



中南大學
CENTRAL SOUTH UNIVERSITY

本科毕业设计(论文)

GRADUATION DESIGN(THESIS)

题 目： 基于平均图的深度神经
网络可解释性方法研究

学生姓名： 王心远

指导教师： 赵颖

学 院： 计算机学院

专业班级： 计算机科学与技术
1807 班

本科生院制

2022 年 6 月

基于平均图的深度神经网络可解释性方法研究

摘要

深度神经网络（Deep Neural Networks,简称 DNN）是一种至少具备一个隐层的神经网络，其中卷积神经网络（Convolutional Neural Network, CNN）是拥有卷积结构的深度神经网络，它拥有强大的空间信息提取能力，目前被广泛应用在目标识别、目标检测等计算机视觉任务中。目前新颖而有效的 CNN 结构层出不穷，但它们的共性是计算复杂度极高、结果难以解释，有着“黑盒”属性，这也为其特征理解和进一步的网络优化带来了巨大的困难。CNN 的可解释性关乎其应用的安全性、公平性和可信性，对神经网络发展的至关重要。

目前对卷积神经网络可解释性的研究十分丰富，提供了不同的理解 CNN 的角度，如基于梯度分析的解释方法、基于扰动分析的解释方法、基于语义分析的解释方法和基于层关联的解释方法等。可视化作为一种帮助研究人员理解的工具，被应用到了多种解释方法之中，是理解 CNN 特征的重要方式之一。本研究——基于平均图的深度神经网络可解释性方法研究的目的是基于数据类平均图像，使用可视化方法以及可解释性算法等工具对 CNN 的特征和性能进行分析，增强 CNN 的可解释性。

本研究提出了神经元相似值分析算法，通过比较神经元可视化图像和类平均图像之间的相似度，将可视化图像量化为可衡量的指标——相似值。其背后的原理是类平均图作为数据类别的“代表”，包含了数据的类公有信息，因此可以作为衡量神经元信息提取能力的标准。本研究提出了多种计算类平均图的技巧以及相似值的计算方法，并在多个预训练模型上对算法进行实验分析。除此之外，本研究还从神经网络剪枝的角度设计实验，验证了神经元相似值的有效性。

关键词： 深度学习 卷积神经网络 可视化 可解释性

The Research on The Interpretability Method of Deep Neural Network Based on Average Image

ABSTRACT

Deep Neural Networks (DNNs) are neural networks with at least one hidden layers, among which Convolutional Neural Networks (CNNs) are neural networks with convolution structures. This special structure brings powerful spacial information abstracting ability, so CNNs are widely applied in computer vision tasks like object recognition and object detection. For now, many noble and effective CNN structures have popped out, but with a common problem of high computational complexity and low interpretability of their results, or "Black Box" attribute, which cause dramatic challenge of CNN feature understanding and further network optimization. The interpretability of CNNs are concerned with the safety, fairness and credibility of their applications. Therefore, delving deeper into the interpretability of CNNs is essential for the future of neural networks.

Currently, there are various researches of CNN interpretability, including interpretation methods based on gradient analysis, perturbation analysis, semantic analysis, layer correlation and so on. Visualization, as a tool of understanding complex data, is applied in many interpretation methods, so it is an important way to understand the features of CNNs. The goal of this research is, based on data average image, using visualization methods, interpretability algorithms and other tools to analyze the characteristics and performance of CNNs and improve the interpretability of CNNs.

This research proposes neural similarity analysis algorithm. It quantify the visualization results as a comparable metric, "similarity value", by computing the distance of neural visualization images and class average image. The reason behind it is that class average image contains public information in one class, so it could be used to measure the information extraction power of neurons. His research also proposed three methods to compute class average image and two ways to compute similarity value. They

are tested on multiple pre-trained models. Besides, neural network pruning experiments are also conducted to validate the efficiency of neural similarity value.

Key words: Deep Learning Convolutional Neural Network Visualization

Interpretability

目录

第 1 章 绪论	1
1.1 研究背景与意义	1
1.2 国内外研究现状的调研	2
1.2.1 卷积神经网络可解释方法研究现状	2
1.2.2 卷积神经网络剪枝研究现状	5
1.3 任务概述	7
1.4 论文组织结构	9
第 2 章 神经元可视化算法	10
2.1 神经元可视化算法	10
2.1.1 基于反卷积的神经元可视化方法	10
2.1.2 基于反向传播的神经元可视化方法	11
2.1.3 基于导向反向传播的神经元可视化方法	12
2.1.4 基于积分梯度的神经元可视化方法	14
2.1.5 神经元可视化方法比较	15
2.2 最大化激活算法	16
2.2.1 网络驱动的最大化激活算法	17
2.2.2 数据驱动的最大化激活算法	17
2.3 图像相似度指标	18
2.3.1 基于传统方法的图像相似度指标	18
2.3.2 图像相似度指标比较结论	19
2.4 本章小结	19
第 3 章 基于类平均图的可视化图像相似值计算方法研究	20
3.1 总体设计	20
3.2 类平均图的设计与计算方法	20
3.2.1 简单类平均图的计算方法	21

3.2.2 数据驱动激活最大化的类平均图计算方法	23
3.2.3 网络驱动激活最大化的类平均图计算方法	25
3.3 神经元可视化图像的相似值计算方法	26
3.3.1 方法一：基于类别平均图的相似度	26
3.3.2 神经元相似值计算方法二：基于类置信度可视化图像的相似度	27
3.4 本章小结	28
第 4 章 基于平均图的神经元可视化图像相似值的实验与分析	29
4.1 预训练数据与模型准备	29
4.2 类平均图的计算结果	30
4.2.1 简单类平均图的计算	30
4.2.2 数据驱动激活最大化的类平均图像的计算	32
4.2.3 网络驱动激活最大化的类平均图像的计算	35
4.3 神经元相似值计算结果与相似值曲线分析	36
4.3.1 层内神经元相似值分析	37
4.3.2 层间神经元相似值分析	39
4.4 本章小结	43
第 5 章 基于神经元相似值的剪枝验证实验与分析	44
5.1 基于相似值的网络层剪枝实验	44
5.2 基于相似值的神经元剪枝实验结果	45
5.3 本章小结	47
第 6 章 总结与展望	48
6.1 研究总结	48
6.2 未来展望	48
致谢	50
参考文献	51

第1章 绪论

1.1 研究背景与意义

随着大数据时代的来临，互联网中的数据量呈爆炸式增长，与此同时，计算机计算能力随着计算机硬件的发展不断提高，由此由数据驱动的深度学习走上了历史的舞台，在计算机视觉、自然语言处理等领域大放异彩。深度学习是机器学习的一个分支，其算法的主体是深度神经网络（Deep Neural Networks,简称 DNN）——是一种至少具备一个隐层的神经网络对数据进行表征学习的算法。其中卷积神经网络（Convolutional Neural Network, CNN^[1]）是一种特殊的深度神经网络，它是使用卷积结构进行特征学习的深度神经网络。AlexNet^[2]网络模型在 ImageNet 2012 目标识别竞赛中取得了远超传统算法的准确率，使卷积神经网络算法走入了大众的视野。卷积计算的序列型计算方式为 CNN 带来了强大的空间信息提取能力，非常适合被应用到图像、时序数据等拥有空间结构的数据上，因此 CNN 目前被广泛应用在目标识别、目标检测、图像分割等计算机视觉任务中，并取得了优秀的性能。

目前新颖而有效的 CNN 结构层出不穷，但它们的共性是计算复杂度极高，一个复杂的深度学习模型通常有百万级以上的参数量，同时模型中各结构的连接方式也十分复杂，由此模型推理得到预测结果的过程难以解释。这带来了几个问题：第一，网络可靠性难以保证——网络模型在应用中可能产生许多无法预测的结果，这为 CNN 在许多领域的应用带来了致命性的缺陷，例如在自动驾驶、医疗影像等领域，这样的错误结果通常是不可接受的；第二，模型的可解释性限制了其泛化能力，例如 Szegedy 等^[3]发现只要在图像中进行微小随机改动就可以改变网络的预测结果，而 Nguyen 等^[4]则用一张完全无法辨认的图像在网络中得到了 99.99% 的置信率，人们却无法解释这样的意外情况；第三，道德与法律要求——通常在模型的训练过程中，由于训练数据集的有限性，得到的模型通常会有一定的“偏见”，因此不适用于在保险风险评估、信用等许多领域对算法公平性要求很高的领域；第四，对模型结构、特征、推理方式的有限理解限制了网络的进一步优化，很多情况下研究人员

只能凭借经验构建神经网络，而不能真正理解其组成部分在网络学习或推理中的作用。CNN 的可解释性关乎其应用的有效性、安全性、公平性和可信性，所以目前神经网络的“黑盒”属性限制了其应用和优化，加深对 CNN 的理解对神经网络未来的至关重要。

目前对卷积神经网络可解释性的研究种类繁多，包括基于梯度分析的解释方法、基于扰动的解释方法、基于语义分析的解释方法和基于层关联的解释方法等。可视化作为一种重要的工具，被应用到了多种解释方法之中，是理解 CNN 特征的重要方法之一。本研究——基于平均图的深度神经网络可解释性方法研究，基于数据平均图像，使用可视化方法以及可解释性算法等工具对 CNN 的特征和性能进行分析。目前，最大激活法、导向反向传播等可视化方法，都在一定程度上直观地展示和解释了网络内部特征的学习情况。但是可视化方法在解释特征时缺乏量化的指标，由此带来两个影响：第一，由于 CNN 内部特征数量多，特征可视化结果相似性高，人眼难以区别；第二，可视化可以增强研究人员对网络的理解，但是这样的理解难以以为网络的优化给出直观的建议。本次毕业设计利用类平均数据为 CNN 卷积核特征可视化方法提供量化指标，在卷积核特征和可视化结果之间建立量化联系，使研究人员可以从多个角度对 CNN 的特征进行分析和理解，获取有关网络特征学习的更多信息，加深对 CNN 的认知。

1.2 国内外研究现状的调研

1.2.1 卷积神经网络可解释方法研究现状

卷积神经网络（CNN）在 1998 年由 LeCun 等[1]首次被提出，Krizhevsky 等[2]在 2012 年提出的 AlexNet 在 ImageNet 2012 目标识别竞赛中取得了远超传统算法的准确率，使卷积神经网络在图像上的应用算法走入了大众的视野，从此多种 CNN 结构被提出，如 2016 年的残差网络（He 等^[5]），2017 年的 DenseNet（Huang 等^[6]），卷积神经网络的性能不断提高，被应用到了多个计算机视觉领域，如目标识别模型 YOLO^[7]，RCNN^[8]；图像分割模型 U-Net^[9]等。但是 CNN 复杂的结构、庞大的参数量和端到端的训练方式使 CNN 成为了一个难以理解的黑盒，除了网络的输入与输出结果外，研究人员无法理解网络内部的推理逻辑。由此产生出 CNN 可

解释性研究领域，希望通过从 CNN 的结构、特征、学习特征等角度进行解析，提高对 CNN 的理解能力，以此推动 CNN 的应用与发展。

目前对卷积神经网络可解释性的研究层出不穷，包括基于梯度分析的解释方法、基于扰动的解释方法、基于语义分析的解释方法和基于层关联的解释方法等^{[10][11][12]}。其中 CNN 卷积核特征可视化方法是本研究的核心，因此在第二章中详细阐述。CNN 可解释方法的分类有多种依据，本文根据^[34]，以不同的解释角度将这些方法分下列类别：

（1）CNN 卷积核特征可视化

将 CNN 的卷积核可视化是探索单个神经元学习内容最直接的方法。第一，基于梯度的方法，其基本原理是对一个给定的 CNN 卷积核，计算其关于一个输入图像的激活特征的梯度，使其在输入层重构可视化图像，该图像显示出输入对梯度影响最显著的区域。Zeiler 和 Fergus^[13]在 2014 年提出了反卷积网络（DeconvNet），将池化、卷积等操作进行转置，从单个神经元逐步上采样重构图像；Simonyan 等^[14]在 2013 年提出了通过单次反向传播计算梯度得到单个分类或者神经元的显著性图像，Springenberg 等^[15]在 2015 年在它的基础上提出了导向反向传播，对整流非线性激活层的反向传播方式进行限制，该方法可以在更深层的网络中计算出更清晰易理解的可视化图像。第二，上卷积网络，Dosovitskiy 和 Brox^[16]在 2016 年提出的上卷积网络通过将 CNN 特征映射翻转至图像，可以展示一个特征映射的图像外观，但相比基于梯度的方法，它不能从数学角度完全反应网络特征的真实情况；2017 年 Nguyen 等^[17]在此基础上引入了带有语义信息的附加先验，可以控制合成图像的语义信息。第三，2015 年 Zhou 等^[18]提出了一种精确计算在一个特征映射中的神经元激活的图像分辨率级别的感受野，实际的感受野要比理论上使用卷积核尺寸计算得到更小，对感受野的估计可以帮助人们理解一个卷积核的特征表征。

（2）CNN 表征分析

一些方法在 CNN 可视化的基础之上，对 CNN 的特征表征进行分析，主要思想是分析 CNN 不同目标分类对应的特征空间，或者探索卷积层的潜在错误特征表征，主要有五个方向。第一，从全局的视角观察 CNN 特征：Szegedy et al., 2014^[19]

探索了每个卷积核的语义信息，Yosinski 等^[20]在 2014 年分析了卷积层中卷积核特征的迁移能力。第二，直接提取影响一个标签/分类的图像区域，来解释单个标签/分类的 CNN 特征表征，比如 Fong 和 Vedaldi^[21]，Selvaraju 等^[22]在 2017 年提出了对网络损失进行反向传播来估计图像区域；Ribeiro 等^[23]在 2016 年提出 LIME 模型用于提取对网络输出最敏感的图像区域；Zintgraf^[24]于 2017 年提出了在输入图像中对 CNN 决策影响最大的部分的可视化方法。第三，对一个 CNN 网络的特征空间中的敏感点分析也是一个重要的方向，Su 等^[25]于 2017 年提出了对一个 CNN 计算对抗样本的方法，其基本思路是在输入图像中寻找最小的扰动来改变网络的输出结果；Koh 和 Liang^[26]于 2017 年提出了影响方程来计算对抗样本，使用这种方法可以创建攻击 CNN 的训练样本，弥补训练集的不足，提高 CNN 的鲁棒性。第四，通过对网络特征空间分析来改善网络的特征表征情况。Lakkaraju 等^[27]于 2017 年提出了一种在弱监督条件下发掘 CNN 的盲点（未知模式）的方法，这种方法通过将 CNN 空间中的所有样本点进行分组获得几千个拟类，CNN 通过这些子空间对一个特定的类进行表征。

（3）解构 CNN 卷积核中混合的模式

将 CNN 卷积层的复杂特征表征解构，将网络特征表征转化为可理解的图像。相比前两种方法，这种方法为网络特征表征提供一种更全面的解释。Zhang 等^[28]在 2018 年提出了对一个预训练 CNN 的卷积层特征的解构方法。一个 CNN 网络的高层中的卷积核通常包含着多种模式，比如一个目标的首部和尾部，因此该工作致力于回答以下几个问题：1) 卷积层的卷积核学习到了多少视觉模式；2) 对于一个目标部分有多少模式被同时激活；3) 两个同时激活的模式的空间关系是什么。它们提出了解释图（explanatory graph）来解释隐藏在 CNN 中的语义信息，使用每个图节点表示一个语义部分，每个边表示共同激活关系，从而将每个神经元的特征映射进行解构。Zhang 等在后续工作中进一步提出了使用决策树来编码全连接层的决策模式，这个决策树并非用来分类，而是解释每个 CNN 预测的逻辑，可以解释哪些神经元在每次预测中起作用。

（4）构建可解释的网络

前面的方法都是理解一个预训练的网络，本部分的工作从网络结构层面构建自身可解释的模型，网络中间层不再是黑盒，而有着清晰的语义信息。Zhang 等在 2017 年提出了通过对卷积层的每个卷积核添加一个损失来解构高层的特征表征，这个损失用来将特征映射正规化为一个特定的目标部分的特征表征，卷积层中每个卷积核被赋予一个特定类别，如果输入图像属于这个类别，那么损失就期望这个卷积核的特征映射匹配这个类别，也就是说卷积核的特征映射的特定位置会被激活。Wu 等^[29]在 2017 年基于目标检测模型 R-CNN 提出了质量上可解释的模型，它的目的是在目标检测过程中自动展开目标的潜在部分。Sabour 等^[30]在 2017 年设计了一种新颖的神经元——胶囊，来代替传统神经元，得到的网络称为胶囊网络，每个胶囊的输出是一个活动向量，而不是一个标量，活动向量的长度代表胶囊的激活长度，低层的激活胶囊回想邻接的高层胶囊发送信息，它们使用路由协议机制为胶囊赋予权重。在 MNIST 数据集上的实验显示，活动向量的不同维度可以控制不同的特征，包括大小、长度、位置、笔宽等等。Chen 等^[31]提出了 InfoGAN，也就是信息最大化生成对抗网络，它最大化潜在特征表征的特定维度和观察图像得到交互信息。

（5）通过人机交互进行语义层面的学习

一个已解构的 CNN，它的特征表征语义的清晰，可以进一步通过弱监督学习令网络进行“中到端”的学习。Zhang 等在^[32]提出了一种通过主动问答来对卷积层神经元模式进行语义化的方法，并且构建了分层对象理解的模型。这个方法的目标是通过提取一个可解释的与或图（AOG）来解释隐藏在 CNN 中的层次语义信息，在 AOG 中与节点表示一个部分的组成区域，或节点表示当一个局部部分的一系列可替代的模板或变形候选。每一个潜在模式（或节点）都自动对应于一个卷积核特征映射的特定范围的单元。为了学习得到 AOG，在每次问答过程中，AOG 在所有未标注图像中定位目标部分，当当前 AOG 不能解释神经元模式时可以通过计算机主动识别和询问目标。

1.2.2 卷积神经网络剪枝研究现状

为了获得更优秀的性能，越来越多复杂的神经网络结构被提出，但这对网络的实时性和计算资源提出了更大的挑战，网络压缩（network compression）方法因此

应运而生。模型压缩通常将模型的体量缩减，其代价通常是准确率的微小损失，在个别情况下可以提升准确率。剪枝（pruning）是模型压缩的重要方法之一，它包括静态剪枝和动态剪枝，其区别是是否在模型运行时完成。卷积神经网络剪枝有多种尺度，包括元素尺度、通道尺度、形状尺度、滤波器尺度、层尺度和网络尺度。本部分主要调研了静态剪枝，静态剪枝是通过在训练后或推理前的网络移除部分神经元对网络进行优化的技术。静态剪枝通常包括三个部分：1) 剪枝参数的选择；2) 神经元剪枝方法；3) 微调或重训练（可选）。重新训练剪枝后的网络可能达到和剪枝前的网络相近的性能，但是会消耗大量算力。

早期的神经网络剪枝方法是暴力剪枝，对网络逐元素遍历，对准确率没有影响的权重将被移除。这个方法的缺点是需要遍历的解空间过大，一个衡量权重剪枝的经典指标是 l_p 正则化， $p \in \{N, \infty\}$ ， p 是自然数，由 n 个元素组成的向量 x 的 l_p 正则化表示为：

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}} \quad (1-1)$$

其中 l_1 正则化被称为曼哈顿正则化， l_2 正则化被称为欧几里得正则化，二者对应的正规化方法被称作 LASSO 正规化和 Ridge 正规化^[33]。LASSO 正规化可以表示为公式一，对于包含 N 个实例的样本，每个实例包含 p 个协变量和一个单独的输出 y_i 。令 $x^i = (x_{i1}, \dots, x_{ip})^T$ 表示神经网络第 i 个输入特征的标准协变量矩阵，那么 $\sum_i x_{ij}/N = 0$, $\sum_i x_{ij}^2/N = 1$, β 表示系数（权重）， t 是预定义的参数，用来决定稀疏性。LASSO 会当 y 的均值为 0 时将 α 估计为 0。如果限制是 $\sum_j \beta_j^2 \leq t$ ，那么则是 Ridge 正规化。

$$\arg \min_{\alpha, \beta} \left\{ \sum_{i=1}^N (y_i - \alpha - \sum_{j=1}^p \beta_j x_{ij})^2 \right\} \quad (1-2)$$

$$\text{对于 } \sum_j |\beta_j| \leq t$$

基于幅度的剪枝和基于惩罚的剪枝可能令权重产生 0 值或近 0 值。基于幅度的剪枝的原理是训练得到的权重中值更大的权重比值更小的权重有更高的重要性^[34]，基于幅度的剪枝的目标是去识别不需要的权重和特征，并将它们在运行评估时移除。

其中最为直接的方法是将全 0 的权重或者绝对值在特定阈值范围的权重剪枝。

LeCun 在 1990 年提出了 Optimal Brain Damage (OBD) 来剪枝单个不重要的权重^[35]。通过使用损失函数的二阶导 (汉森矩阵)，这种静态剪枝技巧可以将网络参数量缩减四分之一。OBD 在三个假设下起作用：1) 二次：损失函数是接近二次的；2) 极值：剪枝在网络收敛后进行；3) 对角线：通过由错误结果的共同结果将错误的单个权重相加。后来的 Optimal Brain Surgeon (OBS)^[36] 在 OBD 的基础上扩展出一个类似的二阶方法，移除了 OBD 的对角线假设。OBS 认为在大多数的应用中海森矩阵是非对角的。OBS 在 XOR 网络上除去了大约 90% 的权重冗余。

基于滤波器尺度的剪枝^[37] 使用 l_1 正则化来移除不会影响分类准确率的滤波器，剪去整个滤波器和它们的特征映射，令 CIFAR-10 数据集上训练的 VGG-16 和 ResNet-110 的推理消耗分别降低了 34% 和 38%，分别带来了 0.75% 和 0.02% 的准确率提升。多数方法的衡量标准是权重，而平均 0 百分比 (Average Percentage Of Zeros, APoZ)^[38] 提出了使用输出激活映射对结果的贡献来进行衡量，作者通过统计网络中所有验证图像的 ReLU 激活值的平均数量为每个神经元提供衡量指标。

基于元素尺度的剪枝会导致非结构化的网络组织，它会导致在指令集处理器不能高效执行的离散权重矩阵，同时它们难以被有效的压缩或者加速。Group LASSO^[39] 通过一种结构化的剪枝方法——将神经元按组剪枝，一定程度减小了这些问题的影响。类似的还有，基于组尺度的 brain damage^[40]，将 group LASSO 应用到滤波器。

基于通道尺度的剪枝：Network slimming^[41] 将 LASSO 应用到了 BN 层，通过将 BN 参数置零，可以实现通道尺度的剪枝。在 ILSVRC-2012 上训练的 VGG 上，达到了 82.5% 的尺寸缩减，30.4% 的计算压缩，并且没有带来准确率损失。同样使用 BN 的参数，Zhang 等^[42] 使用一种聚类方法计算特征映射通道的距离，相近的特征可以被修改。He 使用 LASSO 回归而不是贪心算法来估计通道，在每一次迭代中，首先使用 l_1 正则化估计最重要的通道，然后使用平均平方损失衡量通道，最小的将被剪枝。

1.3 任务概述

本研究——基于平均图的深度神经网络可解释性方法研究的目的是基于数据平均图像，使用可视化方法以及可解释性算法等工具对 CNN 的特征和性能进行分析。目前，最大激活法、导向反向传播等可视化方法，都在一定程度上直观地展示和解释了网络内部特征的学习情况，但是多数可视化方法在解释特征时缺乏量化的指标，由此带来两个影响：第一，由于 CNN 内部特征数量多，特征可视化结果相似性高，难以分辨；第二，可视化依赖人的观察分析，难以对网络的优化给出直观的建议。本研究利用平均数据为 CNN 提出特征可视化的量化指标，在特征和可视化之间建立量化联系，使研究人员可以从多个角度对 CNN 的特征进行分析和理解，获取有关网络特征学习的更多信息，加深对 CNN 的认知。具体任务包括：

预训练数据集与模型设计： CNN 模型结构和训练数据集种类繁多，需要挑选适合本研究的数据集和模型结构。预训练数据集要求有一定的代表性，同时满足物理设备的限制，如主流的图像识别数据集：MNIST 手写数字数据集、CIFAR-10 数据集，ImageNet 数据集等。预训练模型结构要求满足 CNN 网络的典型特征，有一定的代表性，如目前开源的网络模型结构：VGG、AlexNet 等，并对模型结构进行调整，令其满足实验要求。在选取的数据集上训练模型，使其拟合数据集并保持稳定，记录其的测试准确率和损失变化。

调研 CNN 神经元可视化方法： 调研卷积神经网络的可视化方法，并在预训练模型和数据集上测试，从中选取合适的卷积核可视化方法。可视化算法要求可以以可视化图像的形式反映特定神经元的特征表征。具体任务为：在预训练网络上测试神经元可视化算法，观察可视化图像，选取适合实验的可视化方法。

设计 CNN 可视化特征的量化方法： 本研究的目的是衡量神经元对平均图中信息的提取程度，因此需要选取合适的相似度计算方法，相似度方法包括欧式距离、余弦相似度等。具体任务为：选取合适的相似度指标，在预训练网络上进行测试，获取网络的量化指标。

对 CNN 特征的量化指标和可视化图像进行联合分析，分析网络学习情况，挖掘网络信息，具体任务为：对神经元可视化图像的量化结果从不同角度进行分析，测试算法效果；设计量化结果验证实验，验证量化指标的有效性。

1.4 论文组织结构

本论文首先调研了相关领域的研究背景与研究现状，包括卷积网络可解释方法研究和卷积网络剪枝研究，然后着重调研了本研究可能使用的相关技术：1) 4种神经元可视化算法，经过对比，本研究选择导向反向传播算法；2) 激活最大化算法，包括网络驱动和数据驱动两种角度；3) 基于传统方法的图像相似度指标，经过对比，本研究选择余弦相似度。本论文的第三章详细阐述了基于类平均图的神经元相似值算法的设计，本部分提出了三种类平均图的计算方法和两种神经元可视化图像的相似值计算方法。本论文的第四章详细介绍了基于类平均图神经元相似值算法的实验内容，首先介绍了实验的模型与数据准备，然后从网络层内、层间两种角度对神经元相似值的计算结果和应用效果进行了分析。本论文的第五章设计了基于神经元相似值的剪枝实验，包括网络层剪枝实验和单层神经元剪枝实验，二者都从侧面验证了相似值的有效性。第六章对本论文进行了总结与展望。

第 2 章 神经元可视化算法

2.1 神经元可视化算法

2.1.1 基于反卷积的神经元可视化方法

在 Zeiler and Fergus, 2014^[13]中，作者提出了一种 CNN 可视化方法——反卷积（DeconvNet），来观察中间层特征和分类器的作用，它可以看做一个卷积网络的反向网络，将特征映射回图像像素，其基本思想是将池化、卷积等操作进行转置，从单个神经元逐步上采样重构图像。反卷积网络为目标卷积网络的每一层添加上转置操作，提供了一个通向输入的通道。其基本过程是首先为卷积网络输入一张图像，计算出网络中每层的特征映射，将每层的特征映射传入的反卷积网络层，然后对应地进行上池化、整流、卷积来重构输入图像。

上池化：在卷积网络中，最大池化操作是不可逆的，但是通过记录池化区域中最大值的位置——“路径”变量，可以对池化的转置进行拟合。在反卷积网络中，上池化操作根据“路径”将重构值置于特定的位置，以此还原上层的输入。

整流：卷积网络使用 ReLU 非线性激活层，将特征映射整流为非负值，为了获得每层有效的重构特征，反卷积网络也使用了 ReLU 层保证重构特征的非负性。

卷积：卷积网络使用卷积核来对上层特征映射进行卷积，反卷积网络使用相同卷积核的转置来拟合它的翻转版本，也就是对每个卷积核进行水平和垂直翻转，在整流映射的输出上进行卷积操作。反卷积网络结构如下图 3-1；反卷积可视化结果如下图 3-2 所示。

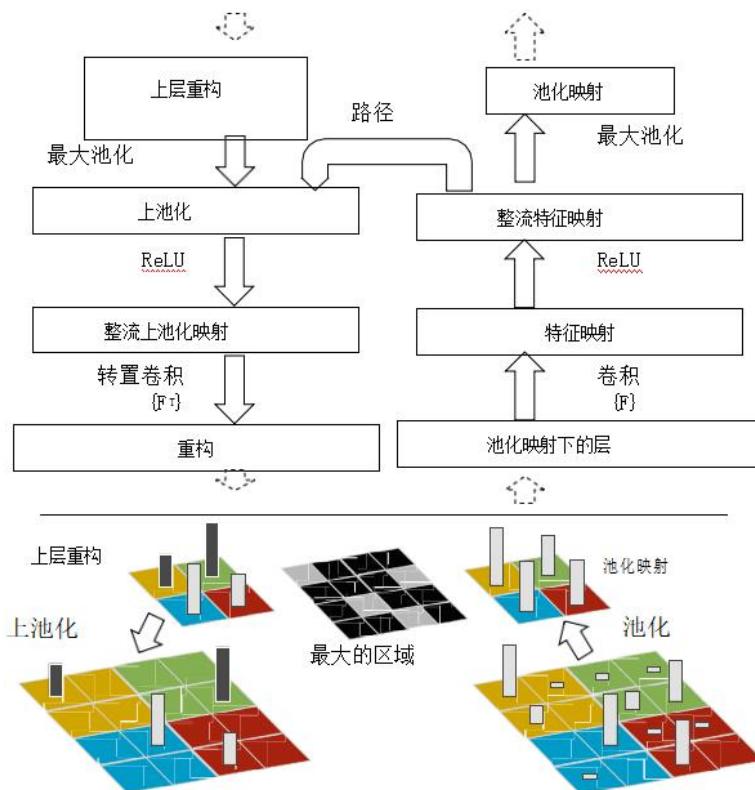


图 3-1 反卷积网络 (DeconvNet) 模型结构



图 3-2 反卷积网络 (DeconvNet) 可视化效果

2.1.2 基于反向传播的神经元可视化方法

Simonyan 等人^[14]提出了一种给定图像和类别计算类显著性图像的方法。对于给定图像 I_0 , 类别 c , 一个分类卷积网络的类别得分函数 $S_c(I)$, 可以根据图像中每个像素对 $S_c(I)$ 的影响对像素进行排序。 $S_c(I)$ 的计算是关于 I_0 的一个高度非线性的方程, 通过计算一阶泰勒展开可以用一个线性方程对 $S_c(I)$ 进行拟合:

$$S_c(I) \approx w^T T + b \quad (2-1)$$

其中 w 是 S_c 在 I_0 对 I 的导数:

$$w = \frac{\partial S_c}{\partial I} |I_0 \quad (2-2)$$

w 也被称作显著性图像, 因为它反应了对哪些像素进行微小的改动就可以对分类得分带来最大的改变。具体在卷积网络中, 计算类显著性映射 $M \in \mathbb{R}^{m \times n}$ 的计算过程如下: 1) 输入图像, 使用反向传播得到导数 W , 显著性映射通过对 W 内元素重新排列, 对于一个灰度图像, W 中的元素数量等于输入图像 I_0 的像素数量, 显著性映射可以通过 $M_{ij} = |w_{h(i,j)}|$, $h(i,j)$ 为输入图像 i 行 j 列位置对应的 W 元素的下标。对于彩色图像, 选取同位置所有通道的最大值 $M_{ij} = \max |w_{h(i,j,c)}|$ 。

得到的可视化图像如图 3-3:

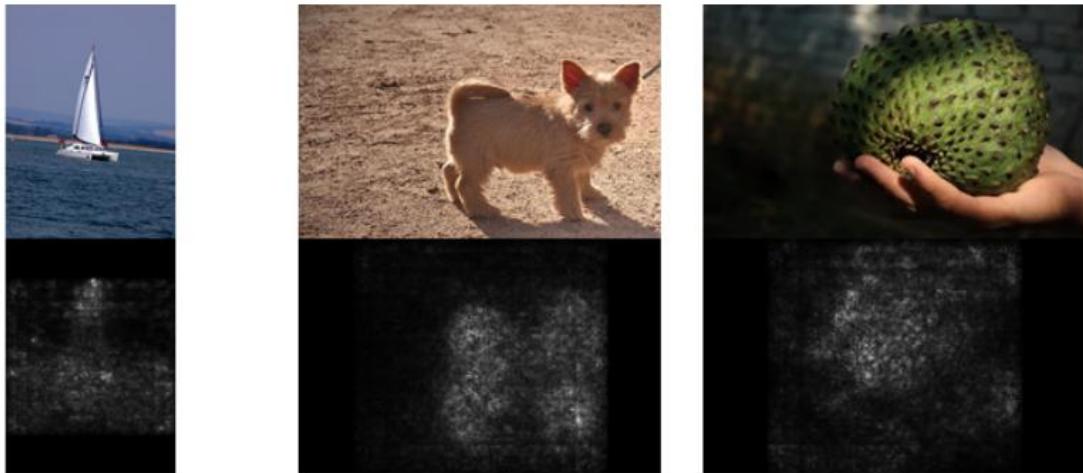


图 3-3 反向传播可视化效果

2.1.3 基于导向反向传播的神经元可视化方法

Springenberg 等人^[15]在提出全卷积网络的同时, 为了测试网络性能提出了一种对反卷积可视化方法的改进——导向反向传播, 用于对 CNN 卷积核特征可视化。

为了在没有池化层的深层网络中重建一张更锐利、可识别的特征可视化图像，作者提出了“反卷积”的一种改进，主要过程是网络的一次反向传播，它与反卷积的不同点在于当传播到 ReLU 非线性激活层时，反卷积方法的梯度仅会基于顶部梯度负信号进行整理，而忽略底部输入的负信号，因此重构的特征有大量噪声；而在反向传播中则仅根据底部特征的负信号进行整理，而忽略顶部负信号。作者结合这两种方法，不仅仅将上层梯度（“反卷积”）的负项置零，还将低层输入的负项置零。这种方法称为导向反向传播，因为它在常规的反向传播上增加了一个额外的引导信号，这阻止了对应神经元的负梯度的传播，这些负梯度可能会减小在可视化的高层神经元的激活，因此导向反向传播在更深层的网络中的效果要比前两种方法更好。

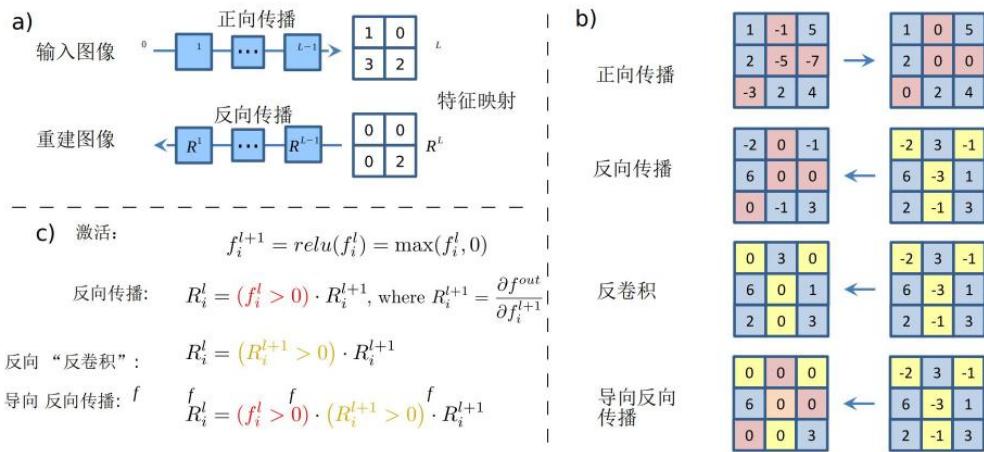


图 3-4：反向传播、导向反向传播、反卷积计算方式的区别。a) 给定一个输入图像，进行正向传播到达可视化的目标网络层，将其他的激活值置 0 后，反向传播至输入层得到重构图像。b) 反向传播时通过 ReLU 非线性层的不同方法。c) 在层 l 通过 ReLU 单元对一个激活进行反向传播得到激活输出的定义；注意“反卷积”方法和导向反向传播不计算真正的梯度，而是一个估算的版本。

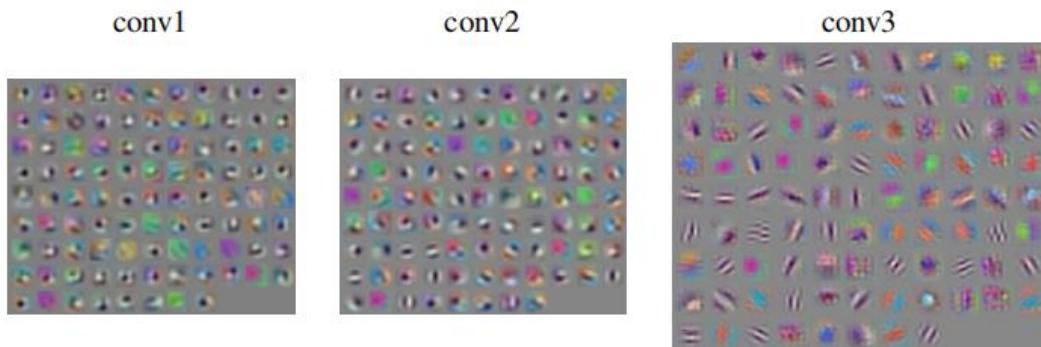


图 3-5：导向反向传播可视化示例：对 ImageNet 上训练的网络的低层（conv1–conv3）学习到的模式进行可视化，每个图块对应与一个滤波器。有趣的是 Gabor 滤波器只在第三层出现。

2.1.4 基于积分梯度的神经元可视化方法

Sundararajan 等人^[16]对基于梯度的 CNN 可视化方法提出了两点性质：1) 敏感性（Sensitivity）：输入和基准有一个特征不同，而二者的预测结果却不同，那么这个特征应该有非零的归因（梯度不应该为 0）；2) 实现不变性：当两个网络对所有的输入，结果对应都相同时，两个结构不同的网络达到功能一致性。一个正确的特征可视化方法应满足这两点，而之前的基于梯度的方法不能满足敏感性，DeepList 和 LRP 等方法又不能满足实现不变性，因此作者提出了积分梯度（Integrated gradient）同时满足这两点。

积分卷积定义，假设有一个方程 $F: R^n \rightarrow [0,1]$ 代表神经网络，令 $x \in R^n$ 为输入， $x' \in R^n$ 为基准输入，对于图像网络，基准输入可以是一个全黑的图像，对于文字模型，基准输入可以是全 0 的向量。通过计算基准输入 x' 到输入 x 的直线路径上所有点的梯度，并将它们累加，可以获得积分梯度。也就是说，积分梯度是从基准输入到输入的直线路径上的梯度积分。对输入 x 和基准输入 x' 的第 i 维的积分梯度可以被表示为：

$$\text{IntegratedGrads}_i(x) := (x_i - x'_i) \times \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \quad (2-3)$$

在实现中通过沿着基准输入到输入的直线路径上插值进行近似计算，积分梯度近似表示如下：

$$\text{IntegratedGrads}_i(x) := \frac{(x_i - x'_i)}{m} \times \sum_{k=1}^m \frac{F(x' + \frac{k}{m} \times (x - x'))}{x_i} \quad (\text{公式 2-4})$$

积分梯度可视化结果如下图：

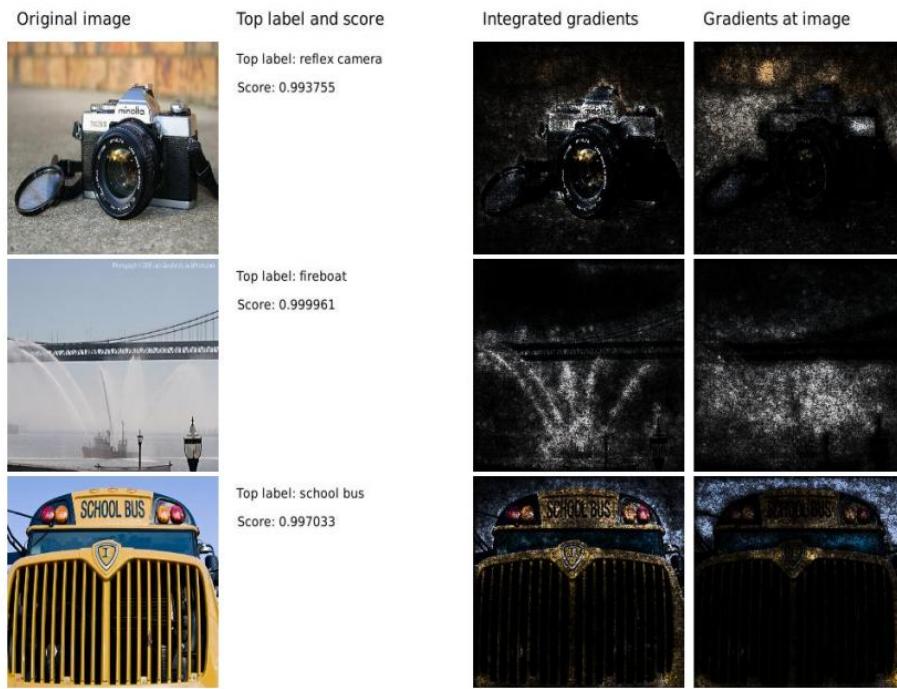


图 3-6: 积分梯度可视化结果, 三列图像分别为原始图像、积分梯度、图像梯度。

2.1.5 神经元可视化方法比较

本节对上述四种神经元可视化方法进行原理和效果两种比较。原理的比较如表 3-1 所示。表中对它们的原理、优缺点进行比较。

表 3-1: 神经元可视化方法原理、优点、缺点比较表格。

可视化方法	原理	优点	缺点
反卷积	将池化、卷积等操作进行转置，从单个神经元逐步上采样重构输入图像。	基于转置卷积，是原卷积网络的逆网络，可以获得任意特征映射的显著性图像。	仅滤去顶层梯度的负值，显著性图像噪声较多，有栅格现象。
反向传播	基于输入图像对特定类别计算其反向传播梯度作为该类别的显著性图像。	计算简单，仅需令网络对输入图像进行反向传播就能得到可视化图像。	仅滤去低层激活的负值，显著性图像噪声较多。
导向反向传播	在常规的反向传播上增加了一个额外的引导信号，这阻止了对应神经元的负梯度的传播。	同时滤去了顶层梯度和低层激活的负值，噪声较少，图像更锐利可识别。	不同类别的显著性图像区分度不大。
积分梯度	计算基准输入 x' 到输	同时满足了敏感性和	需要对输入进行插值

	入 x 的直线路径上所有点的梯度，并将它们累加，可以获得积分梯度。	实现不变性，显著性图像归因更准确，可辨识特征更明显。	计算多次梯度，计算量较大。
--	-------------------------------------	----------------------------	---------------

为了比较四种可视化方法的效果，选择数据集：MNIST, CIFAR-10, ImageNet 数据集；选择模型 ResNet-18；实验的过程为分别输入数据集中的单个样本，对其分类层使用简单反向传播、反卷积、导向反向传播、积分梯度可视化方法。四种可视化效果图，如图 3-7 所示。三行图像分别代表 ImageNet、CIFAR-10、MNIST 三种数据集的原始样本图像、简单反向传播、反卷积、导向反向传播、积分梯度可视化结果。通过对比观察可以发现，简单反向传播、反卷积、导向反向传播、积分梯度可视化方法的可视化效果递增，简单反向传播目标主体较不清晰，反卷积方法得到的图像存在一定栅格状特征，导向反向传播和积分梯度方法结果较为相近，积分梯度算法叠加了多张图像，主体更为清晰。

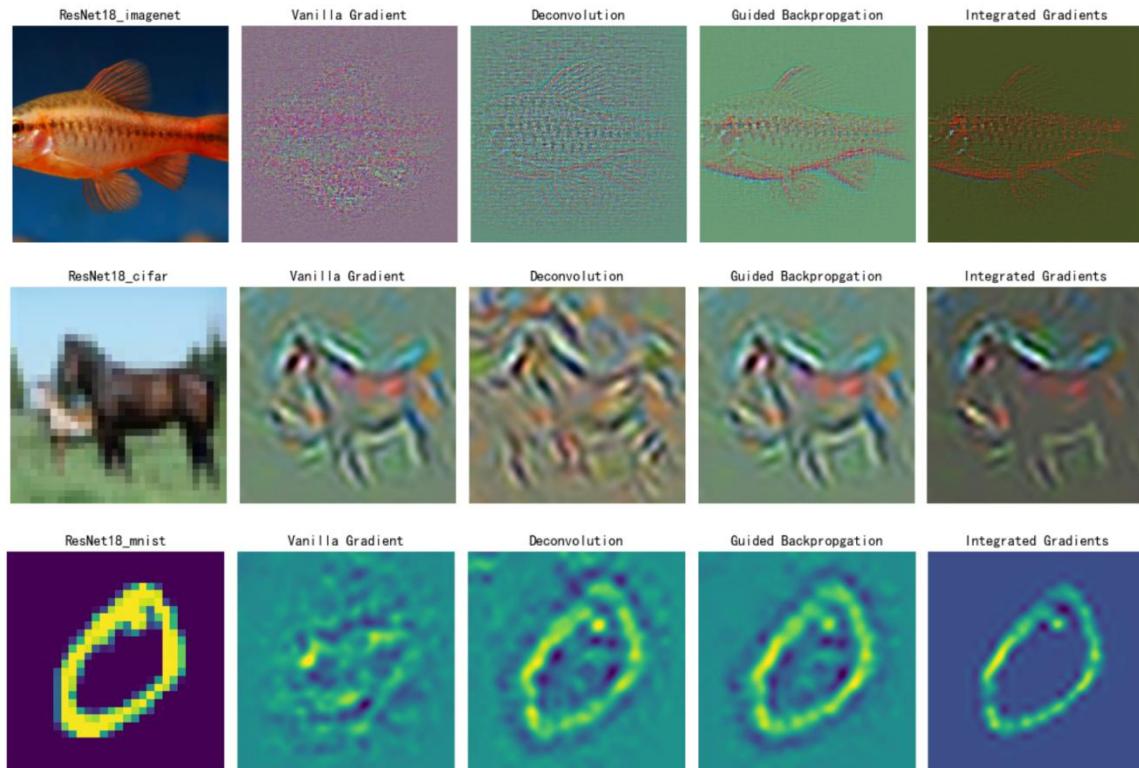


图 3-7：简单反向传播、反卷积、导向反向传播、积分梯度神经元可视化方法效果。

2.2 最大化激活算法

最大化激活的思想来自于医学家 David Hubel 和 Torsten Wiesel 对一种被称为方向选择性细胞的研究，当动物被展示拥有特定边缘的物体时，神经系统中特定的方向选择性细胞就会兴奋，也就是说边缘中的特定特征激活了这些细胞。对于一个卷积神经网络的神经元，也可以寻找令其激活最大的图像，该图像中的语义信息可以用来解释该神经元学习到了哪些特征。深度神经网络的最大化激活可视化方法总体上可以被分为两大类：网络驱动的方法和数据驱动的方法。网络驱动的方法令一个训练完毕的网络直接生成令对应单元激活最大的数据，这不基于数据集中的数据。而数据驱动的方法需要让网络遍历整个数据集，寻找令单元激活最大或最小的数据。

2.2.1 网络驱动的最大化激活算法

最大化激活的目标是寻找可以令一个给定隐层神经元激活最大的有界范数的输入模式，这个想法背后的逻辑是，令一个单元响应最大的模式可能可以很好地解释这个单元的作用。完成这个想法的最简单的方法是给定一个单元，然后在所有样本中寻找令这个单元激活最高的一些样本。但是，应该保留多少样本、如何组合这些样本难以决定。因此网络驱动的方法不会从训练集或测试机中寻找样本，而是用一种更通用的方法——从一个优化问题的角度令单个单元激活最大。令 θ 表示网络参数（权重和偏置），令 $h_{ij}(\theta, x)$ 表示一个给定网络层 j 的单元 i 的激活， h_{ij} 是参数和输入样本 x 的函数。假设参数 θ 是固定的，优化的目标是：

$$x^* = \operatorname{argmax}_{x \text{ s.t. } \|x\|=\rho} h_{ij}(\theta, x) \quad (2-5)$$

这在本质上是一个非凸优化问题，但是至少可以寻找到一个局部最小值，通过对输入空间使用简单的梯度下降方法做到这一点：令 x 向 $h_{ij}(\theta, x)$ 的梯度方向移动， $h_{ij}(\theta, x)$ 的值将逐渐增大。当使用不同的初始化时存在两种情况，第一种是不同的初始化的优化结果都是同样的最小值，另一种是找到不同的局部最小值。对于后一种情况，一种方法是将结果进行平均处理，另一种是选取其中令激活最大的局部极小值，或者将所有的局部极小值都用来描述这个神经元。这个优化方法被称作最大化激活，可以被应用于任意允许计算梯度的神经网络，所需的超参数包括学习率和一个停止指标，否则激活将一直增大。

2.2.2 数据驱动的最大化激活算法

上述的方法以网络为中心，由网络参数优化得到一个生成样本，不需要实际训练数据或测试数据参与。而数据驱动的方法则是一个更简单直观的方法：令网络遍历数据集，捕捉目标单元的激活值，寻找令目标单元激活值最高的样本。这个技巧被广泛应用到卷积网络可解释研究中，例如将不同颜色、纹理、方向的数据集输入网络后，观察神经元对应的激活情况，使用令其最大激活的单张/top K 张图像描述该神经元学习到的特征。

2.3 图像相似度指标

在很多情况下，我们需要计算两张图像的相似度，图像相似度的计算方法主要包括传统算法和基于深度学习的算法。在本项目中，对于相似度指标的要求是：1) 计算快捷，对算法的时空复杂度低；2) 准确，尽可能反映两张图像的相似程度。因此本部分基于时空复杂度不考虑使用深度学习算法，只探究了传统的图像相似度指标。

2.3.1 基于传统方法的图像相似度指标

(1) 欧式距离 (Euclidean Distance)

指在 m 维空间中两个点之间的真实距离或者向量的自然长度，公式如下：

$$d(A, B) = \sqrt{(A - B)^2} \quad (2-6)$$

(2) 余弦相似度

通过计算两个向量的夹角的余弦值来度量它们之间的相似性，向量方向相同余弦值为 1，相反为 -1，余弦相似度更注重两个向量在方向上的差异性，而非在距离或长度上，常用于高维正空间，公式如下：

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n A_i^2 \times \sum_{i=1}^n B_i^2}} \quad (2-7)$$

(3) 汉明距离

在数据传输差错控制编码中，表示两个长度相同的字符串对应位置的不同字符的数量，对两个字符编码进行疑惑运算，统计 1 的个数，即汉明距离。

(4) 直方图距离

计算图像灰度值的直方图，计算两张图像的直方图的距离。缺点：不包含图像

的空间信息，仅仅是明暗差异的图像，直方图距离却很大，纹理不同，明暗相近的图像，相似度却很高。

(5) 感知哈希算法

对每张图像缩小尺寸，将每个像素与平均值比较形成一系列 0 与 1，将这个结果组合构成了一组整数，得到哈希值，也就是图像的“指纹”。对比不同图像指纹的差异来比较图像间的相似性。

2.3.2 图像相似度指标比较结论

欧式距离是最常用的距离指标，但是在图像中很多情况下由于图像亮度不同，可能出现非常相似的图像的欧式距离却很高的情况。汉明距离将每个像素进行比较相同才为 1，在图像中像素完全相同的情况很少。直方图距离信息包含量不足，忽略了图像的空间信息。感知哈希算法需要计算图像指纹，相比余弦相似度更复杂，并且其结果不够直观。

根据比较，余弦相似度最适合本项目，因为 1) 算法简单，计算量小；2) 余弦相似度对向量绝对值不敏感；3) 相似值区间为 -1~1，易直观理解与比较。

2.4 本章小结

本章对本项目可能使用的各类算法、方法进行了调研。本项目的核心部分为神经元可视化方法，2.1 节部分阐述了四种神经元可视化方法的算法原理，展示了其可视化效果，并在最后对比分析了四种方法的优缺点，并在实际预训练模型中进行了对比测试，本项目选取导向反向传播作为神经元可视化方法的基础算法。2.2 节阐述了两类最大化激活算法的原理——网络驱动的最大化激活算法和数据驱动的最大化激活算法，二者都是解释神经网络的重要方法，并作为一种可视化技巧被应用到多重可解释方法中。2.3 节调研了基于传统方法的图像相似度指标，并针对本项目的需求进行了对比分析，最终得出选择余弦相似度的结论。

第3章 基于类平均图的可视化图像相似值计算方法研究

3.1 总体设计

本研究的核心思想是利用数据类平均图为神经元可视化方法提供衡量指标，将可视化结果量化，促进对可视化结果的理解和进一步的应用。具体地，类平均图由数据集提炼而出，包含数据集中的信息，可以作为该数据类别的“代表”。将类平均图作为神经元可视化的输入模式，可以获得每个神经元对该数据类别“代表”的特征可视化图像，神经元特征可视化结果中包含着该神经元对类别“代表”提取的信息，可视化结果的好坏反应该神经元对该数据类别信息提取程度的水平，因此可视化结果与数据类别“代表”的相似度可以用来估计该神经元对该数据类别的特征学习情况。

本研究的算法主体部分为：计算数据类平均图像、基于类平均图像计算神经元可视化图像、比较类平均图像与神经元可视化图像的相似度、基于神经元相似值对网络的性能进行分析。本研究提出了多种计算类平均图像的技巧：简单类平均图像、数据驱动和络驱动激活最大化的类平均图；同时，本研究也提出了神经元可视化图像相似值的两种比较计算方式：基于类别平均图的相似度和基于类置信度可视化图像的相似度，两种相似度对信息提取分析的角度不同。具体内容将在下文中阐述。

3.2 类平均图的设计与计算方法

“平均脸”是对一批相同角度、光照条件下的人脸图像计算其平均值得到的一张图像，这张合成的人脸包含了这批人脸图像的共同特征，显得精致而美观。在计算机视觉的人脸识别领域，Sirovich 和 Kirby 使用 PCA 降维方法使用一个低维度的特征脸（Eigenface）代表一张特定的人脸图像，在算法中，他们对特定的图像减去人脸数据集的均值，以此除去人脸数据的共性信息，得到其独特信息来进行后续的识别。基于这些平均图像的应用可以了解到平均数据可以由一个数据集计算得到，其中蕴含了这批数据的共同信息，因此可以作为这批数据的代表图像。

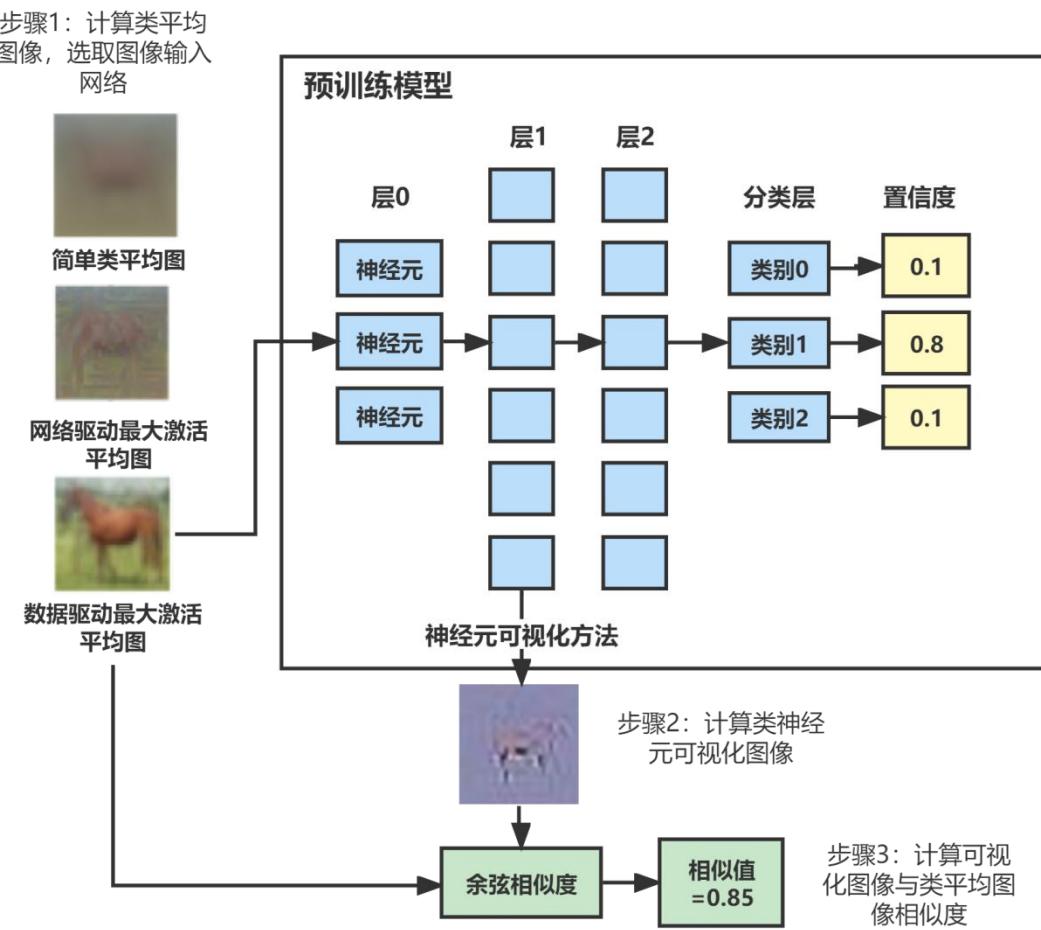


图 3-8：神经元相似值算法流程图。

在基于平均图的神经元相似值分析算法中，类平均图像起到了关键作用——它是神经元对数据类信息提取能力学习情况的衡量基准。数据集本身的性质影响着类平均图像的效果，数据类内方差小的图像——主体清晰、相似、目标占比大的图像得到的平均图像更清晰可辨，而类内方差大的图像——主体较小，背景占比多的图像得到的平均图主体难以辨别，并且尤其当数据集中图像数量较多时，得到的平均图像十分模糊，难以起到代表共同特征的作用。因此本部分除了简单类平均图像外，还设计了基于网络驱动和数据驱动的激活最大化方法来获取对网络价值更高的类平均图像。

3.2.1 简单类平均图的计算方法

通常图像分类任务的数据集包含每个图像类别的数据子集。类平均图像的目的

是用作单类数据子集的“代表图像”，涵盖该数据子集的共有信息。简单类平均图像是一种直观的方法，即对单个数据子集中所有的样本计算均值图像。对于共有 N 类的数据集， $C_N = \{C_i | i = 0, 1, \dots, N\}$ 表示所有类的数据， C_i 表示第 i 类数据子集，每个数据类都包含 n 个样本，每个数据子集的简单类平均图像定义为：

$$\Phi_N = \{\Phi_i | \Phi_i = \frac{1}{n} \sum_{j=1}^n x_j, x_j \in C_i\} \quad (3-1)$$

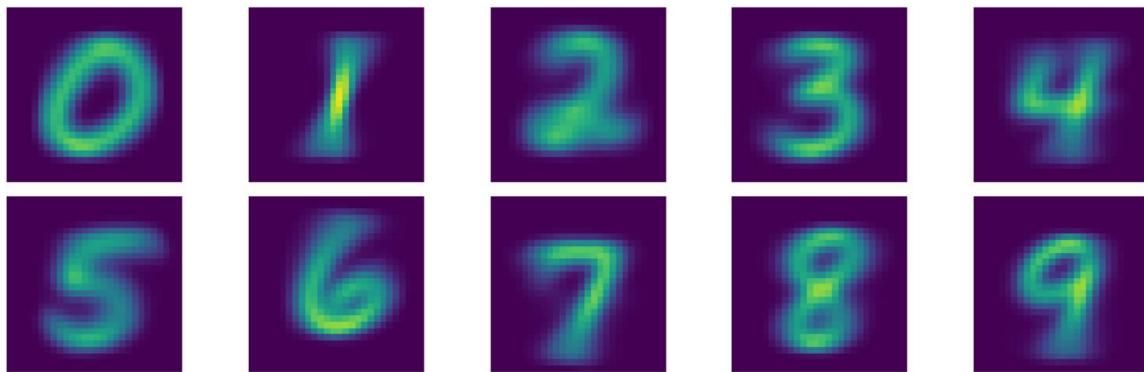


图 3-9：MNIST 数据集的简单类平均图像。

在某些情况下，如数据子集包含的样本较多且类内方差较大，例如 ImageNet 训练集，共包含 1300 个训练样本，其中目标主体位置、大小差异较大，计算得到的简单平均图像出现过平均化的问题——图像模糊、主体不明显、图像像素值相比原数据整体偏低。针对图像像素值整体偏低的情况，使用最小最大正则化方法（min-max normalization）将类平均图像的每个通道按原数据子集通道的最值与当前通道最值的比例进行放大。由此可以避免出现因为类平均图像像素值过低导致神经网络难以激活的问题。



图 3-10: ImageNet 数据集的简单类平均图像。

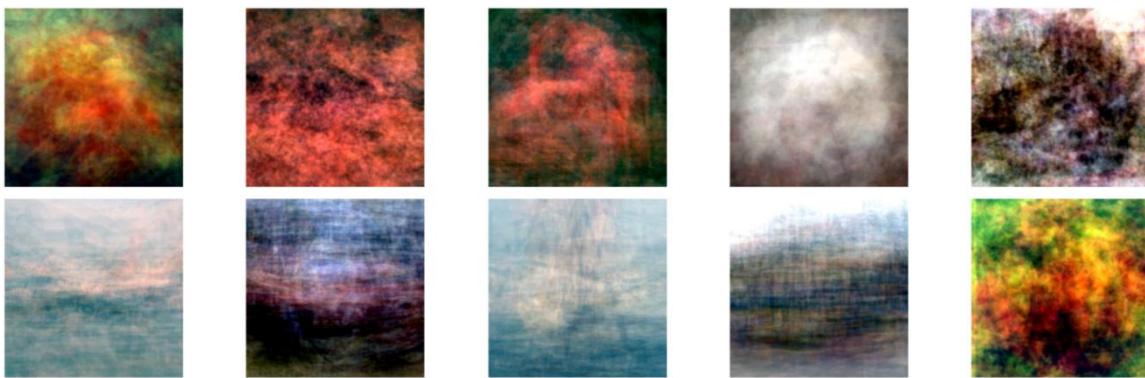


图 3-11: ImageNet 数据集的简单类平均图像经过最小最大正则化后的结果: 相比正则化前, 图像主体更明显, 并且像素值被放大了近 10 倍。

3.2.2 数据驱动激活最大化的类平均图计算方法

理论上简单类平均图像代表了一个数据子集的公有特征, 但是在实际应用中, 简单类平均图作为一批数据的均值图像, 很容易出现过平均化的问题, 尤其在如 ImageNet 这样的真实世界数据集中, 目标的形状、大小、位置多变, 直接计算得到的均值图像并不能令神经网络良好地激活。理论上, 类平均图像体现了一类数据样本的公有特征, 如果将其输入用于分类的卷积神经网络, 网络将以较高的置信度将其判断为对应的类别。但是在实际情况中, 网络并不能很好地将类平均图像判断为响应的类别, 而在很多情况下将其判断为其它类别, 或是将多个类平均图像判断为某一个类别。例如, 将 MNIST 数据的 10 张类平均图像输入 AlexNet 预训练模型后, 观察每张图像输入后网络的 10 个类别的分类置信度热力图(图 3-12), 可以发现网络将多个平均图判断为了类别 1 和类别 8, 而不是将类平均图像判断为对应的类别(即热力图对角线上置信度最高)。

数据驱动的激活最大化算法的基本原理是使用令网络单元激活值最高的一批数据样本描述该网络学习到的特征。神经元相似值分析算法的目标是利用平均图像对数据的代表能力分析神经网络对数据信息的学习情况, 而非简单地寻找令网络激活最大的图像, 因为单张图像是特异的, 不具备足够的代表性。因此, 本部分基于数据驱动的激活最大化类平均图的计算方式是寻找令网络激活最大和最小的前 K 张图像, 分别均值化生成两张类均值图像。具体地, 将单类数据子集的所有样本输入

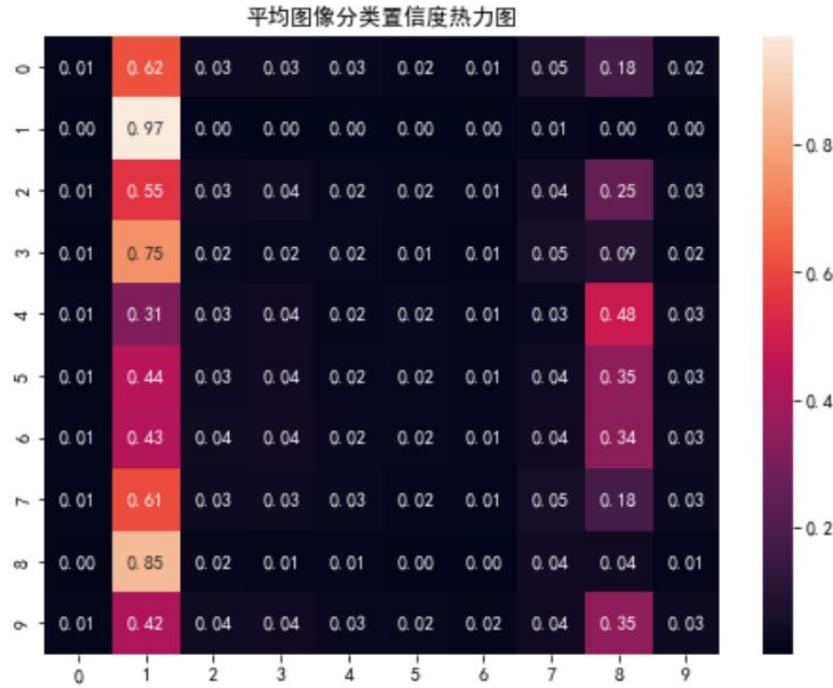


图 3-12：分类置信度热力图：MNIST 数据的 10 张类平均图像在 AlexNet 预训练模型的分类置信度热力图，纵坐标为类平均图像编号，横坐标为网络的分类置信度，例如位置(0, 1)=0.62 表示网络将第 0 张类平均图像判断为类别 1 的置信度为 0.62。

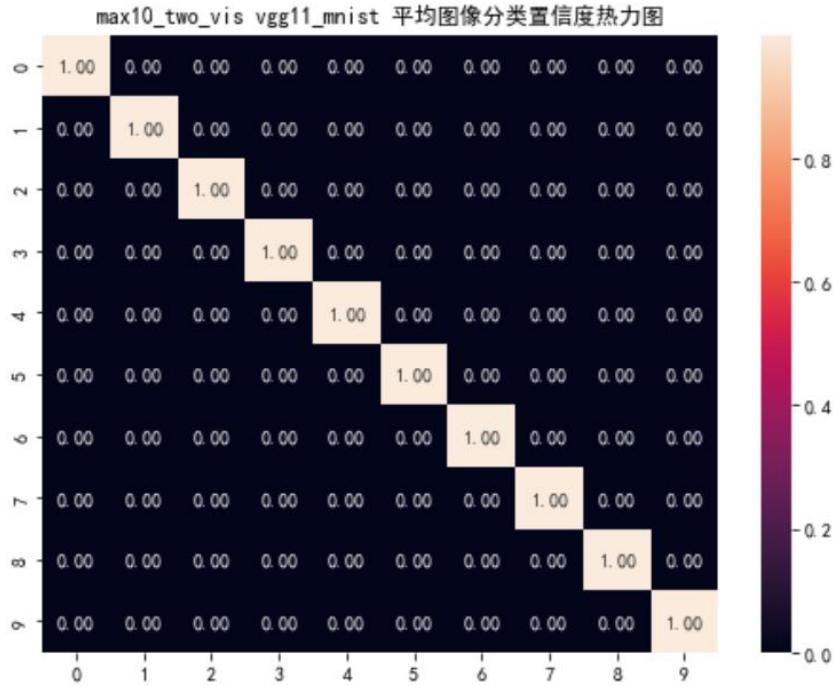


图 3-13：分类置信度热力图：MNIST 数据的 10 张最大激活类平均图像在 AlexNet 预训练模型的分类置信度热力图。

网络，获取其对应类别的分类置信度，排序后分别选取置信度最小和最大的前 K 张图像，分别将其均值化，得到最小激活类平均图和最大激活类平均图。将使用数据驱动激活最大化优化的类平均图像输入网络，其分类置信度热力图如图 3-13 所示，热力图对角线都为 1.00，这代表网络可以良好地将激活类平均图判断为对应的类别。其原理是最大激活图像代表网络充分学习到的特征，而最小激活图像代表网络难以充分学习的特征，将二者进行对比分析可以更全面地分析网络的性能。其中超参数 K 用来控制激活类平均图的信息覆盖范围，K 过小，激活类平均图只能代表特殊的图像；K 过大，激活类平均图将与简单类平均图相近。

3.2.3 网络驱动激活最大化的类平均图计算方法

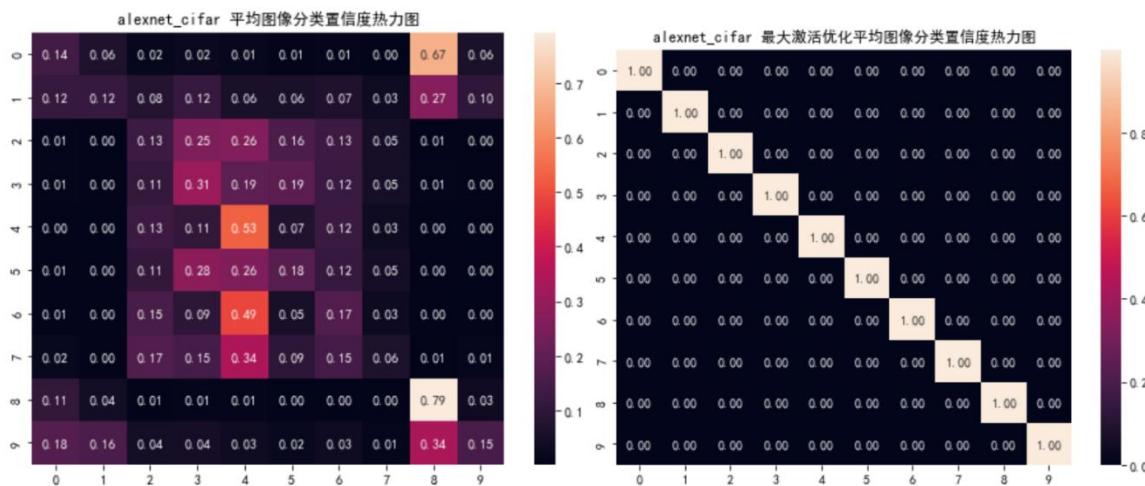


图 3-14：CIFAR-10 数据的 10 张类平均图像，最大激活优化前后在 AlexNet 预训练模型的分类置信度热力图（左图为优化前，右图为优化后）。

网络驱动的激活最大化算法同样利用了激活最大化的思想，其区别是网络驱动的方法以网络为中心，以令网络的某个单元的激活值增大为目标，通过反向的梯度下降优化输入的单个样本，直到优化达到特定次数或者激活值达到某一阈值为止。由于简单类平均图的过平均问题，数据类中的信息被平均后过于分散，像素值较低而且相近，通过网络驱动的激活最大化算法可以通过网络将分类所需要的低频信息（纹理等细节信息）补充回原图像，由此网络可以将其正确地分类，如图 3-14 所示，经过优化，网络可以良好地判断平均图的类别。但是本研究的目标是将类平均图像用作解释网络特征的指标，网络驱动的激活最大化方法存在一系列问题不能达

到这一目的：如果使用以网络为中心优化得到的图像则发生网络解释自身的逻辑问题；而如果像 3.2.2 小节分别使用最大激活和最小激活两种类平均图像，优化得到的最小激活图像将出现大量负值或者零值的情况，不具备解释能力；优化过程需要较长时间，本研究不具备大量生成优化图像的条件。因此本部分的类平均图像近作为参考，而不真正应用到解释网络中。

3.3 神经元可视化图像的相似值计算方法

3.2 部分提出了三种计算类平均图像的方法，基于梯度的神经元可视化方法要求输入一张基础图像用作计算梯度的依据，因此选择图像尤为重要，在过去的方法中，研究人员通常通过基于梯度的神经元可视化结果——显著图（saliency map）分析原输入图像中的高价值部分，由于单张图像的局限性，难以对网络的性能进行更多的分析。类平均图像作为数据类的代表图像，有丰富的类公有信息，因此可以作为神经元可视化方法的良好输入。然后通过比较神经元可视化结果和代表数据的类平均图像之间的相似度，可以获得衡量该神经元对该数据类别的信息提取能力指标——相似值。

3.3.1 方法一：基于类别平均图的相似度

在 2.1 部分对神经元可视化算法的调研中，经过对四种可视化算法的比较，本研究确定使用导向反向传播神经元可视化算法。导向反向传播算法首先正向传播输入图像，到达被可视化神经元后，进行反向传播，在这个过程中负梯度等非必要信息会被 ReLU 层过滤，最终到达输入层得到与输入图像尺寸相当的可视化图像。对于一个图像分类预训练网络，共有 N 张类平均图像 I_N ，网络层数为 L ，每个网络层的滤波器数量为 F （非定值），GBP 表示导向反向传播算法，基于平均图的神经元相似值算法将计算出所有神经元对应类平均图像的可视化图像：

$$\Omega = \{ \omega_{ijk} \mid \omega_{ijk} = \text{GBP}(I_i, j, k), i = 0, 1, \dots, N, j = 0, 1, \dots, L, k = 0, 1, \dots, F \} \quad (3-2)$$

在获得了类平均图像 I_i 和神经元可视化图像 ω_{ilf} 后，需要计算二者的相似度，经过 2.3 部分对图像相似度指标的调研，选择余弦相似度指标，相似值 S 定义为：

$$S = \{ S_{ijk} = \frac{I_i \cdot \omega_{ijk}}{|I_i| \times |\omega_{ijk}|} \mid i = 0, 1, \dots, N, j = 0, 1, \dots, L, k = 0, 1, \dots, F \} \quad (3-3)$$

由此每个网络层中的每个神经元都被赋予了 N 个相似值来描述它们对数据集信息的学习情况。

3.3.2 神经元相似值计算方法二：基于类置信度可视化图像的相似度

虽然余弦相似度有着计算复杂度低、不受比较对象绝对值大小影响的优点，但是由于类平均图像与神经元可视化图像之间差异很大，在部分情况下，较简单的余弦相似度算法可能难以准确地衡量两个差异巨大的图像之间的相似性，即使二者可能有相近的高语义信息。因此本小节通过改变余弦相似度的比较对象，提出了另一种神经元相似值计算方法：比较类平均图像基于网络分类层置信度的可视化图像与神经元可视化图像之间的余弦相似度，二者都是使用相同可视化算法得到的可视化结果，这减小了由于图像本身种类差异带来的相似度比较难度。

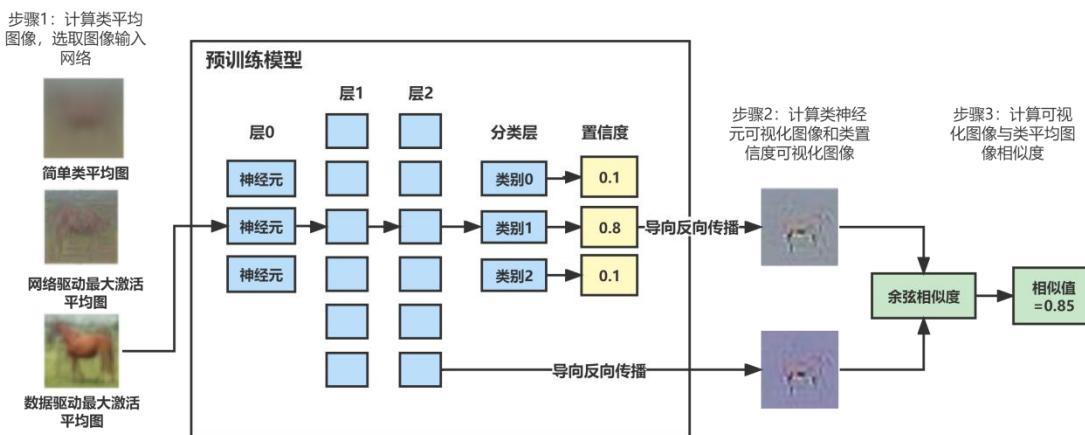


图 3-15：基于相似值计算方法二的神经元相似值算法流程图（与方法一的区别在于步骤 2 将会计算两个导向反向传播可视化图像）。

本部分的神经元可视化图像计算方式与方法一相同，所有神经元对应类平均图像的可视化图像表示为：

$$\Omega = \{ \omega_{ijk} \mid \omega_{ijk} = GBP(I_i, j, k), i = 0, 1, \dots, N, j = 0, 1, \dots, L, k = 0, 1, \dots, F \} \quad (3-4)$$

而相似值比较的另一个对象更改为类平均图像基于网络分类层置信度的可视化图像 ϕ_i ，其原理是对原始的第 i 类别的类平均图像 I_i 计算网络最后的分类层中计算对应类别置信度的神经元 i 的可视化图像：

$$\phi_i = \{ GBP(I_i, N, i) \mid i = 0, 1, \dots, N \} \quad (3-5)$$

由此，网络层 j 的第 k 个神经元对应类平均图 I_i 的可视化图像 ϕ_i 的余弦相似度表示为：

$$S = \{s_{ijk} = \frac{\phi_i \cdot \omega_{ijk}}{|\phi_i| \times |\omega_{ijk}|} \mid i = 0, 1, \dots, N, j = 0, 1, \dots, L, k = 0, 1, \dots, F\} \quad (3-6)$$

相比方法一，由于方法二是比较两个由相同算法得到的可视化图像，相似值的计算不会受到图像本身性质差异的影响，所以理论上计算得到的神经元相似值整体更高。从更深层次的意义来看，输入图像经过神经网络层层的特征提取，网络最后一层包含的特征支持着网络最终的分类功能，而类置信度的可视化图像将这一部分特征展示出来，将其他层的神经元可视化图像与其比较，可以判断网络在哪些层已经达到了支持分类的信息条件，由此可以判断网络的冗余性。

3.4 本章小结

本章节介绍了基于类平均图的可视化图像相似值算法的总体设计以及具体流程。首先提出了类平均图像的三种设计，简单类平均图理论最适合本研究，但是在真实数据集中易出现过平均化导致图像主体不清晰的问题；数据驱动的激活最大化类平均图像从最小激活和最大激活两个角度描述数据子集；网络驱动的激活最大化类平均图像虽然可以带来极高的网络分类置信度，但是却不适合用于解释网络性能。接着，本章节提出了两种神经元可视化图像的相似值计算方法，二者都基于类平均图像，但是方法一直接使用类平均图像与神经元可视化图像计算相似度，而方法二使用类平均图像基于分类置信度的可视化图像与神经元可视化图像计算相似度。方法二弥补了图像性质差异带来的影响，理论上会得到数值更高、更准确的相似值。

第4章 基于平均图的神经元可视化图像相似值的实验与分析

上一章节介绍了基于类平均图的可视化图像相似值算法，本章节将该算法应用到多组预训练模型上，并对结果进行分析。

4.1 预训练数据与模型准备

本研究的目的是分析卷积神经网络的可解释性，因此本实验部分选择了三个图像分类的经典数据集：MNIST 手写数字数据集、CIFAR-10 图像分类数据集、ImageNet 图像分类数据集。MNIST 手写数字数据集包含 70000 张手写数字图像，训练集与测试集大小分别为 60000, 10000，每张图像为 32×32 大小的单个手写数字。CIFAR-10 数据集包含 10 个类别，每类 6000 张 32×32 的彩色图像，每类训练集与测试集大小比例为 5:1。ImageNet 是一个大型图像分类数据集，包含超过 1400 万张图像，由于硬件限制，本实验选取了其中 10 个类别，每类数据的训练集和测试集大小分别为 1300 和 50，每张图像在实验中尺寸被设定为 $3 \times 224 \times 224$ 。

本实验的预训练网络结构选取了 CNN 的经典模型结构：AlexNet 和 VGG11。AlexNet 由 Hinton 等人设计，共有 5 个卷积层，模型的分类部分为两个全连接层。VGG 由 Oxford 的 Visual Geometry Group 的组提出，其卷积核大小统一被设置为 3×3 ，共包含 8 个卷积层以及用于分类的两个全连接层，由于 VGG11 网络较深下采样层更多，对于图像尺寸更小的 MNIST 和 CIFAR-10 数据集需要移除一个最大池化层。

本部分在三组数据上分别训练了两个模型结构，共获得了六个预训练模型。训练使用 Pytorch 框架，优化器为 Adam 优化器，学习率为 1×10^{-3} ，权重衰减为 1×10^{-5} 。训练时的超参数设置为：每个模型训练 30 个 epoch，CIFAR-10 和 MNIST 数据集批大小（batch size）为 128，ImageNet 数据集的批大小为 32。因为本实验的目的是分析网络性能，而非得到一个最好的网络，所以只保存 30 个 epoch 后的模型作为实验预训练模型，经过 30 个 epoch 的训练这些模型通常会发生过拟合问题，参考图 4-1，左为训练损失曲线，中为测试准确率曲线，右为测试损失曲线，

观察图像可以发现，在训练中训练损失逐步降低，而测试损失在 7 个 epoch 后逐渐上升，测试准确率除开始的上升外始终在 0.73 左右波动，网络发生了过拟合问题：

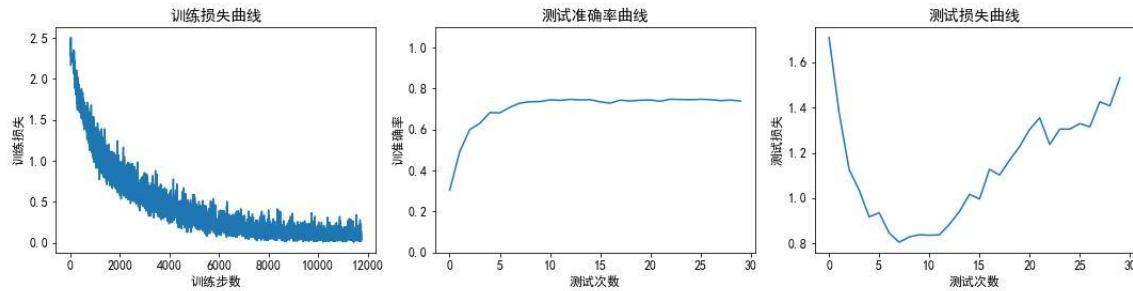


图 4-1：VGG11 在 CIFAR-10 数据集的训练过程曲线。

经过训练，六个预训练模型 AlexNet_MNIST, AlexNet_CIFAR, AlexNet_ImageNet, VGG11_MNIST, VGG11_CIFAR, VGG11_ImageNet 的训练损失、测试损失、测试准确率如下表所示：

表 4-1：AlexNet_MNIST, AlexNet_CIFAR, AlexNet_ImageNet, VGG11_MNIST, VGG11_CIFAR, VGG11_ImageNet 预训练结果表格。

预训练模型	训练损失	测试损失	测试准确率
AlexNet_MNIST	0.001	0.056	99.1%
VGG11_MNIST	0.004	0.049	99.1%
AlexNet_CIFAR	0.198	1.638	67.9%
VGG11_CIFAR	0.129	1.555	73.9%
AlexNet_ImageNet	0.125	1.002	81.4%
VGG11_ImageNet	0.312	0.993	73.0%

4.2 类平均图的计算结果

3.2 部分提出了三种类平均图像的设计方法，本部分分别将其应用到了 MNIST、CIFAR-10 和 ImageNet 数据集上。

4.2.1 简单类平均图的计算

简单类平均图像通过对单个类别的数据集计算均值图像，然后按照原数据集数据的最大最小值将均值图像按比例放大得到。通过对比观察可以发现，在图 4-2 中，

由于 MNIST 手写数字图像之间相似性高，目标主体占比大，由此计算出的简单类平均图像主体清晰可辨，且形状标准。而对于 CIFAR-10 和 ImageNet 数据集，二者的图像来自真实世界，原图像中主体状态角度多变，由此计算出的简单类平均图像十分模糊，个别图像可以勉强观察到主体，如图 4-3 第二行第三列的 CIFAR-10 简单类平均图像，为类别“马”。而在图 4-4 中，ImageNet 数据集的部分简单类平

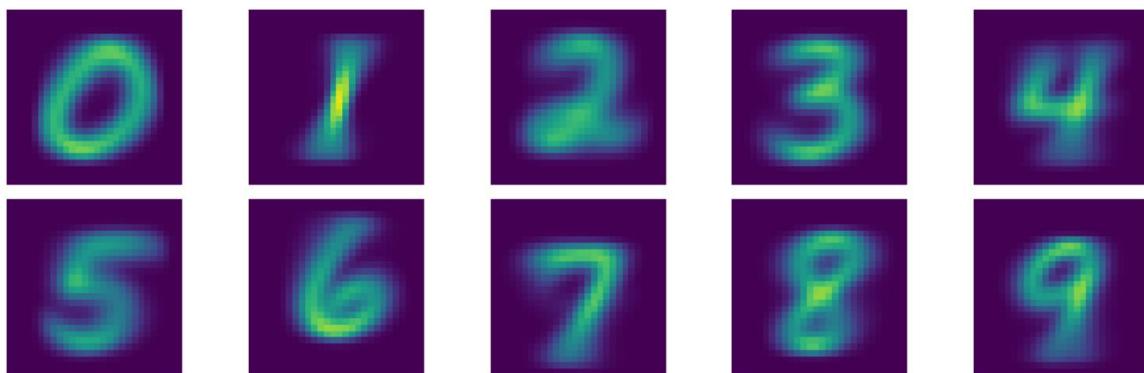


图 4-2：MNIST 数据集简单类平均图。

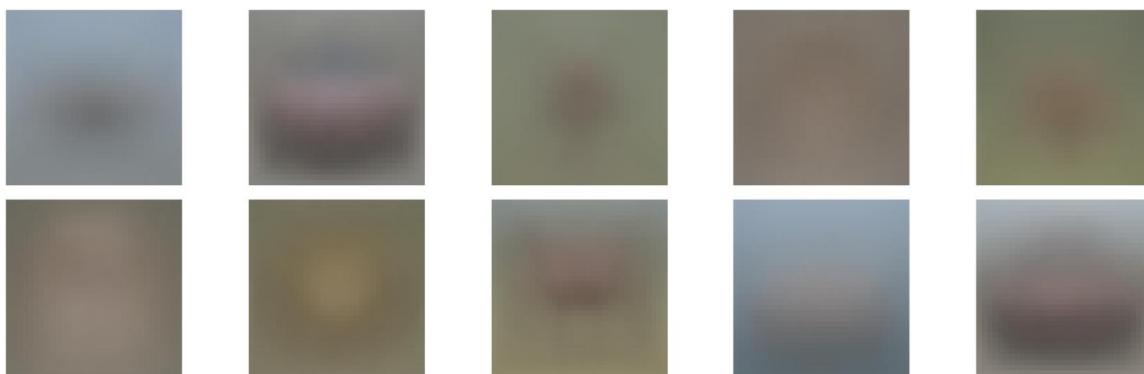


图 4-3：CIFAR-10 数据集简单类平均图。



图 4-4：ImageNet 数据集简单类平均图。

均图呈现出带颜色的圆形目标，这是因为目标在图像中心的概率更高，由于过度平均化，图像被叠加成为原目标颜色的圆形，正则化并不能改善这种问题。

4.2.2 数据驱动激活最大化的类平均图像的计算

数据驱动激活最大化的类平均图的计算方法是令网络遍历单类数据集，记录其对应类别的置信度，选取置信度最高和最低的前 K 张图像，分别计算它们的均值图像，得到最大激活类平均图和最小激活类平均图。本实验求取了六个预训练模型对应的最大和最小激活类平均图，由于篇幅限制，本部分仅展示和分析 MNIST 和 CIFAR-10 数据集由 VGG11 模型得到的类平均图像，以及 MNIST 类平均图在 VGG11 预训练模型的置信度热力图。

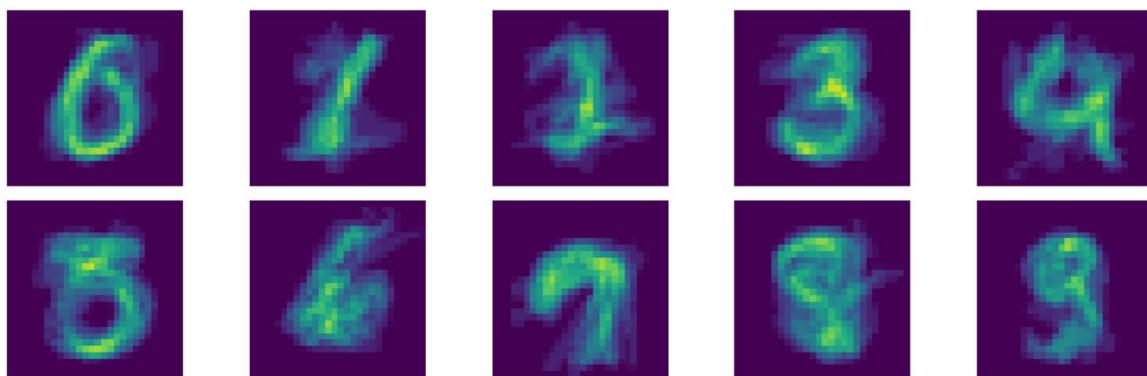


图 4-5：MNIST 数据集在 VGG11 预训练模型上的最小激活类平均图像。

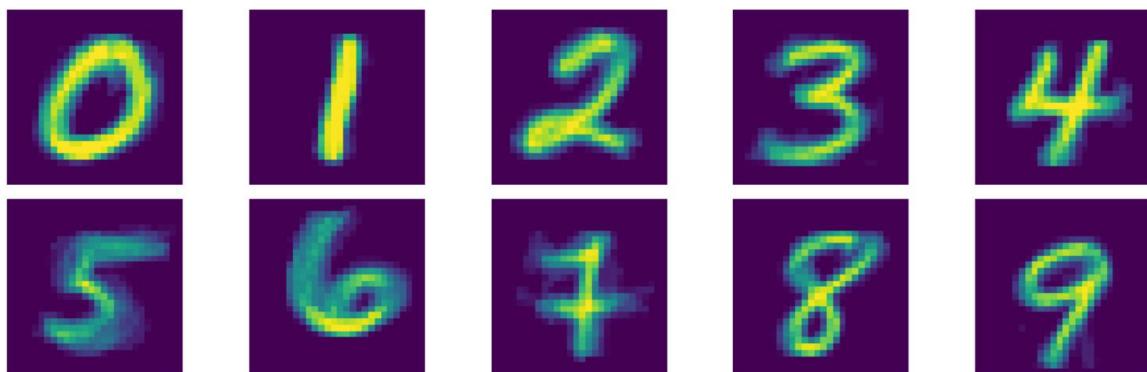


图 4-6：MNIST 数据集在 VGG11 预训练模型上的最大激活类平均图像。



图 4-7: CIFAR-10 数据集在 VGG11 预训练模型上的最小激活类平均图像。



图 4-8: CIFAR-10 数据集在 VGG11 预训练模型上的最大激活类平均图像。

通过对比最大和最小激活类平均图像可以发现，对于 MNIST 数据集，最小激活类平均图的数字较扭曲，而最大激活类平均图更标准，平均后也更清晰；对于 CIFAR-10 数据集最小激活类平均图并不能识别出主体，而最大激活类平均图中可以观察到主体轮廓，并且主体的位置较固定，呈现出一定的对称效果，例如图 4.4 第一行第三列的类别“鸟”。

进一步将得到的 MNIST 数据集得到的类平均图像输入 VGG11 预训练网络，测试其对应类别的置信度，可以发现：对于最小激活类平均图多数不能被网络正确地识别，只有数字 1 的类平均图像拥有较高的置信度；而最大激活类平均图多数可以被网络以高置信度判别，只有数字 7 的类平均图被判断为类别 1，这是因为数字 7 本身与数字 1 有较多的公有特征。

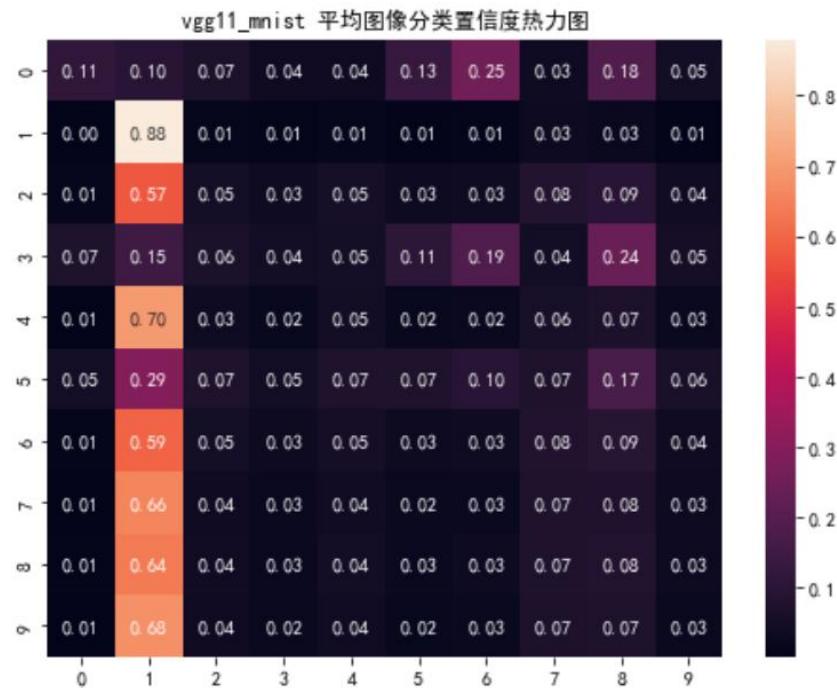


图 4-9: MNIST 数据的最小激活类平均图像输入 VGG11 预训练模型后, 网络的分类置信度热力图: 纵坐标为类平均图像编号, 横坐标为网络的分类置信度。

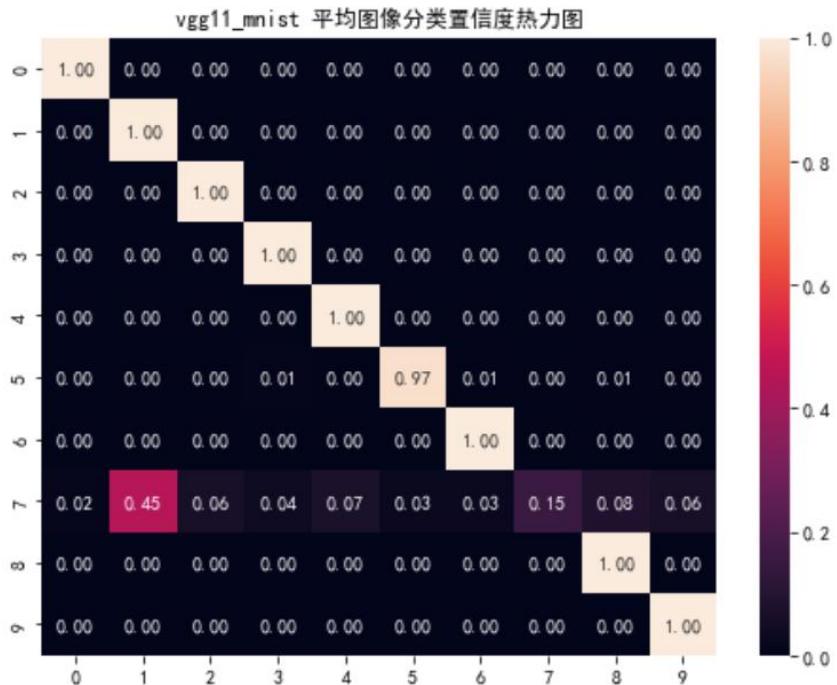


图 4-10: MNIST 数据的最大激活类平均图像输入 VGG11 预训练模型后, 网络的分类置信度热力图: 纵坐标为类平均图像编号, 横坐标为网络的分类置信度。

4.2.3 网络驱动激活最大化的类平均图像的计算

网络驱动的激活最大化类平均图是以分类置信度为目标，通过梯度上升方法优化简单类平均图，最终使网络可以以高置信度将该类平均图判断为对应类别。在这个过程中，网络将自身的信息补充到了简单类平均图上，因此不具备作为指标解释网络的能力。本实验选取了六个预训练模型优化得到的最大激活类平均图，此部分仅展示使用 VGG11 预训练网络优化得到的 CIFAR-10 类平均图。通过将优化前后的类平均图像输入网络后观察它们的分类置信度可以发现，最大激活优化过程的确令每张图像都可以被良好地识别：

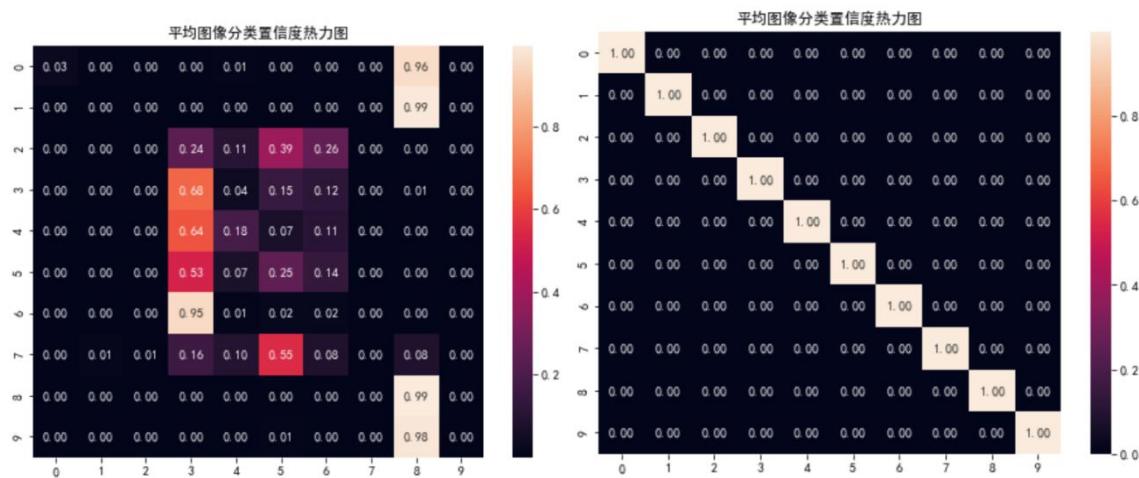


图 4-11：CIFAR-10 数据的简单类平均图像使用基于 VGG11 预训练模型的最大激活前（左）后（右），网络的分类置信度热力图：纵坐标为类平均图像编号，横坐标为网络的分类置信度。

但是，通过对比优化前后的类平均图可以发现，基于网络驱动的最大激活算法并没有对简单类平均图补充高语义信息。图像中的主体依然较模糊，而是补充了低语义信息——特定的纹理、高亮的像素等信息。这并不符合表示数据类公有信息的目的。



图 4-12: CIFAR-10 数据集的简单类平均图像。



图 4-13: 使用 VGG11 预训练网络优化得到的 CIFAR-10 类平均图。

4.3 神经元相似值计算结果与相似值曲线分析

基于 3.3 节神经元可视化图像的相似值计算算法，本部分实验对预训练网络的所有神经元使用基于类平均图像的神经元可视化方法，由此得到每个神经元的相似值。对于所有神经元的相似值，可以从两种粒度进行分析：1) 层内神经元相似值分析：通过对基于类平均图的神经元可视化图像和神经元相似值联合分析，可以观察单个网络层内神经元对于类平均图像的激活情况，例如有些神经元无法被激活，从而产生全黑的可视化图像，导致其相似值为 0，有些神经元的可视化图像则没有包含足够的类别信息，导致其相似值接近 0；2) 层间神经元相似值分析：通过对一个网络层内所有神经元的相似值计算均值和标准差，可以绘制整个神经网络的层相似值均值曲线和层相似值标准差曲线，由此可以分析神经元相似值在整个网络中的变化情况，以此分析网络对特征的学习情况和其他性能。

4.3.1 层内神经元相似值分析

神经元基于类平均图的相似值代表了其对该数据图像特征信息的提取情况，在单个网络层中每个神经元的相似值各不相同，基于相似值对神经元排序后，可以对比分析对应神经元的可视化图像，寻找其低相似值的原因，从而对网络神经元有更深入的理解。因为层内神经元相似值分析对于基于不同方法生成的类平均图使用相同的分析方法，因此本部分仅展示 CIFAR-10 数据集的数据驱动的最大激活类平均图在 VGG11 预训练模型的层内分析结果，本部分选取了其中的类别 2 “鸟”。

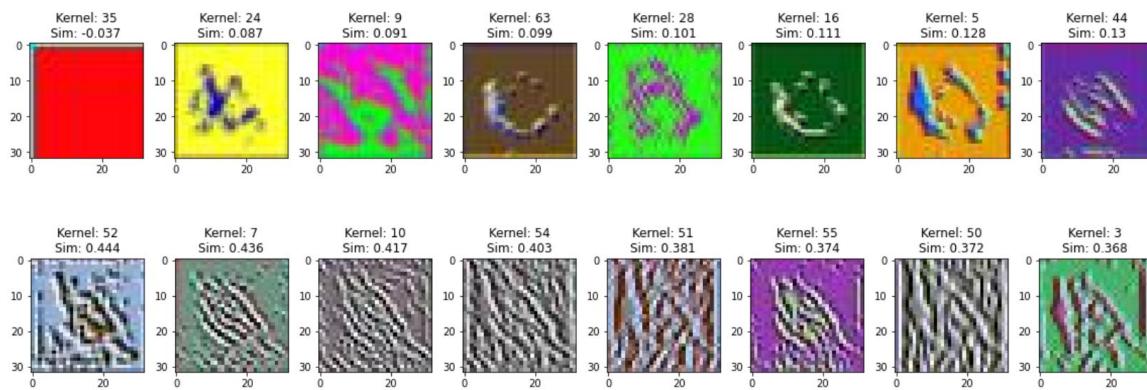


图 4-14：第 0 层神经元可视化图像和相似值：第一行为相似值最低的 8 个神经元，第 2 行为相似值最高的 8 个神经元。

对比观察第 0 层内的神经元可视化图像和相似值可以发现，第 0 层神经元主要用于提取图像的纹理、颜色信息，对于低相似值神经元可视化图像中主要是颜色特征，“鸟”的高语义信息不明显；高相似值神经元主要描述了“鸟”的纹理特征，同时也包含了其轮廓信息。

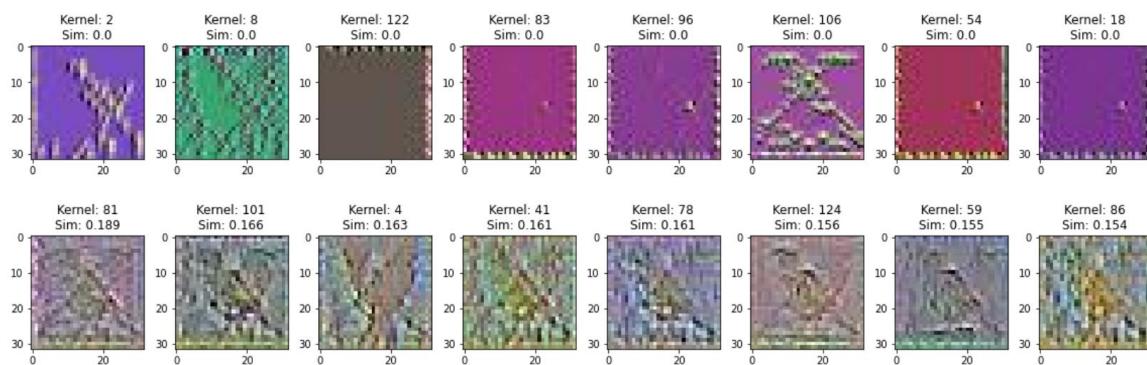


图 4-15：第 1 层神经元可视化图像和相似值：第一行为相似值最低的 8 个神经元，第 2 行为相似值最高的 8 个神经元。

对于第 1 层内的神经元可视化图像和相似值，相似值最低的 8 个神经元并未被激活，可视化图像为 0 值图像，因此图像第一行展示的是非全 0 的最小相似值神经元。观察可以发现，与第 0 层类似低相似值图像包含了多种的颜色、纹理信息，语义信息不强，而高相似值神经元的可视化图像中“鸟”的轮廓更加明显，细节上高相似值神经元的可视化图像颜色、细节各不相同。

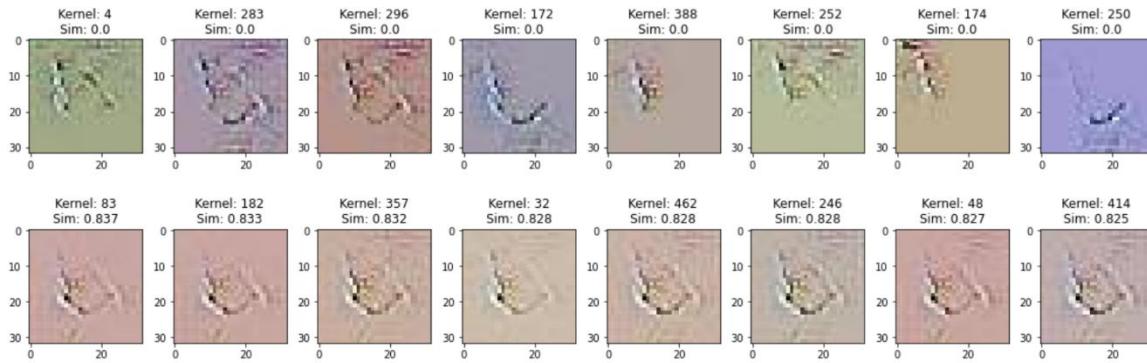


图 4-16：第 4 层神经元可视化图像和相似值：第一行为相似值最低的 8 个神经元，第 2 行为相似值最高的 8 个神经元。

对于第四层的神经元可视化图像，前 24 个神经元都未被激活，为全 0 图像。观察低相似值图像，它们的颜色差距明显，并且目标位置各不相同，如 388、174 号神经元的可视化图像中目标在左上，而 172、250 号神经元的可视化图像中目标在下部分。在第 4 层中高相似值神经元可视化图像颜色相近，目标居中，高语义信息描述的主体是“鸟”的身体部分。

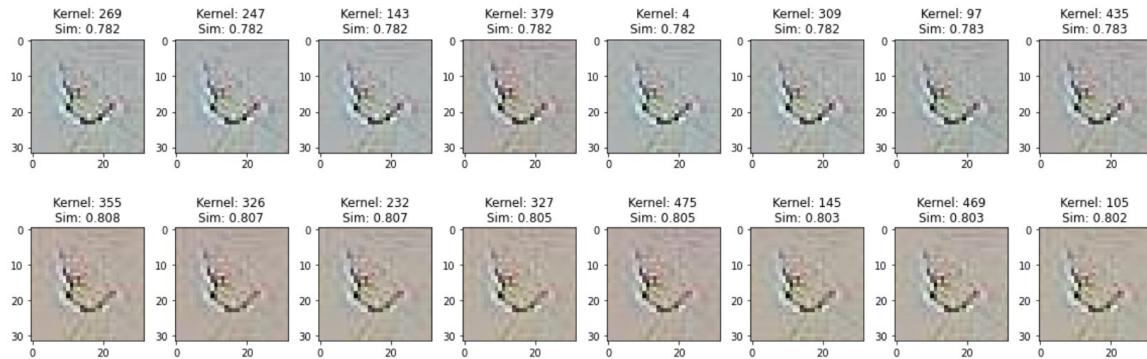


图 4-17：第 7 层神经元可视化图像和相似值：选取了其中的 16 个神经元。

在后面的网络层中，神经元可视化图像之间的相似性逐渐上升，虽然相似值不

同，但是通过观察并不能区分出其中的差异。

通过观察不同层内神经元的可视化图像和相似值可以发现，网络的底层主要用于提取图像中的颜色、纹理等低语义信息，可视化图像差异明显；而网络的深层中，有一批神经元无法被类平均图激活，网络越深不能被激活的神经元数量也逐渐上升，其他神经元的可视化图像中语义信息更加明显，主要为目标的核心部分，神经元的相似值也相比前面层更高。

4.3.2 层间神经元相似值分析

由于每个网络层中神经元数量非常多，直接对比层与层之间的相似值差异与变化十分困难，所以本实验利用相似值的均值和标准差来概括每一个网络层的相似值的大小与波动情况。由此可以对基于每个类平均图像得到的相似值计算均值和标准差，将它们绘制成折线图，通过观察两个折线图可以更好地理解相似值在网络中的变化情况。本部分使用类 Plotly 可视化库，Plotly 很好地支持了查询、选取等交互操作。本实验完成了基于三种类平均图像计算方法、两种相似值比较方法和 6 个预训练模型的全部相似值计算实验，由于篇幅限制，本部分仅展示一组图像结果和一组基于准确率的分析实例。

1) 使用简单类平均图和相似值计算方法二的 VGG11_ImageNet 预训练模型相似值曲线。

观察基于 VGG11_ImageNet 的相似值的均值曲线与标准差曲线（图 4-18 与图 4-19，即使用简单类平均图和相似值计算方法二的 VGG11_ImageNet 预训练模型相似值均值/标准差曲线，横坐标为由浅至深的卷积层编号，纵坐标为相似值/标准差大小，共包含 11 条曲线，其中前 10 条曲线代表由 10 类平均图像得到的相似值的均值/标准差曲线和它们的均值曲线），可以发现相似值在由浅入深的网络层中逐渐上升并趋于稳定，由于类平均图像代表着类公有信息，因此可以通过曲线了解到，网络的第 0,1 层主要用于提取低语义信息，因此相似值较低，而后层中的语义信息逐步增加；网络在 4~7 层的相似值逐渐趋于稳定，不再上升这代表着参数的冗余情况，也就是网络训练的过拟合，网络在前半部分已经提取到了足够的语义信息，而后半部分则变化较小。通过相似值的标准差曲线也可以发现，网络在 4~7 层的标准

差逐步减小，趋近于 0，这说明层相似值波动很小，神经元之间的信息差异不明显。

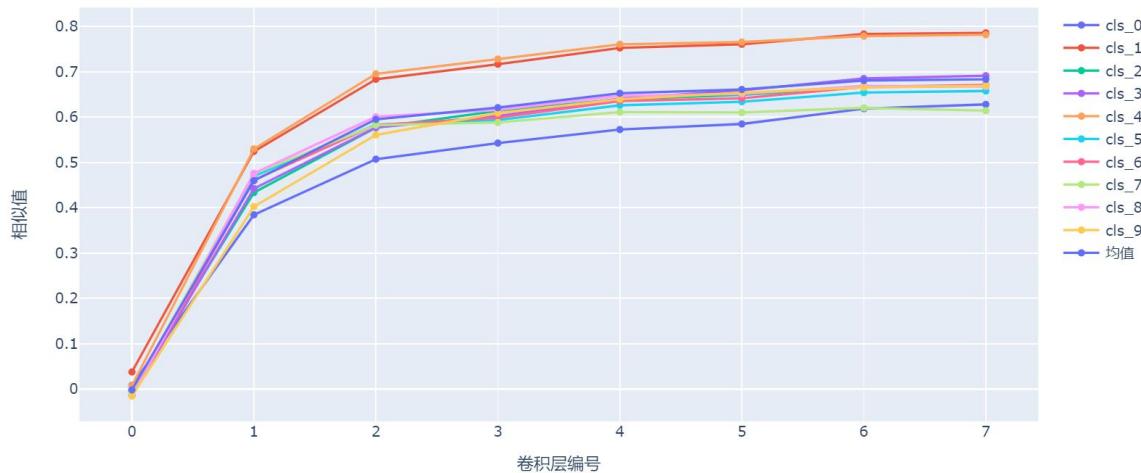


图 4-18: VGG11_Imagenet 神经元相似值均值曲线。

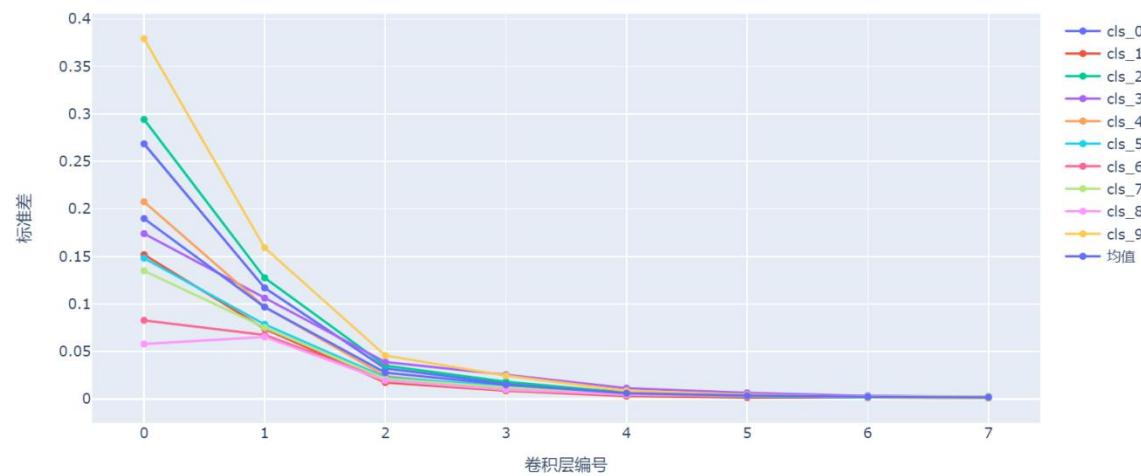


图 4-19: VGG11_Imagenet 神经元相似值标准差曲线。

2) 使用数据驱动最大激活的类平均图和相似值计算方法二的 VGG11_CIFAR 预训练模型的实例分析过程。

表 4.2: VGG11_CIFAR 预训练模型每个数据类别的测试准确率表格。

数据类别	类 0	类 1	类 2	类 3	类 4
测试准确率	0.790	0.883	0.518	0.559	0.639
预训练模型	类 5	类 6	类 7	类 8	类 9

测试准确率	0.627	0.852	0.829	0.831	0.866
-------	-------	-------	-------	-------	-------

观察 VGG11_CIFAR 在 CIFAR-10 数据集的每个类别的测试准确率可以发现，类别 2 的测试准确率明显低于其它类别，因此本部分着重分析了该网络类别 2 的相似值均值曲线和相似值方差曲线，尝试通过基于最小激活和最大激活类平均图的相似值均值、方差曲线理解类别 2 测试准确率更低的原因。

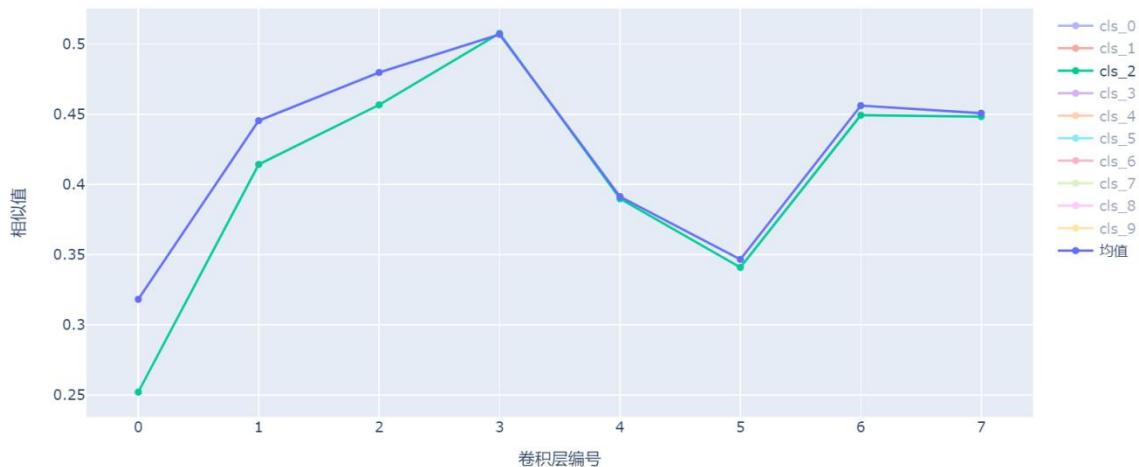


图 4-20：基于最大激活的 VGG11_CIFAR 神经元相似值均值曲线：展示了类别 2 的相似值曲线和 10 类相似值曲线的均值曲线。

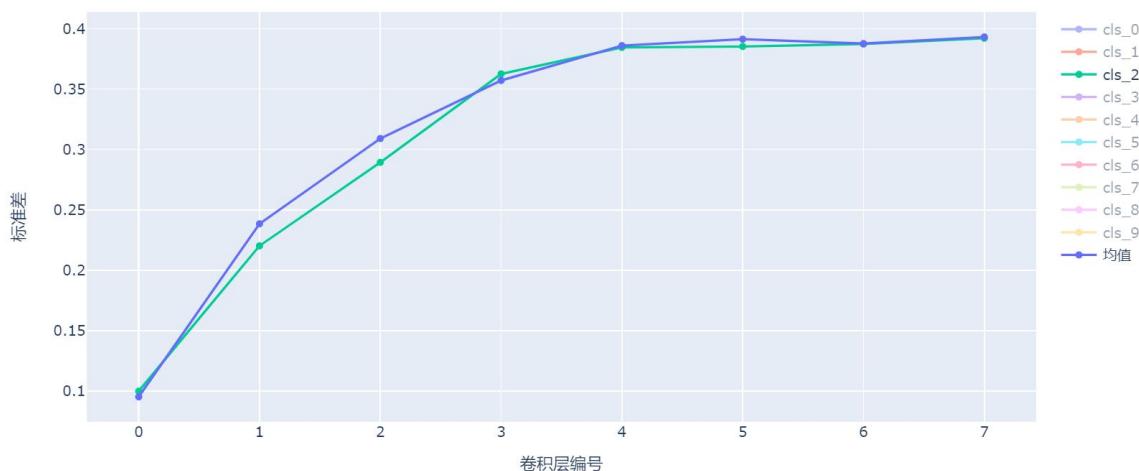


图 4-21：基于最大激活的 VGG11_CIFAR 神经元相似值标准差曲线：展示了类别 2 的相似值标准差曲线和 10 类相似值曲线的均值曲线。

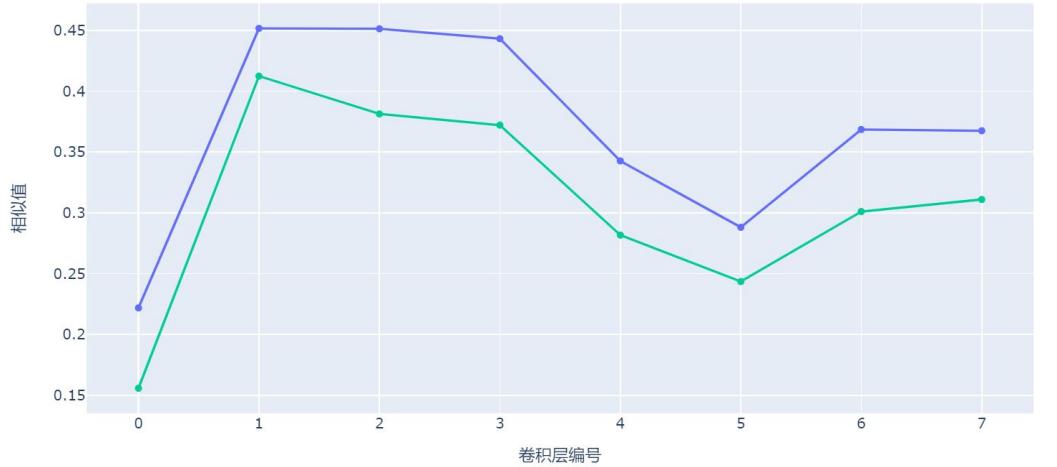


图 4-22：基于最小激活的 VGG11_CIFAR 神经元相似值均值曲线：展示了类别 2 的相似值曲线和 10 类相似值曲线的均值曲线。

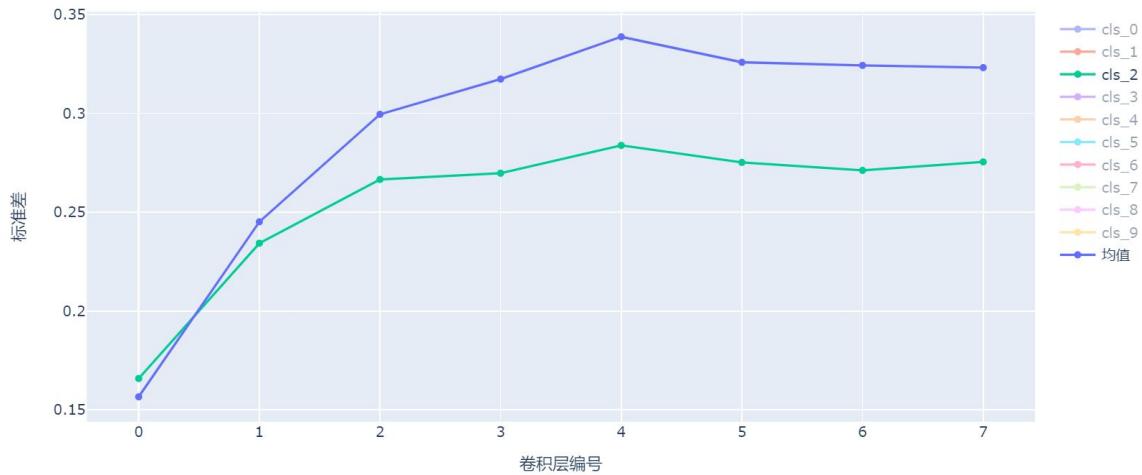


图 4-23：基于最大激活的 VGG11_CIFAR 神经元相似值标准差曲线：展示了类别 2 的相似值标准差曲线和 10 类相似值曲线的均值曲线。

观察基于最大激活类平均图的曲线（图 4.11 和 4.12）可以发现，基于最大激活类平均图的相似值均值和标准差曲线在 0~3 层略低于所有类别的均值曲线，4 层之后两个曲线几乎重合，这表示网络较好地学习到了最大激活类平均图中的特征信息，因此这并不是令其准确率较低的原因。观察基于最小激活类平均图的曲线图（图 4.13 和 4.14），类别 2 的相似值均值曲线明显低于类别均值曲线，这代表着相比其他类别，网络并没有很好地学习到最小激活类平均图中的信息，因而网络对包含这些信息的样本图像的识别能力不足；相似值的标准差曲线中第二层之后类别

2 相似值的标准差明显低于类别均值曲线，这可能代表着深层网络中特征的鲁棒性不足。这两个原因可能导致了网络对类别 2 图像识别的准确率低于其它类别。

4.4 本章小结

本章介绍了神经元相似值分析算法的主要实验部分。首先，本章解释了应用与后续实验的六个预训练模型的模型结构、参数设置与训练过程。接着，文章展示了类平均图的计算结果，包括简单类平均图、数据驱动的激活最大化类平均图（包括最大激活和最小激活类平均图）和网络驱动的激活最大化类平均图像。数据驱动的激活最大化类平均图在保持其解释性的基础上，较少地受到过平均化问题的影响，是后续实验的良好输入。然后，4.3 部分从不同的粒度对网络神经元的相似值进行了分析。层内神经元相似值分析发现，神经网络靠前的网络层主要用于提取图像的纹理、颜色等低语义特征，可视化特征变化丰富，而更深层的网络着重提取图像的高语义信息，可视化特征仅在细节上有一些不同，并且出现 0 激活的概率也更高。层间神经元相似值分析主要依据层相似值均值曲线和层相似值标准差曲线，本部分展示了神经元相似值分析法在 ImageNet 大型数据上的曲线效果，除此之外还从最大激活和最小激活曲线两个角度解释了 VGG11_CIFAR 类别 2 的低准确率原因。

第 5 章 基于神经元相似值的剪枝验证实验与分析

第四章介绍了神经元的相似值实验与分析，并对预训练模型的性能从相似值的角度提出了新的解释，本部分希望通过基于相似值的神经元剪枝实验对相似值的性能进行验证，分析相似值的作用。

5.1 基于相似值的网络层剪枝实验

在 4.1 部分的模型预训练过程中，观察网络模型测试损失曲线，可以发现多数预训练模型发生了过拟合问题。基于相似值的网络层剪枝的主要思想是通过观察模型的层相似值均值曲线，判断模型冗余情况，直接对网络层剪枝，通过缩减特定的网络层数，观察网络过拟合问题的改善情况来验证层相似值对网络性能的指导作用。

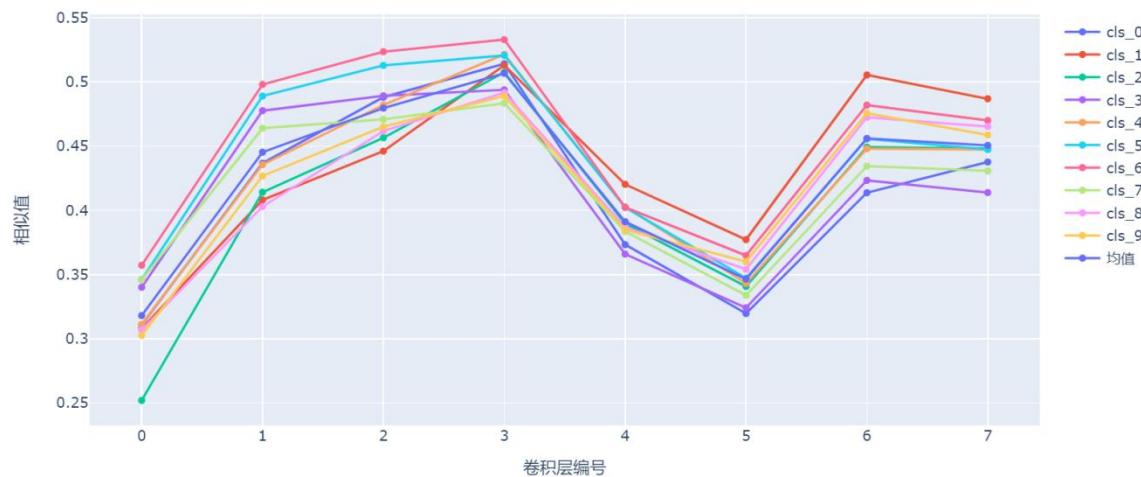


图 5-1：VGG11_CIFAR 神经元相似值均值曲线。

本实验延续了 4.2 部分使用的 VGG11_CIFAR 预训练模型，观察曲线可以发现网络在第 3 层层相似值达到了最大值，因此判断网络在第 3 层已经有足够的信息提取能力，而后部分的网络层属于冗余部分。因此从结构上，剪枝模型只保留前 4 个卷积层与相关的激活层，由于网络分类器的全连接层对输入向量尺寸有要求，为了避免全连接层参数量增加对实验的影响，使用全局平均池化将输出维度降低，并将全连接层输入尺寸从 512 缩减为 256，网络其他部分不做调整。然后将网络按照原训练的超参数进行 30 epoch 的训练，训练过程如图 5-2：

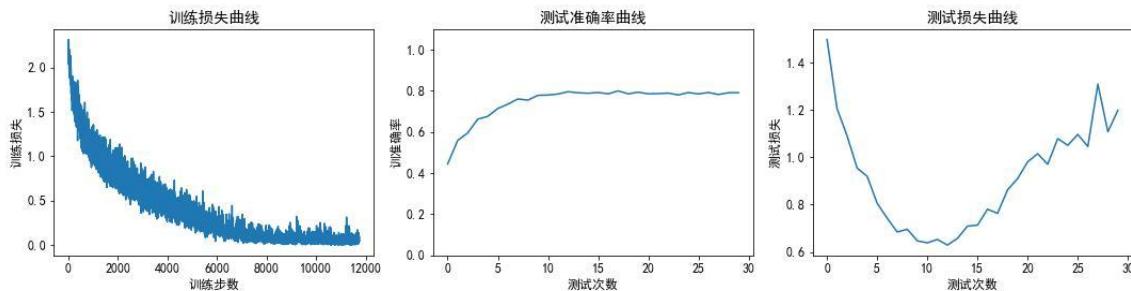


图 5-2: VGG11_CIFAR 剪枝模型 (4 卷积层) 的训练过程曲线: 左为训练损失曲线, 中为测试准确率曲线, 右为测试损失曲线。

值得注意的是, 网络的参数量从 9,488,266 缩减到 1,097,610, 减小了近 10 倍, 而网络的测试准确率从原本的 73.9% 上升至 80.0%, 约上升了 6%, 测试损失曲线中最低点从 7 epoch 延后至 12 epoch 附近。

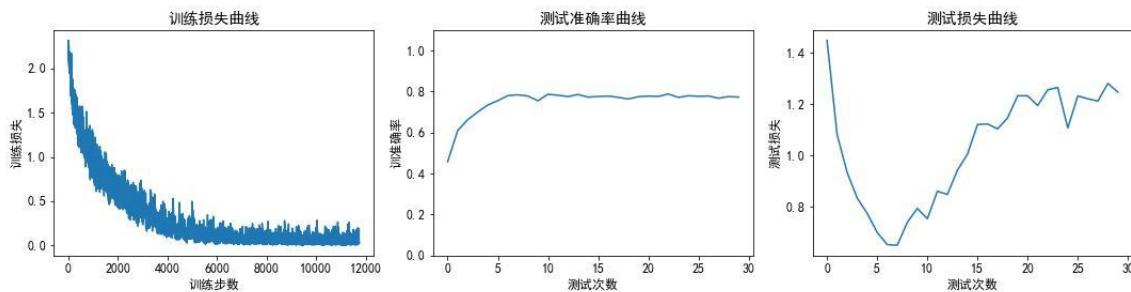


图 5-3: VGG11_CIFAR 剪枝模型 (5 卷积层) 的训练过程曲线: 左为训练损失曲线, 中为测试准确率曲线, 右为测试损失曲线。

然而如果选择保留层相似值曲线下降趋势的第 4 号卷积层 (从 4 个卷积层上升至 5 个卷积层), 虽然网络仅增加一层, 但是准确率却从 80.0% 下降至 77.8% (21 epoch 处, 最终准确率为 76.8%)。而测试损失最低值也从 12 epoch 提前至 6 epoch。这表明只有按照层相似值指示的位置剪枝才有效果, 从一定程度上体现了层相似值的指导作用。

5.2 基于相似值的神经元剪枝实验结果

上部分从层相似值的角度对网络进行剪枝实验, 本部分通过层内神经元相似值的角度对单层神经元的部分神经元进行剪枝。本部分神经元剪枝实验的思路是, 将单层网络内神经元的相似值进行排序, 然后分别选择相似值最高的 40% 神经元和相似值最低的 40% 神经元进行剪枝, 剪枝的方法是基于 1.2.2 部分的静态剪枝: 将

对应神经元的参数直接置 0。因为本部分的目标并非优化网络，而是测试神经元相似值的作用，所以剪枝后的网络不会进行微调（fine-tune）。通过分别剪去最高和最低相似值的神经元，可以对比剪枝后网络测试准确率的变化，理论上相似值更高的神经元拥有更高的重要性，因此剪枝后测试准确率的变化也更明显。本部分选取了基于简单类平均图的 VGG11_CIFAR 预训练模型进行单层神经元剪枝实验，实验分别对网络的每一层分别剪去相似值最大和最小的 40% 神经元（将 40% 神经元参数置零），然后观察网络的准确率变化，如表 5-1，剪枝最大相似值测试准确率表示对最大相似值的 40% 神经元进行剪枝，剪枝最小相似值测试准确率表示对最小相似值的 40% 神经元进行剪枝，测试准确率差值表示二者的差值。

表 5-1：VGG11_CIFAR 预训练模型单层神经元剪枝实验结果表格。

层	神经元数	剪枝比例	剪枝最大相似值测试准确率	剪枝最小相似值测试准确率	测试准确率差值
0	64	40%	0.615	0.547	0.068
1	128	40%	0.407	0.705	-0.298
2	256	40%	0.141	0.739	-0.598
3	512	40%	0.198	0.740	-0.541
4	512	40%	0.100	0.739	-0.639
5	512	40%	0.146	0.739	-0.593
6	512	40%	0.618	0.739	-0.122
7	512	40%	0.653	0.739	-0.086

通过对表格的观察可以发现，除第 0 层和第 7 层外，剪枝最大相似值准确率的值远低于剪枝最小相似值的准确率，这代表着相似值最大神经元中包含着用于图像分类的重要特征，将它们剪枝会对模型的分类性能带来严重损害。同时我们也可以发现，剪枝最小相似值神经元对于网络带来的影响很小，如第 2 至第 7 层，网络的准确率几乎不会变化，这代表着这一部分神经元属于冗余部分，为网络带来过拟合问题。此部分同样验证了神经元相似值的对神经元的解释意义。

5.3 本章小结

本部分通过基于相似值的层剪枝实验和神经元剪枝实验，分别验证了相似值对于网络层和网络神经元的解释意义，相似值一定程度代表了神经元或网络层的重要性。在层剪枝实验中，对于 VGG11_CIFAR 预训练模型，只保留层相似值更高的网络层的模型性能要比保留其他层的模型更好。在神经元剪枝实验中，剪去相似值最大的神经元对网络的影响要远大于剪去相似值最小的神经元。这两个实验都从侧面验证了神经元相似值对神经元的重要性的指示作用。

第 6 章 总结与展望

6.1 研究总结

本研究的核心思想是将类平均图作为单个数据类别特征的集合，使用类平均图中蕴含的公有特征衡量神经元可视化图像中包含的特征信息，通过比较类平均图和神经元可视化图像的相似度，将难以直观理解的可视化图像量化，由此赋予网络中每个神经元一组相似值指标。通过对相似值的层内分析可以观察可以分析层内神经元学习到的特征，它们在网络层中起到的作用，也可以根据一些不能被激活的无效神经元，分析网络的冗余情况。相似值的层间分析通过相似值均值和方差曲线分析神经元相似值在网络中的变化与波动情况，可以分析网络对类平均图中信息的学习情况，信息在网络中的变化，也可以以层的角度分析网络的冗余情况。本研究的实验部分将神经元的相似值分析算法应用到了多个预训练模型上，并进行了实例分析，对算法进行测试。在最后的部分，本研究基于层剪枝和单层神经元剪枝对神经元相似值的衡量指标意义进行了测试，两个测试都肯定了相似值对神经元重要性的衡量作用。

6.2 未来展望

卷积神经网络的可解释性是一个十分广阔而且意义重大的领域，当前的神经网络主要以数据驱动，可解释性不足，因此其应用仍然十分受限。本研究以神经元可视化为切入点，对 CNN 特征的可解释性进行了研究。虽然本研究一定程度地证明了神经元相似值对于神经元的重要性衡量价值，但是在部分实验中，仍然存在相似值均值、方差曲线难以分析，可视化图像难以理解的情况，在剪枝实验中，一批网络呈现出不论以何种方式剪枝 40% 神经元，网络准确率都几乎不会改变的情况。可见，卷积神经网络是一个复杂的模型，从神经元相似值的单一角度并不能全面地分析网络的性能或者为其带来极强的解释意义。对于卷积神经网络的解释需要从计算方式的角度进行剖析，最终才有可能打开神经网络的“黑盒”。对于本研究来说，未来可以通过寻找更有效的类平均图、神经元可视化方法、相似值计算方法得到更

准确的神经元相似值，因为本研究提出的是使用包含类公有信息的图像衡量神经元可视化图像的思路，而非固定的一套算法，因此当更好的指标或可视化算法被提出时，神经元相似值也将可以更有效地衡量神经元的价值。

致谢

大学生活已经接近尾声，这四年是我人生的宝贵经历，在这个过程中我学习了专业知识，发掘了我的研究兴趣，我也变得更加成熟，有更多勇气面对未来的困难。

首先，我想感谢我的导师赵颖老师和师母周芳芳老师。我在 2019 年申请加入可视化训练营，2020 年加入可解释机器学习小组，在老师的指导下，我学习了研究和思考的方法，令我受益匪浅。虽然我只是本科生，导师和师母对我十分关照，我在 2021 年还得到了和学长学姐一同前往北京出差的有趣经历。同时，两位老师对我的出国申请也帮助良多，不厌其烦地帮我处理推荐信，我简历中的经历与成功绝大多数也来自与赵老师的实验室。总之，两位老师对我的生涯发展帮助非常多。除此之外，我还要感谢梁毅雄老师，他同样为我写了推荐信，并且他讲授的机器学习和计算机视觉是我最喜欢的课程。

其次，我想感谢自己在大学中的认真态度和出国深造的决定。我早在 19 年就做出了出国读研的决定，此后一直为这个目标努力，从学习英语，到后续加入赵老师的实验室。出国的决定也令我有更多的时间去发展自己，探索自己的兴趣。在准备的过程中以及未来，我会面对各种各样的挑战，我想我有勇气面对它们，它们会帮助我成长。

最后，我想感谢我的家人。感谢父母的养育之恩，感谢我的哥哥，他是我生涯的引路人，感谢我的女朋友葛同学的陪伴。

感谢中南大学，愿中南越来越好！感谢祖国，愿祖国繁荣富强！

参考文献

- [1] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278–2324.
- [2] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84–90.
- [3] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. arXiv:1312.6199 [cs], 2014.
- [4] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 427–436Boston, MA, USA: IEEE, 2015: 427–436.
- [5] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 770–778Las Vegas, NV, USA: IEEE, 2016: 770–778.
- [6] Huang G, Liu Z, Van Der Maaten L, et al. Densely Connected Convolutional Networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 2261–2269Honolulu, HI: IEEE, 2017: 2261–2269.
- [7] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 779–788Las Vegas, NV, USA: IEEE, 2016: 779–788.
- [8] Girshick, R, Donahue, J, Darrell, T, & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR). 2014: 580–587.
- [9] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[M]. Navab N, Hornegger J, Wells W M, et al., eds.//Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. 2015, 9351: 234–241Cham: Springer International Publishing, 2015: 234–241.
- [10] 白林亭, 海钰琳. 基于梯度分析的卷积神经网络可视化方法[J]. 信息技术与信息化, 2021 (04) :61–63.

- [11] 司念文, 张文林, 屈丹, 等. 卷积神经网络表征可视化研究综述[J]. 自动化学报, 2021:1-31.
- [12] 司念文, 常禾雨, 张文林, 等. 基于注意力机制的卷积神经网络可视化方法[J]. 信息工程大学学报, 2021, 22(3):7.
- [13] Zeiler M D, Fergus R. Visualizing and Understanding Convolutional Networks[M]. Fleet D, Pajdla T, Schiele B, et al., eds.//Computer Vision - ECCV 2014. 2014, 8689: 818–833Cham: Springer International Publishing, 2014: 818–833.
- [14] Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps[J]. arXiv:1312.6034 [cs], 2014.
- [15] Springenberg J T, Dosovitskiy A, Brox T, et al. Striving for Simplicity: The All Convolutional Net[J]. arXiv:1412.6806 [cs], 2015.
- [16] Dosovitskiy A, Brox T. Inverting Visual Representations with Convolutional Networks[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 4829–4837Las Vegas, NV, USA: IEEE, 2016: 4829–4837.
- [17] Nguyen A, Clune J, Bengio Y, et al. Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 3510–3520Honolulu, HI: IEEE, 2017: 3510–3520.
- [18] Zhou B, Khosla A, Lapedriza A, et al. Object Detectors Emerge in Deep Scene CNNs[J]. arXiv:1412.6856 [cs], 2015.
- [19] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. arXiv:1312.6199 [cs], 2014.
- [20] Yosinski J, Clune J, Bengio Y, et al. How transferable are features in deep neural networks?[J]. arXiv:1411.1792 [cs], 2014.
- [21] Fong R C, Vedaldi A. Interpretable Explanations of Black Boxes by Meaningful Perturbation[C]//2017 IEEE International Conference on Computer Vision (ICCV). 2017: 3449–3457Venice: IEEE, 2017: 3449–3457.
- [22] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization[J]. International Journal of Computer Vision, 2020, 128(2): 336–359.

- [23] Ribeiro M T, Singh S, Guestrin C. “Why Should I Trust You?” : Explaining the Predictions of Any Classifier[J]. arXiv:1602.04938 [cs, stat], 2016.
- [24] Zintgraf L M, Cohen T S, Adel T, et al. Visualizing Deep Neural Network Decisions: Prediction Difference Analysis[J]. arXiv:1702.04595 [cs], 2017.
- [25] Su J, Vargas D V, Kouichi S. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5) : 828–841.
- [26] Koh P W, Liang P. Understanding Black-box Predictions via Influence Functions[J]. arXiv:1703.04730 [cs, stat], 2020.
- [27] Lakkaraju H, Kamar E, Caruana R, Horvitz E. Identifying unknown unknowns in the open world: Representations and policies for guided exploration[C]. 31st AAAI Conference on Artificial Intelligence (AAAI). 2017: 618563937.
- [28] Zhang Q, Yang Y, Ma H, et al. Interpreting CNNs via Decision Trees[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 6254–6263Long Beach, CA, USA: IEEE, 2019: 6254–6263.
- [29] Wu T, Sun W, Li X, et al. Towards Interpretable R-CNN by Unfolding Latent Structures[J]. arXiv:1711.05226 [cs], 2018.
- [30] Sabour S, Frosst N, Hinton G E. Dynamic Routing Between Capsules[J]. arXiv:1710.09829 [cs], 2017.
- [31] Chen X, Duan Y, Houthooft R, et al. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets[J]. Advances in Neural Information Processing Systems (NIPS). 2016: 10495258.
- [32] Zhang Q, Cao R, Wu Y N, et al. Mining Object Parts from CNNs via Active Question-Answering[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 3890–3899Honolulu, HI: IEEE, 2017: 3890–3899.
- [33] Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective: Regression Shrinkage and Selection via the Lasso[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2011, 73(3) : 273–282.
- [34] Lei W, Chen H, Wu Y. Compressing Deep Convolutional Networks Using K-means Based on Weights Distribution[0]//Proceedings of the 2nd International Conference on Intelligent Information Processing – IIP’ 17. 2017: 1–6Bangkok, Thailand: ACM

Press, 2017: 1–6.

- [35] Cun Y, Denker J, Solla S. Optimal Brain Damage. *Advances in Neural Information Processing Systems (NIPS)*, 1990, 2(1): 1098–6596.
- [36] Hassibi B, Stork D G, Wolff G J. Optimal Brain Surgeon and general network pruning[C]//IEEE International Conference on Neural Networks. 1993: 293–299San Francisco, CA, USA: IEEE, 1993: 293–299.
- [37] Li H, Kadav A, Durdanovic I, et al. Pruning Filters for Efficient ConvNets[J]. arXiv:1608.08710 [cs], 2017.
- [38] Hu H, Peng R, Tai Y-W, et al. Network Trimming: A Data-Driven Neuron Pruning Approach towards Efficient Deep Architectures[J]. arXiv:1607.03250 [cs], 2016.
- [39] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables[J]. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2006, 68(1): 49–67.
- [40] Lebedev V, Lempitsky V. Fast ConvNets Using Group-Wise Brain Damage[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 2554–2564Las Vegas, NV, USA: IEEE, 2016: 2554–2564.
- [41] Liu Z, Li J, Shen Z, et al. Learning Efficient Convolutional Networks through Network Slimming[C]//2017 IEEE International Conference on Computer Vision (ICCV). 2017: 2755–2763Venice: IEEE, 2017: 2755–2763.
- [42] Zhang Y, Zhao C, Ni B, et al. Exploiting Channel Similarity for Accelerating Deep Convolutional Neural Networks[J]. arXiv:1908.02620 [cs, stat], 2019.