基于互信息自编码和变分路由的胶囊网络结构优化

鲍静益[①] 徐 宁*[②] 尚蕴浩[②] 楚 昕[②]

[①](常州工学院 常州 213032)

[②](河海大学常州校区 常州 213022)

摘 要：胶囊网络是一类有别于卷积神经网络的新型网络模型。该文尝试提高其泛化性和精准性：首先，利用变分路由来缓解经典路由对先验信息依赖性强、易导致模型过拟合的问题。通过使用高斯混合模型(GMM)来拟合低级矩阵胶囊，并利用变分法求取近似分布，避免了参数最大似然点估计的误差，用置信度评估来获得泛化性能的提高；其次，考虑到实际数据大多无标签或者标注困难，构建互信息评价标准的胶囊自编码器，实现特征参数的有效筛选。即通过引入局部编码器，只保留胶囊中对原始输入识别最有效的特征，在减轻网络负担的同时提高了其分类识别的精准性。该文的方法在MNIST, FashionMNIST, CIFAR-10和CIFAR-100等数据集上进行了对比测试，实验结果表明：该文方法对比经典胶囊网络，其性能得到显著改善。

关键词：胶囊网络；变分路由；基于互信息评价的胶囊自编码器

中图分类号：TP181; TN911.73 **文献标识码**：A **文章编号**：1009-5896(2021)11-3309-10

DOI: 10.11999/JEIT201094

Optimization in Capsule Network Based on Mutual Information Autoencoder and Variational Routing

BAO Jingyi[①] XU Ning[②] SHANG Yunhao[②] CHU Xin[②]

[①](*Changzhou Institute of Technology*, *Changzhou* 213032, *China*)

[②](*Hohai University Changzhou Campus*, *Changzhou* 213022, *China*)

Abstract: Capsule network is a new type of network model which is different from convolutional neural network. This paper attempts to improve its generalization and accuracy. Firstly, variational routing is used to alleviate the problem of classic routing that is highly dependent on prior information and can easily lead to model overfitting. By using the Gaussian Mixture Model (GMM) to fit the low-level matrix capsule and using the variational method to fit the approximation distribution, the error of the maximum likelihood point estimation is avoided, and the confidence calculation is used to improve the generalization performance; Secondly, considering that the actual data is mostly untagged or difficult to label, a capsule autoencoder with mutual information evaluation criterion is constructed to achieve effective selection of feature parameters. That is, by introducing a local encoder, only the most effective features in the capsule for identifying and classifying the original input are retained, which reduces the computational burden of the network while improving the accuracy of classification and recognition at the same time. The method in this paper is compared and tested on datasets such as MNIST, FashionMNIST, CIFAR-10, and CIFAR-100. The experimental results show that the performance of the proposed method is significantly improved compared with the classic capsule network.

Key words: Capsule network; Variational routing; Capsule autoencoder based on mutual information

收稿日期：2020-12-30；改回日期：2021-07-01；网络出版：2021-07-08

*通信作者： 徐宁 20101832@hhu.edu.cn

基金项目：国家自然科学基金(61872199)，中央高校基本业务费(B210202083)

Foundation Items: The National Natural Science Foundation of China (61872199), The Fundamental Research Funds for the Central Universities (B210202083)

1 引言

人工智能领域(Artifical Intelligence, AI)经久不衰的一个研究话题是基于机器视觉的图像理解与分类识别。不可否认，卷积神经网络(Convolution Neural Network, CNN)在其中扮演了重要的角色，被一致公认为主流方法之一。然而，CNN亦存在本质缺陷：无法识别物体的姿势和形变。因此，文献[1]提出创新的胶囊网络(Capsule Network, CN)来代替CNN，并取得了令人鼓舞的效果：2017年的向量胶囊网络刷新了MNIST数据集的最高准确率；2018年的矩阵胶囊网络在Smallnorb数据集上达到了仅仅1.8%的错误率[2]。更有研究人员将胶囊网络的应用从图像分类扩展到文本分类、自然语言处理以及对抗网络等领域，并且在学术研究和实际应用场景中证明了胶囊网络的表现普遍优于当前场景最优的神经网络模型[3,4]。

目前来看，胶囊网络亦存在一些不足：一是经典路由使用前必须告知输入数据的类别总数，过度依赖先验知识的人工确定，不具备自主学习的能力；二是主流胶囊网络本质上均为监督学习，强烈依赖数据标定，在面对无标注数据时，缺乏提炼抽象本征特征参数的能力。

针对第1个不足，文献[5]提出利用变分路由来克服。首先，基于变分路由的胶囊网络作用于矩阵胶囊，与向量胶囊相比，有特征区别度高和计算量小的优势；其次，通过使用对数据先验干涉少的高斯混合模型(Gaussian Mixture Model, GMM)来拟合低级胶囊，满足自动确定数据类别数的要求，增强了网络的鲁棒性；最后，利用变分法拟合近似分布，避免了参数最大似然点估计，置信度计算提高了泛化性能[6]。本文在前期工作的基础上补充了更广泛的实验，用来展示变分路由的性能和优势，并对算法进行了更为详尽的阐述和分析。

针对第2个不足，本文尝试构建一个能作用无监督学习的新型胶囊网络结构，提出了一种基于互信息评价的胶囊自编码器。该模型有如下优势：(1)引入局部编码器。使网络拥有从局部特征到全局特征的认知过程；(2)保留了对象空间特征。编码器的输出为胶囊结构，这样的矢量结构能够在保留原始空间特征的同时增强网络鲁棒性；(3)实现了特征的筛选。利用基于矢量重构的互信息作为损失函数修正网络参数，使得网络具备对编码胶囊的筛选能力，只保留最特别的编码特征。

本文结构安排如下：第2节阐述了经典胶囊网络的实现原理；第3节给出变分路由的详尽推导以及基于互信息的胶囊自编码器模型；第4节对上述创新工作进行了实验验证和分析；第5节给出本文的总结。

2 经典胶囊网络

2.1 胶囊

胶囊的灵感来源于大脑视觉皮层中的微柱体，定义为一定数量的单神经元以某种形式的组合[7]。经典胶囊的组成方式有向量和矩阵两种，向量胶囊以列向量形式存在，而矩阵胶囊则包含了神经元、姿势矩阵和一个标量激活值。若胶囊网络的输入数据类别为K，那么最终会得到K个高级胶囊，每个高级胶囊中包含的不同特征值代表输入数据的不同属性，比如手写数字线条的粗细、倾斜程度和大小等。高级向量胶囊的长度表示网络将该输入判断为当前胶囊所对应类别的概率，因此模长最长的胶囊决定了网络对当前输入的预测输出。高级矩阵胶囊中姿势矩阵的不同元素对应网络提取的不同特征，激活值大小表示使用对应低级胶囊的姿势矩阵激活高级胶囊的概率。

2.2 路由

胶囊网络使用路由算法将初始胶囊分组形成高级胶囊，使网络各层之间能够更好地传递数据，经典路由算法有动态路由和期望最大化(Expectation Maximization, EM)路由两种。

EM路由作用于矩阵胶囊，使用GMM分布对初始胶囊间的特征进行拟合，然后利用 EM算法迭代计算所需特征服从的各分布函数的最佳拟合参数，其中特征所属类别以分布函数相关参数(均值、方差等)来决定 [8]。EM路由是两阶段的迭代算法，可分为E步和M步。步骤E计算初始胶囊i间特征符合高级胶囊j分布的先验概率；步骤M在掌握先验概率的基础上，将先验分布的期望最大化，计算得到胶囊j和后验概率值(将胶囊i分配给胶囊j的概率)，分别代表矩阵胶囊结构中的姿势矩阵和激活值。然后将胶囊j和概率值代入E步计算，同样迭代3次完成EM路由。最终输出高级胶囊的姿势矩阵和激活值，其中4×4姿态矩阵是由GMM的16个期望值构成的，代表将给定特征分配给当前高级特征后所有给定特征的平均值，激活值表示给定特征被当前高级胶囊激活的概率。

2.3 损失函数

网络模型参数的初始化具有随机性，因此需通过网络的反向传播，以最小化损失函数为目的，不断修正这些参数，使网络的预测输出更准确。经典胶囊网络有传播和边缘两种损失函数。若将高级胶囊j预测为当前图像的高级抽象特征，那么传播目标函数公式为

$$L_j = (\max(0, m - (a_t - a_j)))^2 \tag{1}$$

其中，a_t是标签对应的正确激活值，a_j是除标签外对应的其他错误的激活值，如果a_t和a_j的边距小于m，则通过$m-(a_t-a_j)$的平方惩罚它。一般将m初始化为0.2，在每一次迭代训练后线性增加0.1。当m达到最大值0.9后会停止增长。其中从较低的边距开始训练使网络惩罚比较宽松，能够避免在早期阶段出现太多的死胶囊。如果网络以0.9或更高的概率预测正确的类别，则函数将返回0。否则，如果置信度小于0.9，则返回0～1的数字。网络的总传播目标函数可以表示为

$$L = \sum_{j \neq t}^{K} L_j \tag{2}$$

边缘目标函数表示为

$$\begin{aligned} L_j =& T_j \max(0, m_1 - |v_j|)^2 \\ &+ \lambda(1 - T_j) \max(0, |v_j| - m_2)^2 \end{aligned} \tag{3}$$

其中，T_j表示对象j的存在与否，如果对象j存在，则$T_j = 1$，否则$T_j = 0$。$|v_j|$表示对高级胶囊j取模长，λ是调整左右部分的比重系数。m_1和m_2是对网络分别表示识别出错和未识别出来的惩罚参数。通常设$m_1 = 0.9$, $m_2 = 0.1$, $\lambda = 0.5$。也就是如果分类正确应该满足两个条件：(1)高级胶囊j的模长不应该小于0.9；(2)其他高级胶囊的模长都应该小于0.1，其中条件(1)的重要性高于条件(2)两倍。若输入数据集类别数为K，那么所有高级胶囊的总损失函数为

$$L = \sum_{j}^{K} L_j \tag{4}$$

3　改进方法

3.1　基于变分路由的胶囊网络

3.1.1　变分路由

变分路由是可以在不计算最大似然解的情况下，完成对初始胶囊特征间的聚合过程，同时还能自适应高级胶囊类别数，因此网络具有一定的抗过拟合能力。变分路由将潜在变量和未知参数都作为不可观测变量，使用$\theta = \{\theta_1 \cdots \theta_i \cdots \theta_k\}$表示，$k$表示不可观测变量数，$X = \{x_1 \cdots x_i \cdots x_m\}$表示可观测变量的集合，$m$表示可观测变量的个数。假设不可观测变量都存在各自的先验概率分布，且互相独立，根据平均场理论[9]，概率分布$q(\theta)$可以分解表示为

$$q(\theta) = \prod_{i=1}^{M} q_i(\theta_i) \tag{5}$$

其中，$q_i(\theta_i)$为θ_i的概率分布。考虑所有可观测变量和不可观测变量联合概率分布的对数，就可以得到$q_i(\theta_i)$的最优解的对数[10]

$$\ln q_i^*(\theta_i) = \mathrm{E}_{(j \neq i)}[\ln p(X, \theta_i)] + \text{const} \tag{6}$$

其中，$p(X, \theta_i)$是输入数据X和变量θ_i的真实联合概率分布，$q_i^*(\theta_i)$表示不可观测变量θ_i的近似分布。变分路由实则基于各个独立分布形成的变分分布来近似隐藏变量的条件分布，然后最优化每个独立分布来达到混合分布的最优化。

变分路由使用GMM对高级胶囊建模，然后将初始胶囊中的特征视为拟合GMM的数据点，计算拟合分布的过程就是计算高级胶囊的过程。矩阵胶囊中姿势矩阵代表的是对数据提取的抽象特征，通过对姿势矩阵进行分组、聚合来实现特征间的聚合。设初始胶囊位于网络的L层，高级胶囊位于网络的$L+1$层，$n \in \text{layer}_l, k \in \text{layer}_{l+1}$。将初始胶囊的姿势矩阵$\boldsymbol{M}_n$乘以一个$4 \times 4$的视角不变转换矩阵$\boldsymbol{W}_{nk}$，然后得到一个$4 \times 4$投票矩阵$\boldsymbol{V}_{nk}$，其中$\boldsymbol{W}_{nk}$通过网络的反向传播学习更新。那么初始胶囊$n$被分组整合到高级胶囊$k$的概率，是基于投票矩阵$\boldsymbol{V}_{nk}$与其他初始胶囊对高级胶囊$k$的投票$\{\boldsymbol{V}_{ik}, i \neq n\}$的接近程度。本文将初始胶囊的投票矩阵$\boldsymbol{V}_{nk}$作为可观测变量，由$V = \{\boldsymbol{\nu}_1 \cdots \boldsymbol{\nu}_m \cdots \boldsymbol{\nu}_M\}$表示，其中$M = N \times K$，表示投票矩阵的数量，每个矩阵$\boldsymbol{\nu}_m$具有16个神经元，对应于被提取的16个原始图像特征数据，该元素表示为ν_{md}。对于每个观察量$\boldsymbol{\nu}_m$，本文都设定一个对应的潜在变量θ_i，表示为$\theta = \{\theta_1 \cdots \theta_n \cdots \theta_N\}$，变量$\theta_n$有$k$个维度，对应数据集类别数，$\boldsymbol{\theta}_n$的数据形式是one-hot向量(只有类别$k$对应的元素为1，其余元素均为0)，元素表示为$\theta_{nk}$。

投票矩阵$\boldsymbol{\nu}_m$符合的高斯混合概率分布公式为

$$p(\boldsymbol{\nu}_m) = \sum_{k=1}^{K} \pi_k N(\boldsymbol{\nu}_n | \boldsymbol{\mu}_k, \Lambda_k) \tag{7}$$

其中，$\pi = \{\pi_k\}$表示高斯混合分布中不同分布的占比大小集合，$\mu = \{\boldsymbol{\mu}_k\}$是高斯混合分布中各分布均值的集合，$\Lambda = \{\Lambda_k\}$是各分布的协方差集合，下标$k$表示第$k$个混合分布的相关参数，$p(\nu_m)$表示$\nu_m$所属的高级胶囊分布。为了获得完整数据集联合分布$p(\boldsymbol{\nu}, \theta)$，需计算后验概率$p(\theta/v)$，本文使用变分推断的方法计算后验概率$p(\theta/v)$的近似解$q(\theta)$，根据式(6)可以得

$$\ln q^*(\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \theta_{nk} \ln \rho_{nk} + \text{const} \tag{8}$$

其中

$$\ln\rho_{nk}=\mathrm{E}\left[\ln\pi_k\right]+\frac{1}{2}\mathrm{E}\left[\ln|\Lambda|\right]-\frac{D}{2}\ln(2\pi)-\frac{1}{2}\mathrm{E}_{\mu_k,\Lambda_k}\left[(\boldsymbol{\nu}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}\Lambda_k(\boldsymbol{\nu}_n-\boldsymbol{\mu}_k)\right] \tag{9}$$

$\mathrm{E}[*]$表示求期望。对式(9)两侧取指数，并归一化可得到正比关系为

$$q^*(\theta)\propto\prod_{n=1}^{N}\prod_{k=1}^{K}r_{nk}^{\theta_{nk}} \tag{10}$$

$$r_{nk}=\frac{\rho_{nk}}{\sum\limits_{j=1}^{k}\rho_{nj}} \tag{11}$$

r_{nk}满足$r_{nk}>0,\ \sum_{k=1}^{K}r_{nk}=1$。为方便计算，为$r_{nk}$定义了3个统计信息

$$N_k=\sum_{n=1}^{N}r_{nk} \tag{12}$$

$$\tilde{\boldsymbol{\nu}}_k=\frac{1}{N_k}\sum_{n=1}^{N}r_{nk}\boldsymbol{\nu}_n \tag{13}$$

$$\boldsymbol{S}_k=\frac{1}{N_k}\sum_{n=1}^{N}r_{nk}(\boldsymbol{\nu}_n-\tilde{\boldsymbol{\nu}}_k)(\boldsymbol{\nu}_n-\tilde{\boldsymbol{\nu}}_k)^{\mathrm{T}} \tag{14}$$

若将r_{nk}视为后验概率，则N_k表示第k类数据所占的比例，$\tilde{\boldsymbol{\nu}}_k$表示属于第$k$类$\boldsymbol{\nu}$的均值向量，$\boldsymbol{S}_k$表示属于第$k$类$\boldsymbol{\nu}$的协方差。

混合系数π的先验概率$q(\pi)$的最优分布$q^*(\pi)$服从狄利克雷分布[11]，即

$$q^*(\pi)=\mathrm{Dir}(\pi|\alpha) \tag{15}$$

其中，α为狄利克雷系数

$$\alpha_k=\alpha_o+N_k \tag{16}$$

先验概率$q(\boldsymbol{\mu},\Lambda)$的最优分布$q^*(\boldsymbol{\mu}_k,\Lambda_k)$服从独立的高斯-Wishart分布[12]

$$q^*(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)=\prod_{k=1}^{K}N\left(\boldsymbol{\mu}_k|\boldsymbol{m}_k,(\beta_k\boldsymbol{\Lambda}_k)^{-1}\right)\boldsymbol{W}(\boldsymbol{\Lambda}_k|\boldsymbol{w}_k,\boldsymbol{h}_k) \tag{17}$$

这里定义了

$$\boldsymbol{m}_k=\frac{1}{\beta_k}(\beta_o\boldsymbol{m}_o+N_k\tilde{\boldsymbol{\nu}}_k) \tag{18}$$

$$\beta_k=\beta_o+N_k \tag{19}$$

$$\boldsymbol{W}_k^{-1}=\boldsymbol{W}_o^{-1}+N_k\boldsymbol{S}_k+\frac{\beta_oN_k}{\beta_o+N_k}\sum_{n=1}^{N}r_{nk}(\tilde{\boldsymbol{\nu}}_k-\boldsymbol{m}_o)(\tilde{\boldsymbol{\nu}}_k-\boldsymbol{m}_o)^{\mathrm{T}} \tag{20}$$

$$\boldsymbol{h}_k=\boldsymbol{h}_o+N_k+1 \tag{21}$$

为了使网络更简洁，实验中令$\boldsymbol{S}_k$和$\boldsymbol{W}_k^{-1}$为对角矩阵，但是在图像数据中，计算的$\boldsymbol{W}_k$不能够确保是否可逆，为防止$\boldsymbol{W}_k^{-1}$不可逆，会在$\boldsymbol{W}_k^{-1}$的对角线上加一个很小的扰动，从而变得可逆

$$\boldsymbol{W}_k=\mathrm{inv}\left(\boldsymbol{W}_k^{-1}+\epsilon I\right) \tag{22}$$

其中，ϵ为扰动强度系数，这里设$\epsilon=0.0001$，以较小的ϵ值尽量减少人为添加的扰动对图像特征计算的影响。

表1给出了变分路由算法的伪代码。变分路由是由VBE步和VBM步构成的两阶段迭代优化算法，VBE步根据当前参数计算先验分布表达式，VBM步根据求得的后验分布将期望最大化，然后更新参数。其中$r_{nk}\leftarrow r_{nk}\cdot a$表示用$a$与$r_{nk}$的点乘对$r_{nk}$进行修正，$N_k$表示每个图像中投票矩阵$\boldsymbol{V}$对类别归属的总和，$m_k$表示每个图像中姿势矩阵对原始图像的高级抽象特征值的平均值。通过伪代码可以更加清晰地展示变分路由算法的整体思想，在初始胶囊层和高级胶囊层间以迭代计算VBE步和VBM步的方式，将初始胶囊分配到对应高斯分布中。VBE步确定初始胶囊分配到高级胶囊的概率为r_{nk}，并更新先验分布的各个参数。VBM步基于r_{nk}重新计算GMM的各参数N_k, $\tilde{\nu}_k$和S_k等。迭代结束后得到

表 1 变分路由算法伪代码

输入：投票矩阵$\boldsymbol{\nu}_n$，激活值a，迭代次数T

(1) 初始化：令$\alpha_0=0.001,\boldsymbol{m}_0=\boldsymbol{0},r_{nk}=1/k,\boldsymbol{W}_0$为单位矩阵，$\beta_0,\nu_0$为常数。
(2) VBM 步:
(3) 更新 $r_{nk}\leftarrow r_{nk}\cdot a$
(4) 更新 $N_k,\tilde{\boldsymbol{\nu}}_k,\boldsymbol{S}_k$(通过式(12)—式(14))
(5) 更新 α_k (通过式(16))
(6) 更新 $\boldsymbol{m}_k,\beta_k,\boldsymbol{W}_k^{-1},\boldsymbol{\nu}_k,\boldsymbol{W}_k$(通过式(18)—式(22))
(7) T=T–1
(8) VBE 步：
(9) 更新 $\ln\rho_{nk}$(通过式(9))
(10) 其中 $\ln\tilde{\pi}_k=\varphi(\alpha_k)\varphi\left(\sum_{i=1}^{k}\alpha_k\right)$
(11) $\ln\tilde{\Lambda}=\sum_{i=1}^{D}\varphi\left(\frac{\boldsymbol{\nu}_k+1-i}{2}\right)+D\ln2+\ln|\boldsymbol{W}|$
(12) $E_{Mk,\Lambda_K}\left[(\boldsymbol{\nu}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}\Lambda(\boldsymbol{\nu}_n-\boldsymbol{\mu}_k)\right]=D\beta_k^{-1}+\boldsymbol{\nu}_k(\boldsymbol{\nu}_n-\boldsymbol{m}_k)^{\mathrm{T}}\boldsymbol{W}_k(\boldsymbol{\nu}_n-\boldsymbol{m}_k)$
(13) $\boldsymbol{M}_k=\mathrm{squeeze}(\boldsymbol{m}_k)$(squeeze为维度转换函数)
(14) $a=\mathrm{squeeze}(N_k)$
(15) 输出：$\boldsymbol{M}_k,a$

首先完成(1)～(2)中输入和初始化步骤，然后开始迭代(4)～(8)的VBM步和(10)～(13)的VBE步，直到T为0时停止更新，然后计算(13)～(14)的$\boldsymbol{M}_k$和a，并完成(15)。

的m_k和N_k分别经过维度转换函数，计算高级胶囊的姿势矩阵与标量激活值。然后使用姿势矩阵重构原始图片，使用标量激活值预测原始图像类别。

3.1.2 网络架构与实现

图1给出了基于变分路由的胶囊网络模型示意图。网络由6层组成，分别为输入层、普通卷积层、初始胶囊层、卷积胶囊层和分类胶囊层。网络的预测输出根据分类胶囊层里高级胶囊中的激活值a决定，每一个高级胶囊对应一个类别，拥有最大激活值a的高级胶囊对应类别为网络的预测输出。高级胶囊中姿势矩阵元素值由该类图像特性的平均值组成。

在普通卷积层中，设定使用32个5×5的卷积核，以2为步长，对图像进行卷积操作，然后激活特征，最后可以得到尺寸为(12×12×32)的输入。其中，可以使用ReLU作为激活函数激活特征，因为相比Sigmoid和tanh，ReLU避免了指数的运算。但ReLU的稀疏激活性会直接导致一些胶囊死亡、失去响应，所以本文使用ReLU的改进版—Leaky ReLU作为激活函数，即$y(x)=\begin{cases}x, & x\geq 0\\ 0.02x, & x<0\end{cases}$，使函数在$x<0$时的水平线变为略微倾斜的直线，保证梯度不为0，使网络可以在训练中逐渐恢复。

初始胶囊层中用17个1×1的卷积核，以1为步长，将32个输入通道转换为32个胶囊特征图，每个胶囊包含一个4×4的姿势矩阵和一个激活值，共有17维。使用S形曲线函数激活得到输出，命名为初始胶囊。网络总共输出12×12×32个初始胶囊，包含12×12×32个4×4的姿势矩阵和12×12×32×1个激活值，表示为(12, 12, 32, 17)。

卷积胶囊层中实现了胶囊版的卷积，对姿势矩阵和激活值分别以卷积核为5×5，步长为2的滑动窗口做卷积，得到4×4×32个中级胶囊，表示为(4, 4, 32, 17)。

分类胶囊层，这一层使用变分路由算法将中级胶囊以部分—整体的方式形成高级胶囊，这些高级胶囊即为网络对原始输入图像的高级特征抽象表示。网络最终输出10个高级胶囊，对应数据的10个类别，表示为(10, 17)。

输出层被设置为经典的Softmax函数。

3.2 基于互信息的胶囊自编码器

3.2.1 胶囊自编码器

胶囊自编码器结构如图2所示，包含以下部分：局部编码器$H(x)$、全局编码器$G(h)$、解码器和路由算法。首先$H(x)$通过计算得到原始输入x的局部编码胶囊h，接着h利用网络的分支结构对应两个输出，一个输出编码h，一个将h作为输入连接到$G(h)$，经过全局编码后输出全局编码胶囊g。然后将h和g进行拼接得到初始编码胶囊，再经过路由计算得到高级编码胶囊。最后将高级编码胶囊输入解码器中重构原始输入图像。

3.2.2 互信息评价准则

模型需要一个损失函数来训练网络参数。文献[13]通过最大化互信息来学习数据的高效表征[14]。

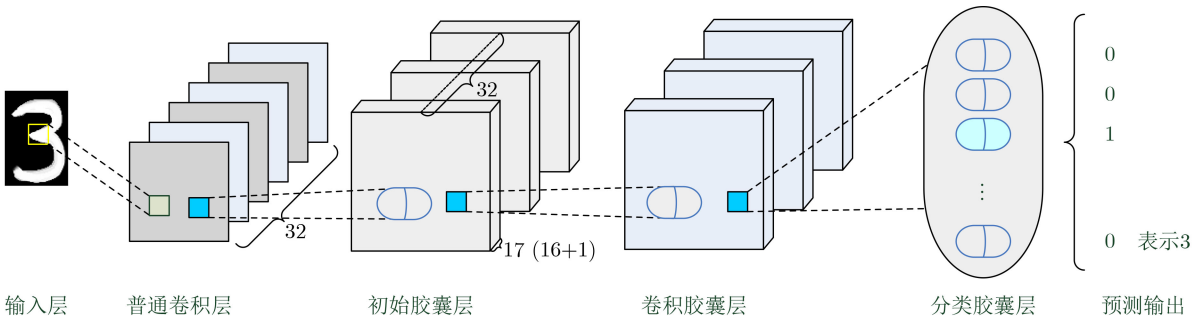


图 1 基于变分路由的胶囊网络模型示意图

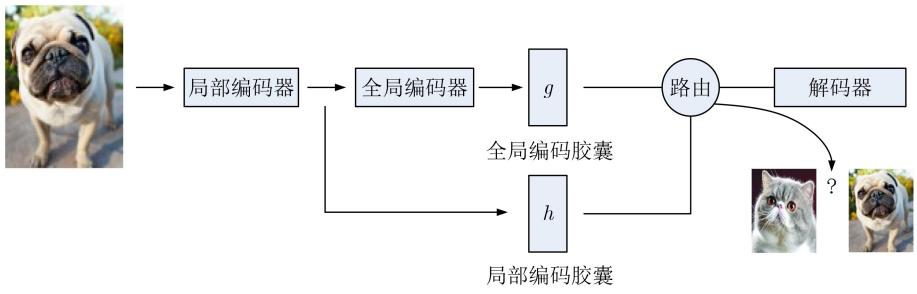


图 2 胶囊自编码器结构示意图

本文受到启发，提出使用基于互信息评价准则的损失函数来训练胶囊自编码器。

假设X表示原始输入图像x的集合、G表示全局编码胶囊g的集合、H表示局部编码胶囊h的集合。首先考虑X和G之间的全局互信息Loss_l的计算。按照基本定义，两者的互信息可表示为

$$I(X,G)=\sum_{x\epsilon X}\sum_{g\epsilon G}p(x,g)\lg\left(\frac{p(x,g)}{p(x)p(g)}\right) \tag{23}$$

其中，$p(x,g)$是联合概率函数，$p(x)$和$p(g)$分别是x和g的边缘概率密度函数。在训练网络时一般以最大化互信息为目标[15]

$$\mathrm{Loss}_g=\max_{g\epsilon G}I(X,G) \tag{24}$$

另一方面，假定g服从先验高斯分布$q(g)$(选择高斯分布可以使编码胶囊的规整度更高且人为干预少)，那么，人为假设的分布$q(g)$与真实分布$p(g)$之间就会存在偏差，而这种偏差应越小越好，KL散度适合用来量化这种偏差[16]，其优化目标为

$$\begin{aligned}\mathrm{Loss}_p&=\min_{g\epsilon G}\left(\mathrm{KL}\left(p(g)\,||\,q(g)\right)\right)\\&=\min_{g\epsilon G}\left(\sum_{g\epsilon G}p(g)\lg\frac{p(g)}{q(g)}\right)\end{aligned} \tag{25}$$

最终优化目标为式(24)和式(25)的加权和，即全局互信息Loss_l经整理后可表示为

$$\begin{aligned}\mathrm{Loss}_l&=m\cdot\mathrm{Loss}_g+n\cdot\mathrm{Loss}_p\\&=\min\left\{\begin{array}{c}-(m+n)\,\mathrm{KL}\left(p(g,x)\,||p(g)p(x)\right)\\+n\cdot\mathrm{E}_{p(x)}\left[\mathrm{KL}\left(p(g/x)\,||q(g)\right)\right]\end{array}\right\}\end{aligned} \tag{26}$$

其中，$\mathrm{E}_{p(x)}(\cdot)$表示关于$p(x)$的数学期望，m和n分别为加权系数。互信息项需要通过最大化KL距离完成最大化互信息，但是KL函数值域为$[0,+\infty)$，无法实现最大化。另外，KL函数关于自变量不对称，在训练中会因为输入数据顺序的不同而得到不同的结果。因此可以选取JS散度来表示互信息的最大化[17]，JS散度定义为

$$\begin{aligned}\mathrm{JS}(P_1,P_2)=&\frac{1}{2}\mathrm{KL}\left(P_1||\frac{P_1+P_2}{2}\right)\\&+\frac{1}{2}\mathrm{KL}\left(p_2||\frac{p_1+p_2}{2}\right)\end{aligned} \tag{27}$$

JS散度的优势在于它有上界$\lg2/2$，且基于P_1,P_2对称。加入了JS散度后的损失函数为

$$\mathrm{Loss}_l=\min\left\{\begin{array}{c}-(m+n)\cdot\mathrm{JS}(P_1,P_2)\\+n\cdot\mathrm{E}_{p(x)}\left[\mathrm{KL}\left(p(g/x)\,||q(g)\right)\right]\end{array}\right\} \tag{28}$$

使用负采样的方法对JS散度进行计算[18]，得到

$$\mathrm{Loss}_l=\min\left\{\begin{array}{c}-(m+n)\,\mathrm{E}_{p(x)}\left[\lg S(\psi(x,g))\right]\\+\mathrm{E}_{p(x)}\left[\lg\left(1-S(\psi(x,g))\right)\right]\\+n\cdot\mathrm{E}_{p(x)}\left[\mathrm{KL}\left(p(g/x)\,||q(g)\right)\right]\end{array}\right\} \tag{29}$$

其中，$S(\cdot)$表示判别网络$\psi(x,g)$的激活函数。

同理可得h与g的局部互信息Loss_h为

$$\mathrm{Loss}_h=\min\left\{\begin{array}{c}-(m+n)\,\mathrm{E}_{p(h)}\left[\lg S(\psi(h,g))\right]\\+\mathrm{E}_{p(h)}\left[\lg\left(1-S(\psi(h,g))\right)\right]\end{array}\right\} \tag{30}$$

综上所述，总损失函数可以表示为Loss_l与Loss_h之和

$$\mathrm{Loss}=\min\left\{\begin{array}{c}\mathrm{E}_{p(x)}\left[\lg\left(1-S(\psi(x,g))\right)\right]\\-(m+n)\,\mathrm{E}_{p(x)}\left[\lg S(\psi(x,g))\right]\\+\mathrm{E}_{p(h)}\left[\lg\left(1-S(\psi(h,g))\right)\right]\\-(m+n)\,\mathrm{E}_{p(h)}\left[\lg S(\psi(h,g))\right]\\+n\cdot\mathrm{E}_{p(x)}\left[\mathrm{KL}\left(p(g/x)\,||\,q(g)\right)\right]\end{array}\right\} \tag{31}$$

3.2.3 网络架构与实现

为了能够直观衡量网络特征提取质量，本文在模型中设计了分类计算模块，如图3所示。网络获取局部和全局编码胶囊后，将其结合成初始编码胶囊，接着利用动态路由算法对初始编码胶囊进行特征聚合得到高级编码胶囊，并使用高级编码胶囊的长度代表对应类别的概率，此时需要将高级编码胶囊经过Squash函数得到归一化概率值，并视最大概率值所在位置的类别为预测输出，最后根据数据标签计算网络预测准确率。在进入解码器之前，我们对高级编码胶囊进行掩码操作(将非预测类别对应特征置为0，只保留预测类别对应的特征)，然后解码器使用全连接网络重构输入图像，以判断网络学习到的高级编码胶囊是否提取到足够的能重构原始输入的特征。

表2给出了胶囊自编码器中的动态路由算法的伪代码。其中b为网络初始编码胶囊的初始化偏置，c为初始编码胶囊投票是否分配给高级编码胶囊的权重系数，c通过对b进行softmax得到。h和g分别为局部和全局编码胶囊、$H(x)$和$G(h)$分别为对应的局部和全局编码器。$\boldsymbol{u}$为投票矩阵，表示网络对当前特征是否聚集到某高级胶囊的投票。$\boldsymbol{w}$是视角不变矩阵，网络通过$\boldsymbol{w}$获得视点等变性。concat$(*)$表示对$*$中元素进行拼接。s为高级编码胶囊的集合，设定s中每个高级编码胶囊的长度表示对应投票正确的概率，因此需要对s进行归一化处理。本文使用Squash函数来完成归一化，然后得到输出编码胶囊v。那么v中胶囊的模长将被压缩为0～1，模长最大的胶囊所在位置的对应类别被网络认为是最正确的投票，即为网络对输入的预测输出。路由算

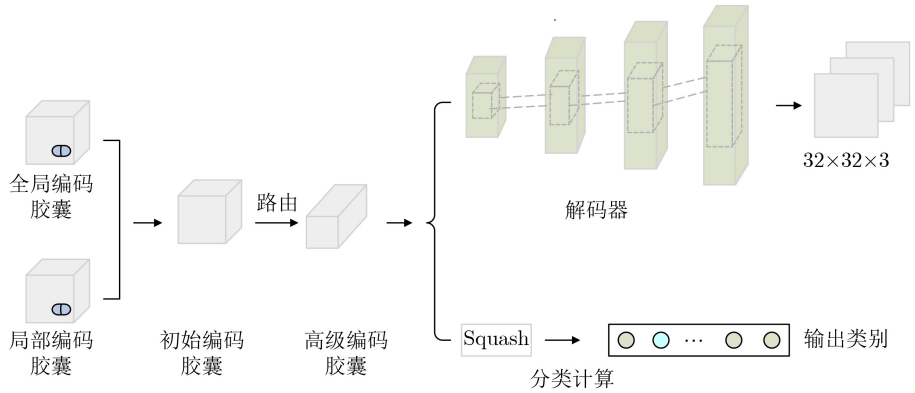


图 3 基于互信息的胶囊自编码器评估模型

表 2　基于编码胶囊的路由伪代码

输入x，t=3，初始化$b=0$
步骤1　计算c，$c \leftarrow \mathrm{Softmax}(b)$
计算h，$h \leftarrow H(x)$
计算g，$g \leftarrow G(h)$
计算u，$u = w \cdot \mathrm{concat}(g,h)$
更新s，$s \leftarrow \sum(c \cdot u)$
更新v，$v \leftarrow \mathrm{Squash}(s)$
更新t，$t \leftarrow t-1$
步骤2　更新b，$b \leftarrow b+(g+h)\cdot v$
步骤3　输出v
完成步骤1，当t不为0时，完成步骤2更新b，并将b代入步骤1计算v和t；当t为0时，结束迭代计算，完成步骤3。

法中的各种参量都需要通过网络的反向传播来进行学习。本文使用互信息作为反向传播的损失函数，由局部互信息Loss_h、全局互信息Loss_l和先验损失Loss_p的加权和组成，损失函数表示为

$$\mathrm{Loss} = \varepsilon \cdot \mathrm{Loss}_l + \tau \cdot \mathrm{Loss}_p + \rho \cdot \mathrm{Loss}_h \tag{32}$$

4　实验

4.1　数据库

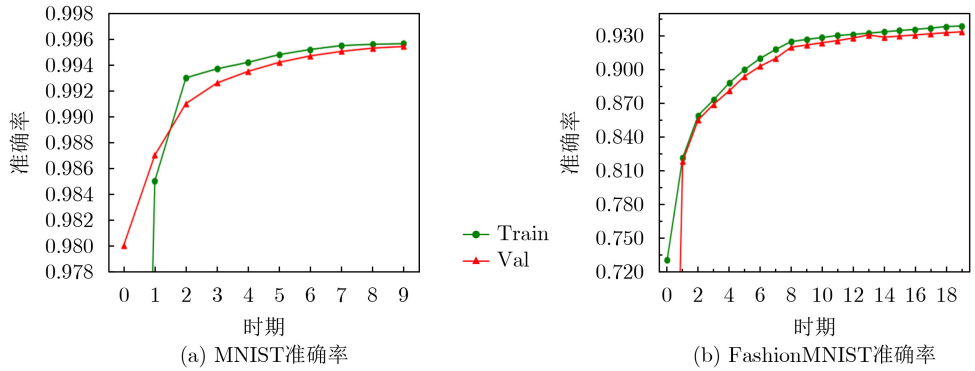
本文实验使用了4个数据集，分别是MNIST[19], FashionMNIST[20], CIFAR-10[18]和CIFAR-100[18]。MNIST数据集由200多个不同职位的人的手写阿拉伯数字构成(10类)。数据集总共包含70000个样本，每个样本都带有标签。FashionMNIST数据集则是由德国的一家时尚科技公司旗下部门提供，由日常衣物鞋类构成。与MNIST数据集一样，总共有70000个样本图片，分为10类，每类有7000张图片。CIFAR-10数据集是深度学习研究中使用最广泛的数据集之一，由60000张普适物体图片集组成。每张图片有RGB(红、绿、蓝)3个通道，包含鸟类、狗、轮船和卡车等生活常见物体，一共10个类别，每个类别分别有6000张图像。CIFAR-100数据集是一个相对比较复杂的数据集，由60000张图片组成，共100个类别，每个类别包含600张图片。

4.2　基于变分路由的胶囊网络实验分析

4.2.1　分类准确率评估

图4展示了采用变分路由算法的胶囊网络分类准确性随着处理批次(epoch)增加的变化曲线。其中，分别将MNIST和FashionMNIST中的55000张图片作为训练集，剩余15000张作为测试集。

从图4中可以发现，MNIST由于图像呈现形式简单，因此经过9个epoch，性能基本稳定；相比较之下，FashionMNIST中图像类别差异性较大，更为复杂，导致模型收敛速度相对较慢。在MNIST数据集上，变分模型平均分类准确率可以达到99.50%；



(a) MNIST准确率　　(b) FashionMNIST准确率

图 4 变分路由胶囊网络分类准确性

在FashionMNIST数据集上达到了93.50%，验证了数据集复杂度不同对模型预测能力的影响。

表3对比了经典CNN, ResNet[21]和Inception-V3[22]、基本胶囊网络(Capsule Network, CN)和变分胶囊网络(Variational Bayesian Capsule Network, VBCN)在上述两个数据集上的测试效果。可以发现：在MNIST上，VBCN比CN提升了0.2个百分点的准确率，在FashionMNIST前者比后者提升了1%左右的准确率。相比较经典CNN而言，胶囊网络具有明显分类增益，体现了潜在结构的优势。相比较于ResNet和Inception-V3这些改进后的CNN，VBCN在MNIST上的准确率要高于这两个网络，但是在FashionMNIST数据集上要低于这两个网络。对于MNIST这种简单的数据集，胶囊网络通过向量建模的方式要优于CNN网络。但是对于复杂的数据集，因为路由网络需要消耗巨大计算资源的限制，无法像CNN一样堆叠很多层去提取特征，因此性能表现暂时劣于CNN的改进模型。这一结果给后期胶囊网络的改进提出了新的思路和启示。

4.2.2 模型泛化能力评估

为了测试和验证模型对数据异构性的适应能力，即模型泛化性能，本文仿照文献[2]设计了两个扩展数据集：Two_MNIST和Two_FashionMNIST。这两个数据集分别由原始MNIST和FashionMNIST样本在垂直方向上拼接构成。标签为包含两个元素的数组构成，依次记录被拼接的两个样本标签。

表4给出了分类准确率的对比结果。其中标签“Two_MNIST”表示用MNIST数据集作为训练数据，人工生成的Two_MNIST数据集作为测试集的结果。标签“Two_FashionMNIST”具有类似含义。训练数据为55000张，测试数据为15000张，图样与训练数据不重叠。从表4中可以看出：得益于CN的优秀结构化设计，CN比CNN, ResNet, Inception-V3等CNN有大幅的性能增益，从另一个侧面反映了CNN对形变和图像内容姿势变化的敏感性，而这点恰好是CN的长处，即CN就是为了克服CNN的该缺陷所提出来的。另外，VBCN比CN也有一定程度的效果提升，这主要得益于VB-VN的变分估计是“集合”估计，而CN的模型估计是“点”估计，前者精准性更佳。

表 3 分类准确率对比(%)

模型	MNIST 准确率	FashionMNIST 准确率
CNN	98.00	90.30
ResNet	99.27	94.90
Inception-V3	99.29	94.97
CN	99.30	92.50
VBCN	99.50	93.50

4.3 基于互信息的胶囊自编码器实验分析

4.3.1 基于最邻近样本的互信息机制评估

为了可视化基于互信息机制的特征分类能力，通过设置式(32)中不同的ε和ρ的值来调节局部互信息和全局互信息在网络中比重，从而来观察两者对最终效果的影响。为保证实验的公平性，保持先验损失Loss_p的权重τ不变，只交替的改变局部互信息Loss_h和全局互信息Loss_l的权重。

图5—图7分别给出了CIFAR-10数据集在下述3种不同情况下的最邻近样本图：情况1下设$\varepsilon=2$, $\rho=1$,$\tau=0.01$；情况2下设$\varepsilon=1$, $\rho=0$, $\tau=0.01$；情况3下设$\varepsilon=0$, $\rho=1$, $\tau=0.01$。实验测试发现当将ε设置为与ρ相等时能达到最好的效果，将τ设为较小的数是为了降低先验信息对网络的影响。待这3种情况达到收敛状态后，实验使用欧氏距离衡量当前测试样本和其他测试样本的相似程度，即图5—图7，其中第1列为10个随机抽取的原始样本，其余

表 4 泛化性对比(%)

模型	Two_MNIST 准确率	Two_FashionMNIST 准确率
CNN	45.30	41.40
ResNet	89.09	59.60
Inception-V3	77.35	68.45
CN	93.15	82.60
VBCN	95.65	86.20

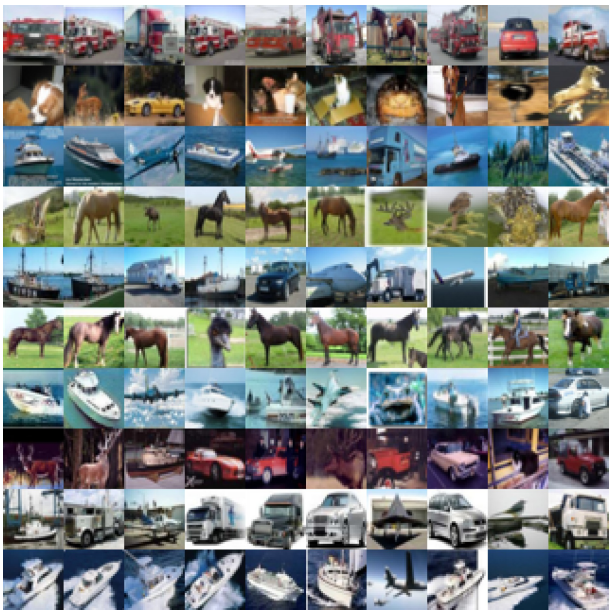


图 5 情况1下的最邻近样本示意图

9列同排样本表示网络计算所得最邻近的9个测试样本，按欧氏距离由大到小排列。

观察图5可以发现，网络能够通过不同测试数据的高级编码胶囊的相似度，判断输入是否为同一类别。如第1行，最左侧为原始样本卡车，然后通过计算相似度得到的9个最邻近样本都同为卡车。说明高级编码胶囊能够较为理想地代表原始输入图像。

经对比观察图6和图7，可以了解到全局互信息和局部互信息损失的系数ε和ρ值会对准确率产生一定影响，良好的分类性能高度依赖于局部项，局部互信息的缺少会导致网络的分类能力骤降。实验验证了网络通过引入局部编码器和局部互信息，不仅获得了从局部看待整体的能力，还能够提升网络的分类能力。

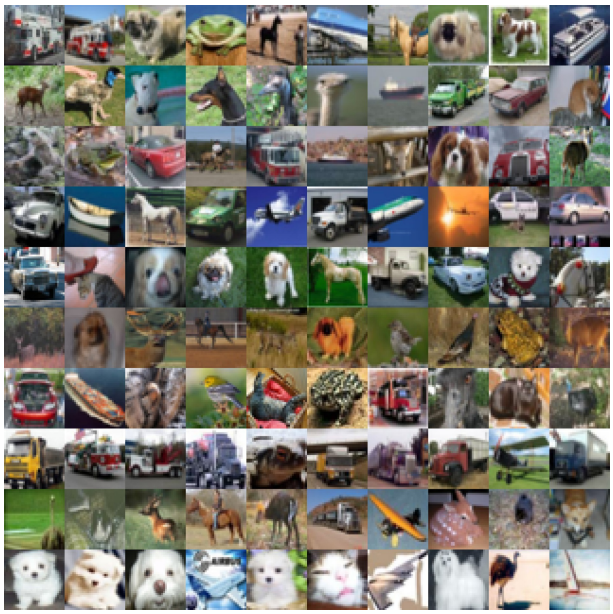


图 6 情况2下的最邻近样本示意图

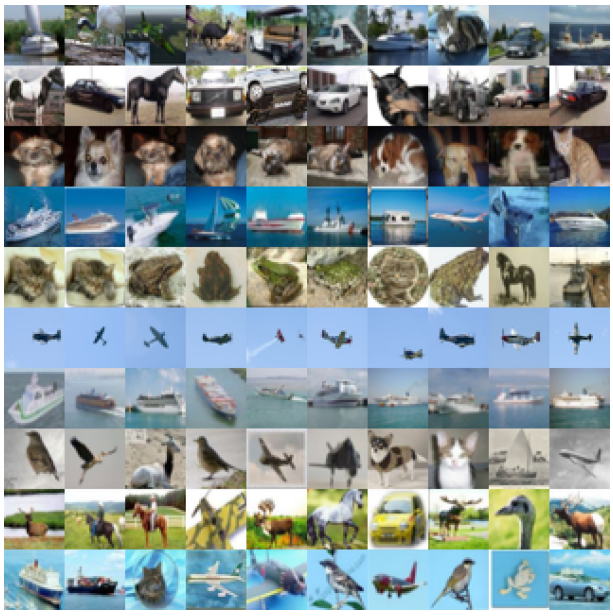


图 7 情况3下的最邻近样本示意图

4.3.2 分类准确率评估

本实验通过计算分类计算准确率来定量分析和比较改进模型和经典模型的优劣。在实验中发现[1]，如果向损失函数中添加边缘目标函数，会提高网络的分类准确率。因此，本次实验在损失函数中加入边缘目标函数，并给边缘目标函数添加一个较小的系数，以使互信息损失占主导地位。表5给出了两种对比方法针对每个类别测试时的准确率结果，其中训练数据为各类标签总共50000张图片，测试数据为各自类别剩余的总共10000张图片。

对比分析表5中数值，改进的CN模型在测试集上达到了平均72%的准确率，相比经典CN模型准确率提高了7%左右。此实验进一步证明本文模型提取的编码胶囊，能够高效代表输入图像特征，且能够直接通过分类计算模块，实现高准确率分类，无需重新使用其他网络模型对其训练分类。

除了在CIFAR10数据集上测试我们模型的性能，本文还在CIFAR100数据集上做了实验。对比分析表6中数值，改进CN模型比经典CN模型准确率提高了5.35%左右。这说明了在复杂数据集的情况下，互信息损失函数能够实现对编码胶囊的有效挑选，将最具辨别性的特征编码成胶囊去做分类，从而提高了网络的识别效果。

表 5　CIFAR-10测试准确率对比

标签	类别名称	经典CN准确率	改进CN准确率
0	飞机	0.73	0.81
1	汽车	0.76	0.87
2	鸟	0.71	0.74
3	猫	0.45	0.54
4	鹿	0.66	0.76
5	狗	0.55	0.60
6	青蛙	0.58	0.64
7	马	0.77	0.80
8	船	0.59	0.67
9	卡车	0.71	0.77
均值	---	0.65	0.72

表 6　CIFAR-100测试准确率对比(%)

模型	CIFAR-100准确率
经典CN	46.98
改进CN	52.33

5 结论

本文针对胶囊网络的基本结构进行了研究，提出了提高其特征提取能力和泛化能力的优化方法。本文的主要贡献如下：

(1)提出了基于变分路由的胶囊网络，通过实验验证了其分类、特征表示和泛化的能力，证明了基于变分路由的胶囊网络在迁移学习和特征整合上优于基本模型CNN；

(2)通过对基于矢量重构的互信息损失函数的推导，使胶囊自编码器获得了对编码胶囊的筛选能力，只保留胶囊中对原始输入进行识别分类最有效的特征，在减轻网络计算负担的同时提高了网络分类识别的能力。

参 考 文 献

[1] SABOUR S, FROSST N, and HINTON G E. Dynamic routing between capsules[C]. The 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 3856–3866.

[2] HINTON G E, SABOUR S, and FROSST N. Matrix capsules with EM routing[C]. International Conference on Learning Representations, Vancouver, Canada, 2018.

[3] GOLHANI K, BALASUNDRAM S K, VADAMALAI G, *et al*. A review of neural networks in plant disease detection using hyperspectral data[J]. *Information Processing in Agriculture*, 2018, 5(3): 354–371. doi: 10.1016/j.inpa.2018.05.002.

[4] PAOLETTI M E, HAUT J M, FERNANDEZ-BELTRAN R, *et al*. Capsule networks for hyperspectral image classification[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2019, 57(4): 2145–2160. doi: 10.1109/TGRS.2018.2871782.

[5] CHU Xin, XU Ning, LIU Xiaofeng, *et al*. Research on capsule network optimization structure by variable route planning[C]. 2019 IEEE International Conference on Real-time Computing and Robotics (RCAR), Irkutsk, Russia, 2019: 858–861.

[6] AUBERT G and VESE L. A variational method in image recovery[J]. *SIAM Journal on Numerical Analysis*, 1997, 34(5): 1948–1979. doi: 10.1137/S003614299529230X.

[7] 李速, 齐翔林, 胡宏, 等. 功能柱结构神经网络模型中的同步振荡现象[J]. 中国科学C辑, 2004, 34(4): 385–394. doi: 10.3321/j.issn:1006-9259.2004.04.012.

[8] MOON T K. The expectation-maximization algorithm[J]. *IEEE Signal Processing Magazine*, 1996, 13(6): 47–60. doi: 10.1109/79.543975.

[9] 西广成. 基于平均场理论逼近的神经网络[J]. 电子学报, 1995(8): 62–64. doi: 10.3321/j.issn:0372-2112.1995.08.016.
XI Guangcheng. Neural network based on mean-field theory approximation[J]. *Acta Electronica Sinica*, 1995(8): 62–64. doi: 10.3321/j.issn:0372-2112.1995.08.016.

[10] BISHOP C M. Pattern Recognition and Machine Learning[M]. New York: Springer, 2006: 293–355.

[11] GÖRÜR D and RASMUSSEN C E. Dirichlet process Gaussian mixture models: Choice of the base distribution[J]. *Journal of Computer Science and Technology*, 2010, 25(4): 653–664. doi: 10.1007/s11390-010-9355-8.

[12] SHRIBERG E, FERRER L, KAJAREKAR S, *et al*. Modeling prosodic feature sequences for speaker recognition[J]. *Speech Communication*, 2005, 46(3/4): 455–472.

[13] HJELM R D, FEDOROV A, LAVOIE-MARCHILDON S, *et al*. Learning deep representations by mutual information estimation and maximization[C]. 7th International Conference on Learning Representations, New Orleans, USA, 2019: 1–24.

[14] BELGHAZI M I, RAJESWAR S, BARATIN A, *et al*. MINE: Mutual information neural estimation[J]. arXiv: 1801.04062, 2018: 531–540.

[15] 徐峻岭, 周毓明, 陈林, 等. 基于互信息的无监督特征选择[J]. 计算机研究与发展, 2012, 49(2): 372–382.
XU Junling, ZHOU Yuming, CHEN Lin, *et al*. An unsupervised feature selection approach based on mutual information[J]. *Journal of Computer Research and Development*, 2012, 49(2): 372–382.

[16] 姚志均, 刘俊涛, 周瑜, 等. 基于对称KL距离的相似性度量方法[J]. 华中科技大学学报: 自然科学版, 2011, 39(11): 1–4, 38.
YAO Zhijun, LIU Juntao, ZHOU Yu, *et al*. Similarity measure method using symmetric KL divergence[J]. *Journal of Huazhong University of Science and Technology*: *Nature Science*, 2011, 39(11): 1–4, 38.

[17] PATHAK D, KRÄHENBÜHL P, DONAHUE J, *et al*. Context encoders: Feature learning by inpainting[C]. The IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 2536–2544.

[18] KRIZHEVSKY A and HINTON G E. Learning multiple layers of features from tiny images[R]. Technical report, 2009.

[19] LECUN Y, CORTES C, and BURGES C J C. MNIST handwritten digit database. 2010[OL]. http://yann.lecun.com/exdb/mnist, 2010, 7: 23.

[20] XIAO H, RASUL K, and VOLLGRAF R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv: 1708.07747, 2017.

[21] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. The IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778.

[22] SZEGEDY C, VANHOUCKE V, IOFFE S, *et al*. Rethinking the inception architecture for computer vision[C]. The IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 2818–2826.

鲍静益：女，1984年生，讲师，研究方向为模式识别与现代信号处理.
徐 宁：男，1981年生，副教授，研究方向为模式识别与现代信号处理.
尚蕴浩：男，1997年生，硕士生，研究方向为图像处理和深度学习.
楚 昕：女，1995年生，硕士生，研究方向为图像处理和深度学习.

责任编辑：余 蓉