山西大学学报(自然科学版)47(4):761—766,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2023069

基于D-ResNeXt骨干网络的小样本图像分类算法

杨红菊[1,2*],翟艳峰[1]

(1.山西大学 计算机与信息技术学院,山西 太原 030006;
2.山西大学 计算智能与中文信息处理教育部重点实验室,山西 太原 030006)

摘 要:小样本图像分类目前是人工智能领域中非常重要的方向之一,其中基于度量学习的方法具有简洁高效的特点。针对目前图像分类中特征提取阶段所使用的骨干网络问题,现有工作大多使用传统残差网络,受数据集的影响,对类内差异大的图片特征提取效果不佳。ResNeXt为传统残差网络ResNet的升级版本,优化了在特征提取阶段准确度不高,误差较大的问题。根据其网络特点,本文设计出一种适用于小样本模型的网络变体,运用其变体作为骨干网络,提高其特征提取能力,同时结合两种注意力模块,进一步提升对图像类内相似性以及类间差异性的识别效果,减少无关因素影响,有效提升整体分类精度。

关键词:小样本学习;图像分类;注意力机制;度量学习;残差网络
中图分类号:TP391 **文献标志码**:A **文章编号**:0253-2395(2024)04-0761-06

Few-shot Image Classification Algorithm Base on D-ResNeXt Backbone Network

YANG Hongju[1,2*], ZHAI Yanfeng[1]

(1. *School of Computer and Information Technology, Shanxi University, Taiyuan* 030006*, China;*
2. *Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan* 030006*, China*)

Abstract: Few-shot image classification is currently one of the most important directions in the field of artificial intelligence. In this area, the method based on metric learning is concise and efficient. To address the problem of the backbone network used in the feature extraction stage of current image classification, most existing works use traditional residual networks, which extracts poorly the features of images with large intra-class differences as the method is influenced by the dataset. ResNeXt is an upgraded version of the traditional residual network ResNet, optimizing the problem of low accuracy and large errors in the feature extraction stage of the traditional network. According to its network characteristics, this paper designs a network variant suitable for small sample models, which uses its variant as a backbone network to improve its feature extraction ability, and combines two attention modules to further improve the recognition effect of intra-class similarity and inter-class variability of images, reduce the influence of irrelevant factors, and effectively improve the overall classification accuracy.

Key words: few-shot learning; image classification; attention mechanism; metric learning; residual network

收稿日期:2023-03-10;**接受日期**:2023-03-28
基金项目:国家自然科学基金(61976128);山西省回国留学人员科研资助项目(2022-008)

* **通信作者**:杨红菊(1975-),女,山西临汾人,博士,副教授,研究方向为中文信息处理、计算机视觉。E-mail:yhju@sxu.edu.cn

引文格式:杨红菊,翟艳峰.基于D-ResNeXt骨干网络的小样本图像分类算法[J].山西大学学报(自然科学版),2024,47(4):761-766. DOI:10.13451/j.sxu.ns.2023069

0 引言

深度学习通常需要学习大量有标记的图片,然而对于某些数据需要耗费较大精力或是本身就很少的情况下,传统的大规模深度学习图像分类算法因为其需要大量的训练数据,从而不再适用。小样本[1]理念的出现则是为了解决样本不足的问题,其目标为使计算机做到学习少量样本便可以准确认识一个新的样本类型。小样本图像分类[2]是小样本任务中一个重要的研究方向。它的目的为仅学习有限数量的标记图片数据后可以对新的图像类别进行对比分类。

在解决小样本图像分类问题的众多方法中,基于度量的方法[3-5]是使用范围较广的一种。基于度量的方法通常使用卷积神经网络来学习图像特征表示,并且使用距离函数来直接计算测试图像与查询图像的特征表示之间的距离,随之通过预测图像标签从而进行分类。Koch等[6]提出的深度卷积孪生网络,是通过图像之间的互相验证进行训练。Vinyals等[7]在之后提出了一种新的方法Match-Net(即匹配网络),该方法采用图片数据之间的特征距离匹配的方式进行小样本图像分类的任务。Snell等[8]提出了原型网络(Prototypical Network),基于匹配网络进行改进,引入聚类的思想,根据共有特征的中心想法构造模型。Sung等[9]则提出了关系网络(Relation Network)。关系网络在原型网络模型的基础上对度量方式进行了改变,使用神经网络构造出一个带有参数的相似性度量函数,Hou等[10]提出了一种新颖的交叉注意网络(Cross Attention Network)来解决小样本图像分类问题。尽管这些方法取得了一定的效果,但混乱的背景图片或者较大的类内外观的变化可能会使相同类别的图像在骨干网络提取特征时计算的距离相差偏大。在大规模的数据训练下,深度神经网络可以很大程度上缓解这个问题,但在小样本任务中,这显然成为一个不可忽视的因素,因其可能对图像分类的精度产生不利的影响。在现有的工作中,大多使用ResNet[11]作为骨干网络[12-14],因网络无法大量堆叠,而训练数据又无法大规模增加,所以其有时的表现不尽如人意。针对上述问题,本文结合ResNeXt[15],设计一种新的结构,可以在不需要堆叠多层,不需要大规模增加参数量的前提下,提升网络性能,而后基于一种改进特征提取网络的图像分类算法,采用端到端的方式对整个网络进行训练。通过提升对图像的特征提取能力,并运用注意力模块,来减少同类图像的差异性,达到有效提升图像分类准确度的目的。

1 基本原理

ResNeXt的提出原因有多方面,在当时的背景下,传统的网络要提高模型的准确率,都是加深或加宽网络,这会导致参数数量巨额增加,但随着参数数量的增加,网络设计的难度和计算开销也会增加。文献[15]提出的ResNeXt结构可以在不增加参数复杂度的前提下提高准确率。因为其拥有拆分的模块思想,使得在多个分支下提取更多有效的特征,避免了由于背景混乱而导致的特征提取有效性不足的问题。小样本任务的突出特点是可供支持的图片较少,所以提升骨干网络特征提取能力,在有限的图片下最大程度地获得有效特征,便是针对小样本图像分类任务中至关重要的问题,因此本文提出了适用于小样本的D-ResNeXt(Double module-ResNeXt,双模块ResNeXt)网络模型。

本文所提出模型的主要结构如图1所示。其最主要的模块分为两大块,即特征提取阶段所使用的D-ResNeXt骨干网络模块与突出有效特征的注意力模块。

1.1 D-ResNeXt模块

神经网络中最简单的神经元执行内积,是由全连接层和卷积层完成的基本变换。内积可以被认为是一种聚合转换的形式:

$$\sum_{i=1}^{D}\omega_i x_i, \tag{1}$$

其中$\boldsymbol{x}=[x_1,x_2,\cdots,x_D]$是神经元的$D$通道输入量,$\omega_i$是第$i$个通道的滤波器权重。这种操作被称为“神经元”。上述操作也可以重构为拆分,转换和聚合的组合。鉴于上述对简单神经元的分析,考虑将基本变换($\omega_i x_i$)替换为更通用的函数,该函数本身也可以是一个网络。而聚合转换将表示为:

$$F(\boldsymbol{x})=\sum_{i=1}^{C}T_i(\boldsymbol{x}), \tag{2}$$

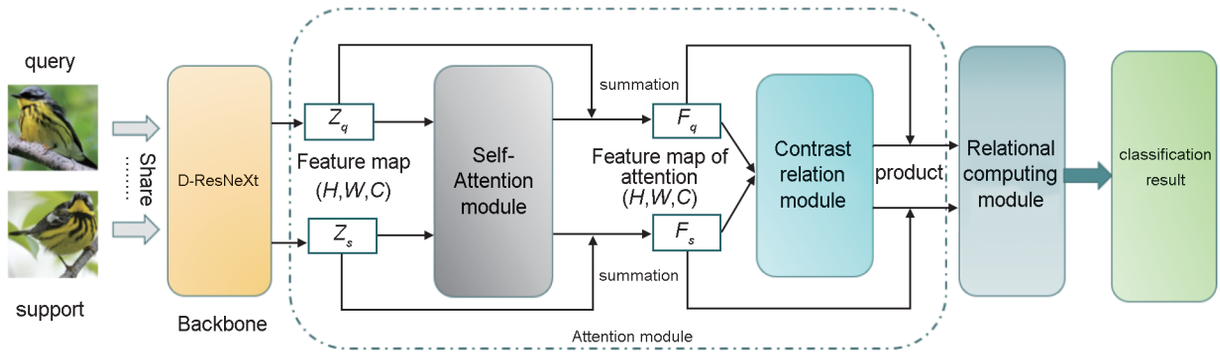

图1 网络模型主要结构

Fig. 1 Main structure of the network model

其中$T_i(\boldsymbol{x})$可以是任意函数。而在公式(2)中,C是要聚合的变换集的大小。其中C称为"cardinality",该数值的含义代表了最终所分支的数值。在公式(2)中,C的位置类似于公式(1)中的D,可以是任意数。公式(2)中的聚合变换用作残差函数时:

$$y=\boldsymbol{x}+\sum_{i=1}^{C}T_i(\boldsymbol{x}), \tag{3}$$

其中y是输出。

本文架构同ResNeXt类似(如图2),最主要的贡献便是参照其结构设计出如图2所示的适用于小样本的任务的模型,使用了一个模块执行一组转换,其输出通过求和进行聚合变换,其变换都是相同的拓扑结构。本文采用的模块化规则与ResNet12相同,设计一个模板板块,就可以扩展到所有的模块。由于其独特的结构,在不会大量增加参数量的前提下可以有效提升网络的提取能力。

1.2 注意力模块

注意力模块是由自注意和互相关两个阶段组成。它进一步从骨干网络提取的特征做出处理,使得模型把重点关注到图片的相同类之间相似性和不同类之间的差异性。

自注意模块中,自注意力计算是给定一个基本特征图Z,计算C维向量在每个位置$\boldsymbol{x}\in[1,H]\times[1,W]$及其邻域的Hadamard积,并将它们收集到自注意力中张量$\boldsymbol{R}\in\mathbb{R}^{H\times W\times U\times V\times C}$,张量$\boldsymbol{R}$可以表示为具有$C$维向量输出的函数:

$$\boldsymbol{R}(x,p)=\frac{Z(\boldsymbol{x})}{\|Z(\boldsymbol{x})\|}\odot\frac{Z(\boldsymbol{x}+p)}{\|Z(\boldsymbol{x}+p)\|}, \tag{4}$$

其中$p\in[-d_U,d_U]\times[-d_V,d_V]$,对应于邻域窗口中的相对位置,使得$2d_U+1=U,2d_V+1=V$,包括中心位置。如图3(a)结构,本文使用了一

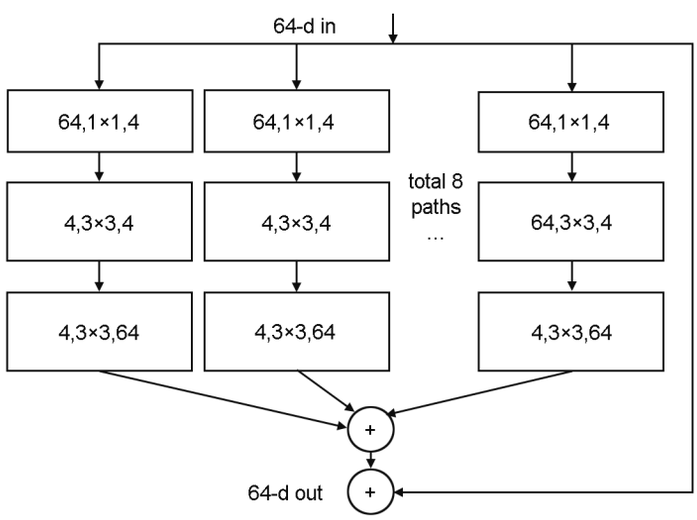

图2 cardinality数为8的一个D-ResNeXt块

Fig. 2 A D-ResNeXt block with cardinality number of 8

个卷积块用于减小通道尺寸,两个用于转换的3×3卷积层和另一个1×1卷积层用于通道大小恢复。而在卷积之间,插入批量归一化和ReLU。卷积块$g(\cdot)$将它们的空间维度从$U\times V$减少到1×1,使得输出$g(\boldsymbol{R})$具有与Z相同的大小,即$g:\mathbb{R}^{H\times W\times U\times V\times C}\rightarrow\mathbb{R}^{H\times W\times C}$。自注意力特征表示$F\in\mathbb{R}^{H\times W\times C}$:

$$F=g(\boldsymbol{R})+Z。\tag{5}$$

互相关模块中,输入一对查询集表示F_q和支持集表示F_s,生成相对应的注意力图,A_q和A_s,如图3(b)所示。这些注意力图最终将聚合到一个嵌入向量,帮助最终分类。互相关计算将F_q和F_s输入计算模块当中,得到张量$\boldsymbol{Q}\in\mathbb{R}^{H\times W\times H\times W}$,且有:

$$\boldsymbol{Q}(X_q,X_s)=\mathrm{sim}\left(F_q(X_q),F_s(X_s)\right),\tag{6}$$

其中X表示特征图上的空间位置,$\mathrm{sim}(\cdot,\cdot)$表示两个特征之间的余弦相似度。互注意力计算张量$\boldsymbol{Q}'$生成了共同注意图,它们展现了两个特征图之间的差异性。最终在关系模块整合,获得最后的关系嵌入,之后通过度量距离的相似度计算,最终完成分类。

2 实验与分析

为了评估本文提出的算法性能,在两个经典数据集上进行了广泛的实验。首先介绍数据集和一些实现细节,之后做消融实验来验证方法的有效性,并将模型与标准数据集上的其他方法进行比较。

2.1 数据集

本文所使用的数据集为两个标准数据集:CUB-200-2011与miniImageNet。

CUB-200-2011(CUB)[16]是一个用于鸟类细粒度分类的数据集,其中包含来自200个类别的11 788张图片,其由100个训练对象类,50个验证对象类,50个测试对象类组成。

miniImageNet最初由Vinyals等提出[7]。其为ImageNet[17]的一个子集,由60 000张图片组成。它包含100个类别,每个类别有600张图像。其中的64类用于训练,16类用于验证,20类用于测试。

2.2 实验细节

本文的主干网络为D-ResNeXt12,主干网络将空间大小为84×84的图像作为输入,并提供基本特征$Z\in\mathbb{R}^{5\times 5\times 640}$。一个episode中为每个类别设置测试集中的15个查询样本,并报告其平均分类精度,其中在随机抽样的2 000个测试样本中添加95%的置信区间。为了公平对比,本文设置标准与其他团队在小样本图像分类任务中所通用的设置标准相同,采用5-way 1-shot和5-way 5-shot的形式来衡量本文模型

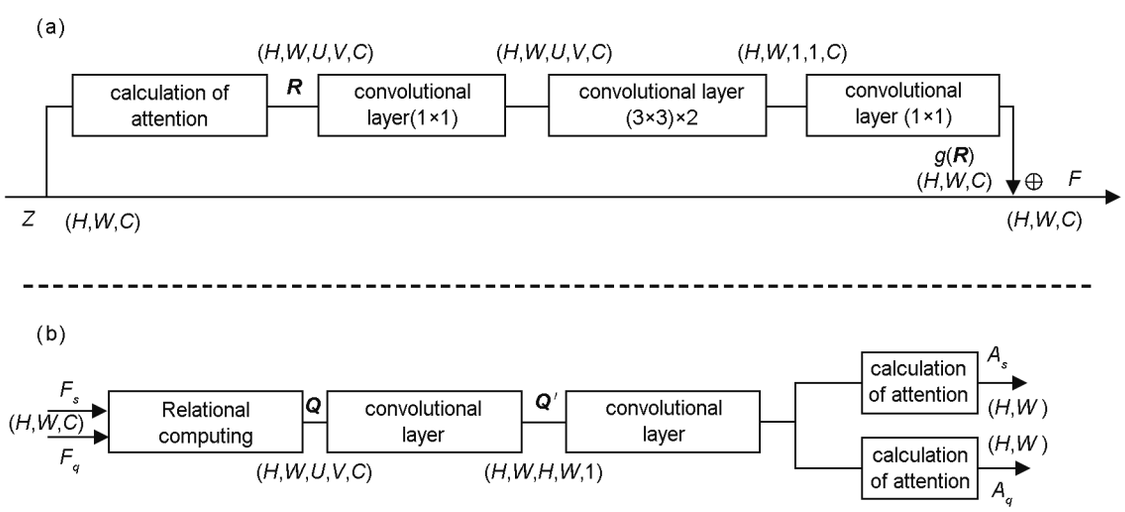

图3 模块图示

(a)自注意力模块;(b)互相关模块

Fig. 3 Module diagram

(a) Self-attention module; (b) Cross-correlation module

准确率,统一使用60个epoch迭代。

2.3 实验结果

为了验证本文模型效果,本文对两个公开数据集上的实验结果与其他先进的工作进行比较,最终结果如表1—表2所示。本文所提出的D-ResNeXt与其他的小样本分类模型所相比,在CUB数据集与miniImageNet数据集下精度分别都取得了较明显的提升。通过分析可得,本文所设计出的网络结构在整体上相对传统残差网络具有良好的适应以及处理能力,通过设计的模块,使得骨干网络提升了对图像有效特征提取的能力,不仅可以对如CUB这样的细粒度鸟类图片做出良好的分类,突出其在相似特征中的不同点,用以区分所属种类的细微差异,还可以对miniImageNet这种多种类数据集提供更为有效的识别能力,增加相同种类的图片的类内相似度,使得网络可以获得提取表示图片中种类的突出特征的能力。在实验结果中的优异表现也验证了本文所使用模型的有效性。

表1 在CUB数据集上的分类精度对比

Table 1 Comparison of classification accuracy in CUB datasets

模型	骨干网络	5-way 1-shot	5-way 5-shot
ProtoNet[6]	ResNet12	66.09±0.92	82.50±0.58
cosine classifier[18]	ResNet12	67.30±0.86	84.75±0.60
MatchNet[7]	ResNet12	71.87±0.85	85.08±0.57
DeepEMD[4]	ResNet12	75.65±0.83	88.69±0.50
RENet[13]	ResNet12	79.49±0.44	91.11±0.24
ours	D-ResNeXt12	80.33±0.43	91.20±0.25

表2 在miniImageNet数据集上的分类精度对比

Table 2 Comparison of classification accuracy in miniImageNet datasets

模型	骨干网络	5-way 1-shot	5-way 5-shot
cosine classifier[18]	ResNet12	55.43±0.81	77.18±0.61
ProtoNet[6]	ResNet12	62.39±0.21	80.53±0.14
MatchNet[7]	ResNet12	63.08±0.80	75.99±0.60
DeepEMD[4]	ResNet12	65.91±0.82	82.41±0.56
RENet[13]	ResNet12	67.60±0.44	82.58±0.30
ours	D-ResNeXt12	68.64±0.43	82.92±0.30

2.4 消融实验

与基线模型相比,或去掉所有模块,或在每个模块做出改变(消除或恢复),并比较在数据集准确度上的结果。现选择这两种数据集的5-way 1-shot情况下的结果作为展示。通过表3可以得出,本文所提出的模型相比传统模型有了较为明显的提升,并且在每一个模块都表现出相应的预期效果。其中基础模型(baseline)为ResNeXt不做变动直接用于骨干网络应用于图像分类任务。实验结果表示,在miniImageNet数据集中更改骨干网络使得最终准确度相比基础模型提升1.81%(原始骨干网络模型参数量为203 k,更改为本文网络后参数量变为268 k),之后取消网络变化,验证两个独立模块的效果,分别做到了1.38%与1.16%的提升,将三者结合,达到所提出模型的最终表现。在CUB数据集中仍然做到了这点,在分别使用三者的情况下达到79.06%、78.75%与78.69%的准确率,更是在三者结合的情况下达到了80.33%的准确度,证明了本文所提出各个模块的有效性,优于当前先进算法,取得良好的效果。

表3 文内各模块单独使用时的消融研究

Table 3 Ablation studies of each module used separately in this paper

模型	miniImageNet/%	CUB/%
baseline	65.38	77.54
D-ResNeXt	67.19	79.06
自注意模块	66.76	78.75
互相关模块	66.54	78.69
自注意模块+互相关模块	67.50	79.50
ours	68.64	80.33

通过实验证明,D-ResNext可以提升特征提取的表征能力,注意力模块进一步将前一阶段所提取出的特征表示来做出更为细化的标识分类,提高模型的泛化能力。因此,本文从骨干网络出发,并结合注意力模块,提升了对小样本图像分类任务的准确度,验证了所提出的模型对于该任务是有效果的。

3 结论

本文提出了一个新的模型来实现小样本的图像分类,解决了传统模型中所使用的残差网络不易堆叠,或多层堆叠后参数量过大的问题。通过研究ResNeXt的优势以及注意力机制对小样本图像分类的效果表现,设计一种新的骨干网络,在不需要堆叠多层的情况下提高网络提取能力,结合注意力模块,有针对性地提升模型提取能力,帮助最后的度量距离计算

提供良好的图像嵌入。通过广泛的实验,在公开的标准数据集上提升了小样本图像分类的准确度,证明了所提出理论的有效性。未来将考虑对提取特征阶段做进一步的研究拓展,通过结构设计,提升模型对于关键特征的提取能力,对数据图片更为复杂的情况下也具有很好的分类能力,进一步提升整体分类准确率。

参考文献:

[1] WANG Y Q, YAO Q M, KWOK J T, *et al*. Generalizing from a few Examples: A Survey on Few-shot Learning [J]. *ACM Comput Surv*, 2021, **53**(3): 1−34. DOI: 10.1145/3386252.

[2] LI F F, FERGUS R, PERONA P. One-shot Learning of Object Categories[J]. *IEEE Trans Pattern Anal Mach Intell*, 2006, **28**(4): 594−611. DOI: 10.1109/TPAMI.2006.79.

[3] FINN C, ABBEEL P, LEVINE S. Model-agnostic Meta-learning for Fast Adaptation of Deep Networks[C]//Proceedings of the 34th International Conference on Machine Learning − Volume 70. New York: ACM, 2017: 1126−1135. DOI: 10.5555/3305381.3305498.

[4] ZHANG C, CAI Y J, LIN G S, *et al*. DeepEMD: Few-shot Image Classification with Differentiable Earth Mover's Distance and Structured Classifiers[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 12200−12210. DOI: 10.1109/CVPR42600.2020.01222.

[5] YE H, HU H, ZHAN D, *et al*. Learning Embedding Adaptation for Few-shot Learning[EB/OL]. arXiv Preprint: 1812.03664, 2018. https://arXiv.org/abs/1812.03664.

[6] KOCH G, ZEMEL R, SALAKHUTDINOV R. Siamese Neural Networks for One-shot Image Recognition[C]//ICML Deep Learning Workshop. 2015, 2 (1).

[7] VINYALS O, BLUNDELL C, LILLICRAP T, *et al*. Matching Networks for One Shot Learning[J]. *Adv Neural Inf Process Syst*, 2016, **29**: 3637−3645. DOI: 10.5555/3157382.3157504.

[8] SNELL J, SWERSKY K, ZEMEL R. Prototypical Networks for Few-Shot Learning[J]. *Adv Neural Inf Process Syst*, 2017, **30**: 4080−4090. DOI: 10.5555/3294996.3295163.

[9] SUNG F, YANG Y X, ZHANG L, *et al*. Learning to Compare: Relation Network for Few-shot Learning[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1199−1208. DOI: 10.1109/CVPR.2018.00131.

[10] HOU R, CHANG H, MA B, *et al*. Cross Attention Network for Few-shot Classification[EB/OL]. arXiv Preprint: 1910.07677, 2019. https://arXiv.org/abs/1910.07677.

[11] HE K M, ZHANG X Y, REN S Q, *et al*. Deep Residual Learning for Image Recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 770−778. DOI: 10.1109/CVPR.2016.90.

[12] KWON H, KIM M, KWAK S, *et al*. Learning Self-similarity in Space and Time as Generalized Motion for Video Action Recognition[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 13045−13055. DOI: 10.1109/ICCV48922.2021.01282.

[13] KANG D, KWON H, MIN J H, *et al*. Relational Embedding for Few-shot Classification[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021: 8802−8813. DOI: 10.1109/ICCV48922.2021.00870.

[14] YANG L, LI L L, ZHANG Z L, *et al*. DPGN: Distribution Propagation Graph Network for Few-shot Learning [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 13387−13396. DOI: 10.1109/CVPR42600.2020.01340.

[15] XIE S N, GIRSHICK R, DOLLÁR P, *et al*. Aggregated Residual Transformations for Deep Neural Networks [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 5987−5995. DOI: 10.1109/CVPR.2017.634.

[16] WAH C, BRANSON S, WELINDER P, *et al*. The Caltech-Ucsd Birds-200-2011 Dataset[EB/OL]. *Caltech*, 2011:16119123. https://api.semanticscholar.org/Gorpus ID: 16119123.

[17] RUSSAKOVSKY O, DENG J, SU H, *et al*. ImageNet Large Scale Visual Recognition Challenge[J]. *Int J Comput Vis*, 2015, **115**(3): 211−252. DOI: 10.1007/s11263-015-0816-y.

[18] LUO X, WU H, ZHANG J, *et al*. A Closer Look at Few-shot Classification Again[EB/OL]. arXiv Preprint: 2301.12246, 2023. https://arXiv.org/abs/2301.12246.