Research on Improving the Performance of Open-Vocabulary Object Detection

Report Advisor: Bumsub Ham

December 2024

Jung Hyun Park, Hyukjin Kim
School of Electrical and Electronic Engineering
College of Engineering
Yonsei University

Contents

ABSTRACT

Recent advances in pre-trained vision-language models (VLMs) have spurred efforts to enable object detection systems to generalize beyond training categories, effectively identifying novel objects in diverse contexts. These models have demonstrated remarkable potential in bridging visual and semantic representations. Traditional knowledge-distillation based open-vocabulary object detection methods primarily distill knowledge by aligning region embeddings with features extracted from pre-trained VLMs. BARON extended this approach by introducing the concept of bag of region and aligns bag-of-region embeddings with corresponding features extracted from frozen VLMs, exploiting the compositional structure of semantic concepts within scenes. Building upon the Mask R-CNN framework, this evolution utilized region proposals to align both individual and grouped regions, thereby capturing the rich visual-semantic relationships embedded in large-scale datasets.

Building on the BARON framework, we extend this approach to further refine the representation of bag-of-region embeddings. By incorporating label smoothing and similarity-preserving knowledge distillation, our method enhances the alignment and fully leverages the compositional structure of semantic concepts. Integrated into the BARON architecture, the proposed approach achieves a significant improvement of 1.2 box AP_{50} on the open-vocabulary COCO benchmark. This performance gain demonstrates that even with minimal architectural modifications, our method effectively enhances the alignment process, validating our method's ability to more effectively utilize the visual-semantic knowledge inherent in VLMs.

Key words : Open-Vocabulary Object Detection, Vison-Language Models (VLMs), BARON, Knowledge Distillation, Label Smoothing, Similarity-Preserving Knowledge Distillation, Object Detection, Out-of-Distribution, Region Proposal

1. Introduction

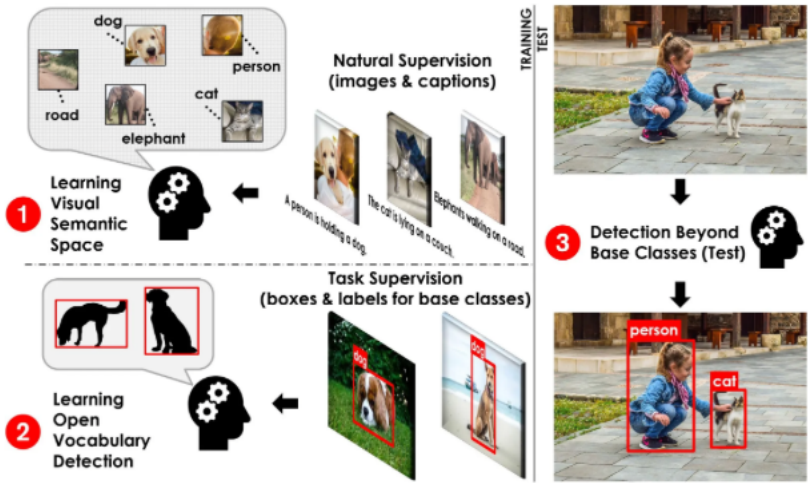

Figure 1 | Open Vocabulary Object Detection

Object detection is one of the most fundamental and essential tasks in the field of computer vision, serving as a cornerstone for scene understanding. It is widely applied across various real-world domains, including autonomous vehicles, security and surveillance systems, medical imaging analysis, sports science, and logistics management. The field has continuously evolved over its long history, driven by the rapid advancements in deep learning technologies. However, traditional object detection models face a critical limitation: they rely on the availability of extensive bounding box annotations for training data, which is both time-consuming and costly to obtain.

To overcome these constraints, methodologies such as zero-shot object detection[1][2][3] and weakly-supervised object detection[4][5] have been actively explored. These approaches aim to achieve robust performance in out-of-distribution settings without requiring full supervision. Among these, Open-Vocabulary Object Detection (OVOD) aims to enable models to effectively detect objects belonging to novel categories, going beyond the base

categories, by utilizing only bounding box annotations of base categories along with weak supervision (e.g., visual grounding data, image captions, and image labels). Notably, following the advancements in large-scale vision-language models (VLMs) such as CLIP[6], there has been an ongoing effort to efficiently utilize these pre-trained models for OVOD.

Currently, open-vocabulary object detection aims to achieve high object detection performance by utilizing the image-text embeddings from pre-trained vision-language models (VLMs). The methods actively researched in open-vocabulary object detection can be broadly categorized into four approaches: region-aware training, pseudo-labeling, knowledge distillation, and transfer learning. **1) Region-aware training**[7][8][9] attempts to implicitly align image-text pairs without using the image encoder from the VLM. The goal is to align regions within the image directly with the associated text descriptions. **2) Pseudo-labeling** [10][11][12] works by explicitly constructing pseudo-labels for novel classes using image-text pairs. This helps to mitigate the degradation in learning caused by insufficient annotations. Common techniques include generating pseudo region-word or region-caption pairs for novel classes and leveraging these for model training. **3) Knowledge distillation** [13][14][15] transfers knowledge from pre-trained VLMs, like CLIP[6], by learning region embeddings from region proposals. Since region proposals generated by Region Proposal Networks (RPNs) in models like Faster R-CNN[16] may contain objects from novel classes, it is crucial to correctly classify them into their respective novel classes. In knowledge distillation approaches, both the detector backbone and the image encoder from the VLM are used during training. However, the image encoder is not used during inference, which helps avoid increased computational overhead, though the training process remains computationally intensive. **4) Transfer learning**[17] addresses this overhead by using only the VLM's image encoder in both the training and testing phases, thereby reducing computational costs while still maintaining effective performance. These four approaches represent the core strategies being explored to improve open-vocabulary object detection, each with its own balance of computational efficiency and performance in recognizing novel object categories.

In this study, we aim to enhance the performance of BARON[14], a representative model of the knowledge distillation approach. BARON was designed to overcome the limitations of traditional knowledge distillation based object detection models that align region proposals independently. By grouping region proposals predicted by the Region Proposal Network (RPN) into “bags of regions”, BARON not only distills knowledge for individual region proposals but also learns embeddings for these bags. To achieve this, BARON introduced a method that samples neighborhood regions with a certain probability to form multiple sampled bags of regions centered on each region proposals. These sampled bag embeddings are then aligned with pre-trained image features through contrastive learning. This approach significantly improves the detection performance for novel categories by better understanding the relationships among visual concepts.

However, BARON encounters a limitation in the alignment process of bag-of-region embeddings, where the knowledge distillation fails to adequately capture the tendencies inherent in the teacher embeddings. Specifically, BARON mistakenly treats different sampled bags of regions centered on the same region proposal as negative pairs, learning to indiscriminately push them apart in the student embedding space. However, an analysis of the self-similarity among the pre-trained CLIP image features of bag-of-region embeddings reveals a distinct pattern: bags of regions centered on the same region proposal exhibit significantly higher similarity to each other compared to those centered on different region proposals.

Knowledge distillation assumes that the teacher model possesses the ability to excel at the target task and aims to transfer that knowledge effectively to the student model. In this context, the goal is to align the space of pre-trained CLIP embeddings (teacher model) with the space of region embeddings (student model). If the teacher model exhibits distinctly high similarity scores between certain bag-of-region embeddings, the student model should strive to replicate this alignment. To address this, we propose an auxiliary loss term that reduces the Euclidean distance between the self-similarity matrices of bag-of-region embeddings obtained

from the teacher and student models during training. Experimentally, our approach demonstrates an improvement of 1.2 AP_{50} on the novel categories compared to BARON under the same conditions on the COCO dataset for conventional open-vocabulary object detection settings.

2. Related Works

2.1. Faster R-CNN

Faster R-CNN[16] represents a groundbreaking advancement in the field of object detection, introducing the Region Proposal Network (RPN) to efficiently generate object proposals from feature maps. Unlike previous methods that required a separate region proposal step, Faster R-CNN integrates this process within a single architecture, enabling rapid generation of object regions using the RPN. This innovation allowed the entire object detection pipeline to be trained end-to-end, resulting in significant improvements in both accuracy and speed.

Technically, Faster R-CNN combines two key stages of object detection: first, the RPN generates object proposals, and then these proposals are classified and refined in terms of their locations through a regression mechanism. This unified approach substantially improved both the efficiency and the speed of object detection, establishing Faster R-CNN as a foundational framework in the field.

In this study, we employ a modified version of the Faster R-CNN architecture to perform open vocabulary object detection. By adjusting and extending specific components of the original network, our approach aims to maximize detection performance for a diverse range of classes that are not pre-defined, pushing the boundaries of open vocabulary detection.

2.2. CLIP

Previous computer vision systems were primarily limited to learning predefined object categories. This approach made it difficult for these systems to adapt to new visual concepts without additional labeled data, thus reducing their flexibility and usefulness. These limitations arose because they relied heavily on strict forms of supervision. However, with the introduction of CLIP[6], the gap between weakly supervised approaches and zero-shot learning using raw text has been significantly narrowed.

CLIP learns by processing text and image data together, mapping them into single embedding space. This allows the model to understand how images and text related to one another, which is at the core of its multimodal training. CLIP achieves this through contrastive learning, using text embeddings generated by a text encoder and image embeddings produced by an image encoder. By calculating cosine similarity, CLIP ensures that the similarity score is maximized for positive pairs (matching image-text pairs) and minimized for negative pairs. This process aligns the embeddings of positive pairs closely while separating those of non-matching pairs. As a result, CLIP learns a cohesive embedding space that bridges the gap between text and visual information.

One of CLIP's major strengths is its ability to perform open-vocabulary recognition, meaning it can handle new tasks or categories without needing additional fine-tuning on specific datasets or objects.

Additionally, CLIP[6] supports zero-shot learning, which allows it to tackle completely new tasks with no extra training. This capability is incredibly useful when dealing with situations where labeled data is limited or unavailable, making it a highly versatile model.

In our research, we take advantage of CLIP's pre-trained text and image encoders to perform open-vocabulary object detection (OVOD). Without retraining the encoders, we

directly use the embeddings they generate, aligning them through a contrastive learning framework similar to CLIP's original training process. By leveraging this alignment, we aim to improve the object detection model's ability to generalize to novel classes that were not seen during training, ultimately achieving better performance in OVOD tasks.

2.3. Open-Vocabulary Object Detection

Open Vocabulary Object Detection (OVOD) is a task in the object detection field that aims to detect not only predefined classes but also new and unseen objects. While traditional object detection models can only accurately identify objects within the trained classes, OVOD is designed to detect and classify objects that the model has not been explicitly trained on. The core technologies enabling this capability are zero-shot learning and vision-language models.

OVOD leverages large-scale image-text pairs to connect visual features with linguistic meanings. This allows the model to combine textual descriptions with visual features when detecting unfamiliar objects. Notable models like CLIP (Contrastive Language-Image Pretraining)[6] and ALIGN (Vision-Language Pretraining)[18] use this approach to align image embeddings with text embeddings, enabling detection of new objects.

OVOD has broad applications, such as in autonomous vehicles, where it helps detect novel road signs or obstacles in real-time, or in robotic vision, where it aids in recognizing untrained objects in various environments. It can also enhance image search by identifying and classifying previously undefined objects, expanding search capabilities.

Driven by advancements in OVOD, research is utilizing large datasets like COCO[19] and LVIS (Large Vocabulary Instance Segmentation)[20] to improve model generalization and explore more flexible methods for detecting a wider variety of objects.

2.4. OVR-CNN

OVR-CNN(Open Vocabulary Region-based Convolutional Neural Networks)[7] played a key role in advancing open-vocabulary object detection (OVOD). It was one of the first models to tackle the limitations of traditional object detection, which only focused on a fixed set of categories. OVR-CNN leveraged multimodal learning, combining text embedding with visual features, enabling the model to detect objects beyond the predefined classes, a foundational step in open-vocabulary detection.

One of the critical contributions of OVR-CNN was its use of contrastive learning to align text and image embeddings. This alignment improved the model's ability to generalize to unseen categories, influencing later models like CLIP[6] that further refined this approach. OVR-CNN also introduces zero-shot learning to object detection, demonstrating that a model could detect novel objects without needing additional training on those categories, an idea that has become central to modern OVOD tasks.

Although OVR-CNN is no longer in widespread use, its innovations in text-image alignment and zero-shot learning have greatly shaped current research. These principles continue to inform the development of more advanced models that push the boundaries of object detection in open-world scenarios.

2.5. ViLD

ViLD(Vision-Language Model Distillation)[13] has made significant strides in the field of open-vocabulary object detection by leveraging knowledge distillation from large-scale vision language models, such as CLIP[6]. ViLD transfers the generalization ability of these models, which have been trained on extensive datasets of image-text pairs, into a region-based object detection framework. Through this distillation process, ViLD enables object detection models to recognize novel categories not seen during training, overcoming the limitations of

traditional closed-set detection.

The key innovation in ViLD is its use of multimodal embedding and contrastive learning to align visual regions with textual descriptions. This alignment allows ViLD to detect objects from a broader, open vocabulary without the need for explicit fine-tuning on new object categories. ViLD's knowledge distillation approach is particularly relevant to our research, as we also employ distillation techniques to transfer the generalization capabilities of a vision-language model into our object detection framework.

By incorporating the knowledge distillation methodology, our work builds on the principles established by ViLD, using pre-trained vision-language models to improve the generalization of open-vocabulary object detection. This allows us to handle novel classes in a similar fashion, enhancing the model's performance in recognizing objects beyond the predefined categories.

3. Preliminaries

3.1. Open-Vocabulary Object Detection

Open Vocabulary Object Detection (OVOD) leverages pre-trained Vision-Language Models (VLMs) to detect novel categories (unseen data) by learning from base categories (seen data). OVOD primarily relies on the image-text embedding capabilities of VLMs, enabling it to excel in scenarios with limited or no annotations for novel categories.

In OVOD, the process begins with generating embeddings for both images and text using the image encoder and text encoder of a pre-trained VLM. Object regions are proposed through a Region Proposal Network (RPN), and features are extracted for these regions. These

features are then projected into the VLM's embedding space, allowing the model to establish meaningful relationships between the visual features of the image and the semantic features of the text. By utilizing the generalized representation power of the VLM, the model is capable of detecting unseen categories, even in the absence of novel class annotations. This makes OVOD a powerful approach for extending object detection capabilities to scenarios with limited or no labeled data for novel categories.

In contrast, applying Knowledge Distillation to OVOD introduces a different training paradigm. This approach frames the pre-trained VLM as a teacher model and the object detector as a student model, with the goal of transferring knowledge from the teacher to the student.

Specifically, the teacher model uses the pre-trained VLM's image encoder to extract powerful feature embeddings, creating a robust embedding space. Then, the RPN generates region proposals from the image, and these proposals are converted into embeddings by the student model's feature extractor. A similarity matrix is constructed between the embeddings of the teacher and the student, where positive pairs are trained to be closer and negative pairs are trained to be farther apart.

Through this process, the rich representation capabilities of the teacher model are effectively transferred to the student model, enabling the student to achieve strong detection performance not only on base categories but also on novel categories.

3.2. BARON

In this paper, for simplicity, we build upon the ideas of the existing BARON[14]. Various methods, including BARON, are built upon Faster R-CNN[16] to perform object detection. To classify objects from novel classes that were not seen during the training process, BARON adds a linear layer that projects each region proposal's region embedding into the word

embedding space. The projected region embeddings are referred to as **pseudo words**, and these pseudo words are then passed through a text encoder. The text embeddings obtained from the text encoder are compared with pre-trained category embeddings by calculating their similarity, ultimately yielding the classification results. Category embeddings primarily use pre-trained text embeddings from models like CLIP[6].

The text encoder T maps the pseudo word w to the pre-trained text embedding, assigning the pseudo word to a specific object category based on the distribution of probabilities as follows.

$$p_c = \frac{\exp(\tau \cdot \langle \mathcal{T}(w), f_c \rangle)}{\sum_{i=0}^{C-1} \exp(\tau \cdot \langle \mathcal{T}(w), f_i \rangle)} \quad (1)$$

$\langle \cdot \rangle$ denotes the cosine similarity between the two embeddings, and τ is a hyperparameter representing the temperature. For the total of C object categories provided, each category is embedded by modifying the 'category' part of a specific prompt (e.g., **'a photo of {category} in the scene'**), and the corresponding class embedding f_c is obtained.

In this process, open-vocabulary object detection relies on the base categories that are annotated only with bounding boxes during training, and other training processes and loss calculations follow methods similar to Faster R-CNN[16].

BARON[14] highlights a critical issue in existing knowledge distillation methods, where region embeddings are learned individually from the features of vision-language models (VLMs), such as CLIP. The authors of BARON argue that while existing methods align region embeddings with individual features, they fail to fully exploit the compositional structures present in the image. Therefore, BARON aims to go beyond learning the visual information of individual objects within an image by leveraging compositional structures. It further seeks to learn the co-existence of visual concepts, capturing the relationships between them.

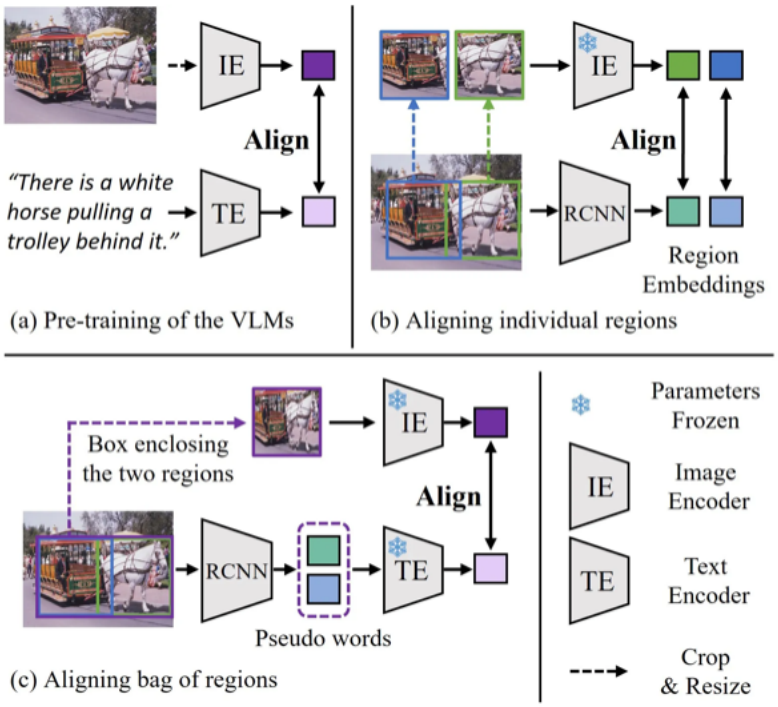

Figure 2 | (a) Typical vision-language models. (b) Existing knowledge distillation-based object detectors. (c) BARON method aligns the embedding of bag of regions.

BARON's training process builds upon the widely studied knowledge distillation approach[13][15]. Similarly, it trains the model to classify region embeddings using pre-trained text embeddings, while also ensuring that the region proposals are learned to resemble the image embeddings obtained by feeding them into a frozen pre-trained image encoder. This enables the model to transfer the visual knowledge embedded in the pre-trained image encoder.

The BARON model learns to represent and align "bags of regions" (groups of object regions) between a student model (the open-vocabulary object detector) and a teacher model (vision-language models, VLMs). The learning process consists of three main steps:

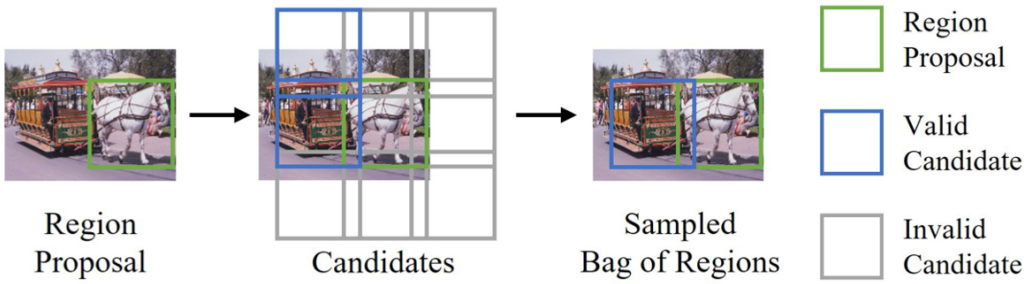

Figure 3 | Forming Bag of Regions

1) Forming bag of regions : BARON creates bag of regions to capture the relationships between multiple objects appearing together in an image. To do this, it groups adjacent regions with similar sizes. This method focuses on multiple areas rather than a single regions. The Region Proposal Network(RPN) is used to select the primary region of interest, and surrounding regions are included to form a bag.

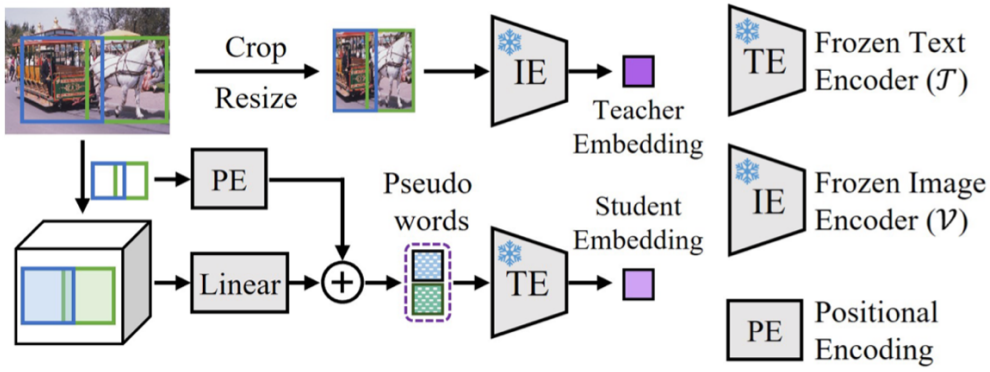

Figure 4 | Representing Bag of Regions

2) Representing bag of regions: In BARON, representing the bag of regions involves two key models: the student model and the teacher model. Each region is denoted as b_j^i, where j represents the j-th region within i-th bag of regions. These bags are used to capture the

spatial and semantic relationships between multiple regions in an image.

For the student model, the regions are first converted into pesudo words, denoted as w_j^i. To ensure that spatial information, such as relative box positions and sizes, is not lost, positional embeddings p_j^i are added to theses pseudo words. This helps the student model understand the spatial relationships between regions in the bag. The combined pseudo words and positional emebeddings are then fed intp the text encoder $\mathcal{T}$, producing the student's bag-of-reigons emebedding f_t^i :

$$f_t^i = \mathcal{T}\left(w_0^i + p_0^i\ , w_1^i + p_1^i\ , \cdots\ , w_{N-1}^i + p_{N-1}^i\right) \tag{2}$$

For the teacher model, the regions in the bag, represented as b_j^i, are enclosed within an image crop. This crop is then passed through the image encoder $\mathcal{V}$ of the vision-language model (VLM). Any content outside the defined regions is masked out, ensuring that the features focus only on the relevant areas. The teacher's bag-of-regions emebedding f_v^i is calculated as :

$$f_v^i = \mathcal{V}(b_0^i\ , b_1^i, \cdots\ , b_{N-1}^i) \tag{3}$$

These emebeddings f_t^i and f_v^i are the foundational representations used for alignment in the next step.

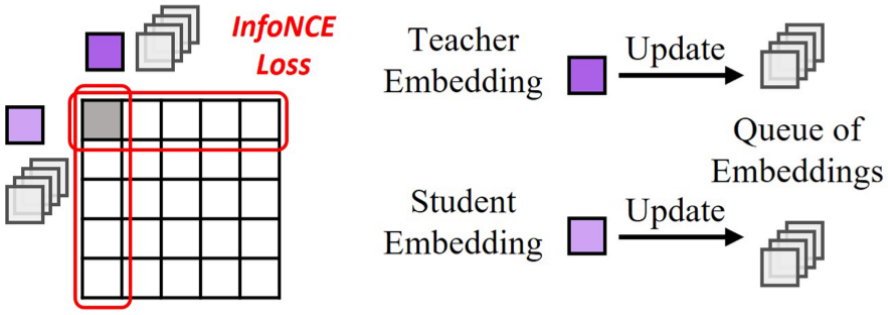

Figure 5 | Aligning Bag of Regions

3) Aligning bag of regions : To align the embeddings from the student and teacher models, BARON employs a contrastive learning approach. This process ensures that the student model learns to encode the coexistence of multiple regions and their relationships effectively. The alignment process uses the InfoNCE loss function, which encourages positive pairs (similar embeddings) to be closer while pushing negative pairs (dissimilar embeddings) apart. The loss for aligning the embeddings is calculated as :

$$\mathcal{L}_{bag} = -\frac{1}{2}\sum_{k=0}^{G-1}\left(\log\left(p_{t,v}^{k}\right) + \log(p_{v,t}^{k})\right) \tag{4}$$

Here, $p_{t,v}^{k}$ and $p_{v,t}^{k}$ represent the probabilities of the student and teacher embedddings aligning correctly for the k-th bag of regions. These probabilities are defined as :

$$p_{t,v}^{k} = \frac{\exp\left(\tau_{bag} \cdot \langle f_t^k, f_v^k \rangle\right)}{\sum_{l=0}^{G-1} \exp\left(\tau_{bag} \cdot \langle f_t^k, f_v^l \rangle\right)} \tag{5}$$

$$p_{v,t}^{k} = \frac{\exp\left(\tau_{bag} \cdot \langle f_v^k, f_t^k \rangle\right)}{\sum_{l=0}^{G-1} \exp\left(\tau_{bag} \cdot \langle f_v^k, f_t^l \rangle\right)} \tag{6}$$

In these equations, τ' is the temperature parameter used to rescale cosine similarity values and G denotes the total number of bags of regions. This alignment process helps the student model not only learn the features of individual regions but also capture the spatial and conceptual relationships between multiple regions within an image.

4. Methods

4.1. Limitations of BARON

Fig. 6 shows that the image and region proposal are effectively aligned in the final stages of knowledge distillation. This method allows the student embedding to fully capture the visual knowledge transferred from the teacher embedding.

The pairs indicated by green arrows are recognized as positive pairs in contrastive learning and are trained to become closer, while the pairs marked by red arrows are identified as negative pairs and are trained to become farther apart. On the surface, this may seem unproblematic, but when considering that contrastive learning is being applied between different bags of regions sampled around the same origin region proposal, an issue arises. Furthermore, since these student embeddings are simply concatenated features, and they likely share similar patches, it becomes problematic that the student embeddings obtained from the same origin region proposal (yellow arrows) are being trained to move farther apart. While in practice, enough negative samples are provided, it is clear that there is still room to further improve the training process.

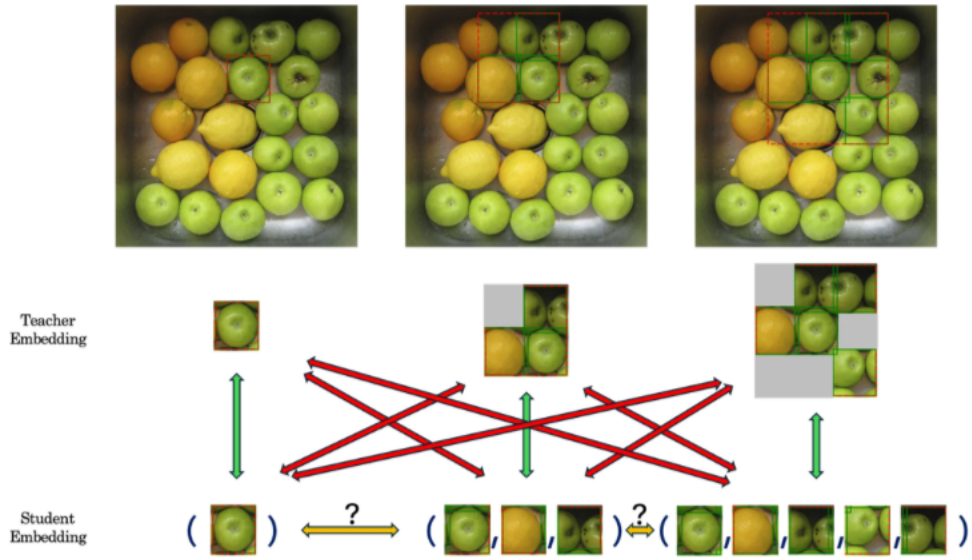

Figure 6 | Problem of aligning bag of regions

The Fig. 7 below presents the self-similarity matrices of bag-of-region features within a batch, obtained from both the trained BARON model (left) and the pre-trained CLIP image encoder (right). When generating the self-similarity matrices, certain cases were masked out: self-comparisons and instances where the permutation of candidate neighbor regions was identical, resulting in the same bag of regions with similarity scores of 1. For the remaining feature pairs, cosine-similarity was calculated and normalized using a Softmax function. Bag of regions derived from the same region proposal are enclosed in red boxes for clarity.

In the case of the pre-trained CLIP image encoder (teacher model), bag-of-region embeddings derived from the same region proposal exhibit consistently high similarity scores. In contrast, the trained BARON model fails to sufficiently capture this pattern. Since these bags of regions are centered on the same region proposal, they are formed using similar patches and focused on overlapping areas, which intuitively suggests higher cosine similarity. However, this expected behavior is not reflected in the trained BARON model's outputs.

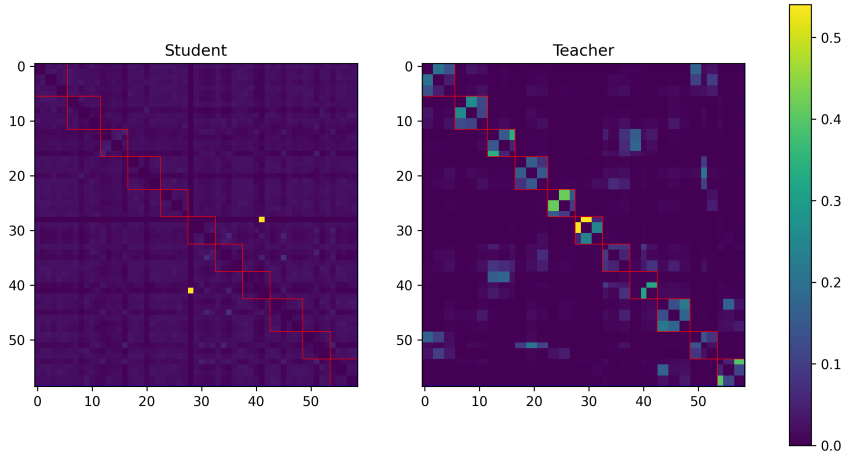

Figure 7 | Self-similarity matrices of bag-of-region features within a batch, obtained from both the trained BARON model (left) and the pre-trained CLIP image encoder (right)

4.2. Label Smoothing

As the first solution, we applied the Label Smoothing[22] technique, which introduces a small degree of uncertainty to the ground truth labels to prevent the model from becoming overly confident about a specific class. In the original approach, contrastive learning was performed using an affinity matrix represented as a one-hot vector. This matrix was used to calculate the similarity matrix between the teacher and student embeddings, ensuring positive pairs are brought closer together while negative pairs are pushed farther apart. However, embeddings from different bags within the same region proposal were often treated as completely negative pairs, causing them to diverge excessively during training.

To address this issue, we introduced a smoothing parameter, α, to adjust the separation between negative pairs. For example, with $\alpha = 0.1$, the probability of the correct class is reduced from 1 to 0.9, and the remaining probability is evenly distributed across other classes. This adjustment forces the model to maintain a small degree of uncertainty, thereby reducing overfitting and mitigating the vulnerabilities caused by overconfidence.

However, the previous approach had a limitation of applying the same α value to all negative pairs, failing to account for the relationships between components of the bag. To address this, we explored a new approach that incorporates Positional Similarity. Using Jaccard Similarity, we calculated the similarity between region embeddings and applied higher smoothing values for closer embeddings and lower smoothing values for more distant embeddings. This adjustment ensures that highly similar negative pairs are not pushed too far apart, reflecting the inclusion relationships among the components within the bag.

Unfortunately, the performance with Positional Similarity was lower than when using a fixed \alpha. This result suggested that relying on a fixed degree of adjustment for negative pairs might not be effective. Consequently, we shifted our focus to Similarity-Preserving Knowledge Distillation, which preserves the relative distances between embeddings during training. This method ensures that the student model learns to maintain the spatial and semantic relationships between visual concepts, leading to more robust and generalized performance.

4.3. Similarity-Preserving Knowledge Distillation

As the second solution, we adopted the Similarity Preserving Knowledge Distillation[23] approach. This method emphasizes the importance of not only improving the accuracy of individual embeddings but also preserving the relative relationships between embeddings during the learning process. By learning the self-similarity matrix of the teacher and student embeddings, similarity-preserving knowledge distillation ensures that the structural relationships between different embeddings, as captured in the teacher model, are effectively transferred to the student model. This approach is grounded in research that highlights the significance of capturing these relative similarities as an integral part of knowledge distillation.

The core idea of similarity-preserving knowledge distillation is straightforward yet

powerful. First, the embeddings from both the teacher and student models are transformed into self-similarity matrices, which encode the pairwise similarities among embeddings in their respective spaces. These matrices capture the interrelationships within the embedding space. To ensure comparability, the matrices are normalized to align their scales. The divergence between the teacher and student self-similarity matrices is then quantified using the Frobenius Norm, which calculates the Euclidean distance by summing the squared differences of corresponding elements.

To minimize this divergence, the Frobenius Norm is incorporated into the training process as an auxiliary loss term, added to the existing loss function. This auxiliary loss encourages the student model to mimic the teacher's self-similarity matrix, thereby preserving the relative relationships between embeddings. This approach goes beyond simply refining individual embedding quality; it enables the student model to better capture the structured relationships inherent in the teacher model's embeddings.

By preserving these relational structures, similarity-preserving knowledge distillation enhances the student model's ability to generalize, particularly when dealing with unseen data or novel categories. This method ensures that the student model inherits not just the feature extraction capabilities of the teacher but also the critical inter-embedding relationships that underpin robust generalization.

Ultimately, similarity-preserving knowledge distillation leverages the relational structure of embeddings to improve the quality of knowledge transferred during training, resulting in a model that delivers superior performance even in challenging, unseen scenarios. This approach demonstrates the profound impact of incorporating structural alignment into knowledge distillation.

5. Experiments

5.1. Dataset and Evaluation Metrics

Dataset Our proposed method is evaluated on two renowned object detection benchmarks, with primary emphasis on the COCO dataset. Adopting the experimental framework introduced by OV-RCNN, we categorize the object classes into 48 base categories and 17 novel categories. This partitioning enables a thorough examination of our model's capacity to generalize to unseen categories, while simultaneously leveraging the rich annotations and diverse visual representations inherent in the COCO dataset. By employing this setup, we aim to underscore the strength of our approach in delivering robust performance across both base and novel categories—a pivotal challenge in the domain of open-vocabulary object detection.

Evaluation Metrics To thoroughly evaluate detection performance, we measure our model's effectiveness on both base and novel categories. This strategy provides a comprehensive perspective on the model's ability to generalize to unseen categories while maintaining high accuracy on familiar ones. For the COCO dataset, within the scope of open-vocabulary object detection (OVCOCO), we adhere to the evaluation protocol outlined by OV-RCNN. In line with this framework, we report the box Average Precision (AP) at an Intersection over Union (IoU) threshold of 0.5, commonly referred to as AP50. This metric serves as a robust indicator of the precision-recall balance, offering a precise assessment of detection performance across both base and novel categories.

5.2. Implementation Details

We use BARON, built on Faster R-CNN with ResNet50-FPN, as our baseline model. To ensure a fair comparison with previous methods, the backbone network is initialized with weights pre-trained by SOCO, and we incorporate synchronized Batch Normalization (SyncBN),

as suggested by DetPro. All experiments were conducted exclusively on the COCO dataset, and the model was trained for 90k iterations using a 1× training schedule. For the Vision-Language Model (VLM), we employed the CLIP model with a ViT-B/32 architecture.

Methods	AP_{50}^{base}	AP_{50}^{novel}
BARON (Baseline)	0.3390	0.2310
Label Smoothing [pos_sim=False, $\alpha = 0.005$]	**0.3520**	**0.2430**
Label Smoothing [pos_sim=True, $\alpha = 0.005, \beta = 0.015$]	0.3330	**0.2310**
Similarity-Preserving KD ($\gamma = 2500$)	**0.3420**	**0.2430**
Similarity-Preserving KD ($\gamma = 2500$) + Label Smoothing [pos_sim=False, $\alpha = 0.01$]	**0.3410**	**0.2320**
Similarity-Preserving KD ($\gamma = 2500$) + Label Smoothing [pos_sim=True, $\alpha = 0.005, \beta = 0.015$]	0.3310	**0.2420**

Table 1 | Comparison between the original BARON & Ours

5.3. Quantitative Results

The comparison between the original BARON and our proposed methods under the same experimental settings is reported in Table 1. First, we observe that applying label smoothing leads to an improvement of 1.2 AP_{50} on novel categories compared to the original BARON. However, empirically, the performance exhibited significant sensitivity to the hyperparameter α, varying considerably with random seeds. When incorporating positional similarity, this sensitivity to the α value became even more pronounced.

Similarly, applying similarity-preserving knowledge distillation, which encourages the alignment of the student model with the teacher model's self-similarity matrix, also demonstrated performance gains of 1.2 AP_{50} on novel categories.

While combining both similarity-preserving knowledge distillation and label smoothing resulted in additional performance improvement, it was not significantly greater than using similarity-preserving knowledge distillation alone.

5.4. Qualitative Results

Fig. 8 illustrates the self-similarity matrices obtained from the teacher and student models after applying similarity-preserving knowledge distillation. Similar to Fig. 7, masking was performed to exclude self-comparisons and identical permutations, and Softmax was applied to the cosine-similarity of the remaining embeddings. As a result, we observe a notable increase in similarity between bag-of-region embeddings centered on the same region proposal, highlighted by the red boxes.

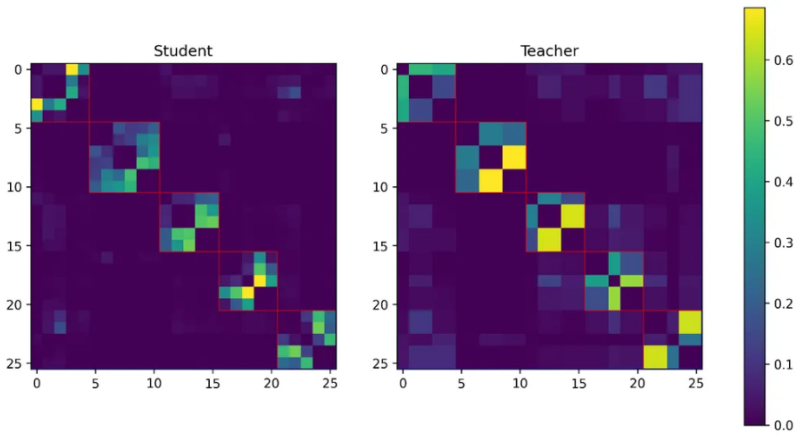

Figure 8 | Self-similarity matrices of bag-of-region features within a batch, obtained from both the similarity-preserved BARON model (left) and the pre-trained CLIP image encoder (right)

6. Conclusion

The original BARON identified the limitation of most knowledge distillation-based Open-Vocabulary Object Detection methods, which focus solely on learning individual regions. It further expanded this paradigm by introducing learning through bags of regions. This approach is significant as it attempts to leverage the potential ability of Vision-Language Models (VLMs) to learn the compositional structure among multiple objects. However, at the same time, it falls short of fully utilizing the characteristics of the proposed bag-of-region representations.

By demonstrating both intuitively and quantitatively the inherent relationships among bags of regions centered on the same region proposal, we emphasized the importance of aligning these bags more effectively during contrastive learning. Our experiments revealed that accounting for these relationships leads to significant improvements in novel category detection.

Ultimately, our approach refines the recognition of visual concepts surrounding a region proposal while more precisely capturing the relationships between neighboring patches that should have been aligned more closely. This enables the learning of a more sophisticated visual embedding space that captures the co-existence of objects—an essential capability for addressing the novel category learning challenges inherent in open-vocabulary object detection tasks.

Reference

[1] A. Bansal, K. Sikka et al., “Zero-shot object detection,” in ECCV, 2018

[2] S. Rahman, S. Khan et al., “Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts,” in ACCV, 2019.

[3] P. Zhu, H. Wang et al., “Zero shot detection,” TCSVT, 2019.

[4] H. Bilen and A. Vedaldi, “Weakly supervised deep detection networks,” in CVPR, 2016.

[5] D. Zhang, J. Han et al., “Weakly supervised object localization and detection: A survey,” TPAMI, 2021.

[6] A. Radford, J. W. Kim et al., “Learning transferable visual models from natural language supervision,” in ICML, 2021.

[7] A. Zareian, K. D. Rosa et al., “Open-vocabulary object detection using captions,” in CVPR, 2021.

[8] R. Krishna, Y. Zhu et al., “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” IJCV, 2017.

[9] N. Carion, F. Massa et al., “End-to-end object detection with transformers,” in ECCV, 2020.

[10] Y. Zhong, J. Yang et al., “Regionclip: Region-based language-image pretraining,” in CVPR, 2022.

[11] X. Zhou, R. Girdhar et al., “Detecting twenty-thousand classes using image-level supervision,” in ECCV, 2022.

[12] M. Gao, C. Xing et al., “Open vocabulary object detection with pseudo bounding-box labels,” in ECCV, 2022.

[13] X. Gu, T.-Y. Lin et al., “Open-vocabulary object detection via vision and language knowledge distillation,” arXiv, 2021.

[14] S. Wu, W. Zhang et al., “Aligning bag of regions for open-vocabulary object detection,” in CVPR, 2023.

[15] Y. Du, F. Wei et al., “Learning to prompt for open-vocabulary object detection with vision-language model,” in CVPR, 2022.

[16] S. Ren, K. He et al., “Faster r-cnn: Towards real-time object detection with region proposal networks,” NeurIPS, 2015.

[17] M. Minderer, A. Gritsenko et al., “Simple open-vocabulary object detection with vision transformers,” arXiv, 2022.

[18] C. Jia, Y. Yang et al., “Scaling up visual and vision-language representation learning with noisy text supervision,” in ICML, 2021.

[19] T.-Y. Lin, . Maire et al., “Microsoft coco: Common objects in context,” in ECCV, 2014.

[20] A. Gupta, P. Dollar et al., “Lvis: A dataset for large vocabulary instance segmentation,” in CVPR, 2019.

[21] A. v. d. Oord, Y. Li et al., “Representation learning with contrastive predictive coding,” arXiv, 2018.

[22] C. Szegedy, V. Vanhoucke et al., “Rethinking the Inception Architecture for Computer Vision”, in CVPR, 2016.

[23] F. Tung, G. Mori, “Similarity-Preserving Knowledge Distillation”, in ICCV, 2019.