DOI:10.16452/j.cnki.sdkjzk.2021.06.008 文章编号:1672-3767(2021)06-0061-08

基于文本信息补充的图像描述模型

花 嵘[1],仪秀龙[1],郑杜磊[1],王志余[2]
(1.山东科技大学 计算机科学与工程学院,山东 青岛 266590;2.山东省青岛市黄岛区第一人民医院,山东 青岛 266555)

摘 要:基于 encoder-decoder 的深度神经网络在图像描述任务中获得了很好的表现,LSTM 解决梯度消失的良好能力使其成为解码器的主流。LSTM 的门控机制较好地解决了 RNN 的长期依赖问题,但该机制对信息的筛选导致信息缺失,使得 LSTM 隐藏单元表达能力不足,出现 LSTM 输入信息缺失、预测信息不充分问题。为解决这两个问题,提出两种基于文本信息补充的图像描述模型:输入信息补充(IIS)模型通过信息提取函数提取更多的文本信息作为输入,解决 LSTM 输入信息缺失问题;输出信息补充(OIS)模型通过信息提取函数提取多个时间步的隐藏单元信息进行预测,解决 LSTM 预测信息不充分问题。实验证明,在 AI CHALLENGER 测试集中,两种模型均显著地提高了各项评价指标。

关键词:长短时记忆网络;图像描述;文本信息补充;信息提取函数;信息缺失
中图分类号:Q936 **文献标志码**:A

Image captioning models based on text information supplement

HUA Rong[1], YI Xiulong[1], ZHENG Dulei[1], WANG Zhiyu[2]
(1. College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, Shandong 266590, China;
2. Huangdao District First People's Hospital, Qingdao, Shandong 266555, China)

Abstract: The deep neural network based on encoder-decoder has achieved good performance in the image captioning task, long short term memmory(LSTM) has an excellent ability to solve gradient disappearance, making it the mainstream of decoders. The gating mechanism of LSTM better solves the long-term dependence of recurrent neural network(RNN). However, the gating structure screening leads to the lack of information, which makes the expression ability of the LSTM hidden unit insufficient, which in turn makes the LSTM input information missing and insufficient prediction information. To solve these two problems, this paper proposes two LSTM image captioning models based on text information supplements. The input information supplement (IIS) model uses the information extraction function to extract more text information as input to solve the missing of LSTM input information. The output information supplement (OIS) model uses the information extraction function to extract the hidden unit information of multiple time steps for prediction to solve the insufficient LSTM prediction information. Experiments show that both modelshave significantly improved the various evaluation indicators in the AI CHALLENGER test set.

Key words: LSTM; image caption; information supplement; information extraction function; information missing

自动生成图像的描述是计算机视觉的一项基础任务,其目的是识别图像内突出的目标、理解目标之间的关系,最终以人类可以理解的自然语言对其进行表达。自然语言与机器语言的巨大差异使得图像描述成为

收稿日期:2021-05-21
基金项目:国家重点研发计划项目(2016YFB020803);山东省重点研发计划项目(2019GGX101066)
作者简介:花 嵘(1969—),男,江苏常州人,副教授,博士,主要从事高性能计算、机器学习等方面研究.E-mail:hrfy@263.net
仪秀龙(1996—),男,山东潍坊人,硕士研究生,主要从事机器学习等方面研究.

一项困难的任务,但其在图像视频检索、协助视障群体感知环境等领域具有广泛的应用价值,吸引了学术界和工业界的广泛兴趣。图像描述任务作为跨学科领域的交叉研究问题,将计算机视觉与自然语言处理联合起来,其目标是自动生成图像的描述,难点在于要使计算机"看到"可见的目标并"理解"不可见的目标关系,难度超过图像分类和目标检测。

机器翻译任务采用编码解码的框架,利用循环神经网络(recurrent neural network,RNN)进行编码及解码,最大化 $P(S \mid T)$,将源语言中的语句 T 转化为目标语言的语句 S。受其启发,图像描述任务考虑到卷积神经网络强大的图像特征提取能力,选择卷积神经网络(convolutional neural network,CNN)作为编码器,RNN 作为解码器。长短时记忆(long short term memory,LSTM)[1]因具有良好的解决梯度消失的能力成为解码器的首选。LSTM 创造性提出了"门控"思想,依靠记忆单元和遗忘门可以有选择地记忆和遗忘信息,但门控结构对信息的筛选会导致信息的遗失,使 LSTM 隐藏单元的表达能力不足。LSTM 隐藏单元表达能力不足会产生两个问题:一是输入信息缺失,二是预测信息不充分。本研究针对 LSTM 对文本信息的提取问题,提出两种文本信息补充模型,两种模型均强调文本信息在模型中起到的重要作用。

1 相关工作

近年来,研究者们提出了很多生成图像描述的方法,主要分为以下 3 类:①基于模板的方法。利用固定的模板和空白槽来生成标题,把检测到的对象、动作、属性对空白槽进行填充。例如,Kulkarni 等[3]在填补空缺之前利用条件随机场来预测对象以及属性,Li 等[4]利用提取与检测到的对象属性及相关关系的句子来生成图像的描述。该类方法可以生成语法正确的标题,但由于模板是预定义的,因此生成标题较为死板,不具备良好的泛化性;②基于检索的方法。把视觉上相似图像的标题作为候选标题,从候选标题中选择并进行简单调整以生成目标标题[5-7]。此类方法可以生成较为灵活的描述,但过于依赖现有的人工描述,难以生成新颖的描述,同样不具备良好的泛化性,并且该方法需要收集大量且全面的人工描述,训练集也需要多样化;③基于神经网络的方法。该类方法受到机器翻译的启发,将图像描述视为从图像到文本的翻译任务,利用 LSTM 作为解码器,克服上述两类方法的局限性。深度神经网络在计算机视觉、自然语言处理等领域得到广泛应用,并取得了突出的成果。

注意力机制在目前主流的深度神经网络方法中得到了广泛应用,其核心目标是从众多信息中选择出对当前任务目标最关键的视觉信息。Xu 等[8]将 Soft-Attention 应用在图像描述任务中,让模型在预测某个单词时,将视觉的重点放在图像的某一部分而不是整幅图像。Lu 等[9]提出的 Adaptive Model 让模型进行预测时判断依赖文本信息或是视觉信息。Anderson 等[10]提出的 Bottom-Up and Top-Down Attention 将候选图像特征变为用目标检测之后得到的属性特征。Wang 等[11]提出的 Hierarchical Attention 使注意力可以同时在多个特征上进行层次计算。You 等[12]提出 Semantic Attention,使模型能够最大化的获取语义信息。以上模型基于 LSTM 提取文本信息,使用注意力机制对模型注入视觉信息,模型性能取决于对视觉信息的利用,但均忽视了文本信息的重要作用。

为了解决 LSTM 隐藏单元表达能力不足导致的两个问题,本研究提出两种基于文本信息补充的图像描述模型:一种输入信息补充(input information supplement,IIS)模型,利用信息提取函数提取更多的文本信息作为输入,赋予 LSTM 更多的输入信息,解决 LSTM 的输入信息缺失问题;另一种输出信息补充(output information supplement,OIS)模型,通过信息提取函数,在多个时间步的隐藏单元中获取所需要的预测信息,解决 LSTM 预测信息不充分问题。最后,在 Neural Image Caption[2]的基础上,实现了上述两种模型并评估了两种模型的有效性,实验证明,两种模型均可以明显提高各项评价指标。

2 基于文本信息补充的图像描述模型

为解决现有的 LSTM 存在的输入信息缺失和预测信息不充分问题,本研究在 Neural Image Caption[2]的基础上,提出两种基于文本信息补充的图像描述模型。

2.1　图像描述的传统 encoder-decoder 架构

先介绍用于图像描述的传统 encoder-decoder 架构,模型结构见图 1。主要计算公式如下:

$$x_{-1}=CNN(I)\ , \tag{1}$$

$$x_t=W_eS_t\ , \tag{2}$$

$$P_{t+1}=LSTM(x_t)\ 。 \tag{3}$$

其中:I 代表输入的图像,图像经过卷积神经网络得到特征向量,并作为解码器的第一次输入,用来告诉解码器图像的内容;每个词汇 S_t 用 one-hot 向量表征,向量的维度等于词典的大小,但由于词汇向量的维度太大,研究者将词汇通过词嵌入 W_e 映射到低维度空间,得到每个时间步的输入 x_t ;P_{t+1} 为模型在每个时间步得到所有单词的概率分布。

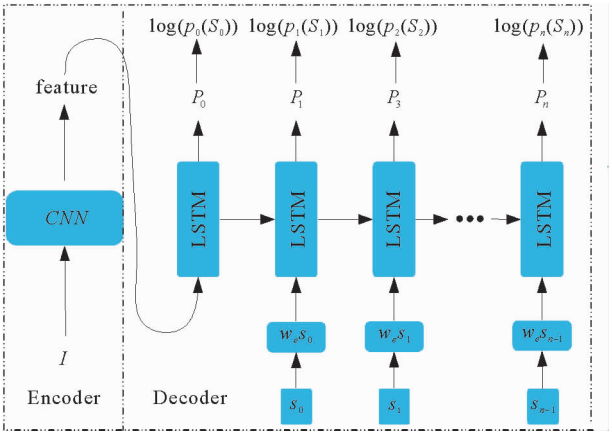

图 1　图像描述的传统编码解码模型

Fig. 1　Traditional encoding and decoding model of image description

采用如下损失函数描述预测标题与人工标题的差别:

$$L(I,S)=-\sum_{t=1}^{n}\log P_t(S_t)\ 。 \tag{4}$$

通过对卷积神经网络、词嵌入、LSTM 的所有参数进行优化,使上述损失最小。

2.2　基于文本信息补充的图像描述模型

通过公式(5)用端到端的方式最大化给定图像的正确描述概率:

$$\theta^*=\arg\max_{\theta}\sum_{(I,S)}\log p(S\mid I;\theta)\ , \tag{5}$$

其中:S 代表图像的描述,I 代表图像,θ 代表需要学习的参数。理论上图像描述的生成过程如下:

$$\log P(S\mid I;\theta)=\sum_{i=1}^{n}\log P(S_i\mid S_0,\cdots,S_{i-1};I;\theta)\ 。 \tag{6}$$

以链式规则计算 $S_0,\cdots,S_n$ 上的联合概率。实际操作中使用 RNN 对式(6)进行建模:

$$h_t=RNN(S_{t-1},h_{t-1})\ , \tag{7}$$

$$p(S_t\mid S_{t-1}\cdots S_0)=p(S_t\mid h_t)\ 。 \tag{8}$$

即

$$p(S_t\mid S_{t-1}\cdots S_0)=p(S_t\mid S_{t-1},h_{t-1})\ 。 \tag{9}$$

RNN 在时刻 t 得到了 S_{t-1} 的信息,$S_{t-1},\cdots,S_0$ 的信息是通过 RNN 的隐藏状态 h_{t-1} 进行表达的。考虑到 LSTM 具有良好的解决梯度消失的能力,研究者将 RNN 替换为 LSTM。LSTM 提出了“门控”的思想,通过输入、输出、遗忘门来获得所需要的信息。LSTM 核心公式如下:

$$i_t=\sigma(W_{ix}x_t+W_{ih}h_{t-1})\ , \tag{10}$$

$$f_t=\sigma(W_{fx}x_t+W_{fh}h_{t-1})\ , \tag{11}$$

$$o_t=\sigma(W_{ox}x_t+W_{oh}h_{t-1})\ , \tag{12}$$

$$c_t=f_t\odot c_{t-1}+i_t\odot\tanh(W_{cx}x_t+W_{cf}h_{t-1})\ , \tag{13}$$

$$h_t=o_t\odot c_t\ 。 \tag{14}$$

LSTM 模拟人类大脑的遗忘记忆过程,其记忆由前一时刻的记忆经过遗忘处理和当前时刻的输入信息组成。门控结构使得 LSTM 可以对信息进行有选择的筛选,解决了梯度消失和长期依赖问题,但门控结构对信息的筛选会导致信息遗失,使得 LSTM 隐藏单元表达能力不足,进而使得 LSTM 输入信息缺失和预测信息不充分。事实上,在 t 时刻可以直接得到前 K 个状态的信息而不需要通过 LSTM 的记忆单元。通过信息提取函数 f_1 提取 K 个状态的信息,作为对记忆单元的补充信息输入 LSTM,图像描述的 IIS 模型为:

$$S_i=W_eS_i\ , \tag{15}$$

$$x_t = f_1(S_{t-k}, \cdots, S_{t-1}) , \tag{16}$$

$$h_t = LSTM(x_t) 。 \tag{17}$$

式(16)中采用的信息提取函数 f_1 为拼接函数,即:

$$x_t = [S_{t-k}, \cdots, S_{t-1}] 。 \tag{18}$$

本研究提出的 IIS 模型结构见图 2。考虑到 t 时刻的输出最可能与 t 时刻之前的 K 个状态有关,将这 K 个状态用信息提取函数 f_1 提取所需要的信息后作为输入。通过使用更多时间步的文本信息,解决由门控结构导致的 LSTM 输入信息缺失问题。IIS 证明了记忆单元会遗忘之前时刻的记忆信息,使得 LSTM 隐藏单元信息缺失,那么利用 LSTM 的隐藏单元进行预测必然是不合理的。为此考虑通过信息提取函数提取更多时间步的隐藏单元信息来进行信息补充,为模型的预测提供充足的信息,保证模型预测的准确性。

OIS 模型结构图见图 3,具体计算方式如下:在 t 时刻得到当前时刻及之前时刻共 L 个隐藏单元值,通过信息提取函数 f_2 来获得这 L 个时间步的信息,得到 OIS 补充模型:

$$\hat{h}_t = f_2(h_{t-L+1}, \cdots, h_t) , \tag{19}$$

$$p(S_t) = soft\max(\hat{h}_t) 。 \tag{20}$$

式(19)采用的信息提取函数 f_2 也是拼接函数,即:

$$\hat{h}_t = [h_{t-L+1}, \cdots, h_t] 。 \tag{21}$$

可见,OIS 模型通过更多的隐藏单元进行预测,可以较好地解决 LSTM 预测信息不充分问题。

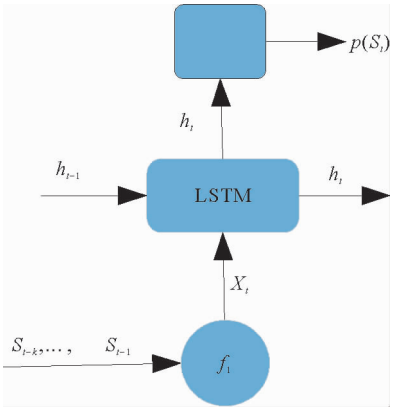

图 2 输入信息补充模型结构图

Fig. 2 Structure diagram of IIS model

3 实验

3.1 实验细节

在 2017 年提出的人为标注的 AI CHALLENGER 大规模中文数据集上分别评估了本研究提出的两种模型。该数据集有训练集 21 万张图片,验证集 3 万张图片,每张图片有 5 个描述。去除出现次数低于 2 次的词汇,最终得到 9 813 个词汇,使用 BLEU1-4[13]、CIDER[14]、ROUGE-L[15]等不同的度量指标来评估,并与其他经典模型进行比较。为更好地与各种经典模型对比,所有实验都采用同样的参数。用预训练的 Resnet-50[16]来获得图像的 2 048 维特征向量,并将其投影到一个新的维数为 256 的空间,这也是解码器双层 LSTM 的隐藏单元维数,在训练过程中使用的目标函数为交叉熵损失函数[17],使用 ADAM 优化器,学习率设置为 0.001,权重衰减设置为 0.000 1,批量大小设置为 64,Epoch 设置为 40。两种模型在单个 NVIDIA-Tesla K80 GPU 上训练大约 34 h。

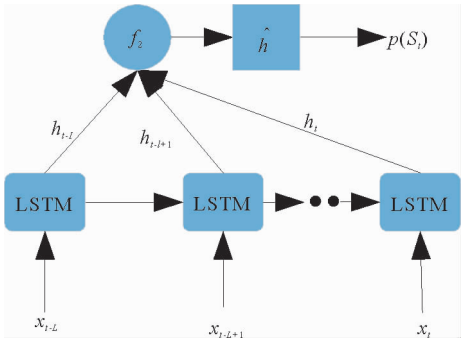

图 3 输出信息补充模型结构图

Fig. 3 Structure diagram of OIS model

3.2 定量分析

3.2.1 IIS 模型

针对由遗忘门所导致的 LSTM 输入信息缺失问题,上节提出了一种 IIS 模型,利用信息提取函数提取 K 个词向量的文本信息作为 LSTM 的输入。为找到最理想的 K,进行了实验,结果见表 1。

表 1　LSTM 各阶输入信息补充模型指标

Tab. 1　Input information of each order of LSTM complements the model indexes

K	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDER
1(NIC)	0.682	0.510	0.378	0.275	0.360	1.677
2	0.689	0.519	0.385	0.282	0.363	1.718
3	0.691	0.519	0.385	0.282	0.365	1.718
4	0.696	0.523	0.388	0.284	0.366	1.727
5	0.692	0.520	0.385	0.282	0.366	1.744
6	**0.705**	**0.531**	**0.395**	**0.290**	**0.370**	**1.758**
7	0.695	0.523	0.388	0.286	0.368	1.752
8	0.699	0.529	0.394	0.289	0.369	1.730
9	0.699	0.529	0.394	0.290	0.369	1.720
10	0.697	0.525	0.390	0.285	0.368	1.745
11	0.696	0.524	0.391	0.287	0.369	1.744
12	0.695	0.526	0.391	0.286	0.367	1.741
13	0.699	0.527	0.392	0.288	0.370	1.751
14	0.699	0.527	0.393	0.290	0.368	1.755

数据集的句子长度大多在 15 左右,为避免取到局部极值,进行了所有时间步的实验。由表 1 看出,随着 K 的增加,LSTM IIS 模型在各指标上的表现越来越好,证明 LSTM 存在重要信息缺失,通过增加更多文本信息作为 LSTM 的输入,可以解决 LSTM 输入信息缺失问题;$K=6$ 时,模型性能达到顶峰;当增加过多的文本信息即 $K>6$ 时,信息冗余对模型产生误导,导致 K 取更大值时,模型在各指标上的表现反而有所下降。故取 $K=6$。

3.2.2　OIS 模型

第二节针对用来预测的 LSTM 隐藏单元存在的信息缺失问题,提出了 LSTM OIS 模型,利用信息提取函数提取 L 个时间步的隐藏单元信息进行预测。通过实验找到最理想的 L,完整的实验数据见表 2。

表 2　LSTM 各阶 OIS 模型指标

Tab. 2　Output information of each order of LSTM complements the model indexes

L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDER
1(NIC)	0.682	0.510	0.378	0.275	0.360	1.677
2	0.685	0.511	0.378	0.275	0.360	1.675
3	0.689	0.516	0.382	0.278	0.364	1.693
4	0.692	0.522	0.387	0.285	0.364	1.714
5	0.692	0.520	0.385	0.282	0.366	1.744
6	0.688	0.515	0.378	0.275	0.362	1.690
7	**0.696**	**0.524**	**0.390**	**0.287**	**0.365**	**1.740**
8	0.692	0.516	0.381	0.276	0.364	1.690
9	0.691	0.519	0.384	0.280	0.364	1.689
10	0.692	0.519	0.384	0.280	0.365	1.697
11	0.687	0.517	0.383	0.279	0.362	1.705
12	0.690	0.515	0.378	0.275	0.361	1.711
13	0.694	0.519	0.384	0.280	0.362	1.719
14	0.691	0.519	0.382	0.278	0.362	1.701

为避免取到局部极值,与 IIS 模型类似,OIS 模型实验也取了所有时间步的值。由表 2 可以看出,当 $L=7$时,OIS 模型可以获得最好的实验效果。实验结果证明,通过采用更多的隐藏单元信息可以解决 LSTM 用来预测的隐藏单元信息缺失问题。

3.2.3 实验结果与分析

将本研究提出的模型与几种经典的图像描述模型进行性能对比,结果见表 3。

表 3 与各种经典模型对比

Tab. 3 Comparison of performance of various classic models

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDER
LRCN 2u[18]	0.660	0.498	0.373	0.261	—	—
LRCN 2f[18]	0.664	0.493	0.374	0.267	—	—
NIC[2]	0.682	0.510	0.378	0.275	0.360	1.677
Soft-Att[8]	0.685	0.518	0.380	0.280	0.363	1.710
Spatial[9]	0.694	0.524	0.388	0.281	0.365	1.726
Sentinel[9]	0.699	0.527	0.393	0.289	0.368	1.753
OIS Model	0.696	0.524	0.390	0.287	0.365	1.740
IIS Model	0.705	0.531	0.395	0.290	0.370	1.758

由表 3 可以看出,IIS 模型和 OIS 模型在原模型 NIC 的基础上,性能都有了较大的提升。其中 IIS 模型效果更是超越了几种经典的注意力机制模型。因此,图像描述同时依赖文本信息与视觉信息,LSTM 的信息缺失问题带来的性能瓶颈可以通过文本信息补充模型解决。

3.3 定性分析

为了对本研究得到的模型效果进行定性分析,对 6 个图像分别利用 3 种模型进行对比(如图 4)。

由图 4 可见,本研究提出的 IIS 模型由于采用更多的文本信息作为输入,模型获得更丰富的输入信号,生成的描述更加精细、饱满。图片中的草原、大厅、足球场、球场、男人等词汇前都有形容词来修饰。本研究提出的 OIS 模型,由于用更多时间步的隐藏单元信息进行预测,预测更加准确,对图片中的典型目标如草原、帽子、骑马、挎包、大厅、女人、四个、足球场、踢足球等都很好进行了识别。

3.4 讨论

将 IIS 模型与 OIS 模型联合起来得到联合模型,模型结构图见图 5。

将表现最好的 IIS 模型($K=6$)分别与各阶 OIS 模型联合起来进行实验,结果见表 4。由表 4 可以看出,6 阶 IIS 模型与 5 阶 OIS 模型联合可以获得最好的效果,该结果比单独使用 OIS 模型略好,比 IIS 模型略差,说明联合模型并没有起到很好的促进作用。分析原因是当采用 LSTM 输入信息补充模型后,由于获得了充足的输入信息,可以较好地进行预测,若此时再使用多个隐藏单元值进行预测,其他时间步提供的无用信息会大于有用信息,对模型产生误导,导致模型的效果不佳。实验结果再次证明信息冗余会导致模型的性能下降。

图 4　三种模型生成的描述对比

Fig. 4　Comparison of the descriptions generated by the three models

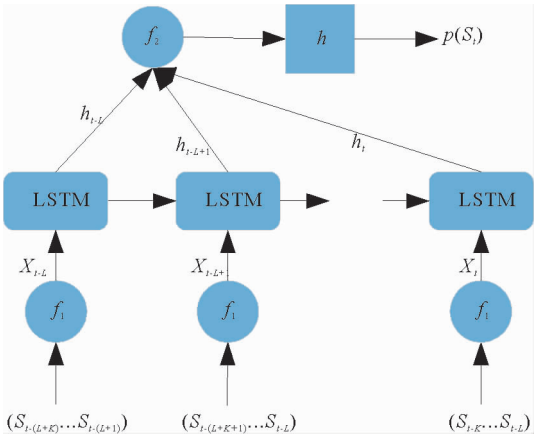

图 5　IIS+OIS 联合模型结构图

Fig. 5　IIS+OIS joint model structure diagram

表 4　IIS+OIS 联合模型实验结果

Tab. 4　IIS+OIS joint model experiment results

K	L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGLE	CIDER
1(NIC)	1(NIC)	0.682	0.510	0.378	0.275	0.360	1.677
2	6	0.694	0.523	0.389	0.285	0.367	1.736
3	6	0.697	0.525	0.390	0.288	0.370	1.741
4	6	0.695	0.521	0.385	0.284	0.366	1.713
5	6	0.699	0.527	0.391	0.289	0.367	1.742
6	6	0.693	0.521	0.386	0.285	0.365	1.729
7	6	0.697	0.525	0.389	0.285	0.366	1.704
8	6	0.693	0.522	0.388	0.281	0.364	1.730

4　结论

针对视觉信息的缺失问题,目前已提出了众多基于注意力机制的图像描述模型,本研究证明以 LSTM 作为图像描述的解码器存在文本信息缺失问题,提出了两种基于文本信息补充的 LSTM 图像描述模型—IIS 模型以及 OIS 模型,用来解决由门控结构所导致的输入信息缺失与预测信息不充分问题。实验结果表明,增加补充信息后模型性能得到提高,但同时,LSTM 对补充信息的利用存在上限,在输入词向量达到 6 个和预测对于隐藏状态的个数依赖达到 7 个以后,模型性能不再提升反而有所下降,这表明冗余的信息会对模型的学习过程产生误导。该结论在两个模型的结合实验中得到了再次证实。

参考文献:

[1]HOCHREITER S,SCHMIDHUBER J.Long short-term memory[J].Neural computation,1997,9(8):1735-1780.

[2]VINYALS O,TOSHEV A,BENGIO S,et al.Show and tell:A neural image caption generator[C]// IEEE Conference on Computer Vision and Pattern Recognition.2015:3156-3164.

[3]KULKARNI G,PREMRAJ V,ORDONEZ V,et al.Babytalk:Understanding and generating simple image descriptions[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,35(12):2891-2903.

[4]LI S,KULKARNI G,BERG T L,et al.Composing simple image descriptions using web-scale n-grams[C]// Proceedings of the 15th Conference on Computational Natural Language Learning.2011:220-228.

[5]GONG Y,WANG L,HODOSH M,et al.Improving image-sentence embeddings using large weakly annotated photo collections[C]// European Conference on Computer Vision.Springer International publishing,2014:529-545.

[6]HODOSH M,YOUNG P,HOCKENMAIER J.Framing image description as a ranking task:Data,models and evaluation metrics[J].Journal of Artificial Intelligence Research,2013,47(1):853-899.

[7]ORDONEZ V,KULKARNI G,BERG T L.Im2text:Describing images using 1 million captioned photographs[C]// International Conference on Neural Information Processing Systems.2011:1143-1151.

[8]XU K,BA J,KIROS R,et al.Show,attend and tell:Neural image caption generation with visual attention[C]// International Conference on Machine Learning.2015:2048-2057.

[9]LU J,XIONG C,PARIKH D,et al.Knowing when to look:Adaptive attention via a visual sentinel for image captioning[C]// IEEE Conference on Computer Vision and Pattern Recognition.2017:3242-3250.

[10]ANDERSON P,HE X,BUEHLER C,et al.Bottom-up and top-down attention for image captioning and visual question answering[C]// IEEE Conference on Computer Vision and Pattern Recognition.2018:6077-6086.

[11]WANG W,CHEN Z,HU H.Hierarchical attention network for image captioning[C]// AAAI Conference on Artificial Intelligence.2019,33:8957-8964.

[12]YOU Q,JIN H,WANG Z,et al.Image captioning with semantic attention[C]// IEEE Conference on Computer Vision and Pattern Recognition.2016:4651-4659.

[13]PAPINENI K,ROUKOS S,WARD T,et al.BLEU:A method for automatic evaluation of machine translation[C]// 40th Annual Meeting of the Association for Computational Linguistics.2002:311-318.

[14]VEDANTAM R,ZITNICK C L,PARIKH D.Cider:Consensus-based image description evaluation[C]// IEEE Conference on Computer Vision and Pattern Recognition.2015:4566-4575.

[15]LIN C Y.Rouge:A package for automatic evaluation of summaries[C]// Text Summarization Branches Out.2004:74-81.

[16]HE K,ZHANG X,REN S,et al.Deep residual learning for image recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition.2016:770-778.

[17]ZHANG Z,SABUNCU M.Generalized cross entropy loss for training deep neural networks with noisy labels[C]// International Conference on Neural Information Processing Systems.2018:8792-8802.

[18]DONAHUE J,ANNE HENDRICKS L,GUADARRAMA S,et al.Long-term recurrent convolutional networks for visual recognition and description[C]// IEEE Conference on Computer Vision and Pattern Recognition.2015:2625-2634.

(责任编辑:吕海亮)