Fine-Tuning없이 Multimodal Few-shot Visual Grounding에 효과적인 모델 아키텍처 연구

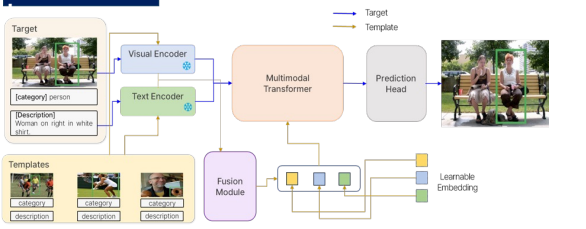
박은주(이화여자대학교) | 황영서(숙명여자대학교) | 김준섭(성균관대학교) | 양희재(성균관대학교)

ABSTRACT

This research presents a novel approach to Visual Grounding, addressing the challenge of handling new classes with minimal data. Traditional models often rely on extensive fine-tuning for new tasks, which is time-consuming and inefficient. To overcome this, we propose a few-shot learning architecture that eliminates the need for fine-tuning by incorporating a template-based multimodal prompt with learnable embeddings. Additionally, our model integrates a fusion module and contrastive loss to enhance generalization across unseen classes. Our approach achieves 83.6% accuracy on the RefCOCOg dataset, demonstrating significant improvements in performance on novel classes.

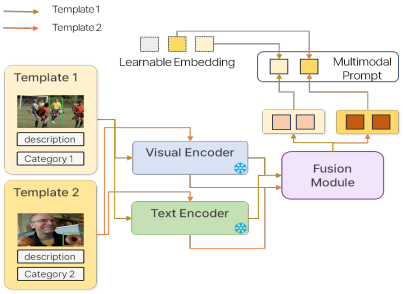
METHOD

Framewor



본 연구는 Few-shot Visual Grounding 문제를 해결하기 위해 template 기반 Multimodal prompt 를 제안

- 상호 Cross-attention을 통해 template 간 상호작용 강화
- Contrastive Loss를 통해 클래스 간 차별성을 학습



Multimodal prompt

- **Templates Learnable Embedding**

템플릿 별로 이미지와 텍스트의 상호작용을 바탕으로 학습 가능한 임베딩을 도입하여 특정 클래스에 대한 고유한 표현을 학습

템플릿은 동일 클래스뿐만 아니라 다른 클래스에서 선택해 분별력 ↑

- **Templates Fusion**

Image-text, text-image 간의 상호 cross attention을 통해 특징 정교화

- **Contrastive learning**

서로 다른 클래스 간의 차이를 극대화하기 위해 contrastive loss를 사용

$$L(z_i, z_j) = -log \frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} exp(sim(z_i, z_j)/\tau)}$$

RESULT

Comparison Accuracy

Methods	Backbone	Support set	Accuracy
TransVG	ResNet-101		67.02
TransVG	ResNet-50		66.56
TransVG++	ResNet-50		73.86
GroundVLP	Vin-VL		74.73
Dynamic MDETR	ResNet-50		69.43
Dynamic MDETR + FS-learnable embedding(ours)	ResNet-50	✓	83.6

- 템플릿 기반 멀티모달 프롬프트와 학습 가능한 임베딩이 모델 성능에 미치는 영향 분석
- 제안된 모델은 83.6%의 정확도를 기록 하여 기존 모델 대비 17% 이상의 성능 향상
- 이는 템플릿 활용이 모델의 세밀한 객체 구별 능력을 크게 향상시켰음을 입증

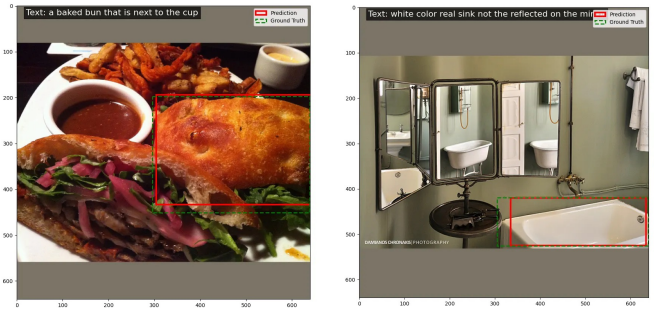
Novel Data Performance

Methods	Backbone	Accuracy	AP
Ours	ResNet-50	0.30	0.53
Ours + Fu*	ResNet-50	0.39 (+0.09)	0.58 (+0.05)
Ours + CI**	ResNet-50	0.38 (+0.08)	0.60 (+0.07)
Ours + Fu* + CI**	ResNet-50	0. 39 (+0.09)	0.60 (+0.07)

*Fu: Fusion Module **CI: Contrastive Loss

- Few-shot Visual Grounding에서 새로운 클래스를 다루는 모델의 성능을 평가
- Fusion Module과 Contrastive Loss 결합 시 정확도 0.39, AP 0.60으로 가장 높은 성능 기록
- 이는 템플릿 기반 학습과 Fu, CI 두가지 모듈이 새로운 클래스에서도 효과적임을 입증

Visualization



- 예측된 바운딩 박스 결과를 통해 제안된 모델이 텍스트 설명에 기반한 객체를 매우 정확히 식별할 수 있음을 확인
- 왼쪽 그림에서는 "a baked bun that is next to the cup"라는 텍스트에 대해 Prediction(빨간색)과 Ground Truth(초록색) 간의 차이가 거의 없는 정확한 바운딩 박스를 생성
- 오른쪽 그림은 "white color real sink not the reflected on the mirror"라는 복잡한 텍스트 설명을 기반으로, 실제 싱크대와 거울에 반사된 싱크대를 정확히 구분한 바운딩 박스를 보여줌
- 이러한 결과는 제안된 모델이 텍스트와 시각적 정보를 효과적으로 통합하여 텍스트 기반 객체 식별 및 위치 예측에서 높은 성능을 입증했음을 시사

CONCLUSION

- 본 연구는 Fine-Tuning 없이 Few-shot Visual Grounding을 가능하게 하는 멀티모달 아키텍처를 제안
- 템플릿 기반 멀티모달 프롬프트, Fusion Module, Contrastive Learning을 결합하여 RefCOCOg와 novel 데이터셋에서 높은 성능을 입증
- 제안된 모델은 템플릿 학습을 통해 시각적 세부 사항을 정교하게 파악하며, 새로운 클래스에 대한 일반화 성능을 크게 향상
- 제안된 접근법은 모델이 텍스트와 이미지 간의 복잡한 관계를 효과적으로 처리할 수 있도록 설계되어, 정밀한 추론과 응답을 가능하게 함
- 그러나 다양한 데이터셋에서의 범용성 검증과 템플릿 구조의 최적화 연구는 진행되지 않음
- 향후 연구에서는 템플릿 수와 구조를 다양화하고, 복잡한 시나리오에서도 높은 성능을 유지할 수 있도록 아키텍처를 확장할 예정

REFERENCE

[1] Shi, Fengyuan, et al. "Dynamic mdetr: A dynamic multimodal transformer decoder for visual grounding," IEEE Transactions on Pattern Analysis and Machine Intelligence., pp.1181-1198, Feb 2024.

[2] Bulat, Adrian, Ricardo Guerrero, Brais Martínez and Georgios Tzimiropoulos. "FS-DETR: Few-Shot DEtection TRansformer with prompting and without re-training," IEEE/CVF International Conference on Computer Vision (ICCV), pp.11759-11768, Oct 2023.

[3] Deng, Jiajun, et al. "Transvg: End-to-end visual grounding with transformers." Proceedings of the IEEE/CVF International Conference on Computer Vision. pp.1769-1779, 2021.