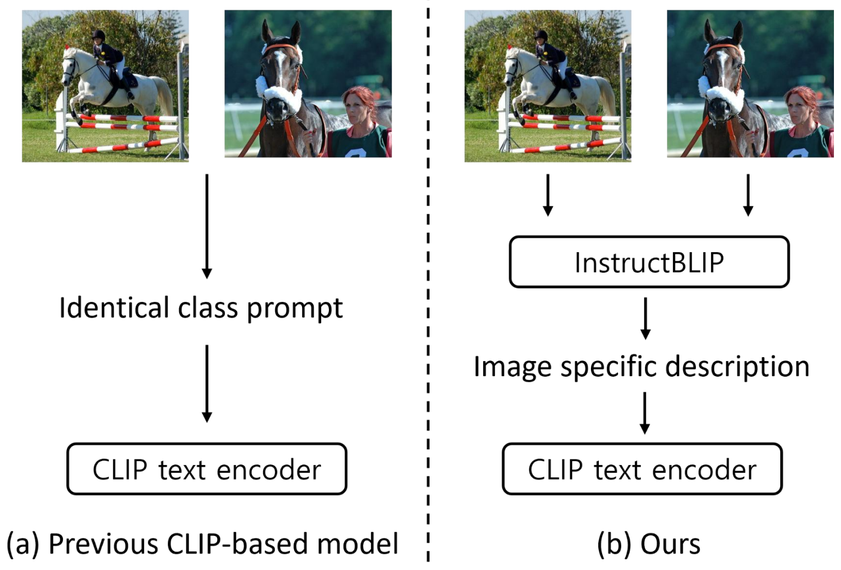
Improving CLIP-based WSSS with Captions

Minseo Kim, and Dong-jin Kim
School of Data Science, Hanyang University

Research Motivations



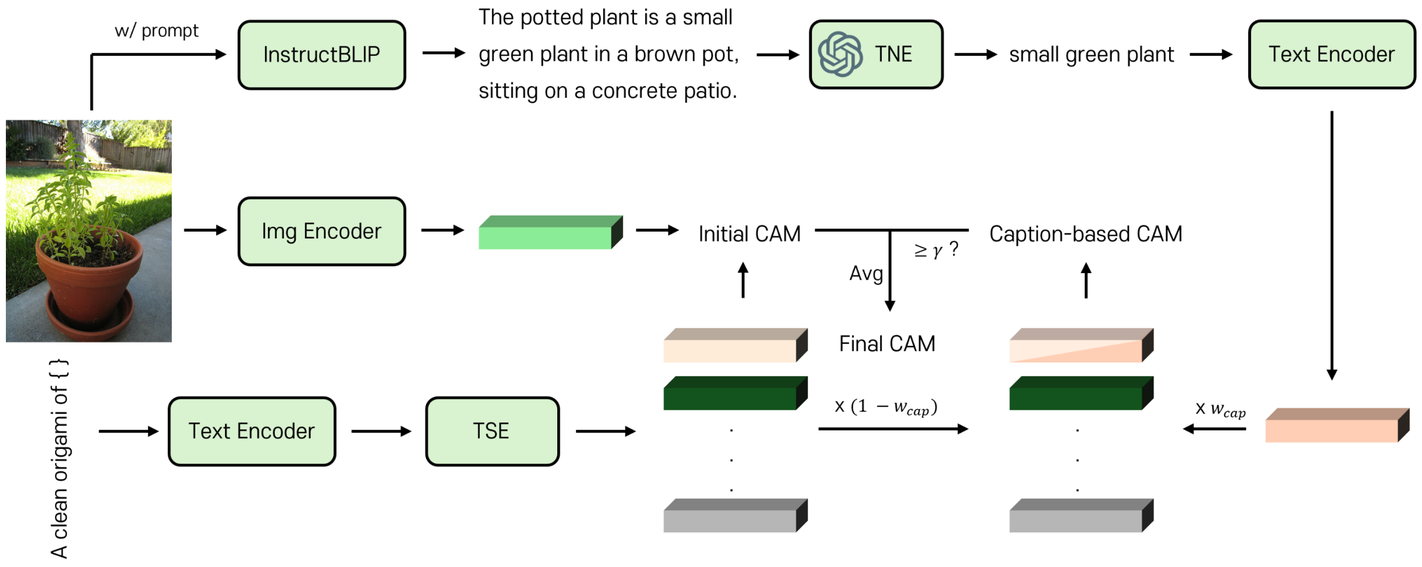
Weakly Supervised Semantic Segmentation (WSSS) is a task aiming to produce a pixel-level mask given class labels of an image to reduce annotation cost.

Several studies leverage CLIP to utilize textual information of class labels and extract mask based on the CLIP embedding.

However, the identical class prompt over different images overlooks intra-class variation

We propose integrating a captioning model to address intra-class variation.

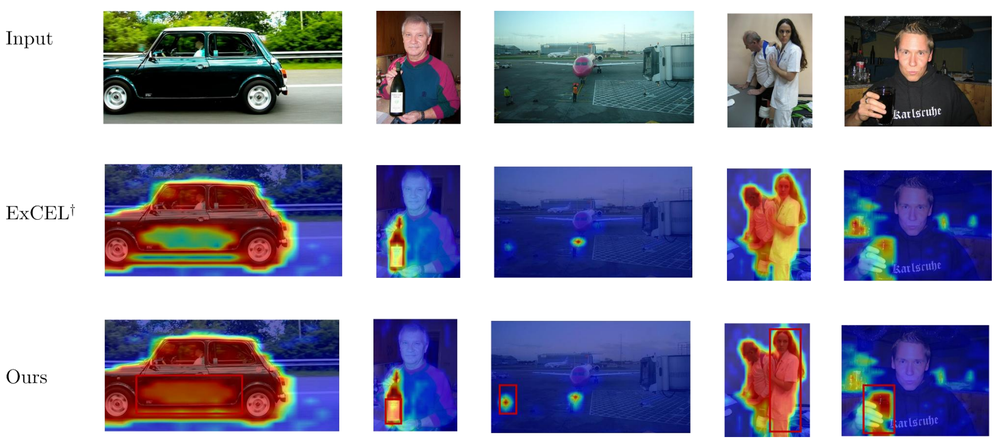
Methodology



- Uses InstructBLIP to get class-specific caption for each class in an image
- Extract a target noun from the original caption to filter superfluous information
- Deploy confidence-based fusion to alleviate under-activation in caption-based CAM

Experiment

Qualitative Results



Quantitative Results & Ablation Study

	CAM	Seg	Seg w/ CRF
ExCEL†	75.6	76.1	77.1
Ours	76.8(+1.2)	77.0(+0.9)	78.1(+1.0)

	CAM	Seg	Seg w/ CRF
Ours	76.8	77.0	78.1
- Confidence based Fusion	76.1(-0.7)	76.2(-0.8)	77.3(-0.8)
- Target Noun Extraction	76.0(-0.8)	76.1(-0.9)	77.3(-0.8)
-TSE	75.4(-1.4)	76.1(-0.9)	77.2(-0.9)

Conclusion & Future Work

- Refining the original class text embedding with captions enables the model to better handle intra-class variation
- Future research should explore a generalizable approach that enhances performance across a wide range of CLIP-based WSSS models