Cap2Det: Learning to Amplify Weak Caption Supervision for Object Detection

Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, Jesse Berent
Department of Computer Science, University of Pittsburgh, USA Google Research Zurich, Switzerland

Introduction

- Fully-supervised object detection requires instance-level annotations, which are labor-expensive

Supervised detection: bowl, bottle, person

WSOD: There are bowl, bottle, person in the image.

Free-form caption: A man is in a kitchen making pizzas.

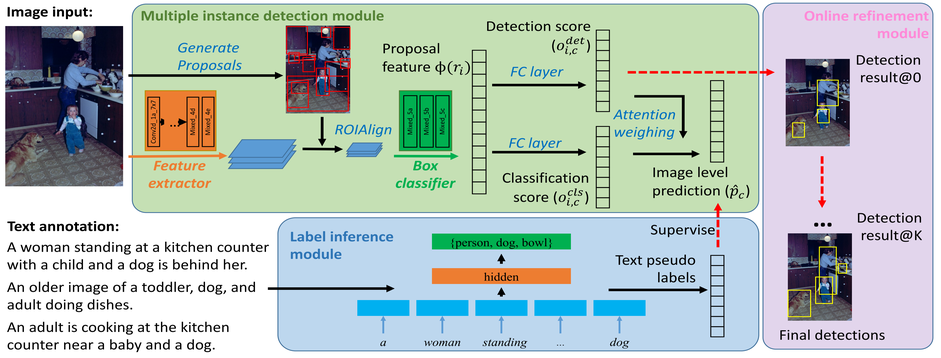
- Weakly-supervised object detection (WSOD) still requires an unnatural, crowdsourced environment
  - ➢ It requires only image-level annotations, which alleviates the burden to a certain extent
  - ➢ Its use of Multiple Instance Learning (MIL) requires precise labels, but in the wild, some objects in the image may not be mentioned
- Our proposed method utilizes free-form captions; these pose a challenge:

What humans mention vs What truly is in an image

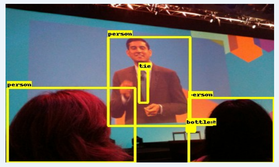
- Contributions
  - ➢ New task: Learning from noisy caption annotations
  - ➢ Benchmark and baseline: We show that predicting what truly is in an image (by training a robust text classifier) is a good way to mediate the reporting bias [Misra 2016], as compared to text matching

Method

- Use pseudo labels extracted from the free-form text as supervision



- Label inference module
  - ➢ It amplifies the supervision signal that captions provide, and squeezes more accurate information out of them
  - ➢ It performs basic reasoning based on the textual context



- *People watch a man **delivering a lecture** on a screen*
- *A large screen showing a person **wearing a suit***
- *An audience is looking at a film of a man talking that is projected onto a wall*

GROUNDTRUTH objects: person, tie, bottle

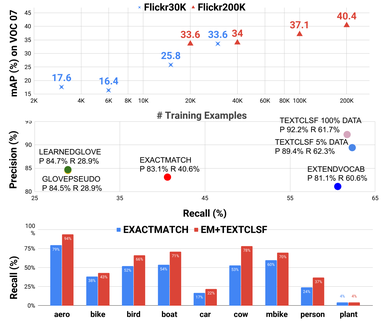
In this example, the object "tie" cannot be extracted using a lexical matching method, but it can be inferred through reasoning (ties are worn at formal events)

- Multiple instance detection module
  - ➢ It predicts detection / classification scores based on proposal features:
    - Detection score – weight of the *i*-th proposal for predicting class *c*
    - Classification score – probability that the proposal *i* belongs to class *c*
  - ➢ It aggregates image-level prediction using an attention mechanism
    - Attention: focus more on the regions with high detection scores
- Online refinement module [Tang 2017]
  - ➢ Iterative refining – previous instance predictions are used as ground-truth to supervise learning in the next iteration

Experiments

- Benchmark
  - ➢ Training on COCO (118,287 images, 591,435 captions)
  - ➢ Training on Flickr30K (31,783 images, 158,915 captions)
  - ➢ Evaluate on Pascal VOC and COCO (mAP@0.5)
- Baselines
  - ➢ GT-LABELS (upper bound): Using ground-truth labels
  - ➢ EXACTMATCH: Lexical matching method
  - ➢ EXTENDVOCAB: Using a manually constructed, hence expensive COCO synonym mapping dictionary
  - ➢ GLOVEPSEUDO: Assigning pseudo-labels based on word embedding distance
  - ➢ LEARNEDGLOVE: Same as the previous one, but we learn the word embedding based on an image-text ranking loss
  - ➢ TEXTCLSF: Using the label inference module trained on COCO

	Training on COCO		Training on Flickr30K	Training on Flickr200K
EVALUATE ON (mAP@.5)	VOC 07	COCO 17	VOC 07	VOC07
GT-LABEL	46.3	23.4	-	-
EXACT MATCH(EM)	39.9	19.7	31.0	-
EM + EXTENDVOCAB	42.5	19.4	29.3	-
EM + GLOVEPSEUDO	40.5	19.0	-	-
EM + LEARNEDGLOVE	41.7	19.7	-	-
EM + TEXTCLSF	**43.1**	**20.2**	**33.6**	**40.4**



Our label inference module generalizes well across domains

Our model benefits from larger-scale annotations from Flickr images

A good text classifier is achieved even with 5% of COCO caption-label annotations

Lexical matching cannot provide reliable supervision; it is still a challenge to recall non-mentioned classes such as "plant"

NSF Grant No. 1566270 Google Faculty Research Award NVIDIA hardware grant