

结合注意力混合裁剪的细粒度分类网络

白瑜颖,刘宁钟,姜晓通

(南京航空航天大学 计算机科学与技术学院,江苏 南京 211106)

摘要:细粒度图像识别旨在区分同属某一大类下更为精细的子类,具有类间差距小和类内差距大的特点。同时细粒度数据集往往种类多,而数据量较少,容易产生训练时的过拟合。针对上述问题,文中提出了一种结合注意力混合裁剪的细粒度分类网络,利用注意力机制指导改进的混合裁剪数据增强。首先使用 ResNet50 作为基础网络提取图像特征,之后利用 1×1 卷积获取注意力图,再通过双线性注意力池化操作将特征图与注意力融合拼接成特征矩阵,最后利用注意力图进行改进的混合裁剪数据增强。其中改进的混合裁剪数据增强是交换两张图片的注意力高峰区域,同时交换两张图片的标注信息,之后再两张图片重新送入网络再次进行学习,以达到强化局部特征学习和丰富训练集背景的效果。实验在 4 个通用细粒度数据集上与弱监督数据增强网络(WS-DAN)和目前主流先进方法进行了比较,取得了具有竞争力的效果,相比 WS-DAN 分别提升了 0.5% (鸟类)、0.4% (车型)、0.6% (狗类)、0.4% (飞机),验证了方法的有效性。

关键词:细粒度;卷积神经网络;弱监督;注意力机制;混合裁剪;数据增强

中图分类号:TP391.4

文献标识码:A

文章编号:1673-629X(2021)10-0038-05

doi:10.3969/j.issn.1673-629X.2021.10.007

Fine Grained Image Classification Network Combined with Attention CutMix

BAI Yu-ying, LIU Ning-zhong, JIANG Xiao-tong

(School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: Fine-grained image recognition aims to distinguish the finer subclasses belonging to a large category, and has the characteristics of small inter-class gap and large intra-class gap. At the same time, fine-grained data sets tend to have more types and less data, which is easy to cause over fitting during the training process. To solve the above problems, we propose a fine-grained image classification network combined with attention CutMix, which uses attention mechanism to guide the improved CutMix data-augmentation. Firstly, ResNet50 is used as the backbone to extract image features, and then multiple 1×1 convolution kernels are used to obtain attention maps. Then, bilinear attention pooling operation is used to fuse the feature map and attention into a feature matrix. Finally, the improved CutMix is performed by using the attention map. The improved attention-CutMix is to exchange the attention peak regions of two images, and exchange the annotation information of the two images at the same time, and then send the two images back to the network for learning again, so as to achieve the effect of strengthening local feature learning and enriching the training set background. Experiments on four general fine-grained data sets are carried out with the weak supervised data enhancement network (WS-DAN) and the current mainstream advanced methods. Compared with WS-DAN, the proposed method improves by 0.5% (cub200-2011), 0.4% (Stanford cars), 0.6% (Stanford dogs), and 0.4% (FGVC aircraft), respectively, which verified the effectiveness of the proposed method.

Key words: fine-grained; convolutional neural network; weak supervision; attention mechanism; CutMix; data augmentation

0 引言

近年来,基于深度学习的图像识别技术迅猛发展,研究人员也不再局限于将目光放在通用物体分类如车和猫,转而向细粒度图像分类发起了挑战。细粒度图像分类旨在区分同属某一大类的物体的更加精细的子

类,因而具有更高的识别难度^[1]。对于细粒度图像分类而言,首先,类间差距大类内差距小,如何发掘图像中具有判别性的局部区域进行分类成为了关键问题;其次,细粒度数据集常常存在类别多,而数据量较少的问题,容易产生训练时的过拟合;最后,为了降低标注

成本,易于实际应用,如何利用图片级别的弱监督方法进行细粒度分类,也是需要解决的问题。

针对上述问题,文中提出一种结合改进混合裁剪的弱监督注意力数据增强网络。通过基于弱监督注意力机制的混合裁剪数据增强解决过分拟合背景的问题,通过改进混合裁剪解决混合背景的问题。一方面避免网络过分拟合背景,另一方面增强网络对局部特征的学习。该方法仅需图像级别标注信息,同时可进行端到端训练。为验证方法的有效性,在四个细粒度公开数据集上进行了验证。

1 相关工作

(1) 细粒度图像识别。目前主流的基于深度学习的细粒度图像识别方法大致分为四类:基于部件级别标注信息的强监督方法如借鉴了目标检测领域的 R-CNN^[2] 方法的 Part-based R-CNN^[3] 方法通过强监督信息提升性能;基于端到端特征编码的方法如双线性卷积神经网络 B-CNN^[4]、kernel pooling^[5]、hierarchical bilinear pooling^[6] 和 MC_Loss^[7] 等方法通过获取高阶特征或者设计新的损失函数进行细粒度识别;基于弱监督局部定位的方法如 NTS-net^[8] 以及结合非局部和多区域注意力的改进方法,它结合了目标检测领域的 RPN 方法^[9] 进行值得关注区域的定位和 MA-CNN^[10] 方法通过通道聚类进行部件检测从而进行细粒度特征提取;基于注意力机制的方法如循环注意力卷积神经网络 RA-CNN^[11] 和基于多尺度特征融合与反复注意力机制的细粒度图像分类算法^[12]。

(2) 混合裁剪数据增强。Sangdoo Yun 等人提出了一种训练具有局部特征的强分类器正则化策略^[13],称之为混合裁剪。具体做法是在 A 图片中随机裁剪出一个矩形,之后在数据集中随机选择 B 图片,并将 B 图片对应位置的像素填充到 A 图片裁剪掉的区域。而新图片的标记由加权求和得到。这个策略可以显著地增强网络对局部特征的学习,同时丰富背景,增强模型的泛化性能。但在细粒度图片中容易混合到背景。

(3) 弱监督注意力数据增强网络 WS-DAN^[14]。该方法借鉴了端到端特征编码方法中的双线性池化操作,先通过多个 1×1 卷积操作获取注意力特征图,之后再特征图和注意力特征图进行双线性池化获取特征矩阵,同时进一步利用注意力的位置信息进行裁剪和丢弃进行数据增强。但是双线性操作在带来高维特征的同时会有过拟合风险,同时基于注意力的丢弃操作虽然能使网络关注次要特征,同样的也可能使得网络过分拟合背景。

文中结合弱监督数据增强网络(WS-DAN)和混合裁剪数据增强,针对细粒度数据集种类多数据量少

的特点,提出了基于注意力图的混合裁剪数据增强,避免网络过分拟合图片中的背景等干扰信息。

2 文中算法

2.1 模型概述

算法流程如图 1 所示。首先将图片预处理为 448×448 大小,然后通过 ResNet-50 对输入图像进行特征提取,获得 $2048 \times 14 \times 14$ 的特征图;其次,利用 M 个 1×1 卷积得到 M 个带有位置信息的注意力图,之后一方面利用双线性注意力池化对注意力图与原来的特征图进行融合,再通过全连接层并计算交叉熵损失;另一方面利用注意力图中的位置信息,对图像进行改进的混合裁剪数据增强,并重新送入网络中进行训练。同时利用类似中心损失的注意力正则化来对注意力图进行规范。

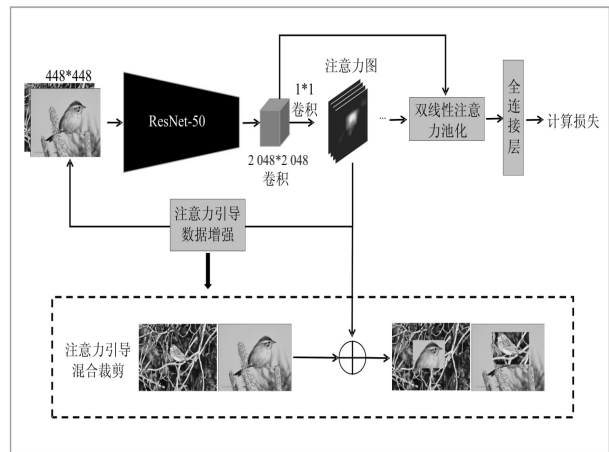


图 1 算法模型

2.2 弱监督注意力数据增强网络

2.2.1 双线性注意力池化

首先,通过卷积神经网络提取到输入图像 I 的特征图 $F \in R^{H \times W \times C}$,其中 H 、 W 表示特征图的长和宽, C 表示特征图的通道数。之后通过 M 个 1×1 卷积核将 F 转化为注意力图 $A \in R^{H \times W \times M}$ 。 M 的值为超参数,代表注意力图的数量。公式如下:

$$A = f(F) = \bigcup_{k=1}^M A_k \quad (1)$$

其中, $f(\cdot)$ 即指代 1×1 卷积操作。

在获取到注意力图集合 A 之后,利用双线性池化操作将注意力图集合 A 与原本的特征图 F 进行汇合。对于每一个注意力图 A_k ,将其逐元素乘到原本的特征图 F 之上,得到 M 个强化局部特征的双线性特征图 $F \in R^{I \times N}$,达到增强细粒度识别的效果。同时为了降低特征维度,利用全局平均池化或者全局最大值池化对 M 个 f_k 进行判别性局部特征提取,获得 M 个局部特征向量。最后将这些局部特征拼接起来得到最后的特征矩阵。该步骤如公式(2):

$$\mathbf{P} = \Gamma(\mathbf{A}, \mathbf{F}) = \begin{pmatrix} g(\mathbf{A}_1 \odot \mathbf{F}) \\ g(\mathbf{A}_2 \odot \mathbf{F}) \\ \dots \\ g(\mathbf{A}_M \odot \mathbf{F}) \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \dots \\ f_M \end{pmatrix} \quad (2)$$

其中, \mathbf{P} 表示最后拼接得到的特征矩阵 $\mathbf{P} \in R^{M \times N}$, \odot 符号表示逐元素乘积, $g(\cdot)$ 指代全局池化操作, $\Gamma(\mathbf{A}, \mathbf{F})$ 表示对注意力图 \mathbf{A} 的原特征图 \mathbf{F} 的双线性池化操作。

2.2.2 弱监督注意力学习

借鉴中心损失思想,引入弱监督注意力学习正则化方式。具体做法是,对于每次通过模型得到的特征图 f_k ,都与该类别的特征中心 $C_k \in R^{1 \times N}$ 计算均方误差作为中心损失,见式(3)。模型即会倾向于对每一个类别学习到相似的特征,对于注意力图的每个通道亦会倾向于响应各自固定的部件。

$$L_{\text{center}} = \sum_{k=1}^M \|f_k - c_k\|_2^2 \quad (3)$$

而特征中心 C 在最初被初始化为全零,之后在训练过程中不断地根据训练中的特征图 f_k 来更新其标记所属类的特征中心,如式(4)。

$$c_k = c_k + \beta(f_k - c_k) \quad (4)$$

其中, β 为超参数,文中按照原文建议设置为 0.05。

2.3 注意力引导混合裁剪

对于细粒度识别而言,采用随机混合裁剪的方式进行数据增强,往往会裁剪到背景,无法带来正向的收益,因此提出改进的基于注意力的混合裁剪算法。详细算法介绍如下。

令 $x_1, x_2 \in R^{H \times W \times C}$ 分别为两张图片, y_1, y_2 分别对应两张图片的标记。 x_1, x_2 在经过弱监督注意力网络之后,会得到各自的注意力图 $A_1, A_2 \in R^{H \times W \times M}$ 。

对于各自的注意力图 A_i 的 M 个通道上的 $A_k \in R^{H \times W \times 1}$, $k < m$, 分别以式(5)进行正则化以增强区域的响应对比,转化为特征热力图 A_i^* 。

$$A_k^* = \frac{A_k - \min(A_k)}{\max(A_k) - \min(A_k)} \quad (5)$$

式中,对于特征热力图 A_i^* ,在 32 个通道求平均值,得到对于该图片响应最强烈的位置信息,之后根据给定阈值 θ 计算出裁剪区域掩膜 M_{ci} ,具体来说对于特征热力图 A_i^* 的每个位置响应值若其大于 θ ,则掩膜 M_{ci} 对应位置为 1,反之则为 0,如式(6):

$$M_{ci}(m, n) = \begin{cases} 1, & \text{if}(A_i^*(m, n) > \theta) \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

其中, (m, n) 表示特征热力图或者掩膜的横纵坐标值。

之后可以根据掩膜 M_{ci} 求出一个能够包围所有大

于阈值区域的包围框 B_i ,而根据此包围框坐标即可从原图中裁剪出目标图片 x_{ci}, x_{c2} 。之后将 x_{ci} 调整大小为 x_{c2} 的大小得到 \tilde{x}_{c1} , x_{c2} 调整大小为 x_{ci} 的大小得到 \tilde{x}_{c2} 。最后分别将 \tilde{x}_{c1} 填充到 x_2 的 B_2 位置,将 \tilde{x}_{c2} 填充到 x_1 的 B_1 位置。即完成了对图像的混合裁剪。公式如下:

$$\begin{cases} \tilde{x}_1 = R(x_1 \odot M_{c1}, M_{c2}) + x_2 \odot (1 - M_{c2}) \\ \tilde{x}_2 = R(x_2 \odot M_{c2}, M_{c1}) + x_1 \odot (1 - M_{c1}) \end{cases} \quad (7)$$

其中, \odot 表示图片与掩膜逐元素乘积, $R(\mathbf{A}, \mathbf{B})$ 表示将 \mathbf{A} 图片调整为 \mathbf{B} 图片(掩膜)的大小, \tilde{x}_1 和 \tilde{x}_2 分别表示由 x_1, x_2 得到的新图片。

而对应的,不同于原本的混合裁剪数据增强采用根据面积比值求加权的方式,笔者认为根据注意力引导的混合裁剪会将图片最主要最具判别性的特征全部裁剪掉,并进行交换,因此将两张图片的真实标记 y_1, y_2 进行交换,如式(8):

$$\begin{cases} \tilde{y}_1 = y_2 \\ \tilde{y}_2 = y_1 \end{cases} \quad (8)$$

此处 \tilde{y}_1 和 \tilde{y}_2 分别对应 \tilde{x}_1 和 \tilde{x}_2 的图像标记。

之后将得到的 $(\tilde{x}_1, \tilde{y}_1), (\tilde{x}_2, \tilde{y}_2)$ 重新送入网络进行训练,增强网络对局部区域的学习,同时降低网络对环境的过拟合可能性。算法流程如图 2 所示。

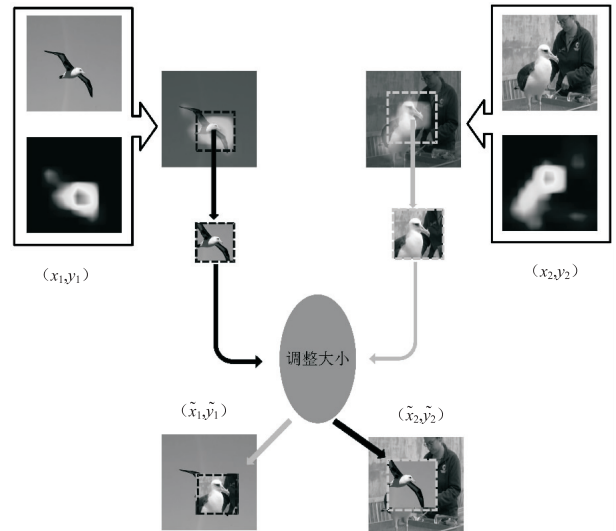


图 2 改进的注意力混合裁剪算法流程

3 实验结果及分析

3.1 实验设置

3.1.1 数据集

实验在四个公开细粒度数据集 CUB200-2011、Stanford Dogs, Stanford Cars, FGVC Aircraft 上进行,数据集详细信息见表 1,数据集部分图片示例见图 3。

表 1 四个常见公开细粒度数据集

数据集	类别数量	训练集 图片数量	测试集 图片数量	目标 种类
CUB200-2011	200	5 994	5 794	鸟类
Stanford Cars	196	8 144	8 041	车辆
FGVC Aircraft	100	6 667	3 333	飞机
Stanford Dogs	120	12 000	8 580	狗



图 3 常用数据集图片示意

3.1.2 实验细节

在接下来的实验中,使用去除全连接层的残差网

络 Resnet-50^[15] 作为基础网络进行特征提取,注意力图的数量 M 设置为 32,即使用 32 个 1×1 大小的卷积进行注意力图的获取。对于裁剪阈值 θ ,选取了 (0.4, 0.6) 之间的一个随机实数。

对于模型优化方法选择随机梯度下降 SGD,动量参数设置为 0.9,最大迭代次数设置为 80,权重衰减设置为 0.000 01,同时将每个批次的大小设置为 12。初始学习率设置为 0.001,每两次迭代将学习率缩放为 0.9 倍。实验在 RTX2080Ti 11G 显存上进行,实现框架为 pytorch。

3.2 对比实验

首先在四个数据集上与现有的先进算法进行了实验对比,对于基准算法 WS-DAN,使用了 pytorch 的复现版本,在表格中同时将原文结果与复现结果进行展示。评价指标使用 Top-1 准确率(表 2 中将同一数据集下最好的结果进行加黑,将第二好的结果添加下划线以便查阅)。

表 2 对比实验结果

算法	基础网络	CUB200-2011	Stanford Cars	Stanford Dogs	FGVC Aircraft
ResNet-50 ^[15]	ResNet-50	83.2	90.7	85.1	87.0
InceptionV3 ^[16]	InceptionV3	83.7	90.8	88.9	87.4
BCNN ^[5]	VGG-16	84.1	91.3	—	84.1
RA-CNN ^[11]	VGG-16	85.4	92.5	87.3	88.4
NTS-Net ^[8]	ResNet-50	87.5	93.9	—	91.4
MC_Loss ^[7]	BCNN	87.3	94.4	92.9	—
WS-DAN(原文) ^[14]	InceptionV3	<u>89.4</u>	<u>94.5</u>	<u>92.2</u>	<u>93.0</u>
WS-DAN	ResNet-50	89.2	94.4	88.7	92.7
WS-DAN+随机混合裁剪	ResNet-50	89.3	94.4	88.9	92.9
文中算法	ResNet-50	89.7	94.8	89.3	93.1

可以看到,文中方法在 CUB200-2011 鸟类数据集,Stanford Cars 车辆数据集和 FGVC Aircraft 飞机数据集上,均取得了最好的效果,分别达到了 89.7%, 94.8% 和 93.1%, 优于近年来的先进方法;同时相较于基准方法 WS-DAN(ResNet-50),在使用相同的基础骨架网络的基础上,在四个数据集上的精度分别提升了 0.5%, 0.4%, 0.6%, 0.4%, 同时与随机混合裁剪方法相比,也有较为明显的提升,证明了文中方法的有效性。

3.3 可视化

为了进一步对比与随机混合裁剪的效果,在 cub200-2011 数据集上进行了可视化实验,直观展示了基于改进的注意力混合裁剪算法的效果。

如图 4 所示,可以看到利用注意力机制引导的混合裁剪避免了混合到背景的问题,同时,将具有判别性

的特征混合裁剪到其他图片的背景中,大大丰富了训练数据的背景,降低了网络对于背景的过拟合的可能性,同时强化了网络对局部特征的学习。

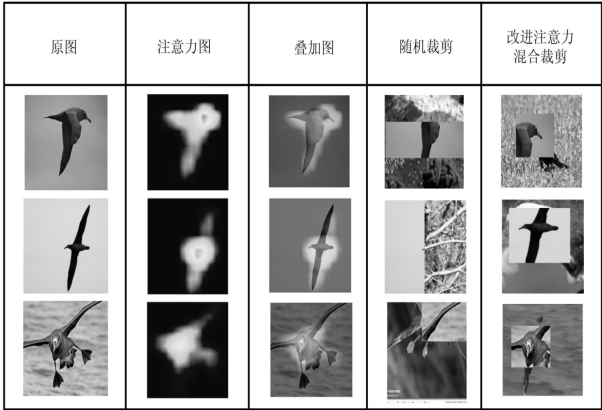


图 4 可视化效果

4 结束语

文中创新性地提出了基于注意力机制的混合裁剪数据增强方法。利用注意力网络在弱监督学习到的位置信息引导混合裁剪数据增强,一方面利用混合裁剪丰富训练数据背景,同时避免随机混合裁剪混合到背景的问题;另一方面增强网络对局部特征的学习,避免网络对背景的过拟合。实验结果表明,该方法相对于基准方法 WS-DAN 在四个细粒度数据集上的精度均有明显提升,并且在其中的鸟类、车型和飞机数据集上展现了很强的竞争力。该方法简单高效,仅需图像级标注信息,可端到端训练,有着良好的应用价值。但目前方法的耗时较高,在今后工作中将把工作中心放在提升识别速度和提升精度上。

参考文献:

- [1] 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.
- [2] GIRSHICK R,DONAHUE J,DARRELL T,et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE conference on computer vision and pattern recognition. Columbus,OH:IEEE,2014:580-587.
- [3] ZHANG N,DONAHUE J,GIRSHICK R,et al. Part-based RCNNs for fine-grained category detection[C]//Computer vision - ECCV 2014. Zurich, Switzerland: Springer, 2014: 834-849.
- [4] LIN T Y,ROYCHOWDHURY A,MAJI S. Bilinear CNN models for fine-grained visual recognition[C]//Proceedings of the IEEE international conference on computer vision. Santiago,Chile:IEEE,2015:1449-1457.
- [5] CUI Y,ZHOU F,WANG J,et al. Kernel pooling for convolutional neural networks[C]//IEEE conference on computer vision & pattern recognition. Hawaii,America:IEEE,2017.
- [6] YU C,ZHAO X,ZHENG Q,et al. Hierarchical bilinear pooling for fine-grained visual recognition[C]//Computer vision - ECCV 2018. Munich,Germany:Springer,2018:595-610.
- [7] CHANG D,DING Y,XIE J,et al. The devil is in the channels:mutual-channel loss for fine-grained image classification[J]. IEEE Transactions on Image Processing,2020,29:4683-4695.
- [8] YANG Z,LUO T,WANG D,et al. Learning to navigate for fine-grained classification[C]//Computer vision - ECCV 2018. Munich,Germany:Springer,2018:438-454.
- [9] REN S,HE K,GIRSHICK R,et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(6):1137-1149.
- [10] ZHENG H,FU J,MEI T,et al. Learning multi-attention convolutional neural network for fine-grained image recognition[C]//2017 IEEE international conference on computer vision (ICCV). Venice,Italy:IEEE,2017.
- [11] FU J,ZHENG H,MEI T. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition[C]//IEEE conference on computer vision & pattern recognition. Hawaii,America:IEEE,2017.
- [12] 何凯,冯旭,高圣楠,等. 基于多尺度特征融合与反复注意力机制的细粒度图像分类算法[J]. 天津大学学报,2020,53(10):1077-1085.
- [13] YUN S,HAN D,OH S J,et al. Cutmix: regularization strategy to train strong classifiers with localizable features[C]//Proceedings of the IEEE international conference on computer vision. Seoul,South Korea:IEEE,2019:6023-6032.
- [14] HU T,QI H,HUANG Q,et al. See better before looking closer: weakly supervised data augmentation network for fine-grained visual classification[J]. arXiv:1901.09891,2019.
- [15] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]//2016 IEEE conference on computer vision and pattern recognition (CVPR). Las Vegas,NV:IEEE,2016.
- [16] SZEGEDY C,VANHOUCKE V,IOFFE S,et al. Rethinking the inception architecture for computer vision[C]//2016 IEEE conference on computer vision and pattern recognition (CVPR). Las Vegas,NV:IEEE,2016:2818-2826.