

基于深度学习的交通场景语义描述

曲仕茹, 席玉玲, 丁松涛

(西北工业大学 自动化学院, 陕西 西安 710129)

摘要:对复杂交通场景进行准确的语义描述,一直是图像视觉领域的难题。交通场景复杂多变,对图像场景的理解容易受到光线变化、物体遮挡等因素的干扰。针对这一问题,提出了一种基于注意力机制的交通场景语义描述方法。使用卷积神经网络(CNN)和循环神经网络(RNN)相结合的方式,产生对交通场景的端对端描述。交通目标种类繁多,为了产生带有明显区分度的场景描述,在语言模型中引入了注意力机制。为了验证新算法的有效性,分别在 Flickr8K、Flickr30K 和 MS COCO 3 个基准数据库上进行了实验。结果表明,在不同评估方法下,算法准确率分别提升了 8.6%, 12.4%, 19.3% 和 21.5%。同时,通过定性分析验证了算法在光线变化、异常天气环境、道路显著目标和多种交通工具等 4 种不同的复杂交通场景下,都具有良好的鲁棒性。

关键词:智能交通;深度学习;神经网络;交通场景语义描述;注意力机制

中图分类号:U495

文献标志码:A

文章编号:1000-2758(2018)03-0522-07

图像语义描述,是在理解图像的基础上,将图像内容用自然语言合理表达出来。在交通场景中,使用自然语言描述场景中的物体以及物体间的相互关系,在辅助视觉障碍人群、安全辅助驾驶等方面,具有重要的意义。由于交通场景受光线变化、角度变化、背景杂乱等影响较大,准确描述场景中的物体和相互关系,具有较大挑战。随着深度学习的快速发展,一种结合深度卷积神经网络与循环神经网络的图像语义生成方法,在图像语义描述问题上取得了显著的成绩。基于图像特征的语义描述,不仅要对象目标进行准确识别,还需要具备检测图像中不同目标相互关系的能力。但是在复杂场景下,尤其是在交通场景中,包含多目标的图像语义描述依然存在描述冗余以及描述不准确的问题。

目标检测、目标识别是图像语义描述的基础,图像语义描述是对图像更抽象的一种表达方式。Li 等^[1]在目标检测方法的基础上,通过使用不同的训练模型分别提取目标及其相互关系,再通过训练语言模型,得到对图像内容的表达。采用这种方法的缺点是计算量大、图像描述生硬。Kuzbetsova 等^[2]提出了一种基于语言解析的语言模型,能够对图像

中存在的自然景象进行描述,但缺点是描述生硬,并且模型需要手动设计。Kiros 等^[3]首先使用基于神经网络的对数双线性模型来生成描述,这种方法忽略了已生成词汇对后生成词汇的影响,所生成的描述中词汇关联性较差。随着循环神经网络(RNN)的发展,因其具有记忆存储功能的特点,被大量使用在文本处理过程中^[4]。Vinyal 等^[5]采用基于长短期记忆单元(long-short term memory, LSTM)的循环网络模型,解决了 RNN 训练过程中梯度消失和梯度爆炸等问题。在图像语义描述生成过程中,只需要在开始阶段输入一次图片进行特征提取,并生成对应的特征向量,便实现了端对端的描述生成。本文在预训练阶段使用 CNN 分类模型对图片进行特征编码,将最后一层隐含层的编码结果作为 LSTM 的输入,取得了理想的结果。

为了解决图像语义描述中存在的上述问题,本文对基于深度学习的特征提取方法做出了改进,以获得更丰富的特征信息,例如颜色、位置等。这些信息用来辅助语言模型产生更精确和生动的语义描述。同时将注意力机制引入到图像语义描述中,提出了一种在注意力作用下,基于 LSTM 变体的语言

模型。该模型对不同目标特征赋予不同的权值影响因素,对复杂背景下多个目标具有更强的表述能力,并能够更好地描述图像中所存在的显著目标。在语言模型上采用端对端的“编码-解码”方法得到完整并且符合语法规则的描述,为了验证算法的有效性,在基准数据库 Flickr8k^[10]、Flickr30K^[11]和 MS COCO^[12]上进行了实验。

1 图像特征提取

以往图像的描述工作多是将输入图片送入训练好的 CNN 模型中,然后将其输出作为 LSTM 网络的输入。本文算法受文献^[13]的启发,使用监督学习和多标签分类方法来预测属性集合,通过训练最小化损失函数所对应的深度卷积神经网络,来进行属性预测。对于每张图片产生固定长度的向量,向量的长度是属性集的大小,每个维度包含了特定属性的预测概率。提取交通场景中显著目标的特征信息,为后续工作奠定基础。注意力是一个抽象的概念,它存在于人类的感知中,影响着人类生活中的各个方面。注意力的产生与作用可分为3个阶段^[14]:第一阶段,显著颜色和物体会对人脑产生对应的生理性刺激;第二阶段,大脑过滤感知到的信息,并重建三维图像;第三阶段,大脑确认关注点。这种由底层到高层的刺激引发模式,展示了人类注意力的工作方式,并启发我们在图像描述中使用注意力机制。

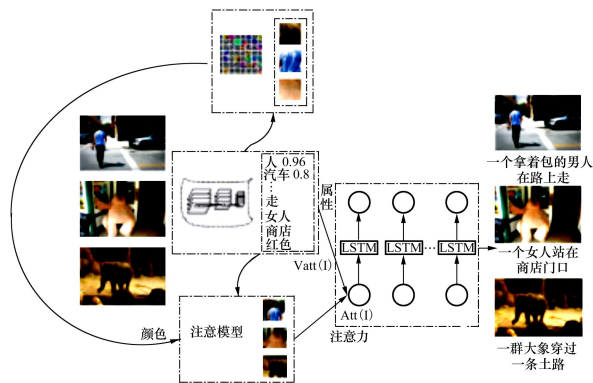


图1 模型结构概览

Bahdanau 等^[6]最早在图像分类中提出注意力机制,并开创性地将注意力机制应用于机器翻译任务。Graves 将注意力机制引入到图像语义描述中,继对复杂场景中的多目标采用有选择的语义描述,且成为当前图像描述领域的一个发展趋势^[7]。注

意力机制的引入,还提升了复杂图像的表达效果。Xu 等^[8]提出2种注意力机制:“硬随机”和“软决定”。“硬随机”通过最大化近似变分下限获得了更好的成绩。

交通场景中,红绿灯、路标路障等重要标志都具有显著的颜色。针对这一类场景,我们在模型中加入了受显著颜色刺激的注意力因子。颜色特征描述子参考 CN(color names)^[15]的使用,通过赋给某些图像区域的颜色标签的概率,来影响图像表达。CN 是人赋予颜色的语义标签,包括11个基本颜色项:黑、蓝、棕、灰、绿、橙、粉、紫、红、白和黄。在计算机视觉中,CN 是将 RGB 值和语义标签进行对应的一个过程。给定图像区域,通过 PLSA 方法,学习得到颜色子概率向量。

$$CN = \{p(cn_1 | B), p(cn_2 | B), \dots, p(cn_{11} | B)\} \quad (1)$$

$$p(cn_i | B) = \frac{1}{P} \sum_{x \in R} p(cn_i | f(x)) \quad (2)$$

式中, cn_i 代表第*i*个颜色子, x 为 bounding boxes 中像素 P 的空间坐标。 $P(cn_i | f)$ 是给定像素值的颜色子概率,每种颜色概率使用 PLSA 模型,由 100 个 google 的图片训练得到。颜色描述子对光线变化具有较好的鲁棒性,同时还能够对黑色、灰色和白色编码,具有较高的辨别力。

2 语言模型

2.1 LSTM 模型

语言模型(language model)是一种概率分布模型,是自然语言处理领域里一个重要的基础模型。目前最成功的语言模型是 LSTM 模型。本文提出了一种注意力机制作用下的 LSTM 变体模型。LSTM 是 RNN 网络的一种特殊形式。由于其结构中特有的记忆单元,多用于处理和预测时间序列中间隔和延迟非常长的重要事件。在每个时间节点,记忆单元决定了什么内容需要被忘记,什么需要被记住,这可由4个门控制,如图2所示。在门控制阶段,文本尝试加入具有特征作用的注意力门控制单元,其输入取决于前一时间节点的隐含层状态和 CNN 提取的图像特征,刺激引起的视觉特征可以通过计算机进行分析。

在编码阶段,模型将图片和词汇记为隐藏层状态向量。给定一张图片,将其输入到训练好的 VGG-

16 模型中,并提取图像特征。同时,将词汇转换为“独热码”向量,通过矩阵变换至 512 维的嵌入空间,并将其输入至 LSTM 模型中。

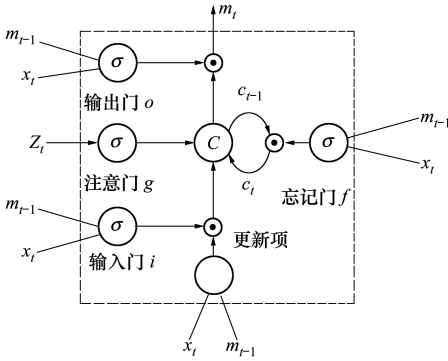


图 2 LSTM 变体结构图

图 2 描述 1 个 LSTM 细胞单元,每个单元包含 4 个门和 1 个记忆细胞 C。通常,LSTM 细胞通过忘记门控制“记忆”和“忘记”当前细胞状态,通过输入门和输出门控制细胞状态的输入和输出。我们在结构中加入“注意门”来调整权值,决定强调的内容。每个细胞单元包含 3 个输入, m_{t-1}, m_t 分别代表 t 时刻和 $t-1$ 的隐藏层状态。 x_t 代表 t 时刻的词汇向量, x_0 为 CNN 最开始输出的图像特征。 Z_t 代表了注意力向量。 c_{t-1} 和 c_t 为 $t-1$ 时刻和 t 时刻的记忆细胞状态。LSTM 的定义公式如下:

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1} + b_i) \quad (3)$$

$$g_t = \sigma(W_{gz}z_t + b_g) \quad (4)$$

$$f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1} + b_f) \quad (5)$$

$$o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1} + b_o) \quad (6)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{cm}m_{t-1} + b_c) + g_t \quad (7)$$

$$m_t = o_t \odot c_t \quad (8)$$

$$p_{t+1} = \text{softmax}(m_t) \quad (9)$$

式中, \odot 代表记忆细胞的输入或输出与每个输出门的乘积。不同的 W 和 b 为训练的权值矩阵和偏置。注意力门从区域选择中提取注意力信息。 σ 为 sigmoid 激活函数, h 为 tanh 函数。在最后一个公式中, softmax 产生概率分布。

2.2 注意力机制

注意力项 Z_t 为颜色刺激下,卷积特征的加权值,由隐含层状态和卷积特征共同决定,本文借鉴了 Zhu 等人^[16]提出的方法,具体公式如下:

$$e_t = w_a^T \tanh(W_{hc}h_{t-1} + W_{ce}C(I) + b_a) \quad (10)$$

$$a_t = \text{softmax}(e_t) \quad (11)$$

$$Z_t = a_t^T C(I) \quad (12)$$

式中, $C(I)$ 为颜色刺激下的卷积特征图。 a_t 为 t 时刻结合注意力的卷积特征向量。 W 和 b 分别为训练的权值和偏置参数。在解码阶段,取最后一层隐藏层状态与第七层全连接层视觉的特征点乘获得对数概率的最大值。由图 3 可以看出,在模型中取样注意力的权值,注意力的引发受颜色影响显著。



图 3 注意力可视化

3 实验结果分析

为了验证本文算法的有效性,设计的实验包含不同光照变化及复杂场景下的图像语义描述。选择的测试数据库主要来源于: Flickr8K、Flickr30K 和 MS COCO。通过将本文方法与 Google NIC^[5], Log Bilinear^[3], Xu 等^[8]方法对比,通过定性和定量的方法验证本文算法的性能。

本文使用 Flickr8K、Flickr30K 和 MS COCO 3 个基准数据库,包含上万张图片及对应的标注。训练集用来调整神经网络的权值。验证集用来决定网络结构和参数复杂性。测试集用来测试最终模型的表现。特别选取交通场景图片作为测试集。

Flickr8K 数据库包含来自网络上的 8 000 多张图片,主要针对人类和动物的行为表现,通过人工标注的方法,每幅图片由 5 个不同的人员产生 5 个对应的标注。每个标注者被要求描述图片中的人员、物体、场景以及活动内容。Flickr30K 和 MS COCO 数据库针对场景理解,图片主要来自于日常场景,图像中的目标通过精确分割进行校准。本文使用的是第一版数据库。

在数据预处理阶段,本文转换了所有的标注为小写,舍弃了非字母单词,过滤发生超过 5 次的词汇,因此 Flickr8K, Flickr30K 和 MS COCO 的词汇数分别为 2 538, 7 414 和 8 791。对所有的数据库,本文使用的词典大小为 10 000。

本文选用当前主流的评估策略对实验结果进行评测。目前最常使用的评估方法为 BLEU,由 n-gram 概率模型计算,通常使用于机器翻译中。这种准则用于计算生成和参考词汇之间不同词汇长度相关性的分数。给定一张图片 I_i ,模型产生一句对应描述,自动评估准则基于人类标注的参考描述集给出分数。

3.1 定量分析

本文通过一系列实验,验证了注意力模型的有效性。表 1~3 分别代表不同数据库下,用不同算法产生的实验结果。通过对比可知,本文算法在 Flickr8K, Flickr30K 和 MS COCO 数据库上相比较于其他模型表现更好。

表 1 Flickr8K 数据库上实验结果比较
($B-n$ 代表 BLEU 分数, $n=1,2,3,4$)

Model	B-1	B-2	B-3	B-4
Google NIC	63	41	27	-
Log Bilinear	65.6	42.4	27.7	17.7
Xu	67	45.7	31.4	21.3
本文	67.6	46.1	32.2	21.5

表 2 Flickr30K 数据库上实验结果比较

Model	B-1	B-2	B-3	B-4
Google NIC	66.3	42.3	27.7	18.3
Log Bilinear	60	38	25.4	17.1
Xu	66.9	43.9	29.6	19.9
本文	67.3	45.1	29.9	21.1

表 3 MS COCO 数据库上实验结果比较

Model	B-1	B-2	B-3	B-4
Google NIC	66.6	46.1	32.9	24.6
Log Bilinear	70.8	48.9	34.4	24.3
Xu	71.8	50.4	35.7	25.0
本文	72.3	51.8	37.1	25.1

3.2 定性分析

为了验证算法的有效性,设计的实验包括光照剧烈变化,异常天气影响,交通显著物体的识别和不同交通工具的检测。选择的图片主要来源于 MS COCO 数据库,通过定性方法分析算法的性能,如图 4~图 7 所示。

实验 1 夜间交通场景交通指示灯描述。如图 4 所示,在夜晚光线昏暗的道路场景中,灯光的散射效应较强并且夜间交通目标的识别难度较高。由于本文注意力算法对醒目颜色敏感,在夜晚能够有效识别出交通信号灯并加以描述。同时,在夜晚强光照下导致的颜色失真也不影响对交通目标和场景的识别和描述。



图 4 夜晚交通场景语义描述

实验 2 雨、雪等异常天气。在雨、雪等异常天气情况下,交通目标变得模糊,影响识别的准确性。本文方法能够准确识别和表达图片中的雨、雪天气,如图 5 所示,本文方法在雨伞遮挡的情况下,识别行人,并对雨伞和行人在具体天气下加以描述。在雨雪天气,算法对其他目标的识别效果较好,受天气变化的影响小。



图 5 雨雪天气交通场景语义描述

实验 3 显著交通目标描述。由于在特征提取中加入颜色描述子,算法在交通场景中提取显著颜色区域,根据注意力规则优先描述显著目标。在图 6 中,算法检测出信号灯、红色巴士和路牌并准确描述。

实验 4 不同交通工具区分。算法在检测不同交通工具上有优秀的区分性。如图 7 所示,算法能

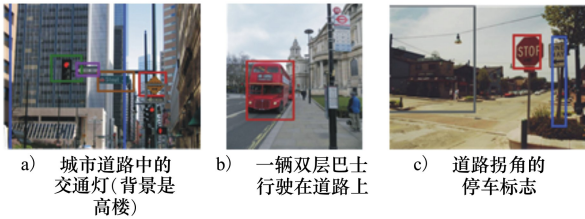


图 6 交通显著物体语义描述

够检测出不同图片中的摩托车、自行车和滑板,根据不同交通目标的特点,方便在应用中做出合理选择。



图 7 不同交通工具的语义描述

4 结 论

本文通过研究高层语义概率分布和引入注意力机制,使模型能够对复杂交通场景进行准确识别和描述。通过改进语言模型,使得图像内容表述简洁通畅。实验结果表明,本文算法在场景光线发生剧烈变化、异常天气等干扰情况下,仍然具有较好的描述效果。下一步研究工作的重点是提高算法运行的实时性和交通目标相互关系的准确描述。

参考文献:

[1] Li S, Kulkarni G, Berg T L, et al. Composing Simple Image Descriptions Using Web-Scale N-Grams[C]//Proceedings of the Fifteenth Conference on Computational Natural Language Learning Association for Computational Linguistics, 2011: 220-228

[2] Kuznetsova P, Ordonez V, Berg T, et al. Treetalk: Composition and Compression of Trees for Image Descriptions[J]. Transactions of the Association of Computational Linguistics, 2014, 2(1): 351-362

[3] Kiros R, Salakhutdinov R, Zemel R S. Multimodal Neural Language Models[C]//International Conference on Machine Learning. 2014: 595-603

[4] Donahue J, Hendricks L A, Guadarrama S, et al. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 2625-2634

[5] Vinyals O, Toshev A, Bengio S, et al. Show and Tell: a Neural Image Caption Generator[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3156-3164

[6] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. Computer Science, 2014(9): 0473-0482

[7] Graves A. Generating Sequences with Recurrent Neural Networks[J]. Computer Science, 2013(8): 0850-0863

[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] Ba J, Mnih V, Kavukcuoglu K. Multiple Object Recognition with Visual Attention[J]. Computer Science, 2014(12): 7755-7771

[10] Rashtchian C, Young P, Hodosh M, et al. Collecting Image Annotations using Amazon's Mechanical Turk[C]//NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010: 139-147

[11] Young P, Lai A, Hodosh M, et al. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions[J]. Transactions of the Association for Computational Linguistics, 2014(2): 67-78

[12] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context[C]//European Conference on Computer Vision Springer Cham, 2014: 740-755

[13] Wu Q, Shen C, Liu L, et al. What Value Do Explicit High Level Concepts Have in Vision to Language Problems[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 203-212

- [14] Corbetta M, Shulman G L. Control of Goal-Directed and Stimulus-Driven Attention in the Brain[J]. *Nature Reviews Neuroscience*, 2002, 3(3): 201-215
- [15] Van De Weijer J, Schmid C, Verbeek J, et al. Learning Color Names for Real-World Applications[J]. *IEEE Trans on Image Processing*, 2009, 18(7): 1512-1523
- [16] Zhu Y, Groth O, Bernstein M, et al. Visual7w: Grounded Question Answering in Images[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 4995-5004

Image Caption Description of Traffic Scene Based on Deep Learning

Qu Shiru, Xi Yuling, Ding Songtao

(School of Automation, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: It is a hard issue to describe the complex traffic scene accurately in computer vision. The traffic scene is changeable, which causes image captioning easily interfered by light changes and object occlusion. To solve this problem, we propose an image caption generation model based on attention mechanism. Combining convolutional neural network (CNN) and recurrent neural network (RNN) to generate an end-to-end description for traffic images. To generate a semantic description with distinct degree of discrimination, the attention mechanism is applied to language model. Using Flickr8K, Flickr30K and MS COCO benchmark datasets to validate the effectiveness of our method. The accuracy is promoted maximally by 8.6%, 12.4%, 19.3% and 21.5% in different evaluation metrics. Experiments show that our algorithm has good robustness in four different complex traffic scenarios, such as light change, abnormal weather environment, road marked target and various kinds of transportation tools.

Keywords: intelligent transportation; deep learning; neural network; image captioning; attention mechanism; design of experiments; reliability analysis