

# 小样本图像分类的注意力全关系网络

李晓旭<sup>1)</sup> 刘忠源<sup>1)</sup> 武继杰<sup>1)</sup> 曹洁<sup>1)</sup> 马占宇<sup>2)</sup>

<sup>1)</sup>(兰州理工大学计算机与通信学院 兰州 730050)

<sup>2)</sup>(北京邮电大学人工智能学院模式识别与智能系统实验室 北京 100876)

**摘要** 传统的基于深度学习的图像分类方法在大样本分类任务中具有较好的分类效果,但在小样本分类任务中却存在较大的挑战,为此,小样本图像分类获得了研究人员的广泛关注.基于度量的方法是解决小样本图像分类的一种简单有效方法,它利用可学习的映射函数将分类任务中的所有样本映射到一个特征空间中,然后基于某种度量标准对查询特征进行分类.由于分类任务中不同类的两个图像有可能包含较多的相似性区域,导致特征空间中某些查询特征与异类的类原型特征的距离较近,较难学习到大的分类边界.为了解决上述问题,本文提出了注意力全关系网络(Total Relation Network with Attention, TRNA),该网络通过计算特征对的全关系和特征对的注意力来实现大边界的特征空间.具体地,在计算出所有的查询特征和类原型后,提出的网络利用特征对全关系拼接操作将特征空间中的任意两个特征在通道方向上进行拼接得到特征对矩阵,然后利用特征对注意力机制将特征对矩阵中不同类间难区分的特征对挑选出来并给予大的权重,最后将特征对矩阵输入卷积网络和全连接网络得到一个相似得分矩阵.实验结果表明本文的方法与关系网络相比,在数据集 mini-ImageNet、Stanford-Dogs、Stanford-Cars、CUB-200-2011的1-shot和5-shot分类任务中分别有2.67%和1.71%、8.31%和3.92%、14.99%和8.00%、4.41%和4.42%的性能提升.

**关键词** 小样本图像分类;基于度量的方法;类原型;注意力机制;大边界学习

**中图法分类号** TP391 **DOI号** 10.11897/SP.J.1016.2023.00371

## Total Relation Network with Attention for Few-Shot Image Classification

LI Xiao-Xu<sup>1)</sup> LIU Zhong-Yuan<sup>1)</sup> WU Ji-Jie<sup>1)</sup> CAO Jie<sup>1)</sup> MA Zhan-Yu<sup>2)</sup>

<sup>1)</sup>(School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050)

<sup>2)</sup>(Pattern Recognition and Intelligent System Laboratory, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876)

**Abstract** With the continuous development of deep learning, image classification methods based on deep learning have achieved excellent classification performance in large sample classification tasks, but face significant challenges in few-shot classification tasks. It is difficult to obtain a large number of labeled samples to train deep learning models in many real-world scenarios. This means that it is of great practical importance to improve the classification performance of deep learning-based image classification methods in few-shot classification tasks. To this end, a growing number of researchers are focusing on few-shot image classification, which aims to complete the classification of unlabeled query samples based on a small number of labeled support samples, that is, to learn new category concepts through a small number of labeled samples. The metric-based method is simple yet effective method for solving the few-shot image classification. It uses

收稿日期:2021-11-02;在线发布日期:2022-10-10. 本课题得到国家自然科学基金(62176110,62111530146,61906080,61922015,U19B2036,62225601)、甘肃省青年博士基金(2021QB-038)、北京市自然科学基金(Z200002)、兰州理工大学红柳杰出青年基金资助. 李晓旭,博士,教授,中国计算机学会(CCF)会员,主要研究领域为机器学习基础,重点是图像和视频理解的应用. E-mail: lixiaoxu@lut.edu.cn. 刘忠源,硕士研究生,主要研究方向为机器学习和小样本学习. 武继杰,博士研究生,主要研究方向为机器学习和小样本学习. 曹洁,博士,教授,主要研究领域为机器学习、模式识别、语音和说话人识别、信息融合和计算机视觉. 马占宇(通信作者),博士,教授,中国计算机学会(CCF)高级会员,国家杰出青年科学基金入选者,主要研究领域为模式识别和机器学习基础,重点是计算机视觉、多媒体信号处理和数据挖掘方面的应用. E-mail: mazhanyu@bupt.edu.cn.

the learnable mapping function to map all samples in a few-shot classification task to feature space and then classifies the query features according to some metric standard. However, two images of different classes in the classification task may contain more similar regions, so there may be situations in the feature space where the distances of certain query features and heterogeneous class prototypes are closer to each other, resulting in few-shot image classification networks that are more difficult to learn large classification margin. In other words, it is more difficult to have a clear classification margin between feature clusters of different classes and significant compactness between features in feature clusters of the same class in the feature space. To solve this problem, this paper proposes a Total Relation Network with Attention (TRNA), which realizes the feature space with a large margin by calculating the Total Relation of Feature-Pair and the Attention of Feature-Pair. Specifically, after calculating all the query features and class prototypes, the proposed network uses the Total Relation Concatenation Operation of Feature-Pair to concatenate any two features in the feature space in the channel direction to obtain the feature pair matrix and then uses the Attention Mechanism of Feature-Pair to select the feature pairs that are difficult to distinguish between different classes in the feature pair matrix and give them large weights. Finally, the feature pair matrix is fed into a convolutional network and a fully connected network to obtain a similarity score matrix. We conduct 5-Way 1-shot and 5-Way 5-shot experiments on the commonly used few-shot image classification dataset mini-ImageNet and three few-shot fine-grained image classification datasets Stanford-Dogs, Stanford-Cars and CUB-200-2011. The experimental results show that compared with the Relation Network (RN), the performance of the proposed method is improved by 2.67% and 1.71%, 8.31% and 3.92%, 14.99% and 8.00%, 4.41% and 4.42%, respectively, on the 1-shot and 5-shot classification tasks of mini-ImageNet, Stanford-Dogs, Stanford-Cars and CUB-200-2011, which shows the obvious effectiveness of our method. Furthermore, ablation experiments demonstrate that both the Total Relation Concatenation module and the Attention Mechanism module in the proposed TRNA play a key role in improving classification performance.

**Keywords** few-shot image classification; metric-based learning; class prototype; attention mechanism; large margin learning

## 1 引 言

传统的基于深度学习的图像分类方法<sup>[1-3]</sup>在图像分类任务上的识别精度已经超越人类,但这些方法通常需要使用大量的带标签图像对网络进行训练,而现实世界中标注数据的获取往往代价昂贵.相反,人类通过很少的实例样本便可以学会对新类的识别,例如人类只需要看几张猫的图片就可以认识猫,即使猫的颜色、大小和姿态等与所看的图片有较大的差异,也毫不影响识别效果.受人类学习模式的启发,“小样本学习”(Few-Shot Learning, FSL)<sup>[4-7]</sup>被提出,用于解决小样本场景下分类任务的挑战,比如,森林中小样本鸟类的识别问题:有些鸟类的标注数据稀缺,称为测试类别,例如每类仅有 1 个或 5 个

标注样本.除了测试类别仅有的少量标记样本之外,常常还可获得其它一些鸟类的标注图像,问题是如何利用这些已知信息,对测试类别中的未标注图像进行预测.

为了解决小样本图像分类,一些基于度量的方法被提出,这类方法通常由一个特征提取器和一个度量模块组成,特征提取器一般由卷积神经网络组成,而度量模块由某种度量函数组成,例如,余弦距离、欧式距离等.对于给定的一个分类任务,特征提取器为分类任务中的所有图像生成特征表示,然后利用度量模块计算出查询特征和所有类原型之间距离,最后依据距离对查询样本做出分类选择.这种基于度量的方法相对简洁高效,受到了越来越多研究者的关注.然而,分类任务中不同类的两个图像间可能具有较多的相似区域,故特征空间中可能会出现

某些查询特征和异类类原型的距离较为接近的情况,如图 1(a)所示,导致小样本分类网络较难实现大边界的特征空间.换言之,较难让特征空间中不同类特征簇之间具有明确的分类边界,且同一类特征簇中的特征之间具有明显的紧凑性,如图 1(c)所示.

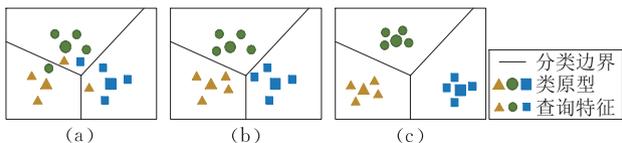


图 1 动机图((a)图中的一些查询特征和异类类原型之间的距离较为接近,不同类之间存在难区分的特征对,导致特征空间中并没有明显的类间可区分性和类内紧凑性.提出特征对注意力机制,使网络重点关注不同类之间难区分特征对的相似性学习,以增大类间可区分性,如(b)图.提出特征对全关系拼接,使不同类之间的距离最大化,进一步增大类间可区分性和类内紧凑性,如(c)图)

为了解决上述问题,本文在关系网络(Relation Network)<sup>[8]</sup>的基础上提出了注意力全关系网络(Total Relation Network with Attention, TRNA),该网络利用特征对全关系拼接操作和特征对注意力机制来实现大边界分类.本文的主要贡献有如下三点:

(1) 提出了特征对全关系拼接,即在分类时,将分类任务中的任意两个特征在通道方向上进行拼接并计算它们之间的相似性,以使特征空间中不同类之间的距离最大化.

(2) 提出在特征对上引入注意力机制,该机制自适应地为分类任务中的每个特征对计算一个注意力权重并执行加权操作,使网络重点关注不同类之间难区分特征对的相似性学习.

(3) 在小样本图像分类数据集 mini-ImageNet、Stanford-Dogs、Stanford-Cars 和 CUB-200-2011 上,本文提出的注意力全关系网络的分类性能相比于原始关系网络具有明显的提升,且相比于现有的最佳小样本图像分类方法,也具有一定的竞争力.

## 2 相关工作

### 2.1 小样本图像分类方法

现有的小样本图像分类方法主要包括基于迁移学习的方法、基于元学习的方法和基于度量的方法:

(1) 基于迁移学习的方法<sup>[9-12]</sup>. 迁移学习方法认为源域和目标域之间存在相互关联的知识,在源域数据上学习到的知识可直接应用于相关的目标域中. Sun 等人<sup>[9]</sup>提出了一种基于迁移学习的算法:元迁移学习(Meta-Transfer Learning, MTL),该算法

在网络训练和测试阶段使用共享的参数,降低了模型学习的难度. Ye 等人、Zhang 等人<sup>[10-11]</sup>在训练过程中使用迁移的预训练参数,极大的减小了模型训练的代价,同时达到了让模型快速适应于小样本数据的目的. Hinton 等人<sup>[12]</sup>通过引入教师网络来诱导学生网络进行学习,实现知识迁移. 这类方法通过知识共享可使小样本分类模型获得较好的性能,并能减少构建模型所需的训练数据和计算成本. 但是,当源域与目标域的数据分布相差较大时,迁移学习将不能提高目标域的分类性能.

(2) 基于元学习的方法<sup>[13-17]</sup>. 元学习(Meta Learning)又叫“学会如何学习”,它是针对传统神经网络模型泛化能力不足、对新类任务适应性较差等问题而提出的. 元学习的训练和测试过程可以想象成人类在掌握一些基础知识后能够将这些基础知识迅速应用于新的任务中,这与小样本学习的过程十分相似,因此元学习的方法经常被用于小样本学习的问题中. Ravi 等人<sup>[13]</sup>利用一种基于长短期记忆网络(Long Short-Term Memory, LSTM)的元学习器来学习精确的优化算法,以用于训练小样本学习中的分类器. Finn 等人<sup>[14]</sup>提出了模型无关的元学习算法(Model-Agnostic Meta-Learning, MAML),该算法为模型学习一个全局最优的初始化参数,模型在这个初始化参数的基础上经过一步或多步随机梯度下降即可快速适应新任务. Rusu 等人<sup>[15]</sup>在 MAML<sup>[14]</sup>的基础上提出了一种基于参数优化的元学习算法:潜在嵌入优化(Latent Embedding Optimization, LEO),该算法学习模型参数的潜在生成表示,并在此低维潜在空间中执行基于梯度的元学习. Sun 等人<sup>[16]</sup>引入少量的随机变量作为潜在类原型,用平摊推理的方式来学习这些潜在原型的后验分布,然后通过几步随机梯度下降来生成新任务上潜在原型的近似后验分布. Li 等人<sup>[17]</sup>提出了一种类似随机梯度下降的元学习器(Meta-SGD),该元学习器同时对初始化参数、学习率和更新方向进行学习,训练得到的模型经过微调后可以容易地适应新的任务. 这类方法在任务分布更广泛的情况下可以获得较优性能,但需要为每个任务优化基类学习器,导致计算成本昂贵<sup>[18]</sup>.

(3) 基于度量的方法<sup>[19-23]</sup>. 基于度量的方法将分类任务中的所有图像映射到一个特征表示空间中,然后通过比较查询特征和类原型之间的距离或相似性来完成小样本图像分类. 匹配网络(Matching Network)<sup>[19]</sup>使用余弦距离对样本进行分类. 原型网络(Prototypical Network)<sup>[20]</sup>使用欧式距离来度量

查询特征与类原型之间的距离,并选择距离最近的类原型的类别作为查询样本的预测类别. Liu 等人<sup>[21]</sup>认为原型网络<sup>[20]</sup>中生成的类原型与真实的类原型之间存在偏差,提出利用未标记的查询样本对类原型进行矫正,最后利用矫正后的类原型来完成查询样本的分类. 关系网络(Relation Network)<sup>[8]</sup>没有采用某个固定的度量函数作为度量标准,而是利用卷积神经网络来自适应地度量样本之间的相似关系. Li 等人<sup>[22]</sup>提出了一种双相似性的度量网络(Bi-Similarity Network, BSNet),试图从两个不同的角度对样本进行度量. Zhang 等人<sup>[23]</sup>提出了一种新的距离度量:推土距离(Earth Mover's Distance, EMD)度量,该度量通过计算查询集和支持集图像的各个图块之间的最佳匹配代价来确定图像的相关性. 这类方法的优点是训练模型不需要针对测试任务进行调整,但是当测试与训练任务距离较远时,效果不佳.

本文提出的方法建立在关系网络<sup>[8]</sup>基础上,与关系网络不同的是,我们提出了特征对注意力模块以及特征对全关系拼接模块,充分利用数据信息,以致获得大的分类边界.

## 2.2 注意力机制

当人的大脑接收到外部信息,如视觉信息、听觉信息时,往往不会对全部信息进行处理和理解,而只会将注意力集中在部分显著或者感兴趣的信息上,这样有助于滤除不重要的信息,提升信息处理的效率. 为了让图像处理系统在处理图像时能够像人类一样忽略图像特征中的无关信息而只关注重点信息,注意力机制的概念被提出并应用于图像处理任务中.

目前的注意力机制主要包括软注意力机制和硬注意力机制. 软注意力机制主要关注区域或者通道,并用 0 到 1 的数值来表示每个区域被关注的程度,是一种确定性的注意力机制;硬注意力机制更加关注点,图像中的每个点都有可能延伸出注意力,是一个随机的预测过程. 目前大多数的图像处理系统采用软注意力机制<sup>[24-27]</sup>. Jaderberg 等人<sup>[24]</sup>认为传统卷积神经网络中的池化操作没有真正地实现模型的空间不变性,故提出了一个空间变换网络(Spatial Transform Network),该网络不需要关键点的标注便能够自适应地学到对于不同数据的空间变换方式. Wang 等人<sup>[25]</sup>提出了一种使用注意力机制的卷积神经网络:残差注意力网络(Residual Attention Network),该网络采用基于注意力的残差方式来进

行学习,可以得到更加丰富和关键的特征表示,另外该网络能够在一个前向过程中就提取出模型的注意力,而不需要新增一个分支来提取注意力,使得模型训练更加简单. Woo 等人<sup>[26]</sup>为前馈卷积神经网络提出了一种简单且高效的注意力模块:卷积块注意力模块(Convolutional Block Attention Module, CBAM),CBAM 对输入的特征图沿着通道(Channel)和空间(Spatial)两个单独的维度依次推断注意力图,然后将注意力图和输入特征图相乘以进行自适应特征细化. Hu 等人<sup>[27]</sup>提出了一个经典的通道注意力方法:压缩-激励块(Squeeze-and-Excitation Block),该模块通过显式建模通道之间的相互依赖来自适应地重新校准通道的特征响应.

为了提升小样本图像分类网络的分类性能,目前一些小样本图像分类方法也引入了注意力机制<sup>[28-32]</sup>. 空间-任务注意力网络(Spatial-Task Attention Network, STANet)<sup>[28]</sup>利用空间注意力来定位相关对象区域并利用任务注意力选择相似的训练数据进行标签预测. Ren 等人<sup>[29]</sup>提出了一种基于注意力的元学习方法:注意力吸引网络(Attention Attractor Networks),该网络对新类的学习起到了正则化作用. 位置感知关系网络(Position-Aware Relation Network, PARN)<sup>[30]</sup>使用双重相关注意力机制提取位置已知的细粒度特征,以解决由于位置偏移和细粒度特征不一致引起的相似性较差问题. 语义对齐度量学习网络(Semantic Alignment Metric Learning, SAML)<sup>[31]</sup>应用基于激活的注意力机制<sup>[33]</sup>来挑选语义相关的局部特征并对它们赋予更大的权重,从而达到将目标对象对齐的目的. 矫正度量传播网络(Rectified Metric Propagation, ReMP)<sup>[32]</sup>将类原型和所有查询特征的相似度组成注意力图,以对类原型进行矫正.

上述小样本图像分类方法使用的注意力图是基于不同图像之间局部或全局特征的余弦相似度或者内积生成的. 与现有的注意力机制方法不同的是,本文的注意力机制是基于特征对的注意力,目的是使网络重点关注不同类之间难区分特征对的相似性学习.

## 3 提出的方法

本节首先介绍小样本图像分类的问题定义,接着介绍关系网络,然后介绍提出的注意力全关系网络,最后对提出的网络进行分析与讨论.

### 3.1 问题定义

为了方便阐述,现给出小样本图像分类的问题定义.小样本图像分类数据集通常由训练集、验证集和测试集组成,我们将这三个数据集分别记为

$$\mathcal{D}_{\text{train}} = \{(X_i, Y_i), Y_i \in \ell_{\text{train}}\}_{i=1}^{N_{\text{train}}} \quad (1)$$

$$\mathcal{D}_{\text{val}} = \{(X_j, Y_j), Y_j \in \ell_{\text{val}}\}_{j=1}^{N_{\text{val}}} \quad (2)$$

$$\mathcal{D}_{\text{test}} = \{(X_z, Y_z), Y_z \in \ell_{\text{test}}\}_{z=1}^{N_{\text{test}}} \quad (3)$$

$X_i, X_j, X_z$  分别代表训练集、验证集和测试集中的第  $i$ 、第  $j$  和第  $z$  张图像,  $Y_i, Y_j, Y_z$  代表对应的类别标签;  $N_{\text{train}}, N_{\text{val}}, N_{\text{test}}$  分别代表训练集、验证集和测试集中样本的总数;  $\ell_{\text{train}}, \ell_{\text{val}}, \ell_{\text{test}}$  分别代表训练集、验证集和测试集中样本类别的标签集合,且这三个类别标签集合是不相交的,即:

$$\{\ell_{\text{train}} \cap \ell_{\text{val}}, \ell_{\text{train}} \cap \ell_{\text{test}}, \ell_{\text{val}} \cap \ell_{\text{test}}\} = \emptyset \quad (4)$$

为了模拟真实场景,小样本图像分类中所有的训练、验证和测试均以基于任务的方式来实现<sup>[19]</sup>,每一个任务包含一个支持集和一个查询集.从训练集  $\mathcal{D}_{\text{train}}$  中生成任务  $\xi^{(u)}$  的过程是:首先从训练集对应的类别标签集合  $\ell_{\text{train}}$  中随机采样  $C$  个类别,接着从每个类别的样本中随机采样  $D$  个样本,最后从每类的  $D$  个样本中随机抽取  $K$  个样本作为支持集样本,剩余的  $D-K$  个样本作为查询集样本.我们将分类任务  $\xi^{(u)}$  中的支持集和查询集分别记为

$$S^{(u)} = \{(X_s, Y_s), Y_s \in \zeta\}_{s=1}^{C \times K} \quad (5)$$

$$Q^{(u)} = \{(\hat{X}_q, \hat{Y}_q), \hat{Y}_q \in \zeta\}_{q=1}^{C \times (D-K)} \quad (6)$$

$X_s, \hat{X}_q$  分别代表支持集和查询集中的第  $s$  和第  $q$  张图片,  $Y_s$  和  $\hat{Y}_q$  分别代表对应的类别标签;  $\zeta$  代表支持集或者查询集中样本类别的标签集合.这里采样到的分类任务  $\xi^{(u)}$  即为小样本图像分类中的一个  $C$ -Way  $K$ -shot 分类问题,依据分类问题中的支持集  $S^{(u)}$  将查询集  $Q^{(u)}$  中的每个查询样本  $\hat{X}_q$  正确分类是小样本图像分类的目的.

### 3.2 关系网络

关系网络<sup>[8]</sup>是一个经典的基于度量的小样本图像分类方法,它采用卷积神经网络和全连接层来度

量两个样本特征之间的相似性,这种可训练的动态度量机制可以适应不同任务之间的差异,具有较大的灵活性.关系网络由嵌入模块和关系模块组成,嵌入模块的作用是为分类任务中的支持样本和查询样本提取特征表示并依据支持特征计算类原型,关系模块的作用是为特征对计算相似得分.利用关系网络为查询集  $Q^{(u)}$  中的每一个查询样本  $\hat{X}_q$  预测类别的流程如下:

**嵌入模块  $f_\theta$ .**对于随机采样生成的分类任务  $\xi^{(u)}$ ,它里面的查询样本  $\hat{X}_q$  和支持样本  $X_s$  输入嵌入模块  $f_\theta$  后得到相应的查询特征和支持特征:

$$T_q = f_\theta(\hat{X}_q) \in R^{V \times H \times W}, q \in \{1, \dots, C \times (D-K)\} \quad (7)$$

$$E_s = f_\theta(X_s) \in R^{V \times H \times W}, s \in \{1, \dots, C \times K\} \quad (8)$$

$V, H, W$  分别代表支持特征或查询特征三个维度的大小.当  $K > 1$ ,即支持集中每类样本的数量大于 1 时,将属于同一类的所有支持特征按位置求平均,得到支持集中每类样本的类原型  $O_c$ :

$$O_c = \frac{1}{K} \sum_{k=1}^K f_\theta(X_{c,k}) \in R^{V \times H \times W}, c \in \{1, \dots, C\} \quad (9)$$

其中,  $X_{c,k}$  代表第  $c$  个支持类中的第  $k$  个样本,当  $K=1$  时,  $O_c = f_\theta(X_{c,1})$ .

**特征拼接.**在利用关系模块确定查询特征  $T_q$  和类原型  $O_c$  的相似度之前,先将  $T_q$  和每一个类原型  $O_c$  沿通道方向上拼接,得到初始特征对  $P_{T_q, O_c}$ :

$$P_{T_q, O_c} = [T_q \| O_c] \in R^{2V \times H \times W} \quad (10)$$

“ $\|$ ”代表拼接操作.

**关系模块  $g_\phi$ .**将拼接得到的特征对  $P_{T_q, O_c}$  送入关系模块得到查询特征  $T_q$  和所有类原型  $O_c$  的相似度,选择相似度最大的类原型的类别作为查询样本  $\hat{X}_q$  的预测类别.

### 3.3 注意力全关系网络

本文在关系网络的基础上提出了注意力全关系网络.如图 2 所示,提出的注意力全关系网络由嵌入模块  $f_\theta$ 、特征对全关系拼接模块  $\psi$ 、特征对注意力模块  $h_\gamma$  以及关系模块  $g_\phi$  组成.嵌入模块为分类任务中

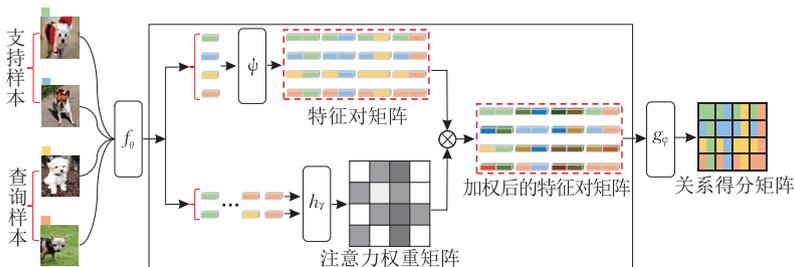


图 2 注意力全关系网络结构图(图中展示了 2-Way 1-shot 小样本图像分类流程,其中提取特征的颜色与输入样本的颜色保持一致)

的所有图像提取特征表示并依据支持特征计算类原型;特征对全关系拼接模块将任意两个特征在通道方向上拼接;特征对注意力模块为每一个特征对计算一个注意力权重;关系模块为每一个加权后的特征对计算关系得分。

### 3.3.1 特征对全关系拼接

为了让特征空间中不同类之间的距离最大化,我们在关系网络中加入了特征对全关系拼接操作,即在特征拼接时,将分类任务中的任意两个特征在通道方向上进行拼接,得到特征对  $P_{O_{c_1}, O_{c_2}}, P_{O_c, T_q}, P_{T_q, O_c}, P_{T_{q_1}, T_{q_2}}$ :

$$P_{O_{c_1}, O_{c_2}} = [O_{c_1} \parallel O_{c_2}] \in R^{2V \times H \times W} \quad (11)$$

$$P_{O_c, T_q} = [O_c \parallel T_q] \in R^{2V \times H \times W} \quad (12)$$

$$P_{T_q, O_c} = [T_q \parallel O_c] \in R^{2V \times H \times W} \quad (13)$$

$$P_{T_{q_1}, T_{q_2}} = [T_{q_1} \parallel T_{q_2}] \in R^{2V \times H \times W} \quad (14)$$

其中,  $T_q, T_{q_1}, T_{q_2}$  均表示查询特征,且  $q, q_1, q_2$  的取值范围均为  $\{1, \dots, C \times (D-K)\}$ ;  $O_c, O_{c_1}, O_{c_2}$  均表示类原型,且  $c, c_1, c_2$  的取值范围均为  $\{1, \dots, C\}$ 。并将这些特征对按顺序组合得到特征对矩阵  $\mathbf{M}$ :

$$\mathbf{M} = \begin{pmatrix} P_{O_1, O_1} & \cdots & P_{O_1, T_{C \times (D-K)}} \\ \vdots & \ddots & \vdots \\ P_{T_{C \times (D-K)}, O_1} & \cdots & P_{T_{C \times (D-K)}, T_{C \times (D-K)}} \end{pmatrix} \quad (15)$$

矩阵  $\mathbf{M}$  中行列的数量一样,记为  $n$ ,且  $n = C + (C \times (D-K))$ ;假设一个分类任务中有 5 类样本,每类中有 1 个支持样本和 15 个查询样本,即  $C=5, D=16, K=1$ ,此时  $n=80$ 。

我们的全关系拼接操作类似于图神经网络(Graph Neural Network, GNN)<sup>[34]</sup>将图中的任意两个特征结点相互连接。不同的是,我们的全关系拼接实际是通过充分利用特征信息实现了不同类之间距离的最大化,而 GNN 利用这种全连接的方式对图结点进行循环更新,以实现将支持样本的标签信息传递到无标签的查询样本上。

### 3.3.2 特征对注意力机制

为了让网络重点关注不同类之间难区分特征对的相似性学习,我们在关系网络的嵌入模块和关系模块之间增加了一个由两个卷积块、两个最大池化层和一个余弦度量层组成的特征对注意力模块。余弦度量层用于为特征对矩阵  $\mathbf{M}$  中的每一个特征对计算一个特征对注意力权重,以突出特征对矩阵中不同类之间难区分的特征对。另外,为了增加模型弹性,我们在余弦度量之前加入了两个卷积块。

**特征对注意力模块  $h_r$ .** 假设  $P_{A,B}$  是特征对矩阵  $\mathbf{M}$  中的一个特征对,将组成特征对  $P_{A,B}$  的两个特征  $A$  和  $B$  分别输入注意力模块  $h_r$  的卷积块中,得到新的特征  $A'$  和  $B'$ ,然后计算  $A'$  和  $B'$  的余弦距离  $d_{A',B'}$ :

$$d_{A',B'} = \cos(A', B') = \frac{A' \cdot B'}{\|A'\| \times \|B'\|} \in [0, 1] \quad (16)$$

“ $\cdot$ ”表示内积操作,  $\|\cdot\|$  表示取向量模的操作;另外,卷积块中 ReLU 激活函数的存在使得卷积块的输出特征  $A'$  和  $B'$  均在第一象限,故  $d_{A',B'}$  的取值范围为  $[0, 1]$ 。

将  $d_{A',B'}$  作为特征对  $P_{A,B}$  的注意力权重  $\omega_{A,B}$ ,得到加权后的特征对  $P'_{A,B}$ :

$$P'_{A,B} = \omega_{A,B} \times P_{A,B} \quad (17)$$

特征对  $P'_{A,B}$  用于送入关系模块  $g_\varphi$  得到关系得分  $r_{A,B}$ :

$$r_{A,B} = g_\varphi(P'_{A,B}) \quad (18)$$

依据上述描述可知 3.3.1 节中的特征对矩阵  $\mathbf{M}$  经过特征对注意力模块后可得到加权的特征对矩阵  $\mathbf{M}'$ :

$$\mathbf{M}' = \begin{pmatrix} P'_{O_1, O_1} & \cdots & P'_{O_1, T_{C \times (D-K)}} \\ \vdots & \ddots & \vdots \\ P'_{T_{C \times (D-K)}, O_1} & \cdots & P'_{T_{C \times (D-K)}, T_{C \times (D-K)}} \end{pmatrix} \quad (19)$$

$\mathbf{M}'$  用于送入关系模块得到关系得分矩阵  $\mathbf{G}$ :

$$\mathbf{G} = g_\varphi(\mathbf{M}') = \begin{pmatrix} r_{O_1, O_1} & \cdots & r_{O_1, T_{C \times (D-K)}} \\ \vdots & \ddots & \vdots \\ r_{T_{C \times (D-K)}, O_1} & \cdots & r_{T_{C \times (D-K)}, T_{C \times (D-K)}} \end{pmatrix} \quad (20)$$

### 3.3.3 注意力全关系网络的训练和测试

**模型训练.** 本文使用均方误差损失来计算关系得分与真值之间的损失  $l_{A,B}$ :

$$l_{A,B} = (r_{A,B} - y_{A,B})^2 \quad (21)$$

$r_{A,B}$  表示特征  $A$  和  $B$  的关系得分;  $y_{A,B}$  代表真值,如果  $A, B$  同类,则  $y_{A,B}$  为 1, 否则  $y_{A,B}$  为 0。在训练阶段,利用式(21)将 3.3.2 节中的关系得分矩阵  $\mathbf{G}$  转化为损失矩阵  $\mathbf{L}$ :

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_1 & \mathbf{L}_2 \\ \mathbf{L}_3 & \mathbf{L}_4 \end{pmatrix} \quad (22)$$

$\mathbf{L}_1, \mathbf{L}_2, \mathbf{L}_3, \mathbf{L}_4$  分别对应为特征对  $P_{O_1, O_1}, P_{O_c, T_q}, P_{T_q, O_c}, P_{T_{q_1}, T_{q_2}}$  处的损失矩阵;然后利用损失矩阵  $\mathbf{L}$  求得总损失  $F_{loss}$ :

$$F_{loss} = \sum_{i=1}^4 \sum_{l_{A,B} \in L_i} l_{A,B} \quad (23)$$

网络的具体训练流程可见算法 1。

**算法 1.** 注意力全关系网络的训练流程.

输入:  $\theta_0, \gamma_0, \varphi_0, \mathcal{D}_{\text{train}}, N_{\text{episode}}$  (任务量)

输出:  $\theta, \gamma, \varphi$

1.  $\theta, \gamma, \varphi \leftarrow \theta_0, \gamma_0, \varphi_0$  //模型参数初始化
2. FOR  $u$  FROM 1 to  $N_{\text{episode}}$  DO
3. 从  $\mathcal{D}_{\text{train}}$  中随机采样一个任务  $\xi^{(u)} = \{S^{(u)}, Q^{(u)}\}$
4. 查询特征:  $T_q$ , 类原型:  $O_c \leftarrow f_{\theta}(\xi^{(u)})$
5. 特征对矩阵  $M \leftarrow \psi(T_q, O_c)$
6. 加权的特征对矩阵  $M' \leftarrow h_{\gamma}(M)$
7. 关系得分矩阵  $G \leftarrow g_{\varphi}(M')$
8.  $L \leftarrow 0$  //初始化第  $u$  个分类任务的损失矩阵
9.  $F_{\text{loss}} \leftarrow 0$  //初始化总损失
10. FOR  $r_{A,B}$  IN  $G$  DO //  $r_{A,B}$  表示特征  $A$  和特征  $B$  的关系得分
11.  $l_{A,B} = (r_{A,B} - y_{A,B})^2$  //利用均方误差求损失
12. 将  $l_{A,B}$  加入  $L$
13. END FOR
14.  $F_{\text{loss}} \leftarrow$  式(23) //更新总损失
15.  $\theta, \gamma, \varphi \leftarrow \arg\min_{\theta, \gamma, \varphi} F_{\text{loss}}$  //更新网络参数
16. END FOR

**模型测试.** 利用训练阶段学习到的最佳模型来对测试集上的分类任务进行测试. 假设在训练集上学习到的最佳模型为  $\mathfrak{S}(\cdot | \eta)$ , 那么依据支持集  $S^{(u)}$  来预测查询集中查询图像  $\hat{X}_q$  类别的过程可公式化为

$$\hat{Y}_q = \mathfrak{S}(\hat{X}_q, S^{(u)} | \eta), \eta = \{\theta, \gamma, \varphi\} \quad (24)$$

其中,  $\mathfrak{S}$  表示模型,  $\eta$  表示模型所有参数的集合. 需要注意的是, 测试阶段不需要进行全关系拼接, 只需要将查询特征和每一类原型拼接.

### 3.4 分析与讨论

本文提出方法中的特征对全关系拼接模块主要致力于实现不同类特征间距离的最大化. 现将 3.3.3 节中的式(23)进行展开:

$$\begin{aligned} F_{\text{loss}} &= \sum_{i=1}^4 \sum_{l_{A,B} \in L_i} l_{A,B} \\ &= \sum_{i=1}^4 \sum_{l_{A,B} \in L_i} (g_{\varphi}(\cos(A', B')) \times P_{A,B}) - y_{A,B})^2 \\ &= \sum_{i=1}^4 F_{\text{loss}i} \end{aligned} \quad (25)$$

式中的  $F_{\text{loss}3}$  实际上是关系网络采用的损失, 为了约束查询特征和类原型之间的距离, 而  $F_{\text{loss}1}$ 、 $F_{\text{loss}2}$ 、 $F_{\text{loss}4}$  分别是为了约束类原型和类原型、类原型和查询特征、查询特征和查询特征之间的距离. 可见, 本文采用的损失可看作是对关系网络增加了正则化项, 增加的正则化项可使特征空间中同类特征间的距离更近, 不同类特征间的距离更远, 从而获得大边界的特征空间.

本文提出方法中的特征对注意力模块主要致力于让网络重点关注不同类之间难区分特征对的相似性学习. 式(25)中的特征  $A'$  和  $B'$  如果属于不同类且较难区分, 即它们是异类但在特征空间中具有较小的距离, 则  $A'$  和  $B'$  的余弦相似度是大的. 由此可见, 提出的特征对注意力机制能为不同类之间难区分的特征对增加大的权重, 着重关注不同类之间难区分特征对的相似性学习.

## 4 实验

### 4.1 实验设置

**数据集.** 本文使用的实验数据集是小样本图像分类中的四个常用公共数据集: mini-ImageNet、Stanford-Dogs、Stanford-Cars 和 CUB-200-2011. mini-ImageNet<sup>[19]</sup> 是 ImageNet<sup>[35]</sup> 的一个子集, 该数据集有 100 个类别, 每类有 600 个图像, 共有图像 60 000 张. 参照文献[13]的做法, 我们将 mini-ImageNet 的 100 个类别划分成分别由 64、16、20 类组成训练集、验证集和测试集. Stanford-Dogs<sup>[36]</sup>、Stanford-Cars<sup>[37]</sup>、CUB-200-2011<sup>[38]</sup> 均是用于细粒度小样本图像分类的基准数据集. 其中, Stanford-Dogs 数据集包含了 120 种狗类, 共有图像 20 580 张, 我们将该数据集的 120 个类别划分为分别由 60、30 和 30 类组成的训练集、验证集和测试集. Stanford-Cars 数据集包括 196 类汽车, 总共有 16 185 张图像, 我们将数据集的 196 个类别划分为分别由 98、49 和 49 类组成的训练集、验证集和测试集. CUB-200-2011 数据集包含了 200 种鸟类, 共有图像 11 788 张, 且每张图像均提供了图像类标记信息. 参照文献[39]的做法, 我们将数据集的 200 个类别划分为分别由 100、50 和 50 类组成的训练集、验证集和测试集. 上述四个数据集的训练集、验证集和测试集中分别含有的样本总数可见表 1.

表 1 数据集划分表

数据集	划分	类别数	样本数
mini-ImageNet	训练集	64	38 400
	验证集	16	9 600
	测试集	20	12 000
Stanford-Dogs	训练集	60	10 337
	验证集	30	5 128
	测试集	30	5 115
Stanford-Cars	训练集	98	8 203
	验证集	49	4 004
	测试集	49	3 978
CUB-200-2011	训练集	100	4 719
	验证集	50	4 715
	测试集	50	2 354

**网络结构.** 为了公平比较,本文使用和文献[8]中一样的嵌入模块和关系模块.如图3(a)所示,嵌入模块由四个卷积块和两个最大池化层组成.如图3(c)所示,关系模块由两个卷积块、两个最大池化层和两个全连接层组成;且第一个全连接层的输入维度为576,输出维度为8,其后接一个非线性激活函数ReLU;第二个全连接层的输入维度为8,输出维度为1,其后接一个Sigmoid激活函数.本文的特征对注意力模块位于嵌入模块和关系模块的中间,由两个卷积块、两个最大池化层和一个余弦距离度量层组

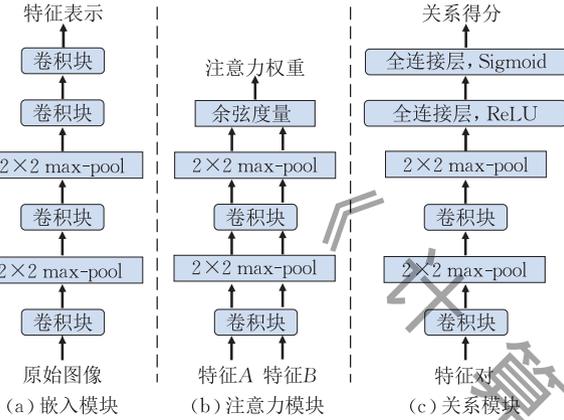


图3 注意力全关系网络中嵌入模块、特征对注意力模块和关系模块的详细结构

成,如图3(b)所示.另外,三个模块中的卷积块均由卷积、批次归一化和非线性激活函数ReLU组成,且卷积中卷积核的大小是 $3 \times 3$ ,卷积的输出通道数为64、卷积步长为1,卷积过程中未进行边界值填充.

**实现细节.** 我们在 mini-ImageNet、Stanford-Dogs、Stanford-Cars 和 CUB-200-2011 四个数据集上进行了 5-Way 1-shot 和 5-Way 5-shot 两种设置的实验,且实验中每类图像的查询样本数设置为 15. 在训练阶段,我们在 1-shot 和 5-shot 两种实验设置下分别随机采样了 60 000 和 40 000 个任务来对模型进行迭代训练,训练过程中采用 Adam 优化器<sup>[40]</sup>在均方误差损失下对模型参数进行优化,初始学习率设置为 $10^{-3}$ ,权重衰减率设置为 0. 在测试阶段,我们对测试集上随机采样的 600 个任务进行测试,然后报告了 95% 置信区间的平均分类准确率. 我们的模型以端到端的方式进行训练,没有预训练过程,并且测试时也不需要微调.

## 4.2 小样本图像分类

我们将提出的方法在 mini-ImageNet 数据集上的实验结果与主流的小样本图像分类方法进行了对比,以此来验证我们提出方法的可靠性,实验结果见表 2.

表 2 mini-ImageNet 数据集上的小样本图像分类准确率

对比方法	骨干网络	类型	5-Way 分类准确率/%	
			1-shot	5-shot
Meta-Learner LSTM <sup>[13]</sup> #	Conv4-32	元学习	43.44±0.77	60.60±0.71
MAML <sup>[14]</sup> *	Conv4-64	元学习	46.47±0.82	62.71±0.71
Meta-SGD <sup>[17]</sup> #	Conv4-64	元学习	50.47±1.87	64.03±0.94
Reptile <sup>[41]</sup> #	Conv4-32	元学习	47.07±0.26	62.74±0.37
Reptile+Transduction <sup>[41]</sup> #	Conv4-32	元学习	49.97±0.32	65.99±0.58
Baseline <sup>[39]</sup> *	Conv4-64	迁移学习	42.11±0.71	62.53±0.69
Baseline++ <sup>[39]</sup> *	Conv4-64	迁移学习	48.24±0.75	66.43±0.63
Matching Network <sup>[19]</sup> *	Conv4-64	度量学习	48.14±0.78	63.48±0.66
Prototypical Network <sup>[20]</sup> *	Conv4-64	度量学习	44.42±0.84	64.24±0.72
Relation Network <sup>[8]</sup> ##	Conv4-64	度量学习	49.33±0.85	65.44±0.69
GNN <sup>[34]</sup> #	Conv4-256	度量学习	50.33±0.36	66.41±0.63
PARN <sup>[30]</sup> ##	Conv4-64	度量学习	50.33±0.89	<b>67.48±0.67</b>
SAML <sup>[31]</sup> ##	Conv4-64	度量学习	50.81±0.83	<b>68.51±0.70</b>
TRNA(Ours)	Conv4-64	度量学习	<b>52.00±0.90</b>	<b>67.15±0.71</b>

注:“\*”表示文献[39]中报道的实验结果,“#”表示原论文中报道的实验结果,“##”表示基于文献[39]的代码复现的实验结果.第2列表示对应方法使用的骨干网络,Conv4-32表示32通道的四层卷积层、Conv4-64表示64通道的四层卷积层、Conv4-256表示256通道的四层卷积层.

首先,表2中基于元学习的方法主要关注于如何实现“学会学习”,却忽略了模型的训练复杂度,导致模型很难收敛,整体分类准确度较低.其次,基于迁移学习的两个方法在新类数据上主要采用传统大样本分类模型,过拟合现象仍比较严重.再次,与表2中基于度量的 Matching Network、Prototypical Network、

Relation Network、GNN 相比,我们的方法在 mini-ImageNet 的 1-shot 和 5-shot 实验中均有较高的分类性能,且与使用了注意力机制的小样本图像分类方法 PARN 和 SAML 相比,我们的方法也具有一定的竞争力.原因在于,提出的方法使用了特征对注意力机制和特征对全关系拼接,获得了高辨识度的分类特征.

### 4.3 小样本细粒度图像分类

为了进一步验证本文提出方法的泛化性能,我们在三个小样本细粒度图像分类数据集 Stanford-Dogs、Stanford-Cars 以及 CUB-200-2011 上进行了 5-Way 1-shot 和 5-Way 5-shot 的分类实验,并将实验结果与 5 个主流方法 MAML、Matching Network、Prototypical Network、Relation Network 以及 DN4 进行了对比,实验结果见表 3。

由表 3 可以看出,我们提出的方法在数据集 Stanford-Dogs、CUB-200-2011 的 1-shot、5-shot 以

及 Stanford-Cars 的 1-shot 实验中均有最高的分类准确度,在 Stanford-Cars 的 5-shot 实验中的分类准确度略低于 DN4。另外,与关系网络相比可发现,我们的方法在 Stanford-Dogs 的 1-shot 和 5-shot 实验中分别有 8.31% 和 3.92% 的性能提升,在 Stanford-Cars 的 1-shot 和 5-shot 实验中分别有 14.99% 和 8.00% 的性能提升,在 CUB-200-2011 的 1-shot 和 5-shot 实验中分别有 4.41% 和 4.42% 的性能提升,这证明我们提出的方法在细粒度分类数据集上具有较高的分类性能。

表 3 Stanford-Dogs、Stanford-Cars 以及 CUB-200-2011 数据集上的小样本图像分类准确率

对比方法	5-Way 分类准确率/%					
	Stanford-Dogs		Stanford-Cars		CUB-200-2011	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MAML <sup>[14]</sup> ##	46.67±0.87	62.56±0.80	48.37±0.81	65.41±0.77	55.92±0.95	72.09±0.76
Matching-Network <sup>[19]</sup> ##	44.88±0.84	61.22±0.73	45.29±0.82	64.00±0.74	61.16±0.89	72.86±0.70
Prototypical Network <sup>[20]</sup> ##	39.67±0.77	60.14±0.71	36.96±0.71	63.82±0.82	51.31±0.91	70.77±0.69
Relation Network <sup>[8]</sup> ##	47.11±0.90	65.56±0.74	45.83±0.87	68.01±0.78	62.67±0.98	76.94±0.66
GNN <sup>[34]</sup> *	46.98±0.98	62.27±0.95	55.85±0.97	71.25±0.89	—	—
DN4 <sup>[42]</sup> *	45.41±0.76	63.51±0.62	59.84±0.80	<b>88.65±0.44</b>	—	—
TRNA(Ours)	<b>55.42±0.92</b>	<b>69.48±0.74</b>	<b>60.82±0.99</b>	<b>76.01±0.74</b>	<b>67.08±0.91</b>	<b>81.36±0.61</b>

注:表中所有方法使用的骨干网络均为 Conv4-64。“\*”表示文献[42]中报道的实验结果,“##”表示基于文献[39]的代码复现的实验结果。

### 4.4 消融实验

为了充分验证提出的注意力全关系网络(TRNA)的有效性,我们对网络中的特征对注意力机制和特征对全关系拼接操作进行了裁剪。如果裁剪了特征对全关系拼接操作,则变为注意力关系网络(Relation Network with Attention, RNA);如果裁剪了特征对注意力机制,则变为全关系网络(Total Relation Net-

work, TRN);如果将特征对注意力机制和特征对全关系拼接操作全部裁剪,则恢复为原始关系网络(Relation Network, RN)。我们基于四层卷积层(Conv4)和 12 层残差层(ResNet12)两种骨干网络比较了关系网络、注意力关系网络、全关系网络以及注意力全关系网络在 mini-ImageNet 和 CUB-200-2011 两个数据集上的分类性能,见表 4。

表 4 mini-ImageNet 和 CUB-200-2011 数据集上的消融实验结果

对比方法	骨干网络	5-Way 分类准确率/%			
		mini-ImageNet		CUB-200-2011	
		1-shot	5-shot	1-shot	5-shot
RN	Conv4-64	49.33±0.85	65.44±0.69	62.67±0.98	76.94±0.66
RNA	Conv4-64	49.76±0.87	66.47±0.71	66.14±1.01	81.04±0.63
TRN	Conv4-64	50.81±0.84	66.22±0.72	66.95±0.94	78.00±0.66
TRNA	Conv4-64	<b>52.00±0.90</b>	<b>67.15±0.71</b>	<b>67.08±0.91</b>	<b>81.36±0.61</b>
RN	ResNet12	51.36±0.87	69.90±0.68	72.94±0.90	85.58±0.54
RNA	ResNet12	52.27±0.88	<b>72.39±0.70</b>	75.11±0.89	86.39±0.50
TRN	ResNet12	54.78±0.91	71.20±0.71	<b>75.54±0.93</b>	<b>86.51±0.54</b>
TRNA	ResNet12	<b>55.74±0.87</b>	70.52±0.69	74.94±0.90	85.84±0.52

注:表中 RN 表示关系网络, RNA 表示注意力关系网络, TRN 表示全关系网络, TRNA 表示注意力全关系网络。

通过观察表 4 中前四行的消融实验结果,可发现在 mini-ImageNet 的 1-shot 和 5-shot 实验中,相比于原始关系网络的分类准确率,注意力关系网络有 0.43% 和 1.03% 的提升,全关系网络有 1.48% 和 0.78% 的提升,注意力全关系网络有 2.67% 和 1.71% 的提升;而在 CUB-200-2011 的 1-shot 和

5-shot 实验中,相比于原始关系网络的分类准确率,注意力关系网络有 3.47% 和 4.10% 的提升,全关系网络有 4.28% 和 1.06% 的提升,注意力全关系网络有 4.41% 和 4.42% 的提升。这证明本文提出的特征对注意力机制和特征对全关系拼接在 Conv4 骨干网络下均是有效的,且在该骨干网络下将两者结合

可获得更好的实验性能。

通过观察表 4 中后四行的消融实验结果,可发现在 mini-ImageNet 的 1-shot 和 5-shot 实验中,相比于原始关系网络的分类准确率,注意力关系网络有 0.91% 和 2.49% 的提升,全关系网络有 3.42% 和 1.30% 的提升,注意力全关系网络有 4.38% 和 0.62% 的提升;而在 CUB-200-2011 的 1-shot 和 5-shot 实验中,相比于原始关系网络的分类准确率,注意力关系网络有 2.17% 和 0.81% 的提升,全关系网络有 2.60% 和 0.93% 的提升,注意力全关系网络有 2.00% 和 0.26% 的提升. 可看出,在 ResNet12 骨干网络下将特征对注意力机制和特征对全关系拼接结合后的实验性能有时比两者单独引入时高,有时也比单独引入时低. 原因可能在于,相比于 Conv4, ResNet12 具有深层结构和残差连接特性,更容易学习不同类图像之间的区分性特征,导致在此基础上同时使用特征对注意力机制和特征对

全关系拼接较难进一步提升性能. 但在 ResNet12 骨干网络下将两者结合后的实验性能均高于原始关系网络,这表明本文提出的特征对注意力机制和特征对全关系拼接在 ResNet12 骨干网络上也是有效的.

#### 4.5 训练和测试中任务类别数对分类性能的影响

为了观察训练和测试中任务类别数  $C$  对测试阶段分类性能的影响,我们在 mini-ImageNet 和 CUB-200-2011 两个数据集上基于提出的注意力全关系网络做了训练和测试中不同任务类别数  $C$  的实验. 首先将训练阶段的任务类别数分别设置为 3、4、5、6、7,测试阶段的任务类别数设置为 5,以此来观察训练阶段的任务类别数对测试性能的影响,实验结果见表 5;然后将训练阶段的任务类别数设置为 5,测试阶段的任务类别数分别设置为 3、4、5、6、7,以此来观察测试阶段的任务类别数对测试性能的影响,实验结果见表 6.

表 5 测试阶段的任务类别数为 5 时,训练阶段不同任务类别数下的实验结果

训练阶段 任务类别数的值	骨干网络	分类准确率/%			
		mini-ImageNet		CUB-200-2011	
		1-shot	5-shot	1-shot	5-shot
3-Way	Conv-64	49.52±0.84	65.84±0.69	65.80±0.90	77.12±0.70
4-Way	Conv-64	51.43±0.81	65.84±0.69	66.81±0.79	80.19±0.63
5-Way	Conv-64	52.00±0.90	67.15±0.71	67.08±0.91	81.36±0.61
6-Way	Conv-64	54.50±0.92	68.44±0.74	68.10±0.92	81.41±0.60
7-Way	Conv-64	<b>55.40±0.93</b>	<b>68.88±0.74</b>	<b>68.44±0.97</b>	<b>81.52±0.62</b>

表 6 训练阶段的任务类别数为 5 时,测试阶段不同任务类别数下的实验结果

测试阶段 任务类别数的值	骨干网络	分类准确率/%			
		mini-ImageNet		CUB-200-2011	
		1-shot	5-shot	1-shot	5-shot
3-Way	Conv-64	<b>64.63±1.08</b>	<b>78.16±0.81</b>	<b>77.33±1.06</b>	<b>88.54±0.65</b>
4-Way	Conv-64	57.28±0.96	71.94±0.77	71.59±1.01	84.26±0.67
5-Way	Conv-64	52.00±0.90	67.15±0.71	67.08±0.91	81.36±0.61
6-Way	Conv-64	46.27±0.77	63.11±0.61	65.13±0.86	78.21±0.60
7-Way	Conv-64	43.04±0.71	60.07±0.59	60.97±0.81	76.05±0.56

通过观察表 5 和表 6 可以发现,当测试阶段的任务类别数固定为 5 时,训练阶段的任务类别数越大,测试性能越高;当训练阶段的任务类别数固定为 5 时,测试阶段的任务类别数越大,测试性能越低. 原因在于若训练阶段的任务类别数越大,支持集中的不同类别数就越多,有利于提高模型对样本区分性特征的学习能力,使得分类准确性越高;若测试阶段的任务类别数越大,支持集中的不同类别数就越多,则分类难度越大,使得分类准确性越低.

#### 4.6 实验细节分析

为了直观地体现本文提出方法的有效性,我们

基于训练好的最佳模型绘制了关系得分预测图、原始图像的特征可视化图、样本对得分样例图.

**关系得分预测图.** 首先从 CUB-200-2011 和 mini-ImageNet 的测试集中分别为 5-Way 1-shot 和 5-Way 5-shot 的实验随机采样分类任务(为了达到清晰的图像效果,分类任务中每类图像的查询样本数设置为 5),然后将采样到的四个分类任务输入关系网络和我们提出网络的最佳训练模型中,得到 8 个  $30 \times 30$  的关系得分预测矩阵,并将这 8 个  $30 \times 30$  的关系得分预测矩阵可视化为 8 个关系得分预测图,如图 4 前四列所示.

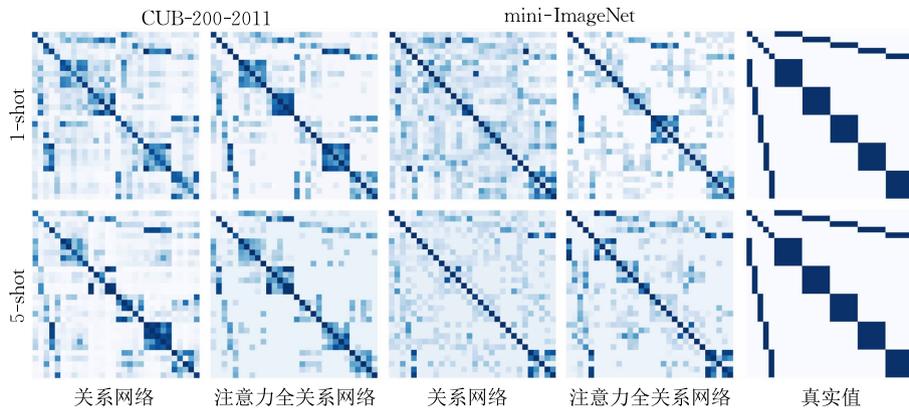


图 4 关系得分预测图

图 4 中的每一个关系得分预测图均由 900 个小方格组成, 方格的颜色越深说明对应的得分越高. 得分预测图的前 5 行代表分类任务中的 5 种类原型和 30 个特征(5 种类原型加 25 个查询特征)的相似得分, 而前 5 行后面的每个 5 行依次代表分类任务中某类查询样本对应的 5 个查询特征和 30 个特征的相似得分. 由 8 个关系得分预测图的颜色变化趋势可知, 我们提出的注意力全关系网络在 mini-ImageNet 和 CUB-200-2011 的 1-shot 和 5-shot 实验中的分类准确率均好于原始的关系网络. 另外, 对比图 4 中 CUB-200-2011 数据集 5-shot 实验下的两个关系得分预测图, 发现有一些负类样本对应的方格颜色有

所提升. 原因可能在于, 随机挑选的分类任务中可能存在一些相似性过高的异类样本, 导致类内紧凑性变大的同时减小了这些不同类的类间可区分性.

**特征可视化图.** 为了证明提出的注意力全关系网络学习到的特征分布在较小的特征空间中且更具辨别力, 我们使用基于梯度的技术 Grad-CAM<sup>[43]</sup>对原始图像中的重要区域进行了可视化.

我们从 CUB-200-2011 的测试集中随机选择 9 张原始图像并将挑选出的原始图像调整为与嵌入模块的输出特征一样的大小; 然后基于匹配网络<sup>[19]</sup>、原型网络<sup>[20]</sup>、关系网络<sup>[8]</sup>以及我们提出的注意力全关系网络来绘制 Grad-CAM 图像, 如图 5 所示. 从图中

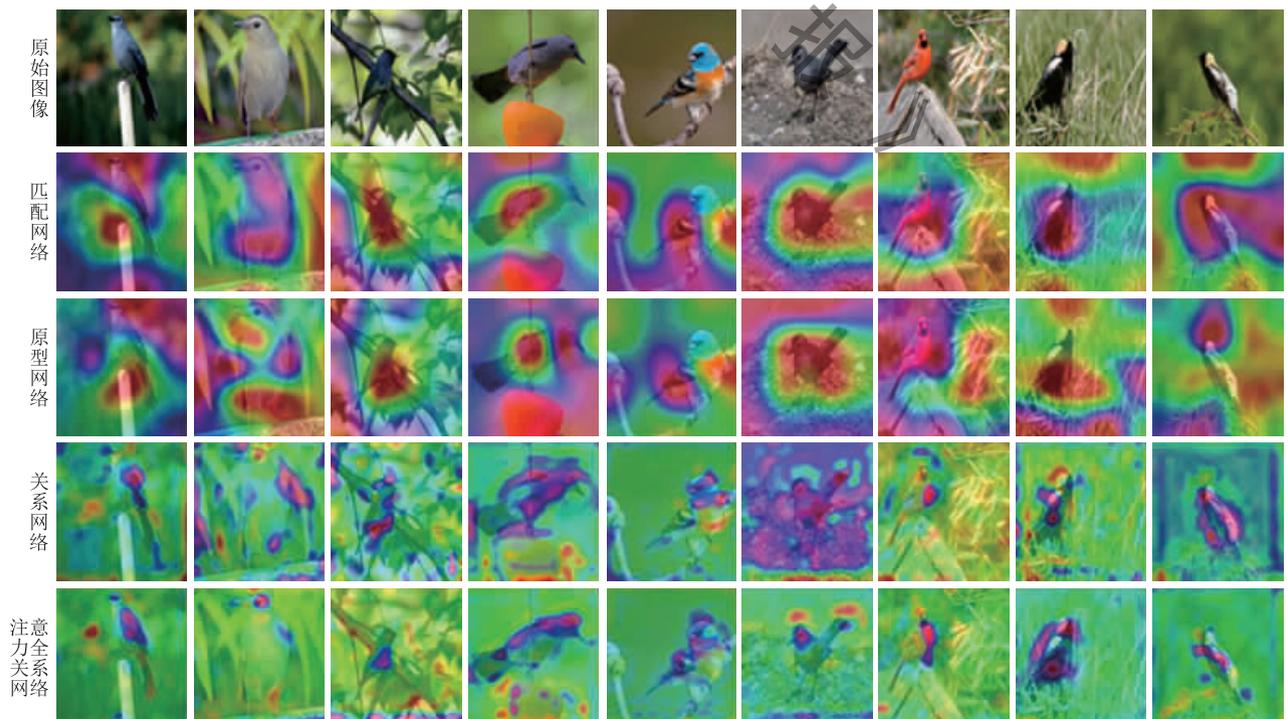


图 5 匹配网络、原型网络、关系网络以及提出的注意力全关系网络的特征可视化图(区域越红, 越具有类可区分性)

可看出本文提出的注意力全关系网络的类别判别区域大多集中在目标对象身上,而其它方法的类别判别区域在背景处也有较多分布,这证明本文的方法可以学习到更加稳健和高效的特征表示。

**样本对得分样例图。**为了进一步说明提出的特征对注意力机制和特征对全关系拼接可提高模型对难区分特征对的辨别能力,我们在 CUB-200-2011 数据集中挑选了四组难区分的样本对,并展示了关系网络和提出网络在这些样本对上的关系得分,如图 6 所示。

查询样本	a类支持样本	关系得分	查询样本	b类支持样本	关系得分
		0.0355 0.0945			0.1477 0.0740
		0.0555 0.1293			0.0906 0.0829
		0.0586 0.1875			0.1177 0.0477
		0.0432 0.2043			0.1580 0.0650

图 6 样本对得分样例图(蓝色框代表关系网络的得分,红色框代表注意力全关系网络的得分)

图 6 中每一行的查询样本和 a 类支持样本属于同一类,但却和不属于同一类的 b 类支持样本也具有较多的相似性区域。对比图中的关系得分可知,相比于关系网络,我们提出的网络在同类样本对上的关系得分均较高,在异类样本对上的关系得分均较低,这说明了我们的网络实现了对难区分样本对相似性的侧重学习。

## 5 总 结

本文在关系网络的基础上提出了一种面向小样本图像分类的新网络:注意力全关系网络。具体地,本文首先引入特征对全关系拼接操作将分类任务中的任意两个特征在通道方向上拼接,以使特征空间中不同类之间的距离最大化;然后,本文利用特征对注意力机制将特征对矩阵中不同类之间难区分的特征对挑选出来并给予大的权重,使网络侧重关注不同类之间难区分特征对的相似性学习。消融实验证明本文提出的特征对注意力机制和特征对全关系拼

接操作均可以改善原始关系网络的分类性能,且在四层卷积层的骨干网络中,两者结合可以实现更好的性能。实验细节分析也体现了本文提出的注意力全关系网络的分类性能且证明了本文的方法可以为原始图像提取到更加关键的特征表示。

## 参 考 文 献

- [1] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014
- [2] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA, 2016: 770-778
- [3] Szegedy C, Liu Wei, Jia Yangqing, et al. Going deeper with convolutions//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA, 2015: 1-9
- [4] Li Feifei, Fergus R, Perona P. One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2006, 28(4): 594-611
- [5] Dong Xuanyi, Zheng Liang, Ma Fan, et al. Few-example object detection with model communication. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2019, 41(7): 1641-1654
- [6] Rahman S, Khan S H, Porikli F. A unified approach for conventional zero-shot, generalized zero-shot and few-shot learning. IEEE Transactions on Image Processing (TIP), 2018, 27(11): 5652-5667
- [7] Huang Kai, Geng Ji, Jiang Wen, et al. Pseudo-loss confidence metric for semi-supervised few-shot learning//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Montreal, Canada, 2021: 8651-8660
- [8] Sung Flood, Yang Yongxin, Zhang Li, et al. Learning to compare: Relation network for few-shot learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018: 1199-1208
- [9] Sun Qianru, Liu Yaoyao, Chua Tat-Seng, et al. Meta-transfer learning for few-shot learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 403-412
- [10] Ye Hanjia, Hu Hexiang, Zhan Dechuan, et al. Few-shot learning via embedding adaptation with set-to-set functions//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA, 2020: 8805-8814
- [11] Zhang Jian, Zhao Chenglong, Ni Bingbing, et al. Variational few-shot learning//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Seoul, Korea, 2019: 1685-1694

- [12] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015
- [13] Ravi S, Larochelle H. Optimization as a model for few-shot learning//Proceedings of the International Conference on Learning Representations (ICLR). Toulon, France, 2017: 1-11
- [14] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks//Proceedings of the International Conference on Machine Learning (ICML). Sydney, Australia, 2017: 1126-1135
- [15] Rusu A A, Rao D, Sygnowski J, et al. Meta-learning with latent embedding optimization. arXiv preprint arXiv:1807.05960, 2018
- [16] Sun Zhuo, Wu Jijie, Li Xiaoxu, et al. Amortized Bayesian prototype meta-learning: A new probabilistic meta-learning approach to few-shot image classification//Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS). San Diego, USA, 2021: 1414-1422
- [17] Li Zhenguo, Zhou Fengwei, Chen Fei, et al. Meta-SGD: Learning to learn quickly for few-shot learning. arXiv preprint arXiv:1707.09835, 2017
- [18] Huisman M, van Rijn J N, Plaata A. A survey of deep meta-learning. Artificial Intelligence Review, 2021, 54(6): 4483-4541
- [19] Vinyals O, Blundell C, Lillicrap T, et al. Matching networks for one shot learning//Proceedings of the Conference on Neural Information Processing Systems (NIPS). Barcelona, Spain, 2016: 3630-3638
- [20] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning//Proceedings of the Conference on Neural Information Processing Systems (NIPS). Long Beach, USA, 2017: 4077-4087
- [21] Liu Jinlu, Song Liang, Qin Yongqiang. Prototype rectification for few-shot learning//Proceedings of the European Conference on Computer Vision (ECCV). Glasgow, UK, 2020: 741-756
- [22] Li Xiaoxu, Wu Jijie, Sun Zhuo, et al. BSNet: Bi-similarity network for few-shot fine-grained image classification. IEEE Transactions on Image Processing (TIP), 2021, 30: 1318-1331
- [23] Zhang Chi, Cai Yujun, Lin Guosheng, et al. DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA, 2020: 12200-12210
- [24] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks//Proceedings of the Conference on Neural Information Processing Systems (NIPS). Montreal, Canada, 2015: 2017-2025
- [25] Wang Fei, Jiang Mengqing, Qian Chen, et al. Residual attention network for image classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017: 6450-6458
- [26] Woo S, Park J, Lee J-Y, et al. CBAM: Convolutional block attention module//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 3-19
- [27] Hu Jie, Shen Li, Albanie S, et al. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2020, 42(8): 2011-2023
- [28] Yan Shipeng, Zhang Songyang, He Xuming. A dual attention network with semantic embedding for few-shot learning//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Hawaii, USA, 2019: 9079-9086
- [29] Ren Mengye, Liao Renjie, Fetaya E, et al. Incremental few-shot learning with attention attractor networks. arXiv preprint arXiv:1810.07218, 2018
- [30] Wu Ziyang, Li Yuwei, Guo Lihua, et al. PARN: Position-aware relation networks for few-shot learning//Proceedings of the International Conference on Computer Vision (ICCV). Seoul, Korea, 2019: 6658-6666
- [31] Hao Fusheng, He Fengxiang, Cheng Jun, et al. Collect and select: Semantic alignment metric learning for few-shot learning //Proceedings of the IEEE International Conference on Computer Vision (ICCV). Seoul, Korea, 2019: 8459-8468
- [32] Zhao Yang, Li Chunyuan, Yu Ping, et al. ReMP: Rectified metric propagation for few-shot learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville, USA, 2021: 2581-2590
- [33] Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928, 2016
- [34] Garcia V, Bruna J. Few-shot learning with graph neural networks. arXiv preprint arXiv:1711.04043, 2017
- [35] Deng Jia, Dong Wei, Socher R, et al. ImageNet: A large-scale hierarchical image database//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Miami, USA, 2009: 248-255
- [36] Khosla A, Jayadevaprakash N, Yao Bangpeng, et al. Novel dataset for fine-grained image categorization: Stanford dogs //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Colorado Springs, USA, 2011: 1-2
- [37] Krause J, Stark M, Deng Jia, et al. 3D object representations for fine-grained categorization//Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW). Sydney, Australia, 2013: 554-561
- [38] Wah C, Branson S, Welinder P, et al. The Caltech-UCSD Birds-200-2011 dataset. Computation & Neural Systems Technical Report: CNS-TR-2011-001, 2011
- [39] Chen Weiyu, Liu Yencheng, Kira Zsolt, et al. A closer look at few-shot classification. arXiv preprint arXiv:1904.04232, 2019
- [40] Kingma D P, Ba J L. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014

- [41] Nichol A, Achiam J, Schulman J. On first-order meta-learning algorithms. arXiv preprint arXiv:1803.02999, 2018
- [42] Li Wenbin, Wang Lei, Xu Jinglin, et al. Revisiting local descriptor based image-to-class measure for few-shot learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA,

2019; 7253-7260

- [43] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization //Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017: 618-626



**LI Xiao-Xu**, Ph. D. , professor. Her research interests include machine learning fundamentals with a focus on applications in image and video understanding.

**LIU Zhong-Yuan**, M. S. candidate. His research interests include machine learning and few-shot learning.

**WU Ji-Jie**, Ph. D. candidate. His research interests include machine learning and few-shot learning.

**CAO Jie**, Ph. D. , professor. Her research interests include machine learning, pattern recognition, speech and speaker recognition, information fusion, and computer vision.

**MA Zhan-Yu**, Ph. D. , professor. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in computer vision, multimedia signal processing, and data mining.

## Background

The problem studied in this paper is how to learn new class concepts based on a small number of labeled samples, which belongs to the field of few-shot image classification. The metric-based method is currently the common few-shot image classification method. This method puts all samples in a classification task into the convolutional network to extract the feature representations, and then some metric standard such as cosine distance, euclidean distance, etc. are used to determine the distance between the two feature representations, and finally classification is completed based on the calculated distance. This metric-based few-shot classification method provides a new idea to solve the problem of few-shot image classification and achieves better classification results. However, since two images of different classes in a classification task may contain more similar regions, some query features are close to the heterogeneous class prototypes, resulting in few-shot image classification networks that are more difficult to learn large classification margin.

In order to solve the above problem, we propose Total Relation Network with Attention (TRNA) based on Relation

Network. Specifically, the Total Relation Concatenation Operation of Feature-Pair is introduced to concatenate any two features in the classification task in the channel direction to obtain the feature pair matrix, which maximizes the distance between different classes in the feature space; the Attention Mechanism of Feature-Pair is introduced to select the feature pairs that are difficult to distinguish between different classes in the feature pair matrix and give them large weights, so that the network pays more attention to the indistinguishable feature pairs. Experiments show that the classification performance of the Relation Network is significantly improved after adding the Total Relation Concatenation Operation of Feature-Pair and the Attention Mechanism of Feature-Pair.

This work was supported in part by the National Natural Science Foundation of China(Grant Nos. 62176110, 62111530146, 61906080, 61922015, U19B2036, 62225601), in part by the Gansu Provincial Youth Doctoral Foundation(Grant No. 2021QB-038), in part by the Beijing Natural Science Foundation (Grant No. Z200002), in part by the Hongliu Outstanding Youth Fund of Lanzhou University of Technology.