Journal of Computer Applications
计算机应用,2020,40(7):1926 - 1931
ISSN 1001-9081
CODEN JYIIDU
2020-07-10
http://www.joca.cn

文章编号:1001-9081(2020)07-1926-06
DOI:10.11772/j.issn.1001-9081.2019112054

基于深度视觉注意神经网络的端到端自动驾驶模型

胡学敏,童秀迟,郭　琳*,张若晗,孔　力
(湖北大学 计算机与信息工程学院,武汉 430062)
(*通信作者电子邮箱10837330@qq.com)

摘　要:针对现有端到端自动驾驶方法中存在的驾驶指令预测不准确、模型结构体量大和信息冗余多等问题,提出一种新的基于深度视觉注意神经网络的端到端自动驾驶模型。为了更有效地提取自动驾驶场景的特征,在端到端自动驾驶模型中引入视觉注意力机制,将卷积神经网络、视觉注意层和长短期记忆网络进行融合,提出一种深度视觉注意神经网络。该网络模型能够有效提取驾驶场景图像的空间特征和时间特征,并关注重要信息且减少信息冗余,实现用前向摄像机输入的序列图像来预测驾驶指令的端到端自动驾驶。利用模拟驾驶环境的数据进行训练和测试,该模型在乡村路、高速路、隧道和山路四个场景中对方向盘转向角预测的均方根误差分别为0.009 14、0.009 48、0.002 89和0.010 78,均低于对比用的英伟达公司提出的方法和基于深度级联神经网络的方法;并且与未使用视觉注意力机制的网络相比,该模型具有更少的网络层数。

关键词:自动驾驶;端到端;视觉注意力;卷积神经网络;长短期记忆网络
中图分类号:TP391.4　　**文献标志码:**A

End-to-end autonomous driving model based on deep visual attention neural network

HU Xuemin, TONG Xiuchi, GUO Lin*, ZHANG Ruohan, KONG Li
(*School of Computer Science and Information Engineering, Hubei University, Wuhan Hubei* 430062, *China*)

Abstract: Aiming at the problems of low accuracy of driving command prediction, bulky model structure and a large amount of information redundancy in existing end-to-end autonomous driving methods, a new end-to-end autonomous driving model based on deep visual attention neural network was proposed. In order to effectively extract features of autonomous driving scenes, a deep visual attention neural network, which is composed of the convolutional neural network, the visual attention layer and the long short-term memory network, was proposed by introducing a visual attention mechanism into the end-to-end autonomous driving model. The proposed model was able to effectively extract spatial and temporal features of driving scene images, focus on important information and reduce information redundancy for realizing the end-to-end autonomous driving that predicts driving commands from sequential images input by front-facing camera. The data from a simulated driving environment were used for training and testing. The root mean square errors of the proposed model for prediction of the steering angle in four scenes including country road, highway, tunnel and mountain road are 0.009 14, 0.009 48, 0.002 89 and 0.010 78 respectively, which are all lower than the results of the method proposed by NVIDIA and the method based on the deep cascaded neural network. Moreover, the proposed model has fewer network layers compared with the networks without the visual attention mechanism.

Key words: autonomous driving; end-to-end; visual attention; convolutional neural network; long short-term memory network

0　引言

作为人工智能的主要研究领域之一,自动驾驶技术能够有效地减少交通事故的发生,合理利用交通资源,缓解交通压力。传统的基于规则式的自动驾驶方法一般分为感知系统、决策系统和控制系统三大模块[1],其优点在于各个模块分工明确,可解释性强,系统稳定性高。但是由于这类方法在做决策时强烈依赖于设定的规则,因此不具备自主学习的能力。此外,基于规则式的方法中预处理的过程较多,做出决策和控制需要处理的任务也较为繁琐,并且需要诸多昂贵的传感器,其硬件成本较高。而基于深度学习的端到端自动驾驶,将决策过程视为一个黑箱,利用神经网络建立输入到输出的映射。通过模仿人类驾驶行为,输入图像信息,输出汽车转向角等控制信号。相比传统的基于规则式的方法,端到端的方法具备强大的学习能力,能够更有效降低硬件设备成本和减少预处理步骤,因此研究端到端的自动驾驶模型具有重要的学术意义和商业价值。

近年来,研究人员在端到端的自动驾驶方面做了大量的

收稿日期:2019-12-04;**修回日期:**2020-03-27;**录用日期:**2020-04-19。
基金项目:国家自然科学基金青年基金资助项目(61806076);湖北省自然科学基金青年项目(2018CFB158)
作者简介:胡学敏(1985—),男,湖南岳阳人,副教授,博士,主要研究方向:计算机视觉、机器学习;　童秀迟(1996—),女,湖北随州人,硕士研究生,主要研究方向:机器学习;　郭琳(1978—),女,湖北随州人,副教授,博士,主要研究方向:图像处理、机器学习;　张若晗(1997—),女,湖北襄阳人,硕士研究生,主要研究方向:深度学习;　孔力(1995—),男,湖北咸宁人,硕士研究生,主要研究方向:计算机视觉。

工作。Chen等[2]使用AlexNet网络,利用12 h的模拟驾驶数据训练,实现多车道高速公路的自动驾驶,该方法在高速公路数据集上表现良好,但是没有考虑输入图像前后帧之间的时间特征,在复杂路况数据集上测试结果不稳定。NVIDIA公司提出了一种基于卷积神经网络(Convolutional Neural Network, CNN)[3]的端到端自动转向模型,实现了真实道路的自动驾驶路测[4],在多种道路上取得了相对满意的结果,但同样没有利用连续帧的信息,驾驶指令预测准确性有限。文献[5]提出利用CNN和长短时记忆(Long Short-Term Memory, LSTM)网络[6]构成的深度级联神经网络来实现从图像到方向盘转角的端到端的自动驾驶,该方法利用了车辆行驶过程中的时间信息,性能有所改进,但是网络体量大,模型训练需要的迭代次数多。加州大学伯克利分校构建了一种FCN-LSTM(Fully Convolutional Network-Long Short-Term Memory)分支网络结构[7],并引入语义分割方法增强对驾驶场景的理解能力,预测离散或连续的驾驶行为。北京大学提出的ST-Conv+ConvLSTM+LSTM网络[8],利用时空卷积、多尺度残差聚合、卷积长短记忆网络和长短时记忆网络等搭建技巧或模块,预测无人车的横向和纵向控制。另一方面,由于深度强化学习在许多传统游戏中取得了超越人类的成绩,其在自动驾驶方面的应用开始受到越来越多的关注。Mobileye将在指定环境中进行安全的多智能体规划决策应用于自动驾驶,使用策略梯度迭代的方法求解最优策略,将学习目标划分为可学习和不可学习部分保障系统安全,并引入有向无环图降低了模型的复杂度[9]。El Sallab等[10]采用深度确定性策略梯度算法在开源赛车模拟器TORCS(The Open Racing Car Simulator)中训练智能体。深度强化学习方法在模拟环境下取得不错效果,是具有潜力的自动驾驶研究方法之一。

现有基于深度神经网络的端到端自动驾驶方法往往利用CNN提取视觉图像中所有像素点的特征,但是没有考虑图像中冗余信息,存在设计的网络层数多、计算量大等问题。反观人类在驾驶时能够通过快速扫描前方,获取需要重点关注的目标区域,也就是注意力焦点,而后对这一区域投入更多注意力资源,以获取更多所需要关注目标的细节信息,而抑制其他无用信息,这一过程称为生物视觉注意力机制[11]。近年来随着深度学习的不断发展,视觉注意力机制的概念被引入这一领域[12-14]。文献[15]中将视觉注意力机制分为软注意力机制和硬注意力机制。软注意力机制为每一个输入分配一个注意力权值,其选择的信息是所有输入信息在注意力权值分布下的期望。软注意力机制平滑可微,可以被嵌入模型中直接训练,通过梯度下降法反向传播至模型其他部分。硬注意力机制使用最大采样或随机采样选取信息,只关注某一输入向量,其损失函数与注意力分布之间的函数关系不可导,因此难以使用反向传播算法进行训练。此外,Google机器翻译团队提出自注意力模型,一种将单个序列的不同位置联系起来搜索序列内部的隐藏关系的注意力机制,并将其应用于学习文本表示[16]。由于在驾驶过程中驾驶员会重点关注车道线和交通灯等信息,而给予天空、路边的建筑物和植物等背景较少的关注[17],而CNN在提取图像特征时对待每个像素均无差别,存在大量的信息冗余,降低处理效率和准确性。因此,在端到端自动驾驶模型中加入视觉注意力机制,能够选择性提取重要信息,减少模型层数和提高驾驶指令预测的准确性。

针对现有端到端自动驾驶方法中存在的驾驶指令预测准确性不高、模型结构体量大和信息冗余等问题,本文提出一种基于深度视觉注意神经网络的端到端自动驾驶方法。首先提出一种深度视觉注意神经网络(Deep Visual Attention Neural Network, DVANN),该网络由CNN层、视觉注意层和LSTM层构成,分别用于提取单个输入序列的重要空间特征、关注有用信息并减少信息冗余和提取连续序列之间的时间特征。此外,基于DVANN,提出一种端到端的自动驾驶方法,利用前向车载相机获取连续的驾驶序列图像,预测车辆的方向盘转角。实验结果表明,本文方法不仅提高了端到端自动驾驶中动作指令预测的准确度,减少了模型层数,同时也为视觉注意力机制的应用提供了新的思路。

1　基于深度视觉注意神经网络的自动驾驶

本文提出的基于视觉注意机制的端到端自动驾驶模型如图1所示,模型输入为前向车载相机的序列图像,经过网络后输出为当前预测的方向盘转角。DVANN模型由CNN层、视觉注意层和LSTM层三部分组成:CNN层用于对每一帧图像提取空间特征;视觉注意层的作用旨在判别图像的注意力权重,区分图像中各个像素点的视觉重要性;LSTM层用于提取连续帧图像的时间特征。最后输出层为1个节点,即方向盘转向角的预测结果。

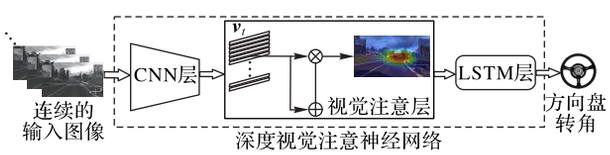

图1　基于深度视觉注意神经网络的自动驾驶模型整体结构

Fig. 1　Overall structure of autonomous driving model based on deep visual attention neural network

1.1　CNN层结构设计

在图像特征提取过程中,CNN能够利用卷积运算操作对原始图像进行高低不同层次的特征表达[18],在诸多领域特别是图像识别等相关任务上表现优异,因此本文设计一个CNN层网络来提取驾驶场景的静态图像特征。

与现有端到端的自动驾驶方法类似,本文采用CNN提取图像空间特征,将高维的输入数据编码成一系列低维的、抽象的特征表达。现有方法要实现准确的驾驶指令预测,需要设计复杂且深的CNN。本文利用注意力机制,减少CNN对网络深度的依赖,设计了一个轻量级的CNN。文献[19]中提出了一个轻量CNN来实现智能体在游戏中与环境交互,并且取得了较好的成果,因此本文以文献[19]为基础来设计本文的CNN层网络结构。如图2所示,该网络由3个卷积层构成。

原始的单帧RGB图像首先通过数据预处理转换成灰度图,并将尺寸缩放为84×84像素。为快速提取不同尺度的特征,本文采用大卷积核的方式,将三个卷积层的卷积核尺寸分别设计为8×8、4×4和3×3,步长分别为4、2和1,卷积核个数分别为32、64和64,每个卷积层后使用修正线性单元作为激活函数,因此输出为7×7像素、64通道的特征向量,作为当前帧驾驶场景的空间图像特征。最后,为将空间特征输入视觉注意层和LSTM层,将特征向量的形状改变为1×49×64的区域向量。

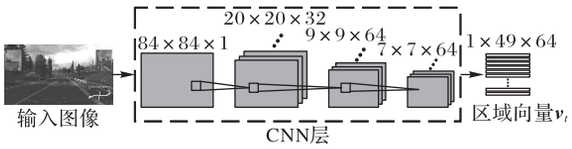

图2 CNN层结构

Fig. 2 Structure of CNN layer

1.2 LSTM层网络设计

本文使用CNN层结构能够有效提取输入图像的空间特征，然而自动驾驶任务的输入不是单帧图像，而是前后关联的图像序列，因此需要提取图像前后帧的时间特征。LSTM是循环神经网络的一种变体，可以学习长期依赖信息[20]，故本文采用LSTM作为端到端自动驾驶模型的时间特征提取层。图3中虚线矩形框展示了LSTM单元内部结构，其中$\boldsymbol{x}_t$表示t时刻LSTM单元的输入；$\boldsymbol{c}_t$表示细胞状态，记录随时间传递的信息；$\boldsymbol{i}_t$表示输入门确定$\boldsymbol{x}_t$输入多少信息给当前细胞状态$\boldsymbol{c}_t$；$\boldsymbol{f}_t$表示遗忘门决定上一时刻细胞状态$\boldsymbol{c}_{t-1}$保留多少信息给$\boldsymbol{c}_t$；$\boldsymbol{o}_t$表示输出门控制$\boldsymbol{c}_t$传递多少信息给当前状态的输出$\boldsymbol{h}_t$；$\boldsymbol{h}_{t-1}$表示$t-1$时刻的输出；$\boldsymbol{m}_t$为状态候选值。LSTM通过门控单元控制细胞状态。首先，遗忘门根据上一时刻输出$\boldsymbol{h}_{t-1}$和当前输入$\boldsymbol{x}_t$通过sigmoid层产生遗忘概率$\boldsymbol{f}_t$，决定从细胞状态中丢弃什么信息。然后分两步产生更新细胞状态的新信息，第一步输入门通过sigmoid层决定需要更新的信息$\boldsymbol{i}_t$，第二步用一个tanh层生成状态候选值$\boldsymbol{m}_t$。将上一时刻的细胞状态乘以$\boldsymbol{f}_t$再加上$\boldsymbol{i}_t \odot \boldsymbol{m}_t$得到新的细胞状态$\boldsymbol{c}_t$。最后决定输出信息，首先输出门通过sigmoid层得到初始输出$\boldsymbol{o}_t$，然后将新的细胞状态$\boldsymbol{c}_t$通过tanh函数处理后与$\boldsymbol{o}_t$相乘得到当前输出$\boldsymbol{h}_t$，其工作原理如式(1)~(6)所示：

$$\boldsymbol{i}_t = \sigma\left(\boldsymbol{W}_{xi}\boldsymbol{x}_t + \boldsymbol{W}_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{b}_i\right) \tag{1}$$

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}_{xf}\boldsymbol{x}_t + \boldsymbol{W}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{b}_f\right) \tag{2}$$

$$\boldsymbol{o}_t = \sigma\left(\boldsymbol{W}_{xo}\boldsymbol{x}_t + \boldsymbol{W}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{b}_o\right) \tag{3}$$

$$\boldsymbol{m}_t = \tanh\left(\boldsymbol{W}_{xm}\boldsymbol{x}_t + \boldsymbol{W}_{hm}\boldsymbol{h}_{t-1} + \boldsymbol{b}_m\right) \tag{4}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \boldsymbol{m}_t \tag{5}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t) \tag{6}$$

其中：$\boldsymbol{W}$与$\boldsymbol{b}$分别表示对应门控单元的权重向量与偏移量；$\sigma(\cdot)$表示sigmoid激活函数；$\tanh(\cdot)$表示双曲正切激活函数；$\odot$表示点乘。

本文设计的LSTM网络层结构如图3所示。LSTM单元的输入$\boldsymbol{x}_t$代表捕捉了特定区域视觉信息的空间特征矢量，这个量由视觉注意层计算得到，将在1.3节中详细介绍。连续的T帧图片经过CNN层和视觉注意层，输出T个在不同时间关注在不同图片区域的空间特征矢量$\boldsymbol{x}_t$。在时刻t，将空间特征矢量$\boldsymbol{x}_t$，上一个LSTM单元的输出$\boldsymbol{h}_{t-1}$和上一时刻的细胞状态$\boldsymbol{c}_{t-1}$输入LSTM单元，得到当前时刻的输出$\boldsymbol{h}_t$，再通过一个全连接(Fully Connected，FC)层得到当前方向盘转向角的预测值。T为历史数据长度，本文中T=10，为经验值。

1.3 视觉注意层设计

在图像特征提取过程中，由于CNN提取特征时无差别对待每个像素，没有考虑视觉冗余情况，造成提取的特征重点模糊，对于复杂的图像则需要通过加大网络深度来改善网络性能[21]。与之相反，人类视觉系统在感知图像信息时，能快速定位重要的目标区域并进行细致的分析。在驾驶过程中，人类往往更关注车道线、道路边缘、前方车辆和行人等障碍物、交通标志、信号灯等，而给予天空、路边建筑物等较少的关注，甚至是忽略。如果对CNN提取的驾驶场景图像特征的不同位置给予不同的权重，让网络更加关注车道线、障碍物等高重要度特征的区域，则可以更有效提取驾驶场景的图像特征，减少视觉冗余，从而更准确预测车辆的动作指令。

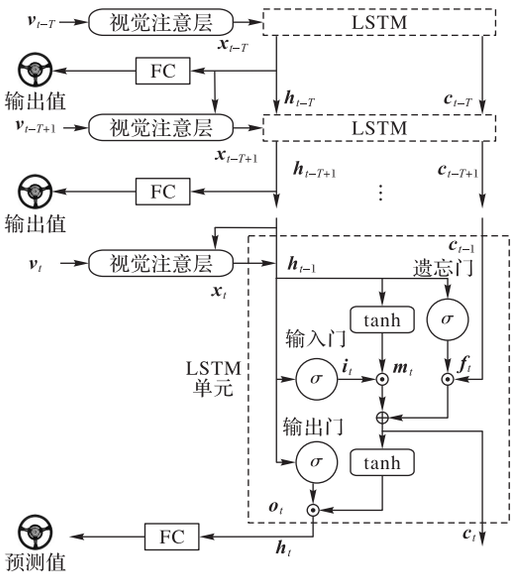

图3 LSTM层结构

Fig. 3 Structure of LSTM layer

在自动驾驶场景中，注意力机制主要用于判断图像不同位置的视觉重要性而不是内部的隐藏关系，驾驶员也不能完全忽略图像某一部分的信息只关注重点信息。由于软注意力机制平滑可微，可以被嵌入模型中直接训练，通过梯度下降法反向传播至模型其他部分，因此本文采用软注意力机制设计深度视觉注意神经网络。本文的视觉注意层结构如图4所示。为了能够更好地描述局部目标，本文针对第三个卷积层输出的特征，通过软注意力机制实现LSTM在预测转向角的不同时刻关注不同的图像区域，进而更准确地输出转向角。因此，视觉注意层的设计有两个关键的量：一个是上一时刻LSTM层产生的隐藏状态$\boldsymbol{h}_{t-1}$，与时间相关；另一个是区域向量$\boldsymbol{v}_t^i$，对应图像的一个区域。假设CNN层网络输出区域向量为$\boldsymbol{v}_t$：

$$\boldsymbol{v}_t = \left\{\boldsymbol{v}_t^1, \boldsymbol{v}_t^2, \cdots, \boldsymbol{v}_t^L\right\}; \quad \boldsymbol{v}_t^i \in \mathbf{R}^D \tag{7}$$

其中D为第三个卷积层生成的特征矢量的维度，每个向量$\boldsymbol{v}_t^i$都对应图像一个区域，表示该区域像素点对应的D维特征矢量。依据上文CNN层的介绍，L=49，D=64。

基于软注意力机制的理论，在时刻t，为输入序列的每个区域计算出一个权重，其中第i个区域的权重$\boldsymbol{e}_t^i$为：

$$\boldsymbol{e}_t^i = f_{\mathrm{FC}}\left(\tanh\left(\boldsymbol{W}_v\boldsymbol{v}_t^i + \boldsymbol{W}_h\boldsymbol{h}_{t-1}\right)\right) \tag{8}$$

其中f_{FC}表示一个节点数为64的全连接层函数，$\boldsymbol{W}_v$和$\boldsymbol{W}_h$表示视觉注意层网络中待优化的权值。采用Softmax函数使输入序列的各个区域的权重归一化，如式(9)所示：

$$\boldsymbol{\alpha}_t^i = \frac{\exp\left(\boldsymbol{e}_t^i\right)}{\sum_{k=1}^{L}\exp\left(\boldsymbol{e}_t^k\right)} \tag{9}$$

其中，$\boldsymbol{\alpha}_t^i$为图像区域向量$\boldsymbol{v}_t^i$在时刻t输入LSTM的信息中所占的权重，即图像中各个区域的重要性的权重，因此将各区域向量$\boldsymbol{v}_t^i$与对应的权重$\boldsymbol{\alpha}_t^i$做加权求和则得到空间特征向量$\boldsymbol{x}_t$，如

式(10)所示:

$$\boldsymbol{x}_t = \sum_{i=1}^{L} \boldsymbol{\alpha}_t^i \boldsymbol{v}_t^i \tag{10}$$

由于$\boldsymbol{\alpha}_t^i$是通过注意力机制计算得到的图像中不同区域的权重,而$\boldsymbol{x}_t$是原始CNN空间特征利用$\boldsymbol{\alpha}_t^i$计算加权的结果,因此$\boldsymbol{x}_t$能够代表图像特定区域中视觉信息的空间特征向量。本文在提取驾驶场景时间和空间特征中引入了视觉注意力机制,因此模型从输入图像中选择和关注某些相对小的图像区域,相比原始的CNN结合LSTM网络结构能够更有效提取图像空间特征,并且可以在设计CNN时减少网络层数,从而提高模型性能和缩短训练时间。

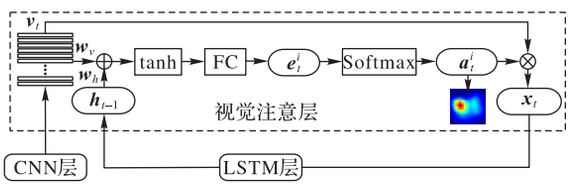

图4　视觉注意层结构

Fig. 4　Structure of visual attention layer

1.4　目标函数与网络的训练

由于本文的预测输出值只有转向角这一个连续的参数,故模型的输出节点数设定为1。在训练过程中为了解决梯度消失和梯度爆炸的问题,将方向盘转角值进行线性变换到40~60(经验值)。50代表直行,60和40分别代表向右和向左打满方向盘。本文中为了清晰直观地显示实验结果,对测试结果进行归一化:0代表直行,1和-1分别代表向右和向左打满方向盘。为了训练神经网络拟合连续值训练样本,本文采用L2范数作为损失函数,如式(11)所示:

$$L(p_g, p;\ \boldsymbol{w}) = \left\| p_g - p \right\|_2 \tag{11}$$

其中:p_g和p分别表示转向角的真实值和预测值;$\boldsymbol{w}$为网络中的参数集合。为了求解损失函数的最小值,本文使用Adam优化算法[22]。因此,本文设计的目标函数更新方法如式(12):

$$\boldsymbol{w}^* \leftarrow \min \frac{1}{n} \sum L(p_g, p;\ \boldsymbol{w}) \tag{12}$$

其中:$\boldsymbol{w}^*$为优化的目标网络权值;n为训练批次大小,本文取值为24。迭代总次数设置为5 000,学习率为0.000 1。网络的训练停止条件为训练的输出误差收敛到9.0。

2　实验结果和分析

本文实验使用Python语言编写程序,深度学习框架采用TensorFlow;硬件CPU为Intel Core i7-7700K(四核4.2 GHz)、GPU为NVIDA GTX 1080Ti、内存为32 GB。

由于自动驾驶训练风险高,以及需要在多种道路上测试,考虑到安全性问题,本文使用模拟驾驶场景数据集。欧洲卡车模拟器具有逼真的画面和丰富的驾驶场景,因此本文使用的数据集是从该模拟器中采集到的约8 h的驾驶数据,帧率为30帧/s,图像的像素尺寸为1 853×1 012。数据集共有约40万幅包含多种驾驶场景的图像,包含了乡村路、高速路、隧道和山路四种驾驶场景,除了采集前向摄像机的视频帧以外,还采集同步的方向盘转向角作为车辆的动作指令。测试时,额外针对每种场景的道路采集一段视频,且测试场景路段未包含在训练集中,四种路段测试集中包含的帧数分别为5 697、6 606、4 909和2 439。

本文旨在构建端到端的自动驾驶模型,利用监督学习的方法让模型从人类驾驶的数据集进行学习。从图像中预测驾驶指令本质上是一个回归问题,因此预测数据跟真实数据的偏差是衡量预测模型好坏的重要标准。本文参照文献[4,6-7]等,采用均方根误差(Root Mean Square Error, RMSE)作为模型准确性评价指标,计算方法如式(13)所示:

$$RMSE = \frac{1}{N} \sqrt{\sum_{i=1}^{N} (p_g(t) - p(t))^2} \tag{13}$$

其中:$RMSE$表示均方根误差,$p_g(t)$和$p(t)$分别为时刻t方向盘转向角的真实值和预测值,N为测试数据帧数。转向角为归一化后的结果。此外,本文利用注意力机制在提高驾驶指令预测的同时,减小网络体量,因此将网络的深度和模型收敛所需的训练时间和迭代次数作为网络体量的衡量标准。

为体现本文方法的有效性,将NVIDIA公司提出的自动驾驶模型[4],以及文献[5]中提出的VGG(Visual Geometry Group)和LSTM构成的深度级联神经网络(Deep Cascaded Neutral Network, DCNN)进行对比。图5、表1~3分别为测试实验结果图、均方根误差对比结果、网络深度对比结果、训练时间和迭代次数对比结果。根据实验结果,可得出如下结论:

1)本文提出的基于DVANN的端到端自动驾驶方法能在不同场景都准确预测驾驶的方向盘转向角。本文方法采用CNN和LSTM的结构能够提取不同驾驶场景序列图像的空间和时间特征,并且视觉注意力机制能够针对不同场景自适应地提取对驾驶有帮助的特征,故能够在不同场景对转向角做出准确预测。从图5可以看出,与其他两种方法相比,本文的预测曲线与真实曲线最为接近。从表1可知在四个场景中本文方法的均方误差均低于文献[4]的NVIDIA的方法和文献[5]的DCNN方法,特别是对于图像特征最不明显的隧道场景(如图5(c)所示),本文方法在准确性方面与其他两种方法相比具有明显的优势。

表1　均方根误差对比结果

Tab. 1　Comparison results of RMSE

场景	NVIDIA/10^{-3}	DCNN/10^{-3}	DVANN/10^{-3}
乡村路	12.66	9.55	**9.14**
高速路	16.05	14.68	**9.48**
隧道	6.61	5.36	**2.89**
山路	20.41	16.48	**10.78**

2)本文方法能够在提取自动驾驶图像特征的时候关注对驾驶更有用的信息。视觉注意力机制根据区域特征和上一时刻LSTM的隐藏状态给各个图像区域赋予不同权重,对需要关注的部分给予较高的权重,对不需要关注的部分给予较低的权重。从图5的视觉注意力分布中可以观察出,本文方法能够提取自动驾驶图像中车道线、车辆、转弯、指示牌等重要信息。比如乡村路由于车道线是重要关注点,山路需要重点关注转弯处,如图5(a)的t_{a2}时刻、图5(d)的t_{d1}时刻视觉注意力分布图所示,车道线和转弯处被赋予较高权重。

3)本文方法能够有效减少端到端自动驾驶中深度神经网络的层数,提高模型收敛速度。由于视觉注意力机制重点关注特征向量中一个较小的位置区域,即使较浅层的卷积神经网络也能够提取有效的视觉特征,因此在设计神经网络时,减少了卷积层的数量和模型的权重参数,加快了模型收敛。表2为各模型的网络深度,表3为训练时间和迭代次数。虽然DVANN增加了视觉注意层,但是本文所采用的模型的网络总

层数相较于文献[4]和文献[5]明显减少，模型收敛迭代次数也大幅度降低，仅为文献[5]的2.5%。轻量级的模型不仅降低了对硬件条件的要求，而且有效地缩短了训练时间，节省计算资源和成本。

表2 网络深度对比结果

Tab. 2 Comparison results of network depth

模型	卷积神经网络	视觉注意层	长短时记忆网络	全连接	总层数
NVIDIA	5	**0**	**0**	3	8
DCNN	16	**0**	1	**1**	18
DVANN	**3**	1	1	**1**	**6**

表3 训练时间和迭代次数对比结果

Tab. 3 Comparison results of training time and iteration number

模型	训练时间/h	收敛迭代次数
NVIDIA	40	100 000
DCNN	90	200 000
DVANN	**30**	**5 000**

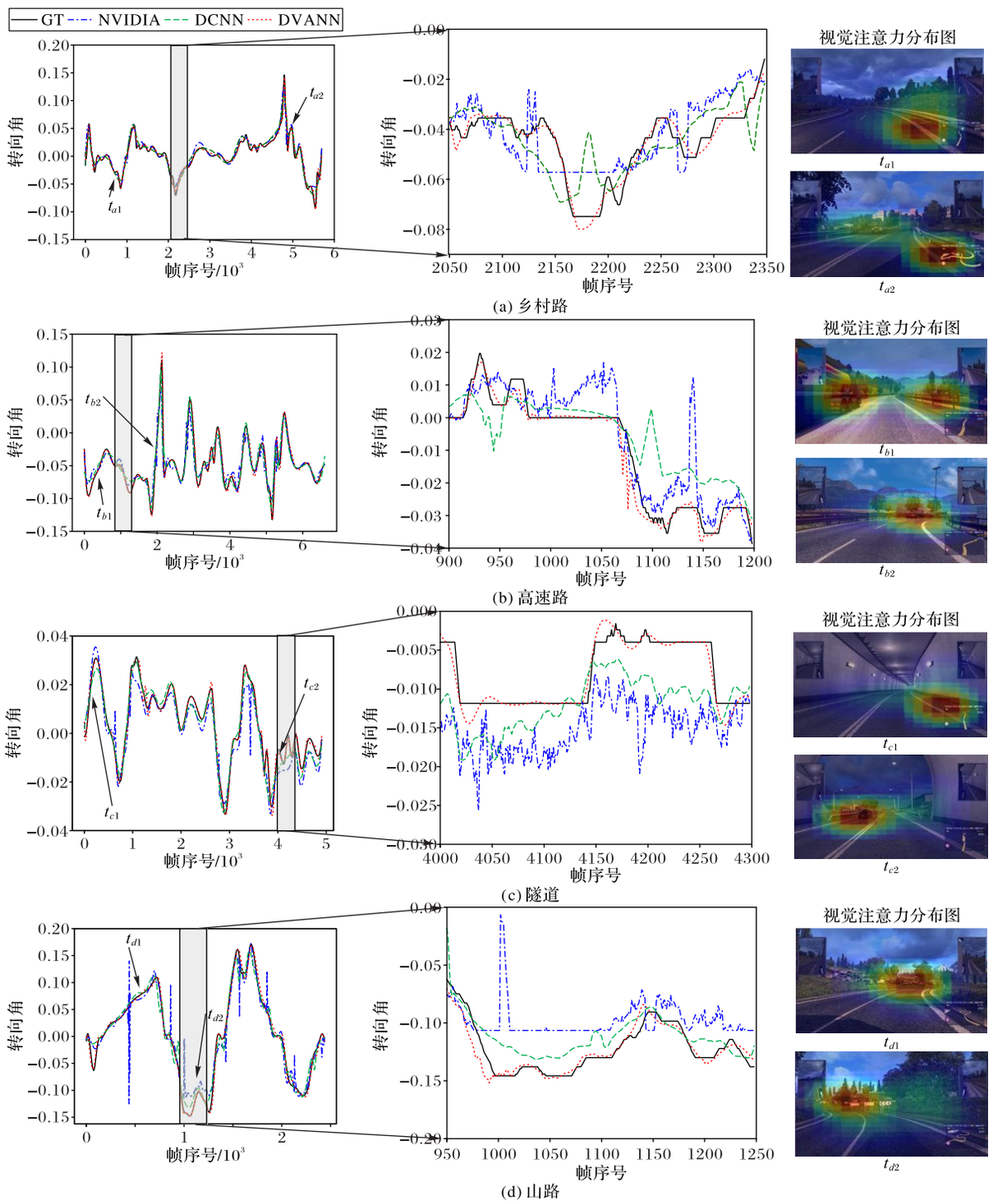

图5 测试实验结果图与代表性场景的视觉注意分布图

Fig. 5 Testing result diagrams and visual attention distribution maps of representative scenes

3　结语

本文提出了一种深度视觉注意神经网络,并基于该网络,利用前向车载相机的序列图像作为输入,实现对自动驾驶车辆方向盘转向角的预测。在设计深度视觉注意网络时,以软注意力机制为原型,将CNN提取的图像特征输入设计的视觉注意层,提取对自动驾驶重要的特征,并将经过视觉注意层加权后的特征输入LSTM提取时间关联性。注意力机制的引入,不仅能够让模型更关注和驾驶相关的特征,提高驾驶指令预测的准确度,并且能够有效降低CNN的层数,减少网络的冗余,提高模型训练速度,节省计算资源。实验结果表明,经过大量数据的训练,该网络在对转向角预测的准确性、网络总层数、训练时间和收敛迭代次数方面相比其他模型有明显的优势。然而,由于本文方法没有考虑复杂的交通规则和全局路径规划,因此无法应用于城市道路。而且由于数据集中缺乏偶然事件样本,本文模型对偶然事件的处理能力不强。未来的工作将集中在如何将交通规则和全局路径规划融入模型,让模型能够适用于更复杂的道路以及如何提高驾驶的安全性。

参考文献 (References)

[1] BROGGI A, CERRI P, DEBATTISTI S, et al. PROUD — public road urban driverless-car test[J]. IEEE Transactions on Intelligent Transportation System, 2015, 16(6):3508-3519.

[2] CHEN C, SEFF A, KORNHAUSER A, et al. DeepDriving: learning affordance for direct perception in autonomous driving[C]// Proceedings of the 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015:2722-2730.

[3] LECUN Y L, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.

[4] BOJARSKI M, DEL TESTA D, DWORAKOWSKI D, et al. End to end learning for self-driving cars[EB/OL]. [2019-02-23]. https://arxiv.org/pdf/1604.07316.pdf.

[5] 白丽赟,胡学敏,宋昇,等. 基于深度级联神经网络的自动驾驶运动规划模型[J]. 计算机应用, 2019, 39(10):2870-2875. (BAI L Y, HU X M, SONG S, et al. Motion planning model based on deep cascaded neural network for autonomous driving[J]. Journal of Computer Applications, 2019, 39(10):2870-2875.)

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

[7] XU H, GAO Y, YU F, et al. End-to-end learning of driving models from large-scale video datasets[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 3530-3538.

[8] CHI L, MU Y. Deep steering: learning end-to-end driving model from spatial and temporal visual cues[EB/OL]. [2018-08-12]. https://arxiv.org/pdf/1708.03798.pdf.

[9] SHALEV-SHWARTZ S, SHAMMAH S, SHASHUA A. Safe, multi-agent, reinforcement learning for autonomous driving[EB/OL]. [2018-10-11]. https://arxiv.org/pdf/1610.03295.pdf.

[10] EL SALLAB A, ABDOU M, PEROT E, et al. Deep reinforcement learning framework for autonomous driving[EB/OL]. [2019-01-10]. https://arxiv.org/pdf/1704.02532.pdf.

[11] ITTI L, KOCH C. Computational modelling of visual attention[J]. Nature Reviews Neuroscience, 2001, 2(3):194-203.

[12] MNIH V, HEESS N, GRAVES A, et al. Recurrent models of visual attention[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014:2204-2212.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2017:6000-6010.

[14] LIANG J W, JIANG L, CAO L, et al. Focal visual-text attention for memex question answering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8):1893-1908.

[15] XU K, BA J L, KIROS R, et al. Show, attend and tell: neural image caption generation with visual attention[EB/OL]. [2018-12-09]. https://arxiv.org/pdf/1502.03044v3.pdf.

[16] LIN Z, FENG M W, DOS SANTOS C N, et al. A structured self-attentive sentence embedding[EB/OL]. [2018-12-09]. https://arxiv.org/pdf/1703.03130.pdf.

[17] UNDERWOOD G. Visual attention and the transition from novice to advanced driver[J]. Ergonomics, 2007, 50(8):1235-1249.

[18] 胡学敏,易重辉,陈钦,等. 基于运动显著图的人群异常行为检测[J]. 计算机应用, 2018, 38(4):1164-1169. (HU X M, YI C H, CHEN Q, et al. Abnormal crowd behavior detection based on motion saliency map[J]. Journal of Computer Applications, 2018, 38(4):1164-1169.)

[19] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540):529-533.

[20] WOJNA Z, GORBAN A N, LEE D S, et al. Attention-based extraction of structured information from street view imagery[C]// Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition. Piscataway: IEEE, 2017: 844-850.

[21] 张盼盼,李其申,杨词慧. 基于轻量级分组注意力模块的图像分类算法[J]. 计算机应用, 2020, 40(3):645-650. (ZHANG P P, LI Q S, YANG C H. Image classification algorithm based on lightweight group-wise attention module[J]. Journal of Computer Applications, 2020, 40(3):645-650.)

[22] KINGMA D P, BA J L. Adam: a method for stochastic optimization[EB/OL]. [2018-12-09]. https://arxiv.org/pdf/1412.6980.pdf.

This work is partially supported by the Youth Program of National Natural Science Foundation of China (61806076), the Youth Program of the Hubei Provincial Natural Science Foundation (2018CFB158).

HU Xuemin, born in 1985, Ph. D., associate professor. His research interests include computer vision, machine learning.

TONG Xiuchi, born in 1996, M. S. candidate. Her research interests include machine learning.

GUO Lin, born in 1978, Ph. D., associate professor. Her research interests include image processing, machine learning.

ZHANG Ruohan, born in 1997, M. S. candidate. Her research interests include deep learning.

KONG Li, born in 1995, M. S. candidate. His research interests include computer vision.