Model Card - Multimodal Clustering Model

Model Details

- Developed by Uri Berger[a b], Gabriel Stanovsky[a], Omri Abend[a] and Lea Frermann[b]
- Published May 2022
- A visual ResNet50 encoder and a simple probabilistic text encoder mutually trained to predict matching clusters for $(image, caption)$ pairs
- For more details, we refer to our paper [Berger et al., 2022]
- For questions contact uri.berger2@mail.huji.ac.il

Intended Use

- This model can be utilized in one of several use cases: Word categorization, Word concreteness estimation, and Image zero-shot classification and segmentation

Factors

- While the model is not human-centric, it is sensitive to biases embedded in the caption annotators and therefore factors highly depend on the training data. For example, in MSCOCO [Lin et al., 2014], gender is a relevant factor [see Zhao et al., 2017]

Metrics

- The model can be measured for taxonomic or syntagmatic word categorization. Words with taxonomic relations are words that can occur in similar contexts (e.g., *dog*, *giraffe*, *elephant*). Words with syntagmatic relations are words that are likely to occur together in the same sentence (e.g., *dog*, *bark*, *collar*)
- For taxonomic categorization, we use F-Score, which is the harmonic mean of precision and recall of the model's produced clusters when compared to ground-truth categories
- For syntagmatic categorization we use mean association strength computed across all word pairs in which both words were assigned the same cluster by this clustering solution. The association strength of a pair of words is extracted from the Small World of Words (SWOW) dataset [De Deyne et al., 2019]
- For concreteness estimation we compute the pearson correlation coefficient of the estimation of all words with ground-truth concreteness values
- For image classification and segmentation we use F-Score (the harmonic mean of precision and recall)

Evaluation Data

- Taxonomic categorization is evaluated against the categorization dataset by Fountain and Lapata [2010], transformed into hard categories by assigning each noun to its most typical category as extrapolated from human typicality ratings
- Concreteness estimation is evaluated against a concreteness dataset by Brysbaert et al. [2013]
- Image classification and segmentation is evaluated on the MSCOCO test split [Lin et al., 2014]

Training Data

- Model is trained on $(image, caption)$ pairs taken from the MSCOCO train split [Lin et al., 2014]
- Preprocessing includes tokenization of captions

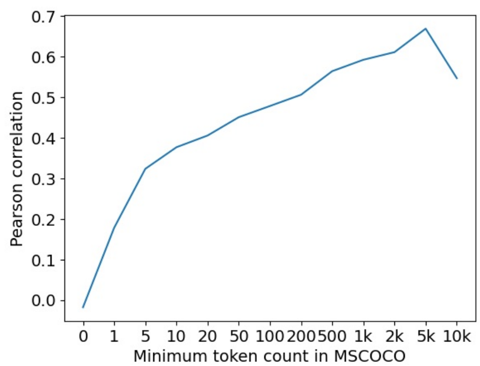
Ethical Considerations

- The model was trained on the publicly available MSCOCO [Lin et al., 2014] and may capture social biases which manifest in its training data

Quantitative Analyses

Taxonomic categorization F-Score	0.33 ± 0.0109
Syntagmatic categorization MAS	7.45 ± 0.33
Image classification F-Score	0.28 ± 0.01
Image segmentation F-Score	0.178 ± 0.01

Table 1: Model results on Taxonomic/Syntagmatic categorization, and Image classification/Segmentation.



Pearson correlation with ground-truth concreteness values, as a function of the word frequency in the training set.

---

[a]The Hebrew University of Jerusalem
[b]University of Melbourne

References

U. Berger, G. Stanovsky, O. Abend, and L. Frermann. A computational acquisition model for multimodal word categorization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, 2022.

M. Brysbaert, A. B. Warriner, and V. Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46(3):904–911, 2013. doi: 10.3758/s13428-013-0403-5.

S. De Deyne, D. J. Navarro, A. Perfors, M. Brysbaert, and G. Storms. The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, 51(3):987–1006, 2019.

T. Fountain and M. Lapata. Meaning representation in natural language categorization. In *Proceedings of the 32nd annual conference of the Cognitive Science Society*, pages 1916–1921, 2010.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. *Computer Vision – ECCV 2014 Lecture Notes in Computer Science*, page 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1323. URL https://aclanthology.org/D17-1323.