Evaluating Semantic Relations in Predicting Textual Labels for Images of Abstract and Concrete Concepts

Tarun Tater, Sabine Schulte im Walde, Diego Frassinelli
tarun.tater@ims.uni-stuttgart.de

CMCL

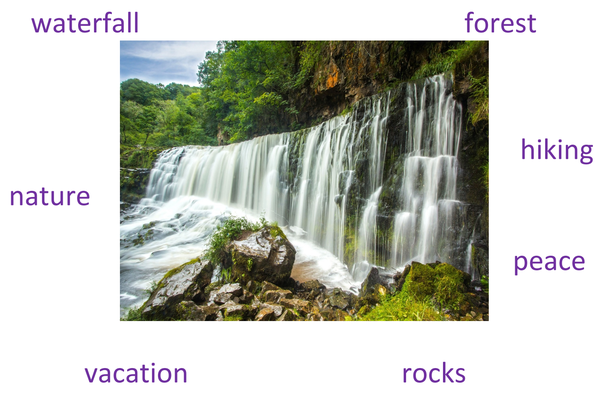
ACL 2024

Abstract vs Concrete

- Concrete concepts can be strongly experienced through human senses i.e., things that can be seen, heard, felt, smelled, or tasted as opposed to abstract concepts.

Interplay Between Concepts and Images

- An image can be associated with multiple concepts.
- **YFCC100M Datatset[1]** – User-tagged dataset of around ~100 million images. Each image has tags provided by users (user tags) when uploading the image.



RQ – How well do VLMs, specifically SigLIP, predict these user tags for abstract and concrete concepts?

- In this study, we perform multi-label classification using SigLIP.

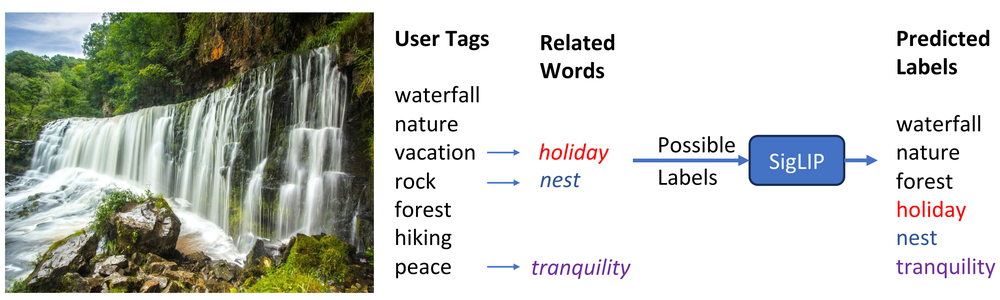
Concept class	Avg. number of user tags	Avg. number of noun user tags	Avg. % of tags pred. as labels	Avg. % images where no label was pred.	Avg. % of images where target concept not pred.
Abstract	8.69	5.40	54.41	8.06	48.87
Concrete	6.85	4.58	62.17	4.11	26.93

- What did we find?
- The model identified a higher percentage of labels for images associated with concrete concepts than for abstract concepts.
- A higher percentage of images associated with abstract concepts did not have the target concept predicted (which we associated the image with) than for concrete concepts.

So, should VLMs accept a wider selection of relevant labels for multimodal representations?

Semantic Relation Prediction

RQ - How do different semantic relations of user tags affect the prediction of image labels for abstract and concrete concepts?



Example of an image and the corresponding user tags. Here *holiday* is a synonym of *vacation*, *nest* is a co-hyponym of *rock*, and *tranquility* is a hypernym of *peace*.

- We analyzed 1,278 concepts with 300 images each, using the SigLIP Vision-Language Model (VLM) to predict semantically related words of user tags (synonyms, hypernyms, co-hyponyms).

What did the model predict?

Semantic Relation	Concept class	Avg. number of user tags with semantic relations	Avg. % of labels pred.	Avg. % images where ≥ 1 tag not pred. but their relation pred.	Avg. % user tag not pred. but their semantic relation pred.	Avg. % of images where no label was pred.
Hypernym	Abstract	80.67	27.96	76.40	40.06	3.39
	Concrete	71.48	28.66	66.94	32.02	1.39
Co-hyponym	Abstract	684.04	26.51	70.25	33.23	1.56
	Concrete	608.55	27.44	55.28	22.97	1.00
Synonym	Abstract	41.15	38.00	53.24	18.60	7.14
	Concrete	33.73	44.26	38.61	12.00	4.26

Table 2: SigLIP prediction (pred.) results when considering semantic relations of user tags as labels.

- Synonyms had the highest percentage of labels predicted, especially for concrete concepts, indicating that the model better captures meaning variations for concrete nouns.
- Abstract concepts had a higher percentage of images where at least one user tag was not predicted but a co-hyponym or hypernym was.
- Surprise – Surprise!!! - Both abstract and concrete concepts had a high percentage of labels where hypernyms or co-hyponyms were predicted but original user tags were not.

Relationship Between Association Norms & User Tags

RQ – How do user tags given a visual cue (image), and word associations given a linguistic cue, differ in characterizing abstract versus concrete concepts?

- 682 concepts (527 concrete and 155 abstract) with 300 images and 100 annotations for associations[2].
- For each image where the target concept was one of the user tags, we evaluated how well SigLIP predicts the associated words of the target concept as labels.

Class	Avg. number of unique user tags	Avg. number of unique associations	Association not in user tags	Association predicted	Association predicted for at least one image
Abstract	751	36.00	64.96%	27.89%	99.67%
Concrete	747	33.75	46.79%	38.68%	99.66%

- What did we find?
- The average number of unique user tags for abstract and concrete concepts across 300 images is similar.
- Abstract concepts have a slightly higher average number of unique associations than concrete concepts, indicating a slightly greater associative diversity for abstract concepts.
- Surprise – Surprise!!! - Despite the directly perceivable nature of concrete concepts, they evoke different personal or contextual mental associations that may not directly translate into visual depictions and vice-versa, similar to abstract concepts.

Takeaways and Future Work

- Integrating diverse semantic relationships has the potential to improve the representations in Vision-Language Models (VLMs), particularly SigLIP, for abstract and concrete concepts.
- SigLIP often predicts semantically related words such as synonyms, hypernyms, and co-hyponyms of a user tag for images associated with both abstract and concrete concepts, even when the user tag itself is not predicted as a label.
- The distinction between visual and linguistic associations shows the differences in how these concepts are perceived and described.
- Ongoing - Probing VLMs for preferences of captions containing more abstract or concrete nouns.

References

[1] Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., ... & Li, L. J. (2016). Yfcc100m: The new data in multimedia research. *Communications of the ACM*, *59*(2), 64-73.
[2] De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior research methods*, *51*, 987-1006.