Improving language-supervised object detection with linguistic structure analysis

Arushi Rai and Adriana Kovashka

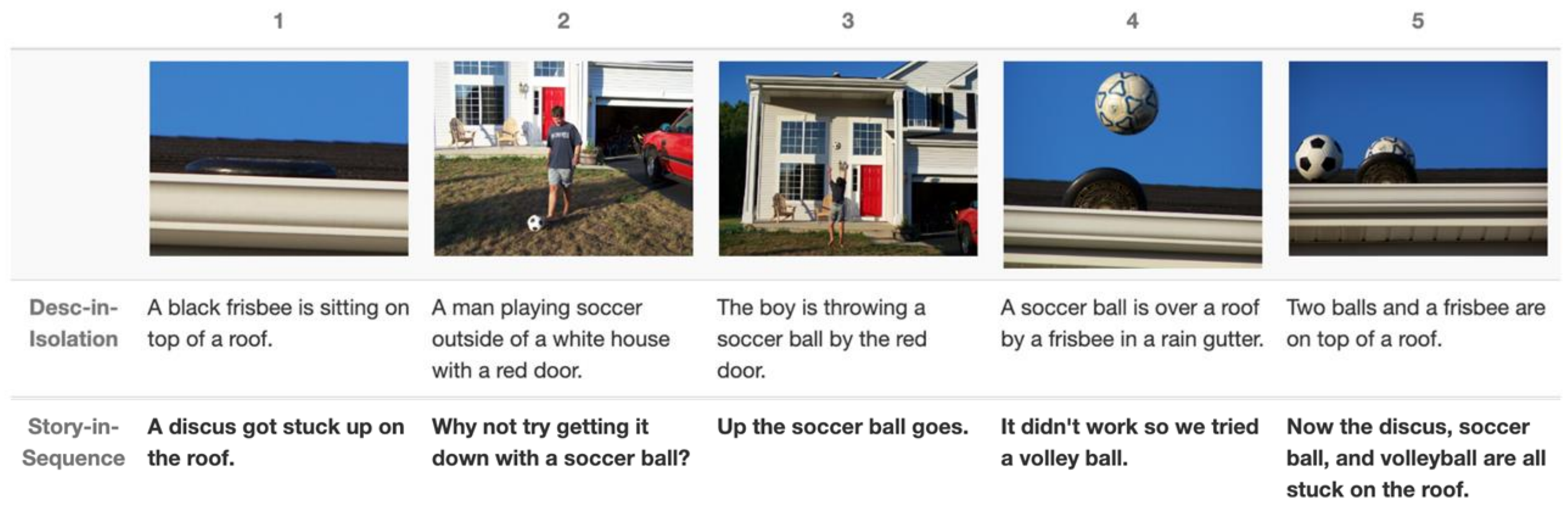
University of Pittsburgh School of Computing and Information

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction

- To overcome annotation cost for object detection, image-level labels extracted from user-uploaded captions can weakly supervise an object detection model [2].
- Previous Work: Focused on extracting labels from *descriptive* captions [5] found in COCO and Flickr30K
- Descriptive captions aim to describe all visible elements within an image. In contrast, abundant user-uploaded captions may extend beyond the temporal boundaries of the image, creating **narratives** that encompass a larger time frame.
- However, this source can extract mentions of objects not visible in image.
- Our Work: Find impact of extracting labels from narrative captions for WSOD and comparing label/caption selection strategies for WSOD.

Identifying Differences in Descriptive and Narrative Captions in the Visual Story Telling Dataset (VIST) [3]:



What's in a Caption?

DII or descriptive captions contain more nouns, prepositions, and adjectives.

Past and present tense within the same sentence is twice as prevalent in narrative captions.

Example:
"Afterwards, we take a couple photographs because we paid the photographer to do so."
Implication: An aligned image would either show the couple posing for a photo or the transaction.

Analysis of Caption Structure using RST

RST provides a taxonomy to define the relationship between spans of text.
"Employees are urged to complete new beneficiary designation forms for retirement or life insurance benefits whenever there is a change in marital or family status"

(b) Distribution of RST tags (top 8).

RST Tag	DII (%)	SIS (%)
Attribution	1.3	**4.8**
Background	1.3	**2.2**
Contrast	1.1	**1.7**
Enablement	0.6	**2.3**
Joint	4.1	**6.0**
Temporal	**2.2**	1.2
Elaboration	**21.4**	10.9
None	**65.1**	64.1

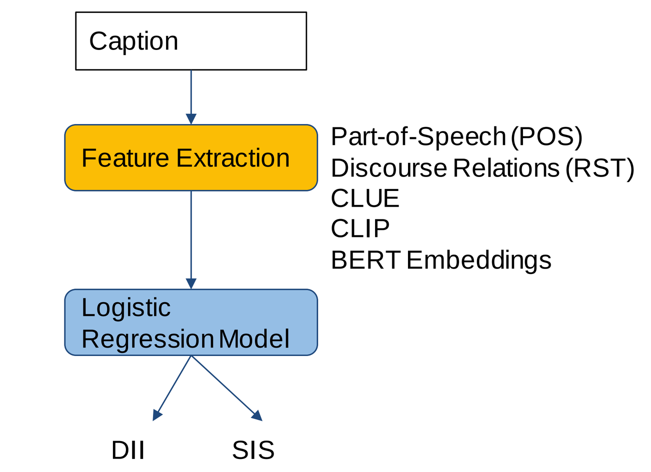
Lexical differences are important; "before" and "while" frequently occur in temporally flagged SIS captions and only "while" frequently occurs in the DII. Labels extracted from Temporal-SIS captions could contain either a future or past reference in either nucleus or satellite clauses, and therefore are not currently visible.

Relationships between a caption and its corresponding image?

CLIP [4]: DII caption-image embeddings have statistically significant ($p < 0.0001$) higher (0.30 ± 0.03) cosine similarity than SIS (0.26 ± 0.04).

CLUE [1] Discourse Relations: DII captions have slightly more "Visible, Action" (0.5%) and "Story" (0.3%) tags. However, 50% of SIS captions have no CLUE prediction.

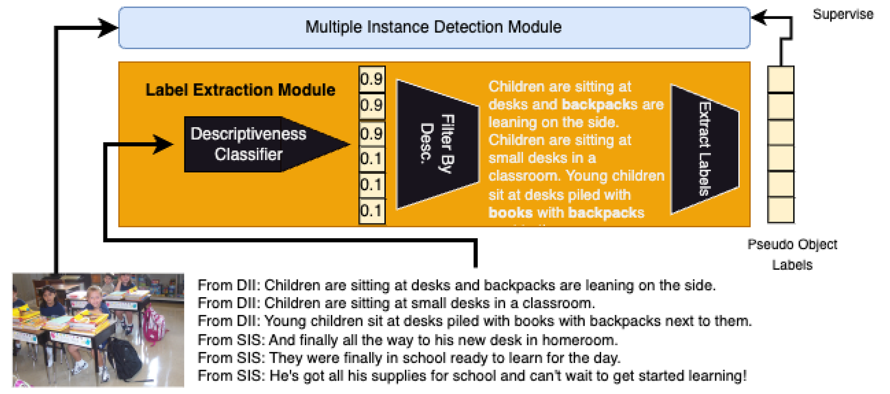
Training a Binary Descriptiveness Classifier



Linguistic features	PREC	REC
POS	**0.8823**	**0.8782**
RST	0.6281	0.5170
CLUE	0.6813	0.6411
CLIP [26]	0.7215	0.7170
POS+RST	0.8838	0.8787
POS+CLUE	0.8911	0.8879
POS+RST+CLUE	0.8916	0.8884
CLIP+POS	0.8925	0.8893
CLIP+RST	0.7360	0.7312
CLIP+CLUE	0.7523	0.7480
CLIP+POS+RST+CLUE	**0.8982**	**0.8960**
BERT	**0.9570**	**0.9560**

Table 4. We evaluate precision/recall of DII/SIS classifiers on a VIST holdout.

Impact on WSOD #1: Selection Based on Descriptiveness (Global View)



Descriptive Classifier	DII	Random	SIS
GT Labels	**0.0195**	0.0105	0.0050
POS	**0.0304**	0.0105	0.0047
POS+RST	**0.0187**	0.0105	0.0046
POS+RST+CLUE	**0.0187**	0.0105	0.0038
CLIP	**0.0173**	0.0105	0.0043

Table 5. Effect of descriptive classifiers for filtering on WSOD. We evaluate mAP performance of WSOD on VOC-07 [9]. The random classifier (trained once) is independent of descriptiveness classifiers. Bold indicates best performance in row.

Analysis of Visually Absent Extracted Labels (VAEL)

Type of VAEL	Definition
Atypical Instance	Object is present in an atypical form (e.g. clay model, toy)
Inside/On Top Of	Image is taken inside or on top of the object
Occluded	Full view of object is limited because object is occluded
Part of Phrase (Rel)	Extracted label is part of a related phrase ('car' in 'car show')
Part of Phrase (Unrel)	Extracted label is part of unrelated phrase ('dog' in 'hot dog')
Missing From Scene (Narrative Artifact)	Object completely missing from the scene but was mentioned in the one of the captions to further the story told in the caption, e.g. "She was returning from the **car** when she pets the dog"
Missing (Other)	Missing from scene for none of the reasons described above

Table 3. Types of visually absent extracted labels.

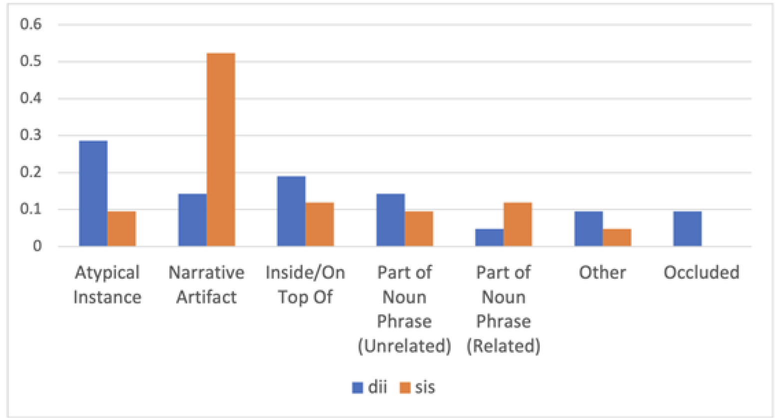


Figure 2. The distribution of the type of VAEL. VAEL from SIS are mainly comprised of narrative artifacts (53%) compared to DII (14%).

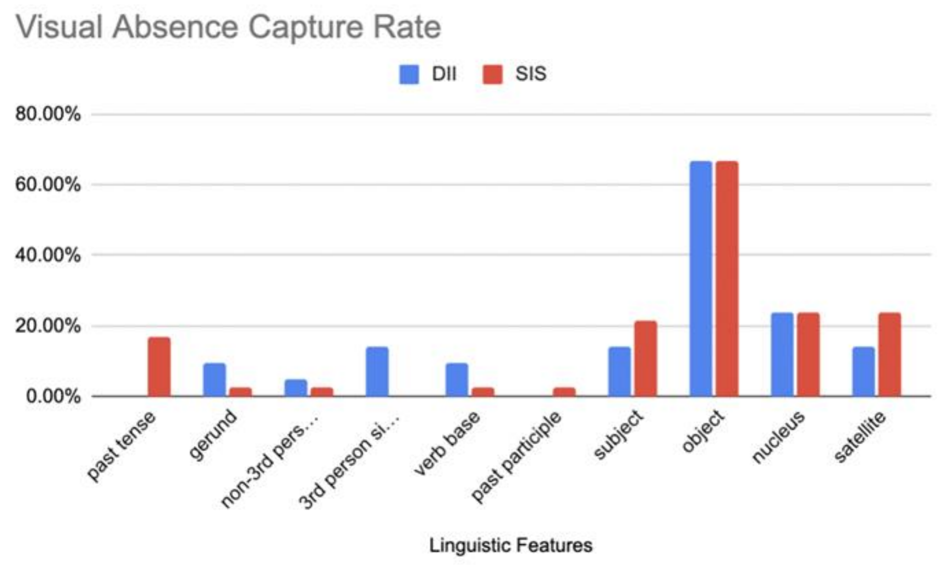
Takeaways:

- SIS contains mainly **narrative artifacts**
- The noise in DII captions comes from **atypical objects**, **occlusion**, and the photo captured while **inside or on top** of the object
- Both styles suffer from extracting labels from **noun phrases**

Impact on WSOD #2: Proximity to Linguistic Features for Extracted Label Selection (Local View)

In this experiment, labels are extracted from windows centered around particular verb tenses and other linguistic features (subject/object)

Example:
Linguistic Feature: Verb Gerund
Caption: "an image of a person smiling at a party"
Window Centered around "Smiling" (Gerund): "a person smiling at a"
Extracted Label: {person}



Using a small annotated set, we evaluate how many visual absent extracted labels are captured when extracted from windows centered around each linguistic feature.

Takeaway: Most visually absent labels appear to be objects (both) and found in the nucleus (DII).

Verb	$D^* \cup S^*$ (Count)	D^* (Count)	S^* (Count)
baseline	0.0047 (583)	0.0026 (456)	0.0024 (145)
verb base	0.0014 (128)	0.0010 (42)	0.0014 (86)
past tense	0.0031 (461)	0.0015 (83)	**0.0033** (403)
gerund	**0.0053** (976)	**0.0046** (909)	0.0014 (136)
non-3rd person sing pres	0.0013 (94)	0.0012 (77)	0.0011 (18)
3rd person sing pres	**0.0068** (985)	**0.0067** (888)	0.0016 (193)
past participle	0.0028 (332)	**0.0028** (270)	0.0014 (100)

Table 7. mAP on VOC-07 [9]. Bold signifies mAP higher than baseline (top row).

References

[1] Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. Cross-modal coherence modeling for caption generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6525–6535, 2020.

[2] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2846–2854, 2016.

[3] Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016), 2016.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In ICML, 2021.

[5] Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Cap2det: Learning to amplify weak caption supervision for object detection.