使用 Vision Transformer 进行图像分类

1、前言

Transformer 模型被大量应用在 NLP 自然语言处理当中，而在计算机视觉领域，Transformer 的注意力机制 attention 也被广泛应用，Vision Transformer 将 CV 和 NLP 领域知识结合起来，将具有自注意力的 Transformer 架构应用于图像块序列，而不使用卷积层。

2、实验步骤

(1)首先准备数据，可使用 CIFAR-100 效果较好，本人因电脑配置所以是自定义的数据集。

(2)因为图片较小，所以进行了一下数据增强，以有利于模型训练，之后使用 data_augmentation 里的 layers.Normalization()对训练数据进行归一化。

(3)实现 patch 并创建为一个层。这个 patch 就是我们要把一个图片分成大小相同的几块，相当于在 nlp 里面把一句话进行分词处理。

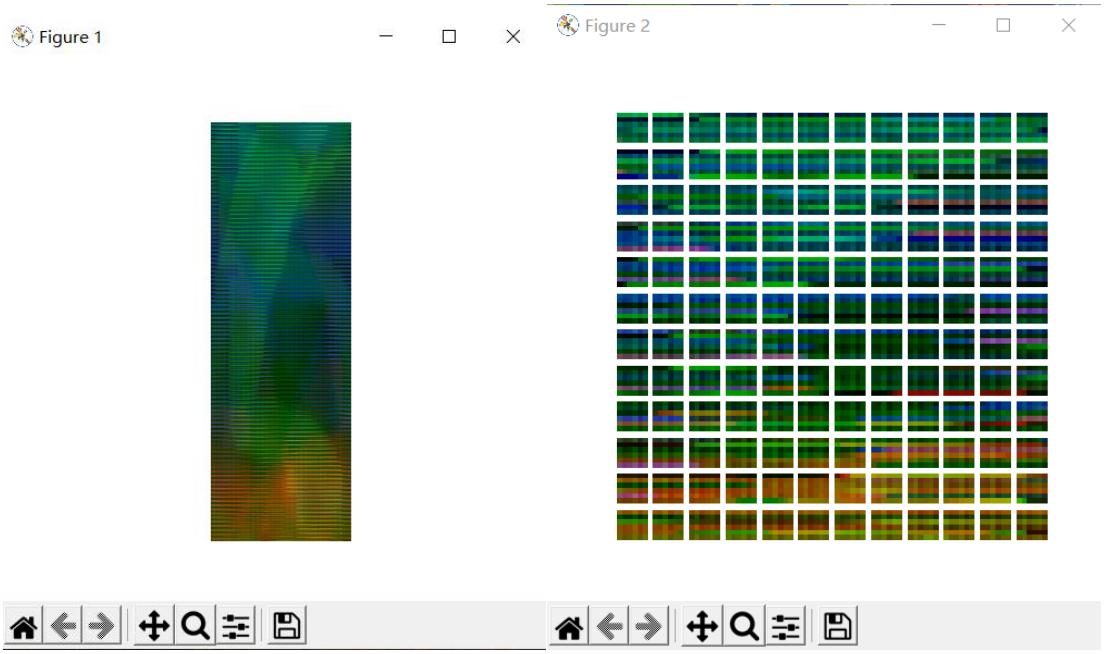
(4)实现 patch 编码层。PatchEncoder 层将通过将 patch 投影到一个大小为 projection_dim 的向量来线性变换 patch。此外，该算法还在投影向量上添加了一可学习的位置嵌入。patch 编码层就是实现 patch_embedding+positions_embedding。

(5)构建 ViT 模型。ViT 模型由多个使用层的 Transformer 块组成。multihead 注意层作为一种自注意机制应用于 patch 序列。它将可学习的嵌入预先添加到编码补丁序列中以用作图像表示，最终 Transformer 块的所有输出都被重新整形 layers.Flatten()并用作分类器头的图像表示输入。

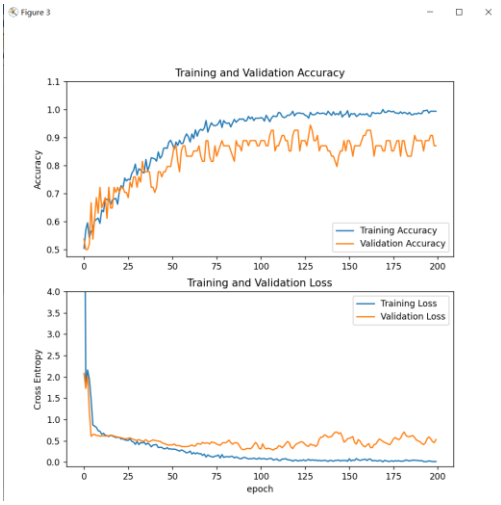
(6) 最后便是编译、训练和评估模式。

3、实验结果

由 patch 层随机生成的图像块



可视化结果:自定义小数据集测试结果　　使用数据 CiFar10 的测试结果



4、结论

本实验只是很简单的实现一个用 ViT 模型实现图片分类的例子，实验结果并不优势。实际上 ViT（或者说 Transformer）需要大数据量做预训练，随着预训练数据量增加，ViT 的准确率也会一直增加。当预训练图片数据量超过 1 亿张时，ViT 图片分类效果将会优于效果最好的卷积神经网络 ResNet。

我认为要提高模型质量，可以尝试将模型训练更多 epoch、使用更多的 Transformer 层、调整输入图像的大小、更改补丁大小或增加投影尺寸。