

Michele Cafagna¹ Kees van Deemter² Albert Gatt^{1, 2}

¹University of Malta, Institute of Linguistics and Language Technology, Malta

²Universiteit Utrecht, Information and Computing Sciences, The Netherlands

Motivation

Current captioning datasets focus on **object-centric** captions, describing the visible objects in the image, e.g. "people eating food in a park".


Image	Axis	Caption
	scene	the picture is shot in a ski resort
	action	they are just relaxing after a round of skiing
	rationale	they want to have a good time together
	object-centric (COCO)	a woman and a boy sitting in the snow outside of a cabin.

Figure 1. Example of High-Level captions. It is shown one of the three captions available for the three axes collected: *scene*, *action*, *rationale*, combined with the object-centric captions from COCO.

However, people often describe images based on the type of scene they depict and the actions they perform ('people having a picnic). **Such descriptions draw on personal experience and world knowledge.** They are grounded in the image at a higher level than object-centric captions

Connecting High- and Low-level Descriptions

HL Dataset allows us to identify concrete objects (mentioned in COCO (Lin et al. 2014) captions) in images that provide 'support' to infer high-level descriptions such as scenes, actions, and rationales.

- We collect the captions by asking the participants to answer three questions (**What, Why, Where**).
- We explicitly ask to **rely on their personal interpretation** of the scene
- We collect confidence scores by asking **independent participants to score the likelihood of a high-level description** given the image and the corresponding question on a Likert scale from 1 to 5.

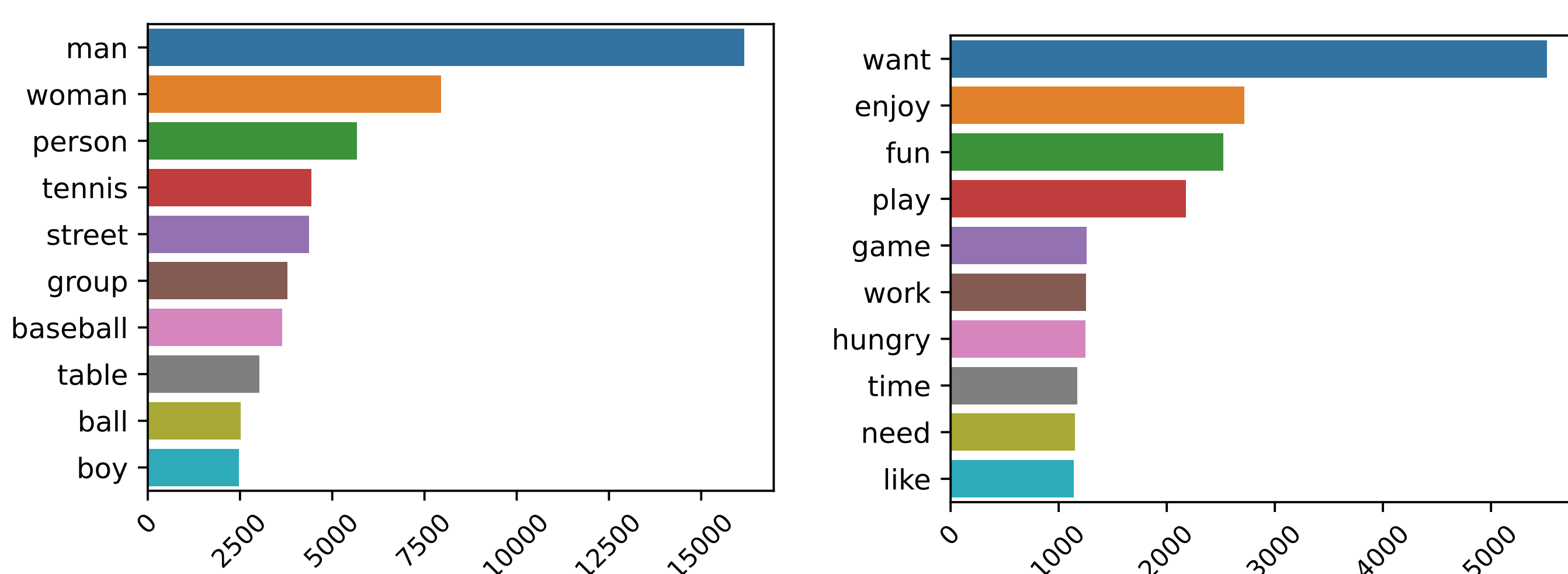


Figure 2. Most frequent lemmas in the COCO captions (left) and in the *rationale* axis (right).

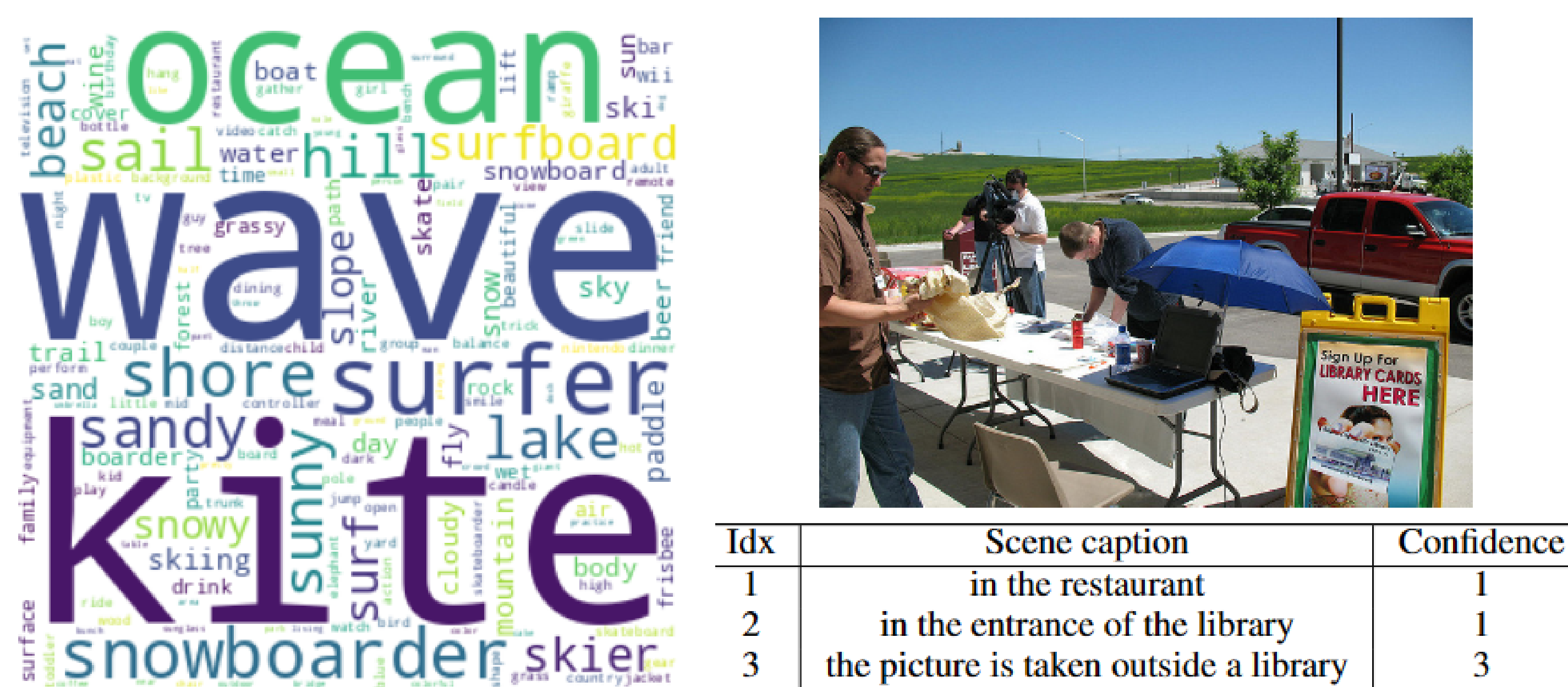


Figure 3. (Left) Most informative objects for the word *enjoy* in the *rationale* axis. (Right) An example of a 'hard' sample in the HL dataset where the scene captions have low confidence scores.

High-level Caption Generation

Model	Axis	Cider	SBLEU	Rouge-L
GIT	action	110.63	15.21	30.43
	rationale	42.58	5.90	18.57
	scene	103.00	24.67	33.92
BLIP	action	123.07	17.16	32.16
	rationale	46.11	6.21	19.74
	scene	116.70	26.46	35.30
ClipCap	action	176.54	27.37	39.15
	rationale	78.04	11.71	25.76
	scene	145.93	36.73	42.83



Scene	the picture is taken in a field.
Action	they are riding a horse
Rationale	they want to win the match.

Figure 4. (Left) Automatic metrics for baselines (GIT, BLIP, and ClipCap) fine-tuned along the three axes (scene, action, and rationales) of the HL dataset. (Right) High-level captions generated by axis-wise fine-tuned ClipCap models.

Narrative Generation

We extend the dataset to combine the three axes to compose a short 'narrative', which describes the scene, action and rationale in tandem. We call this new dataset **HL Narratives**.



GIT (PRE): two girls looking at their cell phones
GIT (FT): they are reading a text message outside on the street, waiting for their friend.

Figure 5. Comparison between the object-centric captions generated by GIT pre-trained (PRE) and the high-level caption generated by the fine-tuned (FT) model. The generated high-level caption embeds high-level information regarding action, rationale, and scene, depicted in the visual content

We use T5 fine-tuned on paraphrase generation to generate the data; We fine-tune three baselines on the synthetic dataset.

Findings and Contributions

Our main findings and contributions are:

- 1 We release the HL Dataset, a new VL resource, grounding high-level captions in images along three axes, aligned with existing object-centric captions;
- 2 We describe the collection protocol and provide an in-depth analysis of the data;
- 3 We present baselines for the High-Level Captioning task and describe further potential uses for our data

