Enhancing the Role of Context in Region-Word Alignment for Object Detection

Kyle Buettner[1], Adriana Kovashka[1,2]

[1]Intelligent Systems Program, [2]Department of Computer Science, University of Pittsburgh, PA, USA

buettnerk@pitt.edu, kovashka@cs.pitt.edu

University of Pittsburgh School of Computing and Information

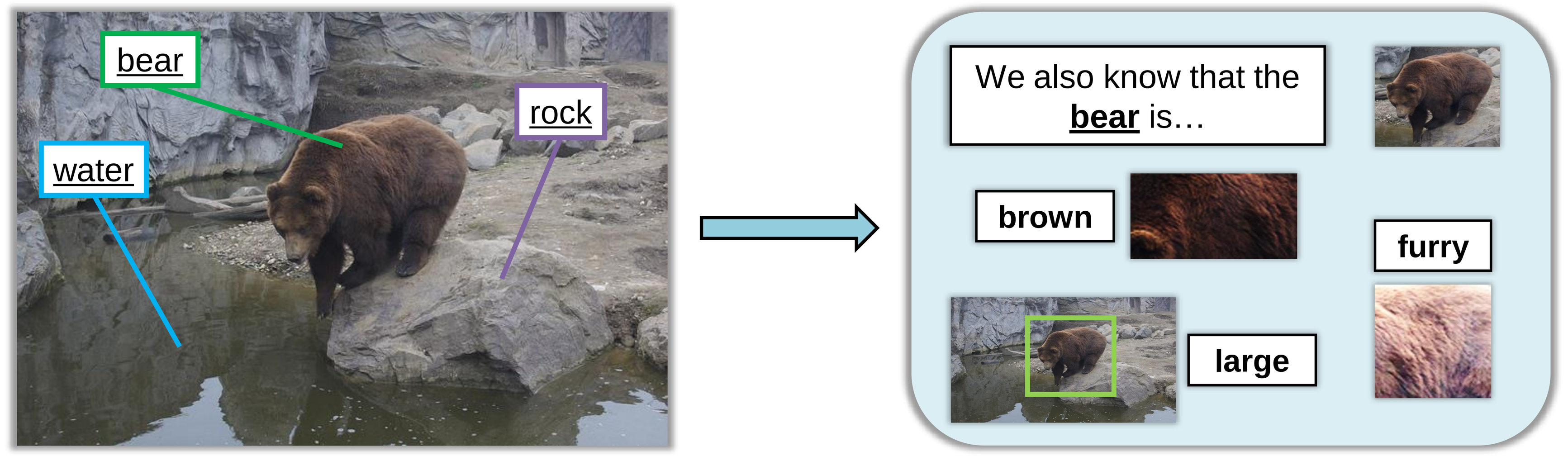
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Background and Motivation

- Learning vision-language alignment with contrastive learning and image-caption pairs has propelled **open-vocabulary** recognition and detection
- Object detectors trained with **region-word grounding** are typically evaluated with respect to how well **object nouns** are learned
- The impact and utility of other rich language context, especially **object attributes**, are underexplored

Example Context in Captions

a very **large furry brown** bear on a rock by the water.

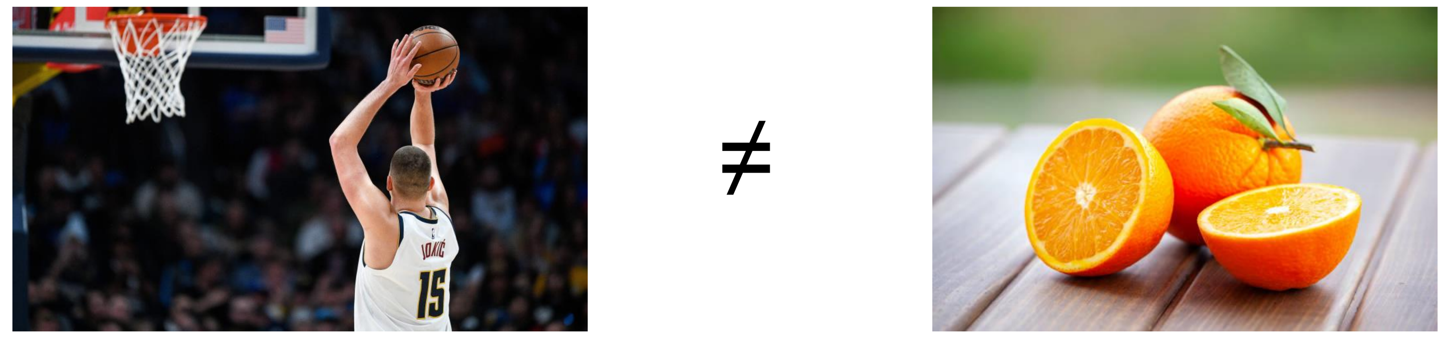


- **Research questions**
  - Does the existence of language context (**adjectives, verb phrases, prepositional phrases**) in vision-language pretraining help object detection?
  - How can object detection effectively leverage **contextualized word embeddings**?
  - Do learned object groundings capture **attribute meaning** from captions (*i.e. has the model learned what a red car is*)?
  - Can **contrastive negative caption sampling** be used as a method to enhance attribute sensitivity?
- To answer these questions, we conduct a case study of **OVR-CNN**, a region-word pretraining framework for open-vocabulary detection

Context Enhancement Strategies of Exploration

- **A contextualized grounding objective** to learn better alignment

He is shooting an orange basketball. **There are oranges on the table.**



- With a **training recipe** to maximize effectiveness **in detection**
  - Unfreezing the **language encoder in PT** and **vision-to-language projection layer in FT**
  - Using a contextualization **prompt** in class embeddings

- **Contrastive negative caption sampling** to add attribute sensitivity
  - When learning to match images to captions, for a given attribute-object pair, add two negatives, one with a **plausible adjective** (appearing with concept in dataset) and one with a **random noun**

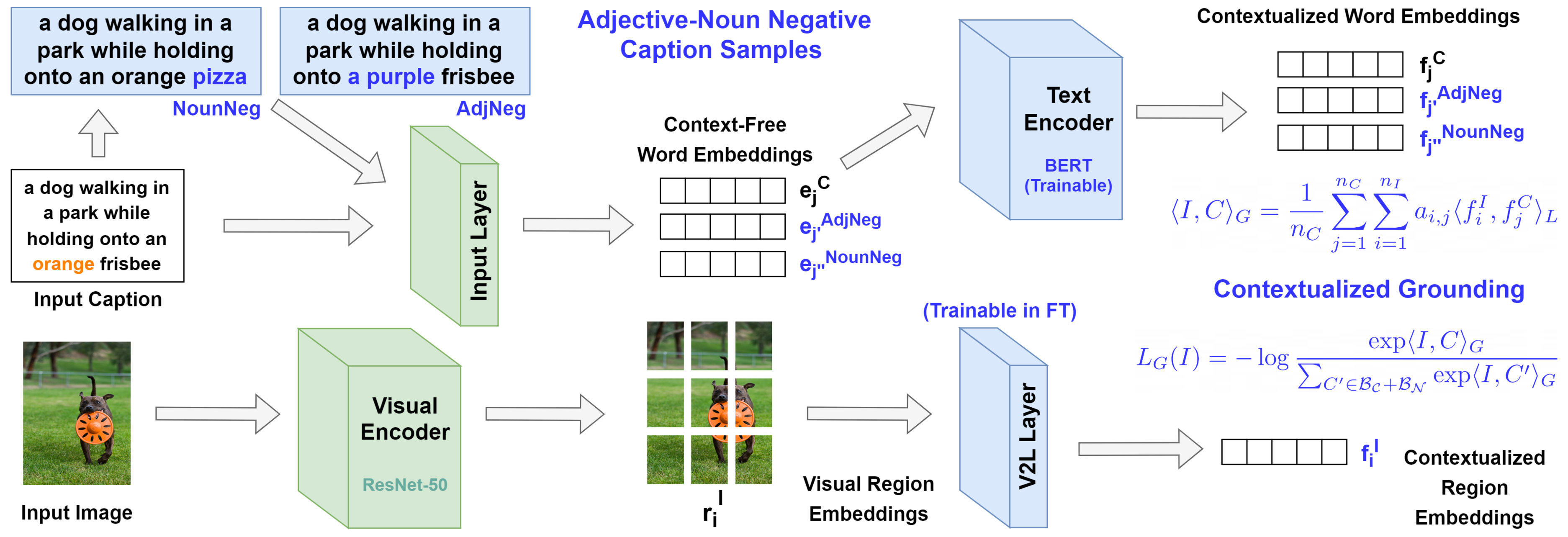
Caption: A red car is on the road.

Negatives Added to Batch:

A blue car is on the road.

A red animal is on the road.

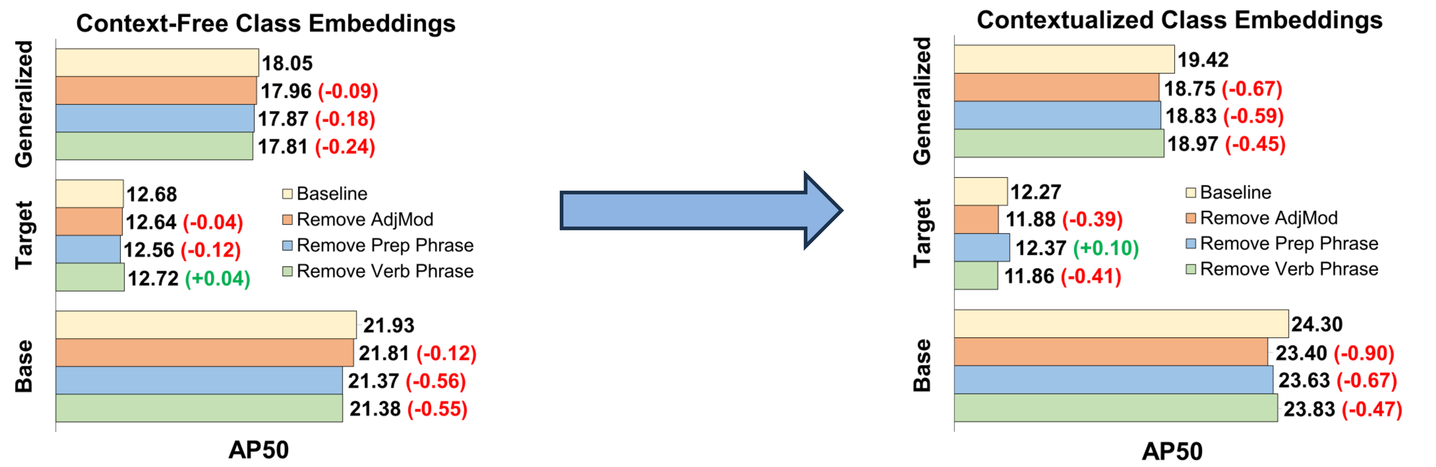
Methodology as Part of OVR-CNN Framework



Results and Analysis

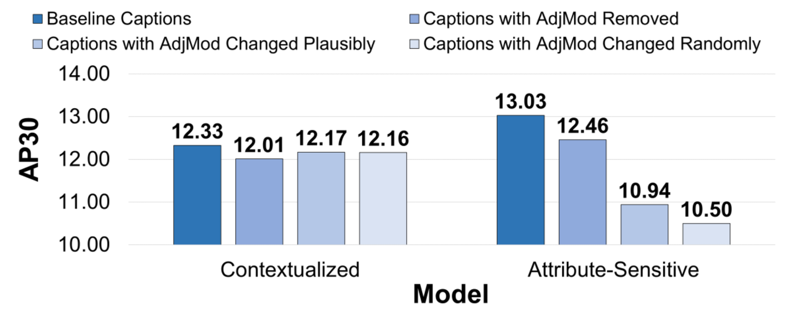
- Context is largely **ignored** in region-word pretraining for detection
  - Replacing **context-free** with **contextualized** embeddings in the grounding objective makes context more impactful

Detection AP on MiniCOCO



- Object alignment learned with contextualized word embeddings is not sensitive to **attribute meaning**
  - **Attribute negatives** teach model to learn attribute-object concepts

Unsupervised Phrase Grounding on COCO



- Context enhancement strategies are especially effective in **base** and **generalized** settings for open-vocabulary object detection

Open-Vocabulary Detection on COCO (3 trials)

Pretraining Method	Base-Only AP_{50}	Base-Only Δ	Target-Only AP_{50}	Target-Only Δ	Generalized All AP_{50}	Δ	Generalized Base AP_{50}	Δ	Generalized Target AP_{50}	Δ
Attribute-Sensitive OVR-CNN (our top method)	**35.81** ± 0.09	**+3.0**	17.68 ± 0.38	+1.9	**28.79** ± 0.17	**+2.5**	**33.94** ± 0.24	**+2.6**	14.24 ± 0.34	+2.4
w/o Plausible Adjective Negative (noun neg. only)	35.25 ± 0.19	+2.5	17.79 ± 0.18	+2.0	28.33 ± 0.12	+2.1	33.31 ± 0.13	+2.0	14.24 ± 0.13	+2.4
w/o Random Noun Negative (context only)	35.18 ± 0.13	+2.4	16.67 ± 0.26	+0.9	28.26 ± 0.20	+2.0	33.62 ± 0.16	+2.3	13.12 ± 0.30	+1.3
w/o Contextualized Embeddings (best context-free)	34.08 ± 0.01	+1.3	**19.09** ± 0.72	**+3.3**	28.28 ± 0.27	+2.0	33.19 ± 0.12	+1.8	**14.42** ± 0.70	**+2.6**
w/o BERT/V2L Training (original OVR-CNN) [35]	32.78 ± 0.08	–	15.80 ± 0.11	–	26.25 ± 0.04	–	31.36 ± 0.15	–	11.82 ± 0.28	–

Conclusion

- We illustrate strategies to effectively use context for detection (contextualized grounding/adjective-noun negative sampling)
- Future work may consider methods to improve target performance or better leverage object relations and actions for detection

References

- Zareian, Alireza, et al. "Open-vocabulary object detection using captions." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.