GroupViT: Text-Based Training for Visual Segmentation Without Pixel-Level Labels

Nisarg Patel - 25PGAI0104
Jaydutt Desai - 25PGAI0017
Nishant Pandey - 25PGAI0042

Agenda

- Introduction
- Problem Statement
- Proposed Solution: GroupViT
- GroupViT Architecture
- Training and Methodology
- Experiments and Results
- Comparative Analysis
- Future Directions

What is Semantic Segmentation?

• Dividing an image into meaningful parts and labeling each part with a category (e.g., car, person, background).

• Traditionally relies on detailed pixel-level annotations.

Underlying Need for the Study

• High cost and effort involved in creating pixel-level labels.

• Limitations of traditional models that can’t generalize well to unseen objects.

• New approach using text supervision (text descriptions paired with images).

Problem Statement

Challenge: Can we train a model for semantic segmentation using only image-text pairs and no pixel-level labels?

Solution: Introduce a grouping mechanism into deep networks to allow semantic segments to emerge automatically using text supervision.

What is GroupViT?

- A new model called the Grouping Vision Transformer (GroupViT).
- Uses hierarchical grouping to create arbitrary-shaped image segments.
- Trained with a text encoder using image-text pairs.

Key Features:

- **No pixel-level labels required.**
- **Capable of zero-shot learning (generalizes to unseen categories).**
- **Uses a hierarchical grouping method to identify semantic segments.**

Overview of GroupViT Architecture

Vision Transformer (ViT) Basics

• Breaks images into smaller patches.

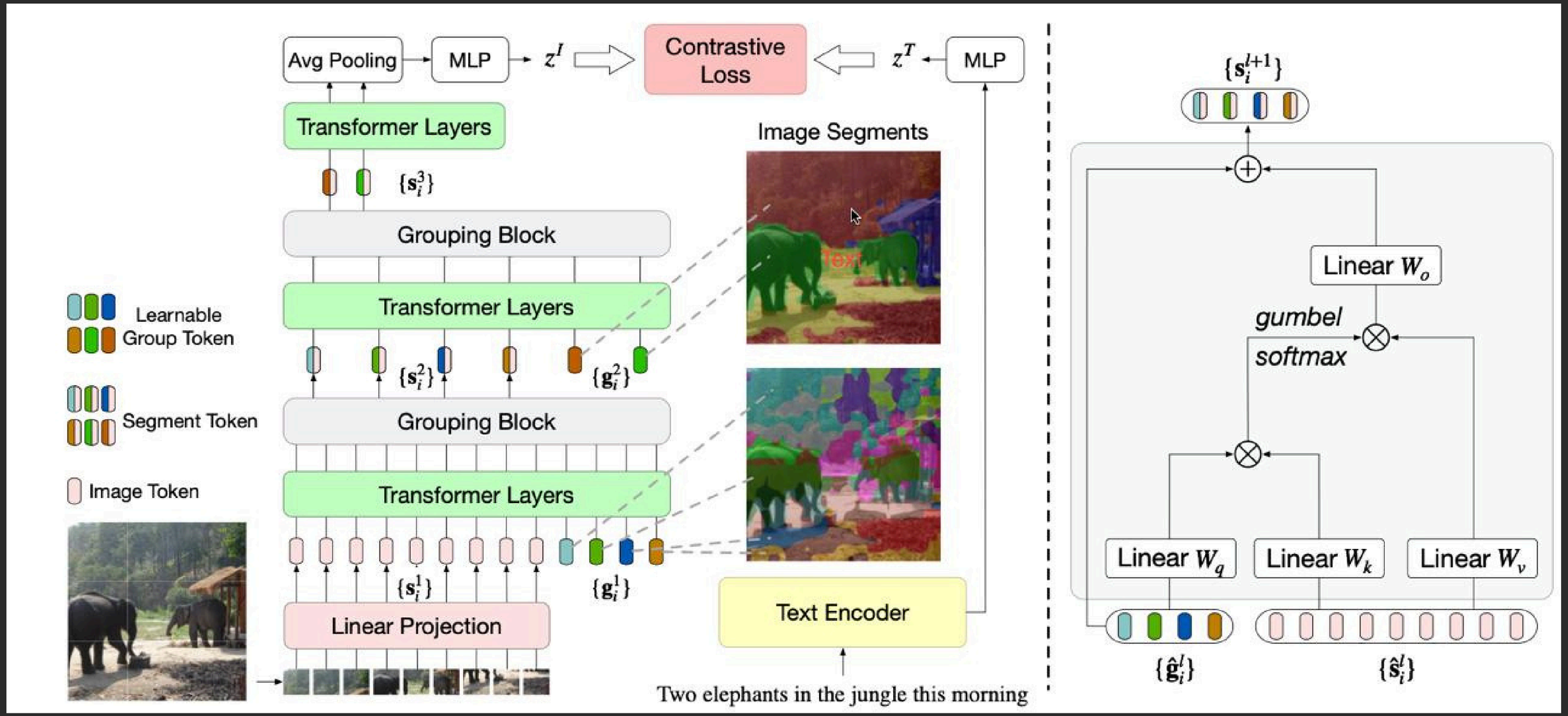
• Uses self-attention to analyze and compare different parts of the image.

Hierarchical Grouping in GroupViT

• Progressive grouping: small patches grouped into segments, segments merged into larger objects.

• Uses a Grouping Block to merge segments based on their visual similarity.

GroupViT architecture



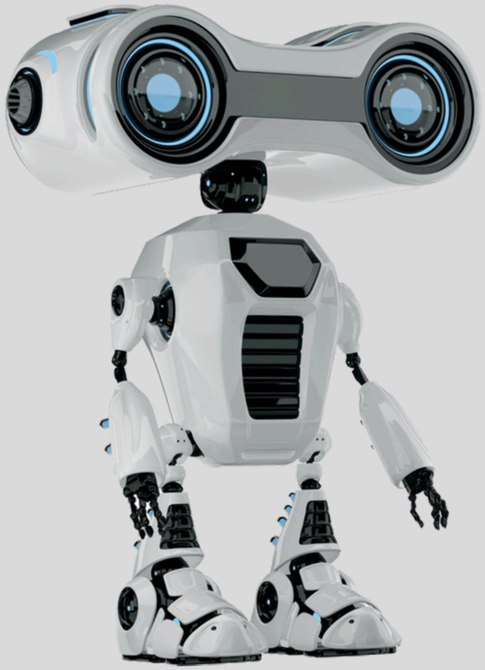
Training GroupViT

Training Data: Image-text pairs from large-scale datasets like CC12M and YFCC.

Contrastive Learning:

- Positive pairs: Correctly matched image and text.
- Negative pairs: Incorrectly matched pairs.

Objective: Bring the positive pairs closer together and push the negative ones apart in the model's understanding.



How GroupViT Learns Segmentation

- **Patches to Segments:**

• Patches are organized into segments using a hierarchical structure.
• Uses Gumbel-Softmax for decision-making in grouping.

- **Zero-Shot Semantic Segmentation:**

• Can perform segmentation without further training or fine-tuning.
• Matches segments with the closest description in the learned text embeddings.

Experiments and Evaluation

Datasets Used:

- PASCAL VOC 2012: 20 object categories.
- PASCAL Context: 59 object categories.

Key Metrics:

- Mean Intersection over Union (mIoU): Measures accuracy by comparing predicted and actual segments.

Ablation Study Results

- Soft Assignment: 12.0% mIoU
- Hard Assignment: 36.7% mIoU

Impact of Group Tokens

- 16 Group Tokens: 28.6% mIoU
- 64 Group Tokens: 39.3% mIoU

Comparison with Zero-Shot Baselines

- ViT with Pixel-Wise Classification: 20.1% mIoU
- ViT with K-means Clustering: 25.0% mIoU
- GroupViT: 52.3% mIoU (Highest)

Comparison with Fully-Supervised Models

Fully-Supervised ViT on PASCAL VOC 2012:

- Achieves 53.0% mIoU with detailed pixel-level labels.

GroupViT's Zero-Shot Performance:

- Achieves 52.3% mIoU, which is nearly equivalent to the fully-supervised method.

Future Directions

Enhancements:

• Improve handling of background categories using specialized techniques.
• Integrate segmentation-specific tools like **Dilated Convolutions** or **U-Net** for better accuracy.

Broader Impact:

• Moving toward more accessible and scalable training methods.
• Potential to apply text supervision to other detailed visual tasks.

Thank You