GroupViT: Semantic Segmentation Emerges from Text Supervision

Jiarui Xu[1]*, Shalini De Mello[2], Sifei Liu[2], Wonmin Byeon[2], Thomas Breuel[2], Jan Kautz[2], Xiaolong Wang[1]

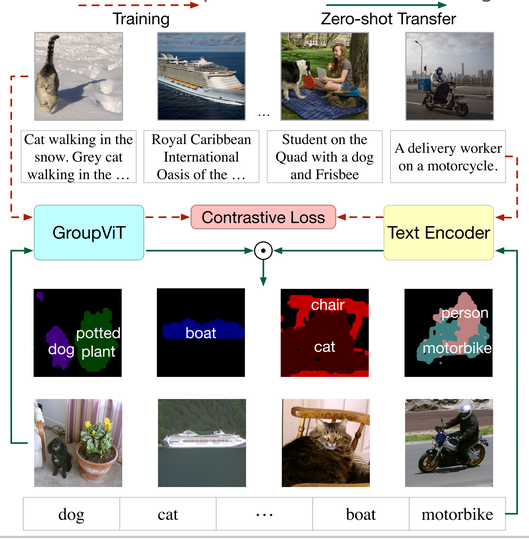
[1]UC San Diego, [2]NVIDIA, *work conducted at an internship at NVIDIA

Goal

To learn semantic segmentation with text supervision and **without** mask labels.

Method Overview

First, we jointly train GroupViT and a text encoder using paired image-text data. With GroupViT, meaningful semantic image groups automatically emerge. Then, we transfer the rained GroupViT to zero-shot semantic segmentation.



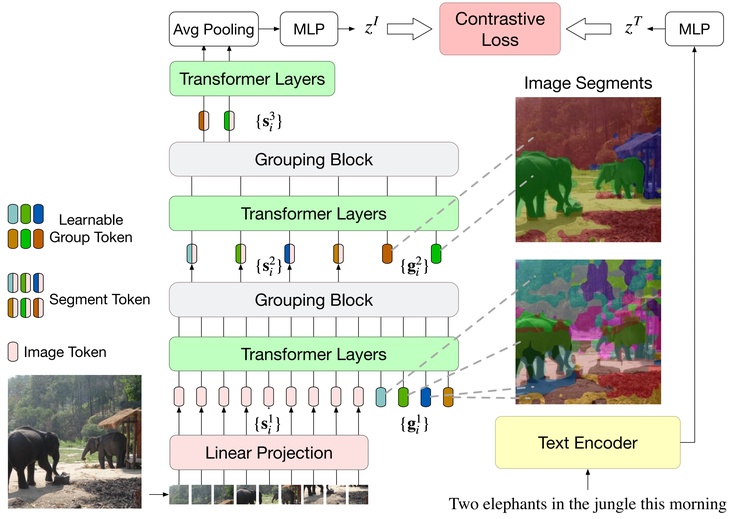
Main Contributions

- Moving beyond regular-shaped image grids in deep networks and performing bottom-up grouping of visual concepts
- Zero-shot transfer to semantic segmentation without any pixel-level labels

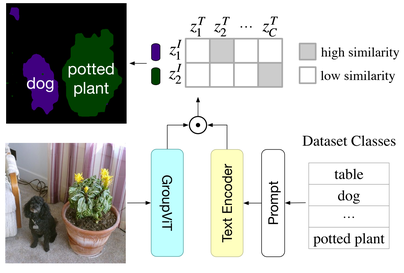
Code and demo!

Architecture and Training Pipeline of GroupViT

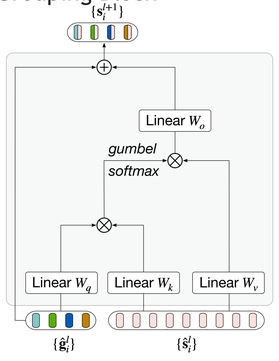
GroupViT contains Transformer layers with a hierarchy of grouping stages. Each stage groups image pixels into progressively larger visual segments. The images on the right show visual segments that emerge in the different grouping stages.



Zero-shot Transfer to Semantic Segmentation



Grouping Block



Visualization Results without Training on any Mask annotations