基于深度学习的细粒度分类综述 [1]

林愈凯[1]，刘丁烨[1]，宋昕帅[1]，孙广岩[1]，廖子睿[1]，李云飞[1]

1. 中山大学智能工程学院，广东，深圳

摘　要：细粒度视觉分类，是在分出基础类别后对子类进行细分，如区分鸟的品种。相较于粗粒度视觉分类，它拥有更接近的外观和特征，再加上采集中的干扰，导致分类更加具有难度。早期的细粒度视觉分类算法基本基于特征提取的算法。随着研究的深入，两种深度学习方法：强监督算法、弱监督算法横空出世，加速了这一领域的进步。本文将着重介绍两种方法在分类网络上的流程、各模块及内在工作原理。

关键词：细粒度分类；深度学习；卷积神经网络

强监督 – 细粒度图像分类方法

细粒度图像分类任务（FGVC）的重点是，细粒度对象的差异只反映在细微之处，比如鸟嘴的形状和尾巴的毛色。如何有效地检测目标并从中发现重要的局部信息，已成为细粒度图像分类算法需要解决的关键问题。因此，通用 FGVC 算法的过程如下（FGVC 过程）：首先，定位对象的重要部分，对齐这部分的位置，并从中提取特征进行分类。

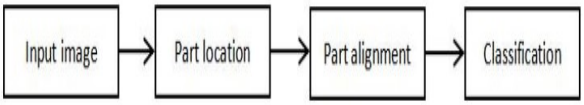

图 1

根据模型在训练中使用的训练数据标签，细粒度视觉分类可分为两类：强监督 FGVC 和弱监督 FGVC。强监督细粒度视觉分类是指，为了在训练过程中获得更好的分类精度，除了类别标签之外，还使用了其他注释，例如对象边界框和零件注释图像。在本节中，我们将介绍最近主流的强监督细粒度分类算法。

1 R-CNN

近年来，大多数图像识别方法都依赖于强大的卷积特征，这比使用常规手动特征要有效得多，细粒度图像识别也适用。研究者 Zhang et al. [1]提出了基于部件的 R-CNN 算法，该算法使用 R-CNN[2]来检测对象。它需要四个过程。

首先，通过选择性搜索生成对象建议和部分建议[3]。随后，与 R-CNN 类似，使用注释对对象测器和部分检测器进行训练。第三，通过对检测器得到的检测盒添加几何约束，获得最佳的目标和部分检测。最后，输入图像块训练 CNN，CNN 学习对象和部分的特征，并将这三个特征连接起来作为细粒度图像的表示。培训期间仅提供对象边界框和部分标注，但 R-CNN 通常要求测试图像具有边界框才能达到令人满意的精度，这在实际应用中是不现实的。

原理：

如图 2 所示系统(1)获取一幅输入图像，(2)提取大约 2000 个自下而上的区域建议，(3)使用大型卷积神经网络(CNN)计算每个建议的特征，然后(4)使用特定类的线性支持向量机（SVM）对每个区域进行分类。

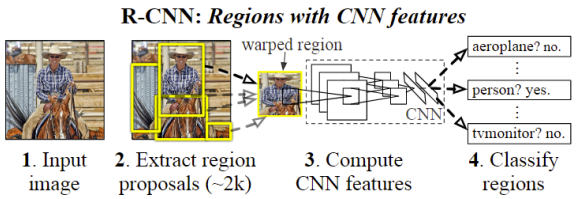

图 2

在特征提取，即 CNN 的构建上，R-CNN 使用 Krizhevsky 等人描述的 CNN 的 Caffe 实现从每个区域建议中提取一个 4096 维的特征向量。特征通过 5 个卷积层和 2 个全连通层向前传播平均减去 227 × 227 RGB 图像来计算。

在测试时，对测试图像进行选择性搜索，提

[1] 指导老师：赵屾，刘梦源

取大约 2000 个区域。卷曲每个区域以满足 CNN 输入格式，并通过CNN向前传播，以计算特征。然后，对于每个类，使用为该类训练的SVM对提取的每个特征向量进行评分。给定图像中所有得分区域，用贪心非最大抑制(对每个类独立)，如果一个区域与一个得分高于学习阈值的选定区域有交集-过并集(IoU)重叠，则拒绝该区域。

2 Pose Normalized CNN

受到上述 R-CNN 方法启发，S.Branson 等人提出的姿势归一化 CNN[4]（Pose Normalized CNN），它应用 DPM 算法[5]预测部分标注，以获得对象和部分的检测框。与之前的工作不同，它对齐部分水平图像块的姿态。然后，随着 CNN 层数的增加，提取的特征具有更高的语义，因此该模型针对细粒度图像的不同局部区域提取不同层的卷积特征，并将其连接起来作为图像的表示。

原理：

给定一幅测试图像，使用检测到的关键点组来计算与原型模型对齐的多个卷曲图像区域。每个区域通过深度卷积网络提供信息，并从多个CNN层中提取特征。特征被连接并提供给分类器。流程如图 3 所示：

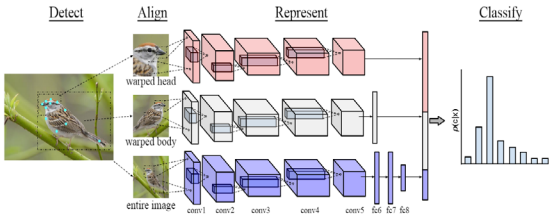

图 3

多层CNN特征不同的对齐模型，即对于对象以及不同的部分类型，将不同的层数CNN提取的特征用于细粒度分类。通过 8 层 CNN 网络的递进可以看作是一个从低特征到中特征到高特征的递进。后面的层在更大的尺度上聚合了更复杂的结构信息——用 max-pooling 交织的卷积层序列能够捕获可变形的部分，而全连接层可以捕获复杂的共现统计。另一方面，后期层保留的空间信息越来越少，因为每个卷积层之间的 max-pooling 依次降低了卷积输出的分辨率，并且层之间完全连通将降低空间位置的语义。因此，该网络的不同层更适合于不同的对齐模型，将多个层次的对齐组合起来可以产生更好的性能。

最终的特征空间将来自多个区域和层的特征连接起来，并使用单对全线性支持向量机来学习每个特征的权重。使用 SVM(而不是 CNN 使用的多类物流损失)主要是为了在合并多个区域时的技术方便。为了处理不同量级的层，在特征提取过程中，每一个CNN层输出都被独立地归一化。

3 end-to-end Mask-CNN

Wei 等人[6]提出了一种新的端到端掩码 CNN（end-to-end Mask-CNN）模型，用于定位部分并选择鸟类 FGVC 的描述符（descriptors）。结构如图 4 所示：

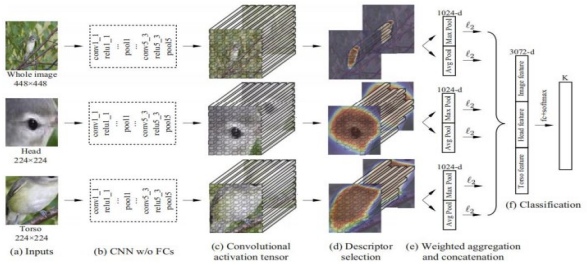

图 4

该模型分两个阶段建立，一个阶段是生成对象/部分掩码，用于定位对象/部分和选择深层描述符。在第一阶段，学习基于部分的分割模型，以使用完全卷积网络（FCN）[7]， 依靠通过部分标注获得的头部和躯干的最小外围矩形，定位对象的头部和躯干。在第二阶段，基于这些掩码，构建掩码 CNN，用于联合训练和捕获对象级和部分级信息。

原理：

上图4为四流Mask-CNN的架构。这四个流分别对应于整个图像、头部、躯干和物体图像。删除了完全连接层。由于采用了描述符选择方案，M-CNN 可以丢弃大量背景对应的描述符，有利于细粒度识别。

进一步使用掩码从分割中选择有用的深度卷积描述符。将 15 个部分关键点周围的图像裁剪处理为 15 个分割前景类，并使用 FCN 解决 16 类分割任务（因为使用了 CUB200-2011 数据集）。在得到训练好的 FCN 后，将这些部分的点位置定位在最后一层卷积中。然后，将对应于 15 个部分和整个物体的深度激活叠加在一起。分类采用全连接层。,M-CNN 只需要定位两个主要的部分(头部和躯干)，这就造成了分割问题更容易，更

准确。网络构建如下图：

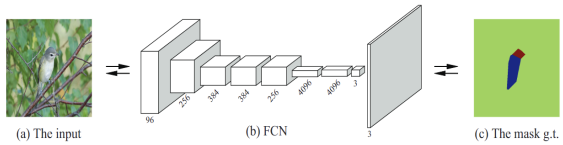

图 5

全卷积网络(FCN)设计用于像素级标记。FCN 可以获取任意分辨率的输入图像，并产生相应尺寸的输出。研究者使用 CUB200-2011 数据集进行鸟类细粒度图像分类，在此基础上，生成了两个局部掩码的真值（ground truth）。一种是头罩，另一种是躯干掩码，图片的其余部分是背景。因此，将部分掩码学习过程建模为三级分割问题。为了有效的训练，所有的训练和测试细粒度图像都保持原来的分辨率。然后，在原始图像中间裁剪一个 384 × 384 的图像补丁作为输入。掩码学习网络结构如图 5 所示。

在 FCN 推断期间，不使用任何注释，为每个图像返回三个类热图(大小与原始输入图像相同)。如图 6 所示。在这些图像中，学习过的掩码被覆盖在原始图像上。头部部分用红色标出，躯干部分用蓝色标出。预测的背景像素为黑色。

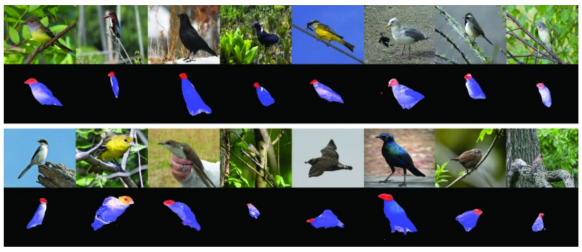

图 6

这两个部分的掩码，如果准确预测，将有利于随后的深度描述符选择过程和最终的细粒度分类。因此，在训练和测试过程中，使用预测的掩码在 M-CNN 中进行部分定位和描述符选择。研究者还将这两个掩码组合起来，形成整个对象的掩码，称为对象掩码。

在获得物体和部分掩模后，构建四流 M-CNN 进行联合训练。以整个图像流为例来说明 M-CNN 中每个流的流程。整个图像流的输入是用 h × h 调整大小的原始图像。输入图像被送入传统的卷积神经网络，但完全连接层被丢弃。也就是说，M-CNN 中使用的 CNN 模型只包含了 convolutional, ReLU 和 pooling 三层，这大大降低了 M-CNN 模型的规模。

具体来说，如果使用 VGG-16 作为基线模型，如流程图 4（b），保留 pool5 之前的层(包括 pool5)。当输入图像为 224 × 224 时，在 pool5 中得到一个 7 × 7 × 512 的激活张量因此，如流程图 4（c）（d），得到 49 个 512-d 的深度卷积描述符，它们也对应于输入图像中的 7×7 空间位置。然后这些掩码用于选择有用和有意义的深度描述符。

对于这些选定的描述符，在端到端 M-CNN 学习过程中，分别取平均值和最大值池化将其放入两个 512-d 的特征向量中。然后，每一个都遵循 l_2-范化。之后，将它们拼接成 1024-d 的特征，作为整个图像流的最终表示。如流程图 4（e）所示。

头部和躯干流的处理步骤与整体图像的处理步骤相似。但是，与整个图像流的输入不同，生成头部和躯干流的输入图像如下：在得到头部和躯干两部分的掩码后，将部分掩码作为部分检测器，在输入图像中对头部和躯干部分进行定位。对于每个部分，我们返回包含部分掩码区域的最小矩形边框。在矩形边框的基础上，裁剪作为零件流输入的图像补丁。

在流程图 4(f)所示的分类步骤中，最终的 4096 d 图像表示是整个图像、头部、躯干和物体特征的拼接。M-CNN 的最后一层是一个 200 路分类(fc+softmax)层，用于对 CUB200-2011 数据集进行分类。四个流 M-CNN 是端到端学习的，同时学习四个 cnn 的参数。在训练 M-CNN 时，学习到的 FCN 分割网络的参数是固定的。

4 HSnet framework

此外，为了更好地搜索 CNN 深度特征图上的信息部分，Michael 等人[8]在 2017 年提出了 HSnet 框架（HSnet framework）来评估边界框候选。该框架使用长-短期记忆（LSTM）[9]记录所有访问的边界框，在这之前由启发式函数对状态评分，并使用后续函数用于回归，从而不断优化局部细节的边界框，最终得到评分最高的细节部分的边界框。

原理：

给定一幅图像，我们使用 HSnet 顺序搜索图像中的判别边界框，并融合所有未暴露的图像部

分进行细粒度识别。HSnet 提供了一个统一的框架，共同学习评估搜索状态的启发式函数和在搜索空间中提出新状态的后继函数。如图 7 所示。图像定义了图像边界框的搜索空间，由卷积神经网络(CNN)的深度特征表示。在这个搜索空间中，运行一个搜索算法，该算法对给定的搜索状态提出并移动到一个新的状态，直到一个时间限制。给定时间的搜索状态由在此时间之前访问过的边界框建议定义。搜索由两个函数定义。当前状态的后继函数在搜索空间中提出一个新状态。启发式函数对图像中所有访问过的边界框的状态进行评分，从而引导搜索到最适合识别的图像部分。当搜索时间过期时，使用最后一个状态上的分类器进行识别。

Michael 等人的主要贡献是制定了一种新的深度架构，称为 HSnet，用于计算图像中顺序搜索的上述启发式和后续函数。HSnet 通过 CNN 连接到图像，由三个部分组成:h 层用于计算启发式函数，s 层用于实现后续函数，Long Short - term Memory (LSTM)用于捕获搜索轨迹上的远程依赖关系。因此，HSnet 的作用是双重的:评估候选边界框和提出新的候选边界框。由于 LSTM 有内存，顺序搜索不是贪婪的。也就是说，LSTM 的内存允许将搜索状态累积定义为在该状态之前访问过的所有边界框的集合。因此，HSnet 有一个内置的健壮机制来处理不确定性(例如遮挡、缺失部件、形状变形)，因为识别不完全依赖于搜索结束时未覆盖的最后一组边界框。

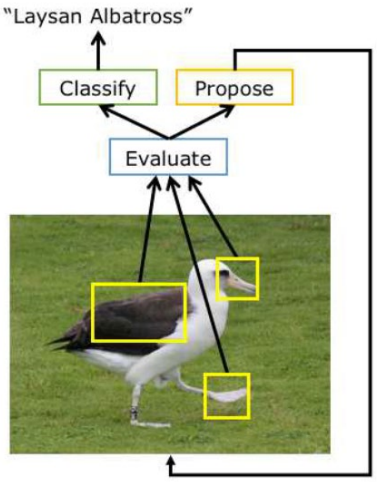

图 7

给定时间的搜索状态由图片中在此时间之前访问过的边界框建议定义。搜索由启发式函数 H 和后继函数 S 引导，通过 HSnet 统一和共同学习。S 提出新的状态，并对状态进行评分，直到时间界限 τ 。HSnet 的一个组成部分是 LSTM，它的内存融合了沿着搜索轨迹访问的所有候选边界框。HSnet 的 softmax 输出最终识别。如图 8 所示：

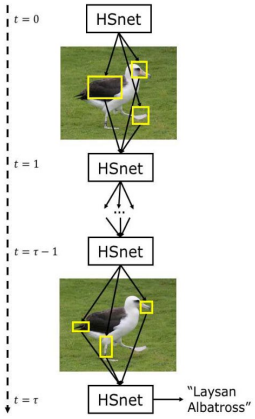

图 8

如图 9 所示，HSnet 由 h 层、s 层和 LSTM 组成。CNN 从图像中提取一个深度特征图x。H 层实现 H。它从k个现有边界框$[x^{(1)} \dots x^{(k)}]$ (红色标志)计算启发式得分φ。这些边界框也被称为感兴趣区域(regions of interest, ROIs)。每个 ROI 被传递到一个兴趣池层(ROIP)区域，以获得一个固定大小的向量表示。然后所有的 ROI 被连接并通过一个多层感知器(MLP)产生φ作为输出。S 层实现 S。它获取φ， LSTM 内存和边界框的位置$[l^{(1)} \dots l^{(k)}]$作为输入，提出k个相对于$[l^{(1)} \dots l^{(k)}]$的空间偏移$[o^{(1)} \dots o^{(k)}]$。它们被循环链接反馈回图像中来定义新的边界框。经过搜索步骤τ后，使用 soft-max 层 C 进行细粒度识别$\hat{y}$。SM 是 softmax 层，R 是回归。

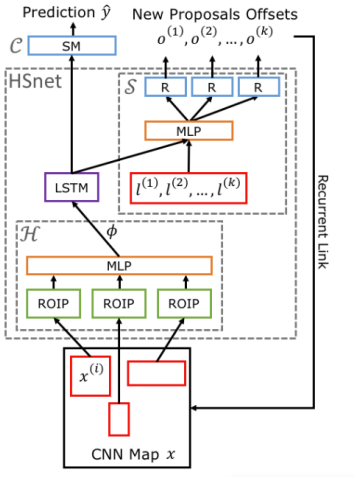

图 9

LSTM 是一种带有记忆单元的递归神经网络。实验中使用 1 层 LSTM 架构。考虑在两种情况下学习 HSnet:(1)可以访问部分位置的注释，(2)训练数据中不提供部分注释。在这两种设置中，HSnet 中所有三个组件的端到端学习都是使用基于梯度的时间反向传播(BPTT)进行的，这通常用于训练 LSTM。BPTT 反向传播了搜索到达时间节点 τ 时训练数据产生的标准交叉熵损失，分类损失通过额外的损失函数进行规范，对上述两种设置分别定义不同的损失函数。

当部分注释可用时，能够对 HSnet 进行规则学习，以预测边界框的位置，从而使它们更好地与真实部分注释对齐。具体来说，正则化学习与欧几里得距离之间的预测边界框和最近的部分真值。对于k个部分，正则化是k个欧氏距离的和。在每个搜索步骤t计算这个和，并用正则化参数λ_t对其进行加权。因此，在这种情况下的正则化损失被定义为

$$\mathrm{L} = -\log p(y) + \sum_{t=1}^{\tau} \lambda_t \sum_{i=1}^{k} \left\| l^{(i)} - \hat{l}_t^{(l)} \right\|^2$$

其中第一项是交叉熵损失，第二项是正则化。y 表示真值类标签，$p(y)$表示真值类 HSnet 的 soft-max 得分，$l(i)$是部分i的真值位置，$\hat{l}_t^{(l)}$是在搜索步骤t预测到离$l(i)$最近的边界框的位置(贪心算法)，λ_t是在搜索步骤t正规化超参数。

当训练数据中没有提供基本事实部分的标注时，研究者试图将 HSnet 的学习正规化，以预测边界框的位置，使它们在视觉上是不同的。为此，用一个由行列式点过程(DPP)表征的项来正则化交叉熵损失。DPP 已被广泛用于学习的正规化。在这种情况下，规则损失被定义为

$$\mathrm{L}(\hat{y},\mathrm{y}) = -\log p(y) - \sum_{t=1}^{\tau} \lambda_t \log P_t$$

其中第一项是交叉熵损失，第二项是 DPP 正则化。超参数λ_t控制 DPP 正则化的大小。P_t是在搜索步骤t具有不同边界框的概率，定义为

$$P_t = \frac{det|\Omega_k|}{det|\Omega + I|}$$

Ω是所有可能的边界框之间有亲和力的半正定核矩阵，Ω_k表示Ω对k个选定边界框的限制。亲和度被指定为位置之间的欧几里得距离的倒数。行列式$det|\Omega_k|$量化了k个位置的多样性。因此，多样性越高，P_t就越高。

5 总结

强监督算法利用附加标注（如边界框和部分标注）来获取对象的位置和大小，这有利于改善局部和全局之间的关联，因此不难获得更高的分类精度。然而，带注释的方法效率低，注释昂贵，这在一定程度上限制了强监督算法的实用性。因此，强监督 FGVC 算法逐渐失去了研究价值。

弱监督-细粒度图像分类方法

由于强监督 FGVC 算法依赖于人工标注，而人工标注是十分昂贵而且耗时的，在实际应用中受到限制。近年来，随着深度学习的发展，许多弱监督细粒度分类算法被提出，它们不需要对图像一部分进行手工标注，只需要使用弱监督，例如图像类别标签或文本描述。

为了解决细粒度图像的挑战，弱监督细粒度分类算法也致力于检测物体的关键部位（如头部和躯干）并提取鉴别特征。近年来，弱监督细粒度图像分类算法发展迅速，大致可分为三类，即：(a)基于部分的方法；(b)基于端到端视觉编码的方法；(c)使用外部先验知识。这些方法在细粒度图像数据集上表现良好，在 CUB200-2011 数据集上的最佳分类精度达到 90.3%，在一定程度上超过了强监督分类算法。本节根据上述三种类型介绍了最近相关的弱监督 FGVC 算法，以及它在 CUB200-2011 上的性能[10]。

1 Part-based

基于部分的方法遵循通用 FGVC 算法（图 2）的过程，该算法将细粒度识别分为两部分，即区分部分位置和从部分中学习细粒度特征。利用深卷积神经网络的卷积特性对判别区域进行定位。在细粒度特征学习过程中，从每个区域提取特征，并将其组合在一起，最后进行分类。

这些方法通常通过注意机制或聚类自动发现区分区域。 其中两级注意[11]是没有附加注释，仅使用类别标签进行细粒度图像分类的第一篇工作。

1.1 背景

大多数细粒度分类系统都遵循以下流程：查找前景对象或对象部分（where）以提取有区分度的特征（what）。

本论文的设计十分直观：执行细粒度分类需要首先“看到”对象，然后是对象中最具辨别力的部分。在图像中找到吉娃娃需要先看到一只狗，然后专注于它的重要特征，这些特征将它与其他品种的狗区分开来。

1.2 贡献

本论文将一个在 ILSVRC2012 的 1K 类别上预先训练的卷积神经网络（CNN）转化为一个 FilterNet。FilterNet 选择与基本级别类别相关的图像块，从而处理对象级别的注意。选定的图像块将另一个 CNN 训练成一个称为 DomainNet 的 Domain 分类器。根据经验，本论文观察了 DomainNet 内部隐藏表示中的聚类模式。神经元群对辨别部位表现出高度的敏感性。因此，本论文选择相应的滤波器作为部分检测器来实现部分级注意。

1.3 网络各模块的组成及功能

1.3.1 Object-Level Attention Model

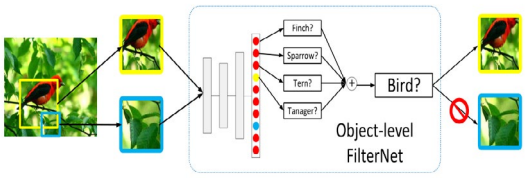

图 10：对象级别的自上而下注意

引入了一个对象级 FilterNet，以决定是否将自下而上方法提出的图像块进行到下一步。FilterNet 只关心图像块是否与基本级别类别相关，目标是过滤掉背景图像块。

这一步的目标是去除与对象无关的噪声块。本论文通过将在 1K 类 ILSVR2012 数据集上训练的 CNN 转换为对象级 FilterNet 来实现这一点。本论文总结属于细粒度类别的父类（例如吉娃娃的父类是狗）的所有 softmax 神经元的激活情况，作为选择信心分数，然后在分数上设置阈值，以决定是否应选择给定的图像块。如图 10 所示。

FilterNet 选择的图像块用于在适当图像处理后从头开始训练新的 CNN。本论文将第二个 CNN 称为 DomainNet，因为它提取与属于特定领域（例如狗、猫、鸟）的类别相关的特征，从一张图像中可以获得许多这样的图像块，其净效果是数据增强。与其他数据扩充（如随机裁剪）不同，本论文对图像块的相关性有更高的置信度。数据的规模也推动了更大网络的训练，使其能够构建更多功能。

使用对象级注意的图像块选择可以应用于测试阶段。为了获得图像的预测标签，本论文向 DomainNet 提供 FilterNet 选择的图像块以进行前馈。然后计算所有补丁的 softmax 输出的平均分类分布，最后可以得到平均 softmax 分布的预测。

内在工作原理：自下而上的过程具有很高的召回率，但精确度很低。如果对象相对较小，大多数图像块都是背景，根本无助于对对象进行分类。这给 pipeline 的 where 部分带来了问题。通过 FilterNet 的构建，本论文可以有效的过滤掉本论文不需要的图像块，FilterNet 会把本论文需要的符合的大类（鸟，狗，猫）的图像块选择出来，而会把无用的如背景图像块去掉，这样我们保留了多尺度，多视角的优点，从而有效的降低了不必要的噪声。

同时 DomainNet 的设置有两个好处。首先，DomainNet 本身就是一个很好的细粒度分类器。其次，它的内部功能现在允许本论文构建部分检测器，这个将在下一步进行解释。

1.3.2 Part-Level Attention Model

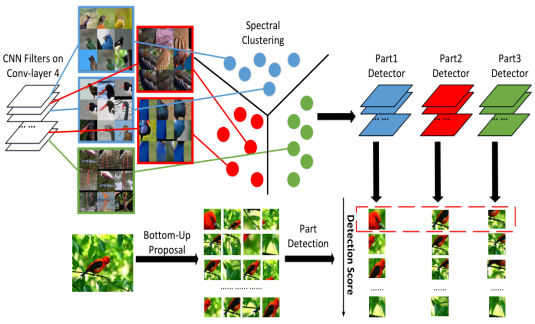

图 11：部分级自上而下注意

DomainNet 中的过滤器对特定的对象部分显示出更多关注，可以根据其关注的部分在过滤器之间找到聚类模式。本论文使用谱聚类来分组，然后使用组中的滤波器作为部分检测器。在这个图中，中间层 CNN 过滤器可以用作鸟类的头部探测器、身体探测器和腿部探测器。

Part-RCNN[1]和 DPD[12]的研究强烈表明，某些有区别的局部特征（例如头部和身体）对细粒度分类至关重要。本论文从 DomainNet 的隐藏层有一些聚类的模式这一事实中受到启发。例如，不同神经元对控制鸟的姿势有种种影响。

图 11 显示了该步骤的执行情况。本质上，本论文对相似矩阵 S 进行谱聚类，将中间层中的滤波器划分为 k 组，其中$S(i，j)$表示域网中两个中间层滤波器F_i和F_j的权重的余弦相似性。在本论文的实验中，本论文的 CNN 网络基本上与 AlexNet 相同[13]，本论文从第四个卷积层选取神经元，k 设置为 3。每个簇都充当部分检测器。

当使用聚类过滤器从区域建议中检测 part 时，步骤是：1）将建议的贴片扭曲为 conv4 过滤器输入图像上的感受野大小。2）将补丁转发给 conv4，为每个过滤器生成激活分数。3）将一个聚类中过滤器的得分相加，得到聚类得分。4）为每个簇选择簇分数最高的补丁作为部分补丁。

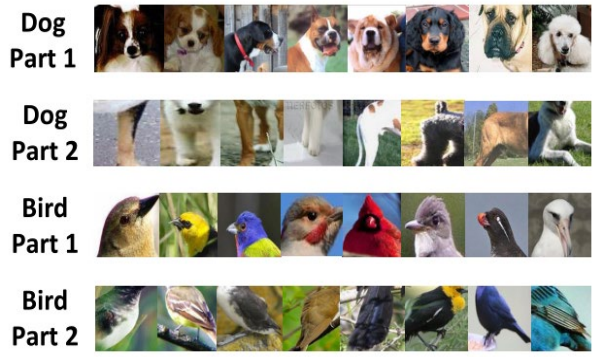

图 12：显示了狗和鸟类的一些检测结果

很明显，鸟 DomainNet 中的一组过滤器特别关注鸟头，另一组关注鸟的身体。同样，对于狗的 DomainNet，一组过滤器关注狗头，一组关注狗腿。然后，部分检测器选择的图像块被改回 DomainNet 的输入大小，以生成激活结果。本论文将不同部分的激活结果与原始图像连接起来，然后训练一个支持向量机作为基于部分的分类器。

内在工作原理：利用了 CNN 的性质，可以不同的层抽取不同的局部特征，同时通过聚类的方法避免了强监督标号的繁琐。

1.4 网络处理数据的整体流程

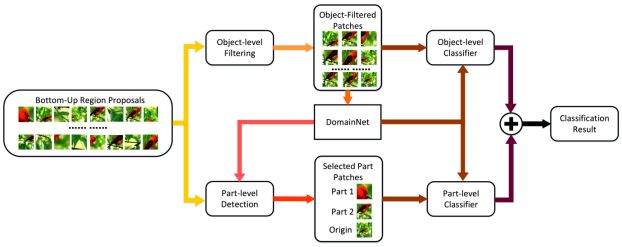

图 13：本论文的方法的完整分类 pipeline

箭头越暗，执行该操作的时间越晚。自上而下的两个层次的注意力应用于自下而上的建议。一种是进行对象级过滤，以选择与鸟类相关的斑块，并将其输入分类器。另一个进行部分级检测，以检测不同部分进行分类。DomainNet 可以为部分级方法提供部分检测器 s，也可以为这两个级别的分类器提供特征提取器。两个分类器的预测结果在后期合并，以结合两个层次的优点。

DomainNet 分类器和基于部分的分类器都是细粒度分类器。然而，它们的功能和强度不同，主要是因为它们允许不同性质的图像块。最后，本论文合并了两种水平注意方法的预测结果，以利用两个分类器的优势。图 13 显示了合并两级注意分类器结果的完整 pipeline。

此后，一些算法将递归神经网络（RNN）、LSTM[9]或 FCN[7]作为注意机制引入弱监督方法中，并采用了复杂的训练过程，如使用附加部分检测器的多阶段训练，以改进定位。最近的一些算法结合了一些想法，例如，用神经树来改进注意机制，或者使用图传播、挖掘补充信息等来更有效地检测关键部分。以下是一些最新进展。

2. 注意力卷积二叉神经树方法

2.1 背景

受[11]的启发，Ji 等人[14]提出了一种基于注意卷积二叉神经树结构的弱监督 FGVC 新方法 ACNet。它结合了沿树结构边缘的卷积运算，并使用每个节点中的路由函数来确定树内根到叶的计算路径。这种设计的结构使得本论文的方法继承了深度卷积模型的表示学习能力，以及从粗到细的分层特征学习过程。同时，文章使用 attention transformer 来强化树网络，以捕获区分性特征。采用负对数似然损失，通过带反向传播的随机梯度下降，以端到端的方式对整个网络进行训练。

2.2 贡献

（1） 提出了一种新的 FGVC 注意卷积二叉神经树结构。

（2） 引入了 attention transformer，以便于在树网络中从粗到细的分层特征学习。‘

2.3 网络各模块的组成及功能

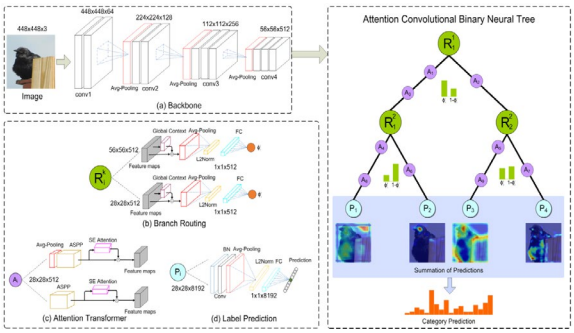

图 14：ACNet 模型概述，由（a）主干网络模块，（b）分支路由模块，（c）attention transformer 模块和（d）标签预测模块组成。

2.3.1 主干网络模块

在论文中使用了 VGG-16[10]（保留 conv1_1 到 conv4_3 的层）和 ResNet-50[15]（保留 res_1 到 res_4 的层）网络作为本工作的主干网络。

内在工作原理：因为很多特征都是在图片里比较小的一片区域（如猫的头，猫的腿等），所以对卷积和池化层使用了相对更小的感受野来可以更好的提取特征。

2.3.2 分支路由模块

如上所述，我们使用分支路由模块来确定样本将发送到哪个子树（即左或右子树）。具体来说，如图 14(b)所示，第 k 层的第 i 个路由模块 R_i^k (•)使用一个内核大小为 1×1 的卷积层，然后是一个全局上下文模块[16]。全局上下文块是简化的非局部(NL)块[17]和压缩激励(SE)块[18]的改进，在上下文建模和融合步骤上与简化版 NL 块共享同样的实现，并与 SE 块共享转换步骤。之后，我们使用全局平均池化、按元素平方根和 L2 归一化，以及带有 sigmoid 激活函数的完全连接层，在[0,1]之间生成一个标量值，指示样本被发送到左或右分支的概率。

内在工作原理：通过使用全局上下文模块，上下文信息被整合，以更好地描述对象。

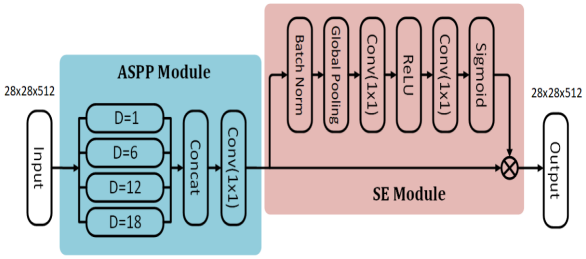

图 15：attention transformer 模块的架构

2.3.3 Attention transformer

内在工作原理：ASPP 模块提供了不同的特征图，每个特征图都有不同的尺度/感受野和注意模块。然后，通过四个不同扩张率的平行扩张卷积，即 1、6、12、18，生成多尺度特征图。在并行扩展的卷积层之后，级联的特征映射由一个核为 1×1、步长为 1 的卷积层进行融合。这样可以获得更大的感受野。通过加入注意力模块这种方式可以引导网络关注有意义的特征，以获得准确的结果。

2.3.4 标签预测模块

内在工作原理：标签预测模块用来预测对象 x_j的从属类别。这为标签预测提供了数学基础。

2.4 网络处理数据的整体流程

ACNet 结合了树边缘的注意卷积，在每个节点使用路由函数来定义从根节点到叶节点的计算路径，类似于神经网络。这种结构使得该算法具有类似于神经网络的表示能力。由于不同的分支集中在不同的局部区域，ACNet 从粗到细学习层次特征。最后，结合所有叶片的预测来预测标签，同时边缘节点使用 attention transformer 来加强网络对关键特征的捕捉，以进行分类。

总结：总的来说，大多数基于部分的方法利用检测或分割技术、深度过滤器和注意力机制来定位细粒度对象的区分性语义部分，然后从这些部分学习详细的表示。在这三种方法中，基于视觉注意的方法一直是研究的热点。

3 Bilinear CNN

3.1 背景

由于图像不同部分之间存在复杂的关系，基于部分的方法难以对图像的特定特征进行建模。为了处理这种复杂的交互，基于高阶图像特征的方法被广泛采用。它主要通过设计高阶图像特征的表示方法代替传统的全局平均池化来进行 CNN 特征的高阶混合，最终可以使分类中使用的图像特征更好地表示细粒度图像的细化差异，进一步丰富图像特征所表示的内容。而 Bilinear CNN [1] 则是其中的一个经典模型。

3.2 网络各模块的组成及功能

对于用于细粒度图像分类的模型B，可以表示为由四个部分构成：$B = (f_A, f_B, P, C)$，其中f_A和f_B为特征函数，P是池化函数，而C是分类函数。特征函数可以实现由$L \times I$到$R^{c \times D}$的维度转换。其中I表示图片而L为包含图片对应着位置和比例的信息。$c \times D$即为输出特征的尺寸。特征输出在每个位置通过矩阵外积进行组合，即在每个对应位置的f_A和f_B通过以下表达式进行组合，输出结果（双线性特征）：$(l, I, f_A, f_B) = f_A(l, I)^T f_B(l, I)$。

为了获得图像描述子，池化函数P聚合图像中所有位置的双线性特征。可以通过简单的对特征进行求和，也可以使用最大池化的方法。这两种方法都忽略了要素的位置，因此是无序的。池化输出结果再输入到分类函数C中，最终得出结果。在这种处理方式下，可以考虑到两个特征输出f_A和f_B的所有成对交互作用（类似于二次核展开）来实现相互制约。简单的结构可以如图 16 所示：

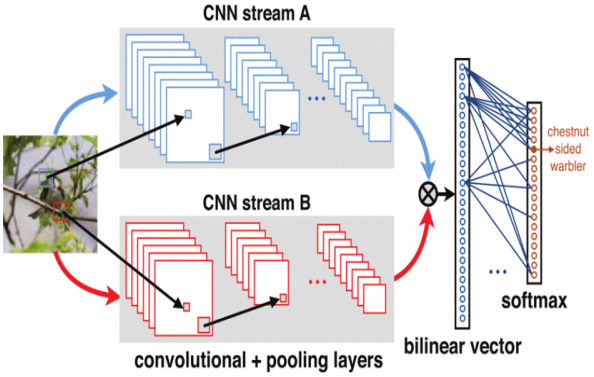

图 16 用于图像分类的 Bilinear CNN 模型。在测试时，图像通过两个 CNN，A 和 B，它们的输出在图像的每个位置使用外部积相乘，并池化以获得双线性矢量。这将通过分类层进行传递以获取预测。

Bilinear CNN 模型可以通过外积捕获特征通道之间的成对相关性，并且可以对部分特征交互进行建模。这使得细粒度分类的精度可以进一步提高。同时由于分别使用两个单独的 CNN，对特定领域微调的梯度计算大幅简化。

3.2.1 Bilinear CNN Models

在确定了双线性模型的框架后，需要对模型中每一部分的具体结构进行定义。

在特征函数部分，采用的是一个由卷积层和池化层组成的 CNN。在论文作者的实验中，他们使用在 ImageNet 数据集上预训练的 CNN，并且将该 CNN 在卷积层上截断。

这种方法的优点在于，当数据集中特定领域的数据稀缺时，通过预训练，可以从额外的训练数据中受益；另一方面，仅使用卷积层时，生成的 CNN 可以在单个前向传播步骤中处理任意大小的图像，并产生由图像和特征通道中的位置索引的输出。

在接下来的部分，作者采用 sum-pooling 对图像特征进行聚合。得到双线性向量后，通过带符号的平方根步骤$(\mathbf{y} = sign(\mathbf{x})\sqrt{|x|})$，随后进行

标准化($\mathbf{z} \leftarrow \mathbf{y}/\| \mathbf{y} \|_2$)。对于分类函数$C$可以使逻辑回归或线性 SVM。在非线性数据上，这部分可以用多层神经网络代替。

由于整体构架是有向无环图，在进行端到端的训练时，可以通过反向传播分类损失的梯度来计算梯度。假设两个特征网络输出为两个矩阵A和B，那么池化后的特征输出为$\mathbf{x} = \mathbf{A}^T\mathbf{B}$，$(dl/d\mathbf{x})$为损失函数对其梯度，根据链式法则，有下式：

$$\frac{d\ell}{dA} = B\left(\frac{d\ell}{d\mathbf{x}}\right)^T, \quad \frac{d\ell}{dB} = A\left(\frac{dl}{d\mathbf{x}}\right).$$

这样正好利用了双线性模型对池化层梯度的简化。

3.2.2 Relation to Orderless Texture Descriptors

各种无序纹理描述子可以用来构建双线性模型。他们通常从图像中密集地提取局部特征，例如 SHFT，并将它们传递给非线性编码器η。一种流行的编码器是高斯混合模型（GMM）。它基于特征的的 GMM 后验将其分配给k个中心。这些编码的描述符在图像中进行汇总时，可以得到 Bag-of-Visual-Words (BoVW) 模型，用于构建双线性模型。

除此之外，还有局部聚合描述符向量 (VLAD) 描述子、Fisher 向量 (FV)可以用于构建非线性模型。它们可以通过微调来适应各种任务，进而提高分类整体性能。

在 VLAD 和 FV 中，非线性编码器η可以看作一个部位检测器。它们可以同时定位部位并使用编码$\eta(x)$和特征x的联合统计数据来描述它们的特征，这可以解释它们在细粒度识别任务中的有效性。

3.3 网络处理数据的整体流程

在实验中，作者选取了两个用于提取双线性模型中的特征的 CNN—— M-Net 和由 16 个卷积层和池化层组成的较深的网络 D-Net。通过对这两个网络的不同组合，设计了三个不同的双线性 CNN 模型——(i) 用 B-CNN [M, M] 表示的以两个 M-net 初始化，(ii) 用 B-CNN [D, M] 表示的以 D-Net 和 M-Net 初始化以及 (iii) 用 B-CNN [D, D] 表示的以两个 D-net 进行初始化的模型。

图像输入时，首先将其大小调整为 448 × 448，再输入特征提取模型，即之前定义的双线性 CNN。D-Net 产生稍大的输出 28×28，而 M-Net 产生 27×27 的输出。通过忽略行和列来简单地对 D-Net 的输出进行下采样。特征输出在每个位置通过矩阵外积进行组合，随后进入池化层，产生的池化双线性特征的大小 512×512。最后的 k-way softmax 层用于对特征进行微调。

在实验中，作者采用两步训练过程，首先使用逻辑回归训练最后一层，这是一个凸优化问题；然后使用反向传播以相对较小的学习率 (η=0.001)对整个模型 epoch 进行微调（大约 45 - 100 取决于数据集和模型）。 在整个数据集中，作者证明用于微调的超参数相当一致。

一旦完成微调，训练集和验证集就会结合起来，通过设置学习超参数来训练提取特征上的一对多线性 $SVM\ Csvm = 1$。训练好的分类器通过缩放权重向量进行校准，使得正样本和负样本的中值分数分别为 +1 和 -1。对于每个数据集，通过翻转图像使训练数据加倍。在测试时，对图像及其翻转副本的预测进行平均，并将其分类为得分最高的类。与线性 SVM 相比，直接使用 softmax 预测会导致精度略有下降。最终的性能测量为所有数据集的正确图像预测的百分比。

4 MC-Loss

4.1 背景

在前述的方法中，都是通过使网络更改以实现部分定位或判别特征学习。而在 Chang 等人的论文[19]，他们设计了 MC-Loss 损失函数，利用单个损失函数通过端到端的训练实现弱监督下的细粒度分类。

在当今的计算机视觉领域，设计特定任务的损失函数，以增强具有强判别信息的 CNN 变得愈发流行。直观地说，当类内紧凑性和类间可分离性同时最大化时，提取的特征最具区分性，即 Fisher 准则。Liu 等人引入了 A-softmax 损失[20]来学习在深度超球嵌入流形上进行图像分类的角度判别特征。Wang 等人根据 Fisher 准则的思想，提出大余量余弦损失（LMCL）[21]来学习用于图像识别的高度判别深度特征。

尽管上述所有损失函数都可以在一定程度上获得判别特征，但它们并没有明确鼓励网络关注局部判别区域。相比之下，MC-Loss 函数强制网络识别多个判别区域，这减轻了对不同的复杂网络设计的需求，使框架易于实施和易于解释。

4.2 网络各模块的组成及功能

MC-loss 由两个部分组成。设输入的特征通道为F，F分成两流进入 。其中，交叉熵流将F作为具有传统 CE 损失的全连接 (FC) 层的输入，得到交叉熵损失$L_{CE}(F)$。在这里，交叉熵损失鼓励网络提取主要集中在**全局判别区域的信息特征**。另一方面，MC-Loss 流监督网络以聚焦**不同的局部判别区域**。然后将 MC-Loss 加以权重μ添加到 CE 损失中。因此，整个网络的总损失函数可以定义为

$$Loss(\mathbf{F}) = L_{CE}(\mathbf{F}) + \mu \times L_{MC}(\mathbf{F}).$$

其中，MC-Loss 是一个判别性分量$L_{dis}(\mathbf{F})$和另一个多样性分量$L_{div}(\mathbf{F})$的加权和。可以将 MC-Loss 定义为

$$L_{MC}(\mathbf{F}) = L_{dis}(\mathbf{F}) - \lambda \times L_{div}(\mathbf{F}).$$

4.2.1 The Discriminality Component

在模型框架中，每个类都由一定数量的分组特征通道表示。判别性分量强制特征通道是类对齐的，并且对应于特定类的每个特征通道应该具有足够的判别力。辨别力分量$L_{dis}(\mathbf{F})$可以表示为

$$L_{dis}(\mathbf{F}) = L_{CE}\left(\mathbf{y}, \underbrace{\frac{\left[e^{g(\mathbf{F}_0)}, e^{g(\mathbf{F}_1)}, \cdots, e^{g(\mathbf{F}_{c-1})}\right]^{\mathrm{T}}}{\sum_{i=0}^{c-1} e^{g(\mathbf{F}_i)}}}_{\text{Softmax}}\right),$$

其中 GAP、CCMP 和 CWA 分别是全局平均池化、跨通道最大池化和通道关注的简写。$L_{CE}(\cdot,\cdot)$是真实类标签 y 和 GAP 输出之间的交叉熵损失。$[\frac{\xi}{2}]$和操作$diag(\cdot)$将向量放在对角矩阵的主对角线上。图 17 中的左侧模块显示了判别性组件的流程图。

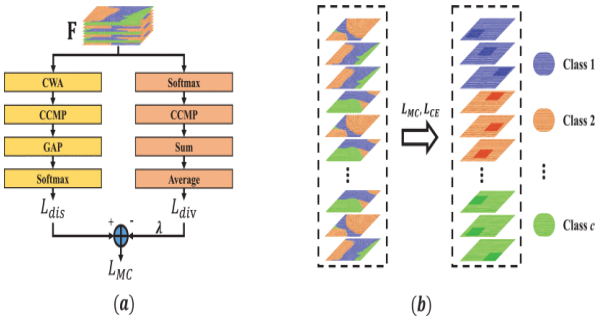

图 17 (a) MC-loss 概述；(b) 应用 MC-Loss 之前（左）和之后（右）的特征图比较

在 CWA 部分，作者提出了通道关注操作，以强制网络在对应于特定类的所有ξ信道中平等地捕获判别信息，确保每个特征信道包含足够的判别信息。

在 CCMP 部分，产生了与一个特定类一致的大小为 WH 的一维向量。CCMP 可以保留局部区域的注意力，并且被证明有利于细粒度分类。

4.2.2 The Diversity Component

多样性分量是特征通道的近似距离测量，用于计算所有通道的总相似度。图 17 右侧块所示的多样性分量通过训练驱动F_i中的特征通道变得彼此不同。换句话说，使一个类的不同特征通道关注图像的不同区域，而不是所有通道都关注最具辨别力的区域。因此，它通过使每个组的特征通道多样化来减少冗余信息，并有助于发现关于图像中每个类的不同判别区域。此操作可以解释为跨通道去相关，以便从图像的不同显着区域捕获细节。在 softmax 之后，通过引入 CCMP 和空间维度求和来直接对卷积滤波器进行监督以测量相交程度。多样性特定损失分量$L_{div}(\mathbf{F})$可以定义为

$$L_{div}(\mathbf{F}) = \frac{1}{c}\sum_{i=0}^{c-1} h\left(\mathbf{F}_i\right),$$

其中，$h(\cdot)$定义为

$$h(\mathbf{F}_i) = \sum_{k=1}^{WH} \underbrace{\max_{j=1,2,\cdots,\xi}}_{\text{CCMP}} \underbrace{\left[\frac{e^{\mathbf{F}_{i,j,k}}}{\sum_{k'=1}^{WH} e^{\mathbf{F}_{i,j,k'}}}\right]}_{\text{Softmax}}$$

函数 softmax 是对空间维度的归一化，这里的 CCMP 与它在判别性组件中的作用相同。

4.3 网络处理数据的整体流程

在实验中，结合 MC-loss 的网络如下图 18 所示。给定输入图像，它首先通过将图像输入基础网络来提取特征图。该基础网络可以是 VGG16 或者 ResNet18 等。图像通过基础网络提取的特征图可以表示为：$\mathfrak{f} \in \mathbf{R}^{N\times W\times H}$，其中 H 是高，W 是宽，而 N 为通道数。在 MC-loss 部分，需要设置 N 的值等于$c \times \xi$，其中c是数据集中类的数目而ξ是用于表示每个类的特征通道数。$\mathfrak{f}$的第 n 个矢量特征通道表示为$\mathfrak{f}_n \in \mathbf{R}^{WH}, n = 1,2,\ldots,N$。在这里通道矩阵$\mathfrak{f}$的维度由$W \times H$被重塑$WH$。对应于第 i 类的分组特征通道由$F_i \in \mathbf{R}^{\xi\times WH}, i = 0,1,2\ldots,c-1$表示。在数学上，它可以被表示为：

$$\mathbf{F}_i = \{\mathcal{F}_{i\times\xi+1}, \mathcal{F}_{i\times\xi+2}, \cdots, \mathcal{F}_{i\times\xi+\xi}\}.$$

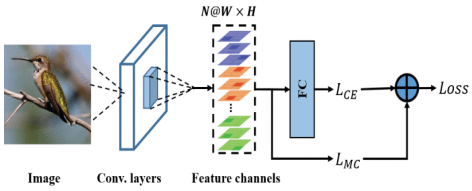

图 18　使用 MC-Loss 的典型细粒度分类网络的框架

随后，$F_n = \{F_0, F_1, F_2, \ldots, F_{c-1}\}$ 进入 MC-loss 的两个流，一个通过全连接层（FC）得到交叉熵子损失$L_{CE}(F)$，而另一个通过 MC-Loss 流监督网络得到子损失$L_{MC}(F)$，具体细节见上一部分。总损失则为这两个子损失的加权和。得到总损失后，通过反向传播等算法不断训练调整参数，最终得到所需的细粒度分类模型。

5 Approaches with External Information（额外信息）

与仅使用图像类别标签的传统弱监督 FGVC 方法不同，最近通过使用网络图像、知识图或文本等额外的信息来协助 FGVC 任务的方式被提出。

考虑到由于精细图像的下级类别标签经常需要专家知识而导致训练图像的标注困难和不精确，针对这一问题，Sun 等人提出了一种对抗判别损失的方法，为 FGVC 训练了一个端到端多任务学习框架，克服了来自标准数据集的标准数据与网络数据之间的差距。随后，Zhang 等设计了一种新的方法，在模型训练时，从网络图像中去除不相关的样本，以减少网络数据噪声带来的有害影响。在三种常用的细粒度数据集(训练集为网络图像，测试数据为 CUB200-2011[22]、FGVC-aircraft[23]和 Cars[24])上进行的实验表明，该方法优于 SOTA webly 监督方法。

5.1 网络图像信息：用于细粒度视觉分类的具有软更新丢弃训练的 Web 监督网络

Schroff, criisi 和 Zisserman 在 2011 年的实验所指出的，谷歌图像搜索引擎的 18 个类别的前 1000 张图片的平均精度仅为 32%。

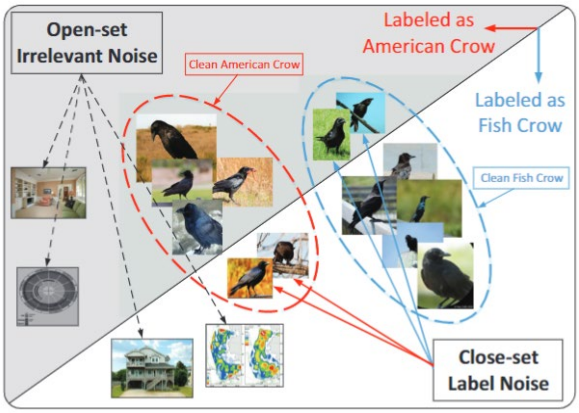

图 19 对有噪声的网页图像进行细粒度分类可以分为两类:闭集噪声和开集噪声

闭集中的噪声图像在数据集中有其真实的标签，但有的图像会被错误标记，如图 19 中的部分“Fish Crow”被错误地标记为“American Crow”。开放集由不相关的噪声图像组成。由于深度神经网络拟合噪声数据的能力也很强，直接用这些嘈杂的网络图像训练细粒度神经网络模型将导致性能较差。

为了克服图像噪声问题，学者们进行了大量的研究，主要可分为两类:损失校正方法和样本选择方法。

1. 现有的损失校正方法试图估计标签噪声转移，如(Goldberger 和 Ben-Reuven 2016)无法处理开放集无关噪声的网络图像。这是因为无关噪声样本的真实标签在训练标签集之外。
2. 样本选择方法不考虑无关噪声样本的真实标签，只是从噪声样本中选择干净的实例。由于这一优势，样本选择方法在开放集场景中更加实用。

5.2 主要工作

5.2.1 软最大化概率的交叉熵

软标签包含比一个热点标签更多的信息，特别是在细粒度分类任务，其中子类别有明显的相似性。因此，本方法我们利用相邻 epoch 之间的软最大化概率的交叉熵代替损失函数来寻找噪声样本(我们在下文中将其命名为概率交叉熵)。该方法可以很好地利用编码在软标签中的信息，并能够度量网络的预测变化。

此外，开集无关噪声样本比纯净样本更难拟合。它们的预测是不稳定的，在训练过程中变化很快，导致概率交叉熵大。因此，通过计算概率交叉熵，可以将不相关的噪声样本从有用的训练集中区分出来并去掉。这样，网络可以缓解开集噪声的有害影响，达到更好的性能。

5.2.2 网络训练方法

通过计算概率交叉熵从训练集中选取有用的样本，交叉熵较大的图像被认为是不相关的噪声图像，将排除在外。

Algorithm 1: Softly Update-Drop Training

Input: Initialized network f, training set $\mathcal{D}$, maximum drop rate τ, epoch T_k and $T_{\max}$.
for $T = 1, 2, ..., T_{\max}$ **do**
  for each instance x_i in training set $\mathcal{D}$ **do**
    if $T > 2$ **then**
      Compute $C(x_i)^T$ according to Eq. (2)
      Obtain $\widehat{\mathcal{D}}^T$ according to Eq. (4)
    else
      $\widehat{\mathcal{D}}^T = \mathcal{D}$
    end
    Compute $p(x_i)^T$ according to Eq. (1)
  end
  Update f and $r(T)$
end
Output: Updated network f

概率交叉熵计算方法如下：

$$p_j(x_i)^T = \frac{\exp(f_j(x_i))}{\sum_{s=1}^{M}\exp(f_s(x_i))}. \tag{1}$$

$$C(x_i)^T = -\sum_{j=1}^{M} p_j(x_i)^{T-1}\log p_j(x_i)^{T-2}. \tag{2}$$

其中 T 为 epoch 编号，$f_j(x_i)$为第 i 个样本在第 j 个网络的输出结果，$C(x_i)$为第 i 个样本的概率交叉熵。

每轮选择的样本数量将动态更新，Drop Rate 由下式决定：

$$r(T) = \tau \cdot \min\{\frac{T}{T_k}, 1\}, \tag{3}$$

最后得到的新训练集$\widehat{\mathcal{D}}^T$由下式决定：

$$\widehat{\mathcal{D}}^T = \underset{\mathcal{D}':|\mathcal{D}'|\geq(1-r(T))|\mathcal{D}|}{\arg\min} \sum_{x\in\mathcal{D}'} C(x)^T. \tag{4}$$

5.2.3 全局采样

按批次进行小批次采样可能导致不同批次的噪声率不平衡的问题。在每个 epoch 中的下降率 r(T)是固定的情况下，纯净样本可能必须在一些小型批次中被丢弃，而噪声样本则用于其他小型批次的训练。

本方法从整个训练集选择样本，通过使选择结果更加稳定，可以达到更好的性能。

5.2.4 动态更新 Drop Rate

深度神经网络具有在早期训练阶段使用其损失值过滤出噪声实例的能力。然后，随着时代的增加，深度神经网络最终会对噪声样本进行过拟合。为了利用这个属性，本方法动态地增加 Drop Rate $r(T)$，在噪声被记忆之前逐渐减少噪声图像。

弱监督 FGVC 算法是细粒度图像研究的发展趋势，但与一般图像识别算法相比，就在最具代表性的 FGVC 数据集 CUB200-2011[21]上获得的分类精度而言，监管薄弱的 FGVC 方法仍有很大的改进空间。

5.3 用于细粒度图像识别的知识嵌入表示学习

人类不仅根据物体的外观，而且根据日常生活或职业中获得的知识来进行物体识别。，我们可以先回忆这个知识，注意相应的部分，看它是否具有这些属性，然后进行推理。

这项工作研究了如何将丰富的专业知识与深度神经网络架构相结合，并提出了一个知识嵌入表示学习(KERL)框架来处理细粒度图像识别问题。在知识的引导下，这个框架可以学习属性感知的特征映射，通过有意义且可解释的配置，突出显示的区域与图中相关属性有良好的相关性，这也可以解释性能的提高。

用于视觉推理的学习知识表示越来越受到关注，最近，一系列的努力致力于使神经网络处理图结构数据。例如[Niepert et al.， 2016]根据图的边缘对图 20 中的节点进行排序，形成规则序列，直接将节点序列送入标准 CNN 进行特征学习。

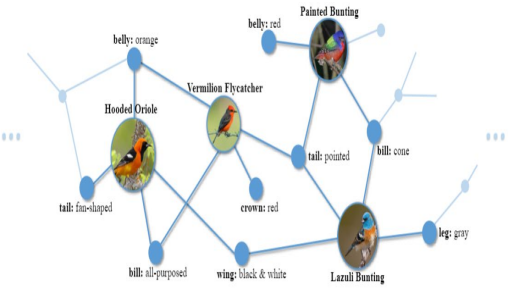

图 20 一个在加州理工大学鸟类数据集上建模类别-属性相关性的知识图示例。

GGNN [Li 等人，2015]是一种用于图结构数据的完全可微递归神经网络体系结构，它递归地将节点消息传播给它的邻居，以学习节点级特征或图级表示。一些工作已经开发了一系列图神经网络变体，并成功地应用于各种任务。作者开发了一种新的门控机制与 GGNN 结合，将知识表示嵌入到图像特征学习中，以增强特征表示。

5.3.1 KERL 网络框架

5.3.1.1 GGNN 介绍

GGNN 是一种递归神经网络结构，可以通过迭代更新节点特征来学习任意图结构数据的特征。输入图 G=（V,A）,其中 V 是节点集合，A 是邻接矩阵。基本的循环过程表述如下：

$$
\begin{aligned}
\mathbf{h}_v^0 &= \mathbf{x}_v \\
\mathbf{a}_v^t &= \mathbf{A}_v^\top [\mathbf{h}_1^{t-1} \ldots \mathbf{h}_{|V|}^{t-1}]^\top + \mathbf{b} \\
\mathbf{z}_v^t &= \sigma(\mathbf{W}^z \mathbf{a}_v^t + \mathbf{U}^z \mathbf{h}_v^{t-1}) \\
\mathbf{r}_v^t &= \sigma(\mathbf{W}^r \mathbf{a}_v^t + \mathbf{U}^r \mathbf{h}_v^{t-1}) \\
\widetilde{\mathbf{h}_v^t} &= \tanh\left(\mathbf{W} \mathbf{a}_v^t + \mathbf{U}(\mathbf{r}_v^t \odot \mathbf{h}_v^{t-1})\right) \\
\mathbf{h}_v^t &= (1 - \mathbf{z}_v^t) \odot \mathbf{h}_v^{t-1} + \mathbf{z}_v^t \odot \widetilde{\mathbf{h}_v^t}
\end{aligned}
\tag{1}
$$

其中h_v^t是第 v 个节点在第 t 个时间步长的隐藏状态，h_v^0由输入特征向量x_v初始化。σ和$tanh$分别表示 sigmoid 函数和双曲正切函数，通过重复循环，可以得到最终的隐藏状态$(h_1^T, h_2^T, \ldots, h_{|V|^T})$。上述过程可以简化表示如下：

$$h_v^t = GGNN(h_1^t - 1, \ldots, h_{|V|}^{t-1}; A_v)$$

5.3.1.2 知识图结构

知识图指的是包含类别标签和部分级别属性的可视化概念存储库的组织，节点表示可视化概念，边表示它们的相关性。

1. 视觉概念(Visual concepts) 视觉概念指的是类别标签或属性。给定一个包含 C 个分类和 A 个属性的数据集，对应构造的知识图图有一个包含 C+A 个元素的节点集 V。
2. 相关性(Correlation) 类别标签和属性之间的相关性表明这个类别是否拥有相应的属性。因此，可以使用一个分值来表示该实例具有某属性的可能性，然后将属于特定类别的所有实例的属性/对象实例对的分值相加，得到一个分值，表示该类别拥有该属性的置信度。所有的分值都被线性归一化到[0,1]，从而得到一个 C × A 矩阵。注意，两个对象类别节点之间或两个属性节点之间不存在连接;因此完全邻接矩阵可以表示为：

$$\mathbf{A}_c = \begin{bmatrix} \mathbf{0}_{C\times C} & \mathbf{S} \\ \mathbf{0}_{A\times C} & \mathbf{0}_{A\times A} \end{bmatrix},$$

这样就构造了一个知识图$G = (V, A_C)$

5.3.1.3 知识表示学习（Knowledge Representation Learning）

使用 GGNN 通过图传播节点消息，并计算每个节点的特征向量。然后将所有特征向量连接起来，以生成知识图的最终表示。

用表该类别的置信度的分数s_i来初始化引用类别标签 i 的节点。得分向量$s = (s_0, s_1, \ldots, S_{C-1})$由预先培训的分类器进行评估,每个结点的输入要素可以表示为:

$$\mathbf{x}_v = \begin{cases} [s_i, \mathbf{0}_{n-1}] & \text{if node } v \text{ refers to category } i \\ [\mathbf{0}_n] & \text{if node } v \text{ refers to an attribute} \end{cases},$$

通过 GGNN 的计算过程，使用A_c来将消息从某个节点传播到其邻居，而我们使用矩阵A_c^T来进行反向消息传播。因此，邻接矩阵$A = [A_c \ A_c^T]$。

在每次迭代中，每个节点的隐藏状态由其历史状态和邻居发送的消息确定。通过这种方式，每个节点可以聚合来自其邻居的信息，并同时将其消息传输到其邻居。这一过程如图 21 所示:

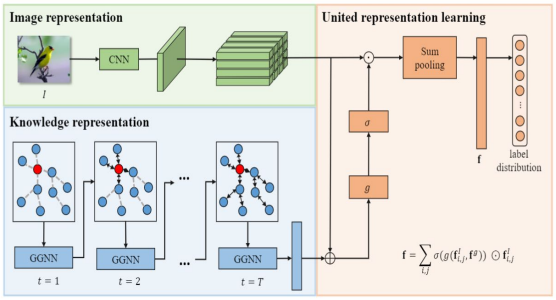

图 21

经过 T 次迭代，每个节点的消息都已经在图中传播，可以得到图中所有节点的最终隐藏状态，即$(h_1^T, h_2^T, \ldots, h_{|V|^T})$。最后节点级特征表达如下：

$$\mathbf{o}_v = o(\mathbf{h}_v^T, \mathbf{x}_v), v = 1, 2, \ldots, |\mathbf{V}|,$$

其中o是由全连接层实现的输出网络。最后，将这些特征连接起来以产生最终的知识表示f^g。

5.3.1.4 图像特征提取

紧致双线性模型 l Gao etal., 2016] 在细粒度图像分类方面效果很好。直接将该模型应用于图像特征提取。

将该表示嵌入到图像特征学习中，以学习与该属性相对应的特征。引入了一种门控机制，在知识的指导下，有选择地允许信息性特征通过，而抑制非信息性特征，这可以表示为

$$\mathbf{f} = \sum_{i,j} \sigma\left(g\left(\mathbf{f}_{i,j}^I, \mathbf{f}^g\right)\right) \odot \mathbf{f}_{i,j}^I,$$

其中$f_{i,\ j}^I$是位置(i,j)处的特征向量。$\sigma\left(g\left(f_{i,\ j}^I, f^g\right)\right)$充当确定哪个位置更重要的选通机制。g 是一个神经网络，它以$\left(f_{i,\ j}^I, f^g\right)$的级联为输入，并输出一个 c 维实数向量。然后将特征向量 f 输入到简单的全连通层，以计算给定图像的分数向量 s。

总结及展望

以上便是我们对这两种深度学习方法的研究。我们可以看到，现在已经有许多优化方法来获得更好的结果。现在已经有很多领域需要细粒度视觉分类：如图片搜索鞋子品牌、人脸部位识别、商店商品识别……未来，细粒度模型也能和现实结合，进行迁移学习从而得到更好的算法。

参考文献：

[1] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Partbased r-cnns for fine-grained category detection. In ECCV. 2014.

[2] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition(pp. 580-587)

[3] Uijlings,J. R., Van De Sande， K. E., Gevers, T., & Smolders, A. W.(2013) .Selective search for object recognition. International journal of computer vision, 104(2), 154-171.

[4] Branson, S., Van Horn, G., Belongie, S., & Perona, P. (2014). Bird species categorization using pose normalized deep convolutional nets.arXiv preprint arXiv:1406.2952.

[5] Branson, S., Beijbom, O., & Belongie, S. (2013). Efficient large-scale structured learning. In Proceedings of the IEEEConference onComputer Vision and Pattern Recognition (pp. 1806-1813).

[6] Wei, X. S., Xie, C. W., Wu, J., & Shen, C. (2018). Mask-cnn: Localizing parts and selecting descriptors for fine-grained bird species categorization. Pattern Recognition, 76, 704-714.

[7] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[8] Lam, M., Mahasseni, B., & Todorovic, S. (2017). Fine-grained recognition as hsnetsearch for informative image parts. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2520-2529).

[9] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory.Neural computation, 9(8), 1735-1780.

[10]Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014. 3

[11]The Application of Two-level Attention Models in Deep Convolutional Neural Network for Fine-grained Image Classification

[12] N. Zhang, R. Farrell, F. Iandola, and T. Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In ICCV, 2013.

[13]Zhang, N., Donahue, J., Girshick, R., & Darrell, T. (2014, September).Part-based R-CNNs for fine-grained category detection. In European conference on computer vision (pp. 834-849). Springer, Cham

[14]Attention Convolutional Binary Neural Tree for Fine-Grained Visual Categorization

[15]Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016. 3

[16]Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. CoRR, abs/1904.11492, 2019. 3, 8

[17]Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In CVPR, pages 7794–7803, 2018. 3

[18]Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In CVPR, pages 7132–7141, 2018. 3

[19] D. Chang, Y. Ding, J. Xie, A. K. Bhunia, X. Li, Z. Ma, et al., "The devil is in the channels: Mutual-channel loss for fine-grained image classification", *IEEE Transactions on Image Processing*, vol. 29, pp. 4683-4695, 2020.

[20] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj and L. Song, "SphereFace: Deep hypersphere embedding for face recognition", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 212-220, Jul. 2017.

[21] H. Wang et al., "Cosface: Large margin cosine loss for deep face recognition", *Proc. IEEE CVPR*, pp. 5265-5274, 2018.

[22]Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011).The caltech-ucsd birds-200-2011 dataset.

[23] Maji, S., Rahtu, E., Kannala, J., Blaschko, M., & Vedaldi, A. (2013).Fine-grained visual classification of aircraft. arXiv preprintarXiv:1306.5151.

[24] Krause, J., Stark, M., Deng, J., & Fei-Fei, L. (2013). 3d object representations for fine-grained categorization. In Proceedings of the IEEE international conference on computer vision workshops (pp.554-561).