# Context matters: evaluation of target and context features on variation in object naming


**Nikolai Ilinykh and Simon Dobnik**

Department of Philosophy, Linguistics and Theory of Science
Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
{name.surname}@gu.se


GÖTEBORGS UNIVERSITET

## 1 Aims

- Explore feature representations for the object naming task
- Investigate if target object and/or context representations encoded with either linguistic or visual information or both capture variation and uncertainty in human object naming

## 2 Question

- What set of features enables a computational model of object naming most closely capture variation in human object naming?

## 3 Task formulation

Dataset: Many Names (Silberer et al., 2020)

Example image with the target object in the red box

Frequencies of names with which humans described the target object:

`car`: 32,
`vehicle`: 2,
`automobile`: 1

Task: given a feature representing either the target or context objects, predict the most likely name for the target object

## 4 Model and input features

We use CLIP (Radford et al., 2021) to encode both linguistic and visual features.

Our model is a simple linear classifier, which takes input features and predicts a single name.

$$\hat{\mathbf{y}} = \sigma\left( \left( \mathrm{f}_2 \left( \mathrm{f1} \left( \mathbf{x} \right) \right) \right) \right), \tag{1}$$

where

$$\mathrm{f}_1(\mathbf{x}) = \mathrm{ReLU}\left( \mathrm{BN}\left( \mathbf{W}_1\, \mathbf{x} + \mathbf{b}_1 \right) \right), \tag{2}$$

$$\mathrm{f}_2(\mathbf{x}') = \mathrm{Dropout}\left( \mathbf{W}_w\, \mathbf{x}' + \mathbf{b}_2 \right) \tag{3}$$

**Input features:**
**`Target`, `Context-Obj`, `Context-Scene`**

Vision

Language
from Visual Genome (Krishna et al., 2017)

sedan

black car, big van, ...

man on the street, car next to the street, ...

## 5 Results

| | Condition | Mode | Accuracy (%) ↑ @1 | @5 | @10 | AMR ↓ | PP ↓ | H ↓ | $\rho$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | TEXT | 69.15 | 87.68 | 89.94 | 41.45 | 4.745 | 0.210 | 0.540* |
| 2 | `Target` | VISION | 56.70 | 81.09 | 86.34 | 52.87 | 7.199 | 0.266 | 0.485* |
| 3 | | VISION-TEXT | 70.02 | 90.99 | 92.30 | 33.77 | 3.740 | 0.178 | 0.574* |
| 4 | | TEXT | 40.90 | 67.58 | 76.73 | 52.13 | 14.924 | 0.365 | 0.343* |
| 5 | `Context-Obj` | VISION | 49.14 | 75.14 | 83.20 | 40.79 | 10.360 | 0.315 | 0.328* |
| 6 | | VISION-TEXT | 46.48 | 72.98 | 81.04 | 45.87 | 11.531 | 0.330 | 0.321* |
| 7 | | TEXT | 4.09 | 16.85 | 31.80 | 59.00 | 51.111 | 0.531 | -0.024 |
| 8 | `Context-Scene` | VISION | 47.93 | 73.51 | 81.42 | 60.73 | 9.116 | 0.298 | 0.410* |
| 9 | | VISION-TEXT | 53.34 | 77.91 | 83.98 | 38.87 | 8.281 | 0.285 | 0.424* |
| | `Human` | | | | | | **1.623** | **0.065** | **1.000** |

## 6 Conclusions

- Both language and vision contribute to human-like object naming variation
- Knowledge about the target object (specifically, background knowledge) is highly relevant
- Context representations matter: full image content is better than object-level segmented contexts (for more, check the paper)

## References


Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Carina Silberer, Sina Zarrieß, Matthijs Westera, and Gemma Boleda. 2020. Humans meet models on object naming: A new dataset and analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1893–1905, Barcelona, Spain (Online). International Committee on Computational Linguistics.




CLASP