CVPR 2023

Token Contrast for Weakly-Supervised Semantic Segmentation

Lixiang Ru[1] Heliang Zheng[2] Yibing Zhan[2] Bo Du[1]*
[1] School of Computer Science, Wuhan University, China.
[2] JD Explore Academy, China
{rulixiang, dubo}@whu.edu.cn {zhengheliang,zhanyibing}@jd.com

- Problem / objective
  - Over-smoothing issue in ViT for WSSS

- Contribution / Key idea
  - Patch Token Contrast (PTC) module
    - Role: Supervise the final tokens with intermediate knowledge
    - Reason: Intermediate layers of ViT retain semantic diversity
  - Class Token Contrast (CTC) module
    - Role: Contrasts the representation of global foregrounds and local uncertain regions (background)
    - Reason: Class token of ViT capture high-level semantics

RU, Lixiang, et al. Token contrast for weakly-supervised semantic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. p. 3093-3102.

CVPR 2023

- **WSSS w/ image-level labels**
- **ViT for WSSS**
  - 문제: CAM only identifies the most discriminative semantic regions
  - 원인: 그동안 CNN을 통해 CAM을 만들어서. CNN이 local features에 집중하니까
  - 해결: ViT사용. ViT는 self-attention block들을 통해 global feature interactions을 모델링함
- **Over-smoothing issue for using ViT for WSSS**
  - 원인: ViT의 self-attention block들이 LPF 역할을 함. Spatial smoothing 역할. 패치 토큰들을 uniform하게 만듬.
  - Fig 2: 1) 뒤의 레이어로 갈수록 패치 토큰들간 유사도 굉장히 증가: Over-smoothing issue
    2) 초기 레이어들은 여전히 semantic diversity 보존

  -> Motivation to address the over-smoothing issue
  by supervising the final layer tokens with knowledge from intermediate layers.

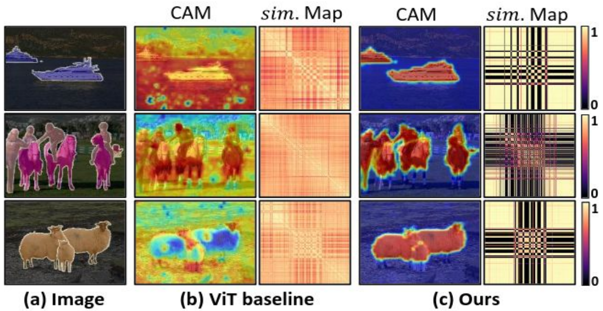

Figure 1. **The generated CAM and the pairwise cosine similarity of patch tokens (*sim.* map).** Our method can address the over-smoothing issue well and produce accurate CAM. Here we use ViT-Base.

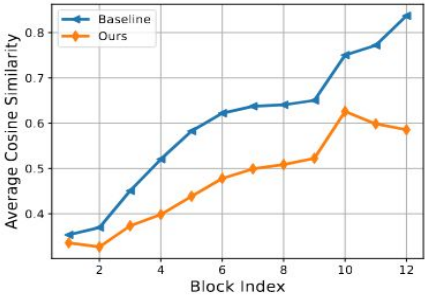

Figure 2. **The average pairwise cosine similarity of patch tokens in each Transformer block.** The cosine similarity is computed on the VOC `train` set. Here we use the ViT-Base (ViT-B) [12] architecture which includes 12 Transformer blocks.

전유진

RU, Lixiang, et al. Token contrast for weakly-supervised semantic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. p. 3093-3102.

CVPR 2023

- **Contribution 1. Patch Token Contrast (PTC) module**
    - 문제: Over-smoothing issue in ViT
    - 사실: Learned representations in intermediate layers still preserve the semantic diversity
    - 해결: Supervise the final tokens with intermediate knowledge
    - 효과: PTC counter the patch uniformity and significantly promote the quality of pseudo labels of WSSS
- **Contribution 2. Class Token Contrast (CTC) module**
    - 목적: Differentiate the uncertain regions in generated CAM
    - 사실: Class token in ViT inherently aggregate high-level semantics
    - 해결: Contrasts the representation of global foregrounds and local uncertain regions (background)
    - 효과: Facilitates the object activation completeness in CAM

전유진