DetermiNet: A Large-Scale Diagnostic Dataset for Complex Visually-Grounded Referencing using Determiners

Clarence Lee[1], M Ganesh Kumar[2,3], Cheston Tan[3]
Design and Artificial Intelligence, SUTD[1], School of Engineering and Applied Sciences, Harvard University[2], Centre for Frontier AI Research, A*STAR[3]

Scan me to:
- Access preprint
- Leave your comments
- Connect

ICCV23 PARIS

Summary

Definition: A **determiner** is an English part-of-speech that quantifies or references the noun following it. (For instance, "my apple" vs "your apple" and "some apples" vs "all apples").

Motivation: Determiners are important word classes to increase the accuracy of reference e.g. Human-Robot Collaboration, ("pass me my screwdriver and some screws", "those screws are faulty, but these are fine"). These concepts need to be learned rather than hardcoded as the referencing of determiners changes according to the context.

Problem: In current datasets, coverage of determiners are limited and the semantics of determiners are not fully captured. Visual language models also fail to learn determiner semantics (see Fig. 3)

Contribution: We created DetermiNet, a visuo-linguistic dataset covering 25 determiners and all 4 determiner classes (Articles, Demonstratives, Possessives, Quantifiers) comprising of 250,000 samples (10,000 image-caption pairs per determiner).

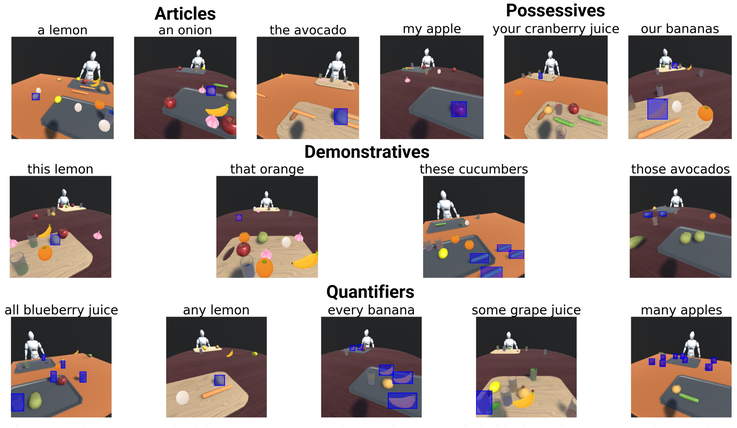
Task: Given an image and caption, predict *N* number of bounding boxes to correctly identify the object referenced and quantified by the determiner as defined in Fig. 1.

Table 1. Comparison of datasets for referring expressions. A, P, D, Q, Exo and Ego stand for Articles, Possessives, Demonstratives, Quantifiers, Exocentric and Egocentric respectively

Datasets	A	P	D	Q	View	Images	Type
RefCOCO [1]	Y	N	N	N	Exo	19,994	Real
RefCOCO+ [1]	Y	N	N	N	Exo	19,992	Real
RefCOCOg [2]	Y	N	N	N	Exo	26,711	Real
CLEVR-Ref+ [3]	Y	N	N	N	Exo	99,992	Synth
YouRefIt [4]	Y	N	N	N	Exo	497,348	Real
DetermiNet	Y	**Y**	**Y**	**Y**	**Ego**	250,000	Synth

Examples Per Determiner Classes

Determiner classes:
- Articles: identify nouns which the speaker is referring to (a, an, the)
- Possessives: signify ownership of the noun (my, your, our)
- Demonstratives: isolate nouns that are being referred to (this, that, these, those)
- Quantifiers: describe the amount of the referred noun (each, neither, every, either, half, several, few, many, both, some, no, all, any, little, much)



Organisation Of Determiners

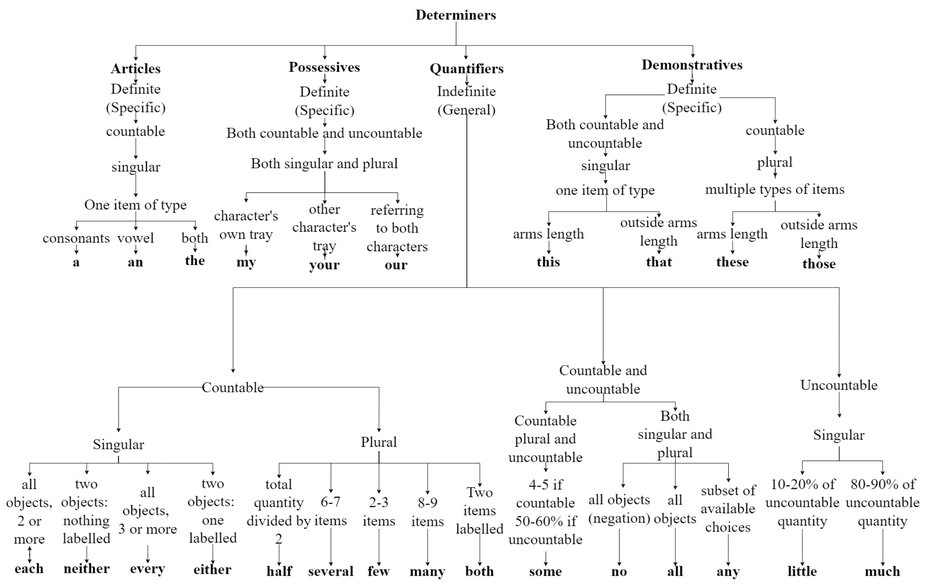


Figure 1. Organization and characteristics of the 25 determiners in DetermiNet.
There are 44 determiners in the English corpus of which DetermiNet covers 25. We omit gender-specific (e.g. his, her), comparison (e.g. more, most, lesser etc.) and interrogative determiners (which, what, whose, whichever).

Evaluation On DetermiNet

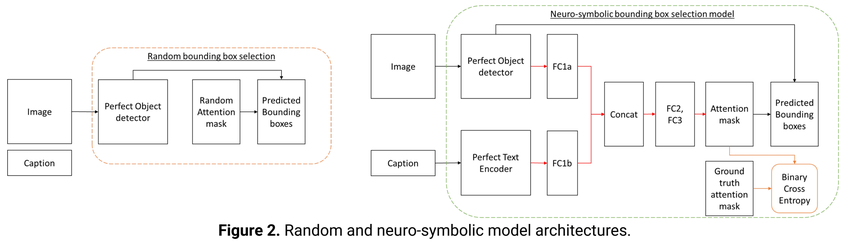


Figure 2. Random and neuro-symbolic model architectures.

Table 1. Model performance (AP@IoU=0.5:0.95). Right column indicates model predictions constrained to single bbox

Models	AP(multiple bbox)	AP(single bbox)
Random	9.8	1.6
Neuro-Symbolic	93.5	34.7
OFA [5]	-	20.6
GLIP [6]	55.0	14.3
MDETR [7]	70.6	29.7

Table 2. Ablation study with masked captions. Performance reported AP@IoU=0.5:0.95

Ablation condition	Oracle	MDETR
Noun+ / Det+	93.5	70.6
Noun+ / Det -	71.3	56.3
Noun - / Det+	11.3	11.3
Noun - / Det -	9.8	0.2

Embedding Analysis

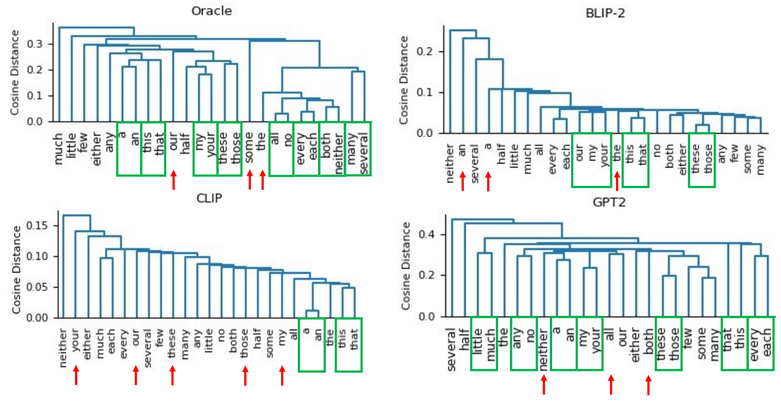


Figure 3. Cosine distance of the 25 determiner embeddings from the text encoders of the Oracle, CLIP, BLIP-2 and GPT-2 models. Pretrained text encoders of VLMs do not show determiner organization like the Oracle.

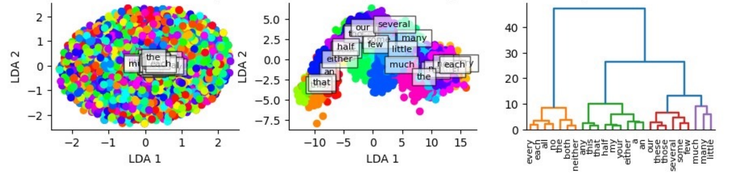


Figure 4. Oracle learns the DetermiNet organization, represented as LDA clusters and a dendrogram.

Evaluation On Real Dataset

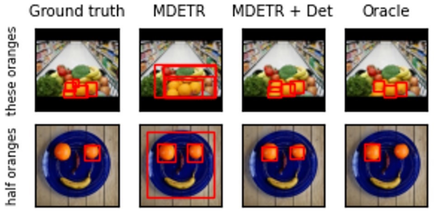


Figure 5. MDETR and Oracle prediction on real images from COCO dataset

Table 3. Oracle & MDETR performance (AP@IoU=0.5:0.95) on 100 real images.

Models	AP (multiple bbox)
Oracle	78.1
MDETR	10.4
MDETR + DetermiNet	19.5

References

[1] Kazemzadeh et al. (2014) Referitgame: Referring to objects in photographs of natural scenes
[2] Mao et al. (2016) Generation and comprehension of unambiguous object descriptions
[3] Liu et al. (2019) CLEVR-ref+: Diagnosing visual reasoning with referring expressions
[4] Chen et al. (2021) YouRefIt: Embodied reference understanding with language and gesture
[5] Wang et al. (2022) OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework
[6] Li et al. (2022) Grounded language-image pre-training
[7] Kamath et al.(2021) MDETR -- modulated detection for end-to-end multi-modal understanding