



JOURNAL OF IMAGE AND GRAPHICS

主办: 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2021
07
VOL.26

ISSN1006-8961
CN11-3758/TB



中国多媒体大会(ChinaMM2020)专刊



第26卷第7期（总第303期）

2021年7月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

版权声明

凡向《中国图象图形学报》投稿，均视为同意在本刊网站及CNKI等全文数据库出版，所刊载论文已获得著作权人的授权。本刊所有图片均为非商业目的使用，所有内容，未经许可，不得转载或以其他方式使用。

Copyright

All rights reserved by Journal of Image and Graphics, Institute of Remote Sensing and Digital Earth, CAS. The content (including but not limited text, photo, etc) published in this journal is for non-commercial use.

主管单位 中国科学院

主办单位 中国科学院空天信息创新研究院

中国图象图形学学会

北京应用物理与计算数学研究所

主 编 吴一戎

编辑出版 《中国图象图形学报》编辑出版委员会

通信地址 北京市海淀区北四环西路19号

邮 编 100190

电子信箱 jig@aircas.ac.cn

电 话 010-58887035

网 址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号

总 发 行 北京报刊发行局

订 购 全国各地邮局

海外发行 中国国际图书贸易集团有限公司

（邮政信箱：北京399信箱 邮编：100048）

印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian

Monthly, Started in 1996

Superintended by Chinese Academy of Sciences

Sponsored by Aerospace Information Research Institute, CAS

China Society of Image and Graphics

Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong

Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics

Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China

Zip code 100190

E-mail jig@aircas.ac.cn

Telephone 010-58887035

Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals

Domestic All Local Post Offices in China

Overseas China International Book Trading Corporation

(P.O.Box 399, Beijing 100048, P.R.China)

Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB

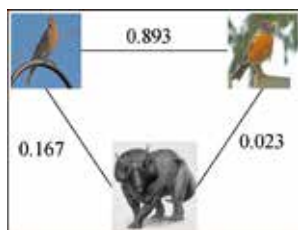
ISSN 1006-8961

CODEN ZTTXFZ

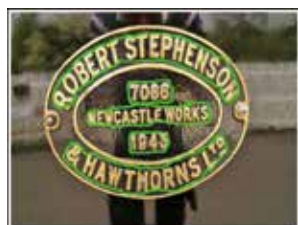
国外发行代号 M1406

国内邮发代号 82-831

国内定价 60.00元



类别语义相似性监督的小样本图像识别(第1594页)



像素聚合和特征增强的任意形状场景文本检测(第1614页)



时序特征融合的视频实例分割(第1692页)

综述

面向上行流媒体的压缩感知视频流技术前沿

刘浩, 黄荣, 袁浩东 1545

图像处理和编码

耦合保持投影哈希跨模态检索

闵康凌, 张国宾, 王磊, 李丹萍 1558

面向视觉搜索的空间局部敏感哈希方法

黄小燕, 孙彬, 杨展源, 朱映映, 田奇 1568

图像分析和识别

融合视觉关系检测的电力场景自动危险预警

高明, 左红群, 柏帆, 田清阳, 葛志峰, 董兴宁, 甘甜 1583

类别语义相似性监督的小样本图像识别

徐鹏帮, 桑基韬, 路冬媛 1594

局部双目视差回归的目标距离估计

张羽丰, 李昱希, 赵明璧, 喻晓源, 占云龙, 林巍峭 1604

像素聚合和特征增强的任意形状场景文本检测

师广琛, 巫义锐 1614

相位一致性指导的全参考全景图像质量评价

夏雨蒙, 王永芳, 王闯 1625

图像理解和计算机视觉

特征金字塔结构的时序行为识别网络

何嘉宇, 雷军, 李国辉 1637

多模态多层次事件网络的谣言检测

李莎, 张怀文, 钱胜胜, 方全, 徐常胜 1648

多模态零样本人体动作识别

吕露露, 黄毅, 高君宇, 杨小汕, 徐常胜 1658

足球视频球员感知跟踪算法

冯思佳, 宋子恺, 于俊清, 何云峰, 管涛 1668

融合时空图卷积的多人交互行为识别

成科扬, 吴金霞, 王文杉, 荣兰, 詹永照 1681

时序特征融合的视频实例分割

黄泽涛, 刘洋, 于成龙, 张加佳, 王轩, 漆舒汉 1692

医学图像处理

多尺度深度特征提取的肝脏肿瘤CT图像分类

毛静怡, 宋余庆, 刘哲 1704

3D多尺度深度卷积神经网络肺结节检测

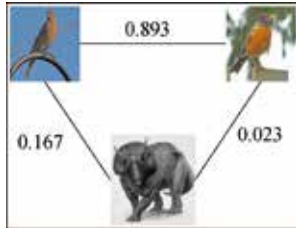
孙华聪, 彭延军, 郭燕飞, 张晓庆 1716

结合多通道注意力的糖尿病性视网膜病变分级

顾婷菲, 郝鹏翼, 白琮, 柳宁 1726

CONTENTS

JOURNAL OF IMAGE AND GRAPHICS



Few shot image recognition based on class semantic similarity supervision(P1594)



Arbitrary shape scene-text detection based on pixel aggregation and feature enhancement (P1614)



Video instance segmentation based on temporal feature fusion (P1692)

Review

Survey on compressive sensing video stream for uplink streaming media

Liu Hao, Huang Rong, Yuan Haodong 1545

Image Processing and Coding

Structure-preserving hashing with coupled projections for cross-modal retrieval

Min Kangling, Zhang Guobin, Wang Lei, Li Danping 1558

Locality-sensitive hashing approach based on semantic space for visual retrieval

Huang Xiaoyan, Sun Bin, Yang Zhanyuan, Zhu Yingying, Tian Qi 1568

Image Analysis and Recognition

Visual relationship detection-based emergency early-warning description generation in electric power industry

Gao Ming, Zuo Hongqun, Bai Fan, Tian Qingyang, Ge Zhifeng, Dong Xingning, Gan Tian 1583

Few shot image recognition based on class semantic similarity supervision

Xu Pengbang, Sang Jitao, Lu Dongyuan 1594

Object distance estimation based on stereo regional disparity regression

Zhang Yufeng, Li Yuxi, Zhao Mingbi, Yu Xiaoyuan, Zhan Yunlong, Lin Weiyao 1604

Arbitrary shape scene-text detection based on pixel aggregation and feature enhancement

Shi Guangchen, Wu Yirui 1614

Phase consistency guided full-reference panoramic image quality assessment algorithm

Xia Yumeng, Wang Yongfang, Wang Chuang 1625

Image Understanding and Computer Vision

Temporal action detection based on feature pyramid hierarchies

He Jiayu, Lei Jun, Li Guohui 1637

Multi-modal multi-level event network for rumor detection

Li Sha, Zhang Huaiwen, Qian Shengsheng, Fang Quan, Xu Changsheng 1648

Multimodal-based zero-shot human action recognition

Lyu Lulu, Huang Yi, Gao Junyu, Yang Xiaoshan, Xu Changsheng 1658

Players-aware tracking algorithm in soccer video

Feng Sijia, Song Zikai, Yu Junqing, He Yunfeng, Guan Tao 1668

Multi-person interaction action recognition based on spatio-temporal graph convolution

Cheng Keyang, Wu Jinxia, Wang Wenshan, Rong Lan, Zhan Yongzhao 1681

Video instance segmentation based on temporal feature fusion

Huang Zetao, Liu Yang, Yu Chenglong, Zhang Jiajia, Wang Xuan, Qi Shuhan 1692

Medical Image Processing

CT image classification of liver tumors based on multi-scale and deep feature extraction

Mao Jingyi, Song Yuqing, Liu Zhe 1704

3D multi-scale deep convolutional neural networks in pulmonary nodule detection

Sun Huacong, Peng Yanjun, Guo Yanfei, Zhang Xiaoqing 1716

Diabetic retinopathy grading based on multi-channel attention

Gu Tingfei, Hao Pengyi, Bai Cong, Liu Ning 1726

中图法分类号: TP399 文献标识码: A 文章编号: 1006-8961(2021)07-1594-10

论文引用格式: Xu P B, Sang J T and Lu D Y. 2021. Few shot image recognition based on class semantic similarity supervision. Journal of Image and Graphics, 26(07): 1594-1603 (徐鹏帮, 桑基韬, 路冬媛. 2021. 类别语义相似性监督的小样本图像识别. 中国图象图形学报, 26(07): 1594-1603)
[DOI: 10.11834/jig.200504]

类别语义相似性监督的小样本图像识别

徐鹏帮¹, 桑基韬¹, 路冬媛^{2*}

1. 北京交通大学计算机与信息技术学院, 北京 100044; 2. 对外经济贸易大学信息学院, 北京 100029

摘要: 目的 现有的深度学习模型往往需要大规模的训练数据, 而小样本分类旨在识别只有少量带标签样本的目标类别。作为目前小样本学习的主流方法, 基于度量的元学习方法在训练阶段大多没有使用小样本目标类的样本, 导致这些模型的特征表示不能很好地泛化到目标类。为了提高基于元学习的小样本图像识别方法的泛化能力, 本文提出了基于类别语义相似性监督的小样本图像识别方法。**方法** 采用经典的词嵌入模型 GloVe (global vectors for word representation) 学习得到图像数据集每个类别英文名称的词嵌入向量, 利用类别词嵌入向量之间的余弦距离表示类别语义相似度。通过把类别之间的语义相关性作为先验知识进行整合, 在模型训练阶段引入类别之间的语义相似性度量作为额外的监督信息, 训练一个更具类别样本特征约束能力和泛化能力的特征表示。**结果** 在 miniImageNet 和 tieredImageNet 两个小样本学习基准数据集上进行了大量实验, 验证提出方法的有效性。结果显示在 miniImageNet 数据集 5-way 1-shot 和 5-way 5-shot 设置上, 提出的方法相比原型网络 (prototypical networks) 分类准确率分别提高 1.9% 和 0.32%; 在 tieredImageNet 数据集 5-way 1-shot 设置上, 分类准确率相比原型网络提高 0.33%。**结论** 提出基于类别语义相似性监督的小样本图像识别模型, 提高小样本学习方法的泛化能力, 提高小样本图像识别的准确率。

关键词: 小样本学习; 图像识别; 特征表示; 类别语义相似性监督; 泛化能力

Few shot image recognition based on class semantic similarity supervision

Xu Pengbang¹, Sang Jitao¹, Lu Dongyuan^{2*}

1. School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China;

2. School of Information Technology and Management, University of International Business and Economics, Beijing 100029, China

Abstract: **Objective** Deep learning has made remarkable achievements in many fields such as image recognition, object detection, and speech recognition. However, most of the extraordinary achievements of these models depend on extraordinary data size. Existing deep-learning models often need large-scale training data. Building large-scale training data sets not only necessitates a large amount of manpower and material resources but are also not feasible in scenarios such as obtaining a large number of rare image class data samples. Inspired by the fact that human children can learn how to distinguish an object through a small number of samples, few-shot image classification aims to identify target categories with only a few labeled samples. Image recognition based on few-shot learning solves the problem in which a deep learning model needs

收稿日期: 2020-08-22; 修回日期: 2021-01-25; 预印本日期: 2021-02-01

* 通信作者: 路冬媛 ludy@uibe.edu.cn

基金项目: 中央高校基本科研业务费专项资金资助(2020JBM016, 2018JBZ006, 2019JBZ110); 国家自然科学基金项目(61832002, 61672518); 北京市杰出青年基金(JQ20023); 对外经济贸易大学优秀青年学者资助项目(17YQ21)

Supported by: National Natural Science Foundation of China(61832002, 61672518)

large-scale training data. At present, the mainstream methods of few-shot image recognition are based on meta learning, which mainly includes three methods: meta learning based on metric, meta learning based on optimization, and meta learning based on model. The method of meta learning is divided into two stages: training and testing. However, most of the metric-based meta-learning methods do not use few shots of the target class in the training stage, which leads to a lack of good generalization ability of these models. These metric-based meta-learning models often show high accuracy in the training stage, but the recognition effect for few-shot image categories in the test stage is poor. The deep feature representation learned by the models cannot be effectively generalized to the target class. To improve the generalization ability of the few-shot learning image recognition method, this study proposes a few-shot learning method based on class semantic similarity supervision. **Method** The method proposed in this paper mainly includes two parts: the first step is to obtain the class similarity matrix between the image dataset classes, and the second step is to use the class similarity matrix as additional supervision information to train the few-shot image recognition model. The details are as follows: a common crawl database containing one billion level webpage data is used to train an unsupervised word-vector learning algorithm GloVe model (global vectors for word representation), which generates 300 dimensional vectors for every word. For classes whose names contain more than one word, we match all the words in the training GloVe model and find their word-embedding vectors. By averaging these word-embedding vectors, we obtain the word embedding vector of the class name. Then, the cosine distance between the word-embedding vectors of classes is used to represent the semantic similarity between classes. In addition to the negative logarithm loss caused by the category labels of the original prototypical networks, this study introduces the semantic similarity measure between categories as the extra supervision information in the training stage of the model to establish the implicit relationship between the source class and few-shot target class. This condition enables the model to have better generalization ability. Furthermore, the loss of class semantic similarity can constrain the features of samples within and between classes learned by the model so that the sample features within each class are more similar, and the distribution of sample features between different classes is more consistent with the semantic similarity between categories. By introducing the loss of class semantic similarity to supervise the training process of the model, our proposed model can implicitly learn the relationship between different classes and obtain a feature representation with more constraint and generalization abilities of class sample features. **Result** This study compared the proposed model with several state-of-the-art few-shot image classification models, including prototypical, matching, and relation networks and other classic methods. In this study, a large number of experiments are conducted on miniImageNet and tieredImageNet. The results show that the proposed method is effective and competitive with the current advanced methods. To ensure fair comparison with the advanced methods, the classical paradigm of meta learning is used to train and test the model, and many experiments are conducted on the widely used 5-way 1-shot and 5-way 5-shot settings. The experimental results show that on the 5-way 1-shot and 5-way 5-shot settings of the miniImageNet dataset, the classification accuracy of the proposed method is improved by 1.9% and 0.32%, respectively, compared with the classical few-shot image recognition meta-learning method prototypical networks. In the tieredImageNet dataset on the 5-way 1-shot setting, the classification accuracy rate is improved by 0.33% compared with that in the prototypical networks. On the 5-way 5-shot setting of the tieredImageNet dataset, the proposed model achieves a competitive result compared with the prototypical networks. At the same time, several ablation experiments are conducted to verify the effectiveness of the key modules of the proposed method, and the influence of prior information of class semantic similarity on the experimental results is analyzed from multiple perspectives. **Conclusion** This study proposed a few-shot image recognition model based on class semantic similarity supervision, which improves the generalization ability and class-feature constraint ability of the few-shot image recognition model. Experimental results show that the proposed method improves the accuracy of few-shot image recognition.

Key words: few shot learning; image recognition; feature representation; class semantic similarity supervision; generalization ability

0 引言

深度学习在诸如图像识别和自然语言处理等各种认知任务中取得了巨大成功(Szegedy 等,2016)。然而,由于模型的参数众多,需要为深度神经网络提供每个类的大量带标签的数据进行训练。对于许多罕见类(如稀有鸟类)来说,构建大型训练样本数据集并不可行,这严重限制了以往模型的可扩展性。相比之下,人类的视觉系统往往只需要看到一个或几个目标类的训练实例就能够很容易地识别这些罕见类。受人类小样本学习能力的启发,人们对小样本学习(Santoro 等,2016)越来越感兴趣。在小样本学习问题中,模型拥有一组每个类都有足够训练样本的源类和一组每个类只有少数几个样本的目标类。小样本学习的目标是通过从源类中迁移知识来为只有几个样本的目标类学习得到分类器。

小样本分类问题的解决方法大多基于元学习(Snell 等,2017)。在元学习阶段,模型对源类进行采样,以模拟目标类的小样本学习条件。然后从源类中学习良好的初始条件(Finn 等,2017)、嵌入(Oreshkin 等,2018)或优化策略(Ravi 和 Larochelle, 2016)等形式的可迁移知识。元学习阶段结束后,使用学习到的优化策略通过微调(Rahman 等, 2018)或者在不更新网络权值的情况下通过前馈过程(Vinyals 等,2016)计算的方式解决目标类的小样本学习问题。然而,这些基于元学习的方法存在一个基本的局限性:模型的学习过程大多只使用源类数据,这导致模型不能很好地对目标类数据进行泛化。

基于度量的元学习方法由于其简单有效而受到广泛关注。基于度量的方法由特征编码器和度量函数两个部分组成,一是利用已有样本学习到一个特征编码器,二是利用欧氏距离等作为度量函数,预测无标签查询图像的类别。然而,基于度量学习的元学习方法在训练阶段大多也没有使用小样本目标类的样本,这就导致尽管在训练阶段模型取得了较好的结果,但是这些方法通常不能很好地泛化到目标类。之前的工作也发现从不同的域提取的图像特征任务有着显著差异。在训练阶段,特征编码器可能会从已看到的源类数据分布上过拟合编码特征,从而无法很好地泛化到未看到的目标类数据分布。同

时,本文还发现目前大多数方法没有直接对类样本特征分布进行约束,本文尝试通过对类样本特征分布进行约束去获得一个更好的特征表示。

本文提出了一种新的方法,解决在小样本图像分类设置下识别目标新类别时的特征泛化能力问题以及类样本特征分布问题。本文使用类别名称的词嵌入向量度量数据集中源类和源类、源类和目标类之间的相似性。在模型的训练过程中引入类别之间的相似性作为监督信息,类别样本的特征距离用类别之间的语义相似度进行约束,使得相同类样本的特征更紧凑,不同类样本的特征分布更符合现实世界。类相似性信息在约束样本特征学习器的同时,也在训练过程中引入了目标类的语义信息,使得模型学习的特征编码器既具有更好的类样本特征约束能力也对不可见类具有很好的泛化能力。

本文的主要贡献有:

1)提出将类别相关性作为先验知识进行整合,引入类别之间的相似性度量作为监督信息,以规范化和指导分类器权值信息的传递训练一个更具类别样本特征约束能力和泛化能力的特征表示。

2)在小样本基准数据集 miniImageNet 和 tiered-ImageNet 上进行了大量的实验,实验结果证明了提出模型的有效性。

1 相关工作

近年来,小样本图像识别成为热门话题。随着基于深度学习的方法在数据丰富的多样本设置中取得成功(He 等,2016),人们对将这种深度学习方法推广到小样本学习设置中的兴趣激增,以便视觉识别系统能够真正扩展到大量的类(例如,数百万个类)。小样本分类旨在通过学习识别仅有少量带标签样本的新类别。目前,小样本学习大多采用基于元学习的策略。

基于元学习解决小样本分类问题的方法主要分为3类。第1类,基于循环框架顺序地处理和编码新类别的少数标记图像。第2类,基于优化的方案(Tseng 等,2020)通过在元训练阶段集成微调过程,学会用几个示例图像微调模型。这些模型试图修改经典的基于梯度的优化(例如,随机梯度下降)以适应元学习场景(Munkhdalai 和 Yu,2017)。对于每一个任务,学习器从训练数据的分布中采样一组数据,

并执行少量迭代次数的随机梯度下降或其他的权重更新策略,以便为特定的任务调整参数化模型。Ravi和Larochelle(2016)提出学习器模型通过一个循环元学习器产生有效的参数更新,从而适应一个新的批次任务。Finn等人(2017)提出的MAML(model-agnostic meta-learning)及其变体(Rusu等,2019;Nichol等,2018)已经展示了令人印象深刻的结果。在这些工作中,学习器模型的参数被优化,以便能够快速适应特定的任务。第3类,基于度量的方法(Sung等,2018)通过计算查询图像和新类别的少量带标签图像之间的相似性来对查询图像进行分类。通过“学习一个好的比较指标”来解决小样本分类问题。简单来说,就是如果一个模型能够度量两幅图像的相似性,就可以用带标签的样本对一幅不带标签的输入图像进行分类。为了学习一个有效的度量模型,这些方法在训练过程中根据查询图像到少数带标签图像的距离对查询图像进行预测。

在这3类基于元学习的方法中,基于度量的方法因其简单有效吸引了大量关注。基于度量的小样本分类方法由两部分组成:1)从带标签和未带标签的图像中提取特征的特征编码器,2)将图像特征作为输入并预测未带标签图像类别的度量函数。例如,Vinyals等人(2016)提出的匹配网络(matching network, MN)将余弦相似性与循环网络一起应用,原型网络(Snell等,2017)使用欧几里德距离,Sung等人(2018)提出的关系网络(relation network, RN)使用深度卷积网络模块学习一个隐式的度量函数,Garcia和Bruna(2018)的工作中使用图神经网络(graph neural network, GNN)模块作为度量函数。然而,这些基于度量的工作大多把重心放在了度量函数上,而忽略了特征编码器提取的特征可能并没有良好的泛化性,无法推广到目标类上。Tseng等人(2020)最近的研究表明,从一个任务中提取的图像特征在不同域中的分布可能截然不同,现有的小样本分类方法的性能在域偏移的情况下会显著降低。本文的工作重点是提高基于度量的元学习小样本图像分类模型对未知域的特征提取的泛化能力,即对目标类的泛化能力;同时本文还特别关注了特征编码器的类别样本特征约束能力,即对类内样本特征分布和类间样本的特征分布进行约束,使其更符合现实环境下的类分布。

2 方法

2.1 问题定义

目前,小样本学习的主流方法是元学习。元学习又称学习去学习。在元训练阶段将训练数据集采样成不同的元任务,去学习类别数据变化时模型的泛化能力,在元测试阶段,面对新的小样本类别,不需要调整训练好的模型,就可以完成小样本类别的分类。

小样本图像识别的训练数据集中包含了许多有充足样本的类别,这些类别称为源类,在训练阶段,会从训练集中随机抽取 N 个类别,每个类别抽取 K 个样本(总共 $N \times K$ 个数据),构建一个元任务,是模型的支持集输入;再从这 N 个类别的剩余数据中分别为每个类随机抽取 m 个样本组成查询集,是模型的预测对象。这样的训练方式使得模型能够从 $N \times K$ 个数据中学会识别这 N 个类别,这样的任务形式又称为 N -way K -shot问题。

训练过程中,每次训练都会采样得到不同元任务,所以总体来看,训练包含了不同的类别组合,这种机制使得模型学会不同元任务中的共性部分,比如如何提取重要特征等,忘掉元任务中特定任务的相关部分。使用这种学习策略获得的模型,能够在很大程度上成功识别之前未见过的新类别。

2.2 模型

本文提出的用于小样本分类的类相似性监督模型是一个简单的端到端神经网络架构,总体框架如图1所示。本文的小样本图像识别方法包括两个阶段:视觉特征学习和具有学习特征的查询图像标签预测。

2.2.1 类别相似度定义

首先提供一种数据驱动的方法来生成类的相似性矩阵。具体地说,首先使用一个300维的词向量来表示每个源/目标类别名称。为了获得这些词向量,使用一个包含了十亿级别网页数据的Common Crawl数据库训练一个无监督词向量学习算法GloVe模型(Pennington等,2014),这将生成300维的向量。对于名称包含多个单词的类,本文匹配训练模型中的所有单词并找到嵌入。通过对这些词嵌入进行平均,得到类名称的词嵌入向量。

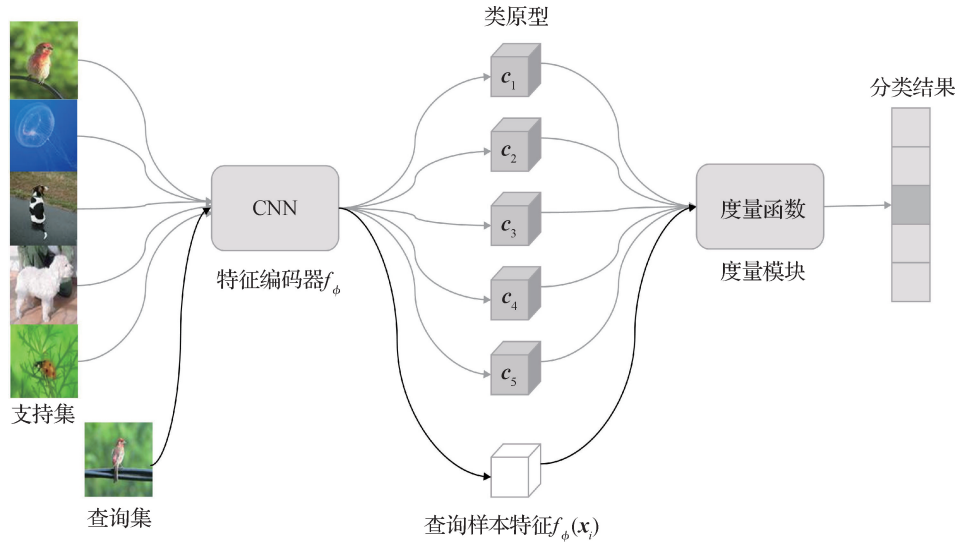


图1 模型总体框架

Fig. 1 Overall framework of the model

对于类别语义相似度矩阵,本文使用余弦相似度计算目标类与目标类、目标类与源类的相似度,以此得到类别语义相似度矩阵 $S_{[C][C]}$, C 表示数据集中目标类和源类的总个数。

2.2.2 原型网络

原型网络是一种学习器体系结构,设计用于使用很少的训练样本学习新类别。学习器使用特征编码器将带有标签的支持集图像和无标签的查询图像嵌入到特征空间中。原型网络通过一个嵌入函数 f_ϕ 将样本数据映射到 M 维的特征空间上,继而为每个类别计算出一个原型表示 c_k , k 表示类别序号。

类别的原型表示 c_k 是对支持集中类别的所有样本通过嵌入函数获得的特征向量取平均值得到

$$c_k = \frac{1}{N} \sum_{i=1}^N f_\phi(x_i) \quad (1)$$

式中, N 表示支持集的样本数量, x_i 表示支持集的样本图像, f_ϕ 表示嵌入函数, c_k 为类的原型表示。

然后,根据嵌入函数 f_ϕ 获得查询集样本 x_j 的特征 $f_\phi(x_j)$,通过计算每个类原型和查询集样本特征的欧氏距离对查询样本的类别标签进行预测。

通过原型网络可以获得查询图像 x 为每一个类的概率分布 $p_\phi(\hat{y} = k | x)$, $\hat{y}$ 表示图像 x 的预测结果。原型网络通过优化最小化负对数概率分布对模型进行训练,目标函数式为

$$L_c(y_j, f_\phi(x_j)) = -\log p_\phi(\hat{y} = k | x) \quad (2)$$

式中, L_c 表示分类概率分布的负对数, y_j 表示查询

集样本 x_j 的真实标签。

2.2.3 模型损失函数

原型网络使用单纯的最小化负对数损失并没有刻画目标类和源类的关系,由此导致了在源类数据集上训练的特征编码器并不能很好地泛化到目标类。当用词向量表示的方法来编码类与类之间的语义关系之后,本文提出了一种类相似性监督损失,将类相似性中的先验知识集成起来,用于约束小样本图像识别中的可迁移视觉特征并提高模型特征表示的泛化能力。为此,引入如下的类相似度监督损失项

$$L_s = \sum_{i=1}^N \sum_{j=1}^K |d_c(V_i - V_j) - d_t(f_\phi(x_j) - f_\phi(x_i))| \quad (3)$$

模型总的目标函数为

$$L_a(x, y) = L_c(y_i, f_\phi(x_i)) + \lambda L_s \quad (4)$$

式中, L_s 表示类相似度损失, L_a 表示模型的总损失,超参数 λ 对类相似度损失进行加权。 V_i 和 V_j 表示样本 i 和样本 j 的所属类别的词嵌入向量。 $f_\phi(x_i)$ 和 $f_\phi(x_j)$ 表示样本 x_i 和 x_j 的特征向量。 d_c 表示两个词嵌入向量的余弦相似度, d_t 表示两个样本特征向量的余弦相似度。模型在训练阶段引入类别之间的语义关系,通过类相似性损失监督项,对类内样本特征距离和类间样本特征距离进行约束。

值得强调的是,这种改变除了超参数 λ 外不添加任何参数。模型性能的所有提升都来自于增强的

特征表示,而不是来自于增加的网络容量。

2.3 标签预测

由于分类过程是非参数化的,所以训练提出的模型就相当于设置特征编码器的参数。通过对支持集图像和查询集图像进行采样形成元任务数据集,继而对模型在元任务数据集上进行训练。

一旦特征学习模型使用源类数据进行训练,就可以用于从目标类中提取图像样本的特征。利用这些视觉特征,可以直接使用一种简单的最近邻搜索方法,从中推断出目标类查询样本的标签。对于每个目标类,计算其几个样本 $\mathbf{x}_i, i \in \{1, 2, \dots, K\}$ 视觉特征的平均值作为其所属类别的原型 $\mathbf{c}_n, n \in \{1, 2, \dots, N\}$ 。在测试时,给定一幅测试图像 $\mathbf{x}_q$,首先通过预先训练的网络提取查询图像的特征表示 $f_\phi(\mathbf{x}_q)$,然后计算其与每个类原型 $\mathbf{c}_n$ 的欧氏距离,并将其类标签预测为距离最小的类标签,即

$$y_q = \{y_n | \min(d_o(f_\phi(\mathbf{x}_q), \mathbf{c}_n)), n \in \{1, \dots, N\}\} \quad (5)$$

式中, d_o 表示两个特征之间的欧氏距离, $\min$ 表示最小值函数, y_q 表示测试图像 $\mathbf{x}_q$ 的预测标签, y_n 表示类原型 $\mathbf{c}_n$ 的类标签。

3 实验

在两个小样本学习基准数据集上评估并比较了本文提出的方法和选取的基线方法,即 miniImageNet (Vinyals 等, 2016) 和 tieredImageNet (Ren 等, 2018)。将提出的方法与各种先进的小样本识别基线方法进行比较,包括匹配网络、MAML、原型网络、关系网络以及 GNN 方法。

3.1 数据集

数据集 miniImageNet 是在原始 ILSVRC-12 数据集 (Russakovsky 等, 2015) 的基础上提出并广泛使用的小样本学习基准数据集。所有图像均为 RGB 彩色,大小为 84×84 像素,从 100 个不同的类别中进行采样,每个类别均有 600 个样本。本文采用与 Vinyals 等人 (2016) 一样的数据集划分方式,训练、验证和测试集分别包括 64、16 和 20 个类别,16 个类组成的验证集仅用于分析模型的泛化能力。

数据集 tieredImageNet 与 miniImageNet 类似, tieredImageNet 也是 ILSVRC-12 数据集的一个子集。

与 miniImageNet 相比,有更多图像 (超过 700 K),从更多的类别 (608 类而不是 100 类) 中采样。重要的是,与 miniImageNet 不同, tieredImageNet 采用 WordNet 层次 (Deng 等, 2009) 分类结构,其中 608 个类中的每一个都属于从 ImageNet 中的高级节点采样的 34 个高级类别之一,每一个高层次的类别包含 10~20 个类。这 34 个高级类别划分为 20 个训练 (351 个类)、6 个验证 (97 个类) 和 8 个测试 (160 个类) 类别,共有 779 165 幅图像,平均每个类别的图像数是 1 281 幅。与 miniImageNet 数据集一样, 97 个类组成的验证集仅用于分析模型的泛化能力,不参与训练。

3.2 实验设置

3.2.1 网络架构

为了与其他方法进行公平比较,本文采用广泛使用的卷积神经网络 (convolutional neural networks, CNN) 架构 (Finn 等, 2017; Snell 等, 2017) 为特征嵌入函数。由于大多数小样本学习模型都使用 4 个卷积块来做嵌入模块,为了进行公平比较,本文遵循相同的架构设置。更具体地说,每个卷积块分别包含一个 3×3 卷积的 64 个滤波器 ($64 \times 3 \times 3$)、一个批标准化和一个修正线性单元 (rectified linear unit, ReLU) 的非线性层。每个卷积块还包含一个 2×2 的最大池化层。对于 miniImageNet 和 tieredImageNet 数据集,最后一个最大池化层的输出大小都为 $H = 64 \times 5 \times 5$,展开成一个行向量之后的维度为 1 600 维。

3.2.2 评估

与 Snell 等人 (2017) 一样,本文采用批次训练的方式,即抽样一组 N -way K -shot 训练任务来模拟 N -way K -shot 测试问题。对于这两个数据集,本文进行了 5-way 1-shot 和 5-way 5-shot 实验,这是目前主流的小样本学习设置之一。对于每个训练集或测试集批次的 N 个样本类别中的每个类,除 K 个支持样本图像外,5-way 1-shot 实验设置还包含 15 个查询图像,5-way 5-shot 实验设置还包含 10 个查询图像。这意味着在一个训练或测试集小批次中有 $15 \times 5 + 1 \times 5 = 80$ 个图像用于 5-way 1-shot 实验, $10 \times 5 + 5 \times 5 = 75$ 个图像用于 5-way 5-shot 实验。为了进行评估,本文为 1-shot 和 5-shot 分别随机抽取 15 和 10 幅查询图像构建测试批次,通过对从测试集中随机生成的 600 个查询批次的实验结果进行平均,计算出模型的小样本图像识别的分类精度。

3.2.3 训练

利用 Adam 优化器对模型进行训练,初始学习率为 5×10^{-4} ,权值衰减为 10^{-6} 。元训练的任务小批量分别设置为 40 和 20,用于 5-way 1-shot 和 5-way 5-shot 实验。对于 miniImageNet 数据集,每训练 15 000 个元任务批次对学习率减半。而对于 tieredImageNet 数据集,每训练 30 000 个元任务批次对学习率减半,因为它是更大的数据集,需要更多的迭代次数才能收敛。本文所有的代码都是在 TensorFlow 中实现的,并使用 NVIDIA TITAN XP GPU 运行。

3.3 小样本分类

将本文方法与不同设置下的几种基线方法进行比较。实验结果如表 1 所示,对于每个任务,加粗字体显示最好的结果,“-”表示模型没有对应的实验结果。从表 1 可以看出,在 miniImageNet 数据集上,本文模型在 5-way 1-shot 和 5-way 5-shot 设置上均比基线模型取得更好的结果。在 tieredImageNet 数据集上,本文模型在 5-way 1-shot 设置上比基线模型取得更好的性能,在 5-way 5-shot 设置上实现了有竞争力的结果。

表 1 在 miniImageNet 和 tieredImageNet 上的小样本分类结果
Table 1 Small sample classification results on miniImageNet and tieredImageNet

方法	miniImageNet/%		tieredImageNet/%	
	1-shot	5-shot	1-shot	5-shot
匹配网络 (Vinyals 等,2016)	43.56 ± 0.84	55.31 ± 0.73	—	—
MAML (Finn 等,2017)	48.70 ± 1.84	63.11 ± 0.92	51.67 ± 1.81	70.30 ± 0.08
原型网络 (Snell 等,2017)	49.42 ± 0.78	68.20 ± 0.66	53.31 ± 0.89	72.69 ± 0.74
关系网络 (Sung 等,2018)	50.44 ± 0.82	65.32 ± 0.70	—	—
GNN (Garcia 和 Bruna,2018)	50.33 ± 0.36	66.41 ± 0.63	—	—
本文 (类语义相似性模型)	51.32 ± 0.76	68.52 ± 0.84	53.64 ± 0.76	71.45 ± 0.55

注:加粗字体为每列最优值,“-”代表模型没有报告对应的实验结果。

3.3.1 超参数 λ 的取值

探索不同的超参数 λ 的取值对实验结果的影响。为了选择一个较好的超参数值 λ,在 miniImageNet 数据集上进行多次试验,实验结果如表 2 所示。结果表明,在 λ 取 0.1 时,在 miniImageNet 数据集上 5-way 1-shot 实验设置下,模型取得了 51.32% 的最好成绩。

表 2 不同 λ 取值在 miniImageNet 上的结果

Table 2 Comparing the results of different λ values on miniImageNet

超参数 λ	1-shot/%
0.1	51.32
0.2	50.51
0.3	50.16
0.4	49.65
0.5	49.83
0.6	50.37
0.7	49.52
0.8	49.15
0.9	49.45
1.0	49.26

注:加粗字体为最优值。

3.4 类相似性矩阵分析

本节对生成的类别相似性矩阵进行分析,了解其与现实环境的符合程度。如图 2 所示,在本文生成的语义相似度矩阵中,墨西哥红雀和美洲知更鸟的相似度为 0.893,和三角恐龙的相似度为 0.167,而美洲知更鸟和三角恐龙的相似度只有 0.023。分析可知,在本文生成的类语义相似度矩阵中,实际环境中视觉较为相似的两个类别有较大的相似度值,而实际环境中视觉差异较为明显的两个类别有较小的相似度值。所以,本文生成的语义相似度矩阵基本能反映现实情况,在很大程度上能度量两个类别之间的语义相似程度。

3.5 类相似性矩阵对结果的影响

探索不同的类相似性矩阵对模型结果的影响,以证明模型性能的提高是因为本文方法引入了较为

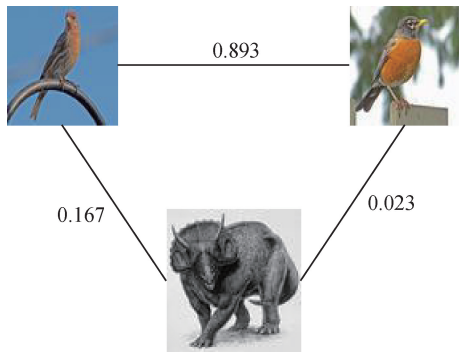


图2 类别相似度示意图

Fig. 2 Schematic diagram of class similarity

符合现实语义环境的类相似性矩阵。除了之前通过使用词向量计算语义相似性获取的相似性矩阵外,本节还分别生成如下两种类相似性矩阵:1)随机初始化的相似性矩阵:为每两个类随机生成一个范围为0~1的值为相似度;2)固定值的相似性矩阵:把每两个类之间的相似度都设为固定值0.5。在 mini-ImageNet 数据集上 5-way 1-shot 实验设置下,3 种相似性矩阵的实验结果如表 3 所示。可以看出,符合现实语义环境的类相似性矩阵能够显著提高实验结果,而随机生成的类相似性矩阵并没有提升模型性能。由此可以证明模型性能的提高是因为本文提出的方法引入了较为符合现实语义环境的类相似性矩阵。

表3 比较不同语义相似性矩阵在 miniImageNet 上的比较结果

Table 3 Comparing the results of different semantic similarity matrices on miniImageNet

类相似性矩阵形式	1-shot/%
随机初始化	48.31
固定值	47.58
语义相似度	51.32

注:加粗字体为最优值。

3.6 消融实验

将提出的模型与一些其他方法进行比较,以评估提出的方法主要部分的有效性。具体来说,比较了3种损失,每种损失都使用原型网络作为基线模型,不同之处在于训练模型使用的具体损失函数。原型分类损失是原型网络使用的负对数损失,三元组分类损失使用的是负对数损失加三元组损失

(Schroff 等,2015),而本文模型使用的是负对数损失加类相似性损失。表4显示了在标准小样本学习 5-way 1-shot 设置下,miniImageNet 数据集的上述损失的比较结果。结果显示本文的类相似性损失与仅有负对数损失的原型网络方法相比有显著改进,这为本文模型主要部分的有效性提供了有力的支持。还与加入三元组损失的模型进行了比较,结果显示提出的类相似性监督损失方法训练的模型性能比三元组损失模型(在 1-shot 设置下,将三元组损失中同类样本的距离默认为 0)略有改善。这意味着通过考虑元训练任务中类别之间的语义相关性,本文提出的类别语义相似性监督模型的结果更好。

表4 miniImageNet 上模型关键部分的消融研究

Table 4 Ablation study for key components on miniImageNet

模型	1-shot/%
原始分类损失	49.42
三元组分类损失	50.58
类相似性损失	51.32

注:加粗字体为最优值。

3.7 样本特征可视化

为了更加直观地说明类相似性监督模块对模型学习到的特征表示的作用,本节使用 van der Maaten 和 Hinton(2008)提出的 t-SNE(t-distributed stochastic neighbor embedding)对查询样本的特征分布进行可视化,结果如图3所示。本节使用 miniImageNet 数据集在 5-way 1-shot 设置下计算目标类查询样本的特征,但每个类仅显示 100 个样本以便于比较。因此,在采样任务以创建这些可视化特征时,首先选择了 5 个类别,并从这 5 个类别数据集中分别选择 100 个样本获取特征并进行可视化。图 3(a)模型在不使用类相似性模块的情况下达到了 49.42% 的精度,改进后的图 3(b)模型在使用类相似度的情况下,性能提高到了 51.32%。

可以清楚地看到,类相似性监督模型具有更紧密和可分离的聚类,这表明模型学习到的特征对任务的区分性更强。这源于类别相似性监督模块的设计,本文提出的模型能够更好地约束类别内的样本以及建立源类和目标类的关系,显著类间差异。如果没有类相似性监督,一些不同类的特征分布会相

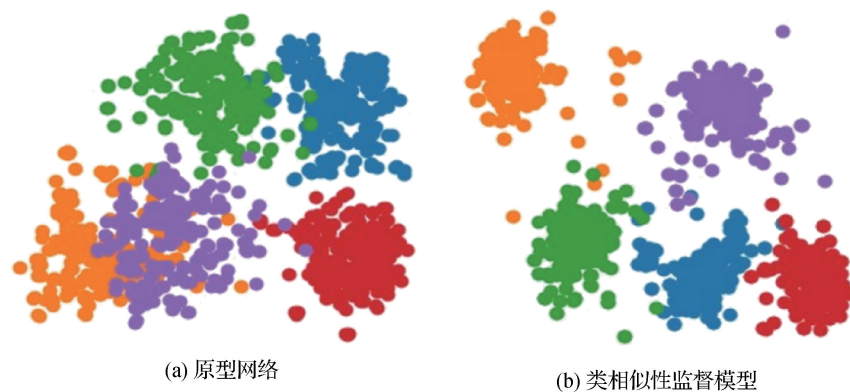


图3 t-SNE 可视化查询样本的特征分布图

Fig. 3 Feature distribution of t-SNE visual query samples ((a) prototype network; (b) class similarity supervision model)

互重叠(例如紫色和橙色),使得度量模块难以区分不同类的样本特征。

4 结 论

针对目前小样本图像识别模型缺乏泛化性这一问题,本文提出了一种类别语义相似性监督的小样本图像识别模型。模型使用类别名称的词向量嵌入作为先验信息度量类别之间的相似性,通过引入类别之间的语义相似性作为额外的监督信息,学习了一个更具泛化能力和类样本特征约束能力的特征表示。在 miniImageNet 和 tieredImageNet 两个小样本学习基准数据集上的实验结果显示提出的方法相比原型网络具有更高的准确率和更好的特征泛化能力。

然而,本文的方法依然存在不足:对于类别语义相似性这一先验信息的利用较为粗糙,以及没有构建一个足够完美的类相似性矩阵。本文的后续工作将会继续探索如何构建一个更好的类相似性矩阵,以及如何更有效地利用类相似性矩阵等先验知识作为监督信息进一步提高小样本学习模型的特征表示能力。

参考文献 (References)

- Deng J, Dong W, Socher R, Li L J, Li K and Li F F. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA; IEEE: 248-255 [DOI: 10.1109/CVPR.2009.5206848]
- Finn C, Abbeel P and Levine S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks [EB/OL]. [2020-05-15]. <https://arxiv.org/pdf/1703.03400.pdf>
- Garcia V and Bruna J. 2018. Few-shot learning with graph neural networks [EB/OL]. [2020-05-15]. <https://arxiv.org/pdf/1711.04043.pdf>
- He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA; IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]
- Munkhdalai T and Yu H. 2017. Meta networks//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia; 2554-2563
- Nichol A, Achiam J and Schulman J. 2018. On first-order meta-learning algorithms [EB/OL]. [2020-05-15]. <https://arxiv.org/pdf/1803.02999.pdf>
- Oreshkin B N, Rodriguez P and Lacoste A. 2018. TADAM: task dependent adaptive metric for improved few-shot learning//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada; Curran Associates Inc.: 721-731
- Pennington J, Socher R and Manning C. 2014. GloVe: global vectors for word representation//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar; Association for Computational Linguistics: 1532-1543 [DOI: 10.3115/v1/D14-1162]
- Rahman S, Khan S and Porikli F. 2018. A unified approach for conventional zero-shot, generalized zero-shot, and few-shot learning. IEEE Transactions on Image Processing, 27(11): 5652-5667 [DOI: 10.1109/TIP.2018.2861573]
- Ravi S and Larochelle H. 2016. Optimization as a model for few-shot learning [EB/OL]. [2020-05-15]. <https://openreview.net/pdf?id=rJY0-Kcll>
- Ren M Y, Triantafillou E, Ravi S, Snell J, Swersky K, Tenenbaum J B,

- Larochelle H and Zemel R S. 2018. Meta-learning for semi-supervised few-shot classification [EB/OL]. [2020-05-15]. <https://arxiv.org/pdf/1803.00676.pdf>
- Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Hong Z H, Karpathy A, Khosla A, Bernstein M, Berg A C and Li F F. 2015. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115(3): 211-252 [DOI: 10.1007/s11263-015-0816-y]
- Rusu A A, Rao D, Sygnowski J, Vinyals O, Pascanu R, Osindero S and Hadsell R. 2019. Meta-learning with latent embedding optimization [EB/OL]. [2020-05-15]. <https://arxiv.org/pdf/1807.05960.pdf>
- Santoro A, Bartunov S, Botvinick M, Wierstra D and Lillicrap T. 2016. One-shot learning with memory-augmented neural networks [EB/OL]. [2020-05-15]. <https://arxiv.org/pdf/1605.06065v1.pdf>
- Schroff F, Kalenichenko D and Philbin J. 2015. FaceNet: a unified embedding for face recognition and clustering [EB/OL]. [2020-05-15]. <https://arxiv.org/pdf/1503.03832v3.pdf>
- Snell J, Swersky K and Zemel R S. 2017. Prototypical networks for few-shot learning [EB/OL]. [2020-05-15]. <https://arxiv.org/pdf/1703.05175.pdf>
- Sung F, Yang Y X, Zhang L, Xiang T, Torr P H S and Hospedales T M. 2018. Learning to compare: relation network for few-shot learning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA; IEEE: 1199-1208 [DOI: 10.1109/CVPR.2018.00131]
- Szegedy C, Ioffe S, Vanhoucke V and Alemi A. 2016. Inception-v4, inception-resnet and the impact of residual connections on learning [EB/OL]. [2020-05-15]. <https://arxiv.org/pdf/1602.07261v2.pdf>
- Tseng H Y, Lee H Y, Huang J B and Yang M H. 2020. Cross-domain few-shot classification via learned feature-wise transformation [EB/OL]. [2020-05-15]. <https://arxiv.org/pdf/2001.08735.pdf>
- van der Maaten L and Hinton G. 2008. Visualizing data using t-SNE. Journal of Machine Learning Research, 9(86): 2579-2605
- Vinyals O, Blundell C, Lillicrap T, Kavukcuoglu K and Wierstra D. 2016. Matching networks for one shot learning//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain; Curran Associates Inc.: 3630-3638

作者简介



徐鹏帮, 1996 年生, 男, 硕士研究生, 主要研究方向为小样本学习。

E-mail: 969847047@qq.com



路冬媛, 通信作者, 女, 副教授, 主要研究方向为人工智能下自然语言处理领域中的应用研究, 如语义分析、情感分析、主题建模、知识感知应用等。

E-mail: ludy@uibe.edu.cn

桑基韬, 男, 教授, 主要研究方向为多媒体计算、社交媒体分析与应用、数据挖掘。E-mail: jtsang@bjtu.edu.cn