Seeing What Tastes Good: Revisiting Multimodal Distributional Semantics in the Billion Parameter Era

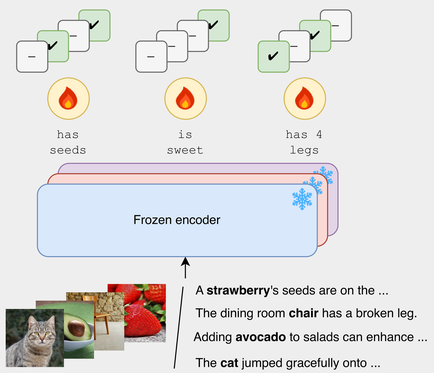
Dan Oneață — Politehnica Bucharest
Desmond Elliott — University of Copenhagen
Stella Frank — Pioneer Centre for AI

PIONEER CENTRE FOR ARTIFICIAL INTELLIGENCE

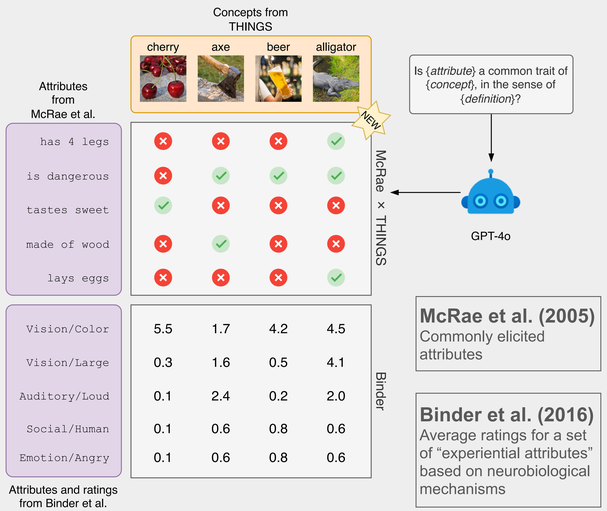
Self-supervised vision models are surprisingly good at predicting the semantic attributes of concepts

Overview

- Revisiting classic work in distributional semantics but with a new twist: **larger models** and **denser data**
- How well do large-scale pretrained models capture the semantic attributes of concepts?
  - ROSE is *red*, *smells sweet*, and *is a flower*



Datasets

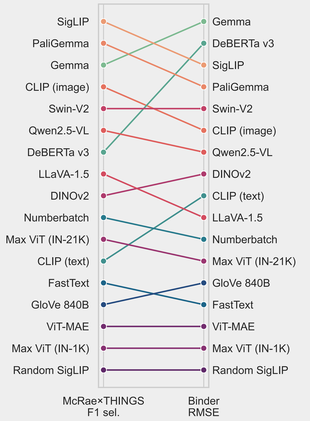


Binder *et al.*, "Toward a brain-based componential semantic representation," *Cognitive Neuropsychology*, 2016
Hebart *et al.*, "THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior," *eLife*, 2023
McRae *et al.*, "Semantic feature production norms for a large set of living and nonliving things," *Behavior Research Methods*, 2005

Results

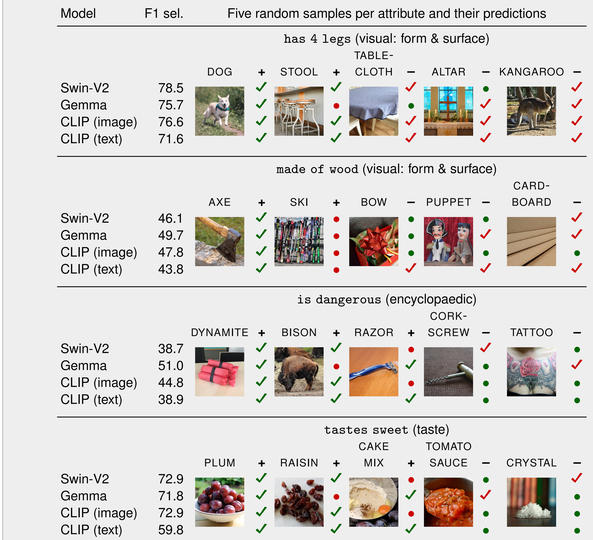
- Multimodally trained vision encoders match largest LLM
- SSL vision model is better than many text-only encoders

Ranking



Model	McRae × THINGS F_1 sel ↑	Binder RMSE ↓
Vision models		
Random SigLIP	15.4	1.43
ViT-MAE	35.6	0.94
Max ViT (IN-1K)	29.0	1.37
Max ViT (IN-21K)	43.3	0.84
DINOv2	44.5	0.80
Swin-V2	**47.0**	**0.74**
Multimodal vision models		
LLaVA-1.5	45.0	0.83
Qwen2.5-VL	46.8	0.79
CLIP (image)	48.4	0.74
PaliGemma	49.9	0.73
SigLIP	**50.1**	**0.71**
Language models		
GloVe 840B	39.1	0.89
FastText	40.2	0.91
Numberbatch	44.1	0.83
CLIP (text)	43.0	0.81
DeBERTa v3	45.5	0.68
Gemma	**49.8**	**0.67**

Example Predictions

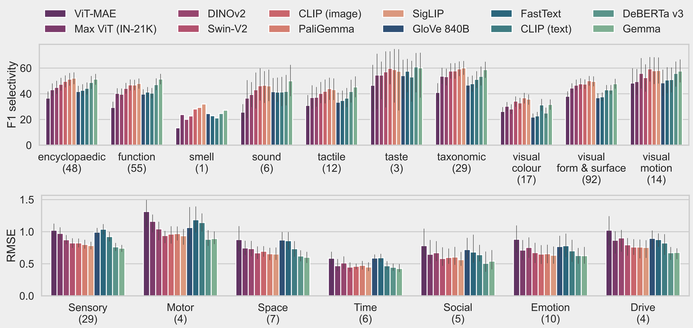
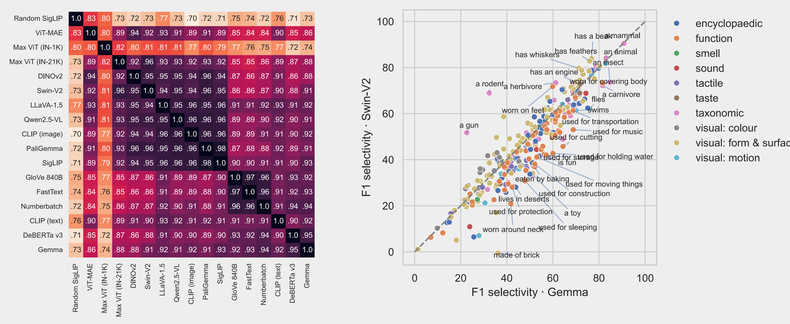


Model Details

- Vision
  - Class-supervised
  - Self-supervised
- Multimodal vision
  - Contrastive
  - Vision + LLM
- Language models
  - Word embeddings
  - Contrastive
  - Encoder-only
  - Generative LLM

Model	Params.	Dataset	Size	Objective	IN-1K
Vision models					
ViT-MAE	304M	ImageNet-1K	1.3M	MSE	85.9
Max ViT (IN-1K)†	212M	ImageNet-1K	1.3M	Classification	85.2
Max ViT (IN-21K)	212M	ImageNet-21K	14M	Classification	88.3
DINOv2	304M	LVD	142M	DINO + iBOT	86.3
Swin-V2†	197M	ImageNet-21K	14M	SimMIM	87.7
Multimodal vision models					
CLIP	304M	Private	400M	Contrastive	83.9
SigLIP	400M	Private	4B	Sigmoid Contr.	83.2
PaliGemma	400M	Private	1B	NLL	N/A
LLaVa-1.5	324M	CC3M, OKVQA, etc.	1.2M	NLL	N/A
Qwen2.5-VL	669M	Private	UNK	NLL	N/A
Language models					
FastText	–	CommonCrawl	840B	NLL	–
GLoVe	–	CommonCrawl	840B	NLL	–
Numberbatch	–	ConceptNet	N/A‡	PPMI	–
CLIP (text)	123M	Private	400M	Contrastive	–
DeBERTa v3	86M	Wiki+Books	3.1B	RTD	–
Gemma	2B	Private	6T	NLL	–

Results by Attribute Type



Cross-Model Correlations