1 任务要求

任务要求：针对已训练好的卷积神经网络，给定一张输入图片，生成该图片对于特定类别的可解释性分析结果。
给出每张输入图片在最后一层卷积层输出的的可视化结果（对输出特征图的每一个通道进行可视化），每张图片分别针对猫和狗两个类别的可解释性分析结果（Grad-CAM 及 LayerCAM），以及对应的实验分析。

2 任务设计

2.1 模型准备

使用载入已给模型。

```
1.  model = torch.load('./experiment4_data/torch_alex.pth')
```

2.2 可视化任务

对于最后一层卷积层输出特征图可视化，特征图上的值应该是越大越重要，负值应该忽略，因为经过 ReLu 或者 Sigmoid 函数之后，负值的影响力变得很小。所以，先将特征图去掉负值，然后使用

```
1.  featuremap = cv2.applyColorMap(np.uint8(255 * mask), colormap)
```

将特征图转化为彩色图，然后透明覆盖在原图上，进行观察。
对于 grad-cam 自己实现了一个简易版本，对于 layer-cam 则直接使用了现有库函数。

3 实验

本实验对可视化结果图像进行分析，试图找到神经网络可解释的理由。

3.1 通道特征图可视化

首先对 dog.jpg 进行可视化，一共 256 个通道，16 个图放一行，结果如图：

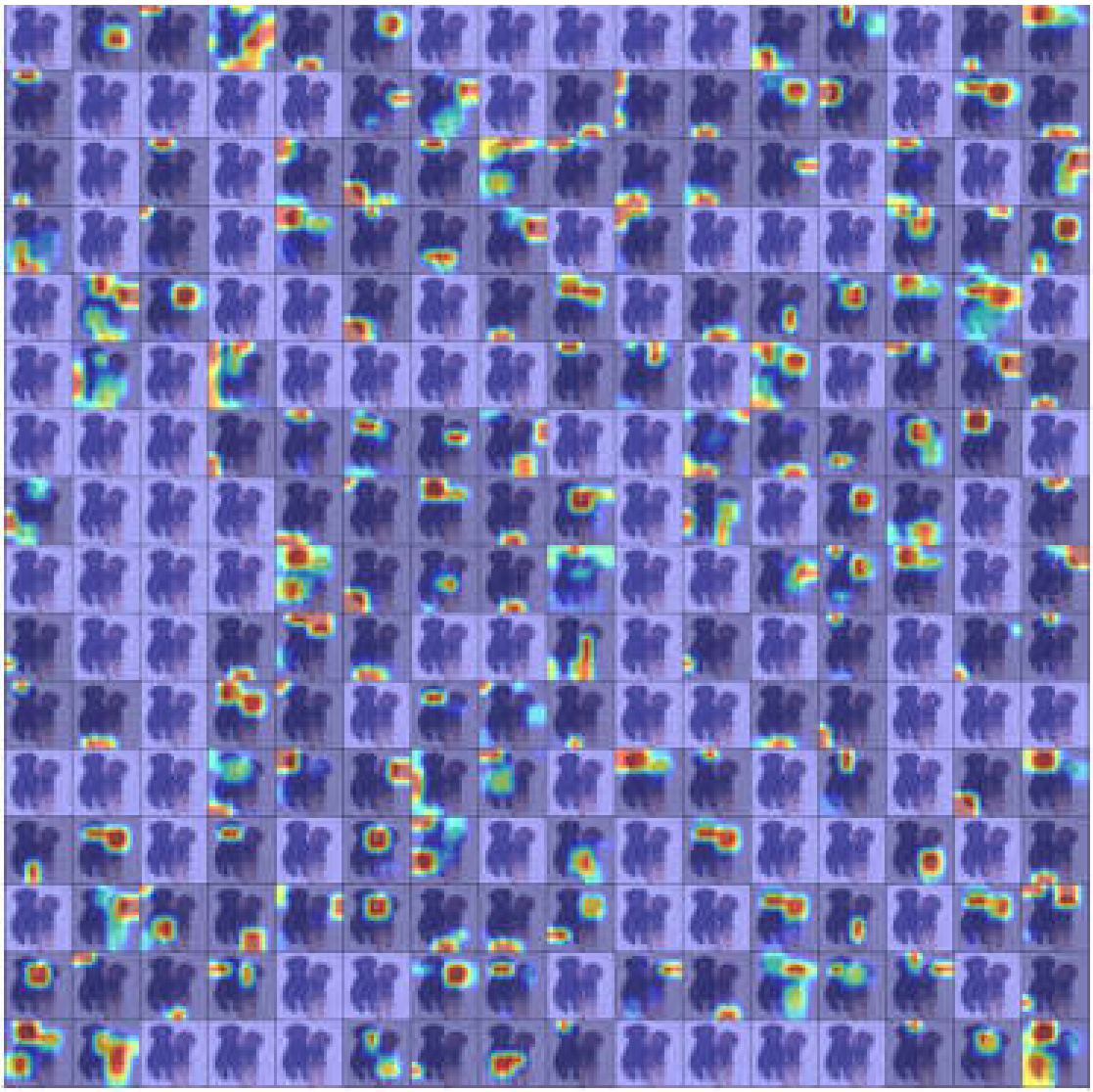


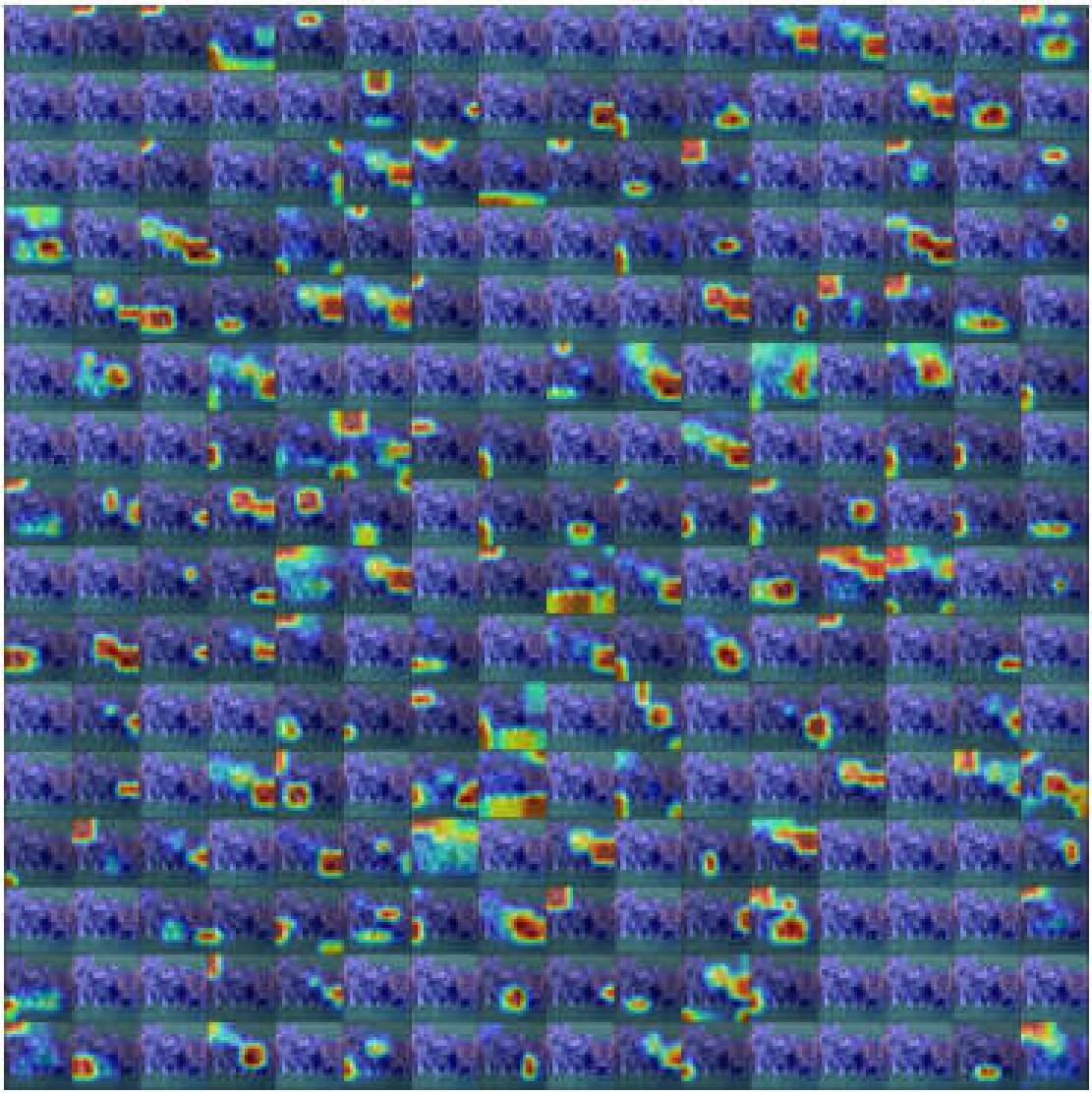
图 1 dog 特征图可视化

顺便给出 model 预测结果：[2.9665e-11, 1.0000e+00]，结果对 dog 的预测值很准确。从特征图可视化中可以看到，数值大的比较集中在 dog 的身体部位上，比如狗头，狗腿，狗的身体等，基本上很少或没有出现空白区域值很大的情况。

下面对 cat：

模型预测结果：[1.0000e+00, 3.3226e-10]

可视化图：

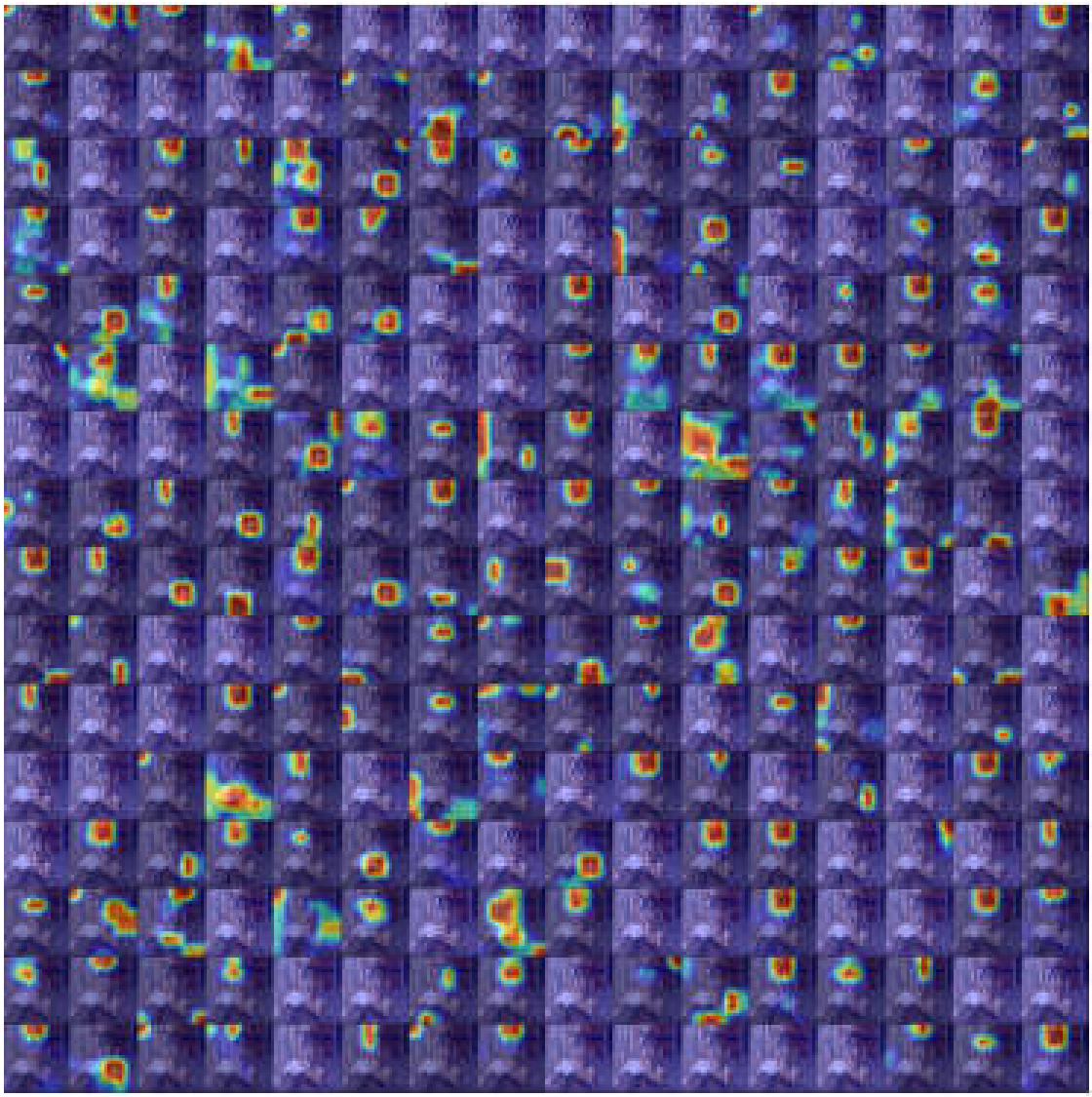


猫可视化，值大的地方差不多也是猫出现的位置，但是也出现较多其他位置值大，预测结果不是的概率也比 dog 的要大一个数量级。

对猫狗的

预测结果：[2.5811e-12, 1.0000e+00]

特征图可视化：



model 将此图识别成了 dog，没有识别出 cat。观察一下特征图，绝大数值出现在了 dog 的区域，很少出现在 cat 的区域。可能这是其中的一个原因。

3.2 grad-cam 分析

一个很浅显的道理，在一个函数中，导数越大的变量对输出结果的影响越大。grad-cam 利用这个原理，把预测类别结果对每一张特征图元素的平均导数作为该特征图的权重，然后将所有特征图乘各自的权重之后叠加，形成热力图。

实验仅针对最后一层卷积层进行可视化。

对 dog：

图 2 dog grad-cacm

对 cat：

图 3 cat grad-cam

对于 both：

图 4 both grad-cam dog

图 5 both grad-cam cat

可以看到，单独而言，对猫和狗热力图均集中在头部，说明该模型是可信的，因为其基于最特别的特征做出了判断。对于 both 这张图片，猫和狗也识别到了正确的区域，但是输出结果为 dog，由图也知，狗头的热力图更明显，或许这是 model 将其识别为 dog 的原因。

3.2 layer-cam 分析

对最后一层卷积层进行 layer-cam 分析。layer-cam 较之 grad-cam，它不采用平均梯度值作为特征图的权重，而是每个元素都有一个权重，该权重就是分类结果对该元素的导数。然后仍是计算出每个通道的特征图后叠加，形成热力图。由于每个元素都有一个权重，其热力图应该更为精细化。

对 dog：

图 6 dog layer-cam

对 cat：

图 7 cat layer-cam

对于 both：

图 8 both layer-cam dog

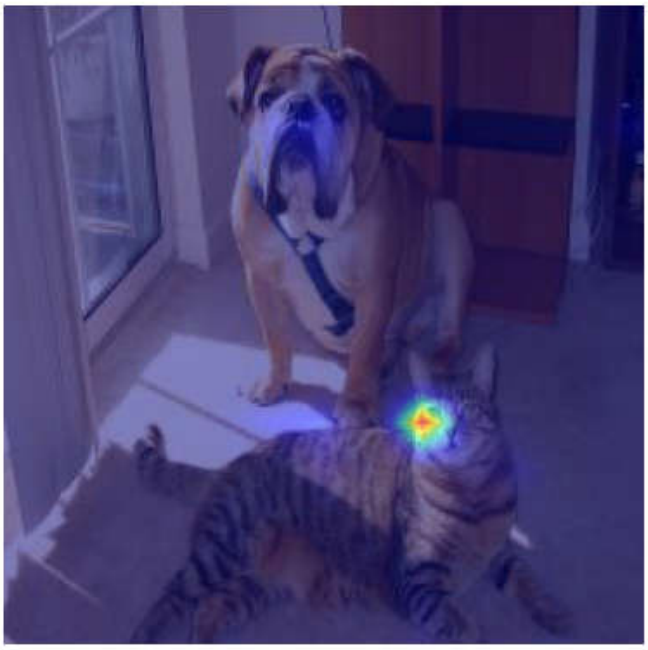


图 9 both layer-cam cat

可以看到对于猫和狗都是其头部对结果影响较大。所以神经网络做出类别的预测应该是可信的。

4 结论

神经网络的过程一直被戏称为“炼丹”，通过上面的可视化实验证明事实并不是如此。实验可知，特征值大的地方，梯度大的地方都集中在了特征显著的区域。这些区域就是 model 做出判断的依据，所以有了这些可视化的可解释性过程，我们可以判断该 model 是否可以被信任。