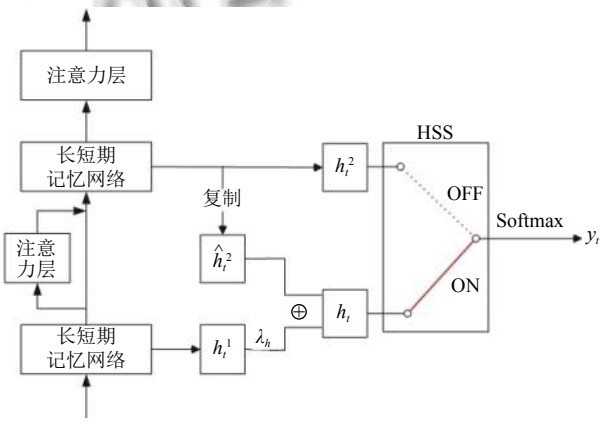

图 3 语言注意力模型

4 实验

4.1 数据集

本文在图像描述领域官方数据集 MSCOCO 上评

估和验证基于类时序注意力机制的图像描述模型 [23]. MSCOCO 数据集包含 123 287 张图片, 其中 82 783 张图片作为训练集, 40 504 张图片作为验证集, 每张图片对应 5 个标签. 此外, 为了方便研究人员评估模型, MSCOCO 官方提供了 40 775 张图片作为在线测试集. “Karpathy”数据集 [24] 用于模型线下评估和测试, 其中 5000 张图片作为验证集, 5000 张图片作为测试集, 其余图片作为训练集. 本文首先将所有标签语句转换为小写, 然后过滤掉出现次数少于 5 次的单词, 最后得到一个含有 9487 个单词的字典. 在实验过程中, 使用领域常用评估策略, 包括 BLEU [25], METEOR [26], ROUGE-L [22], CIDEr [27] 和 SPICE [28], 来评估所提出的方法, 并与其他方法进行比较.

其中, BLEU 为机器翻译中常用的双语精度评估方法, 是用于评估模型生成的句子和实际句子的差异的指标, 取值范围在 0.0 到 1.0 之间, 如果两个句子完美匹配, 那么 BLEU 是 1.0, 反之, BLEU 为 0.0. METEOR 是精度召回率评估方法, 基于单精度的加权调和平均数和单字召回率, 解决一些 BLEU 标准中固有的缺陷, 也是机器翻译常用的评估方法之一. ROUGE-L 是召回率评估方法, 采用召回率作为指标, 将模型生成的句子与实际句子的 n 元组贡献统计量作为评判依据. CIDEr 是基于共识的图像描述评估方法, 将句子看作“文档”, 并将其表示成向量, 然后计算实际句子与模型生成的句子的余弦相似度, 作为打分. SPICE 是基于语义的图像描述评估方法, 以名词为中心, 通过度量实际句子与模型生成句子的场景图相似度来对两个句子做语义匹配.

4.2 实验结果

如图 4 所示, 是本文提出的方法训练的模型与基线模型在 MSCOCO 数据集上的结果比较, 可以看出, 对于同一张图片, 该模型生成的描述与图片内容契合度更高, 语言的准确性和流利性更好.

如表 1 所示, 对于所提出的方法, 本文在 MSCOCO 数据集上进行了离线测试. 实验结果表明, 与基线模型 (Top-Down 模型)[11] 和其他方法相比, 本文训练的模型具有更优越的性能. 从表 1 的离线测试结果中可以看到, 与基线模型相比, 本文的方法训练的模型的评估指标都有所提高, 尤其是 CIDEr 提高了 2.7 个百分点. 本文训练的模型通过 AFS 使注意力机制具有了时序性, 可以与循环神经网络更加紧密地连接, 产生包含更丰富有效信息的隐藏状态向量, 从而生成更高质量的自然语言描述.

图 4 实验结果对比

表 1 MSCOCO 数据集上的实验结果对比

算法	B-1	B-4	METEOR	ROUGE-L	CIDEr	SPICE
LSTM[29]	—	31.9	25.5	54.3	106.3	—
SCST[13]	—	34.2	26.7	55.7	114.0	—
LSTM-A[30]	—	35.5	27.3	56.8	118.3	20.8
Up-Down[8]	79.8	36.3	27.7	56.9	120.1	21.4
RFNet[31]	79.1	36.5	27.7	57.3	121.9	21.2
类时序注意力网络	**79.3**	**36.7**	**27.8**	**57.6**	**122.8**	—

4.3 实验分析

本文在 Top-Down 模型的基础上, 进一步完善了注意力机制, 在 MSCOCO 官方数据集上取得了较好的结果. 在实验过程中, 我们发现语言注意力层的隐藏状态和类时间注意力层的隐藏状态可以按一定比例融合以获得新的状态向量. 此向量生成单词的质量比单独使用语言注意力层的隐藏状态略好. 因此, 我们设计了 HSS 来微调隐藏状态. 表 2 是 HSS 状态对模型性能的影响.

表 2 HSS 对模型的影响

HSS	B-1	B-4	METEOR	ROUGE-L	CIDEr
ON	79.3	36.7	27.8	57.6	122.8
OFF	79.3	36.9	27.7	57.5	122.5

在实验过程中, 如表 3 所示, 本文选择了 4 个模型进行集成实验, 分别为$Model_1$, $Model_2$, $Model_3$, $Model_4$, 相应的集成权重参数分别为 m_1, m_2, m_3, m_4. 调参过程如表 4 所示. 为方便起见, 本文设置 HSS 的状态为“ ON”.

$$Model_a = \{Model_1, Model_2, Model_3\} \quad (14)$$

$$Model_b = \{Model_1, Model_2, Model_4\} \quad (15)$$

$$m_a = \{m_1, m_2, m_3\} \quad (16)$$

$$m_b = \{m_1, m_2, m_4\} \quad (17)$$

其中, $Model_a$表示用$Model_1$, $Model_2$和$Model_3$做集成实验, $Model_b$表示用$Model_1$, $Model_2$和$Model_4$做集成实验, m_a, m_b表示模型对应的权重参数.

从表 4 中不难看出, 对于参与集成的模型, 性能最佳的模型将被赋予最高的权重, 性能稍低的模型将被赋予较低的权重, 这样可以获得比较理想的集成效果.

表 3 选取 4 个模型进行集成实验

模型	B-1	B-4	METEOR	ROUGE-L	CIDEr
$Model_1$	79.3	36.6	27.6	57.4	121.8
$Model_2$	79.1	36.4	27.7	57.4	122.3
$Model_3$	78.8	36.1	27.8	57.3	122.2
$Model_4$	79.2	36.5	27.7	57.3	121.6

表 4 调参过程及实验结果

模型	权重参数	B-1	B-4	ROUGE-L	CIDEr
$Model_a$	{0.3,0.3,0.3}	79.2	36.6	57.5	122.6
	{0.3,0.4,0.3}	79.3	36.7	57.6	122.8
	{0.3,0.5,0.3}	79.2	36.6	57.5	122.6
$Model_b$	{0.3,0.4,0.3}	79.4	27.8	57.5	122.6
	{0.4,0.4,0.3}	79.4	27.8	57.5	122.5
	{0.4,0.3,0.3}	79.4	27.7	57.4	122.3

5 结论与展望

本文提出了一种新型类时序注意力网络用于图像描述, 该网络扩展了传统的注意力机制, 以增强注意力结果与隐藏状态在不同时刻之间的相关性. 此外, 提出“注意力融合槽”(AFS) 的概念, 用于实现不同时刻注意力结果之间的跳跃连接. 设计隐藏状态开关, 用于指导单词的产生, 结合 AFS 在一定程度上解决了累积误差的问题. 同时, 进行了广泛的实验验证分析该方法. 在未来的工作中, 本团队会继续研究注意力机制和模型框架的改进方式, 并考虑引入场景图来进一步提升模型性能.

参考文献

1 Kulkarni G, Premraj V, Ordonez V, *et al*. BabyTalk: Understanding and generating simple image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2891–2903. [doi: 10.1109/TPAMI.2012.162]

2 Yang YZ, Teo CL, Daumé IIIH, *et al*. Corpus-guided sentence generation of natural images. Proceedings of 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK. 2011. 444–454.

3 Mitchell M, Dodge J, Goyal A, *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.

4 Fang H, Gupta S, Iandola F, *et al*. From captions to visual concepts and back. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, UK. 2015. 1473–1482.

5 Yao BZ, Yang X, Lin L, *et al*. I2T: Image parsing to text description. Proceedings of the IEEE, 2010, 98(8): 1485–1508. [doi: 10.1109/JPROC.2010.2050411]

6 Socher R, Li FF. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA. 2010. 966–973.

7 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

8 Wan F, Wei PX, Jiao JB, *et al*. Min-entropy latent model for weakly supervised object detection. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 1297–1306.

9 Wan F, Liu C, Ke W, *et al*. C-MIL: Continuation multiple instance learning for weakly supervised object detection. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA. 2019. 2194–2203.

10 Cho K, van Merriënboer B, Gulcehre C, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar. 2014. 1724–1734.

11 Anderson P, He XD, Buehler C, *et al*. Bottom-up and top-down attention for image captioning and visual question answering. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City,

UT, USA. 2018. 6077–6086.
12 Yang X, Tang KH, Zhang HW, *et al*. Auto-encoding scene graphs for image captioning. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA. 2019. 10677–1068.
13 Rennie SJ, Marcheret E, Mroueh Y, *et al*. Self-critical sequence training for image captioning. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 1179–1195.
14 Rensink RA. The dynamic representation of scenes. Visual Cognition, 2000, 7(1–3): 17–42. [doi: 10.1080/135062800394667]
15 Corbetta M, Shulman GL. Control of goal-directed and stimulus-driven attention in the brain. Nature Reviews Neuroscience, 2002, 3(3): 201–215. [doi: 10.1038/nrn755]
16 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. Lille, France. 2015. 2048–2057.
17 Yang ZC, He XD, Gao JF, *et al*. Stacked attention networks for image question answering. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 21–29.
18 Lu JS, Xiong CM, Parikh D, *et al*. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 3242–3250.
19 Chen L, Zhang HW, Xiao J, *et al*. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 6298–6306.
20 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, CA, USA. 2017. 6000–6010.
21 Huang L, Wang WM, Chen J, *et al*. Attention on attention for image captioning. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea. 2019. 4633–4642.
22 Lin CY, Gao JF, Cao GH, *et al*. Automatic evaluation of summaries: USA, 20080189074. (2008-08-07).
23 Lin TY, Maire M, Belongie S, *et al*. Microsoft COCO: Common objects in context. Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland. 2014. 740–755.
24 Karpathy A, Li FF. Deep visual-semantic alignments for generating image descriptions. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 3128–3137.
25 Papineni K, Roukos S, Ward T, *et al*. Bleu: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, PA, USA. 2002. 311–318.
26 Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, MI, USA. 2005. 65–72.
27 Vedantam R, Zitnick CL, Parikh D. CIDEr: Consensus-based image description evaluation. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 4566–4575.
28 Anderson P, Fernando B, Johnson M, *et al*. SPICE: Semantic propositional image caption evaluation. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands. 2016. 382–398.
29 Vinyals O, Toshev A, Bengio S, *et al*. Show and tell: A neural image caption generator. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA. 2015. 3156–3164.
30 Yao T, Pan Y, Li Y, *et al*. Boosting image captioning with attributes. IEEE International Conference on Computer Vision. IEEE Computer Society, 2017. 4904–4912.
31 Jiang W, Ma L, Jiang YG, *et al*. Recurrent fusion network for image captioning. Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany. 2018. 499–515.