❖特约专栏❖

基于深度学习的图像描述研究

杨 楠[1,2],南 琳[1,2],张丁一[1,2],库 涛[1,2]

(1. 中国科学院沈阳自动化研究所,辽宁 沈阳 110016; 2. 中国科学院大学,北京 100049)

摘 要: 卷积神经网络 (Convolution Neural Networks,CNN) 和循环神经网络 (Recurrent Neural Networks,RNN)在图像分类、计算机视觉、自然语言处理、语音识别、机器翻译、语义分析等领域取得了迅速的发展,引起了研究者对计算机自动生成图像描述的广泛关注。目前图像描述存在的主要问题有输入文本数据稀疏、模型存在过拟合、模型损失函数震荡难以收敛等问题。文中使用 NIC 作为基线模型,针对数据稀疏问题,改变了基线模型中的文本 one-hot 表示,使用 word2vec 对文本进行映射,为了防止过拟合,在模型中加入了正则项和使用 Dropout 技术,并在词序记忆方面取得创新,引入联想记忆单元 GRU,用于文本生成。在试验中使用 AdamOptimizer 优化器进行参数迭代更新。实验结果表明:改进后的模型参数减少且收敛速度大幅加快,损失函数曲线更加平滑,损失最大降至 2.91,模型的准确率比 NIC 提高了接近 15%。实验有效地验证了在模型当中使用 word2vec 对文本进行映射可明显缓解数据稀疏问题, 加入正则项和使用 Dropout 技术可有效防止模型过拟合, 引入联想记忆单元 GRU 能够大幅减少模型训练参数,加快算法收敛速度,进而提高整个模型的准确率。

关键词: 卷积神经网络; 循环神经网络; 门控循环单元; 自然语言处理; 图像描述

中图分类号: TP3 **文献标志码:** A **DOI:** 10.3788/IRLA201847.0203002

Research on image interpretation based on deep learning

Yang Nan[1,2], Nan Lin[1,2], Zhang Dingyi[1,2], Ku Tao[1,2]

(1. Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Convolution Neural Networks (CNN) and Recurrent Neural Networks (RNN) had developed rapidly in the fields of image classification, computer vision, natural language process, speech recognition, machine translation and semantic analysis, which caused researchers´ close attention to computers´ automatic generation of image interpretation. At present, the main problems in image description were sparse input text data, over-fitting of the model, difficult convergence of the model loss function, and so on. In this paper, NIC was used as a baseline model. For data sparseness, one-hot text in the baseline model was changed and word2vec was used to map the text. To prevent over-fitting, regular items were added to the model and Dropout technology was used. In order to make innovations in word order memory, the associative memory unit GRU for text generation was used. In experiment, the AdamOptimizer optimizer was used to update parameters iteratively. The experimental results show that

收稿日期: 2017-08-05; **修订日期:** 2017-10-11

基金项目: 国家科技支撑计划(2015BAF02B01);中国科学院网络化控制系统重点实验室(2015BAF02B00)

作者简介: 杨楠(1994-),男,硕士生,主要从事深度学习、自然语言处理方面的研究。Email:yangnan@sia.cn

导师简介: 南琳(1967-),男,研究员,主要从事企业信息化的应用、软件产品设计开发方面的研究。Email:nl@sia.cn

通讯作者: 张丁一(1981-),女,副研究员,硕士生导师,博士,主要从事深度学习、模式识别、自然语言处理方面的研究。Email:Dy202@sia.cn

the improved model parameters are reduced and the convergence speed is significantly faster, the loss function curves are smoother, the maximum loss is reduced to 2.91, and the model accuracy rate increases by nearly 15% compared with the NIC. Experiments validate that the use of word2vec to map text in the model obviously alleviates the data sparseness problem. Adding regular items and using Dropout technology could effectively prevent over-fitting of the model. The introduction of associative memory unit GRU could greatly reduce the model trained parameters and speed up the algorithm of convergence rate, improve the accuracy of the entire model.

Key words: convolution neural networks; recurrent neural networks; gated recurrent unit; natural language processing; image description

0 引 言

计算机自动生成图像描述是当前计算机研究领域的热点和难点，其根本任务是让计算机正确描述所感知的图像场景以及场景中的内容，相对于图像分类和目标检测，该任务涉及图像识别和自然语言处理两大交叉领域，因而极具挑战性。计算机处理图像描述任务不仅仅关注于图像中的物体识别[1]，更要关注图像中各物体之间的相互联系，并且使用逻辑清晰的语言描述图像[2-3]。

文中使用的是单向端到端的连接模型，设计思想来源于循环神经网络(Recurrent Neural Networks，RNN)在机器翻译中的重大突破[4]，通过将待翻译句子 T 改变为图片 I 并作为原始输入，训练模型目标句子单词 $S=\{S_1,S_2,\cdots\}$ 的最大似然概率 $p(S|I)$[5]，来提高预测下一个单词的准确率，其中每个单词 S_t 都来自于对图像标注的单词字典。图像描述端到端的模型主要将编码端采用卷积神经网络代替循环神经网络，对图像进行向量表示，解码端采用循环神经网络语言模型，用于生成语言描述[7]。整个系统可以端到端进行，并且优于以前的基准测试方法，能够产生更广泛的描述，很快成为该任务的主流方法。因此，将训练图像经由深度 CNN 做预训练，产生固定维度的向量并导入 RNN 解码，训练模型以生成对目标图像的合理性语言描述[5]。但是这种端到端的图像描述方法也存在诸多问题，例如文本在输入模型时使用 one-hot 表示，由于字典中的单词数即是向量长度，造成数据稀疏问题；再者由于模型存在过拟合问题，以及参数过多导致模型不能得到很好地收敛出现震荡的问题。

文中的主要贡献是使用 NIC 作为基线模型，针对数据稀疏问题，改变了基线模型中的文本 one-hot 表示，使用 word2vec 对文本进行映射，为了防止过拟合，在模型中加入了正则项和使用 Dropout 技术，并在词序记忆方面取得创新，引入联想记忆单元 GRU，用于文本生成。验证并调节 GradientDescent 和 AdamOptimizer 优化算法，并最终确定使用 AdamOptimizer 优化器进行参数迭代更新。

1 相关研究

总体来说，现有的解决图像描述任务的算法大致可以分为三类：基于模板式的、基于相似空间检索式的、基于多模翻译式的。现有的生成图像描述的形式大致也可分为三类[8]：单句子式描述，密集型描述，多语言描述。

Farhadi[2]和 Li[7]等人使用模板转换为文本场景的三元组，并使用检测到的对象和物体的关系来生成语言描述。Kulkani 使用了更为复杂的三维检测图，在对模板进行填充之前，采用条件随机场联合对象、属性、介词进行推理，但是依旧是基于模板的文本生成[3]。此外，模板式基于句法树创建的数据驱动模型，基于视觉依存表示来提取对象之间的关系，利用名词，动词，场景和介词组成的四元组描述图像，基于语言解析的表达模型也有广泛应用[11]。基于模板的方法的优点是得到的语言描述更有可能在语法上正确，缺点是它们仍然高度依赖于模板，不适用于所有图像，且限制了输出的多样性。

为此，学者们开发了更强大的方法，以相似性检索的方式[12]，来代替使用固定的模板。对于图像查询方式，是指检索向量空间中最接近的图像描述，以生

成自己的图像描述。通常这个方法确实可以实现人类的描述，因为所有的句子来自现有的人工生成的句子。但该方法需要收集大量人工生成的句子，训练集也需要多样化，并且这个方法一定程度上不能产生一个很新颖的描述。

在文中的研究中，结合卷积神经网络 VGG-16 和处理序列问题的循环神经网络，生成单一端到端的图像描述网络模型[5-7]。与之相近的研究是 Kiros, Mao 等人使用前向传递神经网络在给定图像和之前单词的基础上预测下一个单词。文中所使用的 NIC 模型则使用了循环神经网络模型，直接将可视数据提供给循环神经网络，以便能够使模型更准确追踪图像中的物体并给出物体的文本描述。文中针对 NIC 模型出现的诸多问题，对模型进行了有效地改进。

2 模型架构

文中模型是改进的 NIC 模型，该模型是基于图像到文本描述的神经网络和概率架构模型。模型将图像信息作为原始输入，将标注图像的可变句子映射到和输入图像相同的维度，用以生成对目标图像的描述。

使用如下公式来直接最大化给定图像正确描述的概率：

$$\theta^*=\arg\max\sum_{(I,S)}\log p(S|I;\theta) \tag{1}$$

式中：θ 为模型的参数；I 为输入图片；S 代表图像正确的文本描述，S 的长度是不定的。因此，使用 $S_0,\cdots,S_N$ 的链式规则概率总和来近似模拟联合概率，其中 N 指的是图像标注中特定的描述单词长度。

$$\log p(S|I)=\sum_{t=0}^{N}\log p(S_t|I,S_0,\cdots,S_{t-1}) \tag{2}$$

在训练时 (S,I) 被作为数据对，传入神经网络模型，通过 AdamOptimizer 优化算法，最小化训练数据集的 log 概率总和。RNN 在处理序列时将每一个单词按时刻传入循环神经网络，执行有监督的训练，同时记忆单元会对前一个状态进行选择性的记忆和遗忘并通过非线性激活的函数 f 更新当前的状态，一般为 Sigmoid 或 Tanh 函数：

$$h_{t+1}=f(h_t,x_t) \tag{3}$$

在模型当中，使用联想记忆单元 GRU 作为循环神经网络的记忆模块，不仅可以加快算法的收敛过程，并且模型的表现结果也非常优异。每个单词都是以 256 维的向量表示，输入为随机变量，最终通过模型训练出每个单词在向量空间中的表示，其中相似的单词在空间中的距离也是相近的。

2.1 VGG-16 网络模型

如图 1 所示在文中的模型当中，使用 VGG-16网络的架构，抽取已训练好的包含图像特征的 4 096 维的特征作为图像描述模型的输入，导入循环神经网络进行解码。

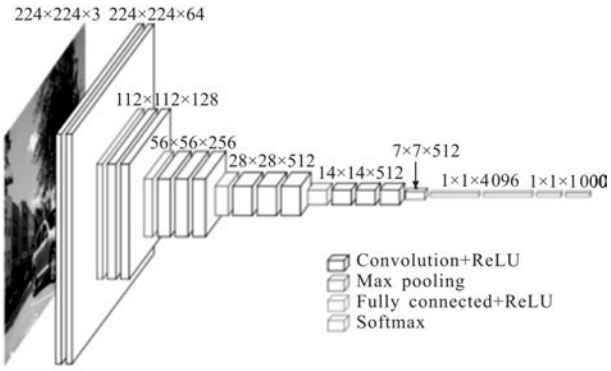

图 1 VGG-16 模型展开图

Fig.1 VGG-16 model expansion diagram

2.2 联想记忆单元 GRU

在设计和训练循环神经网络时，神经元的更新往往面临着诸多挑战，如梯度弥散和梯度爆炸问题。公式(3)中的 h 由于其处理梯度问题优越表现力，而受到研究者的青睐。长短期记忆网络 LSTM 是一种特殊的循环神经网络，在处理机器翻译和序列问题上取得了巨大成功[4]。但是其参数数量繁重，致使算法收敛比较慢，难以快速训练出行之有效模型，因此，在文中引入了联想记忆单元 GRU[13]，并对各轮训练产生的模型进行评估和调优，结果显示通过引入联想记忆单元 GRU 后，模型使用的优化算法能够快速收敛，并能高效的产生图像的合理描述，如图 2 所示。

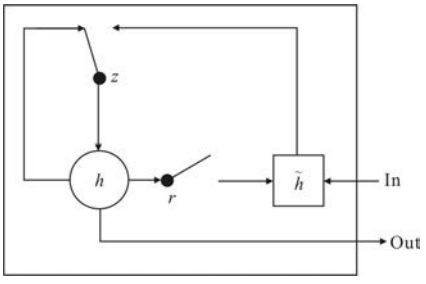

图 2 GRU 单元图[13]

Fig.2 GRU unit diagram[13]

联想记忆单元 GRU,将 LSTM 的遗忘门和输入门合并为一个更新门,大大减少了训练参数,提高了算法的收敛速度, 并对 RNN 的两个问题做了改进:一是改进了句子中位置在前的单词对当前隐藏层节点的影响会越小问题; 二是改进了对于反向传播时的误差可能是由某几个单词诱发的问题,所以 GRU 只对产生误差的部分权重进行局部更新参数, 提高了反向传播算法的工作效率。以下是 GRU 的内部更新公式:

$$h_t^j=(1-z_t^j)h_{t-1}^j+z_t^j\tilde{h}_t^j \tag{4}$$

$$z_t^j=\sigma(W_z x_t+U_z h_{t-1})^j \tag{5}$$

$$\tilde{h}_t^j=\tanh(Wx_t+U(r_t\cdot h_{t-1}))^j \tag{6}$$

$$r_t^j=\sigma(W_r x_t+U_r h_{t-1}) \tag{7}$$

对以上公式进行详细的分析:

(1) 公式(6)New memory,新的记忆是由输入 x_t 和先前的隐藏状态 h_{t-1} 得到的,简言之,该结构能够让先观察到的信息和历史隐藏状态 h_{t-1} 合并;

(2) 公式(7)Reset gate,重置门信号 r_t 会判定 h_{t-1} 对结果 h_t 的重要性, 如果 h_t 和新的记忆计算不相关,那么重置门就可以完全的消除过去的隐藏状态信息;

(3) 公式(5)Update gate,更新门 z_t 会决定以多大程度将 h_{t-1} 向下一个状态传递, 比如 z_t 近似等于1,则几乎完全传递给 h_{t-1},相反的,如果 z_t 约等于 0 向前传递给下一个隐藏层;

(4) 公式(4)Hidden state,通过使用过去的隐藏层输入 h_{t-1},最终得到了隐藏层的状态 h_t,新的记忆会根据更新门的判定产生 h_t。

以上给出了联想记忆单元 GRU 的内部更新公式,显然 GRU 单元在将遗忘门和输入门合并为更新门之后,单元内部通过更新法则,自动的寻找输入数据哪些部分选择记忆,哪些部分选择遗忘,最后通过更新门单元选择产生误差的局部进行参数更新,另外,GRU 有着比 LSTM 更少的参数,可加速在训练过程中反向传播算法的工作效率, 使得算法收敛速度大大加快,提高了整个模型的性能。

3 模型与训练

3.1 Flickr30K 数据集

Flickr30k 数据集已成为基于句子的图像描述的基准,文中介绍了 Flickr30k 实体,其增加了 Flickr30k 的 158 k 字幕, 其中包含 244 k 的互相链接, 给予同一图像连接不同字幕上相同实体提示, 并将其与 276 k 手动注释的边界框相关联[24],如图 3 所示。

数据集示例:(1. A girl with a ponytail is tying her shoes with a bent knee while on a grassy field. 2. A female athlete ties the laces of one of her cleats on the field. 3. A female soccer player crouches to put on her shoes. 4. A girl tying her shoe in a large sports field. 5. Soccer player kneeling down to tie her shoe)

图 3 示例图片及描述

Fig.3 Examples of pictures and descriptions

3.2 CNN 图像编码阶段–Image embedding

VGG–16 网络接受 224×224 像素的输入图像,将在 ImageNet 上预训练好的权重,作为初始值导入笔者的模型,以便加速训练。利用 VGG–16 的特征提取功能, 抽取网络最后生成的对图像 4 096 维特征,并将此高维特征作为该图像的特征描述。

笔者的模型关注的图像的全局信息, 因此每张图像最终使用 4 096 维向量表示, 但是 4 096 维的高维数据构成的特征在向量空间中表示, 造成数据稀疏的危险,因此进行了降维处理,在模型的 CNN 输出阶段将 4 096 维向量映射到和文本相同的 256 维空间中, 笔者的模型在特征空间中学习图像特征和文本之间拟合曲线。

3.3 文本编码阶段–Word embedding

计算机处理文本的信息, 都是需要进行数字化处理的。相比之下,最为简单的是词袋子模型,也就是 0/1 编码的独热向量表示。将所有的单词都使用很长的 n 维向量表示, 其中 n 代表字典中单词的数量。在 Flickr30k 数据集当中,总共有 2 943 个单词,因此 n 为 2 943。该方法的不足之处在于,其一不能保留词序信息,二是采用 0/1 编码的形式,得到的数

据比较稀疏,容易受到维度灾难的问题。为了解决以上问题,基于神经网络的数据表示模型 word2vec 由 Google2013 年开源,通过训练语料,获取词的多维实数向量。

神经网络的词向量表示是指通过神经网络对上下文,以及上下文与目标词之间的关系进行建模。首先对字典中的每个单词进行编码,然后使用 word2vector 模型对该单词进行训练,最终生成 256维的 word-embedding,该向量不仅保留了词序信息,而且词语具有相近词义的在空间中的距离也是相近的,生成的词向量如图 4 所示。

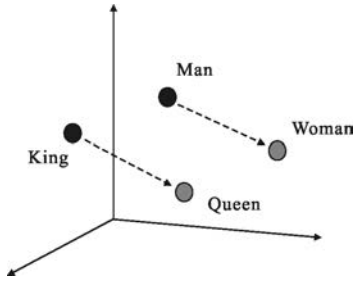

图 4 词向量空间表示图

Fig.4 Word vector space representation

3.4 语言生成模型

改进后的模型结构如图 5 所示,该模型的输出

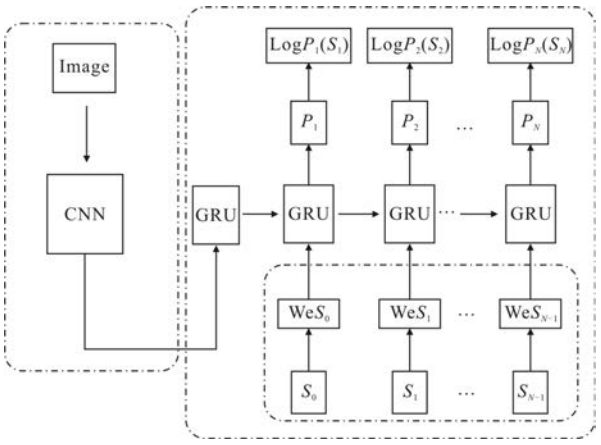

图 5 语言生成模型表示图

Fig.5 Language generation model representation diagram

端是由具有 *Soft*max 层的 GRU 网络单元构成。*Soft*max 层将模型词汇的概率分布输出为[26]:

$$p(S_t|S_{t-1},\cdots,S_0,I)=Softmax(Dy(t)) \tag{8}$$

式中:D 是将具有与 GRU 单元数相同维度的向量 $y(t)$ 映射到输出词汇大小 N 的解码器矩阵。视觉特征 I 通过嵌入矩阵 W,在第 0 时刻保持图像的全局信息并作为 GRU 单元的输入,由该部分对 GRU 隐藏层部分进行初始化,所以此视觉特征输入称之为初始化 GRU 单元。在之后的时间步长 t 中,起始符号后面是参考描述中每个单词的嵌入向量表示(训练模型期间) 或者是前一个生成的词 (在测试阶段) 的向量表示,作为当前时刻 t 的输入 $x(t)$。

在训练阶段,在每个时间步长 t,GRU 单元在给定当前输入和隐藏层状态值的情况下,通过用为损失函数计算梯度的方法 AdamOptimizer,来计算生成下一个词的概率分布,以最高概率对应的 word 作为下一刻的值。结合训练样本的正确图像描述,通过有监督的训练最大似然函数来完成整个模型的训练。

3.5 训练阶段

图 5 中是模型的展开式,通过以下公式对模型进行训练,因为在求解优化问题时,习惯性的将其作为最小优化问题来对待,因此最小化目标函数负对数似然:

$$x_{-1}=\mathrm{CNN}(I) \tag{9}$$

$$x_t=W_eS_t,t\in\{0,\cdots,N\} \tag{10}$$

$$p_{t+1}=\mathrm{GRU}(x_t),t\in\{0,\cdots,N\} \tag{11}$$

$$L(I,S)=-\sum_{t=1}^{N}\log p_t(S_t) \tag{12}$$

训练网络涉及调整语言模型的参数,以最小化公式(12)中所示的负对数似然函数。文中使用的模型通过时间反向传播算法降低损失函数,使用梯度下降调整 GRU 单元中的参数,具体的是使用 AdamOptimizer 优化算法。训练样本用于随机小批量的图像文本对,并在累计每个小批量的损失之后进行梯度下降。

所有 GRU 参数,词嵌入向量和解码矩阵均使用上述方法学习。特征提取模块的参数未经过训练,但是被保持为固定维度,防止发生过拟合。

在生成语言模型中联想记忆单元 GRU 训练期间,使用了 Dropout 进行正则化。Dropout 的原因来自于生物学上神经元的阈值激发和延时性传递特性,使用 Dropout 将网络层中的某些神经元的输出随机丢弃到零,然后再将其输入到下一层,这使得各个神经元之间的有更少的依赖并且网络更加健壮。试验中,发现使用概率为 0.5 的 Dropout 能够最大化的提高模型的泛化能力。

3.6 测试阶段

在测试图像的描述生成阶段,没有可供参考的

描述,需要从 $P(S|I)$中进行采样,以生成描述。与训练阶段相类似,图像的特征向量在时间步长 $t=-1$ 后被送到联想记忆单元 GRU 网络。然后在时间步长 $t=0$,$Soft_{max}$ 位置具有最高概率的词是模型认为最可能的第一个词。选择该词作为第一个单词,然后在时间步长 $t=1$ 中将相应的单词向量作为输入提供给 GRU 单元,该迭代过程重复,直到模型生成完整的句子为止,这种句子生成方式称之为 Sampling 方法。

Sampling 方法的一个问题是只考虑在每个时间步长中最可能的单词,但是不能保证最终得到恰当图像描述合理性句子。在笔者的模型中仍然使用 Beam Search,在每个步骤保持顶部 m 部分的句子,对于这些顶部的 m 个句子,只考虑前 m 个句子的扩展,并对句子进行重新评分,Beam 的大小设定为20。重复该过程,直到所有搜索终止或达到最大允许的句子长度。在这个过程结束时,根据模型分配的对数似然概率生成候选的句子。

4 实验结果分析

4.1 算法收敛速率–模型的损失对比

图 6 所示为 NIC 模型的损失,图 7 所示为改进后模型的损失图,其中下方曲线表示训练集,上方曲线表示测试集。

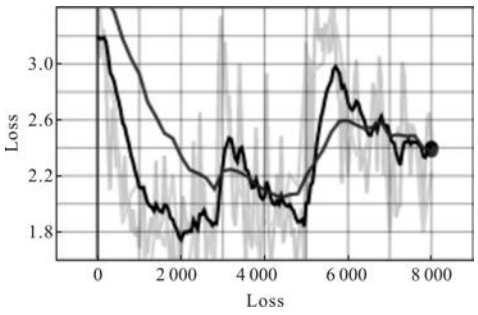

图 6 NIC 模型损失收敛速率图

Fig.6 NIC model loss convergence rate chart

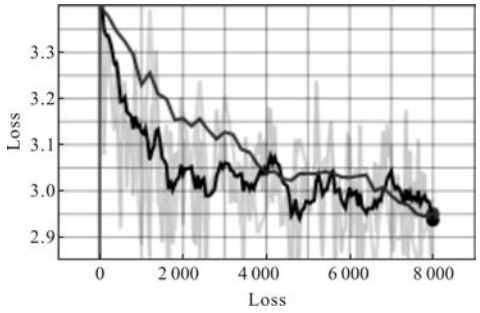

图 7 改进后模型损失收敛速率图

Fig.7 Improved model loss convergence rate chart

将改进后模型与基线模型 NIC 的损失每轮次进行抽样,绘制了上图的收敛速率对比图,通过上图能够清楚的看到笔者的模型损失较原有模型相比下降的很快,并且优化算法最终的收敛效果也较 NIC 有很大的提升。

在模型架构 3.2 节当中,详细分析了联想记忆单元 GRU 的工作原理以及能够加速算法收敛的原因,在本小节中通过实验数据做出的收敛速率比对图更加有力的验证了之前的结论,在引入了联想记忆单元 GRU,并改进了原模型中的文本 one–hot 表示,使用 Word2Vec 对文本进行 256 映射处理,使用为损失函数计算梯度的方法 AdamOptimizer 进行优化之后,笔者的模型在图像描述上取得了更加可观的效果。

4.2 模型的准确率对比

从数据集当中采取了 70%作为训练集,30%作为测试集,在每轮训练的过程当中对笔者的模型进行测试。其中上方曲线表示训练集,下方曲线表示测试集。笔者对 NIC 模型和改进后的模型进行了实验对比,实验对比结果如图 8、9 所示。

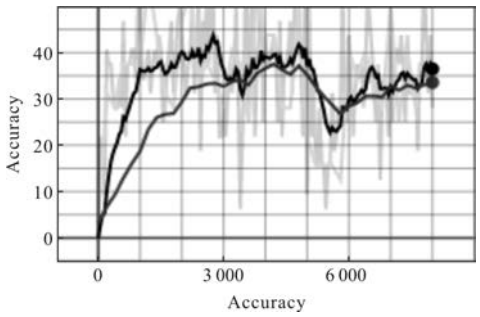

图 8 NIC 模型测试准确率图

Fig.8 NIC model test accuracy chart

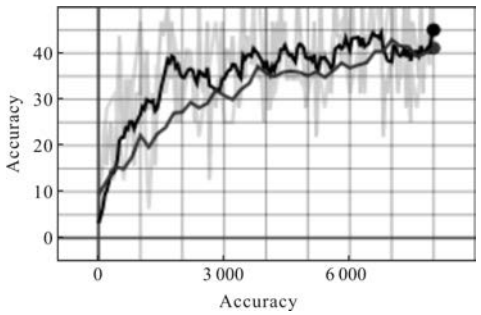

图 9 改进后模型测试准确率图

Fig.9 Improved model test accuracy graph

NIC 模型的准确率起初呈现上升趋势,期间经过不同程度的下降,说明模型存在过拟合,导致测试集准确率急剧下降,最终也并未得到提高。为了提高模型的准确率,在模型当中加入了正则化及使用 Dropout 技术等,通过对模型的改进,笔者的模型不仅在训练集还是在测试集上的表现都非常良好,最终在测试集上的准确率较 NIC 模型有了很大的提升。

4.3 生成文本表示

进行了多次试验,并对每一轮产生的模型进行保存,通过对测试集中图像的验证,发现每一轮模型对图像生成的描述内容不是唯一的,呈现一定的随机性,基本上随着模型训练轮次的增多,对图像的描述就越丰富。但是这也并不是绝对的,在试验测试中也发现,训练轮次很深的模型,反而不能生成对句子的描述。以下给出模型生成文本结果,如图 10 所示((a) A little girl in a pink dress is holding a pink flower, (b) Two dogs are running through the grass. (c) A naked man rides his bike on a Rocky terrain, (d) A child wearing a shirt and blue Headband is holding a toy. (e) A young man is sitting on the Beach With his hand in the air, (f) A man in a white and blue dress is Playing a football)。

图 10 文本生成结果示例

Fig.10 Examples of text generation result

5 结束语

文中的模型基于卷积神经网络对图像编码,循环神经网络对图像和文本信息进行解码,使用 NIC 作为基线模型,针对数据稀疏问题,改变了基线模型中的文本 one-hot 表示,使用 word2vec 对文本进行映射,为了防止过拟合,在模型中加入了正则项和使用 Dropout 技术,并在词序记忆方面取得创新,引入联想记忆单元 GRU,用于文本生成。验证并调节 GradientDescent 和 AdamOptimizer 优化算法,并最终确定使用 AdamOptimizer 优化器进行参数迭代更新。进一步提升了模型的性能,从而生成图像更为准确合理人性化的描述,对图像描述的研究有积极的推动意义。

在实验过程中,也遇到了一些亟待解决的问题。首先,模型是以数据为驱动的,数据量的大小严重影响着模型的描述能力。随着 MSCOCO 数据集的不断扩大,笔者的模型也尝试在更新更大的数据集上进行测试,总体来说,在测试训练数据集上还有很大的提升空间。其次,在图像的编码阶段,该模型是使用的 VGG-16 网络,但是随着更深层次的图像特征提取模型:VGG-19、ResNet、Inception、Xception 等的不断涌现,相信在使用了更为有力的图像特征提取模型之后,的改进之后的模型在图像描述上还会有质的提升。最后,在创建和训练模型的过程当中,存在随机性,这也是可以理解的,因为笔者的模型的最终目的是最小化负对数似然概率,因此笔者的初始化参数对寻找最优解有一定影响,另外笔者通过随机梯度下降算法寻找到的是局部最优解,不能确定是否获得全局最优解也是我们正在面临的一大难题,因而生成的模型也具有随机性,并不是模型的训练的轮次越多,模型的效果就越好,这一点在保存的模型当中通过实验得到了验证。

图像描述研究具有深远的现实意义,其将在早期的幼儿教育,基于内容的图像检索、医学图像分析、辅助导盲、新闻自动化、军事安全等方面发挥重大的作用。目前的研究进展,仅限于对图像的解释亦或描述,尚且存在许多亟待解决的问题,还远远不能称之为图像的理解。真正意义上的图像理解在于用更深层次的语言去理解图像。更深层的图像理解是为了让机器模拟人更好的进行人机交互。尝试去理解图像中的场景、人物、动作、表情、以及所处的状态。而国内外所以做到的是对图像进行简单的描述,描述性的语句单词量不超过 15 个单词,甚至语句不

通顺或逻辑性错误。相对于实现计算机能够对一幅图像做出详尽的描述，甚至能够以生动形象故事形式理解图像，这条路还任重道远。相信在未来，软件、硬件、以及数据量足够的情况下，伴随着巨大的支持和热情的期盼，深度图像理解变为现实指日可待，但前路并非坦途，还需要国内外科研学者的共同努力。

参考文献：

[1] Xu Feng, Lu Jiangang, Sun Youxian. Application of neural network in image processing [J]. ***Chinese Journal of Information and Control***, 2003, 4(1): 344-351. (in Chinese)
许锋, 卢建刚, 孙优贤. 神经网络在图像处理中的应用[J]. 信息与控制, 2003, 4(1): 344-351.

[2] Farhadi A, Hejrati M, Sadeghi M A, et al. Every picture tells a story generating sentences from images [J]. ***ECCV***, 2010, 21(10):15-29.

[3] Kulkarni G, Premraj V, Dhar S, et al. Baby talk: Understanding and generating simple image descriptions [J]. ***CVPR***, 2014, 35(12): 1601-1608.

[4] Cho K, van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [J]. ***EMNLP***, 2014, 14 (6): 1078-1093.

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

[6] Alex Krizhevsky, IIya Sutskever, Geoffrey Hinton. Imagenet classification with deep convolution neural networks [C]// Proceedings of Advances Neural Information Processing Systems(NLPS), 2012: 1097-1105.

[7] Sermanet P, Eigen D, Zhang X, et al. Overfeat: Integrated recognition, localization and detection using convolutional networks [J]. ***Computer Vision and Pattern Recognition***, 2013, arXiv preprint arXiv: 1312.6229.

[8] Gerber R, Nagel H H. Knowledge representation for the generation of quantified natural language description of vehicle traffic in image sequence[C]//Proceeding of the IEEE International Conference on Image Processing, 1996: 805 - 808.

[9] Yao B Z, Yang X, Lin L, et al. I2t: Image parsing to text description[C]//Proceedings of the IEEE, 2010, 98(8): 1485-1508.

[10] Li S, Kulkarni G, Berg T L, et al. Composing simple image descriptions using web-scale n-grams [C]//Proceeding of the Conference on Computational Natural Language Learning, 2011.

[11] Aker A, Gaizauskas R. Generating image descriptions using dependency relational patterns [C]//Proceedings of the Meeting of the Association for Computational Linguistics (ACL), 2010: 49 (9) :1250-1258.

[12] Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics [C]//International Conference on Artificial Intelligence, 2013, 47(1): 853-899.

[13] Wen Ya, Nan Lin. Research on semantic analysis method of image based on natural language understanding [D]. Shenyang: Shenyang Institute of Automation, Chinese Academy of Sciences, 2017. (in Chinese)
温亚, 南琳. 面向自然语言理解的图像语义分析方法研究[D]. 沈阳: 中国科学院沈阳自动化研究所, 2017.