doi:10.11835/j.issn.1000-582X.2018.05.011

基于 Xception 的细粒度图像分类

张　潜[a,b],桑　军[a,b],吴伟群[a,b],吴中元[a,b],向　宏[a,b],蔡斌[a,b]
(重庆大学 a.信息物理社会可信服务计算教育部重点实验室;b.软件学院,重庆 401331)

摘要:细粒度图像分类是对传统图像分类的子类进行更加细致的划分,实现对物体更为精细的识别,它是计算机视觉领域的一个极具挑战的研究方向。通过对现有的细粒度图像分类算法和 Xception 模型的分析,提出将 Xception 模型应用于细粒度图像分类任务。用 ImageNet 分类的预训练模型参数作为卷积层的初始化,然后对图像进行缩放、数据类型转换、数值归一化处理,以及对分类器参数随机初始化,最后对网络进行微调。在公开的细粒度图像库 CUB200-2011、Flower102 和 Stanford Dogs 上进行实验验证,得到的平均分类正确率为 71.0%、89.9%和91.4%。实验结果表明 Xception 模型在细粒度图像分类上有很好的泛化能力。由于不需要物体标注框和部位标注点等额外人工标注信息,Xception 模型用在细粒度图像分类上具有较好的通用性和鲁棒性。

关键词:细粒度图像分类;Xception;卷积神经网络;深度学习
中图分类号:TP311.1　　**文献标志码**:A　　**文章编号**:1000-582X(2018)05-085-07

Fine-grained image classification based on Xception

ZHANG Qian[a,b], *SANG Jun*[a,b], *WU Weiqun*[a,b], WU Zhongyuan[a,b], *XIANG Hong*[a,b], *CAI Bin*[a,b]
(a. Key Laboratory of Dependable Service Computing in Cyber Physical Society, Ministry of Education;
b. School of Software Engineering, Chongqing University, Chongqing 401331, P.R.China)

Abstract: Fine-grained image classification is a more detailed division of the sub-categories of traditional image classification, which achieves a more sophisticated identification of objects. And it is a very challenging research in the field of computer vision. By analyzing the existing fine-grained image classification algorithm and Xception model, we propose to apply the Xception model to the fine-grained image classification task. Initialization of convolution layers uses pre-training model parameters of ImageNet classification. Then we resize images, transform data type, normalize value, and randomly initialize classifier. Finally, the network is fine-tuned. Our method obtains 71.0%, 89.9% and 91.4% per-image accuracy on the CUB200-2011, Flower102 and Stanford Dogs dataset respectively. The experimental results show that the Xception model has good generalization ability in fine-grained image classification. Because it does not need additional annotation information such as object bounding box and part annotation, the Xception model has good versatility and robustness in fine-grained image

收稿日期:2017-12-12
基金项目:国家重点研发计划资助项目(2017YFB0802400)。
Supported by the National Key R&D Program of China (2017YFB0802400).
作者简介:张潜(1992—),男,重庆大学硕士,主要从事深度学习、软件工程等方向研究,(E-mail)zhngqn@cqu.edu.cn。
桑军(联系人),男,重庆大学教授,博士生导师,主要从事机器学习、图像处理、软件工程等方向研究,(E-mail)jsang@cqu.edu.cn。

classification.

Keywords: fine-grained image classification; Xception; convolutional neural network; deep learning

随着深度学习[1-2]技术的发展,神经网络在各个领域取得的成功较为突出。由于深度神经网络具有强大的非线性拟合能力和特征学习能力,同时得益于计算能力的提升和数据量的增大,使得深度神经网络在图像分类问题上取得了突破性的效果[3-6],也为细粒度图像分类带来了新的发展。细粒度图像分类是计算机视觉领域的一个研究热点,是对粗粒度的大类别进行更加细致的子类划分[7]。例如,粗粒度图像分类是区分图像中猫、狗、汽车和飞机等,而细粒度图像分类是区分大类下的子类,比如区分 CUB200-2011 [8]数据库中的 200 种鸟类和 Flower102[9]数据库中的 102 种花类等。由于子类间的差异更加细微,较之普通的图像分类任务,细粒度图像分类难度更大,但它能更有效地应用在生活和实践中。

细粒度图像分类可以分为强监督的细粒度图像和弱监督的细粒度图像。强监督的细粒度图像分类算法,在模型训练时,除了图像的类别标签外,还使用了物体标注框、部位标注点等额外的人工标注信息。文献[10]利用 R-CNN 算法对细粒度图像进行物体级别和局部区域的检测,在训练时需要借助物体标注框和部位标注点,测试图像还需要提供物体标注框。文献[11]提出姿态归一化算法完成对局部区域的定位检测,根据检测的标注框对图像进行裁剪,提取不同层次的局部信息进行姿态对齐,最后得出不同层的卷积特征。文献[12]模型分为局部定位和全局、局部图像块的特征学习 2 个模块。由于标注信息的获取代价十分昂贵,在很大程度上限制了这类算法的实用性。

弱监督的细粒度图像分类算法仅使用标签,而不需要额外的标注信息。局部区域信息对于细粒度图像分类至关重要,要实现更好的弱监督的细粒度图像分类,需要解决的是如何检测并定位这些局部区域。文献[13]算法利用对象级和局部级 2 个层次的特征,首次尝试不依赖额外的标注信息,仅使用类别标签来完成细粒度图像分类。文献[14]利用卷积网络特征本身产生一些关键点,再利用这些关键点来提取局部区域信息。文献[15]提出了双线性 CNN 网络模型,利用 2 个网络完成局部区域检测和特征提取。

卷积神经网络[16](CNNs, convolutional neural networks)已经被广泛应用于各个领域,比如目标检测[17]、语音识别[18]等,在图像分类应用上也取得了显著的成绩。在 ILSVRC(imagenet large scale visual recognition challenge)比赛上研究者提出了很多优秀的卷积神经网络模型,ILSVRC2014 冠军 GoogLeNet[19]采用了在同一层中使用不同大小的卷积核,即 Inception 结构,来获得不同大小的感受野以提高分类效果。继 Inception 之后又提出了一些改进,Xception[20]是 Inception 的极限版本,在 ImageNet[21]数据库上分类的 top-5 正确率是 94.5%。

目前,强监督的和弱监督的细粒度图像分类过程大多都分为 2 个步骤,先提取图像的局部区分性区域,再使用卷积神经网络对这些区域进行特征学习。不同于上述方法,研究采用的是 Xception 模型,不需要分 2 步进行,让神经网络自动学习到局部区分性特征。同时,不需要物体标注框和部位标注点信息,仅使用类别标签信息。使用 Xception 模型在公开的细粒度图像库 CUB200-2011、Flower102 和 Stanford Dogs[22]3 个数据库上进行了实验分析,实验结果验证了 Xception 模型可以很好地应用于细粒度图像分类任务。

1 Xception 模型概述

2014 年,由谷歌研究院提出的 GoogLeNet 在 ILSVRC 图像分类获得冠军。GoogLeNet 对网络中传统卷积层进行修改,提出在同一层中使用不同大小卷积核的 Inception 结构。后续相继提出了改进版本 Inception V_2[23]、Inception V_3[24]、Inception V_4[25]。Inception V_2 一方面使用 3×3 的卷积核代替 5×5 的大卷积核,在降低参数的同时建立更复杂的非线性变换,另一方面使用 Batch Normalization 来减小神经网络的

训练难度，其公式如下

$$y^{(k)}=\gamma^{(k)}\hat{x}^{(k)}+\beta^{(k)}, \tag{1}$$

其中：γ 和 β 为参数；$\hat{x}$ 为一个批次的归一化项。

Inception V_3 引入 Factorization into small convolutions 的思想，将一个较大的二维卷积拆成 2 个较小的一维卷积，比如将 7×7 卷积拆成 1×7 卷积和 7×1 卷积，减少大量参数加速运算并降低过拟合；同时增加一层非线性变换，扩展模型的表达能力。Inception V_4 在 Inception V_3 基础上进一步改进，结合了微软的 ResNet[3] 思想。Xception 也是在 Inception V_3 基础上进行的改进，使用了深度可分离卷积。

Inception 的思想是将一个卷积核需要同时映射跨通道相关性和空间相关性的过程分解成一系列相互独立的操作，即 Inception 模块首先处理跨通道相关性，通过一组 1×1 卷积，将输入数据映射到 3 或 4 个小于原始输入的不同空间；然后处理空间相关性，通过 3×3 或者 5×5 卷积将所有相关性映射到更小的 3D 空间。实际上 Inception 背后的基本的假设是使跨通道相关性和空间相关性充分解耦。在 Inception 假设和思想基础下，对 Inception V_3 继续改进，把 Inception V_3 中标准的 Inception 模块（如图 1 所示）进行简化，只使用一种规格的卷积（例如 3×3），并且不含平均池化，结果如图 2。然后对图 2 中的 Inception 模块重新定义，用一个大的 1×1 的卷积，在不重叠的通道区块上进行空间卷积（如图 3 所示）。很自然地发现通道区块的数量越多，以及跨通道相关性和空间相关性映射完全分开的假设更合理。基于上述的发现与假设提出了“极致”Inception 模块（如图 4 所示），为 Xception 网络中的重要模块，其首先使用 1×1 卷积映射跨通道相关性，然后在每个 1×1 卷积的输出通道上都有一个独立的空间卷积来映射空间相关性。普通卷积把所有输入通道视为单区块情况，深度可分卷积把每个通道当成为一个区块，Inception 模块居于其间，将数百个通道划分成 3 或 4 个区块。而“极致”Inception 是把所有的通道视为一个区块，即是一个可分离的卷积。同时 Xception 加入的类似 ResNet 的残差连接机制显著加快了 Xception 的收敛，并获得了更高的正确率。

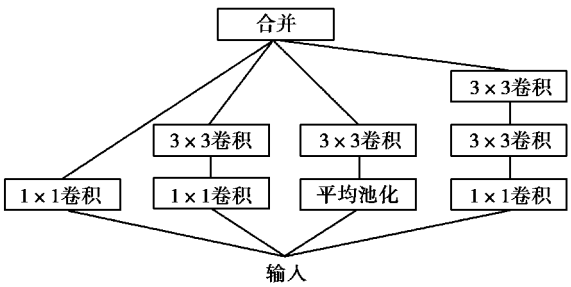

图 1　Inception V_3 中标准的 Inception 模块结构[24]

Fig.1　A canonical Inception module (Inception V_3)[24]

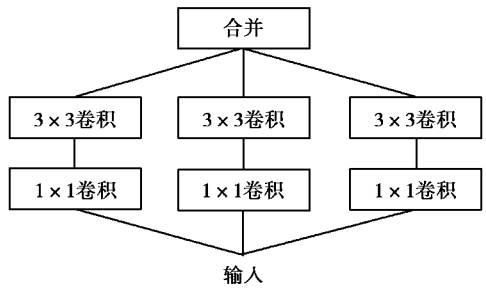

图 2　简化的 Inception 模块结构[20]

Fig.2　A simplified Inception module[20]

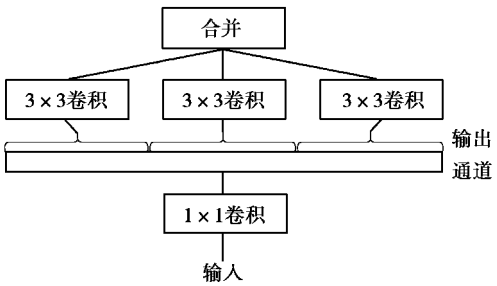

图 3　完全等价的简化 Inception 模块[20]

Fig.3　A strictly equivalent reformulation of the simplified Inception module[20]

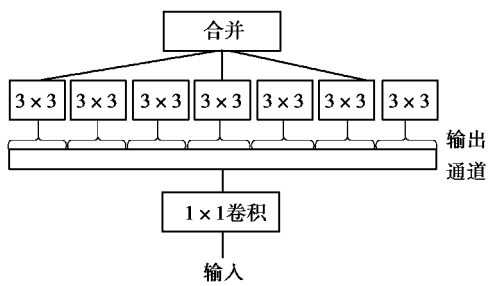

图 4　“极致”版本的 Inception 模块[20]

Fig.4　An “extreme” version of Inception module[20]

Xception模型使用深度可分离卷积增加网络宽度,不仅提升了分类的正确率,也增强了网络对细微特征的学习能力,提供了Xception用在弱监督的细粒度图像分类的可行性。基于上述分析,尝试将Xception应用于细粒度图像分类。

2 实验与结果

2.1 实验数据库

选择CUB200-2011、Flower102和Stanford Dogs 3个公开的细粒度图像库进行分析研究。其中CUB200-2011数据库是细粒度图像分类领域一个经典的数据库,也是最常用的一个数据库,共包含200种不同类别的鸟,每个类别包含41到60张图像不等,共11 788张图像。Flower102数据库分为2种不同规模的版本,分别包含17种类别和102种类别的花。实验用的是102种类别的数据库,每个类别包含了40到258张图像数据,总共有8 189张图像。Stanford Dogs数据包库包含120类狗,每类包含148到252张图像不等,总共有20 580张图像。

图5展示了3种数据库的部分样本,对于每个数据库,随机采集了4张来自不同类别的图像。从这些图片中可以发现,3种数据库的图像背景复杂,同一类别内,目标的姿态多样、光照变化大;不同类别之间的差异十分细微,目标的形状、颜色非常相似。

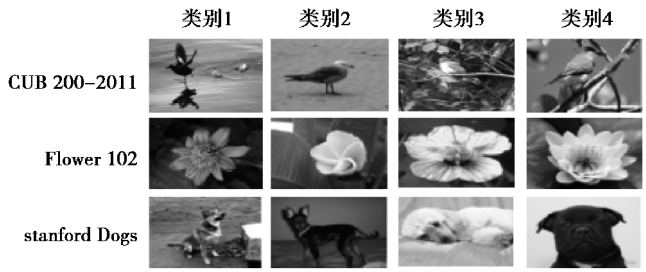

图5 3种数据库样本示例

Fig.5 The example of three datasets

2.2 实验过程

第一步:图像数据归一化。首先将图像缩放到299×299像素,再把图像的每个像素数据类型转化为浮点型,并归一化到[−1,+1],归一化公式如下

$$J=(\boldsymbol{I}/255.0-0.5)\times 2, \tag{2}$$

其中:$\boldsymbol{I}$为图像像素矩阵,J为数据类型转换和归一化的结果;

第二步:模型参数初始化。微调是训练神经网络的一种常用的方式,即用已训练好的公开模型和参数,加上自己的数据,来训练新的模型。微调相当于使用公开预训练好的模型来提取特征,然后再用到自己的分类中。微调不用完全重新训练模型,从而提高效率,能够在比较少的迭代次数得到一个比较好的效果。在数据量不是很大的情况下,微调是一个比较好的选择。本实验选择的数据库中每类的数据不是很多,所以使用ImageNet分类[26]的预训练模型参数作为网络卷积层的初始化,并对softmax分类器随机初始化;

第三步:模型训练。采用有监督学习方式,利用BP算法对卷积层参数进行细微的调整,以及对分类器参数进行训练,即比较分类器的输出值与期望输出的差别,得到误差信号并进行反向传播来微调各层参数,直到损失趋于不变。

训练过程中使用动态调整学习率的Adam优化器,避免对学习率的手动调节,参数更新公式如下

$$\theta_{t+1}=\theta_t-\frac{\eta}{\sqrt{\widetilde{v}_t}}\widetilde{m}_t, \tag{3}$$

其中：θ 为需要更新的参数；$\widetilde{m}_t$ 和 $\widetilde{v}_t$ 分别为梯度第一个时刻平均值和第二个时刻方差；η 为学习率。

为了提高网络的泛化能力，降低网络过拟合，采用 Dropout 正则化技术，值取 0.5，公式如下

$$y_{(l+1)i}=F(W_i^{l+1}y^l r^l+b^{(l+1)}), \tag{4}$$

其中：$y_{(l+1)i}$ 为 $l+1$ 层的第 i 个输出；$l+1$ 层的权重；y^l 为 l 层的输出；r^l 为以 0.5 的概率取值 0 或 1；$b^{(l+1)}$ 为 $l+1$ 层的偏执；

第四步：分类预测。测试数据在已经优化好的神经网络中进行正向传播获得网络输出，并将得到的实际输出与标签数据值进行比较，从而判别测试图像所属的类别，统计正确分类的数量，计算出正确率。

大多数深度学习方法对于数据库的处理是将数据库以一定比例分配生成训练集、验证集和测试集。在实验中，将 CUB200-2011、Stanford Dogs 和 Flower102 的 3 个数据库按 8∶2的比例随机生成训练验证集和测试集；在训练验证集中，将其 80%作为训练数据集，剩余的 20%作为验证数据集。

经实验训练、验证以及测试均在 NVIDIA Tesla K80 GPU 上完成，基于深度学习框架 Keras 进行，通过数次实验择优原则获取结果。

2.3　实验结果分析

在 CUB200、Flower102 和 Stanford Dogs 数据库上的分类结果[27-28]如表 1 所示。

表 1　不同数据库上的细粒度图像分类结果

Table 1　Fine-grained image classification results on different datasets

模型	数据库	分类结果/%
B-CNN[15](AlexNet)	CUB200-2011	72.5
Xception	CUB200-2011	71.0
Murray	Flower102	84.6
Xception	Flower102	89.9
Xiao	Stanford Dogs	88.9
Xception	Stanford Dogs	91.4

从表 1 可以看出，Xception 模型在这些细粒度图像分类数据库上的分类效果整体较好。在 CUB200-2011 数据库上，Xception 模型的正确率略微低于 Biliner-CNN 模型，一方面由于去除验证集和测试集每类能参与训练的数据量比较少，另一方面是由于 Biliner-CNN[15]采用了 2 个特征提取函数，即 2 个网络相互协作完成细粒度图像分类过程中物体、局部区域的检测和特征提取，而 Xception 单模型学习细粒度的区分性特征。在 Flower102 数据库上 Xception 模型正确率比 Murray 方法正确率高出 5.3%。在 Stanford Dogs 数据库上 Xception 模型准确率比 Xiao 方法正确率高出 2.5%。

Xception 模型在 CUB200-2011、Flower102 和 Stanford Dogs 3 个数据库上的正确率都取得较优的效果，说明 Xception 模型在细粒度图像分类上具有很好细微特征学习能力。而在 CUB200-2011 数据库上正确率偏低主要是因为该数据库不仅种类很多、每类的数据量较少，而且有些图像目标对象很小。在 Stanford Dogs 数据库上正确率较高在于该数据库的图像来源于 ImageNet 数据库，而在 Flower102 数据库上进行实验，取得 88.9%的正确率，进一步验证了 Xception 模型对细粒度图像分类的实用性。

3　结　论

细粒度图像分为基于强监督信息和基于弱监督信息 2 类，选择仅使用标签信息的弱监督方式。使用 Xception 模型在 3 个公开的细粒度图像分类数据库上进行研究分析。通过实验对比发现，基于 Xception 方法分类准确率均较为优异，具有很好的泛化能力，用于细粒度图像分类有较好的鲁棒性。由于细粒度图像的

类别精度更加细致,类间差异更加细微,往往只能借助于微小的局部差异才能区分出不同的类别,因此神经网络学习这些细微的特征需要更多的数据。在未来的工作里,将结合双通道思想,一个通道学习局部有区分性信息检测,一个通道学习物体对象级别检测,来减少对数据量的需要,并进一步提升细粒度图像分类的正确率。

参考文献:

[1] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504.

[2] Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[3] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. Las Vegas, USA: IEEE, 2016: 770-778.

[4] Krizhevsky A, Sutskever I, Hinton G E, et al. Imagenet classiflcation with deep convolutional neural networks[C]// Proceedings of the 25th International Conference on Neural Information Processing Systems. Nevada-December 03-06, 2012. Lake Tahoe, USA: ACM, 2012: 1097-1105.

[5] Huang G, Liu Z, Weinberger K Q, et al. Densely connected convolutional networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.USA: IEEE, 2017: 1-4.

[6] Chen Y, Li J, Xiao H, et al. Dual path networks[C]// Advances in Neural Information Processing Systems. Long Beach: IEEE, 2017: 4470-4478.

[7] 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.
LUO Jianhao, WU Jianxin. A Survey on fine-grained image categorization using deep convolutional features[J]. Acta Automatica Sinica, 2017, 43(8): 1306-1318. (in Chinese)

[8] Wah C, Branson S, Welinder P, et al. The Caltech-UCSD Birds-200-2011 dataset[C]// Technical Report CNS-TR-2011-001, USA: California Institute of Technology, 2011: 38-42.

[9] Nilsback M E, Zisserman A. Automated flower classification over a large number of classes[C]// Proceedings of the 6th Indian Conference on Computer Vision, Graphics & Image Processing, 2008. Bhubaneswar, India: IEEE, 2008: 722-729.

[10] Zhang N, Donahue J, Girshick R, et al. Part-based r-CNNs for fine-grained category detection[C]// European Conference on Computer Vision, 2014. Zurich, Switzerland: Springer, 2014: 834-849.

[11] Branson S, Van Horn G, Belongie S, et al. Bird species categorization using pose normalized deep convolutional nets[EB/OL].[2017-11-04].https://arxiv.org/abs/1406.2952.

[12] Wei X S, Xie C W, Wu J, et al. Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization[J]. Pattern Recognition, 2018, 76(2): 704-714.

[13] Xiao T, Xu Y, Yang K, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. Boston, Massachusetts: CVPR, 2015: 842-850.

[14] Simon M, Rodner E. Neural activation constellations: Unsupervised part model discovery with convolutional networks[C]// Proceedings of the IEEE International Conference on Computer Vision, 2015. Santiago, Chile: CVPR, 2015: 1143-1151.

[15] 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, 2015. Santiago, Chile: ICCV, 2015: 1449-1457.

[16] LeCun Y, Boser B E, Denker J S, et al. Handwritten digit recognition with a back-propagation network[C]// Advancesin Neural Information Processing Systems, Denver 1989. Colorado, USA: NIPS Conference, 1989: 396-404.

[17] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[C]// European Conference on Computer Vision, 2014. Zurich, Switzerland: Springer, Cham, 2014: 346-361.

[18] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four resear chgroups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[19] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. Boston, Massachusetts: CVPR, 2015: 1-9.

[20] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2017: 1800-1807.

[21] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database[C]// Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009. Miami, USA: IEEE, 2009: 248-255.

[22] Khosla A, Jayadevaprakash N, Yao B, et al. Novel dataset for fine-grained image categorization: Stanford dogs[C]// Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC).[S.l.]: IEEE, 2011, 3-8.

[23] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]// International Conference on Machine Learning, 2015. Lille, France: ICML, 2015: 448-456.

[24] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. Las Vegas, NV, USA: CVPR, 2016: 2818-2826.

[25] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017. San Francisco, California, USA: AAAI, 2017: 4278-4284.

[26] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3) : 211-252.

[27] Murray N, Perronnin F. Generalized max pooling[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, Columbus, USA: CVPR, 2014: 2473-2480.

[28] Liu X, Xia T, Wang J, et al. Fully convolutional attention localization networks: efficient attention localization for fine-grained recognition[EB/OL]. [2017-11-04]. https://arxiv.org/abs/1603.06765.

(编辑　侯　湘)