1. 大家上午好，今天这节课，由我来介绍CV领域中的细粒度分类方向的相关知识。首先是几句题外话，这次的内容我主要以介绍为主，具体的算法实现及公式讲到的较少，内容上不会很难理解，如果有同学对具体的算法感兴趣，我们可以课下再一起讨论。那么接下来我们就开始我们的课程。

细粒度分类，我们简称为FGVC，Fine Grained Visual Classification。我们又称之为物体精细识别。其实细粒度分类问题并不仅限于图像的细粒度分类，在其他领域，像文本啊，也会用到细粒度分类方法，在这里，我们讲到的细粒度分类就是图像的细粒度分类，后期我说到的细粒度分类，也全部是图像的细粒度分类。（1’10’’）

1. 那么要讨论细粒度分类，我们就必须先了解一下图像分类，因此，这节课的内容分两大部分，第一部分是图像分类的相关内容，在这一部分，首先会介绍一下我们做图像分类用到的数据集，其次就是图像分类领域比较典型的几种网络结构，其中最重点的介绍的就是深度学习方法。第二部分是细粒度分类的相关内容。主要是介绍我们用到的细粒度分类的数据集以及一些到现在为止，典型的，效果比较好的，有代表意义的算法。绝大多数细粒度分类的算法也都是基于分类算法中的深度学习的算法做的。所以，接下来，我们先一起看一下图像分类的相关内容.(1’14’’)
2. 首先，我们来简单介绍下图像分类的相关内容。
3. 什么是图像分类呢？这个概念理解起来比较简单，我也简单百度了一个比较官方的图像分类的定义，图像分类：根据各自在图像信息中所反映的不同特征，把不同类别的目标区分开来的图像处理方法。那么这个定义中有几个关键词，首先，分类的对象是不同类别的目标，分类的依据是图像信息中反映的不同的特征，分类的目的，就是把这些不同类别的给它分开。最后就是图像分类也是一种图像处理的方法。

像下图，我们有9张图片，这9张图片属于三个类别，分别是Cat类，Dog类和Pig类，那么Cat，pig, dog 是他们的类别标签，我们也可以把它们定义为第0类，第1类和第2类。这样他们就组成了一个包含9张图片，共三类的小的数据集，图像分类的目的就是将每张图片正确地分到他们所属的类别。（1’12’’）

1. 图像分类问题，从提出到现在，一直是CV领域中比较主流的问题，这也和图像分类技术的广泛应用是分不开的。首先，公安系统的像指纹识别，人脸识别；还有我们常说的辅助医疗，通常医生判定患者是否得病，会看他拍的片子，那么通过图像分类技术，可以帮助医生诊断。在商业领域也有很广泛的应用，像商标管理系统，判断商标是否已经注册，会用到我们的分类技术，否则那么大的数据库，靠人工管理，会很麻烦。还有我们网上购物，淘宝，我们的搜索引擎，百度，都开始了搜图的功能，比如我们看一件衣服不错，打开淘宝扫一扫，同款就都出来了。还有地理方面，在管理土地资源利用的时候，会经常对拍到的遥感图像进行分类，还有在生态保护方面，我们物种的分类的应用也非常广泛。总之，现在大多数领域都用到图像分类技术，图像分类现在几乎是无处不在。
2. 对于我们而言，把狗和猫分开，再简单不过了，首先，我们大脑的视觉系统的组成非常复杂，而且非常高级，有很多复杂的神经网络；其次，我们见过了，或者说已经学习了太多的猫和狗，因此，把狗和猫分开，对我们来说很容易。但是对计算机来说就很难了。因为计算机见到的，全部都是数字。对于计算机来说，图像是一个由数字组成的巨大的三维数组。现在我们看到的这一小块28\*14的图片在计算机中是一个数字组成的28\*14\*3的二维数组。这里我只把他R通道的取值拿了出来。整张图片大小是317\*488，计算机通过这317\*488\*3=464088个数字来判断这张图片属于Pig类的概率为92%，属于Cat类的概率为4.8%，属于Dog类的概率为3.2%，预测这张图片的分类标签是Pig，从而实现图像的分类。所以说，让计算机能够正确实现分类有一定难度的。
3. 接下来，我们先来看一下，常用的图像分类的数据集。
4. 首先是1998年，LeCun提出的MNIST手写体识别的数据集。大家对这个数据集已经非常熟悉了，这是我们在模式识别课上用到的数据集。这个数据集一共包含60000张训练图片和10000张测试图片，从0-9一共分为10类，这个数据集的标注信息只有类别标签。当时这个数据集的提出时LeCun为应用于美国邮政和支票系统而提出来的，这个数据集也一直沿用到了现在。
5. CIFAR数据集也是图像分类中常用的数据集。Cifar数据集包含两个，cifar10 和cifar100，两个都包含50000张训练图和10000张测试图，cifar10是被标注成10类，cifar100 被标注成了100类，Cifar 数据集是Alex和Hinton等人在得到CIFAR的赞助后做的数据集，标注信息也只有类别标签。
6. Caltech数据集是加利福尼亚理工学院做的256类的图像数据集，其中，Caltech-101包含102类，共9144张图片，Caltech-256包含257类，共30607张图片。这些图片的标注信息也只有类别标签。这三个数据集都是比叫小的数据集，他们最多的也只是几万张图片。
7. 接下来我们来看Pascal voc数据集。从VOC2005-VOC2012，每年都有一次图像识别的挑战赛。Pascal VOC数据集包含11000张图片，一共被分为20类。但是poscal voc数据集的标注信息除了类别标签之外，还标注的27000个目标，以及7000个目标的分割图像，因此，这个数据集不仅可以用于图像分类，还可用于目标检测，图像分割，动作分类以及人体部位检测等。
8. ImageNet数据集是现在常用的数据集。ImageNet数据集是ImageNet大规模视觉识别挑战赛中用到的数据集。ImageNet挑战赛是现在视觉识别领域最有名的一个比赛。值得一提的是，往年都是Google，MSRA等大公司夺冠，16年是中国团队包揽了全部项目的冠军。这也可以看出现在中国计算机视觉的发展也是不落人后的。Imagenet数据集包含130万张训练图像和10万张测试图像，共1000类，几乎成为了目前深度学习图像领域算法性能检验的“标准”数据集。除此之外，Imagenet也有标注Boundingbox信息，因此Imagenet数据集不仅可以用于分类，也可以用于定位和检测等。
9. MS coco 数据集是微软2016年公布的数据集，这个数据集包含了16万多张训练图片，8万多张验证图片和8万多张测试图片。图片一共被分成了102类。COCO数据集类别较少，但是每一类的样本比较多，也不仅仅是做图片分类的数据集，他的标注信息除了类别标签外，还有图像分割的信息和语境信息，因此这个数据集除了做图像分类之外，还可以做分割，目标检测等。
10. Google的open Image dataset 是迄今为止最大的一个数据集，共有900万张，超过6000类的图片。在这个数据集中，一张图片往往对应多个标签，比如这张图片，不仅对应dog标签，还有哺乳动物，脊椎动物等，一张图片一般对应8个标签。其次，他还有场景信息的标签，可以用来做场景识别。
11. 有了数据集之后，那我们再来看一下图像分类问题的评价指标。
12. 我们常用的图像分类中的评价指标主要有四个，分别是全局准确率，top-1错误率，top-5错误率和混淆矩阵，全局准确率比较好理解，全局准确率就是测试集中，正确分类的目标数和测试集全部目标数的比值。
13. Top-n 错误率就是当我们把输出的类别按概率大小排序，若正确类别能够排到前n位，那就默认分类正确。因为之前在数据集部分，我们提到过，有的图片会有不止一个类别，如果我们单纯只看top-1error的话，有时候对网络性能的评价可能不够准确，因此在有些时候我们会把top-1 error 和top-5 error作为评价指标。
14. 还有一个比较重要的评价指标是混淆矩阵，混淆矩阵是一种可视化工具，主要是应用于有监督学习问题，矩阵的每一列代表一个类的预测的类别，而每一行代表一类的真实类别。混淆矩阵，顾名思义，主要功能就是看分类器是否把不同类别给混淆了。我们举一个例子，班级中实际有20名女孩，80名男孩，那么在分类的时候，将27名同学分成女孩，将73名同学分成了男孩，而这27名被分为女孩的同学中，只有18名是真的女孩，其他9名是男孩，被分成男孩的73名同学中，也只有71名是男孩，另外两名同学是女孩。这样根据这样的分类结果，我们可以画出这样一个混淆矩阵。这样，对角线上的就是正确分类的，而在对角线之外的，都是分类错误的。
15. 通过混淆矩阵，我们可以求得其他评价指标的数值，像准确率，精确率和召回率。我们先看一下在分类问题中，可能会出现的几种分类正确和分类错误的情况。我们以二分类为例，我们将第一类标记为正类，将第二类标记为负类。那么就一共有TP，FP，FN，TN四种情况的样本。TP，FP，FN，TN的含义分别为……，我再来举个例子让大家更好的理解这四个定义。准确率我们在刚才已经讲过，这里就不再赘述了。但是分类有不同的情况，有的时候仅准确率高是不够的。比如我们要做一个质量检测的分类器，一批整体的合格率在90%以上，那在有误差的情况下，我们更希望分类器分出的所有的合格产品都是真正的合格产品，这样我们只需要在极少数被分类器分为不合格类别的产品中复检。这种情况下，就有另外一个评价指标：精确率也叫做查准率。精确率是标签为正类并且也被分类器分到正类的样本的数量和所有被分类器分到正类的样本的数量。精确度越大，代表被分类器分出来的正类的可靠性越高。在产品质量检测方面，如果他的准确率不够高，但是精确率足够高，那我们还是可以接受的。还有另外一种情况，比如说我们想做一个是否是癌症的诊断分类器。我们将阳性定义为正类，阴性定义为负类。和上一个情况不同的是，我们更希望所有的正类都被挑选出来，没有漏掉的，将负类误判为正类的代价就会相对较小。这样就会考虑另外一个评价指标：召回率，也叫做查全率。召回率是标签为正类并且被分到正类的样本数和所有标签为正类的样本数的比值。召回率越高，就代表正类这一类被挑选出来的越多。精确率是被分成正类的样本中真正是正类的样本数量和全部被分为正类的样本的数量的比值。召回率是被分成正类的样本中真正是正类的数量和所有真正是正类的样本的比值。因此在这个示例中，我们可以P（讲讲具体步骤）
16. 通过精确率和召回率我们可以求得F-measure，F-measure是精确率和召回率的调和均值，其中a是权重，它调节精确率和召回率的权重。
17. 接下来，我们来一起来了解一下图像分类的方法。
18. 图像分类的方法有传统的图像分类方法，也有近几年发展飞快的深度学习方法，那么我们先来看一下传统的图像分类方法。这个中医治病的漫画，可以帮助我们理解传统分类方法的过程。我们之前也提到过，在图像分类问题中，特征提取至关重要，传统分类方法，我们把它分成两大步，
19. 传统图像分类方法的分类过程主要包括以下几个阶段：首先，我们的图像有可能会存在质量不高，背景过于复杂等问题，因此，有时候会针对不同图像存在的不同的问题有不同的预处理方法。预处理之后，我们会对图像进行特征提取，这一部分也是图像分类过程中最重要的。几个典型的特征提取方法有SIFT，SURF，HOG等。但是提取到的特征的格式未必适合做后续的分类的工作，因此会做一个特征表达。特征表达主要是把特征映射成其他形式从而来更好的处理我们提取到的特征。最常用的特征映射方式是FV，FV甚至在后期的深度学习方法中以及一些细粒度分类的方法中也都还会用到，除了FC还有向量量化，LCC等方法。然而我们提取到的特征可能并不全都是有用的特征，数量太多的特征也会造成分类性能的下降，特征选择是要从所有的特征中，找出主要的特征，去掉大量的冗余和无关的特征。特征选择的方法主要有按搜索测量进行特征选择的方法，按评价准则进行特征选择的方法两种。在传统的分类方法中，降维也是很重要的一部分，特征维度过高会严重影响我们的分类性能，特征降维就是将特征从高维转化成低维的过程，主要的特征降维的方法是主成分分析法（PCA），以及LDA，特征哈希等。在我们把特征提取好，并且处理好之后，我们就要用到分类器来进行分类了。最常见的分类器是SVM，在深度学习之前，SVM是图像分类领域常用的方法，在后期的细粒度分类部分，我们也会看到，依然有很多深度学习的方法也结合了SVM来进行图像分类。
20. 那么介绍了这么多的传统的图像分类的方法，那么我们就很容易感受到传统方法的弊端了。首先，人工成本太高，无论是进行特征提取，还是特征表达，每一部分都要有特定的算法，每一步都需要人工的参与，这样的方法对人力的要求太高。除此之外，最重要的就是传统分类方法的分类效果不够好，分类准确率不够高。还有，随着数据量的增大以及电脑硬件的发展，可处理的数据越来越多，而传统的图像分类方法能够处理的样本数量有限，也一定程度上制约了传统分类方法的发展。
21. 接下来就是我们要讲的图像分类方法中，可以carry全场的深度学习方法。深度学习是一种解决问题的方法，这张漫画是我从中科院的博士刘昕在天津大学深度学习暑期学校中做的报告中截取的。从这幅漫画中我们可以看出，深度学习是一种方法，是一种处理问题的工具，但是它不仅可以用来解决CV问题，现在也有很多其他方面的问题在用深度学习做。目前，深度学习在人工智能领域占据统治地位，但凡有关人工智能的产业报道，必然离不开深度学习。现在深度学习火到什么程度呢？现在像谷歌，Facebook，微软，百度，阿里巴巴，360等，都在搞深度学习，那么接下来，就让我们一起了解一下深度学习的内容。
22. 深度学习是机器学习的一个分支，是一系列用于特征提取和特征转换的非线性单元的级联。深度学习方法一定程度上也可以理解成深度神经网络的方法，网络的特点首先是深，包含多层，通常不少于两个隐含层的网络我们就会称之为深度神经网络。其次神经网络中包含多个非线性单元，有了非线性单元我们才可以进行非线性分类。下面我举了几个深度学习中比较有代表意义的几种深度神经网络，首先是多层感知机MLP，MLP大家再熟悉不过了，MLP是全连接结构，也就是前一层的每个节点都和下一层的每个节点连接到了一起，和MLP结构类似的，还有auto-encoder等。中间这个是现在统治了整个图像分类领域的卷积神经网络CNN，卷积神经网络相对于MLP在特征提取及表达方面有了很多改进，也取得了更好的分类效果，卷积神经网络也是我们接下来要重点讲的深度神经网络。第三个是递归神经网络RNN，递归神经网络有时间递归和结构递归，递归神经网络在图像分类问题中应用的不多，在此也不再详细解释了。
23. 深度学习方法的出现，是计算机视觉领域的革命性事件。通过Revolution我们可以知道深度学习方法带来的重大影响，那么在图像分类领域，深度学习的方法到底都有什么优势呢？首先，用深度神经网络对图像进行分类，取得了更好的分类效果。右图是我们刚才提过的现在最著名的图像分类和识别领域的挑战赛ImageNet比赛的近几年的冠军。10年的冠军由美国NEC研究院和UIUC获得，他们的底层特征采用了密集提取的HOG和LBF特征，特征编码的算法使用了局部坐标编码和超向量编码，最终采用的是基于平均随机梯度下降的大尺度SVM进行分类，在ImageNet数据集上获得了28.2%的top-5错误率。2011年的冠军是施乐欧洲研究中心，他们沿用了2010年的框架，他们主要改进了三个方面，首先采用Fisher向量进行特征编码，其次使用了乘积量化来进行特征压缩，在分类器方面，他们使用的是基于随机梯度下降的SVM。进过一系列的改进，top-5错误率下降了2个百分点。上面这两种方法都是基于我们刚才提到过得传统的图像分类方法。从12年开始，我们就迎来了深度学习方法。12年是加拿大多伦多大学的Hinton教授和他的学生Alex，提出了深度卷积神经网络AlexNet，将分类准确率一下子降到了16.42，通过这个数据，我们一下子就感受到了深度学习的力量，从12年之后，所有的冠军都是深度学习的方法，从12年开始，深度学习几乎是统治了整个图像分类领域。像14年的冠军，Google提出的GoogleNet，14年的亚军，牛津大学视觉组提出的VGGNet以及15年的冠军，何恺明提出的ResNet等，直接把错误率降到了3.6%。这些都是基于深度神经网络的模型。通过这个表格，我们可以直观的看出深度神经网络的分类效果。

除了分类效果之外，深度神经网络还有很好的可移植性。移植性就是说，一个神经网络如果在一个数据集上分类效果不错，那么往往，这个网络在其他数据集上的表现也都不错。我们不再需要针对一个数据集，就提出一种方法，往往提出的神经网络都是可移植的，这样的分类方法有更强的现实意义。

1. 另外，深度学习方法还有一个更为人津津乐道的是end-to-end模式，我们称之为端到端。我用一幅改编自刘忻博士ppt中的漫画来解释下什么是端到端。比如我们做一道可乐鸡翅，通常的流程是，把鸡翅洗干净，各种撩准备好，第二步，放进锅里，煎至金黄；第三步，加入葱姜进行翻炒入味，第四部，加入可乐，收汁，装盘。这样，一道可乐鸡翅就做好了。那么这个过程可以类比到我们的传统方法进行图像分类的过程中，我们要想做图像分类，就得一步一步，一个模块一个模块的进行操作，特征提取，特征选择，特征表达，分类器分类。端到端就是从输入，直接到输出，像是一个box，从原料，直接到最终的成品。深度学习方法的输入就是图片，输出是最终的分类结果，中间无需人工操作。
2. 刚才我们提到过，深度神经网络的深度体现在它层数的多少。深度神经网络的这么多层主要是用来做特征提取的，这种情况下，深度神经网络提取到的特征就具有层级性的特点。刚才已经提到过，计算机看到的图片都是一个个像素点，我们会发现，像素特征的有效性明显不够高，如果直接拿像素特征来进行分类，分类效果不会很好。那深度学习中的层级特征可以有效地改善这一问题。首先什么是层级特征呢？在图像识别领域，我们将特征分为几个层级。最低级的是像素级特征，这个不再解释了，到边缘级特征，边缘级特征已经开始包含边缘信息，那么这样的信息还比较低级，后期到纹理级，我们就可以看到，纹理级的特征比边缘级特征更抽象，更高级，到主题级，再到局部级特征再到全局特征。随着特征层级的升高，特征越来越抽象，越来越高级，这样的用于分类和识别也越来越有效。在深度神经网络中，我们提取到的特征也是从低级到中级到高级不断递进的。像我们的低级特征，可能就只包含一些像素特征或者是边缘特征，我们的中级特征中，就出现了纹理和主题，在高级特征中，就会出现局部甚至全局特征。这样的层级特征使得我们提取到的特征更加有效，因此能够更高程度地提高深度学习方法的图像分类准确率。这也是现在深度神经网络越做越深的原因之一。
3. 虽然深度学习现在火到不行，但是深度学习的发展也经历了漫长而曲折的过程。深度学习是从1986年Rumelhart等人提出了反向传播算法，也就是我们常说的以及现在仍在广泛使用的Bp算法，87年出现的多层感知机以及97年出现的LSTM推动了深度神经网络的发展，一直到98年，Lecun就是我们刚才提到的，提出了MNIST数据集的那位大牛，提出了第一个卷积神经网络LeNet，并将其成功应用于手写体识别问题。但是当时计算机的计算能力有限，LeNet在其他很多问题上的表现也不够好，在此之后，深度学习就陷入了沉寂。一直到2006年，著名学者加拿大多伦多大学教授Hinton在Science上发表的论文中提出了深度信念网路，开启了深度学习的第二春。2009年-2010年，深度学习方法已经成功应用于语音识别，一直发展到2012年，在ImageNet比赛上，Hinton和他的学生Alex提出的AlexNet获得了空前的成功，DeepLearning也被成功得应用于图像处理领域。12年之后，深度学习在图像处理领域呈现出爆炸式发展
4. 深度学习的回春离不开很多科学家的努力。在十年寒冬期，他们仍然坚持。

LeCun：1998年，第一个提出卷积神经网络的人，现在在facebook。

Hinton：提出BP算法，2006年在一篇发表在Science的论文上提到的深度信念网络，是现在的深度学习得开端，为深度学习带来了第二春。最开始，深度学习在语音识别领域大展拳脚，12年，Hinton及其学生提出了AlexNet，将深度学习成功得应用于图像处理领域。现在在谷歌。

Bengio：开创了神经网络做 language model 的先河，推动RNN的发展，至今仍在学术界。

1. 每次讲深度神经网络都是从MLP多层感知机开始，关于多层感知机的内容，在这我就不再赘述了，MLP是上世纪80年代流行的深度神经网络，MLP那么每一层都有非线性的激活函数，使用的是BP算法来对参数进行优化。多层感知机出现之后，解决了非线性分类的问题。
2. 但是MLP结构有其自身的缺点，首先，我们的图像是二维图像，MLP在训练时将整个二维图像拉成一个一维向量，这样就会忽略一部分像素之间的关系。其次，如果始终采用全连接的方式，比如一张1000\*1000的图像，输入层节点数就是1M，那隐藏层节点数也是这样的数量级，这样的话，输入层和隐藏层之间的参数数量就是10^12，这么多的参数，几乎是无法训练的。卷积神经网络就有效地解决了这两个问题：首先卷积神经网络提出了二维卷积核来进行特征提取。这个卷积核是用于进行特征提取的，对于一副图像，我们通常设置不同卷积核的数量，来提取不同的特征，这样就保证了在一个二维的空间中对图像的特征进行提取。那么这个图像是1000\*1000的大小，而我们卷积核只是3\*3的大小，怎么进行卷积呢？那么卷积核是不会一起和整张图片进行卷积的，而是对不同的局部进行卷积。这个特点被称为局部感知。局部感知的理念是通过研究猫的视觉皮层发现的一个现象：神经元在接受一幅图片的时候，并不是对图片的全部区域都很敏感，而是对图片的某个子区域比较敏感，神经元敏感的这个区域我们称之为：感受野。比如这个黑色的卷积核，他从这个位置出发，先卷积当前位置，当前位置就是它的感受野，来提取当前位置的特征，再继续向前滑动，依次类推，在整张图片上不停地滑动，从而来通过提取每个局部的特征来提取到整张图片的特征。除了局部感知，CNN还有另一个神器就是参数共享，刚才也讲到了，卷积核在这个位置以当前参数来提取该位置的特征，在下一位置也以当前参数来提取下一位置的特征，那这种方式是否是有效的呢？该如何理解呢？我们可以将每个卷积核都看成是一种特征提取的方式，这种方式和位置无关，因此，对于这幅图像上的所有位置，我们都可以采用相同的参数。
3. 卷积神经网络主要是由卷积层，池化层（下采样层），激活函数层，全连接层组成。通常我们所说的卷积网络的深度就是卷积层的数量。
4. 卷积层主要是通过使用固定大小的滤波器与整个图像进行卷积，来模拟Hubel和Wiesel提出的简单细胞，池化层通过取卷积得到的特征图中局部区块的最大值、平均值来达到降采样的目的， 并在这个过程中获得一定的不变性。池化层用来模拟Hubel和Wiesel理论中的复杂细胞。那么我们学习MLP得时候，学到了BP反向传播来进行参数优化，在卷积神经网络中也是一样的，卷积神经网络的参数就是卷积核中的参数，也就是图像黄色部分右下角小红字。在这个卷积操作中，我们是5\*5的输入图片，3\*3的卷积核，步长设置为1，得到了一个3\*3的输出特征图。卷积层呢，是用来提取特征的，通过卷积层提取到的特征，我们通常称之为卷积特征，通过对对在ImageNet数据集上所 训练得到的卷积网络模型进行分析, 发现从卷积网络中所取的特征具有更强的语义特性, 比人工特征具有更好的区分度. 而且卷积特征可以迁移到其他任务中，这就是卷积神经网络的迁移性。卷及神经网络在其他领域的任务中，例如场景识别，细粒度分类问题上，都获得了良好的分类性能，从实验中证明了卷积特征具有强大的泛化性。
5. ReLU激活函数是卷积神经网络常用的激活函数，ReLU的特点是，当输入<0时，输出全部是0，当输入>0时，输出等于输入。如右图，当权值全部是二维的时候，经过ReLU激活函数后的效果是这样的。
6. 我们之前学过的Sigmoid激活函数和tanh激活函数，都是指数运算，他们的梯度同样也都是指数运算，相比较之下，ReLU激活函数的运算速度更快，从这幅摘自AlexNet的论文中的图像我们可以看到，在降到25%的错误率时，ReLU激活函数的速度要比tanh激活函数快6倍。ReLU函数以0为中心，作为激活函数，我们更希望输入是0的时候，输出也是0.其次，ReLU激活函数在求解梯度的时候，更简单。最后就是ReLU激活函数在输入小于0时，输出都是零，这样就可以使得输出具有一定的稀疏性。稀疏性是一个非常重要的内容，在涉及到特征的问题中，我们都会强调稀疏性。输出具有一定的稀疏性，会有以下几个优点：首先，具有一定的稀疏性，有利于信息解离。在深度学习中，我们有一个明确的目标是从数据变量中解离出关键因子，因为一幅图片中，通常特征都是很密集的，这些特征也是相互关联的，但是有一些特征是有效的，有一些特征是无效的，如果我们能够解开特征之间复杂的关系，这样就可以提取关键信息，并且将复杂特征转化成稀疏特征，这样可以提高系统的鲁棒性。鲁棒性是指系统在扰动或不确定的情况下，仍能保持他们的某些特性的性能。其次，映射后的特征更容易实现线性分类。
7. softmax layer
8. 了解了卷积神经网络的基本知识之后，接下来我们一起来看几个典型的卷积神经网络。我以时间顺序给向大家介绍用于图像分类的卷积神经网络。第一个就是LeNet-5，刚才已经提到过，LeNet是LeCun和Bengio在1998年提出的第一个卷积神经网络，LeNet-5不加输入层的话，一共有七层。分别是卷积层，池化层或者说下采样层，卷积层，池化层，全连接层，全连接层，输出层。首先我们来看输入。卷积神经网络中，输入图像的大小是要严格限制的，根据有多少个卷积层，多少个池化层，严格计算出来的。在LeNet-5中，输入图片的大小是32\*32，那么这样就要求我们在使用LeNet-5之前，都要对图片做一下预处理，将图片的大小resize成32\*32.这个大小对于当时的MNIST数据集是可以的，在第一层卷积层中，有6个卷积核，卷积核大小为5\*5每一个卷积核和输入图片卷积都会得到一张特征图（feature map），我们现在看一下下面这张图，我把Lenet求解卷积和下采样的过程拿了过来，为什么要拿出来单独讲呢？是因为LeNet是很早，98年就出现的一个卷积神经网络，而后期经典的用于图像分类的卷积神经网络都是在12年到16年之间出现的，所以LeNet相比于这些CNN，在计算上还是有很多的不同之处，和我们刚刚讲到的卷积神经网络的经典计算也有一些不一样的地方，大家注意比较。看一下这个卷积过程，都在进行完卷积操作后，在卷积之后加了一个b，bias偏置，就得到了特征图。这样我们在经过第一层卷积层之后，就得到了6张28\*28的特征图，卷积层后面，紧跟了一层下采样层，它的下采样选用的是2\*2的感受野（receptive fields），所有的感受野都不重叠。感受野不重叠就是……，在下采样过程中，先加一个权重w，再加一个偏置b再经过一个sigmoid激活函数，最终得到输出。C3也是卷积层，包含16个5\*5大小的卷积核，因此也得到了16张特征图，但是需要注意的是，每一张特征图并不是完全输入了前一层的6张图，有的是将六张中相邻的三张作为输入，有的是将六张中的四张作为输入，有的是六张中的六张为输入，具体的方法可以参考LeNet的论文，在这里不再详细讲。为什么要这样做呢？首先，这样的不完全连接可以将连接的数量控制在合理的范围内，其次不同的特征图不同的输入，这样就会抽取不同的特征，这就像是之前提到的生物视觉系统类似，层层深入，提取更抽象，更高级的特征。同样，s4也是下采样层，和S2层一样，C5层也是卷积层，120个5\*5的卷积核，得到120张特征图，值得注意的是，在这一层，每个单元是和s4层全部的16个单元相连的。F6是全连接层，OUTPUT层是用到欧式径向基函数函数单元，这个函数就是用来求解输入向量和目标向量之间的距离，结果越大，证明距离越远。最终通过距离的判断来实现分类。从这我们也知道，这种训练方式属于有监督训练，实际上我现在常用的做图像分类的网络都是有监督训练的。在训练的时候，LeNet也是用到了反向传播算法。讲到这里，我们再回想一下，深度学习得发展，在98年提出了LeNet之后就陷入了寒冬期，那我们看LeNet的网络结构是存在很多问题的，首先，LeNet中用到的是Sigmoid激活函数，sigmoid激活函数有运算复杂，容易出现梯度消失的问题，这样对于LeNet来说，优化起来就很困难，而且LeNet的计算能力有限等，除了这些网络自身的问题，更重要的就是当时硬件条件的限制，当时没有GPU。而我们知道，深度神经网络最需要的就是大量的数据的训练，所以当时硬件的限制，对于卷积神经网络的发展，有很严重的限制。
9. LeNet-5之后，深度神经网络就陷入了很长时间的沉寂。06年Hinton提出了深度信念网络，深度学习开始重新发展起来，但是在当时，这样的方法并没有立即应用到图像处理中，而是在语音识别等问题上取得了重大突破。一直到2012年，在ImageNet 比赛， AlexNet横空出世。AlexNet开启了深度学习在图像分类问题上的应用，从ALexNet开始，深度学习开始真正的应用于图像分类，并且carry了整个图像分类领域。AlexNet是2012年，由Alex和他的导师Hinton在ImageNet比赛上提出来的，他们当时一下子将图像分类的top-5准确率由27%降到了15%，那么现在我们一起来学习一下AlexNet具体的网络结构，来看下相比于LeNet，到底为什么AlexNet会取得这么大的成功。这个图像看着像是只截了一半似的，其实Alex在论文中就是这样画的，整个网络结构分为上下两层是因为当时GPU即使是旗舰GTX580也只有3GB的显存，为了解决训练时间的问题，他们将网络分为上下两层，用两块GPU并行训练。AlexNet是我们需要重点学习的网络，因为他不仅仅是一个开端，同时他自身也有结构简单，分类性能好，准确率高等优点，因此在后续的很多分类的方法，尤其细粒度分类的方法中，都还有用到ALexNet，因此我们把AlexNet拿出来，重点学习。
10. 首先我们来看一下AlexNet的网络结构，Alexnet主要包含卷积层，激活函数层，Pooling层，全连接层，最后一层全连接层是softmax层。其中包含五个卷积层，三个全连接层，共8层。关于卷积神经网络层数的定义，ReLU激活函数层和池化层我们是不计的，一般只数有多少层卷积层，多少层全连接层。接下来我们一起看一下AlexNet的改进和创新。
11. AlexNet首先做了一步图像预处理，叫做数据增强，或者是数据扩张。数据增强或者说数据扩张，顾名思义就是将原来的数据给变多，虽然这样变换得来的数据之间仍然还有一些联系，但是这样依然可以对数据起到一定程度的增强作用。首先，AlexNet用到的第一种方法是对图像进行裁剪，这样在一张大的图片上，可以裁剪出不同的小尺寸的图片，从而可以实现数据的增强。第二种方法是镜像。通过对图片的镜像操作，也可是实现数据增强。除了这两个方法，AlexNet还改变RGB的频谱密度，采用主成分分析的方法，让主成分翻倍，这样也可以实现数据增强。
12. 首先我们先看左边的网络结构图，通过这个图，我们可以看到每一层的参数，比如第一层卷积层，他的卷积核个数是96个11\*11的卷积核，卷积的步长为4，卷积层参数的设置都有所改进。和LeNet不同的是，在卷积之后，我们直接用了ReLU激活函数，ReLU激活函数的使用是ALexNet最重要的创新点，对Alexnet性能的提升起到了至关重要的作用，关于ReLU激活函数的优点，前面已经仔细讲过了，在这我就不再赘述了。
13. AlexNet第三个创新点是增加了LRN结构。LRN全称local response normalization，叫做局部响应归一化。通过实验发现，在ALexNet中添加LRN单元，可以使top-1error和top-5error分别降低0.4%和0.3%。
14. LRN主要是应用了生物脑神经中的“侧抑制”的原理，侧抑制是指相邻神经元之间相互抑制，这样的原理用在神经网络中，可以让神经网络有更好的泛华能力。**泛化能力**是指机器学习算法对新鲜样本的适应能力。学习的目的是学到隐含在数据对背后的规律，对具有同一规律的学习集以外的数据，经过训练的网络也能给出合适的输出，该能力称为泛化能力。在这个部分，论文中说的比较含糊，具体参数的物理意义也没有明确得给出，但是从这个公式中，我们可以看到，对于输入，他的输出和他周围神经元大小的平方是成反比的，这样周围神经元就对它起到了抑制的作用。其实通过这些结构的设计我们可以看到，人工神经网络中很多的单元都是模仿了生物脑神经的功能，这样的设计也使得我们的人工神经网络更加接近生物大脑的功能，更加的智能。
15. AlexNet的第四个创新点在于他对Pooling方式的改进。
16. ALexNet也使用的是Max Pooling的方式，但是它的Max Pooling采用了一种重叠池化的方式。AlexNet中，池化核大小设置为3\*3，步长设置为2，像我们红框的位置，第一次maxPooling我们得到一个池化后的数值3，那么步长为2的时候，向后移两个框，这样我们相邻窗口之间会有重叠区域，这种池化方式就被是重叠池化，通过设置卷积核的大小和步长就可以实现。通过改进池化方式可以降低top-1 和top-5的错误率。
17. AlexNet另一个很大的贡献在于，它加入了Dropout层。在卷积神经网络中，通常都会有全连接层，而且卷积神经网络中的全连接层维度一般很高，比如在AlexNet中，FC6层和FC7层各有4096个节点，参数很多，这样就容易造成过拟合，而且参数过多，也是我们不想看到的。针对这些问题，AlexNet提出了Dropout机制。Dropout随机的将50%的节点的数值设置为0，这些被置0的神经元不再参与前馈和反馈的过程，相当于是被扔掉了。这样的方式就可以使得传递到下一层的节点数减半，从而减少参数。实验证明，dropout层可以有效地避免严重过拟合的情况，而且由于这种随机失活机制，每个神经元都不能依赖于其他某个特定神经元的表现，这样就强迫神经元去学习更具鲁棒性的特征。这对提高神经网络的性能来说非常重要。除了这些创新点之外，AlexNet还实现了多GPU并行训练。由于我们深度神经网络可以处理的样本数量很多，像ImageNet数据集有100多万张图片，这么多的数据就对硬件提出了一定的要求。单个GPU的处理速度不够快，会造成训练一个神经网络可能需要花费几天时间的情况，这样会耗费大量的时间，因此多GPU并行处理，可以节省我们很多的时间。现在我们神经网络的训练也都采用多GPU并行训练的方式。
18. AlexNet是卷积神经网络中的经典中的经典，在AlexNet之后，出现了很多其他卷积神经网路。接下来我们来看一下14年ImageNet比赛的亚军VGGNet。VGGnet的框架和AlexNet的框架非常像，但是VGGnet将AlexNet的5个卷积层发展成了五个卷积组，并分别给出了A-E五个不同的配置，网络层数从11-16递增。其中我们更常用的是D和E网络，因为他们各自是16层和19层，所以我们通常称他们为VGG16和VGG19.VGG在AlexNet基础上有什么改进呢？首先是网络层数增多，总层数增加到了19层。其次，VGG删掉了AlexNet中引以为傲的LRN层，删掉的原因就是他们进行了实验，发现在他们的网络上，没有LRN层的话，分类的准确率可以更高。其实后期很多的卷积神经网络经过测试发现，LRN层并不能改善他们网络的性能，因此后期的卷积神经网络中，都不再使用LRN层。VGG更大的一点改进是，他将卷积核和步长都变小了。在AlexNet中有11\*11的卷积核，也有步长设置为4，还有5\*5的卷积核等，在VGG中，卷积核的大小都是3，步长都是1。
19. 我们在选择卷积核的时候，需要考虑卷积核的大小，大的卷积核可以带来更大的感受野，但是这也意味着会有更多的参数。所以就会考虑：如何才能用参数少的小的卷积核获得更大的感受野？在VGGnet中提出了两个3\*3的卷积核可以替代一个5\*5卷积核的理论。这时候，我们可能会有一个疑问：最小的卷积核不是1吗？2也比三小啊，怎么选3呢？这个问题，在VGG论文中有解释，因为3\*3是可以捕获到上下左右中全部信息的最小的尺寸，所以选择3\*3的卷积核来做。我们看左边这幅图，一个5\*5的卷积核会有一个5\*5的感受野，通过卷积计算，得到一个输出值，那我们再看一下两个3\*3的卷积核是如何代替一个5\*5的卷积核的。同样在一个5\*5的感受野上，在步长为1的情况下，一个3\*3的卷积核首先进行卷积，这样我们可以得到一个初步的3\*3的特征图，接下来，再用一个3\*3的卷积核卷积，这样我们就可以得到一个输出的特征值。通过这种方式，两个3\*3的卷积核就可以代替一个5\*5的卷积核。那我们为什么要用2个3\*3的卷积核代替一个5\*5的卷积核呢？首先，在进入深度神经网络模块，我们就开始强调参数的数量，参数的数量不能过多，那么两个3\*3的卷积核的参数数量要少于一个5\*5的卷积核，这样可以有效地帮我们控制参数的数量。其次，两个3\*3的卷积核在步长为1的情况下可以替代一个5\*5的卷积核，三个3\*3的卷积＆可以替代一个7\*7的卷积核。除此之外，这样的设置会增加卷积层和ReLU层的层数，这样，多个小卷积核的卷积层比一个大卷积核的卷积层具有更强的非线性。
20. 但是这样结构的卷积神经网络依然存在许多问题。首先，通过实验我们发现，太深的网络更容易出现过拟合的现象，而且随着网络层数的加深，消耗的计算资源更多，而且越深的层越难训练，容易出现“训练不动”的情况。针对这些问题，卷积神经网络的结构有了进一步的改进。
21. 接下来让我们一起看一下2014年ImageNet比赛的冠军 GoogLeNet. GoogLeNet是谷歌团队提出的，有一句题外话，不知道大家知不知道，为什么谷歌net是把L写成大写。把L改成大写主要是为了向leNet致敬。GoogLeNet考虑到了卷积神经网络中参数多主要是因为最后两层全连接层参数多的问题，Szegedy吸取了教训，为了压缩GoogLeNet的网络参数，他把全连层取消了！GoogLeNet在最后一层用了average Pooling，虽然没有用全连接层，但是在GoogLeNet中依然使用了dropout层，现在取消全连层貌似是个大趋势，近两年的优秀大型神经网络都没有全连层，可能是全连层参数太多，网络深度增加了以后，难以接受吧。
22. GoogLeNet的另一个改进是引入了Inception结构。Ia结构是‘分叉-汇聚’网络，也就是说一层网络中存在多个不同尺度的kernels，分别卷积完之后再汇聚。IA结构能提高网络精度，可能就是归功于它拥有多个不同尺度的kernels，使kernel趋于小型化，多层化，而且是非对称的，每一个尺度的kernel会学习不同的特征，把这些不同kernels学习到的特征汇聚给下一层，能够更好的实现全方位的深度学习！
23. 另外，我们刚才提到过，随着网络层数的加深，深度神经网络出现了深层训练不动的情况，为解决这个问题，GoogLeNet加了多个softmax层，由于这个网络结构很大，看不太清楚，但是黄色框中的是softmax层，我们可以看到，GoogLeNet一共有三个softmax层，这是为了避免出现梯度消失的情况，网络额外增加了2个辅助的softmax用于向前传导梯度。

除了Alexnet, VGGnet, 以及GoogLeNet之外，还有很多其他的用于分类的卷积神经网络，像2015年获得第二次CVPR best paper的何恺明提出的ResNet，将卷积神经网络一下子做到了100多层，还有一些其他的可用于分类的卷积神经网络，在这就不过多介绍了。

1. 之所以讲这么多图像分类，卷积神经网络的内容，是因为在细粒度分类中，也沿用了卷积神经网络的方法，主要是用到了Alexnet和VGGNet，但是对卷积神经网络又有了新的改进。那接下来我们一起来看细粒度分类的相关内容。
2. 首先，我们来看一下什么是细粒度分类。细粒度分类是近年来的热门话题，细粒度分类关注的是子类别的分类，比如我们在分类问题中研究把不同类别的目标正确分类，在细粒度分类问题中，我们关注的是同一类别下的诸多子类别，以狗的种类为例，我们看哈士奇和阿拉斯加，这样的分类任务比普通分类的难度更大。细粒度分类问题难度大的原因主要是，他们有区别性的特征往往只存在于细小的局部。这样的特征就对特征的提取和处理方法提出了更高的要求。细粒度分类问题从提出到现在，已经经历了很长的时间，但是早期的基于人工特征的算法，由于特征提取和表达的能力有限，分类也面临着很大的局限性，因此，细粒度分类的问题几乎是没什么涉及。但是近年来，随着深度学习的兴起，我们找到了更为有效的深度卷积特征，这也促进了细粒度分类领域的发展，尤其是近几年视觉类的顶级会议CVPR，ICCV等，已经有很多论文是关于细粒度分类问题。细粒度分类问题也引起了越来越多的关注。
3. 首先，一起来看一下细粒度分类的应用。
4. 细粒度分类问题相对于分类问题来讲，有更大的挑战，首先细粒度分类的数据集较小，包含的训练样本的图像更少，对于深度学习来讲，大量的训练样本是必须的，那么样本数量过少会很大程度上限制细粒度分类效果的提升。其次，细粒度分类作为子类别的分类，他不同的类别之间的差异性较小，像我们刚才提到的哈士奇和阿拉斯加，类别与类别之间极小的差异会使得分类的难度加大。
5. 除此之外，类内的差异很大
6. 其次，姿态问题也是细粒度分类中常见的问题
7. 还有就是复杂的背景的问题
8. 因此，我们要再次强调局部信息在细粒度分类问题中的重要性, 包含足够区分度的信息往往只存在于很细小的局部区域中. 像这两只表的不同之处在于他们的嘴巴，还有腹部的颜色。因此,如何找到并有效利用这些有用的局部区域信息,成为了决定细粒度图像分类算法成功与否的关键所在. 那么通过后期算法的学习，我们也会体会得到，细粒度分类问题的重点就在于如何获取更有效的局部信息。
9. 无论是细粒度分类问题，还是分类问题，数据集都是至关重要的。细粒度分类问题的数据集和分类的数据集有相同之处，也有不同之处，接下来我们就一起看一下细粒度分类的数据集。首先介绍CUB-200-2011数据集，这个数据集是细粒度分类领域最常用的数据集，包含200类，共11788张图片，从这个图片量上，我们就可以感受得到细粒度分类问题的数据集大小的问题。仅包含1万多张图片的数据集在分类问题上已经没怎么在用了。从数据集的示例图片中，我们也可以看出来，鸟类的形态各异，不同的姿态，不同的背景等，分类的难度非常大。CUB数据集的人工标注信息，除了类别标签之外，还有局部信息，和boundingbox信息。
10. boundingbox的功能主要是框出前景图像，也就是找到鸟在整张图片的位置，减少背景的干扰。局部信息主要是标注出了几个重要的局部。CUB数据集是最常用的，同时也是分类难度最高的数据集。在后续我们讲到的很多算法都是在这个数据集上测试他们的性能的。
11. StandFord Dog数据集
12. UCEFOOD数据集
13. FGVC-Aircraft数据集
14. Cars
15. Oxford flower。
16. 那么从数据集中我们可以看到，很多数据集都提供了boundingbox信息或是局部的标注信息。细粒度分类问题，简单讲，可以理解成分类问题的加强版，和分类相同的是，细粒度分类也需要进行特征提取和分类。和分类不同的是，细粒度分类问题更关注的是细小的局部的特征，因为不同子类别之间，他们的区别主要在细小的局部。随着深度学习以及图像分类的发展，我们通过卷积神经网路提取到的卷积特征有效性已经很高，在细粒度分类问题中，很多方法关注的是如何定位更好的局部，如何对提取到的卷积特征做更优的处理，从而来提高细粒度分类的分类性能。那么在细粒度分类部分，卷积神经网络更多的被当成了特征提取器。或者说，几乎全部的细粒度分类的方法都是基于卷积神经网络做的，接下来，让我们一起了解一下不同的细粒度分类方法。
17. 传统分类模型17.3，AlexNet 61
18. 细粒度分类问题在Alexnet上表现的不好，一个重要的原因是细粒度分类数据集的图片数量过少，为增加图片数量…
19. 其次，为解决样本数量过少的问题，我们还有另外一个进行迁移学习的方法。迁移学习就是：顾名思义就是就是把已学训练好的模型参数迁移到新的模型来帮助新模型训练数据集。我们学习的时候，如果我们已经认识了狗……考虑到大部分数据或任务是存在相关性的，所以通过transfer learning我们可以将已经学到的parameter 分享给新模型从而加快并优化模型的学习不用像之前那样learn from zero.
20. 算法主要分为两种：强监督和弱监督。监督学习训练数据有标签，无监督是无标签。在细粒度分类问题中，又分为强监督和弱监督两种方式。强监督是…弱监督是……
21. 我们首先来看强监督分类方法，强监督分类方法中用到的人工标注信息一般是boundingbox和Part信息
22. 首先，我们先来看一下R-CNN网络结构。R-CNN全称是Region-CNN，R-CNN方法的实现主要包含以下几步：
23. 首先让我们来看一下Selective Search方法。Selective search算法的主要功能是在图像中确定一个物体，找到他的具体位置。Selective search意为选择性搜索，在选择性搜索之前，常用的方法主要是穷举搜索（Exhaustive search），就是我用一个窗口去扫描整张图像，扫面完之后，再把窗口调小点，再次扫描整张图像，用这种方法不断地试。这样的方法，十分耗时，而且得到的结果很杂。Selective Search的作者突破了穷举搜索的思维定式，找到了一种新的方法。接下来我们一起看一下，Selective Search是如何实现的。首先，Selective Search用了一种图像分割的方法来获得所有的区域，把这些区域都划分了出来，这样形成了很多过分割的区域，然后，计算所有区域之间的两两相似度。然后将两个相似度最高的区域合并。合并后呢，这个区域就作为一个区域，再来计算相似度，再进行合并，直到得到最终的候选区域。
24. 区域融合之后呢，在这幅图像中就是红框部分，同时呢，会有一个Ground truth图像，ground truth是由人标注的，真实的目标区域的位置，图中绿框部分。之前在分割问题中我们就讲到过，分割实际上可以看成是一个二分类问题，一类是确获得目标区域，也就是标注框刚好标注的是我们的目标区域，另一类是标注框标注的并不是正确的目标区域，我们把正确标注目标区域的称之为正样本，没有正确标注目标区域的称之为负样本，正样本呢，是我们ground truth值，那负样本怎么来的呢？负样本主要是通过对我们之前确定的候选区域的筛选得来的。负样本需要有两个特点：1. 并不是正确的区域，2. 不能重复。因此在筛选负样本的时候，首先是计算所有的候选区域和groundtruth之间的相似度。论文中将相似度控制在20%-50%，否则就被舍去。其次候选区域要和已经选出来的负样本比较，若候选区域和已经选出来的负样本之间的相似度超过70%，那么也将被舍去。相似度关注的是color，texture，size和fit四个方面的特征。在正样本和负样本都选好之后呢，这就变成了一个有类别标签的二分类的问题。我们用它来训练一个SVM分类器。一般如果概率大于0.5将被认为是目标,否则被认为是非目标,如果完全分类正确,所有的正样本的SVM输出概率都大于0.5,所有负样本的SVM输出概率都小于0.5,但是最常见的情况是有一部分的负样本的输入概率也是大于0.5的,我们会错误地将这样样本认为是目标,这些样本就称之为"False Positives".把这些"False Positives"收集起来,以刚才训练得到的SVM的权值作为其初始权值,对SVM进行二次训练,经过二次训练的SVM的分类准确度一般会有一定的提升;测试的过程基本和训练过程相同: 首先用SS方法得到测试图像上候选区域 --> 然后提取每个区域的特征向量 --> 送入已训练好的SVM进行软分类 --> 将这些区域按照概率值进行排序 --> 把概率值小于0.5的区域去除 --> 对那些概率值大于0.5的,计算每个区域与比它分数更高的区域之间的重叠程度,如果重叠程度大于30%,则把这个区域也去除了 --> 最后剩下的区域为目标区域.
25. 在获取了局部区域之后，我们要提取这些局部区域的特征，和很多后期的方法一样，R-CNN也是用卷积神经网络的卷几层来进行特征提取。之前说过了，每个卷积神经网络对输入的大小有严格的限制，把他们都warp成224\*224的大小。然后用在ImageNet数据集上训练好的AlexNet做特征提取器。来提取目标区域中的特征。最后，R-CNN采用的SVM做分类器来进行分类。
26. Part R-CNN就是利用R-CNN算法进行对象(鸟)与局部区域(头、身体等)的检测。同R-CNN一样, Part R-CNN也使用自底向上的区域算法(如Selective Search[47])来产生区域候选, 如图左上角所示. 之后, 利用R-CNN算法来对这些 区域候选进行检测, 给出评分分值. 在这里, Part R- CNN只检测前景对象(鸟)和两个局部区域(头、身 体). 之后, 根据评分分值(图4中间)挑选出区域检测 结果(见图上方中间). R-CNN给 出的评分分值并不能准确地反映出每个区域的好坏. 例如, 对于头部检测给出的标注框可能会在对象检 测的标注框外面, 身体检测的结果与头部检测的结 果可能会有重叠等. 这些现象都会影响最终的分类 性能. 因此, 需要对检测区域进行修正.因此，作者在这篇文章中提出了两个约束条件：边框约束: 所有的局部区域的范围不能超出对象区域，将阈值设置为10个像素点，也就是说如果局部区域的框框，超出了对象区域框框十个像素点，就会被舍弃。还有一个更强的几何约束：part相对于root应该是有一个“默认”的位置的（比如鸟头应该在上方等），在论文中提出了基于混合高斯模型和最近邻的两种几何约束方法。用如上所述的约束条件对R-CNN检测的位置信息进行修正之后, 再分别对每一块区域取卷积特征,将不同区域的特征相互连接起来,构成最后的特征表示, 用来训练SVM分类器.
27. 经过几何约束后图像是这样的，也就是说，几何约束的效果就是限定了局部相对整体的位置。
28. Part R-CNN就是利用R-CNN算法进行对象(鸟)与局部区域(头、身体等)的检测。同R-CNN一样, Part R-CNN也使用自底向上的区域算法(如Selective Search[47])来产生区域候选, 如图左上角所示. 之后, 利用R-CNN算法来对这些 区域候选进行检测, 给出评分分值. 在这里, Part R- CNN只检测前景对象(鸟)和两个局部区域(头、身 体). 之后, 根据评分分值(图4中间)挑选出区域检测 结果(见图上方中间). R-CNN给 出的评分分值并不能准确地反映出每个区域的好坏. 例如, 对于头部检测给出的标注框可能会在对象检 测的标注框外面, 身体检测的结果与头部检测的结 果可能会有重叠等. 这些现象都会影响最终的分类 性能. 因此, 需要对检测区域进行修正.因此，作者在这篇文章中提出了两个约束条件：边框约束: 所有的局部区域的范围不能超出对象区域，将阈值设置为10个像素点，也就是说如果局部区域的框框，超出了对象区域框框十个像素点，就会被舍弃。还有一个更强的几何约束：part相对于root应该是有一个“默认”的位置的（比如鸟头应该在上方等），在论文中提出了基于混合高斯模型和最近邻的两种几何约束方法。用如上所述的约束条件对R-CNN检测的位置信息进行修正之后, 再分别对每一块区域取卷积特征,将不同区域的特征相互连接起来,构成最后的特征表示, 用来训练SVM分类器.
29. Part R-CNN的进步是明显的. 从局部区域 的检测定位, 到特征的 取, 该算法均基于卷积神经 网络, 并针对细粒度图像的特点进行改进优化, 以改 进通用物体定位检测算法在该任务上的不足, 达到 了一个相对比较高的准确度. 同时, 该算法进一步放 松了对标记信息的依赖程度, 在测试时无需 供任 何标记信息, 大大增强了算法的实用性. 其不足之处 在于, 利用自底向上的区域产生方法, 会产生大量无 关区域, 这会在很大程度上影响算法的速度. 另一方 面, 该算法本身的创新性十分有限, 既然局部区域对 于细粒度图像而言是关键所在, 那么对其进行定位 检测则是必要的途径. 只是引入现有的通用定位算 法, 似乎并不能很好地解决该问题.
30. 要进行局部区域的对齐需要首先要解决如何检测局部区域的问题.对于输入图像,Branson等人利用预先训练好的DPM算法完成关键点的检测.之后,利用这些关键点来和原型进行姿态对齐操作，它在原有的局部区域模型的基础上, 进一步考虑 了鸟类的不同姿态的干扰, 减轻了类内方差造成的 影响, 从而取得了较好的性能表现.同时，文章还提出了不同网络层 取的特征包含不同的语义信 息, Branson等人认为应该针对不同的局部区域 取 不同网络层的卷积特征. 为了证明这一点, 他们比 较了不同的局部区域在各个网络层 取的特征所能 达到的分类准确度. 实验结果表明, 对于低层对齐图 像(原始图像与前景对象)而言, 后层的卷积特征更具 区分度, 能够实现更高的准确度, 相对浅层特征具有 绝对的优势. 但对于高层对齐图像(头部图像)来说, 情况却恰恰相反. 因此, 对于不同的局部区域应当 取不同网络层的特征.
31. Posenormal效果如下
32. Location网络由三个卷积层和两个全连接层组成，他输出的是区域框图的左上角和右下角。对齐网络提供进行了平移，缩放和旋转的操作，从而实现了姿态对齐。这一步操作对于提高分类正确率来讲，至关重要。除了对齐操作，同时这一步也是做反向传播时，整个LAC网络的关键部分，通过他可以在反向传播时调整对齐以及定位的结果。
33. Part-stack CNN 采用了FCN来提取局部特征，two-stream的分类网络来分别处理全局特征和局部特征，
34. Coarse-to-fine就是从粗到细的特征，他们认为最后层的卷积层，更多的对应的是全局特征而忽略了局部，因此最后一层卷积层所表达的特征对传统特征分类有效，但是对细粒度分类的有效性不够。因此，他们设置第五层卷积层和前面的第2，3，4层卷积层都交叉相关，这样的设置可以使得整个卷积神经网络不仅仅关注全局信息，也会关注局部信息。
35. Mix DCNN方法，训练了K个CNN网络,图片会被输入到这K个网络中，这样，我们每一个网络都会有一个分类结果，在整合所有分类结果的时候，他并不是所有的分类结果求平均，而是采用了注意力机制：Ck是最佳的分类结果，Cc是每一类的分类结果，这样，对于分类效果更好的子网络，我们可以给他更高的权重，这个网络中，用softmax进行分类。还有值得一提的是，即使是有K个网络，但是他们都可以共用一个标签，所有K个网络可以一起进行端到端训练。
36. 几种强监督算法在CUB数据集上的分类效果入下，效果最好的是……准确率依然不高。
37. 仅仅依赖于类别标签完成分类是近年来细粒度 图像研究的一大趋势.得益于深度学习的发展, 以 及相关研究工作的深入, 不借助人工标注信息, 也 能实现良好的分类性能从前文的讨论中可以看出, 对于细粒度图像分 类算法而言, 局部区域信息是至关重要的, 这也正是 大多数算法依赖于标注信息的一大原因. 因此, 要实 现更好的弱监督的细粒度图像分类, 首先要解决的 就是如何检测并定位这些局部区域.
38. 两级注意力(Two Level Attention)算法[56]是第一 个尝试不依赖额外的标注信息, 而仅仅使用类别标 签来完成细粒度图像分类的工作。顾名思义, 该模型主要关注 两个不同层次的特征, 分别是对象级(Object-Level)和 局部级(Part-Level),与R-CNN相类似，使用一个卷积网络来对Selective Search产生的所有区域候选进行筛选, 检测该区域的图像中是否包含鸟类.
39. 这张图应该是论文的精髓了，网络跟AlexNet是一样的，选择第四个卷集成聚类成3类，每个聚类看作是一个部分检测器。对于每一聚类，选择生成score最高的patch，把这个patch当做part patch。每张图应该是有聚类个数3个part patch。（就相当于另外一次对原始selective search产生的region proposal进行删选，这次一张图片只保留3个最有影响力的patch，
40. 这张图应该是论文的精髓了，网络跟AlexNet是一样的，选择第四个卷集成聚类成3类，每个聚类看作是一个部分检测器。对于每一聚类，选择生成score最高的patch，把这个patch当做part patch。每张图应该是有聚类个数3个part patch。（就相当于另外一次对原始selective search产生的region proposal进行删选，这次一张图片只保留3个最有影响力的patch，
41. 把一个在ILSVRC2012上面预训练过的CNN转变成一个filterNet。FilterNet可以选择跟基准类别很相近的patches，因此可以处理object-level attention。这个选择到的patches用来训练另外一个CNN，训练成一个domain分类器，称DomainNet。
42. 将不同局部区域的特征级联成一个特征向量, 用来训练SVM, 作为局部级模型给出的分类器. 最 后, 将对象级模型的预测结果与局部级模型的结果 相结合, 作为模型的最终输出, 达到了69.7%的精度. 需要说明的是, 这是在Alex-Net[22]上的实验结果, 如 果采用更强大的网络结构如VGG-Net[39], 则能将分 类准确率 升到77.9%. 这也从另一个角度说明了特 征对于图像分类算法的重要性.总体上来看, 两级注意力模型较好地解决了在 只有类别标签的情况下, 如何对局部区域进行检测 的问题. 但是, 利用聚类算法所得到的局部区域, 准 确度十分有限. 在同样使用Alex Net的情况下, 其分 类精度要低于强监督的Part R-CNN算法.
43. 都只是把卷积网 络当做一个特征 取器, 各个步骤之间的处理仍然是一个分散的过程, 并未从整体上进行端到端(end- to-end)的训练优化.B-CNN完全实现了端到端。
44. 一个双线性模型B由一个四元组 组成: B = (fA,fB,P,C). 其中, fA,fB代表特征 取函数, 即图7中的网络A、网络B, P 是一个池化函数(pooling function), C则是分类函数.
45. 网络A的作用 是对物体进行定位, 即完成传统算法的对象与局部 区域检测工作, 而网络B则是用来对网络A检测到的 物体位置进行特征 取. 两个网络相互协调作用, 完 成细粒度图像分类过程中两个最重要的任务: 区域 检测与特征 取.
46. B-CNN的两个网络分别是……
47. B-CNN 的创新点
48. B-CNN在三个数据集上进行测试
49. B-CNN 分类结果，有无fine-tune，不同的网络组合
50. Constellations：2015 ICCV

正如我们在上一节曾经 到的, 基于Selective Search 产生区域候选的方法, 尽管有效, 却面临巨 大的计算代价和资源浪费. 在这篇文章中他们尝试采用其他方式来产生足够的局部区域.他们利用卷积网络特征产生一些关键点, 并基于这些关键点来 取局部区域信 息. 通过对卷积特征进行可视化分析, Simon等人发 现响应比较强烈的区域往往对应于原图中一些潜在 的局部区域点.至于前景对象, Simon等人并未 出更好的解决 方案, 他们采用的仍是传统的局部区域候选的方法, 即利用Selective Search[47]产生候选区域, 再利用卷积 神经网络对其进行分类, 取置信度最高的区域作为 前景对象.

最后的特征向量由三部分信息构成: 原图 的特征、前景对象的特征以及局部区域的特征. 在训 练时, 对VGG-Net[39]进行了微调, 并将所有训练数据 进行水平翻转, 用来进行数据增强, 最终结果为81%。

1. ST: 在前馈过程中，把空间转换网络应用到特征图上，对局部特征图的角度进行调整，上图是应用两个空间转换网络部分，关注的part只有头部和身体，如果使用4个空间转换网路，可以关注四个局部，这样会一定程度上提高我们的分类准确率。
2. CNN-tree在分类问题上，我们经常会遇到像左图这样的情况：有一些样本更容易错分到其他类别中，我们把这些样本称为干扰项，在CNN-tree提出的方法中，他用到了树状的结构，在第一个卷积神经网络训练完成后，我们将错误分类的干扰项，全部收集起来，再将这些所有的干扰项打成不同的包，再以这些数据为一个节点，继续进行分类，这样循环
3. DVAN 算法 网络包含四个部分：注意局部获取，卷积特征提取，多元化注意力，分类。

在获取多个局部区域之后，将他们全部输入到卷积神经网络中来提取卷积特征，这里是用VGG来提取卷积特征。根据提取到的卷积特征。DVAN网络的改进就在于多元化注意力机制，它不是仅关注一个最重要的区域，而是同时识别多个区域，用LSTM递归单元来学习这些所有局部的注意或是歧视，最终的预测结果是所有预测结果的平均值。在这个网络中，用到了LSTM单元，LSTM并不是卷积神经网络，它是递归神经网络的一种。递归神经网络在分类问题上应用的极少，但是在细粒度分类问题上，注意力机制对于细粒度分类问题比较重要，因此有几种方法都用到了递归神经网络。

1. 弱监督算法在CUB数据集上的结果