Egoshots, An Ego-Vision Life-Logging Dataset And Semantic Fidelity Metric To Evaluate Diversity In Image Captioning Models

Machine Learning In Real Life ICLR 2020 Workshop

Pranav Agarwal*   Alejandro Betancourt   Vana Panagiotou   Natalia Díaz-Rodríguez**

*pranav2109@hotmail.com   **natalia.diaz@ensta-paris.fr

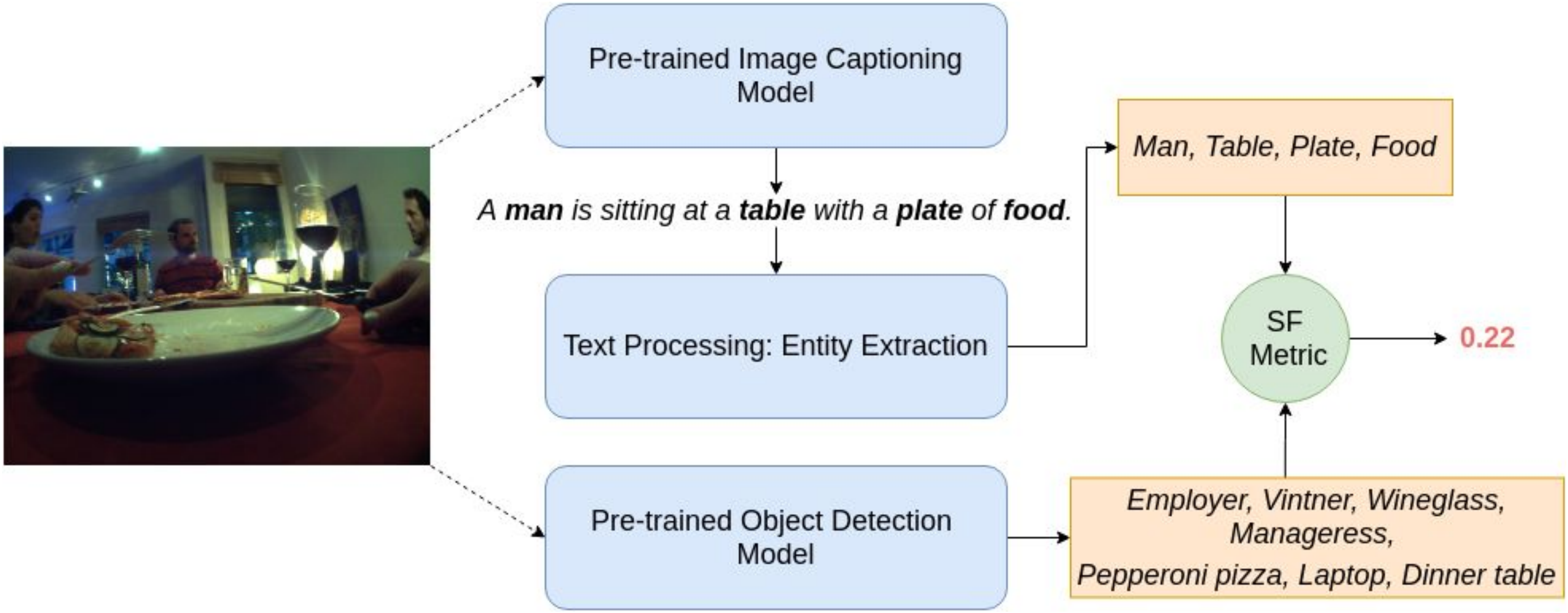
Image captioning in the wild

A man holding a child in a park with a kite

Egoshots Dataset: a lifelogging ego-vision 2 month dataset

- 978 images with captions predicted using pre-trained models.
- Images are collected by 2 female PhD and Master students for one month each.
- *Autographer camera is used* (takes photos automatically based on interestingness):
  - Images range from indoor to outdoor scenes.
  - Day to day activities: biking, socializing, office work...

Low-cost Dataset Annotation Pipeline



Semantic Fidelity Metric

$$SF_i = s_i \cdot \frac{\#N}{\#O}$$

- s_i: Semantic similarity
- *#N:* Count of Nouns in caption generated by image captioning model
- *#O:* Count of Objects detected by object detector
- SF_i and s_i: in [0, 1]

- Each word is mapped to their corresponding *embedding*.
- Average embedding for all the nouns and average embedding for objects are calculated.
- s_i: *cosine similarity* between the noun and the objects embeddings.
- Ratio of count of nouns and objects: penalizes the caption for the objects not mentioned in the caption

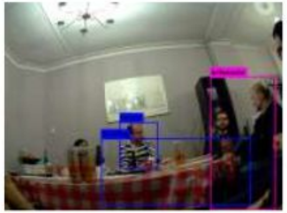
Assessing different captioning models through Semantic Fidelity

Image Captioning Method	**S**-V	**S**-Co	**Y3**-V	**Y3**-Co	**C**-V	**C**-Co	**Y9**
Show Attend And Tell	0.35	0.34	0.34	0.33	0.30	0.36	**0.28**
Novel Object Captioning at Scale	0.40	0.39	0.39	0.37	0.34	0.40	**0.33**
Decoupled Novel Object Captioner	0.41	0.41	0.40	0.39	0.35	0.44	**0.32**

Table 1: Mean Semantic Fidelity of different image captioning models using various object detectors: S: *SSD* (Liu et al., 2016), Y3: *YOLOv3* (Redmon & Farhadi, 2018), C: *Center Net* (Duan et al., 2019), Y9: *YOLO9000* trained on *ImageNet* and *COCO*, V: trained on *VOC*, Co: trained on *COCO*.

- Object detectors mostly use MSCOCO (80 classes) or PASCAL-VOC (20 classes) datasets.
- YOLO-9000 is trained for 9000 classes.
- Low SF for Y9 reflect its ability to better penalize the caption.
- Diverse and robust object detectors make SF more reliable.

Examples



YOLO9000	Model	Caption	SF
[Panelist, Ambassador, Furnishing]	**SAT**	*A man is standing in front of a television.*	**0.31**
	NOC	*A man in a kitchen with a large mirror.*	0.22
	DNOC	*A man in a kitchen with a bottle.*	0.19

YOLO9000	Model	Caption	SF
[Entrepreneur, Wineglass, Vintner, Dinner table]	**SAT**	*A group of people sitting at a table with wine glasses.*	0.36
	NOC	*A group of people sitting at a table with food.*	0.27
	DNOC	*A man and woman sitting at a table with food.*	**0.38**

YOLO9000	Model	Caption	SF
[Entrepreneur, Background, Laptop, Camp Chair, Settler]	**SAT**	*A man sitting at a table with a laptop.*	0.42
	NOC	*A man in a kitchen with a large display of food..*	0.44
	DNOC	*A man in a suit and tv standing in front of a tv.*	**0.62**

Future Work

- Extending SF to include other syntactic elements.
- Dependency on using robust object detectors matching human level accuracy.
- SF: the only metric able to rank image captionings in the wild when no labels are available
- Using SF to improve captions for different applications such as life-logging by the blind, autonomous driving or telepresence robotics.