VEIL: Vetting Extracted Image Labels from In-the-Wild Captions for Weakly-Supervised Object Detection

Arushi Rai and Adriana Kovashka

University of Pittsburgh | School of Computing and Information

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction

- Weakly supervised object detection has primarily been applied to smaller-scale, paid-for crowdsourced vision-language datasets like COCO and Flickr30K.
- Larger VL-datasets contain a significant number of **objects mentioned** in captions that are **not present in the corresponding image**.
- We **introduce the task of vetting labels** extracted from captions.
- **VEIL** is introduced to vet each extracted label from a caption and is trained by bootstrapping visual presence info from pretrained object detectors.
- We **compare our method to eleven baselines**.
- VEIL **improves weakly-supervised detection by 80% compared to no vetting** (16.0 to 29.1 mAP) and surpasses Large Loss Matters by +11 mAP and CLIP filtering by +18 mAP on PASCAL VOC.

Label Noise

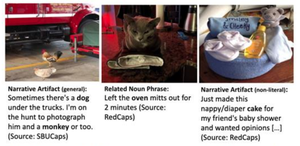

Figure 1. A naive label extraction process, e.g. substring matching between class names and caption, can lead to visually absent labels.

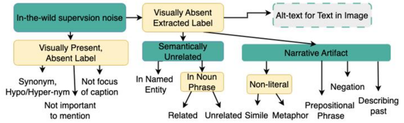

Figure 2. Label noise taxonomy outlining common reasons for visual absence in extracted labels from in-the-wild captions.

Baselines

No Vetting. Accept all extracted labels (perfect recall).
Global CLIP. Calculate the cosine similarity between image and text with the prompt "A photo depicts" prepended. We use a GMM for predictions to handle cosine similarity distribution differences between datasets.
CLIP-E. We curate multiple prompts, prepend them to the caption, and use the score from the highest-scoring prompt.
Local CLIP follows a similar process as GlobalCLIP but computes cosine similarity between the image and the prompt "this is a photo of a" followed by the extracted label. Extracted labels are filtered by Local CLIP, not entire captions.
Reject Large Loss. Large Loss Matters is language-agnostic adaptive noise rejection and correction method. To test its vetting ability, we simulate five epochs of WSOD training and consider label targets with a loss exceeding the large loss threshold as "predicted to be visually absent" after the first epoch.
Accept Descriptive / Narrative. We train a logistic regression model to predict whether a VIST caption comes from the DII (descriptive) or SIS (narrative) split.
Reject Noun Mod. (Adj/Any). We reject labels that are noun modifiers ("car park"). The first noun modifier rule rejects an extracted label if the POS label is an adjective or is followed by a noun. The second rule rejects if the extracted label is not a noun.
Cap2Det. We reject a label if it is not predicted by the Cap2Det classifier.

Model Architecture

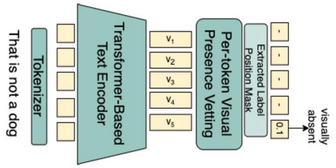

Figure 3. VEIL model architecture. A pretrained tokenizer parses the input caption into tokens which is passed to a transformer-based language model. Finally, the vetting and masking layer only predict visual presence of tokens corresponding to a label.

1. Use pretrained object detection models (VinVL C4 and YOLOv5) to get image-level object predictions
2. Use these predictions to create visual presence pseudo-labels for *each* extracted label
3. Split into train-test (8:2)

Extracted Label Vetting

Extracted Labels from Each Image	{boat}	{orange}	{cake}
No Vetting (Same as Above)	{boat}	{orange}	{cake}
Reject Noun Modifier	{boat}	{}	{cake}
LocalCLIP-E	{boat}	{orange}	{}
VEIL-SBUCapsCC	{}	{}	{}
VEIL-RedCaps	{}	{}	{}

Figure 6. Qualitative examples of extracted labels after vetting on RedCaps-Test. These are all label noise types defined in label noise taxonomy, and only VEIL-based methods is able to overcome these three commonly found noise types.

	Method	SBUCaps	RedCaps	CC	VIST	VIST-DII	VIST-SIS	COCO	AVG
	No Vetting	0.633	0.747	0.849	0.853	0.876	0.820	**0.973**	0.822
Image + Lang.	Global CLIP [27]	0.604	0.583	0.569	0.668	0.625	0.683	0.662	0.628
	Global CLIP - E [27]	0.594	0.569	0.534	0.654	0.613	0.660	0.640	0.609
Image	Local CLIP [27]	0.347	0.651	0.363	0.427	0.476	0.418	0.464	0.449
	Local CLIP - E [27]	0.760	0.840	0.597	0.759	0.695	0.812	0.788	0.750
	Reject Large Loss [18]	0.667	0.790	0.831	0.782	0.794	0.743	0.896	0.786
Lang.	Accept Descriptive	0.491	0.413	0.740	0.687	0.844	0.264	0.935	0.625
	Accept Narrative	0.470	0.645	0.383	0.487	0.154	0.757	0.143	0.434
	Reject Noun Mod. (Adj)	0.618	0.703	0.814	0.823	0.847	0.788	0.906	0.786
	Reject Noun Mod. (Any)	0.616	0.689	0.812	0.821	0.842	0.782	0.900	0.780
	Cap2Det [44]	0.639	0.758	0.846	0.826	0.854	0.774	0.964	0.809
	VEIL-Same Dataset	**0.809**	**0.890**	**0.909**	0.871	**0.892**	0.816	**0.973**	**0.884**
	VEIL-Cross Dataset	0.716	0.793	0.828	**0.875**	**0.892**	**0.830**	0.958	0.842

Table 2. Extracted Label Vetting F1 Performance. Visual presence ground truth is estimated by an object detection ensemble, X152-C4 [49] and YOLOv5-XL [17], on all datasets except for COCO, where we use existing annotations. **Bold** indicates the best performance in each column, and underlined denotes the second-best performance.

Impact of Vetted Labels on Weakly Supervised Object Detection:

Method (Train Data Size in Thousands)	VOC Det. mAP (Δ)	VOC Rec. mP (Δ)	COCO 0.5:0.95 mAP
No Vetting (19)	16.0 (-45%)	58.0 (-15%)	1.9
GT* (17)	8.2 (-72%)	77.8 (13%)	1.2
Large Loss [18] (19)	18.1 (-38%)	63.2 (-8%)	1.7
LocalCLIP-E [27] (18)	11.0 (-62%)	**81.9** (19%)	1.4
$VEIL_{ST}$-R,CC (18)	26.8 (-8%)	61.2 (-11%)	2.9
VEIL-SBUCaps (16)	**29.1** (–)	68.3 (–)	**3.1**

Table 3. Impact of vetting on WSOD performance on VOC-07 and COCO-14 datasets. There is a significant difference in detection and recognition on VOC-07 illustrated by Δ, relative performance change w.r.t. VEIL-SBUCaps on the same column. This highlights that VEIL variants filter out labels harmful to localization. (GT*) directly vets labels using the pretrained object detectors which were used to train VEIL.

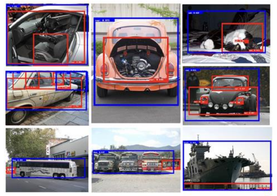

Figure 4. Part and contextual bias found in detections from LLM [18] WSOD model (red boxes) compared to full object localized with high confidence by WSOD model trained on data vetted by VEIL-SBUCaps (blue boxes). The categories shown are (clockwise from top-left): car, car, cat, car, boat, bus, bus, car.

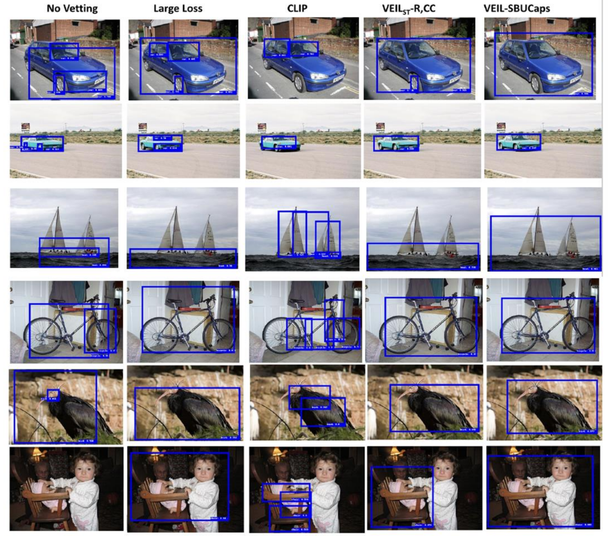

Figure 7. Detections (blue bounding box) from WSOD models trained with various vetting methods (top row) indicate that training with either VEIL filtering method leads to similar detection capability on VOC-07 [8], and show a strong part and contextual bias in Large Loss and No Vetting. The categories shown by row (from top to bottom) are: car, car, boat, bicycle, bird, chair.

See references in paper.