摘要：

大型卷积网络模型最近在ImageNet基准Krizhevsky等人证明了令人印象深刻的分类性能。[18]。 然而，对于他们为什么表现如此出色以及他们如何改进这一点，并不清楚。 在本文中，我们探讨了这两个问题。 我们介绍一种新颖的可视化技术，可以深入了解中间要素图层的功能和分类器的操作。 用于诊断角色，这些可视化使我们能够找到胜过Krizhevsky等人的模型架构。 在ImageNet分类基准上。 我们还进行消融研究，以发现不同模型层的性能贡献。 我们展示了我们的ImageNet模型很好地推广到其他数据集：当softmax分类器被重新训练时，它令人信服地击败了Caltech-101和Caltech-256数据集的当前最先进的结果。

1.介绍

自LeCun等人介绍他们以来， 20世纪90年代初期，卷积网络（convnets）在手写数字分类和人脸检测等任务中表现出色。在过去的18个月里，一些论文表明，他们也可以在更具挑战性的视觉分类任务中表现出色。 [4]证明了最先进的NORBandCIFAR-10数据库的性能。最值得注意的是，Krizhevsky等人[18]在ImageNet 2012分类基准上表现出创纪录的表现，其中convnet模型的错误率为16.4％，而第二名的结果为26.1％。继此之后，Girshick等人[10]显示了PASCAL VOC数据集的领先检测性能。有几个因素造成了这种显着的性能改善：（i）可用更大的训练集，数百万个标记的例子; （ii）强大的GPU实现，使大型模型的训练更加实用，（iii）更好的模型正则化策略，如Dropout [14]。

尽管取得了令人鼓舞的进展，但对于这些复杂模型的内部运作和行为，或者如何取得如此好的表现，仍然没有多少见识。 从科学的角度来看，这是非常令人不满的。 如果没有清楚的理解他们如何和为什么他们的工作，更好的模型的发展被减少到反复试验。 在本文中，我们介绍一种可视化技术，揭示在模型中的任何层激发单个特征映射的输入激励。 它还使我们能够观察训练过程中的特征演变，并用模型诊断潜在的问题。 我们提出的可视化技术使用Zeiler等人提出的多层去卷积网络（deconvnet）。 [29]，将特征激活投影回输入像素空间。 我们还通过遮挡输入图像的一部分来对分类器输出进行灵敏度分析，揭示场景的哪些部分对于分类很重要。

使用这些工具，我们从Krizhevsky等人的体系结构开始。 [18]并探索不同的架构，发现在ImageNet上表现优于其结果的架构。 然后我们探索模型对其他数据集的泛化能力，只需对softmax分类器进行再训练。 因此，这是一种有监督的预训练形式，与Hinton等人推广的无监督预训练方法形成鲜明对比。 [13]等[1,26]。

相关工作：

可视化：

可视化特征以获得关于网络的直觉是常见的做法，但大多局限于第一层，其中像素空间的投影是可能的。在更高层中，必须使用替代方法。 [8]通过在图像空间中执行梯度下降以最大化单元的激活来找出每个单元的最佳刺激。这需要仔细的初始化，并没有提供有关单位不变的任何信息。受到后者短促的影响，[19]（通过[2]扩展了一个想法）显示了一个给定单位的Hessian可以如何在最佳响应周围进行数值计算，从而对不变性有一定的了解。问题在于对于更高层次而言，不变性是非常复杂的，所以通过简单的二次近似很难被捕获。相比之下，我们的方法提供了不变性的非参数化视图，显示来自训练集的哪些模式激活特征映射。我们的方法与Simonyan等人的当代工作类似。他们演示了如何通过从网络的完全连接层反射回来从一个小圆点获得显着性图，而不是我们使用的卷积特征。 Girshick等人[10]显示可视化，识别数据集内的补丁，负责在模型中较高层的强烈激活。我们的可视化效果不同，因为它们不仅仅是输入图像的作物，而是自上而下的投影，揭示每个补丁内的结构，以刺激特定的特征图。

特征泛化：

我们在Donahue等人的同时工作中也探讨了convnet特征的泛化能力。 [7]和Girshick等。[10]。 他们使用convnet特性在前一种情况下在Caltech-101和Sun场景数据集上获得最新的性能，在后者中使用PASCAL VOC数据集上的对象检测。

途径：

在整个论文中，我们使用标准的完全监督的模型，正如LeCun等人所定义的那样。 [20]和Krizhevsky等。[18]。 这些模型将彩色2D输入图像xi经由一系列层映射到C个不同类别上的概率矢量yi。 每一层由（i）前一层输出（或者在第一层的情况下，输入图像）与一组学习滤波器的卷积组成; （ii）通过整数线性函数（relu（x）= max（x，0））传递响应; （iii）[可选地]在本地邻居上的最大汇集，以及（iv）[可选地]局部对比度操作，其在特征地图上对响应进行归一化。 有关这些操作的更多详细信息，请参见[18]和[16]。 网络的前几层是传统的全连接网络，最后一层是softmax分类器。 图3显示了我们许多实验中使用的模型。

我们使用大量的N个标记图像{x，y}来训练这些模型，其中labely是一个离散变量，表示真实的类别。 交叉熵损失函数适用于图像分类，用于比较yi和yi。 网络参数（卷积层中的滤波器，完全连接层中的权重矩阵和偏差）通过相对于整个网络中的参数向后传播损耗的导数并通过随机梯度下降来更新参数来训练。 第三部分详细介绍了培训内容。

用卷积进行可视化

理解一个convnet的操作需要解释中间层的特征活动。我们提出了一种将这些活动映射回输入像素空间的新方法，显示了最初在特征映射中引起给定激活的输入模式。我们用去卷积网络（deconvnet）Zeiler等人进行这种映射。 [29]。一个deconvnet可以被认为是一个使用相同的组件（过滤，合并）但是相反的convnet模型，所以不是将像素映射到要素而是相反。在Zeiler等人[29]，deconvnets被提议作为执行无监督学习的一种方式。在这里，他们没有被用于任何学习的能力，只是作为一个已经训练的转折点的探索。为了检查一个定位点，在每个图层上都附加一个去卷积，如图1（顶部）所示，为图像像素提供一条连续的路径。首先，将输入图像呈现给convnet，并在整个图层中计算特征。为了检查给定的定位点激活，我们将图层中的所有其他激活设置为零，并将该功能图作为输入传递给附加的解卷层。然后我们先后（ⅰ）解开，（ⅱ）整顿，（ⅲ）过滤重建在下面层的活动，导致选择的激活。然后重复这个操作直到输入像素空间到达。

非池化

：在最小化池中，最大池操作是不可逆的，但是我们可以通过在一组开关变量中记录每个池区中的最大值的位置来获得近似的逆。 在deconvnet中，解开操作使用这些开关将来自上层的重建放置到合适的位置，从而保持刺激的结构。 参见图1（下图），了解该过程。

矫正：

使用relu非线性的convnet，纠正特征映射从而保证特征映射总是正面的。 为了在每一层获得有效的特征重建（也应该是正的），我们将重构的信号通过非线性非线性1。

过滤：

使用已知滤波器来对前一层的特征映射进行卷积。 为了近似地反转这个，deconvnet使用相同过滤器的转置版本（和其他自动编码器模型，如RBM），但是应用于整形图，而不是下面图层的输出。 在实践中，这意味着垂直和水平方向各个过滤器。

请注意，在此重建路径中，我们不使用任何对比度归一化操作。从较高层向下投影使用由convnet中的最大池所产生的交换设置。由于这些开关设置对于给定的输入图像是独特的，因此从单个激活获得的重建因此类似于原始输入图像的一小块，结构根据它们对特征激活的贡献而加权。由于模型是有区别地训练的，它们隐含地显示输入图像的哪些部分是有区别的。请注意，这些预测不是来自模型的样本，因为不涉及生成过程。整个过程类似于反向传播一个强烈的激活（而不是通常的梯度），即计算∂h∂Xn，其中h是强激活特征映射的元素，Xn是输入图像。但是，（i）relu是独立执行的，（ii）没有使用正常化操作。我们的方法的一个普遍缺点是它只能看到一个单一的激活，而不是一层中的联合活动。然而，正如我们在图6中所显示的那样，这些可视化是对模型中的给定特征映射进行刺激的输入模式的精确表示：当对应于该模式的原始输入图像的部分被遮挡时，在特征地图内的活动。