【技术综述】你真的了解图像分类吗？

全能言有三

作者 | 言有三

编辑 | 言有三

图像分类是计算机视觉中最基础的任务，基本上深度学习模型的发展史就是图像分类任务提升的发展历史，不过图像分类并不是那么简单，也没有被完全解决。

01

什么是图像分类

图像分类是计算机视觉中最基础的一个任务，也是几乎所有的基准模型进行比较的任务。从最开始比较简单的10分类的灰度图像手写数字识别任务mnist，到后来更大一点的10分类的 cifar10和100分类的cifar100 任务，到后来的imagenet 任务，图像分类模型伴随着数据集的增长，一步一步提升到了今天的水平。现在，在imagenet 这样的超过1000万图像，超过2万类的数据集中，计算机的图像分类水准已经超过了人类。

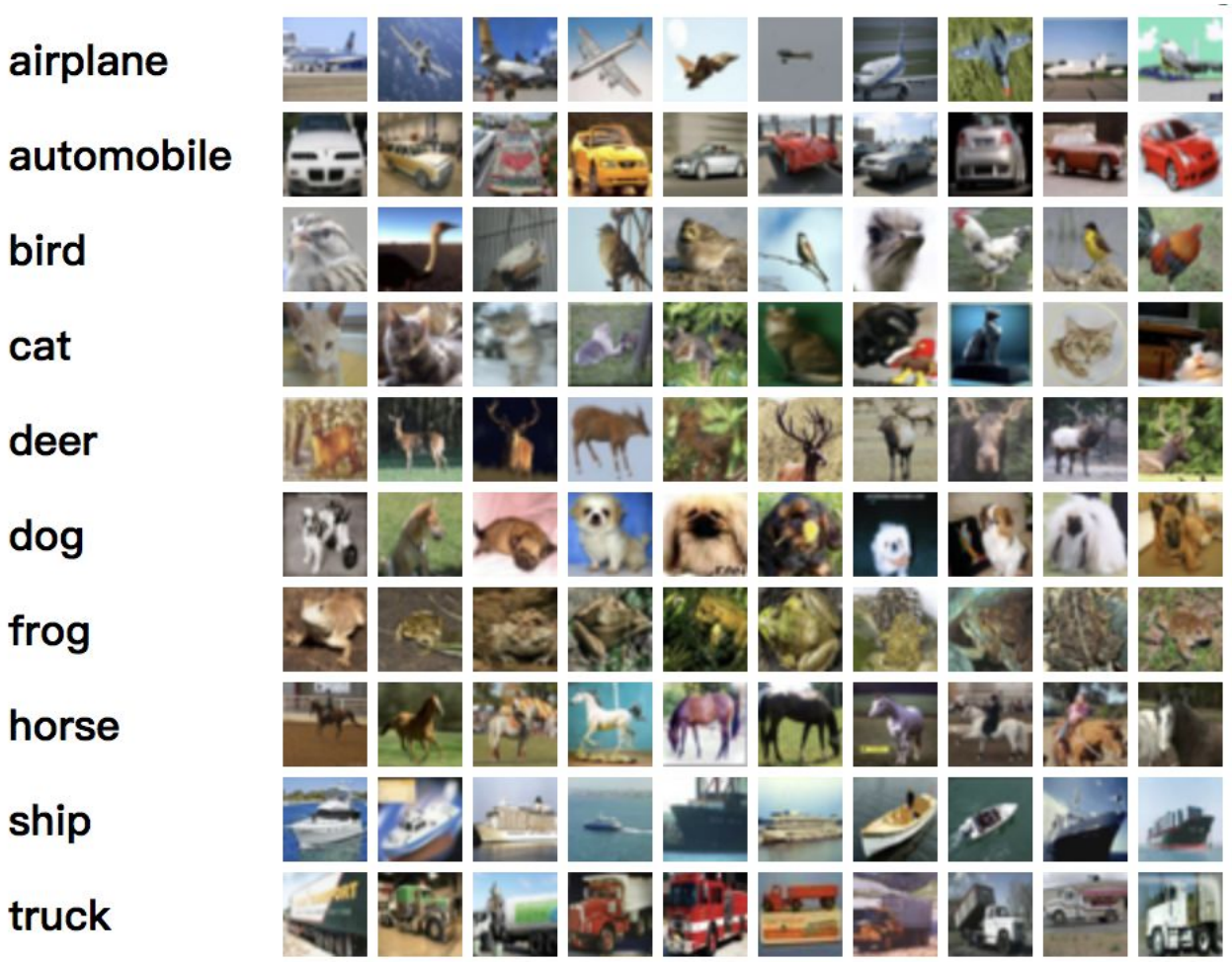
不过，不要把图像分类任务想的过于简单。

图像分类顾名思义就是一个模式分类问题，它的目标是将不同的图像，划分到不同的类别，实现最小的分类误差。总体来说，对于单标签的图像分类问题，它可以分为跨物种语义级别的图像分类，子类细粒度图像分类，以及实例级图像分类三大类别。

1.1 跨物种语义级别的图像分类

所谓跨物种语义级别的图像分类，它是在不同物种的层次上识别不同类别的对象，比较常见的包括如猫狗分类等。这样的图像分类，各个类别之间因为属于不同的物种或大类，往往具有较大的类间方差，而类内则具有较小的类内误差。

下面是cifar10 中的10个类别的示意图，这就是一个典型的例子。



cifar包含10个类别，分别是airplane，automobile，bird，cat，deer，dog，frog，horse，ship，truck，其中airplane，automobile，ship，truck都是交通工具，bird，cat，deer，dog，frog，horse都是动物，可以认为是两个大的品类。而交通工具内部，动物内部，都是完全不同的物种，这些都是语义上完全可以区分的对象，所以cifar10的分类任务，可以看作是一个跨物种语义级别的图像分类问题。类间方差大，类内方差小。

1.2　子类细粒度图像分类

细粒度图像分类，相对于跨物种的图像分类，级别更低一些。它往往是同一个大类中的子类的分类，如不同鸟类的分类，不同狗类的分类，不同车型的分类等。

下面以不同鸟类的细粒度分类任务，加利福尼亚理工学院鸟类数据库-2011，即Caltech-UCSD Birds-200-2011为例。这是一个包含200类，11788张图像的鸟类数据集，同时每一张图提供了15个局部区域位置，1个标注框，还有语义级别的分割图。在该数据集中，以woodpecker为例，总共包含6类，即American Three toed Woodpecker，Pileated Woodpecker，Red bellied Woodpecker，Red cockaded Woodpecker，Red headed Woodpecker，Downy Woodpecker，我们取其中两类各一张示意图查看如图。

从上图可以看出，两只鸟的纹理形状都很像，要像区分只能靠头部的颜色和纹理，所以要想训练出这样的分类器，就必须能够让分类器识别到这些区域，这是比跨物种语义级别的图像分类更难的问题。

1.3 实例级图像分类

如果我们要区分不同的个体，而不仅仅是物种类或者子类，那就是一个识别问题，或者说是实例级别的图像分类，最典型的任务就是人脸识别。

在人脸识别任务中，需要鉴别一个人的身份，从而完成考勤等任务。人脸识别一直是计算机视觉里面的重大课题，虽然经历了几十年的发展，但仍然没有被完全解决，它的难点在于遮挡，光照，大姿态等经典难题，读者可以参考更多资料去学习。

02

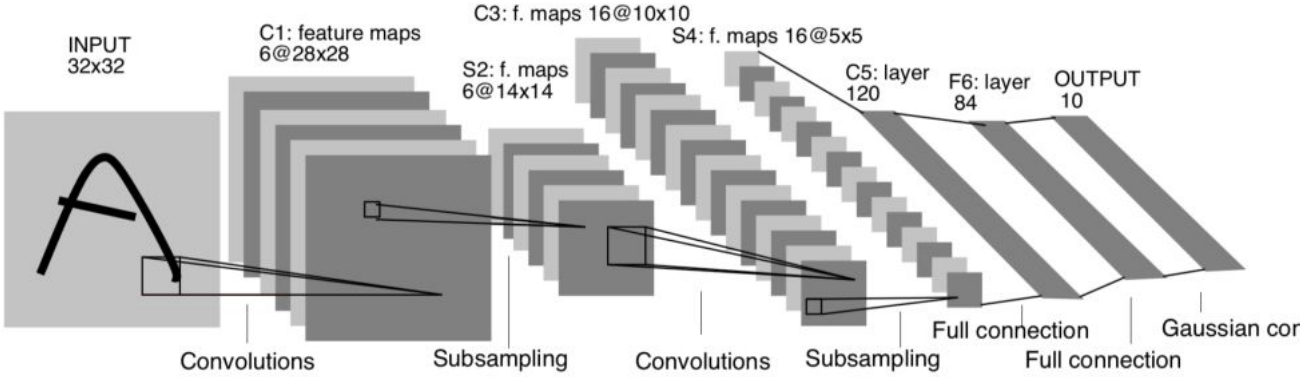
图像分类模型

图像分类任务从传统的方法到基于深度学习的方法，经历了几十年的发展。这里只关注于深度学习的进展，下面重点讲述几个重要的节点。

2.1 MNIST与LeNet5

在计算机视觉分类算法的发展中，MNIST 是首个具有通用学术意义的基准。这是一个手写数字的分类标准，包含 60000 个训练数据，10000 个测试数据，图像均为灰度图，通用的版本大小为 28×28。

在上个世纪90年代末本世纪初，SVM and K-nearest neighbors方法被使用的比较多，以SVM为代表的方法，可以将MNIST分类错误率降低到了0.56%，彼时仍然超过以神经网络为代表的方法，即LeNet系列网络。LeNet网络诞生于1994年，后经过多次的迭代才有了1998年的LeNet5，是为我们所广泛知晓的版本。



这是一个经典的卷积神经网络，它包含着一些重要的特性，这些特性仍然是现在CNN网络的核心。

- 卷积层由卷积，池化，非线性激活函数构成。从1998年至今，经过20年的发展后，卷积神经网络依然遵循着这样的设计思想。其中，卷积发展出了很多的变种，池化则逐渐被带步长的卷积完全替代，非线性激活函数更是演变出了很多的变种。
- 稀疏连接，也就是局部连接，这是以卷积神经网络为代表的技术能够发展至今的最大前提。利用图像的局部相似性，这一区别于传统全连接的方式，推动了整个神经网络技术的发展。

虽然LeNet5当时的错误率仍然停留在0.7%的水平，不如同时期最好的SVM方法，但随着网络结构的发展，神经网络方法很快就超过了其他所有方法，错误率也降低到了0.23%，甚至有的方法已经达到了错误率接近0的水平。

2.2 ImageNet与AlexNet

在本世纪的早期，虽然神经网络开始有复苏的迹象，但是受限于数据集的规模和硬件的发展，神经网络的训练和优化仍然是非常困难的。MNIST和CIFAR数据集都只有60000张图，这对于10分类这样的简单的任务来说，或许足够，但是如果想在工业界落地更加复杂的图像分类任务，仍然是远远不够的。

后来在李飞飞等人数年时间的整理下，2009年，ImageNet数据集发布了，并且从2010年开始每年举办一次ImageNet大规模视觉识别挑战赛，即ILSVRC。ImageNet数据集总共有1400多万幅图片，涵盖2万多个类别，在论文方法的比较中常用的是1000类的基准。

在ImageNet发布的早年里，仍然是以SVM和Boost为代表的分类方法占据优势，直到2012年AlexNet的出现。

AlexNet是第一个真正意义上的深度网络，与LeNet5的5层相比，它的层数增加了3层，网络的参数量也大大增加，输入也从28变成了224，同时GPU的面世，也使得深度学习从此进行GPU为王的训练时代。

AlexNet有以下的特点：

- 网络比LeNet5更深，包括5个卷积层和3个全连接层。
- 使用Relu激活函数，收敛很快，解决了Sigmoid在网络较深时出现的梯度弥散问题。
- 加入了Dropout层，防止过拟合。
- 使用了LRN归一化层，对局部神经元的活动创建竞争机制，抑制反馈较小的神经元放大反应大的神经元，增强了模型的泛化能力。
- 使用裁剪翻转等操作做数据增强，增强了模型的泛化能力。预测时使用提取图片四个角加中间五个位置并进行左右翻转一共十幅图片的方法求取平均值，这也是后面刷比赛的基本使用技巧。
- 分块训练，当年的GPU计算能力没有现在强大，AlexNet创新地将图像分为上下两块分别训练，然后在全连接层合并在一起。
- 总体的数据参数大概为240M，远大于LeNet5。

2.3 分类模型的逐年进步

2013年ILSVRC分类任务冠军网络是Clarifai，不过更为我们熟知的是zfnet。hinton的学生Zeiler和Fergus在研究中利用反卷积技术引入了神经网络的可视化，对网络的中间特征层进行了可视化，为研究人员检验不同特征激活及其与输入空间的关系成为了可能。在这个指导下对AlexNet网络进行了简单改进，包括使用了更小的卷积核和步长，将11x11的卷积核变成7x7的卷积核，将stride从4变成了2，性能超过了原始的AlexNet网络。

2014年的冠亚军网络分别是GoogLeNet和VGGNet。

其中VGGNet包括16层和19层两个版本，共包含参数约为550M。全部使用3×3的卷积核和2×2的最大池化核，简化了卷积神经网络的结构。VGGNet很好的展示了如何在先前网络架构的基础上通过简单地增加网络层数和深度就可以提高网络的性能。虽然简单，但是却异常的有效，在今天，VGGNet仍然被很多的任务选为基准模型。

GoogLeNet是来自于Google的Christian Szegedy等人提出的22层的网络，其top-5分类错误率只有6.7%。

GoogleNet的核心是Inception Module，它采用并行的方式。一个经典的inception结构，包括有四个成分。1×1卷积，3×3卷积，5×5卷积，3×3最大池化，最后对四个成分运算结果进行通道上组合。这就是Inception Module的核心思想。通过多个卷积核提取图像不同尺度的信息然后进行融合，可以得到图像更好的表征。自此，深度学习模型的分类准确率已经达到了人类的水平(5%~10%)。

与VGGNet相比，GoogleNet模型架构在精心设计的Inception结构下，模型更深又更小，计算效率更高。

2015年，ResNet获得了分类任务冠军。它以3.57%的错误率表现超过了人类的识别水平，并以152层的网络架构创造了新的模型记录。由于ResNet采用了跨层连接的方式，它成功的缓解了深层神经网络中的梯度消散问题，为上千层的网络训练提供了可能。

2016年依旧诞生了许多经典的模型，包括赢得分类比赛第二名的ResNeXt，101层的ResNeXt可以达到ResNet152的精确度，却在复杂度上只有后者的一半，核心思想为分组卷积。即首先将输入通道进行分组，经过若干并行分支的非线性变换，最后合并。

在ResNet基础上，密集连接的DenseNet在前馈过程中将每一层与其他的层都连接起来。对于每一层网络来说，前面所有网络的特征图都被作为输入，同时其特征图也都被后面的网络层作为输入所利用。

DenseNet中的密集连接还可以缓解梯度消失的问题，同时相比ResNet，可以更强化特征传播和特征的复用，并减少了参数的数目。DenseNet相较于ResNet所需的内存和计算资源更少，并达到更好的性能。

2017年，也是ILSVRC图像分类比赛的最后一年，SeNet获得了冠军。这个结构，仅仅使用了“特征重标定”的策略来对特征进行处理，通过学习获取每个特征通道的重要程度，根据重要性去降低或者提升相应的特征通道的权重。

至此，图像分类的比赛基本落幕，也接近算法的极限。但是，在实际的应用中，却面临着比比赛中更加复杂和现实的问题，需要大家不断积累经验。

03

总结

虽然基本的图像分类任务，尤其是比赛趋近饱和，但是现实中的图像任务仍然有很多的困难和挑战。如类别不均衡的分类任务，类内方差非常大的细粒度分类任务，以及包含无穷负样本的分类任务。

- 不是所有的分类任务，样本的数量都是相同的，有很多任务，类别存在极大的不均衡问题，比如边缘检测任务。图像中的边缘像素，与非边缘像素，通常有3个数量级以上的差距，在这样的情况下，要很好的完成图像分类任务，必须在优化目标上进行设计。
- 虽然前面我们说过图像分类可以分为3大类，对于猫狗分类这样的语义级别的问题，算法已经达到或超越人类专家水平，但是对于如何区分不同种类的猫这样的细粒度分类问题，算法仅仅在某些数据集上勉强能突破90%，远未超越人类专家，还有非常大的发展空间。
- 另外前面所说的分类，全部都是单标签分类问题，即每一个图只对应一个类别，而很多的任务，其实是多标签分类问题，一张图可以对应多个标签。多标签分类问题，通常有两种解决方案，即转换为多个单标签分类问题，或者直接联合研究。前者，可以训练多个分类器，来判断该维度属性的是否，损失函数常使用softmax loss。后者，则直接训练一个多标签的分类器，所使用的标签为0,1,0,0...这样的向量，使用hanmming距离等作为优化目标。

历史推荐

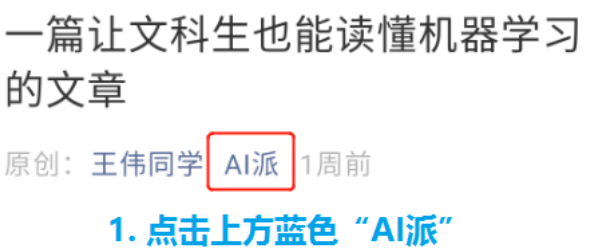
人人都是数据分析师，人人都能玩转Pandas | Numpy 精品系列教程汇总 | 我是如何入门机器学习的呢 | 谷歌机器学习43条黄金法则

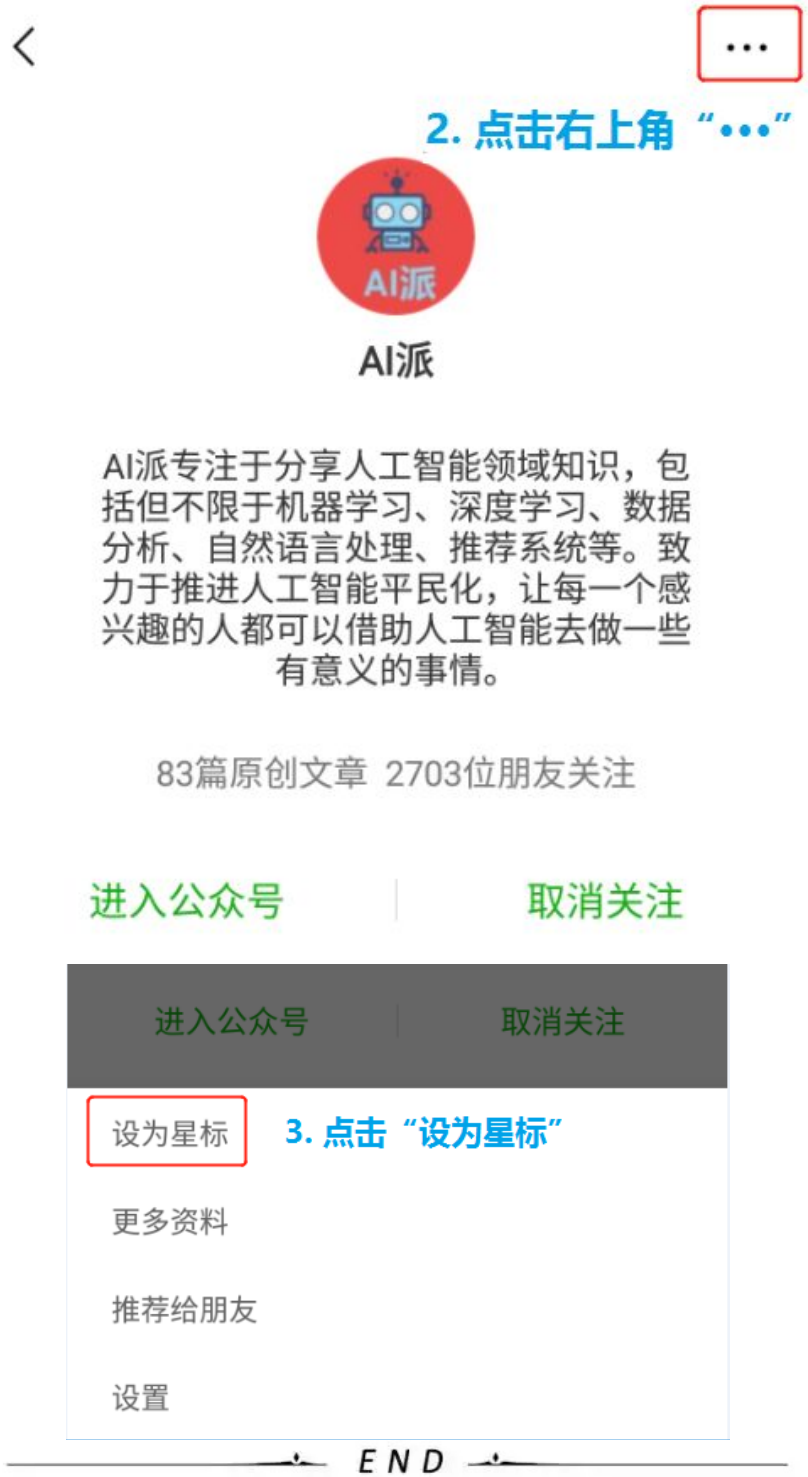
现在微信改版了

很多人说找不到AI派了

其实只需要设为星标就可以了

只需三步，轻松搞定星标设置




2. 点击右上角“•••”
AI派
AI派
AI派专注于分享人工智能领域知识，包括但不限于机器学习、深度学习、数据分析、自然语言处理、推荐系统等。致力于推进人工智能平民化，让每一个感兴趣的人都可以借助人工智能去做一些有意义的事情。
83篇原创文章 2703位朋友关注
进入公众号
取消关注
进入公众号
取消关注
设为星标
3. 点击“设为星标”
更多资料
推荐给朋友
设置

END

长按，识别二维码，加关注

文章转载自公众号

与有三学AI

与有三学AI