華中科技大學

计 算 机 视 觉 课 程 报 告

题目： 卷积神经网络可解释性分析

学　　号　U202115638

姓　　名　马耀辉

专　　业　数据科学与大数据技术

班　　级　大数据 2101 班

指 导 教 师　杨卫

计算机科学与技术学院

目　录

一　实验要求

卷积神经网络(Convolutional Neural Network, CNN)在计算机视觉领域取得了显著的成就。然而,CNN 的高效性也使得其内部决策过程难以解释,这给其在一些关键应用领域的可信度带来了挑战。为了增强对 CNN 模型的解释性,一种被广泛研究的方法是通过可解释性分析来理解网络在特定输入图像上的决策依据。

本实验旨在针对已经训练好的卷积神经网络,在给定一张输入图像的情况下,生成该图像对于特定类别的可解释性分析结果。实验提供了基于 PyTorch 和 TensorFlow 两个不同深度学习框架的二分类模型,其中 PyTorch 使用的网络架构是 AlexNet,而 TensorFlow 使用的是 VGG16。

在实验中,我们将使用三张输入图像,并针对每张图像的猫和狗两个类别进行可解释性分析。我们将采用两种常用的可解释性方法,即 Grad-CAM 和 LayerCAM,对最后一层卷积层的输出特征图进行可视化。通过分析这些可解释性分析结果,我们可以更好地理解网络对不同类别的判别依据,进而提高对模型决策的可解释性和可信度。

二　实验内容

2.1　模型结构

我们使用的模型是 VGG16(VGGNet) 网络, 它是一个经典的卷积神经网络架构, 在计算机视觉任务中具有广泛的应用。

下面是该模型的结构和参数信息:

1. 输入形状(Input shape): (None, 224, 224, 3), 表示输入图像的尺寸为 224x224, 通道数为 3(RGB 颜色通道)。
2. 输出形状(Output shape): (None, 2), 表示模型的输出是一个 2 类分类问题, 输出为 2 维向量。
3. 模型结构:
    - vgg16: VGG16 网络的主体部分, 包含多个卷积层和池化层, 用于提取图像特征。这部分网络的输出形状为 (None, 7, 7, 512), 表示经过卷积和池化操作后, 特征图的尺寸为 7x7, 通道数为 512。
    - classifier: VGG16 网络的分类器部分, 包含多个全连接层和激活函数, 用于将提取的特征映射到最终的输出类别。这部分网络的输出形状为 (None, 2), 表示模型的输出是一个 2 维向量, 对应于两个类别的概率分布。
4. 模型参数:
    - 总参数数量(Total params): 134,268,738, 表示模型的总参数数量为 134,268,738 个, 约占 512.19 MB 的内存空间。
    - 可训练参数数量(Trainable params): 119,554,050, 表示模型中可以通过训练进行学习和更新的参数数量为 119,554,050 个, 约占 456.06 MB 的内存空间。
    - 不可训练参数数量(Non-trainable params): 14,714,688, 表示模型中不参与训练的固定参数数量为 14,714,688 个, 约占 56.13 MB 的内存空间。
    - 损失函数(Loss function): 我们使用的损失函数是 Sparse Categorical Crossentropy, 这是一种常用于多类分类问题的损失函数。
    - 优化器(Optimizer): 我们使用的优化器是 Adam optimizer, 它是一种常用的梯度下降优化算法, 用于更新模型参数以最小化损失函数。

模型层级结构(Layers): VGG16 网络由两个主要部分组成: vgg16 和 classifier。vgg16 部分包含了卷积层和池化层, 用于提取图像特征。classifier 部分包含了全连接层和激活函

数,用于进行分类。具体的层级结构如下所示:

1. vgg16:
    - input_6
    - block1_conv1
    - block1_conv2
    - block1_pool
    - block2_conv1
    - block2_conv2
    - block2_pool
    - block3_conv1
    - block3_conv2
    - block3_conv3
    - block3_pool
    - block4_conv1
    - block4_conv2
    - block4_conv3
    - block4_pool
    - block5_conv1
    - block5_conv2
    - block5_conv3
    - block5_pool
2. classifier
    - flatten_5
    - dropout_10
    - dense_15
    - dropout_11
    - dense_16
    - dense_17

以上是关于我们使用的 VGG16 模型的详细介绍。在接下来的实验中,我们将使用该模型进行可解释性分析,并观察其在猫和狗分类任务上的表现。

2.2 CNN 可解释性分析

2.2.1 Grad-CAM

Grad-CAM 是一种技术,用于可视化卷积神经网络在进行特定类别的预测时关注的图像区域。它的主要思想是利用梯度信息突出显示对模型决策最重要的区域。

工作原理:

1. 选择层: Grad-CAM 通常在最后一个卷积层上应用, 因为这个层保留了空间信息(即图像中不同区域的位置信息),同时也捕获了高层次特征。
2. 计算梯度:对于给定的类别,计算这个类别在模型输出上的分数相对于选定层的特征图的梯度。
3. 池化梯度:对这些梯度进行全局平均池化,得到一个权重向量。这个权重向量代表了不同特征图对类别分数的重要性。
4. 生成热力图:将权重向量与特征图相乘,然后对所有特征图进行求和,得到最终的热力图。热力图上高强度区域表明这些区域在模型进行类别预测时起到了关键作用。

2.2.2 LayerCAM

Layer-CAM 是一种新的可解释性技术, 它不仅聚焦于最后一个卷积层, 而是综合考虑多个层次的特征来生成更精细的热力图。

工作原理:

1. 多层分析:Layer-CAM 在多个卷积层上进行分析,提取各层的特征图。
2. 计算分层权重: 对于每个层, Layer-CAM 计算类别分数相对于该层特征图的梯度,并使用这些梯度来评估每个层对最终决策的贡献度。
3. 融合特征图:Layer-CAM 将不同层的特征图按照它们的贡献度加权融合,形成一个综合的、表现多层特征影响的热力图。

Grad-CAM 专注于最后一个卷积层, 易于实现和理解, 但有时可能无法捕获所有相关的视觉特征,特别是在较深的网络中。

Layer-CAM 提供了更细致的视角, 可以捕捉到多个层次的特征对模型决策的影响,但实现起来更复杂,需要处理更多的层次和特征图。

2.3 生成热力图

```
1 import tensorflow as tf
2 from tf_explain.core.grad_cam import GradCAM
```

```
import cv2
import numpy as np

# 加载模型
model = tf.keras.models.load_model('./tf_vgg16')
model = model.get_layer("vgg16")

# 选择要解释的图像
image_path = "./data4/cat1.jpg"

# 加载图像并进行预处理
image = cv2.imread(image_path)
image = cv2.cvtColor(image, cv2.COLOR_BGR2RGB)
image = cv2.resize(image, (224, 224))
image = np.expand_dims(image, axis=0)

# 创建Grad-CAM对象并进行解释
explainer = GradCAM()
for channel_index in range(7):
    grid = explainer.explain((image, 'cat'), model, layer_name="
        block5_conv3", class_index=channel_index,image_weight=0.1)

    # 可视化Grad-CAM结果
    explainer.save(grid, "./grad_cam", "grad_cam_{}.jpg".format(
        channel_index))
```

三　实验结果

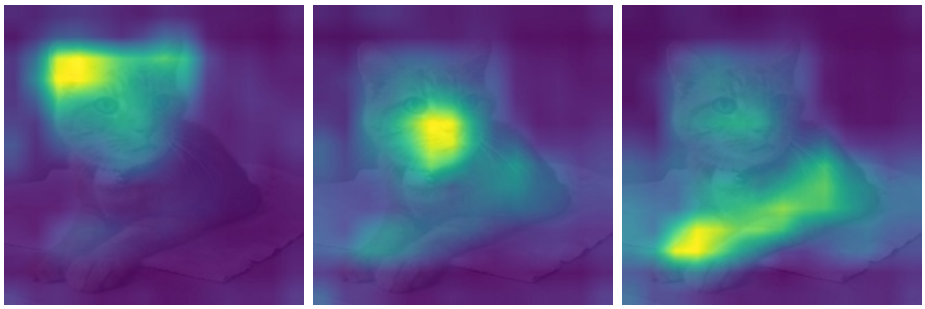

图 3-1　cat 的热力图

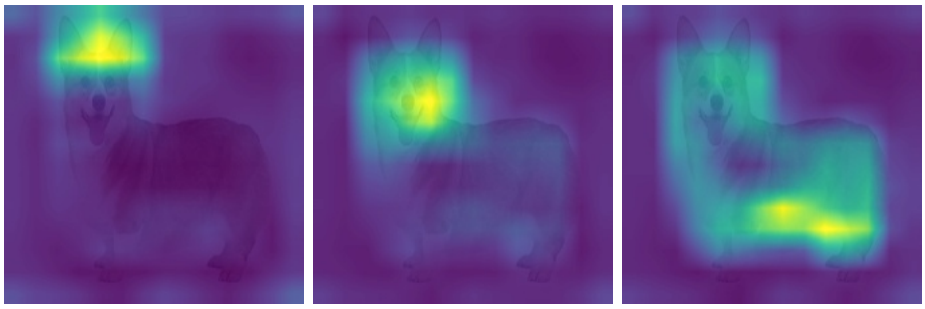

图 3-2　dog 的热力图

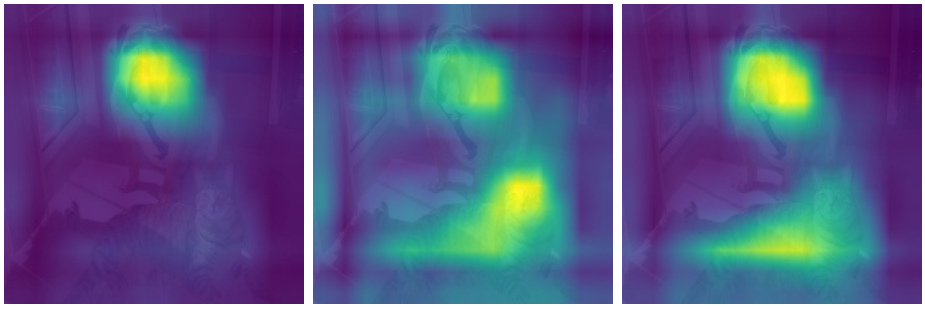

图 3-3　both 的热力图

某些通道的热力图可能显示出对猫和狗耳朵的高关注度。这表明模型在区分猫和狗时认为耳朵是重要的特征。耳朵的形状、大小、位置等特征可能对模型的分类决策起到一定的作用。

某些通道的热力图可能显示出对猫和狗身体的高关注度。这表明模型可能将身体的

特征视为区分猫和狗的重要依据。身体的形状、纹理、颜色等特征可能在模型的分类决策中扮演关键角色。

某些通道的热力图可能显示出对猫和狗面部的高关注度。这可以解释为什么模型能够正确识别猫和狗，因为它们的面部通常具有明显的特征差异。例如，猫和狗的眼睛、鼻子、嘴巴等部位可能在模型的决策中起到重要作用。

当猫和狗同时出现在一张图片中时，模型可能会专注于它们的面部和身体。这是因为猫和狗的面部和身体通常是它们的重要特征，能够帮助区分它们的类别。模型可能会通过学习猫和狗的面部特征(如眼睛、鼻子、口部)以及身体形态(如身体轮廓、四肢位置)来对它们进行区分。

这种观察结果进一步证明了 Grad-CAM 方法的有效性。通过生成热力图，我们可以直观地观察到模型对于不同类别的关注区域，从而理解模型在决策过程中的注意力分布。在这个案例中，我们可以看到模型对于猫和狗的不同部位有明显的关注，这与我们对于猫和狗的视觉感知和区分有一定的一致性。

四　总结

通过本次实验，我们使用了 Grad-CAM 和 LayerCAM 这两种可解释性技术，对卷积神经网络（CNN）的注意力分布进行了可视化，并专注于研究猫和狗图像分类任务。以下是我在实验过程中的一些心得体会：

1. 可解释性的重要性：对于深度学习模型而言，可解释性是一个重要的研究领域。通过了解模型关注的区域和特征，我们可以更好地理解模型的决策过程，并验证模型是否根据我们期望的特征进行分类。
2. Grad-CAM 和 LayerCAM 的应用：Grad-CAM 和 LayerCAM 是两种流行的可解释性技术，通过生成热力图可视化模型的注意力分布。这些技术能够帮助我们定位模型关注的区域，发现模型用于分类的关键特征。
3. 观察结果的解读：在实验中，观察到模型在猫和狗同时存在的图像中，更多地关注它们的面部和身体部位。这符合我们的直观认知，因为面部和身体特征对于猫和狗的识别是非常重要的。
4. 模型的学习能力：通过观察模型的注意力分布，我们可以发现模型学习到了与不同类别相关的特征。模型能够自动关注这些特征，并基于它们进行分类决策，这体现了模型在学习任务中的有效性。

总之，通过可解释性技术的应用，我们能够更好地理解 CNN 模型的决策过程，并验证模型是否根据我们期望的特征进行分类。这些技术为深度学习模型的可解释性提供了有力的工具，对于进一步探索和改进模型性能具有重要意义。