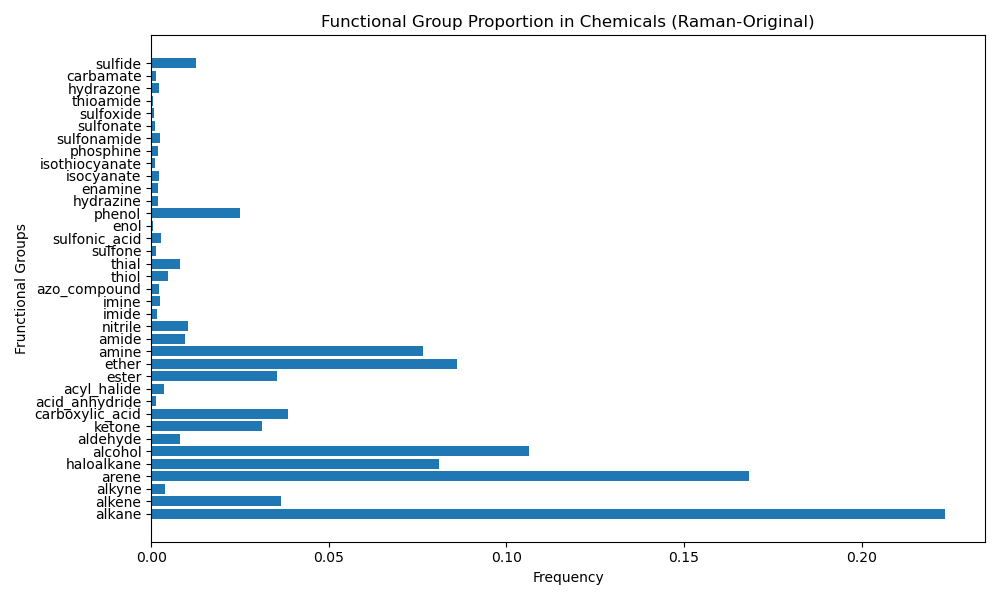
2023年10月月报

本月的主要工作有：确定了用于拉曼光谱官能团识别任务的标签；搭建了具有初步识别性能的CNN分类模型；对样本进行了显著性分析。

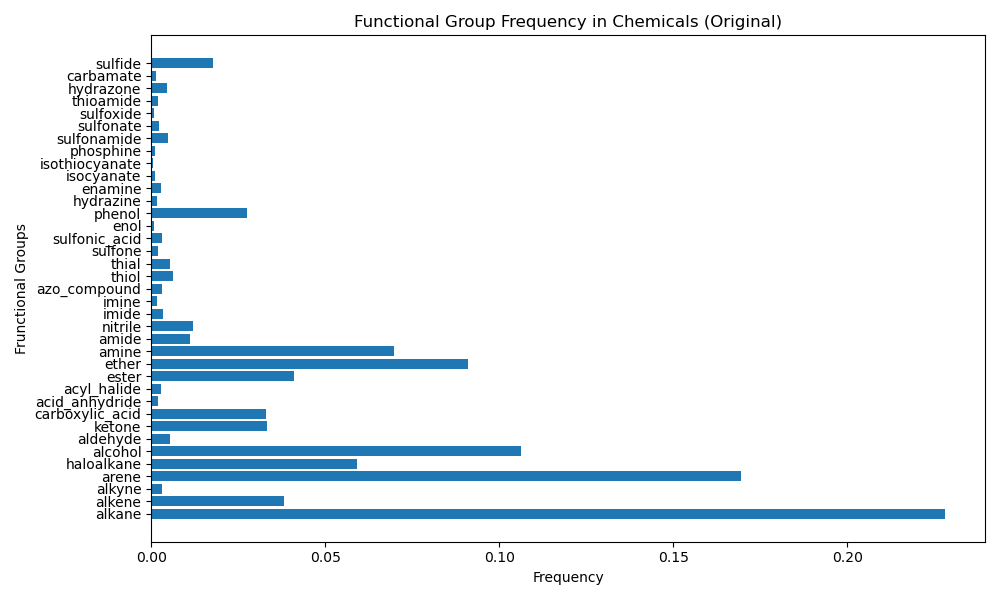
（一）已进行工作

1. 确定官能团标签

经计算，Wiley拉曼数据集官能团标签的频数分布图如下（总数22411）：



此前报道中[1]用于官能团识别任务的红外数据集，官能团频数分布如下：

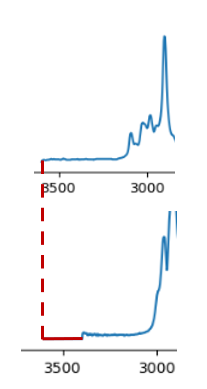


可以看出，两个数据集的官能团数量随种类的分布近似，考虑到两个数据集都在万级别规模，可以认为图中所示的37种官能团，具有很好的代表性，即使用此前报道文献中的组成标签，并应用于我们的拉曼数据集是可行的。

1. 数据预处理

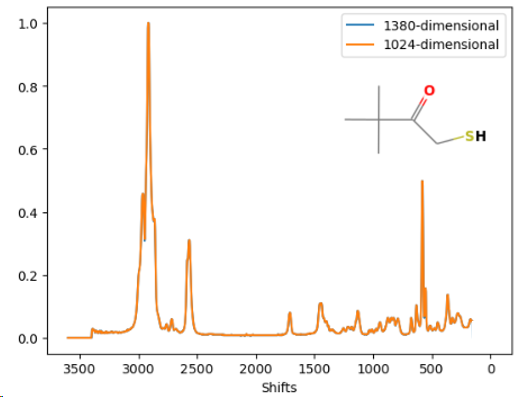
2.1 最大波数对齐

此前采集的原始数据由于最大波数不同，导致维数并不相同，因此对波数较低的数据，统一对信号值补零至最大波数，使各样本维数一致。具体来说，本数据集中含有最大波数为3600、3500和3400cm-1的样本，将最大波数为后两者的数据信号值补零到3600 cm-1。以3400cm-1样本为例，如下图所示：

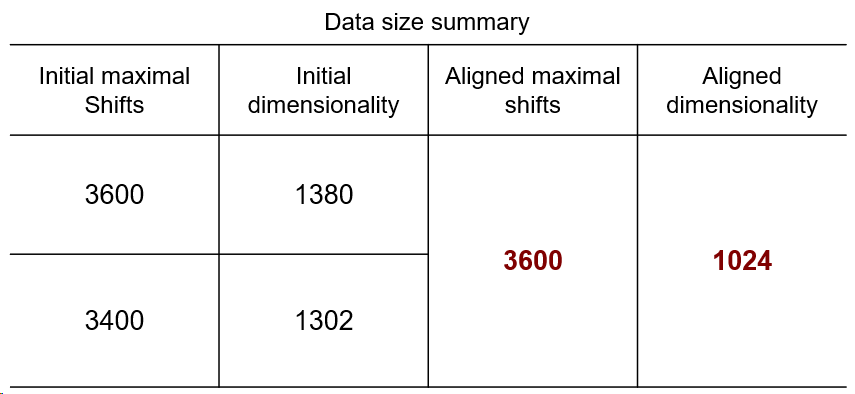


2.2 插值法降维

数据维数过高，会导致计算成本增加；数据维数过低则会导致局部信息缺失，模型性能差。所以在确保性能不受明显影响的前提下，适当降低维数，可以节约计算成本。在最大波数对齐后，数据维数达到1380维，采用插值法降低至1024维（2的倍数便于后续计算），结果发现降维前后数据基本信息仍然较好地得到保留，如下图：



数据预处理前后的基本尺寸指标如下表（以3400cm-1为例）：



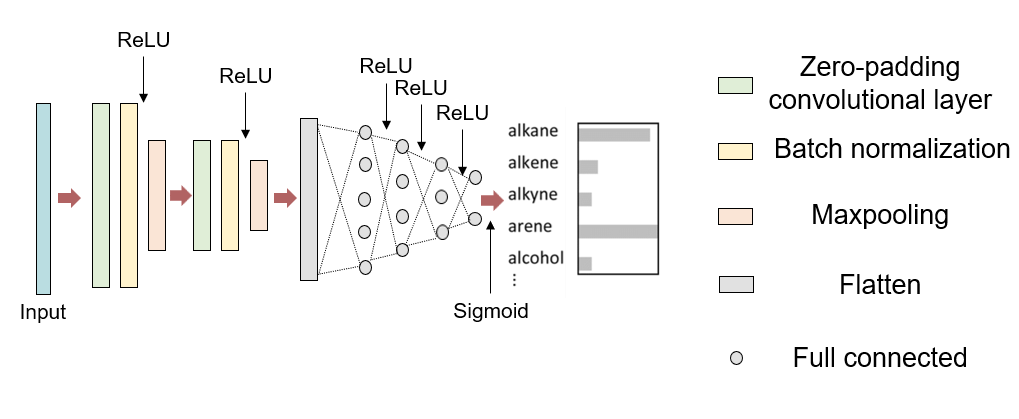
1. 搭建分类器模型并评估性能

设计并搭建用于官能团识别任务的分类器模型，是为了初步获得拥有一定识别性能的分类器，通过对其进行性能评估，可以为后续数据、模型架构、训练策略等方面的优化提供依据。

3.1 CNN架构

卷积神经网络（CNN）是应用最为广泛的ANN模型之一，它使用卷积操作生成输入数据的特征映射（feature map）。具体来说，一个或多个具有一定尺寸的权重矩阵（称为“核”）以一定的步长在输入矩阵上滑动，每滑动一次，核与输入矩阵中相同位置的元素相乘并相加，得到的值称为特征映射，所有特征映射形成一个张量。CNN还包括最大池化、dropout、全连接等操作，以实现不同的功能。例如在光谱识别任务中，卷积操作可以捕捉到谱线的走势、峰值等形状的特征映射，最大池化可以从特征映射中提取局部信息最丰富的部分。相比MLP，CNN通过卷积操作实现了权值共享，考虑了光谱数据的局部相关性并减少了参数量。

初步设计的CNN架构，依次包含两个“zero padding卷积层-归一化-ReLU激活-最大池化”组合（输出通道数分别为32和64），以及四个带dropout的全连接-激活函数组合：



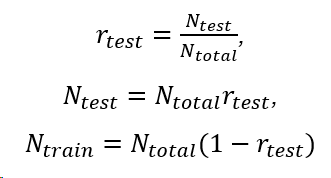
3.2 数据集划分

采用分层抽样的方法对数据集进行划分，相比完全随机划分，分层抽样可以保证每类标签在三个数据集中占比一致。以划分测试集和验证集为例：

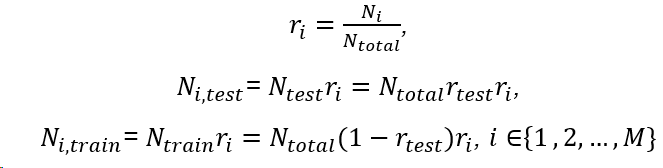
多输出二分类问题的标签是多维的，当两个标签的元素完全相同时，称这两个标签属于同一类：



首先，根据训练集和测试集的比例确定各自的数据量：



然后统计每类标签占数据总量的比例*ri*，并用该比例分别乘以测试集和验证集的数据量*Ni*，就可以得到该类标签在两个数据集中被分配的数量：

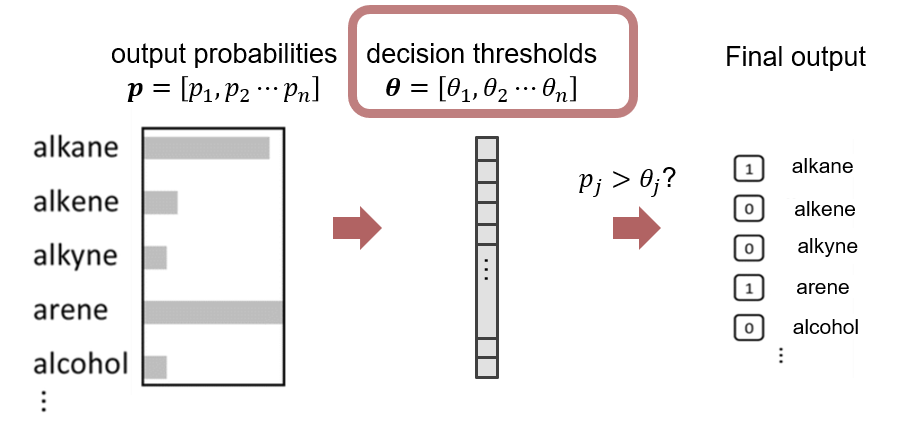


划分情况见下表：

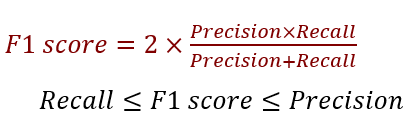
|  |  |
| --- | --- |
| 训练集 | |
| 数据量 | 总数60% |
| 单批次数据量 | 32 |
| 验证集 | |
| 数据量 | 总数20% |
| 单批次数据量 | 16 |
| 测试集 | |
| 数据量 | 总数20% |
| 单批次数据量 | 16 |

3.3 决策阈值优化

CNN经过sigmoid层输出一个37维向量，每一维代表一种官能团在输入样本中存在的概率。此时为每个官能团设定一个阈值，组成向量***θ***。比较每个概率与对应阈值的大小，若大于阈值，视为该官能团存在，输出标签的对应维度上的元素置为1，反之则认为不存在，对应元素置为0。然而，由于官能团在样本中的数量分布不平衡，导致模型在不同维度上输出的概率值大小差异很大，因此，我们希望为每个官能团找到一个合适的决策阈值，使模型拥有最好的分类效果。



为了评估决策阈值的优化情况，选取37类官能团的F1分数作为优化指标：



F1分数介于精确率和召回率之间，其值可以视为对二者的权衡，能够综合体现模型对某一类官能团的识别性能。

优化步骤为：(1) 将训练集和验证集输入训练过的模型，输出一组概率值。(2) 为每个输出概率从0到1选取一系列决策阈值，分别与输出概率比较，得到一系列的精确率和召回率，组成一条精确率-召回率曲线。(3) 计算每组精确率和召回率对应的F1值，选取其中的最大值，其所对应的决策阈值就是优化后的决策阈值。

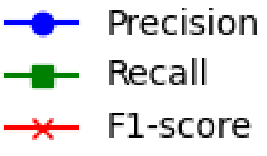
我们试图比较将所有阈值全部设为0.5（优化前）以及优化后模型的宏平均F1（Macro F1）。 宏平均F1是各类官能团F1的算术平均值，该值不对数据量作加权，对离群的F1值较为敏感，因此可以直观体现模型在少数类样本上分类性能的好坏：

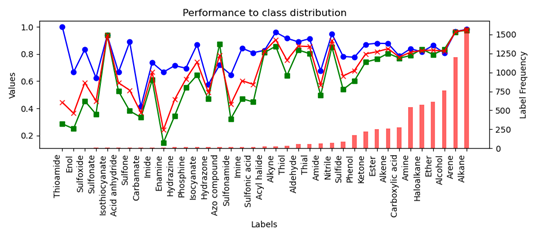


宏平均F1将作为后续实验中评价模型性能的重要指标。

3.4 CNN性能评估

为了测试模型的性能上限，训练时将测试集作为验证集，使测试集既用于验证又用于测试，而不采用验证集。训练持续100个epoch，期间保存验证准确率最高的模型。在决策阈值优化之前，使所有决策阈值全部为0.5。模型在测试集上识别各类官能团的精确率、召回率、F1分数以及各类官能团的频数分布如下：



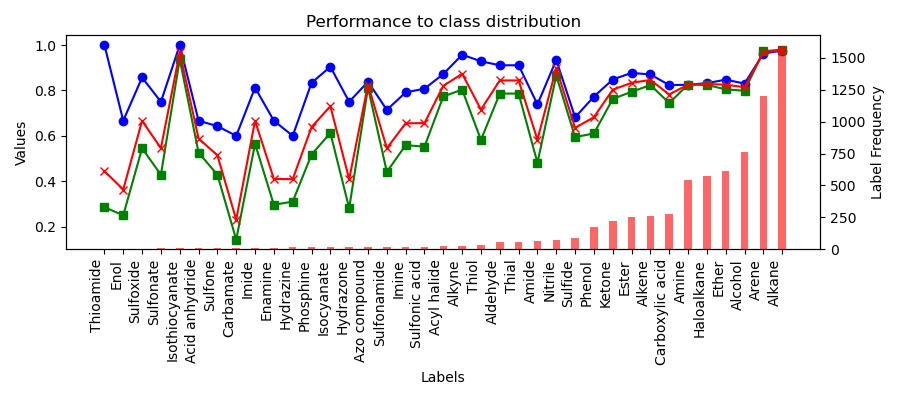


宏平均F1为0.6938。

可以看到，模型预测大部分官能团的精确度尚可，对少数类官能团的召回率较低。另外，对于多数类官能团，模型已经具备了较好的分类性能。

3.5 决策阈值优化

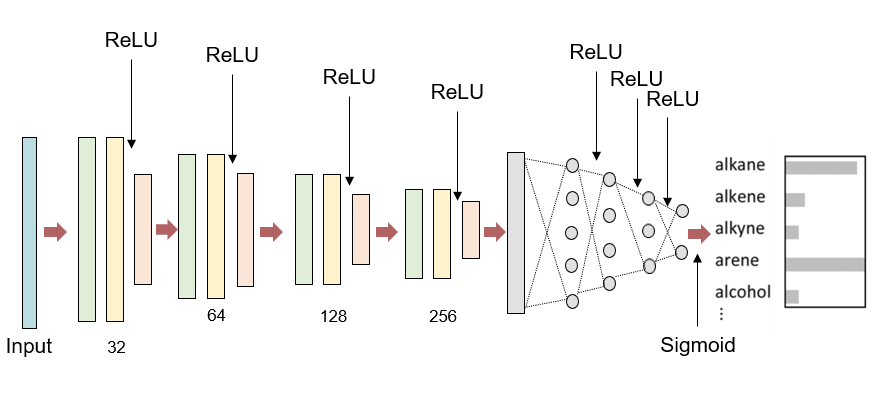
决策阈值优化后，测试结果如下：



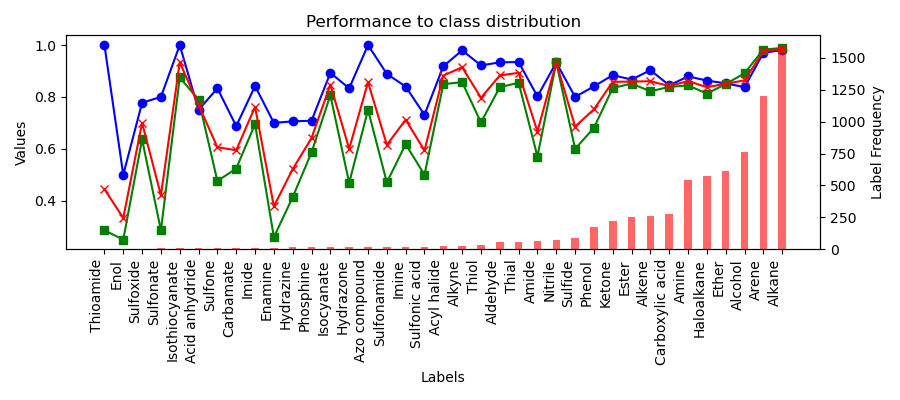
宏平均F1为0.6938。模型的性能没有显著提升。

1. 网络架构优化

目前模型识别性能较差，可能是由于其网络结构不能有效地从样本中提取特征信息。我们增加了两个卷积层，并通过每次卷积运算逐步增加输出通道数量，从而尽可能多地为输入样本生成特征图，如下图所示（下方的数字代表通道数）：



此时，我们采用完整的数据集，即训练、验证和测试集分别进行训练、验证和测试，考察模型在改进结构并优化决策阈值后的性能：



此时模型的宏平均F1为0.7441。可以看到，模型的整体性能有了较为明显的提升。

1. 显著性分析

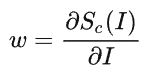
我们认为，官能团在样本中数量分布的不平衡，仍然是限制模型性能，特别是使模型对少数类官能团识别性能较差的根本原因。显著性分析有助于让我们了解模型对样本的识别依据，最好可以与现实中已有的理论（如振动光谱的谱图解析理论）相对应，从而提高模型的可解释性，这个过程也叫归因（attribution）。

* 1. 传统的显著性分析方法Saliency Map

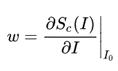
显著性图（Saliency Map）旨在将对于图像矩阵***I***的复杂分类识别模型*Sc*(***I***)进行一阶泰勒展开，获得一个线性模型：



其中***ω****c*是*Sc*(***I***)对图像各样本点的一阶导数（梯度）：

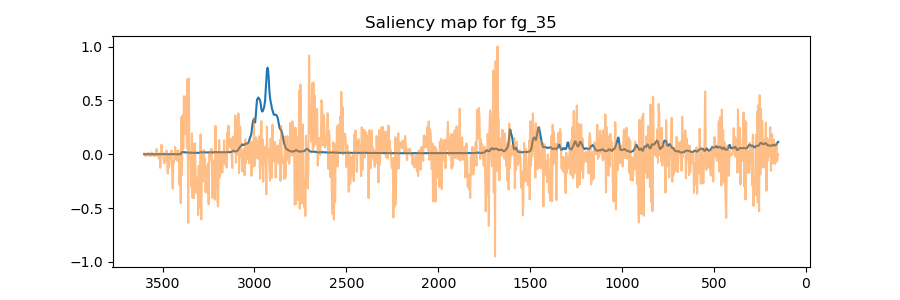


那么对于某一样本点***I0***，定义*Sc*(***I***)对***I***的各像素点在***I0***处（即输入位置）的梯度为模型对***I0***的显著性：

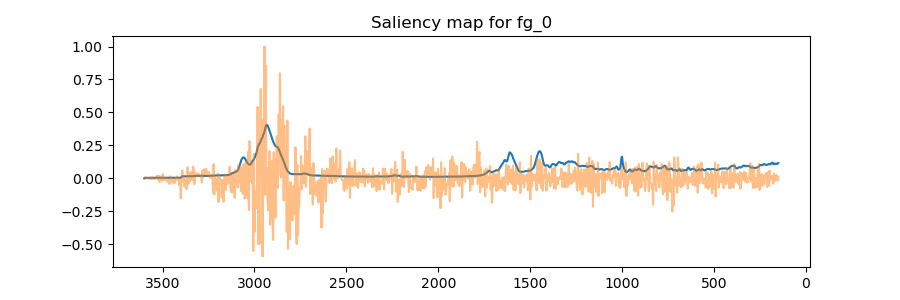


该值可以通过反向传播（即链式求导法则）计算。显然Saliency Map是对模型的简化估计，对于较为复杂的模型，由于其中各种参数的约束，当特征值增大到一定程度后，输入对输出的贡献将趋于饱和，梯度趋近于0，产生梯度消失现象，此时Saliency Map输出将趋近于0，变得没有意义。

以少数类官能团异硫氰酸酯（isothiocyanate）为例，计算测试集中含有该官能团的样本且识别正确的平均光谱（蓝）和平均显著性图（黄），如下：

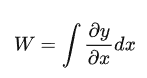


可以看到，梯度的分布较为散乱，但是1500-1700cm-1、2500-2700cm-1以及3300-3500cm-1处有相对突出的梯度分布。这可能是由于样本数量过少，模型对该官能团的识别仍处于欠拟合水平造成的。对比烷基的显著性图，可以发现烷基的梯度分布相对集中，其对应的波数恰好与烷基的C-H特征峰较为吻合，说明模型对于烷基这种多数类样本识别能力较强：

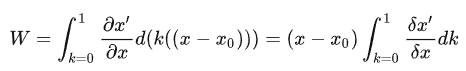


* 1. 基于积分梯度法（Integrated Gradients）的显著性

积分梯度法选取一个基线（一般是0），对梯度从基线到目标样本求积分，这样就能保证在梯度趋近于0的位置上，仍然返回一个较大的值作为模型的显著性：

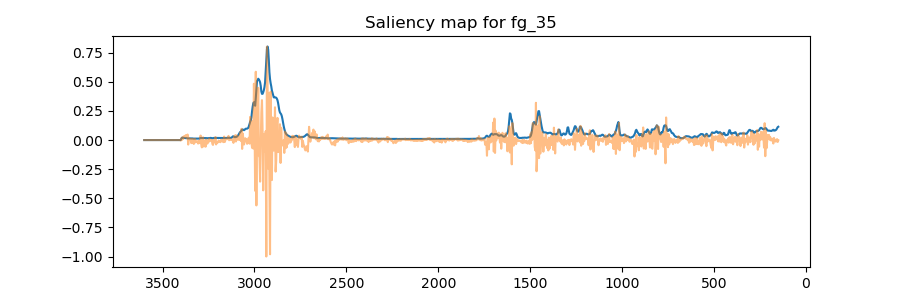


设选取的基线为*x*0，目标样本为*x*，可以通过线性插值构造变分积分：

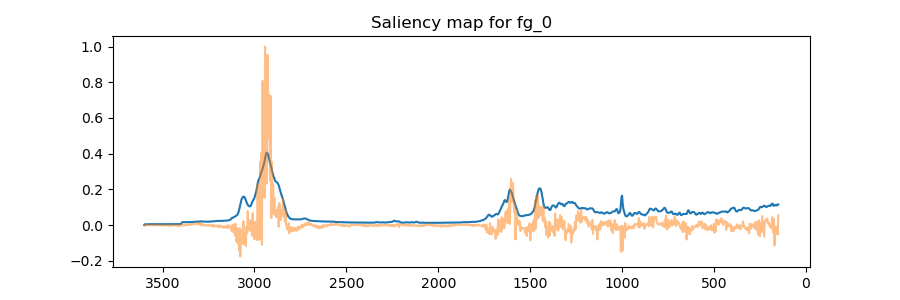


其中*x’ = x*0 + *k*(*x- x*0)。可见，基线的选择能够决定积分梯度法的输出值。

设定基线为0，异硫氰酸酯的平均光谱和平均积分梯度计算结果如下：



梯度集中了很多，但是大部分的波数位置与烷基特征峰重合，可能是烷基数量过多，影响了积分梯度的归因所致。根据烷基的结果，同样可以作证这一点：



这不禁让我们想到，既然烷基在有机物中普遍存在，那么可否将烷基的谱图作为积分梯度的基线进行归因？这样做是否能够降低烷基特征峰对于少数类模型归因的影响？

（二）后续安排

(1) 使用烷烃、烯烃等多数类作为基线进行积分梯度归因，进一步确定少数类样本的梯度分布。

(2) 结合拉曼光谱的解谱方法，对归因结果进行解释。