Note: GroupViT: Semantic Segmentation Emerges from Text Supervision

Sinkoo

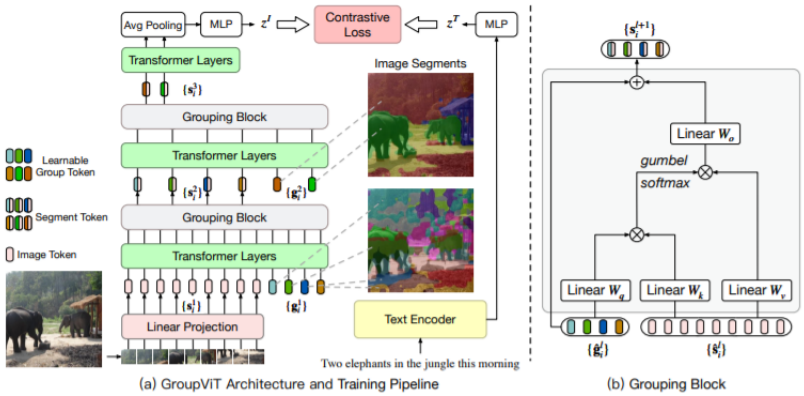
April 3, 2022

1 Abstract

Semantic segmentation is commonly achieved via a FCN. There are 2 limitation: 1, Learning is limited by high-cost of per-pixel human labels; 2, Can not generalize unseen categories.
The paper developed CroupViT(Group Vision Transformer) trained with text supervision only, and it is Bottom-up. This work is the first to perform semantic segmentation on different label vocabularies in a zero-shot manner with text supervision alone, without requiring any pixel-wise labels.

2 Method

2.1 Grouping Vision Transformer



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

Figure 2. (a) **The Architecture and Training Pipeline of GroupViT.** GroupViT contains a hierarchy of Transformer layers grouped into stages, each operating on progressively larger visual segments. The images on the right show visual segments that emerge during different grouping stages. The lower stage groups pixels into object parts, e.g., noses and legs of elephants; and the higher stage further merges them into entire objects, e.g., the whole elephant and the background forest. (b) **The Architecture of Grouping Block**. Each grouping stage ends with a grouping block that computes the similarity between the learned group tokens and segment (image) tokens. The assignment is computed via gumbel softmax over group tokens and converted into one-hot hard assignment. The segment tokens assigned to the same group are merged together and represent new segment tokens that are input to the next grouping stage.

Architecture First split a image into N non-overlapping patches and linearly project them to a latent space, denote them as $\{p_i\}_{i=1}^{N}$. In each grouping stage concatenate a set of learnable group tokens.

Multi-stage Grouping Each stage there is a Grouping Block to merge small groups to larger ones.
Information propagation:

$$\{\hat{g}_i^l\}, \{\hat{s}_i^l\} = Transformer([\{\hat{g}_i^l\}; \{\hat{s}_i^l\}])$$

Grouping Block:

$$\{s_i^{l+1}\} = GroupingBlock(\{\hat{g}_i^l\}, \{\hat{s}_i^l\})$$

g is the group token(learnable), and s is the segment token.
After final Grouping stage, apply transformer layer on all segment tokens and average their outputs to obtain the final global image representation.

Grouping Block Figure2. ($W_o, W_q, W_k, W_v$are all learnable) Formula:

$$A_{i,j}^l = \frac{exp(W_q \hat{g}_i^l \dot{W}_k \hat{s}_j^l + \gamma_i)}{\sum_{k=1}^{M_l} exp(W_q \hat{g}_k^l W_k \hat{s}_j^l + \gamma_k)}$$

$$\hat{A}^l = one - hot(A_{argmax}^l) + A^l - sg(A^l)$$

$$s_i^{l+1} = \hat{g}_i^l + W_o \frac{\sum_{j=1}^{M_l} \hat{A}_{i,j}^l W_v \hat{s}_j^l}{\sum_{j=1}^{M_l} \hat{A}_{i,j}^l}$$

2.2 Learning from Image-Text Pairs

This model employs carefully-designed contrastive losses between image-text pairs.

Image-Text Contrastive Loss Train a dual-encoder architecture via an image-text contrastive loss: GroupViT to process image and a transformer to process the text. The Final image embedding is the average embedding of all GroupViT output segment tokens, the final text embedding is the Transformer embedding of the last token. Consider all matched image-text pairs as positive pairs, and all other unmatched ones as negative ones. The objective is to get right pair closer and wrong pair far.

$$\mathcal{L}_{I \leftrightarrow T} = \mathcal{L}_{I \to T} + \mathcal{L}_{T \to I}, \tag{6}$$

which is composed of an image-to-text contrastive loss defined as

$$\mathcal{L}_{I \to T} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(z_i^I \cdot z_i^T / \tau)}{\sum_{j=1}^{B} \exp(z_i^I \cdot z_j^T / \tau)},$$

and a text-to-image contrastive loss defined as

$$\mathcal{L}_{T \to I} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(z_i^T \cdot z_i^I / \tau)}{\sum_{j=1}^{B} \exp(z_i^T \cdot z_j^I / \tau)},$$

where τ is a learnable temperature parameter to scale the logits.

Multi-Label Image-Text Contrastive Loss Use the "prompting engineering" to generate additional text labels by randomly selecting K noun words from a sentence and prompt each of them with a set of handcrafted sentence templates.
And analog Image-Text contrastive pairs:

$$\mathcal{L}_{I \leftrightarrow \{T_k\}_{k=1}^K} = \mathcal{L}_{I \rightarrow \{T_k\}_{k=1}^K} + \mathcal{L}_{\{T_k\}_{k=1}^K \rightarrow I}, \tag{7}$$

which is a sum of two two-way contrastive losses

$$\mathcal{L}_{I \rightarrow \{T_k\}_{k=1}^K} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\sum_{k=1}^{K} \exp(z_i^I \cdot z_i^{T_k} / \tau)}{\sum_{k=1}^{K} \sum_{j=1}^{B} \exp(z_i^I \cdot z_j^{T_k} / \tau)}$$

and

$$\mathcal{L}_{\{T_k\}_{k=1}^K \rightarrow I} = -\frac{1}{KB} \sum_{k=1}^{K} \sum_{i=1}^{B} \log \frac{\exp(z_i^{T_k} \cdot z_i^I / \tau)}{\sum_{j=1}^{B} \exp(z_i^{T_k} \cdot z_j^I / \tau)}.$$

Finally, the total image-text contrastive loss for training GroupVIT is defined as

$$\mathcal{L} = \mathcal{L}_{I \leftrightarrow T} + \mathcal{L}_{I \leftrightarrow \{T_k\}_{k=1}^K}. \tag{8}$$

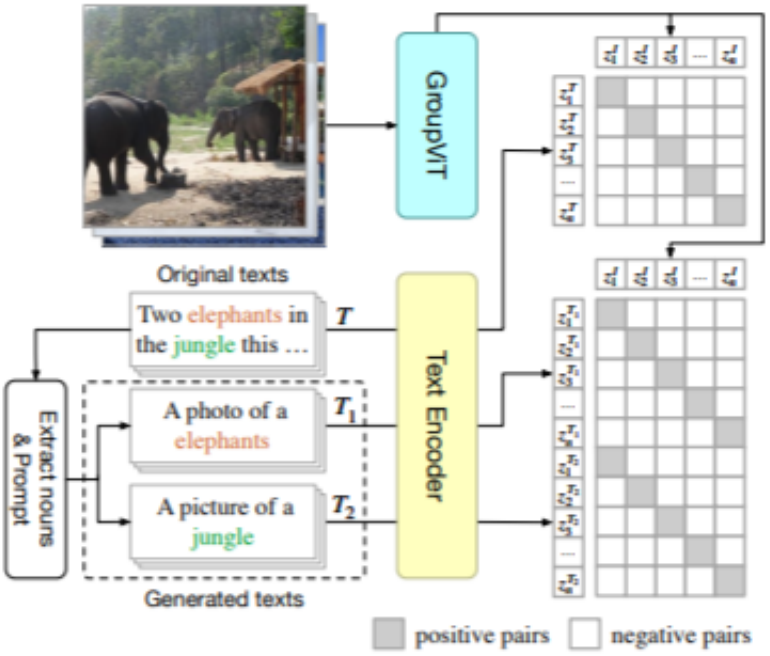


Figure 3. **Multi-label Image-Text Contrastive Loss**. Given an input image-text pair, we generate new text from the original text by extracting its nouns and by prompting them with several sentence templates. For constrastive learning, we treat only matched image and text pairs as positive ones. We train GroupViT and the text encoder to maximize the feature similarity between positive image-text pairs and minimize it between the negative pairs.

2.3 Zero-Shot Transfer to Semantic Segmentation

GroupViT's output can be easily zero-shot transferred to semantic segmentation without any further fine-tuning.

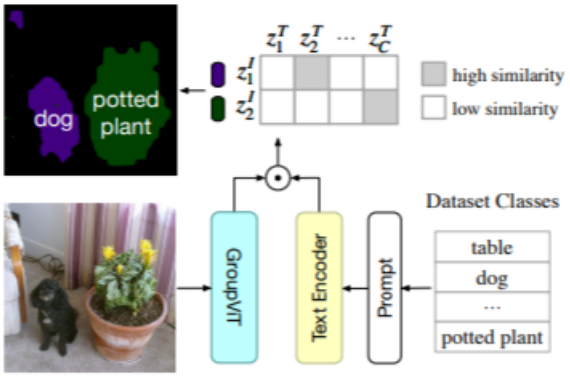


Figure 4. **Zero-Shot Transfer of GroupViT to Semantic Segmentation**. Each output segment's embedding from GroupViT corresponds to a region of the image. We assign each output segment to the object class with the highest image-text similarity in the embedding space.

Each segment token corresponds to an arbitrarily-shaped region of the input image. We then compute the similarity between the embedding of each segment token and the text embedding of all the semantic classes present in the dataset. We assign each image segment to the semantic class with the highest image-text embedding similarity.

3 Ablation Study

1, hard assignment improves over soft assignment by a large margin.
2, Multi Label Contrastive Loss helps GroupViT better classify the learned image segments.
3, increasing group tokens consistently improves performance.
4,2-stage GroupViT generates smoother segmentation maps compared to its 1-stage variant.

4 Summary

GroupViT is the first model to use pure text to learn semantic segmentation with grouping, and successfully transferred to zero-shot. But there are two potential improvements of GroupViT: its performance is lower versus PASCAL VOC. And GroupViT's architecture currently doesn't integrate segmentation-specific

enhancements(like dilated Convolution).

5 Other Knowledge

[1], Superpixels: combine some pixels which share similar properties to a large pixel.