



中山大學 网络空间安全学院

SUN YAT-SEN UNIVERSITY SCHOOL OF CYBER SCIENCE AND TECHNOLOGY

人工智能安全实验报告

图像分类及可视化

第 1 次实验项目

姓名:	林浩
学号:	21312596
学院:	网络空间安全学院
专业:	网络安全

2024 年 2 月 28 日

1.实验要求

- 主要任务:

基于 Fashion-MNIST 数据集上实现图片分类,数据集 包含了 60,000 个训练图像和 10,000 个测试图像, 每个图像都是 28x28 像素的灰度图, 这些图像代表了不同的服装, 如衬衫、裤子、鞋子等 10 种不同类别的服装。

- 任务要求:

- 1.使用两种常见的 CNN 分类模型 (包含两种模型, 如 VGG16 或 ResNet 等)
- 2.尝试对代码进行调优, 在验证集上获得尽可能高的准确率
- 3.可视化中间结果(如特征图、卷积、梯度、准确率)
- 4.在实验报告中展示训练后模型在验证集上准确率以及训练过程中损失值的变化图, 训练完成后将模型 **checkpoint** 文件保存, 随代码和实验报告一同提交

2.实验原理

- 整体程序框架:

使用自定义的图像分类模型在 Fashion-MNIST 数据集上进行图像分类的同时, 捕获并可视化模型中第一个卷积层输出的特征图。

自定义的 ResNet 类通过继承 PyTorch 的 ResNet 模型并修改第一层卷积以适应单通道的 Fashion-MNIST 图像实现。数据加载部分使用了 PyTorch 的数据加载器和转换器来加载 Fashion-MNIST 数据集并进行必要的预处理, 如转换为张量和标准化。

为了可视化特征图的输出, 代码中定义了一个钩子函数, 它在第一个卷积层的前向传播过程中被触发, 捕获并存储了特征图。在每个训练 epoch 结束后, 这些特征图被可视化并保存为 PNG 文件, 以便于分析模型是如何识别和处理不同的图像特征的。

卷积核是 CNN 中的核心, 负责提取输入数据的特征。为了可视化这些卷积核, 我首先访问了模型的第一个卷积层的权重 (`model.conv1.weight.data`), 这是一个四维张量, 包含了卷积核的实际数值。通过选择特定数量的卷积核, 并将它们的权重绘制为图像, 我们可以直观地看到模型是如何“看待”输入数据的。在每个训练 epoch 结束后, 我调用 `visualize_kernels` 函数, 它提取指定卷积层的权重, 选择前几个卷积核, 并将它们作为图像保存下来。

梯度反映了权重对损失函数的敏感度, 是模型训练过程中优化的关键。通过可视化权重的梯度, 我们可以理解哪些权重在训练过程中被大幅调整。这种理解可以帮助我们调试模型, 了解训练过程中可能的问题。与卷积核可视化类似, 梯度可视化是通过捕获特定卷积层权重的梯度 (`model.conv1.weight.grad`), 并在每个 epoch 结束后通过 `visualize_gradients` 函数绘制出来。

训练过程本身遵循标准的深度学习训练流程, 包括前向传播、损失计算、反向传播和参数更新。在训练过程中, 代码记录了损失和准确率, 并在每个 epoch 结束时输出这些信息。此外, 还在每个 epoch 结束时保存了模型的状态, 这允许训练中断后可以从最后保存的状态恢复, 或者用于以后的模型评估和预测。

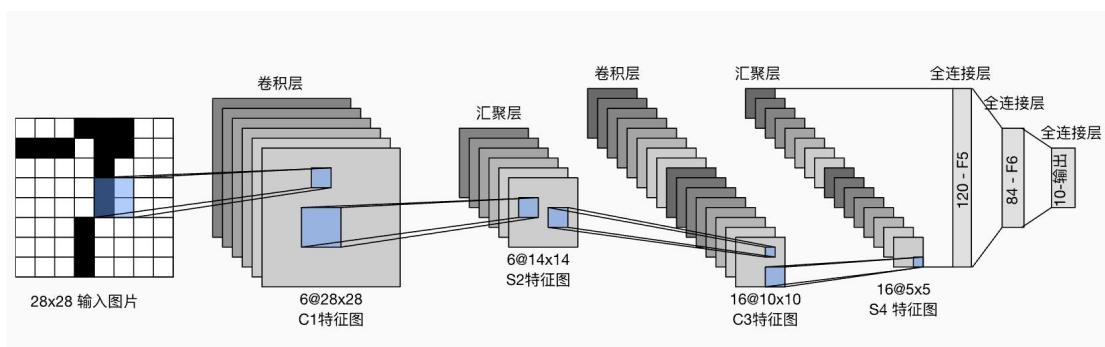
最后, 在所有训练完成后, 代码绘制并保存了一个包含训练损失和准确率随 epoch 变化的曲线图, 提供了一个直观的方式来评估模型性能的改进。

整个过程展示了图像分类模型的训练和评估过程, 并通过特征图的可视化提供了对模型内部工作机制的直观观察。

- 图像分类模型选择:

选择自定义的 Resnet 和 Alexnet 模型进行图像分类训练, 两者都是在图像分类表现优异的 CNN 模型, 能够较好的提取图像特征。

同时由于 MNIST 属于相对简单的图像数据集, 不需过于复杂的模型也足以胜任该分类任务。项目自定义实现了层数较小的轻量化 Resnet 和 Alexnet 模型。



基于 MNIST 数据集进行图像分类

ResNet 的核心思想是引入了“残差块”（Residual Block），每个残差块都尝试学习输入和输出之间的残差（即差异）。这是通过将输入直接连接到输出（通过跳跃连接或快捷连接）来实现的，这样模型就可以只学习输入和输出之间的差异，而不是试图直接学习输出。

深度网络常常面临梯度消失或梯度爆炸的问题，这会使得网络难以训练。通过残差块中的跳跃连接，梯度可以直接流过多个残差块，不仅帮助减轻了梯度消失的问题，还使得深度模型更容易训练。

尽管 Fashion-MNIST 图像的原始尺寸是 28x28，但 ResNet 通常需要较大尺寸的输入（如 224x224）。因此，图像需要被上采样或通过其他预处理方法调整尺寸。

图像经过多个残差块处理，每个块通过卷积层、批归一化（Batch Normalization）和 ReLU 激活函数提取特征，同时使用跳跃连接帮助梯度流动。

在通过所有残差块后，使用全局平均池化（Global Average Pooling）减少每个特征图的维度，最后通过全连接层（Fully Connected Layer）输出每个类别的预测概率。

AlexNet 由 5 个卷积层和 3 个全连接层组成，这种深度架构使得网络能够学习复杂的特征表示。

AlexNet 是首次在 CNN 中广泛使用 ReLU（Rectified Linear Unit）作为激活函数的网络。ReLU 相比传统的 Sigmoid 或 Tanh 激活函数在训练深度网络时有更快的收敛速度，并且能有效减轻梯度消失问题。

AlexNet 使用的池化层步长小于池化核的大小，这种重叠池化策略旨在减少池化操作的信息丢失。

为了减轻过拟合，AlexNet 在其全连接层中引入了 Dropout 技术，随机丢弃一部分神经元，使模型在训练过程中变得更加健壮。

2.实现细节

● 环境记录:

序号		
1	系统	MAC
2	硬件配置	M1
3	python 版本	3.9
4	torch 版本	2.1.1

● 增强分类效果手段:

通过引入残差连接来解决深度网络中的梯度消失问题,使得网络能够成功训练更深的模型。这种架构允许网络学习恒等映射,保证了深层信号的有效传递,从而提高了模型的学习能力和最终的分类性能。

使用 `transforms.Compose` 进行数据预处理,包括转换为张量和标准化。这些步骤是准备数据以适应模型输入的标准做法,有助于加快收敛速度,提高模型训练效率。标准化可以使输入数据在同一尺度上,减少模型训练时的偏差,帮助模型更快地学习数据的特征,从而提高分类准确率。

`Dropout` 是一种有效的正则化技术,通过在训练过程中随机丢弃一部分神经元的激活值,可以防止模型对训练数据过拟合。

● 模型具体参数设置:

序号		
1	epoch	5
2	batch	128
3	learning rate	0.001

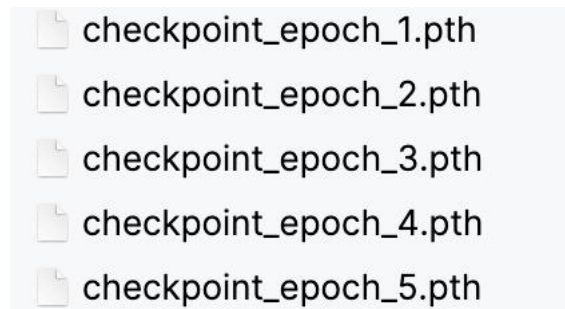
● 训练加速策略:

增大 `batch` 数量 (由 64 到 128), 减少 `epoch` 次数 (由 10 到 5), 增强 `lr` 的变动速率 (加大扰动)。

使用 `MPS` 加速。`MPS` 加速的原理基于利用 `GPU` 的并行处理能力来执行大量的数据运算。与 `CPU` 相比, `GPU` 拥有更多的执行单元,能够同时处理大量的数据操作。`MPS` 库中的算法都经过了高度优化,以充分发挥 `Apple GPU` 的性能,从而为应用程序提供实时或接近实时的处理速度。

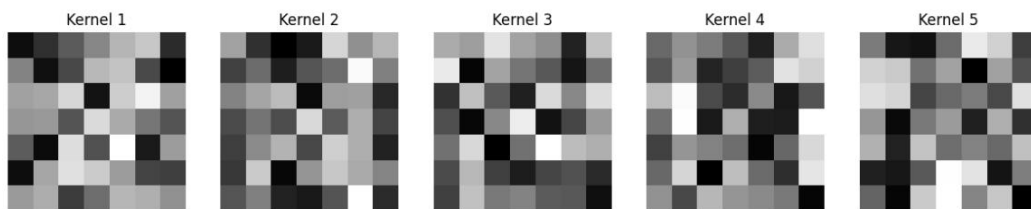
3.实验分析

- checkpoint 文件保存:



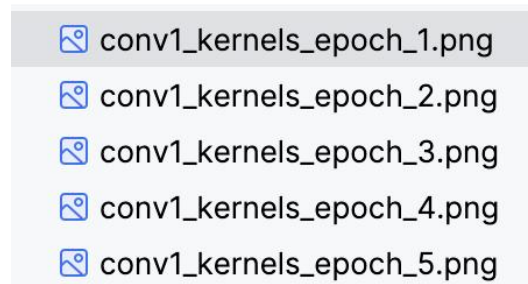
保存模型训练过程的模型参数作为备份

- Resnet 模型卷积核可视化输出:



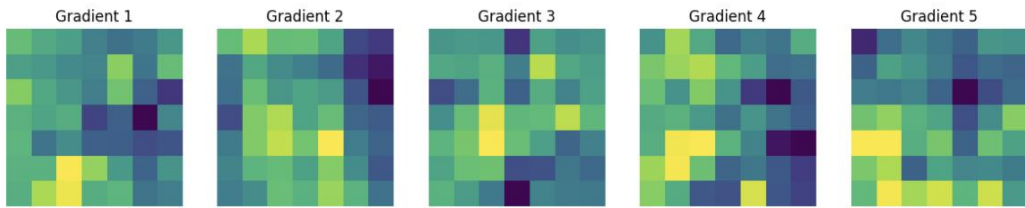
卷积核可视化信息示例

卷积核的可视化展示了网络第一层中每个卷积核的权重布局。在这个上下文中，每个卷积核都是学习到的滤波器，它们负责捕捉图像中的基础模式，如边缘、角点或更复杂的纹理。通过观察这些卷积核的可视化图像，我们可以看到模型如何从原始像素中提取有用的信息来进行进一步的处理。例如，一些卷积核可能专门用于检测衣物的边缘，而另一些则可能识别特定的纹理或形状。卷积核的多样性直接影响了网络的学习能力和最终的分类性能。



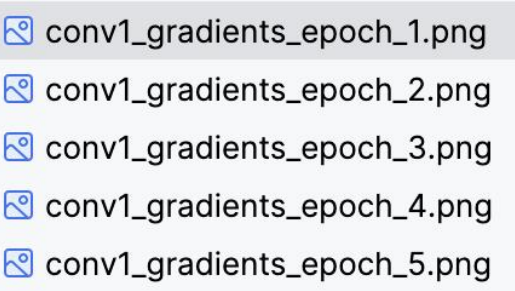
保存卷积核可视化输出

● Resnet 模型梯度可视化输出:



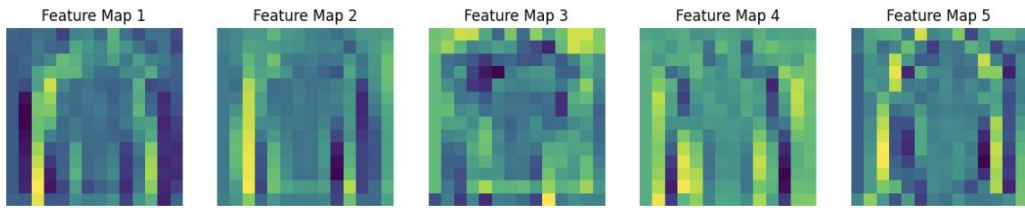
梯度可视化信息示例

梯度的可视化提供了关于模型训练过程的重要信息。在每次迭代中，通过反向传播算法计算得到的梯度反映了模型权重应该如何调整以减少预测误差。观察这些梯度图可以帮助了解模型在训练过程中的学习动态。如果某些特征的梯度始终很小，这可能意味着模型没有从这些特征中学习到有价值的信息，或者这些特征对于任务来说并不重要。相反，具有大梯度值的特征表明它们对于减少预测误差至关重要，因此模型在这些方面投入了更多的学习努力。




保存梯度可视化输出


● Resnet 模型特征图输出:



特征图信息示例

特征图的可视化展现了网络中间层的激活情况，这些激活反映了网络对输入图像的内部表示。通过查看不同层的特征图，可以理解网络是如何逐步提取和组合更高级别的特征的。在初始层，特征图可能主要包含关于图像边缘和纹理的信息。而在更深层，特征图可能会聚焦于更复杂的模式，如特定类型的衣物或其部件。这些深层特征是网络进行最终分类决策的基础。通过分析这些特征图，可以验证网络是否成功地学习到了区分不同类别所需的特征。

 feature_maps_epoch_1.png

 feature_maps_epoch_2.png

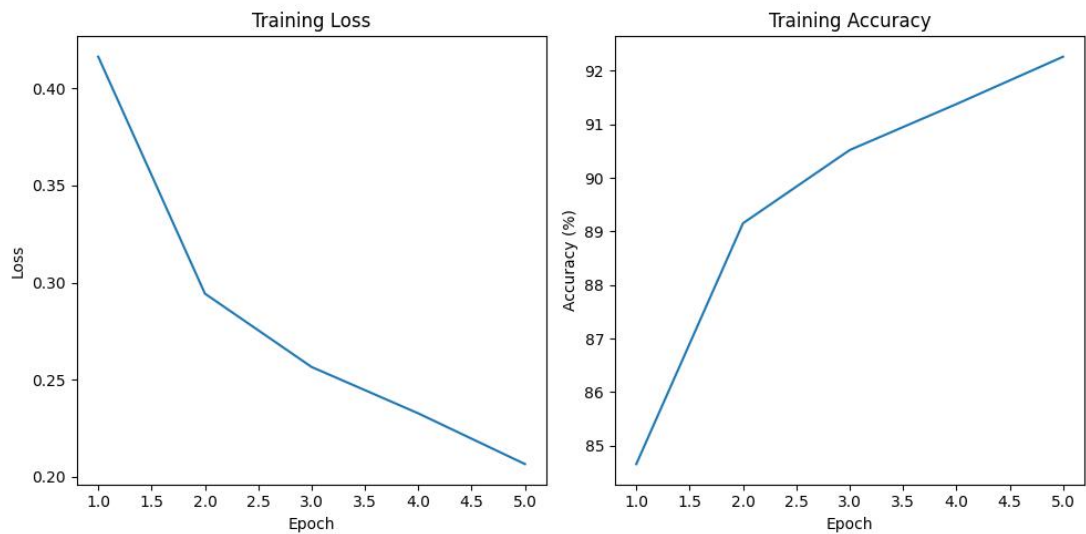
 feature_maps_epoch_3.png

 feature_maps_epoch_4.png

 feature_maps_epoch_5.png

保存特征图输出

- 模型分类效果:



Resnet 模型分类 loss 和 accuracy 变化图

```
Train Epoch: 5 [0/60000 (0%)]    Loss: 0.169982
Train Epoch: 5 [6400/60000 (11%)]  Loss: 0.212265
Train Epoch: 5 [12800/60000 (21%)] Loss: 0.277310
Train Epoch: 5 [19200/60000 (32%)] Loss: 0.243431
Train Epoch: 5 [25600/60000 (43%)] Loss: 0.368379
Train Epoch: 5 [32000/60000 (53%)] Loss: 0.434586
Train Epoch: 5 [38400/60000 (64%)] Loss: 0.332634
Train Epoch: 5 [44800/60000 (75%)] Loss: 0.318252
Train Epoch: 5 [51200/60000 (85%)] Loss: 0.341239
Train Epoch: 5 [57600/60000 (96%)] Loss: 0.197777
```

Test set: Average loss: 0.0003, Accuracy: 9001/10000 (90%)

Alexnet 模型分类训练完成的 loss 和 accuracy

4.实验拓展

- 拓展任务说明:

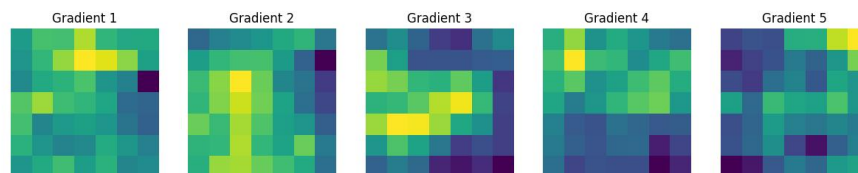
为了进一步研究 NLP 的工作原理，尝试换用更深层次的 Resnet 模型进行训练和可视化模型信息并进行对比分析。

原有的 ResNet 模型是基于 ResNet18 的配置，具有[2, 2, 2, 2]的结构，意味着在每个残差阶段有 2 个残差块。现在，我们将这个结构扩展为类似 ResNet34 的结构，即[3, 4, 6, 3]，这样可以使网络更深，从而有能力学习更复杂的特征。

- Resnet 模型梯度可视化输出对比分析:



Resnet34 梯度可视化信息示例



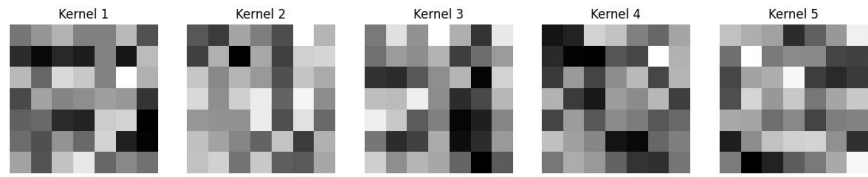
Resnet18 梯度可视化信息示例

在 ResNet18 中，梯度可视化揭示了相对集中和明确的反馈信号，指示模型在调整权重时关注于较为明显的特征。

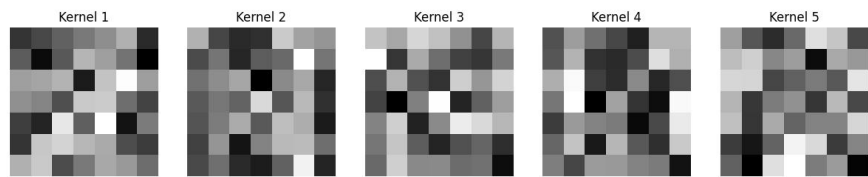
而在 ResNet34 中，梯度分布更为分散。在一些层中，梯度可能表现得更加稀疏或分布广泛，这表明网络在更广泛的特征上进行调整，包括那些更加抽象和复杂的特征。

这种差异反映了深层网络在训练过程中如何通过其复杂的层次结构来捕捉和利用高层次的特征。深层网络在其多个层次上对特征的细微变化更为敏感，导致梯度的分布和响应更加复杂

● Resnet 模型卷积核可视化输出对比分析:



Resnet34 卷积核可视化信息示例



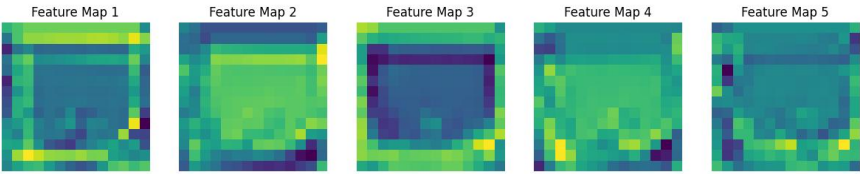
Resnet18 卷积核可视化信息示例

ResNet18 的卷积核主要捕捉基础和简单的图像特征，如边缘、角点和纹理。这些特征相对容易解释，因为它们直接与输入图像的基本视觉属性相关。

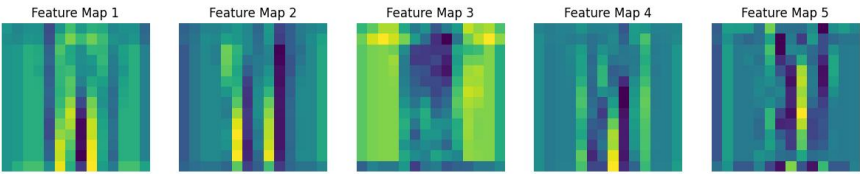
相比之下，**ResNet34** 中的卷积核除了捕捉这些基础特征外，还展示了对更复杂模式的敏感性。在网络的更深层次，卷积核反映了对特定纹理、形状或部分对象的响应，尽管这些特征更难以直接解释。

ResNet34 比 **ResNet18** 拥有更多的层，这意味着网络可以构建更加复杂和抽象的特征表示。随着层数的增加，网络能够从简单特征逐渐过渡到复杂特征的学习。

● Resnet 模型特征图可视化输出对比分析:



Resnet34 特征图可视化信息示例



Resnet18 特征图可视化信息示例

ResNet18 的特征图主要展示了对低级和中级特征的活性化响应，如边缘、颜色块和基本形状。

对于 ResNet34，特征图展示了更深层次的网络能够捕捉到的高级特征，如特定类型的纹理、复杂的形状和可能的对象局部。这些特征图在网络的不同层次之间展示了更大的多样性和特异性。

随着网络深度的增加，模型能够利用更多的层来逐步提取和综合更高级的特征。这一点在特征图的变化中得到了体现，深层网络通过其更加复杂的架构能够表示更为复杂的视觉模式和对象特征。

5.实验总结

总的来说，通过可视化卷积核、梯度和特征图，我们不仅可以验证模型是否按照预期学习，还可以识别学习过程中可能存在的问题。例如，如果某些卷积核没有显示出明显的模式，或者某些特征图在所有输入上几乎没有激活，这可能表明网络结构需要调整，或者训练过程需要改进以确保模型能够有效地从数据中学习。这些可视化技术是理解和优化深度学习模型的强大工具。

同时，对于实验拓展，从浅层到深层网络，可视化结果从直观、简单的特征过渡到复杂、抽象的特征表示。深层网络的可视化揭示了深度学习的强大之处——能够从数据中自动学习出复杂的特征层次结构。然而，这也带来了解释性的挑战，因为深层的特征和梯度变得难以用简单的视觉概念来解释。这正是为什么深度学习模型被认为是“黑盒”的一个原因，同时也推动了可解释性研究的发展，以更好地理解 and 解释深层网络的决策过程。