Textual Attention RPN for Open-Vocabulary Object Detection

Tae-Min Choi[1,2] Inug Yoon[2] Jong-Hwan Kim[2] Juyoun Park[1]

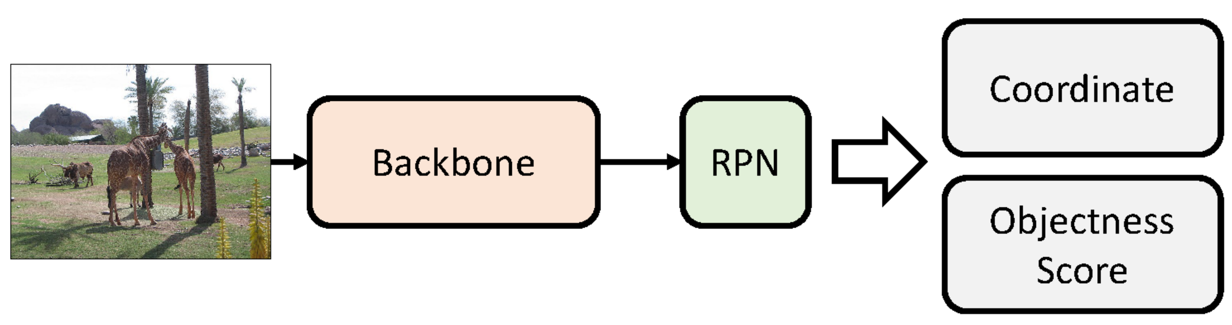
[1] Korea Institute of Science and Technology (KIST), Seoul, Republic of Korea

[2] Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea

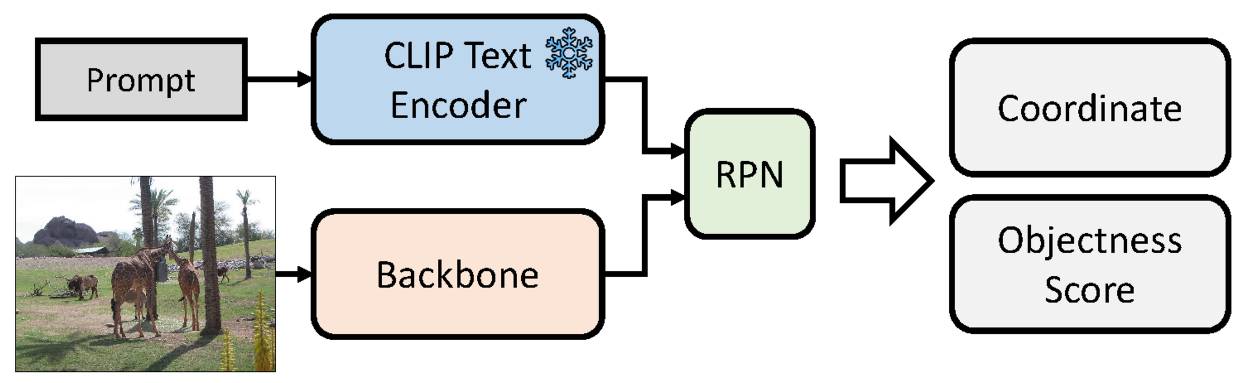
TL; DR: Make powerful Region Proposal Network (RPN) with textual and visual feature fusion for open-vocabulary object detection

Introduction

- **Open-Vocabulary Object Detection (OVD)** is a method that allows a model to identify objects it hasn't specifically seen or trained on before, using **external textual information**.
- However, **conventional RPNs focus solely on visual features**, which limits their effectiveness for novel objects in OVD..
- Our proposed **Textual Attention RPN (TA-RPN)** integrates both visual and textual features from CLIP's text encoder, employing prompt learning to enhance localization.
- TA-RPN **improves proposal generation performance** through pixel-wise attention and prompt learning.



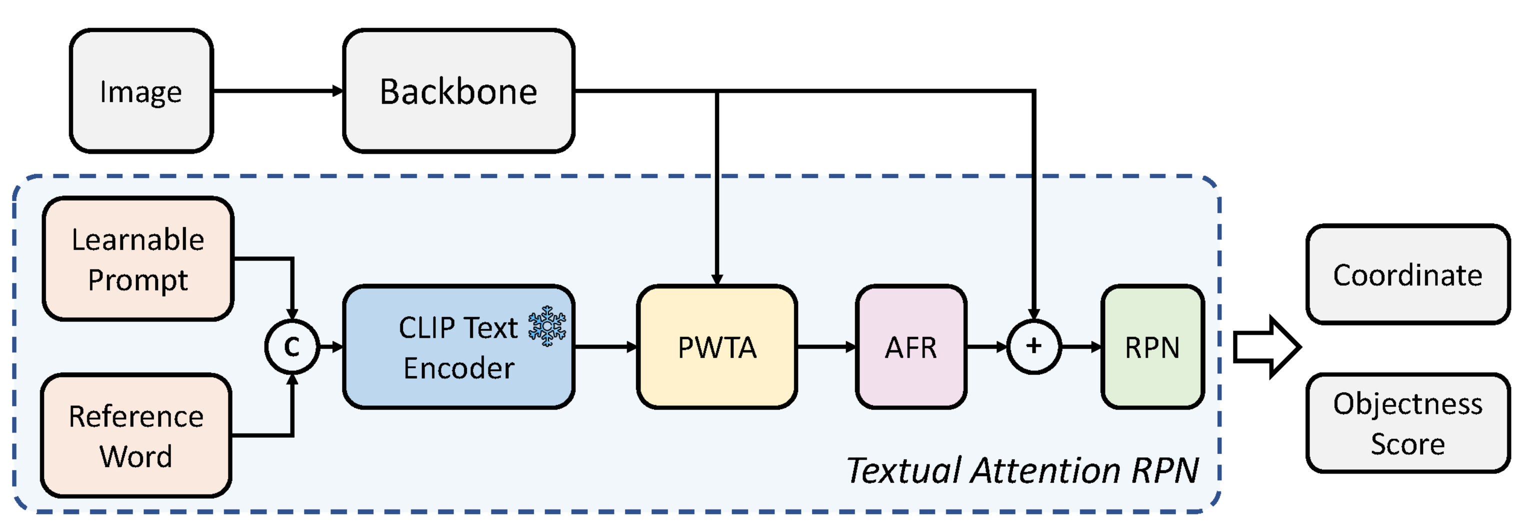
(a) Conventional RPN



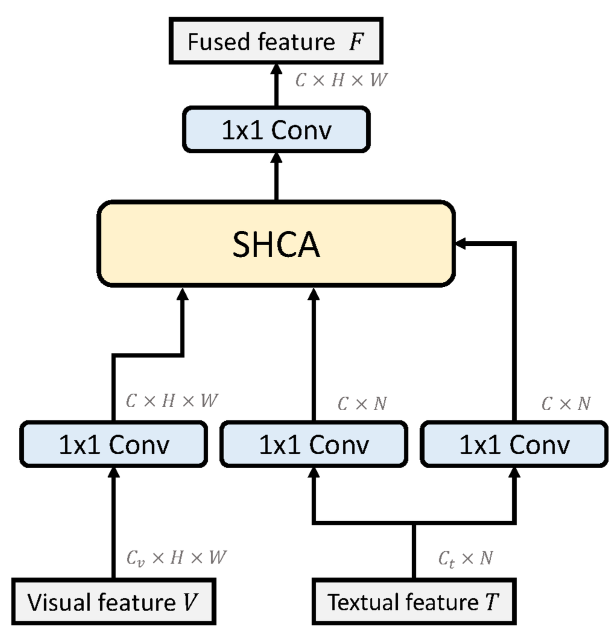
(b) Textual Attention RPN (ours)

Method

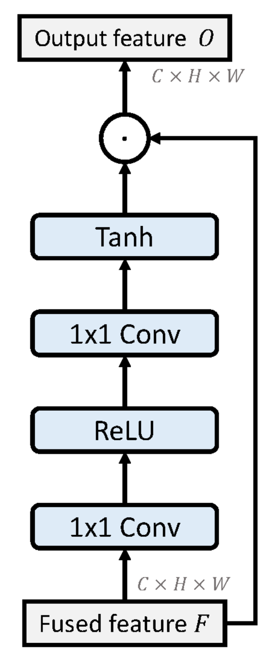
- TA-RPN integrates information from CLIP's text encoder using **Pixel-Wise Textual Attention (PWTA), Adaptive Feature Refinement (AFR), and prompt learning.**
- **Pixel-Wise Textual Attention (PWTA): Fuses visual and textual features at each pixel** to capture detailed, context-aware information across the entire image.
- **Adaptive Feature Refinement (AFR):** Uses an **attention map** to enhance visual relevance **while preventing textual information from overwhelming the localization process.**
- **Prompt Learning:** Employs dynamic prompts to tailor textual features for effective localization of novel objects.



(a) Textual Attention RPN architecture



(b) PWTA design



(c) AFR design

Experiments

- **TA-RPN is built on the BARON [1] framework** with Faster R-CNN [2] and ResNet50-C4/FPN initialized with SoCo pre-trained weights.
- CLIP (ViT-B/32) is used to generate text embeddings, which are fused with visual features to enhance proposal generation.
- On **COCO**, TA-RPN achieves a **36.1 AP** for novel classes, surpassing existing methods like BARON by **2.1 AP**.
- On **LVIS**, TA-RPN improves the **average precision for rare classes (AP_r) to 21.5.**
- Each module (PWTA, AFR, and prompt learning) was evaluated independently, with all components together resulting in a **4.0 AP increase for novel classes**.
- **The highest average recall (ARN@100) for novel classes** was achieved when all modules were active.

#	PWTA	AFR	Prompt learning	AP^N_{50}	AP^B_{50}	$AR^N@100$
1	✗	✗	✗	32.1	50.7	35.5
2	✓	✗	✗	34.3 (+2.2)	53.5	35.9 (+0.4)
3	✓	✓	✗	35.2 (+3.1)	53.1	36.1 (+0.6)
4	✓	✗	✓	34.7 (+2.6)	48.8	35.3 (-0.2)
5	✓	✓	✓	**36.1** (+4.0)	53.1	**36.4** (+0.9)

Ablation study

Method	Backbone	Extra Data	AP^N_{50}	AP^B_{50}	AP_{50}
OVR-CNN	ResNet50-C4	✓	22.8	46.0	39.9
ViLD	ResNet50-FPN	✗	27.6	59.5	51.3
RegionCLIP	ResNet50-C4	✓	26.8	54.8	50.4
Detic	ResNet50-C4	✓	27.8	47.1	45.0
VLDet	ResNet50-C4	✓	32.0	50.6	45.8
F-VLM	ResNet50-FPN	✗	28.0	-	39.6
OADP	ResNet50-FPN	✗	30.0	53.3	47.2
CoDet	ResNet50-C4	✗	30.6	52.3	46.6
ProxyDet	ResNet50-C4	✗	30.4	52.6	46.8
BARON	ResNet50-FPN	✗	34.0	60.4	53.5
Ours	ResNet50-C4	✗	**36.1**	53.1	43.3

Performance comparison on the COCO dataset

Method	Backbone	Object Detection				Instance Segmentation			
		AP_r	AP_c	AP_f	AP	AP_r	AP_c	AP_f	AP
ViLD	ResNet50	16.1	20.0	28.3	22.5	16.3	21.2	31.6	24.4
RegionCLIP	ResNet50	17.1	27.4	34.0	28.2	-	-	-	-
Detic	ResNet50	-	-	-	-	17.8	26.3	31.6	26.8
DetPro	ResNet50	20.8	27.8	32.4	28.4	19.8	25.6	28.9	25.9
F-VLM	ResNet50	-	-	-	-	18.6	24.0	26.9	24.2
ProxyDet	ResNet50	-	-	-	-	18.9	-	-	30.1
BARON*	ResNet50	20.4	30.9	33.4	30.1	19.8	28.6	30.2	27.7
Ours	ResNet50	**21.5**	30.4	33.7	30.2	**20.7**	28.6	30.3	27.9

Performance comparison on the LVIS dataset

[1] Wu, Size, et al. "Aligning bag of regions for open-vocabulary object detection." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023.
[2] Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." IEEE transactions on pattern analysis and machine intelligence 39.6 (2016): 1137-1149.