**基于知识度量的图像多粒度识别与检索的论文研究**

**丁磊 西安科技大学**

1. **引言**

细粒度识别，即在普通的图像识别的基础上再精细化一些，比如识别狗，他不仅要求要识别出这是一条狗，还要说出他是泰迪还是哈士奇或是其他某一种狗。这类问题由于类别之间的区别较小，因此人类对于该工作的识别准确率普遍较低。在数据集上，目前常被使用的是CUB-200-2011,Standford-Cars,FGVC-Airplane这三个数据集。

目前我看了三篇文献，大部分都是采用了深度神经网络提取特征，然后再通过诸如强化学习或者注意力机制甚至是通过强堆数据来提高其网络识别的有效性。通过百度、知乎、CSDN等，我了解到，在细粒度识别方面，还存在通过高阶Pooling把两个网络的特征做一个outer product，从而达到一定效果上的提高；通过先做姿态识别，找到相关部件，然后再对部件进行特征提取，然后将他们送到分类器在进行分类的办法等等。但目前大部分都还是依托于深度学习中神经网络的特征提取来做相关工作。

1. **复现文献算法详解**

下面，介绍一下，我进行复现的基于注意力机制的细粒度识别[3]。在该文中，作者采用了基于弱监督学习的图像增强方法，结合注意力机制，使得网络可以在不需要额外标注信息的情况下聚焦到图像中有重要信息的部分，从而提高了细粒度分类的有效性。

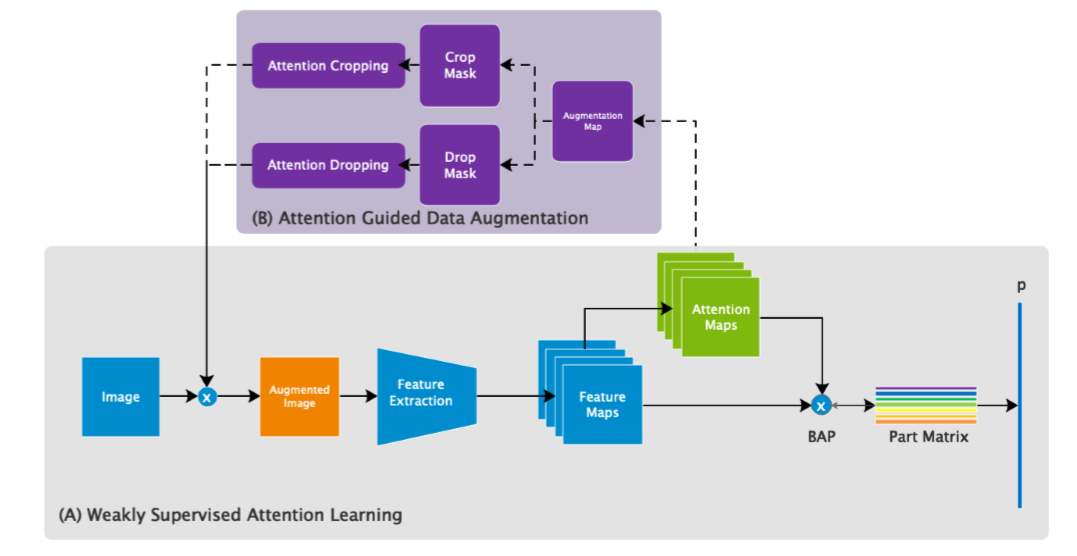


图1. 训练过程

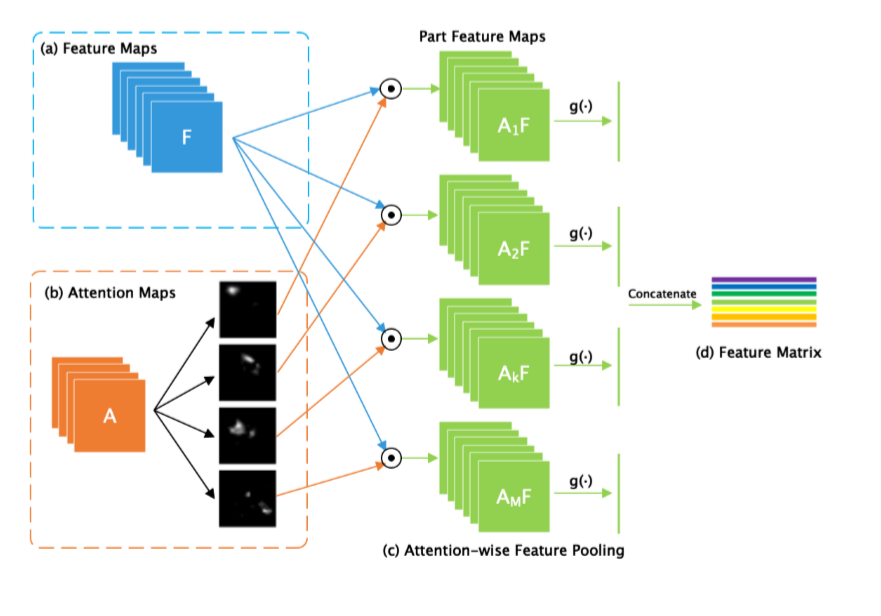
首先可以从图1中看到，训练过程分成了（A）弱监督注意力区域学习及（B）基于注意力机制的数据增强两部分。下面我们先看A部分。

* （A）**Weakly Supervised Attention Learning**

这一部分首先运用网络对原始图片进行基础的CNN特征提取，特征提取网络文中用的InceptionV3。提取到的信息文中作者称之为**Feature maps**，随后对**Feature maps**进行一个kernel size为1的卷及运算得到**Attention maps**，即**Attention maps**是由**Feature maps**降维得到的。根据作者描述，M个**Attention maps**中每一个都代表物体的一个位置例如鸟的头部等。在得到**Feature maps**和**Attention maps**后，作者采用了一种名叫**BAP（Bilinear Attention Pooling）**的方法，将**Feature maps**与每个channel的**Attention map**按元素相乘，如公式（1）所示:

(1)

相乘之后在经过池化降维以及拼接操作得到最后的**Feature Matrix**作为最后线性分类层的输入。

图2. BAP操作

* **(B)Attention-Guided Data Augmentation**

这里是用之前得到的Attention maps来做数据增强，即将Attention maps提取到的部位放大作为增强后的数据进行训练，从而增强其聚焦能力。在上一步中，我们获得了M个Attention map，网络会在M个中随机地选取一个作为后续数据增强的依据，而选取到一个Attention map之后，先对其进行归一化。

然后根据Attention map生成Crop Mask，Crop Mask是一种截图策略，强制模型注意这些细节区域。而Attention Dropping与Attention Cropping类似，是考虑到不同的Attention map可能聚焦了物体相同部位，该操作是为了让模型可以注意到其他部位。训练过程中还加入了一个损失函数的设计，除了计算预测结果的交叉熵损失之外，作者为了每次各个Attention map可以找到相同的物体部位，还加入了特征图与部位中心的平方差之和作为作为惩罚项，如下公式（2）：

(2)

* **预测过程**

预测过程同样分2部分，最终预测结果是两个子结果的平均值。预测过程如下:

1. 将原始图片输入训练好的模型中得到属于各个类别的概率，以及Attention Maps。
2. 将1中得到的M个Attention Maps取平均值，根据Attention Maps的平均值Am得到出截取框，将截取框上采样再放入训练好的网络中，得到注意力区域属于各个类别的概率。
3. 将上面两步的结果取平均值得到最后的分类结果。
4. **Others works**

接下来，我们简单的介绍一下其他2篇文献中的工作。

Ming Sun等人采用了一个基于注意力模型的网络加度量学习（即相似度学习）的方法，具体方法如下：首先，将不同种类图片输入CNN网络中学习注意力区域，然后将学习到的这个注意力区域特征带入到一个度量学习框架中，该框架会使相同注意区域的相同类别的特征更近，而不同注意力区域的不同特征会更远，从而有效地将不同类别的图片区分开[2]。值得注意的是，该算法，也是一个基于弱监督的学习算法，这对于数据集的要求将大大降低。

Abhimanyu Dubey等人则采用一种称为pairwise confusion正则化的方法，通过欧氏距离度量两个分布的相似度，并否认了KL散度的可行性，在文中给出了证明过程。值得注意的是，在本文中，作者采用了权重共享的孪生网络，每个网络都有其单独的交叉熵损失，两个网络之间有混淆损失，具体的，在训练时，作者将batch size分为2部分，一部分输入网络1中，另一部分输入网络2中，然后判断标签是否相同，如果相同则不存在混淆损失，若不相同，则计算混淆损失，该方法网络情况较为复杂，但引入孪生网络和欧氏距离度量相似度，较为特别[1]。

1. **参考文献**

[1] Dubey, A. , Gupta, O. , Guo, P. , Raskar, R. , & Farrell, R. . (2017). Pairwise confusion for fine-grained visual classification.

[2] Sun, M., Yuan, Y., Zhou, F., & Ding, E. (2018). Multi-Attention Multi-Class Constraint for Fine-grained Image Recognition. european conference on computer vision.

[3] Hu, Tao & Qi, Honggang. (2019). See Better Before Looking Closer: Weakly Supervised Data Augmentation Network for Fine-Grained Visual Classification.