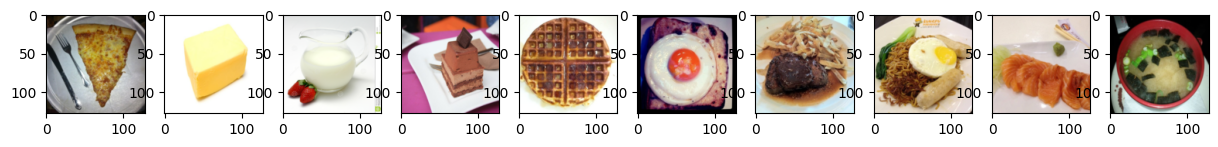
# 作业要求

HW9的主要任务是可解释AI，是对模型的可解释性的研究，没有评测上面的要求，需要完成GradeScope上面的一些选择题，但是网站无法进入，在网上也没有找到题目，所以主要是对试验过程和结果在本文档中进行分析。

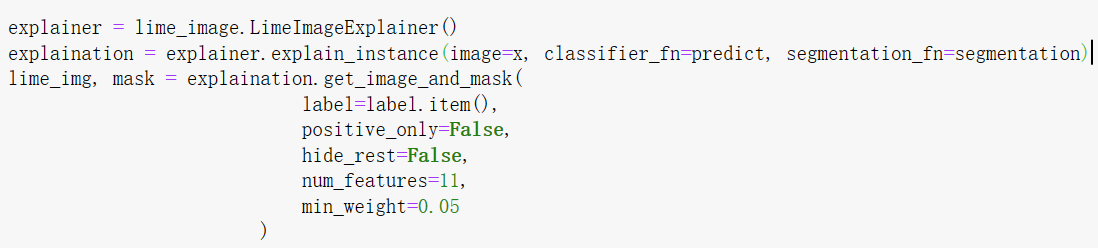
# CNN模型的可解释性

在这一部分主要使用了五种方法，Lime，Saliency Map，Smooth Grad，Filter Explanation，Integrated Gradients。使用这五种方法对下图中的十个图片进行实验，这些食物的分类依次是：面包，乳制品，乳制品，甜食，甜食，鸡蛋，肉，面食，海鲜，汤。接下来分别介绍这些方法。

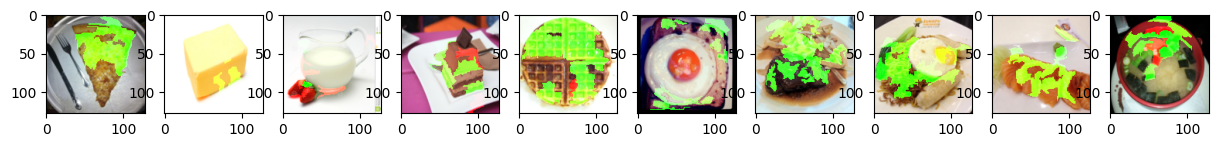


## Lime

Lime\_image模块是LIME库中用于解释图像分类模型的模块。它的主要思想是在给定图像的局部区域内，使用一个可解释的模型来近似原始模型的行为。这个可解释的模型可以是线性回归、决策树等简单模型，它能够更好地反映原始模型的行为，并且可以提供更好的解释和可视化结果。在这部分的核心代码部分如下图所示，主要通过创建解释器，对每个图像进行解释，将解释结果可视化三个步骤来实现解释模型。其中相关参数里面的classifier\_fn是模型的推理过程的函数，而segmentation\_fn是对输入图像进行分割成超像素的函数。



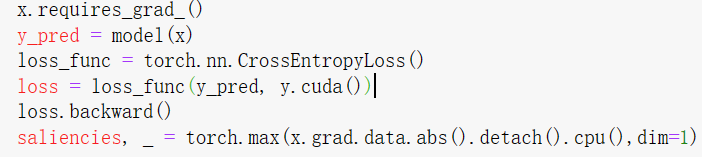
该解释方法最后的结果如下图所示，可以看到，对于大部分的食物，模型都可以抓住主要的部分，忽略周围的环境，但是对于第三个土，似乎更多关注牛奶旁边的草莓。



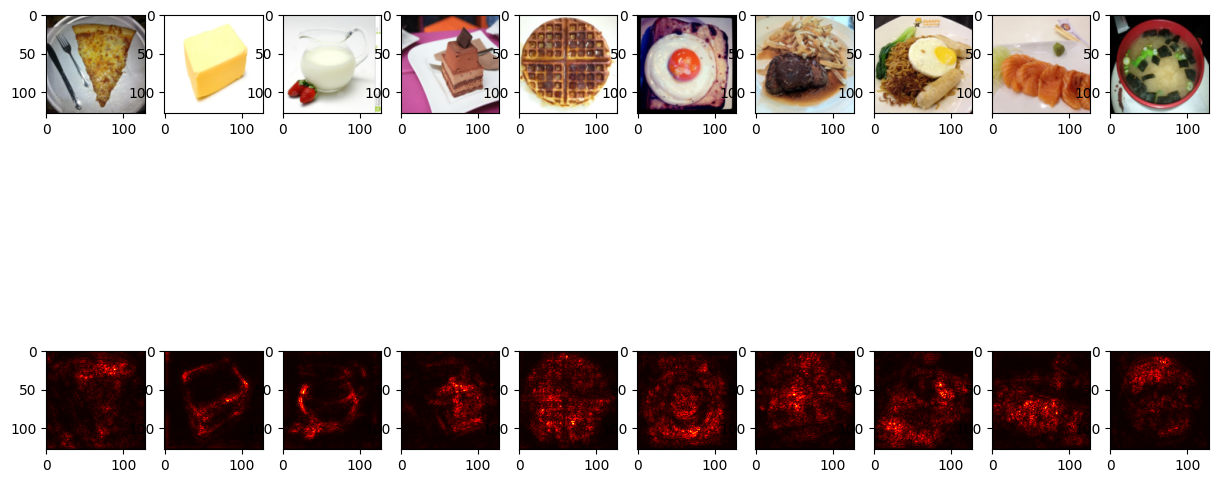
## Saliency Map

Saliency Map（显著性图）是一种用于可视化深度学习模型的输入数据对于输出结果的相对重要性的技术。它可以帮助我们理解模型是如何决策的，以及哪些输入特征对于模型的决策起到了关键作用。Saliency Map的基本思想是，通过计算输入数据对于输出结果的梯度来确定输入数据中哪些特征对于输出结果最为重要。通过计算输入数据的梯度，并对其进行归一化处理，得到一个与输入数据相同形状的显著性图。显著性图中的每个像素表示该位置的输入特征对于输出结果的相对重要性，值越大表示越重要。

方法的主要代码部分如下图所示，主要是计算输入的梯度，计算输入图像带来的损失。



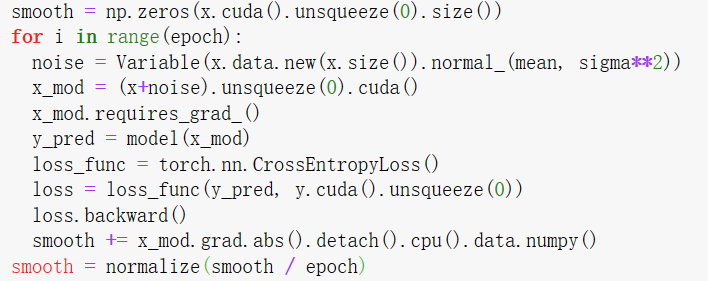
最终的结果如下图所示，可以看到对于梯度的热力图，可以明显看出每个食物的轮廓，解释了模型可以“看到”一个食物需要被关注的点。



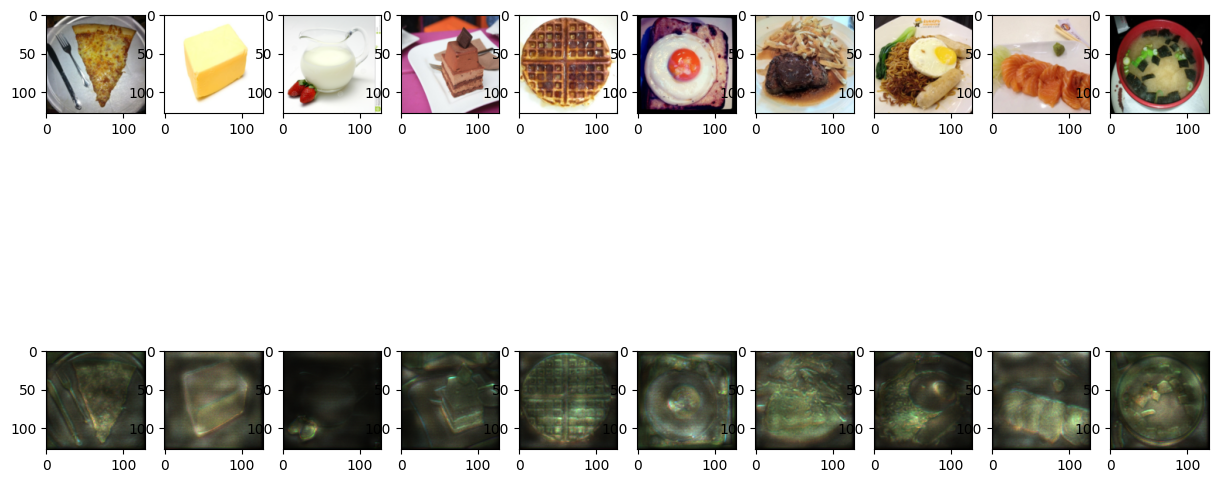
## Smooth Grad

Smooth Grad是对Saliency Map的改进，通过在输入数据上添加噪声来减少显著性图中的噪声，并使可视化结果更加平滑和可靠。Smooth Grad的基本思想是在输入数据中添加一些随机噪声，并计算每个像素对于输出结果的梯度。这个过程会重复进行多次，然后将得到的显著性图进行平均处理，得到最终的Smooth Grad图像。Smooth Grad图像中的每个像素表示该位置的输入特征对于输出结果的相对重要性，值越大表示越重要。

该方法的主要代码如下图所示，与基本思想相吻合，首先是随机制造一个noise，添加到图像之后，执行和Saliency Map相同的操作，最后归一化梯度值，得到最终的图像输出。



实验结果如下图所示，和上一种方法的图像相类似，但是可以明显看到，食物的轮廓更加清晰。



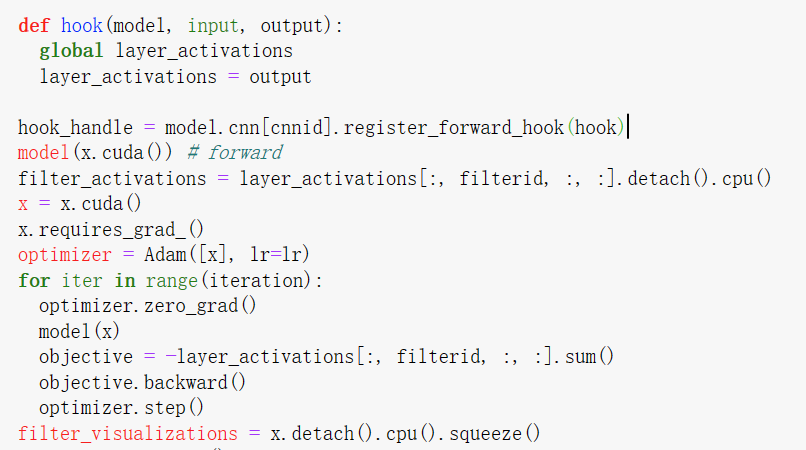
## Filter Explanation

在这个部分主要包括两个方法，Filter activation：挑几张图片出来，看看图片中哪些位置会 activate该filter。Filter visualization：找出怎样的image可以最大程度的activate该filter。

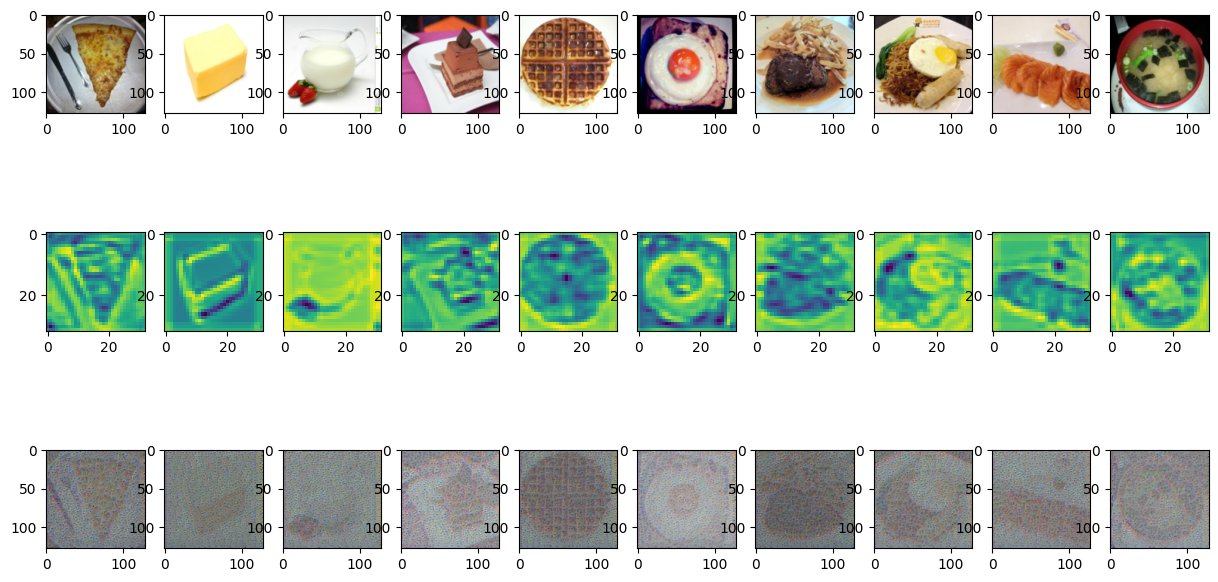
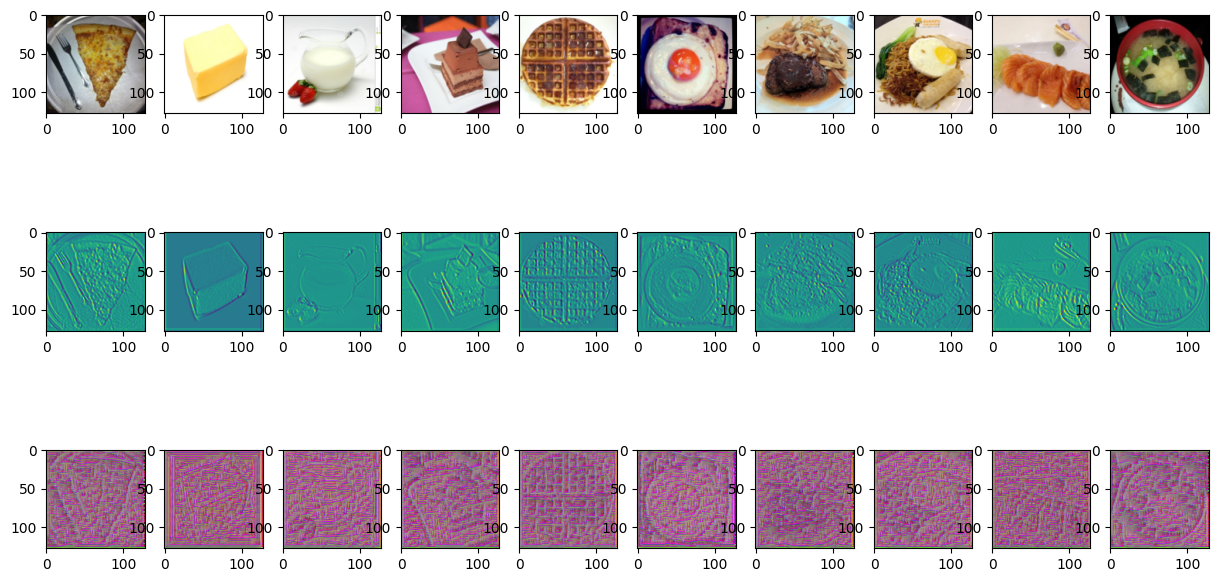
Filter Visualization是一种用于可视化深度学习模型中的卷积层滤波器的技术。它可以帮助理解卷积层滤波器是如何捕捉输入数据中的特征的，并帮助我们调整模型的参数以提高性能。Filter Visualization的基本思想是对卷积层滤波器进行可视化，以便观察滤波器在输入数据中的响应。可以通过最大化滤波器在输入数据中的响应来得到最佳的可视化结果。这个过程可以通过梯度上升来实现，即通过反向传播计算滤波器对于输入数据的梯度，并对输入数据进行更新，以最大化滤波器在输入数据中的响应。

Filter activation也是一种用于可视化深度学习模型中卷积层滤波器的响应的技术。它可以帮助理解卷积层滤波器是如何捕捉输入数据中的特征的。Filter activation的基本思想是通过可视化卷积层滤波器在输入数据中的响应来观察滤波器是如何捕捉输入数据中的特征的。具体来说，我们可以将输入数据传递到卷积层，并记录每个滤波器在输入数据中的响应值。通过观察Filter activation图像，我们可以了解每个滤波器所捕捉到的输入特征是什么。

这种方法的主要代码部分如下图所示，hook在通过注册之后，可以在正向传播到某一层的时候，自动调用hook函数，从而获得某一层的输出，这样可以获得activation的结果，而对于visualization，则需要对模型进行一些次数的训练，更改的不是模型参数，而是输入图像的参数。



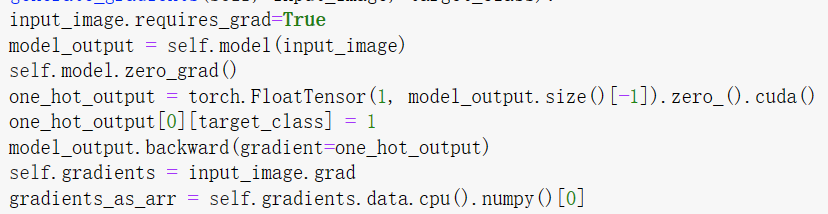
该方法的结果展示最终如下图所示，上面部分是模型第六层的结果，下面部分是模型的第二十三层的结果，对于activation的结果，随着层数的增加，获取到的特征更加的抽象，整体感觉更加的粗粒度。但是对于visualization而言，层数增加带来的结果是对模型想要的图像与真实图像之间更加接近。



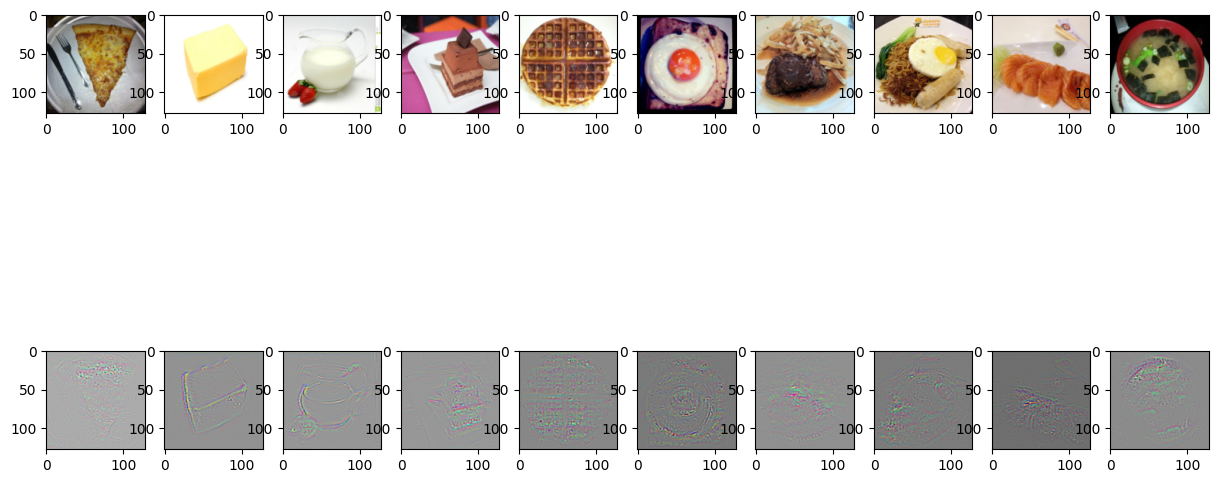
## Integrated Gradients

和上述方法类似，Integrated Gradients也是一种用于可视化深度学习模型的输入数据对于输出结果的相对重要性的技术Integrated Gradients的基本思想是，在图片和空白图片之间做线性插值产生多个图片，然后通过图片的模型输出对图片求导，计算输入数据对于输出结果的梯度，并对其进行积分，以便确定输入数据中哪些特征对于输出结果最为重要。这个过程可以通过一些数值方法来实现，例如龙格-库塔法（Runge-Kutta method）。

主要代码如下图所示，与前几种方法不同的是，计算输入图像对于输出结果的梯度，而saliency map是对loss函数求导。



实验结果如下图所示，同样可以从这些图像中分辨出食物的轮廓，验证了模型是可解释的。



# Bert的可解释性（使用Embedding Visualizaiton）

Embedding Visualization是可视化嵌入层中每个离散输入特征对应的向量表示。将嵌入层中每个离散输入特征对应的向量表示降维为二维或三维，并将其可视化为散点图或者其它形式的可视化图像。通过观察Embedding Visualization图像，我们可以了解每个离散输入特征在向量空间中的位置和关系，以及哪些特征之间存在相似性或者差异性。这些信息可以帮助我们更好地理解模型是如何对输入数据进行表示和学习的，并帮助我们调整模型的参数以提高性能。

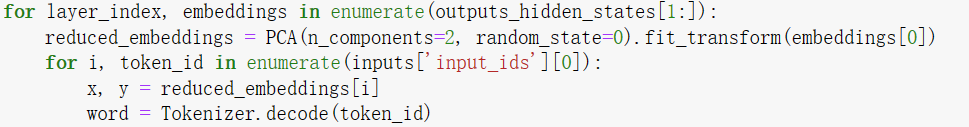
作业中提供了3个阅读理解，每一个包括context， question和answer三部分。其中可视化的问题是下面这个问题。

context:"Nikola Tesla (Serbian Cyrillic: Никола Тесла; 10 July 1856 – 7 January 1943) was a Serbian American inventor, electrical engineer, mechanical engineer, physicist, and futurist best known for his contributions to the design of the modern alternating current (AC) electricity supply system."

questions: "In what year was Nikola Tesla born?"

answers: "1856"

主要处理的代码如下图所示，使用HW7的bert模型生成context中单词的embedding，并使用PCA工具将embedding变为2维，然后对每个单词的2维信息可视化。



最终在第一层，第五层，第十层，第十二层的结果如下面的图所示，可以看到随着层数的增加，文本中与问题相关的部分与问题之间的距离越来越近，而答案与他们之间的距离越来越远，在文本中找到了问题的出处，并可以得到答案。

