文章编号：2095-2163(2020)03-0111-04 中图分类号：TP183 文献标志码：A

深度学习在图像自动标注中的应用初探

魏珺洁
（西安石油大学 计算机学院，西安 710065）

摘 要：近几年，随着人工智能深度学习的不断发展，计算机视觉领域也逐渐发展扩大，先后出现了图像检索、图像自动标注等新的研究方向。最初为支持图像检索而逐渐兴起的图像自动标注技术，可以在一定程度上跨越“语义鸿沟”，让计算机自动给图像加上能够反映图像内容的语义描述，从而减少人工标注成本。深度学习作为人工智能领域的新技术，其复杂的神经网络结构能够在学习到图像特征后快速输出结果，如果将深度学习应用于图像自动标注，将大大节约人工标注时间，降低人工标注成本。文章为探究深度学习在图像自动标注上的可行性，将以作者的生活照为样本数据，使用深度卷积神经网络与深度循环神经网络进行图像处理，最后输出图像的文字描述。

关键词：深度学习；深度卷积神经网络；深度循环神经网络；图像自动标注

Preliminary study on the application of deep learning in automatic image annotation

WEI Junjie
（College of Computer, Xi´an Shiyou University, Xi´an 710065, China）

【Abstract】 In recent years, with the continuous development of artificial intelligence deep learning, the field of computer vision has gradually developed and expanded, and new research directions such as image retrieval and automatic image annotation have emerged. The automatic image annotation technology, which was originally developed to support image retrieval, can cross the “semantic gap” to a certain extent, allowing the computer to automatically add a textual description of the image content to the image, thereby reducing the cost of manual labeling. As a new technology in the field of artificial intelligence, the complex neural network structure of deep learning can quickly output results after learning image features. If applied to automatic image annotation, deep learning will greatly save manual labeling time and reduce manual labeling cost. In order to explore the feasibility of deep learning in automatic image annotation, the article will take the author´s photos of life as sample data, use deep convolutional neural network and deep recurrent neural network for image processing, and output the text description of the image.

【Key words】 deep learning; deep convolutional neural network; deep recurrent neural network; automatic image annotation

0 引 言

深度学习是一种试图使用包含复杂结构或由多重非线性变换构成的多个处理行对数据进行高层抽象的算法，深度神经网络能够通过多层网络进行特征学习，从大量的数据中学习规律，从而实现预测、识别等结果[1]。在计算机视觉领域，深度学习不仅能够实现图像分割[2]、图像分类[3]及图像识别[4]，还可以用于图像检索[5]、图像超分辨率重建[6]、目标检测[7]等方面。

图像自动标注就是让计算机自动地给输入图像生成能够反映图像内容的语义描述。在此过程中，是利用已经标注的图像作为训练数据，将训练数据输入模型中，使模型在图像的高层语义信息和低层特征之间建立一种映射关系，从而使用此模型对未知语义的图像进行自动标注[8]。文章使用的是 Vinyals 等人[9]提出的 Encoder-Decoder 模型，该模型中的 Encoder 为编码器，是一个深度卷积神经网络（Deep CNN），常用于图像识别，目标检测等领域；Decoder 为解码器，是一个深度循环神经网络（Deep RNN），常用于语言模型或机器翻译等领域。文章将图像数据输入深度学习 Encoder-Decoder 模型后，由编码器负责提取图像特征，解码器负责获取并输出图像的文字描述，从而实现图像自动标注。

将深度学习应用于图像自动标注技术，可以有效节约人工标注的成本，减少标注时间，提高图像标注效率。对此可展开研究论述如下。

1 深度学习基础

1.1 深度卷积神经网络

卷积神经网络（Convolutional Neural Network，CNN）是一种前馈人工神经网络，由输入层、卷积层、池化层、全连接层、输出层组成，主要用于图像识别。相比于浅层卷积神经网络，深度卷积神经网络结构较复杂，一般会有几十个神经层，每一层又会有数百个神经元。深度卷积神经网络通过将输入图像嵌入

作者简介：魏珺洁（1994-），女，硕士研究生，主要研究方向：智能计算、深度学习、图像处理。

收稿日期：2019-11-15

到固定长度的向量中生成输入图像的丰富表示,具有超强的图像处理能力。

文章使用的深度卷积神经网络模型为 GooLe Net 网络中的 Inception v3 模型,GoogLe Net 中的 Inception v1 模型[10]通过采用全局平均池化层取代全连接层,极大地降低了参数量,是非常实用的模型。随后的 Inception v2 模型[11]中,引入了 Batch Normalization 方法,加快了训练的收敛速度。在 Inception v3 模型[12]中,通过将二维卷积层拆分成 2 个一维卷积层,不仅降低了参数数量,同时减轻了过拟合现象。

深度卷积神经网络在 Encoder-Decoder 模型中充当"编码器",先对其进行训练以完成图像分类任务,然后将其作为下一个隐藏层(即用作生成语句的解码器)的输入,见图 1,Vinyals 等人[9]将此结构称为 NIC 模型。

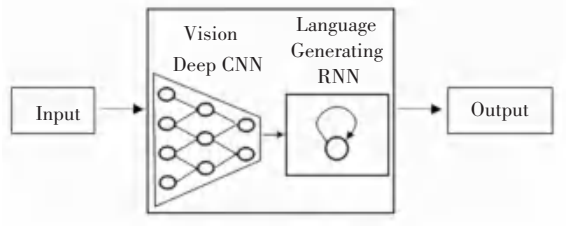


图 1 图像自动标注中的 NIC 模型

Fig. 1 NIC model in automatic image annotation

1.2 深度循环神经网络

循环神经网络(Recurrent Neural Network,RNN)是一种递归神经网络,包括输入层、隐藏层和输出层,主要用于处理序列数据,在该网络中一个序列当前的输出与之前的输入有关,其最大的特点就是神经元在某时刻的输出可以作为输入再次输入到神经元[13]。

深度循环神经网络如图 2 所示,这是一种串联的网络结构,其输入为集合$\{x_0,x_1,\ldots,x_t,x_{t+1},\ldots\}$,输出为集合$\{y_0,y_1,\ldots,y_t,y_{t+1},\ldots\}$,隐藏单元的输出为集合$\{s_0,s_1,\ldots,s_t,s_{t+1},\ldots\}$,隐藏层内的节点可以自连、也可以互连。

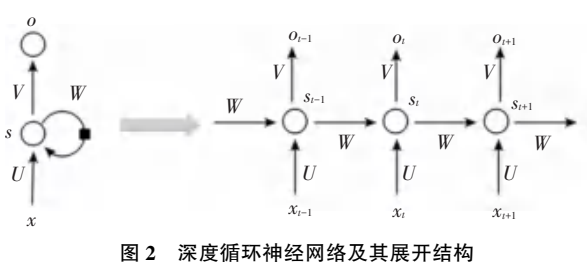


图 2 深度循环神经网络及其展开结构

Fig. 2 Deep recurrent neural network and its unfolded structure

文章使用的深度循环神经网络模型为长短期记忆网络(Long Short-Term Memory networks,LSTM),该网络在诸如翻译之类的序列任务中显示出了较好的性能。在文章的 Encoder-Decoder 模型中,LSTM 网络作为"解码器",主要用来进行翻译和序列生成。研究中用 I 表示输入图像,用 $S=(S_0,\ldots,S_N)$ 表示描述该图像的真实句子,则展开过程为:

$$\boldsymbol{x}_{-1}=CNN(I), \tag{1}$$

$$\boldsymbol{x}_t=\boldsymbol{W}_e\boldsymbol{S}_t,\ t\in\{0\ldots N-1\}, \tag{2}$$

$$\boldsymbol{p}_{t+1}=LSTM(x_t),\ t\in\{0\ldots N-1\}. \tag{3}$$

其中,一维向量 $\boldsymbol{S}_t$ 表示为每个单词,其维数等于字典的大小;S_0 表示一个特殊的开始词;S_N 表示一个特殊的停止词,用来指定句子的开头和结尾。图像和单词都映射到相同的空间,使用视觉 CNN 映射图像,使用单词 W_e 嵌入单词。图像 I 仅在 $t=1$ 时输入一次,以告知 LSTM 有关图像的内容。

2 使用 GoogLeNet+LSTM 实现图像自动标注

2.1 实验原理

本次实验使用的深度学习模型是一个 Encoder-Decoder 模型,如图 3 所示。

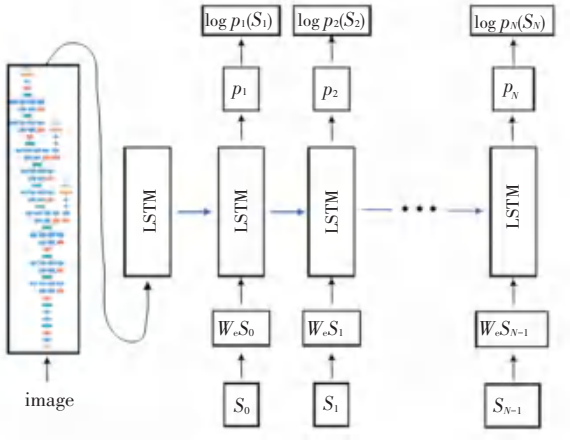


图 3 Encoder-Decoder 模型结构图

Fig. 3 Encoder-Decoder model structure

由图 3 可以看到,图 3 左侧的卷积神经网络是 Encoder 编码器,是使用 ILSVRC-2012-CLS 数据集训练出来的 Inception-V3 模型。把图片输入 Inception-V3 中,可以得到一个固定长度的向量,这个向量可以看成是从图像中提取出来的特征。图 3 右侧的循环神经网络是 Decoder 解码器,这是一个 LSTM 模型,常用于语言模型或机器翻译等领域。文中的实验是把 Encoder 中输出的固定长度的向量输入到 Decoder 中,获得关于图像的描述。

图 3 中的 $\{S_0,S_1,\ldots,S_{N-1}\}$ 表示图像的描述,每个 s 代表一个词,图 3 中的 $\{W_eS_0,W_eS_1,\ldots,$

$W_eS_{N-1}\}$ 是每个词的词向量，比如 word2vec。输出的$\{p_1, p_2, \ldots, p_N\}$ 表示 LSTM 模型预测句子中的下一个词所对应的概率分布。$\{\log p_1(S_1), \log p_2(S_2), \ldots, \log p_N(S_N)\}$ 表示正确词的对数似然估计。

2.2　实验过程

（1）配置实验环境：本次实验工具为 Anaconda3，Spyder（Tensorflow），实验环境为 Windows10，Python3.6，Tensorflow1.12.0。

（2）导入模型：本次实验模型是使用“2015 MSCOCO Image Captioning Challenge”的数据集训练出来的深度学习模型，即上文提到的 Encoder-Decoder 模型。该模型分为 Encoder 编码器和 Decoder 解码器两部分，编码器负责图像特征提取的工作，输入的图像在 Inception v3 网络中能够被转化为一个固定长度的向量。通过 NIC 模型，在编码器中得到的固定长度的向量将作为解码器的输入，最终通过训练好的 LSTM 网络生成对向量的文字描述。

编码器是 Inception v3 模型，共有 47 层，比以往的 CNN 网络计算速度更快，对非线性更鲁棒。解码器是 LSTM 模型，通过输入大量已经标注的图像对其进行训练形成字典，训练集中的单词每个至少出现 5 次，从而使其具有捕获语义的能力。

（3）导入数据集：从该模型的训练集中抽取的图像数据进行实验，效果较为客观，所以文中收集了一些生活照作为输入进行训练。

（4）运行：在 Tensorflow 中使用 Encoder-Decoder 模型对输入图像进行编码与解码处理，最终输出输入图像的文字描述。

2.3　实验结果

实验结果见表 1。表 1 的第一列为输入图像，第二列为输出的文字描述。具体来说，第一列从上至下可描述为：第一张图像为一个男人站在石墙旁边；第二张图像为一个女人怀里抱着一只泰迪熊；第三张图像为一个拿着伞的女人站在商店前面；第四张图像为一个年轻的姑娘坐在板凳上。

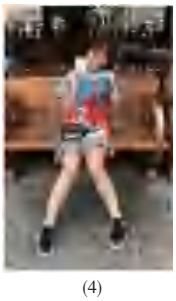
对于实验输出，即仔细观察每张图像的文字描述，分析后可知深度学习模型生成的文字描述能够表达出图像的典型特征，但是由于该模型的字典容量有限，也未能准确识别出一些特殊的物体。例如：第三张图像女生手中的棉花糖由于与伞的形状相似，该模型识别结果为伞，女生身后的装饰心愿墙识别为商店。

从实验结果不难发现，Encoder-Decoder 这一深度学习模型，确实能够实现图像自动标注。

本次实验能够证明深度学习在图像自动标注上的可行性，只是图像标注的精确程度还有待提高。

表 1　图像自动标注结果

Tab. 1　Results of automatic image annotation

输入	输出
 (1)	0) a man standing next to a stone wall. 1) a man standing next to a stone wall next to a stone wall. 2) a man standing next to a stone wall next to a wall.
 (2)	0) a woman holding a teddy bear in her arms. 1) a woman is holding a teddy bear in her arms. 2) a woman holding a teddy bear in her arms.
 (3)	0) a woman holding an umbrella in front of a store. 1) a woman holding a pink umbrella in front of a building. 2) a woman holding a pink umbrella in front of a store.
 (4)	0) a young girl sitting on a wooden bench. 1) a young girl sitting on a wooden bench. 2) a woman sitting on a bench with a book.

3　结束语

图像自动标注技术是提高图像检索效率的重要突破，同时也是人们快速获取图像信息的技术手段，而使用先进的深度学习技术来对图像进行高效自动标注就能够推动图像检索领域的发展。文章实验使用深度卷积神经网络与深度循环神经网络相结合的 Encoder-Decoder 模型，证明了深度学习实现图像自动标注的可行性，因此，深度学习能够实现图像的自动标注。

（下转第 118 页）