第一章　可解释性应用研究

1.1　利用 VGG19 的可解释性进行剪枝

近年来，卷积神经网络作为一种非常强大的图像分析和处理模型得到了广泛的应用。随着网络研究的不断深入，模型的精度越来越高，但是同时也表现出计算量大、能耗高等特点。为了将模型应用到计算能力弱的边缘计算节点或者部署于即时推理情景中，需要将模型尺寸进一步压缩。受到本文对于可视化及聚类方面的研究的启发，本文进一步提出可解释的剪枝方案。

研究表明，卷积神经网络在特定的应用领域中存在大量的冗余结构。一个在 ImageNet 上良好训练的模型，它可以分离 1000 个不同的类，而在现实世界或不同使用条件下，我们只需要对特定几个类型进行分类。在这种情况下，能够定位冗余结构并对神经网络进行裁剪，将节省大量的计算量并且使得卷积神经网络更加灵活。

基于这种动机，本节利用可解释的方法，从网络剪枝的角度实现对于模型的裁剪。本节对于模型的裁剪分为卷积核簇级别和单个卷积核级别，并且创新地根据网络中特定层单个卷积核对于推理结果的贡献进行优先级排序实现更为均衡的剪枝。三种剪枝在大幅减少神经网络特定卷积层参数的同时，对于预测准确率没有显著的影响。

1.1.1　卷积核簇级别的裁剪

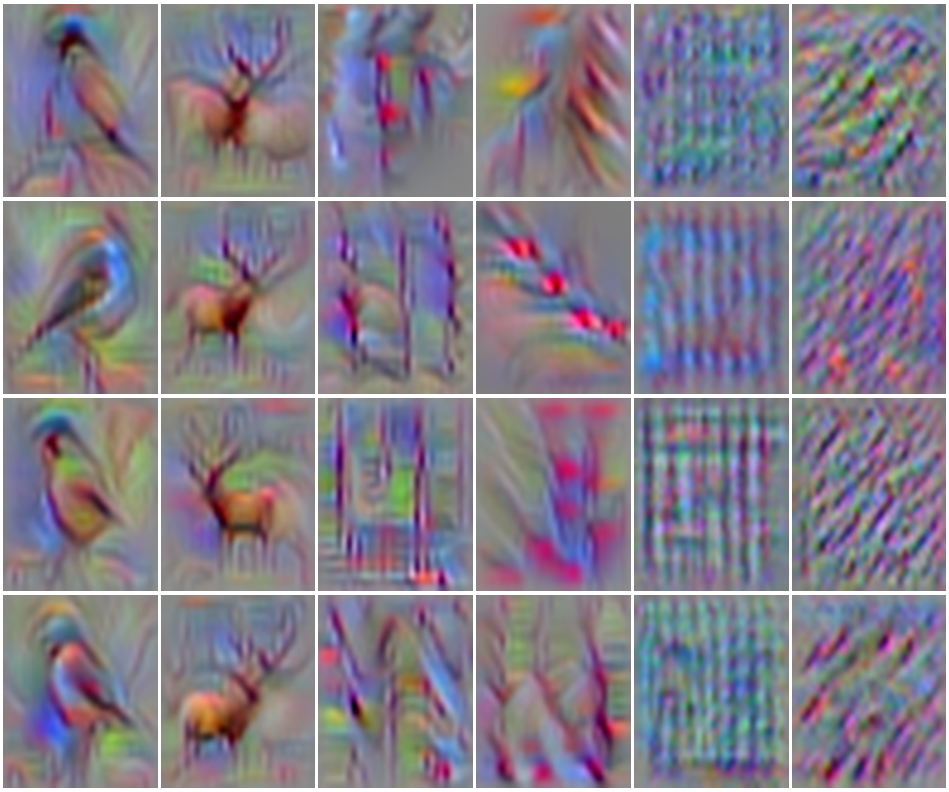
在卷积核簇级别的裁剪过程中，本文根据特定层卷积核输出的特征兴趣图案对卷积核进行聚类，然后对聚类后的结果进行裁剪。

本节采用的是卷积神经网络可视化技术，激活最大化（Activation Maximization）的方法来分析卷积核的特征兴趣的特性兴趣。在激活最大化技术中，卷积核的特征兴趣表现为一种特性兴趣图案，可以量化的记为 $P(F_i^l)$，其中 l, i 分别表示第 l 个卷积层中的第 i 个卷积核。生成 $P(F_i^l)$ 的过程可以用以下公式表述：

$$P(F_i^l) = \arg\max_X A_i^l(X), X \leftarrow X + \eta \cdot \frac{\partial A_i^l(X)}{\partial X}, \tag{1–1}$$

在这个公式中，$A_i^l(X)$ 是输入图像 X 经过卷积核 F_i^l 后输出的激活图像，η 是梯度上升的步长。原始输入 X，在这个过程中得到了最大的 $A_i^l(X)$。由以上公式可以说得到了卷积核 F_i^l 的特征兴趣图案 $P(F_i^l)$。

为了获得这种最大的激活程度，本文选择随机噪声作为输入。这种思想可以归结为使用良好训练的模型来训练图像，而不是传统的神经网络训练中多次迭代图像来训练模型的参数。相较于传统的对于参数执行梯度下降的过程，本文的方式是对于图像进行梯度上升的计算。图像的每个像素根据 $\partial A_i^l / \partial X$ 梯度的引导更新图片，在经过一定次数的迭代后，最后给出卷积核的最大偏好，这个结果是卷积核的特征兴趣图案。对于每个卷积核我们都采用这样的操作，就得到了特定层逐卷积核的特征兴趣图案，如图1–1所示。在得到此卷积神经网络内部信息的基础上，本节继续分析卷积核内在的关联。



a) b) c) d) e) f)

图 1–1 (a) (b) 展示了在 VGG-19 的第 15 个卷积层 $CL5\_3$ 的两个卷积核聚类, (c) (d) 展示了第 6 层 $CL3\_2$ 中的两个聚类，(e) (f) 是来自第二层 $CL1\_2$ 的结果。我们可以根据图像明确找出 (a) (b) 的特征, 即他们激活最大化倾向于什么类别的图像；对于 (c) (d) 只是一些具有相似颜色的轮廓或斑点，表明它们激活并识别这些隐含的图案特征；(e) (f) 比 (c) (d) 更加抽象, 只显示垂直或者水平的边缘。

本节卷积核簇的操作直接对于卷积核特征兴趣图案进行聚类。使用的方式包括 (1) 直接使用原始特征图片，（2）SIFT 提取特征点。考虑到研究的简洁性，后续的操作基于上述获得的原始特征兴趣图案进行操作。对于已经获取的图片信息，本方法使用 K-means 聚类的方式，将其按照图片的特征分成数量特定的簇，针对簇对于模型进行整体裁切。

1.1.2 卷积核级别的裁剪

卷积核级别的裁剪相对于卷积核簇级别的裁剪更加精细地关注定量角度，因而具有更好的可解释性解读。上述的卷积核簇的裁剪，虽然能明确裁剪掉有着相似特定偏好的卷积核簇，但是并不能足够明晰裁剪过程对结果的具体影响。卷积核级别的裁剪，则是量化每一个卷积核对于结果分类概率的贡献，找出卷积核对于结果分类中贡献最大的类作为卷积核的标签。

为了定量地定义每个卷积核的标签，我们通过反向传播梯度来评估每个卷积核的类激活情况。梯度说明了单个卷积核对最终输出的贡献，即对于分类标签概率的贡献。本文使用的反向传播梯度通过以下公式计算：

$$\gamma_{i,y_j} = \frac{1}{N}\sum_{n=1}^{N}\left|\frac{\partial P(y_j)}{\partial W_i(x_n)}\right|, \tag{1–2}$$

其中 γ_{i,y_j} 代表卷积核 i 对于类 y_j 的偏好程度，求平均值的 N 是这个过程中使用的真实图片的数量。

对于训练好的网络，此方法将图片输入进网络中，在反向传播过程中得到卷积核对于各类的梯度。对于特定类，将卷积核经过多张图片计算得到梯度的绝对值求解平均值获取该卷积核对于特定类的敏感程度。该卷积核倾向为所有类别中均值最大的类别。利用上述过程，我们将特定层每个卷积核的倾向求解出来，根据他们的倾向归类。依据此方法实现根据特定类偏好的一致性剪枝。

值得注意的是，仿佛第二节所述的卷积核级别的剪枝和第一节卷积核簇级别的剪枝都是研究卷积核的倾向实现剪枝。两种方法看似有相似之处，实则解决问题的思路完全不同。具体细节如图1–2。卷积核簇中获取卷积核激活最大化时，是应用训练好的模型的前半部分，模型参数不变，对于图片进行梯度上升的多次训练，将图片从噪声图片训练至展示卷积核特征兴趣的图像。整个过程是从前向后反复迭代，使得激活最大化。而卷积核级别获取卷积核对于结果贡献程度是利用完整模型，在反向传播计算过程中获取特定层卷积核对于结果概率的贡献。这个过程是从后向前的梯度传导计算。两者计算的方向和理念不同。

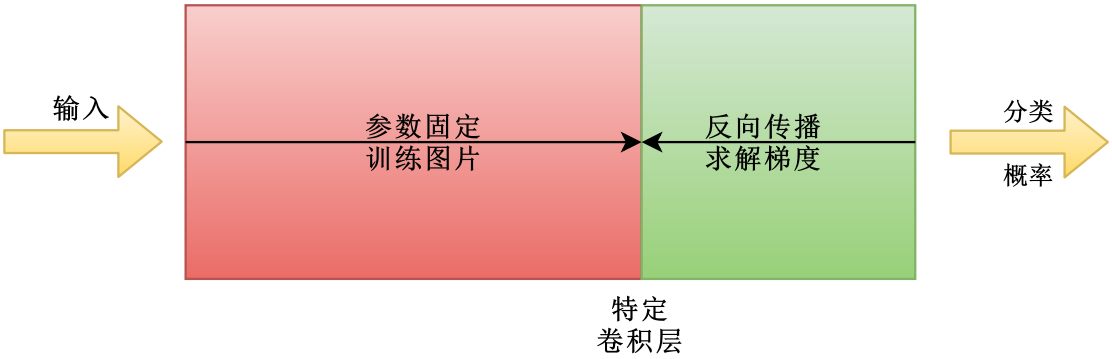


图 1–2 卷积核簇中获取卷积核激活最大化时是从前向后固定模型的参数多次迭代对于图片进行训练，而卷积核级别获取卷积核对于结果贡献最大程度是在反向传播过程中求解梯度。

1.1.3 带有优先级的均衡裁剪

节1.1.2获取了特定层逐卷积核的偏好情况，使得我们可以实现细粒度的卷积核粒度的模型裁剪。在此基础上，本文又提出了带有优先级的均衡裁剪方法。1.1.2节的方法整体裁掉或者保留偏向特定类的全部卷积核，从直观的角度这样的方法对于整体模型类的平衡具有一定的影响。本节提出的带有优先级的均衡裁剪立足于保持倾向于各类的卷积核数目平衡的角度，实现裁剪。

根据前面小节已经可以获得特定卷积核的倾向，并且定量化给出了对于各类倾向的数值。依据此我们将模型中的卷积核分到各类别中，并且依据其倾向数值大小进行排序，得到的顺序作为其在该类别中的优先级。

本节模型裁剪的具体方法是，根据优先级的顺序，逐类保留适当数目的卷积核。如果遇见已经提取的情况则提取该类的下一优先级卷积核。基于此方法可以实现类均衡的模型剪枝，并显著裁剪该层模型的参数量。

1.1.4 实验结果

模型裁剪的实验中，本文对于提出的三种模型裁剪的方法进行了测试。实验是基于预训练好的 VGG19 模型。在本节实验中，VGG19-N 的记号表示不同的裁剪模型。本节实验的裁剪基于 VGG19 模型中的第 15 个卷积层进行。实验使用 CIFAR-10 10000 张图片进行测试。

卷积核簇级别的裁剪 首先是卷积核簇级别的裁剪，实验结果如表1–1所示。实验首先通过前文的方法对 VGG19 的第 15 个卷积层输出卷积核的特征兴趣图案，然后使用 K-means 的方式对于特征图案进行聚类，得到卷积核簇。本实验中选择聚类的数目为 10。聚类后，对于特定聚类中的卷积核进行保留，生成 VGG19-1, VGG19-2 模型。

表 1–1 卷积核簇级别的裁剪

模型	卷积核个数	保留的聚类	飞机	汽车	鸟	猫	鹿	狗	青蛙	马	船	卡车	全部
VGG19	512	全部	86.9%	92.1%	71.3%	65.5%	82.6%	73.8%	88.3%	86.0%	90.0%	89.4%	82.59%
VGG19-1	257	1 2 3 8 9	87.4%	91.7%	72.6%	69.8%	80.0%	71.9%	86.9%	84.2%	89.1%	88.9%	82.25%
VGG19-2	128	1 8 9	86.9%	91.0%	66.8%	83.3%	75.8%	57.2%	84.3%	82.2%	86.9%	85.5%	80.05%

实验结果显示，随着保留的卷积核数目的减少，全部类别的推理准确率逐渐下降。但是由于聚类并没有与推理结果的标签相对应，所以单独类别的准确率并没有很明显的倾向。

卷积核级别的裁剪 第二部分的实验是在卷积核级别进行，通过进一步研究卷积核对于结果概率贡献的程度，将卷积核定量地分配到和标签对应的类别中。实验结果如表1–2所示。

表 1–2 卷积核级别的裁剪

模型	卷积核个数	保留的聚类	飞机	汽车	鸟	猫	鹿	狗	青蛙	马	船	卡车	全部
VGG19	512	全部	86.9%	92.1%	71.3%	65.5%	82.6%	73.8%	88.3%	86.0%	90.0%	89.4%	82.59%
VGG19-3	114	0,1,2,4,7	86.3%	91.5%	69.2%	79.8%	80.1%	63.8%	82.9%	83.0%	88.0%	89.0%	81.36%
VGG19-4	116	3,5,8	88.7%	91.3%	69.9%	79.1%	74.3%	60.8%	85.3%	83.7%	87.9%	88.9%	80.99%
VGG19-5	251	9	85.9%	90.9%	63.9%	84.3%	76.0%	54.3%	85.5%	81.0%	85.3%	**90.8%**	79.79%
VGG19-6	261	排除 9	87.7%	91.8%	71.6%	67.8%	81.8%	72.5%	87.2%	85.6%	90.0%	88.7%	82.47%
VGG19-7	352	1,8,9	87.3%	91.6%	70.0%	70.7%	81.0%	70.7%	87.7%	85.1%	88.4%	90.2%	82.27%

VGG19-3 和 VGG19-4 的设置是基于保留卷积核数目大致相同的情况分析模型裁剪。VGG19-5 和 VGG19-6 是对照组，区别是保留 9 和排除 9 的其他类别。VGG19-7 作为保留卷积核数目更多的级别，分析其推理情况。

VGG19-3 和 VGG19-4 实验结果表示，即使保留有相近数量的卷积核，推理准确率也会出现较大的差异。VGG19-5 中只保留了倾向于标签 9 的卷积核，在预测中对于卡车（标签 9）的推理结果有着超越裁剪前网络的推理准确率。这说明排除倾向于其他标签的卷积的干扰，卷积核可以实现更加专注的推理。VGG19-5 和 VGG19-6 的对照试验中可以发现，在卷积核数目接近的情况下，保留倾向性更全面的卷积核可以使得全部的推理结果更好，甚至接近裁剪前的网络。VGG19-7 的实验数据表明，单纯提高卷积核的数目，对于网络的推理结果没有帮助。

带有优先级的均衡裁剪 第三部分的实验是利用本文所述的方法，对于每类别中的卷积核按照其倾向程度进行排序，然后利用此优先级使用轮询的方式逐类保留高优先级的卷积核，实现模型裁剪。实验数据如表1–3所示。

表 1–3 带有优先级的均衡裁剪

模型	卷积核个数	飞机	汽车	鸟	猫	鹿	狗	青蛙	马	船	卡车	全部
VGG19	512	86.9%	92.1%	71.3%	65.5%	82.6%	73.8%	88.3%	86.0%	90.0%	89.4%	82.59%
VGG19-8	128	85.7%	90.6%	71.3%	75.5%	79.7%	66.3%	85.9%	82.4%	89.3%	89.7%	81.64%
VGG19-9	256	86.8%	91.7%	70.8%	68.6%	81.8%	72.6%	87.4%	85.5%	89.7%	89.3%	82.42%
VGG19-10	384	86.9%	92.0%	71.7%	66.8%	82.4%	72.5%	88.1%	85.7%	89.7%	89.3%	82.51%

VGG19-8, VGG19-9 和 VGG19-10 三个类别分别保留 128, 256, 384 个卷积核。实验结果显示，由于使用了优先级的均衡裁剪方式，随着保留的卷积核数量的增多，模型的准确率愈发趋近于裁剪前的模型。并没有出现表1–2中 VGG19-6 和 VGG19-7 的数据显示的卷积核数量增加，反而模型性能明显降低的情况。同时，表1–1中 VGG19-2 和 VGG19-8 都保留 128 个卷积核，但是模型的准确率从 VGG19-2 的 80.05% 提升至了 82.27%。这三个现象，说明优先级均衡裁剪的有效性。

表 1–4 带有优先级的均衡剪枝得到的模型

模型	卷积核个数	全部准确率	裁剪层参数	减少比例
VGG19	512	82.59%	2,359,296	0
VGG19-8	128	81.64%	589,824	75%
VGG19-9	256	82.42%	1,179,648	50%
VGG19-10	384	82.51%	1,769,472	25%

在表1–4中，记录了模型裁剪后的模型尺寸和参数量的变化情况。通过模型裁剪后，得到的模型在不损失太多精确程度的前提下，参数量大幅减少。

本实验从神经网络可视化出发，基于神经网络可解释性，提出的剪枝方案经实验验证有效。

同时本文受到论文的启发：裁剪得到的网络结构比参数更加重要。本文还对于得到的模型进行重新训练，实现了比原模型更高的精度。