Open-vocabulary Semantic Segmentation with Frozen Vision-Language Models

Chaofan Ma, Yuhuan Yang, Yanfeng Wang, Ya Zhang, Weidi Xie

Coop. Medianet Innovation Center, Shanghai Jiao Tong University, China
Shanghai AI Laboratory

ABSTRACT

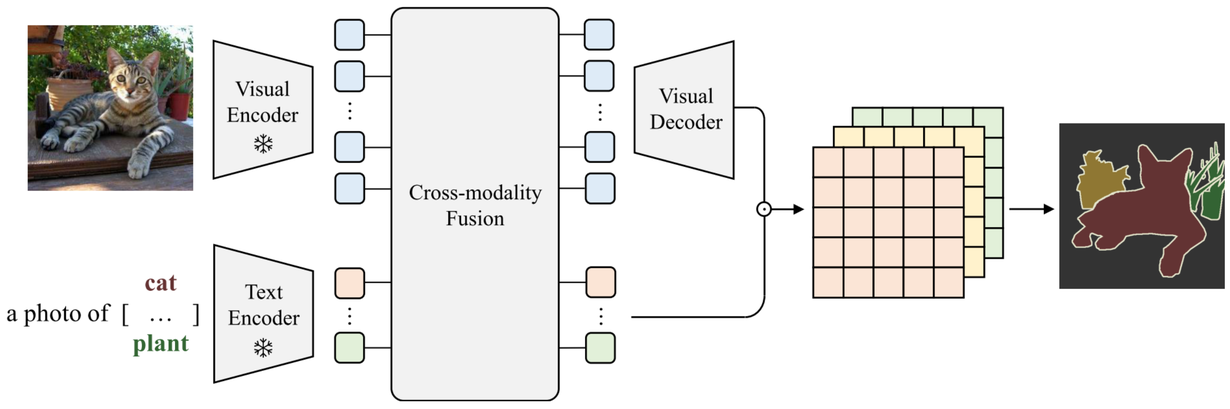
When trained at a sufficient scale, self-supervised learning has exhibited a notable ability to solve a wide range of visual or language understanding tasks. In this paper, we investigate simple, yet effective approaches for adapting the pre-trained foundation models to the downstream task of interest, namely, open-vocabulary semantic segmentation. To this end, we make the following contributions: (i) we introduce *Fusioner*, with a lightweight, transformer-based fusion module, that pairs the frozen visual representation with language concept through a handful of image segmentation data. As a consequence, the model gains the capability of zero-shot transfer to segment novel categories; (ii) without loss of generality, we experiment on a broad range of self-supervised models that have been pre-trained with different schemes, e.g. visual-only models (MoCo v3, DINO), language-only models (BERT), visual-language model (CLIP), and show that, the proposed fusion approach is effective to any pair of visual and language models, even those pre-trained on a corpus of uni-modal data; (iii) we conduct thorough ablation studies to analyze the critical components in our proposed *Fusioner*, while evaluating on standard benchmarks, e.g. PASCAL-5i and COCO-20i , it surpasses existing state-of-the-art models by a large margin, despite only being trained on frozen visual and language features; (iv) to measure the model's robustness on learning visual-language correspondence, we further evaluate on a synthetic dataset, named Mosaic-4, where images are constructed by mosaicking the samples from FSS-1000. *Fusioner* demonstrates superior performance over previous models.

Motivation

- Self-supervised representation learning on *classification*
  - Vision: MoCo, DINO, …
  - Language: BERT, …
  - Multi-Modal: CLIP, ALIGN, …
- Segmentation: a more challenging task
  - the conventional strategy: the model can only segment objects of the training set classes, and lacks the ability to handle samples from novel (unseen) classes
  - open-vocabulary semantic segmentation: segmenting objects from any categories by their textual names or description
- *Explore efficient ways to **bridge** the powerful pre-trained vision-only, language-only or visual-language models, for **open-vocabulary semantic segmentation***

Methodology

We first encode the input image and category names with frozen, self-supervised visual and language models. The computed features are then concatenated and passed to a transformer-based fusion module, that enables the features to be iteratively updated, in condition of the other modality through self-attention. As to obtain the segmentation mask, we further measure the cosine similarity between each pixel and the textual features of each category, *i.e.*, computing dot product between the L2-normalized visual and language embeddings.

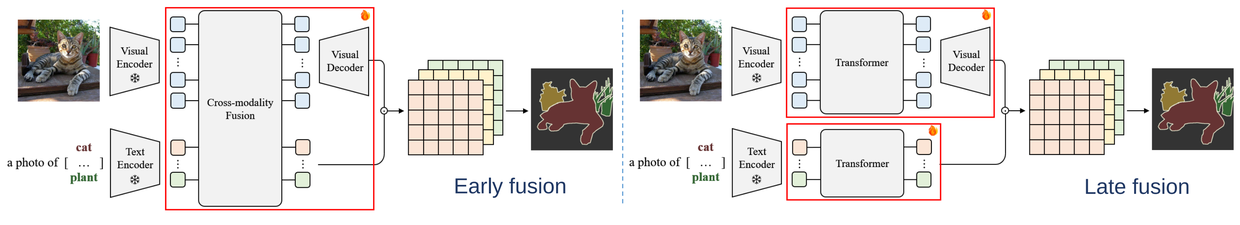


Comparison with State-of-the-art

We compare our method with various open vocabulary segmentation methods: ZS3Net, SPNet and LSeg on PASCAL-5i and COCO-20i.

Model	Backbone	PASCAL-5^i					COCO-20^i				
		5^0	5^1	5^2	5^3	mIoU	20^0	20^1	20^2	20^3	mIoU
SPNet [51]	ResNet101	23.8	17.0	14.1	18.3	18.3	-	-	-	-	-
ZS3Net [6]	ResNet101	40.8	39.4	39.3	33.6	38.3	18.8	20.1	24.8	20.5	21.1
LSeg [27]	ResNet101	52.8	53.8	44.4	38.5	47.4	22.1	25.1	24.9	21.5	23.4
LSeg [27]	ViT-L/16	61.3	63.6	43.1	41.0	52.3	28.1	27.5	30.0	23.2	27.2
Fusioner (ours)	ResNet101	46.8	56.0	42.2	40.7	46.4	26.7	34.1	26.3	23.4	27.6
Fusioner (ours)	ViT-L/14	**61.9**	**70.0**	**51.2**	**52.7**	**59.0**	**31.7**	**35.7**	**34.9**	**31.8**	**33.5**

Ability to Bridge Different Pre-trained Backbones



Model	Visual Encoder	Text Encoder	Early Fusion					Late Fusion				
			5^0	5^1	5^2	5^3	mIoU	5^0	5^1	5^2	5^3	mIoU
A	CLIP-B	CLIP-B	50.2	62.4	**51.5**	44.4	52.1	46.3	55.2	36.4	36.7	43.6
		BERT-B	46.6	61.3	44.6	43.7	49.1	47.5	58.5	38.5	39.1	45.9
		BERT-L	48.5	59.3	47.5	43.8	49.7	47.0	56.6	41.4	38.8	46.0
B	CLIP-L	CLIP-L	**61.9**	**70.0**	51.2	52.7	**59.0**	52.6	58.4	45.9	43.7	50.1
		BERT-B	56.9	66.0	45.9	49.6	54.6	56.5	60.2	44.6	47.0	52.1
		BERT-L	56.0	64.4	43.7	52.0	54.0	55.3	60.5	46.3	46.3	52.1
C	DINO-B	CLIP-B	56.4	67.1	49.8	47.5	55.2	57.1	64.6	48.5	46.2	54.1
		BERT-B	56.7	65.1	48.5	45.2	53.9	54.2	61.4	48.7	48.8	53.3
		BERT-L	56.1	65.8	48.7	44.2	53.7	57.4	65.1	48.5	46.3	54.3
D	MoCo-B	CLIP-B	58.9	67.0	47.7	51.7	56.3	56.4	64.9	47.1	49.5	54.5
		BERT-B	59.7	65.3	47.3	**53.4**	56.4	57.4	66.1	46.6	51.2	55.3
		BERT-L	59.1	65.7	49.5	53.1	56.8	54.9	65.0	49.8	50.6	55.0

Early fusion refers to our proposed fusion approach (left), while late fusion denotes similar the idea as LSeg (right).

Observations:

- Pairing the frozen visual and language models can be surprisingly powerful, even for models that are pre-trained on a corpus of uni-modal data
- Early fusion consistently outperforms the late fusion, that validates the conjecture that visual-language correspondence can be better captured by allowing the feature of one modality to be updated in condition to the other
- The segmentation performance tends to improve with the model scale

Qualitative Results

The subimages in each row include the original image, and segmentation results of seen and unseen (marked as *) categories.

Reference

LSeg: Boyi Li et al. "Language-driven Semantic Segmentation." ICLR 2022

Acknowledgments

This work is supported by the National Key R&D Program of China (No. 2020YFB1406801), National Natural Science Foundation of China (62271308), 111 plan (No. BP0719010), and STCSM (No. 18DZ2270700, No. 22511105700), and State Key Laboratory of UHD Video and Audio Production and Presentation.