Open-Vocabulary Semantic Segmentation via Attribute Decomposition-Aggregation

Chaofan Ma, Yuhuan Yang, Chen Ju, Fei Zhang, Ya Zhang, Yanfeng Wang

SHANGHAI JIAO TONG UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

Closed-Set to Open-Vocabulary
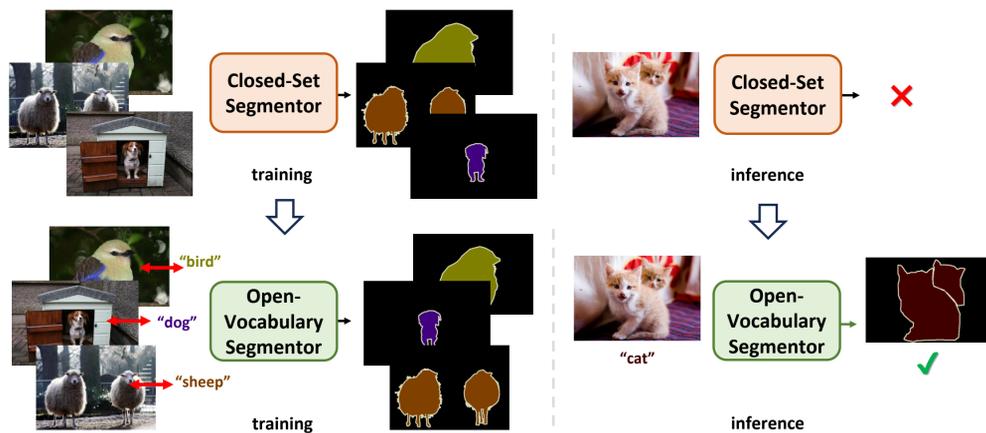
Problems

Unrealistic assumptions in real-world scenarios:

the given new textual categories should *accurate*, *complete*, and *exist* in the pre-trained lexicons.

"cat"?
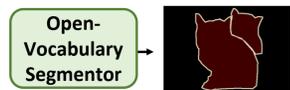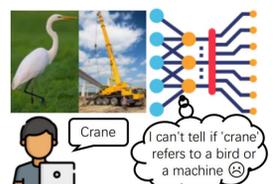
(1) Ambiguity:
brief or incomplete class names bring lexical ambiguity
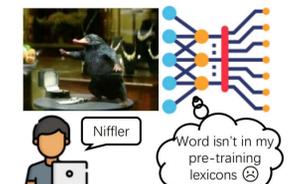
(2) Neologisms:
new words may not in the lexicons during vision-language pre-training
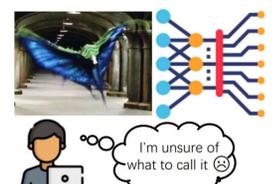
(3) Unnameability:
difficult-to-describe categories create labeling problems for users

Motivation

If we and our small child went to the zoo and asked for the *flamingo*…

Small children may be confused about the word *Flamingo*

But they are familiar with easy and detailed descriptions (aka *attributes*)
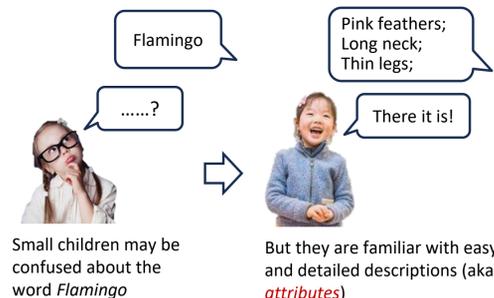
Cognitive psychology
when human understanding new concepts

Attributes:
detailed descriptions from multiple distinct or complementary perspectives

Attributes can …

(1) For ambiguity
make up for missing context information to achieve completeness

(2) For neologisms
transformed into known attributes, easily interpreted by pre-trained language models

(3) For unnameability
used for replace in a more detailed manner

Overview

Decomposition Stage

Bowtruckle
- twig-like
- small eyes
- green body
- sharp claws

Demiguise
- ape-like
- fluffy creature
- large, round eyes
- long, silky hair

Mooncalf
- large eyes
- a round head
- pale-grey skin
- long neck

Aggregation Stage

attribute embeds

aggregation ×L

mask

interactions

visual embeds

Framework

Building Attribute by Large Language Models

What are useful visual features for distinguishing *flamingo* in the scene?

Pink feathers; long, thin legs; large, distinctive wingspan; …

Building Attribute by Manual Collections

data collection → manual annotation

(a) Decomposition Stage

(b) Aggregation Stage

Clustering Module: Mixer, Slot Attn, Cross Attn, Cross Attn

Mask, Clustering Module, Fusion Layer, ×L Stages, Clustering Module, Fusion Layer

visual tokens, attr tokens, learnable cluster tokens

(a) **Decomposition stage**: to build various attribute descriptions (by querying LLM for existing datasets and manually collecting a dataset of rare categories)

(b) **Aggregation stage**: to combine separate attribute pieces into an integrated global description, which then serves as a classifier to differentiate the target object from others

Optional Designs

(a) direct aggregation
(b) pre-aggregation
(c) post-aggregation
(d) hierarchy aggregation
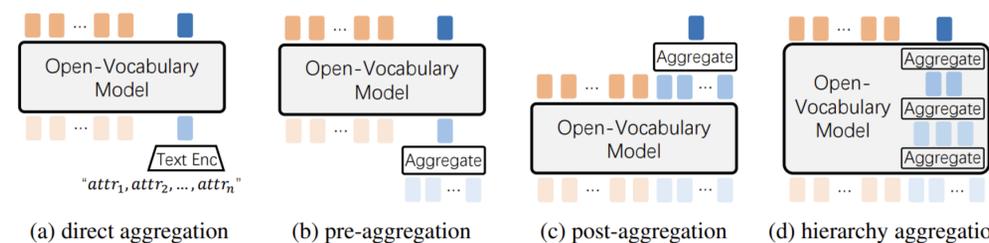
Experimental Results

Table 1: **Evaluation on PASCAL-5^i and COCO-20^i.** For textual inputs, we compare "cls name" (category names) and "attr" (attributes). "direct", "pre-", "post-" and "hrchy" are four aggregation strategies in Sec. 3.4.4.

Model	Settings	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 [46]	cls name	RN101	23.8	17.0	14.1	18.3	18.3	-	-	-	-	-
ZS3Net [7]	cls name	RN101	40.8	39.4	39.3	33.6	38.3	18.8	20.1	24.8	20.5	21.1
LSeg [25]	cls name	RN101	52.8	53.8	44.4	38.5	47.4	22.1	25.1	24.9	21.5	23.4
LSeg [25]	cls name	ViT-L	61.3	63.6	43.1	41.0	52.3	28.1	27.5	30.0	23.2	27.2
LSeg [25]	attr (direct)	RN101	48.6	51.2	39.7	36.2	44.0	20.9	23.6	20.8	18.4	20.9
LSeg [25]	attr (pre-)	RN101	49.1	51.4	40.9	35.9	44.3	21.7	24.5	22.0	19.6	22.0
LSeg [25]	attr (post-)	RN101	50.0	52.2	42.1	36.8	45.3	21.2	24.0	22.5	19.2	21.7
AttrSeg (Ours)	attr (hrchy)	RN101	52.9	55.3	45.0	43.1	49.1	27.6	28.4	26.1	22.7	26.2
AttrSeg (Ours)	attr (hrchy)	ViT-L	61.5	67.5	46.1	50.5	56.4	34.8	32.6	31.6	24.2	30.8

Table 3: **Evaluation on Fantastic Beasts (transferred from PASCAL-5^i and COCO-20^i).** For textual inputs, we compare "cls name" (category names) and "attr" (attributes). "direct", "pre-", "post-" and "hrchy" are four aggregation strategies. 5^i and 20^i refer to the best checkpoints of ith fold of two datasets used for evaluation.

Model	Settings	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
LSeg [25]	cls name	RN101	<10	<10	<10	<10	<10	<10	<10	<10	<10	<10
LSeg [25]	attr (direct)	RN101	44.8	46.6	46.4	46.2	46.0	50.1	51.2	51.7	49.3	50.6
LSeg [25]	attr (pre-)	RN101	46.4	47.7	48.2	47.3	47.4	51.3	52.7	52.9	50.9	52.0
LSeg [25]	attr (post-)	RN101	46.9	48.3	48.8	47.7	47.9	52.8	52.5	52.4	51.2	52.2
AttrSeg (Ours)	attr (hrchy)	RN101	50.7	53.4	53.6	51.3	52.3	56.4	54.9	55.7	55.3	55.6
AttrSeg (Ours)	attr (hrchy)	ViT-L	54.1	55.8	55.4	54.5	55.0	59.2	58.8	58.3	58.5	58.7

Visualizations

(a) **Category:** Erumpent. **Main Attributes:** "large creature", "rhinoceros-like creature", "a thick hide", "a horn on its nose", "huge, grey beast", "grey beast", "resembles a rhinoceros", "roundish body", "single long horn", "thick tail"

(b) **Category:** Doxy. **Main Attributes:** "fairy-like creature", "small creature", "thick hair", "black hair", "four wings", "a pair of sharp pincers", "four arms and four legs", "beetle-like wings", "two rows of teeth"

(c) **Category:** Grindylow. **Main Attributes:** "small creature", "octopus-like creature", "long tentacles with suckers", "blue-green skin", "bulbous eyes", "large eyes", "pale-green", "pointed fangs", "long, brittle fingers"

Feel free to contact: *chaofanma@sjtu.edu.cn*