SCLIP: Rethinking Self-Attention for Dense Vision-Language Inference

Feng Wang, Jieru Mei, and Alan Yuille

Johns Hopkins University

- Problem / objective
  - CLIP 사용해서 zero-shot semantic segmentation
    - Self-attention 과정에서 location-misalignment 이슈
- Contribution / Key idea
  - **SCLIP**: Segmentation-adapted CLIP
    - CLIP의 vision encoder의 기존 self-attention 대신 Correlative Self-Attention (CSA) 적용
    - Training-free

WANG, Feng; MEI, Jieru; YUILLE, Alan. Sclip: Rethinking self-attention for dense vision-language inference. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024. p. 315-332.

ECCV 2024

- **Analysis of CLIP's zero-shot semantic segmentation results**

[문제] 이미지 내 객체들 recognize는 잘하지만, localization은 못함

Fig. 1: Open-vocabulary semantic segmentation examples. We evaluate on two images from COCO [5] (the 3rd and the 5th examples) and three high-resolution images in the wild, where our SCLIP consistently generates high quality segmentation masks yet the original CLIP fails to correctly localize objects. We display the corresponding text query of each segmentation mask, where *"g. retriever"* and *"b. collie"* in the first example denote golden retriever and border collie, respectively.

전유진

WANG, Feng; MEI, Jieru; YUILLE, Alan. Sclip: Rethinking self-attention for dense vision-language inference. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024. p. 315-332.
ECCV 2024

- **Analysis of CLIP's zero-shot semantic segmentation results**

[원인] CLIP이 spatial-invariant한 visual features를 학습함
[근거] CLIP의 attention map이 주요 객체 형상은 어느정도 알지만, 다양한 소스 포인트들에 대하여 attention map이 비슷하게 나옴

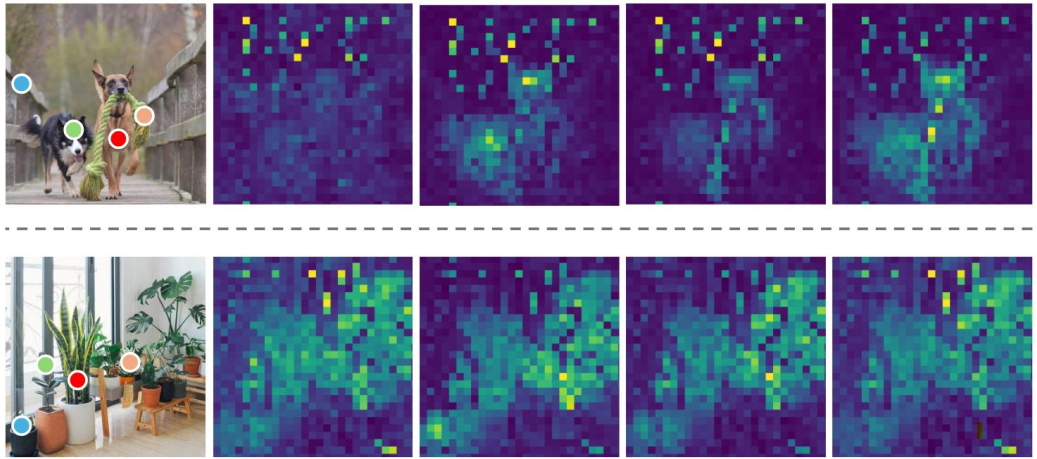


Fig. 2: Final layer attention maps of vanilla CLIP with a ViT-Base/16 image encoder. We display the attention maps of four points (marked in different colors) for each example. It shows that each local visual token attends to a wide range of positions and the attention maps often share similar patterns, indicating that CLIP learns spatial-invariant visual features.

WANG, Feng; MEI, Jieru; YUILLE, Alan. Sclip: Rethinking self-attention for dense vision-language inference. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024. p. 315-332.

- **Analysis of CLIP's zero-shot semantic segmentation results**

[결론] CLIP이 semantic segmentation을 잘하려면, spatial-covariant visual features가 필요하다.
[해결] Covariant visual features를 이용하는 새로운 Correlative Self-Attention (CSA) 메커니즘 제안

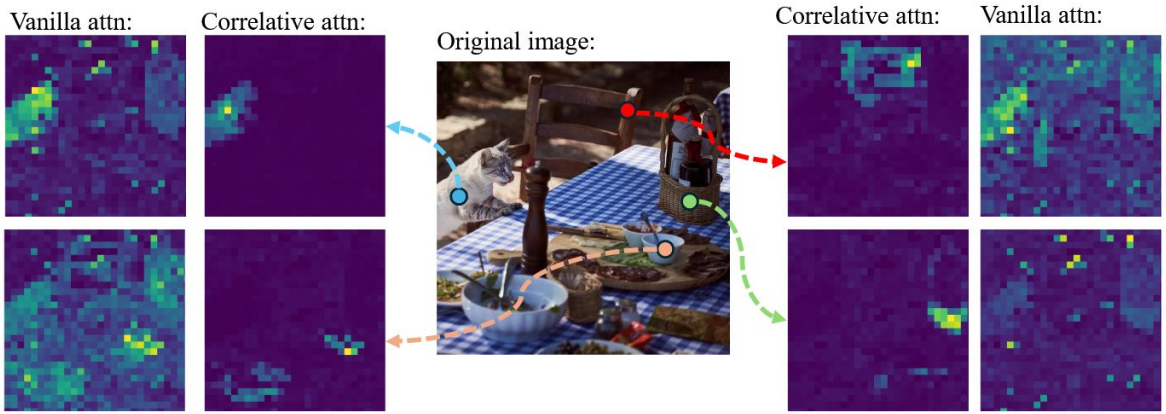


Fig. 4: Comparison of attention maps. We show the attention maps of the last transformer layer in CLIP vision encoder equipped with the original self-attention (right) and our correlative self-attention (left). Our correlative self-attention exhibits spatially covariant patterns as the attention maps are distinct to different source points and show clear boundaries of semantic objects (*e.g.*, the chair and the cat).

WANG, Feng; MEI, Jieru; YUILLE, Alan. Sclip: Rethinking self-attention for dense vision-language inference. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024. p. 315-332.
ECCV 2024

- **Method**

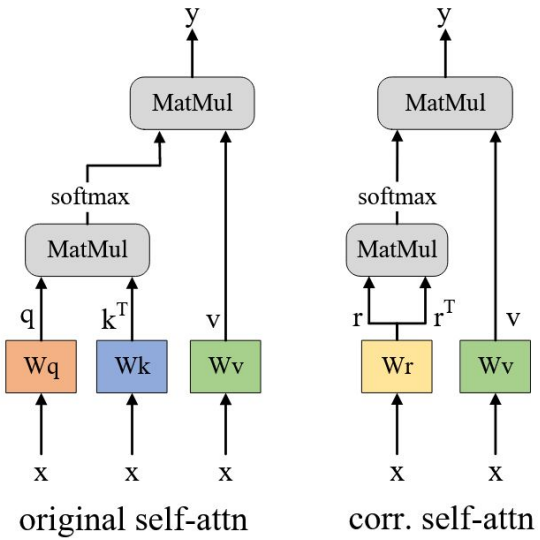


Fig. 3: An architectural comparison between the original self-attention and our correlative self-attention mechanism. Our method determines attention scores by pairwise correlations between the local tokens.

WANG, Feng; MEI, Jieru; YUILLE, Alan. Sclip: Rethinking self-attention for dense vision-language inference. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024. p. 315-332.
ECCV 2024

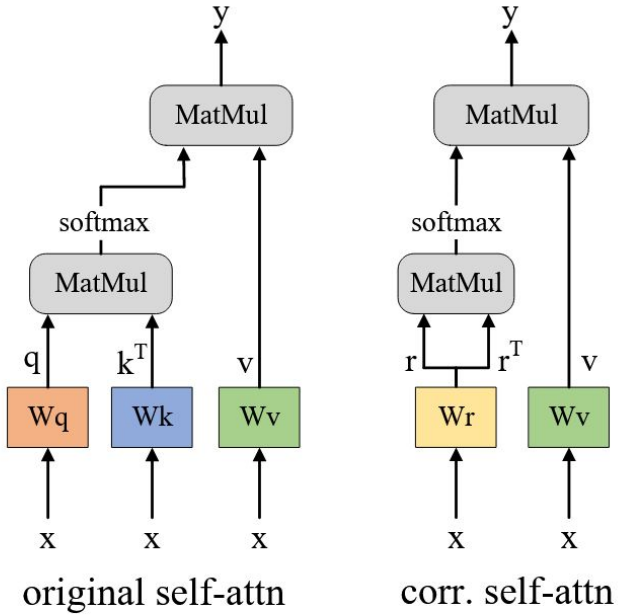
- **Correlative Self-Attention**

[방법] 각 로컬 토큰 간에 correlation 점수를 attention 점수로 사용
[효과] Self-attention이 각 토큰 간의 거리를 더 반영하게 만듦

$$\boldsymbol{Attn} = \text{Softmax}\left(\boldsymbol{X}\boldsymbol{W}_q\boldsymbol{W}_k^T\boldsymbol{X}^T/\sqrt{d}\right), \quad (1)$$

$$\boldsymbol{Attn} = \text{Softmax}\left(\boldsymbol{X}\boldsymbol{W}_r\boldsymbol{W}_r^T\boldsymbol{X}^T/\tau\right), \quad (2)$$

$$\boldsymbol{Attn} = \text{Softmax}\left(\boldsymbol{X}\boldsymbol{W}_q\boldsymbol{W}_q^T\boldsymbol{X}^T/\tau\right) + \text{Softmax}\left(\boldsymbol{X}\boldsymbol{W}_k\boldsymbol{W}_k^T\boldsymbol{X}^T/\tau\right), \quad (3)$$

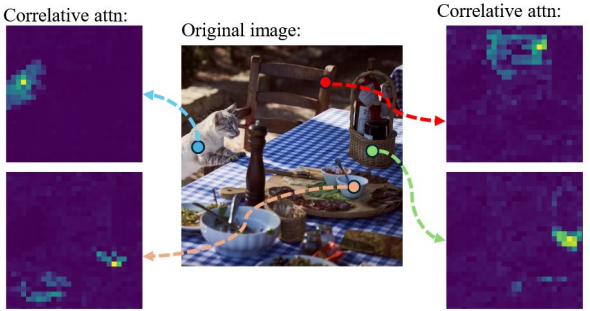


WANG, Feng; MEI, Jieru; YUILLE, Alan. Sclip: Rethinking self-attention for dense vision-language inference. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024. p. 315-332.
ECCV 2024

- **Correlative Self-Attention이 dense prediction에 더 적합한 이유**

1. Feature localization 우수
- Attention map의 대각 성분의 크기는 각 토큰이 자기 자신의 위치에 집중하는 정도를 나타냄.
- CSA에서 projection된 벡터 간에 correlation은 i=j일때 최대가 되어, 자연스럽게 대각 성분이 강화되어 위치 정보 보존 우수
- cf) MaskCLIP[1]: 인위적으로 attention map의 대각성분만 1로 남기고, 나머지는 0으로 채움.(I.E., 항등행렬)

2. Semantic correlation 반영
- 자기 자신 뿐만 아니라 의미적으로 유사한 feature에도 높은 attention을 부여
- Figure4: CSA의 attention map은 같은 객체 내 의미적으로 관련된 위치들만 집중
- 그래서, segmentation 시 객체 단위로 부드럽고 일관된 마스크 생성 가능



3. Zero-shot 가능
- CSA에서 사용되는 projection matrix는 feature간 거리 측정 용도여서 파라미터에 민감하지 않음.
- Projection matrix를 학습 없이 고정하거나 랜덤으로 설정해도 성능 유지
- 따라서, CLIP을 fine-tuning 없이 zero-shot으로 dense prediction task에 직접 활용 가능

WANG, Feng; MEI, Jieru; YUILLE, Alan. Sclip: Rethinking self-attention for dense vision-language inference. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024. p. 315-332.

- **Experiments**

Table 1: Evaluation results (mIoU, %) of our method and the baseline models on eight semantic segmentation benchmarks. The methods with an asterisk symbol * denote using a PAMR [1] post-processing strategy which introduces heavy computation cost so we de-emphasize these results. Our results are marked in gray. The best results on each dataset are **bolded**.

Method	*With a background category*			*Without background category*					Avg.
	VOC21	Context60	Object	VOC20	City.	Ctx59	ADE20k	Stuff.	
CLIP [44]	18.8	9.9	8.1	49.4	6.5	11.1	3.1	5.7	14.1
MaskCLIP [71]	43.4	23.2	20.6	74.9	24.9	26.4	11.9	16.7	30.3
GroupViT [61]	52.3	18.7	27.5	79.7	18.5	23.4	10.4	15.3	30.7
ReCo [49]	25.1	19.9	15.7	57.7	21.6	22.3	11.2	14.8	23.5
TCL [8]	51.2	24.3	30.4	77.5	23.5	30.3	14.9	19.6	33.9
CLIP-Surg [33]	-	-	-	-	31.4	29.3	-	21.9	-
OVSeg. [63]	53.8	20.4	25.1	-	-	-	5.6	-	-
SegCLIP [39]	52.6	24.7	26.5	-	-	-	-	-	-
SCLIP (ours)	**59.1**	**30.4**	**30.5**	**80.4**	**32.2**	**34.2**	**16.1**	**22.4**	**38.2**
Approaches with pamr post-processing:									
CLIP*	19.8	8.7	10.4	54.2	7.0	11.7	3.6	5.9	15.2
MaskCLIP*	52.0	28.2	22.6	72.1	30.1	31.5	14.0	20.0	33.8
GroupViT*	52.7	19.5	27.9	81.5	21.7	24.4	11.8	16.9	32.1
ReCo*	27.2	21.9	17.3	62.4	23.2	24.7	12.4	16.3	25.7
TCL*	55.0	30.4	31.6	83.2	24.3	33.9	17.1	22.4	37.2
SCLIP* (ours)	**61.7**	**31.5**	**32.1**	**83.5**	**34.1**	**36.1**	**17.8**	**23.9**	**40.1**

WANG, Feng; MEI, Jieru; YUILLE, Alan. Sclip: Rethinking self-attention for dense vision-language inference. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024. p. 315-332.

- **Experiments**

Table 2: Ablation results (mIoU, %) of projection matrices in correlative self-attention. n denotes the number of random projection matrices used in this experiment. Our default setting is marked in gray. The best result on each dataset is **bolded**.

Mode	PASCAL VOC	PASCAL Context	COCO-Stuff
Single projection matrix for CSA:			
Identity projection	57.5	33.0	21.5
W_q projection	58.2	33.5	21.7
W_k projection	58.4	33.1	21.8
Learned projection	**60.4**	**34.7**	**22.6**
Random projection matrices (average of 5 trials):			
$n = 1$	57.1	32.4	20.6
$n = 4$	58.0	32.7	20.9
$n = 16$	58.1	32.7	21.2
Default	59.1	34.2	22.4

WANG, Feng; MEI, Jieru; YUILLE, Alan. Sclip: Rethinking self-attention for dense vision-language inference. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024. p. 315-332.

- **Experiments**

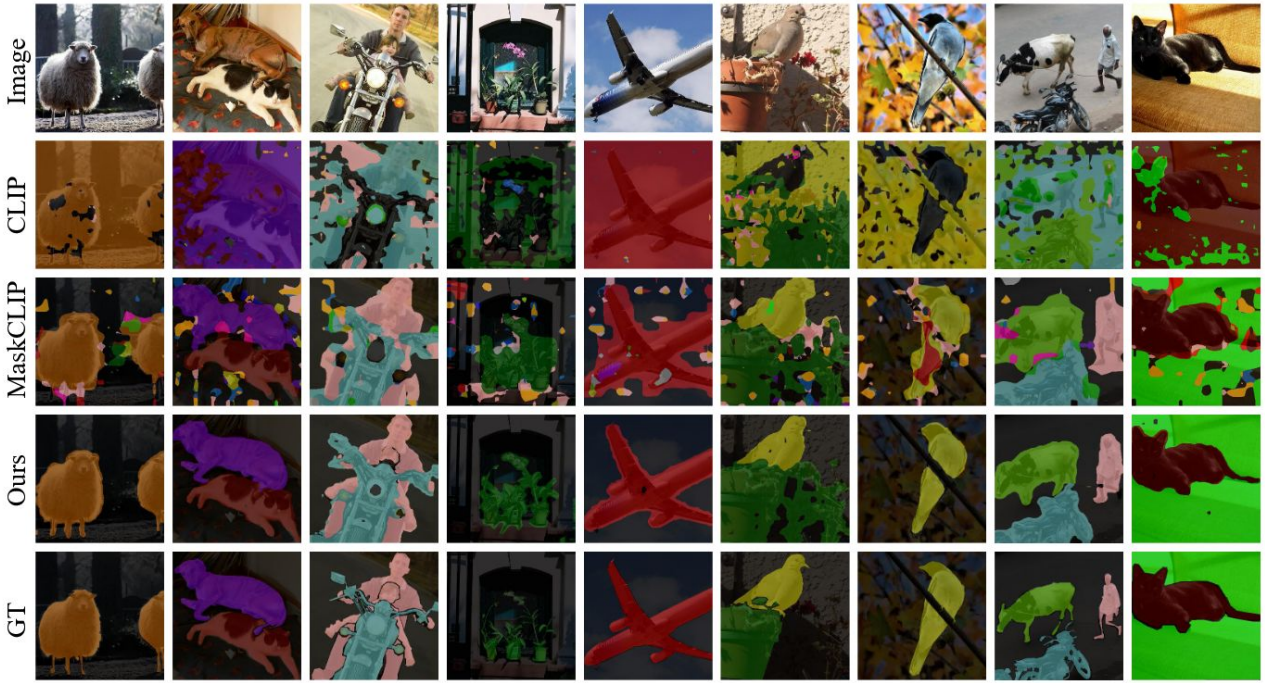


Fig. 5: Additional visualization results on PASCAL VOC. "GT" denotes ground truth.

WANG, Feng; MEI, Jieru; YUILLE, Alan. Sclip: Rethinking self-attention for dense vision-language inference. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024. p. 315-332.
ECCV 2024

- **Experiments**

Fig. 6: Additional visualization results on COCO-Object. "GT" denotes ground truth.