

基于生成对抗网络的图像识别

程焕新, 张志浩, 刘文翰, 郭占广
(青岛科技大学 自动化学院, 山东 青岛 265200)

摘要:随着深度学习的迅速发展,图像识别技术也在日益提高。但在日常的人脸识别、物体识别的应用中常有识别内容错误、识别准确率过低的问题。对此,提出了一种基于生成对抗网络的图像问答模型(GAN-QA)。首先生成对抗网络显示了强大的图像识别能力,通过生成对抗网络的生成器、判别器原理可以更好地提取图像特征,显著提高了图像识别的准确率。同时根据视觉识别的自然语言处理(NLP)也取得了极大的提升。该模型通过长短期记忆网络(LSTM)将两者结合起来,通过生成对抗网络识别图像,而后问题和视觉信息被输入到长短期记忆网络中,通过模型的训练可以对图像上的问题给出答案。在数据集 DAQUQR 上的验证结果表明,所提出的基于生成对抗网络的图像问答模型能够有效地提高对带问题图像的识别问答能力,由此明显提升了图像问答的准确度。

关键词:自然语言处理;生成对抗网络;深度学习;图像识别;准确性

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2021)06-0175-06

doi:10.3969/j.issn.1673-629X.2021.06.031

Image Recognition Based on Generative Adversarial Network

CHENG Huan-xin, ZHANG Zhi-hao, LIU Wen-han, GUO Zhan-guang
(Qingdao University of Science and Technology, Qingdao 265200, China)

Abstract: With the rapid development of deep learning, image recognition technology is also increasing. However, in the daily application of face recognition and object recognition, there are often problems of wrong recognition content and low recognition accuracy. Therefore, we propose an image question answering model (GAN-QA) based on generative adversarial network. First of all, the generative adversarial network shows a strong image recognition capability. By generator and discriminator principle of generative adversarial network, image features can be better extracted, which significantly improves the accuracy of image recognition. At the same time, natural language processing (NLP) based on visual recognition has also been greatly improved. The model combines the two through a long-term short-term memory network (LSTM). The image is recognized by generative adversarial network, and then the questions and visual information are input into the long-term and short-term memory networks. Through model training, answers to the questions on the images can be given. The verification on the data set DAQUQR shows that the proposed image question answering model based on generative adversarial networks can effectively improve the ability to identify question and answer images with questions, thereby significantly improving the accuracy of image question answering.

Key words: natural language processing; generative adversarial network; deep learning; image recognition; accuracy

0 引言

随着自然语言处理^[1]和图像理解的进步,更复杂和更苛刻的任务已经触手可及。该文的目标是利用最新的发展来改变现实世界中回答自然语言问题的现状。这个任务结合了对问题意图的推断和视觉场景理解与单词序列预测任务相结合。最近,基于分层的、端到端可训练的人工神经网络架构,已经在不同任务中改善了技术水平。最显著的是生成对抗网络提高了图

像识别的准确率,而长短期记忆网络在一系列序列预测任务(如机器翻译)中占主导地位。近年来,这两种神经结构已卓有成效地与方法相结合以生成图像和视频描述。两者都是针对源自深度学习架构的视觉特征,并使用递归神经网络方法来产生描述。

为了进一步拓展深度学习架构的边界并探索其局限性,该文提出了一种解决图像问题的架构。与之前的工作相比,这项任务需要语言和视觉输入的训练。

收稿日期:2020-06-28

修回日期:2020-10-29

基金项目:国家海洋局重大专项项目(国海科学[2016]494号 No. 30)

作者简介:程焕新(1966-),男,博士,教授,研究方向为人工智能、图像识别等;通讯作者:张志浩(1994-),男,硕士研究生,研究方向为计算机视觉。

这两种模式都必须被解释,并共同表示为一个答案,这取决于问题的推断意义和图像内容。

1 生成对抗网络

生成对抗网络如图 1 所示,是由 Ian Goodfellow 等人于 2014 年首次提出的神经网络模型,是一种深度学习模型,也是近年来复杂分布上无监督学习最具前景的方法之一。模型通过框架中(至少)两个模块—生成模型(generative model)和判别模型(discriminative model)的互相博弈学习产生相当好的输出。原始 GAN 理论中,并不要求 G 和 D 都是神经网络,只需要能拟合相应生成和判别的函数即可。但实用中一般均使用深度神经网络作为 G 和 D。一个优秀的 GAN 应用需要有良好的训练方法,否则可能由于神经网络模型的自由性而导致输出不理想。

GAN 的核心思想源于博弈论的纳什均衡。设定参与游戏的双方分别为一个生成器(generator)和一个判别器(discriminator),生成器捕捉真实数据样本的潜在分布,并生成新的数据样本;判别器是一个二分类器,判别输入是真实数据还是生成的样本。为了取得游戏胜利,这两个游戏参与者需要不断优化,各自提高自己的生成能力和判别能力,这个学习优化过程就是寻找二者之间的一个纳什均衡。

同时需要注意的是生成模型与对抗模型是完全独立的两个模型,它们之间没有什么联系。那么训练采用的大原则是单独交替迭代训练。因为是两个网络,不方便一起训练,所以才交替迭代训练。

GAN 的强大之处在于能自动学习原始真实样本

集的数据分布,不管这个分布多么复杂,只要训练得足够好就可以学出来。

传统的机器学习方法,一般会先定义一个模型,再让数据去学习。比如知道原始数据属于高斯分布,但不知道高斯分布的参数,这时定义高斯分布,然后利用数据去学习高斯分布的参数,得到最终的模型。再比如定义一个分类器(如 SVM),然后强行让数据进行各种高维映射,最后变成一个简单的分布, SVM 可以很轻易地进行二分类(虽然 SVM 放松了这种映射关系,但也给了一个模型,即核映射),其实也是事先知道让数据该如何映射,只是映射的参数可以学习^[2]。

以上这些方法都在直接或间接地告诉数据该如何映射,只是不同的映射方法能力不一样。而 GAN 的生成模型最后可以通过噪声生成一个完整的真实数据(比如人脸)^[3],说明生成模型掌握了从随机噪声到人脸数据的分布规律。GAN 一开始并不知道这个规律是什么样,也就是说 GAN 是通过一次次训练后学习到的真实样本集的数据分布^[4]。因此生成对抗网络在计算机视觉的图像生成和 NLP 的生成式对话内容等方面表现得非常好。简单说:就是机器可以根据需要生成新的图像和对话内容^[5],生成对抗网络(GAN)通过生成器和判别器的机制可以更好地通过图像内容和问题来推断含义^[6]。有大量关于自然语言理解的工作已经解决了基于语义解析、符号表示和演绎系统的文本问答,使得将自然语言理解用在图像问答上成为了可能^[7],因为需要通过工作来寻求端到端的架构,这些架构学习在一个单一的整体和单一的模型中回答问题。

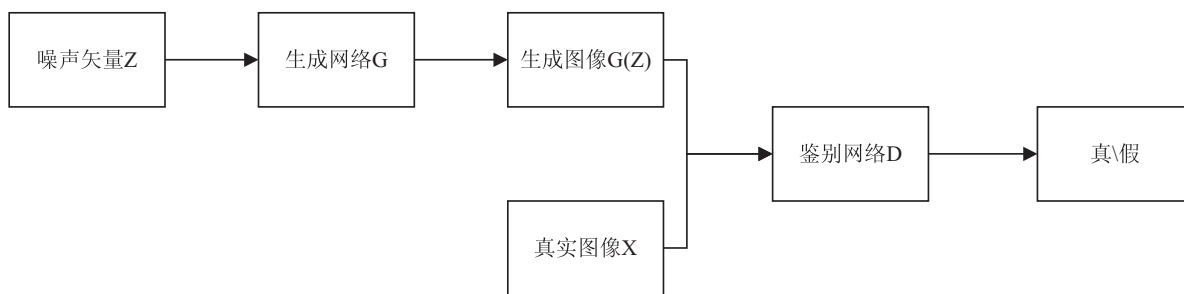


图 1 生成对抗网络模型

该文提出了“生成对抗-图像问答”(generative adversarial-image-QA),一种解决图像问答问题的神经网络模型,网络结构如图 2 所示。图像通过生成对抗网络(GAN)进行分析,问题和图像表示一起输入到长短时记忆(LSTM)网络中。该系统经过训练,能够对图像上的问题给出正确的答案。GAN 和 LSTM 是从单词和像素开始的端到端的联合训练。

由于该方法涉及机器学习、计算机视觉和自然语言处理的不同领域,所以通过以下方式组织了相关工

作:生成对抗神经网络用于视觉识别。最近,生成对抗神经网络(GAN)在视觉识别方面取得了成功,故而在此基础上进行了研究。生成对抗网络通过不断生成和原始数据相似的图像,通过不断训练、不断逼近真实图像,从而提高了图像的识别准确度。生成对抗网络在过去两年中取得了迅速进展,因此图像识别方面可以使用一组准确的模型^[8]。

递归神经网络(RNN)用于序列建模。递归神经网络允许神经网络处理灵活长度的序列。一种称为长

短期记忆(LSTM)的特殊变体在自然语言任务(例如 机器翻译)上显示出近期的成功。

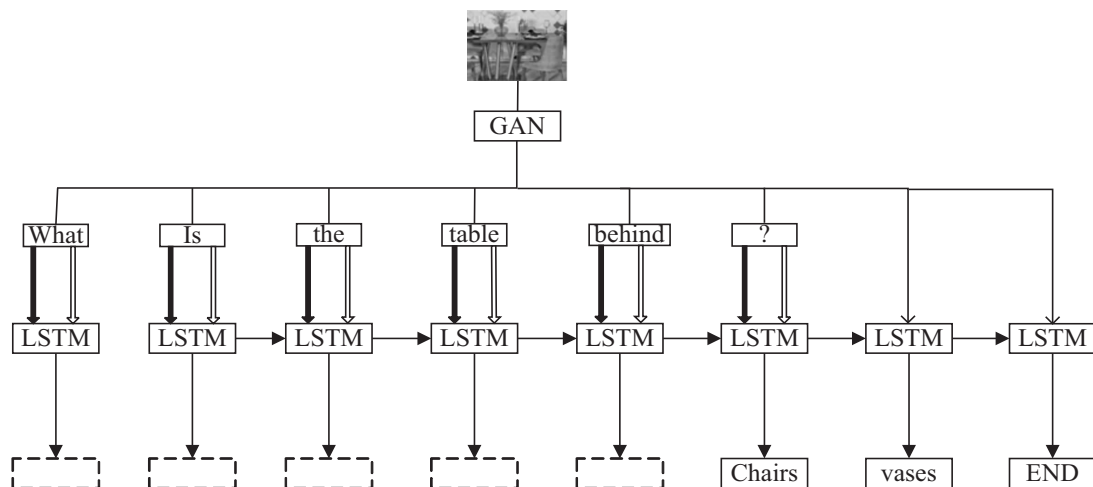


图2 生成对抗-图像问答模型网络结构

结合 GAN 和 LSTM 来描述视觉内容。描述先前的两个想法的任务已成功解决了描述静态内容以及视频之类的视觉内容的任务^[9]。这是通过使用 RNN 类型的模型来实现的。该模型首先可以观察视觉内容,并经过训练可以事后预测代表视觉内容的单词顺序。文中的工作是将这一思想扩展到问题回答,在那里制定了一个经过训练的模型来生成一个基于视觉和自然语言输入的答案^[10]。

在处理自然语言输入时,确实涉及到单词与意义的联系^[11]。这通常被称为接地问题—特别是如果“意义”与感官输入相关。遵循这样的思想,即不强制或评估任何特定的“意义”在语言或图像形态上的表现,从而将其视为潜在的,并将其留给联合训练方法来为问题回答任务建立适当的内部表示。

文本问题回答。对纯文本问题的回答已经在 NLP 社区中进行了研究^[12],并且最先进的技术通常使用语义解析来获得捕获预期含义并推断相关答案的逻辑形式^[13]。直到最近,前面提到的神经序列模型才延续到这项任务中。更具体地说,使用依赖树递归神经网络代替 LSTM,将问答问题简化为分类任务。

视觉图灵测试。最近有几个方法被提出来接近视觉图灵测试,即回答关于视觉内容的问题。例如,D. Geman、S. Geman 在计算机视觉系统的视觉图灵测试中提出了一个二进制(是/否)版本的可视化图灵测试合成数据。在 M. Malinowski and M. Fritz 的一种基于不确定输入的关于真实世界场景的多世界问题回答方法中^[14],提出了一个基于语义解析器的问题回答系统。该语义解析器基于一组更多样化的人类问题-答案对。

相比之下,在这项工作中,文中方法是基于神经结构的,是端到端的训练,直接通过图像来进行问题回答,因此该方法将问答系统从语义解析器中解放出来。

2 语言处理

在图像上回答问题是根据参数概率测度预测给定图像 x 和问题 q 的问题:

$$\hat{a} = \operatorname{argmax} p(a | x, q; \theta)$$

$$a \in A$$

所有参数 θ 表示一个向量的学习,是一组所有的答案。后面描述如何代表 x, a, q 和 $p(\cdot | x, q; \theta)$ 更多的细节。语言问答模型如图3所示。

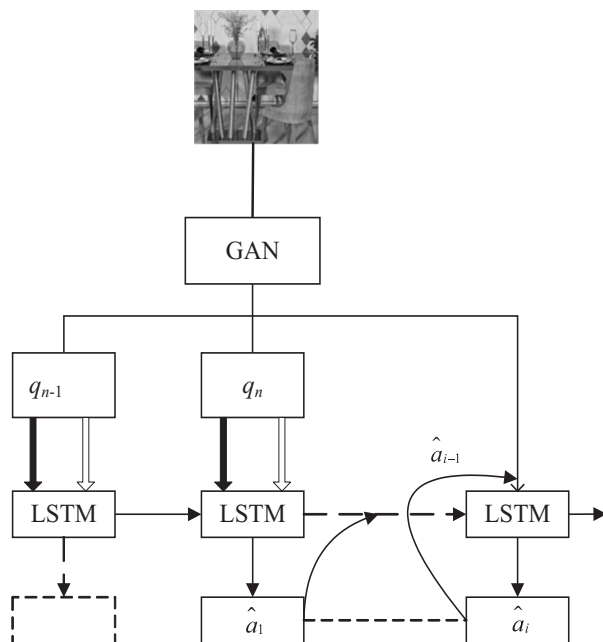


图3 语言问答模型

在文中场景,问题可以有多个单词答案,因此将问题分解为预测一组答案单词 $a_{q,x} = \{a_1, a_2, \dots, a_{N(Q,X)}\}$, 其中 a_i 是有限词汇表中的单词, V 和 $N(Q, X)$ 是给定问题和图像的答案词数。要预测多个单词,需要根据词汇表 $v: = v' \cup \{\$ \}$ 来预测单词的顺序,其中额外的标记 $\$$ 表示答案序列的结束,并指出问题已

经完全回答完毕。因此,递归地建立了预测过程:

$$\hat{a} = \operatorname{argmax} p(a | x, q, \hat{a}_{1:t-1}; \theta)$$

$$a \in v$$

其中, $\hat{a}_{1:t-1} = \{\hat{a}_1, \dots, \hat{a}_{t-1}\}$ 是上一个词的集合, 当文中方法到目前为止没有给出任何答案时, $\hat{a}_0 = \{\}$ 开始, 当 $\hat{a}_t = \$$ 时, 方法终止。仅基于预测的答案词来评估该方法, 却忽略了额外的词汇 $\$$ 。为了确保预测答案词的唯一性, 当要预测答案词集时, 可以通过最大化 $v/\hat{a}_{1:t-1}$ 来简单地更改预测过程。但是, 在实践中, 该算法会预测任何先前预测的单词。由于问题的形式是可变长度的输入/输出序列, 因此使用递归神经网络和 softmax 预测层对生成对抗—图像问答的参数分布 $p(\cdot | x, q; \theta)$ 进行建模。

如图 1 和图 2 所示, 通过向生成对抗—图像问答模型输入一个由单词组成的问题, 即 *i. e.* $q = [q_1, \dots, q_{n-1}, [?]]$, 每个 q 是第 t 个单词问题, $[?] = q_n$ 是问题的结尾。由于该问题是一个变量输入/输出序列,

可以用递归神经网络和 softmax 预测层对生成对抗—图像问答的参数分布 $p(\cdot | x, q; \theta)$ 进行建模。更准确地说, 生成对抗—图像问答模型是由 GAN 和 LSTM 构建的深度网络。最近, LSTM 在学习可变长度序列到序列映射方面被证明是有效的。

问答词均采用单热向量编码 (在词汇表中表示该词的索引位置上有一个非零项的二进制向量) 表示, 并嵌入到较低维空间中, 采用联合学习的潜在线性嵌入。在训练阶段, 通过增加相应的疑问词序列 q 与相关的答案序列 a , 即 *i. e.* $\hat{q}_t = [q, a]$ 。在测试期间, 在预测阶段, 在 t 时间步先增加问前预测答案单词 $\hat{a}_{1:t-1} = [\hat{a}_1, \dots, \hat{a}_{t-1}]$, *i. e.* $\hat{q}_t = [q, \hat{a}_{1:t-1}]$ 。这意味着问题 q 和之前的答案被隐式编码在 LSTM 的隐藏状态中, 同时学习潜在的隐藏表示。使用 GAN 对图像 x 进行编码, 并在每个时间步中将其作为输入提供给 LSTM。设置输入区域 $[x, \hat{q}_t]$ 的连接。

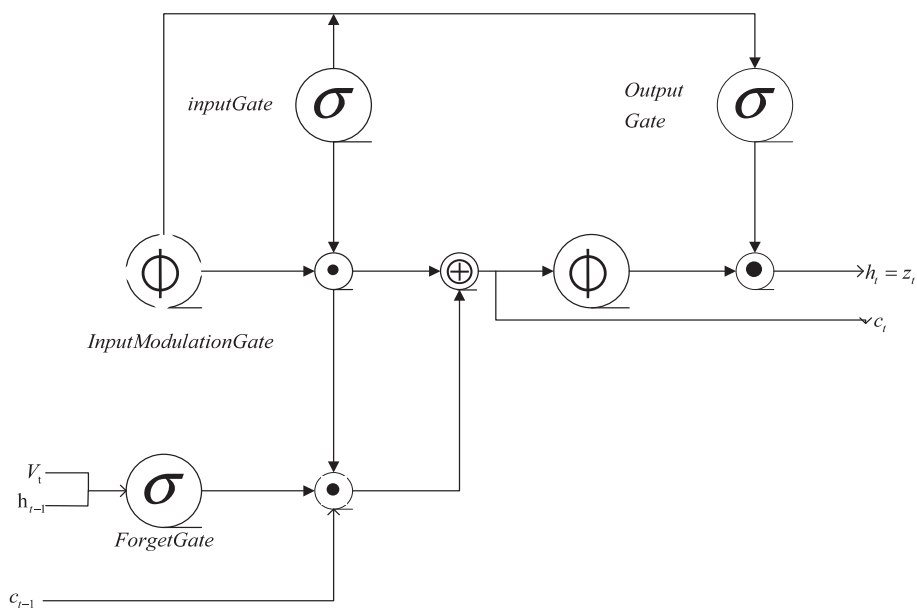


图 4 LSTM 网络单元

如图 4 所示, LSTM 单元在每个时间步长 t 中取一个输入向量 v_t , 并预测一个输出词 z_t 等于其潜在隐藏状态 h_t 。如上所述, z_t 是一个相应的线性嵌入答案词。与一个简单的 RNN 单元相比, LSTM 单元额外维护了一个内存单元 c 。这允许更容易地学习长期动态, 并显著减少消失和爆炸梯度问题。如图 2 和图 3 所示, 所有出现在问号之前的输出预测都被排除在损失计算之外, 因此模型仅根据预测的答案词进行惩罚。

通过设定 LSTM 和 GAN 的默认超参数, 所有 GAN 模型首先在 ImageNet 数据集上进行预训练, 然后在任务上随机初始化和训练最后一层和 LSTM 网络。结果发现这一步对获得良好的成绩至关重要。尽管已经探索了使用 2 层 LSTM 模型, 但始终性能较差。

3 实验与分析

3.1 实验设置

在本次实验中, 将以回答关于图像的问题为任务对文中的模型方法进行基准测试。通过将该模型的不同变体与之前的工作进行比较, 从而观测该模型的图像问答的准确率。此外, 作为对比分析了在不使用图像的情况下如何很好地回答问题, 以先验知识和常识的形式来理解偏差。为这项任务提供了一个新的人类基线。同时将讨论问题回答任务中的歧义, 并通过引入对这些现象敏感的度量来进一步分析它们。特别是, WUPS 评分被广泛扩展为考虑多种人类答案的共识度量。

3.2 WUPS 分数

文中实验和共识度量都基于 WUPS 得分。该度量是解释答案单词中单词级歧义的准确性度量的一般化。例如,“纸箱”和“盒子”可以和一个类似的概念联系起来,因此,模型不应该因为这种类型的错误而受到严厉的惩罚。正式:

$$WUPS(A, T) = \frac{1}{N} \sum_{i=1}^N \min \left\{ \prod_{t \in T_i} \max_{a \in A_i} \mu(a, t), \prod_{a \in A_i} \max_{t \in T_i} \mu(a, t) \right\}$$

为了包含上述的歧义,建议对 μ 使用基于阈值分类的 Wu-Palmer 相似度。门槛越小,衡量标准就越宽容。在本实验中,采用的 WUPS 处于两个极端,0.0 和 0.9。

3.3 数据集

该文采用了 DAQUQR 数据集,它是 VQA 第一个重要的数据集。在 DAQUAR 数据集上进行了实验。该数据集在室内场景图像上提供了 12 468 个人类问题答案对,并通过提供准确性和 WUPS 得分为 {0.9, 0.0} 的结果,遵循相同的评估协议。通过对整个数据集及其缩减集进行实验,该缩减集将输出空间限制为仅 35 个对象类别和使用 30 个测试图像。此外,还评估了仅存在 1、2、3 或 4 个单词答案的 DAQUAR 的不同子集的方法。

3.4 实验结果分析

表 1 显示了整套(“多个单词”)上的 Generative adversarial-Image-QA 方法的结果,其中包含 645 张图像和 5 200 个问答对。另外,评估一种经过训练只能预测单个单词(“单个单词”)的变体以及不使用视觉功能的变体(“仅语言”)。与先前的工作(在表 1 中显示)相比,发现该模型准确性提高了 9% 以上,WUPS 分数提高了 11% 以上(表 1 中的第二行对应“多个单词”)。请注意,尽管事实是唯一可用于整套比较的已发布数字使用了真实对象注释,但仍实现了这一改进—使文中方法处于不利地位。当仅对单个单词的答案进行训练时,就会观察到进一步的改进,这会使先前工作中获得的准确性提高一倍。将此归因于语言和视觉表示以及数据集偏差的联合训练,其中约 90% 的答案仅包含一个单词。

表 1 生成对抗-图像问答模型不同的 CMC 比较

	准确率	0.9 的 WUPS	0.0 的 WUPS
Malinowski et al.	7.86	11.86	38.79
生成对抗模型 多个单词	27.52	33.15	67.82
单个单词	29.46	35.33	72.12
回答	60.32	60.79	77.42

续表 1

	准确率	0.9 的 WUPS	0.0 的 WUPS
仅语言输入 多个单词	27.12	32.42	66.37
单个单词	27.24	32.42	68.25
回答无图片	17.44	23.24	45.75

根据答案中的单词数(由于性能下降而被截断为 4 个单词),显示了该模型方法的性能(“多个单词”)。单字子集上“单个字”变体的性能显示为水平线。尽管对于较长的答案,准确性会迅速下降,但是文中模型能够产生大量正确的两个单词的答案。“单个单词”变体在单个答案上有优势,并受益于数据集对此类答案的偏见。

表 2 显示了对 DAQUAR 的单词答案子集的“单词”模型的定量结果。尽管文中与先前的工作相比有了实质性的进步,但仍然可以提高 30% 的人类准确度和 25% 的 WUPS 评分(表 1 中的“回答”)。

表 2 单词对生成对抗-图像问答模型的影响

	准确率	0.9 的 WUPS	0.0 的 WUPS
生成对抗-图像 问答模型	42.57	47.87	85.24
模型仅语言输入	39.22	45.25	80.45

同时为了与 M. Malinowski 中所提出的多世界方法相比较,还在缩减集上进行了模型的测试,在测试时,该缩减集包含 35 个对象类和仅包含 298 个问题-答案对的 30 幅图像。如表 3 所示,生成对抗图像问答模型在缩减的 DAQUAR 集上也有改进,准确率达到了 45.12%,在 0.9 的 WUPS 也达到了 51.67%,大大优于 M. Malinowski 的 12.73% 和 18.10%。与之前的实验相似,使用“单字”变体获得了最佳性能。

表 3 生成对抗-图像问答模型在缩减数据集的表现

	准确率	0.9 的 WUPS	0.0 的 WUPS
Malinowski et al.	12.73	18.10	51.47
生成对抗模型 多个单词	39.45	47.50	89.54
单个单词	45.12	51.67	89.67
仅语言输入 多个单词	42.23	49.42	91.14
单个单词	41.56	48.25	90.19

为了研究问题中已经包含了多少信息,训练了一个忽略视觉输入的模型版本。结果显示在表 1 和表 3 下的“仅语言输入”。单个单词的“仅语言输入”的模型(27.24% 和 41.56%)在准确性方面与包括视觉的最佳模型相比表现还是不错的。后者在完整数据集和

缩减数据集上分别达到 29.46% 和 45.12%。

4 结束语

该文提出了一种神经结构,用于回答关于图像的自然语言问题,与之前基于语义分析的工作形成对比,并在这个具有挑战性的任务中加倍的表现,使之前的工作表现更好。在同样的条件下,一个不使用图像来回答问题的模型只比文中提出的模型表现略差。从而得出的结论是,该模型已经学会了偏见和模式,这些可以被看作是人类用来完成这项任务的常识和先验知识的形式。同时这个模型还有许多不足之处,可观察到,室内场景统计、空间推理和小物体并没有被 GAN 的全局表示很好地捕捉到,这种表示的真正局限性只能在更大的数据集上探索。

参考文献:

- [1] COLLOBERT R, WESTON J, LÉON B, et al. Natural language processing (almost) from scratch[J]. *Journal of Machine Learning Research*, 2011, 12(1): 2493–2537.
- [2] 徐一峰. 生成对抗网络理论模型和应用综述[J]. *金华职业技术学院学报*, 2017, 17(3): 81–88.
- [3] 曹志义, 牛少彰, 张继威. 基于半监督学习生成对抗网络的人脸还原算法研究[J]. *电子与信息学报*, 2018, 40(2): 323–330.
- [4] 戴臣超, 王洪元, 倪彤光, 等. 基于深度卷积生成对抗网络和拓展近邻重排序的行人重识别[J]. *计算机研究与发展*, 2019, 56(8): 1632–1641.
- [5] 唐贤伦, 杜一铭, 刘雨微, 等. 基于条件深度卷积生成对抗网络的图像识别方法[J]. *自动化学报*, 2018, 44(5): 855–864.
- [6] 孙 钰, 李林燕, 叶子寒, 等. 多层次结构生成对抗网络的文本生成图像方法[J]. *计算机应用*, 2019, 39(11): 3204–3209.
- [7] 牛 斌, 吴 鹏, 马 利, 等. 一种基于生成对抗网络的行为数据集扩展方法[J]. *计算机技术与发展*, 2019, 29(7): 43–48.
- [8] 徐 戈, 王厚峰. 自然语言处理中主题模型的发展[J]. *计算机学报*, 2011, 34(8): 1423–1436.
- [9] 王灿辉, 张 敏, 马少平. 自然语言处理在信息检索中的应用综述[J]. *中文信息学报*, 2007, 21(2): 35–45.
- [10] 奚雪峰, 周国栋. 面向自然语言处理的深度学习研究[J]. *自动化学报*, 2016, 42(10): 1445–1465.
- [11] KANTOR P B. Foundations of statistical natural language processing[J]. *Information Retrieval*, 2001, 4(1): 80–81.
- [12] ADAM L B T, PIETRA V J D, PIETRA S A D. A maximum entropy approach to natural language processing[J]. *Computational Linguistics*, 2002, 22(1): 5–8.
- [13] COHEN K B, DOLBEY A. Foundations of statistical natural language processing (review)[J]. *Language*, 2002, 78(3): 599.
- [14] TELLER V. Review of “speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition” by Daniel Jurafsky and James H. Martin. Prentice Hall 2000 [J]. *Computational Lingus*, 2000, 26(4): 638–641.
- [15] WANG Q, GAO J, XING J, et al. DCFNet: discriminant correlation filters network for visual tracking[C]//IEEE international conference on computer vision and pattern recognition. Hawaii: IEEE, 2017: 1121–1127.
- [16] DANELLJAN M, HAGER G, KHAN S F, et al. Learning spatially regularized correlation filters for visual tracking [C]//2015 IEEE international conference on computer vision. Santiago, Chile: IEEE, 2015: 4310–4318.
- [17] WU Y, LIM J, YANG M H. Object tracking benchmark[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(9): 1834–1848.
- [18] WU Y, LIM J, YANG M H. Online object tracking: a benchmark[C]//Proceedings of IEEE conference on computer vision and pattern recognition. Portland, OR: IEEE, 2013: 2411–2418.

(上接第 174 页)

classification with deep convolutional neural networks[J]. *Advances in Neural Information Processing Systems*, 2012, 25(2): 1097–1105.

[12] WU Y, LIM J, YANG M H. Object tracking benchmark[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(9): 1834–1848.

[13] WU Y, LIM J, YANG M H. Online object tracking: a benchmark[C]//Proceedings of IEEE conference on computer vision and pattern recognition. Portland, OR: IEEE, 2013: 2411–