日期：2022.10.28
报告人：李号元
记录人：李号元
分享论文：TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization

Q&A：

1. Q(陈匡义)： 这个网络的输入和输出是什么?

   A：网络的输入是一张全景图或者是卫星图（比如640 x 320 x 3的RGB图像），输出是一个1xD的嵌入特征。通过计算特征之间的距离（比如余弦距离或者L2距离），将距离最小的特征作为检索的结果。

2. Q(陈匡义)：第一阶段输出的atten map是怎样得到的？是类似分类对应的某一类最大的的那种feature map吗?

   A：atten map并不是对应了哪一类，而是将注意力权重进行了求和，得到的一个数值作为对应patch的权重。

   原文论文中提及到具体的计算方式：在最后一层，由图像patch的tokens和class token计算多重注意力权重，并对多重注意力进行求和得到的每一个patch的权重。

   > we select the correlation between class token and all other patch tokens as the attention map and reshape to the original image shape.

   对应代码如下

```
if i == len(self.blocks)-1:  # len(self.blocks)-1:
        y = blk.norm1(x)
        B, N, C = y.shape
        qkv = blk.attn.qkv(y).reshape(B, N, 3, blk.attn.num_heads, C //
blk.attn.num_heads).permute(2, 0, 3, 1, 4)
        q, k, v = qkv[0], qkv[1], qkv[2]  # make torchscript happy
(cannot use tensor as tuple)

        att = (q @ k.transpose(-2, -1)) * blk.attn.scale
        att = att.softmax(dim=-1)

        last_map = (att[:, :, :2,
2:].detach().cpu().numpy()).sum(axis=1).sum(axis=1) #atten between
class token and other patch tokens
        last_map = last_map.reshape(
            [last_map.shape[0],x_shape[2] // 16, x_shape[3] // 16])
```

这段代码意思是：在第一阶段的最后一层中，用token计算query、key和value，并计算注意力权重。将class token（`q[:, :, :2,:]`)和其他token的注意力进行求和作为每一个patch的权重，保存下来用于第二阶段的cropping。

3. Q（陈匡义）：中间的transformer有什么作用？CNN也可以起到类似的效果，也能生成相似的atten map。如何监督学到的?

   A：网络的监督只有最后面的Triplet Loss做监督，中间的atten map可以看成是网络生成的中间结果，并没有直接的监督。CNN可能也能起到对应的结果，但可能操作流程不如transformer来的自然，其中包括位置编码、注意力和token，用transformer的框架很多现成的代码，比如作者就是直接在vision transformer上继承的。

4. Q（杨老师）：你这个工作主要的应用场景是什么？是对目标进行定位吗？要说清楚这个工作的使用场景。

   A：这个工作主要用于对拍摄者（行人）自身的定位，还没有做对场景目标的定位。跟无人机拍摄图像的差距是：这个工作针对的应用是在行人获取街景照片时，通过图像检索的方式对行人所在区域进行定位。

5. (杨老师)： 在介绍工作的时候要清晰明了地介绍应用场景和要解决的问题。并且针对不同的听众，讲解的侧重点要有针对性。对不是很熟悉这个方向的听众，要着重介绍这个领域要做什么，有哪些问题。而在组会中以及一些比较熟悉这方向的会中，可以把重心放在算法的讲解中。