多阶段注意力胶囊网络的图像分类

宋 燕 [1] 王 勇 [1]

摘 要 针对传统的胶囊网络 (Capsule network, CapsNet) 特征提取不充分的问题, 提出一种图像分类的多阶段注意力胶囊网络模型. 首先, 在卷积层对低层特征和高层特征分别采用注意力 (Spatial attention, SA) 和通道注意力 (Channel attention, CA) 来提取有效特征; 然后, 提出基于向量的注意力 (Vector attention, VA) 机制作用于动态路由层, 增加对重要胶囊的关注, 进而提高低层胶囊对高层胶囊预测的准确性; 最后, 在五个公共数据集上进行图像分类的对比实验. 结果表明, 所提出的 CapsNet 模型在分类精度和鲁棒性上优于其他胶囊网络模型, 在仿射变换图像重构方面也表现良好.

关键词 图像分类, 胶囊网络, 注意力机制, 多阶段, 鲁棒性

引用格式 宋燕, 王勇. 多阶段注意力胶囊网络的图像分类. 自动化学报, 2024, **50**(9): 1804−1817

DOI 10.16383/j.aas.c210012

Multi-stage Attention-based Capsule Networks for Image Classification

SONG Yan[1] WANG Yong[1]

Abstract Aiming to address the inadequate feature extraction problems in the traditional capsule networks (CapsNets), a multi-stage attention-based CapsNet model is proposed in this paper for image classification. Firstly, spatial attention (SA) and channel attention (CA) are used to extract effective features in the convolutional layer from low-level features and high-level features, respectively. Then, attention mechanism based on vector direction is introduced into the dynamic routing layer to enhance the focus on the important capsules, thereby improving the prediction accuracy of the low-layer capsules to the high-layer capsules. Finally, the comparison experiments on image classification are carried out on five public datasets. The experimental results show that the proposed CapsNet outperforms other CapsNets at the classification accuracy and the robustness, and its shows a good performance on the image reconstruction for affine images.

Key words Image classification, capsule network (CapsNet), attention mechanism, multi-stage, robustness

Citation Song Yan, Wang Yong. Multi-stage attention-based capsule networks for image classification. *Acta Automatica Sinica*, 2024, **50**(9): 1804−1817

图像分类是指根据图片中的信息将图片划分到某一类别, 因此对图像进行特征信息提取是图像分类的重要研究内容. 传统的图像分类主要采用机器学习方法来提取特征, 随着深度学习的不断发展, 各种深度学习算法逐渐应用到图像分类当中. 2012 年, AlexNet[1] 神经网络在图像分类效果上超越了传统方法, 在 AlexNet 之后, 涌现出一系列改进的卷积神经网络 (Convolutional neural network, CNN) 模型[2−4], 不断地提高分类精度.

然而, CNN 的模型也存在一些缺陷. 首先, CNN 的池化层会导致大量有价值的特征信息丢失, 从而对分类精度产生影响. 其次, 由于 CNN 对位置信息不敏感, 这将导致 CNN 对物体之间的空间关系的识别能力不强[5]. 随后提出的胶囊网络[6] 则能够较好地处理上述问题, 具体地, 胶囊网络摒弃了 CNN 的池化层, 保留了大量的图片信息, 这使得胶囊网络运用较少的训练数据就能达到理想的效果. 此外, 胶囊网络是部分对整体的预测, 在预测的过程中能够较好地保留特征的姿态, 如位置、大小、方向等信息, 这使得胶囊网络不仅能够进行更加精确的分类, 还能够有效地识别出经过仿射变换等一系列空间变换的图像.

近年来, 胶囊网络成为图像领域的一大研究热点. Sabour 等[6] 首先提出胶囊网络并且应用到图像分类任务, 作者基于公共数据集研究了胶囊网络的图像识别能力, 实验结果表明胶囊网络在图像分类

收稿日期 2021-01-05 录用日期 2021-05-12
Manuscript received January 5, 2021; accepted May 12, 2021
国家自然科学基金 (62073223), 上海市自然科学基金 (22ZR 1443400), 航天飞行动力学技术国防科技重点实验室开放课题 (6142210200304) 资助
Supported by National Natural Science Foundation of China (62073223), Natural Science Foundation of Shanghai (22ZR 1443400), and Open Project of Key Laboratory of Aerospace Flight Dynamics and National Defense Science and Technology (6142210200304)
本文责任编委 杨健
Recommended by Associate Editor YANG Jian
1. 上海理工大学控制科学与工程系 上海 200093
1. Department of Control Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093

的问题上可以成功地超越 CNN. 然而, 由于胶囊网络在计算和内存负载方面的代价较高, 所以该胶囊网络的结构相对较浅, 主要适用于简单数据集, 在处理复杂数据时表现不佳. 随后, Deliege 等[7] 提出一种名为 HitNet 的深度学习网络, 该网络的核心思想是使用由胶囊组成的“Hit-or-Miss”层, 假定给定类的所有图像都具有各类特有的特征, 当给定一个输入图像, 强制 HitNet 瞄准真实类的胶囊所在的特定空间的中心, 而其他类的胶囊则被发送到远离各自特征空间中心的地方. 虽然该方法的收敛速度有大幅度提升, 但是复杂数据集上的精度有所降低. 针对该问题, 文献 [8] 提出多种改进的胶囊网络, 例如堆叠更多胶囊层、增加初始胶囊的数量、增加卷积层的层数或者更换其他激活函数等. 然而, 在复杂数据集上, 改进的胶囊网络在分类精度上没有明显的提升. 文献 [9] 中将原始胶囊网络中用向量表示的胶囊替换为矩阵表示, 同时将动态路由中的聚类算法思想换成高斯混合模型 (Gaussian mixture model, GMM). 实验结果表明该模型仅在 smallNORB 数据集上有较小的提升, 复杂数据集上的效果依然不尽如人意.

在注意力胶囊网络的研究方面, 相对于注意力卷积神经网络丰富的研究成果而言, 还有待进一步深入开展. 文献 [10] 在胶囊网络的卷积层中, 针对低层特征添加空间注意力机制, 虽然有效提取了特征之间的空间位置信息, 但缺乏对高层特征所描述的重要语义信息的特别关注, 同时也没有充分考虑低层胶囊对高层胶囊的影响. 文献 [11] 通过采用注意力路由来调整训练参数的大小进而改变不同空间位置上胶囊的权重, 虽然在一定程度上增加了对重要胶囊的关注, 但没有充分考虑从低层胶囊到高层胶囊的预测过程中低层胶囊的影响.

由上述分析可见, 尽管胶囊网络是近年来模式识别领域的一大研究热点, 已经取得一些研究成果, 但目前仍处于起步阶段, 有很多尚待完善之处, 例如特征提取不充分、在复杂数据集上的分类效果较差等. 针对以上问题, 本文提出了一种改进的胶囊网络模型, 主要贡献如下:

1) 提出一种多阶段注意力胶囊网络的新模型, 该模型分别在卷积层和动态路由层中引入了注意力机制, 这使得模型的参数可以根据与给定任务相关的图像区域进行更新. 注意力机制考虑了特征之间的相关性, 保证能够学习到更多和任务相关的重要特征, 从而提升了效率.

2) 为充分提取特征信息以及特征之间的空间位置信息, 在卷积层中引入注意力机制. 具体地: 对于高层特征, 重点考虑其包含的高度抽象语义, 因此采用通道注意力 (Channel attention, CA) 机制; 对于低层特征, 重点考虑特征之间的空间位置信息, 因此采用空间注意力 (Spatial attention, SA) 机制.

3) 为提高对仿射变换图像的鲁棒性, 提出基于向量的注意力 (Vector attention, VA) 机制并且应用到胶囊网络动态路由层中的低级胶囊中, 充分考虑初始胶囊 (即低级胶囊) 之间的相关性, 从而加大对任务相关的初始胶囊的关注, 为高级胶囊的准确预测提供帮助.

4) 传统胶囊网络由于网络架构较浅, 不能充分提取有效特征, 因而在如 CIFAR10 这样的复杂数据集上效果不好. 针对该问题, 本文提出的多阶段注意力的胶囊网络具有更深的网络架构, 在复杂数据集上也能获得比较满意的结果. 大量的实验结果表明, 改进的胶囊网络模型能够在不同数据集中得到更加准确的分类结果, 明显优于几类常用的胶囊网络模型. 并且, 所提出的胶囊网络在图像重构方面也表现良好.

1 胶囊网络

最近崛起的胶囊网络代表了在神经网络方面的巨大突破. 胶囊网络主要包含三种不同类型的网络层: 卷积层、初始胶囊层和分类胶囊层[6], 如图 1 所示. 与 CNN 相比, 胶囊网络主要包含以下两大优点: 1) 摒弃了 CNN 中的池化层, 在初始胶囊层和分类胶囊层之间添加动态路由层, 以便于在低层胶囊中选择合适的低层胶囊对高层胶囊进行准确的预测. 对于每个高层胶囊来说, 胶囊网络可以增加或者减少低层胶囊和高层胶囊之间的连接强度. 因此, 胶囊网络能够保持图像内部目标之间的相关性. 2) 将 CNN 中用标量表示的特征替换为用向量表示的胶囊特征. 胶囊是一组神经元, 可以捕捉图像的各种属性, 如位置、大小、纹理等. 同时, 分类胶囊层输出的胶囊经压缩后可以较好地表示输入图像中出现对象的概率, 进而为图像分类任务的完成提供有效的帮助.

2 注意力机制

注意力机制能够帮助模型聚焦于图像中与任务相关的区域, 从而提升模型的性能. 除此之外, 注意力机制还能够学习到对象之间更深层次的关联以及不同区域之间的依赖. 目前注意力机制已经成功地应用到各个领域中, 包括机器翻译[12−14]、家庭活动识别[15]、图像字幕[16−18]、显著性检测[19]、视觉问题回答[20−21]、行为检测[22−23]、文本分类[24]、图像分类[25]、自然语言

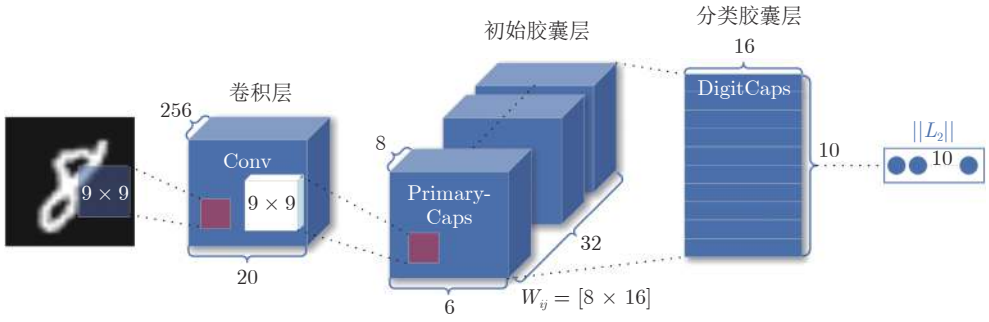


图 1 胶囊网络结构图

Fig. 1 The structure of CapsNet

处理[26] 等. 在图像分类中, 注意力机制大致分为 SA 机制和 CA 机制. SA 机制主要用于捕获特征之间的位置关系, 提出基于空间的注意力机制模型主要有 Non-local[27] 和 CBAM (Convolutional block attention module)[28]. CA 机制主要用于获得不同通道间特征之间的相关性, 提出基于通道的注意力机制模型主要有 SENet[29]. 本文除了在特征提取部分分别采用 SA 和 CA 机制外, 还在动态路由部分提出一种 VA 机制, 通过给与任务相关的胶囊分配更多的权重来加大对重要胶囊的关注.

3 本文模型

在本文中, 提出了一种多阶段注意力的胶囊网络, 并且在图像分类上进行了应用. 该网络包括三个注意力机制模块, 分别为 SA 模块[27]、CA 模块[29] 和动态路由中低级胶囊层的 VA 模块. 其中 SA 机制模块和 CA 机制模块分别加在低层特征和高层特征中, 并且将低层特征和高层特征进行融合, 既保留了低层特征的位置信息和细节信息等, 又得到了高层特征的语义信息. 这使得胶囊网络不仅能够得到有效特征, 同时特征中保留的位置信息也有助于胶囊网络对真实的类进行分类. 动态路由层中的 VA 模块则加在低层胶囊和高层胶囊之间, 动态路由中包括低层胶囊对高层胶囊的预测, 所以注意力机制可以更多地考虑低层胶囊中与分类任务相关的低层胶囊, 加大与分类任务相关的低层胶囊的权重, 进而增加低层胶囊对高层胶囊预测的准确性, 最终提高分类精度. 总体网络模型如图 2 所示.

3.1 特征提取

标准 CNN 使用的是卷积池化的组合操作, 并且一般在卷积的时候使用的是大小相同的卷积核, 由此得到特征的感受野大小是相同的. 本文在卷积的过程中使用大小不同的卷积核来提取特征, 进而增加特征的多样性.

首先, 本文对输入的图片进行四层卷积, 然后

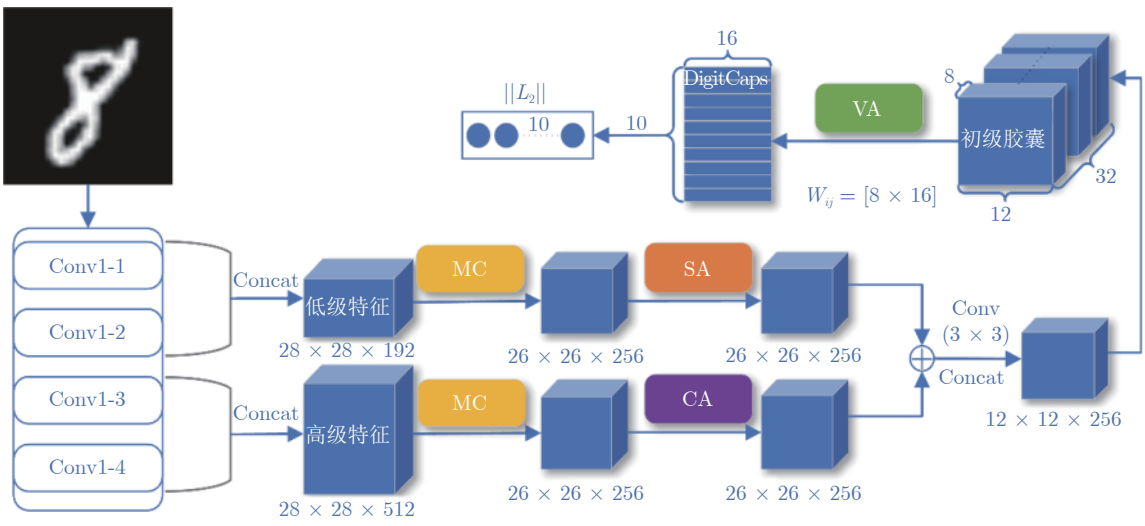


图 2 多阶段注意力的胶囊网络模型

Fig. 2 A capsule network model of multi-stage attention

将其中的前两层特征进行融合作为低层特征; 后两层特征进行融合作为高层特征. 随后分别对低层特征和高层特征使用多卷积 (Multiple convolution, MC) 操作, 即分别使用两个不同大小的卷积核对特征进行卷积, 获得不同大小的感受野, 经过测试后本文使用的是 3×3 和 5×5 大小的卷积核. 最后将得到的两个特征进行融合 (对应元素相加), 并输出融合后的特征.

3.2　注意力模块

神经网络中随着卷积层数的增加, 得到的特征的语义性也会越来越高级. 现有的方法大多是没有区分地集中多尺度特征, 这将导致信息冗余, 从而降低模型的性能. 针对该问题, 本文根据不同层次的特征的特点, 对高级特征采用 CA 机制[29], 对低层特征采用 SA 机制[27], 进而选择有效特征. 此外, 对高层特征不使用 SA 机制, 因为高层特征包含高级的抽象语义, 不需要过滤空间信息; 而对于低层特征, 不使用 CA 机制, 因为低层特征的不同通道上几乎没有语义上的区别. 同时本文在动态路由层中添加向量注意力机制, 增加和分类任务相关的低层胶囊的权重, 进而提高分类效率.

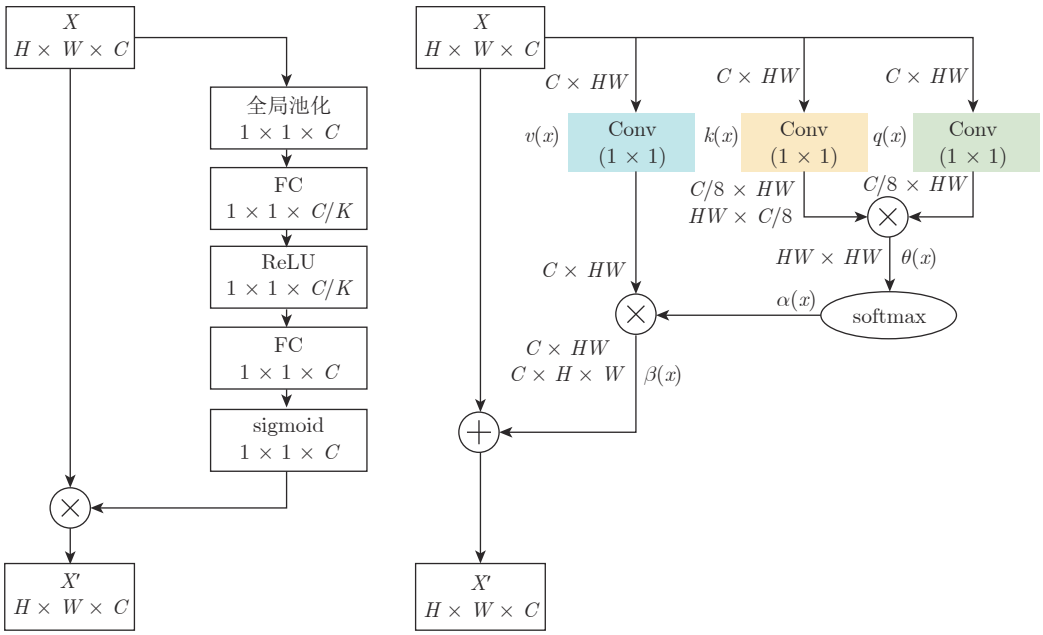
3.2.1　通道注意力机制模块

在 CNN 中, 不同通道上的特征代表着不同的语义信息. 低层特征中不同通道之间的语义性没有太大的差别, 而高层特征中不仅拥有丰富的语义信息, 不同通道之间的语义性也有较大的差异. 本文在融合后的高层特征中加入 CA 模块[29] 来给每个通道上的特征分配不同的权重, 加大与分类任务相关的通道特征的权重, 进而增加与分类任务相关的特征的关注, 提高分类效率.

具体地, 将融合后的高层特征 $f^h\in\mathbf{R}^{W\times H\times C}$ 展开为 $f^h=[f_1^h,f_2^h,\cdots,f_C^h]$, 其中, $f_i^h\in\mathbf{R}^{W\times H}$ 代表高层特征 f^h 中第 i 个通道上的特征, C 代表高层特征的通道数. 首先, 对每个通道上的特征 f_i^h 采用平均池化 (Average pooling), 进而得到基于通道特征的向量 $\boldsymbol{v}^h\in\mathbf{R}^C$. 紧接着, 将得到的向量输入两个连续的全连接层 (Full connection, FC) 来捕捉特征通道间的依赖关系 (如图 3(a) 所示), 其中 K 为降维参数, 用于降低 FC 的参数量, 两个全连接层中的 ReLU 激活函数既可以限制模型的复杂性, 又可以增加模型的非线性拟合能力. 然后, 通过式 (1) 的 sigmoid 运算将已经映射到的特征进行归一化处理, 即

$$CA=F_{\rm se}(\boldsymbol{v}^h,W)=\sigma(fc_2(\delta(fc_1(\boldsymbol{v}^h,W_1)),W_2)) \tag{1}$$

其中, $F_{\rm se}$ 表示通道注意力机制操作, W 为 CA 机制模块的参数, σ 为 sigmoid 操作, fc 代表 FC 操



(a) CA 机制
(a) CA mechamism

(b) SA 机制
(b) SA mechamism

图 3　CA 和 SA 机制

Fig. 3　CA mechanism and SA mechanism

作, δ 代表 ReLU 激活函数. 最后, 使用 CA 模块对输入 f^h 的不同通道特征进行加权得到 $\hat{f}^h$:

$$\hat{f}^h = CA \cdot f^h \tag{2}$$

3.2.2 空间注意力机制模块

CNN 中低层特征的语义性较低, 所以一般情况下图像分类模型都会选择增加网络的层数来得到更高的语义信息, 进而使用高层特征来进行分类. 虽然低层特征的语义性较低, 但是低层特征包含更多的位置和细节信息, 这些信息正是高层特征不具有的. 在低层特征中加入 SA 机制[27] 可以选择性地考虑空间位置, 通过分配不同的权重来更多地关注和分类任务相关的区域, 如边缘信息、纹理等.

SA 机制模型如图 3(b) 所示, 设融合后的低层特征为 $f^l \in \mathbf{R}^{W^1 \times H^1 \times C^1}$, 其中, H、W 和 C 分别为特征高度、宽度和通道的数量. 我们将其定义为 $x \in \mathbf{R}^{N^1 \times C^1}$, $v(x)$, $k(x)$, $q(x)$ 分别为从低层特征 f^l 中提取出的特征的特征提取器. 其中, $v(x)$ 和 f^l 具有相同的通道数 (C^1), 这里综合考虑实验精度和速度后选取通道数为 256, $k(x)$, $q(x)$ 用于计算注意力机制分布图的位置模块, $k(x_i)$ 和 $q(x_j)$ 分别为输入特征映射中的第 i 和第 j 个位置. 与 $v(x)$ 相比, $k(x_i)$ 和 $q(x_j)$ 的通道数减少到 $C^1/8$, 这使得能够过滤掉输入通道中的噪声, 进而保留与注意力机制相关的特征. 在 SA 机制模块中, 使用 1×1 大小的卷积核和 non-local 算法, 通过对图像特征的所有位置进行加权求和, 帮助模型建立位置特征之间的长距离依赖关系, 使得模型即使在浅层网络中依然能够捕获全局的感受野. 这里 non-local 算法定义为

$$\theta_{ij}(x) = k^{\mathrm{T}}(x_i) q(x_j) \tag{3}$$

其中, $k(x_i) = W_k x_i$, $q(x_j) = W_q x_j$, $W_k \in \mathbf{R}^{C^1 \times C^1}$, $W_q \in \mathbf{R}^{C^1 \times C^1}$ 为学习到的权重矩阵. 接下来, 我们对 θ_{ij} 进行如下所示的 softmax 归一化:

$$\alpha_{ij} = \frac{\exp(\theta_{ij})}{\sum\limits_{i=1}^{N} \exp(\theta_{ij})} \tag{4}$$

得到注意力机制权重分布图. 为了得到最终的注意力机制特征图, 将 α_{ij} 和 $v(x_i)$ 进行矩阵乘法, 即

$$\beta_j = \sum_{i=1}^{N} \alpha_{ij} v(x_i) \tag{5}$$

其中, $v(x_i) = W_h x_i$ 是第三个特征提取器, 其通道数为 C^1. 与 W_q 和 W_k 相似, W_h 也是一个学习过的权重矩阵. 通过这个矩阵乘法, β 中的每个位置都是图像特征中所有位置的一个加权和, 将以上所有运算归为 SA 模块, 可以得到最终的输出, 即

$$\hat{f}^l = SA \cdot f^l + f^l \tag{6}$$

3.2.3 向量注意力机制模块

胶囊网络中的动态路由是低层胶囊对高层胶囊的预测. 一方面由于胶囊网络在预测的过程中对每个低层胶囊都是等价处理的, 所以会导致低层胶囊中的一些冗余信息包括背景也以等价的形式参与训练, 致使训练效率下降; 另一方面低层胶囊对高层胶囊单独进行预测, 每个胶囊在训练过程中都忽略了其他胶囊对自身的影响. 我们在动态路由层中加入向量注意力机制, 可以对低层胶囊先进行一次筛选, 降低与分类任务无关或者关联较小的胶囊的权重, 提高与分类任务相关的胶囊权重.

如图 4 所示, 设低层特征为 $U \in \mathbf{R}^{H^2 \times W^2 \times C^2 \times L^2}$, $\boldsymbol{f} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3, \cdots, \boldsymbol{f}_{N^2}]$, 其中 $\boldsymbol{f}_s$ 表示第 s 个低层胶囊, N^2 表示低层胶囊的个数. 将其沿着向量方向进行压缩, 得到 N^2 ($N^2 = H^2 \times W^2 \times C^2$) 个 $1 \times L^2$

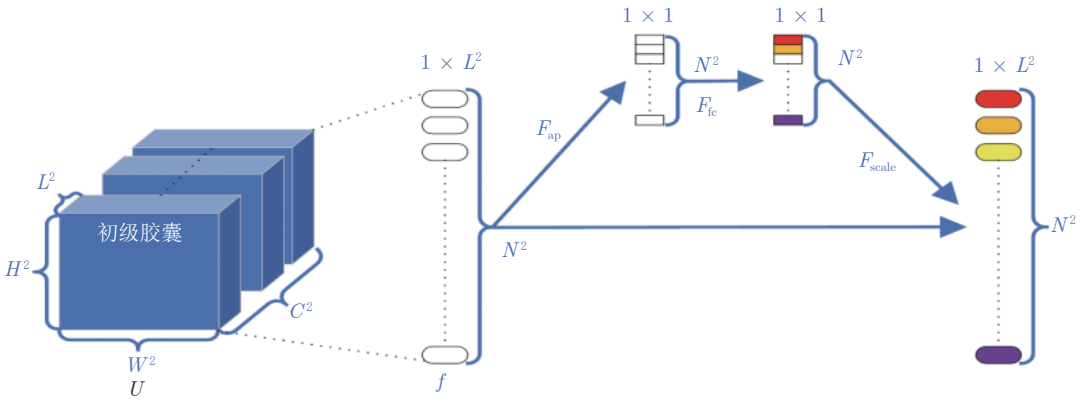


图 4 向量注意力机制

Fig. 4 Vector attention mechanism

大小的低层胶囊, 定义为

$$z_s = F_{\text{ap}}(f) = \frac{1}{L^2}\sum_{i=1}^{L^2} f_s(i) \tag{7}$$

其中, $z \in \mathbf{R}^{N^2}$ 代表压缩后的特征, z_s 表示压缩第 s 个胶囊后的标量, F_{ap} 代表胶囊压缩操作, L^2 表示胶囊的长度.

为了利用压缩操作中聚集的信息, 接下来进行第二个操作, 用于捕获低层胶囊之间的依赖关系, 即

$$o = F_{\text{fc}}(z, W^2) = \sigma({W_2}^2, \delta({W_1}^2 z)) \tag{8}$$

其中, F_{fc} 表示两层全连接层, δ 代表 ReLU 激活函数, σ 代表 sigmoid 激活函数, ${W_1}^2 \in \mathbf{R}^{\frac{C^2}{r}\times C^2}$, ${W_2}^2 \in \mathbf{R}^{C^2\times\frac{C^2}{r}}$, r 为降维参数, 用于降低两层全连接层的参数量. 首先将压缩后的胶囊特征放入两层 FC 中, 进而实现以下四种功能: 1) 两层 FC 能够捕获低层胶囊之间的线性关系; 2) ReLU 激活函数能够增加模型的非线性拟合能力; 3) 减少隐藏层的参数量, 降低模型的复杂度; 4) 对输出使用 sigmoid 激活函数将参数归一化, 方便后续处理. 最后将输出 o 和输入的低层胶囊 f 相乘, 即

$$\hat{\boldsymbol{f}}_{N^2} = F_{\text{scale}}(\boldsymbol{f}, o_{N^2}) = \boldsymbol{f}\cdot o_{N^2} \tag{9}$$

其中, F_{scale} 代表逐胶囊相乘, $\hat{\boldsymbol{f}} = [\hat{\boldsymbol{f}}_1, \hat{\boldsymbol{f}}_2, \cdots, \hat{\boldsymbol{f}}_{N^2}]$ 表示添加注意力机制后输出的初始胶囊. 将上述过程用 VA 表示, 则有

$$\hat{\boldsymbol{f}} = VA\cdot\boldsymbol{f} \tag{10}$$

3.3　动态路由层

将添加了注意力机制的初始胶囊 $\hat{\boldsymbol{f}}_i$ 送入动态路由层. 设 $\tilde{\boldsymbol{f}}_j$ 为胶囊 j 的输出向量, 向量的长度表示特定对象位于图像中给定位置的概率, 因而其取值范围应在 0 到 1 之间. 为保证这一条件成立, 运用一个压缩函数来保存对象的位置信息. 短向量可以压缩到接近 0, 长向量则可以延伸至接近 1, 压缩函数定义为

$$\tilde{\boldsymbol{f}}_j = \frac{\left\|\sum_i c_{ij}W_{ij}^3\hat{\boldsymbol{f}}_i\right\|^2}{\left(1+\left\|\sum_i c_{ij}W_{ij}^3\hat{\boldsymbol{f}}_i\right\|\right)}\frac{\sum_i c_{ij}W_{ij}^3\hat{\boldsymbol{f}}_i}{\left\|\sum_i c_{ij}W_{ij}^3\hat{\boldsymbol{f}}\right\|_i} \tag{11}$$

其中, W_{ij}^3 是低层胶囊和高层胶囊中的权重矩阵, c_{ij} 是第 i 个低层胶囊与所有第 j 个高层胶囊之间的耦合系数, 由如下定义的迭代动态路由过程确定, 即

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_j \exp(b_{ij})} \tag{12}$$

其中, b_{ij} 是第 i 个低层胶囊和第 j 个高层胶囊耦合的先验概率.

3.4　图像重构

胶囊网络还有一个典型特征是能够进行较好的图像重构, 其实现架构如图 5 所示.

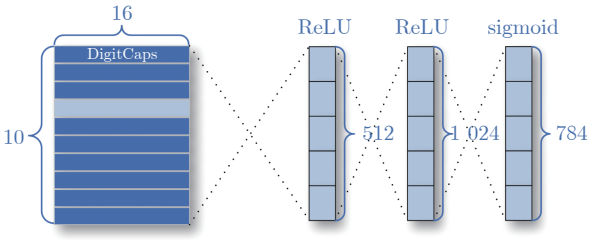


图 5　图像重构

Fig. 5　Image reconstruction

为了在训练过程中得到一幅重构的图像, 先使用 c_{ij} 中对应的耦合系数最高的向量 $\tilde{\boldsymbol{f}}_j$, 然后使用两个完全连接的 ReLU 提供正确的 $\tilde{\boldsymbol{f}}_j$. 重构的损失函数定义为

$$L_R(I, \hat{I}) = \left\|I-\hat{I}\right\|_2^2 \tag{13}$$

其中, I 是原始输入图像, $\hat{I}$ 是重构图像. $L_R(I, \hat{I})$ 需要根据选择的 $\tilde{\boldsymbol{f}}_j$ 和输入来重构图像, 这使得胶囊网络在学习的过程中会尽量选择对重构图像有用的特征, 进而降低重构损失. 将重构损失函数添加到间隔损失函数 L_M 中, 则有

$$L_M = \sum_K\left(T_K\max\left(0, m^+ - \left\|\tilde{f}_K\right\|\right)^2\right) + \sum_K\left(\lambda(1-T_K)\max\left(0, \left\|\tilde{f}_K\right\| - m^-\right)^2\right) \tag{14}$$

其中, T_K 表示对应的样本标签, 若输入图像中的对象属于类别 K, 则 $T_K = 1$, max 是最大值函数, 参数 $\lambda = 0.5$. 参照文献 [6], 令 $m^+ = 0.9$, $m^- = 0.1$, 使用总损失函数 L_T 对模型进行评估, 即

$$L_T = L_M + \varepsilon I_{\text{size}} L_R \tag{15}$$

其中, $\varepsilon = 0.000\,5$ 是每个通道像素值的正则化因子, 保证了在训练过程中重构损失 L_R 不高于 L_M, $I_{\text{size}} = H^4\times W^4\times C^4$ 是输入值的数量.

4　实验结果

4.1　实验数据

本文借助于 MNIST、Fashion-MNIST、CIFAR-10、SVHN 和 smallNORB 五个数据集来验证提出模型的有效性. MNIST 是一个包含数字 0 ~ 9

的手写体数字数据集, 大小为 28×28 像素的黑白图片, 包含 60 000 幅训练样本和 10 000 幅测试样本; Fashion-MNIST 与 MNIST 相似, 但是种类为 10 种衣物; CIFAR-10 是包含 10 类 RGB、大小为 32×32 像素图片的真实世界对象的数据集, 包括交通工具和动物, 含有 50 000 幅训练样本和 10 000 幅测试样本; SVHN 包含从谷歌街景中房屋数字号码截取的经过裁剪的 RGB 图像, 大小为 32×32 像素, 与 MNIST 一样为数字样本, 但是因为有不同的颜色和样式, 单个样本中还包含多个数字, 所以更加复杂, 其拥有 73 257 幅训练样本, 26 032 幅测试样本; smallNORB 是一个包含 5 类样本不同角度图片的数据集, 单个样本为 96×96 像素大小的灰度图片, 本文使用 24 300 幅图片作为训练集, 24 300 幅图片作为测试集.

4.2 消融实验

本文对原始的胶囊网络做了很多改进, 主要包括添加卷积层中的注意力机制模块来提取有效特征; 添加动态路由层中的向量注意力机制模块来提高分类的准确率; 采用交叉验证来说明添加注意力机制模块的有效性. 实验结果如表 1 和图 6 所示, 其中, (SA + CA) 为卷积层中的注意力机制, (VA) 为动态路由层中的向量注意力机制.

实验结果表明, 传统的胶囊网络[6] 虽然在 MNIST 上具有非常好的分类精度, 但是在复杂数据集, 如 CIFAR-10 上的分类效果较差, 而增加注意力机制后的胶囊网络不仅可以提升简单数据集的精度, 在复杂数据集上的实验效果也大大超过原始的胶囊网络. 对于 MNIST、Fashion-MNIST、CIFAR-10、SVHN 和 smallNORB 这五个数据集, 本文的模型比原始的胶囊网络分别提高了 0.16%、2.48%、11.22%、1.04% 和 0.73%.

4.3 分类对比实验

本文使用交叉验证证明了提出模型的有效性, 同时与几个常用胶囊网络, 包括 Prem Nair et al.′s CapsNet[5], HitNet[7], Matrix Capsule EM-routing[9], SACN[10], AR-CapsNet[11], DCNet[30], MS-CapsNet[31], VB-routing[32], Aff-CapsNets[33] 在五个公共数据集上进行了分类对比实验, 实验结果如表 2 所示.

由表 2 可得, 本文提出的模型在五个数据集上的分类错误率都低于其他的胶囊网络模型, 在 MNIST、Fashion-MNIST、CIFAR-10、SVHN 和 smallNORB 这五个数据集上的分类错误率分别为 0.22%, 4.63%, 9.99%, 4.08%, 4.89%, 实验结果证明了本文模型的先进性.

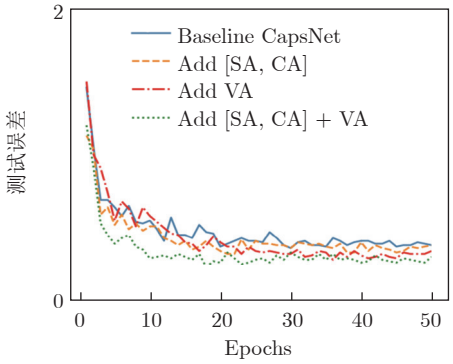
4.4 鲁棒性对比实验

为了验证模型的鲁棒性, 本文将 MNIST 数据

表 1 不同改进模块在五个数据集上的分类错误率 (%)

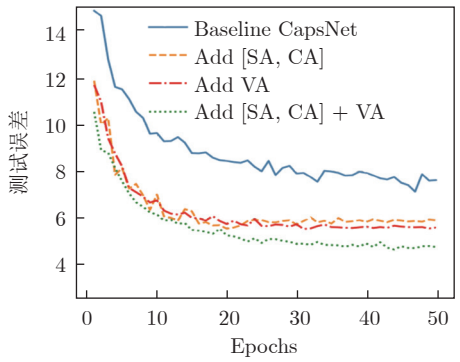
Table 1 Classification error rates of different improvement modules on five datasets (%)

模型	MNIST	Fashion-MNIST	CIFAR-10	SVHN	smallNORB
Baseline	0.38	7.11	21.21	5.12	5.62
Baseline + (SA + CA)	0.32	5.54	11.69	4.61	5.07
Baseline + VA	0.28	5.53	14.65	4.99	5.21
Baseline + (SA + CA + VA)	0.22	4.63	9.99	4.08	4.89



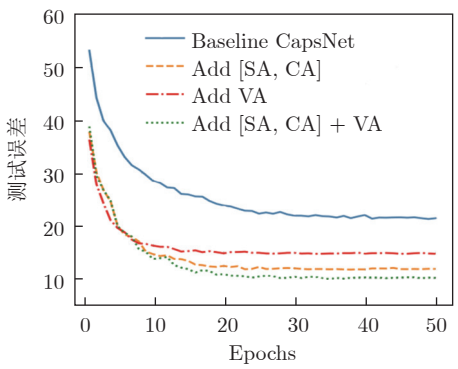
(a) MNIST 迭代曲线

(a) Iteration curves of MNIST

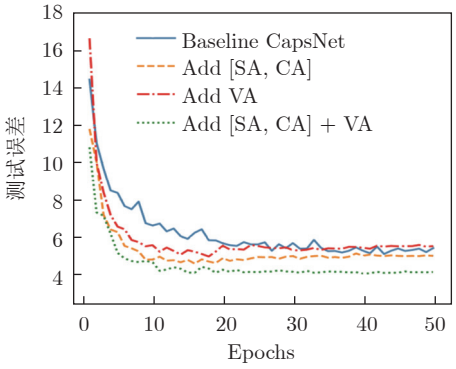


(b) Fashion-MNIST 迭代曲线

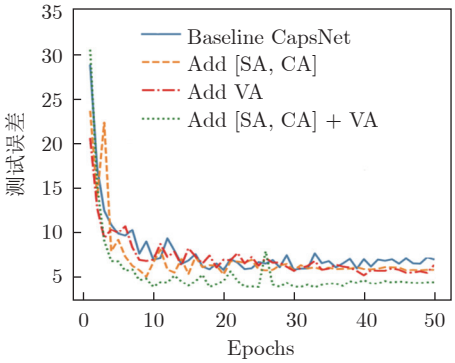
(b) Iteration curves of Fashion-MNIST



(c) CIFAR-10 迭代曲线
(c) Iteration curves of CIFAR-10



(d) SVHN 迭代曲线
(d) Iteration curves of SVHN



(e) smallNORB 迭代曲线
(e) Iteration curves of smallNORB

图 6　不同改进模块在五个数据集上的迭代曲线
Fig. 6　Iteration curves of different improvement modules on five datasets

表 2　不同模型在五个数据集上的分类错误率 (%)
Table 2　Classification error rates of different models on five datasets (%)

模型	MNIST	Fashion-MNIST	CIFAR-10	SVHN	smallNORB
Prem Nair et al.′s CapsNet[5]	0.50	10.20	31.47	8.94	—
HitNet[7]	0.32	7.70	26.70	5.50	—
Matrix Capsule EM-routing[9]	0.70	5.97	16.79	9.64	5.20
SACN[10]	0.50	5.98	16.65	5.01	7.79
AR-CapsNet[11]	0.54	—	12.71	—	—
DCNet[30]	0.25	5.36	17.37	4.42	5.57
MS-CapsNet[31]	—	6.01	18.81	—	—
VB-routing[32]	—	5.20	11.20	4.75	1.60
Aff-CapsNets[33]	0.46	7.47	23.72	7.85	—
本文模型	0.22	4.63	9.99	4.08	4.89

集的测试集在 [−25°, −15°, 0°, 15°, 25°] 之间进行随机旋转, 旋转结果如图 7 所示, 然后将训练好的模型在旋转过后的测试集上进行验证. 同时, 本文还与文献 [6] 和文献 [9] 提出的 CapsNet 和 EM-routing, 以及与本文模型具有相同层数的 CNN 进行鲁棒性对比实验, 对比结果如表 3 和图 8 所示. 由表 3 可得, CNN 在处理旋转图像时的分类精度降低了 4.78%, 文献 [6] 的胶囊网络降低了 1.73%,

(a) MNIST 数据集
(a) MNIST dataset

(b) 旋转后的 MNIST 数据集
(b) MNIST dataset after rotation

图 7 原图和仿射变换图
Fig. 7 Raw image and affine transformation image

表 3 不同模型的鲁棒性对比实验 (%)
Table 3 Robustness comparison test of different models (%)

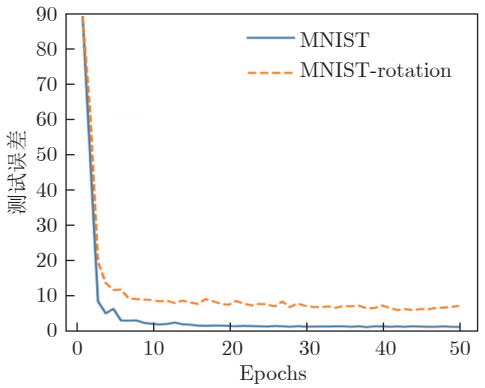
模型	MNIST	MNIST-rotation
CNN	0.74	5.52
CapsNet[6]	0.38	2.11
EM-routing[9]	0.43	2.65
本文模型	0.22	0.63

EM-routing 的降低了 2.22%, 而本文提出的模型在旋转数据集上精度只降低了 0.41%. 实验结果不仅证明了胶囊网络与 CNN 相比, 对仿射变换图像具有更强的鲁棒性, 同时验证了本文提出的胶囊网络在鲁棒性方面是传统胶囊网络的进一步提升和改善.

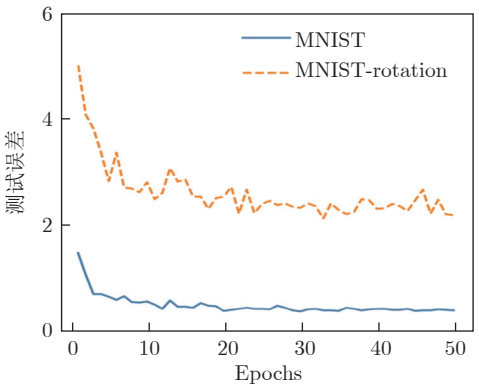
4.5 重构对比实验

模型重构的结果也是衡量模型的评判标准, 通过可视化模型产生的重构结果, 可以更加直观地对不同模型进行对比. 在图 9 ~ 13 中, 本文分别展示了 100 个真实图像、原始胶囊网络重构出的 100 个图像和本文模型重构出的 100 个图像. 通过获取 100 个图像中的部分样本进而将原始的胶囊网络与本文提出的模型进行比较, 图 9 ~ 13 中, 子图 (a) 代表 100 个真实图像的部分图像; 子图 (b) 代表原始的胶囊网络重构的 100 个图像中的部分图像; 子图 (c) 代表本文模型重构的 100 个图像中的部分图像.

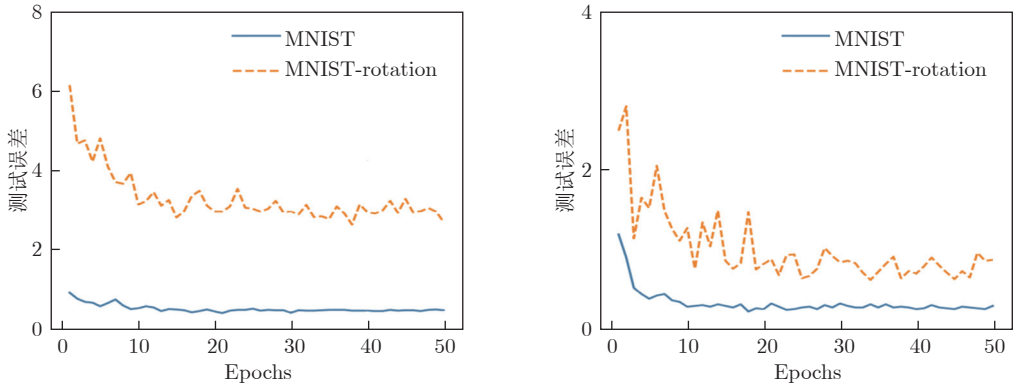
MNIST 的重构图相比于真实图片的数字边缘更宽, 类似于图像膨胀的效果, 可以将数字之间断开的部分进行连接. 由图 7 可得, 原始的胶囊网络在重构的时候容易将数字 2 重构成数字 7, 而本文的模型则能够正确地重构出与真实图片相对应的结果; Fashion-MNIST 中无论是原始的胶囊网络还是本文模型, 重构结果都与原图十分相似, 但仔细观察可以发现, 本文模型能够重构出原始图像中衣服上的褶皱, 而原始的胶囊网络则不能. 对比重构图和原图易见, 重构图像中并没有捕捉到精细的特征, 如衣服的标志和鞋子上的图案, 这可能与原始胶囊模型重构的网络太浅有关; CIFAR-10 的重构图几



(a) CNN 的鲁棒性
(a) Robustness of CNN



(b) CapsNet 的鲁棒性
(b) Robustness of CapsNet



(c) EM-routing 的鲁棒性
(c) Robustness of EM-routing

(d) 本文模型的鲁棒性
(d) Robustness of our model

图 8　不同模型的鲁棒性对比实验

Fig. 8　Comparison of robustness of different models

(a) 真实图像
(a) Real image

(b) 胶囊网络重构
(b) Capsule network reconstruction

(c) 本文模型重构
(c) Our model reconstruction

图 9　比较 MNIST 数据集中的真实图像、传统胶囊网络的重构图像以及本文模型的重构图像

Fig. 9　Comparison of the real images from the MNIST dataset, the reconstructions from a conventional capsule network, and the reconstructions from our model

(a) 真实图像
(a) Real image

(b) 胶囊网络重构
(b) Capsule network reconstruction

(c) 本文模型重构
(c) Our model reconstruction

图 10　比较 Fashion-MNIST 数据集中的真实图像、传统胶囊网络的重构图像以及本文模型的重构图像

Fig. 10　Comparison of the real images from the Fashion-MNIST dataset, the reconstructions from a conventional capsule network, and the reconstructions from our model

(a) 真实图像 (b) 胶囊网络重构 (c) 本文模型重构

(a) Real image (b) Capsule network reconstruction (c) Our model reconstruction

图 11 比较 CIFAR-10 数据集中的真实图像、传统胶囊网络的重构图像以及本文模型的重构图像

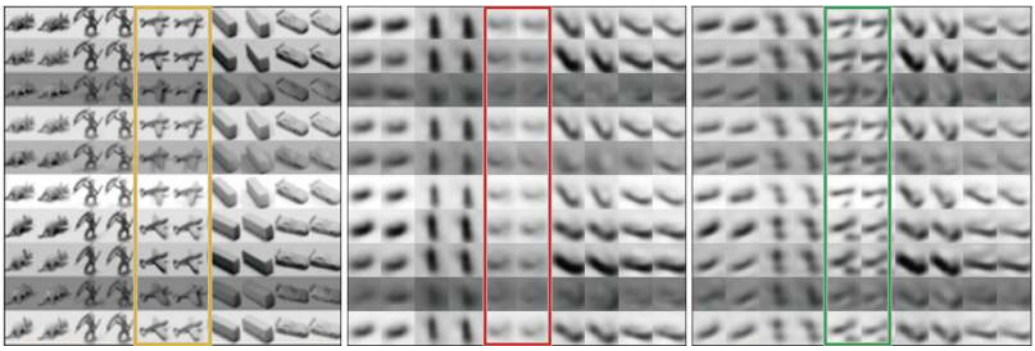
Fig. 11 Comparison of the real images from the CIFAR-10 dataset, the reconstructions from a conventional capsule network, and the reconstructions from our model

(a) 真实图像 (b) 胶囊网络重构 (c) 本文模型重构

(a) Real image (b) Capsule network reconstruction (c) Our model reconstruction

图 12 比较 SVHN 数据集中的真实图像、传统胶囊网络的重构图像以及本文模型的重构图像

Fig. 12 Comparison of the real images from the SVHN dataset, the reconstructions from a conventional capsule network, and the reconstructions from our model



(a) 真实图像 (b) 胶囊网络重构 (c) 本文模型重构

(a) Real image (b) Capsule network reconstruction (c) Our model reconstruction

图 13 比较 smallNORB 数据集中的真实图像、传统胶囊网络的重构图像以及本文模型的重构图像

Fig. 13 Comparison of the real images from the smallNORB dataset, the reconstructions from a conventional capsule network, and the reconstructions from our model

乎难以辨认, 但是仔细观察还是能够发现本文的模型在色彩的重构方面强于原始的胶囊网络; SVHN 中原始的胶囊网络将数字 0 重构成了数字 6, 而本文的模型则能够正确地重构; 在 smallNORB 数据集的重构中, 能够很明显地看到本文模型重构的图片在清晰度上远远高于初始胶囊网络的重构图. 以上实验结果充分说明了本文模型的有效性.

4.6　仿射图像重构对比实验

为了进一步验证本文提出的多阶段注意力胶囊网络针对仿射变换图像的重构性能, 我们将 MINST 数据集上的原始图片分别旋转+25°和−25°生成仿射变换图像, 如图 14 所示. 然后分别使用文献 [10] 的 CapsNet 和本文模型进行测试并输出重构图片如图 15 和图 16 所示, 同时采用均方误差 (Mean square error, MSE) 损失函数来计算模型重构图片与真实图片的差值, 实验结果如图 17 所示. 由对比重构实验结果可见, 本文提出的多层注意力胶囊网络在仿射变换图像的重构上效果更好, 具有更好的鲁棒性.

5　结束语

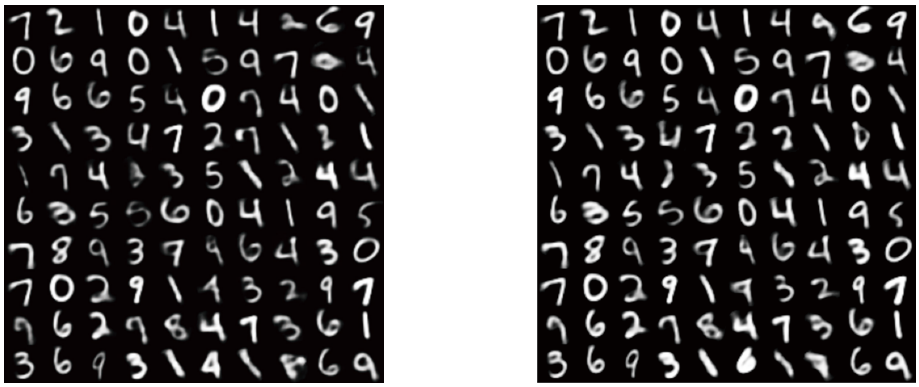
本文提出的多阶段注意力胶囊网络模型能够有效地解决原始胶囊网络特征提取不充分, 在复杂数据集上表现欠佳的问题. 在特征提取过程中, 我们通过在卷积层中对低层特征采用 SA 机制, 对高层特征采用 CA 机制来捕捉有效特征; 在计算效率方面, 我们在动态路由中添加 VA 机制来更多地考虑和分类任务相关的胶囊; 此外, 胶囊网络能够较好地学习特征间的空间相关性, 从而解决 CNN 特征间的空间关系难以捕获的问题. 通过实验可以看出, 本文的模型无论在简单数据集还是复杂数据集上都明显优于其他的胶囊网络模型. 未来的工作将专注于更加复杂的数据集以及模型中注意力机制模块的

(a) 真实图像　(b) 旋转 +25°　(b) 旋转 −25°

(a) Real image　(b) +25 degrees rotation　(c) −25 degrees rotation

图 14　MINST 数据集原图和仿射变换图

Fig. 14　Original image and affine transformations images of MINST dataset



(a) CapsNet 重构　(b) 本文模型重构

(a) Reconstruction by CapsNet　(b) Reconstruction by our model

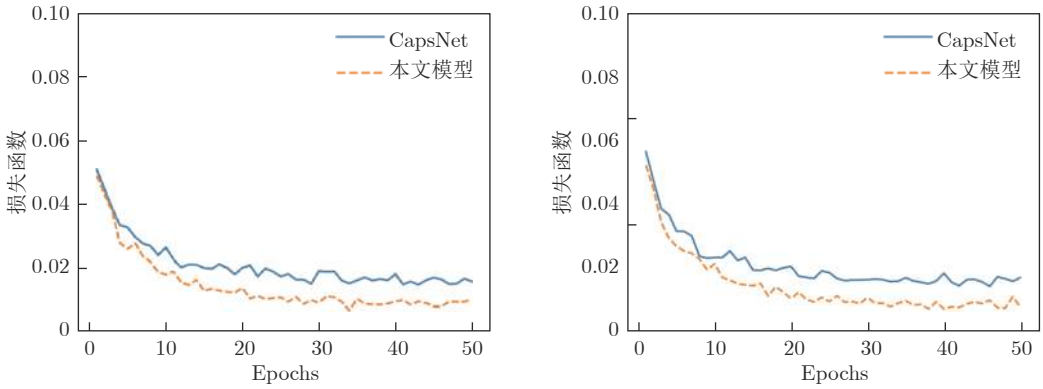
图 15　图 14(b) 的重构实验对比图

Fig. 15　Comparison of reconstructions to Fig. 14(b)

(a) CapsNet 重构
(a) Reconstruction by CapsNet

(b) 本文模型重构
(b) Reconstruction by our model

图 16 图 14(c) 的重构实验对比图
Fig. 16 Comparison of reconstructions to Fig. 14(c)



(a) 旋转 +25°
(a) +25 degrees rotation

(b) 旋转 −25°
(b) −25 degrees rotation

图 17 本文模型与文献 [10] 的 CapsNet 重构损失对比曲线
Fig. 17 Comparison of reconstruction loss curves between our model and CapsNet in [10]

优化, 同时改进图像重构的模型, 得到还原度更高的重构图像, 进而用于模型训练.

References

1 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the Conference on Neural Information Processing Systems. Lake Tahoe, USA: NIPS, 2012. 1097−1105

2 Simonyan K, Zisswerman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations. San Diego, USA: ICLR, 2015. 1−14

3 Howard A G, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv: 1704.04861, 2017.

4 Huang G, Liu Z, Van Der Maaten L, Weinberger K Q. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 2261−2269

5 Nair P, Doshi R, Keselj S. Pushing the limits of capsule networks. arXiv preprint arXiv: 2103.08074, 2021.

6 Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules. In: Proceedings of the Neural Information Processing Systems. Long Beach, USA: NIPS, 2017. 3856−3866

7 Deliege A, Cioppa A, Van Droogenbroeck M. HitNet: A neural network with capsules embedded in a hit-or-miss layer, extended with hybrid data augmentation and ghost capsules. arXiv preprint arXiv: 1806.06519, 2018.

8 Xi E, Bing S, Jin Y. Capsule network performance on complex data. arXiv preprint arXiv: 1712.03480, 2017.

9 Hinton G E, Sabour S, Frosst N. Matrix capsules with EM routing. In: Proceedings of the International Conference on Learning Representations. Vancouver, Canada: ICLR, 2018. 1−15

10 Hoogi A, Wilcox B, Gupta Y, Rubin D L. Self-attention capsule networks for object classification. arXiv preprint arXiv: 1904.12483, 2019.

11 Choi J, Seo H, Im S, Kang M. Attention routing between capsules. In: Proceedings of the IEEE International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 1981−1989

12 Wang X, Tu Z, Zhang M. Incorporating statistical machine translation word knowledge into neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Proceeding*, 2018, **26**(12): 2255−2266

13 Zhang B, Xiong D, Su J. Neural machine translation with deep attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **42**(1): 154−163

14 Zhang B, Xiong D, Xie J, Su J. Neural machine translation with gru-gated attention model. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, **31**(11): 4688−4698

15 Wang Jin-Jia, Ji Shao-Nan, Cui Lin, Xia Jing, Yang Qian. Identification of family activities based on attention capsule network. *Acta Automatica Sinica*, 2019, **45**(11): 2199−2204
(王金甲, 纪绍男, 崔琳, 夏静, 杨倩. 基于注意力胶囊网络的家庭活动识别. 自动化学报, 2019, **45**(11): 2199−2204)

16 Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, et al. Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the International Conference on Machine Learning. Lugano, Switzerland: ICML, 2015. 2048−2057

17 Gao L, Li X, Song J, Shen H T. Hierarchical lstms with adaptive attention for visual captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, **42**(5): 1112−1131

18 Lu X, Wang B, Zheng X. Sound active attention framework for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 2019, **58**(3): 1985−2000

19 Wang X, Duan H. Hierarchical visual attention model for saliency detection inspired by avian pathways. *IEEE/CAA Journal of Automatica Sinica*, 2017, **6**(2): 540−552

20 Xu H, Saenko K. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands: ECCV, 2016. 451−466

21 Liang J, Jiang L, Cao L, Kalantidis Y, Li L J, Hauptmann A G. Focal visual-text attention for memex question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, **41**(8): 1893−1908

22 Xiao Jin-Sheng, Shen Meng-Yao, Jiang Ming-Jun, Lei Jun-Feng, Bao Zhen-Yu. Abnormal behavior detection algorithm with video-bag attention mechanism in surveillance video. *Acta Automatica Sinica*, 2022, **48**(12): 2951−2959
(肖进胜, 申梦瑶, 江明俊, 雷俊峰, 包振宇. 融合包注意力机制的监控视频异常行为检测. 自动化学报, 2022, **48**(12): 2951−2959)

23 Zhao X, Chen Y, Guo J, Zhao D. A spatial-temporal attention model for human trajectory prediction. *IEEE/CAA Journal of Automatica Sinica*, 2020, **7**(4): 965−974

24 Wang Ya-Kun, Huang He-Yan, Feng Chong, Zhou Qiang. A study of conceptual sentence embedding based on attentional mechanism. *Acta Automatica Sinica*, 2020, **46**(7): 1390−1400
(王亚珅, 黄河燕, 冯冲, 周强. 基于注意力机制的概念化句嵌入研究. 自动化学报, 2020, **46**(7): 1390−1400)

25 Feng Jian-Zhou, Ma Xiang-Cong. Research on fine-grained entity classification method based on transfer learning. *Acta Automatica Sinica*, 2020, **46**(8): 1759−1766
(冯建周, 马祥聪. 基于迁移学习的细粒度实体分类方法的研究. 自动化学报, 2020, **46**(8): 1759−1766)

26 Wang Xian-Xian, Yu Long, Tian Sheng-Wei, Wang Rui-Jin. Independent RNN and CAPE networks were populated with missing elements of Uyghur events. *Acta Automatica Sinica*, 2021, **47**(4): 903−912
(王县县, 禹龙, 田生伟, 王瑞锦. 独立 RNN 和胶囊网络的维吾尔语事件缺失元素填充. 自动化学报, 2021, **47**(4): 903−912)

27 Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake, USA: IEEE, 2018. 7794−7803

28 Woo S, Park J, Lee J Y, Kweon I S. Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. Munich, Germany: ECCV, 2018. 3−19

29 Hu J, Shen L, Sun G, Wu E. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, **42**(8): 2011−2023

30 Phaye S S R, Sikka A, Dhall A, Bathula D. Dense and diverse capsule networks: Making the capsules learn better. arXiv preprint arXiv: 1805.04001, 2018.

31 Xiang C, Zhang L, Tang Y, Zou W, Xu C. MS-CapsNet: A novel multi-scale capsule network. *IEEE Signal Processing Letters*, 2018, **25**(12): 1850−1854

32 Ribeiro F D S, Leontidis G, Kollias S. Capsule routing via variational bayes. In: Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020. 3749−3756

33 Gu J, Tresp V. Improving the robustness of capsule networks to image affine transformation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 7283−7291

宋 燕 上海理工大学教授. 2001 年获得吉林大学学士学位, 2005 年获得电子科技大学硕士学位, 2013 年获得上海交通大学博士学位. 主要研究方向为模式识别, 数据分析和预测控制. 本文通信作者.
E-mail: sonya@usst.edu.cn
(**SONG Yan** Professor at University of Shanghai for Science and Technology. She received her bachelor degree from Jilin University in 2001, the master degree from University of Electronic Science and Technology of China in 2005, and the Ph.D. degree from Shanghai Jiao Tong University in 2013. Her research interest covers pattern recognition, data analysis, and predictive control. Corresponding author of this paper.)

王 勇 上海理工大学硕士研究生. 2019 年获得皖西学院学士学位. 主要研究方向为图像处理.
E-mail: 18856496454@163.com
(**WANG Yong** Master student at University of Shanghai for Science and Technology. He received his bachelor degree from Western Anhui University in 2019. His main research interest is image processing.)