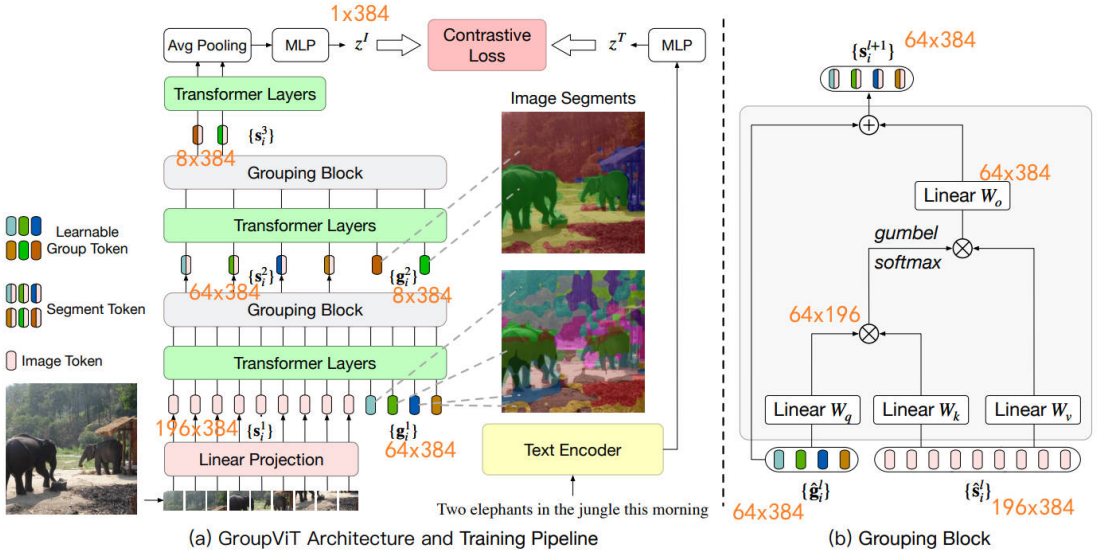
GroupViT: Semantic Segmentation Emerges from Text Supervision

动机：

semantic segmentation mask的标注，是一件十分“贵”的事情，在end-to-end deep networks出现之前的时代，在visual sense understanding任务中，图像grouping和recognition是两个很重要的手段。为了解决这个问题，并受到grouping的启发，提出GroupViT，它只用Text来做监督，用对比学习来训练。

模型框架：



(a) GroupViT Architecture and Training Pipeline (b) Grouping Block

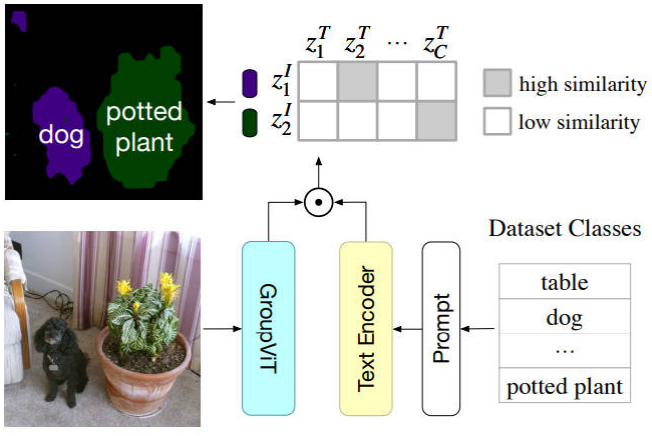
Grouping Block：用到了矩列中心操作，其不可导，用Gumbel Softmax这个trick，就使得过程可导了，便可通过反向传播进行训练了

训练：

$\{g_i\}$的个数表示这张图片将被分为几个grouping

Transformer Layers学习到底哪个$\{S_i\}$属于$\{g_i\}$，从而通过Grouping Block将属于同一个$\{g_i\}$的$\{S_i\}$合并

Zero-Shot 推理：



z_1^I和z_2^I是模型架构图，中的$\{s_i^3\}$中的前两个，表示这个图中的两个grouping

下标C表示从原始text labels中提取出来的名词数量

两种层次，1. 几何学层次，分割的轮廓正确；2. 语义层次，对应的类别正确

Limitations：

1. 不能对背景分类，这个和CLIP的特性很像（很擅长学习明确的概念，但是对于模糊的概念则不太擅长），当下的权宜之计是设定一个阈值（只有similarity值最大且超过了该阈值才认为是前景类，否则认为是背景），未来的工作可以是将背景的学习融入到模型的训练过程中。

2. 还未将语义分割中的强大模型融入到GroupViT模型中