Synonym relations affect object detection learned on vision-language data

Anonymous ACL submission

Abstract

We analyze whether object detectors trained on vision-language data learn effective visual representations for synonyms. Since many current vision-language models accept user-provided textual input, we highlight the need for such models to learn feature representations that are robust to changes in how such input is provided. Specifically, we analyze changes in synonyms used to refer to objects. Here, we study object detectors trained on vision-language data and investigate how to make their performance less dependent on whether synonyms are used to refer to an object. We propose two approaches to achieve this goal: data augmentation by back-translation and class embeddings enrichment. We show the promise of such approaches, reporting improved performance on synonyms from mAP@0.5=33.87% to 37.93%.

1 Introduction

In recent years, we have witnessed increased interest in vision-language models (Radford et al., 2021; Yuan et al., 2021) that learn joint image and text representations in a self-supervised way, and that can later be used as building blocks for models fine-tuned on downstream tasks (Wu et al., 2023; Kuo et al., 2022; Kim et al., 2023). In addition, recent models such as GPT-4 (OpenAI, 2023) and DALL-E 3 (Betker et al., 2023) are built to accept image and text input provided by end users, with no set constraints on such inputs. Thus, models must be robust to variations in how input is provided.

We analyze how vision-language models handle the variability in textual inputs. Specifically, we investigate variations in synonyms used to refer to objects. We show how such variability negatively affects performance for open-vocabulary object detection, and we propose two ways to help vision-language detectors learn better representations for synonyms: data augmentation by back-translation and class embeddings enrichment.

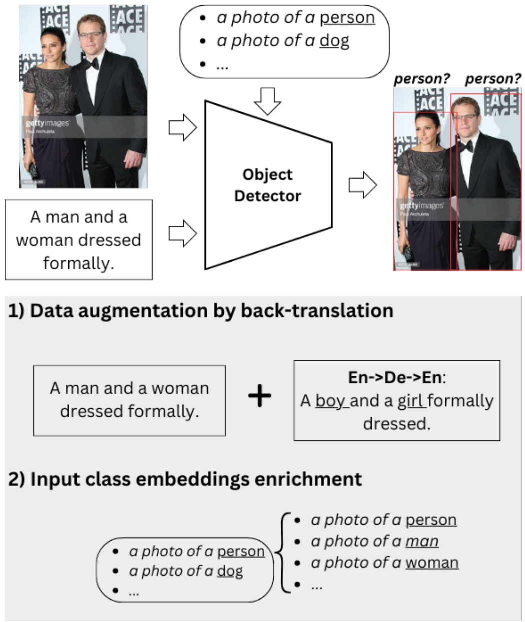

Figure 1: Top: input to an open-vocabulary object detector: images, class embeddings, and captions; and its output: bounding boxes with associated labels. Bottom: our approaches. 1) Data augmentation by back-translation: add captions back-translated from a foreign language; 2) Class embeddings enrichment: consider synonyms when extracting class embeddings.

Figure 1 illustrates our proposed approaches. With back-translation, we use a machine translation model to translate captions from English to another language, and then we translate them back to English. Because the back-translation is not perfect, the original caption and the back-translated one are not going to be the same: they will show changes, for instance, in which nouns are used to refer to objects (i.e., synonyms). We hypothesize that adding more synonyms to the captions used for training will help a model learn better representations for them. With class embeddings enrichment, we modify the class embeddings that open-vocabulary object detectors (Wu et al., 2023; Gu et al., 2021; Minderer et al., 2022) use to match visual embeddings learned for image regions. Furthermore, when training with enriched class embeddings, we experiment with enriching them throughout the whole training process, or using a Curriculum Learning approach: start training with the original embeddings, and finish with the enriched ones.

In both our approaches, we modify *inputs* to the training process (i.e., captions and class embeddings), making them generalizable to different model architectures and training strategies. We show promising results with improved performance on synonyms from mAP@0.5=33.87% to 37.93%.

In summary, our contribution is twofold: (1) we identify an issue with current state-of-the-art (SOTA) vision-language object detector models (namely, difficulty in detecting objects referred to by synonyms), and (2) we propose two generalizable strategies to train vision-language object detectors to learn better representations for synonyms.

2 Related Work

Vision-language (VL) models for open-vocabulary detection. Open-vocabulary object detection refers to training a detector model on a set of classes and testing it also on a separate set of classes unseen during training (Gu et al., 2021; Gao et al., 2022; Minderer et al., 2022; Kim et al., 2023; Wu et al., 2023). Many methods take advantage of large pre-trained VL models (Radford et al., 2021; Jia et al., 2021; Lu et al., 2019) that are generally trained to recognize which image-caption pairs match and which do not. In this work, we use BARON (Wu et al., 2023): a state-of-the-art (SOTA) open-vocabulary object detector making use of the CLIP (Radford et al., 2021) pre-trained VL model.

Concept relationships. Text embeddings have been shown to encode relationships between concepts such as synonyms and antonyms (Lu et al., 2018; Gokhale et al., 2022). At the same time, studies on adversarial attacks have highlighted how performance of language models varies when the input is changed, even when preserving the semantic meaning of the input text (Jia et al., 2019; Zhu et al., 2019; Ribeiro et al., 2018). Unsurprisingly, when such language models are combined with vision models, similar problems arise, with performance on VL tasks varying under perturbations of text input (Tascon-Morales et al., 2023; Gokhale et al., 2022; Sheng et al., 2021; Gokhale et al., 2020). Our work is related to such studies since we aim to make VL models more robust to text input variations, although we differ from previous work in target task (object detection vs. visual reasoning). Further, we do not require changes in how a model is trained, for instance, by defining a new loss function (Gokhale et al., 2022; Tascon-Morales et al., 2023); we simply modify inputs to the model, making our approach more general.

Curriculum learning. Curriculum learning (Bengio et al., 2009) (CL) refers to training a deep learning model by ordering the training samples; a model can learn better if the training samples are chosen following a schedule (i.e., a curriculum) rather than randomly selected. Previous work has shown the promise of CL for tasks such as machine translation (Liu et al., 2023; Qian et al., 2021), automated text scoring (Zeng et al., 2023), and common sense reasoning (Maharana and Bansal, 2022). We apply the idea of changing the input a model is trained on, but, instead of changing the training images, we change what class embeddings the model is trained on.

3 Methods

3.1 Object detection: BARON

We choose BARON (Wu et al., 2023) as our vision-language open-vocabulary detector since it achieves SOTA results on the task of open-vocabulary detection. BAg of RegiONs (BARON) is based on Faster R-CNN (Ren et al., 2015), where the classification layer is replaced by a linear layer so that its output is an embedding (or pseudo-words), rather than a class label. The key novelty of this method is the introduction of bags of regions: embeddings are extracted for a set of bounding boxes around each region proposal, not for a single proposal only. This is to model the co-occurrence of bags of visual concepts. BARON is trained from images and captions, and it requires a list of class embeddings (extracted from object names) to classify each region proposal. At test time, an image is fed to the model and bounding boxes are classified by comparing the extracted visual embeddings with the provided class embeddings. If we change such class embeddings by extracting them with synonyms, detection performance significantly drops (Table 2), motivating our work.

3.2 Evaluating using synonyms

To evaluate the ability of a model to detect objects when using synonyms, we change the class em-

Original	German	Russian
A **skate board rider** does a trick in front of a building.	A **skateboarder** does a trick in front of a building.	A **skater** does a trick in front of the building.
Three adults help a **youngster** follow a sheet of instructions.	Three adults help a **teenager** follow a sheet of instructions.	Three adults help the **teenager** follow instructions.

Table 1: Examples of (left) original COCO captions, (middle) captions back-translated from German, and (right) captions back-translated from Russian.

beddings during inference by replacing each class name with one of its synonyms and computing the class embeddings using such synonym. Since we have multiple synonyms per class, we repeat this process 5 times (with 5 different synonyms), and we compute the mean and standard deviation of the detection performance across these five runs. The mean measures how well the downstream task is performed when varying input synonyms, the standard deviation measures how variable performance is: if a model learned all synonyms as well as class names, standard deviation would be 0 (i.e., performance does not depend on the input synonym).

3.3 Augmentation by back-translation

In our first approach, we apply a machine translation model from English to another language to the input captions, and then translate the translated caption back to English. This approach has been successfully used as a data augmentation strategy on NLP tasks (Edunov et al., 2018; Xie et al., 2020; Sennrich et al., 2016) but it is less explored for VL models. Back-translation (BT) is a form of data augmentation because the BT process is imperfect: the back-translated caption will not be the same as the original one. There can be changes in, for instance, words used to refer to objects (i.e., synonyms), which is our motivation for proposing this method: we hypothesize that the increased variability in the vocabulary used to describe objects is beneficial to learn robust feature representations.

3.4 Class embeddings enrichment

In our second approach, we enrich the class embeddings BARON is trained with by incorporating synonyms. Class embeddings are matched to region proposals to assign a class to each region proposal: the class whose embedding is most similar to that predicted for the region proposals. We compute class embeddings off-line using a CLIP Text Encoder (TE): for each class (e.g., person), we process a list of prompts through the TE (e.g., “A picture of a person”, “A photo of a person”), returning one embedding per prompt; their average is taken as the overall class embedding. When enriching the class embeddings, we do not only add the class name (e.g., “person”) in the prompts, but also each synonym for that class (e.g., “man”, “woman”). The enriched class embedding is the average of the resulting text embeddings for prompts with the class name and its synonyms.

3.5 Curriculum learning

A potential issue with our embeddings enrichment approach is that, when training on enriched embeddings and testing on object names, the shift in training vs. test embeddings may cause a decrease in performance. We propose curriculum learning to train with both the original class embeddings and our enriched version: we start training on the former, and finish training on the latter. By seeing both sets of embeddings during training, we hypothesize a model will perform competitively when evaluated both on object names and synonyms.

4 Results

4.1 Implementation

We train models on COCO Captions (Chen et al., 2015) and evaluate them on COCO Objects (Lin et al., 2014), and we use the list of synonyms made available by (Lu et al., 2018) for synonym evaluation (e.g., “ship, motorboat” for “boat”, “plane, aircraft” for “airplane”). In this list, only 44 of the 80 COCO class have at least one synonym, so we limit evaluation to this subset of classes.

For machine translation, we use the Facebook FAIR WMT2019 models (Ng et al., 2019).

To train and evaluate BARON[1], we reduce batch size from 16 to 12 due to hardware constraints. For curriculum learning experiments, we train with one set of class embeddings for half of the training process, and we finish with the other set.

[1] `https://github.com/wusize/ovdet/tree/main`, last accessed October 10th, 2023

Captions	COCO names	Synonyms mean (std)	Avg.
Original	**44.45**	*33.87* (5.94)	*35.63*
Back-translation			
German	*44.23*	**34.25** (5.32)	**35.91**
Russian	43.89	33.67 (5.99)	35.37
Both	42.92	32.89 (5.97)	34.56

Table 2: Back-translation: mAP@0.5 (as %) evaluated on COCO class embeddings ("COCO names") and on synonyms embeddings ("Synonyms"). "Avg.": mean performance across the 5 synonyms and the COCO name. **Bold**: highest performance, *italics*: second-best.

Captions	COCO names	Synonyms mean (std)	Avg.
Class embeddings: COCO names			
Original	**44.45**	33.87 (5.94)	35.63
BT: German	*44.23*	34.25 (5.32)	35.91
Class embeddings: enriched			
Original	43.58	*37.25* (4.56)	*38.31*
BT: German	37.48	36.75 (4.56)	36.87
Curriculum	43.49	**37.93** (3.22)	**38.85**

Table 3: Class embedding enrichment: mAP@0.5 (as %) evaluated on COCO class embeddings ("COCO names") and on synonyms embeddings ("Synonyms").

4.2 Evaluating using synonyms' embeddings

We now evaluate models on synonyms used as test class embeddings. As a baseline, we train a model on the original COCO captions and COCO class name embeddings, and we compare it with models trained using back-translation or class embeddings enrichment. In Table 2, we see performance greatly drops when using synonyms as opposed to COCO names (mAP@0.5=44.45% vs. 33.87% when training with original captions). This corroborates the need to better learn synonyms during training.

4.3 Augmentation by back-translation

We qualitatively verify that back-translation increases the use of synonyms by showing examples of original COCO captions and their back-translated versions with two languages: German and Russian. From Table 1, we see that back-translation is successful at introducing synonyms: "skateboarder" or "skater" in the first caption and "teenager" in the second. In addition, we compute the ratio between the number of mentions of an object using a synonym divided by the total number of mentions (synonyms and verbatim mention of the COCO object name). We compare such ratio computed from the original captions and from the back-translated (BT) ones, obtaining 0.317 for original captions, 0.326 for BT: German, 0.344 for BT: Russian, and 0.343 for BT: Both. These results corroborate our assumption that back-translation increases variability in synonyms usage.

From Table 2, adding back-translated captions from German improves mean performance on synonyms (with a slight decrease in performance on class names), as well as decreases variability in performance (from 5.94% to 5.32%), showing improved robustness to variations in input synonym.

4.4 Class embeddings enrichment

Table 3 shows increased mean performance on synonyms when enriching class embeddings (mAP@0.5=37.25% vs. 33.87%, and std=4.56% vs. 5.94%, respectively), as well as increased overall average performance (38.31% vs. 35.63%). These results show the promise of enriching class embeddings, although we notice a small decrease in performance when evaluating on COCO names when training with original captions (larger when comparing BT with/without enrichment). When evaluated on synonyms, combining back-translation and embedding enrichment yields an improvement over using back-translation (mAP@0.5=34.25% to 36.75%).

4.5 Curriculum learning

In Table 3 (bottom), we notice how curriculum learning improves performance on synonym evaluation compared to COCO embeddings and enriched embeddings, while performance on COCO names decreases only slightly. Average performance improves (mAP@0.5=38.31% to 38.85%). To our knowledge, this is one of the first results demonstrating curriculum learning for object detection using VL data for training.

5 Conclusions

In this work, we considered variations in nouns used to refer to objects (i.e., synonyms), and how they affect performance of vision-textual object detectors. We highlighted how detecting objects when synonyms are used as input is challenging, and we introduced two approaches to ameliorate this issue, which proved successful at boosting detection performance on synonyms.

6 Limitations

In this work, we show the promise of altering the training process of vision-language object detectors to help learn more robust representations that better adapt to variations in textual input in terms of synonyms used to refer to objects. Despite such promise, our study has some limitations. First, we only evaluate on object detection; further studies on other vision and language tasks (e.g., visual question answering) are needed to fully characterize the problem and evaluate the proposed solutions. Second, we evaluate only on synonyms provided by (Lu et al., 2018). Although the used synonyms allow us to show our main points, more comprehensive synonyms' lists can be tested. Third, we show the impact of our approaches on one model (i.e., BARON (Wu et al., 2023)); while this is a SOTA open-vocabulary object detection model whose overall design is similar to that of other detectors (Minderer et al., 2022; Gu et al., 2021), repeating our experiments with other models would better show the generalizability of our proposed strategies. Finally, our approaches to better learn synonyms focus on changing the input to the model (whether it being the captions or the class embeddings it is trained with). While such a choice makes our approach independent of the model's inner architecture (e.g., how features are extracted and combined) or the training process (e.g., how a batch is constructed), more individualized approaches are worth investigating to solve the observed trade-off between performance on synonyms and on object names.

Ethical considerations. In our work, we use a machine translation model to augment captions with synonyms. Such models may have learned gender-related biases (e.g., doctor/man, nurse/woman) that, in turn, could be passed on to the object detector (making it easier for the model to detect people in a certain profession if they are of a specific gender). The fact that we keep the original captions and add the back-translated one should offer some safeguards against this issue.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *International Conference on Machine Learning*, pages 41–48.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, page 489. Association for Computational Linguistics.

Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. 2022. Open vocabulary object detection with pseudo bounding-box labels. In *European Conference on Computer Vision*, pages 266–282. Springer.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Vqa-lol: Visual question answering under the lens of logic. In *European conference on computer vision*, pages 379–396. Springer.

Tejas Gokhale, Abhishek Chaudhary, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2022. Semantically distributed robust optimization for vision-and-language inference. *ACL 2022 Findings*.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4129–4142.

Dahun Kim, Anelia Angelova, and Weicheng Kuo. 2023. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11144–11154.

Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. 2022. Open-vocabulary object detection upon frozen vision and language models. In *International Conference on Learning Representations*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Min Liu, Yu Bao, Chengqi Zhao, and Shujian Huang. 2023. Selective knowledge distillation for non-autoregressive neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13246–13254.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228.

Adyasha Maharana and Mohit Bansal. 2022. On curriculum learning for commonsense reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 983–992.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. 2022. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's wmt19 news translation task submission. *WMT 2019*, page 314.

OpenAI. 2023. Gpt-4 technical report.

Lihua Qian, Hao Zhou, Yu Bao, Mingxuan Wang, Lin Qiu, Weinan Zhang, Yong Yu, and Lei Li. 2021. Glancing transformer for non-autoregressive neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1993–2003.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 856–865.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. 2021. Human-adversarial visual question answering. *Advances in Neural Information Processing Systems*, 34:20346–20359.

Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. 2023. Logical implications for visual question answering consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6725–6735.

Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. 2023. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15254–15264.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Zijie Zeng, Dragan Gasevic, and Guangliang Chen. 2023. On the effectiveness of curriculum learning in educational text scoring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14602–14610.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2019. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*.