Computer Science and Application 计算机科学与应用, 2019, 9(12), 2218-2230
Published Online December 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.912247

Hans汉斯

Research of Improved Fine-Grained Image Classification Based on Saliency

Huishi Wu[1], Lianglun Cheng[1], Fangxiong Chen[2]

[1]Department of Computer, Guangdong University of Technology, Guangzhou Guangdong
[2]Shenzhen H&T Intelligent Control Co., Ltd., Shenzhen Guangdong
Email: whsgdut1214@163.com

Received: Nov. 13th, 2019; accepted: Nov. 25th, 2019; published: Dec. 2nd, 2019

Abstract

In view of large intra-class differences, small differences between classes and the problems of dependency on data annotation in fine-grained images, an algorithm based on saliency fusion to improve fine-grained image classification is proposed. This paper introduced a two-input deep neural network, which integrated two components in a single framework: the salient feature fusion structure and the feature extractor. Firstly, the SALICON saliency detection algorithm is used to generate the saliency map. The original RGB image is fused with the saliency map according to the fusion network structure. Secondly, in order to make full use of higher resolution, the modulation potential of the salient features, maximum pooling operation is used to reduce the dimensionality of the data space so that the modulation potential of higher resolution salient features can be fully utilized. Finally, with the help of migration learning, the deep neural network model Inception_V3.0 pre-trained on the ImageNet dataset is used as the basic feature extraction model to extract high-level semantic features. The comparison experiments in the public datasets CUB200-2011 and Stanford Dogs show that the classification accuracy of the algorithm is 84.36%, 84.94%, compared with Part R-CNN, LRBP and other mainstream fine-grained classification algorithms, this method can achieve better classification results.

Keywords

Fine-Grained Image Classification, Convolutional Neural Network (CNN), Saliency Detection Algorithm, Saliency Map, Feature Fusion

基于显著性融合的细粒度图像分类方法研究

吴慧诗[1]，程良伦[1]，陈仿雄[2]

[1]广东工业大学计算机学院，广东 广州
[2]深圳和而泰家居在线网络科技有限公司，广东 深圳
Email: whsgdut1214@163.com

收稿日期：2019年11月13日；录用日期：2019年11月25日；发布日期：2019年12月2日

文章引用: 吴慧诗, 程良伦, 陈仿雄. 基于显著性融合的细粒度图像分类方法研究[J]. 计算机科学与应用, 2019, 9(12): 2218-2230. DOI: 10.12677/csa.2019.912247

摘 要

针对细粒度图像存在的类内差异大、类间差异小和依赖数据标注的问题，提出了一种基于显著度融合改进细粒度图像分类的算法。该算法基于一种双输入的深度神经网络，包括显著性特征融合结构和特征提取网络两个部分。首先，根据Fusion层网络结构将原RGB图与显著图进行特征融合，显著图是由SALICON显著性检测算法计算产生；其次，为充分利用更高分辨显著特征的调制潜力，利用最大池化操作对数据空间进行降维操作；最后，借助迁移学习思想，把在ImageNet数据集上预训练好的深度神经网络模型Inception_V3.0作为基础特征提取模型，进一步提取高层语义特征。在公开数据集CUB200-2011和Stanford Dogs中进行对比实验，结果表明，该算法的分类准确率分别达到84.36%、84.94%，相较于Part R-CNN、LRBP等多个主流细粒度分类算法，本文方法能取得更好的分类效果。

关键词

细粒度图像分类，卷积神经网络，显著性检测算法，显著图，特征融合

1. 引言

细粒度图像分类又被称为子类别图像分类，目的是对粗层级的大类别进行更加细致的子类划分。与之相关的研究课题主要包括识别不同种类的鸟[1]，犬[2]，飞机模型[3]、花种[4]及车[5]等。细粒度图像容易受姿态、光照、遮挡、背景干扰等诸多不确定因素的影响[6]，使得子类别具有类内相似度大而类内相似度较小的特点。而且细粒度图像的信噪比很小，不同子类别间的差异主要表现在细微的局部区域上，若想要挖掘足够区分度的信息，通常需要借助人工数据标注和图像类别标签。面对复杂、耗时、昂贵、易错的标注程序，进一步研究如何降低人工标注成本与有效利用局部区域信息成为细粒度图像分类算法未来发展的趋势。

细粒度图像分类的研究经历了传统人工设计阶段和深度学习阶段。基于人工特征的算法最先采用视觉词袋模型[7]。Wah 等人[1]提出基准方法，定位图像局部区域并将视觉词模型编码后的特征输入到 SVM 分类器进行训练，但图像分类的准确率仅仅只达到 10.3%，主要是局部定位不准和人工设计的特征表达能力不强。因此 POOF [8]，Fisher-encoded SIFT [9]，KDES [10]等新的特征描述子被提出，分类准确度提高到 50%~62%左右。同时也有针对局部区域的算法研究，如尝试使用模板匹配的方法来减少滑动窗口的计算代价。这一阶段的研究不仅受限于局部定位以及人工特征的表达能力，而且严重依赖于人工标注信息。昂贵的标注成本和弱泛化能力限制了基于人工特征的算法在细粒度图像分类中的实际应用。

近年来，卷积神经网络的提出促进了图像分类领域的快速进步，深度神经网络可以提取图像中的高层语义特征以弥补人工设计在特征表达能力上的不足，其研究可以分为基于强监督学习的算法和基于弱监督学习的方法。基于强监督信息的图像分类包括几个代表性模型：PartR-CNN [11]、Posenormalized CNN [12]、MASK-CNN [13]、HSnet [14]等。Zhang 等人[11]提出的 Part R-CNN 模型利用 R-CNN 算法进行对象与局部区域的检测，然后利用局部特征训练分类器，在 CUB200-2011 数据集上获得 73.89%的准确度。Branson 等人[12]提出姿态归一化 CNN 算法，增加局部图像的姿态对齐操作来解决类内方差大的问题。

最终分类准确度达到 75.7%。WEI [13]等提出掩码 CNN (mask-CNN)算法，通过使用全连接网络定位目标并将目标分割为两个 mask 联合判别。这类算法虽然分类精度比较高，但需要依赖人工标注数据实现局部特征的获取，实用性不强。因此细粒度图像分类的研究开始转向仅依赖于图像类别标签的弱监督学习算法。Xiao 等人[15]类比强监督学习的对象类别(object-level)和局部(part-level)类别，提出一种名为两级注意力(Two level attention)模型，获得了 77.9%的准确度，已好于一般的强监督学习算法。Simon 等人[16]提出一种新颖的局部区域定位方法，在 CUB200-2011 数据集上达到 81.01%的准确率。以上方法都是先检测局部区域进行训练后分类，没有做到端对端(end-to-end)的训练和优化[6]。为此，Lin 等人[17]提出一个实现端到端训练的 B-CNN 网络结构，分上下两层网络完成定位和特征提取步骤，分类精度很高，且仅依赖类别标签，而无需借助其他的图像标注信息。2017 年 FU [18]等提出了循环注意 CNN (RA-CNN)用以解决细粒度图像集的分类问题，该网络在 Stanford Dogs 等三个主流细粒度图像集上准确率提升近 4%。然而，该网络不能很好利用图像全局信息，选取的特征区域形状较为单一。

通过上述文献研究发现，主要存在以下问题：在不需要人工标签的情况下，如何精准定位局部区域并获得具有判别性的局部特征。针对这个问题，本文提出一种新的双输入的深度神经网络，主要贡献有两个方面：1) 选择一种由图像数据指导的显著性检测优化算法及特征提取的基础优化模型，以获取图像具有判别性的视觉特征；2) 证明显著性检测算法选择的区域涵盖图像中最丰富的信息，利用基于图像数据的显著图可强化显著区域对分类结果的影响，从而提高细粒度图像分类的性能。

2. 相关基本原理

2.1. 显著性检测原理

图像的显著性凸显人类对信息丰富区域的专注度，表示图像中的不同区域所呈现的不同视觉感观。一般来说，人类的视觉注意力焦点通常集中在某一局部区域，且这些区域往往能够很好的展现出图像所表达的最重要的内容。根据引起注意的原因，显著性检测主要包括基于图像数据的自下而上注意力机制和基于任务的自上而下注意力机制。自下而上模型是一个低级认知过程，自动地识别给定场景中最为显著的一个或者多个对象区域；自上而下模型将注意力集中在特定类别对象并自动定位目标所在的位置。本文采用自下而上显著检测算法自适应地检测图像中最为显著的局部区域，最终产生图像的显著图。

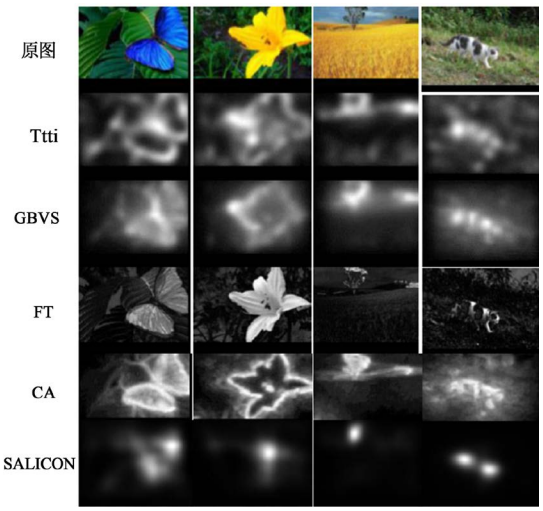

Figure 1. Saliency diagram generated by different significance detection algorithms
图 1. 不同显著性检测算法产生的显著图

图 1 描述的是常见的显著性检测算法所呈现的不同显著图。本文选取的是 SALICON 算法，该算法有两个关键要素：1) 基于显著性评估指标的目标函数；2) 在不同图像尺度上整合信息。不同于传统的显著性检测算法，其充分利用在 ImageNet 上预训练的 CNN 高级语义的强大功能，明显减少计算模型和人类感知之间的语义差距，获取图像中更加符合人类的视觉关注区域。

2.2. 迁移学习

迁移学习(Transfer Learning)实现不同但相关领域之间的知识迁移，它是解决带标签训练样本数据难获取，模型训练成本高这一基础问题的重要手段。如图 2 所示，传统的机器学习，面向不同的任务都需要重新进行学习训练。图 3 表示在迁移学习的学习过程中可以将源任务所存储的知识迁移到新的任务中，以减少训练的次数。

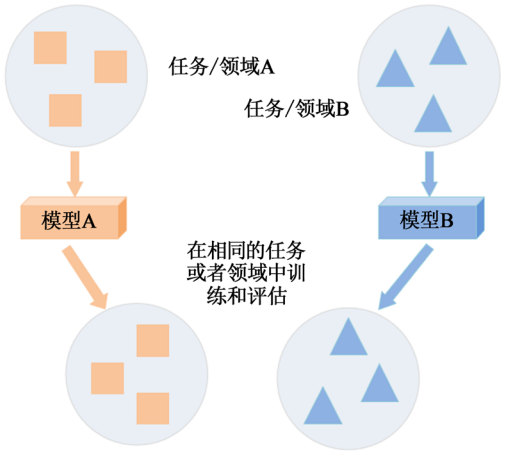

Figure 2. Traditional machine learning
图 2. 传统机器学习

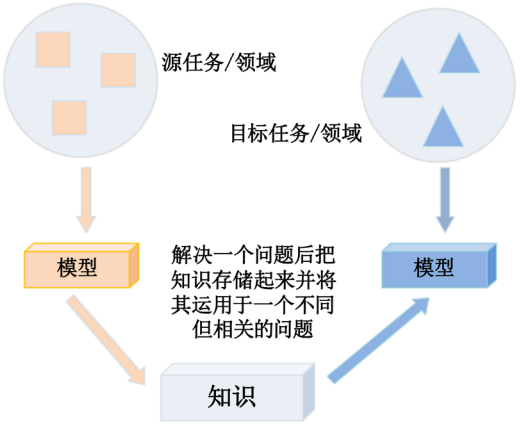

Figure 3. Migration learning
图 3. 迁移学习

深度学习中常用的特征提取网络由于网络层次深，神经元数量大，所以在进行网络训练时需要足够可利用的数据样本。对于小样本数据，即使辅以大量的、长时间的训练，也难以取得良好的训练模型和网络收敛效果。为验证在小样本数据下，显著图能够改进分类模型，本文借助迁移学习思想，利用其强大的泛化性能初始化基础特征提取网络模型，其中基础特征提取网络预先在大型图像数据集 ImageNet 上

训练好，然后通过细粒度图像小样本数据对网络模型进行微调训练，修改全连接层的类别数完成图像分类任务。实验证明迁移学习在细粒度图像分类效果稳定的同时，可显著降低深度学习所需的硬件资源。

3. 基于显著性融合的双输入深度神经网络模型

为验证在细粒度图像分类任务中，显著性能够自动捕获和感知图像中具有判别性的局部区域，本研究提出一种基于显著性融合的双输入深度神经网络模型。如图 4 所示，该网络包括显著性特征融合结构和特征提取网络两个部分。Fusion 网络结构融合显著图信息和原 RGB 图特征信息，显著图作为引导特征提取过程的注意力机制，预先通过显著性检测算法 SALICON 计算产生。两次融合网络不仅可以增加局部特征的权重，还能够保留原图的整体特征，融合后的特征对图像某个具有判别性的局部区域有高度响应。经最大池化层对数据空间进行降维操作后，采用 Inception_V3 作为基础的特征提取网络模型，进一步提取图像中的高层语义特征，得到最终的分类结果。

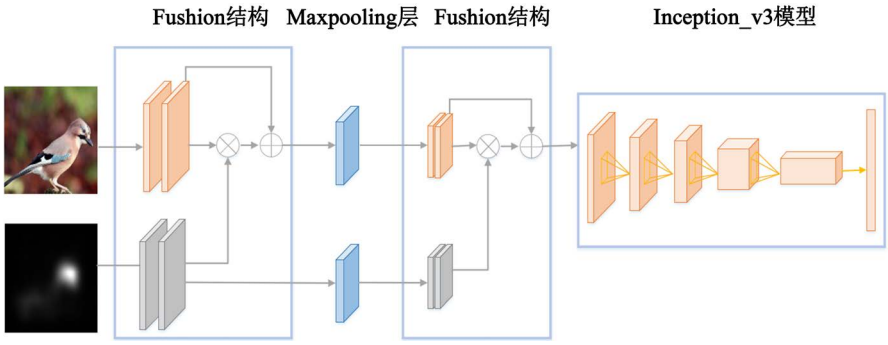

Figure 4. The overall structure of a two-input deep neural network based on saliency fusion
图 4. 基于显著性融合的双输入深度神经网络的整体结构

图 5 描述提出的模型训练算法流程图。首先令 $x_1, x_2, \cdots x_n$ 为 n 个图库数据集，相应的类别标签为 $L_\varphi \in [1, 2, \cdots n]$。采用显著性检测算法 SALICON 产生 x_i 对应的显著图 Y_i。由于图像数据集的图像大小可能不完全相同，而 Inception_V3 的输入图像尺寸固定为 $299 \times 299 \times 3$，因此将输入网络模型的图像统一裁剪为 299×299 大小。在 ImageNet 数据集上预训练 Inception_V3 网络后初始化本文中的模型。设置每个数据集训练时的批量大小 m、学习率 α。用 n 个图库图像集中的所有图像集进行训练，且用权值 W 初始化卷积神经网络，通过正向传播，得出每一个类别的预测值，最后通过最小化遍历所有图库集实例上的重建误差反向传播来执行随机梯度下降算法，当损失函数值低于某个阈值时停止训练。

基础特征提取的卷积神经网络正反向传播过程如下：

a) 正向传播。采用激励－响应机制计算每个神经元的输入，进行基础特征提取卷积神经网络的正向传播，每个神经元的输出如公式(1)所示：

$$x_n^i = f\left(y_n^i\right) = f\left(\sum_{j=0}^{C_{n-1}} w_n^{ji} x_{n-1}^j\right) \tag{1}$$

其中：x_n^i 表示第 n 层的第 i 个神经元的输出；w_n^{ji} 为第 $n-1$ 层的第 j 个神经元到第 n 层的第 i 个神经元的网络权值；C_{n-1} 为第 $n-1$ 层神经元的个数。

b) 反向传播。对于一系列神经元互联形成的卷积神经网络，反向传播用于学习网络的权值。

$$\Delta w_n^{ji} = -\alpha \frac{\partial L(y,t)_n}{\partial w_n^{ji}} \tag{2}$$

其中：Δw_n^{ji} 表示每次反向传播后权值的变化量；$L(y,t)$ 表示第 n 层的输出误差，具体计算步骤已在公式(3)中详细说明。

计算出第 n 层权值 w_n^{ji} 的增加量，第 $n-1$ 层 $\frac{\partial L(y,t)_{n-1}}{\partial x_{n-1}^i}$ 计算公式为：

$$\frac{\partial L(y,t)_{n-1}}{\partial x_{n-1}^i} = \sum w_n^{kj} \frac{\partial L(y,t)_n}{\partial x_n^i} \tag{3}$$

按以上公式类推计算每一层，可以得到：

$$w_n^{now} = w_n^{pre} - \alpha \frac{\partial L(y,t)_n}{w_n} \tag{4}$$

其中：w_n^{now} 表示更新后的权值，w_n^{pre} 表示更新前的权值。

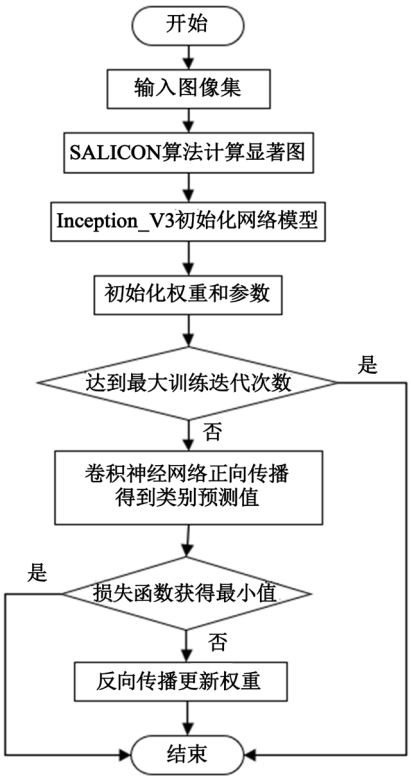

Figure 5. Flow chart of the network model training algorithm
图 5. 网络模型训练算法流程图

3.1. Fushion 层网络结构

显著性特征融合结构的一个输入为原始图像 $I(x,y,z)$，另一个输入为由显著性检测算法 SALICON 所产生的 $I(x,y,z)$ 相对应显著图 $S(x,y)$，其中 x 和 y 指空间坐标，$z=\{1,2,3\}$ 代表图像的三个颜色通道。融合到图像分类网络后，图像的颜色通道为四个，即 $I(x,y,4)=S(x,y)$。Murabito 等人曾将两个 CNN 网络组合，一个网络用来计算 RGB 图像的自顶向下显著图，另一个网络将生成的显著图附加到 RGB 图像通道以执行图像分类。分类网络只需训练第一层的权重，便可使用预训练网络初始化网络权重。这称为显著性和图像数据的早期融合。

本研究进一步提出显著图和原图像内容融合结构，使用显著性特征图映射来调制中间网络层的特征，对应的标签类别作为输出而进行训练。显著性分支和原图分支使用类似的体系结构以确保显著图具有与原图相同的空间维度。主要差异体现在通道尺寸的大小。假设第 h_i 层网络的尺寸为 $W_i * H_i * C_i$，为避免显著图与原图特征融合后原图分支的特征被完全忽略，在进行乘法操作时再次强调原图的特征。由于显著图与特征图的数量无关，只考虑单个显著图 S。融合层方程式计算如下：

$$h_i = f\left(h_i(w,h,c) * \left[S(x,y)+1\right]\right) \tag{5}$$

其中：f 为 ReLU 激活函数，与原图网络结构采用 ReLU 激活函数不同，本研究选择在显著性分支的末尾使用 Sigmoid 激活函数，使 $0 \le S(x,y) \le 1$，从而为特征调制提供合适的范围。

在一般的 CNN 结构中，最大池化操作在第二个卷积层后执行，但这没有考虑显著图能否在特征调制层上执行最大池化。因此，本研究将最大池操作推迟至显著性和原图这两个分支的特征融合后，以便充分利用更高分辨显著特征的调制潜力。

3.2. 目标函数优化

在多分类任务中，一般选择归一化指数函数即 Softmax 函数作为分类器，Softmax 利用指数函数的映射特性，把函数取值范围映射到 $(0,1)$ 区间。在最后选取输出节点时，通过选取概率最大的结点来获得分类结果，计算公式如下：

$$p(i) = \frac{e^{W^T x_i + b}}{\sum_{i=1}^{n} e^{W^T x_i + b}} \tag{6}$$

其中：W 表示网络中神经元的权重，n 表示图像类别总数，b 表示偏差值。

本研究网络模型的损失函数由 L_c 和 L_s 线性组合而成。在分类器输入处添加批量归一化模块，实施零均值和单一标准偏差分布，以便提高网络训练和收敛速度，获得多重损失函数的最小值。网络模型的损失函数计算公式如下：

$$\begin{aligned} L(y,t) &= L_c + L_s \\ &= -r\sum_{i=1}^{n} I(i=t)\log(p_i) + \mu\frac{1}{h\times w}\sum_{i=1}^{h}\sum_{j=1}^{w}\left(Y_{i,j} - T_{i,j}\right)^2 \end{aligned} \tag{7}$$

其中：n 表示图像样本数据中的类别数目，$I(i)$ 表示指标函数，若 i 取值为真则函数 $I(i)$ 返回 1，L_c 表示 Softmax 输出向量 p_i 与正确类 t 的交叉熵分类损失值；L_s 则是显著图 Y 与真实图像 t 之间的均方误差(MSE)损失值，h 和 w 是显著图的大小指标。r 和 μ 为权重参数，取值范围为 $[0,1]$。

4. 实验及结果分析

本文程序由 Python3.6 编写完成。实验采用的计算机硬件环境为 Intel(R) Core(TM) i9-7900X CPU@3.30GHz，内存大小 64GB，GTX1080 Ti 显卡。软件环境选择 Ubuntu16.04 系统和 TensorFlow 深度学习框架。

4.1. 数据集与性能评价

本文算法选择在两个公开的细粒度图像数据集上进行实验：加利福尼亚理工学院鸟类数据库 Caltech-UCSD Birds200 (CUB-200-2011)和斯坦福大学的犬类图片数据集 Stanford Dogs。CUB-200-2011 包含 200 不同种鸟类，共计 11788 张鸟类图片。每张图片都有不同的姿态，但差异往往只存在于比如鸟嘴、翅膀颜色等局部区域。因此在模型训练和测试仅仅采用图像的标签数据，如图 6 所示。Stanford Dogs

数据集提供了来自世界各地的 120 种狗的共 20580 张图片。包括各种不同视角和姿态，由 ImageNet 的图像和标注构建。如图 7 所示，列举两个不同品种的狗，从图中可知这类数据集存在背景复杂的特点，需要通过一定的数据预处理以降低背景的影响。

(a) Black_footed_Albatross

(b) Sooty_Albatross

Figure 6. Dataset of CUB-200-2011
图 6. CUB-200-2011 数据集

(a) Chihuahua

(b) Maltese_dog

Figure 7. Dataset of Stanford Dogs
图 7. Stanford Dogs 数据集

基于上述不同的数据集，本文使用分类准确度 accuracy 指标来检验细粒度图像分类模型的训练效果，其计算公式如下：

$$accuracy = \frac{n_t}{n} \tag{8}$$

其中：n 表示总测试样本数量，n_t 表示分类预测正确的图像个数，能够比较直观地反映出模型的分类性能。

4.2. 数据预处理

由于细粒度图像样本少，为避免出现过拟合的现象，实验开始前对图像进行预处理(去噪、降维、归一化、标准化等)以及数据扩充操作。

1) 裁剪相同尺度。不同的细粒度图像数据集图像大小不同，由于本文采用了 Inception_V3 网络在

ImageNet 数据上进行预训练，而 Inception_V3 的输入图像尺寸固定为 $299\times299\times3$，因此将输入网络模型的图像统一裁剪为 299×299 大小。

2) 数据归一化。由于数据每一维度的数值变化范围有所差异，为降低模型的分类误差，需要事先对原始图像序列进行处理，将这些像素值都除以 255，使得每一个通道的数值变化范围缩放到 0~1，且每一个维度具有零均值和单位方差。这不仅可以加快神经网络的收敛速度，还能够防止梯度弥散。

3) 数据扩充。考虑到网络参数的数量巨大，可能出现过拟合现象，但受限于图像数据量，往往需要对现有数据进行扩充处理。在本实验中对细粒度图像数据集采用多种方法来扩充数据，包括随机翻转及扭曲图像、随机裁剪图像、随机添加噪音、随机修改图像的对比度和饱和度等。使得细粒度图像每一类的训练样本数量保持相对均衡。

同时把每个数据集训练时的基本学习率的初始值 ℓ 设置为 0.001，在迭代的过程，根据衰减函数 $d=1/t$ 进行逐渐减少，其中 t 的计算公式为 $t=\dfrac{\ell}{\left(1+10^{-5}\cdot i\right)}$，$i$ 为训练迭代次数。

4.3. 实验结果与分析

4.3.1. 不同分类算法的比较

考虑到现有细粒度图像分类算法在设计相关算法模型时存在针对不同数据集表现相差较大的问题，为验证本文提出的模型在不同的细粒度数据集上具有鲁棒性和先进性。将本文算法分别在 Stanford Dogs 和 CUB-200-2011 数据集上，与其他算法的分类精度进行对比。由于 CUB-200-2011 数据集中的数据提供了关键点信息，因此在进行对比实验时与强监督学习方法进行比较。实验结果如表 1 和表 2 所示。

Table 1. Classification accuracy on the Stanford Dogs dataset
表 1. Stanford Dogs 数据集上的分类准确度

不同算法	Accuracy (%)
PD [19]	75.00
Part R-CNN [11]	73.89
姿态归一化 CNN [12]	75.70
B-CNN [17]	82.66
RA-CNN [18]	85.05
DVAN [20]	80.23
FCAN [21]	81.57
本文算法	84.36

Table 2. Classification accuracy on the CUB-200-2011 dataset
表 2. CUB-200-2011 数据集上的分类准确度

不同算法	标注框	关键点	Accuracy (%)
Part R-CNN [11]	√	√	79.89
姿态归一化 CNN [12]	√	√	80.73
Two level attention [13]			82.95
B-CNN [17]			84.10
HSnet [14]	√	√	85.69
LRBP [22]			83.57
本文算法			84.94

从表 1 可以看出，在 Stanford Dogs 数据集上，本文算法的分类精度远高于 PD、Part R-CNN 和姿态归一化 CNN 算法，与目前在该数据集上分类精度最高的 RA-CNN 算法相当。

从表 2 可以看出，在 Stanford Dogs 数据集上，本文算法比 Two level attention、B-CNN、LRBP 等基于弱监督学习算法的分类精度分别高出 1.99%，0.84%，1.37%，和基于强监督信息模型的 B-CNN 分类精度相当，比目前最新的基于强监督学习算法的 HSnet 算法仅低 0.75%。同时，较之 R-CNN 这一基于强监督算法，分类精度提高 5%左右，因此可以得出结论，融合显著性特征，在整体上优于单独使用某一特征。即显著图可以提高模型的性能。

4.3.2. 不同 CNN 基础模型的识别效果分析

本实验分别采用目前广泛使用的深度网络结构 VGG16，ResNet50，Inception_V3，Inception_ResNet 和 DenseNet 作为原图分支的基础网络结构，目标是从中选取对分类效果影响最大的 CNN 基础网络模型。25,000 个训练样本从包含 250 类汽车共 25500 张图像的 Cars 数据集中选取，每个类别车型的数量大致为 100。在 CUB-200-2011 数据集中随机选取 500 个测试图像作为测试样本。采用在 500 个测试样本检测速度的平均值与分类准确度进行性能评价。

Table 3. Classification accuracy of different CNN models on different datasets
表 3. 不同 CNN 模型在两种数据集上的分类准确度

网络结构	Cars	Birds	测试速度(pic/s)
VGG16	0.581	0.682	1.34
VGG16 + Salienc 分支(SALICON)	0.603	0.721	2.78
ResNet50	0.642	0.755	3.12
ResNet50 + Saliency 分支(SALICON)	0.687	0.768	4.23
Inception_V3	0.701	0.761	3.85
Inception_V3 + Saliency 分支(SALICON)	0.724	0.789	4.43
Inception_ResNet	0.718	0.770	6.16
Inception_ResNet + Saliency 分支(SALICON)	0.741	0.793	7.91
DenseNet	0.731	0.781	6.72
DenseNet + Saliency 分支(SALICON)	0.753	0.802	8.31

如表 3 所示，随着网络深度加深，图像分类准确度存在不同程度的提升。整体上网络结构 DenseNet 的分类性能最好，Inception_V3 和 ResNet50 网络结构的分类性能相差不大，在分类精度上次之，VGG16 网络结构性能最差。但网络深度加深的同时，测试时间明显增加，本研究将识别速度和分类准确度作为考量指标，最终选择 Inception_V3 作为基础网络。

4.3.3. 不同显著性检测算法对图像分类的影响

关于不同的显著性检测算法产生的显著图，其侧重的显著区域差异较大，已在 1.1 节中详细阐述。为选择出显著性检测的优化算法，本文采用 Inception_V3 作为基础特征提取网络，在鸟类数据集 CUB-200-2011 上进行各种不同的对比实验组。

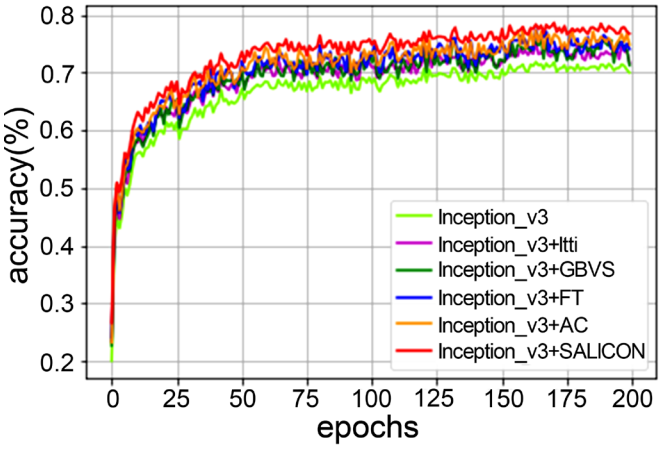

Figure 8. Comparison of different significance detection algorithms
图 8. 不同显著性检测算法的比较

实验结果如图 8 所示，当迭代次数小于 50 时，随着次数的增加，各种显著性检测算法提升图像分类准确度的速率比较大；当迭代次数大于 50 时，图像准确度的提升速度趋于平缓。与其他传统的显著性检测算法较之，SALICON 显著性检测算法相对更集中显著区域信息，在提高图像分类准确度上表现最优，大约可提升 5%。

4.3.4. 融合网络结构变体的选取

为证明本文模型中所选择的结构为最优网络结构，实验同时提取多个融合结构变体并进行对比实验，其中一个结构体只是采用一层的融合结构用 Fusion1_S1 进行表示，采用两层的融合结构用 Fusion2__S2 表示，以此类推。另外其中一个结构体没有采用原图分支跳过连接点，采用显著性分支信息集成到深度神经网络的结构用 Fusion0 表示。比较项中还包括将基础 Inception_V3 网络作为原始网络结构，只包括原图分支，不添加任何显著信息的情况。在 Stanford Dogs 数据集上进行实验，实验结果如图 9 所示。

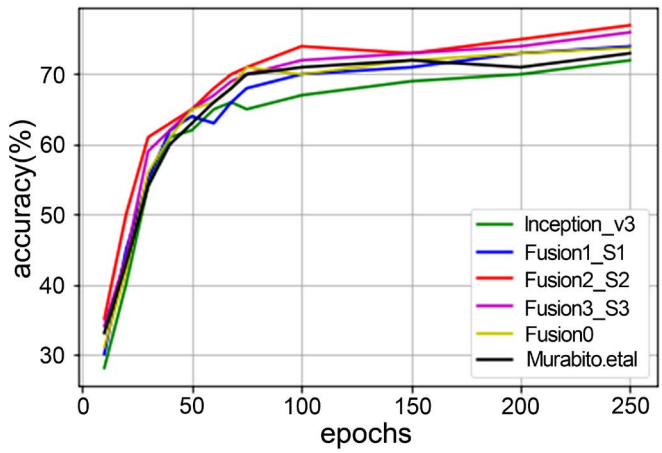

Figure 9. The influence of different fusion structures on classification
图 9. 不同融合结构对分类的影响

分析可知，在 Stanford Dogs 数据集上，相比 Murabito 等人提出的早期融合方法，本文提出的显著性融合方法在分类准确度上更加优异，且 2 层的融合结构较之 3 层或者 1 层融合结构，可获得更优实验性能，本文采取的是 2 层的融合结构。

5. 总结与展望

本文提出一种基于显著性融合进行细粒度图像分类的网络模型，该方法利用 SALICON 显著性检测算法计算产生的显著图作为引导特征提取过程的注意力机制，选取图像中的显著区域。然后融合原 RGB 图像特征，输入基础的网络分类模型 Inception_V3 中提取高层语义特征，完成图像分类。在公开的细粒度图像数据集 Stanford Dogs 和 CUB-200-2011 上进行对比实验，结果表明，本文算法的分类精度分别为 84.36%、84.94%，优于 Part R-CNN、LRBP 等算法。且该算法基于弱监督学习模型，无需额外的标注信息，实用性和鲁棒性较强。考虑到模型需要分步提取出显著区域位置信息，再进行网络训练。今后将从提高检测识别运行时间方面来改善模型，例如将显著区域定位和特征提取融入到整个网络模型实现端对端细粒度图像分类，进一步提升算法性能，同时为领域带来良好的贡献价值。

基金项目

本文得到国家重点研发计划项目(2016YFC0800506)；国家自然科学基金广东联合基金(U1801263, U1701262)；国家自然科学基金青年项目(61702111)的资助。

本文得到广东省科技计划项目(No.2016B030301008，2014B090904079)的资助。

参考文献

[1] Wah, C., Branson, S., Welinder, P., *et al*. (2011) The Caltech-UCSD Birds-200-2011 Dataset.

[2] Khosla, A., Jayadevaprakash, N., Yao, B. and Li, F.-F. (2011) Novel Dataset for Fine-Grained Image Categorization: Stanford Dogs. *Proceedings of CVPR Workshop on Fine-Grained Visual Categorization*, 1-2.

[3] Maji, S., Rahtu, E., Kannala, J., Blaschko, M. and Vedaldi, A. (2013) Fine-Grained Visual Classification of Aircraft. ArXiv Preprint ArXiv: 1306.5151.

[4] Nilsback, M.E. and Zisserman, A. (2008) Automated Flower Classification over a Large Number of Classes. 2008 *Sixth Indian Conference on Computer Vision*, *Graphics & Image Processing*, Bhubaneswar, India, 16-19 December 2008, 722-729. https://doi.org/10.1109/ICVGIP.2008.47

[5] Krause, J., Stark, M., Deng, J. and Li, F.-F. (2013) 3D Object Representations for Fine-Grained Categorization. 2013 *IEEE International Conference on Computer Vision Workshops*, Sydney, Australia, 2-8 December 2013, 554-561. https://doi.org/10.1109/ICCVW.2013.77

[6] 罗建豪, 吴建鑫. 基于深度卷积特征的细粒度图像分类研究综述[J]. 自动化学报, 2017, 43(8): 1306-1318.

[7] 张琳波, 王春恒, 肖柏华, 等. 基于 Bag-of-Phrases 的图像表示方法[J]. 自动化学报, 2012, 38(1): 46-54.

[8] Berg, T. and Belhumeur, P.N. (2013) POOF: Part-Based One-vs-One Features for Fine-Grained Categorization, Face Verification, and Attribute Estimation. 2013 *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, 23-28 June 2013, 955-962. https://doi.org/10.1109/CVPR.2013.128

[9] Daniilidis, K., Maragos, P. and Paragios, N. (2010) Improving the Fisher Kernel for Large-Scale Image Classification. *Proceedings of the* 11*th European Conference on Computer Vision* (*ECCV*), Crete, Greece, 5-11 September 2010, 143-156. https://doi.org/10.1007/978-3-642-15561-1

[10] Wang, P., *et al*. (2013) Supervised Kernel Descriptors for Visual Recognition. 2013 *IEEE Conference on Computer Vision and Pattern Recognition*, 23-28 June 2013, Portland, OR, 1828-1830. https://doi.org/10.1109/CVPR.2013.368

[11] Zhang, N., Donahue, J., Girshick, R. and Darrell, T. (2014) Part-Based R-CNNs for Fine-Grained Category Detection. In: Fleet, D., Pajdla, T., Schiele, B. and Tuytelaars, T., Eds., *Computer Vision-ECCV* 2014. *Lecture Notes in Computer Science*, Volume 8689, Springer, Cham, 834-849. https://doi.org/10.1007/978-3-319-10590-1_54

[12] Branson, S., Belongie, S., Van Horn, G. and Perona, P. (2014) Bird Species Categorization Using Pose Normalized Deep Convolutional Nets. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Columbus, OH, 594-605. https://doi.org/10.5244/C.28.87

[13] Wei, X.-S., Xie, C.-W., Wu, J.X. and Shen, C. (2018) Mask-CNN: Localizing Parts and Selecting Descriptors for Fine-Grained Bird Species Categorization. *Pattern Recognition*, **76**, 704-714. https://doi.org/10.1016/j.patcog.2017.10.002

[14] Lam, M., Todorovic, S. and Mahasseni, B. (2017) Fine-Grained Recognition as HSnet Search for Informative Image

Parts. 2017 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 21-26 July 2017, 6497-6506. https://doi.org/10.1109/CVPR.2017.688

[15] Xiao, T.J., *et al*. (2015) The Application of Two-Level Attention Models in Deep Convolutional Neural Network for Fine-Grained Image Classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, MA, 842-850.

[16] Simon, M. and Rodner, E. (2015) Neural Activation Constellations: Unsupervised Part Model Discovery with Convolutional Networks. *Proceedings of the* 15*th IEEE International Conference on Computer Vision* (*ICCV*), Santiago, Chile, 7-13 December 2015, 1143-1151. https://doi.org/10.1109/ICCV.2015.136

[17] Lin, T.Y., Roychowdhury, A. and Maji, S. (2015) Bilinear CNN Models for Fine-Grained Visual Recognition. *Proceedings of the* 15*th IEEE International Conference on Computer Vision* (*ICCV*), Santiago, Chile, 7-13 December 2015, 1449-1457. https://doi.org/10.1109/ICCV.2015.170

[18] Fu, J.L., Zheng, H.L. and Mei, T. (2017) Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition. 2017 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 21-26 July 2017, 4438-4446. https://doi.org/10.1109/CVPR.2017.476

[19] Zhang, X.P., Xiong, H., Zhou, W., Lin, W. and Tian, Q. (2016) Picking Deep Filter Responses for Fine-Grained Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, NV, 27-30 June 2016, 1134-1142. https://doi.org/10.1109/CVPR.2016.128

[20] Zhao, B., Wu, X., Feng, J.S., Peng, Q. and Yan, S. (2017) Diversified Visual Attention Networks for Fine-Grained Object Classification. *IEEE Transactions on Multimedia*, **19**, 1245-1256. https://doi.org/10.1109/TMM.2017.2648498

[21] Liu, X., Xia, T., Wang, J., *et al*. (2016) Fully Convolutional Attention Localization Networks: Efficient Attention Localization for Fine-Grained Recognition. https://arxiv.org/pdf/1603.06765.pdf

[22] Kong, S. and Fowlkes, C. (2017) Low-Rank Bilinear Pooling for Fine-Grained Classification. 2017 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 21-26 July 2017, 365-374. https://doi.org/10.1109/CVPR.2017.743