ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 6,Nov. 2023,pp. 1392－1405
DOI:10. 16337/j. 1004-9037. 2023. 06. 014

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

MAFDNet:复杂环境下图像自适应分类新方法

叶继华[1]，黎　欣[1]，陈　进[1]，江爱文[1]，化志章[1]，万文涛[2]

(1. 江西师范大学计算机信息工程学院,南昌 330022;2. 江西师范大学教育学院,南昌 330022)

摘　要：复杂环境下,往往困难样本和简单样本并存,现有分类方法主要针对困难样本进行设计,所构建网络用于分类简单样本时会造成计算资源的浪费;而网络修剪和权重量化等方法则不能同时兼顾模型的准确度和存储开销。为提升计算资源的使用效率并有更好的准确率,本文着眼于输入样本的空间冗余,提出了复杂环境下图像自适应分类网络MAFDNet,并引入置信度作为分类准确性的判断,同时提出了由内容损失、融合损失和分类损失组成的自适应损失函数。MAFDNet由3个子网组成,输入图像首先被送入到低分辨率子网中,该子网有效提取了低分辨率的特征,具有高置信度的样本先被识别并从网络中提前退出,低置信度的样本则需要依次进入更高分辨率的子网中,而网络中的高分辨率子网具有识别困难样本的能力。MAFDNet将分辨率自适应和深度自适应结合在一起,通过实验表明,在相同计算资源条件下,MAFDNet在CIFAR-10、CIFAR-100和ImageNet这3个复杂环境数据集上的top-1准确率均得到提升。

关键词：MAFDNet;复杂环境;自适应分类;自适应损失;置信度

中图分类号：TP391　　**文献标志码**:A

MAFDNet:A New Method of Image Adaptive Classification in Complex Environment

YE Jihua[1], LI Xin[1], CHEN Jin[1], JIANG Aiwen[1], HUA Zhizhang[1], WAN Wentao[2]

(1. College of Computer Information Engineering, Jiangxi Normal University, Nanchang 330022, China; 2. College of Education, Jiangxi Normal University, Nanchang 330022, China)

Abstract: In complex environments, difficult samples and simple ones often coexist. The existing classification methods are mainly designed for difficult samples, and the constructed network causes a waste of computing resources when it is used to classify simple ones. However, network pruning and weight quantization couldn't take into account both accuracy and storage cost. To promote the efficiency of computing resources with better accuracy, focusing on the spatial redundancy of input samples, this paper proposes an adaptive image classification network MAFDNet in complex environment, introduces the confidence as the classification accuracy of judgment, and puts forward the adaptive loss function composed of the content loss, fusion loss and classification loss at the same time. MAFDNet consists of three subnets. The input images are first sent to the low-resolution subnet, which effectively extracts low-resolution features. Samples with high confidence are first identified and removed from the network in advance, while samples with low confidence need to enter the subnet with higher resolution in turn. The high resolution subnet in the network has the ability to identify difficult samples. MAFDNet combines resolution adaptive

基金项目:国家自然科学基金(62167005,61966018);江西省教育厅重点科研项目(GJJ200302)。
收稿日期:2023-03-05;**修订日期**:2023-04-06

and depth adaptive. Through experiments, the top-1 accuracy of MAFDNet is improved in CIFAR-10, CIFAR-100 and ImageNet data sets under the same computing resource conditions.
Key words: MAFDNet; complex environment; adaptive classification; adaptive loss; confidence

引　言

在实际场景中,因光照不足、物体遮挡和雨雾模糊等恶劣环境导致获取的图像模糊不清晰、严重退化,难以被正确分类,此类图像被称为困难样本;简单样本则为较容易正确分类的图像。针对简单和困难样本同时存在的情况,即复杂环境下,简单的网络只可正确分类简单样本,很难正确分类困难样本,复杂的网络虽然可以正确分类困难样本,但是对于简单样本来说无疑存在较大的计算资源的浪费。尽管计算机硬件的进步使人们能够训练非常深的卷积神经网络(Convolutional neural network, CNN),例如ResNet[1]和DenseNet[2],但深层CNN带来的大量计算消耗在许多应用中仍然无法承受。为了既可以正确分类困难样本,又可以较少的资源正确分类简单样本,国内外许多学者已做了很多研究,例如,网络修剪[3]、权重量化[4]和自适应网络[5-7]等。其中,网络修剪首先将低于某个阈值的权重连接全部剪除,之后对剪枝后的网络进行微调以完成参数更新。这种方法的不足之处在于,剪枝后的网络是非结构化的,即被剪除的网络连接在分布上没有任何连续性,这种稀疏的结构导致CPU高速缓冲与内存频繁切换,从而限制了实际的加速效果,同时稀疏数据结构将会需要额外存储开销,而且目前对稀疏操作加速支持的库非常有限;权重量化则通过将float32格式的数据转变为int8格式,虽然可以降低内存和存储的开销,但在一定程度上降低了模型的精度;自适应网络旨在通过动态调整网络结构来减少简单样本上的计算冗余,已显示出很好的性能,但是仍存在模型准确率和计算消耗不能同时兼顾的问题,本文基于已有的自适应网络模型,利用数据样本中的信息冗余,克服了之前自适应网络的不足。

现有的大多数关于自适应的工作都集中在通过容易识别的特征来减少网络的深度或宽度。研究表明,不同样本的分类难度大不相同:简单样本可以通过较少层数或通道较小的网络进行正确分类,而另一些样本可能需要较复杂的网络。例如多尺度密集网络(Multi-scale dense network, MSDNet)[6]允许一些样本在达到其预测置信度时从辅助分类器中退出,没有考虑通过利用图像中的空间冗余来设计自适应模型。本文与现有工作侧重于网络结构中的计算冗余相反,目的是利用数据样本中的信息冗余。由于低分辨率特征图足以对简单样本进行分类,而使用高分辨率特征图来探测细节对于准确识别某些困难样本是必要的。从信号频率的观点[8],可以使用包含在低分辨率特征中的低频信息正确地对简单样本进行分类,当无法精确预测具有低分辨率特征的样本时,高频信息仅用作识别困难样本的补充。

为加快网络速度,目前主要通过修改网络模型来实现。如设计轻量级模型MobileNet[9]和ShuffleNet[10]、修剪冗余网络连接[3]或量化网络权重[4]、知识蒸馏[11]等方法,这些方法可以看作是静态模型加速技术,通过整个网络推断所有输入样本。相反,动态自适应网络可以根据输入复杂度策略性地分配适当的计算资源,以对输入图像进行分类,该研究方向近年来受到越来越多的关注。最直观的是集合多个模型,并以级联或混合方式有选择地执行模型的子集。最近的工作还提出自适应地跳过层或块[12],或动态选择通道[13],分类器也可以附加在深度网络的不同位置,以允许尽早分类成功简单样本。然而,这些现有技术中的大多数集中于利用网络的架构冗余来设计自适应网络。由于输入图像的空间冗余已在最近的工作中得到证明[8],因此本文提出了一种新颖的自适应学习模型,该模型同时利用了神经网络的结构冗余和输入样本的空间冗余。

由于单尺度网络中的下采样操作可能会限制网络识别物体的能力,因此最近的研究提出在网络中

采用多尺度特征图以同时利用粗略和精细特征，可显著改善许多视觉任务的网络性能，包括图像分类[14]、目标检测[15]、语义分割[16]和姿态估计[17]，并且多尺度结构在自适应网络[6]和高效存储网络[18]中体现了显著的效果。虽然通过深层神经网络保持高分辨率特征图对于识别某些非典型困难样本或某些特定任务（例如姿态估计）[17]是必要的，但对高分辨率特征的频繁卷积操作通常会导致模型的资源匮乏。可以观察到，对于所有具有低分辨率输入的样本，轻量级网络可以产生较低的错误率。ADASCALE[19]还自适应地选择了输入图像的比例，该比例提高了视频对象检测的准确性和速度。

基于上述分析，本文提出了基于DenseNet融合的自适应分类网络MAFDNet（Multi-resolution image adaption classification network based on dense network fusion），该网络实现了在深度CNN中分辨率自适应的思想。MAFDNet由具有不同输入分辨率的子网组成，简单样本通过最低分辨率的子网进行分类，当先前的子网无法达到给定阈值时，将使用分辨率更高的子网，同时，来自先前子网的粗略特征将被重用并融合到当前子网中。当使用低分辨率子网可以准确预测样本时，则避免对高分辨率特征执行不必要的卷积操作，MAFDNet的自适应机制减少了冗余的计算，从而提高了网络的计算效率。

本文参考文献[6]中单张图像预测和多张图像预测的设置，在3个图像分类数据集（CIFAR-10，CIFAR-100和ImageNet）上评估MAFDNet性能。

1 MAFDNet网络

1.1 分辨率自适应设置

将MAFDNet设置为具有K个分类器的网络，分类器连接在模型的不同深度上。给定一个输入图像$\boldsymbol{x}$，第k个分类器（$k=1, 2, \cdots, K$）的输出可以表示为

$$\boldsymbol{p}^k = f_k(\boldsymbol{x}, \theta_k) = [p_1^k, p_2^k, \cdots, p_C^k]^{\mathrm{T}} \in \mathbf{R}^C \tag{1}$$

式中：θ_k表示与第k个分类器相对应的子网的参数，每个元素$p_C^k \in [0,1]$为第C个分类器的预测置信度，θ_k在此处具有共享参数。

自适应模型通过样本的复杂度动态分配适当的计算资源，样本将在其输出满足输出条件的第一个分类器处退出网络。由于softmax作为输出层，结果可以直接反映概率值，并且避免了负数和分母为0的情况，因此在本文中，将softmax输出的概率最大值作为分类正确的置信度，这意味着最终输出将是第一个分类器的最大输出大于给定阈值ε的值。这可以表示为

$$\begin{aligned} k^* &= \min\{k | \max_C p_C^k \geqslant \varepsilon\} \\ \hat{y} &\in \arg\max_C p_C^{k^*} \end{aligned} \tag{2}$$

式中阈值ε控制了测试时分类精度和计算成本之间的平衡。

1.2 总体框架

图1为MAFDNet的总体架构。它包含一个初始层和对应于不同分辨率的H个子网。每个子网在最后几个块中都有多个分类器。与MSDNet相似，本文采用多尺度体系结构和密集连接。尽管MAFDNet和MSDNet具有类似的多尺度结构，但是它们的详细体系结构设计和计算却有很大差异。最突出的区别是MAFDNet需要首先提取低分辨率特征，而这不遵循经典的深层CNN（包括MSDNet、ResNet和DenseNet等）中首先提取高分辨率特征的传统设计。MSDNet和MAFDNet之间差异的更多详细信息将在1.4节中讨论。

MAFDNet的基本思想是，网络首先使用最低空间分辨率的特征图即第一个子网预测样本，以避免因对大型特征进行卷积而导致较高的计算成本。如果第一个子网对样本的预测不可靠，则第一个子网的中间特征将以更高的分辨率融合到下一个子网中。然后，由具有较大特征的下一个子网执行分类任

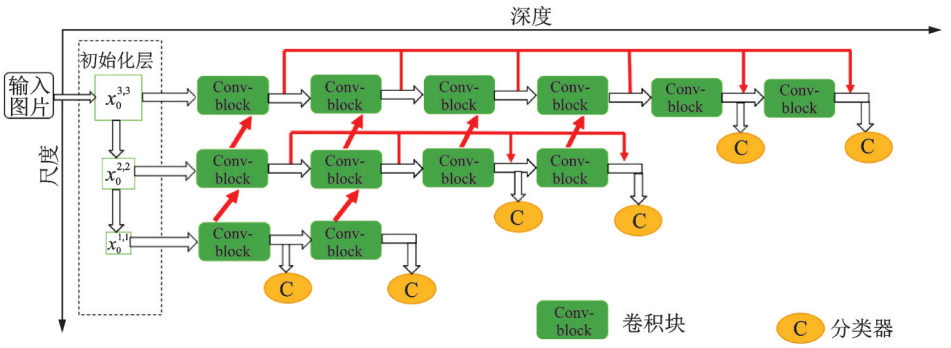


图1 本文提出的MAFDNet整体框架

Fig.1 Overall framework of the proposed MAFDNet

务。重复此过程,直到一个子网产生可靠的预测,或者进入最后一个子网。在每一个子网中添加了密集跳跃连接,就是把密集块看成一个整体,第一个卷积层的输出以及每个密集块的输出,都输入给之后的所有密集块。由于这样做,所有的特征都串联起来,这样直接输入全连接层会产生巨大的计算开销,因此添加了一个核大小为1×1的卷积层来减小特征数量,这个卷积层称为瓶颈层。

算法1进一步说明了MAFDNet的分辨率自适应过程。

算法1 MAFDNet的分辨率自适应过程

输入:待分类图像$\boldsymbol{x}$

(1)初始化层生成S个大小的H个基本特征图(例如,图1中3个标度,$s=1$代表最低分辨率),对应于子网h中尺度s的基本特征可以表示为$x_0^{s,h}$;

(2)子网i(i初始值为1)使用低层次的特征$x_0^{i,i}$进行分类;

(3)如果子网i中的分类器预测置信度(使用softmax概率的最大值作为置信度度量)超过预先确定的阈值ε,则退出网络;

(4)如果子网i无法以高置信度获得分类结果,子网i中的中间特征被依次融合到子网$i+1$中;

(5)使用较大特征($x_0^{i+1,i+1}$)的子网$i+1$进一步对样本$\boldsymbol{x}$进行分类;

(6)如果中间子网的分类器预测置信度无法达到阈值ε,则返回步骤(3)直到分类器的预测置信度达到预先确定的阈值ε或者达到整个网络的最后一个分类器。

输出:步骤(3)中的分类器预测结果。

值得注意的是,即使是MAFDNet通常也会由粗到细的处理输入图像,MAFDNet中的每个子网在正向传播期间仍会对特征进行下采样,直到达到最低分辨率($s=1$),并且所有分类器仅在每个子网的最后几个卷积块之后添加。

上述分辨率自适应过程符合图像识别的常规认知。即使仅提供低分辨率输入,也能以高置信度对具有代表性特征的简单样本进行正确分类。具有非典型特征的困难样本只能从基于高分辨率特征图中提取带有精细细节的信息来正确分类。

1.3 网络细节

1.3.1 *初始化层*

初始化层用于生成包含S个尺度的H个基本特征,在图1中其仅包含垂直连接,可以将其垂直布局视为一个微型H层卷积网络(H为网络中基本特征的数量)。图1显示了具有3个尺度的3个基本特征的MAFDNet。具有最大比例的第一个基本特征是从常规卷积(此处的常规卷积层由一个瓶颈层和一

个规则的卷积层组成。每一层由BN层、ReLU层和1个卷积层组成。)产生的,而粗略特征是从以前的高分辨率特征中通过步长卷积(即步长为2的卷积)得到的。值得注意的是,这些基本特征的比例可以相同。例如,一个MAFDNet可以具有3个比例的4个基本特征,其中最后两个基本特征具有相同的分辨率。

1.3.2 不同尺度的子网

因为初始化层生成H个基本特征,可将MAFDNet分为H个子网络,这些子网络由不同的卷积块组成。除第一个子网外,每个子网都由其对应的基本特征图和上一层子网的中间特征图组成。

子网1处理最低分辨率的图像,输入为输入$x_0^{1,1}$。子网1中的密集块有t层,如图2(a)所示。每个密集块中第i层的输出$x_i^{1,1}(i=1,2,\cdots,t)$也将传播到子网2,以重用子网1的特征。通常,可以将子网1视为具有多个分类器的DenseNet,从而处理分辨率最低的特征图。

尺度为s的子网$h(h>1)$处理基本特征$x^{s,h}$,并从子网$(h-1)$融合特征。同时在参考文献[20]在子网中添加了密集跳跃连接,就是把密集块看成一个整体,第一个卷积层的输出以及每个密集块的输出,都输入给之后的所有密集块。由于这样做,所有的特征都串联起来,这样直接输入全连接层会产生巨大的计算开销,因此添加了一个核大小为1×1的卷积层来减小特征数量。将具有特征融合功能的密集块称为融合块,如图2(b,c)所示,假设子网$(h-1)$具有b_{h-1}个块,则子网h中的前b_{h-1}个块都将是融合块。

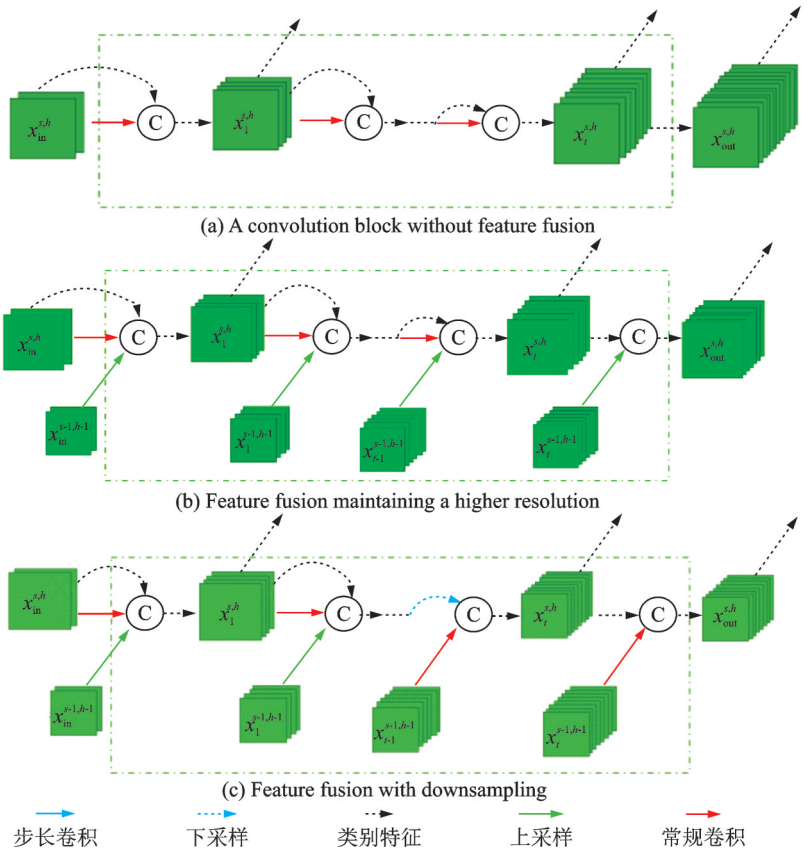


图2 MAFDNet的卷积块:密集块和融合块

Fig.2 MAFDNet's convolution blocks: Dense block and fused blocks

本文设计了两种不同的特征融合方法。一种保持输入图像的分辨率不变,如图2(b)所示,而另一种则通过步长卷积减小特征图的大小,如图2(c)所示。为了生成分辨率更高的新特征图作为输入,图2(b)中的融合块首先生成带有常规卷积层的$x_{\text{in}}^{s,h}$。上一个子网的特征尺度$(s-1)$由Up-Conv层处理,该层由常规卷积层和上采样双线性插值组成,这样可以确保产生的特征具有相同的空间分辨率。然后,通过密集连接将所生成的特征融合在一起。

如图2(c)所示,具有下采样功能的融合块使用步长卷积层来降低该块末尾的空间分辨率。如蓝色虚线箭头所示,在合并操作之后,还会进行密集连接的连接。由于当前子网中的特征尺度减小了,因此前一个子网的特征将由常规卷积层进行处理以保持较低的分辨率,然后在图2(c)的方框末尾通过拼接进行融合。

尺度为s的子网h:对于具有b_h块的子网,块1到块$b_{h-1}(b_{h-1}<b_h)$都是融合块,而其余的都是常规密集块。在正向传播过程中对特征图进行了s次下采样,依次在$b_{h-s},b_{h-s+1},\cdots,b_{h-1}$上进行。这样可以确保连接分类器每个子网的末尾特征具有最低的分辨率。

与文献[1,6]中的架构设计相似,本文加入过渡层以进一步压缩每个子网中的特征图。过渡层的设计与文献[1,6]中的过渡层完全相同,由1×1卷积操作以及BN层和ReLU层组成。过渡层进一步保证了MAFDNet的计算效率。为简单起见,在图1中省略了这些过渡层。

每个分类器由两个卷积层、1个平均池化层和1个线性层组成。分类器在不同子网的最后几个密集块中实现,$f_k(\cdot)$表示第k个分类器。在单张图像的测试过程中,该图像依次通过每个子网络,直到耗尽预算并输出最新的预测。在多张图像的测试过程中,如果一个分类器的预测置信度(本文使用softmax概率的最大值作为置信度度量)超过预先确定的阈值ε,则退出网络。在训练前,先计算网络进入到第k个分类器所需要的资源消耗C_k,本文用$0<q\leqslant 1$表示一个图像达到退出网络的概率。假设在所有层中都有概率p的可能退出网络,则图像在第k个分类器处退出的概率为

$$q_k=z(1-q)^{k-1}q \tag{3}$$

因为要保证所有层的退出概率之和为1,故此处z是一个归一化常数,满足$\sum_k p(q_k)=1$。在测试时,需要确保D_{test}中所有样本的总消耗不超过预算B,由此产生约束条件$|D_{\text{test}}|\sum_k q_k C_k\leqslant B$,在验证集上,由约束中的$q$来确定阈值$\varepsilon$,以使大约$|D_{\text{test}}|q_k$数量的验证样本在第$k$个分类器处退出。

1.4 损失函数

本文的自适应损失函数Loss由交叉熵损失Loss_{CE}、融合损失Loss_{fu}和内容损失$\text{Loss}_{\text{content}}$组成。交叉熵损失$\text{Loss}_{\text{CE}}$用来训练分类器,融合损失$\text{Loss}_{\text{fu}}$用来削弱不同子网融合带来的损失,使得不同子网的图像融合具有更完整的信息,内容损失$\text{Loss}_{\text{content}}$确保处理后的图像内容保持不变。

1.4.1 交叉熵损失

在训练过程中,所有的分类器使用交叉熵损失函数$L(f_k)$,并且最小化权重累加损失,即

$$\text{Loss}_{\text{CE}}=\frac{1}{|D|}\sum_{(x,y)\in D}\sum_k w_k L(f_k) \tag{4}$$

式中:D代表训练集,$w_k\geqslant 0$表示第k个分类器的权重。

经验发现,对所有损失函数使用相同的权重(即满足$\forall k:w_k=1$)效果更好。

1.4.2 融合损失

值得注意的是,对图像进行不同层级之间的融合会造成原图像信息的部分损失。针对此问题,定

义了图像融合损失Loss_{fu}。令f_h为图像原始向量表示，f_h'为融合后的向量表示，则图像融合损失函数Loss_{fu}可定义为

$$\text{Loss}_{\text{fu}}=\begin{cases}0 & |f_h-f_h'|\leqslant\phi_1\\ \dfrac{1}{2}(f_h-f_h')^2 & \phi_1<|f_h-f_h'|<\phi_2\\ \phi_2|f_h-f_h'|-\dfrac{1}{2}\phi_2^2 & |f_h-f_h'|\geqslant\phi_2\end{cases}\tag{5}$$

式中ϕ_1、ϕ_2为超参数。图像融合损失的原则为计算融合后的图像向量与融合前图像向量间的信息损失，目的在于降低一部分融合损失的同时，保证融合图像在模型中的重要意义。所以，图像融合损失定义为当融合图像向量与融合前图像向量差异小于某一超参数时，损失为0，当二者差异在两个超参数之间时，通过二次函数的形式缓慢降低误差，而对于二者差异过大，超过了ϕ_2的情况，则采用线性的方式快速降低误差。通过图像融合损失，可以得到图像表示向量f_h'，有效地削弱了图像信息融合过程中损失带来的影响，模型可以更有效地对数据进行训练。

1.4.3 内容损失

为了初始化层可以生成效果较好的不同分辨率的图像，加入内容损失函数$\text{Loss}_{\text{content}}$。$V_j(\hat{\boldsymbol{Y}})$表示第$j$层网络在处理图像$\boldsymbol{Y}$时的激活情况，其形状为$(C_j, H_j, W_j)$。使用$L2$损失的平方，根据图像的形状归一化，比较ground truth图像$\boldsymbol{Y}$和预测图像$\hat{\boldsymbol{Y}}$的激活情况，即

$$L_j=\frac{\|V_j(\hat{\boldsymbol{Y}})-V_j(\boldsymbol{Y})\|_2^2}{C_j\cdot H_j\cdot W_j}\tag{6}$$

n表示输入图像的数量，m表示初始化层网络的层数，内容损失$\text{Loss}_{\text{content}}$为

$$\text{Loss}_{\text{content}}=\frac{1}{n}\sum_{j=1}^{mn}L_j\tag{7}$$

将式(6)代入式(7)可得

$$\text{Loss}_{\text{content}}=\frac{1}{n}\sum_{j=1}^{mn}\frac{\|V_j(\hat{\boldsymbol{Y}})-V_j(\boldsymbol{Y})\|_2^2}{C_j\cdot H_j\cdot W_j}\tag{8}$$

综上，本文的自适应损失函数可表示为

$$\text{Loss}=\text{Loss}_{\text{CE}}+\text{Loss}_{\text{fu}}+\text{Loss}_{\text{content}}\tag{9}$$

1.5 分辨率和深度自适应

本文提出的MAFDNet可以同时实现MSDNet中采用的深度自适应和分辨率自适应的思想。图3说明了MSDNet和MAFDNet的主要区别。在MSDNet中，分类器位于最低分辨率子网中，如果中间的预测置信度无法达到阈值，则将执行所有尺度的所有层。但是，在MAFDNet中，具有最小尺度的密集块首先被激活，并且深度适应在单个尺度内进行。如果先前的子网无法做出可靠的预测，则输入样本将传播到下一个子网，并重复深度适应过程，直到预测的可信度大于阈值，或者达到整个网络的最后一个分类器。这样的自适应方案将分辨率和深度适应结合在一起，与MSDNet相比有了显著改进。

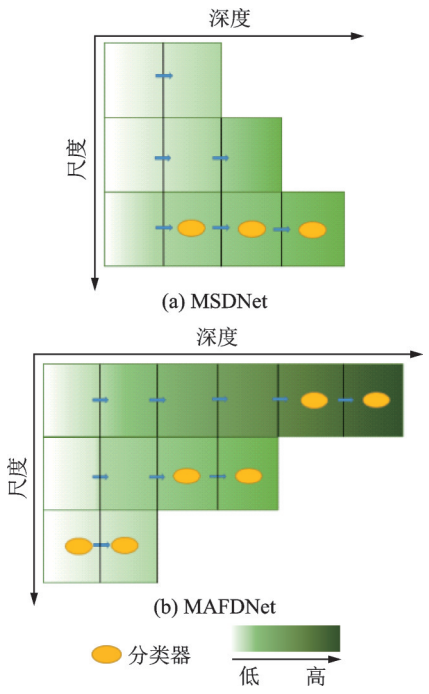


图3 MSDNet和MAFDNet

Fig.3 MSDNet and MAFDNet

2 实验和分析

由于CIFAR和ImageNet数据集中的图像均包含简单样本和困难样本,本文在CIFAR和ImageNet数据集上进行了实验。CIFAR-10和CIFAR-100这两个数据集都包含50 000张训练图像和10 000张测试图像,本文在训练集中拿出5 000张图像作为寻找最佳置信度阈值的验证集;ImageNet数据集包含120万张1 000个类别的图像和50 000张验证图像,本文使用数据集的原始验证集作为测试集,从训练集中选择50 000张图像作为验证集。

使用随机梯度下降训练网络,动量梯度下降法是计算梯度的指数加权平均数,并利用该数值来更新参数值,与原始的梯度下降相比,不但能使用较大的学习率,其迭代次数也减少,神经网络模型中常用的动量值为0.9,因此本文设置动量为0.9。由于权重衰减率的大小直接影响了复杂模型损失函数的值,为防止过拟合,调节模型复杂度对损失函数的影响,设置权重衰减率为1×10^{-4}。

对于CIFAR数据集,考虑到数据集的大小和图片的尺寸,设置批训练数量为64,模型的初始学习率大小设定为0.1,最大迭代次数为300,在150到225次迭代时,学习率除以10,参照MSDNet的参数设置,设置网络的子网个数为3,从低层次子网到高层次子网的卷积块个数分别为2、4和6,特征图的尺度分别为8×8、16×16和32×32,通道个数分别为64、32和16,对于每个融合块,压缩系数为0.25,这意味着新增通道中的75% 由当前的子网络产生,其余25% 由之前特征分辨率较低的子网络和当前子网络前面卷积层计算得到。通过以下两步来控制每个卷积块的层数:(1) 每个卷积块的层数设置为4;(2) 每个卷积块的层数在之前卷积块层数的基础上加2。

ImageNet数据集也用相同的训练策略,批训练数量为256,最大迭代次数为90,在30到60次迭代时,学习率除以10。参照MSDNet的参数设置,设置网络的子网个数设置为4,从低层次子网到高层次子网的卷积块个数分别为2、4、6和8,每个卷积块的层数为8,特征图的尺度分别为56×56、112×112、168×168和224×224,通道个数分别为128、64、64和32, 对于每个融合块,压缩系数为0.25。

本文参考文献[6],在CIFAR和ImageNet数据集上使用标准数据增强方法。在CIFAR-10和CIFAR-100两个数据集上,先对图像每个边界进行4个像素的零填充,然后把图像随机裁剪为32×32像素的大小。图像以0.5的概率水平翻转,RGB通道归一化通过减去相关的通道平均值然后除以其标准偏差的方式实现。在ImageNet数据集上,训练时本文参考文献[6]中训练的数据增强方案,测试时图像采取中心裁剪的方式裁剪为224×224像素的大小。

2.1 任意时间分类

在任意时间分类中,存在一个可用于每个测试样本的计算资源B,输入样本依次通过网络直到B全部用完并且输出最接近的预测值。本文用浮点运算数(Floating point operations, FLOPs)衡量网络的计算资源。 按照参考文献[6]中的设置,除了MSDNet,还选取了几个效果不错的模型作为Baseline,包括ResNet和DenseNet。

表1显示了本文模型和其他Baseline的分类准确率。总体而言,MSDNet的效果明显优于ResNet和DenseNet,而本文模型MAFDNet更优于MSDNet,特别在总体计算资源很低的时候。

在CIFAR-10(CIFAR-100)上,当计算资源为0.2×10^8至0.8×10^8FLOPs时,MAFDNet比DenseNet、ResNet和MSDNet的准确率分别高5%(9%~10%)、3.6%~6%(10%~21%)和1%(2%~5%)左右。与MSDNet相比,MAFDNet以较少的计算资源(0.2×10^8FLOPs左右)实现了更高的分类准确率。在ImageNet上,当计算资源为0至2×10^9 FLOPs时, MAFDNet准确率比MSDNet高1%~5%。虽然最终的分类器准确率MSDNet和MAFDNet相差不多(75% 左右),但是MAFDNet比MSDNet少了约25% 的计算资源。

表1 不同模型在各数据集上任意时间分类的top-1准确率

Table 1 Top-1 accuracy for classification of different models at anytime on each dataset %

数据集	计算资源/10^8	DenseNet	ResNet	MSDNet	MAFDNet
CIFAR-10	0.2～0.4	84.5	85.3	89.5～91.2	91.5～92.2
	0.4～0.6	84.5	85.3	91.2～92.8	92.2～93.1
	0.6～0.8	87.8	89.7	92.8～93.3	93.1～93.3
	0.8～1.0	87.8	90.1	93.3～93.5	93.3～93.4
	1.0～1.2	89.9	91.2	93.5～93.8	93.4～93.7
	1.2～1.4	91.2	91.8		
CIFAR-100	0.2～0.4	57.2	45.0～47.8	61.4～67.8	66.4～68.7
	0.4～0.6	57.2	47.8～61.4	67.8～69.7	68.7～72.0
	0.6～0.8	57.2～61.4	61.4～63.7	69.7～72.0	72.0～72.2
	0.8～1.0	61.4	63.7	72～73.8	72.2～73.8
	1.0～1.2	61.4～67.2	63.7～67	73.8～74.0	
	1.2～1.4	67.2～70.1	67.0～69.8		
ImageNet	0～10	54.0	54.0	54～62.6	54.0～67.4
	10～20	54.0～62.1	54.0～62.8	62.6～70.0	67.4～71.2
	20～30	62.1～66.5	62.8～69.7	70.0～74.3	71.2～75.1
	30～40	66.5	69.7	74.3～74.8	
	40～50	66.5～71.4			

可以观察到,在较少的计算资源下,MAFDNet达到了比MSDNet、ResNet和DenseNet更好的实验效果。这是因为模型的预测性能首先取决于第一层轻量级的网络,它可以很好地降低网络的计算资源。在所有数据集上,MAFDNet分类性能始终优于ResNet,这符合子网1专门用来识别简单样本的期望。因为子网1直接在低分辨率的特征图上进行操作,避免了对高分辨率特征图进行卷积,第一个分类器从而获得较高的计算效率。因为子网1被视作MAFDNet中专门优化计算资源的轻量级模型,所以MAFDNet中早期的分类器在分类任务中显示了其优势。与重复计算相似低层次特征图的MSDNet不同,MAFDNet将之前轻量级网络的特征图融合到大型网络中,从而充分获取之前的特征。当有更多的计算资源时,这种机制有效地提高了分类的准确率。

2.2 预算批分类

在预算批分类中,在预先知道有限的计算资源B的情况下,模型需要先分类一个样本集$D_{\text{test}}=\{x_1, x_2, \cdots, x_M\}$,预算批分类可能会花费少于$B/M$的计算资源用于简单样本分类,同时,花费大于$B/M$的计算资源用于困难样本分类,因此,当拥有多个测试样例,预算B就被认为是一个软约束。本文根据不同的计算消耗设置了一系列阈值,对于给定的输入图像,让它依次通过网络中的每个分类器,在分类器的输出置信度达到阈值时即退出网络,然后将该分类器的输出作为该图像的最终分类结果。

在CIFAR和ImageNet数据集上,Baseline采用ResNet和DenseNet。表2总结了实验的结果。本文选择在测试集的每个预算批分类上准确率最高的模型。在两个CIFAR数据集上的实验结果表明,MAFDNet在所有的计算资源下始终优于DenseNet、ResNet和MSDNet。一般情况下,具有多尺度密集连接结构的网络在相同计算资源条件下比其他模型的准确率更高。对于少计算资源(0.2×10^8FLOPs),在CIFAR-100上,本文模型仅使用58%左右的计算资源即可达到MSDNet在此计算资源下的分类准确率。在CIFAR-10和CIFAR-100数据集上,相比于DenseNet和ResNet,特别是在计算资源较少时,本文模型分类准确率遥遥领先。当计算资源为0.1×10^8至0.3×10^8 FLOPs时,虽然本文模

表2 不同模型在各数据集上预算批分类的top-1准确率

Table 2 Top-1 accuracy of budget batch classification for different models on each dataset %

数据集	计算资源/10^8	DenseNet	ResNet	MSDNet	MAFDNet
CIFAR-10	0.1～0.3	91.0	91.0	91.0～93.5	91.0～93.8
	0.3～0.5	91.0～91.7	91.0～92.1	93.5～93.7	93.8～94.2
	0.5～0.7	91.7～92.8	92.1～92.8	93.7～93.8	94.2～94.5
	0.7～0.9	92.8～93.3	92.8～93.3	93.8	94.5
CIFAR-100	0～0.2	60.0～61.7		63.8～68.9	65.8～71.4
	0.2～0.4	61.7～67.8	66.0～67.2	68.9～73.2	71.4～74.6
	0.4～0.6	67.8～70.5	67.2～68.8	73.2～74.6	74.6～76.0
	0.6～0.8	70.5～72.2	68.8～72.2	74.6～74.9	76.0
	0.8～1.0	72.2～73.2	72.2～73.1	74.9	
ImageNet	5～10			68.0～71.1	68.0～72.5
	10～15	68.0～69.0		71.1～73.2	72.5～74.8
	15～20	69.0～71.8	68.0～69.8	73.2～74.1	74.8～75.6
	20～25	71.8～73.1	69.8～70.7	74.1	75.6
	25～30	73.1～74.6	70.7～72.0		
	30～35		72.0～72.9		
	35～40		72.9～74.7		
	40～45		74.7～75.0		

型和MSDNet在CIFAR-10上达到的性能相近,但是在CIFAR-100数据集上中高计算资源区间(超过0.2×10^8 FLOPs)时,MAFDNet的分类准确率始终高于MSDNet 1.5%左右。当计算资源高于0.5×10^8 FLOPs时本文模型的准确率可达94.5%,高于其他3个模型的准确率。

在ImageNet上,实验结果与在CIFAR上类似。可以观察到MAFDNet的实验效果一直优于MSDNet,其top-1准确率在计算资源为1×10^9、1.5×10^9、2×10^9 FLOPs时分别高出1.4%、1.6%和1.5%。实验结果表明,随着计算资源的增加,MAFDNet比MSDNet表现更好。在同样的计算资源下,本文模型比这些深度神经网络的分类准确率更高。在同样的分类准确率下,ResNet和DenseNet相比,本文模型所消耗的的计算资源分别为其计算资源的56%和44%左右。所有的这些结果都表明本文模型性能的优越性。

2.3 消融实验

不同模型在各数据集上任意时间分类和预算批分类的top-1准确率分别如表3和表4所示,其中MAFDNet-BF为MAFDNet去除不同尺度之间的融合,MAFDNet-NF为MAFDNet去除相同尺度之间的融合。从表3、4可看出,任意时间分类和预算批分类两种情况下,MAFDNet-BF在计算资源较少(CIFAR-10、CIFAR-100和ImageNet分别不高于0.4×10^8、0.4×10^8和1×10^9 FLOPs)的时候,分类准确率与MAFDNet相比并无太大的差异。这是因为较少的计算资源即可成功分类的一般为简单样本,其在低层次网络被成功分类的概率比较大,而第一层网络并没有融合到其他层的特征,所以MAFDNet-BF在少量计算资源时与MAFDNet性能相近;在计算资源较多(CIFAR-10、CIFAR-100和ImageNet分别高于0.4×10^8、0.4×10^8和1×10^9 FLOPs)时,此时分类样本为困难样本的可能性较大,所以其在高层次网络被成功分类的概率比较大。由于MAFDNet中来自先前低层次子网的粗略特征将被重用并融合到高层次

表3 不同模型在CIFAR-10、CIFAR-100、ImageNet数据集上任意时间分类的top-1准确率

Table 3 Top-1 accuracy of classification for different models at anytime on CIFAR-10, CIFAR-100, and ImageNet datasets %

数据集	计算资源/10^8	MAFDNet-BF	MAFDNet-NF	MAFDNet
CIFAR-10	0.2～0.4	91.5～92.2	91.2～92.0	91.5～92.2
	0.4～0.6	92.2～93.1	92.0～92.7	92.2～93.1
	0.6～0.8	93.1～93.2	92.7～93.0	93.1～93.3
	0.8～1.0		93.0～93.3	93.3～93.4
	1.0～1.2			93.4～93.7
CIFAR-100	0.2～0.4	66.4～68.7	66.1～68.5	66.4～68.7
	0.4～0.6	68.7～71.4	68.5～71.5	68.7～72.0
	0.6～0.8	71.4～71.8	71.5～71.9	72.0～72.2
	0.8～1.0	71.8～72.8	71.9～73.4	72.2～73.8
ImageNet	0～10	54.0～67.1	53.6～67.1	54.0～67.4
	10～20	67.1～70.1	67.1～70.8	67.4～71.2
	20～30	70.1～73.2	70.8～74.8	71.2～75.1

表4 不同模型在CIFAR-10、CIFAR-100、ImageNet数据集上预算批分类的top-1准确率

Table 4 Top-1 accuracy of budget batch classification for different models on CIFAR-10, CIFAR-100, and ImageNet datasets %

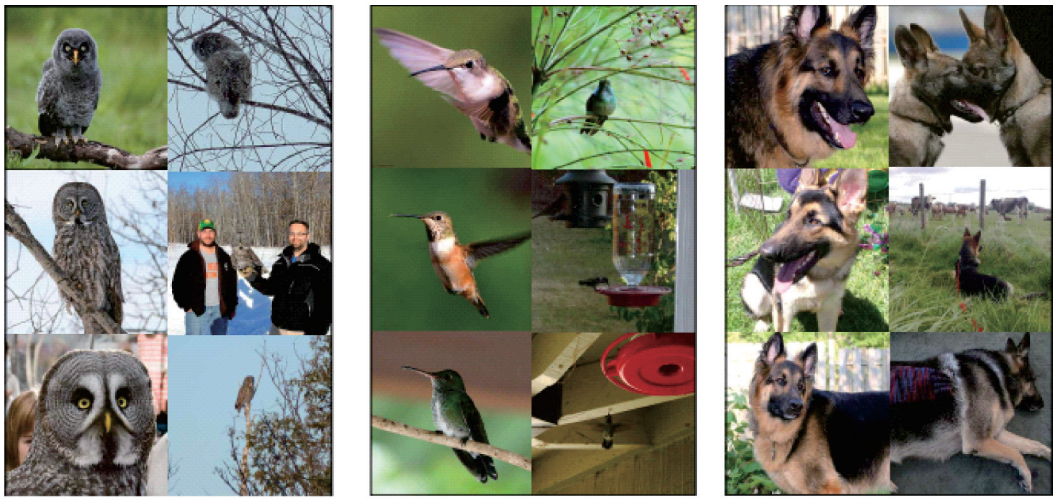
数据集	计算资源/10^8	MAFDNet-BF	MAFDNet-NF	MAFDNet
CIFAR-10	0.1～0.3	91.0～93.8	91.0～93.6	91.0～93.8
	0.3～0.5	93.8～94.1	93.6～94	93.8～94.2
	0.5～0.7		94～94.2	94.2～94.5
	0.7～0.9		94.2	94.5
CIFAR-100	0～0.2	65.8～71.4	65.6～71.2	65.8～71.4
	0.2～0.4	71.4～74.0	71.2～74.3	71.4～74.6
	0.4～0.6	74.0～75.2	74.3～75.6	74.6～76
	0.6～0.8			76.0
ImageNet	5～10	68.0～72.2	67.6～72.2	68.0～72.5
	10～15	72.2～73.9	72.2～74.6	72.5～74.8
	15～20	73.9～74.5	74.6～75.2	74.8～75.6
	20～25			75.6

子网中，所以在相同的计算资源条件下，MAFDNet-BF识别准确率会低于MAFDNet。

相较于MAFDNet-BF，MAFDNet-NF在3个数据集上各个计算资源区间的分类准确率均低于MAFDNet。这是因为MAFDNet在每一个子网中添加了密集跳跃连接，即第一个密集块的输出以及每个密集块的输出都输入给之后的所有密集块。由于MAFDNet中每一个子网在前向传播的过程中都会进行下采样，由此造成输入样本的特征丢失，严重影响了输入样本最终的分类正确率，而密集块之间的跳跃连接则将当前子网每一个密集块下采样的结果均进行融合，由此可解决下采样过程中造成的特征丢失的问题。此外，由于密集块之间的跳跃连接将所有的特征都串联起来，这样直接输入分类器会产生巨大的计算开销，因此添加了一个核大小为1×1的卷积层来减小特征数量，以降低计算消耗。

2.4 可视化

图4显示了MAFDNet识别不同难度图像的能力。在每个分图中,左侧显示了简单样本,其可被低层次的分类器正确分类;右侧显示了困难样本,其很难在低层次的分类器中达到足够的分类置信度而被正确分类,需要进入高分辨率的子网中才可被正确分类。



(a) Owl (b) Hummingbird (c) German shepherd

图4 ImageNet的可视化样本

Fig.4 Visualization of ImageNet samples

2.5 结果分析

对于MAFDNet而言,一张包含多个目标的图像可能会造成目标的遮挡从而被认为是困难样本,而不同目标的共性可能会影响特征图,容易混淆低分辨率网络的分类器,从而需要进入高分辨率的子网才能正确分类。例如图4(a)为ImageNet数据集中的一张图像,该图像为两个男人的手上有一只猫头鹰,虽然该图像中有两个人,但是该图像为猫头鹰类。很显然,下采样会将类别本身淹没在整个图像中从而使图像被网络识别为人,而非猫头鹰。此外,将图像分类为猫头鹰类可能源于人类认为一张图像中心位置的物体为该图像的重要组成成分。这种具有复杂关系的图像,只有具有明显的表征信息才可被强大的网络正确分类。

可以观察到摄像机与被拍摄物体距离较远,即小目标的图像,需要通过整个网络才可被正确分类,可以认为具有小目标的图像是困难样本。对于这一现象的一个解释是,小目标在图像进行快速下采样以后可能被完全丢失,对其分类只能通过高分辨率特征图来实现。例如,在图4(b)第2行,蜂鸟喝水中的蜂鸟很小,所以在进行下采样的时候蜂鸟可能完全消失在特征图中,这使得在网络进入到高层次的高分辨率特征图之前难以正确分类。

MAFDNet的另一种困难样本就是图像中的物体没有代表性特征。因为各种因素(如照明条件和拍摄角度)这种图像很常见。在这种情况下,本模型对其进行分类是利用替代特征而不是代表性特征。例如在图4(c)中,通过比较简单样本和困难样本,网络很容易识别德国牧羊犬,只要它的面部特征完全呈现在图像中。然而,由于没有完整的面部特征,德国牧羊犬只能在最后一个分类器中被正确分类。对于那些困难样本,网络可将德国牧羊犬的皮毛纹理作为判别特征进行分类。因此,在没有完整的面部信息的情况下,网络通过在高分辨率特征图中搜索有用的替代特征从而对德国牧羊犬进行正确的分类。

3 结束语

本文提出了一种基于DenseNet融合的复杂环境下图像自适应分类网络MAFDNet,其设计方式为:首先将处理粗糙特征的轻型子网用于图像分类,具有较高预测置信度的样本将从网络中提前退出,具有更精细细节的大规模特征将进一步用于那些在先前子网中不可靠预测的非典型图像。MAFDNet每个子网中的这种分辨率自适应机制和深度自适应可确保其较高的计算效率。在3个图像分类基准上的实验结果验证了本文提出的MAFDNet在单张图像预测和多张图像预测分类中的有效性。

参考文献:

[1] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Providence, USA: IEEE, 2016: 770-778.

[2] HUANG G, LIU Z, PLEISS G, et al. Convolutional networks with dense connectivity[EB/OL]. (2020-01-08).https://doi.org/10.48550/arXiv.2001.02394.

[3] LIU Z, LI J, SHEN Z, et al. Learning efficient convolutional networks through network slimming[C]//Proceedings of IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 2736-2744.

[4] JACOB B, KLIGYS S, CHEN B, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference[C]//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Providence, USA, 2018: 2704-2713.

[5] VEIT A, BELONGIE S. Convolutional networks with adaptive inference graphs[C]//Proceedings of European Conference on Computer Vision (ECCV). Munich, Germany: [s.n.], 2018: 3-18.

[6] HUANG G, CHEN D, LI T, et al. Multi-scale dense networks for resource efficient image classification[C]//Proceedings of International Conference on Learning Representation. Vancouver, Canada: [s.n.], 2018.

[7] LI S, ZHANG J, MA W, et al. Dynamic domain adaptation for efficient inference[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Providence, USA: IEEE, 2021: 7832-7841.

[8] CHEN Y, FAN H, XU B, et al. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019: 3435-3444.

[9] HOWARD A, SANDLER M, CHU G, et al. Searching for mobilenetv3[C]//Proceedings of 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, USA: IEEE, 2019: 1314-1324.

[10] MA N, ZHANG X, ZHENG H T, et al. Shufflenet v2: Practical guidelines for efficient CNN architecture design[C]//Proceedings of European Conference on Computer Vision (ECCV). Munich, Germany: [s.n.], 2018: 116-131.

[11] DAI X, JIANG Z, WU Z, et al. General instance distillation for object detection[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Providence, USA: IEEE, 2021: 7842-7851.

[12] FIGURNOV M, COLLINS M D, ZHU Y, et al. Spatially adaptive computation time for residual networks[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Providence, USA: IEEE, 2017: 1039-1048.

[13] CAI Shaofeng, CHEN Gang, OOI Bengchin, et al. Model slicing for supporting complex analytics with elastic inference cost and resource constraints[EB/OL]. (2019-04-18). https://doi.org/10.48550/arXiv.1904.01831.

[14] 任永梅, 杨杰, 郭志强, 等. 基于多尺度卷积神经网络的自适应熵加权决策融合船舶图像分类方法[J]. 电子与信息学报, 2021, 43(5): 1424-1431.
REN Yongmei, YANG Jie, GUO Zhiqiang, et al. An adaptive entropy-weighted decision fusion method for ship image classification based on multi-scale convolutional neural network[J]. Journal of Electronics & Information Technology, 2021, 43(5): 1424-1431.

[15] 陈鸿坤, 罗会兰. 多尺度语义信息融合的目标检测[J]. 电子与信息学报, 2021, 43(7): 2087-2095.
CHEN Hongkun, LUO Huilan. Target detection based on multi-scale semantic information fusion[J]. Journal of Electronics & Information Technology, 2021, 43(7): 2087-2095.

[16] ZHAO H, QI X, SHEN X, et al. Icnet for real-time semantic segmentation on high-resolution images[C]//Proceedings of 2018 European Conference on Computer Vision (ECCV). Munich, Germany: [s.n.], 2018: 405-420.

[17] SUN K, XIAO B, LIU D, et al. Deep high-resolution representation learning for human pose estimation[C]//Proceedings of 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, USA: IEEE, 2019: 5693-5703.

[18] VENIAT T, DENOYER L. Learning time/memory-efficient deep architectures with budgeted super networks[C]//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Providence, USA: IEEE, 2018: 3492-3500.

[19] CHIN T W, DING R, MARCULESCU D. Adascale: Towards real-time video object detection using adaptive scaling[EB/OL].(2019-02-08). https://doi.org//10.48550/arXiv.1902.02910.

[20] TONG T, LI G, LIU X, et al. Image super-resolution using dense skip connections[C]//Proceedings of IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 4799-4807.

作者简介:

叶继华(1966-),**通信作者**,男,教授,研究方向:物联网技术、数据融合、图像处理等,E-mail:yjhwcl@163.com。

黎欣(1995-),女,硕士研究生,研究方向:智能信息处理。

陈进(1998-),男,硕士研究生,研究方向:智能信息处理。

江爱文(1984-),男,教授,博士,研究方向:智能信息处理、图像处理等。

化志章(1972-),男,副教授,研究方向:算法形式推导、软件验证等。

万文涛(1963-),男,教授,研究方向:教育领导与管理研究等。

(编辑:王静)