控制与决策
Control and Decision

基于注意力的共享参数胶囊网络

宋燕, 覃俞璋, 曾入

引用本文:
宋燕,覃俞璋,曾入. 基于注意力的共享参数胶囊网络[J]. 控制与决策, 2023, 38(6): 1577−1585.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.1825

基于注意力的共享参数胶囊网络

宋 燕†, 覃俞璋, 曾 入

(上海理工大学 控制科学与工程系，上海 200093)

摘 要: 针对传统胶囊网络特征信息的传播冗余性和解构低效性问题,提出一种共享参数的注意力胶囊网络. 该网络的优点主要体现于以下两方面: 1) 提出注意力机制的动态路由方法,通过计算低级胶囊的相关性,使得在保留特征空间信息的同时更加关注相关性高的特征信息,并完成前向传播; 2) 在动态路由层提出共享转换矩阵,基于低级胶囊投票一致性对高级胶囊激活,并通过共享转换矩阵减少模型的参数量,同时实现改进胶囊网络的稳健性. 首先,通过5个公开数据集的分类对比实验,表明所提出胶囊网络在Fashion-MNIST、SVHN和CIFAR 10数据集上分别取得了5.17%、3.67%和9.35%的最好分类结果,而且在复杂数据集上具有显著的白盒对抗攻击鲁棒性; 然后,通过在基于smallNORB和affNISH公开数据集的仿射变换对比实验,表明所提出的胶囊网络具有显著的仿射变换鲁棒性; 最后,通过计算效率分析对比实验结果,表明所提出共享参数胶囊网络在不增加浮点运算的情况下,参数量比传统的胶囊网络减少4.9%,具有突出的计算量优势.

关键词: 图像分类；胶囊网络；注意力机制；共享参数；鲁棒性；对抗攻击

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.1825

引用格式: 宋燕, 覃俞璋, 曾入. 基于注意力的共享参数胶囊网络[J]. 控制与决策, 2023, 38(6): 1577-1585.

Attention-based capsule network with shared parameters

SONG Yan†, QIN Yu-zhang, ZENG Ru

(Department of Control Science and Engineering，University of Shanghai for Science and Technology，Shanghai 200093，China)

Abstract: Aiming to handle the problem of propagation redundancy and deconstruction inefficiency of features in traditional capsule networks, this paper proposes an attention-based capsule network with shared parameters. The merits of such a network lie mainly in the following two issues: 1) A dynamic routing method based on an attention mechanism is proposed. This method calculates the correlation between low-level capsules to maintain the space information of features and pay more attention to the feature information with a high correlation, thus fulfilling the forward propagation; 2) A shared transformation matrix is developed in the dynamic routing layer. The high-level capsules are activated based on the voting consistency of the low-level capsules. Then, the transformation matrix with shared parameters is used to reduce the parameters of the model and obtain the robustness of the capsule network. Experimental results of comparison classification on five public datasets show that the proposed capsule network achieves the best classification results of 5.17%, 3.67% and 9.35% on the Fashion-MNIST, SVHN and CIFAR 10 datasets, respectively. Moreover, it has significant robustness against the white-box anti-attack. In addition, the transformation experimental results on smallNORB and affNISH public datasets show that the proposed capsule network has obvious robustness to the transformation. Finally, the experimental results of computational efficiency show that the proposed capsule network with shared parameters reduces the parameters of traditional capsule networks by 4.9% without adding floating-point operations and has an overwhelming advantage in computation.

Keywords: image classification；capsule network；attention；shared parameters；robustness；adversarial attacks

0 引 言

卷积神经网络(convolutional neural network, CNN)因其具有良好的表征学习能力而成为图像分类任务中常用的标准模型,但与人类视觉不同的是,CNN在应对仿射变换图像时易判断错误[1],且对输入图像设计微小的扰动噪声,便能够使得训练好

收稿日期: 2021-10-25；**录用日期:** 2022-03-15.

基金项目: 国家自然科学基金项目(62073223)；上海市自然科学基金项目(22ZR1443400)；航天飞行动力学技术国防科技重点实验室开放课题项目(6142210200304).

†通讯作者. E-mail: sonya@usst.edu.cn.

*本文附带电子附录文件,可登录本刊官网该文“资源附件”区自行下载阅览.

的网络受到很大偏差而导致模型性能断崖式的下降[2-4]. 虽然CNN可以通过特定的设计改变模型结构和数据增强等方法解决上述问题,如文献[5]设置特定的群卷积使CNN获得旋转等变性,然而现实中的图像往往不是单一的线性变换,而是存在多种复杂变换的仿射变换,通过设置特定卷积层不能很好地应对仿射变换图像.

胶囊网络[6]以学习符合人类感知的图像特征为目标,通过将图像特征解构为低级胶囊和聚类为高级胶囊来实现对仿射图像的精准分类,同时在应对扰动噪声时也展现了比CNN更好的健壮性[7]. 但是其需要重复迭代计算低级胶囊对高级胶囊的预测一致性,使得在前向传播中的计算量要比CNN更多,且需要对高级胶囊的预测作出假设,使得易受到解构异常的低级胶囊影响而导致路由失败,因此往往不能应对复杂的数据集以及部署至更深层的网络架构中. 为了减少胶囊网络在前向传播的计算量,文献[8]提出了名为HitNet的神经网络,其通过由胶囊组成的“Hit-or-Miss”层将输入特征与相匹配的激活胶囊强制路由到目标类的特征中心而远离其他非激活胶囊以达到快速收敛的效果;文献[9]提出了一种名为MS-CapsNet的胶囊网络,通过多尺度特征提取低级特征和高级特征在主胶囊层编码特征金字塔来增加胶囊网络的效率,但上述改进方法在复杂数据上的性能提升效果仍然不够理想.

为了保持胶囊网络优势,提升胶囊网络内在的特征对象转换能力和分类性能是关键问题. 相较于胶囊网络路由的设计和复杂的胶囊网络而言[10],基于注意力机制的胶囊网络研究仍待进一步开展. 文献[11]提出了一个多专家系统的路由方法,每个胶囊独立地在子胶囊层路由,但是会缺失对高级语义特征信息进行关注,且忽略低级胶囊与高级胶囊之间的部分与整体关系. 文献[12]将注意力机制应用于具有无迭代的前向传播路由的胶囊网络,并设计3D的注意力路由算法,即通过将反向传播参数向量与激活胶囊乘积以选择低级胶囊,但会损失仿射变换性能. 根据上述存在的问题,本文提出一种改进的胶囊网络模型,其通过将注意力机制辅助动态路由进程以提升模型性能,且通过对胶囊网络中转换矩阵的优化使其获得更好的仿射变换鲁棒性和对抗攻击的健壮性,主要内容如下.

1)本文提出了一种融合注意力机制与共享转换矩阵的胶囊网络,称为基于注意力的共享参数胶囊网络(attention routing shared matrixes capsule network, ARWCaps),即在动态路由层分别引入注意力机制和共享转换矩阵,使得胶囊网络可根据给定任务图像区域进行更新,因为注意力机制的引入使得低级胶囊能够考虑胶囊间的相关性,保证能够学习更多与任务相关的重要特征,且由于共享转换矩阵的引入,减少了模型冗余的参数,从而能够提升模型的鲁棒性能和更新效率.

2)为了充分提取特征信息以及考虑特征之间的相关性,本文提出在动态路由层引入注意力机制. 具体地,胶囊中的注意力机制通过计算低级胶囊间的相关性可增强与重要特征相关性高的低级胶囊,使得低级胶囊关注图像特征信息的重要区域而忽略不重要的区域,为准确预测高级胶囊提供帮助. 实验结果表明,在相同条件下,注意力胶囊网络相较于原始胶囊网络,前者不仅拥有更好的分类性能还拥有更好的仿射变换图像的稳健性.

3)为了提高对仿射变换的鲁棒性以及减少模型总体的参数量,本文提出了共享转换矩阵并将其引入至动态路由层. 具体地,转换矩阵将低级胶囊转换为投票对高级胶囊预测,通过共享对每个低级胶囊对高级胶囊的转换值来减少大量模型参数,还能减少噪声扰动带来对模型的影响,且对于预测同类的高级胶囊而言,能够共享不同投票之间仿射变换参数以提高对仿射变换的鲁棒性. 实验结果表明,在相同条件下,引入共享转换矩阵的胶囊网络相较于原始胶囊网络,前者在牺牲一定的参数量下不仅拥有与后者相近的分类性能还拥有更好仿射变换图像的稳健性和对抗攻击的鲁棒性.

4)为了充分研究所提出模型的性能,本文依据其他文献提出的胶囊网络实现的实验设计了广泛的实验,在相同条件下观察不同算法的实验结果,所提出方法展示了更强大的仿射变换鲁棒性和对抗攻击的健壮性.

1 图像的鲁棒性

在图像识别领域中,图像鲁棒性是指输入图像经过仿射变换或对抗攻击后,模型仍然能够正确识别和分类该图像. 但在图像处理中,经常需要对图像进行各种操作,如平移、旋转、放缩和剪切等线性变换. 而仿射变换代表的是两幅图像之间的映射关系,通过线性变换矩阵和平移向量将输入图像变换至另一个向量空间. 设图像原像素点坐标为(x,y),经过仿射变换后像素点坐标为(u,v),其原理如下:

$$\begin{bmatrix}u\\v\end{bmatrix}=M\cdot\begin{bmatrix}x\\y\end{bmatrix}+B\Leftrightarrow\begin{cases}u=a_1x+b_1y+c_1,\\v=a_2x+b_2y+c_2.\end{cases}\tag{1}$$

其中:矩阵M为线性变换矩阵;向量B为平移量;标量$a_i, b_i \in M$;标量$c_i \in B$. 而计算机处理使用矩阵乘法进行运算,所以可通过增加一个维度,将平移与线性变换整合为一个齐次坐标矩阵,即

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = M \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}. \tag{2}$$

式(2)描述了图像在向量空间上的变换,很好地保留了图像的空间信息. 除了对输入图像进行仿射变换扰动模型性能外,网络的预测可在人类视觉系统不易察觉的范围内输入扰动进行操纵. 由于深度学习算法的输入形式是一种数值型向量或矩阵,攻击者会通过设计一种有针对性的数值型向量或矩阵从而令深度学习模型作出误判,称为对抗性攻击. 在构造对抗性数据的过程中,根据攻击者掌握深度学习模型信息的多少,可分为白盒攻击和黑盒攻击. 其中,白盒攻击能够获知深度学习所使用的算法以及算法所使用的参数;而黑盒攻击并不知道深度学习所使用的算法和参数.

设$\mathcal{M}(\cdot)$为目标模型,使$\mathcal{M}(I): I \to \mathcal{I}$,其中$I \in \boldsymbol{R}^m$为自然图像且$\mathcal{I} \in \boldsymbol{Z}^+$为模型的输出. 在常见的对抗攻击形式中,通常寻找一个信号$\rho \in \boldsymbol{R}^m$以实现$\mathcal{M}(I+\rho) \to \tilde{\mathcal{I}} \neq \mathcal{I}$. 为了保证对图像的处理是人类无法察觉的,扰动$\rho$常常是范数有界的,如通过执行$||\rho||_p < \eta$,其中$||\cdot||_p$为向量的$p$范数且$\eta$是预定义标量,即对抗攻击寻求$\rho$满足

$$\begin{aligned} &\mathcal{M}(I+\rho) \to \tilde{\mathcal{I}}; \\ &\text{s.t. } \tilde{\mathcal{I}} \neq \mathcal{I},\ ||\rho||_p < \eta. \end{aligned} \tag{3}$$

式(3)为白盒对抗攻击的一般性公式,本文所使用的对抗攻击算法均以该式为理论基础生成模型的对抗性图像样本.

2　本文模型

本文提出一种融合注意力机制与共享转换矩阵的胶囊网络,并在图像分类、仿射变换和白盒对抗攻击实验上进行应用. 总体模型的框架如图1所示. 其中h、w、c分别为输入特征图的宽度、高度和通道数,且胶囊层$\mathcal{L}$中的胶囊个数为$n_{\mathcal{L}} = h \times w \times c$.

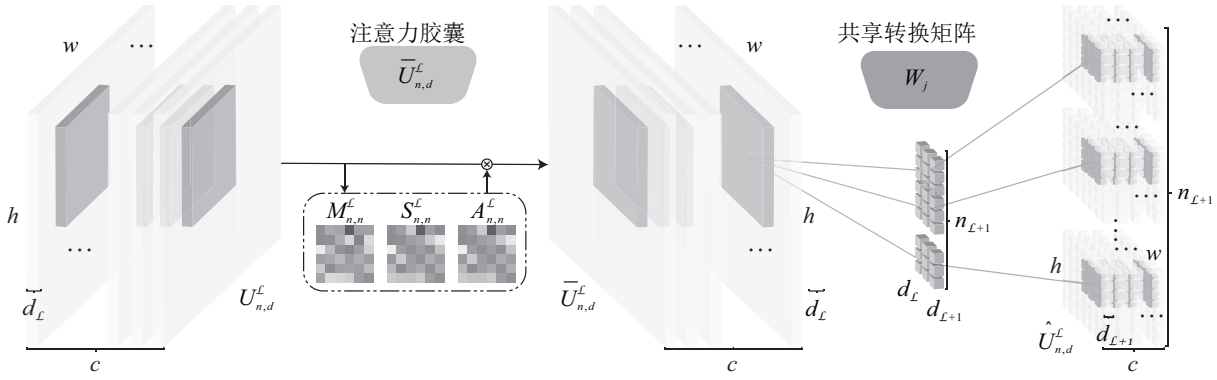

图1　注意力胶囊与共享转换矩阵过程

2.1　注意力胶囊

在动态路由中,所有低级胶囊需要对高级胶囊进行预测,选择一致性程度较高的胶囊进行激活以完成胶囊层之间的信息传播. 但是,一方面不是所有的低级胶囊均存在有效的特征信息,而是可能含有无关的背景、噪声等信息,对高级胶囊的预测可能会产生消极的影响,从而导致训练效率下降和模型性能的损失;另一方面,这些冗余的胶囊会随着输入图像复杂程度、尺寸大小等因素而变化,以CIFAR 10数据集为例,其含有许多无关的背景信息、噪声甚至是歧义性的特征,而对于胶囊网络而言,由于难以从中解构低级胶囊,导致难以获得准确的高级胶囊,从而在模型性能上达不到良好的水准.

针对上述胶囊网络存在的问题,受注意力机制启发,本文提出了注意力胶囊,旨在解决更有效地提取有效的低级胶囊且减少冗余胶囊对胶囊网络的影响. 具体地,通过计算低级胶囊间的相关性,即计算低级胶囊间的向量点积生成注意力权重矩阵,然后引入Softmax函数将注意力权重矩阵进行数值转换,使其所有元素的数值变为权重之和为1的概率分布. 由于胶囊使用向量作为其神经元,为了保证每个胶囊向量的参数压缩至0～1区间,将其在胶囊向量维度进行归一化操作,最后将低级胶囊与注意力权重矩阵进行加权求和得到注意力胶囊.

设胶囊层$\mathcal{L}$的胶囊为$U^{\mathcal{L}}_{n,d}$,其中n和d分别为该胶囊的个数和维数. 在将胶囊进行转换矩阵的线性变换前,先计算出胶囊间的相关性,如下式所示,提取胶囊间的相关性:

$$M_{n,n}^{\mathcal{L}}=\frac{U_{n,d}^{\mathcal{L}}\times U_{n,d}^{T\mathcal{L}}}{\sqrt{n_{\mathcal{L}}}}. \tag{4}$$

式(4)通过计算胶囊的点积关系得到注意力权重矩阵,然后除以胶囊个数,即帮助稳定模型的训练以及平衡输入胶囊的值. 随后,需要将权重分配给各胶囊,以达到突出含有有效特征信息胶囊和减弱含有无效信息胶囊的效果,同时还需要将各胶囊值限定于0∼1区间,以符合胶囊长度代表某类存在的概率的思想,如下式所示:

$$S_{n,n}^{\mathcal{L}}=\frac{\exp\left(M_{n,n}^{\mathcal{L}}\right)}{\sum\limits_{n}\exp\left(M_{n,n}^{\mathcal{L}}\right)}, \tag{5}$$

$$A_{n,n}^{\mathcal{L}}=\frac{s_{ij}}{\sum\limits_{k}s_{ik}}, \tag{6}$$

$$\bar{U}_{n,d}^{\mathcal{L}}=A_{n,n}^{\mathcal{L}}\times U_{n,d}^{\mathcal{L}}. \tag{7}$$

综上所述:式(4)通过计算胶囊间的相关性得到注意力权重矩阵$M_{n,n}^{\mathcal{L}}$;然后式(5)引入Softmax函数对该权重矩阵进行数值范围的压缩以及增强重要特征元素的权重得到$S_{n,n}^{\mathcal{L}}$;接着式(6)将每个胶囊进行归一化操作使得胶囊向量维度参数压缩至0∼1区间,得到注意力权重$A_{n,n}^{\mathcal{L}}$,其中$s_{ij}\in S_{n,n}^{\mathcal{L}}$;最后式(7)将得到的注意力权重与胶囊进行矩阵乘法得到注意力胶囊$\bar{U}_{n,d}^{\mathcal{L}}$. 实现以下功能:1)注意力胶囊能够更有效地提取和利用含有有效特征信息的胶囊,在模型性能上得到提升;2)注意力胶囊能够去除冗余胶囊,减弱无关信息的影响;3)注意力胶囊符合其向量长度代表某实体概率的构想.

2.2 共享转换矩阵

胶囊网络中,胶囊层$\mathcal{L}$中的每个低级胶囊u_i使用转换矩阵W_{ij}对其所属胶囊层$\mathcal{L}+1$中的高级胶囊s_j进行投票,得到对高级胶囊的预测$\hat{u}_{ji}\in\hat{U}_{n,d}^{\mathcal{L}}$,即$\hat{u}_{ij}=W_{ij}\bar{u}_i$,其中$W_{ij}$为一个可学习参数,它将学习低级胶囊与高级胶囊之间线性变换的参数以实现胶囊网络的仿射变换鲁棒性. 但由于W_{ij}参数量过大,通常为$n_{\mathcal{L}}\times d_{\mathcal{L}}\times n_{\mathcal{L}+1}\times d_{\mathcal{L}+1}$数量级,从而导致模型参数量过于庞大使得胶囊网络很难部署至深层网络. 同时为每个低级胶囊分配一个转换矩阵W_{ij}会影响高级胶囊的实例化参数,从而会减弱胶囊网络对仿射变换的鲁棒性. 具体地,低级胶囊与高级胶囊的一致性可用简单地标量积计算,即

$$a_{ij}=v_j\cdot\hat{u}_{j|i}=\mathrm{Sq}\Big(\sum_{i=1}^{n^{\mathcal{L}}}c_{ij}^{(t)}\hat{u}_{j|i}\Big)\cdot\hat{u}_{j|i}=$$

$$\mathrm{Sq}\Big(\sum_{i=1}^{n^{\mathcal{L}}}c_{ij}^{(t)}W_{ij}g(x)\Big)\cdot W_{ij}g(x). \tag{8}$$

其中:t为动态路由迭代的次数;c_{ij}为低级胶囊与高级胶囊间的耦合系数;$g(\cdot)$为胶囊算子,即将输入特征通过卷积和维度变换等线性操作转换为胶囊向量;而Sq$(\cdot)$非线性激活函数“Squash”,其作用是将输入的胶囊的向量长度压缩至0∼1区间并输出压缩后的胶囊,同时该函数不改变胶囊向量的方向,保留对某实体的实例化参数,如下式所示:

$$v_j=\frac{||s_j||^2}{1+||s_j||^2}\cdot\frac{s_j}{||s_j||}. \tag{9}$$

式(8)中对于输入的数据x,通过胶囊算子$g(\cdot)$输出低级胶囊u_i,而转换矩阵W_{ij}将低级胶囊u_i的线性变换映射至预测胶囊$\hat{u}_{j|i}$中. 在模型测试阶段,若x经过仿射变换生成的样本x^{T}输入至胶囊网络模型中,则通常希望x与x^{T}的预测值相近,但是由于x^{T}经过转换矩阵W_{ij}后得不到有效的线性变换,即两者的一致性值相差较大从而导致激活的高级胶囊不同. 换言之,x的转换矩阵W_{ij}与x^{T}的转换矩阵W'_{ij}两者映射的空间不同从而导致转换矩阵W无法作出有意义的线性变换以及激活的高级胶囊不一致而降低预测仿射变换图像的性能.

除了仿射变换影响模型的鲁棒性外,输入的扰动噪声也会影响模型的判断从而导致性能下降. 为了验证模型在对抗攻击上的鲁棒性,本文分别采用白盒攻击中的快速梯度下降法(fast gradient sign method, FSGM)[2]、基本迭代法(basic iterative method, BIM)[3]和黑盒攻击中的单像素攻击(one-pixel attack, OPA)[4]生成含有扰动噪声的图像模拟模型受噪声影响下模型的性能. 由于黑盒攻击无法通过模型参数对对抗攻击进行预测,本文仅针对白盒攻击进行理论分析. 具体地,通过对原始输入图像x添加扰动噪声生成图像x'从而使得网络对生成的图像x'进行误分类,FSGM如下式所示:

$$x'=x+\eta=x+\epsilon\cdot\mathrm{sign}(\nabla_x J(\theta,x,y)). \tag{10}$$

其中:η为输入扰动值,sign$(\cdot)$函数为数值符号的函数,$J(\cdot)$为损失函数,x和y为输入图像及其对应的真实标签,θ为网络参数. 而BIM方法攻击本质上是迭代FGSM算法,重复计算一个对抗性图像. 当存在一个微小噪声$\epsilon>||\eta||_\infty$施加于原始输入图像上生成对抗样本时,希望模型对这两个样本的分类结果应一致,但是引入转换矩阵W_{ij}后,从前向传播的角度而言,存在

$$\hat{u}_{j|i}=W_{ij}u_i=W_{ij}g(x')=W_{ij}g(x)+W_{ij}g(\eta). \tag{11}$$

根据W_{ij}的维度大小计算可知胶囊网络的激活函数会增加维度为$\epsilon n_{\mathcal{L}}d_{\mathcal{L}}n_{\mathcal{L}+1}d_{\mathcal{L}+1}$的扰动噪声,虽然$\epsilon$的值很小,但当$W_{ij}$的维度很大时,会使得扰动产生一个很大的值,从而给模型的预测受到噪声误导产生偏差,即

$$\boldsymbol{E}_{x,y}||v_j|| = \boldsymbol{E}_{x,y}\Big|\Big|Sq\Big(\sum_{i=1}^{n^{\mathcal{L}}} W_{ij}u_i - \epsilon||W_{ij}||_1\Big)\Big|\Big|. \tag{12}$$

从反向传播的角度而言,路由算法过程中的梯度传播会受到转换矩阵维度大小影响,设损失函数为$J(\cdot)$,对预测胶囊$\hat{u}_{j|i}$进行链式求导,得到

$$\frac{\partial J}{\partial \hat{u}_{j|i}} = \sum_{i=1}^{n^{\mathcal{L}+1}}\Big(\frac{\partial J}{\partial v_j^t}\frac{\partial v_j^t}{\partial s_j^t}\Big(c_{ij}^t + \hat{u}_{j|i}\frac{\partial c_{ij}^t}{\partial \hat{u}_{j|i}}\Big)\Big). \tag{13}$$

最优化耦合系数c_{ij}的软化公式将在下节进行推导,$\dfrac{\partial c_{ij}^t}{\partial \hat{u}_{j|i}}$可进行求偏导以求解梯度. 而预测胶囊受转换矩阵的影响,当输入图像存在扰动噪声时,转换矩阵维度过大影响模型判断错误的概率更高.

针对转换矩阵存在的问题,受CNN局部连接和权值共享思想的影响,本文提出共享转换矩阵参数的方法,将所有不同的低级胶囊对某个高级胶囊的投票预测中共用一个转换矩阵W_j进行投票预测的转换,维度为$(d_{\mathcal{L},n_{\mathcal{L}+1},d_{\mathcal{L}+1}})$,即对于每个高级胶囊$v_j$,所有的低级胶囊$u_i$对其的预测投票均使用同一个转换矩阵,使其能够充分学习仿射变换的参数以提高模型对仿射变换图像的鲁棒性,且由于其维度的减少使得模型在对抗攻击上更具稳健性.

2.3　动态路由

将注意力胶囊$\hat{u}_i \in \hat{U}_{n,d}^{\mathcal{L}}$通过共享转换矩阵$W_j$的线性变换输出预测向量$\hat{u}_{j|i}$送入动态路由中,设迭代次数为$t=1,2,\ldots,n$,路由过程为

$$v_j^t = Sq\Big(\sum_{i=1}^{n^{\mathcal{L}}} \mathrm{c}_{ij}^t\hat{u}_{j|i}\Big), \tag{14}$$

$$c_{ij}^{t+1} = \frac{\exp\left(b_{ij}^t + \hat{u}_{j|i}^t\cdot v_j^t\right)}{\sum\limits_{k=1}^{n^{\mathcal{L}+1}}\exp\left(b_{ik} + \hat{u}_{k|i}^t\cdot v_k^t\right)}. \tag{15}$$

其中:b_{ij}为低级胶囊u_i和高级胶囊$v_j \in V_{n,d}^{\mathcal{L}}$的耦合先验概率,其初始化为$b_{ij}^1=0$;耦合系数$c_{ij}$初始化为$c_{ij}^1 = \exp\left(b_{ij}^1\right)\Big/\exp\sum\limits_k(b_{ik}^1)$. 动态路由算法是通过卷积生成的低级胶囊$u_i$乘以转换矩阵$W_{ij}$得到预测胶囊$\hat{u}_{j|i}$送入动态路由中. 但考虑到图像中的像素点之间不是孤立存在的,图像中某一点的像素与别处的像素点一定存在某种关联. 本文通过计算任意两个位置间的交互相关性直接捕捉低级胶囊中非相邻像素点之间的关系,为胶囊空间构建一个更为丰富的结构,并将局部与全局信息相结合. 为了减少计算量和计算复杂度,通过计算低级胶囊间的相关性$M_{n,n}^{\mathcal{L}} = (U_{n,d}^{\mathcal{L}} \times U_{n,d}^{T\mathcal{L}})/\sqrt{n_{\mathcal{L}}}$,对重要的信息进行加权.

前向传播的动态路由可视为一个函数,将低级胶囊u_i映射为耦合系数,即$f: u_i \to C_{ij}^* = \{c_{ij}^{t+1} \in (0,1)\}$,其中索引$i$和$j$为低级胶囊和高级胶囊的个数. 设目标类别索引为$m$,根据上述动态路由算法,高级胶囊耦合先验概率$b_{ij}^{t+1} = b_{ij}^t + \hat{u}_{j|i}\cdot v_j^t \to \infty$,意味着目标类别的高级胶囊的向量模长$||v_m^t||$接近1,其他高级胶囊的向量模长接近0,而对应的耦合系数为$c_{im}^t \to 1, c_{ij}^t \to 0, j \neq m$,因此最优化耦合系数可写为

$$\begin{aligned} &C_{ij}^* = \\ &\max_{c_{ij}} f(u_i) = \max_{c_{ij}}\Big(\sum_{i=1}^{n^{\mathcal{L}}}\sum_{j=1}^{n^{\mathcal{L}+1}} a_{ij}\Big) = \\ &\max_{c_{ij}}\Big(\sum_{i=1}^{n^{\mathcal{L}}}\sum_{j=1}^{n^{\mathcal{L}+1}}\Big(\sum_{i=1}^{n^{\mathcal{L}}} c_{ij}^t W_{ij}A_iu_i\Big)W_{ij}A_iu_i\Big) \approx \\ &\frac{1}{n^{\mathcal{L}}}\ln\sum_{j=1}^{n^{\mathcal{L}}}\exp\Big(n^{\mathcal{L}}\times \\ &\sum_{i=1}^{n^{\mathcal{L}}}\sum_{j=1}^{n^{\mathcal{L}+1}} Sq\Big(\sum_{i=1}^{n^{\mathcal{L}}} c_{ij}^t W_{ij}A_iu_i\Big)W_{ij}A_iu_i\Big). \end{aligned} \tag{16}$$

式中:预测胶囊$\hat{u}_{j|i}$受注意力胶囊的影响,即$\hat{u}_{j|i} = W_{ij}A_iu_i$;式(16)等价于最大化激活目标类别胶囊的耦合系数c_{im};c_{ij}为最小化非目标类别胶囊的耦合系数. 其中最大值函数可软化为一个可导函数,于是最优化耦合系数可通过随机梯度下降法求取,而注意力矩阵将对初级胶囊输入的权重进行重新分配,使得能够对胶囊中重要的信息进行加权,进一步加快算法的收敛,从而防止重要的信息被噪声掩盖,也保证重要的信息在网络中更有效地传递. 同时,动态路由层中的共享转换矩阵能够给胶囊网络带来更优越的仿射变换鲁棒性,通过对预测胶囊$\hat{u}_{j|i}$进行展开分析如下:

$$\hat{u}_{j|i} = W_{ij}u_i = T_{ij}(u_i) = T_{ij}(\sigma(x)). \tag{17}$$

其中:x为输入图像,$\sigma(\cdot)$为卷积过程,$T_{ij}(\cdot)$为仿射变换函数. 设输入图像x的空间域为$\boldsymbol{U}$,将卷积过程$\sigma(\cdot)$展开为

$$\sigma(x) = (x\otimes g)(u,v) =$$

$$\sum_{j=0}^{c}\Big(\sum_{i=0}^{r}x(i,j)g(u-i,v-j)\Big)=$$

$$\mathcal{F}^{-1}(\sqrt{2\pi}\mathcal{F}[x]\mathcal{F}[g]). \tag{18}$$

其中: c和r分别为输入图像的长度和宽度, $\otimes$为卷积算子, $\mathcal{F}$为傅里叶变换, $\mathcal{F}^{-1}$为傅里叶逆变换. 输入图像x属于空间域$\boldsymbol{U}$, 而经过卷积过程$\sigma(x)$输出仍属于空间域$\boldsymbol{U}$, 即卷积过程不改变输入图像的空间信息. 但是, 仿射变换函数T_{ij}会将输入图像x的空间域$\boldsymbol{U}$变换至另一个空间域$\boldsymbol{U}'$, 即通过W_{ij}将低级胶囊u_i仿射变换至另一空间, 但由于每个低级胶囊u_i对应$i\times j$个变换矩阵, 即对于同一个高级胶囊v_j而言, 每个低级胶囊u_i需要经过不同变换矩阵预测高级胶囊v_j以生成预测胶囊$\hat{u}_{j|i}$. 当输入图像x经过仿射变换后$x'\in\hat{\boldsymbol{U}}$, 对于模型识别而言, 经过不同$W_{ij}$会使得预测同一高级胶囊的激活值发生变化, 从而导致模型的识别产生偏差. 同时, 由于$\sigma(x')\notin\boldsymbol{U}$, 因而无法对高级胶囊的计算作出有效的变换, 即

$$T_{ij}(u_i)=W_{ij}u_i\notin\hat{\boldsymbol{U}}. \tag{19}$$

而本文通过引入共享转换矩阵W_j以优化动态路由层中的仿射变换鲁棒性, 即对于每个低级胶囊u_i预测相同高级胶囊v_j时采用相同的转换矩阵, 使得其激活值尽量保持不变, 同时能够保证转换矩阵对低级胶囊作出有效的变换, 即

$$T_{ij}(u_i)=W_ju_i\in\hat{\boldsymbol{U}}. \tag{20}$$

综上所述, 所提出算法通过引入注意力胶囊重新分配低级胶囊的权重, 使得重要的信息能够被突出, 且减小噪声的干扰. 同时引入共享转换矩阵使得动态路由层中的仿射变换鲁棒性得到改善, 也使得在应对对抗攻击时, 减少对抗攻击样本的干扰, 令胶囊网络更具有图像鲁棒性.

2.4 图像重构

胶囊网络在分类胶囊层后还设置了一个解码器[6]用以表征分类胶囊层中的激活胶囊, 通过重构图像计算与原图像间的均方差鼓励分类胶囊层对参数进行有效编码. 重构损失定义如下:

$$L_R=||I-H\times W||_2^2. \tag{21}$$

其中: I为原始输入图像, H和W分别为输入图像的高和宽. 此外, 胶囊网络中还引入边缘损失函数(margin loss)用以训练胶囊网络, 其原理如下:

$$L_M=T_c\max(0,m^+-||v_c||)^2=$$

$$\lambda(1-T_c)\max(0,||v_c||-m^-). \tag{22}$$

其中: $\lambda=0.5; m^+=0.9; m^-=0.1; T_c$表示该类别是否存在, 若存在, 则$T_c=1$, 反之$T_c=0$. 同时为了不令重构损失在训练过程中占主导地位, 将重建损失按比例缩小了0.000 5, 即总损失函数为

$$L_T=L_M+0.000\,5L_R. \tag{23}$$

3 实验设置

在设计实验中, 本文将从多角度比较所提出模型的性能. 首先设置模型的数据集和参数, 然后分别评估模型图像分类性能、仿射变换图像鲁棒性、对抗攻击稳健性以及算法计算效率.

3.1 数据集介绍

本文选择以下5个公开数据集对模型的分类性能进行评估, 即MNIST、Fashion-MNIST、CIFAR 10、SVHN和smallNORB.

MNIST: MNIST是一个包含类别为0～9的灰度手写数字图像数据集, 图像的尺寸大小为28×28, 包含60 000张训练图像和10 000张测试图像.

Fashion-MNIST: Fashion-MNIST是一个由包含10个类别的70 000个时尚产品的灰度图像组成的数据集, 图像的尺寸大小为28×28.

CIFAR 10: CIFAR 10是一个包含10个类别的60 000张尺寸大小为32×32 RGB真实图像组成的数据集, 每类有5 000张训练图像和1 000张测试图像.

SVHN: SVHN是一个以数字分类为基准的数据集, 包含超过600 000张尺寸大小为32×32 RGB打印数字(0～9)的真实图像.

smallNORB: smallNORB是一个包含5个类别的3D玩具图像, 每个对象在18个角度(0～340)和9个水平高度条件下成像, 尺寸大小为96×96.

3.2 模型结构

在胶囊网络的构建中, 本文选择ResNet 18[13]的变体作为胶囊网络的特征提取层, 因为ResNet是CNN网络中性能优越的框架, 能够有效地提取图像的低级特征以帮助胶囊网络性能的提升. 基于ResNet 18模型结构, 将最后的平均池化层和全连接层分别替换为主胶囊层和分类胶囊层. 在主胶囊层中, 胶囊个数和胶囊维度分别设置为128和8; 在分类胶囊层中, 胶囊个数和胶囊维度分别设置为分类类别数和16. 而CNN基线模型同样只替换最后两层, 分别为: 1) 平均池化层和全连接层 —— AvgPool; 2) 卷积层和全连接层 —— Conv.

训练细节: 本文实验配置PC系统硬件如下: 两

张NVIDIA GeForce GTX 1 080显卡; Windows 10操作系统, Intel(R) Xeon(R) Gold 5 120 CPU @ 2.20 GHz 2.19 GHz. 编程语言为Python, 基于Pytorch深度学习框架搭建模型. 数据集划分为训练集、验证集和测试集, 其中验证集由训练集选出20 %的数据样本构成. 超参数设置如下: 训练批次(Bacth size)大小为64, 采用动量(momentum)为0.9的SGD优化器, 初始学习率设置为0.1, 学习率调整策略为每个Epoch降低0.1, 总共迭代次数为300个Epoch, 并使用最好的验证分类精度的模型进行测试

3.3　图像分类结果

本文比较了所提出模型在不同数据集上的图像分类性能, 并设置消融实验以验证模型性能, 其中DRCaps和ARCaps分别为以ResNet为基础结构的原始胶囊网络和仅使用注意力胶囊的胶囊网络. 此外, 本文给出了每种方法的错误率和内存开销, 尽管DRCaps与ARCaps有相似的参数量, 但是ARCaps的分类性能更为优越, 如在SVHN和smallNROB数据集上, ARWCaps取得了比DRCaps更好的分类性能, 即分类错误率分别为3.67 %和2.99 %, 而在MNIST和Fashion-MNIST数据集上, 虽然ARWCaps牺牲了一些参数量, 但是获得了与DRCaps相似的分类性能. 且所提出的胶囊网络模型在MNIST、Fashion-MNIST、SVHN、CIFAR 10和smllNORB数据集上也获得了较为先进的分类性能, 分类错误率分别为0.30 %、5.17 %、3.67 %、9.58 %和2.95 %, 实验结果表明了所提出改进算法的先进性.

虽然在CIFAR 10的数据集上, 胶囊网络的分类性能比不过更深更宽的CNN模型, 但是所提出模型能够有效地提升胶囊网络应对复杂自然背景数据集的分类性能, 也从侧面说明了由于复杂数据集难以提取和解构其特征, 导致胶囊网络分类性能的应用难以扩展. 而在smallNORB数据集和MNIST数据集上, 虽然未能达到或超越最佳的分类性能, 但是基本上已能够接近且提高原胶囊网络的分类性能. 综上所述, 所提出胶囊网络模型能够有效地提升其在不同数据集下的分类性能, 且还能减少部分参数量的同时, 达到甚至超越原胶囊网络的分类性能的效果.

3.4　新视角的泛化性

遵循文献[7]的实验条件, 为了验证模型对视角变换的鲁棒性, 本文设定在相同水平高度条件下, 训练模型在包含6个不同方位角(0, 20, 40, 300, 320, 340)的smallNORB样本, 记为同视角, 并在包含60～280方位角的测试样本上进行模型测试, 记为新视角. 对于水平高度的视角变换, 设定在3个较小的水平高度上进行训练, 并在6个较大的水平高度上进行测试. 在训练期间, 本文使用与训练集相同视角的测试数据进行验证, 并选择验证精度达到基线要求的模型进行新视角的测试, 得到在同视角和新视角上的测试错误率, 其中损失率是由同视角上的测试错误率减去新视角上的测试错误率得到的, 作为衡量网络模型新视角的泛化性能指标. 在实验中, 所提出胶囊网络模型的损失率均低于原始胶囊网络的损失率且优于其他胶囊网络, 表明所提出胶囊网络模型能够取得良好的视角变换的鲁棒性.

3.5　仿射变换图像的鲁棒性

为了验证所提出算法对仿射变换的鲁棒性, 本文在MNIST数据集上训练所有模型, 然后测试模型在affNIST[6]数据集上的性能. 其中训练集是将原始尺寸大小为28×28的MNIST图像用6个像素填充为40×40, 然后在40×40的尺寸大小的背景内随机放置数字. 而测试集是将原始MNIST图像填充为40×40的图像后, 对数字进行仿射变换, 即在20° 内旋转、45° 内剪切、从0.8～1.2间进行水平方向的垂直缩放以及每个方向8个像素内的随机平移, 得到affNIST数据集. 同时为了公平比较, 本文将选择在MNIST数据集上的测试精度达到99.2 %的模型送入affNIST数据集中进行测试. 所提出胶囊网络模型实现了96.73 %的affNIST分类测试精度. 实验结果表明, 本文实现了在测试affNIST数据集上较为先进的泛化性能, 若取消MNIST分类精度的限制, 则所提出模型将达到97.4 %的分类精度.

3.6　识别重叠数字图像的性能

胶囊网络能够通过分割图像识别重叠的数字[6], 本文使用MultiMNIST[6]数据集训练模型. MultiMNIST数据集是通过将一个数字叠加在另一个数字上生成的(类别不同). 具体地, 将MNIST图像首先在每个方向上移动最多4个像素, 从而生成尺寸大小为36×36的图像, 并叠加来自不同类别但为相同数据(训练集或测试集)的另一幅图像. 由于MultiMNIST包含了大量的数据, 即60 M训练集和10 M测试集, 由于使用的实验设备计算性能有限且非常耗时, 仅随机选取了数据集的千分之一以验证模型并取10次随机测试的平均值作为实验结果. 需要说明的是, 尽管本文只选取了部分数据, 但是为了检验所提出方法的显著性, 本文运用了P值统计检验

的方法刻画从随机性抽取的样本获取的结果推论至整个数据集时所犯错的概率,即P值. 若P值小于显著性水平α,此时认为随机选取千分之一数据用于模型检测可以接受,即拒绝(rejected). 反之,P值大于显著性水平α便可认为该方法用于模型检测不合理,接受(accepted). 本文实验设显著性水平$\alpha = 0.05$,由分类实验结果易见,所提出模型实现了最先进的分类性能,达到了96.20%. P值统计检验实验结果表明,本文随机选取千分之一的数据集作为模型的测试结果显著.

3.7 旋转数字图像的鲁棒性

胶囊网络的另一个优势是对图像具有旋转鲁棒性,能够将旋转后的图像识别出来. 为了验证该性能,本文将所有模型在MNIST数据上训练至相同测试MNIST分类精度,然后使测试集在$-30°\sim30°$间随机旋转,从而验证模型的旋转鲁棒性. 实验结果表明,其中ARCaps取得了最好的旋转数字分类性能97.74%,而ARWCaps取得了97.58%分类性能,相较于DRCaps有所提升,因此所提出模型能够较好地提升和保证模型的旋转鲁棒性.

3.8 对抗样本攻击的鲁棒性

为了验证所提出模型对抗性攻击的鲁棒性,本文使用有目标(target)和无目标(untarget)的白盒攻击FGSM方法生成对抗性样本,还采用了BIM方法测试模型且实验设置与FGSM方法相同. 为了公平比较,本文仅攻击每个模型预测正确的图像,且扰动量$\epsilon = 0.1$. 实验结果表明,所提出模型在应对白盒对抗攻击时,攻击成功率较CNN神经网络要低得多,且引入了共享转换矩阵后,能够大幅降低对抗攻击的成功率,同时为了观察扰动量ϵ的影响,本文将逐渐增加扰动量的值以评估模型的对抗攻击鲁棒性. 实验结果表明,随着扰动量的增大,攻击成功率逐渐上升,但是所提出模型的对抗攻击鲁棒性较CNN神经网络要好得多,尤其是引入了共享转换矩阵的胶囊网络,能够降低扰动量ϵ的影响,使得模型应对白盒攻击成功率下降.

同时,为了模拟更真实、自然的对抗攻击,本文采用黑盒攻击中OPA方法测试模型的鲁棒性. 其中,实验设置在CIFAR10数据集上,同时分别以3像素和5像素评估模型,实验生成100个攻击图像,并计算攻击成功率. 为了更好地比较算法的优越性,胶囊网络的特征提取层使用文献[6]的设置,即两层卷积尺寸为9的卷积替代,以$(*)$表示. 实验结果表明,随着攻击像素的增大,模型受攻击成功的概率也增大,而无论是无目标还是有目标对抗攻击,所提出算法均展示了较为优越的效果.

综上所述,本文针对胶囊网络提出的模型能够在胶囊网络应对对抗攻击的基础上大幅度提升模型应对对抗攻击的健壮性,减少对抗攻击的成功率以及更高的置信度概率.

3.9 算法计算效率的对比

为了更全面地评估所提出算法的特性,本文设计了算法计算效率的对比实验. 对于所有胶囊网络模型,分别输入相同尺寸大小和通道数的图像以测试模型的参数量(params)、浮点运算量(floating points of operations, Flops)、乘加运算量(multiply and add operations, Madd)以及显卡内存占用量(memory read and write, MemR+W)评估计算效率上的性能. 由实验结果可知,所提出算法相较于原始胶囊网络,在参数量的使用上具有明显优势,虽然浮点运算量等实验结果相同,但是本文重心在于胶囊网络于图像的鲁棒性,所以计算效率方面未作改进.

4 结 论

本文针对特征信息的传播冗余性和解构低效性问题,提出的基于胶囊网络的算法不仅能够有效提升网络模型的分类性能,还能改善对仿射变换图像的鲁棒性和对抗攻击的健壮性. 在动态路由中,通过使用注意力胶囊,重新为低级胶囊分配注意力权重,增强对有效特征的关注且减少对噪声信息的扰动. 同时在低级胶囊线性转换过程中,通过使用共享转换矩阵参数,在减少参数量的同时,还能有效地保证模型的仿射变换和对抗攻击的稳健性. 实验结果表明,所提出模型首先能够在分类性能得上到有效的保证和提升,然后还能改善模型对仿射变换的分类性能和噪声扰动的稳健性. 未来的工作将专注于模型更轻量化和更深层的优化,得到更具有拓展性的胶囊网络.

参考文献(References)

[1] Aharon A, Weiss Y. Why do deep convolutional networks generalize so poorly to small image transformations?[J]. Journal of Machine Learning Research, 2019, 20(184): 1-25.

[2] Goodfellow I J, Shlens J, Szegedy C. Explainingand harnessing adversarial examples[C]. International Conference on Learning Representations. California, 2015: 1-11.

[3] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world[C]. International

Conferenceon Learning Representations. Toulon, 2017: 1-14.

[4] Su J W, Vargas D V, Sakurai K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.

[5] Lenssen J E, Fey M, Libuschewski P. Group equivariant capsule networks[C]. Advances in Neural Information Processing Systems. Montreal, 2018: 8844-8853.

[6] Sabour S, Frosst N, Hinton G E. Dynamic routing betweencapsules[C]. Advances in Neural Information Processing Systems. LongBeach, 2017: 3856-3866.

[7] Hinton G E, Sabour S, Frosst N. Matrix capsules with EM routing[C]. International Conference on Learning Representations. Vancouver, 2018: 1-15.

[8] Deliège A, Cioppa A, Droogenbroeck M. An effective hit-or-miss layer favoring feature interpretation as learned prototypes deformations[C]. AAAI Conference on Artificial Intelligence. Hawaii, 2019: 1-8.

[9] Xiang C Q, Zhang L, Tang Y, et al. MS-CapsNet: A novel multi-scale capsule network[J]. IEEE Signal Processing Letters, 2018, 25(12): 1850-1854.

[10] 宋燕, 王勇. 多阶段注意力胶囊网络的图像分类[J]. 自动化学报, 2021, 47(x): 1-14.
(Song Y, Wang Y. Multi-stage attention-based capsule networks for image classification[J]. Acta Automatica Sinica, 2021, 47(x): 1-14.)

[11] Hahn T, Pyeon M, Kim G. Self-routing capsule networks[C]. Advances in Neural Information Processing Systems. Vancouver, 2019, 32: 7658−7667.

[12] Choi J, Seo H, Im S, et al. Attention routing between capsules[C]. IEEE/CVF International Conference on Computer Vision Workshop. Seoul, 2019: 1981-1989.

[13] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 770-778.

[14] Phaye S S R, Sikka A, Dhall A, et al. Dense and diverse capsule networks: Making the capsules learn better[J/OL]. 2018, arXiv: 1805.04001.

[15] Nair P, Doshi R, Keselj S. Pushing the limits of capsule networks[J/OL]. 2021, arXiv: 2103.08074.

[16] Zhang S F, Zhao W, Wu X F, et al. Fast dynamic routing based on weighted kernel density estimation[J]. Concurrency and Computation: Practice and Experience, 2021, 33(15): e5281.

[17] Killian T, Goodwin J, Brown O, et al. Kernelized capsule networks[J/OL]. 2019, arXiv: 1906.03164.

[18] Gu J D, Tresp V. Improving the robustness of capsule networks to image affine transformations[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 7283-7291.

[19] Sousa Ribeiro F, Leontidis G, Kollias S. Capsule routing via variational Bayes[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 3749-3756.

[20] Ahmed K, Torresani L. Star-caps: Capsule networks with straight-through attentive routing[C]. Advances in Neural Information Processing Systems. Vancouver, 2019, 32: 9101-9110.

[21] Rawlinson D, Ahmed A, Kowadlo G. Sparse unsupervised capsules generalize better[J/OL]. 2018, arXiv: 1804.06094.

[22] Kosiorek A R, Sabour S, Teh Y W, et al. Stacked capsule autoencoders[C]. Advances in Neural Information Processing Systems. Vancouver, 2019: 15486-15496.

[23] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks[C]. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 4700-4708.

作者简介

宋燕 (1979−), 女, 教授, 博士生导师, 从事模式识别、数据分析和预测控制等研究, E-mail: sonya@usst.edu.cn;

覃俞璋 (1998−), 男, 硕士生, 从事计算机图像处理的研究, E-mail: qinyuzzz@foxmail.com;

曾入 (1998−), 男, 硕士生, 从事计算机图像处理的研究, E-mail: zengru_neo@163.com.