

基于特征变换和度量网络的小样本学习算法

王多瑞^{1,2} 杜杨^{2,3} 董兰芳¹ 胡卫明² 李兵²

摘要 在小样本分类任务中, 每个类别可供训练的样本数量非常有限. 因此在特征空间中同类样本分布稀疏, 异类样本间边界模糊. 提出一种新的基于特征变换和度量网络 (Feature transformation and metric networks, FTMN) 的小样本学习算法用于小样本分类任务. 算法通过嵌入函数将样本映射到特征空间, 并计算输入该样本与所属类别中心的特征残差. 构造一个特征变换函数对该残差进行学习, 使特征空间内的样本特征经过该函数后向同类样本中心靠拢. 利用变换后的样本特征更新类别中心, 使各类别中心间的距离增大. 算法进一步构造了一种新的度量函数, 对样本特征中每个局部特征点的度量距离进行联合表达, 该函数能够同时对样本特征间的夹角和欧氏距离进行优化. 算法在小样本分类任务常用数据集上的优秀表现证明了算法的有效性和泛化性.

关键词 特征变换, 度量学习, 小样本学习, 残差学习

引用格式 王多瑞, 杜杨, 董兰芳, 胡卫明, 李兵. 基于特征变换和度量网络的小样本学习算法. 自动化学报, 2024, 50(7): 1305–1314

DOI 10.16383/j.aas.c210903

Feature Transformation and Metric Networks for Few-shot Learning

WANG Duo-Rui^{1,2} DU Yang^{2,3} DONG Lan-Fang¹ HU Wei-Ming² LI Bing²

Abstract For few-shot classification, training samples for each class are highly limited. Consequently, samples from the same class tend to distribute sparsely while boundaries between different classes are indistinct in the feature space. Therefore, a novel few-shot learning algorithm based on feature transformation and metric networks (FTMN) is proposed for few-shot learning. The algorithm maps samples to the feature space through an embedding function and calculates the residual between the input features and their class center. A feature transformation function is then constructed to learn from the residual, enabling input features to move closer to their class center after transformation. The transformed features are used to update the class centers, increasing the distance between centers of different classes. Furthermore, the algorithm introduces a novel metric function that jointly expresses the metric distances of each point within the features. The metric function simultaneously optimizes both cosine similarity and Euclidean distance. The performance of the algorithm on commonly used datasets for few-shot classification validates its effectiveness and generalization ability.

Key words Feature transformation, metric learning, few-shot learning, residual learning

Citation Wang Duo-Rui, Du Yang, Dong Lan-Fang, Hu Wei-Ming, Li Bing. Feature transformation and metric networks for few-shot learning. *Acta Automatica Sinica*, 2024, 50(7): 1305–1314

收稿日期 2021-09-20 录用日期 2021-12-11

Manuscript received September 20, 2021; accepted December 11, 2021

国家重点研发计划 (2018AAA0102802), 国家自然科学基金 (62036011, 62192782, 61721004), 中国科学院前沿科学重点研究计划 (QYZDJ-SSW-JSC040) 资助

Supported by National Key Research and Development Program of China (2018AAA0102802), National Natural Science Foundation of China (62036011, 62192782, 61721004), and Key Research Program of Frontier Sciences of Chinese Academy of Sciences (QYZDJ-SSW-JSC040)

本文责任编辑 郑伟诗

Recommended by Associate Editor ZHENG Wei-Shi

1. 中国科学技术大学 合肥 230026 2. 中国科学院自动化研究所模式识别国家重点实验室 北京 100190 3. 阿里巴巴科技 (北京) 有限公司 北京 100016

1. University of Science and Technology of China, Hefei 230026
2. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190 3. Alibaba (Beijing) Group, Beijing 100016

近年来, 神经网络相关研究快速发展, 现有的神经网络模型在大规模任务上^[1-4]有着非常可观的表现, 且经过训练的网络在面对新任务时, 通常可对网络模型参数进行微调来完成任务. 但在面对仅有几个样本的分类任务时, 即使对预训练过的神经网络模型进行微调, 也很难达到较好的识别效果. 因此, 近几年小样本学习的研究热度很高, 早期一些小样本学习模型^[5-9]获得了不错的结果, 但这些模型几乎只适用于一些特定问题, 不具备良好的扩展能力. 将在大型数据集上表现良好的深度网络模型迁移至小样本任务并保持原有性能, 是小样本学习的主要研究方向之一^[10-18].

小样本分类任务希望模型具有快速学习能力,

即面对样本数量有限的类别时,也可以达到不错的分类精度.直接使用预训练好的深度网络模型对小样本分类任务进行微调,很容易产生过拟合现象,因此引入基于度量的小样本学习算法.这类算法需要使用由卷积网络构成的嵌入函数对输入数据提取特征,但提取出的特征通常在特征空间内分布稀疏,导致分类任务的难度增大.因此,本文提出一种特征变换策略,可在特征空间中减小同类样本的类内距离,并增大异类样本的类间距离.同时,分析余弦相似度和欧氏距离在特征度量上的局限,提出一种同时考虑两种度量方式影响的特征度量策略.

1 相关研究

目前,在视觉相关领域的小样本学习算法大致可分为基于模型结构的小样本学习、基于度量的小样本学习和基于优化的小样本学习三类.这三类小样本学习算法基本都可以使用元学习的方法实现.基于模型结构的小样本学习主要希望通过设计一种模型,使模型结构支持其在小样本上快速地更新参数,直接建立输入与预测值间的函数关系;基于度量的小样本学习是将样本映射到一个样本空间,通过度量支持样本和查询样本间的距离,应用最近邻思想完成分类预测;基于优化的小样本学习认为一般的梯度下降法很难在小样本的情况下拟合出一个精度较高的映射关系,所以希望通过调整改进优化方法,来实现小样本分类任务.本文着重关注基于度量的小样本学习.

基于度量的小样本学习算法是用于小样本学习的主流方法之一.它借用元学习的思想,将分类任务使用的训练集和测试集进一步划分为支持集和查询集.在基于度量方法中,先验知识是一个嵌入函数^[11-12],其功能是将样本映射到一个特征空间,对样本与空间中的特征中心进行度量,用最近邻方法将样本划分到正确的类别中.基于度量的小样本学习算法目前已经有较多经典模型,例如文献^[11]指出,匹配网络(Matching networks, MN)在大型数据集上的学习速度较慢,是因为网络的参数数量非常巨大,训练样本需要大量时间开销.还有一些传统的非参数模型可以快速地吸收、同化新样本,使训练时间减少,但是这些方法实际依赖所选择的度量.匹配网络希望结合参数化模型和非参数化模型的优点,既具有优秀的泛化能力,又可以快速地在新的样本上得到优秀结果.匹配网络的特点是将使用注意力机制的基于深度网络的度量学习和基于外部存储的记忆增强网络相结合,得到针对小样本任务的快速学习能力.原型网络^[12]将查询样本和支持样本通过同一个嵌入函数映射到一个特征空间中,在特征

空间中利用支持样本计算每个类别的“均值”作为该类“原型”,使用欧氏距离作为度量查询样本特征与各个“原型”间的距离,利用最近邻方法预测查询样本的标签类别.原型网络将“原型”概念与最近邻的分类思想相结合,探索一种简单、直接且有效的小样本学习思路,同时验证了使用欧氏距离作为度量距离的效果要略好于使用余弦相似度.孪生网络(Siamese networks, SN)^[13]通过学习嵌入函数来优化固定的最近邻和一个线性分类器,同时该网络框架的输入作为样本对,一对样本分别经过结构相同和参数共享的嵌入函数,并计算输出结果,判断得到这对输入样本是同类还是异类.所以本质上是在训练一个判断两个样本是否相同的二分类器.由于网络的输入是成对的,所以其有效样本数量非常大,不会存在严重的过拟合问题.关系网络(Relation networks, RN)^[14]具有嵌入函数以及关系模块,该网络的特点是可以学习推理,它的整体结构简单明了,仅将支持样本和查询样本的特征在通道的维度上进行拼接,然后将合并后的特征作为度量函数的输入.关系网络的主要创新之处在于使用了一个卷积网络构造度量函数,与上述的欧氏距离和余弦相似度不同,它可以学习到一种合适的度量距离,自适应地优化模型的表现效果.

局部聚合向量(Vector of locally aggregated descriptor, VLAD)^[19-20]是一种样本特征的表示方法.该方法与现在流行的深度学习网络不同,是一种比较传统的提取图像特征算法,需要先提取局部特征,再将这些局部特征聚类得到VLAD向量.但是,经典的VLAD算法的表征特征是稀疏且不可导的.为使其可以应用到神经网络中进行训练,则需要让它变得平滑.NetVLAD^[19]可以在计算局部特征与其所属的聚类中心差时,在前面乘以一个权值函数 $a_k(x_i)$,该值在局部特征越接近聚类中心时,权值越高;反之,权值越低.与传统VLAD不同的是,NetVLAD的聚类中心是由训练网络得到的,而非通过预先训练好的码本获得.NetVLAD将局部聚合向量作为卷积神经网络的一部分嵌入到网络中,使得VLAD层中所有参数都可以通过端到端的方式学习得到.

目前,很多研究为小样本学习算法研究提供了新思路.使用岭回归方法的元学习算法^[21]是将元学习中学习器的分类方法由最近邻法或全连接层更换为带有闭式解的可微分回归方法,验证了成熟的传统优化方法也可以用于分类器的参数优化.基于边标签的图神经网络(Edge-labeling graph neural network, EGNN)^[22]是基于图卷积网络的小样本学习方法,它利用类内相似性和类间差异来迭代更新

边标签,再用边标签预测节点标签,并用转导方法将支持集和查询集特征同时训练模型,以此优化了小样本分类的效果.这样做的好处是适用于不同类别间的迁移,且迁移后不需要重新进行参数训练.基于因果干预的分类器的小样本学习^[23]设计了一个基于因果推理的分类器,希望通过因果干预手段,去除预训练模型在提取特征时受到混杂因子的影响,提高模型分类的准确率.该方法设计的分类器训练时,不对原预训练的模型参数进行微调,仅需训练分类器的参数即可优化模型的表现,充分发挥了算法模块化的优点.

综上所述,小样本学习的发展由网络结构的整体优化向模块化方向推进.因此,本文以经典的原型网络为基线模型,利用 VLAD 使用局部残差表示图像的方法,按照模块化思路,设计全新的特征变换模块和特征度量模块,对样本特征表达和基于度量的分类器度量方法进行优化.

2 特征变换和度量网络

本文的研究目标是设计用于小样本分类任务的特征变换模块和特征度量模块,特征变换模块在将样本映射进特征空间后,使原本因神经网络过拟合造成的特征聚类分散问题得以缓解,特征度量模块在特征空间内设计出优于常用的余弦相似度或欧氏距离的度量方法,以缓解它们单独存在时的限制.由于在小样本任务中,可用于训练的数据较少,在数据集上使用预训练模型对网络模型参数进行微调很容易造成过拟合;在使用基于度量的小样本学习算法时,提取出的特征聚类分散,呈现同类样本分布较为稀疏的特点,造成不同类别样本难以区分的问题.因此,需要对提取出的局部特征进行变换,使这些特征可以更好地用于小样本的分类任务.

首先,设计特征变换模块.提取输入样本的特征图形成一个特征空间,在该特征空间内,对特征图上的局部特征进行变换,使样本局部特征向同类样本的局部特征中心移动.同一类别中的全部样本局部特征进行变换后,重新计算得到新的局部特征中心,从而达到不同类别的特征中心互相远离的目的,以减小分类难度,提高分类准确率.然后,设计一个特征度量模块,训练时,网络可同时对特征及其特征中心间的欧氏距离和夹角进行优化,以达到比单独使用欧氏距离或余弦相似度的度量方法更好的性能.

在各个模块设计中,本文使用非线性 VLAD 思想,即在嵌入函数提取了样本的特征图后,对特征图的每个位置上的局部特征进行操作.首先,通过计算均值,得到每个局部特征对应类别的支持特征

中心;然后,对每个局部特征进行特征变换以及度量;最后,使用特征度量函数将特征图上得到的全部局部特征度量距离进行联合表达,以此计算特征图的全局特征度量距离,使用该距离对特征图代表的样本进行分类.

2.1 特征变换模块

本节目标是提出特征变换模块,将特征中心移动到更好位置;因此,设计的特征变换模块包含了特征变换函数和迁移得分.

本文使用嵌入函数 θ 将输入网络的样本映射为特征图 $f \in \mathbf{R}^{W \times H \times C}$, 其中 $W \times H$ 为特征图的维度, C 为特征图的通道数.该嵌入函数由卷积神经网络构成,将输入的样本图像进行卷积计算,并通过最大池化层对样本图像进行下采样,得到样本图像的特征图 f . 可以从每个特征图 f 中得到一组局部特征 $\{\mathbf{x}_i \in \mathbf{R}^C\}$, 其中 $i \in W \times H$ 代表特征图上每个局部特征的位置.综上所述,每个支持样本可以得到一组支持特征 $\{\mathbf{x}_{s,i}\}$, 其中 s 代表该组特征属于支持集;每个查询样本可以得到一组查询特征 $\{\mathbf{x}_{q,i}\}$, 其中 q 代表该组特征属于查询集.网络模型提出一种收缩变换,利用特征变换函数对样本的各个局部特征与其正确类别局部特征中心的残差进行学习,利用残差使每个种类的特征向中心收缩.首先,利用支持集中的全部样本,计算它们特征图中位置处属于类别 k 的支持特征的均值:

$$\bar{\mathbf{x}}_{k,i} = \frac{1}{S} \sum_{s \in S} \mathbf{x}_{s,k,i} \quad (1)$$

将均值 $\bar{\mathbf{x}}_{k,i}$ 作为位置 i 处类别 k 的支持特征中心;接着,计算支持特征中心 $\bar{\mathbf{x}}_{k,i}$ 和位置 i 上任意支持样本特征的残差:

$$\Delta_{s,k,i} = \bar{\mathbf{x}}_{k,i} - \mathbf{x}_{s,i} \quad (2)$$

然后,设计一个抽象的特征变换函数 Φ , 为输入的支持特征学习一个对应其正确类别的支持特征中心的残差表达,特征变换函数也由多层卷积神经网络构成:

$$\delta_{s,k,i} = \Phi(\Delta_{s,k,i}), \forall s, k, i \quad (3)$$

该残差表达用于当前支持特征变换的计算,在变换时,希望该特征与所有支持特征中心间的残差都发挥作用,所以需要为残差设置权值并进行软分配,于是对每个支持特征中心定义一个特征迁移得分:

$$a_{k,i} = \frac{\exp(\omega^T \bar{\mathbf{x}}_{k,i} + b)}{\sum_{k' \in K} \exp(\omega^T \bar{\mathbf{x}}_{k',i} + b)} \quad (4)$$

该得分是一个注意力得分,通过一个卷积网络计算,

并通过 Softmax 层输出, 其中 ω 和 b 均为可学习参数. 综上所述, 收缩变换后的特征计算方式为:

$$\tilde{\mathbf{x}}_{s,i} = \mathbf{x}_{s,i} + \sum_{k \in K} a_{k,i} \delta_{s,k,i} \quad (5)$$

最后, 将变换过的支持特征通过一个线性整流函数激活, 并重新计算各类别的支持特征中心. 在理想状态下经过变换的特征会移动到与其特征中心相同的位置, 即学习到的特征变换如下:

$$\tilde{\mathbf{x}}_{s,i} = \mathbf{x}_{s,i} + 1 \times I(\tilde{\mathbf{x}}_{k,i} - \mathbf{x}_{s,i}) = \tilde{\mathbf{x}}_{k,i} \quad (6)$$

式中, k 为 $\mathbf{x}_{s,i}$ 的正确类别, 特征对应的正确特征中心得分 $a_{k,i}$ 为 1, 其余得分均为 0. 特征变换函数 Φ 表现为 $I(\cdot)$ 是一个残差 $\Delta_{s,k,i}$ 的线性函数, 如此待迁移的支持特征中心的位置保持不变. 但若这种情况发生, 将很难为每个正确的类别中心学习到一个最高得分, 因为这些类别中心间本身在特征空间中的位置就可能十分接近. 所以, 还是希望将非线性的 Φ 作为特征变换函数, 来实现类别中心的偏移, 达到增大不同类别间距离的目的. 偏移的支持特征中心通过对收缩变换后的支持特征重新求均值计算得到:

$$\tilde{\mathbf{x}}_{k,i} = \frac{1}{S} \sum_{s \in S} \tilde{\mathbf{x}}_{s,k,i} \quad (7)$$

同理, 查询样本变换后的特征通过使用 q 代替式 (2)、式 (3) 和式 (5) 中的 s 进行计算.

2.2 特征度量模块

本节目标是设计新的度量方式和距离度量函数, 希望该函数可以从数据中学习一种最适合的度量方式. 基于度量的的小样本算法是常用的度量方法, 使用欧氏距离和余弦相似度作为度量, 欧氏距离用来度量特征空间中两点间的几何距离, 余弦相似度度量内积空间中两个非零向量间的相似度, 即它们之间夹角的余弦值. 较小的余弦相似度可以保证样本特征和特征中心的方向基本一致, 但对特征间大小的相似度没有限制可能导致分类错误. 在多数情况下, 使用欧氏距离的性能要好于使用余弦相似度, 较小的欧氏距离可以保证两个特征间的几何距离很近, 但当两个特征间的欧氏距离非零时, 它们之间的夹角并不唯一. 由此可见, 这两种常用的度量方式都有各自的局限性, 为了解决这样的限制, 本文提出一种互补性的距离度量方式.

首先, 计算特征图中位置 i 的查询样本特征与支持特征中心间的距离如下:

$$d_{q,k,i} = \lambda_{q,k,i} \|\tilde{\mathbf{x}}_{q,i} - \tilde{\mathbf{x}}_{k,i}\|^2 \quad (8)$$

式中, $\lambda_{q,k,i}$ 表示一个系数, 用来平衡欧氏距离大小的影响. 使用余弦相似度设计系数 $\lambda_{q,k,i}$:

$$\lambda_{q,k,i} = \text{sigmoid} \left(\frac{1}{\cos \langle \tilde{\mathbf{x}}_{q,i}, \tilde{\mathbf{x}}_{k,i} \rangle + 1} - 0.5 \right) \quad (9)$$

这种方式是希望同时对余弦相似度和欧氏距离进行优化. 这样设计的度量距离希望参数 $\lambda_{q,k,i}$ 在使用欧氏距离的基础上, 对度量距离有一定影响而非具有主导作用. 当余弦相似度趋近于 1 时, 度量距离趋近于特征向量与特征中心欧氏距离的一半; 在特征向量与特征中心间夹角较大时, 则以它们间的欧氏距离作为主导. 使用一个非线性特征度量函数 D 计算特征图中所有位置 $\{d_{q,k,i}\}$ 的联合距离表达, 将其结果作为查询样本和支持样本中心间最终距离度量:

$$d_{q,k} = D(\{d_{q,k,i}\}), \quad d \in \mathbf{R} \quad (10)$$

本文使用 $\lambda_{q,k,i}$ 对欧氏距离进行加权. 当特征空间中查询特征与其所属的类别特征中心间角度相差较小但欧氏距离较大时, 度量距离大小受参数 $\lambda_{q,k,i}$ 影响较大, 它们之间最终距离也因 $\lambda_{q,k,i}$ 的关系变小, 可以减少一些正样本被分类错误情况.

最后, 使用与查询样本具有最小距离 $d_{q,k'}$ 的支持样本中心类别作为模型, 对该输入类别预测 y :

$$y = \text{argmin}(d_{q,k'}) \quad (11)$$

2.3 网络训练方式

计算每个查询样本与支持样本中心间的距离后, 使用距离的负数 Softmax 归一化输出, 生成一个针对查询点跨越各类别的分布:

$$p_{\theta, \Phi, D}(y = k | z_q) = \frac{\exp(-d_{q,k})}{\sum_{k' \in K} \exp(-d_{q,k'})} \quad (12)$$

通过最小化模型对正确类别 k 所对应的负对数预测概率:

$$J_c(\theta, \Phi, D) = -\log_2 p_{\theta, \Phi, D}(y = k | z_q) \quad (13)$$

式 (13) 可以联合地学习嵌入函数 θ 、特征变换函数 Φ 以及特征度量函数 D .

为了使特征点可以向正确的样本类别中心偏移, 在训练过程中, 对每个特征向各个支持特征中心的迁移得分进行了约束:

$$\bar{a}_s(y = k | z_s) = \frac{1}{WH} \sum_{i=0}^{WH} a_{s,k,i} \quad (14)$$

$$\bar{a}_q(y = k | z_q) = \frac{1}{WH} \sum_{i=0}^{WH} a_{q,k,i} \quad (15)$$

为清晰表示, 将 $a_{k,i}$ 写为输入为支持样本时 $a_{s,k,i}$ 或查询样本时 $a_{q,k,i}$ 两种形式. \bar{a}_s 和 \bar{a}_q 分别表示支持和查询样本的约束. 同样取其最小化负对数:

$$J_a(\theta, \Phi) = -\ln \bar{a}_s - \ln \bar{a}_q \quad (16)$$

来约束学习特征的正确迁移. 同时最小化 J_c 和 J_a 训练模型. 首先, 对训练集中的每一次迭代, 随机抽取训练集中的 K 个类, 每个类别中包含几个支持样本; 然后, 再随机抽取每个类别中的一批样本作为查询样本, 即每一次迭代的输入为一组样本. 使用自适应矩估计优化算法训练神经网络模型, 该算法也使用小批量抽取样本方式进行训练, 设置一批数据的大小 n , 即经过 n 次迭代后网络模型的参数会进行一次更新.

2.4 网络模型结构

本文提出的特征变换和度量网络 (Feature transformation and metric networks, FTMN) 模型结构如图 1 所示.

支持样本和查询样本均取自一个基本的数据集, 且支持样本和查询样本的数据类别一致. 将这些样本分别输入嵌入函数中, 提取样本特征. 本文的嵌入函数是使用卷积神经网络实现的, 分别使用 4 层卷积结构和 ResNet-12 结构作为嵌入函数, 其中 4 层卷积结构使用的网络, 与作为基线原型网络使用的嵌入函数结构相同.

接着, 求支持样本中同类样本特征的均值, 得到每个支持样本特征图的每个位置上的局部特征的特征中心. 通过计算得到支持特征中心与支持样本集中所有的样本特征的残差, 并将残差输入特征变换函数 Φ 中, 在使用 4 层卷积结构作为嵌入函数时, 函数的模型结构如图 2(a) 所示; 在使用 ResNet-12

作为嵌入函数时, 函数 Φ 结构减少为 2 层卷积, 以缓解网络的过拟合问题, 其他部分不变. 再将查询特征与各个支持特征中心的残差送入特征变换函数, 对残差进行学习. 支持特征和查询特征的残差都会通过一个单层的注意力网络学习到一个迁移得分, 得分由注意力网络的参数和支持特征中心决定. 最后, 得到特征迁移变换后的支持特征和查询特征, 以此减小同类样本间的距离, 同时变换后的支持特征还要参与支持特征中心的重新计算.

将变换后的查询特征和支持特征送入度量模块, 利用变换后的支持特征重新计算支持特征中心, 实现支持特征中心的移动, 增大不同类别间的距离. 同时, 这个新的支持特征中心将用于与变换后的查询特征间的度量, 计算它们之间的余弦相似度, 并按式 (9) 得到 $\lambda_{q,k,i}$ 值, 再按照式 (8) 计算查询特征图上每个位置的局部度量距离表达, 最终将每张查询特征图上每个位置的局部度量距离送入特征度量函数 D , 得到该查询特征的全局度量距离和分类结果. 特征度量函数 D 的网络结构如图 2(b) 所示.

本文将小样本学习中经典网络模型和一些近期表现良好的网络模型重要结构和使用的嵌入函数结构进行总结和对比, 结果见表 1. 嵌入函数一般选择使用 4 层卷积网络提取特征和使用 ResNet-12 网络提取 2 种, 本文使用这 2 种嵌入函数算法进行实验, 其中使用 4 层卷积网络作为嵌入函数命名为特征变换和度量网络, 使用 ResNet-12 网络作为嵌入函数命名为 FTMN-R12.

相较于 4 层卷积网络, ResNet-12 提取出的特征具有更好效果, 但是其参数量和计算时间也会更大. 所以本文分别使用 2 种嵌入函数进行实验. 在模型的重要结构部分, 本文的重点是对特征变换模

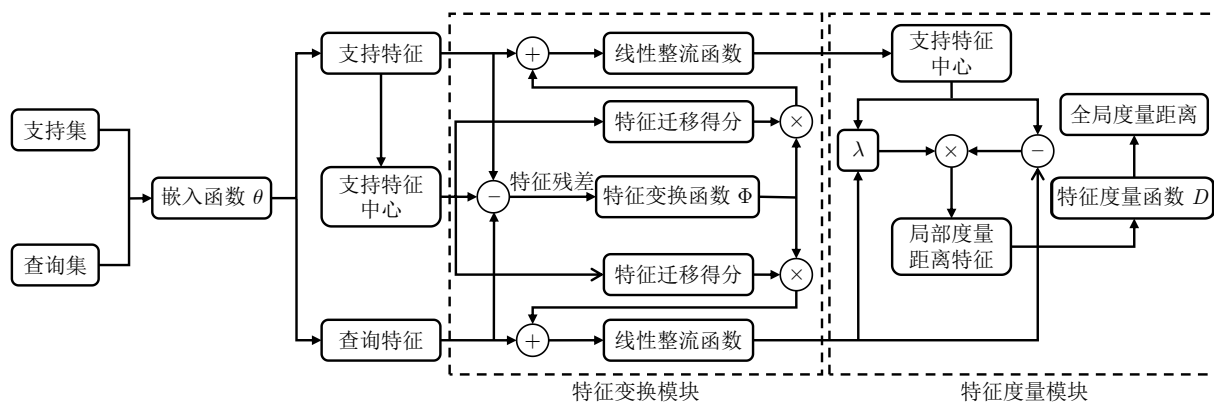


图 1 特征变换和度量网络模型

Fig.1 Model of feature transformation and metric networks

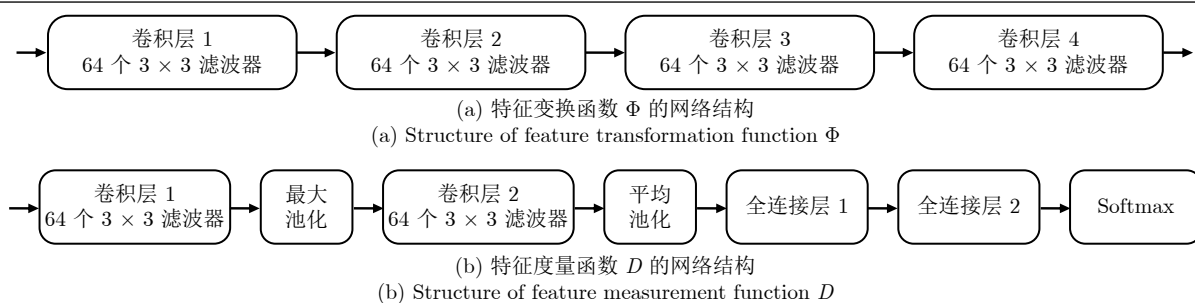


图 2 网络中关键函数的结构

Fig. 2 Structure of important functions of networks

表 1 网络模型的嵌入函数与重要结构
Table 1 Embedding function and important structures of networks

模型名称	嵌入函数	重要结构
MN	4 层卷积网络	注意力长短时记忆网络
ProtoNet ^[12]	4 层卷积网络	“原型”概念、使用欧氏距离进行度量
RN	4 层卷积网络	卷积神经网络作为度量函数
EGNN	4 层卷积网络	边标签预测节点类别
EGNN + Transduction ^[22]	ResNet-12	边标签预测节点类别、转导和标签传递
DN4 ^[24]	ResNet-12	局部描述子、图像与类别间的相似性度量
DC ^[25]	4 层卷积网络	稠密分类
DC + IMP ^[25]	4 层卷积网络	稠密分类、神经网络迁移
FTMN	4 层卷积网络	特征变换模块、特征度量模块
FTMN-R12	ResNet-12	特征变换模块、特征度量模块

块和特征度量模块的设计, 通过主动特征变换来优化特征聚类的效果, 又将 2 种传统的度量方式与卷积神经网络相结合, 得到一个可进行自优化的特征度量模块, 这 2 个模块共同发挥作用, 对本文小样本学习分类效果进行优化. 表 1 中其他模型的重要结构分别为: 1) 匹配网络利用注意力机制和长短时记忆网络表现出了支持特征与查询特征间存在的联系, 以此优化网络效果; 2) 原型网络通过计算类别的原型, 优化特征聚类的效果; 3) 关系网络首次使用卷积网络作为度量函数, 并以此优化了小样本学习的正确率; 4) 深度最近邻网络利用局部描述子描述样本特征, 并将样本与类别的相似性作为度量的目标, 使分类效果有了进一步提升; 5) 稠密分类和神经网络迁移利用深度最近邻思想, 结合迁移学习, 设计一种平行网络结构, 通过大量计算和微调优化分类结果; 6) 基于边标签的图卷积网络利用边标签预测节点标签, 并利用转导方法, 同时利用支持集和查询集特征, 以此优化小样本分类效果. 详细实验结果见第 3 节.

3 实验结果

本文使用 Omniglot 和 miniImageNet 两个数据集, 验证基于度量的特征变换网络模型的有效性; 再通过 CUB-200、CIFAR-FS 和 tieredImageNet 三个数据集, 验证模型的泛化性. 通过将 Omniglot 数据集中的原始图像旋转 90° 、 180° 、 270° 进行数据增强. 原数据集共有 1 623 种手写字符, 分成 1 200 种和 423 种 2 组, 前者及其旋转过的图像用作训练集, 后者及其旋转过的图像用作测试集. 主要使用的小样本任务类型为 5-类 1-样本、5-类 5-样本、20-类 1-样本、20-类 5-样本四种, 其中“类”表示任务中样本种类的数量, “样本”表示每个类别中支持样本的数量. 在训练集中, 以上 4 个任务类型的支持集中每类分别有 1、5、1、5 个图像样本, 查询集中每个类别分别有 15、15、10、5 个图像样本. 训练时, 每个任务类型、每轮训练的总图像数为 80、100、220、200 张. 训练设置每轮完整训练包含 2 000 个阶段, 模型共经过 20 000 次迭代训练, 使用自适应矩估计优化算法, 优化算法中超参数只有学习率

需要进行动态调整, 本文使用模拟退火算法对学习率进行更新.

网络模型在训练完成后, 使用与训练集设置相同的测试集, 对网络性能进行评估.

随机抽取 1 000 次测试样本进行测试, 将得到的结果取平均值, 作为本文模型在 Omniglot 数据集上的分类精度. 实验结果如表 2 所示, 本文模型在 5-类任务上与最优结果相当, 并在 20-类任务上取得了最优结果. 由于 Omniglot 数据集的准确率接近饱和, 因此提高的空间十分有限, 目前多数模型在该数据集上达到了 99% 以上的准确率, 可以作为验证模型有效的最基本判别标准.

miniImageNet 数据集样本丰富且复杂, 样本均取自真实世界, 是小样本学习领域中最具影响力和说服力的数据集之一, 该数据集本身较大, 不需要进行数据增强. 本文对该数据集进行划分, 将其以 80:20 的比例分割为训练集和测试集, 又在训练集中抽取 16% 作为验证集, 验证该模型的通用性. 在该数据集上, 本文按照最常用的小样本学习任务的数据模式对 5-类 1-样本和 5-类 5-样本进行测试. 在 miniImageNet 数据集上, 随机在测试集中抽取 600 个测试样本集, 取测试结果的平均值作为该网络在数据集上的分类精度. 实验结果如表 3 所示, 在使用 ResNet-12 作为模型嵌入函数时, 本文算法得到优良表现, 效果仅次于最优模型. 本文算法的基线模型是使用 4 层卷积网络作为嵌入函数的原型网络, 模型在使用同样嵌入函数时, 在 1-样本任务中, 取得了高于基线模型准确率 10% 左右的结果; 在 5-样本任务中, 取得了高于基线模型准确率 7.7% 的结果. 使用 ResNet-12 作为嵌入函数时, 在 1-样本任务中, 取得了高于基线模型准确率 12% 的结果, 与最优结果相差 1%; 在 5-样本任务中, 取得了

高于基线模型准确率 11% 的结果, 与最优结果相差 0.2%.

为验证模型的泛化性, 选择 3 个小样本分类任务常用的数据集 CUB-200、CIFAR-FS 和 tiered-ImageNet 进行训练和测试, 实验结果如表 4. 由表 4 可知, 本文模型在 CUB-200 上取得了最优结果, 在 CIFAR-FS 上得到的 1-样本、5-样本结果均与最优模型相差 2% 左右. tieredImageNet 和 miniImageNet 是 ImageNet 的子集, 它们的规模比 ImageNet 数据集大, 在 1-样本结果上, 本文方法与 MetaOptNet 相差将近 4%, 但在 5-样本上的结果却与它相差无几, 这说明本文的特征变换模块发挥了很大作用. 由泛化实验可以看出, 本文提出的模型能在绝大多数数据集上表现优秀, 且均可接近或达到最优模型表现.

消融实验测试本文提出的特征变换度量网络中各个模块是否起作用. 首先, 使用基础网络为原型网络, 将具有 8 层卷积网络作为嵌入函数, 为原型网络提取特征. 再使用原型网络的欧氏距离进行度量, 相较于具有 4 层卷积的原型网络, 在 1-样本和 5-样本任务中, 精度分别提升 1.76% 和 2.03%, 将这个结果作为本文变换网络的基准, 并称该网络为 ProtoNet-8C. 将嵌入函数换成 4 个卷积模块, 然后将 4 个卷积模块作为特征变换函数训练, 将输出的结果通过平均池化层, 并使用欧氏距离作为度量, 将 ProtoNet-8C 中由 8 个卷积模块构成的嵌入函数减少为 4 个, 称该网络为 ProtoNet-Trans-4C. 在 ProtoNet-4C 网络中, 使用本文设计的度量函数替换原本的欧氏距离, 称该网络为 ProtoNet-M-4C. 本文再将特征变换函数替换为 VLAD 算法, 使用 VLAD 向量重新计算残差, 保持其他部分不变, 称该模型为 ProtoNet-VLAD-4C. 为了测试缩小类内

表 2 在 Omniglot 数据集上的小样本分类性能 (%)
Table 2 Few-shot classification performance on Omniglot dataset (%)

模型	5-类		20-类	
	1-样本	5-样本	1-样本	5-样本
MN	98.1	98.9	93.8	98.5
ProtoNet ^[12]	98.8	99.7	96.0	98.9
SN	97.3	98.4	88.2	97.0
RN	99.6 ± 0.2	99.8 ± 0.1	97.6 ± 0.2	99.1 ± 0.1
SM ^[15]	98.4	99.6	95.0	98.6
MetaNet ^[16]	98.95	—	97.00	—
MANN ^[17]	82.8	94.9	—	—
MAML ^[18]	98.7 ± 0.4	99.9 ± 0.1	95.8 ± 0.3	98.9 ± 0.2
MMNet ^[26]	99.28 ± 0.08	99.77 ± 0.04	97.16 ± 0.10	98.93 ± 0.05
FTMN	99.7 ± 0.1	99.9 ± 0.1	98.3 ± 0.1	99.5 ± 0.1

表 3 在 miniImageNet 数据集上的小样本分类性能 (%)
Table 3 Few-shot classification performance on miniImageNet dataset (%)

模型	5-类	
	1-样本	5-样本
MN	43.40 ± 0.78	51.09 ± 0.71
ML-LSTM ^[11]	43.56 ± 0.84	55.31 ± 0.73
ProtoNet ^[12]	49.42 ± 0.78	68.20 ± 0.66
RN	50.44 ± 0.82	65.32 ± 0.70
MetaNet ^[16]	49.21 ± 0.96	—
MAML ^[18]	48.70 ± 1.84	63.11 ± 0.92
EGNN	—	66.85
EGNN + Transduction ^[22]	—	76.37
DN4 ^[24]	51.24 ± 0.74	71.02 ± 0.64
DC ^[25]	62.53 ± 0.19	78.95 ± 0.13
DC + IMP ^[25]	—	79.77 ± 0.19
MMNet ^[26]	53.37 ± 0.08	66.97 ± 0.09
PredictNet ^[27]	54.53 ± 0.40	67.87 ± 0.20
DynamicNet ^[28]	56.20 ± 0.86	72.81 ± 0.62
MN-FCE ^[29]	43.44 ± 0.77	60.60 ± 0.71
MetaOptNet ^[30]	60.64 ± 0.61	78.63 ± 0.46
FTMN	59.86 ± 0.91	75.96 ± 0.82
FTMN-R12	61.33 ± 0.21	79.59 ± 0.47

距离是否会对网络的性能产生影响, 将不对网络进行缩小类内距离的操作, 然后进行实验, 称该网络为 Trans*-M-4C. 本文还测试了仅使用余弦相似度或仅使用欧氏距离作为度量函数的结果, 如表 5 所示. 由表 5 可知, 本文设计的每个模块都对提升网络性能有一定影响, 并在组合后相辅相成, 得到了最好效果.

本文使用的特征变换和度量网络是基于度量的

小样本学习模型使用了元学习的方法, 与使用的基线 (原型网络) 相比, 本文模型添加了特征变换模块, 并在特征度量部分将余弦相似度和欧氏距离进行了结合; 在此基础上, 添加了一个特征度量函数. 特征变换模块使用聚类的方法对特征进行迁移变换, 并在整个流程中主要对样本的局部特征进行操作, 因此, 使用度量函数对局部的度量距离进行联合表达, 得到该样本的全局特征度量结果.

本文使用的模型训练方法是基于匹配网络的分段训练形式, 即一个阶段进行一次支持集和查询集的挑选和训练, 相当于元学习中的一次任务. 在每个阶段中, 支持集和查询集中的类别是随机选择的, 因此每次训练选择的类别均为新类别, 导致在训练过程中损失函数的波动较大, 但网络模型整体的准确率在稳步提升.

4 结束语

本文提出一种特征变换度量网络, 该网络模型由嵌入函数模块、特征变换网络模块和特征度量网络模块构成. 特征变换网络能够通过学习残差使特征向量发生偏移, 以此缩小同类支持样本在特征空间中的特征表示间隔, 并能改变不同特征类别中心的位置, 来达到增大特征类间差异的目的. 最后, 设计了新的特征度量距离和度量网络, 同时考虑余弦相似度和欧氏距离对特征度量的影响, 从角度和几何空间距离两种方式同时对特征进行度量, 并通过度量函数将局部对特征的度量距离联合地表达, 作为查询样本的最终度量结果. 本文将提出的模型在小样本通用数据集 Omniglot、miniImageNet、tieredImageNet、CUB-200 和 CIFAR-FS 上进行了测试, 测试结果取得了最优表现或仅次于最优模型的表现, 验证了本文模型的有效性和泛化性.

表 4 在 CUB-200、CIFAR-FS 和 tieredImageNet 数据集上的小样本分类性能 (%)
Table 4 Few-shot classification performance on CUB-200, CIFAR-FS and tieredImageNet datasets (%)

模型	CUB-200 5-类		CIFAR-FS 5-类		tieredImageNet 5-类	
	1-样本	5-样本	1-样本	5-样本	1-样本	5-样本
MN	61.16 ± 0.89	72.86 ± 0.70	—	—	—	—
ProtoNet ^[12]	51.31 ± 0.91	70.77 ± 0.69	55.5 ± 0.7	72.0 ± 0.6	53.31 ± 0.89	72.69 ± 0.74
RN	62.45 ± 0.98	76.11 ± 0.69	55.0 ± 1.0	69.3 ± 0.8	54.48 ± 0.93	71.32 ± 0.78
MAML ^[18]	55.92 ± 0.95	72.09 ± 0.76	58.9 ± 1.9	71.5 ± 1.0	51.67 ± 1.81	70.30 ± 1.75
EGNN	—	—	—	—	63.52 ± 0.52	80.24 ± 0.49
DN4 ^[24]	53.15 ± 0.84	81.90 ± 0.60	—	—	—	—
MetaOptNet ^[30]	—	—	72.0 ± 0.7	84.2 ± 0.5	65.99 ± 0.72	81.56 ± 0.53
FTMN-R12	69.58 ± 0.36	85.46 ± 0.79	70.3 ± 0.5	82.6 ± 0.3	62.14 ± 0.63	81.74 ± 0.33

表 5 消融实验结果 (%)
Table 5 Results of ablation study (%)

模型	5-类	
	1-样本	5-样本
ProtoNet-4C	49.42 ± 0.78	68.20 ± 0.66
ProtoNet-8C	51.18 ± 0.73	70.23 ± 0.46
ProtoNet-Trans-4C	53.47 ± 0.46	71.33 ± 0.23
ProtoNet-M-4C	56.54 ± 0.57	73.46 ± 0.53
ProtoNet-VLAD-4C	52.46 ± 0.67	70.83 ± 0.62
Trans*-M-4C	59.86 ± 0.91	67.86 ± 0.56
仅使用余弦相似度	54.62 ± 0.57	72.58 ± 0.38
仅使用欧氏距离	55.66 ± 0.67	73.34 ± 0.74
FTMN	59.86 ± 0.91	75.96 ± 0.82

References

- Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 1-9
- He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 770-778
- Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: NIPS, 2012. 1106-1114
- Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR, 2015.
- Liu Ying, Lei Yan-Bo, Fan Jiu-Lun, Wang Fu-Ping, Gong Yan-Chao, Tian Qi. Survey on image classification technology based on small sample learning. *Acta Automatica Sinica*, 2021, **47**(2): 297-315
(刘颖, 雷研博, 范九伦, 王富平, 公衍超, 田奇. 基于小样本学习的图像分类技术综述. *自动化学报*, 2021, **47**(2): 297-315)
- Miller E G, Matsakis N E, Viola P A. Learning from one example through shared densities on transforms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hilton Head Island, USA: IEEE, 2000. 464-471
- Li F F, Fergus R, Perona P. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, **28**(4): 594-611
- Lake B M, Salakhutdinov R, Gross J, Tenenbaum J B. One shot learning of simple visual concepts. In: Proceedings of the 33rd Annual Meeting of the Cognitive Science Society. Boston, USA: CogSci, 2011. 2568-2573
- Lake B M, Salakhutdinov R, Tenenbaum J B. Human-level concept learning through probabilistic program induction. *Science*, 2015, **350**(11): 1332-1338
- Edwards H, Storkey A J. Towards a neural statistician. In: Proceedings of the 5th International Conference on Learning Representations. Toulon, France: ICLR, 2017.
- Vinyals O, Blundell C, Lillierap T, Kavukcuoglu K, Wierstra D. Matching networks for one shot learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: 2016. 3637-3645
- Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: Proceedings of the 31th International Conference on Neural Information Processing Systems. Long Beach, USA: 2017. 4080-4090
- Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR, 2015.
- Sung F, Yang Y X, Zhang L, Xiang T, Torr P H S, Hospedales T M. Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 1199-1208
- Kaiser L, Nachum O, Roy A, Bengio S. Learning to remember rare events. In: Proceedings of the 5th International Conference on Learning Representations. Toulon, France: ICLR, 2017.
- Munkhdalai T, Yu H. Meta networks. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: JMLR.org, 2017. 2554-2563
- Santoro A, Bartunov S, Botvinick M, Wierstra D, Lillicrap T. Meta-learning with memory-augmented neural networks. In: Proceedings of the 33rd International Conference on Machine Learning. New York, USA: PMLR, 2016. 1842-1850
- Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: JMLR.org, 2017. 1126-1135
- Arandjelovic R, Gronat P, Torii A, Pajdla T, Sivic J. NetVLAD: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 5297-5307
- Jégou H, Douze M, Schmid C, Pérez P. Aggregating local descriptors into a compact image representation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE, 2010. 3304-3311
- Bertinetto L, Henriques J F, Torr P H, Vedaldi A. Meta-learning with differentiable closed-form solvers. In: Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: ICLR, 2019.
- Kim J, Kim T, Kim S, Yoo C D. Edge-labeling graph neural network for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 11-20
- Yue Z Q, Zhang H W, Sun Q R, Hua X S. Interventional few-shot learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Incorporated, 2020. Article No. 230
- Li W B, Wang L, Xu J L, Huo J, Gao Y, Luo J B. Revisiting local descriptor based image-to-class measure for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 7253-7260
- Lifchitz Y, Avrithis Y, Picard S, Bursuc A. Dense classification and implanting for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 9250-9259
- Cai Q, Pan Y W, Yao T, Yan C G, Mei T. Memory matching networks for one-shot image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 4080-4088
- Qiao S Y, Liu C X, Shen W, Yuille A L. Few-shot image recognition by predicting parameters from activations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 7229-7238
- Gidaris S, Komodakis N. Dynamic few-shot visual learning without forgetting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 4367-4375

- 29 Ravi S, Larochelle H. Optimization as a model for few-shot learning. In: Proceedings of the 5th International Conference on Learning Representations. Toulon, France: ICLR, 2017.
- 30 Lee K, Maji S, Ravichandran A, Soatto S. Meta-learning with differentiable convex optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 10649–10657



王多瑞 2021 年获得中国科学技术大学硕士学位. 主要研究方向为小样本学习, 目标检测.

E-mail: wangduor@mail.ustc.edu.cn

(WANG Duo-Rui He received his master degree from University of Science and Technology of China in

2021. His research interest covers few-shot learning and object detection.)



杜 杨 2019 年获得中国科学院自动化研究所博士学位. 主要研究方向为行为识别, 医学图像处理.

E-mail: jingzhou.dy@alibaba-inc.com

(DU Yang He received his Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences.

His research interest covers action recognition and medical image processing.)



董兰芳 中国科学技术大学副教授. 1994 年获得中国科学技术大学硕士学位. 主要研究方向为图像与视频智能分析, 知识图谱与对话系统, 数值模拟与三维重建.

E-mail: lfdong@ustc.edu.cn

(DONG Lan-Fang Associate pro-

fessor at University of Science and Technology of China. She received her master degree from University of Science and Technology of China in 1994. Her research interest covers image and video intelligent analysis, knowledge mapping and dialogue systems, and numerical simulation and 3D reconstruction.)



胡卫明 中国科学院自动化研究所研究员. 1998 年获得浙江大学博士学位. 主要研究方向为视觉运动分析, 网络不良信息识别和网络入侵检测. 本文通信作者.

E-mail: wmhu@nlpr.ia.ac.cn

(HU Wei-Ming Professor at the

Institute of Automation, Chinese Academy of Sciences. He received his Ph.D. degree from Zhejiang University in 1998. His research interest covers visual motion analysis, recognition of web objectionable information, and network intrusion detection. Corresponding author of this paper.)



李 兵 中国科学院自动化研究所研究员. 2009 年获得北京交通大学博士学位. 主要研究方向为网络内容安全, 智能图像信号处理.

E-mail: bing.li@ia.ac.cn

(LI Bing Professor at the Institute of Automation, Chinese

Academy of Sciences. He received his Ph.D. degree from Beijing Jiaotong University in 2009. His research interest covers the web content security and intelligent image signal process.)